Detection and Impact of Genome Rearrangement in *Salmonella*

Liam Tucker

Masters of Research in Biological Science (MRes)

Quadram Institute Bioscience

University of East Anglia, School of Biological Sciences

June 2021

1 Abstract

The species *Salmonella enterica* subspecies enterica is responsible for both localised gastrointestinal infections (commonly referred to as food poisoning) and severe systemic infections. Salmonella Typhi is the causative agent of Typhoid fever with an estimated 22 million cases annually, maintaining a presence in select regions of the world despite vaccination programs, with certain regions being more endemic due to either remoteness or political issues. Asymptomatic carriers are critical to the proliferation of Typhi through constant bacterial shedding into the environment. S. enterica can rearrange its genome around the seven ribosomal operons found across the genome via homologous recombination, a phenomenon originally discovered in Typhi carriage isolates. This work's hypothesis is that the ability to rearrange affects gene expression via gene dosage and growth rate via *ori-ter* balancing. In this work I establish a method to 1) induce genome rearrangement in bacterial samples within the laboratory, 2) extract high quality DNA and sequence these extracts and 3) analyse the sequence data to determine their genome arrangement. Following this method, I generated a rearrangement (LAT2, a strain derived from BRD948) through long term culturing that I characterised further using growth curve analysis and transcriptomics via RNA sequencing. I analysed the RNA data for differential gene expression which revealed many genes upregulated and downregulated in the LAT2 arrangement, in accordance with their shift in genomic position, relative to *oriC*. Key genes affected in this rearrangement were genes in the rfb and cyo clusters as well as ackA and pta, with these genes having a role in surface adhesion, survival under stress, and global signalling respectively. My work shows that sequencing provides a scalable alternative to PCR-based determination of genome rearrangement, and that rearrangement impacts upon both growth rate and specifically expression of genes that have changed location relative to the origin of replication.

Table of Contents

List of Figures

List of Tables

Glossary

ack – Acetyl kinase

AcP – Acetylphosphate

aroC – Chorismate synthase

BOB – Back-of-bench (long-term culture)

DNA – Deoxyribonucleic acid

DT2 – Definitive Type 2

GAT – Genome Arrangement Type

GS – Genome Structure

HMW – High molecular weight

LB – Luria-Bertani

LB-NaCl – Luria-Bertani, no salt

LB-NaCl+aro – Luria-Bertani, no salt, aro mix added

MDR – Multidrug-Resistant

NTS – Non-Typhoidal *Salmonella*

OD – Optical density

OLC – Overlap-Layout-Consensus assembly

oriC – Origin of replication

Ori-ter – Origin of replication to terminus

PCR – Polymerase chain reaction

pta - Phosphate acetyltransferase

RNA – Ribonucleic acid

rrn – Ribosomal RNA

SCV – Salmonella Containing Vacuole

sdh – Succinate dehydrogenase

SNP – Single-nucleotide polymorphism

SPI – Salmonella Pathogenicity Island

ter – DNA replication terminus binding-site

Typhi – *Salmonella enterica* subsp. *enterica* serovar Typhi

Typhimurium - *Salmonella enterica* subsp. *enterica* serovar Typhimurium

XDR – Extensively Drug Resistant

Acknowledgements

I would like to give my thanks to my supervisors, Gemma Langridge and Gary Rowley, their guidance and support has been critical for me in completing this thesis.

I would also like to extend my thanks to Emma Ainsworth and Cailean Carter for not only providing advice and encouragement when it was needed. I am especially thankful to Emma for being there so often when it comes to learning new techniques or answering the questions I have regarding either theory or laboratory work.

I would also like to give thanks to the Kingsley group, especially Rob Kingsley and Gaetan Thilliez for providing their collection of DT2 samples for the purpose of this paper.

My thanks also goes out to my family who have been supportive towards my endeavours and being there when I needed them. This thanks also goes to many fellow members of staff at the Quadram Institute for making this experience not only welcoming, but also enjoyable.

2 Introduction

2.1 *Salmonella*

The *Salmonella* genus was first discovered by Theobald Smith and Daniel Elmer Salmon, whom the genus is named after, after discovering the bacterium in pigs in 1886 (Jajere, 2019). *Salmonella*, a member of the Enterobacteriaceae, consists of bacteria that are rod-shaped, Gram-negative and facultative anaerobes. *S. enterica* remains a major health concern worldwide in both the developed and developing world, as a leading causative agent for gastroenteritis, as well as being responsible for other types of infection in both humans and animals.

There are two *Salmonella* species, *S. enterica* and *S. bongori*. *S. bongori* is typically found in cold-blooded animals such as reptiles but rarely causes disease in humans. *S. bongori* is distinguished from *S. enterica* through unique genes present in *S. enterica* such as the virulence genes encoding type III secretion systems (Fookes *et al.*, 2011, Helaine *et al.*, 2010). Within *S. enterica* are several subspecies such as *enterica* (*S. enterica* subsp. *enterica*) and *arizonae* (*S. enterica* subsp. *arizonae*) and within said subspecies are serovars such as *S. enterica* subspecies *enterica* serovar Typhi (hereafter Typhi).

Salmonella strains are classified using the White-Kauffmann-Le Minor classification system on top of phylogeny-based subspecies classification (Grimont and Weill, 2007). This system is based on three major antigenic determinants: capsular (K), somatic (O) and flagellar (H). The K antigens are heat-sensitive polysaccharides located on the capsular surface, these antigens are rarely seen among most serovars, with the most common K antigen being the virulence (Vi) antigen, which is found on serovars Typhi, Paratyphi C and Dublin. The heat-stable O antigen makes up the oligosaccharide component of the lipopolysaccharide of the bacteria and is located on the outer cell membrane, with more than one O antigen potentially being expressed by a serovar. The H antigen is a heat-labile antigen involved in host immune response stimulation and is found in the flagella. Most serovars possess genes for two H antigens though typically *Salmonella* only express one at a time (Silverman *et al.*, 1979).

Within the developed world, *Salmonella* infection is typically associated with gastroenteritis, being one of the most common pathogens responsible for foodborne disease (Eng *et al.*, 2015). In the developing world Typhi, responsible for causing typhoid fever, continues to

remain an issue particularly in regions with hygiene and sanitation challenges alongside restricted access to healthcare (Gasem *et al.*, 2001, Antillón *et al.*, 2017). Enteric fever describes typhoid alongside paratyphoid fevers caused by Paratyphi serovars A, B and C which show similar symptoms to typhoid fever. Despite efforts for improved sanitation and vaccine distribution in the case of Typhi, enteric fever remains prevalent worldwide particularly in low to middle income countries (Mogasale *et al.*, 2014, Stanaway *et al.*, 2019).

2.1.1 *Salmonella* Typhimurium

Salmonella enterica subspecies *enterica* serovar Typhimurium is a very diverse serovar featuring a large number of strains (over 300), classified as Definitive Types (DTs) based on their susceptibility to various bacteriophages (Rabsch, 2007). The DTs range from host-generalist pathogens (Kozak *et al.*, 2013) such as DT49 which does not have a definitive reservoir that can be observed, while other strains show adaptation to one specific host reservoir (host-specific) such as DT2 (found in pigeons) (Kingsley *et al.*, 2013), or DT104 in pigs, poultry and other wild animals (Branchu *et al.*, 2018).

For the purposes of identifying and classifying Typhimurium strains, typing is carried out via susceptibility testing to a variety of phages of differing specificity. For public health, phage typing can provide a strong correlation to epidemic source which is useful in tracing outbreak origins (Fisher, 2004).

Typhimurium infection in humans is typically responsible for causing salmonellosis, characterised by intestinal inflammation, fever and diarrhoea 6-12 hours after infection (Crump *et al.*, 2008), and is one of the most common causes of food-borne illness (Gart *et al.*, 2016, Fabrega and Vila, 2013). In mice, however, it is used as a model of Typhi infection, as Typhimurium develops into a similar systemic infection in mice compared to Typhi infection in humans (Sabbagh *et al.*, 2010, Monack *et al.*, 2004). However, caution should be given with the use of mouse models in this regard as the genetic content of Typhimurium does not align very closely to Typhi, so this extrapolation should be taken with care.

2.1.2 *Salmonella* Typhi

Salmonella enterica subspecies *enterica* serovar Typhi (Typhi) is a host-restricted pathogen that causes infection exclusively in humans (Spanò, 2016). It is the causative agent of typhoid

fever, characterised by fever and diarrhoea. If left untreated, intestinal perforation may occur which can be fatal if the bleeding caused is clinically significant (Contini, 2017).

Typhi is typically found in remote regions with poor sanitation (due to faecal-oral route transmission) and difficult to access healthcare. Such regions are present in countries including India, Pakistan and sections of China with a global burden estimated to be 22 million new cases annually with 210,000 estimated deaths caused by typhoid fever, with both Sub-Saharan Africa and South Asia contributing half of these numbers (Buckle *et al.*, 2012). There has also been increasing attention on the incidence of Typhi within urban centres in Asia and Africa likely caused by dense populations and the presence of urban slums (Crump *et al.*, 2004, Breiman *et al.*, 2012, WHO, 2008).

Finding new methods of diagnosis and treatment for Typhi is particularly important with the emergence of antimicrobial resistance (Olarte and Galindo, 1973, Wain and Kidgell, 2004), including resistance to quinolones, fluoroquinolones and more recently ceftriaxone (Rasheed *et al.*, 2020, Djeghout *et al.*, 2018, Klemm *et al.*, 2018). The rise of particularly virulent MDR (multidrug-resistant) and XDR (Extensively Drug Resistant) clades such as H58 has made it a priority to find new avenues of detection and treatment (Wong *et al.*, 2015).

2.2 Pathogenesis

Many genes responsible for virulence in *Salmonella* are located on *Salmonella* Pathogenicity Islands (SPIs); each SPI contains a group of genes that work together to contribute to a step in the infection process (Hansen-Wester and Hensel, 2001). *Salmonella* pathogenesis can be either non-invasive (mild gastrointestinal illness) or invasive (not self-limiting and leads to serious infections). Typhoidal *Salmonella* is usually associated with invasive salmonellosis and non-typhoidal *Salmonella* (NTS) is associated with non-systemic enteric salmonellosis, though the opposite can occur in both cases (non-invasive typhoidal *Salmonella* infection and invasive non-typhoidal *Salmonella* infection).

The SPIs appear to be obtained by horizontal gene transfer events (Ilyas *et al.*, 2017), as a related and phylogenetically older member, *S. bongori*, was shown via DNA hybridization analysis and genome sequencing to contain SPI-1 genes but not SPI-2 (Fookes *et al.*, 2011, Ochman and Groisman, 1996). As *S. bongori* is rarely associated with human disease, it's believed that SPI-2 is responsible for infection of human hosts. With these observations, it is

considered that the initially acquired SPI-1 enables localized gastrointestinal infection, while SPI-2 enables systemic infection by replication within macrophages (Hensel *et al.*, 1998, Figueira and Holden, 2012, Que *et al.*, 2013).

2.2.1 Gastrointestinal Pathogenesis

Non-typhoidal *Salmonella* infection typically takes place upon consumption of water or food contaminated with the bacteria, from here the bacteria survive the hostile environment of the stomach and reach the gut. The *Salmonella* then move towards the Peyer's patches located throughout the gut epithelium and attach to the M cells within these patches. Using a Type III Secretion System (T3SS) encoded by SPI-1, the bacteria injects effector proteins into the attached cell (Coburn *et al.*, 2007b), causing actin cytoskeleton remodelling and consequently membrane ruffling resembling phagocytosis that leads to the host cell engulfing the bacteria (Takaya *et al.*, 2003).

Upon engulfment, *Salmonella* activates genes present in SPI-2 to produce another T3SS and its effectors, this time to remodel the vesicle (known as a *Salmonella* containing vacuole (SCV)) in such a way that lysosomes cannot combine with the vesicle, which would otherwise cause the death of the bacteria inside (Steele-Mortimer, 2008). The *Salmonella* bacteria continue to grow within the SCV, surviving the generally nutrient-poor environment through genes such as *aroC,* which allows for the synthesis of required aromatic compounds (Dougan *et al.*, 1988, Herrmann and Weaver, 1999).

2.2.2 Invasive Pathogenesis

Invasive *Salmonella* take a different route to infection after getting engulfed by the host cell compared to typical non-typhoidal *Salmonella* (NTS) gastrointestinal infections. Upon uptake by M cells, non-invasive *Salmonella* cause significant recruitment of immune cells in the localized region of infection as it also invades them, then escape back to the lumen via shedding and replicate resulting in localized inflammation. In comparison, invasive typhoidal *Salmonella* has limited luminal replication, instead being taken up by macrophages and engaging in intramacrophage replication before disseminating through the reticuloendothelial system (Coburn *et al.*, 2007a), eventually residing in organs such as the liver and gall bladder.

As previously mentioned, while invasiveness is typical for typhoidal *Salmonella*, it can also be seen in NTS infections too (designated as iNTS), which is emerging as a disease of concern in sub-Saharan Africa (Gordon, 2011) while remaining rare in industrialised populations. The occurrence of iNTS in sub-Saharan Africa is linked to diseases in the region such as malaria and HIV (Feasey *et al.*, 2012).

2.3 Chronic carriage

What makes typhoid fever difficult to effectively treat for a population is the ability of Typhi to establish chronic, asymptomatic carriage (Levine *et al.*, 1982) in 3-5% of those that experienced acute infection (Bhan, Bahl and Bhatnagar, 2005, Senthilkumar *et al.*, 2014). These carriers may constantly shed bacteria, particularly in faecal matter, which can cause others to get infected if this faecal matter contaminates food or drinking water.

Carriage is established primarily through colonisation of the gall bladder (Gonzalez-Escobedo, Marshall and Gunn, 2011) which can lead to long-term regular shedding of infected gall bladder cells. Certain conditions such as cholelithiasis (gall stones) serve as risk factors for the development of carriage, as Typhi can establish biofilms on the surface of the gall stones (Crawford *et al.*, 2010).

Carriage remains one of the major difficulties in eradication of typhoid fever, as lack of symptoms make detection difficult, furthermore, treatment and vaccination options for acute disease previously shown to be occasionally effective at treating carriage (Phillips, 1971, Nolan and White, 1978, Brodie *et al.*, 1970, Zavala Trujillo *et al.*, 1991) are no longer recommended especially due to concerns of MDR strains. This may be due to either resistant bacteria being more strongly selected for carriage, or the biofilm formed providing protection (Costerton *et al.*, 1999).

As a result, the main treatment option for treating carriage is the removal of the gall bladder (Main, 1961, Gonzalez-Escobedo *et al.*, 2011), which is both highly invasive and not guaranteed to be effective, as carriage can also be established in the liver and the ducts of the gall bladder (Nath *et al.*, 2011).

Understanding which factors are involved in the establishment of carriage is therefore an important factor in improving disease control and one avenue of research is genome rearrangement, which has been identified in Typhi carriage isolates.

2.4 Genome Rearrangement

Genome rearrangement involves the shuffling of the bacterial genome around long repeat sequences, such as ribosomal *(rrn)* operons (Helm *et al.*, 2003). This rearrangement is achieved through homologous recombination (Darmon and Leach, 2014) between two of these repeat sequences, resulting in the fragment in-between being inverted, or excising and reintegrating at a different position in the genome as a transposable element (seen in Figure 1).

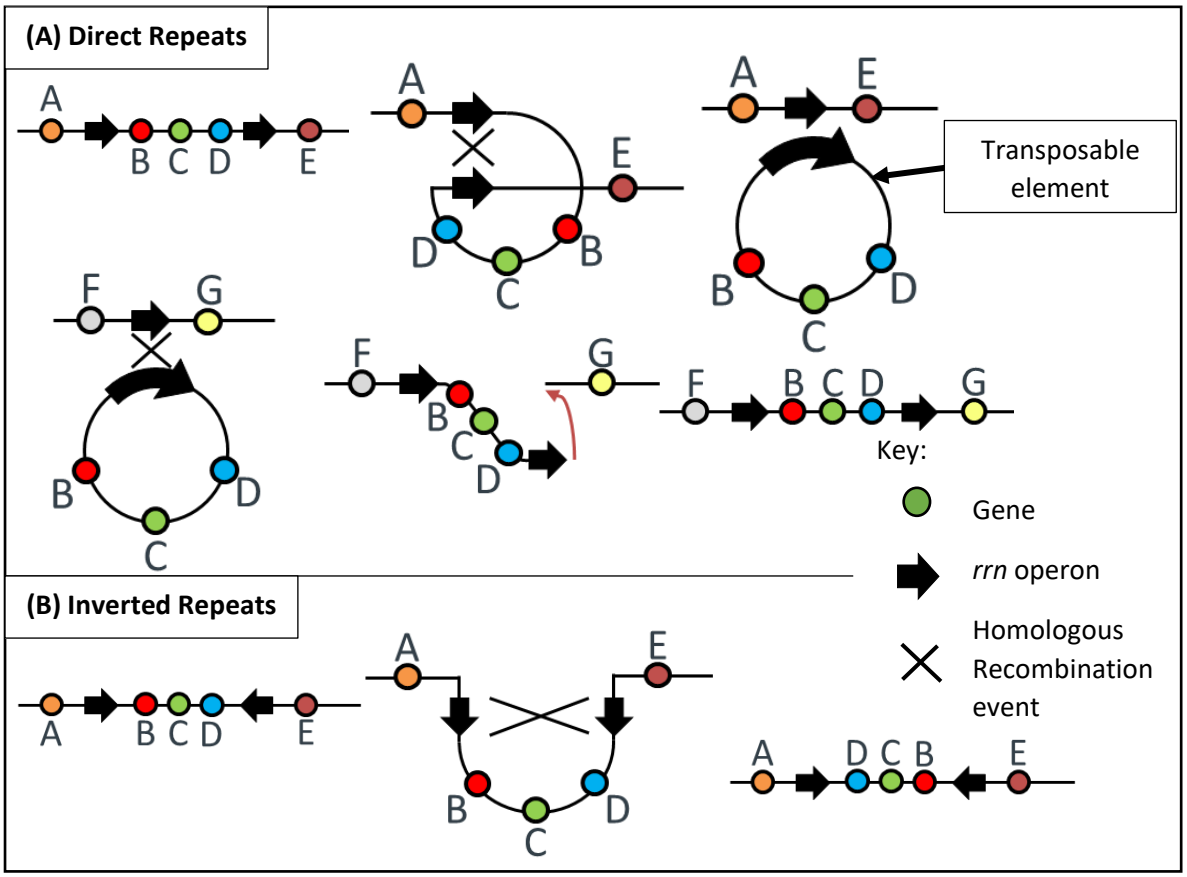


Figure 1 | Homologous recombination events that affect genome arrangement

Rearrangement can occur via either direct (A) or indirect (B) repeats. Coloured circles represent genes, arrows indicate *rrn* operons and their direction, crosses represent homologous recombination events.

The arrangement of the genome is based around the direction of replication from the origin (*oriC*) to the terminus (*ter*), with each fragment having to follow this direction for the arrangement to be valid (5' to 3' starting from *oriC* to *ter*). When *rrn* operons are direct repeats of each other (Figure 1A), where the flanking operons follow the same direction of replication, the fragment between is excised from the genome forming a transposable element that can reintegrate elsewhere in the genome. When *rrn* operons are inverted repeats (Figure 1B), where the operons are in opposing directions, the fragment flanked by the operons can invert, for example fragments 1 and 3 in the case of *Salmonella*.

As demonstrated in Figure 2, fragments being excised from the genome can reintegrate anywhere between other fragments. If the fragment reintegrates along the same side it originally was (relative to *oriC* and *ter*) it can reintegrate without further issue as *rrn* operon directionality is maintained (Figure 2A and B). If the fragment reintegrates along the opposite side relative to *oriC* and *ter* it is unable to do so in its original orientation as this does not follow the direction of replication on that side (Figure 2C). Instead, the fragment must invert (marked as 4' in this case), allowing it to successfully reintegrate onto the opposing side (Figure 2C). Indirect repeat fragments maintain the direction of their *rrn* operons regardless of orientation.

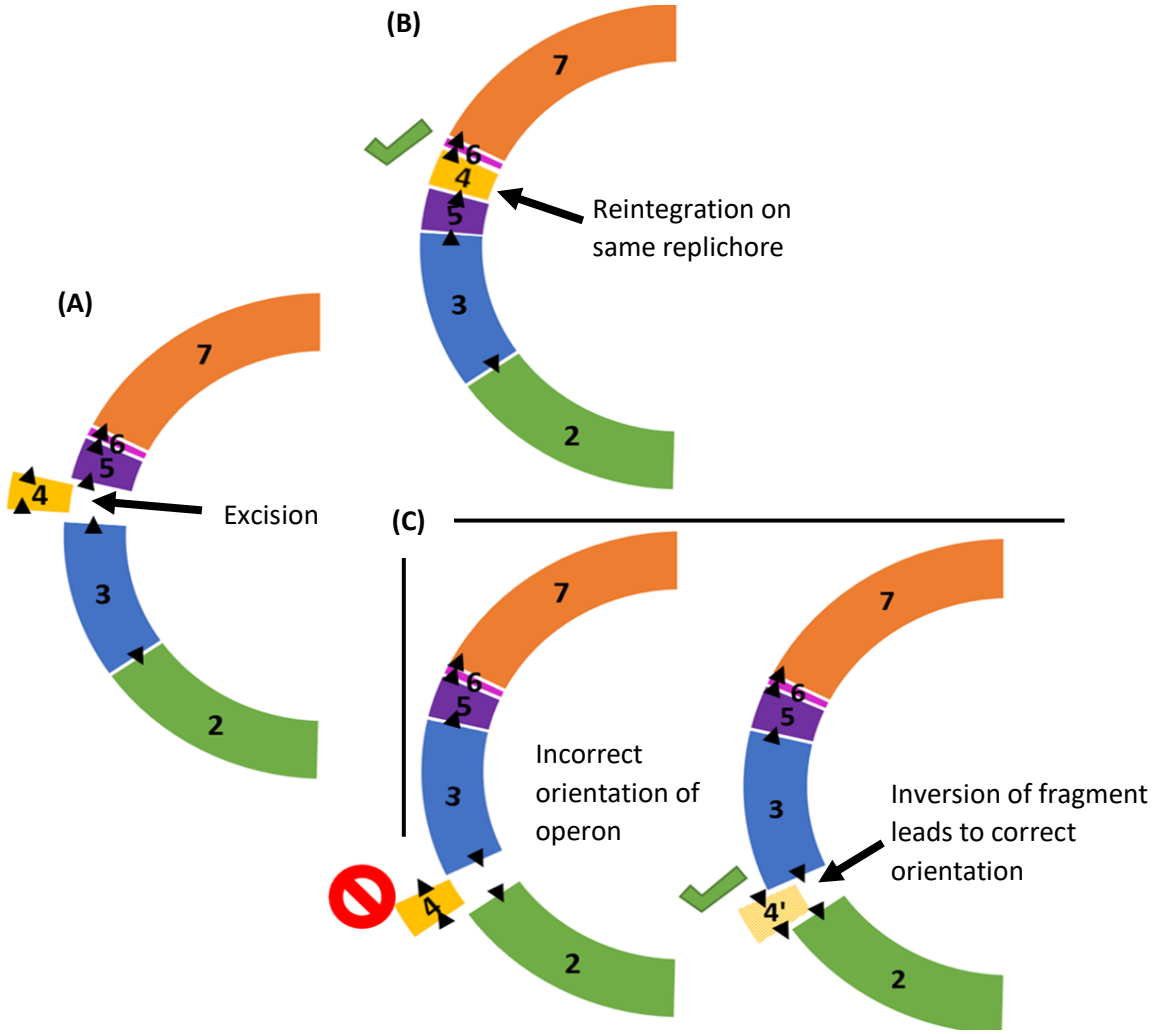


Figure 2 | *rrn* Operon Directionality and Role in Inversions

Direct repeat fragments can reintegrate back into the genome anywhere between other fragments. If a fragment tries to reintegrate onto the opposing side of the genome relative to *oriC* and *ter* it must invert first to maintain correct directionality on that side of the genome. (A) Represents the excision of a fragment out of the genome, with (B) representing the reintegration of this fragment into the same replichore while (C) represents the movement, inversion and integration of this fragment into the opposing replichore. Each coloured and numbered block represents a fragment, with the arrows on the ends of these representing *rrn* operon directionality.

As these sequences are long repeat sequences (approximately 5 kb in length), with short read sequencing it becomes impossible to discern between operons, making any genomes generated fragmented and thus unusable for determining genome arrangements. This is because short read sequences, while highly accurate, can only achieve a maximum length of approximately 500 bp in optimal conditions, which is not sufficient to bridge across the ~5 kb long *rrn* operons. Reads long enough to bridge across the *rrn* operons into the distinct regions that flank them as seen in Figure 3 are needed to piece together fragments into a single contig assembly.

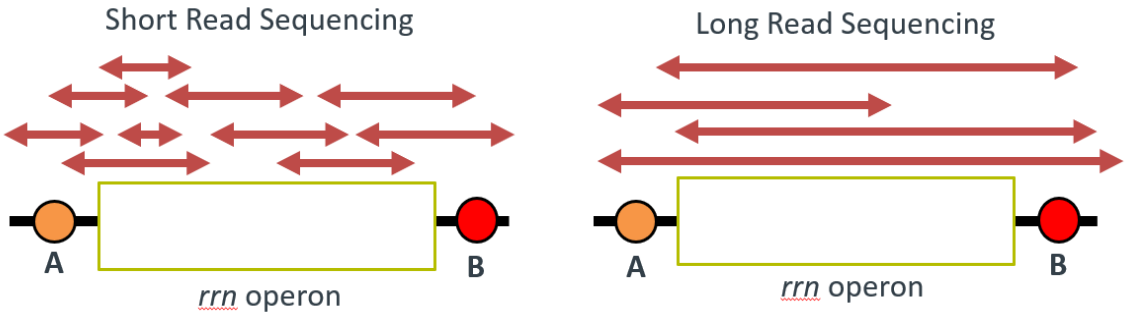


Figure 3 | Coverage of the rrn Operons in Sequencing

A schematic comparison of short and long read sequencing coverage against the *rrn* operon. The short-read sequencing (~500 bp) does not produce reads long enough to span across the operon. Reads produced by long-read sequencing (generally >5 kb) bridge to the flanking regions (indicated by A and B) that can be used to assemble the order and orientation of fragments.

2.4.1 Indirect determination of genome arrangement

One method of determining genome arrangement is with long range PCR which operates in the same way as regular PCR but with introducing modifications (such as deletions improving accuracy)(Barnes, 1992) to the polymerases used to significantly increase the size of the amplicons generated (Jia *et al.*, 2014). This can be used to determine genome rearrangements by designing primers to the genomic regions flanking each of the *rrn* operons in the given genome (in the case of *Salmonella*, requiring 91 PCRs to include all possible fragment pairings) (Haase, 2008) and then by conducting gel electrophoresis on these amplicons and interpreting the bands on the gel (Helm and Maloy, 2001). While this works, it has key drawbacks in being extremely time consuming and resource intensive. Long-range PCR also cannot be effectively scaled as each set of reactions can only be for one sample at a time.

2.4.2 Direct determination of genome arrangement

In the past, it was difficult to assemble complete genome sequences due to the *rrn* operon sequences between fragments. Recently, long read sequencing methods such as PacBio and Nanopore have enabled more routine generation of complete genome sequences (single

contig chromosomal assemblies that can be fully circularized). The significant advantages of these method are that reads for these sequencers can easily reach above 5 kbp, which is the length required to read across the *rrn* operons and into the chromosomal regions flanking them. Furthermore, long read sequencing requires less setup compared to long range PCR and is much more automated, meaning it should take significantly less time to discern genome arrangements than in the past. Furthermore, long read sequencing is also scalable, and since this uses genome sequences instead of bands produced via gel electrophoresis, there is also less risk of subjectivity. The application of this technology for determining genome arrangements is investigated in this thesis.

2.4.3 Impact of genome rearrangement

Many species of bacteria are capable of rearranging their genome (Belda *et al.*, 2005, Chen *et al.*, 2017, Liu *et al.*, 2013, Tsuru *et al.*, 2006). However, most tend to display a conserved arrangement such as GS (Genome Structure) 1.0 (1 to 7, no inversions) in the case of *S. enterica* (Figure 4A). Typhi is an exception, as it displays GS 2.67 (1 7’ 3 5 6 4 2’) as its main type (Figure 4B) (Page *et al.*, 2020).

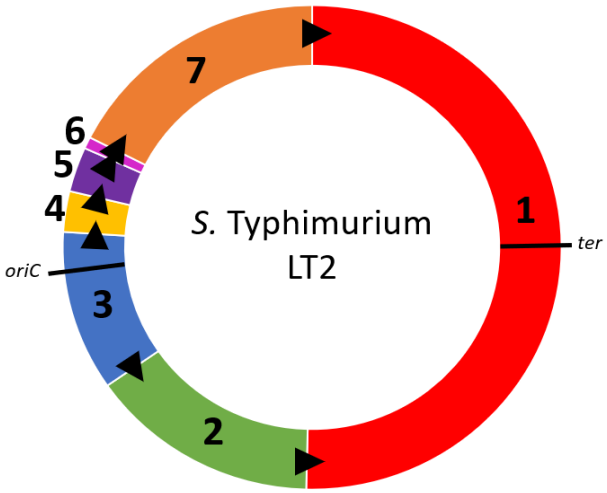


Figure 4 | Example Genome Structures in *Salmonella*

Salmonella genomes are divided into seven numbered fragments (coloured blocks) by the repeat ribosomal operons (white lines). Genome structures (GSs) are shown of (A) Typhimurium LT2 (GS 1.0) and (B) Typhi BRD948 (GS 2.67), with the origin (oriC) and terminus (ter) locations indicated on fragments 3 and 1 respectively. Fragments 2 and 7 have translocated and inverted in Typhi BRD948, with inverted orientation denoted by apostrophes and hashed colour. Black arrows indicate the direction of replication from oriC to ter.

Rearranging the genome has an effect on various aspects of the genome from overall structure to gene expression, relative to placement around the genome (Bryant *et al.*, 2014). The rearrangement of the genome influences the relative positioning of *oriC* and *ter* from

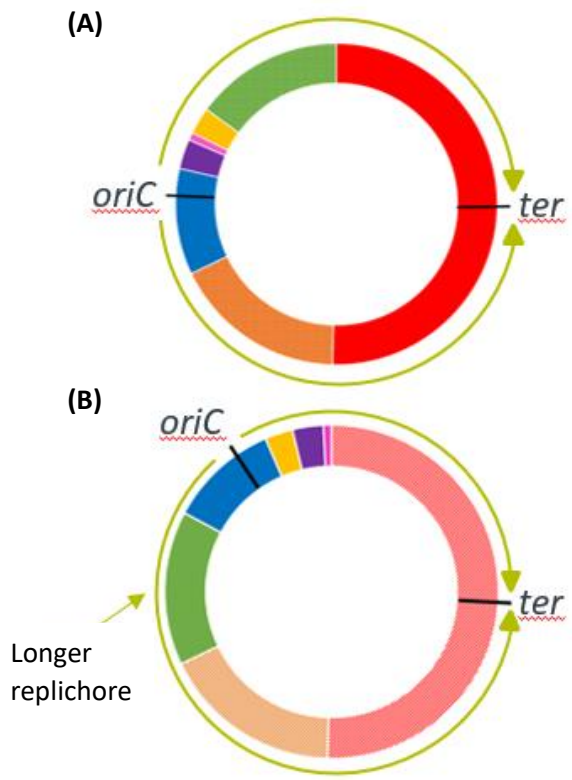


Figure 5 | Movement of *oriC* and *ter* Positioning Through the Genome and the Effect on Replication.

In (A), the genome is arranged in a manner where *oriC* and *ter* are approximately opposite to each other, generating little lag in replication. In (B) the genome has rearranged in a way that generates significant bias in the *oriC*-ter balance, generating one replichore much longer than the other. The replichores in both genomes are indicated with green arrows.

each other, which can generate a bias with one replichore longer than the other (Figure 5). As bacterial replication starts at *oriC* and moves across both replichores to *ter* (Moriya *et al.*, 2009) this bias affects the bacteria's overall growth rate as it takes longer to replicate the longer replichore (Mackiewicz *et al.*, 2004), resulting in a hampered growth rate.

Another effect is likely to be gene dosage, as genes are typically more actively expressed the closer they are to *oriC* (Sousa et al., 1997, Gerganova et al., 2015). This would modulate gene expression as genes closer to *oriC* can have multiple replicates compared to genes further away, increasing availability as there are more templates of these genes to be expressed compared to genes distant from *oriC*. This previous work has focused on the movement of individual genes across the genome to assess gene dosage whereas changes on a genome-wide scale have not been done previously.

Furthermore, this research is also based on previous studies that found that transcription rates of genes in bacterial genomes were affected relative to their distance from *oriC* (Petkov *et al.*, 2005, Soler-Bistué *et al.*, 2017).

2.5 Genome sequencing and assembly

The majority of sequencing used in this work was on the MinION platform from Oxford Nanopore Technologies (ONT), which detects changes in electrical current caused by the passing of DNA through biological pores that pass through a membrane. This technique works by first attaching the DNA strand to a leader adaptor and motor protein, with the leader

adaptor guiding the complex to a pore, and the motor protein unzipping the double stranded DNA and passing it through the pore, causing a change in current as the base passes through (Leggett and Clark, 2017). One of the key advantages of this method is that the sequencer is significantly smaller than other sequencers available, meaning that this device can also be used in the field with greater ease of deployment than others. This is particularly useful for Typhi research as the bacterium is typically endemic in remote regions, thus making it difficult to provide larger equipment for these areas.

