MASSEY UNIVERSITY

COLLEGE OF SCIENCE

SCHOOL OF NATURAL SCIENCES

Characterisation of Epigenomic variation in natural isolates of *E. coli*

A thesis submitted in partial fulfilment
of the requirements for the degree of
Ph.D.
in
Genetics

Submitted by

Georgia Breckell

Supervisor: Olin Silander Auckland 2023

Abstract

DNA methylation is ubiquitous in bacteria and has a range of roles including self versus non-self recognition, DNA repair, and regulation of gene expression in response to internal and external cues. Regulation of gene expression by DNA methylation can lead to the establishment of phenotypic variation in otherwise isogenic populations. Until recently methods for the genome-wide study of DNA methylation in bacteria have been limited and therefore the full extent of DNA methylation's role in bacterial genomes is not well understood.

In this thesis I use Oxford Nanopore Technologies sequencing to investigate the presence and activity of DNA methyltransferase in natural isolates of *E. coli.* The first aim of this thesis is to produce high quality genome assemblies that can be used to determine methylation patterns. To achieve this, in Chapter 2 I first use in silico methods to quantify the effects of different read length characteristics on assembly quality. I then optimise DNA isolation and library prep methods to obtain high quality DNA.

In Chapter 3 I apply the results of Chapter 2 to sequence 49 natural isolates of *E. coli* from across the *E. coli* clade. I next benchmark five genome assembly methods for assembly accuracy. I base accuracy on five metrics designed to measure both the overall structural accuracy and the sequence accuracy of each assembly. The large number of isolates (49) used in this study, allows identification of the strengths associated with each assembly method. These results quantitatively describe best practices for bacterial genome assembly and highlight the current variability in genome assembly accuracy and therefore the importance of tailoring assembly methods to the study objectives.

Finally, in chapter 4 I use the data produced in Chapter 3 to investigate DNA methylation in three *E. coli* natural isolates. After in silico identification of all the methyltransferases in each genome, I show that the activity of all predicted methyltransferases can be detected, as well as the activity of unexpected putative methyltransferases which are present in our isolates. Finally, I show that the genome wide DNA methylation patterns show consistent differences across growth conditions. These results suggest that *E. coli* exhibits transient DNA methylation patterns depending on growth environment and state.

Overall this thesis establishes methods for assessing genome assemblies and broadens our understanding of genome wide DNA methylation patterns and the dynamics of these patterns in *E. coli.* Additionally this work provides insight into the possibility of transient

epigenetic differentiation in *E. coli* which is reflected in the DNA methylation patterns across the genome.

Acknowledgments

Firstly, I would like to thank my supervisor Dr. Olin Silander for his immeasurable support and guidance during my PhD. His comments, feedback and constant support have helped shape me into a better scientist.

I would also like to thank Dr Nikki Freed for her assistance in the lab, particularly getting started with Nanopore sequencing, and for her valuable questions, comments and feedback on results and presentations. I would also like to acknowledge Dr Sebastian Schiemier, and Dr Tim Cooper for their conversations and assistance when I hit roadblocks, particularly early assistance learning computational techniques, and support when growing my isolates in varying environments.

To the past and present members of the Silander Lab, it has been a pleasure to have shared the joys and frustrations of PhD life with you all. Particular thanks go to Marketa for providing the perfect lab book template, and being a constant source of inspiration, William and Bhargava for challenging me computationally, and Stella for her endless moral support though relatable content, coffee and croissants, and passionate conversations about life inside and outside the lab.

Table of contents

		Page
Abbreviations		9
List of Figures		11
List of Tables		13
Chapter 1	**Introduction**	14
1.1	Phenotypic variation via genetic differentiation	14
1.2	Phenotypic variation via epigenetic differentiation	15
1.3	DNA methylation in bacteria	17
1.4	Methods for studying genomic and epigenetic variation in bacterial populations	22
1.4.1	Next-generation-sequencing platforms	22
1.4.2	Whole genome assemblies	23
1.4.3	Genome assembly methods	24
1.4.4	Assessing genome accuracy	24
1.5	Methods for investigating DNA methylation patterns in bacteria	25
1.6	Aims and chapter overview	27
Chapter 2	**Complete Genome sequences of 47 natural isolates of *E. coli***	29
2.1	Preface	30
2.2	Abstract	30
2.3	Introduction	31
2.4	Materials and Methods	32
2.4.1	Sequencing read simulation	32
2.4.2	DNA extraction - Promega Wizard kit	33
2.4.3	DNA extraction - Phenol Chloroform	33
2.4.4	DNA shearing	33

2.4.5 Illumina DNA sequencing 33
2.4.6 Nanopore DNA sequencing 34
2.4.7 Nanopore reads preprocessing and quality control 34
2.4.8 Read length analysis 34
2.4.9 Genome assembly 34
2.5 Results 35
2.5.1 Simulated K12 Unicycler assemblies 35
2.5.2 Optimising DNA extraction methods for high molecular weight DNA 36
2.5.3 Optimising Nanopore Rapid Barcoding kit for high molecular weight DNA 38
2.5.4 Nanopore sequencing 39
2.5.5 Genome assemblies 40
2.6 Discussion 43

Chapter 3 Do You Want to Build a Genome? Benchmarking Hybrid Bacterial Genome Assembly Methods 46

3.1 Preface 47
3.2 Abstract 48
3.3 Introduction 48
3.4 Materials and Methods 50
3.4.1 DNA extraction 50
3.4.2 Library preparation and sequencing 50
3.4.3 Genome assembly and polishing 51
3.4.4 Multiple alignment and phylogenetic reconstruction 51
3.4.5 Quality assessment 51
3.4.5.1 Genome fragmentation 51
3.4.5.2 Plasmid assignment 52
3.4.5.3 Illumina mapping discordancy 52
3.4.5.4 rRNA operon orientation 52
3.4.5.5 Truncated ORFs 52
3.4.5.6 Deletions and SNPs 52
3.4.5.7 Pairwise genome alignment 53

3.5 Results 53
3.5.1 Establishing the validity of proposed metrics 53
3.5.2 Genome sequencing and assembly of phylogenetically diverse *E. coli* strains 56
3.5.3 Structural accuracy and consistency 57
3.5.4 Sequence accuracy 61
3.5.5 Effects of data quality on assembly accuracy 64
3.6 Conclusion 68
3.7 Supplementary Data 69

Chapter 4 Growth Condition Dependent Differences in Methylation Implies Transiently Differentiated DNA Methylation States in *E. coli* 81

4.1 Preface 82
4.2 Abstract 83
4.3 Introduction 83
4.4 Materials and Methods 84
4.4.1 Bacterial growth 84
4.4.2 DNA Extraction and Whole Genome Amplification 85
4.4.3 Library preparation and DNA sequencing 85
4.4.4 Identification of methyltransferase 86
4.4.5 Detection of modified sites using Nanodisco 86
4.4.6 Quantification of methylation at individual sites 86
4.4.7 Correlation in methylation fractions 87
4.4.8 Genome wide methylation Patterns 87
4.5 Results 87
4.5.1 Determination of methylation motifs 87
4.5.2 Quantitative analysis of methylation levels 90
4.5.3 Identification of local and global methylation patterns 92
4.6 Discussion 98
4.7 Supplementary data 101

Chapter 5 **Concluding Remarks** 109

References 113

Appendices 124

List of Abbreviations

4mC	N4-methylcytosine
5mC	N5-Methylcytosine
6mA	N6-methyldeoxyadenosine
ACF	auto-correlation function
BP (kbp, Mbp,Gbp)	basepair(s) - (Kilo base pairs, Million base pairs, Giga base pairs)
cAMP	cyclic adenosine triphosphate
CAP	catabolite activator protein
Dam	DNA adenine methyltransferase
Dcm	DNA cytosine methyltransferase
DNA	deoxyribonucleic acid
EPEC	enteropathogenic *Escherichia coli*
FDR	false discovery rate
ICE element	integrative and conjugative elements
Indel	insertion-deletion
IQR	interquartile range
LB	lysogeny broth
MTase	methyltransferase
OD	optical density
ONT	Oxford Nanopore Technologies
ORF	open reading frame
PacBio	Pacific Biosciences
PCR	polymerase chain reaction
RM systems	restriction modification system
RNA	ribonucleic acid
rRNA	ribosomal RNA
siRNA	short interfering RNA
SMRT sequencing	single molecule real time sequencing

SNP	single nucleotide polymorphism
SNV	single nucleotide variant
SSR	simple sequence repeats
UAS	upstream activating sequence
WGA	whole genome amplification

List of Figures

Chapter 1

		Page
Figure 1.1	The ON-OFF state of the pap operon is mediated by competitive binding of Lrp and Dam methyltransferase	20

Chapter 2

Figure 2.1	Assembly contiguity and total assembly length is influenced by read length	36
Figure 2.2	Phenol-cholorform DNA extractions produced longer ONT sequencing reads than Promega Wizard kit extracted DNA	37
Figure 2.3	Recovery of DNA following AMPure bead cleanup was improved by increasing elution time and temperature	39

Chapter 3

		Page
Figure 3.1	Phylogeny of the *E. coli* strains used to assess assembly accuracy	54
Figure 3.2	ONT read length and read quality score distributions for all filtered ONT datasets, ordered by percentage of reads with a quality score greater than 15	58
Figure 3.3	Genome structural contiguity, consistency and accuracy across assemblers	59
Figure 3.4	Genome sequence accuracy metrics across assemblers	62
Figure 3.5	Whole genome alignments for all assemblies for six example strains	64
Figure 3.6	Assembly accuracy is strain dependent and does not correlate with the quality of ONT data	67
Figure S.3.1	Differences in assembly structure between assemblers	69
Figure S.3.2	Illumina data has systematically lower quality scores for a subset of strains	70
Figure S.3.3	No structural accuracy metrics significantly correlate with read length across strains	71
Figure S.3.4	No sequence accuracy metrics significantly correlate with read length across strains	72
Figure S.3.5	Differences in read quality does not significantly correlate with assembly quality across structural metrics	73
Figure S.3.6	Differences in read quality does not significantly correlate with assembly quality across sequence accuracy metrics	74
Figure S.3.7	Differences in read quality for “high quality” and “low quality” MG1655 read sets	75

Figure S.3.8 The relationship between read identity and quality score is consistent across strains 76

Chapter 4

Page

Figure 4.1 Experimental design for sampling native (possibly modified) and unmodified DNA 88

Figure 4.2 The p-values resulting from Mann-Whitney U-tests for signal deviations at DAM and DCM sites are correlated with the fraction of methylated molecules 92

Figure 4.3 The fraction of DAM 6mA and DCM 5mC methylated sites within 10 kbp windows varies according to strain and growth condition 94

Figure 4.4 (A) The fraction of methylated sites in 10kbp windows across the genome is correlated across growth conditions. Pairwise partial correlations in DAM (B) and DCM (C) methylation patterns between all growth environments accounting for genome coverage 95

Figure 4.5 Genome wide patterns in the fraction of methylated sites 97

Figure S.4.1 Correlation between coverage and the Nanodisco-derived p-values 101

Figure S.4.2 Raw nanopore signal distributions on the forward and reverse strands at identical genomic locations of DCM sites that we inferred as methylated (top panels in each pair) or unmethylated (bottom panels in each pair). 102

Figure S.4.3 Identical DAM sites are inferred as methylated or unmethylated across different growth conditions 103

Figure S.4.4 Cumulative distributions of p-values for DAM sites relative to random (unmethylated) sites 104

Figure S.4.5 Cumulative distributions of p-values for DCM sites relative to random (unmethylated) sites 105

Figure S.4.6 Mean-normalised fractions of modified sites across the genome 106

Figure S.4.7 Global autocorrelation plots for DCM methylation 107

Figure S.4.8 Global autocorrelation plots for DAM methylation 108

List of Tables

Chapter 2

		Page
Table 2.1	Genome statistics for all 47 assemblies	40

Chapter 3

		Page
Table 3.1	Percentage of truncated ORFs and total SNVs in different MG1655 assemblies relative to ground truth	55
Table S.3.1	Mutations in the laboratory strain of *E. coli* K12 MG1655	77
Table S.3.2	Summary of Oxford Nanopore sequencing data for each flow cell	77
Table S.3.3	Filtered and Unfiltered ONT read quality	78

Chapter 4

		Page
Table 4.1	Matches between sequence motifs identified by MEME and REBASE Gold methyltransferases	90

Chapter 1

Introduction

In this introduction I first focus on the advantages of phenotypic variation within populations, and more specifically, non-genetic phenotypic variation (i.e. phenotypic variation between isogenic individuals). This provides the general context for the thesis and sets the stage for Chapter 4, which focuses on non-genetic phenotypic differentiation. Chapters 2 and 3 of this thesis act as a foundation for Chapter 4, where I establish the methods and resources required for the study performed in Chapter 4. I introduce this content in the latter half of the introduction.

1.1 Phenotypic variation via genetic differentiation

Natural selection acts on phenotypes. Phenotypes can vary from the macro-scale (eg body size or coat colour) to the micro-scale (eg cell shape, or protein concentration) and can be determined by both genetic and epigenetic factors. Phenotypes can manifest at the level of the individual, or at the level of the population. Phenotypic variation can increase a population's establishment success and long term survival and therefore be an evolutionary advantage.

Phenotypic variation can be established and maintained in three distinct ways: fixed due to genetic differentiation, plastic phenotypic variation due to responses to environmental cues, and stochastic variation, which occurs without regard to environmental cues and is due to random events during the development or differentiation of individuals (Schlichting and Pigliucci 1998; Pigliucci 2001; Kussell and Leibler 2005; Smits, Kuipers, and Veening 2006). Fixed phenotypic variation between population members results in a range of stable phenotypes maintained throughout an individual’s life and is typically associated with genotypic variation within a population (Bull 1987; Farine, Montiglio, and Spiegel 2015; Lande 1976).

Because bacterial genomes are highly dynamic, frequently gaining and losing chromosomal and extrachromosomal elements, genetic differentiation can be a major driver of phenotypic differentiation (Arber 2000). The most rapid genome dynamics are often observed for repetitive elements (e.g. repetitive extragenic palindromic sequences (rep), IS elements, ICE elements, microsatellite, and homopolymeric regions, or plasmids) (Partridge et al. 2018; Redondo-Salvo et al. 2020; Baltrus 2013; Dimitriu, Matthews, and Buckling, n.d.; Kaufman et al., n.d.). This phenomenon can occur through mutation or through the movement of mobile

genetic elements within and between genomes known as horizontal gene transfer, or HGT. HGT encompasses a range of mechanisms including the transformation of naked DNA from the environment, virus-mediated transduction, or conjugation between two bacterium (Soucy, Huang, and Gogarten 2015; Norman, Hansen, and Sørensen 2009; Johnston et al. 2014). In all cases, the transferred DNA is either incorporated into the chromosome via recombination or maintained in a plasmid form. Genomic structures involved in horizontal gene transfer are collectively referred to as mobile genetic elements. Mobile genetic elements are a ubiquitous feature of bacterial genomes, and most bacterial genomes contain and actively use multiple types of mobile genetic elements. The horizontal transfer of genetic material mediated by mobile genetic elements has been associated with the dissemination of antibiotic resistance, restriction-modification systems, and other genes associated with overcoming environmental challenges (Partridge et al. 2018; Tóth et al. 2021; Birkholz et al. 2022).

Mobile genetic elements play a pivotal role in the establishment of genotypic and therefore phenotypic variation in bacteria. Therefore, tracking the movement of mobile genetic elements throughout populations is important for understanding population genetic processes such as the response to environmental pressures and the action of selection.

1.2 Phenotypic variation via epigenetic differentiation

In addition to genomic variation, bacterial population diversity can be established and maintained via epigenetic variation. Epigenetic variation can be defined as heritable changes to phenotype which are not caused by changes to the DNA sequence (Waddington 1942; Holliday 2006; Jablonka and Lamb 2006; Jablonka and Raz 2009). Epigenetic variation can be further broken down to make divisions between epigenetic mechanisms, which are the processes which generate and maintain variable phenotypes (Nanney 1958; Jablonka and Raz 2009), and epigenetic inheritance, the processes in which the variable phenotypes are transmitted from parent to offspring (Heard and Martienssen 2014; Jablonka and Lamb 2006; Jablonka and Raz 2009). For example histone modifications in eukaryotic organisms can be considered epigenetic mechanisms, as they are non-genetic changes which alter gene expression (C. T. Wu and Morris 2001), however they are not necessarily epigenetically inherited, i.e. carried across generations.