Alongside this, short read sequencing on the Illumina platform is also utilized as the accuracy of this method of sequencing allows for the detection of events such as SNPs (single nucleotide polymorphisms), which are useful for tracking phylogenetic relationships.

For sequence analysis, there is also a need to utilise efficient algorithms by which the data generated by sequencing can be taken and assembled both rapidly and accurately. There are various methods by which to piece together individual sequence reads into larger contiguous sequences, all of which relate back to graph theory and are described below. For assemblers, either greedy algorithms or graph method assembly (utilized by more modern assemblers such as Canu) are used for generating contigs from reads.

2.5.1 Overlap-Layout-Consensus assembly

An algorithm used early on in genome assembly was “overlap-layout-consensus” (OLC) which is considered a “greedy algorithm”, with the assembly working in three steps: reads are compared among each other to search for overlaps (O), the assembler then carries out a layout (L) of said reads and their overlaps into a graph which leads to the consensus (C) sequence being found (Li *et al.*, 2011). OLC is considered to be an intuitive algorithm, as all possible pieces of the sequence are tried among one another (all-against-all) and pieces that match are put together (Pevzner, Tang and Waterman, 2001). This method of assembly is typically seen in early de novo sequence assemblers such as CAP (Contig Assembly Program) (Huang, 1992).

OLC however has several issues, the first being that the layout step is a Hamiltonian path problem wherein each vertex must be visited once. Graphically, reads are represented as vertices on a graph and the alignment between reads is represented as edges. This presents issues in particular with genomes with repeats, as repeat regions will have a large amount of

overlap with other reads, making it computationally intensive to distinguish regions with repeats from each other. Scalability is also an issue as the increase in genome size also means the number of comparisons exponentially increases, as well as comparisons that should match, but are not actually correct matches if compared back to what the assembly should look like.

2.5.2 de Bruijn graph assembly

Another method of assembly is through de Bruijn graphs (DBG), where instead of visiting each vertex once, it conducts a Eulerian cycle where it visits each edge once (Compeau, Pevzner and Tesler, 2011), with edges being *k*-mers and vertices are (*k*-1)-mers (Figure 6).

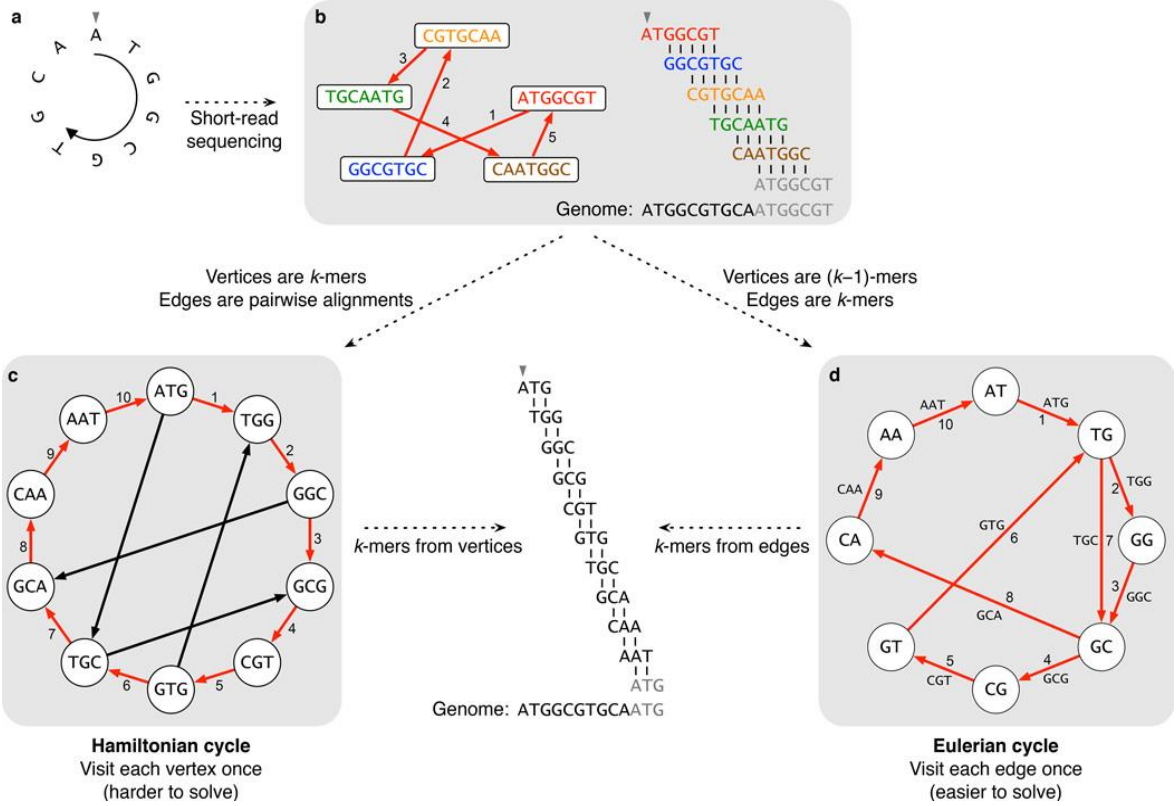


Figure 6 | The Differences in Assembly Approach Between Using Hamiltonian Cycles and Eulerian Cycles

(A) A simplified version of a small circular genome. (B) Traditional Sanger sequencing algorithms present reads as nodes and edges as alignments between the reads. Using a Hamiltonian cycle we can reconstruct the genome by combining alignments, at the end the sequence wraps around to the start of the genome. (C) An alternative technique that splits reads into *k*-mers that follows a Hamiltonian cycle by forming an alignment in which each successive *k*-mer is shifted by one position (OLC). (D) With a de Bruijn graph the edges are represented with *k*-mers representative of the nodes and the Eulerian cycle allows a reconstruction by forming an alignment where each *k*-mer is shifted by one position.
(Source: Compeau, Pevzner and Tesler, 2011)

This avoids the issue of *NP*-Completeness that graph building with Hamiltonian cycles encounters (Skiena, 2008). This means that while solvable, using a Hamiltonian cycle other than for rough approximations would take an unreasonable amount of time to complete.

Despite this, this method of assembly encounters an issue of tolerance for errors which is similarly shared with OLC assembly.

2.5.3 Repeat graph assembly

With this method, graphs are built using approximate sequence matches (unlike de Bruijn graphs, which require exact *k*-mer matches) allowing for greater tolerance of noise in reads. Nodes in repeat graphs represent junctions while edges represent genomic sequences that can be classified as either unique or repetitive (Kolmogorov *et al.*, 2019).

2.6 Assembler choice

Due to the recent development of long-read sequencing, there are a variety of assemblers available for long read assemblies such as Flye (repeat graph assembly), Canu (correct-trim-assembly, derived from OLC assembly) and Raven (OLC assembly). Despite using similar principles for assembling genomes, there are notable differences between each one. These differences may lead to differences in assembly outcome and the GS determined for samples, which means there is a need to investigate which assemblers provide the most accurate assemblies as well as other parameters e.g. which have the fastest runtimes.

Assembly accuracy and the ability to form complete contigs is of particular importance in this project, as these factors are critical in accurately determining GS. Poor genome accuracy can result in mis-assigned fragments which will result in samples being assigned incorrect GSs. The ability to generate single contig assemblies is also important as there can be variation in coverage between sequencing runs, so the ability to form complete genomes even at low coverage is important.

There is previous research on the performance differences between different long read assemblers (Wick and Holt, 2020), focusing primarily on chromosome completion, sequence identity, contig circularisation, and computational resource use. In this work I will also be focusing on differences in performance, focused on those relevant to producing accurate determinations of bacterial genomic arrangements.

2.6.1 Automatic assignment of GS

Using sequencing data provided from long read sequencing, the GS of bacteria can be identified using assembled contigs, producing objective results.

Socru (Page *et al.*, 2020), a software tool developed at QIB, takes fragments provided from a complete genome assembly and compares these to a database of *dnaA* sequences via blastn to identify the fragment containing *oriC*, and to a database of *dif* sequences via blastn to identify the fragment with *ter*. It then identifies ribosomal repeat sequences to identify fragments based on known lengths in order to assign a GS to the sample sequence. Generating single complete contigs in this case is extremely important due to *Socru* being very sensitive to incomplete genomes, generally failing to assign a GS if there are 3 or more contigs present even with all *rrn* operons present.

2.7 Gene Expression

DNA sequencing allows us to observe any changes in genome arrangement, however it does not give insight into any impact these rearrangements have on the cell, such as gene expression.

Previously, gene expression was determined with microarray studies which has disadvantages such as poor quantification of extremes in expression (very low and high expression genes) as well as the presence of cross-hybridisation artifacts which affect the resolution of experiments that make use of this method (Marioni *et al.*, 2008).

Current methodology, RNA sequencing (shortened to RNA-seq), can reveal the effect that genome rearrangement has on the cellular transcriptome by mapping the gene location the RNA is expressed from and quantity of RNA present at a given point within a colony's growth cycle (Kukurba and Montgomery, 2015, Wang *et al.*, 2009). RNA-seq also does not require the prior knowledge of genes and their encoding sequence beforehand to gather data on them (unlike microarrays), giving a far more complete picture to the transcriptome of the bacteria in question (Zhao *et al.*, 2014).

The process of RNA-seq includes the extraction of total RNA from cells, removal of the rRNA (ribosomal RNA) to leave mRNA (messenger RNA, via ribosome depletion) and then using reverse transcriptase to produce cDNA that is put forward for next generation sequencing. In this project, this will allow quantification of gene expression and to compare this to fragment movement between genome arrangements.

2.8 Objectives

Previous studies using long-range PCR have shown *Salmonella enterica* capable of rearranging its genome (Helm and Maloy, 2001), and revealing the existence of multiple arrangements (Matthews *et al.*, 2011). More recent research using long read sequencing has confirmed this observation and has also noted that the different members of the *Salmonella* genus often have different common genome arrangements to each other (Tucker, Ainsworth and Langridge, unpublished). These different arrangements may suggest a potential role in the evolution of each of these serovars and their adaptations.

In this Masters project, my aim was to find ways to optimise DNA extraction as well as improve the setup process towards long-read sequencing for detection of genome arrangements. I also investigated potential effects caused by rearrangement by looking at growth rate and analysing RNA to determine any effect upon gene expression.

My project objectives were:

- To induce genome rearrangement in the laboratory
- To optimize laboratory methods to improve quality and quantity of DNA extraction for long read sequencing
- To scale up methods for long-read sequencing
- To determine which assemblers are best suited to identify genome arrangements
- To investigate the impact of genome rearrangement on gene expression

3 Methods

3.1 Strains used in the work

The strains used in this work consisted of the following, Typhi BRD948 (Tacket *et al.*, 1997) (a vaccine strain that requires aro mix to grow) and genome rearrangements of BRD948 vaccine strain (initial GS: GS 2.67), as well as Typhimurium strains DT2 and SL1344 (GS 1.0) (also seen in Table 1).

Strain	Details	Reference
Typhi BRD948	Ty2 strain that has deletions in the *aroC*, *aroD* and *htrA* genes.	Original: (Tacket *et al.*, 1997) Generation of GS: (Haase, 2008)
Typhimurium SL1344		Complete Genome: (Kröger *et al.*, 2012)
Typhimurium DT2	30 DT2 strains total, inversions previously detected in Helm *et al*.	Discovery of GS: (Helm *et al.*, 2004)

Table 1 | The strains used in this work.

3.2 General Reagent Preparations

For most of this work LB-NaCl (Luria-Bertani, no salt) media was used as the inducements of rearrangements seen in (Haase, 2008) were discovered in samples grown in this media. For the long-term cultures, samples were grown in either LB-NaCl (Oxoid), MOPS EZ Rich (a defined medium, Teknova) or iso-sensitest broth (a semi-defined medium used generally for antimicrobial susceptibility testing, Oxoid).

Aro mix is a mixture of aromatic compounds required for Typhi strain BRD948 WT and variants derived from it. These compounds are ones that BRD948 is not capable of producing by itself due *aroC*, *aroD* and *htrA* deletions which render this strain non-infectious. These deletions are found at 241,977-242,636bp, 554,192-554,844bp, and 1,298,947-1,299,416 bp in the genome.

Aro mix was prepared as two 100X stock solutions (Table 2); Tyrosine was prepared separately due to its low solubility in water (0.45 mg mL^{-1} at 25 °C, pH 7.0). Stock solutions of aro mix were prepared by dissolving the appropriate mass of aromatic compounds in water as shown in Table 2.

Solution	Compound	100X Stock Solution	Final Working concentration
A	2,3-dihydroxybenzoic acid	1 mg mL^{-1}	10 µg mL^{-1}
A	4-aminobenzoic acid	1 mg mL^{-1}	10 µg mL^{-1}
A	L-phenylalanine	4 mg mL^{-1}	40 µg mL^{-1}
A	L-tryptophan	4 mg mL^{-1}	40 mg mL^{-1}
B	L-tyrosine	4 mg mL^{-1}	40 mg mL^{-1}

Table 2 | The composition of the two components of aro mix.

MOPS EZ Rich defined medium (Neidhardt *et al.*, 1974) (hereafter EZ-rich) was prepared with the following: 100 mL 10X MOPS Buffer, 10 mL 0.132 M K_2HPO_4, 100 mL 10X ACGU Solution, 200 mL of 5X Supplement EZ, 10 mL 20% Glucose and 580 mL water. The whole media was filter sterilised after being made.

3.3 Bacterial growth and long-term growth cultures setup

Single colonies of Typhimurium or Typhi BRD948 were prepared by inoculating Luria-Bertani (LB) or LB-NaCl+aro (Luria-Bertani, no salt, aro mix) agar plates, respectively, from frozen glycerol stocks and incubated overnight at 37 °C. Single colonies from these plates were used to inoculate fresh media as required.

The long-term growth cultures were prepared in three different types of broth: LB-NaCl, EZ-rich or Iso-sensitest broth with or without aro mix for Typhi or Typhimurium, respectively. These cultures were started from colonies picked from overnight plates of the respective sample and inoculated into 25 mL of each type of broth. Separate 25 mL cultures were incubated at 37 °C and room temperature for months (and remain ongoing within the same media). These samples were processed monthly for sequencing to identify any genome rearrangements.

3.4 Growth curve generation

Growth curve analysis for each strain from each medium was performed in triplicate using a Bioscreen C (Oy Growth Curves Ab) in LB-NaCl to automatically record optical density (OD) measurements. In brief, overnight broths were inoculated with single colonies from respective agar plates in LB-NaCl+aro broth. After ~ 16 hrs growth the absorbance at 600 nm (OD_{600}) of the overnights were measured using an Eppendorf BioPhotometer spectrophotometer. If any were outside the linear readable range of 3.0 A, the samples were diluted by a known amount and remeasured. The samples were then standardised to an OD_{600}

of 0.6 using the respective broth, before a further 100X dilution was made by adding 4 µL of the 0.6 OD_{600} sample to 396 µL of appropriate broth.

For each prepared sample 100 µL was pipetted into 3 wells in a BioScreen well plate. The plate reader was then setup to take readings at 15 mins intervals over 36 hrs, whilst the plate was incubated at 37 °C with shaking occurring for 30 s before readings were taken.

The data generated from the three technical repeats were used in Excel to generate growth curves for each strain, with error bars shown for standard error.

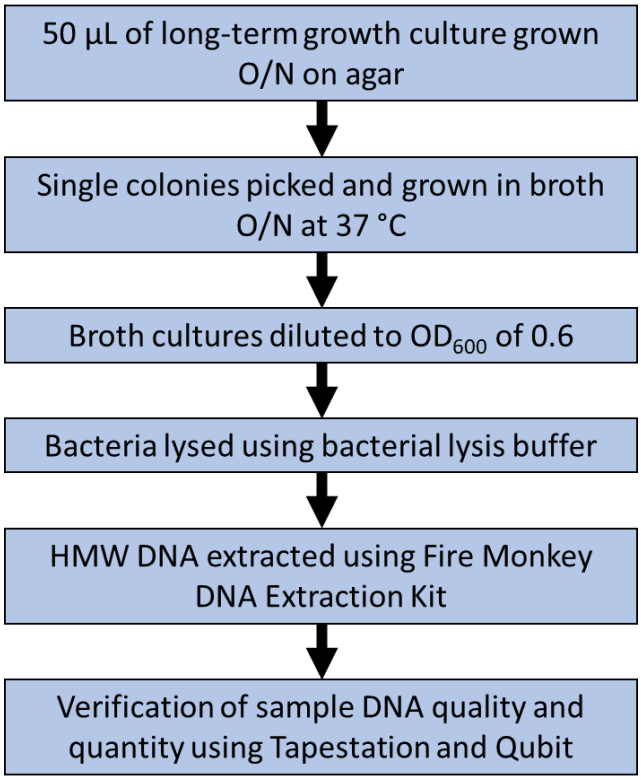


Figure 7 | Process of HMW DNA Extraction

Summarised version of the method taken to extract HMW DNA from bacterial samples taken. The first step from this is in reference to the long-term cultures and can be ignored for extractions other than these.

3.5 DNA extraction and library preparation

3.5.1 DNA extraction from pure isolates and long-term growth cultures

For DNA extraction (Figure 7) from glycerol stocks of Typhimurium and Typhi isolates, samples were grown on LB agar or LB-NaCl+aro agar plates respectively overnight at 37 °C and stored thereafter at 4 °C. Single colonies were used to inoculate 2 mL of LB broth which were incubated at 37 °C, 200 rpm overnight.

For DNA extraction from long-term growth cultures, 50 µL of culture from each medium was spread on respective agar plate before being inoculated at overnight at 37 °C and stored thereafter at 4 °C. Twelve single colonies from these plates were picked based on difference in size and not satellite to larger colonies compared to the original parent in a hope that genomic rearrangement had occurred. Size difference is used as a means of determining potential rearranged colonies as rearrangement can have a negative impact on growth rate (2.4.3 Impact of genome rearrangement). Single colonies were used to inoculate 2 mL of LB broth which were incubated at 37 °C, 200 rpm overnight.

After ~ 16 hrs growth, OD_{600} measurements of the 2 mL cultures were determined using a spectrophotometer. Cultures were adjusted to an OD_{600} of 0.6 for DNA extraction as this provided the cell number of 10^9 cells per 1 mL that was optimal for the extraction of high-molecular weight (HMW) DNA using the RevoluGen PuriSpin Fire Monkey DNA kit.

The bacterial lysis buffer for Fire Monkey DNA extractions was prepared beforehand in batches, 3 mg mL^{-1} lysozyme was prepared before sterile water and Triton X-100 was added to give a concentration of 1.2% Triton X-100 in the solution.

HMW DNA was extracted using a modified protocol: 1 mL of OD_{600} adjusted overnight cell cultures were centrifuged at 14,500 g for 2 mins to pellet. The harvested cells were then resuspended in 100 µL of bacterial lysis buffer (described above) then incubated for 10 mins at 37 °C to lyse the cells. Afterwards 10 µL of 20 mg mL^{-1} RNase A (Sigma-Aldrich) was added to each prep before pipette-mixing and incubating for 5 mins at room temperature. After this 20 µL of 20 mg mL^{-1} Proteinase K stock solution and 300 µL of lysis solution (LSDNA) was added to each sample before being pipette-mixed and incubated for 20 mins at 56 °C.

After incubation, 350 µL of binding solution (BS) was added to each sample and mixed via inversion, before adding 400 µL of freshly prepared 75 % isopropanol and inverted again.

Samples were then added to Fire Monkey spin columns and centrifuged at 4,700 g for 1 min and flow-through was discarded. Samples were firstly washed with 500 µL wash solution (WS) by being centrifuged at 4,700 g for 1 min and the flow-throughs were discarded. Samples were then subsequently washed with 500 µL of freshly prepared 90 % ethanol by being centrifuged at 11,400 g for 3 mins and flow-throughs were discarded. Samples afterwards were then centrifuged again at 11,400 g for 1 min to remove any residual ethanol.

After the wash steps, the spin columns were inserted into clean Lo-Bind Eppendorfs which were preheated at 80 °C before 100 µL elution buffer (EB, also pre-heated to 80 °C) was added. These columns/tubes were heated at 80 °C for 1 min before being centrifuged at 1,200 g for 2 mins to elute the extracted DNA (fraction A). This step was then repeated to produce a second fraction (fraction B). Fraction B was preferentially used in sequencing as this contained longer, better quality fragments of HMW DNA compared to Fraction A. Both fractions were checked for quantity and quality via ThermoFisher Qubit (3.5.4 DNA Quantification) and Agilent Tapestation (3.5.3 DNA Quality).

3.5.2 Sample Validation

To verify that BRD948 colonies grown from the long-term samples were not contaminants after months of being left to grow, a PCR was performed to check for the *aroC* gene deletion. PCR results from such colonies were compared against two positive controls: one that had a complete *aroC* gene from Typhimurium SL1344 and WT Typhi which was confirmed to have *ΔaroC*. Water was used as a negative control to ensure that the PCR reaction components were not contaminated.

The positive controls were picked from weekly prepared LB(-NaCl+aro) plates with a single colony first mixed in 50 µL of water and incubated for 10 mins at 99 °C to lyse the cells. The samples were taken from 50 µL the overnight culture and were also incubated for 10 mins at 99 °C to lyse the cells. Each PCR confirmation required 1 µL DNA, 25 µL of PCR master mix and 0.125 µL of 100 µM of both reverse and forward primer (both listed in Table 3). These were then mixed and briefly centrifuged before being amplified using a preprogrammed thermocycler at 95 °C for 50 s, then 25 cycles of 95 °C for 10 s; 55 °C for 1 min and 72 °C for 1 min, followed by 72 °C for 1 min and then cool to 4 °C to prevent heat degradation.

PCR Master Mix contents (1.125 mL solution, 1.1x Concentration		
Compound	Concentration	
Tris-HCl	22 mM	
KCl	22 mM	
$MgCl_2$	1.65 mM	
Each dNTP	220 µM	
Taq DNA Polymerase	22 U/mL	
PCR Primers (sequences are 5'-3', 100 µM)		
Compound	Sequence	Amount (µL)
Forward (aroC-05)	GTGATCCATCAGTACGATCG	0.125
Reverse (aroC-06)	GACAACTCTTTCGCGTAACC	0.125

Table 3 | Contents of PCR Master Mix and PCR primers used for sample validation.

The reasoning behind this is that the full *aroC* gene, a gene essential for Typhi's pathogenicity, produces a 1010 bp fragment whereas the ΔaroC gene from vaccine strain BRD948 instead produces a fragment of approx. 400bp as this is attenuated by a deletion of 659 bp. The PCR products were put through gel electrophoresis via Tapestation (Agilent) and an image of the gel was assessed to confirm if the samples had *ΔaroC* based on bands matching or similar to the bp values listed above (Figure 8).

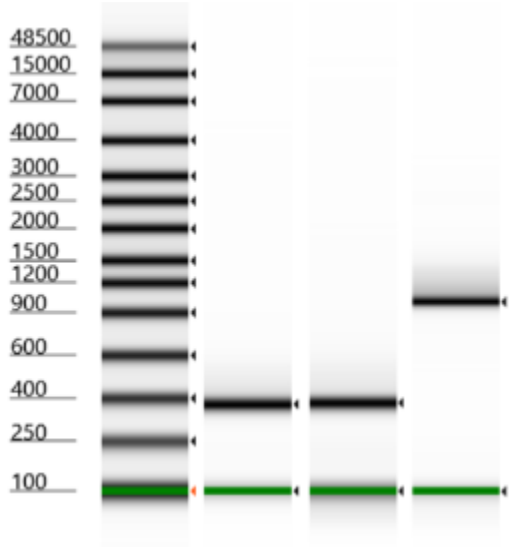


Figure 8 | Example aroC PCR Tapestation

From left to right: Ladder, *ΔaroC* sample, *ΔaroC* positive control, and *aroC* positive control.

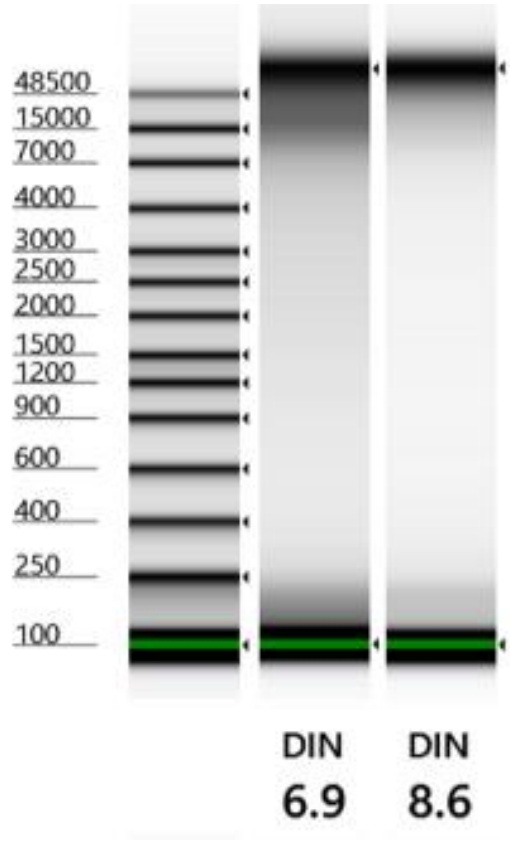


Figure 9 | Example DNA Quality Check Tapestation

DINs as seen can be indicated by the smearing of the bands seen on the gel, with lower DINs showing greater smearing.

3.5.3 DNA Quality

Gel electrophoresis was performed to check the presence and quality of genomic DNA after DNA extraction and to check the results from ΔaroC PCR check. Before preparation, it was ensured that all Genomic Tapestation components used were at room temperature before the process began. Genomic Tapestations were prepared by adding 10 µL of genomic DNA sample buffer to the ladder and sample tubes and 1 µL of genomic DNA ladder was added to the first, ladder tube and 1 µL of DNA sample to the remaining tubes.

When checking for good-quality, HMW DNA after DNA extraction on these Tapestation images, ideally high DIN (DNA Integrity Number) values of 7.5-10 and a single peak ideally >60 kb were satisfactory as this suggests low amounts of sheared DNA with the peaks generated being of extremely high length (Figure 9).

3.5.4 DNA Quantification

The Qubit dsDNA Broad range (BR) assay kit (Invitrogen) was used to quantify the amount of DNA extracted. Before preparation, it was ensured that all Broad Range (BR) Qubit components used were at room temperature before the process began. BR reagent was diluted 200-fold in the BR buffer to prepare the working buffer. Standards were prepared by aliquoting 190 µL working buffer into two Qubit tubes and 10 µL of each BR standard was added, before the standards were mixed and then incubated for 2 mins in the dark. The samples

were prepared by aliquoting 198 µL working buffer into Qubit tubes for each sample and 2 µL of DNA sample was added, before being mixed and incubated for 2 mins in the dark. Qubit functions by using fluorescent dyes that only emit when bound to target molecules, the fluorescence generated by this is detected by the fluorometer used.

Once the standards and samples were prepared the Qubit fluorometer (v3.0) was calibrated using the standards before the samples were then measured. As the sample volume and concentration was known, the total amount of DNA in samples could be quantified. For DNA library preparation, ideally 850 ng of DNA minimum within a sample was the cut-off point as this allowed significant leeway in sample loss due to later processes (such as AMPure bead concentration).

3.5.5 DNA Library Preparation

850 ng of DNA of satisfactory quality (i.e. >7.5 DIN and >60 kb peak length) was firstly required to be concentrated to approximately 80 ng/µL DNA (approx. 600ng of gDNA) using a pre-AMPure XP Bead (0.6X ratio, Agencourt) concentration step. This step both concentrated the sample and also removed low-length fragments from the sample, as the AMPure beads preferentially bind to long fragments. This input is higher than that recommended by the manufacturer as the aim for this method is to produce outputs that can generate complete circularized genomes for GS determination. As the tagmentation process used for this library preparation causes fragmentation of the DNA into smaller fragments whilst being tagged, increasing the DNA input reduces the average number of times a fragment is fragmented during this process (Figure 10).

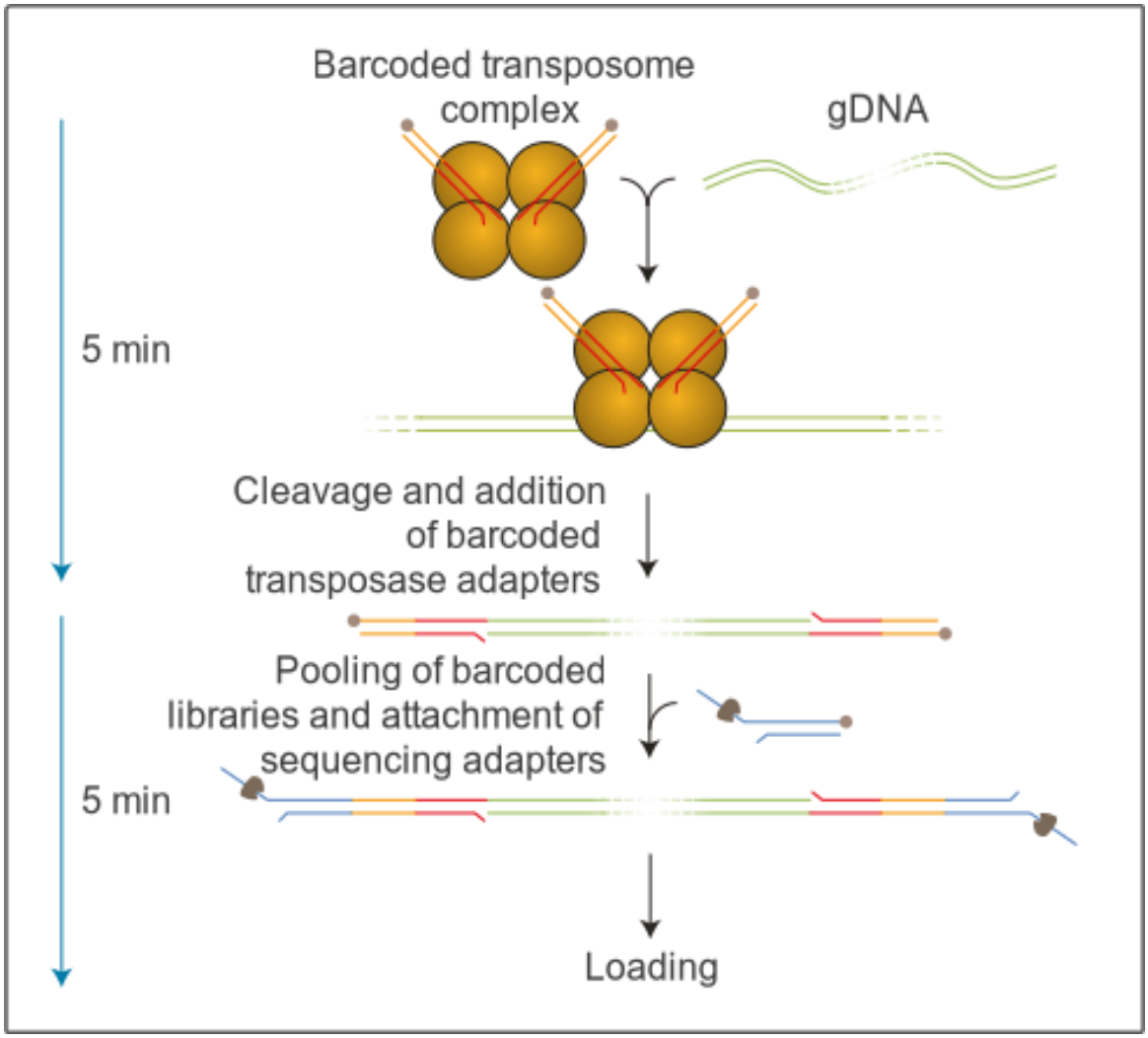


Figure 10 | Function of Tagmentation

Transposome complexes bind to the genomic DNA, cleaving and attaching barcoded adapters to the cleaved DNA. From here sequencing adapters are added which attach to the adapters. Tagmentation is used in this library preparation to both cleave DNA as well as add barcodes that can be detected in basecalling after sequencing to differentiate samples from the pool (Image source: Oxford Nanopore Technologies)

To perform a pre-AMPure XP bead concentration, firstly 6 µL of beads per 10 µL of sample containing 850 ng was added before being vortexed at 1800 rpm for 2 mins and incubated at room temperature for 5 mins. The samples were then placed into a magnetic stand for ~2 mins to separate the beads from the supernatant which was discarded. The beads were then washed twice with 200 µL of freshly prepared 80 % ethanol, with the supernatant removed and discarded each time. After this is beads were then left to air dry for 15 mins before being resuspended in 8.5 µL of water which was then vortexed at 1800 rpm for 2 mins and left to incubate at room temperature for 2 mins. Subsequently the beads were separated using a magnetic stand for approximately 2 mins and 7.5 µL of supernatant containing concentrated DNA was taken and put into a new Lo-Bind Eppendorf.

The samples were then barcoded using a DNA Rapid Barcoding Kit (SQK-RBK004, ONT). 7.5 µL of ~600 ng of DNA sample and 2.5 µL fragmentation mix (one unique mix for each sample) were added together, flick mixed and heated in a preprogrammed thermal cycler at 30 °C for 1 min, 80 °C for 1 min and then on hold at 4 °C. After barcoding, 10 µL of each barcoded sample was pooled together into a pooled library and flick mixed.

The pooled library was then cleaned and purified using a 1X AMPure XP bead cleanup where 10 µL of beads were added per 10 µL of pooled library. This ratio was used to retain as many fragments as possible from the sample. The AMPure XP bead process was the same as the above until after the air-dry step. In contrast to the above AMPure bead process, 12 µL of MinION buffer (10 mM Tris-HCl, 50 mM NaCl, pH 8) was used to resuspend the beads, before 10 µL of supernatant containing cleaned-up, pooled library was taken and retained in a new Lo-Bind Eppendorf. Another 1 µL was taken for analysis via Tapestation to confirm that the pooled library was of high enough quality for sequencing. Ideally the pooled library had a single peak of high length and DIN in the Tapestation image to reflect non-sheared, HMW DNA ready for sequencing.

To this cleaned pooled library, 1 µL of Rapid Sequencing Adapters (RAP) was added and incubated for 5 mins at room temperature, before the sample was then prepared to be loaded into the MinION flow cell.

3.6 MinION Flow Cell loading and Sequencing

The MinION (Figure 11) was plugged into the computer being used for the sequencing process and the flow cell slotted into it. If nothing erroneous has taken place (faulty wiring, etc.) the flow cell should be present on the provided GUI for MinION. Flow cells were checked using the QC DNA present in the internal buffer that was shipped within the flow cells to ensure that the flow cell has a good enough number of pores for use, with anything above 800 pores being acceptable to perform sequencing.

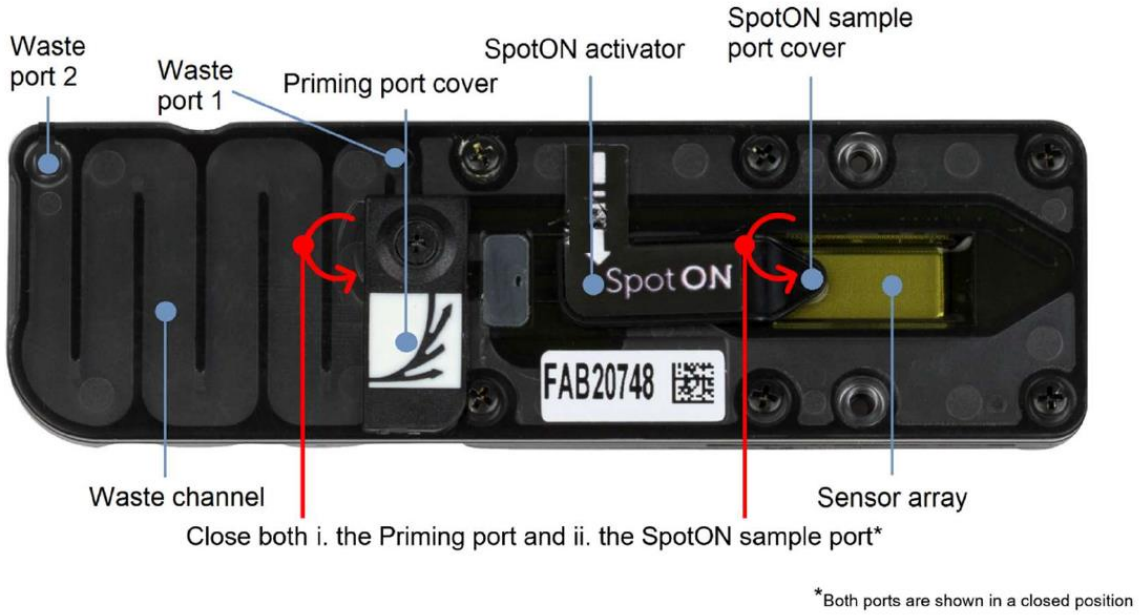


Figure 11 | Example MinION Flow Cell

The layout of a MinION flow cell from above. Note that both the priming port and sample port must be closed before starting MinION sequencing. (Image source: Oxford Nanopore)

Sequencing Buffer (SQB), Loading Buffer (LB), Flush Tether (FLT) and Flush Buffer (FB) were first thawed at room temperature before being placed on ice. The flow cell's priming port was exposed and a small volume, no more than 20 µL, was drawn to remove any bubbles present as these can have a significant negative effect on sequencing efficiency. The flow cell was visually checked to confirm there was continuous buffer from the priming port to across the sensor array. FLT was then pipette mixed and 30 µL of FLT was added to the tube of FLB and pipette mixed to produce the priming mix. The flow cell was primed using 800 µL of this priming mix by loading into the flow cell carefully via the priming port whilst avoiding introducing air bubbles and was left to rest for 5 mins.

The contents of the SQB and LB tubes were thoroughly mixed by pipetting and immediately afterwards 34 µL of SQB, 25.5 µL of LB, 4.5 µL of nuclease-free water and 11 µL of DNA library

were combined in a new tube. The SpotON sample port was opened and another 200 µL of priming mix was added into the flow cell via the priming port while avoiding the introduction of air bubbles. Afterwards the prepared library sample was gently pipette mixed prior to loading and 75 µL of library sample was added to the flow cell via the sample port in a dropwise fashion, making sure no air bubbles were introduced. Once complete both the sample port, priming port and MinION lid were closed.

The sequencing was performed using the following options, the kit selected was SQK-RBK004, live basecalling was disabled, the standard bias voltage (-180 mV) and the run time was set to 72 hrs to get the most sequencing data out of the flow cell as possible.

Once sequenced, the fast5 sequencing data was retrieved and basecalled using Guppy version 2.3.7 to generate fastq files which were first concatenated and then demultiplexed using qcat.

3.7 Bioinformatics workflow for GS Determination

After the Typhimurium and Typhi isolates were sequenced, the sequences were then run through an assembly pipeline (Figure 12) consisting of Porechop (v 0.2.3) to trim reads by 50 bp, to Nanofilt (v 0.1.0) (De Coster *et al.*, 2018) using settings to keep reads with quality greater than 6, and either length greater than 6 kbp or 500 bp filter depending on the genome coverage of the sample to Flye (v 2.5) (Lin *et al.*, 2016, Kolmogorov *et al.*, 2019) for assembly. The purpose of Porechop was to find and remove adapters from ONT Nanopore reads and remove chimeric reads, where an adapter is in the middle of them. Nanofilt was used to filter out poor quality reads as well as short reads. Flye is the *de novo* assembler used for assembly by using repeat graphs built with approximate sequence matches.

After assembly, these sequences were then processed through Prokka (version 1.13) (Seemann, 2014) for manual determination if necessary and *Socru* (version 2.2.2)(Page *et al.*, 2020) for automatic determination if possible. Nanostat (version 0.1.0) was also used at all steps from the start of the pipeline up to Flye to provide us various useful stats and follow the reads filtering process.

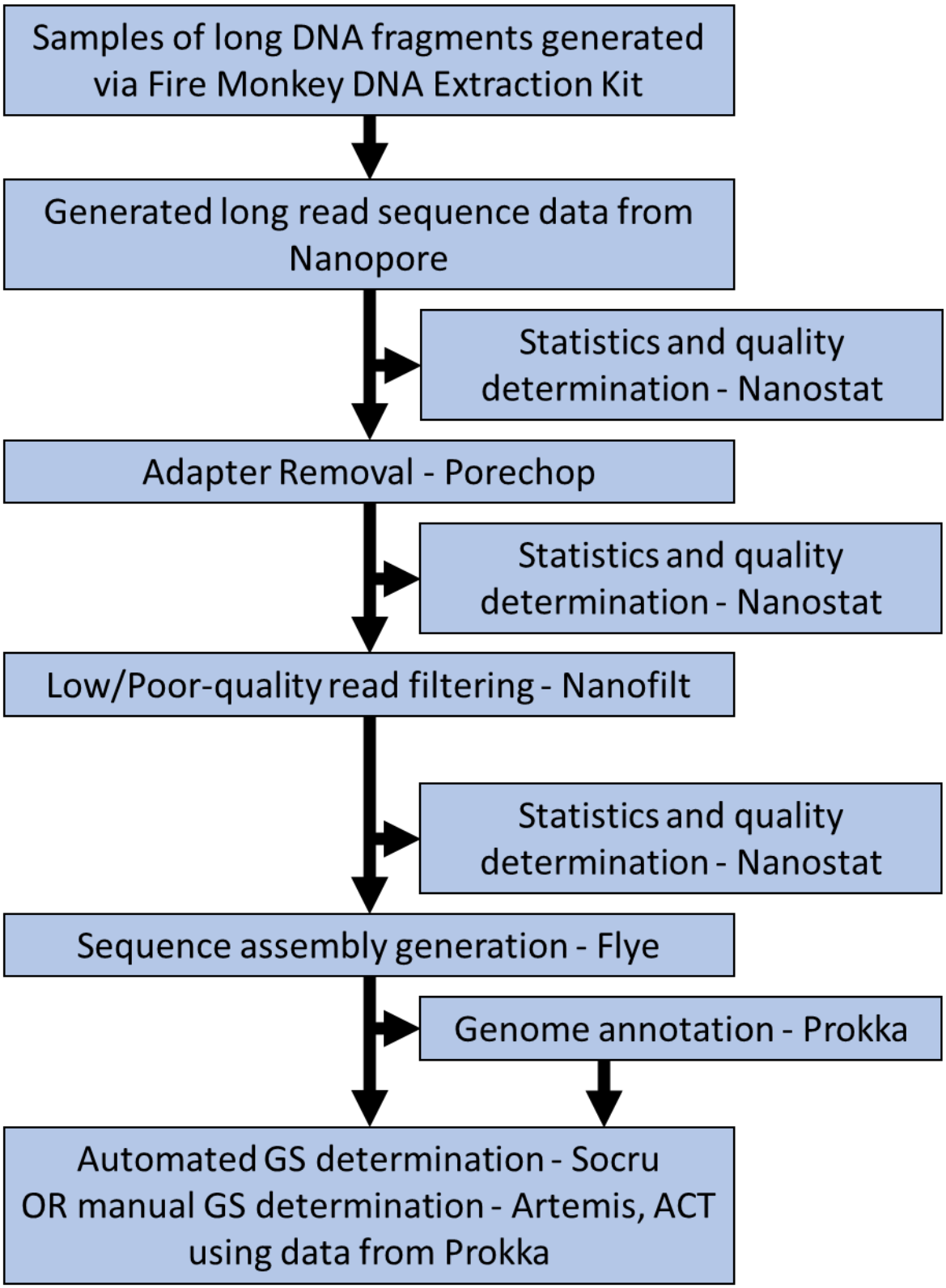


Figure 12 | The Pipeline Used for DNA Assembly

The general method taken from initial extraction to sequencing via Nanopore to defining the GSs of the assemblies.

3.8 Phylogenetic Tree Generation

Short read sequences of the DT2 isolates (36 total, from various regions in Germany from 1997 through 2003, originally provided by the Robert Koch Institute in Wernigerode, Germany) were provided by the Kingsley group and were compared against a reference genome (NC_022544.1 (Kingsley *et al.*, 2013)) using *snippy* (Seemann, 2015) to find SNPs before being processed with *snippy-core* to combine these outputs into a core SNP alignment. This data was then passed to IQ-TREE (Nguyen *et al.*, 2014) and by using the default settings generated a phylogenetic tree of the DT2 samples that could be visualised with iTOL (Letunic and Bork, 2019).

3.9 Long-read assembler comparison

The assemblers compared were Flye (Kolmogorov *et al.*, 2019), Canu (Koren *et al.*, 2017), Miniasm (Li, 2016) and Raven (Vaser and Šikić, 2021), using a hybrid assembly approach (Unicycler) to generate a 'gold standard' assembly for comparison.

All assemblers used in the comparison used the same settings for both Porechop and Miniasm to make the comparisons as fair as possible. Settings for each assembler were all default apart from any settings to be set specifically for Nanopore reads (details given in Chapter 5).

3.10 Hybrid assembly

The short read sequences of the DT2 isolates mentioned previously were also combined with the long-read sequences to generate gold standard references using Unicycler (Wick *et al.*, 2017) using the default settings. Unicycler makes use of SPAdes graphs (Bankevich *et al.*, 2012) to combine large k-mer assemblies provided by long read sequencing (that can solve repeats in the genome) with short reads (a more connected and overall more accurate graph). This approach combines the two primary advantages to both types of reads with long reads able to resolve large structures and short reads with their much greater accuracy compared to long read in order to generate reference genomes.

3.11 RNA Extraction and Sequencing

3.11.1 RNA Extraction

RNA Sequencing requires sufficient cells in a culture (up to 4 x 10^9) as well as the culture to be within the early exponential phase of the culture growth cycle so the quantity of RNA is sufficient for sequencing. The further along in growth bacterial colony has gone through, the

greater variance (and consequently noise) there is between repeats, meaning there needs to be a balance in obtaining enough RNA and keeping noise in samples low.

For the sequencing, samples were inoculated within 2 mL of suitable media, which was then incubated at 37 °C overnight.

100 µL of the overnight samples were used to sub-inoculate 10 mL triplicate repeats of LB-NaCl+aro (Luria-Bertani, no salt, aro mix added) and incubated at 37 °C until the culture reached an OD_{600} of approximately 0.3-0.35. These were conducted as triplicate repeats to provide statistical power for subsequent analysis, as RNA sequencing currently remains volatile in terms of noise present in the data.

The 10 mL cultures were centrifuged at 4,000 g for 10 mins to pellet the cells, the supernatant was removed, and the cell pellet was resuspended in 100 µL of RNAlater RNA stabilization agent.

RNA was extracted and purified using the Qiagen AllPrep DNA/RNA Kit according to the manufacturer's instructions: 600 µL of RLT Plus buffer was added to the resuspended cell pellet and pipette mixed before transferring the sample to an AllPrep DNA spin column and centrifuging at 8,000 g for 30 s. The flow-through containing the RNA was kept and pipette mixed with 700 µL of 70 % ethanol.

The sample was transferred to an AllPrep RNeasy spin column and centrifuged at 8,000 g for 30 s and the flow-through discarded. The RNeasy spin column was washed with 700 µL of RW1 buffer and centrifuged at 8,000 g for 30 s and the flow-through was discarded. 500 µL RPE buffer was added to the spin column which was then centrifuged at 8,000 g for 30 s with the flow-through discarded, this was then repeated. The spin column was then centrifuged at 16,000 g for 1 min to remove residual wash buffer with the flow-through discarded. 30 µL of RNAse-free water was added to the column and then centrifuged at 8,000 g for 1 min into a new Eppendorf to elute the RNA, this is repeated for a second elution fraction.

3.11.2 Ribosomal RNA Depletion

Within total RNA (tRNA), ribosomal RNA (rRNA) makes up an extremely large proportion of the sample which would make up the majority of data produced by RNA sequencing and

therefore would not be useful for picking up differential gene expression. The purpose of RiboCop (Lexogen) in this protocol is to deplete the rRNA.

The extracted RNA fractions were checked with Tapestation (See 3.11.4 RNA and D5000 Tapestation) and Qubit to confirm that rRNA, and assumingly tRNA, was present before having rRNA depleted using RiboCop according to manufacturer's instructions. To 26 µL of total RNA sample (containing an input between 1-1000 ng tRNA), 4 µL Hybridisation Solution (HS) was added alongside 5 µL of G- Probe Mix before the sample was pipette mixed. The sample was denatured using a preprogrammed thermomixer at 75 °C at 1250 rpm for 5 mins before decreasing the temperature of the thermomixer to 60 °C and incubating at 1250 rpm for 30 mins.

Whilst the sample was being denatured, 75 µL of Depletion Beads (DB) were added to a fresh Eppendorf before separating the beads with a magnetic stand for 2-5 mins. The supernatant was discarded and the beads were washed twice with 75 µL Depletion Solution (DS) before the beads were then removed from the magnetic stand and resuspended in 30 µL DS.

Thirty µL of freshly prepared depletion beads were added to the hybridised RNA and pipette mixed before being incubated using a preprogramed thermomixer at 60 °C, 1250 rpm for 15 mins. The beads were then separated using a magnetic stand for 5 mins before 60 µL of the supernatant containing rRNA-depleted RNA was transferred to a fresh tube.

To the supernatant, 24 µL of Purification Beads (PB) and 108 µL of Purification Solution (PS) was added before it was pipette mixed and incubated for 5 mins at room temperature. Afterwards, the purification beads were separated using a magnetic stand for 5-10 mins, the supernatant was then removed before washing the beads twice with 150 µL of 80 % ethanol before the sample was air-dried on the magnetic stand for between 5-10 mins to remove residual ethanol from the beads. The beads were then removed from the magnetic stand and resuspended in 12 µL Elution Buffer (EB) before incubating at room temperature for 2 mins. The beads were then separated using a magnetic stand for 2-5 mins before transferring 10 µL of the supernatant containing the purified rRNA-depleted RNA to a fresh Eppendorf.

Once rRNA depletion was completed, the samples were checked with Tapestation (See 3.11.4 RNA and D5000 Tapestation) to confirm that the rRNA was depleted. If the rRNA was depleted there will be no detectable RNA seen on the Tapestation. This is because within extracted

RNA samples a significant majority of the total RNA present is rRNA, and therefore with the rRNA depleted the remaining RNA is below the detectable range for the instruments used.

3.11.3 RNAseq Library Preparation

RNA libraries were prepared using the Qiagen QIAseq Stranded mRNA Select Kit with modifications to manufacturer instructions. With exception to the adapter plate preparation, 1/5th of input RNA and reagents are used. With the methods described above, fragmentation of extracted RNA was not required as their RINs were below 3.

3.11.3.1 RNA Fragmentation and Reverse Transcription

Within a PCR tube, to 5.8 µL of rRNA-depleted RNA sample (containing an input of 1-100 ng mRNA) was added 1.6 µL 5X RT Buffer, 0.2 µL RT enzyme, 0.2 µL RNase inhibitor and 0.2 µL 0.4 M DTT and this sample was pipette mixed. The sample was then incubated using a preprogrammed thermocycler at 25 °C for 10 mins, 42 °C for 15 mins, 70 °C for 15 mins and then cool to 4 °C before being transferred to a fresh Eppendorf.

To the sample, 11.2 µL of QIAseq beads was added before being pipette mixed and incubated at room temperature for 5 mins. The beads were then separated using a magnetic stand for 5-10 mins and the supernatant was discarded before the beads were washed twice with 40 µL of 80 % ethanol. The beads were subsequently left to air-dry on the magnetic stand for 5-10 mins to remove residual ethanol. The beads were then removed from the magnetic stand and resuspended in 8 µL nuclease-free water before being separated again in a magnetic stand for 2 mins, before transferring 7.7 µL of the supernatant containing the DNA to a fresh PCR tube.

3.11.3.2 Second-Strand Synthesis, End-Repair and A-Addition

To 7.7. µL of DNA sample, 1 µL Second Strand Buffer (10X) and 1.3 µL Second Strand Enzyme Mix were added and pipette mixed before being incubated using a preprogramed thermocycler at 25 °C for 30 mins, 65 °C for 15 mins and then cooled to 4 °C before being transferred to a fresh Eppendorf.

To this sample, 14 µL of QIAseq beads were added, pipette mixed and incubated at room temperature for 5 mins. The beads were then separated using a magnetic stand for 5-10 mins before the supernatant was discarded and the beads were washed twice with 40 µL of 80 % ethanol, with residual ethanol removed from the beads by air drying on the magnetic stand

for between 5-10 mins. The beads were removed from the magnetic stand and resuspended in 10.4 µL of nuclease-free water before the beads were separated with a magnetic stand again for 2 mins, before transferring 10 µL of the supernatant containing the cDNA to a fresh PCR tube.

3.11.3.3 Adapter Plate Preparation for Strand-Specific Ligation

To prepare the Adapter Plate for strand-specific ligation, the plate was first thawed on ice before vortexing and then centrifuged briefly. The protective adapter plate lid was removed, the foil seal was carefully pierced and 115 µL of nuclease-free water were added to each adapter well to be used (each well containing 10 µL neat adapter) before being pipette mixed 6 times before transferring 10 µL to a fresh plate. To the transferred solution 10 µL of nuclease-free water were added to each well in the fresh plate and pipette mixed 6 times resulting in a ready to use 1:25 dilution of the Adapter Plate.

3.11.3.4 Strand-Specific Ligation

To the 10 µL of cDNA produced earlier, 2 µL of 1:25 diluted unique adapter (one unique adapter for each sample), 5 µL 4X Ultralow Input Ligation Buffer, 1 µL Ultralow Input Ligase, 1.3 µL Ligation Inhibitor and 0.7 µL water was added and pipette mixed. The sample was then incubated using a preprogrammed thermocycler at 25 °C for 10 mins and then cooled to 4 °C before being transferred to a fresh Eppendorf.

To the sample, 16 µL of QIAseq beads were added, before being pipette mixed and incubated at room temperature for 5 mins. The beads were separated using a magnetic stand for 5-10 mins before the supernatant was removed and the beads were washed twice with 40 µL of 80 % ethanol and air-dried on the magnetic stand for 5-10 mins. The beads were removed from the stand and resuspended in 18.4 µL nuclease-free water before being separated with a magnetic stand for 2 mins, before transferring 18 µL of the supernatant containing the DNA to a fresh Eppendorf.

To the sample, 21.6 µL of QIAseq beads were added before being pipette mixed and incubated at room temperature for 5 mins and then separated using a magnetic stand for 5-10 mins. The supernatant was then removed and the beads were washed twice with 40 µL of 80 % ethanol, residual ethanol was removed from the beads by air-drying on the magnetic stand for 5-10 mins. The beads are taken out of the magnetic stand and resuspended in 5 µL

nuclease-free water before the beads were separated using a magnetic stand for 2 mins and then 4.7 µL of the supernatant containing DNA was transferred out to a fresh PCR tube.

3.11.3.5 CleanStart Library Amplification

To the 4.7 µL of DNA sample, 0.3 µL of CleanStart PCR Primer Mix and 5 µL 2X CleanStart PCR Mix were added and pipette mixed before being treated with the following PCR program on the preprogramed thermal cycler: preheated lid at 99 °C to prevent precipitation, 37 °C for 15 mins, 98 °C for 2 mins, for 16 cycles of 98 °C for 20 s, 60 °C for 20 s, 72 °C for 30 s, then 72 °C for 1 min before being held at 4 °C.

To the sample, 12 µL of QIAseq beads was added, pipette mixed and incubated at room temperature for 5 mins. Beads were separated on a magnetic stand for 5-10 mins before discarding the supernatant. Beads were then washed twice with 40 µL of 80 % ethanol, with residual ethanol removed by 5-10 mins air-drying. Beads were removed from the magnet and resuspended in 10.4 µL nuclease-free water before a final 2 min magnetic separation. Ten µL of the DNA-containing supernatant were transferred to a fresh Eppendorf tube.

Afterwards, the bead cleanup step above was repeated to remove excess adapters but in this case were resuspended beads in 5 µL nuclease-free water instead of 10.4 µL, and 4 µL of the supernatant was transferred to a fresh PCR tube. At this point, to the 1 µL left in the Eppendorf with the final QIAbeads added another 1 µL of nuclease-free water was added and pipette mixed. One µL of this sample was used to calculate concentration using HS DNA Qubit and the other 1 µL was used to perform a QC RNAseq library check with a high sensitivity D5000 tape with Tapestation (See 3.11.4 RNA and D5000 Tapestation).

The full pipeline of the RNA extraction, rRNA depletion and RNAseq library preparation described above is presented below in Figure 13.

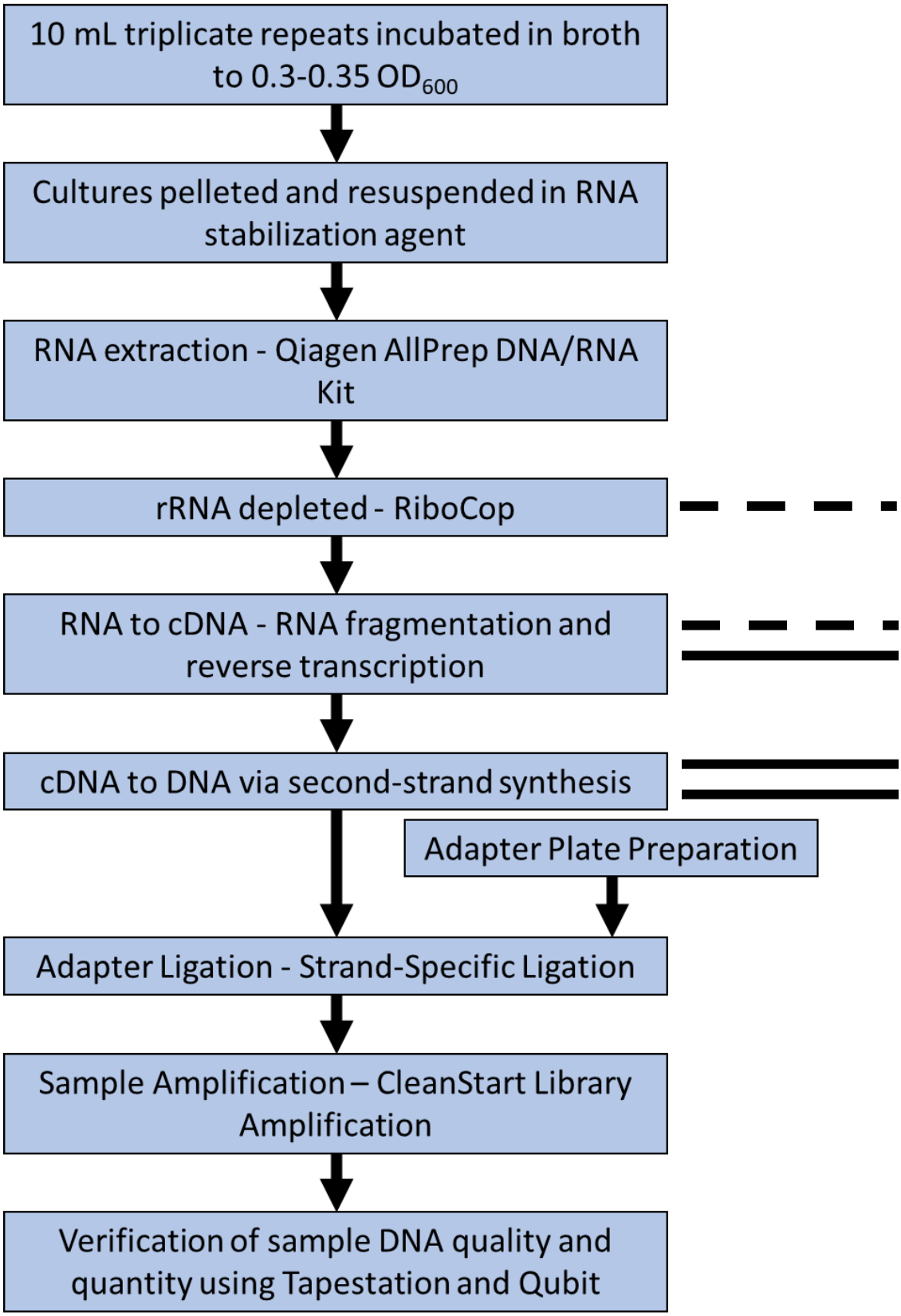


Figure 13 | Pipeline for RNA Extraction and RNASeq

The pipeline used to extract RNA from samples for the purpose of determining differential gene expression. The dashed lines represent RNA strands while full lines represents DNA.

3.11.4 RNA and D5000 Tapestation

This check was employed at three stages. Once after RNA extraction to verify the presence of total RNA and also to determine the RIN of RNA fragments in the event that they may need further fragmenting and once again after Ribocop depletion of rRNA which should mean that there is not any RNA detectable. Both of these were checked using RNA tapestation protocol. A third check was performed after CleanStart Library amplification using the D5000 Tapestation protocol provided to determine that enough DNA library was present in the sample to perform RNAseq.

For the RNA Tapestation, 5 µL of RNA sample buffer were added to the ladder and sample tubes and 1 µL RNA ladder to the first ladder tube as well as 1 µL RNA sample to the remaining tubes. Both the ladder and samples were denatured at 72 °C for 3 mins, cooled at 4 °C for 2 mins and then spun down to move the samples to the bottom of their tubes before being added to the Tapestation.

D5000 Tapestation was conducted by adding 10 µL of the appropriate DNA sample buffer to the ladder and sample tubes and adding 1 µL appropriate DNA ladder to the first, ladder tube and then 1 µL DNA sample to the remaining tubes.

3.11.5 RNASeq Qubit

Qubit is carried out at multiple stages in the RNA sequencing process. In this case RNA High Sensitivity (HS) reagents were used instead, though the process remained the same in terms of preparation of samples for measurements (see 3.5.4 DNA Quantification).

Once RNA was extracted, the process above was carried out to determine the concentration of RNA within samples and to determine if the samples had an adequate concentration and quantity for use in subsequent steps of the pipeline.

3.12 Bioinformatics workflow for RNA Sequencing

The sequenced RNA was processed through an assembly pipeline (Figure 14) (built originally by Ainsworth, E.) that consisted of FastQC (v. 0.72) (Andrews, 2015) to provide quality control checks to the reads generated to detect any potential problems such as sequence quality.

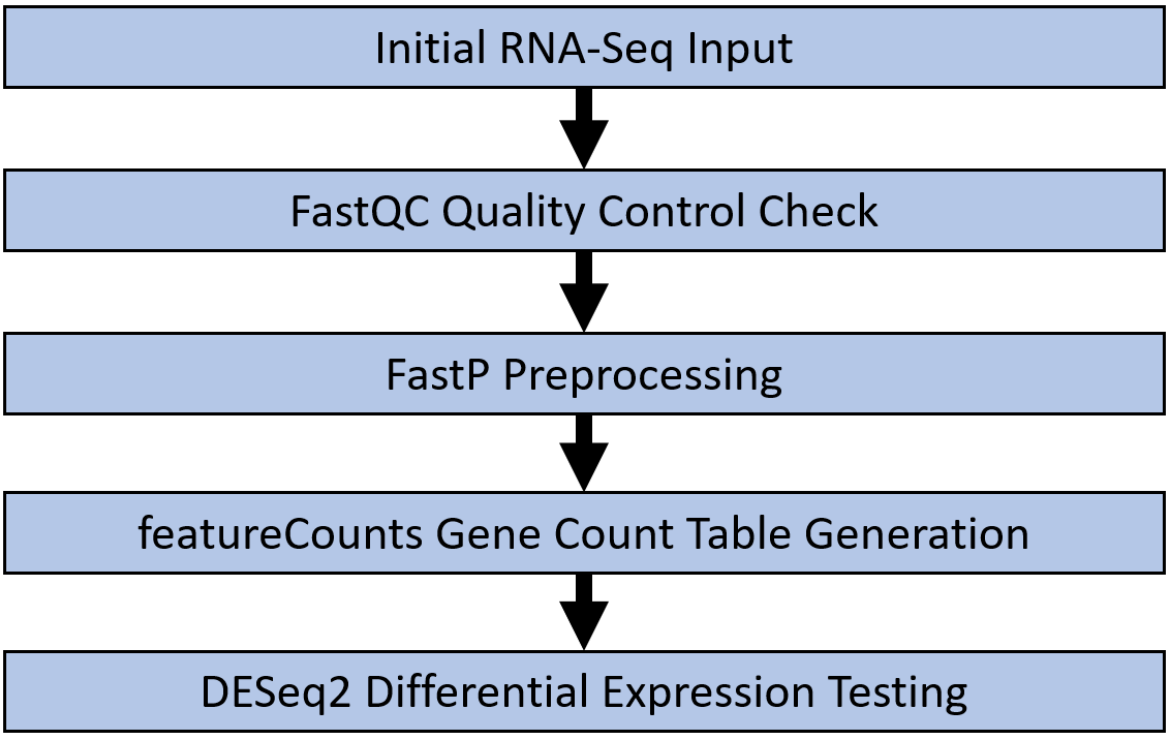


Figure 14 | Pipeline for RNA Sequencing

A generalised pipeline for discovering differential gene expression in rearrangements.

Once reads passed quality control FastP (v. 0.19.5) (Chen *et al.*, 2018) was used as preprocessing, after this the sequences were then aligned to the Typhi reference strain Ty2. The genes of the repeats were counted using featureCounts (v. 1.6.3) (Liao *et al.*, 2013) to measure gene expression against the reference. DESeq2 (v. 2.11.40.4) (Love *et al.*, 2014) then takes the count tables generated from featureCounts to determine differentially expressed genes between the sample and the reference by estimating variance-mean dependence.

To look at how various pathways within the cell were affected by rearrangement, a Typhi-specific Pathway/Genome database (provided by Langridge, G.)(Kingsley *et al.*, 2018) loaded into Pathway Tools (v. 24.0) (Karp *et al.*, 2002) was used in conjunction with the data produced by DESeq2 to produce a pathway map.

3.12.1 SNP and mutation detection

To check for SNPs, *snippy* (v. 4.4.3) (Seemann, 2015) was used to compare to a reference sequence to observe any changes as SNPs can be either synonymous (no change in amino acid) or non-synonymous, the latter of which can result in either missense caused by amino acid changes or nonsense caused by the mutation creating a premature stop codon.