Multiple molecular mechanisms exist for the establishment of epigenetic variation including self sustaining feedback loops, variable protein structures such as prions, and covalent DNA modifications (Jablonka and Raz 2009; Heard and Martienssen 2014; Veening, Smits, and Kuipers 2008). Covalent DNA modification describes the addition of alternate molecules to

DNA, usually at the nucleotide base. The best understood form of covalent DNA modification is DNA methylation by methyltransferases although other forms have been identified (Casadesús and Low 2006; Heard and Martienssen 2014; Jablonka and Raz 2009; X. Wu et al. 2020). DNA methylation has been observed in all kingdoms of life. In eukaryotes, DNA methylation most commonly occurs on the cytosine residue of CpG islands and has been implicated in genomic imprinting and gene silencing (Bannister and Kouzarides 2011; Casadesús and Low 2006; Wilson and Murray 1991; Wion and Casadesús 2006). In bacteria, DNA methylation is ubiquitous and most often occurs at cytosine or adenine residues located within specific methyltransferase recognition sites. Each methyltransferase catalyses the transfer of methyl groups to the DNA from the universal methyl donor S-adenosyl-methionine (AdoMet) at specific nucleotide motifs (Gormley et al. 2005; Jeltsch 2002). Many methylation motifs are palindromic, leading to methylation of both the forward and reverse strand (Jablonka and Raz 2009). Due to the semi-conservative nature of DNA replication, this allows for transmission of methylation patterns through successive generations and the establishment of heritable epigenetic variation (Allshire and Selker 2007; Genereux et al. 2005; Henderson and Jacobsen 2007; Jablonka and Raz 2009).

In addition to differences in molecular mechanisms, there are differences in the proximate causes. The first of these is phenotypic plasticity. Phenotypic plasticity is defined as the ability for a single genotype to express different phenotypes in response to environmental conditions (Schlichting and Pigliucci 1998; Pigliucci 2001). Phenotypic plasticity as an evolutionary strategy relies on a population's ability to sense and respond to environmental cues. The system depends on reliable environmental cues and that environmental fluctuations are timed in such a way that reduces the cost of phenotypic lag. Despite these costs, the adoption of a plastic phenotype allows for a greater proportion for the population to be adapted to the environmental conditions (Dewitt, Sih, and Wilson 1998; Auld, Agrawal, and Relyea 2010; Xue and Leibler 2018). Phenotypic plasticity has been observed in all kingdoms of life, which is evidence of its success as an evolutionary strategy for adapting to environmental change (Whitman, Ananthakrishnan, and Others 2009; West-Eberhard 1989; Sommer 2020). While observed in higher orders, phenotypic plasticity is particularly common in bacterial populations where it offers the opportunity for otherwise isogenic populations to establish and maintain heterogeneity on a population scale.

Finally, phenotypic variation via epigenetic changes can occur via stochastic mechanisms. This leads to individuals with identical genotypes and residing in identical environments to have different phenotypes. This differs from plasticity in that the changes are not a response to an environmental signal, but due only to random fluctuations in the concentrations of specific molecules within cells (Kussell and Leibler 2005; Rainey et al. 2011) . In some

cases, stochastic phenotypic differentiation can increase a population's fitness. There are two well-studied evolutionary strategies that provide insight into the population-level fitness advantages of stochastic phenotypic differentiation: bet-hedging and the division of labour (de Jong, Haccou, and Kuipers 2011). The division of labour describes the partitioning of tasks between individuals such that co-ordinated groups of specialists more efficiently perform each task, and the population can achieve more together than the sum of their parts (Traxler and Rozen 2022; Kearns 2008). In contrast, bet-hedging describes an evolutionary strategy that involves the maintenance of varied phenotypes that are not necessarily optimally adapted to the current environment (Jan Grimbergen et al. 2015; Veening et al. 2008; de Jong, Haccou, and Kuipers 2011; Kussell and Leibler 2005). Although not suited to the current environment, they may be better suited to future environments, offering the population an adaptive advantage when environmental change occurs.

As mentioned above, there are different molecular mechanisms driving epigenetic differentiation. One of the primary mechanisms in bacteria is through covalent modifications of DNA (but not changes in the sequence of base pairs). These covalent modifications are most often methylation of different bases. Below, I cover this mechanism in more detail.

1.3 DNA methylation in bacteria

Bacterial DNA methylation was first identified as part of restriction-modification (R-M) systems where a restriction enzyme and methyltransferase target the same site and methylation of this site prevents restriction. R-M systems can act in self versus non-self recognition, protecting bacteria from phage and other incoming DNA. Incoming foreign DNA is unlikely to be modified at these sites and is therefore a target for restriction (Løbner-Olesen, Skovgaard, and Marinus 2005; D. A. Low, Weyand, and Mahan 2001; Reisenauer et al. 1999; R. J. Roberts et al. 2003; Gormley et al. 2005; Atack et al. 2018; Thomas A. Bickle 2004; Vasu and Nagaraja 2013; Wilson and Murray 1991). Methyltransferases also exist outside of R-M systems. These are termed orphan methyltransferase. Orphan methyltransferase have been implicated in a range of cellular functions which have the capacity to influence phenotype, including the regulation of DNA replication, guiding DNA repair, the timing of the cell cycle, and regulation of gene expression (Atack et al. 2018; Camacho and Casadesús 2005; Collier 2009; Løbner-Olesen, Skovgaard, and Marinus 2005; David A. Low and Casadesús 2008; Marinus and Casadesus 2009). The most well studied orphan methyltransferase are DAM and DCM, found in *E. coli*. Many bacterial species contain multiple methyltransferase, either as part of a R-M system or as an orphan methyltransferase (Vasu and Nagaraja 2013; R. J. Roberts et al., n.d.; Gormley et al. 2005; Kong et al. 2000). DNA methyltransferase are highly varied and are

able to spread between strains via horizontal gene transfer, leading to huge diversity in the type and number of DNA methyltransferase carried by bacterial species. This diversity and prevalence of DNA methylation is evidence of the R-M systems success as a defence mechanism as well as the utility of the additional roles methylation has been found to play in bacterial gene regulation (Casadesús and Low 2013; Blow et al. 2016; Atack et al. 2018; T. A. Bickle and Krüger 1993; Pingoud et al. 2005; Vasu and Nagaraja 2013; R. J. Roberts et al., n.d.).

When bacterial DNA methylation occurs at cytosine and adenine nucleotides, three different types of modifications can occur (Wilson and Murray 1991). Cytosine residues can be methylated at either the C5 position forming C5-methyl-cytosine (5mC) or the N4 position to produce N4-methyl-cytosine (4mC). In contrast, a single adenine methylation structure has been observed, C6-methyl-adenine (6mA) formed through methylation of the C6 position. Of these three structures observed in bacteria, 5mC and 6mA are also found in eukaryotes, while 4mC is thought to be unique to bacteria (Blow et al. 2016; Sánchez-Romero, Cota, and Casadesús 2015).

Bacterial DNA methylation patterns are inherited due to the semi-conservative nature of DNA replication. Hemi-methylated DNA inherited by daughter cells serves as a template for methyltransferase to re-establish a fully methylated state in an accurate manner ensuring the prolonged maintenance of methylation patterns (Jablonka and Raz 2009). While heritable, this type of DNA methylation does not always lead to different phenotypes, meaning that the maintenance of these DNA methylation patterns can not be termed epigenetic (Collier 2009; Skarstad, Boye, and Steen 1986; Waldminghaus and Skarstad 2009). However, the importance of methylation determining phenotypic states within a cell's lifetime has been well established and is frequently linked with DNA replication (Marinus and Casadesus 2009).

One such example is the regulation of DNA replication initiator protein DnaA in *Caulobacter crescentus* and *E. coli* cells. The *dnaA* promoter is most active when fully methylated, but its location next to the origin of replication means it rapidly becomes hemi-methylated after the initiation of DNA replication (Collier 2009). This causes rapid down-regulation of DnaA activity, limiting the reinitiation of DNA replication. In *C. crescentus,* remethylation is prevented until the *ccrM* methyltransferase re-accumulates. This occurs towards the end of the replication cycle due to the regulation of ccrM expression by the ctrA protein. In *E. coli* DnaA remethylation is prevented by SeqA which preferentially binds the hemi-methylated MTase binding site (Collier 2009; Waldminghaus and Skarstad 2009).

A second example of DNA replication linked establishment of replication patterns is the *tra* operon in *Salmonella enterica (*Sánchez-Romero, Cota, and Casadesús *2015).* The *tra* operon is located on a virulence plasmid, and expression is required for the production of conjugative pili as well as products of the DNA transfer system (David A. Low and Casadesús 2008). The plasmid must be activated by the transcription factor TraJ which is controlled by global regulator Lrp. Lrp binds to two sites in the TraJ UAS region, one of which also contains a target site for DNA methylation. When the Lrp site is methylated, Lrp is unable to bind and TraJ transcription is prevented (Camacho and Casadesús 2005). As observed in other cyclic methylation regulation examples, passage of the replication fork results in hemi-methylation of the Lrp binding site and a window of opportunity for Lrp to bind at both of the TraJ UAS binding sites on one copy of the replicated plasmids, therefore inducing *tra* operon expression (Marinus and Casadesus 2009). It is suggested that coupling expression of the *tra* operon with DNA replication, and restricting expression to only one of two daughter plasmids is useful to reduce energetically wasteful gene expression and protein production (Camacho and Casadesús 2005).

In contrast to DNA methylation controlling gene expression through the methylated/hemi-methylated cycles produced during DNA replication. The pattern of methylated and non methylated target sites throughout the genome have been shown to regulate gene expression in a phenotypically heritable manner (David A. Low and Casadesús 2008; Totsika et al. 2008). DNA methylation patterns are established through competition for binding sites with alternative DNA binding proteins such as Lrp, OxyR, GutR and H-NS (Casadesús and Low 2006; D. A. Low, Weyand, and Mahan 2001; van der Woude, Hale, and Low 1998). One example is the Dam methyltransferase mediated ON-OFF switching of the *E. coli pap* operon.

The *pap* operon of uropathogenic *E.coli* regulates the expression of P pili which control adhesion of cells to the mucosa of hosts (Casadesús and Low 2006). The upstream regulatory region of the *pap* operon contains six binding sites for the global regulator Lrp. Two of these sites, the distal and proximal sites relative to the promoter, also contain binding sites for Dam methyltransferase (Peterson and Reich 2006, 2008; van der Woude, Hale, and Low 1998). When the pap operon is off, Lrp is bound to the proximal binding site, preventing DNA methylation at this site. In contrast Lrp has a lower affinity for the distal binding site and methylation is able to concur here, further preventing Lrp binding. Two rounds of DNA replication with Lrp bound at the proximal site results in completely unmethylated DNA at this position. Lrps high affinity for the proximal site, coupled with DNA replication establishes a feedback loop which maintains the operon in the off state. In order for the operon to be activated, Lrp must bind at the distal site. This requires the formation of a PapI/Lrp complex

which has increased affinity for the distal site. PapI/Lrp binding at the distal site, enables methylation of the proximal site therefore maintaining this conformation. Once the operon is activated, PapB is produced which activates PapI expression therefore creating a feedback loop which maintains the on state of the operon (**Figure 1.1.**) (Peterson and Reich 2006, 2008; Totsika et al. 2008; van der Woude, Braaten, and Low 1996; van der Woude, Hale, and Low 1998). The ON-OFF state of the pap operon, mediated by the location of DNA methylation within the promoter region is a clear example of the effects of DNA methylation on phenotypes and the way in which DNA methylation patterns can act as a gene expression switch.

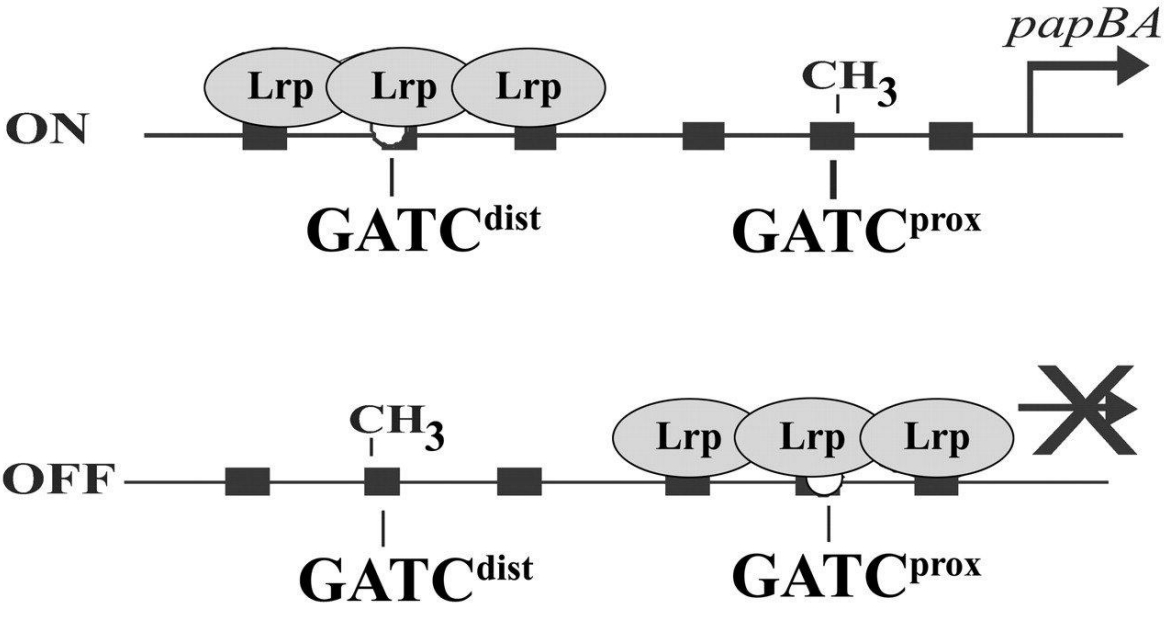


Figure 1.1. The ON-OFF state of the pap operon is mediated by the competitive binding of Lrp and Dam methyltransferase. The ON state is defined by the methylation of the proximal Dam binding site, while the OFF state requires methylation of the distal Dam binding site DNA methylation Taken from (Hernday et al. 2002)) Copyright 2002 National Academy of Sciences

DNA modification patterns can be influenced by environmental conditions. However, whether changes to DNA modification patterns driven by the environment *cause* changes in transcription or are simply an *effect* of changes in transcription that occur as a response to the environment is not always easy to establish. One example where the environmental conditions *cause* changes to transcription is the regulation of the pap operon described above (van der Woude, Braaten, and Low 1996). The switch from the OFF to ON state of the pap operon is mediated by the carbon source available in the environment, signalled by the concentration of catabolite activator protein (CAP) molecules within the cell (Hernday et al. 2002). In order to establish the ON state, Dam mediated 6mA modifications in the promoter

region must be released from the proximal binding site and reformed at the distal binding site. Following this translocation, the cAMP-CAP complexes bind the DNA and interact with RNA polymerase acting to enhance *papAB* transcription (Weyand et al. 2001). Therefore, when environmental conditions lead to an increase in CAP concentration through a decrease in the preferred carbon source, papAB transcription is upregulated (Casadesús and Low 2006).

In contrast, changes in methylation patterns can simply be an effect of transcription factor binding or rates of transcription driven by environmental conditions. This is due primarily to the fact that methylation patterns are maintained by the competitive binding of regulatory proteins whose activity can be environmentally controlled. Methylation enzymes can gain access to DNA when the binding state of transcription factors change. One example of this is the glucitol utilisation (*gut)* operon in *E. coli (*van der Woude, Hale, and Low *1998)*. In the *gut* operon, the upstream regulatory region contains a dam methyltransferase binding site. However, the site is typically unmethylated due to the presence of the GutR repressor, which binds the region blocking methylation (van der Woude, Hale, and Low 1998). Therefore the DNA methylation pattern of the inactive gut operon is characterised by non-methylation at this site. The presence of gluticol in the environment releases GutR from the DNA and DNA methylation can occur, thereby changing DNA methylation pattern (van der Woude, Hale, and Low 1998; D. A. Low, Weyand, and Mahan 2001; Casadesús and Low 2006). Whilst DNA methylation at the gut operon is not directly responsible for the transcriptional state of the operon, it is a clear example of the effect environmental conditions can have on the DNA methylation patterns present throughout the chromosome.

Some DNA methyltransferases are found across multiple species with conserved core roles and species specific roles. An example of this is Dam and CcrM which are found across multiple species throughout Gammaproteobacteria and Alphaproteobacteria respectively (D. A. Low, Weyand, and Mahan 2001). Within each class of bacteria Dam and CcrM are widespread. Dam is found across Enterobacteriaceae*,* while CcrM is known to be present in many alphaproteobacteria, including *C. crescentus*, *Sinorhizobium meliloti, and Agrobacterium tumefaciens* (Wright, Stephens, and Shapiro 1997; Kahng and Shapiro 2001). The two enzymes catalyse the production of 6mA residues at the recognition sites, GATC and GANTC, respectively. Both Dam and CcrM carry out the role of regulating DNA replication initiation but have unique evolutionary origins and varied species specific roles (García-Del Portillo, Pucciarelli, and Casadesús 1999; M. E. Watson Jr, Jarisch, and Smith 2004; Julio et al. 2001). In contrast, identical motifs have been shown to be catalysed by unique DNA methyltransferases across species (Fang et al. 2012).

The wide variety of methylase enzymes observed across bacterial species, coupled with the observation that homologous methylase enzymes can have species specific functions, leads to huge diversity in both methylase presence and methylase function across bacteria. To compound this diversity, variation in the activity of methyltransferases within isogenic populations has been shown to occur.

1.4 Methods for studying genomic and epigenetic variation in bacterial populations

Genetic and epigenetic variation are critical to the establishment of phenotypic heterogeneity in bacterial populations. However, studying the mechanisms by which they occur and are maintained is not trivial. As discussed above, phenotypic heterogeneity established through genetic variation is often due to the movement of mobile genetic elements. Detecting the presence or absence of mobile genetic elements within genomes is not sufficient to fully determine their phenotypic effects and their influence on population genetic processes. In order to fully quantify the role of genome dynamics in establishing and maintaining phenotypic heterogeneity in bacteria, whole genome assemblies which are accurate and fully contiguous are required.

1.4.1 Next-generation-sequencing platforms

Producing accurate whole genome assemblies depends on the characteristics and quantity of available sequencing data. The major next-generation-sequencing (NGS) platform, Illumina sequencing, is characterised by its highly accurate reads. The accuracy of Illumina reads is due to the massively parallel amplification of DNA fragments on the sequencing flowcell and the subsequently high coverage for each read (Loman et al. 2012). Although Illumina reads are highly accurate, they are also relatively short, ranging from 75 to 300 bp. In contrast to Illumina, long read sequencing technologies PacBio SMRT and ONT produce significantly longer but more erroneous reads (Lu, Giordano, and Ning 2016; de Lannoy, de Ridder, and Risse 2017; Rang, Kloosterman, and de Ridder 2018). Both platforms routinely produce reads between 10-30 kbp, however ONT read lengths are determined by the length of input DNA fragments and reads have exceeded 1Mbp (Payne et al. 2019). Although the read lengths for both platforms are similar, the sequencing chemistry and therefore error profiles are distinct for each platform. Errors in PacBio reads tend to be randomly distributed and can therefore be mitigated by increased sequencing depth (Wenger et al. 2019). In contrast, ONT errors typically occur within homopolymeric regions and are not randomly distributed throughout the genome (Dohm et al. 2020). Despite this, advancements in ONT basecalling software have taken this into account and super high accuracy basecalling models, combined with the latest flowcells now achieve raw read accuracy of 99.3% (“ONT Accuracy” n.d.). A key advantage of the ONT platform is accessibility, ONT MinION devices

are affordable and accessible. This has led to a large community of active users across a wide range of disciplines, and the rapid development of both ONT and open source ONT specific software. In contrast, PacBio sequencing has a high capital investment and is not accessible to many smaller research groups. In this work, we make use of the Illumina and ONT sequencing platforms and unless otherwise stated further discussion of long and short reads refers to these platforms respectively.

1.4.2 Whole genome assemblies

Despite their accuracy, Illumina sequencing reads alone are usually unable to produce complete bacterial genomes. Mobile genetic elements frequently contain repetitive sequence structures which can be of variable size, and copy number throughout the genome. When Illumina read lengths are shorter than these repetitive regions it is almost impossible to accurately determine the location, length and copy number of mobile genetic elements throughout the genome (Goldstein et al. 2019; Murigneux et al. 2021; Klassen and Currie 2012). The simplest way to resolve this is by increasing read length (Koren and Phillippy 2015). Specifically, read lengths should exceed the length of the longest repeat region contained in a genome. This has been termed the golden threshold. Empirical investigation has suggested that for bacterial genomes the golden threshold is close to 7kbp, as reads over this length would be expected to resolve almost 80% of currently known bacterial genomes (Koren and Phillippy 2015).

Despite routinely producing reads greater than this golden threshold, long-read-only assemblers also struggle to accurately assemble bacterial genomes, due to high error rates in the sequencing data. In contrast to Illumina errors which are rare, and likely to be single nucleotide polymorphisms (SNP), long read errors are predominantly insertions or deletions (Dohm et al. 2020; Y. Chen et al. 2021). This type of error can lead to frame shifts within the assembled genome and therefore an increase in the annotation of non-functional pseudogenes and an inaccurate representation of the internal structure of the genome (M. Watson n.d.).

Combining short and long read data in a hybrid assembly approach has rapidly emerged as a way to capitalise on the strengths of each data type while reducing the effects of their weaknesses (Sović et al. 2016; Utturkar et al. 2014; Liao, Lin, and Lin 2015; Koren et al. 2012; Vasudevan et al. 2019; De Maio et al. 2019). Hybrid genome assembly methods can be classified into two groups depending on the way in which the long reads are used. In the first method, contigs are produced with Illumina sequencing data and then scaffolded together with long reads (Antipov et al. 2016; R. R. Wick et al. 2017a). Alternatively, long

reads can be used alone to generate a genome assembly (Koren et al. 2017; Kolmogorov et al. 2019). Due to the nature of long reads, these assemblies are often structurally complete but contain errors at a per base level. One way to reduce these errors is to align Illumina reads to the assembly and use a consensus approach to improve the assembled genome. This is referred to as assembly polishing and a variety of tools are available for this purpose (Walker et al. 2014; Vaser et al. 2017; Simpson et al. 2017). Long-read-only assemblies, followed by polishing with Illumina reads can be considered a hybrid assembly strategy and can produce high quality genome assemblies.

1.4.3 Genome assembly methods

Within each category of assembly methods, numerous assembly tools have been developed to address the computational problem of producing accurate hybrid bacterial genomes. The popular assembler Unicycler (R. R. Wick et al. 2017a) approaches the assembly problem by using both short and long reads. First a short read assembly is produced using the SPAdes assembler. The contigs of this assembly are then scaffolded into larger, often complete contigs using the long reads. Within Unicycler the assembly is then polished. In contrast, assemblers such as Raven (Vaser and Šikić, n.d.), Flye (Kolmogorov et al. 2019), Canu (Koren et al. 2017) and RedBean (Ruan and Li 2019) produce long-read-only assemblies using an overlap consensus method, which require additional polishing by software such as Pilon (Walker et al. 2014) or Racon (Vaser et al. 2017). Although each assembler uses a different assembly approach, all are capable of routinely producing assemblies containing a single circular chromosomal contig. With an abundance of choice, deciding which assembly method to use should be determined by the accuracy of the final assembly. However, assessing assembly accuracy for single contig assemblies is not trivial. This is particularly true for de novo genome assemblies where no reference genome is available to act as a benchmark.

1.4.4 Assessing genome accuracy

Prior to the development of hybrid assembly methods, short read only assemblies were typically highly fragmented. Assessing the accuracy of these assemblies relies on manual curation and additional independent data which can be costly to obtain (Ghodsi et al. 2013). In practice gene annotation, the consistency of read mapping and particularly contiguity are frequently used as metrics to compare fragmented short read genome assemblies (Phillippy, Schatz, and Pop 2008; Hunt et al. 2013; Earl et al. 2011). Contiguity is often expressed by the N50 value for an assembly. When summing contig lengths, the N50 value represents the shortest contig length which must be included to total greater than 50% of the total assembly length. N50 gives an indication of contig structure, a larger N50 represents an assembly of

generally larger contigs. However, for single contig genome assemblies, the N50 value is equal to the assembly size and offers no information about the assembly structure or accuracy. While contiguity remains a valid measure for assessing the external structure of an assembly, additional methods are required to evaluate internal structural accuracy.

As with short read assemblies, sequencing reads and genome annotation can be used to assess internal assembly accuracy. Both Illumina and Nanopore reads can be mapped onto the assembly to detect mis-assemblies or small mutations such as SNPs, insertions or deletions (Deatherage and Barrick 2014). Genome annotation can also be used to infer mis-assembly by comparing the presence and length of assembled protein coding regions with known database sequences. Assembly proteins that are shorter than the database sequence suggest mis-assembly (M. Watson n.d.).

Additionally, the annotation and arrangement of known genomic structures such as the rRNA operons can be used to assess an assembly's internal structure (Page, Ainsworth, and Langridge 2020). Unbiologically viable arrangements of these regions are a strong indicator of a misassembly.

Each assembler approaches the assembly problem uniquely and has strengths and weaknesses. Some of these can be mitigated by improving the long read dataset, such as increasing read length or coverage, while others are inherent to the assembly algorithm. The strengths and weaknesses of assembly methods can be chosen to suit downstream analysis. For example, evolve and resequence experiments rely on high per base pair accuracy while quantification of genome dynamics relies on structurally accurate assemblies. However, in all cases an assembly should be assessed for contiguity and internal accuracy, and completeness should be evaluated across both metrics.

1.5 Methods for investigation of DNA methylation patterns in bacteria

Prior to the development of ONT sequencing, methods for detecting DNA methylation included lab based methods such as Restriction-inhibition assays, or bisulfite sequencing (Frommer et al. 1992; Y. Li and Tollefsbol 2011). Restriction-inhibition assays use the protection from restriction conferred by methylation to identify methylate motifs. In contrast, Bisulfite sequencing refers to the treatment of DNA with bisulfite prior to sequencing. Bisulfite converts cytosine to uracil, but does not affect 5-methylcytosine residues. Sequence reads can then be compared to reads from untreated DNA to infer sites of methylation (Krueger et al. 2012). Each of these methods has limitations for studying genome wide methylation patterns. Restriction-inhibition assays, are time consuming and can have high rates of false negatives (Srikhanta et al. 2009; Atack et al. 2018) while bisulfite sequencing is

only able to detect 5mC (Feng et al. 2020; Jaenisch and Bird 2003; Krueger et al. 2012). Although 5mC modifications are present in bacteria, unlike eukaryotes they are not the dominant modification and most bacterial DNA methylation is 6mA.

The development of high throughput long read technology such as SMRT sequencing and Nanopore sequencing, has improved the identification of methylated motifs throughout genomes and is allowing further development of the field of methylomics (Clark et al. 2012; Flusberg et al. 2010; Rand et al. 2017; Simpson et al. 2017). A recent study on the human pathogen Listeria monocytogenes used SMRT sequencing to establish closed genomes for 15 isolates of the strain, these were used to identify the presence of DNA methyltransferase genes and SMRT sequencing data was used to identify the methylation patterns established in each strain (P. Chen et al. 2017). Studies on 3 strains of the human and sheep pathogen *Campylobacter jejuni* also used SMRT sequencing to identify differences in methylation patterns. They suggest that the different methylation patterns may be responsible for the different disease patterns caused by each strain (Mou et al. 2015). Further work carried out on uropathogenic *E. coli* strain EC958, within the sequence type 131 (ST131) used SMRT sequencing to define the complete methylome of the strain.They were able to identify three novel, ST131 methyltransferases, and show that methyltransferase presence and expression varies amongst ST131 strains (Forde et al. 2015). Despite this, there is some evidence that SMRT can only reliably detect 6mA modifications (Timp 2018). In contrast to these approaches, ONT sequencing allows for the detection of any DNA modification from sequencing data (Rand et al. 2017; Stoiber et al. 2016).

ONT sequencing uses nanopore technology to sequence DNA. Native DNA fragments are passed through a protein nanopore embedded in a synthetic polymer membrane. A current maintained across the membrane and translocation of DNA strands through the pore cause characteristic voltage changes directly dependent on the sequence of bases passing through the pore (de Lannoy, de Ridder, and Risse 2017). This voltage pattern is referred to as the “squiggle” and distinct squiggle patterns occur depending on the local sequence context (the kmer within the pore as well as some bases on either side of the pore) (Jain et al. 2016). Preparation of ONT sequencing libraries does not require DNA amplification, therefore DNA can be sequenced in its “native” or unamplified state. When native DNA is sequenced, some bases may contain covalent modifications which can cause a voltage change distinct from unmodified bases (Rand et al. 2017; Simpson et al. 2017).

A number of software packages are available for the detection of methylated bases, with two primary approaches. Some tools rely on parameterized models of the squiggle from modified and unmodified nucleotides. On the basis of this, they then assign the nucleotides as being

modified or unmodified. These methods have proved successful where models are well trained on sufficiently high quality datasets (“Tombo” n.d.; Simpson et al. 2017; Ni et al. 2018; Liu et al. 2019; Bonet et al. 2021; Tourancheau et al. 2021; “Megalodon” n.d.). The majority of this work to date has focused on the development of tools and models for the detection of eukaryotic DNA methylation. In bacteria, models for the detection of 5mC and 6mA have been developed but are most reliable only within the context of specific methyltransferase recognition sites. This limits the utility of model-based approaches for the detection of potentially unknown or novel methyltransferase. In contrast, it is possible to infer methylation by comparing the squiggle data generated from native, methylated DNA, with unmodified DNA (Tourancheau et al. 2021; “Tombo” n.d., “Megalodon” n.d.). To apply this on a genome wide level requires either enzymatic removal of methylation marks or whole genome amplification, which will produce unmodified DNA when performed in the absence of methyltransferases.

To maximise the accuracy of detecting modified and unmodified bases, complete genome assemblies are useful. Complete genome assemblies ensure accurate assignment of sequence reads to the genome, allowing the squiggle signals from different sequence reads to be accurately compared. Establishing complete genomes also facilitates the investigation of relationships or linkages between methylation patterns in different parts of the genome.

1.6 Aims and chapter overview

This thesis aims to expand understanding of the establishment and maintenance of both genomic and epigenetic variation in natural isolates of *E. coli*. In chapter 2 we use in silico techniques to investigate the read length and coverage requirements of long read sequencing data in hybrid genome assembly. We then establish methods for the routine extraction of high molecular weight DNA and the preparation of Oxford Nanopore sequencing libraries. We then use these methods to sequence 47 natural isolates of *E.coli* and produce complete hybrid genome assemblies. In chapter 3 we use the datasets produced for each of the 47 natural isolates to perform a broad scale benchmarking investigation of five hybrid assembly methods. We establish metrics to assess each assembly for contiguity and accuracy on both a per bp and structural level. Using the established assembly and benchmarking pipeline produce high quality de novo genomes for each natural isolate. These genome assemblies are essential for the study of genetic and epigenetic variation in highly dynamic bacterial genomes. Finally, in chapter 4 we use these genomes to predict methyltransferase in each of 3 natural isolates of *E. coli.* We then establish methods to detect the activity of predicted methyltransferase using Oxford Nanopore sequencing technology. We then use this approach to profile genome wide

methylation patterns for the methyltransferases Dam and Dcm across a range of growth conditions. We find predictable variations in methylation patterns which are correlated with growth conditions suggesting distinct methylation states associated with different growth states.