This was also used in conjunction with Breseq (v. 0.34.0) (Deatherage and Barrick, 2014) to observe any mutations over time within samples of the long term experiments, particularly major changes such as large insertions and deletions that *snippy* does not detect. This was carried out by a three-way comparison with the sample of interest, a reference genome of BRD948 and a reference genome of Ty2. These references were employed to use the mutations known in BRD948 strains as a control to indicate accuracy. Changes versus Ty2 from both the sample of interest and the BRD948 reference were compared against each other to detect any new mutations. The purpose of this was to detect changes in the genome that were not rearrangements to see what effects there were on gene expression that do not involve rearrangement (as deletions may knockout entire genes or render them non-functional), or their potential effect on the genome structure (as large deletions or insertions can affect the ori-ter balance of the genome).

4 Optimization of Methods and Initial Assignment of Genome Structure

4.1 Introduction

Genome rearrangements were originally identified in *Salmonella* Typhi, with gene order in Typhi Ty2 being radically different compared to *E. coli* and Typhimurium LT2 both of which have largely conserved gene orders (Liu and Sanderson, 1995). Partial DNA digestion with I-*Ceu* I confirmed the presence of chromosomal rearrangements through homologous recombination in the *rrn* genes and was theorized to have a potential role in virulence.

In previous research, long-range PCR was used to identify genome rearrangements by looking for a PCR product for each possible primer/fragment combination for the tested bacteria (Figure 15), with each primer combination looking for a potential fragment repositioning. While this method is viable for detecting rearrangements, this has several issues associated that makes a scalable alternative desirable. Due to the number of possible fragment orders, 91 PCRs primer combinations are needed for *Salmonella* which requires a large amount of processing, extensive use of gel electrophoresis and furthermore the results can be subjective due to being interpreted from bands of varying brightness on a gel.

Recently, long read sequencing has become more widely available and in theory can achieve sequence reads long enough to span across the *rrn* operon repeat regions, allowing for the assembly of complete genomes. This presents the possibility of leveraging this for detecting genome arrangements, which would greatly improve the rate at which genome arrangements could be detected compared to long range PCR.

Due to this method being a relatively recent advancement, methodologies still need optimisation and tools for long read sequencing are still young compared to short read sequencing. A method of inducing genome rearrangements was first described in Haase, 2008, but was not taken further at the time. Finding a consistent means of inducing genome rearrangement in the laboratory via long-term growth, and media in which these rearrangements can survive in, is key for establishing future work.

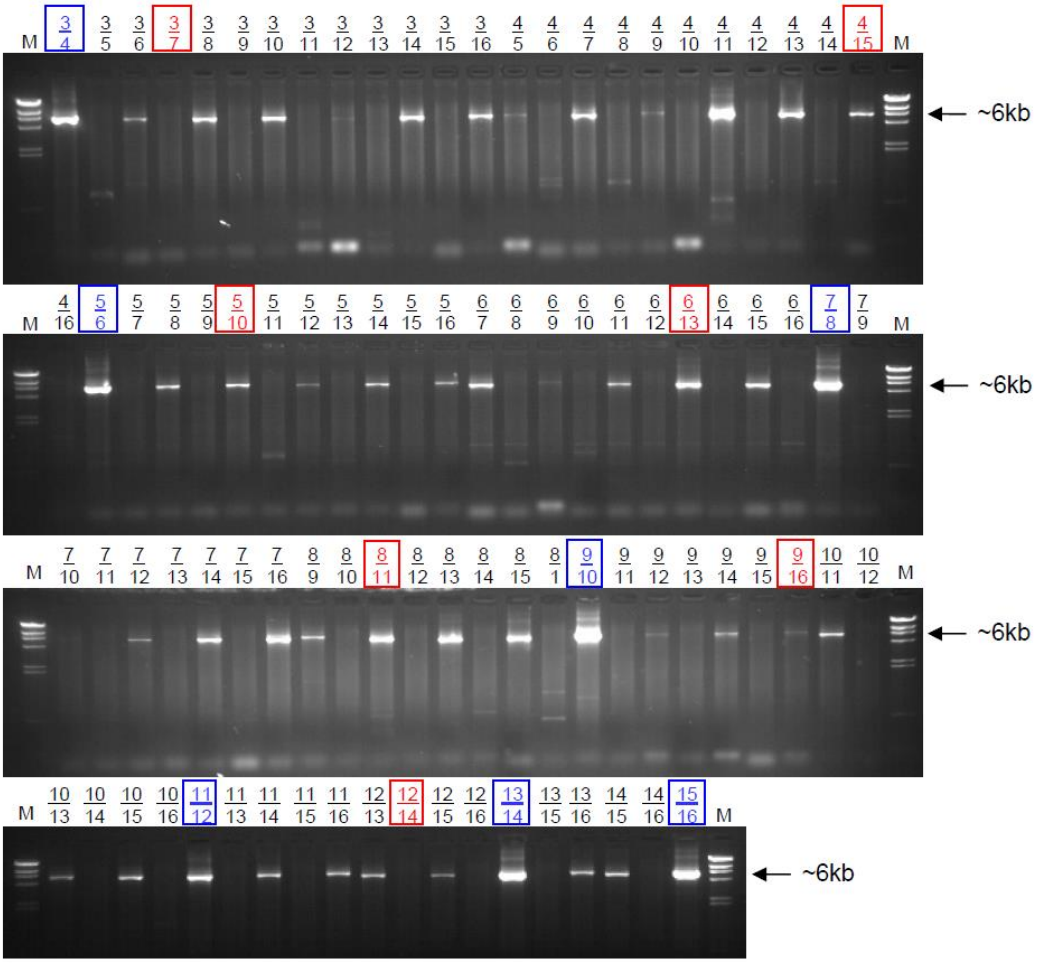


Figure 15 | Example Long-range PCR for GS Determination

Primer combinations are above every well, with combinations in red binding to regions within one I-*Ceu*I fragment and combinations expected to give a PCR product in blue, image source: Haase, 2008.

While rearrangements were first found in clinical samples of Typhi (Matthews *et al.*, 2011), in this work I was unable to work on clinical samples of Typhi due to this being a HG3 (Hazard Group 3) pathogen, which I am not qualified to handle. However, Haase, 2008 demonstrated that rearrangements were also inducible in a vaccine strain of Typhi (BRD948, classed as a HG2 organism) which I used for this work.

Other optimisation points include: parts of the DNA extraction to improve the quality and quantity of DNA acquired; the scale of the sequencing itself via multiplexing; the scale of bioinformatics analysis in order to further ramp up the process so that many samples at once can simultaneously have arrangements identified.

The purpose of this chapter was to take presently available protocols and optimize these alongside creating a full methodological pipeline from bacterial colony to complete genomes,

to culminate in a protocol that produces high quality long sequence reads that enable genome arrangements to be identified.

4.2 Specific Methods

The strains used in this chapter are Typhi BRD948 parent strain (hereby referred to as WT) and variants 7, 8, U and T (derivatives of WT produced previously). Typhimurium strain SL1344 was also used alongside several Typhimurium DT2 strains (Table 6 in Chapter 5).

Salmonella survival is sensitive to salt concentration (4.3.1 Growth Media Assessment). Salt tolerance was explored by creating a series of 10 mL LB-NaCl+aro broths and adding increasing salt concentration from 3 g L^{-1} to 10 g L^{-1}. 2 mL of each broth was taken and put into two new tubes, inoculated with 2.5 µL of T variant (as this had previously displayed high sensitivity to salt) and grown overnight at 37 °C and shaking at 200 rpm.

Cell input optimization was performed to control and optimize the input used at the start of the DNA extraction process. Growth curves were generated to compare CFUs (Colony Forming Units) with $OD_{600.}$ By comparing the two I could then estimate the number of cells in a given sample.

CFU counts were carried out by first standardising broth samples to 0.1 OD_{600} before being grown at 37 °C. Every 15 mins an OD_{600} reading was taken and every 30 mins a dilution series was prepared, with this process running up to 4 hrs. For the dilution series, 2 µL of sample was added to 18 µL of water or PBS (1:10 dilution) from 10^{-1} to 10^{-6} and mixed. For each dilution, three 2.5 µL dots were plated and grown overnight at 37 °C before the CFUs of each dilution were then counted.

The summary of the protocol and the parts that underwent optimisation are shown in Figure 16.

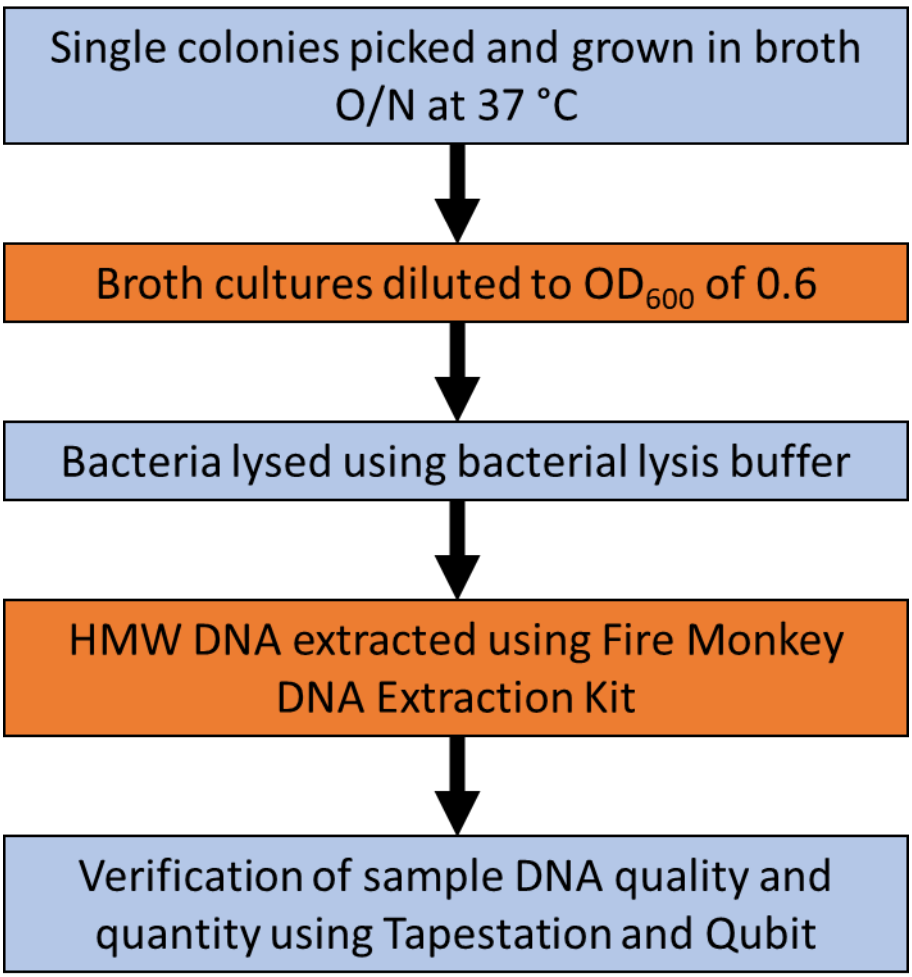


Figure 16 | Summary of the HMW DNA Extraction Method

Sections highlighted in orange are the sections optimised in this chapter.

4.3 Results

4.3.1 Growth Media Assessment

A problem encountered early on was certain variants of BRD948, in particular T, experienced little to no growth in regular LB+aro broth and agar. This presented a need to find a new or modified medium that all rearrangements would grow in so that extreme but biologically viable rearrangements were not lost due to media choice.

One of the first avenues taken was investigating the effect of NaCl content in the media used. *Salmonella* is known for relatively poor salt tolerance (Matches and Liston, 1972) especially when compared to other bacterial species such as *Escherichia coli* (Gibson and Roberts, 1986), so media with low NaCl concentration was investigated.

Results for this experiment were unfortunately inconclusive, with T growing at 10 g L^{-1} when this should be intolerable, and also growing at 4, 5, 7 and 8 g L^{-1}, not showing any distinct pattern of tolerance and also being non-repeatable, meaning no insight could be gathered from this data. Live/dead PCR testing conducted by Ainsworth, E., revealed that there were living cells of T in the glycerol stock meaning that the issue was not due to the absence of live cells in the 10-year-old glycerol stocks. A possible reason for this could be cells in this stock being viable but non-culturable which is a state that has been previously seen in Typhi (Roszak and Colwell, 1987). If so this would mean these cells would need resuscitation which has been conducted previously with Typhi (Zeng *et al.*, 2013) though this is outside the scope of this work.

Outside of these results it was previously shown in (Haase, 2008) that arrangements could be induced using LB-NaCl, which is a standard LB media without NaCI. With this I decided to go forward with using various low/no-NaCl media such as LB-NaCl, Iso-sensitest and MOPS EZ for long-term growth cultures to ensure that any induced rearrangements were not lost due to salt intolerance. In the future salt could be used to see if stress can encourage rearrangement.

4.3.2 Controlling Cell Input

As the optimal cell input for Fire Monkey HMW DNA extraction was stated by the manufacturer to be 1 x 10^9 cells, I needed to know at what Optical Density (OD) this number of cells would be within a given sample.

The manufacturer advised that samples should not exceed the stated input since excess loading into the spin columns of the extraction kit causes a decrease in both the quality and quantity of DNA extracted. By controlling cell input into DNA extraction, I aimed to reduce the issue of inconsistent output and make it less prone to overloading issues which can induce DNA shearing.

The samples used for this were dense overnight cultures in triplicate biological repeats to inoculate fresh media before incubating at 37 °C, the OD_{600} of these were measured at regular timepoints and an aliquot was taken to be plated out in order to calculate CFU.

Both the OD_{600} to achieve this cell quantity and the time taken to reach this OD_{600} was necessary for my work. To find the OD_{600} required to reach this quantity, I established a standard curve between OD_{600} and CFUs to then correlate the two (Figure 17, blue line).

Upon generating a growth curve (Figure 17, orange line), a very close correlation between OD_{600} and CFU counts was observed, which allowed calculation of the approximate number of cells in a given sample based on the OD_{600}.

Using a plot of absorbance (OD_{600}) vs CFU as shown in Figure 18, the following formula was generated:

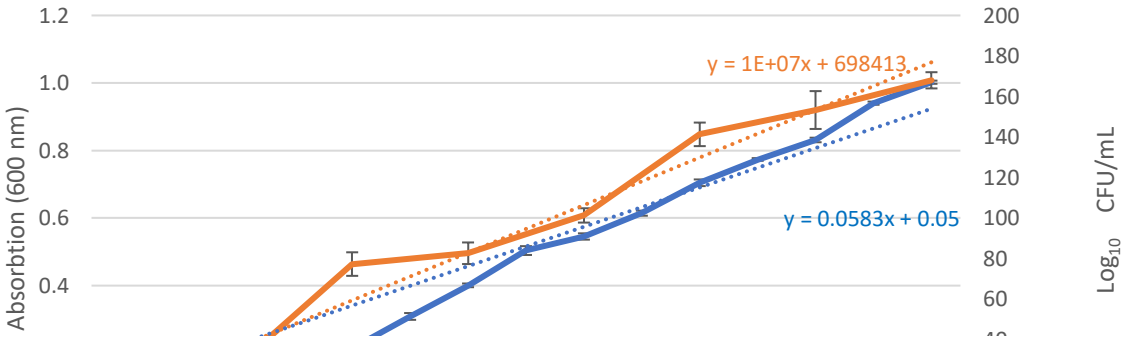


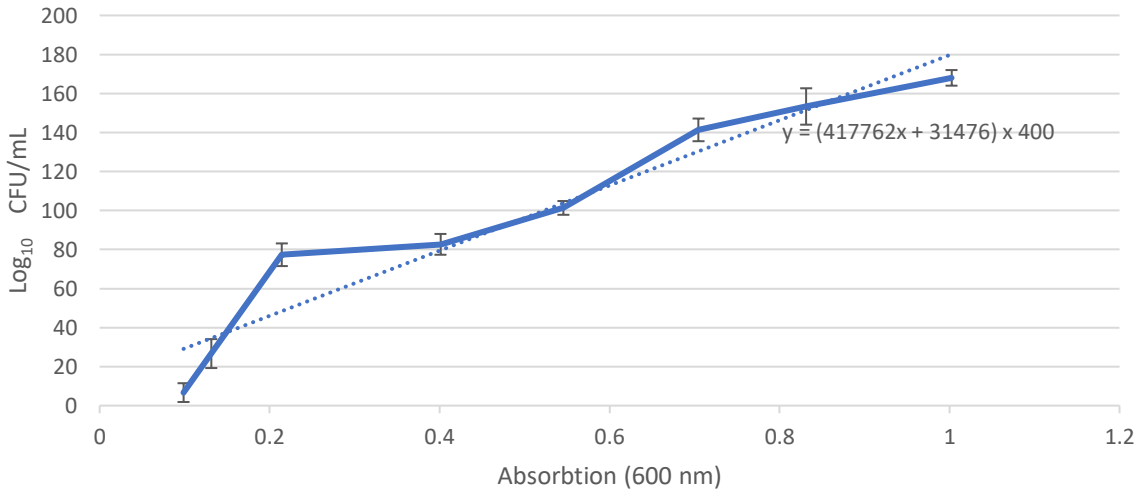
Figure 18 | OD_{600} and CFUs of BRD948 WT measured over time.

The blue line indicates the OD_{600} of WT while the orange line indicates the CFU of WT. The error bars show standard error of the data, n=3. Trendlines (dotted lines) were generated with linear trendlines.

Time (Mins)

$$y = (417762x + 31476) \times 400$$

Where x is OD_{600}, and y is CFU in 1 mL of undiluted WT. While this is far below the initial ideal



number of 1 x 10^9, it was found that 1 x 10^9 is the upper limit of input for the spin columns and any higher would have a negative impact on output. Extractions using this cell input showed consistently good quality and quantity, indicating that 1 x 10^8 is likely close to the optimal input and was used for future extractions.

To achieve this required cell input with overnights, that will go significantly over this desired OD_{600}, samples to go into DNA extraction were diluted to an OD_{600} of 0.6 with purified water before 1 mL was taken forward for DNA extraction.

4.3.3 DNA Concentration

4.3.3.1 Changes in Concentration Steps

DNA concentration is important as ideally 80 ng/µL was needed as the output from DNA extraction for the purpose of sequencing. In total, ~600 ng of genomic DNA was required, but as samples generally needed concentration to achieve the desired concentration, 850 ng DNA was processed through a bead clean-up step, to account for DNA loss during that step.

Attempts to improve sample retention throughout the process were first tested during and immediately after the extraction process (see section 3.5.1).

One potential route was using evaporation to reduce the amount of solution while retaining the DNA, as a result this would concentrate the DNA within the sample while removing the need for the pre-AMPure step. This was conducted by taking eluted samples after extraction with the Fire Monkey Kit and leaving them heated at 40 °C overnight. This temperature was chosen because temperatures above 60 °C could lead to DNA fragmentation and excessive evaporation would lead to precipitation of the DNA.

Unfortunately, I found that eluting with EB has a significant drawback upon evaporation. Evaporating the water from the EB solution causes the DNA to concentrate in the sample, but at the same time causes the salts and other chemicals found in the EB solution to also concentrate. These salts and chemicals at high concentrations induce fragmentation in the DNA sample and as a result drastically worsened the quality of the DNA to an extent that this is no longer HMW and fit for our purposes (Figure 19).

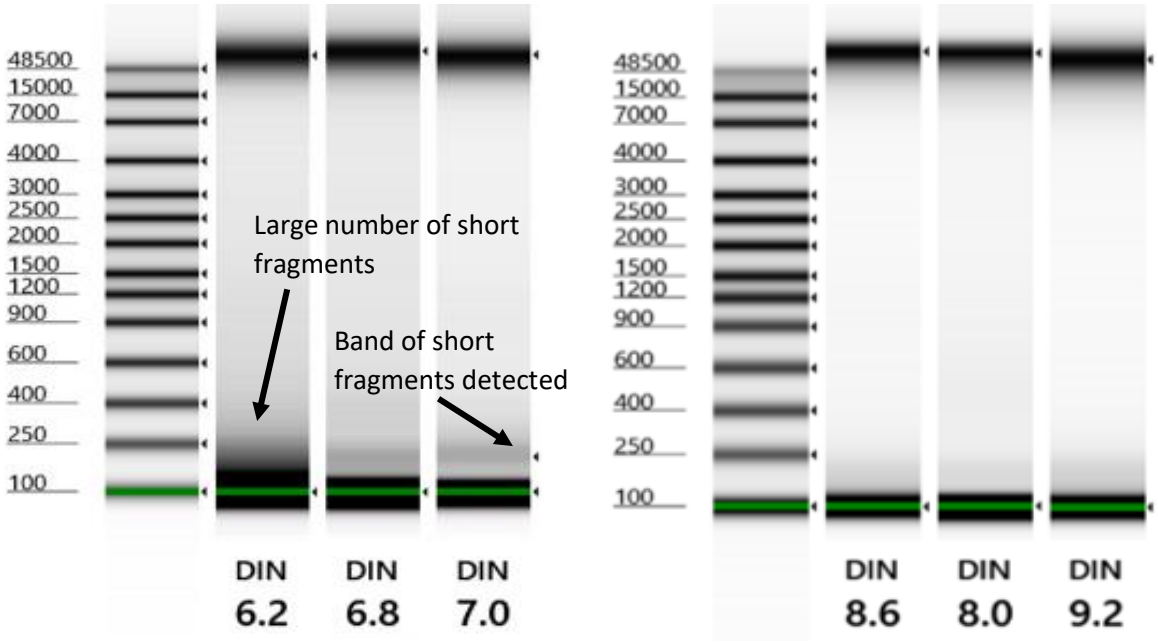


Figure 19 | Example Comparisons between BRD948 WT DNA extractions using evaporation versus conventional extraction.

The Tapestation image on the left is extractions carried out using the evaporation method of concentration compared to the conventional extraction seen in the Tapestation image to the right. The decrease of DIN values seen in the evaporation method demonstrates more shearing and the presence of shorter DNA fragments. These Tapestations images come from runs of two different samples but demonstrate the general result of either approach.

Following a discussion with the kit manufacturer I trialled replacing EB with purified water to elute the DNA from the spin column. This change however worsened the output in terms of quantity as I observed that after evaporation there was white precipitate present in the tubes while Qubit quantitation showed very little change in concentration. What likely occurred was that at high concentrations the solution is saturated with DNA and excess DNA precipitates out of the solution.

Given the significant negative effect on DNA yield, I concluded that we have not yet found a viable alternative to the pre-AMPure step for concentrating the DNA in the solution.

4.3.3.2 Changes in AMPure Bead Ratio

While replacing this step entirely was not possible, other measures could be taken with this step itself to reduce the amount of DNA fragments lost.

In addition to the first AMPure bead step acting as a preconcentrating step it also removes short fragments from the sample which are not desired for the sequencing conducted in this work. This is because the short fragments cannot span across the entire *rrn* operons, making them useless for bridging the DNA either side of said operons to produce complete

circularized genomes. Using a bead-to-sample ratio of 0.6 selectively bound longer fragments of DNA over the shorter ones. The second bead clean-up, which removes the left-over reagents from the tagmentation reaction to ensure the sample is pure for sequencing was also performed at a bead ratio of 0.6.

Through consultation with Ainsworth, E., it was proposed that of the two bead steps, only the first should be used to filter out shorter fragments while the second should aim to retain as much DNA as possible (as the short fragments are already removed by the first bead step). Working from this, the first bead step retained the 0.6X ratio of beads to maintain a filter for removing shorter fragments, but the second bead step instead used a 1X ratio to retain as much DNA as possible while purifying the sample for sequencing.

4.3.4 Changes to Filtering Options in Assembly

An issue found in some samples was that they were unable to be assigned a GS despite being found to have all *rrn* operons present, indicating that all operons are bridged by reads. The problem in these cases was likely due to a large number of reads being lost in the NanoFilt read filtering step that removes reads below 6 kbp in length. This resulted in coverage too low to also assemble the entire genome with the reads provided. By reducing the filter size in such samples, that no longer need a strict length filter due to having all *rrn* operons present at 6 kb, more reads are allowed for assembly, thus increasing coverage.

In samples where these criteria were applied, it was found that changing from a 6 kbp filter to 500 bp greatly reduced the number of contigs in the final assemblies to the extent that these samples could then be either determined by *Socru* or were able to be manually determined as seen in Tables 4 and 5 which display examples of these. Notably, applying this change in general can make some samples assemble more poorly than they would with 6 kbp filtering. This is likely due to too high an input of reads into the assembler used, so this change should only be used under the circumstances described above.

These optimizations were carried out using Typhimurium DT2 isolates, characterisation of which will be described more extensively in Chapter 5.

Sample ID	Post-NanoFilt						Post-Flye	
	Filt. Reads (6 kbp, q>7)	Filt. Read Bases	Mean Read Length	Read Length N50	Theoretical Coverage	Longest Read	Number of Contigs	GS
29	29,633	380,263,097	12,832.4	14,037	79.22	116,292	4	2 2 3 ? 1
40	9,097	98,942,525	10,876.4	11,203	20.61	71,877	31	N/A
41	11,918	126,202,989	10,589.3	10,709	26.29	97,613	17	N/A

Table 4 | Stats of assemblies that failed to determine GS despite having all *rrn* operons sequenced when using a filter of 6 kb

Sample ID	Post-NanoFilt						Post-Flye	
	Filt. Reads (500 bp, q>7)	Filt. Read Bases	Mean Read Length	Read Length N50	Theoretical Coverage	Longest Read	Number of Contigs	GS
29	90,851	531,448,315	5,849.7	10,333	110.72	116,292	3	1 2 3 4 5 6 7
40	65,907	209,514,421	3,178.9	5,580	43.65	71,877	5	1 2 3 4 5 6 7
41	72,883	263,781,206	3,619.2	5,698	54.95	97,613	5	1’ 7’ 3 4 5 6 2’

Table 5 | Stats of assemblies that failed to have a GS determined at 6 kbp reassembled at 500 bp

4.3.6 Full Optimized Protocol for Genome Structure Identification

In summary, the changes applied to the DNA extraction were a pre-extraction step of adjusting the OD_{600} of samples to 0.6 which is around the ideal cell input for the extraction kit. Alongside this was some slight modifications to the bead clean-up to improve DNA fragment retention as well as changes in read filtration for fringe cases in the bioinformatics pipeline.

4.3.7 Evaluation of Optimised Protocol

Back-Of-Bench (BOB) samples were left to grow in the same broth for months at a time which stresses the samples via nutrient depletion which should induce rearrangement. These were grown in various media and incubated at either room temperature or 37 °C.

From the BOB samples, several colonies of varying morphology were observed during monthly plating and picked for arrangement detection. Some of these colonies were noted to be smaller than usual without indication that they were satellites, as they were distant from larger colonies (Figure 20). Alongside these smaller colonies, large colonies were also picked, especially those that looked unusual in terms of appearance (larger than average, difference in colour, etc). RTISL1 was picked due to being the largest colony present and being slightly discoloured compared to the others, while RTISS4 was picked as an extremely small colony that is not a satellite to a larger colony.

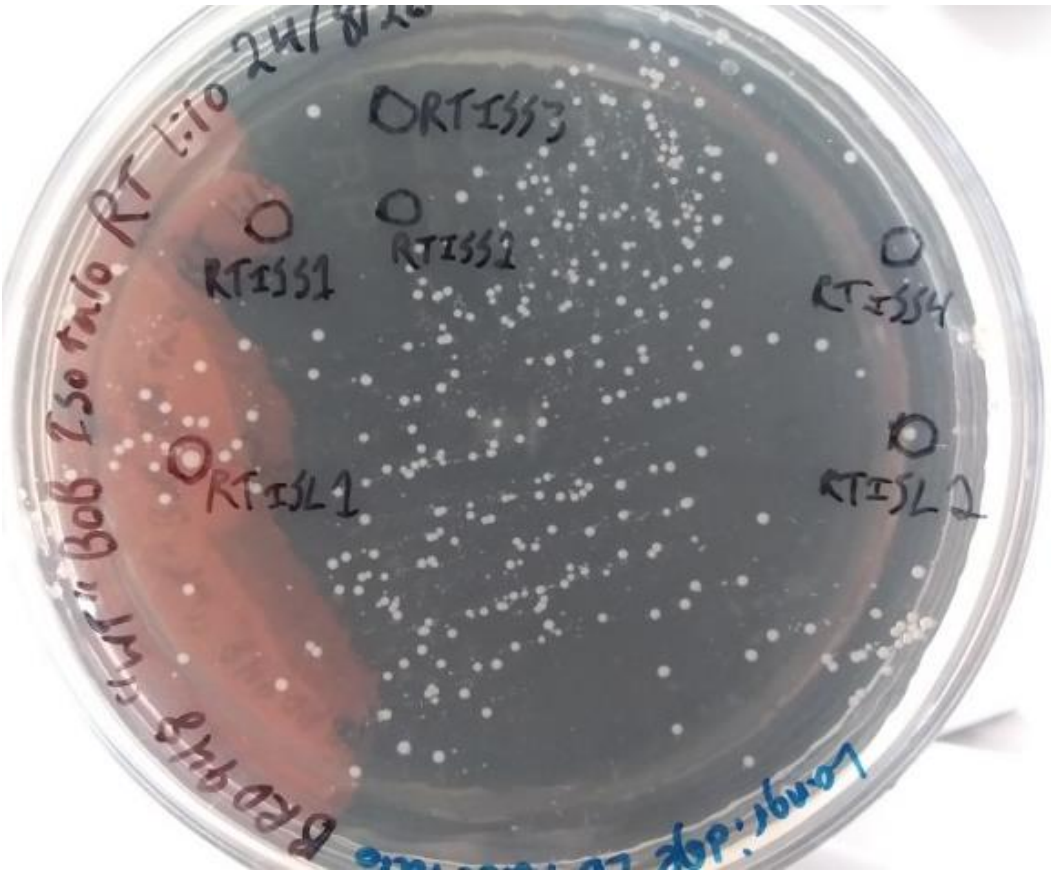


Figure 20 | Plate of origin for LAT1 and LAT2

The plate by which the colonies containing LAT1 (RTISL1) and LAT2 (RTISS4) were picked from, as seen, these were picked as the former appeared unusual (very large with slight discolouration to be more yellow) while the latter was picked for being an extremely small colony (small pale dot in the circle) that was not satelliting a larger colony.

Following the optimised protocol for genome structure identification, I determined that both RTISL1 and RTISS4 had rearranged. The GS of both these colonies was 1' 3 5 6 4 2' 7 (visualised below in Figure 21). Both colonies which have this rearrangement, designated as LAT1 (originating from colony RTISL1) and LAT2 (originating from RTISS4) notably originated from the BOB using Iso-sensitest and grown at room temperature (the plate of which can be seen in Figure 20 above). This rearrangement is notable as the fragments with the origin and

terminus of replication (*oriC* and ter, fragments 1 and 3 respectively) are next to each other rather than on opposing sides. This results in a significant oriC/ter bias within the genome that can be observed phenotypically, as the colonies with this rearrangement are extremely slow growing compared to other rearrangements tested so far.

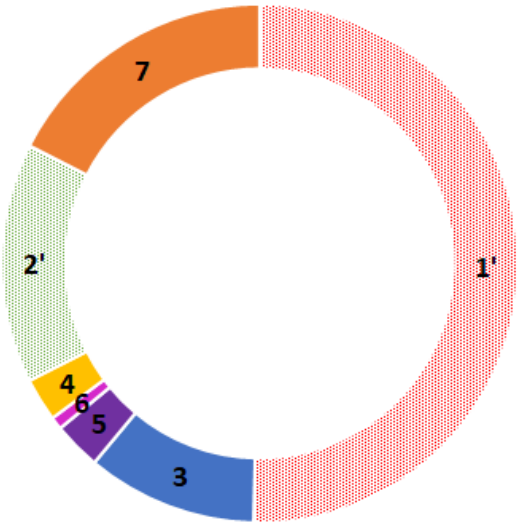


Figure 21 | The arrangement of LAT1 and LAT2 compared to WT (parent).

There is a large amount of rearrangement relative to WT's arrangement of GS 2.66, fragment 7 has excised, inverted and re-inserted between fragments 2 and 1, alongside this fragment 1 has also inverted.

Following sequencing, I observed that there was a contaminant bacterium present in the LAT1 sample which likely explained the differences seen in the initial colonies (larger than usual, slight discolouration), this issue was resolved by streak plating out the LAT1 sample in LB-NaCl+aro to separate colonies of LAT1 and the contaminant; LAT1 was successfully purified after repeating this process a couple of times.

4.4 Discussion

To summarise, I have established several changes to the overall pipeline to improve sample extraction and rearrangement detection speed from a timespan of several weeks to approximately one week, when comparing between long range PCR and this presented method that uses long read sequencing. The scalability of this method is also much greater than long-range PCR going from a single sample at a time to 12 samples per sequencing run, with possible further increases in samples run simultaneously in the future with suggested changes in barcodes used. This new method also has the added benefit of being objective through genome assemblies and fragment sizes, rather than being reliant on the subjectivity of band brightness on a gel.

With this optimised pipeline, the improved scalability allowed me to rapidly sequence the sample collection of Typhimurium DT2s, the process and results of which are described in Chapter 5.

The inducement of a new arrangement demonstrates that the BOB culture method is capable of generating rearrangements and supports previous evidence of inducement using low-salt media. The optimised GS identification protocol was successfully used to rapidly detect the new arrangement in 2 separate colonies. The new arrangement was found to have the same genome arrangement as BRD948 T variant and is more extensively covered in Chapter 6. Considering the extreme oriC/ter imbalance of the rearrangement, this could suggest that the conditions which samples are grown in may encourage the inducement of certain genome structures over others. This potential influence is outside of the scope of this work though it would be something of interest to pursue in future work.