Chapters 2 - 4 are presented as separate publications, with an introduction, methods and discussion section within each chapter. Therefore, there will be some overlap of the concepts and references presented in these chapters. Finally, chapter 5 provides a summary and concluding remarks across the entire thesis.

Chapter 2

Complete Genome Sequences of 47 Environmental Isolates of Escherichia Coli

Georgia Breckell, Olin K. Silander

Article published after peer-review

Microbiology Resource Announcements (Vol.9,No.38, 2020)

Author contributions:

Georgia Breckell: Conceptualization (equal); Methodology (lead); Investigation (lead); Visualisation (lead); Writing-original draft (lead); Writing-review & editing (equal).
Olin K. Silander: Conceptualization (equal); Methodology (supporting); Funding acquisition(lead); Resources (lead); Supervision (lead); Writing-original draft (supporting); Writing-review & editing (equal).

2.1 Preface

Parts of the following chapter have been published following peer review in the journal Microbial Resources Announcements on the 17th September 2020. (DOI:https://doi.org/10.1128/MRA.00222-20)
In order to integrate the results presented in this paper into this thesis, the formatting has been changed significantly from what can be found online. The paper formatted as published is attached in Appendix 1.

Olin and I conceived this project and designed the experiments. I carried out all the sequencing and produced each of the assemblies for this project. This manuscript was written as a collaboration with Olin.

2.2 Abstract

Hybrid genome assemblies, produced with a combination of Illumina short read and long read sequencing data are routinely able to produce complete bacterial genome assemblies. In contrast, neither Illumina reads or Oxford Nanopore long reads are able to produce an accurate assembly alone. Short read Illumina data frequently results in highly fragmented assemblies, due to the presence of repetitive structures in bacterial genomes which are longer than the reads. While long read sequencing reads frequently span these regions, the sequencing data contains many more errors resulting in inaccurate assembly. Combining sequencing data from each platform capitalises on the strengths of each platform while mitigating the effects of their weaknesses. However the coverage and read length requirements for long read sequencing data in the production of accurate hybrid genome assemblies is not well established. Here we use simulated Nanopore and Illumina datasets to explore the effects of read length and coverage on hybrid assembly structure. We show that read length has an effect on contiguity. We found that datasets containing reads greater than 10Kb can produce complete genome assemblies with as little as 5X coverage. We then explore lab based methods to extract DNA and prepare sequencing libraries which will routinely produce reads of this length. We use these techniques to produce and sequence high molecular weight DNA for 47 natural isolates of *E. coli*. Finally, we use the long read assembler Flye, and polish each assembly with Racon and Pilon to produce a complete hybrid genome assembly for each isolate.

2.3 Introduction

The primary focus of this thesis is the investigation of DNA methylation and its potential role in establishing and maintaining phenotypic heterogeneity in bacteria. In this thesis, we use a collection of natural isolates of *E. coli* isolated from the shores of the St Louis River, near Lake Superior (Ishii et al. 2006). In order to investigate DNA methylation in these natural isolates, accurate reference genomes are required. Although high quality reference genomes are available for many lab strains of *E.coli*, the isolates used in this study had not been assembled prior to this work and reference genomes were not available.

Reference genomes facilitate the study of DNA methylation in a variety of ways. First they allow for the identification of methyltransferase genes which are present in each isolate. While not all methyltransferase genes will result in an active enzyme, activity of methyltransferase can be identified by mapping sequencing reads containing methylation data to a reference genome. In addition to confirming the activity of predicted methyltransferase, mapping methylation data to reference genomes also allows for the identification of putative methyltransferase. The genomic context of modified bases can be used to identify candidate modification enzymes which can be used to confirm modifications as DNA methylation. Finally, mapping methylation sequence reads to a reference genome allows for the identification of genome wide methylation patterns, and the investigation of any correlation between methylation patterns across the chromosome.

In order to be used as a reference genome an assembly must accurately represent both the genomic structure (i.e. both the correct number of chromosomal and plasmid contigs, and the correct arrangement of elements) and the genomic sequence (i.e. few errors in genomic sequence) of the isolate. *E. coli* is well known for its dynamic genome, with prolific rates of horizontal gene transfer leading to the loss and gain of mobile genetic elements such as insertion sequences, rep sequences, and plasmids. These mobile genetic elements are known for their repetitive sequences which frequently leads to mis-assembly, particularly for de novo genomes.

To produce an accurate genome assembly, a hybrid approach using both short and long reads has emerged as the most reliable method (R. R. Wick et al. 2017a; De Maio et al. 2019; Hernandez-Beeftink et al. 2018; Goodwin et al. 2015; Rhoads and Au 2015; Sović et al. 2016; Bashir et al. 2012). This approach combines the strengths, and mitigates the weaknesses of each sequencing platform. Illumina reads are highly accurate on a sequence level and standardised in both read length and coverage. However, most Illumina datasets contain reads which are less than 300 bp long.These reads are shorter than the repetitive elements found throughout bacterial genomes (Goldstein et al. 2019). This makes it difficult

to accurately determine the correct placement, length and number of repetitive regions in the genome (Koren and Phillippy 2015; Goldstein et al. 2019; Treangen and Salzberg 2011; Pevzner et al. 2004). In contrast, ONT long reads have higher per base pair error rates, but read lengths are frequently longer than the repetitive regions found in bacterial genomes (Y. Chen et al. 2021; de Lannoy, de Ridder, and Risse 2017). ONT sequencing datasets are directly representative of the extracted DNA and can vary substantially in both coverage and read lengths. This raises the question of how much Oxford Nanopore sequencing data is required to complement Illumina reads and routinely produce a complete assembly.

In this chapter we use in silico analysis to explore the effects of ONT datasets read length and coverage in producing complete hybrid genome assemblies. We establish that datasets with read lengths in excess of 10Kb will routinely produce complete assemblies with as little as 5X coverage, and that coverage requirements increase as read lengths decrease. In order to ensure long read lengths we then establish DNA extraction methods which will routinely extract high molecular weight DNA. To avoid DNA shearing we avoid magnetic bead and spin column based extraction kits and instead test protein precipitation methods. We extract high molecular weight DNA for 47 natural isolates of *E. coli*. We then optimise library preparation protocols to ensure sufficient DNA is retained throughout the library prep for successful sequencing and assembly. Finally we report the sequencing results and the hybrid genome assemblies produced.

The work presented in this chapter lays the groundwork for future chapters. In chapter three we use the sequencing data produced here to evaluate genome assembly methods ensuring accurate assemblies are used as reference genomes. In chapter four we use those assemblies as reference genomes to investigate DNA methylation in a selection of isolates.

2.4 Materials and Methods

2.4.1 Sequencing read simulation

The MG1655 reference genome was obtained from NCBI (NC_000913.3) and dwgsim v0.1.11 (Homer n.d.) was used to simulate Illumina reads. We generated four million 250 bp paired end reads with an error rate uniformly increasing from 0.01% at the start of each read to 0.1% at the end of the read. Nanopore reads were simulated using Nanosim v2.1.0 (C. Yang et al. 2017) with the same MG1655 reference genome as used for the Illumina read simulation. For each read length dataset, reads equivalent to 150x coverage were generated using the inbuilt error profile for *E. coli* reads. Hybrid assemblies of simulated data were produced using Unicycler v0.4.4 (R. R. Wick et al. 2017a) with default settings. We used

Quast (Gurevich et al. 2013) to compare each assembly to the MG1655 reference genome and report on the number of contigs and total assembly length of each assembly.

2.4.2 DNA extraction - Promega Wizard kit

3ml of liquid LB was inoculated with a single colony of each strain and incubated shaking at 37ºC overnight. 1 ml of this culture was used for DNA extraction with the Promega wizard DNA kit following the manufacturer's protocol, with the following changes: Following addition of the protein precipitation solution, centrifugation at 13,000g was repeated (step 12) to ensure complete removal of protein from the sample, additionally steps 16 and 17, washing the DNA in ethanol and centrifugation between washes was repeated to improve the purity of the resulting DNA. Furthermore DNA was rehydrated into water rather than the provided elution buffer to avoid EDTA contamination in downstream uses.

2.4.3 DNA extraction - Phenol Chloroform

5 ml of liquid LB was inoculated with a single colony of each strain and incubated shaking at 37ºC overnight. An adapted version of the Josh Quick phenol extraction DNA protocol (Quick 2018) was then performed with the entire 5ml overnight culture. The original protocol calls for 50 ml of starting culture, however as 5 ml was used all other volumes throughout the protocol were scaled appropriately. We made other minor changes to the protocol including not submerging the coiled DNA “jellyfish” into ethanol as we found this made removing the DNA from the glass hook unnecessarily difficult. As we were extracting large volumes of DNA, DNA was initially rehydrated into 100 ul of water, but additional water was added as and where necessary to produce a homogenous solution. Triton X to a final concentration of 0.02% was added to the final, homogenous solution.

2.4.4 DNA shearing

Two DNA extractions were diluted 10 fold to a total volume of 100ul, from this 25ul was used for each pipette and needling shearing trial. Using a P10 pipette or a 1ml syringe fitted with a 23 gauge needle we passed the DNA solution through the pipette tip 40 times and through the syringe and needle 10, 20 or 40 times. Following shearing, 10ul of each sample was run on a 0.7% agarose gel at 100 volts for 90 minutes with a Lambda DNA ladder.

2.4.5 Illumina DNA sequencing

Illumina sequencing was performed by MicrobesNG. DNA was extracted using the Promega Wizard DNA extraction kit as described above. DNA concentration was standardised to between 40-100ng/ul in a 50 ul sample volume and 250 bp paired end sequencing was

performed for each isolate. Reads were quality controlled by MicrobesNG and we obtained at least 30-fold coverage of 2x250 bp paired end reads for each genome.

2.4.6 Nanopore DNA sequencing

Sequencing was performed using a MinION Mk1B device and R9.4 chemistry. DNA was prepared using the Rapid Barcode Sequencing Kit with DNA barcoding (SQK-RBK004) with between six and twelve strains on each flow cell. The manufacturer's instructions were followed except for the following modification. As multiple samples were pooled, the suggested AMPure bead cleanup was performed on the pooled library following barcode ligation. The AMPure bead protocol was modified to maximise DNA recovery, a 0.7:1 ratio of beads was used to maximise the length of recovered DNA and DNA was eluted into 10mM Tris-HCl pH 8.0 with 50mM NaCl preheated to 50ºC for 10 mins at 50ºC. Sequencing was performed using MinKnow v1.1.15

2.4.7 Nanopore read preprocessing and quality control

Reads sequenced to investigate the effect of extraction method on read length were basecalled by the Albacore ONT basecaller v2.2.7 and were processed by Poretools to convert the fast5 files into the fastq format. Albacore's inbuilt demultiplexing and barcode removal was run and the output of this passed to the third party program Porechop (R. Wick n.d.). Porechop was set to flag and remove reads containing middle adapters, remove barcodes, demultiplex reads and remove an additional 31 base pairs from the end of each read. It is known that due to the barcoding transposase activity a specific 31 bp sequence is present at the end of each read.

Reads sequenced for producing assemblies were basecalled and demultiplexed with Guppy v2.3.7. For all strains with more than 500 Mbp of sequence data, we used Filtlong v0.2.0 (https://github.com/rrwick/Filtlong) to retain only 500 Mbp in total, prioritising read quality over length with the following parameters; min length set to 1000, mean quality weight set to 10 and split set to 500.

2.4.8 Read length analysis

Read length statistics were generated for each dataset with the Stats command from SeqKit (Shen et al. 2016) and visualised in R.

2.4.9 Genome assembly

We used the long read assembler Flye v2.4.2 (Kolmogorov et al. 2019) for genome assembly of our natural isolates. We polished the assemblies using four rounds of long-read polishing with Pilon v1.23 (Walker et al. 2014) followed by two rounds of short-read

polishing with Racon v1.3.2 with the following parameter changes; gap penalty increased to -8, match score increased to 8, mismatch score increased to -6 (Vaser et al. 2017). The parameter changes in Racon were in accordance with ONT guidelines for passing Racon output to Medaka. The contigs were left as linear if not circularised by Flye, with no re-orientation. We confirmed the structural accuracy of each genome using socru v2.1.7 (Page, Ainsworth, and Langridge 2020) to assess the order and orientation of the seven rRNA operons (**Table 1**). All software was run using default parameters, unless otherwise specified.

2.5 Results

2.5.1 Simulated K12 Unicycler assemblies

In order to investigate the coverage and read length requirements for hybrid genome assembly we simulated Nanopore and Illumina read sets for *E. coli* MG1655. We simulated an Illumina dataset containing four million 250 bp paired end reads and five Nanopore datasets each with 150x coverage and different read lengths (< 2.5 kbp, 2.5 kbp - 5 kbp, 5kbp - 10 kbp, > 10 kbp and no min or max length constraints). For each simulated Nanopore dataset, reads were randomly subsampled to each of the 6 coverage levels being investigated (50x, 20x, 15x, 10x, 8x, 5x). This subsampling process was repeated to create replicate datasets for each read length/coverage combination. For each dataset a Unicycler hybrid assembly was produced from the simulated Nanopore and Illumina reads.

We compared the contiguity and total assembly length to the known MG1655 reference genome to assess assembly completeness. We found that across all coverage levels contiguity was greatest for assemblies generated with ONT reads greater than 10 kbp . In datasets where reads were greater than 10kbp or ultra long (no read length limit) we produced single contig assemblies with a maximum of 10x coverage in five out of six attempts. In contrast, assemblies produced with reads less than 5kbp were highly fragmented regardless of coverage levels (**Figure 2.1 A and B**).

We next compared the total assembly length of each assembly to the reference genome. The MG1655 reference genome has a total length of 4.64Mbp (“Escherichia Coli Str. K-12 Substr. MG1655, Complete Genome - Nucleotide - NCBI” n.d.). Assemblies produced from reads less than 5kbp were always shorter than the reference genome. In contrast, assemblies produced with reads greater than 10 kbp were the same length as the reference regardless of coverage levels. We found that assemblies produced with reads less than 10 kbp increased in total length as coverage increased. This suggests that the reduced assembly length is due to misassembly, and the collapse of repeat regions (Tørresen et al.

2019; Treangen and Salzberg 2011; Phillippy, Schatz, and Pop 2008). Reads shorter than 10 kbp were unable to span the entire repeat region and could not provide structural information about the repeat length. Therefore, we propose that the longest repeat region in the *E. coli* MG1655 genome is between 5 and 10 kbp long. Taken together, these results suggest that to achieve complete assemblies of our natural isolates of *E. coli* ONT datasets should contain at least 10x coverage of reads greater than 10kbp.

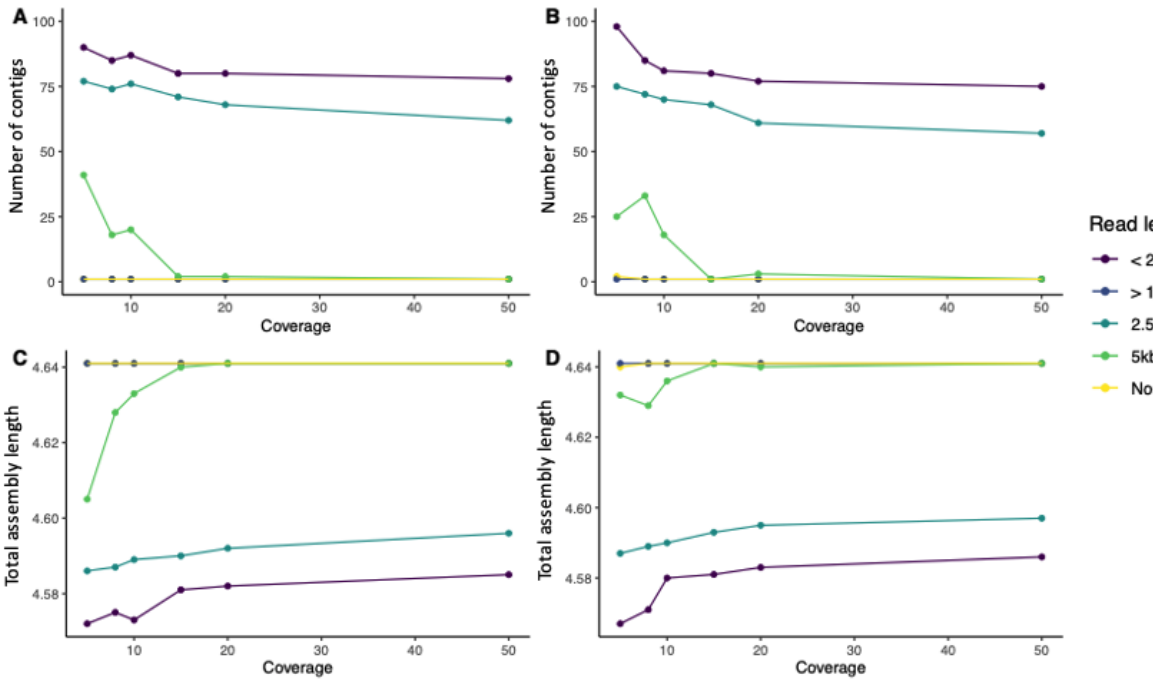


Figure 2.1: Assembly contiguity is influenced by read length. Unicycler hybrid assemblies were produced with simulated Nanopore and Illumina datasets at a range of read length and coverage limits. Generally, increased coverage leads to increased contiguity, but read length has a greater influence on contiguity. A significant decrease in total contigs is observed when read length increases beyond 5 kb. Furthermore, when reads are in excess of 10kb, single contig assemblies were generated with as low as 5x coverage. Each plot represents a replicated set of simulated reads and assemblies.

2.5.2 Optimising DNA extraction methods for high molecular weight DNA

ONT read length is directly related to the length of the DNA molecules prepared in the sequencing library. Therefore, DNA extraction methods which produce sufficient

concentrations of high molecular weight DNA are required. We tested two DNA extraction protocols, the Promega Wizard Genomic DNA extraction kit and a Phenol based DNA extraction protocol (Quick 2018). DNA from four strains was extracted according to each protocol and the resulting eight samples were barcoded, pooled and sequenced on the same flowcell. We used mean length, N10, N25, and the percent of reads >50kbp as metrics to evaluate the ability of each DNA extraction method to produce long read lengths (**Figure 2.2)**. Across the four Phenol extracted DNA datasets the average mean length was 4454bp, while DNA extracted with the Promega Wizard kit produced an average mean read length of 3984bp. Phenol extracted DNA also produced higher N10 and N25 values for each dataset (**Figure 2.2B and 2.2C)**. N10 and N25 values represent the read length which when all reads longer than this value are summed, will equal 10% or 25% respectively of the total number of bases. A larger N10 or N25 value represents a dataset with a greater number of long reads. Finally we quantified the percent of reads in each dataset greater than 50kbp as a direct measure of the proportion of long reads in each dataset (**Figure 2.2B**). Across each of our metrics DNA extracted using the Phenol extraction method produced longer read lengths than DNA extracted using the Promega wizard extraction kit.

As shown in Figure 2.1, longer read lengths are associated with more contiguous assemblies. Here we showed that DNA extracted using the Phenol method was repeatedly able to produce longer sequencing reads than the Promega wizard kit.

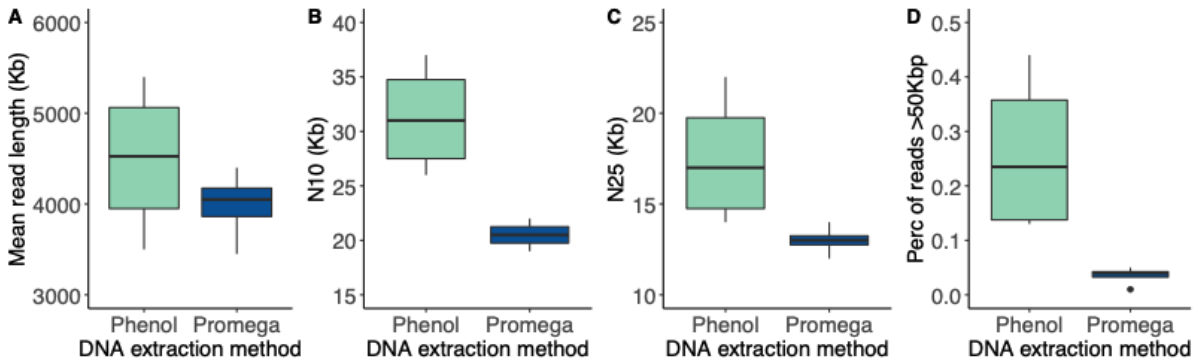


Figure 2.2: Phenol-chloroform DNA extractions produced longer ONT sequencing reads than Promega Wizard kit extracted DNA. DNA from four natural isolates of *E. coli* was extracted using both a Phenol based extraction method, and the Promega Wizard extraction kit.

The eight DNA samples were barcoded and sequenced on the same flowcell. The sequencing reads for each method were assessed to determine which DNA extraction method produces higher molecular weight DNA. For each metric, mean read length **(A)**, N10 **(B)**, N25 **(C)**, and percent of reads >50kbp **(D)** phenol-chloroform extracted DNA outperformed Promega Wizard kit extracted DNA.

2.5.3 Optimising Nanopore Rapid Barcoding kit for high molecular weight DNA

To maximise the number of samples sequenced per flowcell we used the rapid barcoding kit to multiplex up to 12 samples per flowcell. When doing this, it is highly recommended to perform an AMPure bead clean up on the pooled samples (Oxford Nanopore Technologies, n.d.). This removes short fragments from the samples and concentrates the sample into 10ul for loading on the flowcell. We found that our high molecular weight DNA bound to the beads as expected, but that the manufacturer's protocol was not sufficient to release the DNA into the elution buffer. We observed the DNA bound beads formed a viscous gel like structure. We hypothesised that shorter DNA fragments would not bind the beads in this way and would release from the beads easier. Gel electrophoresis showed that the majority of our extracted DNA was approximately 20kb which was in excess of our requirements. To test our hypothesis we attempted to shear the DNA into smaller fragments by passing the DNA repeatedly through a 23 gauge needle (Paul Coupland, n.d.) and a p10 pipette tip.

We established four shearing schemes which subjected each DNA sample to between 10 and 40 passes through the syringe and needle or 40 passes through the pipette tip. Consistent with the results described in the shearing protocol, overall DNA length was reduced from passing the DNA sample through the needle (**Figure 2.3A**). We did not observe a reduction in DNA fragment length from passing DNA through the pipette tip, and did not observe a significant difference in DNA fragment size after passing the sample through the needle more than 10 times.

We next tested if we were able to recover more sheared DNA than native DNA following an AMPure bead cleanup. We tested two DNA extractions from different natural isolates, and sheared the DNA by passing it through the syringe and needle 40 times. We then performed an AMPure bead clean up on the native and sheared DNA for each isolate. We collected both the wash step supernatant and final elute and ran all samples on a 70% agarose gel for 90 minutes at 100 volts to determine DNA recovery (**Figure 2.3B**). Although needle shearing appeared to decrease DNA fragment size, an increase in the total amount of DNA recovered following AMPure bead cleanup was not observed. As no DNA was detected in the wash supernatant we suggest that the DNA remained bound to the AMPure beads.

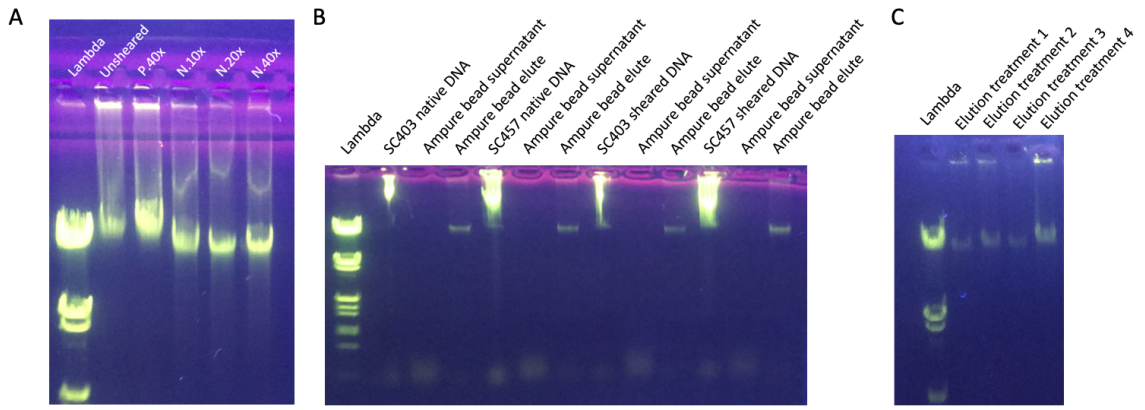


Figure 2.3 Recovery of DNA following AMPure bead cleanup was improved by increasing elution time and temperature. (A) DNA was mechanically fragmented by passing through a p10 pipette tip 40 times (P.40x) or passing through a 40 gauge needle 10, 20 or 40 times (N.10x, N.20x, N.40x). Pipette shearing did not alter DNA fragments, while needle shearing resulted in a small reduction in fragment size. **(B)** Needle sheared and unsheared DNA for two natural isolates were bound to AMPure beads and recovered. Input DNA, AMPure bead washstep supernatant and final elute was run on a 1% gel to visualise DNA recovery. Low amounts of DNA were recovered from both sheared and unsheared DNA, and no HMW DNA was observed in the supernatant, suggesting the DNA remained bound to the AMPure beads. **(C)** Increases to elution time and temperature were used to increase DNA recovery from AMPure beads. The standard protocol of 2 mins elution at rooms temperature (elution treatment 1) was compared with 10 mins elution at room temperature (elution treatment 2), 10 mins elution at 37ºC (elution treatment 3) and 10 mins elution at 50ºC (elution treatment 4). 10 mins elution at 50ºC resulted in the greatest DNA recovery.

We next investigated the effect of elution temperature and time on the DNA recovery rate. The AMPure bead protocol calls for elution at room temperature for 2 minutes (Elution treatment 1). We trialled an increased elution time of 10 mins at room temperature (Elution treatment 2), as well as performing the elution for 10 minutes at 37ºC and 50ºC with pre pre-warmed elution buffer (Elution treatments 3 and 4, respectively). These times and temperatures were chosen as they had been successfully used by other groups for bead elution (Schalamun and Schwessinger 2017). Gel electrophoresis visualisation of the DNA recovered from each set of trial conditions showed that pre-warming elution buffer to 50ºC combined with a 10 min incubation at 50ºC resulted in the greatest DNA recovery **(Figure 2.3C)**.

2.5.4 Nanopore sequencing

We used the protocol modification described above to sequence high molecular weight DNA for 47 natural isolates using MinION R9.4 flowcells and the rapid barcoding library prep kit.

For each isolate except one, we obtained at least 250 Mbp of sequence data (**Table 2.1**), with a median of 1002 Mbp per strain (interquartile range (IQR): 670 Mbp - 1296 Mbp). For all strains with greater than 500 Mbp, we used Filtlong to subsample each read set down to 500 Mbp prioritising read quality over read length. The filtered read sets had a median read N50 of 17.4 kbp (IQR: 13.4 kbp - 20.5 kbp).

2.5.5 Genome assemblies

For each of the 47 natural isolates, we produced a complete genome assembly using Flye v2.4.2 (Kolmogorov et al. 2019), followed by four rounds of long-read polishing with Pilon v1.23 (Walker et al. 2014) and two rounds of short-read polishing with Racon v1.3.2 (Vaser et al. 2017). The genomes ranged in length from 4.4 Mbp to 5.2 Mbp. We assume that any non-chromosomal contigs are plasmids and therefore 17 of our isolates contain no plasmids (a single chromosomal contig), 21 isolates contain a single plasmid and ten isolates contain multiple plasmids (**Table 2.1**). We used Socru, which reports the arrangement of rRNA operons within an assembly, to investigate the structural accuracy of these genomes. In all cases, the rRNA operons were found in standard or known orientations, supporting the structural accuracy of these genomes.

Table 2.1. Genome statistics for all 47 assemblies. Strain IDs are taken from (Ishii et al. 2006). Strains are sorted by chromosome length. Ph:CHCl3 indicate phenol:chloroform extraction was used; Promega indicates that the Promega Wizard DNA extraction kit was used. Filtered read N50 indicates the N50 after Filtlong (https://github.com/rrwick/Filtlong) was used to retain only 500 Mbp from each strain. Chrom. length indicates the length of the longest contig, assumed to be the chromosome. Chrom. circular indicates whether the chromosome is a single circular contig. Genome length indicates the sum of the length of all contigs. rRNA orientation indicates the orientation of the seven ribosomal operons in *E. coli*, as assessed by socru (Page, Ainsworth, and Langridge 2020).

Strain ID	Number of Illumina Reads	ONT extraction method	ONT Total bp	Number of ONT reads	ONT Read N50	ONT filtered read N50	Chrom. length	Chrom. Circular	Genome length	Total contigs	rRNA orientation
SC468	1,288,488	Ph:CHCl3	1,402,515,335	34,337	7,164	14,815	4,426,017	Y	4,426,017	1	standard
SC457	603,943	Ph:CHCl3	465,520,117	51,693	9,209	10,280	4,555,909	Y	4,555,909	1	standard
SC455	819314	Ph:CHCl3	1,033,029,654	30,634	10,560	17,536	4,655,420	Y	4,655,420	1	standard
SC434	618526	Ph:CHCl3	849,800,865	33,880	10,113	16,056	4,658,197	Y	4,658,197	1	standard
SC477	1045031	Ph:CHCl3	843,013,946	44,613	7,856	12,549	4,658,510	Y	4,658,510	1	standard

SC316	3,409,700	Promega	43,287,859	39,445	11,807	13,365	4,663,327	Y	4,681,204	3	standard
SC467	922,562	Ph:CHCl3	1,003,216,478	38,336	8,521	13,963	4,715,938	Y	4,715,938	1	standard
SC423	849,233	Ph:CHCl3	274,260,727	18,522	16,188	18,066	4,716,885	Y	4,716,885	1	standard
SC465	1,577,380	Promega and Ph:CHCl3	1,386,503,691	35,047	9,489	15,195	4,722,586	Y	4,785,019	2	standard
SC452	417,390	Ph:CHCl3	2,207,184,721	16,959	11,909	30,889	4,723,951	Y	4,755,166	2	standard
SC431	722,122	Ph:CHCl3	1,774,332,186	25,144	3,326	21,138	4,727,732	Y	4,866,254	2	standard
SC475	413,920	Ph:CHCl3	907,422,842	36,085	8,263	14,996	4,729,401	Y	4,729,401	1	standard
SC492	544,889	Ph:CHCl3	718,894,150	30,095	13,767	18,501	4,736,913	Y	4,736,913	1	standard
SC480	614,748	Ph:CHCl3	1,009,767,146	25,686	12,892	20,491	4,741,504	Y	4,741,504	1	standard
SC476	661,043	Promega and Ph:CHCl3	1,467,536,905	33,435	6,509	15,758	4,747,946	Y	4,747,946	1	standard
SC479	483,518	Ph:CHCl3	468,589,667	41,638	11,800	13,367	4,762,128	Y	4,913,012	3	standard
SC392	870,803	Promega	366,289,283	44,990	8,646	9,714	4,770,015	Y	4,783,281	2	standard
SC312	3,582,617	Promega	678,059,117	32,708	13,222	17,765	4,775,485	Y	4,878,778	3	standard
SC386	1,917,879	Promega	989,915,389	4,259	7,170	12,507	4,778,381	Y	5,088,094	5	standard
SC456	350,577	Ph:CHCl3	120,521,609	14,793	8,716	9,946	4,790,285	Y	4,885,342	2	standard
SC487	1,289,213	Ph:CHCl3	651,178,805	34,634	13,099	17,546	4,794,586	Y	4,794,586	1	standard
SC433	566,913	Ph:CHCl3	660,660,048	25,781	17,312	23,362	4,797,429	Y	4,961,214	2	standard
SC429	1,110,243	Promega and Ph:CHCl3	1,081,882,331	28,668	9,271	18,780	4,797,468	Y	4,961,244	2	standard
SC430	2,171,974	Promega and Ph:CHCl3	1,749,627,499	27,343	7,478	19,159	4,797,499	Y	4,961,283	2	standard
SC397	1,485,195	Promega	1,001,724,414	28,040	13,643	19,677	4,858,696	N	5,067,247	3	standard
SC411	565,996	Ph:CHCl3	1,360,250,706	25,134	11,224	20,582	4,859,344	Y	5,068,109	3	standard
SC419	793,357	Ph:CHCl3	1,421,369,578	35,942	6,732	14,684	4,859,796	Y	4,916,116	2	standard
SC364			1,077,384,666		13,126	20,928	4,860,085	Y	5,063,812	2	standard