With such an extreme genome rearrangement as LAT2, this presents a valuable opportunity to study how such rearrangements may influence the bacteria beyond replication time. The most promising candidate being the effects of gene expression due to gene dosage effects as described in section 2.4.3. This topic is covered in Chapter 6.

5 Phylogeny of *Salmonella* DT2 and Comparing Genome Assemblers

5.1 Introduction

With an optimized method to extract and sequence DNA for determining GS IDs, I applied this process to an independent collection of samples available for testing. In this case, I had access to a collection of Typhimurium DT2 (Definitive Type 2) isolates which were previously investigated by long-range PCR.

For the DT2 collection, Helm *et al.*, 2004, used long range PCR (Figure 22) and identified these isolates as having GSs either in the conserved or an inverted orientation (Table 6 and Figure 23). As DT2s are confirmed to have rearrangements from this work, this collection provides a case study for rearrangements while also serving as a proof of the functionality for my optimised sequencing method for determining GS.

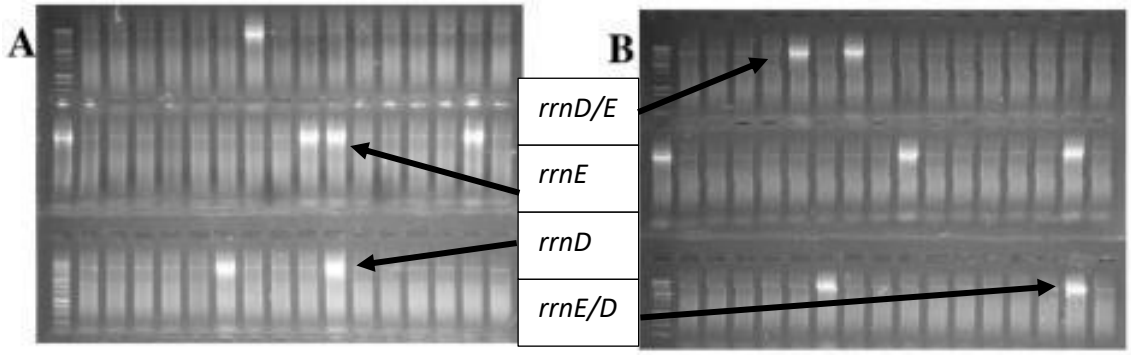


Figure 22 | Long-range PCR for GS Determination

(A) Agarose gel images representing the *rrn* PCR of the conserved arrangement and (B) representing the *rrn* PCR of the inverted arrangement, with two new bands representing *rrnD/E* (top row left) and *rrnE/D* (bottom row far right), image source: Helm *et al.*, 2004.

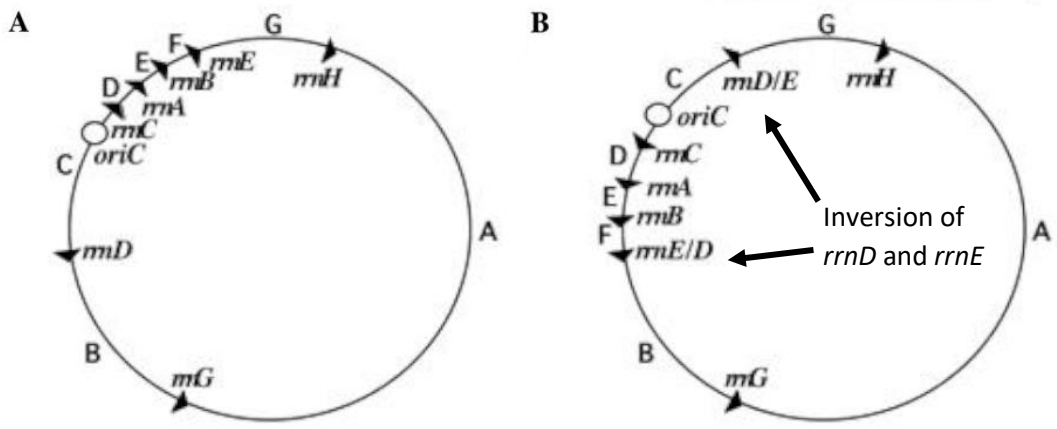


Figure 23 | *rrn* arrangements detected by long range PCR

A) The conserved arrangement and B) the inversion. The inverted *rrn* arrangement seen in B is indicated by the *rrnD/E* and *rrnE/D* (representing an inversion of fragments 1, 2 and 7) products present in long range PCR. Source: Helm *et al.*, 2004.

ID	Strain ID	*rrn* arrangement
10	DT2 99-9034	**Inversion**
11	DT2 99-1870	**Inversion**
12	DT2 99-397	Conserved
13	DT2 00-15	Conserved
14	DT2 00-2141	Conserved
15	DT2 00-4752	Conserved
16	DT2 00-5779	Conserved
17	DT2 01-523	Conserved
18	DT2 01-1025	Conserved
20	DT2 01-6098	Conserved
21	DT2 01-8048	Conserved
22	DT2 01-8664	Conserved
23	DT2 01-8908	Conserved
24	DT2 01-9907	Conserved
25	DT2 02-4155	Conserved
26	DT2 02-4788	Conserved
27	DT2 02-5729	Conserved
29	DT2 03-1253	Conserved
30	DT2 03-2614	Conserved
31	DT2 03-3659	Conserved
32	DT2 97-1797	Conserved
33	DT2 97-5686	Conserved
34	DT2 97-7246	-
35	DT2 97-10215	Conserved
36	DT2 98-652	Conserved
37	DT2 98-3011	Conserved
38	DT2 98-6289	Conserved
39	DT2 00-7941	Conserved
40	DT2 98-12423	Conserved
41	DT2 98-7988	**Inversion**

Table 6 | DT2 Strain Collection. ID34 is not an ID previously sequenced by Helm *et al.*, 2004, and thus its *rrn* arrangement is unknown. These DT2s were originally collected from various regions in Germany between 1997 to 2003.

The inverted arrangement shows differences from the conserved arrangement at two particular points, with the swapping of *rrnD* and *rrnE* resulting in an inversion of fragments C to F. While the DT2s have been identified as either of conserved arrangement or of an inverted arrangement previously in Helm *et al.*, 2004, these have not been translated into GS IDs. When translated to a GS format with fragments A-G renamed to 1-7, the conserved arrangement is GS1.0 and the inverted arrangement is GS 8.67, which has an arrangement of 1’ 7’ 3 4 5 6 2’.

Alongside this, this DT2 collection also provided an opportunity to evaluate the bioinformatics aspect of this work. Due to long read sequencing being new, the tools that work with and assemble these reads are also relatively new and are constantly being updated. The performance and accuracy of these tools are not only variable between the tools themselves,

but also between updates of the same tool, warranting a need to evaluate these to understand which of these is the most capable for the purpose of this work. A more general evaluation of long-read assemblers has been previously undertaken (Wick and Holt, 2021) measured by performance metrics such as structural completeness, accuracy, sequence identity and resources used. The study found that of the assemblers tested, Flye, Raven and Miniasm/Minipolish were overall the best, though no assembler emerged as the ideal choice for long read genome assembly. Building on this type of evaluation, the DT2 collection allowed me to test for specific factors crucial for identifying genome rearrangements. This DT2 collection can be used as a dataset to investigate potential reasons why rearrangement can take place. DT2s are a strain of Typhimurium that is highly associated with disease in pigeons compared to other Typhimurium strains which are instead considered as host generalists (Rabsch *et al.*, 2002). As host-specialized pathogens seem more likely than host-generalist pathogens to rearrange their genomes, there is a possibility that rearrangement may have direct links to bacterial evolution (Langridge *et al.*, 2015).

The aims of this chapter are 1) to demonstrate that the method of extraction and sequencing previously optimised and detailed in chapter 4 can be used as a viable improvement upon the previous method of long-range PCR, 2) to demonstrate that the method can be optimized at the analysis end by evaluating various assemblers to find out which one is most suitable for the purpose of genome structure identification and 3) to determine if there is a link between evolution and genome rearrangement in this DT2 collection.

5.2 Methods

As seen in Table 6, there were a total of 30 DT2 strains studied in this work, with 26 noted to have a conserved arrangement, 3 an inverted arrangement, and 1 of unknown arrangement.

In assembler comparison testing, Nanofilt was kept at 6 kbp length filtering and 500 bp filtering was subsequently used for only specific cases where all *rrn* operons were resolved but required more theoretical coverage to form a complete assembly where GS could be identified. Assemblers, where required, were provided an estimated assembly size of 4.8 Mbp. Each sample was processed through Porechop into Nanofilt (as per 3.7 Bioinformatics workflow for GS Determination) before being assembled by one of the tested assemblers

(Flye (v. 2.5), Canu (v. 1.9) and Raven (v. 1.1.10)), the pipeline of which can be seen in Figure 24.

Long read sequences were originally obtained via HMW DNA extraction via the Fire Monkey DNA Extraction Kit as detailed in 3.5 DNA extraction and library preparation and sequenced as in 3.6 MinION Flow Cell loading and Sequencing. These sequences were processed for the purpose of generating assemblies to detect genome rearrangements via 3.7 Bioinformatics workflow for GS Determination.

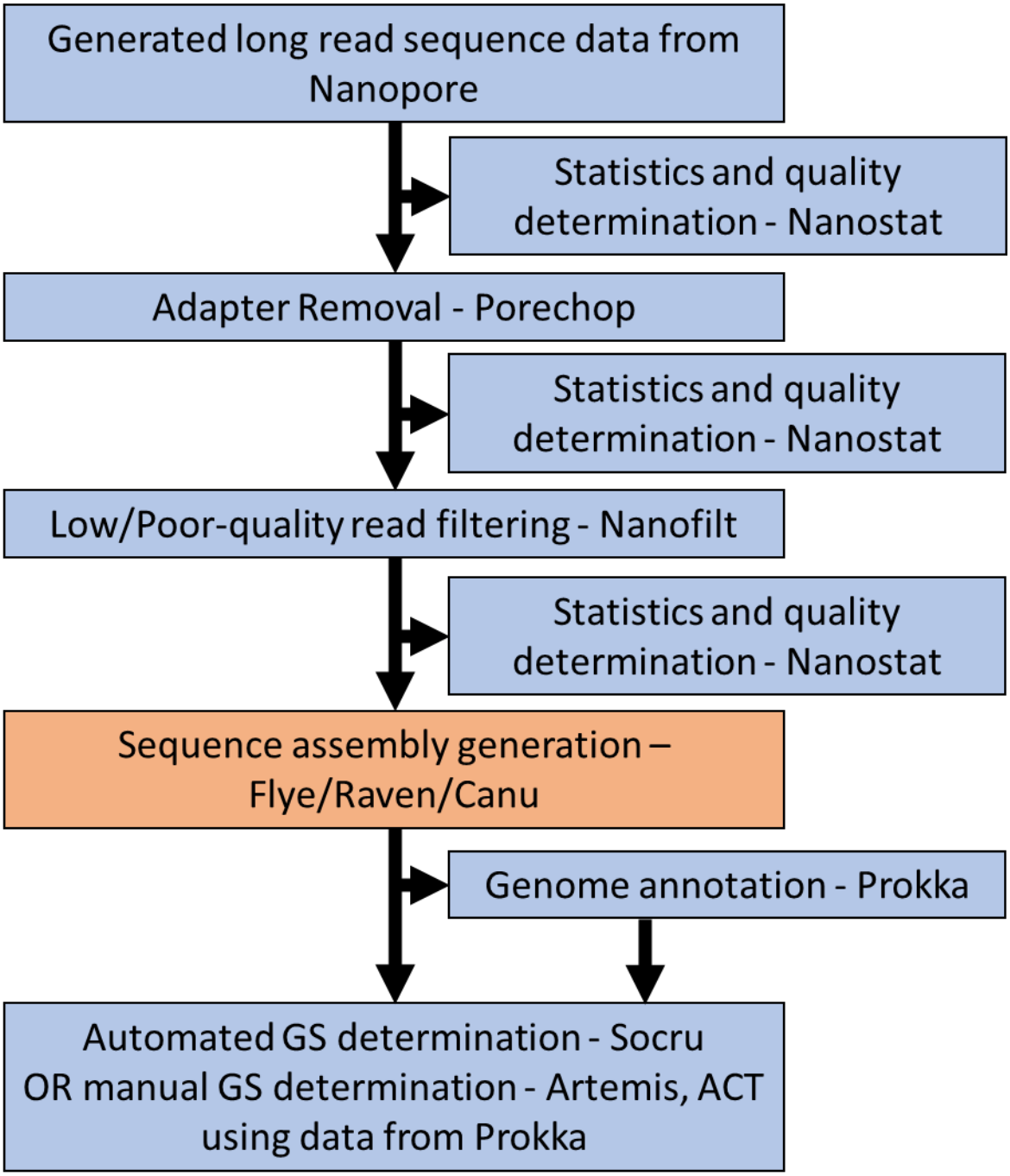


Figure 24 | Pipeline for Assembly Comparison

The pipeline used for genome assembly generation, taking into account the assemblers tested, the section of the pipeline being tested is highlighted in orange.

To investigate a potential link between evolution and genome rearrangement, a phylogenetic tree was reconstructed using short read sequence data of the DT2 collection (provided by Prof. Rob Kingsley) alongside a DT2 reference (NC_022544.1, (Kingsley *et al.*, 2013)). Short read data was used as input through *snippy* (v. 4.4.3) to detect SNPs between these reads and the reference genome which was then combined into a core SNP alignment using *snippy-core* (v. 4.4.3) (Seemann, 2020). This core SNP alignment was then used to construct a maximum

likelihood phylogenetic tree using IQ-TREE (v. 1.6.12) (Nguyen *et al.*, 2014) which was then visualised and annotated using iTOL (v 6) (Letunic and Bork, 2019)(Figure 25), default parameters were used for all of these.

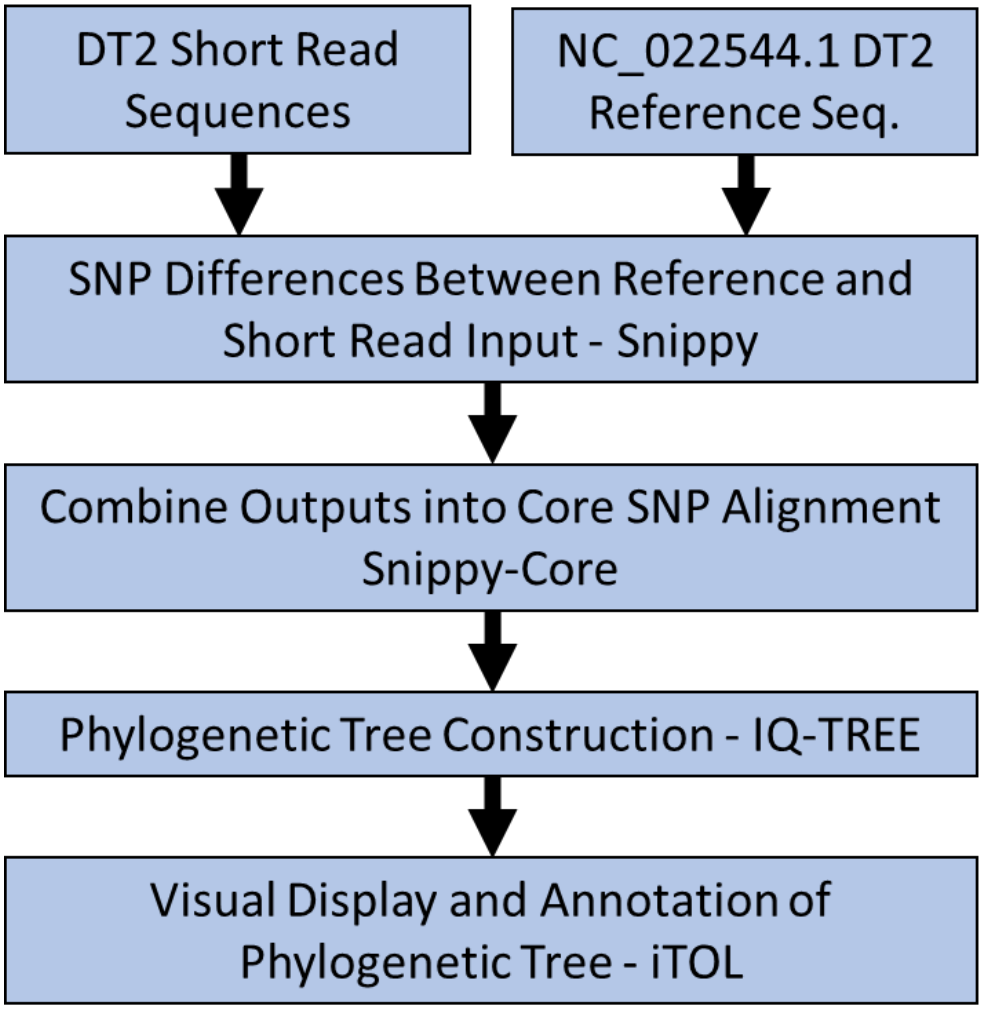


Figure 25 | Pipeline for Phylogenetic Tree Generation

The pipeline used in this work, generating a core SNP alignment that is used to generate a phylogenetic tree

5.3 Results

5.3.1 Sequencing of DT2s and Genome Structure Identification

As stated previously in Chapter 5.1, these DT2s have had their structure identified previously via long range PCR, but not with long read sequencing. Knowing this, I could determine the accuracy of the methods that have been described throughout this work by assembling this collection and comparing the IDs and their arrangements with the data from Helm *et al.*, 2004. Presented in Figure 26 and seen earlier in Figure 23, the genomes marked as "conserved" translate to GS 1.0, with sections A to G being fragments 1 to 7 respectively in these GS figures. This study also identifies a position swap of sections B and G (fragments 2-1-7) which translates to GS 8.67.

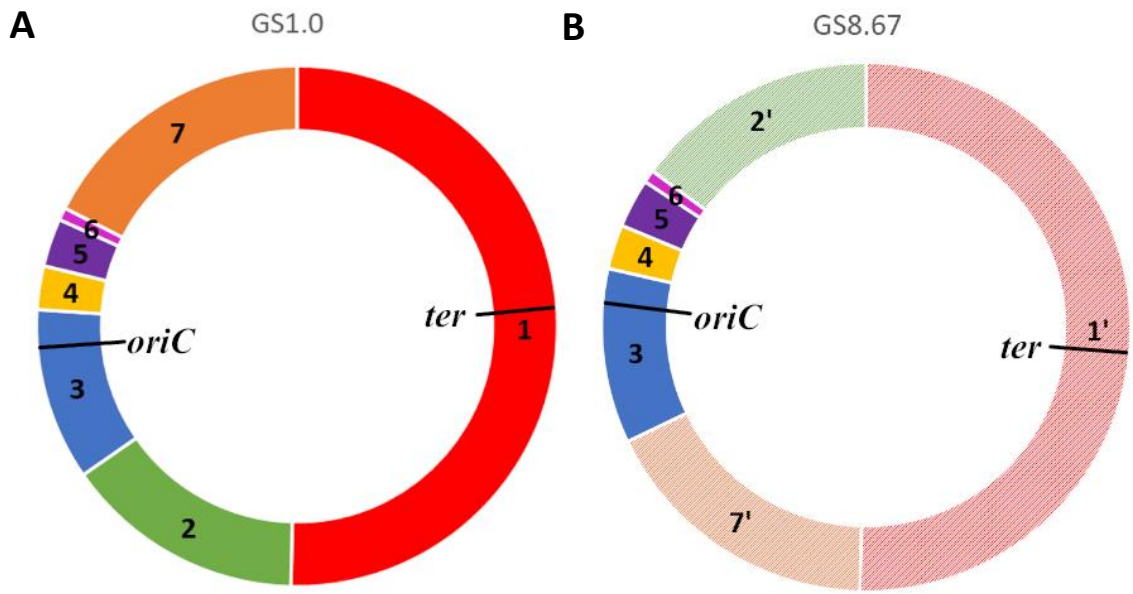


Figure 26 | DT2 arrangements found compared to Helm et al structures.

Images of the genomes found from this DT2 collection from this work (A: GS 1.0, B: GS 8.67).

Thirty DT2 isolates were sequenced across 8 MinION flowcells (repeats included), multiplexed with between 10 and 12 strains per flowcell. Genome coverage ranged from approximately 30x to >200x the expected size of the genome; samples with coverage below 1x were not used.

As seen in Table 7, 18/30 sequenced DT2s were successfully processed through to automatic *Socru* GS determination, and 27/30 strains with the sequencing data I had produced had their arrangement determined with either automatic or manual determination. With this data, I compared these to the samples sequenced in Helm *et al.*, 2004, and confirmed that the arrangements detected by this method were the same.

Sample ID	Number of Contigs	Number of rRNAs	Determined with?	Fragment order	Predicted GS
10	2 (1 is plasmid)	22	Socru	1' 7' 3 4 5 6 2'	GS 8.67
11	2 (1 is plasmid)	22	Socru	1' 7' 3 4 5 6 2'	GS 8.67
12	5 (1 is plasmid)	22	Socru	1 2 3 4 5 6 7	GS 1.0
13	2 (1 is plasmid)	22	Socru	1 2 3 4 5 6 7	GS 1.0
14	6	22	Prokka	1 2 3 4 5 6 7	GS 1.0
15	2 (1 is plasmid)	22	Socru	1 2 3 4 5 6 7	GS 1.0
16	9	22	Prokka	1 2 3 4 5 6 7	GS 1.0
17	2 (1 is plasmid)	22	Socru	1 2 3 4 5 6 7	GS 1.0
18	4	22	Prokka	1 2 3 4 5 6 7	GS 1.0
20	3	22	Socru	1 2 3 4 5 6 7	GS 1.0
21	5	22	Prokka	1 2 3 4 5 6 7	GS 1.0
22	3	22	Prokka	1 2 3 4 5 6 7	GS 1.0
23	2 (1 is plasmid)	22	Socru	1 2 3 4 5 6 7	GS 1.0
24	3 (1 is plasmid)	22	Socru	1 2 3 4 5 6 7	GS 1.0
25	2 (1 is plasmid)	22	Socru	1 2 3 4 5 6 7	GS 1.0
26	2 (1 is plasmid)	22	Socru	1 2 3 4 5 6 7	GS 1.0
27	2 (1 is plasmid)	22	Socru	1 2 3 4 5 6 7	GS 1.0
29	4 (1 is plasmid)	22	Prokka	1 2 3 4 5 6 7	GS 1.0
30	6 (1 is plasmid)	22	Socru	1 2 3 4 5 6 7	GS 1.0
31	3 (1 is plasmid)	22	Socru	1 2 3 4 5 6 7	GS 1.0
32	12 (1 is plasmid)	22	N/A	N/A	N/A
33	3 (1 is plasmid)	22	Prokka	1 2 3 4 5 6 7	GS 1.0
34	2 (1 is plasmid)	22	Socru	1 2 3 4 5 6 7	GS 1.0
35	26 (1 is plasmid)	16	N/A	N/A	N/A
36	2 (1 is plasmid)	22	Socru	1 2 3 4 5 6 7	GS 1.0
37	4 (2 are plasmids)	23	Socru	1 2 3 4 5 6 7	GS 1.0
38	13 (1 is plasmid)	22	N/A	N/A	N/A
39	4 (1 is plasmid)	22	Prokka	1 2 3 4 5 6 7	GS 1.0
40	5 (1 is plasmid)	22	Socru	1 2 3 4 5 6 7	GS 1.0
41	5 (2 are plasmids)	22	Prokka	1' 7' 3 4 5 6 2'	GS 8.67

Table 7 | GS Results for the DT2 Collection

The perfect scenario for determination of GS from assemblies is a single chromosomal contig (the entire bacterial chromosome is assembled as one) though outside of this situation *Socru* can still automatically determine a GS if all 22 *rrn* genes are present on the largest genomic

contig (22 rather than 21, as one 5S gene has been duplicated in *Salmonella*). If *Socru* is unable to automatically determine arrangement, then with few enough contigs the arrangement can be manually determined by taking the contigs and *rrn* operon positions provided by Prokka to piece together the arrangement against known fragment sizes.

ID 37 (DT2 98-3011) interestingly has 23 *rrn* operons instead of the expected 22, this is because in this DT2 there is another duplicated 5S *rrn* gene. Considering the chromosome in this ID is in two pieces, it is likely this duplication is an artifact of non-complete assembly, with the chromosome split at an *rrn* operon.

Some samples required a second extraction and resequencing due to the first run of sequencing producing too few reads to generate assemblies of reasonable contig number. A critical statistic for these assemblies is coverage, which estimates how many times the reads fed into the assembler would cover across the entire genome by taking the total basepairs of all the reads and dividing this by the approximate size of the genome (in this case 4.8 Mbp). In certain cases such as IDs 29, 40 and 41, to increase coverage, I reduced read filtering in Nanofilt from 6 kbp minimum to 500 bp as described in 4.3.4 Changes to Filtering Options in Assembly. Due to time constraints three DT2s (32, 35 and 38) did not have their GSs determined due to poor first assemblies and no laboratory time available to redo these.

5.3.2 Assembler Comparison

With the extraction of HMW DNA and generation of long read sequences optimized, one critical part of the process that had not been tested for potential optimization was the assembler used to take the cleaned and filtered reads to generate an assembly that could then be processed by Prokka and *Socru*. By finding the best assembler for this method I could be confident that I was getting the best assemblies from the data provided.

As mentioned earlier, Wick and Holt, 2020, have previously tested a variety of long-read assemblers for whole genome sequencing but focusing more on computational resources used, circularisation rate, and structural accuracy and completeness. In this work I focus primarily on structural completeness at a variety of genome coverages. This places an emphasis on the ability to generate a low number of contigs despite a low number of reads.

With a proven method and a large collection of data that could be tested, one possible avenue of methods optimization was comparing assemblers used. So far in this work only the

assembler Flye was used but in this chapter I describe testing Canu and Raven as well. These assemblers were chosen as these are the most actively updated long-read assemblers used and are thus the most up-to-date in terms of performance. These were also straightforward to set up on the Galaxy platform these assemblies were carried out on.

The varying quality of sequencing data and coverage seen across the DT2 set also allowed insight into how these assemblers perform at different degrees of genome coverage, allowing for a wider range of performance. This also allowed me to see how these assemblers perform with extremely poor/extremely low coverage samples.

What I was looking for ideally in an assembler for the purpose of this work is genome completeness and robustness at varying degrees of coverage. An ideal assembler should primarily be able to resolve sequences, to complete genomes and to also assemble plasmids alongside this. Said assembler should also be able to resolve assemblies within a reasonable amount of time to account for future upscaling. The results I obtained from Flye, Canu and Raven on sequence data from the 30 DT2 samples are shown in Table 8.

	Input for assembly				Raven (1.1.10) (Vaser and Šikić, 2021)			Canu (1.9) (Koren *et al.*, 2017)			Flye (2.5) (Kolmogorov *et al.*, 2019)		
ID	Filt. Reads (6kb, q>7)	Mean Read Length (bp)	Bases (bp)	Theo. Cov.	# Contigs	Genome size (bp)	# rrn operons	# Contigs	Genome size (bp)	# rrn operons	# Contigs	Genome size (bp)	# rrn operons
10	100,314	14455.3	1450073023	302	2	4914124	22	2	4988222	22	2	4912359	22
11	5454	7665.9	41809936	8.71	108	2984003	16	NA	NA	NA	37	4875648	14
12	6425	11805.2	75848472	15.8	47	3835646	18	42	4548041	22	14	4932926	21
13	63645	12214.1	777368,46	162	1	4819105	22	2	4996120	22	2	4911645	22
14	22454	10105.0	226898729	47.3	11	4863306	22	11	4814868	22	2	4913373	22
15	50605	13512.0	683775275	142	2	4913409	22	5	4964042	22	2	5006460	22
16	7	9959.4	69716	0.0145	NA	NA	NA	NA	NA	NA	NA	NA	NA
17	57162	13932.3	796396507	166	2	4912129	22	3	4940483	22	2	4911787	22
18	10	10452.8	104528	0.0218	NA	NA	NA	NA	NA	NA	NA	NA	NA
20	813	294.8	239666	0.0499	NA	NA	NA	NA	NA	NA	NA	NA	NA
21	3	10533.7	31661	0.00660	NA	NA	NA	NA	NA	NA	NA	NA	NA
22	4	8399.8	33599	0.00700	NA	NA	NA	NA	NA	NA	NA	NA	NA
23	49893	13458.3	671473941	140	2	4956408	25	2	4926172	22	2	4912125	22
24	50929	12832.8	653,562932	136	1	4819639	22	3	4927159	22	2	4911637	22
25	71173	14025.6	998240,831	208	2	4913569	22	3	4919461	22	2	4912385	22
26	47734	14011.1	668807952	139	2	4913028	22	2	4959338	22	2	4912989	22
27	53532	12650.6	677212021	141	2	4913434	22	8	5057334	22	2	4912242	22
29	29633	12832.4	380263097	79.2	3	4907681	22	10	4879610	22	2	4911541	22
30	26059	11615.8	302696343	63.1	5	4906599	22	8	4905996	22	2	4911914	22
31	16499	10675.5	176134449	36.7	18	4807780	22	24	4772100	22	6	4912842	22
32	6245	10511.0	65640949	13.7	68	3725759	16	51	4373911	19	21	4804711	19
33	15117	10726.8	162156427	33.8	21	4744164	22	31	4777025	22	5	4913154	22
34	32168	11825.8	380412987	79.3	2	4909223	22	5	4909196	22	2	4911169	22
35	4006	11221.0	44951406	9.36	50	2760846	14	NA	NA	NA	34	4967625	16
36	23178	12018.8	278571856	58.0	3	4902327	22	5	4918684	22	2	4912209	22
38	5142	11349.4	58358422	12.2	55	3486494	18	52	3892942	17	19	4815384	22
39	15628	10937.2	170926358	35.6	14	4865277	22	19	4815704	22	4	4913876	22
40	9097	10876.4	98942525	20.6	47	4479427	21	24	4675181	22	8	4915109	22
41	11918	10589.3	126202989	26.0	28	4780513	22	11	4895910	22	6	4908753	22

Table 8 | Assembler Performance in Generating Complete Contigs Against Varying Coverage

It should be noted that some results seen in Table 5 are different in quality and thus ability to generate contigs compared to the final DT2 assemblies described in Table 4. Quality in this case includes total read count, mean read length, and coverage, all of which affect the quality of the final assembly. The sequencing runs used to test the assemblers were from single runs whereas in the final assemblies, multiple runs were combined where appropriate.

Generally, Canu struggled to assemble samples that Flye and Raven otherwise were able to assemble, which can be seen in IDs 29 and 15. Canu also struggled the most with low coverage. Seen in IDs 14, 29, 30, 31 and 39, Flye shows extremely good performance with samples with coverage of approximately 70x or less when compared to Raven and Canu. In ID 30 (63x coverage) Flye produces 2 contigs (1 is a plasmid) while Raven produces 5 and Canu produces 8. In ID 36 (58x coverage) Flye once again produces 2 contigs (1 plasmid), while Raven produces 3 and Canu produces 5. Seen in both ID 15 (142x coverage) and ID 27 (141x coverage), both Flye and Raven produce the complete chromosomal contig and plasmid, whereas Canu produces 8 contigs.

As 70x is the lowest value considered to be good coverage, for this work consistently good performance at this coverage and below is of note. Only at less than approximately 35x coverage does Flye fail to produce assemblies that at minimum could be manually determined.

Raven sits between Canu and Flye in terms of ability to generate complete contigs, though it was interestingly the most likely to give the incorrect number of contigs and was also the most prone to fail to give a genome of the anticipated length.

One notable positive for Raven outside of data produced is being able to handle dataset collections properly via Galaxy, which makes it extremely easy to input extremely large collections simultaneously. However, despite Raven being able to complete assemblies rapidly, it produces a non-compressed output file which requires compression before it can be accepted by *Socru*. This means an additional step needs to be added to the pipeline to accommodate this in Galaxy (Afgan *et al.*, 2018).

These results indicate that for the purpose of this work, Flye was the most suitable in terms of complete genome assembly. This was confirmed with the samples below recommended

coverage (70x) having generally far less contigs produced by Flye compared to Raven and Canu. While Flye produced complete chromosomal contigs at as low as 50x and manually determinable assemblies as low as approximately 30x, Raven would sometimes fail to produce complete contigs at 70x coverage and Canu would fail at producing complete contigs consistently at coverage below 100x.

Flye was also better at assembling plasmids (samples with plasmids are shown in Table 7) compared to Raven, as seen in IDs 13 and 24, alongside complete chromosomal contigs, while Raven did not assemble the plasmid for these IDs; Canu also assembled these plasmids. In terms of runtime, Flye was between Canu and Raven in terms of time taken with Canu being the slowest and Raven the fastest.

5.3.3 DT2 Phylogeny

With this DT2 collection, I could start to look at the potential role genome rearrangement has on *Salmonella*. One possibility is that rearrangement is an evolution-linked process, with these large structural changes to the genome being a form of variation like that seen with SNPs, insertions and deletions. If there is an evolutionary link rearrangements may be fixed and passed to following generations to suit the environments they survive in. Alternatively, this process takes could take place independently in clades.

To investigate this, the short read data provided with these samples was combined with the above rearrangement data to produce a phylogenetic tree with these inversions highlighted. If rearrangement is evolutionarily linked these rearrangements would be within the same clade.

Alongside the DT2s used earlier short read sequencing data provided for the rest of the DT2 collection provided was also used for this. These extra strains are referred to as DT2MKs (the GSs of these are reported as GS 1.0).

As seen in Figure 27, the three DT2s with rearrangements are not all within the same clade, indicating that genome rearrangement is not directly linked to evolutionary changes and instead likely occurs independently of this. While it is likely that rearrangement is another means of adaptation for the purpose of survival in constantly changing environments, it is not entirely linked to evolution which shares a similar purpose in surviving in various environments.

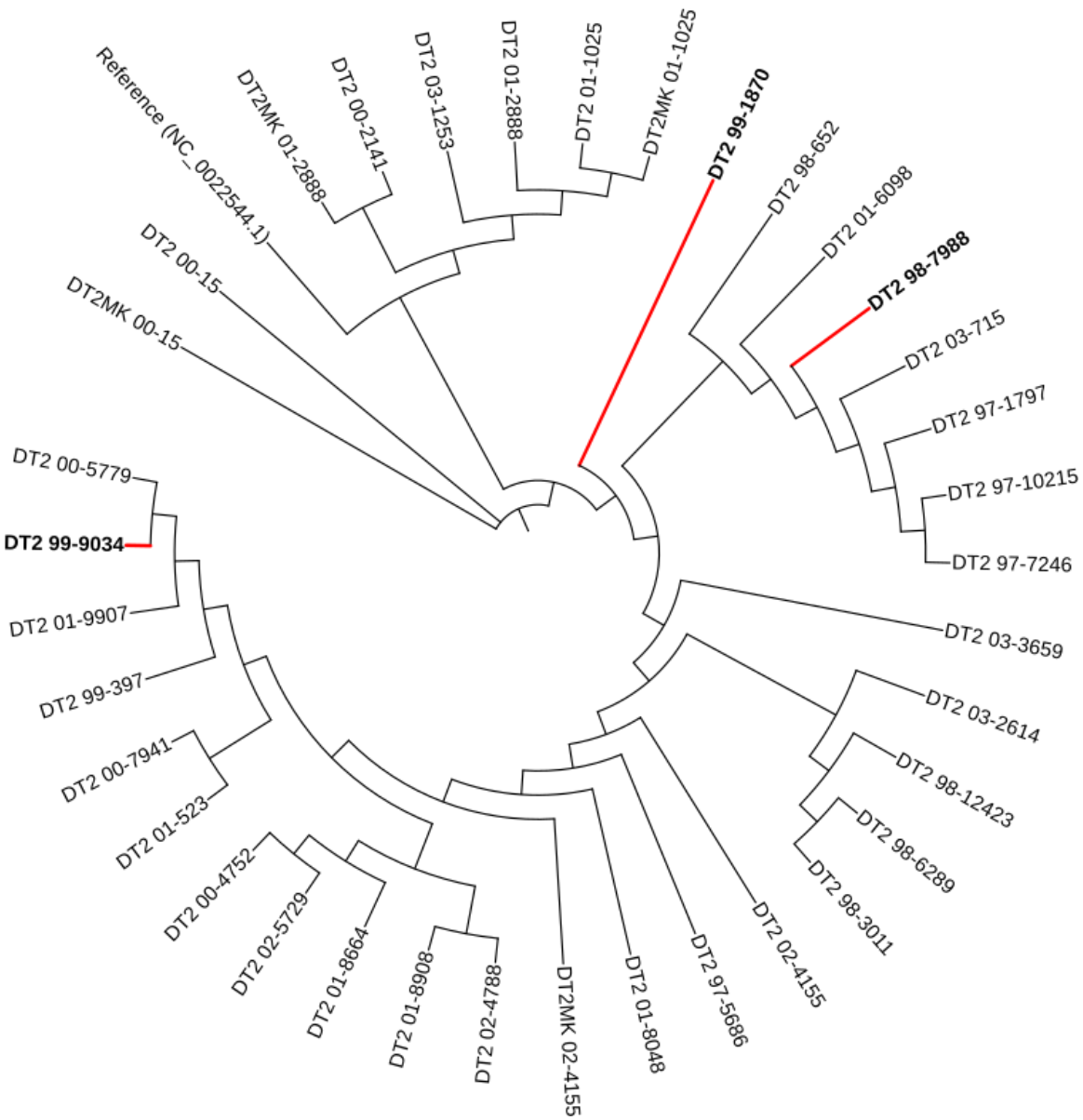


Figure 27 | Phylogenetic Tree of the DT2 Collection

The phylogenetic tree generated from the DT2 collection alongside the DT2 reference used, highlighted in red and in bold are the DT2 samples with a GS of 8.67 instead of GS 1.0. The DT2MK samples were additional samples from the DT2 collection that had previously been sequenced via long-read sequencing and thus were not included in the previous sections.

5.4 Discussion

The 30 GSs identified here and shown in Table 4 demonstrate that the 3 samples sequenced in this work that had rearrangements had “inversions” as described by Helm *et al.*, 2004. As the genome structure predicted by Helm *et al.*, 2004, matches the ones described here this demonstrates that this sequence-based method can reliably detect genome rearrangements.

With Flye being the assembler currently in use for this work, I can conclude that Flye still remains the main choice that I will be using for future assemblies using this complete methodology primarily due to better performance at low read depths. A possible reason why Flye outperforms the other assemblers tested may be due to taking a different approach to assembly. While Canu and Raven make use of OLC graph assembly, Flye first produces disjointigs from combining reads where repetitive sequences are collapsed into a repeat graph that is resolved to make the final contigs. This difference in approach and options that suit uneven coverage depth may be the reasons why this assembler performed better.

The issue with the findings with the assembler data is that these assemblers are constantly being updated, one version of an assembler may have vastly different outcomes compared to another version. In the future such tests can also be expanded from 30 real samples to a much larger set on top of simulated sets using Assembly Dereplicator (Wick and Holt, 2019) and Badread (Wick, 2019).

Consequently, updates to these assemblers can significantly alter their performance particularly in the context of this work. These updates highlight how rapidly the field of bioinformatics can change compared to other fields; programs may rapidly depreciate or become outdated within a year or even a matter of months if not well-maintained. The data presented in this work serves more as a snapshot into the current assemblers provided at the time these measurements were taken, with assemblers being continuously updated or even replaced with newer ones, these conclusions will eventually be no longer accurate. Caution should also be taken for more specific requirements such as the ones detailed in this thesis as updates could also behave unpredictably in terms of results. Thus, updates to this software should be tested against a set of known GSs to demonstrate that performance is not negatively affected. The outcome of this work not only demonstrates that long-term cultures can produce rearrangements but also that the optimised pipeline allows GSs to be determined accurately and reliably.

A suggestion for future work evaluating bioinformatics tools is the concept of “living papers”; papers which are regularly looked back upon and evaluated again to maintain accuracy with the current state of the field. This would allow future researchers to make

an accurate and informed decision with the options they have available to them and what they should pick for their own uses.

Genome rearrangement has been shown with this collection to be independent of evolution. If rearrangement has an influence on gene expression this could act as another mechanism for expression to be altered alongside other changes such as acetylation/methylation, and more permanent changes such as SNPs, insertions and deletions.

Since there is no evidence of an evolutionary link with genome rearrangement, this presents questions to consider outside of this such as what factors or genes drive rearrangement, and how might these be induced and tested within the laboratory? I can now also look at the influence genome rearrangement has on gene expression, which will be explored in the following chapter.

6 Impact of Genome Rearrangement Upon Gene Expression

6.1 Introduction

With the aspects of the method described earlier in this work both optimized and demonstrated to function, I moved on to testing a new genome arrangement for its impact on functions of the bacteria.

As previously mentioned in 2.4.3, there are already known impacts caused by rearrangements. One of these is reduced growth rate directly caused by skewing in the *ori-ter* balance of the genome, as the skewing increases, the maximum length of the replichores also increase. Furthermore, there is a known impact, at the gene level, in gene expression in the form of gene dosage directly based on the distance from *oriC* due to how bacterial replication results in more replicates of genes closer to *oriC* compared to those further away.

There has previously been research in genome rearrangement in clinical strains of *Salmonella* Typhi with indication of rearrangement in long term carriage as well as strains of the same rearrangement passing from a carrier to an acute case, showing the potential for lineage tracing via genome rearrangement (Ainsworth *et al.*, 2021 (Unpublished)-b). Changes to arrangements in long-term carriers suggest a potential role for genome rearrangement in the ability of Typhi to survive and be shed for extremely long periods. Observing the impact to gene expression in rearrangements presents a critical step in determining why these bacteria rearrange and understanding the potential advantages and disadvantages this provides.

The aims of this chapter are to 1) induce and identify new arrangements in long-term growth cultures, 2) conduct RNASeq to measure differential expression of genes and observe the changes rearrangements have on a variety of processes, and 3) determine what context these differentially expressed genes have on the wider function of the bacteria.

6.2 Methods

Rearrangements were induced using the methods previously described in 3.3 Bacterial growth and long-term growth cultures setup. Colonies picked from these long-term cultures were selected using the methods described in 3.5 and 4.3.6 in order to extract high

molecular weight DNA. These were then sequenced using the pipeline described in 3.7 and Chapter 6 to identify if picked colonies had undergone rearrangement. These were normally tested monthly, however, there was a period between 4 - 10 months where rearrangement might have occurred (i.e. between March to August 2020) where I could not conduct monthly testing due to the national lockdown.

The induced rearrangement, LAT2, was tested using RNASeq by the methods described in 3.11 and the sequences obtained were used in the pipeline provided in 3.12, the output of which was used to generate tables of differential gene expression in Excel.

Data from these tables were overlaid onto a Pathway/Genome Database of Typhi (Kingsley *et al.*, 2018) to identify which metabolic pathways were affected. The genome sequence of Typhi Ty2 (parent strain of the Typhi strains in this thesis; accession AE014613.1) was used as the reference for the genome sequences for WT and LAT2 alongside the differential gene expression data to produce figures using BRIG (v. 0.95) (Alikhan *et al.*, 2011).

As described earlier in 3.12.1, *snippy* and Breseq are used to compare WT and LAT2 against each other and both against Ty2 to detect SNPs and other mutations (such as additions or deletions) in the samples.

The overall pipeline from initial colonies to GS discovery and then RNA sequencing and finally, differential gene expression determination, can be seen in Figure 28.

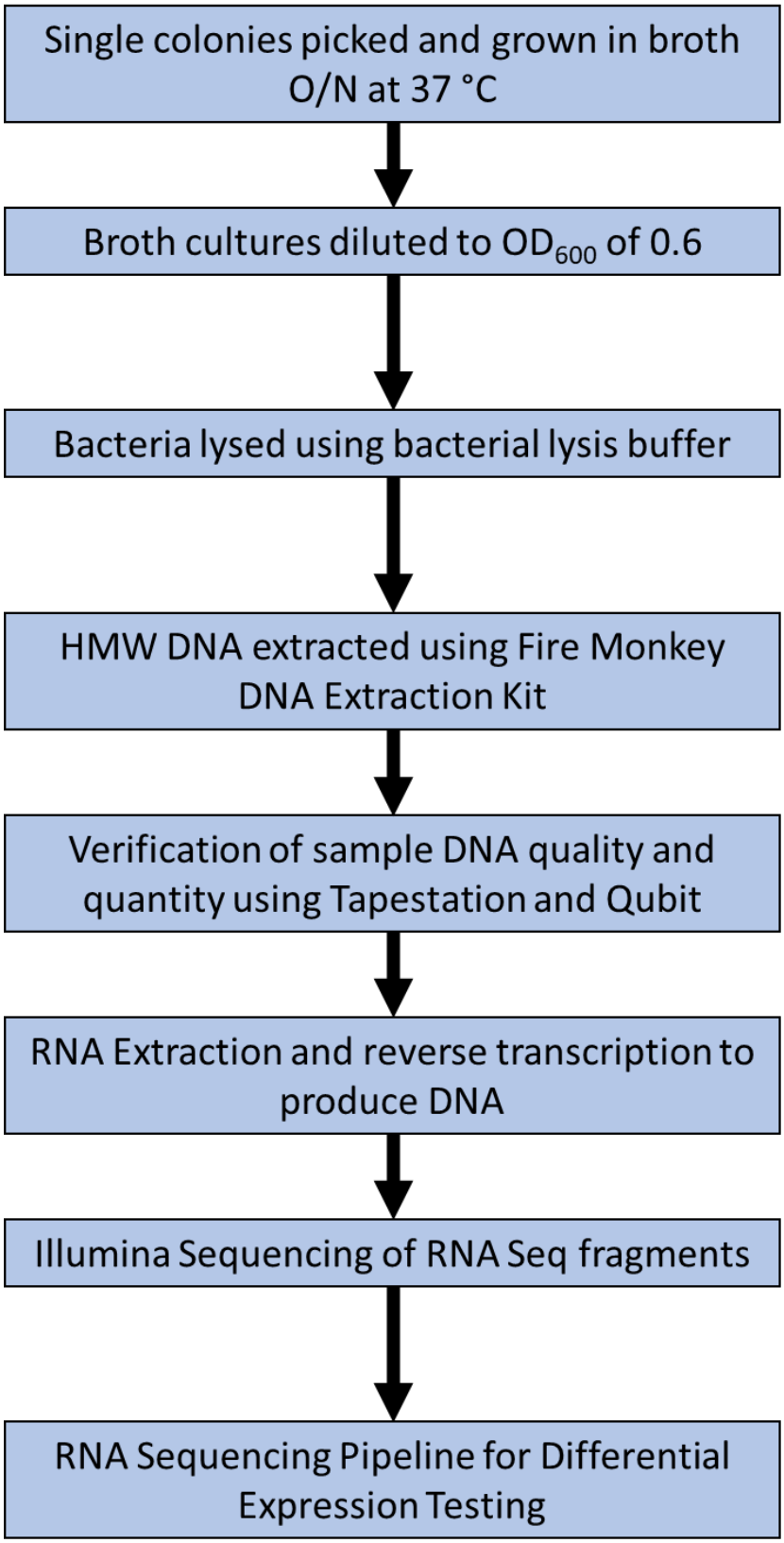


Figure 28 | Overall Pipeline from Initial Colonies to Differential Gene Expression Testing

A summarised pipeline for the entire process of GS discovery and determination of differential gene expression.

6.3 Results

6.3.1 Choice of Strain for Further Investigation

Upon checking for SNPs using *snippy*, LAT2 had no SNP difference compared to WT whereas LAT1 had one. To investigate the impact of variation cause by genome rearrangement alone, I decided to move forward with RNA Sequencing of LAT2, as there should be no impact on gene expression in this strain from SNPs.

6.3.2 Growth Rates for RNA Extraction

RNA extraction required a similar growth curve experiment to determine cell input as done for DNA extraction (Chapter 4.3.3) but for different reasons. Firstly, RNA extraction requires cells to be harvested in the early exponential growth phase for differential gene expression analysis. This is because 1) the exponential phase is where gene expression will be at its highest as replication is taking place at a rapid rate, and 2) the noise present in samples is as low as possible, as the further along in the growth curve the sample is, the more cells within it will start to deviate in terms of expression which generates noise. Furthermore, a balance has to be struck with enough cell input to extract sufficient RNA, but not too much as an input of more than 20 ug cDNA could also result in DNA contamination in the RNA eluate produced by the kit used, negatively affecting results produced.

In Figure 29, WT shows a typical growth curve with an early log phase that moves into a typical log phase, eventually reaching the start of stationary phase at 0.6 OD_{600}.

Unusually, as seen in Figure 29, there is a noticeable and sustained drop in growth rate in LAT2 rearrangement strain that is not seen in WT between approximately 0.2 and 0.28 OD_{600}. Since the growth rate actually increases again after this point, it can be ruled out that this is a transition into the stationary phase, which can be seen at the end of the curve for WT's growth beyond 0.6 OD_{600}.

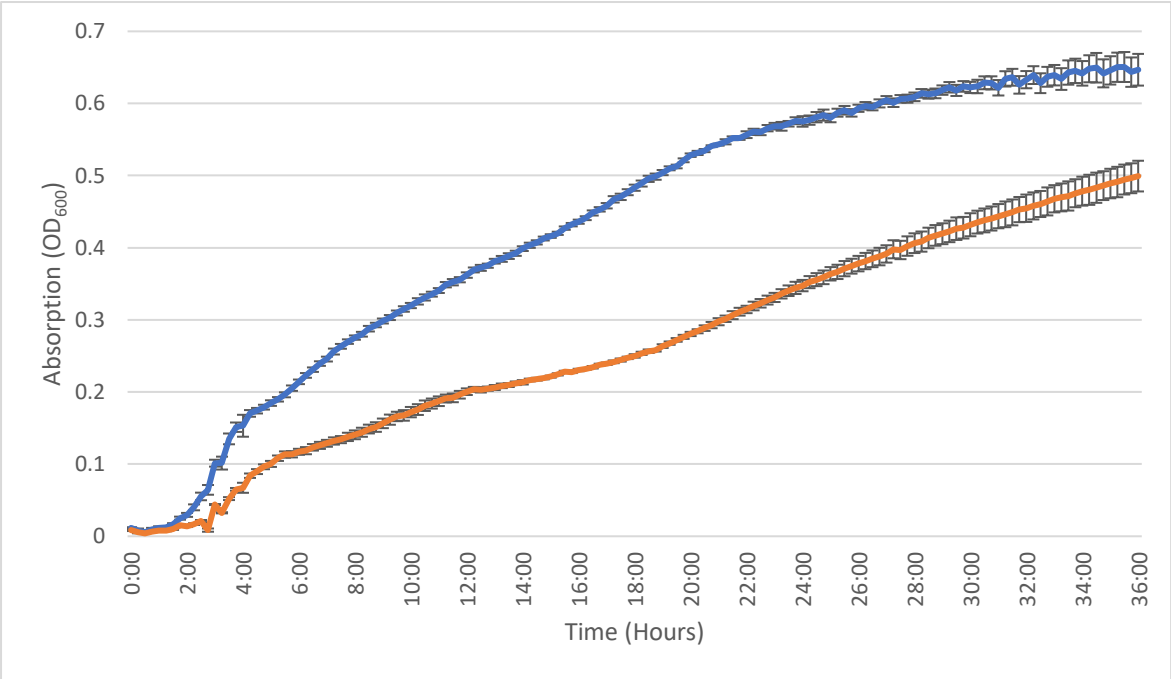


Figure 29 | Growth rates for BRD948 rearrangements WT and LAT2 in LB-NaCl+aro

The error bars present are produced from the standard error using the triplicate biological repeats conducted during this experiment. The blue line represents the average growth of WT and the orange line represents the average growth of LAT2.

From this growth curve, it appeared that conducting RNA extractions at a range of 0.3 to 0.35 OD_{600} would be suitable with the volumes used for RNA extraction, as this was above the region of poor growth and extracting below this would have poor results due to low cell quantity. This was supported by RNA extractions which I trialled at lower OD_{600}, as they typically had extremely poor quantities (Table 9), in agreement with previous literature showing that growth rate generally translates to transcription rate (Klumpp *et al.*, 2009). At 0.2-0.25 OD_{600} the first WT extract and all three LAT2 extracts were below the detectable range, meaning all four had a concentration less than 1 ng mL^{-1}. In comparison, samples extracted at an OD_{600} of ~0.3 to 0.35 all had detectable RNA which was of sufficient quantity for sequencing. More volume can be used instead to extract from the same number of cells at lower OD_{600}, in the future this would be carried out to extract at the early exponential phase seen below 0.2 OD_{600} to detect any differential gene expression that takes place during that period.

RNA quality was assessed using the Tapestation, where the main aspect being looked at was the RIN (RNA Integrity Number) which indicates how fragmented the RNA is based on signal strengths from 5S, 16S and 23S and regions such as the inter and fast regions.

Sample	Concentration of extractions between 0.2 to 0.25 OD (ng mL^{-1})	Concentration of extractions between 0.3 to 0.35 OD (ng mL^{-1})
WT 1	Too Low	14.2
WT 2	12.8	9.01
WT 3	3.99	8.92
LAT2 1	Too Low	5.48
LAT2 2	Too Low	9.46
LAT2 3	Too Low	4.40

Table 9 | Concentration of WT and LAT2 RNA Extractions at varying OD_{600}.

An example of this is seen in Figure 30 where weaker bands relative to the rest of the lane (such as in B1 and D1) give worse RINs than those with low noise (seen in E1 and G1). This high noise is indicative of degradation. All of this confirmed what was observed from the growth curve: that extracting between 0.2 to 0.25 OD is not viable for LAT2, as replication seems to noticeably slow down in this region.

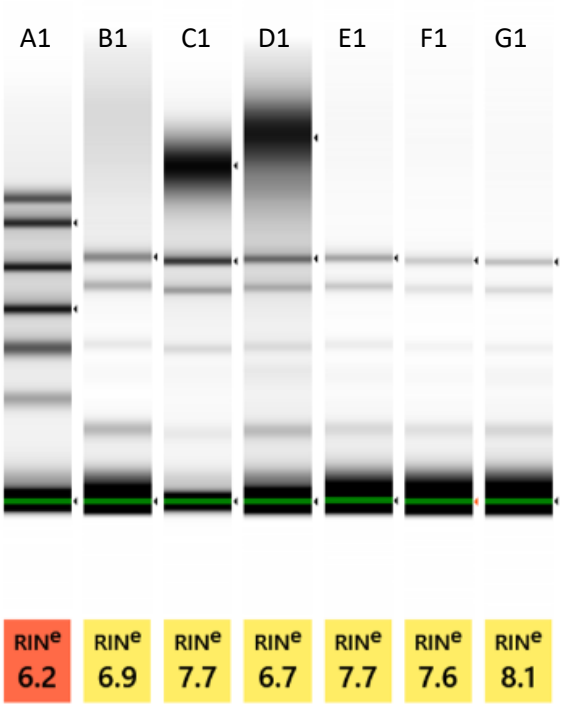


Figure 30 | Example Gel Electrophoresis of WT and LAT2 RNA extracts.

A1 is the RNA ladder, while B1 to D1 are the WT extracts and E1 to G1 are the LAT2 extracts. RIN represents the overall quality of peaks, generally a higher RIN represents a more intact sample.

6.3.3 *oriC* Repositioning and Genome Level Impact

A genome rearrangement was induced in the long-term cultures, after approximately 10 months; it was October 2019 when I first set up the cultures and August 2020 when the

rearranged colonies were identified by plating. This induced rearrangement time is longer than expected as it was theorised it would take approximately 3 to 4 months to occur.

Two different samples (also referred to in 4.3.7 Evaluation of Optimised Protocol) from the monthly tests of the long-term cultures were discovered to be harbouring genome rearrangements. Through DNA sequencing one was discovered to have a contaminant bacterium present, but was eventually purified, and the other was completely pure. These samples were named LAT1 and LAT2, respectively. They were originally picked due to their unusual size compared to typical colonies (LAT1/RTISL1 being much larger than normal, LAT2/RTISS4 being much smaller than normal) which can be seen in Figure 20.

Both LAT1 and LAT2 when compared to WT showed a very unusual configuration for their genome, resulting from fragments 1 and 7 flipping as a pair, leaving fragments 1 and 3 next to each other on the genome. This causes a significant ori-ter bias by a skew of 57.6°, which can be seen in Figure 21.

The radical change in genome arrangement presents an opportunity to see how the repositioning of fragments relative to *oriC* has an impact on genome-wide gene expression.

In theory this is the second most extreme skew possible, the only more extreme skew would have fragment 3 also inverted which would place *oriC* and *ter* as close to each other as possible.

This skew in LAT2 is predictive of a significant negative effect upon growth rate due to this skew producing one extremely long replichore (from fragment 3 into 5, 6, 4, 7 and then into 1). This rearrangement also means a large number of genes have moved either closer to or further away from *oriC*, which could impact their expression due to gene dosage effects.

Through combining the results of both DESeq2 and locating genes with differential expression on the Typhi genome, I have determined that as these fragments move around the genome, the genes located on said fragments have also moved. Seen in Figure 31 an MA-plot (which displays the log 2-fold change (M) against q-value of all the genes sequenced (A)), shows that there are a large number of genes (highlighted in red) that show a significant change in expression between the two strains. Using BRIG, I was then able to

generate a plot comparing both LAT2 and WT while also displaying genes with significant changes in expression and their relative positions (Figure 32).

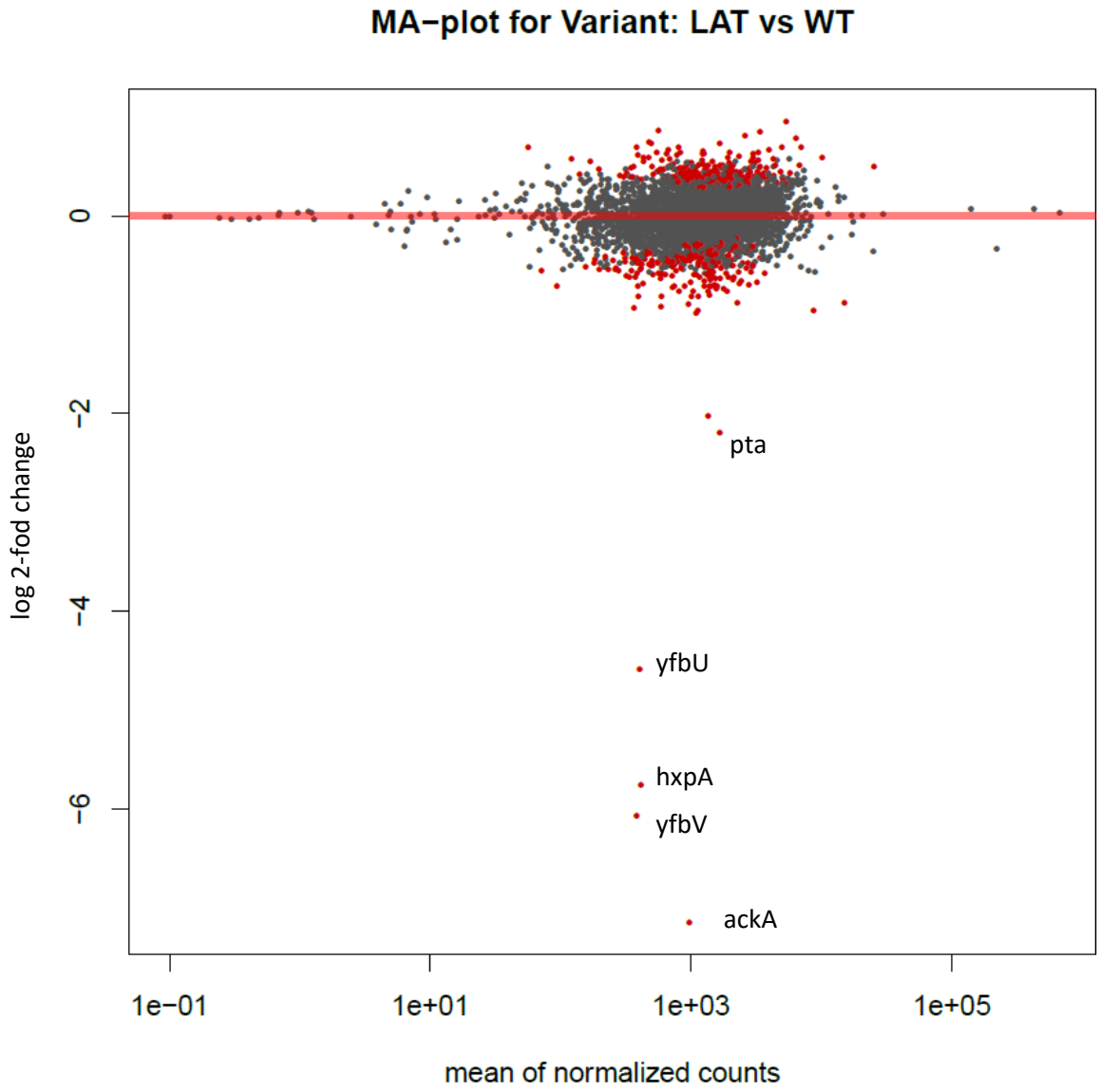


Figure 31 | Plot of Log 2-Fold Gene Expression between WT and LAT2.

MA plot of the two compared strains (WT and LAT2) that displays genes of potential significance by colouring genes red if their *q*-value is less than 0.1

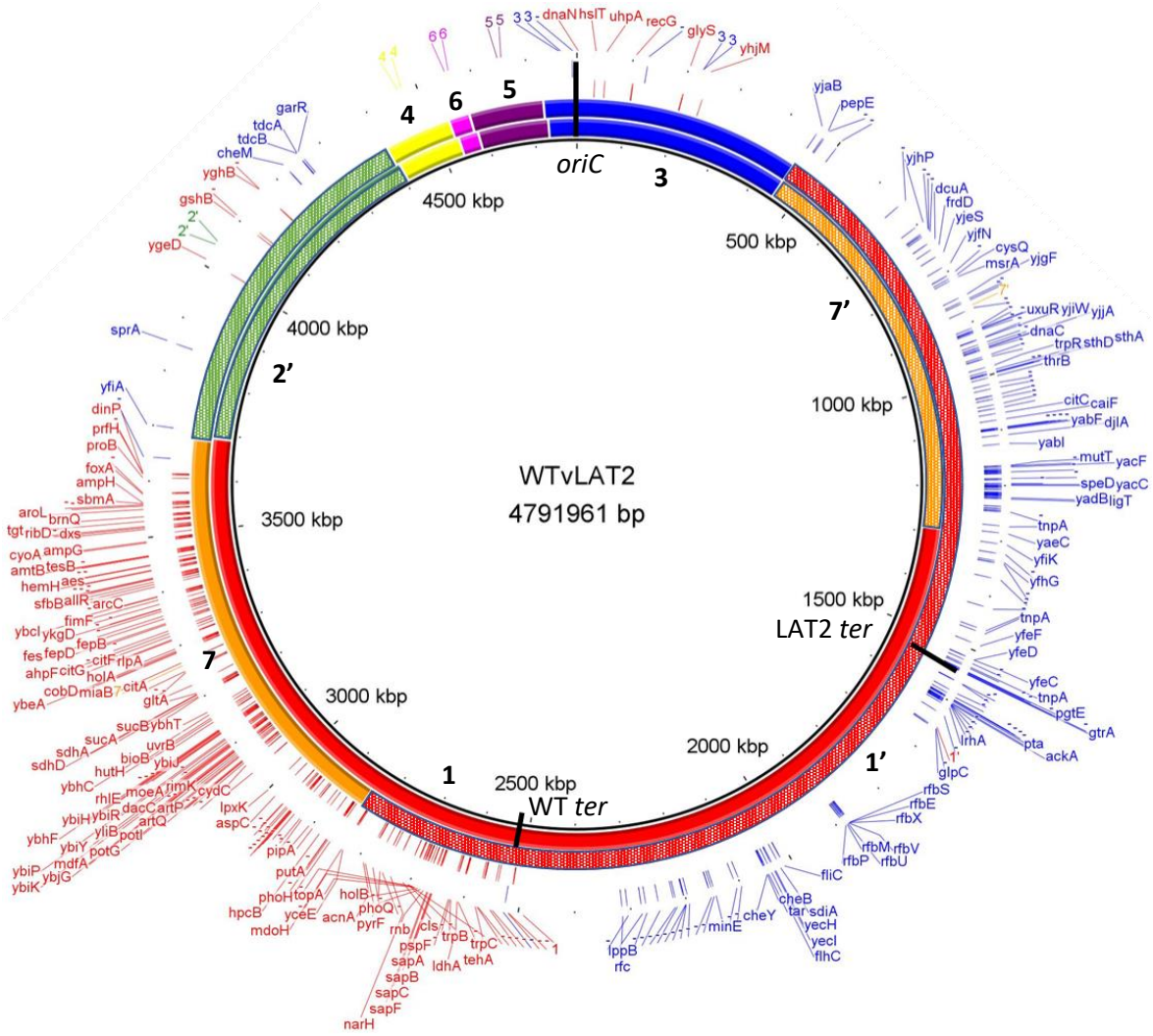


Figure 32 | The Genome of LAT2 against WT and the Overall Impact on Gene Expression

WT (inner ring) and LAT2 (outer ring) arranged against Ty2 (black innermost ring) starting from *oriC*, genes in blue are up regulated in WT compared to LAT2, whereas genes in red are down regulated. The genes shown are against their original positions in WT. Cut-offs are q-value < 0.05.

This overall indicates that the movement of fragments in this rearrangement has caused a large number of genes to experience a significant change in expression. By visualising this data it is shown that fragment 7 (orange) shows downregulation of genes across the entire fragment in LAT2 due to being much further away from *oriC*. Due to inversion of fragment 1 part of this fragment is also further away and also displays this. In contrast the other half of fragment 1 that is now closer to *oriC* displays an upregulation in gene expression. These differences indicate that through the rearrangement of fragments across the genome their distance from *oriC* has a clear impact on gene expression.

6.3.4 Differential Expression of Genes

One of the first significant changes I observed was a cluster of 6 genes with an extreme downregulated change in expression in LAT2 (log2FC > 2 in WT against LAT2, t0526-t0531). These genes consisted of a putative phosphatase, putative sodium/sulphate transporter, two hypothetical proteins and the genes *pta* (phosphate acetyltransferase) and *ackA* (acetate kinase).