	9,348,468	Promega		26,426							
SC453	522,987	Promega and Ph:CHCl3	1,231,172,695	50,059	7,010	11072	4863138	N	5308239	4	standard
SC307	2,670,106	Promega	1,216,450,680	26,661	10,595	19865	4892106	Y	5221106	5	standard
SC400	9,809,258	Promega	1,138,079,055	21,521	16,543	25690	4924724	Y	5065688	2	standard
SC489	779,700	Ph:CHCl3	530,260,839	73,424	5,806	7779	4929025	N	5008168	4	standard
SC469	399,522	Ph:CHCl3	723,447,743	40,512	9,772	13766	4940057	Y	5129818	3	standard
SC402	3,305,579	Promega	1,103,528,651	25,927	14,104	20875	4944324	Y	5085287	2	standard
SC406	700,028	Promega and Ph:CHCl3	2,471,779,990	25,820	8,152	19613	4958102	Y	4958102	1	standard
SC454	1,653,791	Ph:CHCl3	1,010953,794	43,107	7,741	13039	4982834	Y	4988386	2	alternative
SC441	738,159	Ph:CHCl3	988,454,259	46,048	6,225	11435	4986040	Y	5022479	2	alternative
SC446	1,037,860	Ph:CHCl3	451,143,119	35,499	13,071	14580	4986746	Y	4997068	2	alternative
SC445	490,093	Ph:CHCl3	260,701,609	14,322	19,183	21437	4987469	Y	5033261	2	alternative
SC443	841,953	Ph:CHCl3	1,027,640,114	31,687	9,897	17265	4999711	N	5021047	2	alternative
SC410	436,740	Ph:CHCl3	981,338,093	25,375	10,873	22404	5001654	Y	5008324	2	standard
SC422	514,475	Ph:CHCl3	611,712,027	98,763	4,900	5623	5003951	Y	5003951	1	standard
SC464	874,611	Promega and Ph:CHCl3	3,086,301,299	18,303	8,993	27897	5023622	Y	5137983	2	standard
SC407	1,260,500	Promega and Ph:CHCl3	2,703,637,269	24,419	7,660	20853	5088866	Y	5088866	1	standard
SC403	1,257,142	Promega	765,980,798	33,276	12,830	17205	5089116	Y	5145436	2	standard
SC368	4,766,180	Promega	910,447,798	48,174	6,454	11026	5101998	Y	5101998	1	standard
SC418	665,222	Promega and Ph:CHCl3	2,677,454,945	24,282	7,752	21140	5222289	Y	5222289	1	standard

2.6 Discussion

Here we have investigated the long read sequencing data required to produce a complete hybrid genome assembly. We use simulated Nanopore and Illumina read sets for *E. coli* strain MG1655 as a proxy for our *E. coli* isolates and show that a single contig assembly is able to be produced when datasets contain at least 10x coverage of reads greater than 10kbp. We propose that this exceeds the longest repeat region in the *E. coli* MG1655 genome (Koren and Phillippy 2015). We then investigated DNA extraction methods to generate high molecular weight DNA for each of our natural isolates. We extract and sequence DNA from four natural isolates using a phenol chloroform method (Quick 2018) and the Promega Wizard DNA extraction kit. We show that across four metrics of read length, longer sequencing reads are produced for DNA extracted using the phenol chloroform method. Despite this, phenol chloroform DNA extractions require more hands on bench time, use more starting material and require specialist chemicals and appropriate lab and safety equipment. Therefore, due to these limitations, some samples were extracted for sequencing using the Promega Wizard extraction kit.

We then optimised sequencing library prep protocols to accommodate the high molecular weight DNA we extracted. During library prep, DNA bound to AMPure beads was not able to be recovered resulting in loss of sequencing libraries. Our initial hypothesis was that reducing DNA length would increase DNA recovery, however this was not shown. DNA fragmentation is commonly used in the production of Illumina sequencing libraries and many methods are available which can be categorised into three broad groups: mechanical, enzymatic and chemical (Y. Yang and Hang 2013; Poptsova et al. 2014; Knierim et al. 2011). We considered two factors in choosing a shearing method, target fragment size and downstream applications of this sequencing data. Many mechanical shearing approaches are designed to fragment DNA smaller than 5kbp for the preparation of Illumina sequencing libraries (Caruccio 2011; Ambardar et al. 2016; Knierim et al. 2011). In contrast we aimed to slightly reduce fragment lengths while keeping fragments longer than 10 kbp. Secondly, downstream analysis of this sequence data would include the detection of methylated bases. Enzymatic or chemical fragmentation methods may disrupt the methylation patterns established on each fragment, complicating future analysis.

To avoid excessive fragmentation and disruption of possible methylation patterns, we chose to use a needle shearing approach to reduce DNA length(Paul Coupland, n.d.). Although we achieved some reduction in DNA fragment size, this was not sufficient to positively impact DNA recovery. Other methods which reduced DNA fragment size further may have been successful in increasing DNA recovery from AMPure beads. We chose not to investigate this

further as increased fragmentation would likely result in shorter sequencing reads which is shown above to have a negative effect on assembly completion.

In order to increase DNA recovery following AMPure bead cleanup we instead increased both the time and temperature of the elution step from two minutes at room temperature, to ten minutes at 50ºC in prewarmed elution buffer. These changes to the AMPure bead elution resulted in the recovery of sufficient DNA to successfully sequence 47 natural isolates of *E. coli.* We produced sequencing read sets with a median of 1002Mbp (approx 200x coverage) of sequence data per strain, which enabled us to filter each read set to retain only the longest, highest quality reads. Following filtering, our read sets had a median read N50 of 17.4 kbp, significantly longer than the target of 10x coverage of reads exceeding 10kbp. For each of the 47 natural isolates, we produced a complete hybrid genome assembly using the long read assembler Flye and short and long read polishing with Racon and Pilon.

Overall the work presented in this chapter establishes sequencing data requirements for producing a single contig hybrid genome assembly for natural isolates of *E. coli*. We establish DNA extraction and library preparation protocols to routinely produce sequencing data which meets these requirements. We then use this data to produce a hybrid assembly for each natural isolate.

Until the recent development of long read sequencing technologies, genome assemblies have been built from Illumina data. Due to the relatively short read lengths, these assemblies are frequently highly fragmented (R. R. Wick et al. 2017a). This has meant that assembly metrics have focused on structural metrics such as contig N50 and overall contig number. Here we use contiguity to assess assembly completeness, additionally we use Socru to assess the internal structural accuracy of each assembly through the arrangement of rRNA operons. We acknowledge that while these metrics contribute to the completeness of a genome assembly, they do not represent all of the factors which constitute a complete, accurate assembly. Contiguity and rRNA operon arrangement are influenced by the large-scale structural accuracy of the assembly and do not report on sequence level accuracy.

Long read sequence data is known to have higher error rates compared to Illumina sequence data. Therefore, although the addition of long reads in a hybrid assembly approach can increase contiguity, per base sequence level accuracy may be affected. This raises the question of how to accurately assess genome completeness on both a genome wide structural level and a per base pair sequence level particularly for de novo assemblies.

Chapter 3

Do You Want to Build a Genome? Benchmarking Hybrid Bacterial Genome Assembly Methods

Georgia Breckell, Olin K. Silander

Article submitted for peer-review

(Microbial Genomics)

Author contributions:

Georgia Breckell: Conceptualization (equal); Methodology (lead); Investigation (lead); Visualization (lead); Writing-original draft (lead); Writing-review & editing (equal).
Olin K. Silander: Conceptualization (equal); Methodology (supporting); Funding acquisition(lead); Resources (lead); Supervision (lead); Writing-original draft (supporting); Writing-review & editing (equal).

3.1 Preface

The following chapter has been published as a preprint on bioRxiv (DOI:https://doi.org/10.1101/2021.11.07.467652) and was submitted for peer review to the journal Microbial Genomics.

Prior to the development of long read sequencing approaches, assembly accuracy was routinely reported on the basis of contiguity, and fragment length with metrics such as contig N50 and total number of contigs. Furthermore, a range of new assembly methods rapidly became available which leverage long sequencing reads to routinely produce single contig, circularised assemblies. These advancements significantly reduce the value of classical assembly measures. Accounting for the known lower accuracy of long read data we established a set of five metrics to evaluate the structural and sequence accuracy of genome assemblies and used these to characterise differences in assembly accuracy from five hybrid Nanopore and Illumina assembly approaches. We use 49 Natural isolates of *Escherichia coli* which provides us the depth required to observe infrequently occurring incidents of strain specific limitations to the production of a complete assembly.

Olin and I conceived this project and designed the experiments. I carried out all the sequencing and data curation for this project and I performed most of the bioinformatics analysis presented in this chapter. Olin performed the initial testing of our assemblies against K12, however I performed future iterations of this analysis presented in Table 1. Olin produced figure 4.D, and performed the analysis and produced Figure 5, Figure S2, S7 and S8. This manuscript was written as a collaboration with Olin.

The formatting of this paper was modified slightly from the manuscript available at bioRxiv to ensure consistency throughout this thesis.

3.2 Abstract

Long read sequencing technologies now allow routine highly contiguous assembly of bacterial genomes. However, because of the lower accuracy of some long read data, it is often combined with short read data (e.g. Illumina), to improve assembly quality. There are a number of methods available for producing such hybrid assemblies. Here we use Illumina and Oxford Nanopore (ONT) data from 49 natural isolates of *Escherichia coli* to characterise differences in assembly accuracy for five assembly methods (Canu, Unicycler, Raven, Flye, and Redbean). We evaluate assembly accuracy using five metrics designed to measure structural accuracy and sequence accuracy (indel and substitution frequency). We assess structural accuracy by quantifying (1) the contiguity of chromosomes and plasmids; (2) the fraction of concordantly mapped Illumina reads which had been withheld from the initial assembly process; and (3) whether rRNA operons are correctly oriented. We assess indel and substitution frequency by quantifying (1) the fraction of open reading frames that appear truncated and (2) the number of variants that are called using Illumina reads only. Applying these assembly metrics to a large number of *E. coli* strains, we find that different assembly methods offer different advantages. In particular, we find that Unicycler assemblies have the highest sequence accuracy in non-repetitive regions, while Flye and Raven tend to be the most structurally accurate. In addition, we find that there are unidentified strain-specific characteristics that affect ONT consensus accuracy, despite individual reads having similar levels of accuracy. The differences in consensus accuracy of the ONT reads can preclude accurate assembly regardless of assembly method. These results provide quantitative insight into the best approaches for hybrid assembly of bacterial genomes and the expected levels of structural and sequence accuracy. They also show that there are intrinsic idiosyncratic strain-level differences that inhibit accurate long read bacterial genome assembly. However, we also show it is possible to diagnose problematic assemblies, even in the absence of ground truth, by comparing long-read first and short-read first assemblies.

3.3 Introduction

Bacterial genomes are extremely dynamic, undergoing frequent loss and gain of both chromosomal and extrachromosomal elements including insertion sequences, rep sequences, ICE elements, and plasmids (Touchon et al. 2009; Lee et al. 2016; Baltrus 2013; Horesh et al. 2021). Quantifying these rapid genome dynamics is critical for understanding population genetic processes, such as the generation of genetic variation and the action of selection. However, it is difficult to quantify such dynamics without accurate and complete (fully contiguous) genome assemblies. Unfortunately, the most dynamic genome elements

tend to be both repetitive and plentiful in bacterial chromosomes and plasmids, and thus the most likely to lead to genome mis-assemblies.

Until recently, the most common method for assembling bacterial genomes was using short reads only (e.g. Illumina). However, it is usually not possible to achieve completely contiguous genome assemblies using short reads alone. The development of long read technologies (PacBio and Oxford Nanopore) now allows routine fully contiguous assembly of bacterial genomes (Koren and Phillippy 2015; R. R. Wick et al. 2017b). One shortcoming of long read technologies is that they can have low accuracy (although this is rapidly changing). Because of this, one of the most common means of building high-accuracy reference-level genomes is to use hybrid assembly methods, which combine the accuracy and economy of short read technology (e.g. Illumina) with long reads from the economical and accessible Oxford Nanopore (ONT) platform.

Considerable work has been done to address the problem of bacterial genome assembly, using either long-read only or hybrid approaches. Wick and Holt have performed extensive benchmarking of long-read assemblers using both simulated and real ONT read datasets (R. R. Wick and Holt 2019), finding that the long read assemblers Flye and Raven perform well across a number of datasets. However, the number of different genomes used in that study (six) was not extensive, and thus may not tell the full story of the strengths and weaknesses of specific assemblers. De Maio et al. (De Maio et al. 2019) evaluated assemblies for 20 Enterobacteriaceae genomes, including four *Escherichia coli* (*E. coli*) using hybrid data, either Illumina and PacBio, or Illumina and ONT. However, the only long-read assembler assessed was Flye.

Here we use ONT and Illumina data from 49 phylogenetically diverse E. coli strains, with genome sizes ranging from 4.5 Mbp to 5.4 Mbp, to assess differences in hybrid assembly accuracy for five assembly methods (with polishing when necessary): Unicycler (R. R. Wick et al. 2017a), Raven (Vaser and Šikić 2021), Redbean (Ruan and Li 2019), Canu (Koren et al. 2017), and Flye (Kolmogorov et al. 2019). Unicycler is unique in this group in that it first constructs a short read assembly graph, and then resolves ambiguities in the graph using long reads; the latter four are among the most accurate long read assemblers (R. R. Wick and Holt 2019).

To quantify assembly accuracy in the absence of known “ground truth” assemblies, we use five metrics designed to measure both structural accuracy and sequence accuracy. We assess structural accuracy by quantifying (1) the contiguity of chromosomes and plasmids; (2) the fraction of concordantly mapped Illumina reads that have been withheld during assembly; and (3) whether all rRNA operons are correctly oriented. We assess indel and

substitution frequency by quantifying (1) the fraction of open reading frames that appear truncated and (2) the number of variants that are called using Illumina reads only.
Our results highlight the fact that some assemblers perform predictably better than others. In addition we find that significant differences in assembly accuracy arise because of unidentified strain-specific characteristics that affect long read consensus accuracy. This suggests that currently, there are fundamental limitations to assembly accuracy for bacterial genomes when relying on ONT and Illumina data. At the same time, they emphasise the utility of consensus assembly approaches such as those implemented by Tricycler (R. R. Wick et al. 2021a).

3.4 Materials and Methods

3.4.1 DNA extraction

We grew single cell colonies from each of the 49 natural isolates overnight in 3 mL of liquid LB media at 37°C . We extracted DNA using either the Promega Wizard DNA extraction kit (following the gram negative bacterial extraction protocol), or phenol chloroform (Quick 2018). We measured DNA size distributions using gel electrophoresis in a 1% agarose gel, and both a Qubit fluorometer and Nanodrop readings to measure the size and quality of each DNA prep.

3.4.2 Library preparation and DNA sequencing

All Illumina Genome sequencing was performed by MicrobesNG (http://www.microbesng.uk) which is supported by the BBSRC (grant number BB/L024209/1).
We prepared ONT sequencing libraries using 400ng of DNA with the SQK-RBK004 or SQK-LSK109 kits according to the manufacturer's protocol with the following modifications: samples were eluted off Agencourt AMPure XP beads into TE buffer pre-warmed to 50°C; we performed the elution at 50°C; and increased the incubation time in the elution buffer to 10 minutes.

ONT sequencing was carried out on a MinION Mk1b device using R9.4.1 flowcells. 17 flowcells were used in total, with between four and 12 strains run per flowcell (**Table S.3.2**). Several strains were sequenced multiple times to ensure that at least 500 total Mbp (approximately 100X coverage) with a read N50 greater than 5 kb was generated for each strain. For nine strains we achieved complete assemblies with lower coverage (between 24x and 94x), and did not generate additional sequence data for those strains. We defined a complete assembly as having no unidentified contigs, the correct rRNA operon arrangement, greater than 98% illumina read concordancy and fewer than 7% short ORFs.