To investigate the possibility that larger genomic deletions had occurred than could be detected by *snippy*, I processed the genome of LAT2 through Breseq and found that there was a single small section of the genome (6,239 bp in length) deleted in LAT2 when compared to WT. Upon visual conformation with Artemis Comparison Tool (ACT) (Carver *et al.*, 2005) I found that the section missing is where these 6 genes are normally located within WT, explaining the complete loss of expression of these genes. The outmost genes of this segment (*pta* and the putative transporter) are partially intact however the loss of content was so large that these genes have been rendered non-functional.

In this cluster, two particular genes are of note: *ackA* and *pta*. Both genes are needed as part of the phosphoenolpyruvate pathway (Figure 33) and both flank the reversible pathways that lead to the production of acetylphosphate (AcP). Since both of these genes are missing in LAT2 I can speculate that this deletion would prevent production of AcP in samples of LAT2, which has been described in the literature to act as a global signalling molecule in bacteria (covered in more detail in Section 6.4).

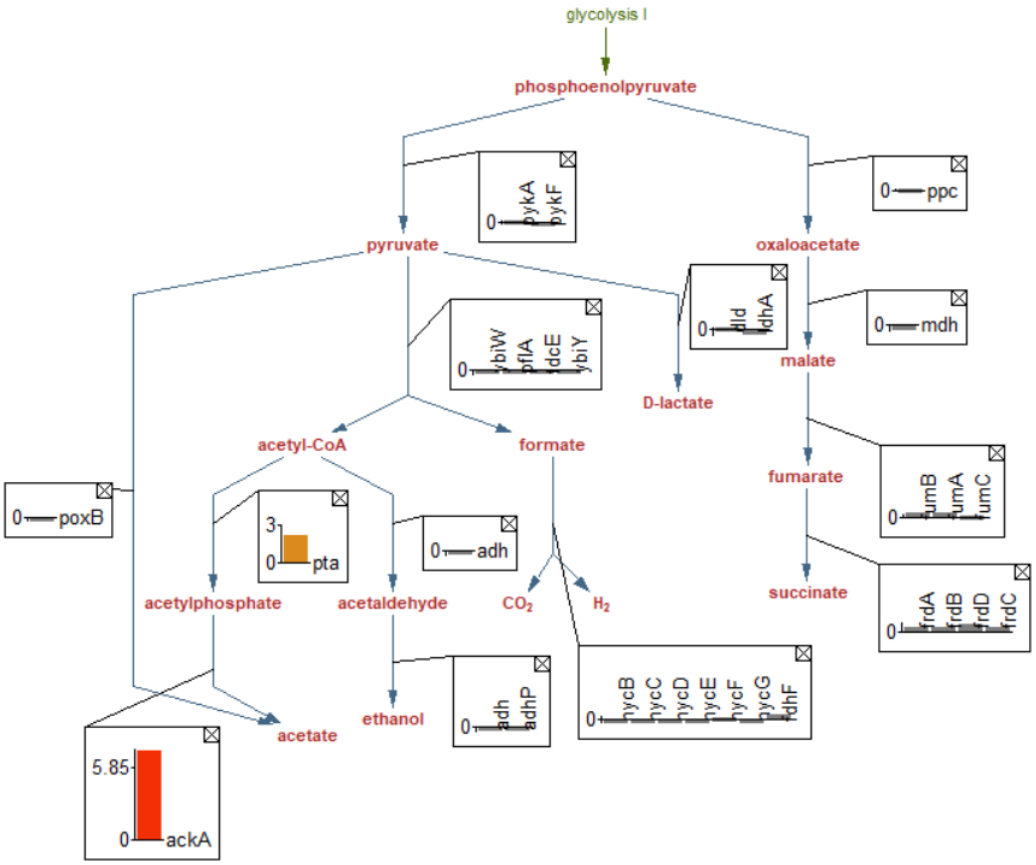


Figure 33 | Role of *ackA* and *pta* in *Salmonella* Metabolism

The role in metabolism that the genes *ackA* and *pta* have, indicated in the log 2-fold change in expression in WT compared against LAT2 (*q*-value < 0.1). The reactions *ackA* and *pta* are involved in are both reversible reactions. All other genes in these pathways showed no differential expression between WT and LAT2.

Outside of this gene cluster, there were a total of 340 other genes with a significant difference in expression (as indicated by *q*-value < 0.05) between LAT2 and WT which is seen in Figure 33 earlier. Due to the large number of genes, I focused upon genes with a log 2-fold change of ±0.58 as well as a q-adjusted < 0.05, as this represents a fold change of 1.5 in expression (Kilpinen *et al.*, 2008, Zhao *et al.*, 2018), which would indicate genes with larger degrees of expression change. From this, 83 genes (including the 6 in the missing section) were found (Appendix 1) and can be seen in Figure 34 and from here a few clusters were observed, such as *rfb* genes (*rfbS*, *rfbE*, *rfbX*), *sdh* (*sdhA, sdhD*) genes and *cyo* (*cyoA, cyoB, cyoC, cyoE*) genes, which will be looked at in more detail in Section 6.4.

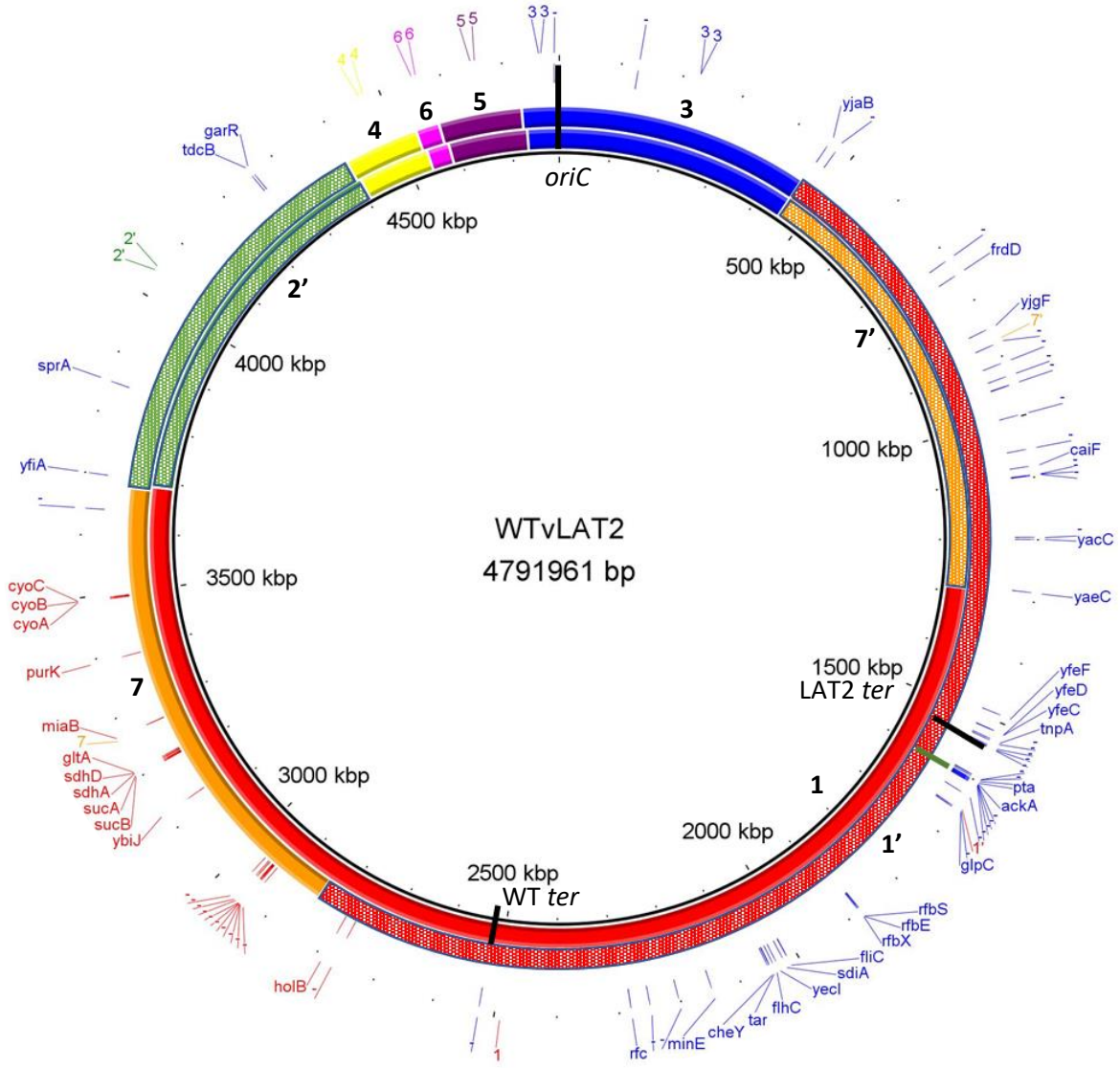


Figure 34 | Log 2-Fold and *q*-significant Genes in Expression Between LAT2 and WT

Log 2-fold (> ±0.58) and *q*-significant (< 0.05) genes in expression between LAT2 and WT. WT (inner ring) and LAT2 (outer ring) arranged against Ty2 (black innermost ring), genes in blue are more highly expressed in WT compared to LAT2, whereas genes in red have reduced expression in WT. The genes shown are against their positions in WT. The green line indicates the region of deletion between t0526-t0531.

Overall, this indicates that genome rearrangements not only have a wider impact on the genome but also have notable effects on gene clusters throughout the genome.

6.4 Discussion

The loss of a 6kb genome segment was identified in the rearranged genome of LAT2, which included *pta* and *ackA*, these two genes being required for the production of AcP.

AcP serves as the high-energy intermediate of the Pta-AckA pathway and serves as a common intermediate in a number of metabolic processes, either resulting in energy production through acetate or conversion back to Acetyl-CoA that can produce energy through the ethanol pathway through alcohol dehydrogenase. AcP is described in bacteria as a global signalling molecule, particularly as a switch for the transition into the stationary phase of bacterial growth (Wolfe, 2005), facilitating slower growth from glucose to acetate scavenging. Most notably, previous experiments that have studied both *ackA* and *pta/ackA* knockout mutants showed reduced growth rates particularly under anaerobic conditions (Ren *et al.*, 2019). As AcP is used to move bacteria from exponential growth into the stationary phase, the inability to produce AcP in LAT2 may be the reason why this strain shows the unusual growth curve seen in Figure 29, as this could affect the ability to transition properly between stages of the bacterial growth curve.

The increased presence of AcP or acetate in mutants likely results in acetylation of proteins as part of post-translational modification which affects their function, this acetylation also has implications in AMR expression in *Salmonella* (Li *et al.*, 2018); future work could investigate this in LAT2.

I have considered that these bacteria might be discarding segments of their genome to try and reduce the *ori-ter* bias. While the deleted ~6 kbp segment is on the longer replichore of the LAT2 genome (Figure 34), the loss of this segment only amounts to 0.13% of the bacterial chromosome (or 0.468°). This therefore provides only a very small adjustment. Beyond this missing gene cluster in LAT2 there are a number of other genes showing significant differential gene expression which are likely affected purely by the rearrangement of genome fragments. Of particular interest are clusters of genes in which most or all of the cluster are simultaneously significantly upregulated or downregulated.

The *rfb* cluster (*rfbS, rfbX, rfbE*) shows significantly reduced expression in LAT2, this cluster partially encodes for the O-antigen in *Salmonella*, and variance in this gene cluster is responsible for O-antigen variation in *Salmonella* (Fitzgerald *et al.*, 2003). This antigen has

a role in pathogenesis through modulation of the antigenic properties of the bacterial surface, as well as improved adhesion and survival under stress (Lerouge and Vanderleyden, 2002, Marshall and Gunn, 2015). These three genes in particular are key for O-antigen production and presentation with *rfbX* transporting the antigen with the *rfbS* and *rfbE* genes catalysing the last two intracellular reactions of the process.

Interestingly, despite the reduced expression of this cluster, *fliC* is also significantly reduced. This is despite O-antigen deficient mutants typically producing exclusively type 1 flagellin (which *fliC* is responsible for producing). *fliC* is also critical in biofilm formation and attachment in *Salmonella*. Flagella also modulate multicellular behaviour in *Salmonella* which could further influence viability and behaviour in biofilms as this functionality would be disrupted in this rearrangement (Römling and Rohde, 1999, Crawford *et al.*, 2010, Wang *et al.*, 2020).

The two genes *sdhA* and *sdhD* appear to have a role in inducing gut inflammation by enabling use of succinate produced by members of the gut microbiota for the TCA cycle, which is also enabled by taking electron acceptors induced by gut inflammation (Spiga *et al.*, 2017). The increased expression of these genes in LAT2 suggests an enhanced ability to induce localised gastrointestinal infection by this rearrangement (if the sample this originated from was capable of infection) (Bowden *et al.*, 2010). This is further supported by the enhanced expression of *sucA* and *sucB* which is used to produce succinyl-CoA, also within the TCA cycle. This is further supported by previous findings displaying the role of central metabolic pathways such as the TCA cycle in *Salmonella* virulence (Bumann, 2009). While this hasn't been explored in Typhi specifically, there is evidence in Typhimurium that mutations deleting genes involved in the succinyl-CoA to succinate and succinate to fumarate pathways results in attenuation (Tchawa Yimga *et al.*, 2006).

The *cyo* gene cluster is involved in the respiratory electron transfer chain and as such can be classified as involved in central metabolism similar to the *sdh* genes and *pta*/*ackA*. Previous literature indicates that these genes experience changes in expression in environments of high stress such as in water (for Typhi) or egg whites (for Enteritidis). This suggests that this arrangement may be under stress but this also seems to be a way in which *Salmonella* are able to survive in these conditions (Baron *et al.*, 2017, Kingsley *et al.*, 2018).

Knowing that Typhi infection and carriage both require survival in low-nutrient environments, the latter is also reliant on biofilm formation. Genes of note such as the *rfb* cluster, *cyo* gene cluster, and *sdhA* and *sdhD* in the LAT2 arrangement are crucial for a number of processes involved in pathogenesis, particularly biofilm formation. Since biofilm production on gallstones is prevalent in carriage, and increased ability to form biofilms is linked to carriage rate (Devaraj *et al.*, 2021) and protection from antimicrobials (González *et al.*, 2018), characterising the ability of various arrangements to produce biofilms is another avenue for future investigation. These experiments could be carried out by conducting motility experiments as well as growing samples in high-cholesterol media, or in broth containing solid surfaces intended to induce biofilms (Merritt *et al.*, 2005, Wijesinghe *et al.*, 2019).

7 Final Discussion

7.1 Implementation of Long-read Sequencing Techniques for Detecting Genome Rearrangement

Through this Masters project, I have curated a means of taking samples of bacteria, inducing rearrangements and then both detecting said rearrangements and finding phenotypic changes that they cause.

Previously, the main method used for discovering genome rearrangements in bacteria was long range PCR which. While functional it had issues that made seeking a more robust alternative desirable. Long-read sequencing has become rapidly prominent over recent years and gradual improvements to the technology have made it viable for use in genome arrangement determination thanks to producing reads sufficiently long to bridge the *rrn* operons of bacterial genomes. Despite this, the relatively short period over which long-read sequencing has become more widely available has meant that various aspects of the process required optimization to improve output and quality.

This thesis has detailed various elements of the process for generating HMW DNA fragments and subsequently producing high quality long-read sequence data, and my efforts to improve these. I implemented a spectrophotometry and dilution step at the beginning to control cell input and to ensure that there was no input overload which would negatively affect DNA quantity and quality. I also made changes in the second of two bead-based steps used to improve the quantity of HMW DNA retained for sequencing, without going against the original intentions of this step (concentration, purification, and removal of shorter fragments).

For the bioinformatics aspect of this method, I also made improvements and conducted tests for confirmation. I found that in certain cases of low coverage, reducing the minimum length filter improved assemblies through reducing the number of contigs to a point where arrangements could be determined. With long-read sequencing being a recent development in research, there are a variety of assemblers available that are being regularly updated. Unlike general purpose sequencing, determining genome rearrangements has an emphasis upon generating complete assemblies and thus this was the main factor considered when comparing these assemblers. By testing the assemblers

against sets of real reads, I confirmed that the assembler Flye was best able to generate complete assemblies, enabling GS determination.

I therefore now have an established method that is a viable alternative to long-range PCR, that avoids the subjectivity of long-range PCR as a result of using DNA assembly instead of bands on an agarose gel. My method also has a much higher throughput as up to 12 samples can be sequenced within a single run, whereas long-range PCR requires samples to be studied one at a time.

Despite this, there are still some potential improvements to be made to this method. Currently the research group is trialling changing the barcode tags used, to increase the number of samples that can be processed in a single run from 12 to 48 which reduce the overall cost per sample (Ainsworth *et al.*, 2021 (Unpublished)-b). This change may present a risk of reducing the average coverage of samples; based on my data, a minimum coverage of 70x should be aimed for, per sample. With samples in general achieving coverage ranging between 80x and 140x, means that a run of up to 24 samples would be ideal as possible for this method. Oxford Nanopore have also recently introduced 96-plex barcode kits and potentially this number will increase later on. Future improvements to the durability of the pores used by the MinION sequencer that means that they lose functionality at a slower rate could mean that we see significant increases in the rate at which long-read sequencing can be conducted, alongside increased data output per flowcell.

Another aspect that could be improved is the selection process for long-term culture of samples. Currently, this process relies upon morphological differences such as colony discolouration and size. Improving the selection process requires a technique that would make it easier to detect colonies that likely have rearranged genomes and reduce the risk of picking colonies that have not rearranged. Higher throughput methods such as microscopy to identify morphological differences, or looking for metabolic differences seen via metabolic testing kits could be tested using a variety of rearrangements derived from the same strain of *Salmonella* to find which differences are indicative of rearrangement.

7.2 Discovery of a Genome Rearrangement in Long-Term BRD948 Cultures

The discovery of a rearrangement from the long-term cultures confirmed that this method can produce rearrangements in samples. The rearrangement generated in this case showed a major movement of the fragments of the genome that lead to fragments 1 and 3 being next to each other. This generated a significant *ori-ter* skew, causing one large replichore to be generated during replication thus slowing down growth significantly. This new rearrangement presented an opportunity to observe how these rearrangements have a wider effects on *Salmonella*.

In the future, this method can be taken from this initial stage and expanded to include more media conditions and a wider variety of strains. Changes to conditions could introduce different stresses which could potentially yield different rearrangements.

7.3 Effects caused by Genome Rearrangement

By taking the collection of DT2s and generating a phylogenetic tree based on their short-read data, I found that strains with rearrangements in this sample set were not of the same clade, which suggests that genome rearrangement is not inherited.

One limitation of the DT2 dataset is that 30 DT2 samples are not truly representative of the population of DT2s nor of *Salmonella*. In this work there were plans to expand to other *Salmonella* collections available at QIB, though this unfortunately could not take place due to the national lockdown caused by the SARS-CoV-2 pandemic which limited my ability to access laboratory resources.

From the RNA sequencing, I found that LAT2, one of the rearranged strains from the long-term growth cultures, had a large proportion of genes significantly changed in expression. This also highlighted a section of the genome that was entirely missing in LAT2 compared to WT; this loss resulted in the pathway to produce AcP in LAT2 being entirely absent which has significant consequences on cell function. Beyond this, a variety of other genes and gene clusters also had reduced expression in LAT2, with these genes being crucial for various aspects of pathogenesis such as gastrointestinal infection, adhesion, survival under stress and even biofilm formation, the last of which is critical for Typhi's ability to survive in the long-term in chronic carriers.

One potential avenue of further research from these RNASeq results is to observe if these changes in expression are due to the genes in particular or due to the movement of promoters as well. This could be tested by cloning downregulated genes to a point nearer to the origin to see if doing so would restore their standard levels of expression.

7.4 Final Remarks

My original aims included optimizing laboratory methods to improve the quality and quantity of DNA extraction for long-read sequencing to serve as a viable alternative to long-range PCR. I also planned to investigate ways in which to improve the bioinformatics pipeline to improve tolerance to lower coverage and poorer quality runs. This work also intended to discover a means by which to induce rearrangements in a laboratory environment by creating long-term growth cultures that could be regularly sequenced to check for rearrangements. With any discovered rearrangements the aim was to investigate the impact of this rearrangement, first by finding the new configuration of the bacterial genome, and then by conducting RNA sequencing to find out how this rearrangement impacts the expression of genes. All of these aims were achieved.

This work produced a method from culturing samples in the long-term to extracting HMW DNA and long-read sequencing. In the process I also tested the bioinformatics pipeline in place when I started, and made key changes to this while also confirming that the assembler in use was best suited for the purpose of this work. From this I discovered a rearrangement generated by long-term culture. Despite taking longer than predicted to induce such rearrangement, it had the 2nd most extreme *ori-ter* skew theoretically possible. Interestingly, this rearrangement displayed a large amount of differential gene expression with a clear change in expression based on the repositioning of fragments of the genome. In summary, I have established an effective method from growing samples for inducing rearrangement to rapid discovery of rearrangements and their effects on the functionality of the bacteria itself.

This work demonstrates a direct link between genome rearrangement and the expression of genes. With this knowledge, there is now an opportunity for exciting steps forward to take this into a clinical context. There is potential that these rearrangements can be linked to pathogenicity, or in the context of Typhi: to carriage. Finding the ways in which carriage

is established could prove critical in modernising diagnostics by being able to predict the risk of carriage from a rearrangement.

7.5 Future Work

There is increasingly clear evidence that genome rearrangement has a role in gene expression and thus likely has a role within the greater context of bacterial pathogenicity. The bioinformatic methods described in this work have been refined to the point they can be used with clinical samples of Typhi and other bacteria such as *Campylobacter*, *Klebsiella* and others for the purpose of detecting genome rearrangements. It can be theorised, with the genes seen to be downregulated in LAT2 including genes involved with biofilm formation, rearrangement likely has a crucial role in Typhi carriage.

In *Bordetella pertussis* it has been found that these bacteria cycle through a series of arrangements which can branch out into more rearrangements (Weigand *et al.*, 2019). While *B. pertussis* is different from *Salmonella* through having an extremely large number of insertion sites where homologous recombination can take place (Weigand *et al.*, 2020), presenting a much greater number of possible rearrangements, this might also take place in other bacteria such as *Salmonella*.

Several clinical samples of Typhi have been shown to have rearrangements and are from both acute and chronic carraige patients (Ainsworth *et al.*, 2021 (Unpublished)-a). Thus these rearrangements can also have their RNA sequenced which could indicate changes occurring in their genome that could tie to their functionality in a clinical setting.

Another avenue of interest is expanding the use of long-term culture collections, using different medias or stressors, such as minimal media or antimicrobials. One theory is that certain rearrangements are selected for based on the environment the bacteria grow in, as a means of adapting to the environment. This could be used to see if rearrangements have an impact on clinically relevant genes such as those involved in antimicrobial resistance, or if the genes impacted in this work are also seen in the rearrangement of such samples.

Other experiments can be conducted with multiple rearrangements to compare their competitive advantage under particular conditions. One such study could investigate the ability to establish biofilms under conditions such as limited media, presence of other bacteria, or the effect upon antimicrobial resistance through the introduction of AMR

genes or bacterial motility with such experiments being able to be conducted in bacterial rearrangements in general, not just *Salmonella*, such as *Campylobacter*, *Bordetella* and *Staphylococcus* (Scott *et al.*, 2007, Weigand *et al.*, 2017, Ziebuhr *et al.*, 2000),of which have extensive variety in antimicrobial resistance (Li *et al.*, 2019, Hull *et al.*, 2021). In theory with rearrangements that downregulate such genes, strains with these rearrangements would perform worse in surviving against antimicrobials or ability to move through semi-solid agar which directs *Salmonella* to the gut epithelia surface for initial intestinal colonisation (Barbosa *et al.*, 2017, Stecher *et al.*, 2004, Burns *et al.*, 2001).

As it stands, compared to the large number of theoretical rearrangements for *Salmonella* there are relatively few GSs found overall as detailed in Page *et al.*, 2020. This may in part be due to there only being approximately a thousand complete *Salmonella* genomes currently available in public databases. With the findings detailed above showing the clear role of genome rearrangement in the overall functionality of bacteria, more work needs to be done to ensure these bacteria are being sequenced in such a way that genome rearrangements can be found. Furthermore, the methods produced in this work could be used on clinical samples in order to translate changes in gene expression caused by genome rearrangement with clinically relevant samples.

References

AFGAN, E., BAKER, D., BATUT, B., VAN DEN BEEK, M., BOUVIER, D.*, et al.* (2018) The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2018 Update. *Nucleic Acids Research,* 46**,** W537-W544.

AINSWORTH, E. V., TUCKER, L. A. & LANGRIDGE, G. (2021 (Unpublished)-a) Differential Gene Expression Analysis of *Salmonella* Typhi BRD948 Genome Rearrangements.

AINSWORTH, E. V., TUCKER, L. A. & LANGRIDGE, G. (2021 (Unpublished)-b) Genome Rearrangement in Clinical *Salmonella* Typhi.

ALIKHAN, N.-F., PETTY, N. K., BEN ZAKOUR, N. L. & BEATSON, S. A. (2011) Blast Ring Image Generator (Brig): Simple Prokaryote Genome Comparisons. *BMC genomics,* 12**,** 402.

ANDREWS, S. (2015) Fastqc.

ANTILLÓN, M., WARREN, J. L., CRAWFORD, F. W., WEINBERGER, D. M., KÜRÜM, E.*, et al.* (2017) The Burden of Typhoid Fever in Low- and Middle-Income Countries: A Meta-Regression Approach. *PLOS Neglected Tropical Diseases,* 11**,** e0005376.

BANKEVICH, A., NURK, S., ANTIPOV, D., GUREVICH, A. A., DVORKIN, M.*, et al.* (2012) Spades: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology,* 19**,** 455-477.

BARBOSA, F. D. O., FREITAS NETO, O. C. D., BATISTA, D. F. A., ALMEIDA, A. M. D., RUBIO, M. D. S.*, et al.* (2017) Contribution of Flagella and Motility to Gut Colonisation and Pathogenicity of *Salmonella* Enteritidis in the Chicken. *Brazilian Journal of Microbiology,* 48**,** 754-759.

BARNES, W. M. (1992) The Fidelity of Taq Polymerase Catalyzing Pcr Is Improved by an N-Terminal Deletion. *Gene,* 112**,** 29-35.

BARON, F., BONNASSIE, S., ALABDEH, M., COCHET, M.-F., NAU, F.*, et al.* (2017) Global Gene-Expression Analysis of the Response of *Salmonella* Enteritidis to Egg White Exposure Reveals Multiple Egg White-Imposed Stress Responses. *Frontiers in microbiology,* 8**,** 829-829.

BELDA, E., MOYA, A. S. & SILVA, F. J. (2005) Genome Rearrangement Distances and Gene Order Phylogeny in Γ-Proteobacteria. *Molecular Biology and Evolution,* 22**,** 1456-1467.

BOWDEN, S. D., RAMACHANDRAN, V. K., KNUDSEN, G. M., HINTON, J. C. & THOMPSON, A. (2010) An Incomplete Tca Cycle Increases Survival of *Salmonella* Typhimurium During Infection of Resting and Activated Murine Macrophages. *PLOS ONE,* 5**,** e13871.

BRANCHU, P., BAWN, M. & KINGSLEY, R. A. (2018) Genome Variation and Molecular Epidemiology of *Salmonella Enterica* Serovar Typhimurium Pathovariants. *Infection and Immunity,* 86.

BREIMAN, R. F., COSMAS, L., NJUGUNA, H., AUDI, A., OLACK, B.*, et al.* (2012) Population-Based Incidence of Typhoid Fever in an Urban Informal Settlement and a Rural Area in Kenya: Implications for Typhoid Vaccine Use in Africa. *PLOS ONE,* 7**,** e29119-e29119.

BRODIE, J., MACQUEEN, I. A. & LIVINGSTONE, D. (1970) Effect of Trimethoprim-Sulphamethoxazole on Typhoid and Salmonella Carriers. *Br Med J,* 3**,** 318-9.

BRYANT, J. A., SELLARS, L. E., BUSBY, S. J. W. & LEE, D. J. (2014) Chromosome Position Effects on Gene Expression in *Escherichia Coli* K-12. *Nucleic Acids Research,* 42**,** 11383-11392.

BUCKLE, G. C., WALKER, C. L. F. & BLACK, R. E. (2012) Typhoid Fever and Paratyphoid Fever: Systematic Review to Estimate Global Morbidity and Mortality for 2010. *Journal of global health,* 2**,** 010401.

BUMANN, D. (2009) System-Level Analysis of *Salmonella* Metabolism During Infection. *Current Opinion in Microbiology,* 12**,** 559-567.

BURNS, D. L., SCHMITT, C. K., IKEDA, J. S., DARNELL, S. C., WATSON, P. R.*, et al.* (2001) Absence of All Components of the Flagellar Export and Synthesis Machinery Differentially Alters Virulence of *Salmonella Enterica* Serovar Typhimurium in Models of Typhoid Fever, Survival in Macrophages, Tissue Culture Invasiveness, and Calf Enterocolitis. *Infection and Immunity,* 69**,** 5619-5625.

CARVER, T. J., RUTHERFORD, K. M., BERRIMAN, M., RAJANDREAM, M.-A., BARRELL, B. G.*, et al.* (2005) Act: The Artemis Comparison Tool. *Bioinformatics,* 21**,** 3422-3423.

CHEN, P., DEN BAKKER, H. C., KORLACH, J., KONG, N., STOREY, D. B.*, et al.* (2017) Comparative Genomics Reveals the Diversity of Restriction-Modification Systems and DNA Methylation Sites in *Listeria Monocytogenes*. *Applied and Environmental Microbiology,* 83**,** e02091-16.

CHEN, S., ZHOU, Y., CHEN, Y. & GU, J. (2018) Fastp: An Ultra-Fast All-in-One Fastq Preprocessor. *Bioinformatics,* 34**,** i884-i890.

COBURN, B., GRASSL, G. A. & FINLAY, B. B. (2007a) *Salmonella* , the Host and Disease: A Brief Review. *Immunology & Cell Biology,* 85**,** 112-118.

COBURN, B., SEKIROV, I. & FINLAY, B. B. (2007b) Type III Secretion Systems and Disease. *Clinical Microbiology Reviews,* 20**,** 535-549.

CONTINI, S. (2017) Typhoid Intestinal Perforation in Developing Countries: Still Unavoidable Deaths? *World journal of gastroenterology,* 23**,** 1925-1931.

COSTERTON, J. W., STEWART, P. S. & GREENBERG, E. P. (1999) Bacterial Biofilms: A Common Cause of Persistent Infections. *Science,* 284**,** 1318-22.

CRAWFORD, R. W., REEVE, K. E. & GUNN, J. S. (2010) Flagellated but Not Hyperfimbriated *Salmonella Enteric*a Serovar Typhimurium Attaches to and Forms Biofilms on Cholesterol-Coated Surfaces. *Journal of bacteriology,* 192**,** 2981-2990.

CRUMP, J. A., KRETSINGER, K., GAY, K., HOEKSTRA, R. M., VUGIA, D. J.*, et al.* (2008) Clinical Response and Outcome of Infection with *Salmonella Enterica* Serotype Typhi with Decreased Susceptibility to Fluoroquinolones: A United States Foodnet Multicenter Retrospective Cohort Study. *Antimicrob Agents Chemother,* 52**,** 1278-84.

CRUMP, J. A., LUBY, S. P. & MINTZ, E. D. (2004) The Global Burden of Typhoid Fever. *Bulletin of the World Health Organization,* 82**,** 346-53.

DEATHERAGE, D. E. & BARRICK, J. E. (2014) Identification of Mutations in Laboratory-Evolved Microbes from Next-Generation Sequencing Data Using Breseq. *In:* SUN, L. & SHOU, W. (eds.) *Engineering and Analyzing Multicellular Systems: Methods and Protocols.* New York, NY: Springer New York.

DEVARAJ, A., GONZÁLEZ, J. F., EICHAR, B., THILLIEZ, G., KINGSLEY, R. A.*, et al.* (2021) Enhanced Biofilm and Extracellular Matrix Production by Chronic Carriage Versus Acute Isolates of *Salmonella* Typhi. *PLoS pathogens,* 17**,** e1009209.

DOUGAN, G., CHATFIELD, S., PICKARD, D., BESTER, J., O'CALIAGHAN, D.*, et al.* (1988) Construction and Characterization of Vaccine Strains of *Salmonella* Harboring Mutations in Two Different Aro Genes. *The Journal of Infectious Diseases,* 158**,** 1329-1335.

ENG, S.-K., PUSPARAJAH, P., AB MUTALIB, N.-S., SER, H.-L., CHAN, K.-G.*, et al.* (2015) Salmonella: A Review on Pathogenesis, Epidemiology and Antibiotic Resistance. *Frontiers in Life Science,* 8**,** 284-293.

FABREGA, A. & VILA, J. (2013) *Salmonella Enterica* Serovar Typhimurium Skills to Succeed in the Host: Virulence and Regulation. *Clinical Microbiology Reviews,* 26**,** 308-341.

FIGUEIRA, R. & HOLDEN, D. W. (2012) Functions of the *Salmonella* Pathogenicity Island 2 (Spi-2) Type III Secretion System Effectors. *Microbiology (Reading),* 158**,** 1147-1161.

FISHER, I. (2004) Dramatic Shift in the Epidemiology of *Salmonella Enterica* Serotype Enteritidis Phage Types in Western Europe, 1998-2003--Results from the Enter-Net International Salmonella Database. *Euro surveillance : bulletin européen sur les maladies transmissibles = European communicable disease bulletin,* 9**,** 43-5.

FITZGERALD, C., SHERWOOD, R., GHEESLING, L. L., BRENNER, F. W. & FIELDS, P. I. (2003) Molecular Analysis of the Rfb O Antigen Gene Cluster of *Salmonella Enterica* Serogroup O:6,14 and Development of a Serogroup-Specific PCR Assay. *Applied and Environmental Microbiology,* 69**,** 6099-6105.

FOOKES, M., SCHROEDER, G. N., LANGRIDGE, G. C., BLONDEL, C. J., MAMMINA, C.*, et al.* (2011) *Salmonella Bongori* Provides Insights into the Evolution of the Salmonellae. *PLoS pathogens,* 7**,** e1002191.

GART, E. V., SUCHODOLSKI, J. S., WELSH, T. H., ALANIZ, R. C., RANDEL, R. D.*, et al.* (2016) *Salmonella* Typhimurium and Multidirectional Communication in the Gut. *Frontiers in microbiology,* 7.

GASEM, M. H., DOLMANS, W. M. V. W. M. V., KEUTER, M. M. & DJOKOMOELJANTO, R. R. (2001) Poor Food Hygiene and Housing as Risk Factors for Typhoid Fever in Semarang, Indonesia. *Tropical Medicine and International Health,* 6**,** 484-490.

GERGANOVA, V., BERGER, M., ZALDASTANISHVILI, E., SOBETZKO, P., LAFON, C.*, et al.* (2015) Chromosomal Position Shift of a Regulatory Gene Alters the Bacterial Phenotype. *Nucleic Acids Research,* 43**,** 8215-8226.

GIBSON, A. M. & ROBERTS, T. A. (1986) The Effect of pH, Water Activity, Sodium Nitrite and Storage Temperature on the Growth of Enteropathogenic *Escherichia Coli* and Salmonellae in a Laboratory Medium. *International Journal of Food Microbiology,* 3**,** 183-194.

GONZALEZ-ESCOBEDO, G., MARSHALL, J. M. & GUNN, J. S. (2011) Chronic and Acute Infection of the Gall Bladder by *Salmonella* Typhi: Understanding the Carrier State. *Nature Reviews Microbiology,* 9**,** 9-14.

GONZÁLEZ, J. F., ALBERTS, H., LEE, J., DOOLITTLE, L. & GUNN, J. S. (2018) Biofilm Formation Protects *Salmonella* from the Antibiotic Ciprofloxacin in Vitro and in Vivo in the Mouse Model of Chronic Carriage. *Scientific Reports,* 8**,** 222.

GORDON, M. A. (2011) Invasive Nontyphoidal *Salmonella* Disease: Epidemiology, Pathogenesis and Diagnosis. *Current opinion in infectious diseases,* 24**,** 484-489.

GRIMONT, P. & WEILL, F.-X. (2007) Antigenic Formulae of the *Salmonella* Serovars, (9th Ed.) Paris: Who Collaborating Centre for Reference and Research on Salmonella. *Institute Pasteur.***,** 1-166.

HAASE, J. (2008) The Impact of Genomic Rearrangement in *Salmonella* Typhi*.* Diploma in Biology, Martin-Luther-University Halle-Wittenberg.

HANSEN-WESTER, I. & HENSEL, M. (2001) *Salmonella* Pathogenicity Islands Encoding Type III Secretion Systems. *Microbes and Infection,* 3**,** 549-559.

HELAINE, S., THOMPSON, J. A., WATSON, K. G., LIU, M., BOYLE, C.*, et al.* (2010) Dynamics of Intracellular Bacterial Replication at the Single Cell Level. *Proceedings of the National Academy of Sciences,* 107**,** 3746-3751.

HELM, R. A., LEE, A. G., CHRISTMAN, H. D. & MALOY, S. (2003) Genomic Rearrangements at Rrn Operons in *Salmonella*. *Genetics,* 165**,** 951-9.

HELM, R. A. & MALOY, S. (2001) Rapid Approach to Determine Rrn Arrangement in *Salmonella* Serovars. *Applied and Environmental Microbiology,* 67**,** 3295-3298.

HELM, R. A., PORWOLLIK, S., STANLEY, A. E., MALOY, S., MCCLELLAND, M.*, et al.* (2004) Pigeon-Associated Strains of *Salmonella Enterica* Serovar Typhimurium Phage Type Dt2 Have Genomic Rearrangements at Rrna Operons. *Infection and Immunity,* 72**,** 7338 LP - 7341.

HENSEL, M., SHEA, J. E., WATERMAN, S. R., MUNDY, R., NIKOLAUS, T.*, et al.* (1998) Genes Encoding Putative Effector Proteins of the Type III Secretion System of *Salmonella* Pathogenicity Island 2 Are Required for Bacterial Virulence and Proliferation in Macrophages. *Molecular Microbiology,* 30**,** 163-174.

HERRMANN, K. M. & WEAVER, L. M. (1999) The Shikimate Pathway. *Annual Review of Plant Physiology and Plant Molecular Biology,* 50**,** 473-503.

HUANG, X. (1992) A Contig Assembly Program Based on Sensitive Detection of Fragment Overlaps. *Genomics,* 14**,** 18-25.

HULL, D. M., HARRELL, E., VAN VLIET, A. H. M., CORREA, M. & THAKUR, S. (2021) Antimicrobial Resistance and Interspecies Gene Transfer in *Campylobacter Coli* and *Campylobacter*

Jejuni Isolated from Food Animals, Poultry Processing, and Retail Meat in North Carolina, 2018–2019. *PLOS ONE,* 16**,** e0246571.
ILYAS, B., TSAI, C. N. & COOMBES, B. K. (2017) Evolution of *Salmonella*-Host Cell Interactions through a Dynamic Bacterial Genome. *Frontiers in Cellular and Infection Microbiology,* 7.
JAJERE, S. M. (2019) A Review of *Salmonella Enterica* with Particular Focus on the Pathogenicity and Virulence Factors, Host Specificity and Antimicrobial Resistance Including Multidrug Resistance. *Veterinary world,* 12**,** 504-521.
KARP, P. D., PALEY, S. & ROMERO, P. (2002) The Pathway Tools Software. *Bioinformatics,* 18**,** S225-S232.
KILPINEN, S., AUTIO, R., OJALA, K., ILJIN, K., BUCHER, E.*, et al.* (2008) Systematic Bioinformatic Analysis of Expression Levels of 17,330 Human Genes across 9,783 Samples from 175 Types of Healthy and Pathological Tissues. *Genome Biology,* 9**,** R139.
KINGSLEY, R. A., KAY, S., CONNOR, T., BARQUIST, L., SAIT, L.*, et al.* (2013) Genome and Transcriptome Adaptation Accompanying Emergence of the Definitive Type 2 Host-Restricted *Salmonella Enterica* Serovar Typhimurium Pathovar. *mBio,* 4.
KINGSLEY, R. A., LANGRIDGE, G., SMITH, S. E., MAKENDI, C., FOOKES, M.*, et al.* (2018) Functional Analysis of *Salmonella* Typhi Adaptation to Survival in Water. *Environ Microbiol,* 20**,** 4079-4090.
KLEMM, E. J., SHAKOOR, S., PAGE, A. J., QAMAR, F. N., JUDGE, K.*, et al.* (2018) Emergence of an Extensively Drug-Resistant *Salmonella Enterica* Serovar Typhi Clone Harboring a Promiscuous Plasmid Encoding Resistance to Fluoroquinolones and Third-Generation Cephalosporins. *mBio,* 9.
KLUMPP, S., ZHANG, Z. & HWA, T. (2009) Growth Rate-Dependent Global Effects on Gene Expression in Bacteria. *Cell,* 139**,** 1366-1375.
KOLMOGOROV, M., YUAN, J., LIN, Y. & PEVZNER, P. A. (2019) Assembly of Long, Error-Prone Reads Using Repeat Graphs. *Nature Biotechnology,* 37**,** 540-546.
KOREN, S., WALENZ, B. P., BERLIN, K., MILLER, J. R., BERGMAN, N. H.*, et al.* (2017) Canu: Scalable and Accurate Long-Read Assembly Via Adaptive K-Mer Weighting and Repeat Separation. *Genome Research,* 27**,** 722-736.
KOZAK, G. K., MACDONALD, D., LANDRY, L. & FARBER, J. M. (2013) Foodborne Outbreaks in Canada Linked to Produce: 2001 through 2009. *Journal of Food Protection,* 76**,** 173-183.
KRÖGER, C., DILLON, S. C., CAMERON, A. D. S., PAPENFORT, K., SIVASANKARAN, S. K.*, et al.* (2012) The Transcriptional Landscape and Small Rnas of *Salmonella Enterica* Serovar Typhimurium. *Proceedings of the National Academy of Sciences of the United States of America,* 109**,** E1277-86.
KUKURBA, K. R. & MONTGOMERY, S. B. (2015) Rna Sequencing and Analysis. *Cold Spring Harbor Protocols,* 2015**,** pdb.top084970.
LANGRIDGE, G. C., FOOKES, M., CONNOR, T. R., FELTWELL, T., FEASEY, N.*, et al.* (2015) Patterns of Genome Evolution That Have Accompanied Host Adaptation in *Salmonella*. *Proceedings of the National Academy of Sciences of the United States of America,* 112**,** 863-8.
LEROUGE, I. & VANDERLEYDEN, J. (2002) O-Antigen Structural Variation: Mechanisms and Possible Roles in Animal/Plant–Microbe Interactions. *FEMS Microbiology Reviews,* 26**,** 17-47.
LETUNIC, I. & BORK, P. (2019) Interactive Tree of Life (Itol) V4: Recent Updates and New Developments. *Nucleic Acids Research,* 47**,** W256-W259.
LEVINE, M. M., BLACK, R. E. & LANATA, C. (1982) Precise Estimation of the Numbers of Chronic Carriers of *Salmonella* Typhi in Santiago, Chile, an Endemic Area. *The Journal of Infectious Diseases,* 146**,** 724-726.
LI, H. (2016) Minimap and Miniasm: Fast Mapping and De Novo Assembly for Noisy Long Sequences. *Bioinformatics,* 32**,** 2103-2110.

LI, L., DENG, J., MA, X., ZHOU, K., MENG, Q.*, et al.* (2019) High Prevalence of Macrolide-Resistant *Bordetella Pertussis* and Ptxp1 Genotype, Mainland China, 2014–2016. *Emerging Infectious Disease journal,* 25**,** 2205.

LI, L., WANG, W., ZHANG, R., XU, J., WANG, R.*, et al.* (2018) First Acetyl-Proteome Profiling of *Salmonella* Typhimurium Revealed Involvement of Lysine Acetylation in Drug Resistance. *Veterinary Microbiology,* 226**,** 1-8.

LIAO, Y., SMYTH, G. K. & SHI, W. (2013) Featurecounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features. *Bioinformatics,* 30**,** 923-930.

LIU, S. L. & SANDERSON, K. E. (1995) Rearrangements in the Genome of the Bacterium *Salmonella* Typhi. *Proceedings of the National Academy of Sciences,* 92**,** 1018-1022.

LIU, W.-Y., WONG, C.-F., CHUNG, K. M.-K., JIANG, J.-W. & LEUNG, F. C.-C. (2013) Comparative Genome Analysis of *Enterobacter Cloacae*. *PLOS ONE,* 8**,** e74487.

LOVE, M. I., HUBER, W. & ANDERS, S. (2014) Moderated Estimation of Fold Change and Dispersion for Rna-Seq Data with DeSeq2. *Genome Biology,* 15**,** 550.

MACKIEWICZ, P., ZAKRZEWSKA-CZERWINSKA, J., ZAWILAK, A., DUDEK, M. R. & CEBRAT, S. (2004) Where Does Bacterial Replication Start? Rules for Predicting the *oriC* Region. *Nucleic Acids Research,* 32**,** 3781-3791.

MAIN, R. G. (1961) Treatment of the Chronic Alimentary Enteric Carrier. *BMJ,* 1**,** 328-333.

MARIONI, J. C., MASON, C. E., MANE, S. M., STEPHENS, M. & GILAD, Y. (2008) Rna-Seq: An Assessment of Technical Reproducibility and Comparison with Gene Expression Arrays. *Genome Research,* 18**,** 1509-1517.

MARSHALL, J. M. & GUNN, J. S. (2015) The O-Antigen Capsule of *Salmonella Enterica* Serovar Typhimurium Facilitates Serum Resistance and Surface Expression of Flic. *Infection and Immunity,* 83**,** 3946-3959.

MATCHES, J. R. & LISTON, J. (1972) Effects of Incubation Temperature on the Salt Tolerance of *Salmonella*. *Journal of Milk and Food Technology,* 35**,** 39-44.

MATTHEWS, T. D., RABSCH, W. & MALOY, S. (2011) Chromosomal Rearrangements in *Salmonella Enterica* Serovar Typhi Strains Isolated from Asymptomatic Human Carriers. *mBio,* 2**,** e00060-e11.

MERRITT, J. H., KADOURI, D. E. & O'TOOLE, G. A. (2005) Growing and Analyzing Static Biofilms. *Current protocols in microbiology,* Chapter 1**,** Unit-1B.1.

MOGASALE, V., MASKERY, B., OCHIAI, R. L., LEE, J. S., MOGASALE, V. V.*, et al.* (2014) Burden of Typhoid Fever in Low-Income and Middle-Income Countries: A Systematic, Literature-Based Update with Risk-Factor Adjustment. *The Lancet Global Health,* 2**,** e570-e580.

MONACK , D. M., BOULEY , D. M. & FALKOW , S. (2004) *Salmonella* Typhimurium Persists within Macrophages in the Mesenteric Lymph Nodes of Chronically Infected Nramp1+/+ Mice and Can Be Reactivated by Ifnγ Neutralization. *Journal of Experimental Medicine,* 199**,** 231-241.

NEIDHARDT, F. C., BLOCH, P. L. & SMITH, D. F. (1974) Culture Medium for Enterobacteria. *Journal of bacteriology,* 119**,** 736-747.

NGUYEN, L.-T., SCHMIDT, H. A., VON HAESELER, A. & MINH, B. Q. (2014) Iq-Tree: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution,* 32**,** 268-274.

NOLAN, C. M. & WHITE, P. C., JR. (1978) Treatment of Typhoid Carriers with Amoxicillin. Correlates of Successful Therapy. *Jama,* 239**,** 2352-4.

OCHMAN, H. & GROISMAN, E. A. (1996) Distribution of Pathogenicity Islands in *Salmonella* Spp. *Infection and Immunity,* 64**,** 5410-5412.

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.

PETKOV, P. M., GRABER, J. H., CHURCHILL, G. A., DIPETRILLO, K., KING, B. L., *et al.* (2005) Evidence of a Large-Scale Functional Organization of Mammalian Chromosomes. *PLOS Genetics,* 1**,** e33.

PHILLIPS, W. E. (1971) Treatment of Chronic Typhoid Carriers with Ampicillin. *Jama,* 217**,** 913-5.

QUE, F., WU, S. & HUANG, R. (2013) *Salmonella* Pathogenicity Island 1(Spi-1) at Work. *Current Microbiology,* 66**,** 582-587.

RABSCH, W. (2007) *Salmonella* Typhimurium Phage Typing for Pathogens. *In:* SCHATTEN, H. & EISENSTARK, A. (eds.) *Salmonella: Methods and Protocols.* Totowa, NJ: Humana Press.

RABSCH, W., ANDREWS, H. L., KINGSLEY, R. A., PRAGER, R., TSCHÄPE, H., *et al.* (2002) *Salmonella Enterica* Serotype Typhimurium and Its Host-Adapted Variants. *Infection and Immunity,* 70**,** 2249-2255.

REN, J., SANG, Y., QIN, R., SU, Y., CUI, Z., *et al.* (2019) Metabolic Intermediate Acetyl Phosphate Modulates Bacterial Virulence Via Acetylation. *Emerging Microbes & Infections,* 8**,** 55-69.

RÖMLING, U. & ROHDE, M. (1999) Flagella Modulate the Multicellular Behavior of *Salmonella* Typhimurium on the Community Level. *FEMS Microbiol Lett,* 180**,** 91-102.

ROSZAK, D. B. & COLWELL, R. R. (1987) Survival Strategies of Bacteria in the Natural Environment. *Microbiological reviews,* 51**,** 365-379.

SABBAGH, S. C., FOREST, C. G., LEPAGE, C., LECLERC, J.-M. & DAIGLE, F. (2010) So Similar, yet So Different: Uncovering Distinctive Features in the Genomes of *Salmonella Enterica* Serovars Typhimurium and Typhi. *FEMS Microbiology Letters,* 305**,** 1-13.

SCOTT, A. E., TIMMS, A. R., CONNERTON, P. L., LOC CARRILLO, C., ADZFA RADZUM, K., *et al.* (2007) Genome Dynamics of Campylobacter Jejuni in Response to Bacteriophage Predation. *PLoS pathogens,* 3**,** e119-e119.

SEEMANN, T. (2015) Snippy: Fast Bacterial Variant Calling from Ngs Reads.

SEEMANN, T. (2020) Snippy. 4.6.0 ed. GitHub.

SENTHILKUMAR, B., SENBAGAM, D. & RAJASEKARAPANDIAN, M. (2014) An Epidemiological Surveillance of Asymptomatic Typhoid Carriers Associated in Respect to Socioeconomic Status in India. *Journal of Public Health,* 22**,** 297-301.

SILVERMAN, M., ZIEG, J., HILMEN, M. & SIMON, M. (1979) Phase Variation in *Salmonella*: Genetic Analysis of a Recombinational Switch. *Proceedings of the National Academy of Sciences,* 76**,** 391-395.

SKIENA, S. S. (2008) The Algorithm Design Manual.

SOLER-BISTUÉ, A., TIMMERMANS, M. & MAZEL, D. (2017) The Proximity of Ribosomal Protein Genes to *oriC* Enhances *Vibrio Cholerae* Fitness in the Absence of Multifork Replication. *mBio,* 8**,** e00097-17.

SOUSA, C., DE LORENZO, V. & CEBOLLA, A. (1997) Modulation of Gene Expression through Chromosomal Positioning in *Escherichia Coli*. *Microbiology,* 143**,** 2071-2078.

SPANÒ, S. (2016) Mechanisms of *Salmonella* Typhi Host Restriction. *In:* LEAKE, M. C. (ed.) *Biophysics of Infection.* Cham: Springer International Publishing.

SPIGA, L., WINTER, M. G., FURTADO DE CARVALHO, T., ZHU, W., HUGHES, E. R., *et al.* (2017) An Oxidative Central Metabolism Enables *Salmonella* to Utilize Microbiota-Derived Succinate. *Cell Host & Microbe,* 22**,** 291-301.e6.

STANAWAY, J. D., REINER, R. C., BLACKER, B. F., GOLDBERG, E. M., KHALIL, I. A., *et al.* (2019) The Global Burden of Typhoid and Paratyphoid Fevers: A Systematic Analysis for the Global Burden of Disease Study 2017. *The Lancet Infectious Diseases,* 19**,** 369-381.

STECHER, B., HAPFELMEIER, S., MÜLLER, C., KREMER, M., STALLMACH, T., *et al.* (2004) Flagella and Chemotaxis Are Required for Efficient Induction of <I>Salmonella Enterica</I> Serovar Typhimurium Colitis in Streptomycin-Pretreated Mice. *Infection and Immunity,* 72**,** 4138-4150.

STEELE-MORTIMER, O. (2008) The *Salmonella*-Containing Vacuole—Moving with the Times. *Current Opinion in Microbiology,* 11**,** 38-45.

TACKET, C. O., SZTEIN, M. B., LOSONSKY, G. A., WASSERMAN, S. S., NATARO, J. P.*, et al.* (1997) Safety of Live Oral *Salmonella* Typhi Vaccine Strains with Deletions in Htra and Aroc Arod and Immune Response in Humans. *Infection and Immunity,* 65**,** 452.

TAKAYA, A., SUZUKI, M., MATSUI, H., TOMOYASU, T., SASHINAMI, H.*, et al.* (2003) Lon, a Stress-Induced Atp-Dependent Protease, Is Critically Important for Systemic *Salmonella Enterica* Serovar Typhimurium Infection of Mice. *Infection and Immunity,* 71**,** 690-696.

TCHAWA YIMGA, M., LEATHAM MARY, P., ALLEN JAMES, H., LAUX DAVID, C., CONWAY, T.*, et al.* (2006) Role of Gluconeogenesis and the Tricarboxylic Acid Cycle in the Virulence of *Salmonella Enterica* Serovar Typhimurium in Balb/C Mice. *Infection and Immunity,* 74**,** 1130-1140.

TSURU, T., KAWAI, M., MIZUTANI-UI, Y., UCHIYAMA, I. & KOBAYASHI, I. (2006) Evolution of Paralogous Genes: Reconstruction of Genome Rearrangements through Comparison of Multiple Genomes within *Staphylococcus Aureus*. *Molecular Biology and Evolution,* 23**,** 1269-1285.

VASER, R. & ŠIKIĆ, M. (2021) Raven: A De Novo Genome Assembler for Long Reads. *bioRxiv***,** 2020.08.07.242461.

WANG, F., DENG, L., HUANG, F., WANG, Z., LU, Q.*, et al.* (2020) Flagellar Motility Is Critical for *Salmonella Enterica* Serovar Typhimurium Biofilm Development. *Frontiers in microbiology,* 11**,** 1695-1695.

WANG, Z., GERSTEIN, M. & SNYDER, M. (2009) Rna-Seq: A Revolutionary Tool for Transcriptomics. *Nature Reviews Genetics,* 10**,** 57-63.

WEIGAND, M. R., PENG, Y., LOPAREV, V., BATRA, D., BOWDEN, K. E.*, et al.* (2017) The History of *Bordetella Pertussis* Genome Evolution Includes Structural Rearrangement. *Journal of bacteriology,* 199**,** e00806-16.

WEIGAND, M. R., PENG, Y., POUSEELE, H., KANIA, D., BOWDEN, K. E.*, et al.* (2020) Genomic Surveillance and Improved Molecular Typing of *Bordetella Pertussis* Using wgMLST. *bioRxiv***,** 2020.10.28.360149.

WEIGAND, M. R., WILLIAMS, M. M., PENG, Y., KANIA, D., PAWLOSKI, L. C.*, et al.* (2019) Genomic Survey of *Bordetella Pertussis* Diversity, United States, 2000-2013. *Emerging infectious diseases,* 25**,** 780-783.

WHO (2008) Typhoid Vaccines: Who Position Paper. *Wkly Epidemiol Rec,* 83**,** 49-59.

WICK, R. & HOLT, K. (2021) Benchmarking of Long-Read Assemblers for Prokaryote Whole Genome Sequencing [Version 4; Peer Review: 4 Approved]. *F1000Research,* 8.

WICK, R. R. (2019) Badread: Simulation of Error-Prone Long Reads. *Journal of Open Source Software,* 4.

WICK, R. R. & HOLT, K. E. (2019) Assembly Dereplicator. 0.1.0 ed. GitHub.

WICK, R. R., JUDD, L. M., GORRIE, C. L. & HOLT, K. E. (2017) Unicycler: Resolving Bacterial Genome Assemblies from Short and Long Sequencing Reads. *PLOS Computational Biology,* 13**,** e1005595.

WIJESINGHE, G., DILHARI, A., GAYANI, B., KOTTEGODA, N., SAMARANAYAKE, L.*, et al.* (2019) Influence of Laboratory Culture Media on in Vitro Growth, Adhesion, and Biofilm Formation of *Pseudomonas Aeruginosa* and *Staphylococcus Aureus*. *Medical Principles and Practice,* 28**,** 28-35.

WOLFE, A. J. (2005) The Acetate Switch. *Microbiology and Molecular Biology Reviews,* 69**,** 12-50.

WONG, V. K., BAKER, S., PICKARD, D. J., PARKHILL, J., PAGE, A. J.*, et al.* (2015) Phylogeographical Analysis of the Dominant Multidrug-Resistant H58 Clade of *Salmonella* Typhi Identifies Inter- and Intracontinental Transmission Events. *Nature Genetics,* 47**,** 632-639.

ZAVALA TRUJILLO, I., QUIROZ, C., GUTIERREZ, M. A., ARIAS, J. & RENTERIA, M. (1991) Fluoroquinolones in the Treatment of Typhoid Fever and the Carrier State. *Eur J Clin Microbiol Infect Dis,* 10**,** 334-41.

ZENG, B., ZHAO, G., CAO, X., YANG, Z., WANG, C.*, et al.* (2013) Formation and Resuscitation of Viable but Nonculturable *Salmonella* Typhi. *Biomed Res Int,* 2013**,** 907170.
ZHAO, B., ERWIN, A. & XUE, B. (2018) How Many Differentially Expressed Genes: A Perspective from the Comparison of Genotypic and Phenotypic Distances. *Genomics,* 110**,** 67-73.
ZHAO, S., FUNG-LEUNG, W.-P., BITTNER, A., NGO, K. & LIU, X. (2014) Comparison of Rna-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLOS ONE,* 9**,** e78644.
ZIEBUHR, W., DIETRICH, K., TRAUTMANN, M. & WILHELM, M. (2000) Chromosomal Rearrangements Affecting Biofilm Production and Antibiotic Resistance in a *Staphylococcus Epidermidis* Strain Causing Shunt-Associated Ventriculitis. *Int J Med Microbiol,* 290**,** 115-20.

Appendix 1 Genes Between WT and LAT2 that Express Significant Difference in Expression (log 2-FC > ±0.58, *q*-adjusted < 0.05)

Gene ID	Log_2 Fold Change	p-value	q-adjusted	Gene	Gene Description
t0527	7.15121	2.75E-300	1.21E-296	ackA	Acetate kinase
t0528	6.07643	3.87E-198	8.54E-195	yfbV	Conserved hypothetical protein
t0530	5.761087	1.25E-162	1.85E-159	hxpA	Putative phosphatase
t0529	4.59295	1.62E-87	1.43E-84	yfbU	Conserved hypothetical protein
t0526	2.19271	1.34E-99	1.48E-96	pta	Phosphate acetyltransferase
t0531	2.032055	1.37E-76	1.01E-73	-	Putative sodium/sulphate transporter
t0069	0.991138	1.54E-08	6.20E-06	caiF	Transcriptional activator caif
t2620	0.959656	6.44E-05	0.003846	yfiA	Putative sigma-54 modulation protein
t0576	0.95352	9.26E-07	0.000248	-	Conserved hypothetical protein
t0521	0.929088	1.13E-05	0.001253	-	Putative sugar phosphotransferase component II B
t0443	0.924784	4.90E-09	2.41E-06	yfeD	Conserved hypothetical protein
t0520	0.892706	8.95E-06	0.001068	-	Putative sugar phosphotransferase component II A
t0918	0.887232	0.000188	0.007987	fliC	Flagellin
t2769	0.882813	1.10E-05	0.001246	sprA	Possible arac-family transcriptional regulator
t0046	0.818601	1.38E-11	7.62E-09	-	Conserved hypothetical protein
t0444	0.817949	0.00041	0.013349	yfeC	Conserved hypothetical protein
t0961	0.815597	7.04E-05	0.003904	cheY	Chemotaxis protein chey
t0524	0.813226	5.71E-06	0.000814	-	Putative transketolase C-terminal section
t0168	0.797662	0.000139	0.006588	-	Pdxa-like protein
t0578	0.76735	0.000106	0.005403	glpC	Anaerobic glycerol-3-phosphate dehydrogenase subunit C
t0953	0.763685	0.00038	0.012706	flhC	Flagellar transcriptional activator
t4360	0.759922	1.43E-06	0.00035	-	Putative membrane protein
t0248	0.758474	0.000528	0.015446	yaeC	Putative lipoprotein precursor
t3673	0.734392	0.00071	0.018446	-	Conserved hypothetical protein
t0549	0.732627	0.000545	0.015548	-	Putative receptor/regulator protein
t0172	0.730577	1.25E-08	5.54E-06	yacC	Conserved hypothetical protein
t4493	0.716374	0.001636	0.029752	yjgF	Conserved hypothetical protein
t1179	0.713683	0.002083	0.033841	-	Putative exported protein
t0945	0.712326	4.03E-05	0.00306	yecI	Ferritin-like protein
t4109	0.708581	0.00114	0.02444	yjaB	Putative acetyltransferase
t3798	0.708422	0.000117	0.005725	-	Hypothetical protein
t0456	0.708148	0.000229	0.009101	tnpA	Transposase for insertion sequence element IS200
t1212	0.707829	0.000159	0.007173	rfc	O-antigen polymerase

t4635	0.706392	5.23E-15	3.30E-12	-	Putative transcriptional regulator protein
t0784	0.705286	0.001354	0.026461	rfbE	CDP-tyvelose-2-epimerase
t4519	0.6986	8.12E-06	0.000997	-	Protein kinase
t4121	0.691786	9.53E-06	0.001108	-	Conserved hypothetical protein
t3164	0.688499	4.48E-05	0.003244	tdcB	Catabolic threonine dehydratase
t0958	0.674172	0.001087	0.023421	tar	Methyl-accepting chemotaxis protein II
t1475	0.670931	0.001806	0.031668	-	Putative secreted protein
t0926	0.66296	0.001999	0.033333	sdiA	Cell-division regulatory protein
t3168	0.661479	0.001341	0.026337	garR	2-hydroxy-3-oxopropionate reductase
t4323	0.647529	0.000411	0.013349	-	Hypothetical protein
t0458	0.638657	9.53E-07	0.000248	-	Conserved hypothetical protein
t0785	0.633235	9.95E-05	0.00517	rfbX	Putative O-antigen transporter
t2564	0.63188	0.003379	0.045507	-	Hypothetical protein
t4572	0.624728	0.000842	0.020641	-	Putative membrane protein
t0466	0.622345	0.002521	0.038281	-	Putative lipopolysaccharide modification acyltransferase
t1061	0.620892	0.000538	0.015524	minE	Cell division topological specificity factor
t4549	0.619915	9.13E-05	0.004921	-	Conserved hypothetical protein
t0413	0.615429	0.003112	0.042968	yfeF	Putative oxidoreductase
t0083	0.612689	4.16E-07	0.000153	-	Probable secreted protein
t4324	0.604069	0.000698	0.018258	-	Hypothetical protein
t0783	0.601003	0.000262	0.010083	rfbS	Paratose synthase
t0084	0.598085	7.07E-05	0.003904	-	Putative membrane protein
t4392	0.597261	0.000864	0.020641	frdD	Fumarate reductase complex subunit D
t1123	0.592153	0.000881	0.020706	-	Putative secreted protein
t0468	0.58081	0.000171	0.007569	gtrA	Bactoprenol-linked glucose transferase
t0085	0.58045	0.003452	0.045636	-	Probable secreted protein
t1894	-0.584	0.00132	0.026276	-	Putative prophage terminase large subunit
t2065	-0.61876	0.001625	0.029752	ybiJ	Putative exported protein
t1719	-0.62379	4.96E-06	0.000756	holB	DNA polymerase III, delta' subunit
t1881	-0.62527	0.002692	0.03965	-	Putative bacteriophage protein
t2203	-0.62666	0.001328	0.026276	miaB	Miab protein
t2419	-0.63349	0.000911	0.021125	cyoC	Cytochrome o ubiquinol oxidase subunit III
t1702	-0.63681	0.001316	0.026276	-	ABC transporter ATP-binding subunit
t1915	-0.64489	0.000215	0.008773	-	Conserved hypothetical bacteriophage protein
t2328	-0.64663	5.03E-07	0.000169	purK	Phosphoribosylaminoimidazole carboxylase atpase subunit

t1873	-0.64728	0.000881	0.020706	-	Putative bacteriophage protein
t1899	-0.65997	0.000774	0.019767	-	Putative bacteriophage protein
t2139	-0.6629	0.003438	0.045636	sucB	Dihydrolipoamide succinyltransferase component
t2421	-0.68845	0.003444	0.045636	cyoE	Cytochrome o ubiquinol oxidase C subunit
t2418	-0.68995	0.001866	0.032464	cyoB	Cytochrome o ubiquinol oxidase subunit I
t1885	-0.69071	0.001607	0.029578	-	Putative bacteriophage protein
t1886	-0.69745	0.002192	0.034955	-	Putative bacteriophage protein
t1883	-0.72988	0.000457	0.014138	-	Putative bacteriophage protein
t1882	-0.7309	0.000838	0.020641	-	Putative bacteriophage protein
t1888	-0.74862	0.001515	0.028797	-	Putative bacteriophage protein
t2140	-0.78879	0.000771	0.019767	sucA	2-oxoglutarate dehydrogenase E1 component
t2146	-0.80806	3.69E-06	0.000652	gltA	Citrate synthase
t2417	-0.84524	0.000307	0.011236	cyoA	Cytochrome o ubiquinol oxidase subunit II
t2143	-0.85699	0.000184	0.007877	sdhD	Succinate dehydrogenase hydrophobic membrane anchor protein
t2142	-0.9525	5.74E-05	0.003562	sdhA	Succinate dehydrogenase flavoprotein subunit

Appendix 2 is a separately provided Excel document consisting of all genes analysed via RNA Sequencing; all genes from this that show only a *q*-adjusted of <0.05, and all genes that meet both this *q*-adjusted criteria and a log 2-FC of > ±0.58.