For each MinION run, we demultiplexed the reads using Deepbinner (Wick 2021a) and basecalled with Guppy v4.2.2.

3.4.3 Genome assembly and polishing

We filtered all Oxford Nanopore reads to retain approximately 500 Mbp using Filtlong (R. Wick n.d.), specifying the option to use Illumina reads to determine which ONT reads are of high or low quality. From all Illumina read sets, 5,000 pairs of Illumina reads were removed and withheld from assembly (see below).
For all five assemblers (Unicycler, Canu, Flye, Raven, and Redbean) we used default settings. We polished each long read-first assembly with both Illumina and ONT data, using four rounds of Racon with ONT reads, followed by Pilon and Racon with Illumina reads. All basecalling, filtering, assembly and polishing pipelines are available on github (https://github.com/GeorgiaBreckell/assembly_pipeline).
For all long read assemblers we also tested the utility of an intermediate polishing step with medaka v1.2.2 (“Medaka — Medaka Documentation” n.d.). However, we obtained inconsistent results: some assemblies decreased in quality while others changed little. Due to these inconsistencies, we do not report any results with medaka-polishing steps here.

3.4.4 Multiple alignment and phylogenetic reconstruction

The alignment was created using REALPHY (Bertels et al. 2014) with default parameters and specifying the *E. coli* strains MG1655, REL606, W, SE11, O157:H7, IAI39, and CFT07 as the references. *E. fergusonii* was also used as a reference to allow rooting. RAxML (Stamatakis 2014) was used for phylogenetic reconstruction.

3.4.5 Quality assessment

All commands, and software versions used for quality assessment are described in the assembly pipeline available on the github repository linked above. A brief explanation of the tools and their usage is given below.

3.4.5.1 Genome fragmentation

A complete E. coli genome should contain a single circular contig representing the chromosome along with a variable number of circularised contigs representing plasmids. We determined fragmentation for each assembly by testing (1) the number of contigs present in each assembly; and (2) whether any additional contigs could be identified as plasmids.

3.4.5.2 Plasmid assignment

We used mlplasmids (Arredondo-Alonso et al. 2018) and Plasmidfinder (Carattoli et al. 2014) to assess whether contigs were of plasmid or chromosomal origin. mlplasmids uses a support vector machine learning approach based on the frequency of 5-mers, with training on taxa-specific chromosomes and plasmids. The result is a posterior probability of a contig belonging to a plasmid or chromosome class. We assigned contigs as plasmids if this posterior probability was greater than 0.5. In contrast, Plasmidfinder uses database matching for contig classification, and returns the identity of the matched plasmid.

3.4.5.3 Illumina mapping discordancy

We withheld 5,000 Illumina paired end reads from each assembly. We then aligned these to assembly using bwa mem (H. Li 2013). We used samtools (H. Li et al. 2009) to extract the fraction of concordantly mapping reads and from this determined the fraction of discordantly mapped reads.

3.4.5.4 rRNA operon orientation

We used Socru (Page, Ainsworth, and Langridge 2020) to determine whether the arrangement of inter-rRNA regions in each of our assemblies was as expected, or had been observed previously. Socru maps inter-rRNA regions against its *E. coli* model to determine their presence and arrangement in each assembly. We ran Socru with default parameters and the *E. coli* model of rRNA structure.

3.4.5.5 Truncated ORFs

To quantify the number of truncated ORFs we followed the approach outlined in ideel (M. Watson n.d.). Briefly, we annotated each assembly with Prodigal (Hyatt et al. 2010) to produce a set of predicted proteins. We aligned to the Uniprot protein database using Diamond (Buchfink, Reuter, and Drost 2021). We then compared the length of each protein and its top hit. We considered any ORF with a length less than 90% of the top hit as truncated.

3.4.5.6 Deletions and SNPs

To quantify the number of SNPs, and deletions in each assembly we used the Breseq pipeline with default settings (Deatherage and Barrick 2014). We ran the pipeline with our Illumina reads for each strain, with the assemblies produced by each assembler.

3.4.5.7 Pairwise genome alignment

We used dnadiff v1.3 (Kurtz et al. 2004) to align whole genomes to quantify differences in SNPs and indels between assemblies of the same strain, using default settings.

3.5 Results

3.5.1 Establishing the validity of proposed metrics

In order to test the utility of the proposed assembly metrics, we calibrated the metrics on a known “ground truth” *E. coli* K12 MG1655 reference genome. We first obtained ONT and Illumina data for our laboratory strain of MG1655. To ensure that our lab strain matched the sequence of the NCBI reference MG1655 genome (and could truly serve as ground truth), we first called variants against the NCBI reference sequence using Illumina reads alone and the Breseq pipeline, which identifies variants with strong statistical support (Deatherage and Barrick 2014). Breseq identified ten differences (including substitutions, indels, and structural changes) between our lab strain and the reference MG1655. Seven of the ten changes were shared with ATCC strain 700926 (**Table S.3.1**), which is provided by ATCC as having the MG1655 genotype, but is known to differ from the NCBI reference sequence (Freddolino, Amini, and Tavazoie 2012). This suggested that our lab strain is derived from ATCC 700926.

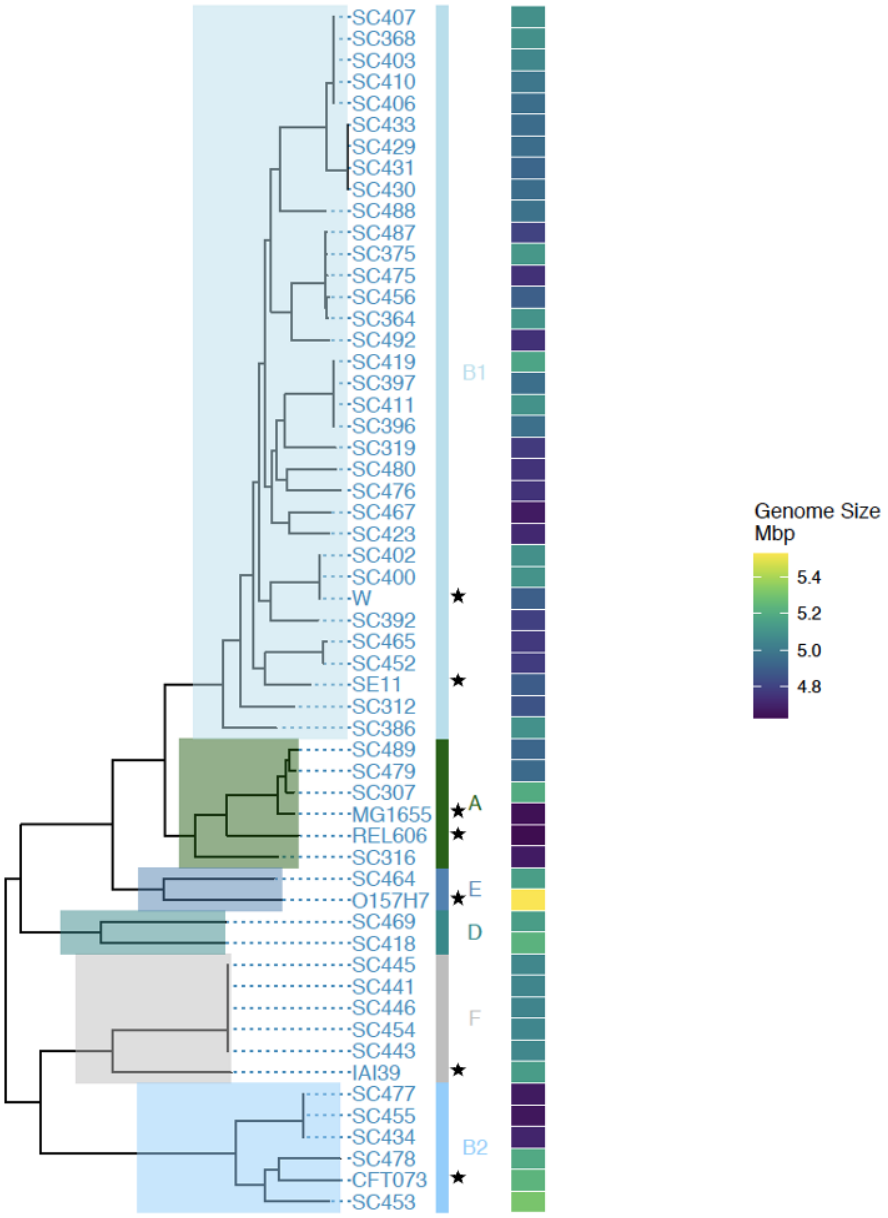


Figure 3.1. Phylogeny of the *E. coli* strains used to assess assembly accuracy. We used 49 strains to test assembly methods, with representatives from all the major clades of *E. coli* (A, B1, B2, D, E, F). For guidance, we have included seven well known *E. coli* strains in the phylogeny (MG1655, REL606, W, SE11, O157:H7, IAI39, and CFT073), indicated by stars. The 49 strains we use have a wide variety of genome sizes, ranging from 4.5 Mbp to 5.4 Mbp, indicated in the heatmap on the right side of the phylogeny (genome sizes are the means of all five assemblers and include all contigs).

We incorporated these ten changes into the reference genome sequence, and designated this as ground truth. We note that given the relatively conservative nature of Breseq, there

may be SNPs present in our lab strain that were not identified (false negatives), although it is unlikely that there are a large number of these.

We next used the MG1655 ONT and Illumina data to produce five hybrid assemblies. These assemblies were constructed with the same Illumina data used for the Breseq analysis above, as well as an additional 500 Mb of ONT data (approximately 100x coverage). We then compared each of these assemblies to our ground truth MG1655 sequence. For all assemblers, we found a single contig, all rRNA operons correctly oriented, and the vast majority of paired-end Illumina reads concordantly mapped (99.4% for all). Thus, using these metrics, we would infer that the correct genome structure had been obtained by all assemblers. However, direct comparison with the reference showed that the Canu assembly had duplicated sequences at both the start (34.3 kbp) and end (63.4 kbp) of its assembled contig, and the Redbean assembly had 3.7 kbp gap at the end of its contig, illustrating that these metrics do not detect all changes to genome structure, especially when present at the ends of contigs.

Table 3.1. Percentage of truncated ORFs and total SNVs in different MG1655 assemblies relative to ground truth.

Assembler	Total Annotated ORFs[1]	Total Truncated ORFs	% Truncated ORFs	Total Breseq SNPs and indels	Total dnadiff SNPs and indels
Reference	4318[2]	142	3.3%	0	0
Unicycler	4312	142	3.3%	4	71
Raven	4372	247	5.6%	115	129
Redbean	4377	255	5.8%	109	116
Canu	4479	267	6.0%	105	114
Flye	4383	268	6.1%	128	139

1 The number of annotated ORFs in each assembly is affected both by the increased (or decreased) numbers of base pairs in each assembly, as well as the number of indels, which will often create new shorter ORFs that may then not be annotated as ORFs at all (e.g. if they are not of sufficient length), or annotated as two or more ORFs.
2 This number of ORFs is not equal to the true number of ORFs in the reference as the genomes are computationally annotated using Prodigal (see **Methods**).

We also inferred SNP and indel frequency in the MG1655 assemblies. We first quantified the fraction of truncated ORFs, which should correlate with the number of indels present in the assembly (see **Methods**). We found between 142 (3.3%, Unicycler) and 268 truncated ORFs (6.1%, Flye) in each assembly (**Table 3.1**). A small number of truncated ORFs is expected, as pseudogenes are common in *E. coli*; performing the same analysis on the MG1655 reference identified 142 out of 4318 ORFs as truncated (3.3%). Finally, we used the Breseq pipeline to call variants in the assemblies using the Illumina reads only. We found between four (Unicycler) and 128 (Flye) SNPs and indels, with the vast majority of errors being deletions. These results suggested that the Unicycler assembly was more than an order of magnitude more accurate in terms of sequence accuracy than the other assemblies.

However, the Breseq pipeline conservatively calls SNPs and indels, so we also compared the assemblies to the original ground truth sequence using dnadiff (Kurtz et al. 2004), which performs a full genome alignment to determine sequence differences. This analysis revealed that all five assemblies contained more SNPs and indels than we discovered using Breseq. The most accurate assembly, Unicycler, contained 9 indels and 62 substitutions relative to the reference. Notably, 69 of the 71 differences were in rRNA operons, and are likely due to the repetitive nature of these regions (there are seven rRNA operons in *E. coli*), and difficulties in accurate polishing. However, this is a difficulty that may be resolved by more carefully considered polishing steps (R. R. Wick and Holt 2021).

Using dnadiff to compare the other assemblies to the ground truth reference, we found they contained between 114 (Canu) and 139 (Flye) SNPs and indels. These results highlight the fact that even for relatively small bacterial genomes, obtaining highly accurate and complete hybrid genome assemblies is not easily done. In addition, they show that one advantage of long read-first assemblers compared to short read-first assembly (i.e. Unicycler) is that there is a considerably lower rate of error within repetitive regions. This is starkly indicated by the fact that the Unicycler assembly had approximately 20-fold more errors called by dnadiff (which relies on whole genome alignment SNP and indel calling, and thus includes repetitive regions) compared to Breseq (which relies on short read variant calling and largely ignores repetitive regions). In contrast, the number of errors in the long read-first assemblies was similar for both dnadiff and Breseq (**Table 3.1**).

3.5.2 Genome sequencing and assembly of phylogenetically diverse *E. coli* strains

To more thoroughly investigate the accuracy and consistency of bacterial genome assembly methods, we produced assemblies for a set of 49 phylogenetically diverse *Escherichia coli*

strains of varying genome complexity and size (Ishii et al. 2006; Breckell and Silander 2020) (**Fig. 3.1**). We obtained at least 30-fold coverage of 250bp paired end Illumina data for 47 of the 49 strains, and more than 30-fold coverage of 100 bp paired end data for the remaining two. We also obtained a median of 770 Mbp of Oxford Nanopore (ONT) data for each strain (interquartile range (IQR): 545 Mbp - 1118 Mbp). The median average read length across all strains was 5.3 kbp (IQR: 3.9kbp - 6.2 kbp), and median quality scores for all reads across all strains was 13.3 (IQR: 12 - 14). We filtered the ONT reads using Filtlong to retain 500 Mbp (approximately 100-fold coverage), prioritising quality over length (**Methods**). The filtered data sets had a median average read length of 7.7 kbp (IQR: 6.3 kbp - 8.0 kbp; **Fig. 3.2A**) and a median quality score of 14.6 (IQR: 14 - 15; **Fig. 3.2B**). Surprisingly, we found that both read length and quality varied considerably between strains, despite the data being produced by identical genomic DNA prep methods, flow cell chemistry, and basecalling software.

We used this data to produce hybrid assemblies with the five assemblers: Canu, Flye, Raven, Redbean and Unicycler. As with the MG1655 assembly, we polished each long read-only assembly (Canu, Flye, Raven, and Redbean) using four rounds of Racon with ONT data, followed by Pilon and Racon with the Illumina reads (see **Methods**). It has been previously reported that Canu assemblies do not necessarily benefit from polishing with long read data (Goldstein et al. 2019). However, we polished all long read assemblies in the same way for consistency. As Unicycler contains a built-in polishing step, we did not polish these assemblies.

3.5.3 Structural accuracy and consistency

We first quantified the total number of contigs in each assembly. Across all assemblers and strains, a median of two contigs were produced (IQR 1-4). Raven consistently produced the fewest contigs, followed closely by Flye. Unicycler, Canu, and Redbean assemblies were the most fragmented (**Fig. 3.3A** and **Fig. S.3.1**).

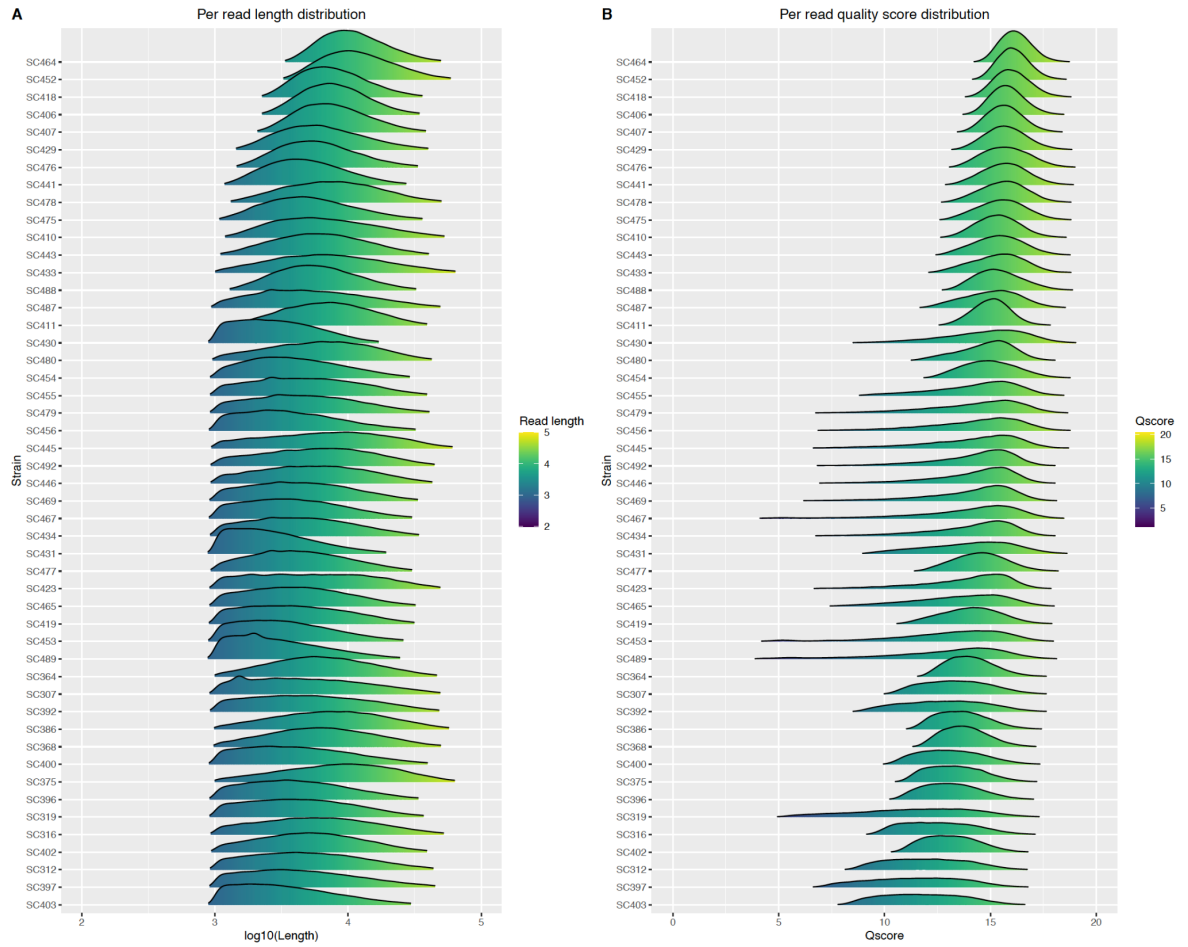


Figure 3.2. ONT read length and read quality score distributions for all filtered ONT datasets, ordered by percentage of reads with a quality score greater than 15. ONT reads were filtered prior to assembly to retain approximately 500 Mbp using filtlong, prioritising read quality over length. **A. Filtered read length distributions for each strain. B. Filtered per read average quality scores for each strain**.

These results suggested that Raven and Flye were optimal for maximising assembly contiguity. However, it is also possible that the additional contigs in the more fragmented assemblies were true (extrachromosomal) plasmid contigs, which for Raven and Flye had been incorrectly incorporated into the chromosome. We thus classified each contig as plasmid or chromosomal using mlplasmids (Arredondo-Alonso et al. 2018) and PlasmidFinder (Carattoli et al. 2014). These methods can identify small contigs as plasmids; however, they cannot identify regions of chromosomal contigs as plasmid. Both classification methods identified many of the small contigs in the Canu, Unicycler, and Redbean

assemblies as chromosomal. In contrast, almost all small contigs in Flye and Raven were identified as plasmids **(Fig. 3.3B)**. This suggests that the additional contigs are in fact unincorporated chromosomal fragments.

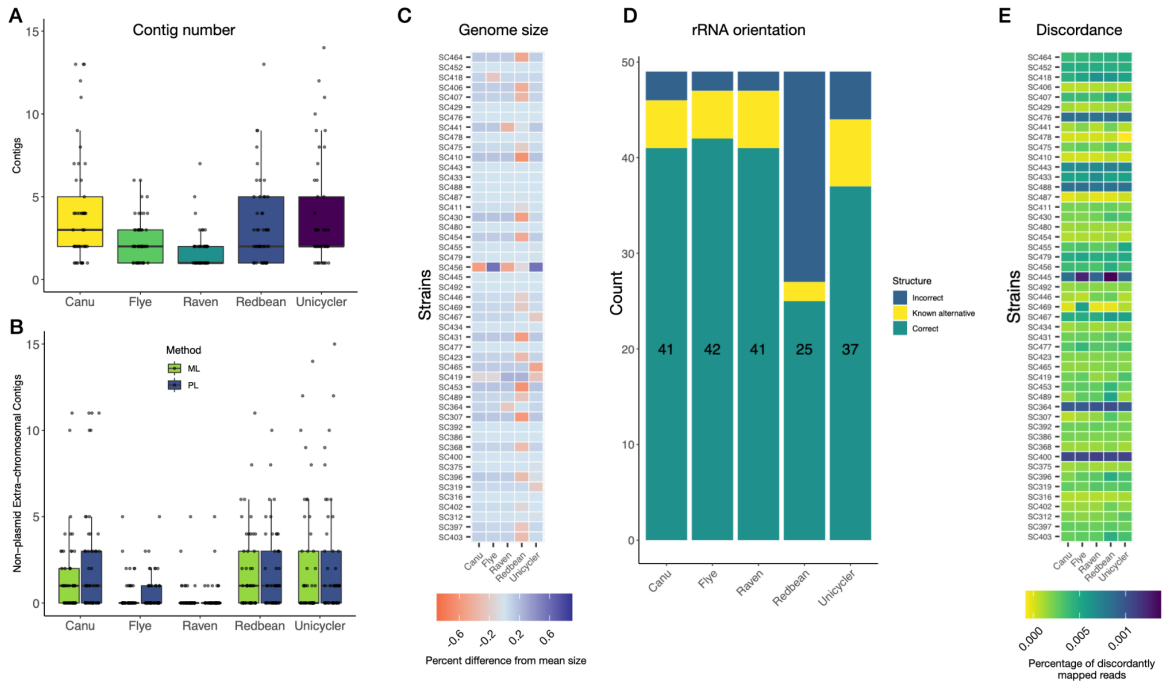


Figure 3.3. Genome structural contiguity, consistency, and accuracy across assemblers. A. Total number of contigs for each strain and assembler. Most strains assembled into five or fewer contigs, with raven consistently having the fewest. **B. Total number of non-plasmid contigs for all strains across assemblers.** We identified plasmid contigs using MLplasmids and Plasmid finder. As in panel **A**, raven assemblies consistently produced the fewest non-plasmid contigs, with the vast majority of assemblies consisting of a single chromosomal contig. **C. Difference in chromosomal assembly size for each assembler.** The difference in assembly size for each assembler from the mean of all assemblers is indicated. Differences in assembly length were generally less than 0.1% of the total genome size, equivalent to 5,000 bp for a 5 Mbp chromosome. Redbean assemblies were often the shortest, in many cases by approximately 0.5% (25 kbp for a 5 Mbp genome). Strains are ranked from top to bottom by the percentage of reads with a quality score greater than 15 (see **Fig. 1**). **D. rRNA operon orientation for all strains across assemblers.** We tested whether each of the seven rRNA operons were in an orientation that has been observed previously in E. coli. Flye, Raven, and Canu consistently produced structurally correct assemblies with expected or known alternative rRNA operon orientations (47 out of 49 genomes in the case of Raven and Flye). In contrast, Redbean produced structurally

correct assemblies for only 27 out of 49 strains. **E. Illumina mapping discordancy for each assembler.** We withheld 5,000 pairs of Illumina reads from each assembly, mapped these back on to the assembled genomes, and calculated the fraction of mapped reads that were discordantly oriented. We observed strain specific trends in discordant mapping, but no assembler-specific trends.

We next examined consistency in the size of each genome assembly across all assemblers. We calculated the mean size of the largest contig for each strain for all assemblers (under the assumption that this was the chromosomal contig) and tested whether there were systematic deviations from this mean size for each assembler. Although this is a relative metric and does not objectively establish whether one assembler is the most accurate, it provides insight into whether specific assemblers tend to arrive at the same or different results - a “wisdom of the crowds” approach. We found that the median difference in assembly size across all strains and assemblers was 39 kbp (0.0078%). The most consistent pattern was that shorter assemblies were frequently produced by Redbean (**Fig. 3.3C**). This is expected, as Redbean generally produced more fragmented assemblies that had parts of the chromosome as unincorporated contigs. Although previous work has suggested that Canu assemblies are often larger as they often contain overlaps at the start and end of chromosomes (R. R. Wick et al. 2017a), we did not observe this as a general pattern.

Single-contig circularised genomes may be structurally inaccurate, containing large inversions, deletions, insertions, or amplifications. One means of identifying structural mis-assemblies is through examining the order and orientation of rRNA operons. *E. coli* contains seven highly conserved rRNA operons. If mis-assembly occurs at these locations, these operons may be joined incorrectly to the neighbouring chromosomal regions (in orientation, chromosomal location, or both). We used Socru (Page, Ainsworth, and Langridge 2020) to assess whether rRNA operons were properly oriented within each assembly. We found that Flye produced the greatest fraction of assemblies having all seven rRNA operons correctly oriented or in a known alternative orientation (42 out of 49 in the expected orientation and five out of 49 in a known alternative orientation). Raven and Canu were similar, with 47 and 46 rRNA operons, respectively, in the expected or known alternative orientations (**Fig. 3.3D**); Unicycler had slightly fewer. However, for the Redbean assemblies, only 27 out of 49 had all operons in an expected or known alternative orientation.

We also assessed structural accuracy by mapping 5,000 paired end Illumina reads (10,000 individual reads) that we withheld during assembly (see **Methods**). We then quantified the fraction of discordantly mapped reads (i.e. the fraction of all mapped paired-end reads that were not concordantly oriented). This should yield insight on whether there are frequent small scale inversions or duplications in the assemblies. We found that across all assemblers, very few mapped discordantly, with a median of 0.26% (approximately 26 reads; IQR 0.14% to 0.36%; **Fig. 3.3E**). Although Redbean performed very slightly worse than the other assemblers, almost all differences in discordance were strain-specific rather than assembler-specific, suggesting that some strains are simply harder to assemble correctly, perhaps due to the arrangement or numbers of repetitive elements in the genome. Alternatively, the Illumina sample libraries may have differed in the fraction of artifactual chimeric reads (Marçais, Yorke, and Zimin 2015).

3.5.4 Sequence accuracy

Large-scale structural accuracy is only one assembly characteristic; accuracy at the base pair level is also a critical aspect, especially as ONT data is prone to systematic indel errors in homopolymeric regions (usually deletions). These indels can lead to frameshifts, resulting in premature stop codons and shortened open reading frames (ORFs). To estimate the extent of indels in each assembly, we quantified the percentage of unexpectedly short ORFs (see **Methods**). In the MG1655 reference genome we found that 3.3% of all ORFs were truncated (**Table 3.1**); repeating this analysis for each of the 49 other *E. coli* strains we observed that Unicycler assemblies consistently had the smallest fraction of truncated ORFs (**Fig. 3.4A**). These results suggest that even if long read-first ONT assemblies are polished with short reads, this does not completely mitigate the problem of frequent indels.

While the fraction of short ORFs gives some insight into the prevalence of deletions, there are also real differences between strains in the number of pseudogenes, which affects the fraction of short ORFs. In an effort to reduce the effects of these differences on our assessment of assembly accuracy, we used the Breseq pipeline to call SNPs and indels in each assembly using Illumina reads only. In all cases, we used the same Illumina reads that were used in the original assembly or subsequent polishing steps. We found that while on average few indels or SNPs were identified (median of 37 deletions, IQR 9 - 81; median of 0 SNPs, mean 2.9), some assemblies had considerably more - between 200 and 300 deletions and up to 40 SNPs **(Figs. 3.4B** - **4D)**. Again we found that Unicycler assemblies contained the fewest deletions across all strains. This is perhaps not surprising, as the initial Unicycler assemblies were made using these same reads. It is critical to note though that both the truncated ORF metric and the Breseq variant calling metric may be prone to

missing real SNPs and indels present in Unicycler assemblies, as they tend to occur in repetitive regions such as rRNA, as observed above for MG1655. These errors do not necessarily affect ORF length, nor are they called by Breseq. However, it is difficult to identify these false negatives with certainty in the absence of a ground truth genome. This contrasts with the long read-first assemblies, which do not appear as prone to false negative errors when assessed for quality using Breseq variant calls.

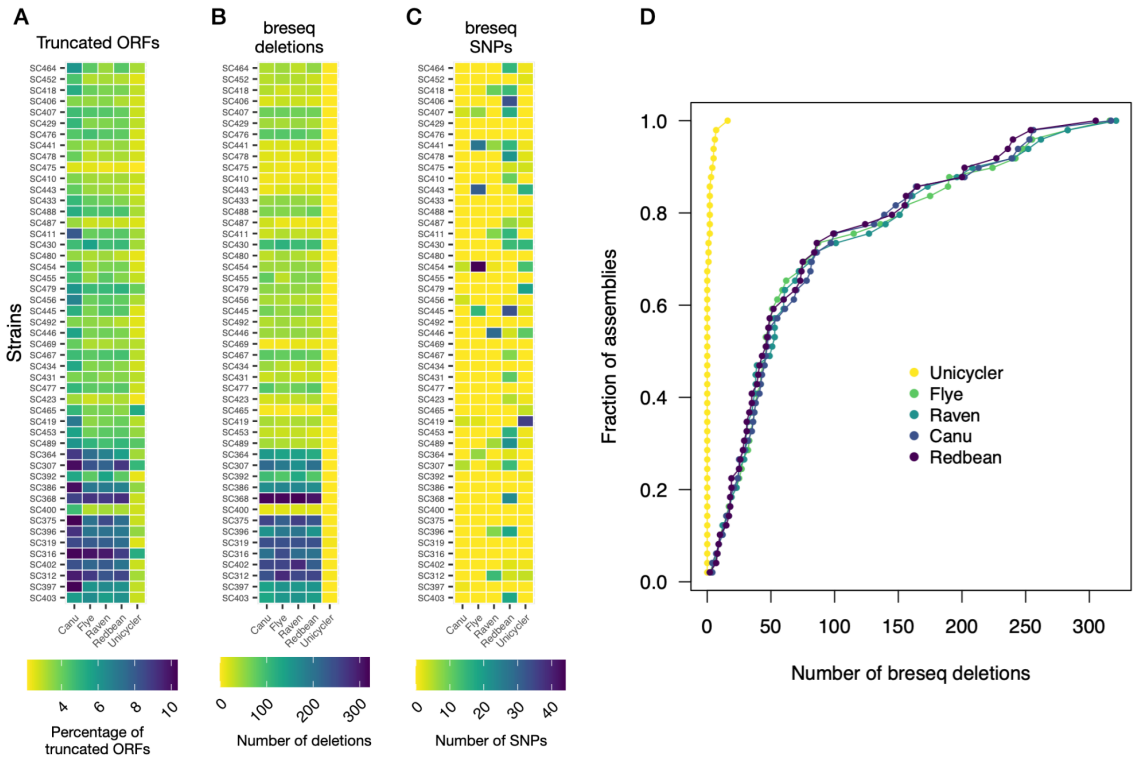


Figure 3.4. Genome sequence accuracy metrics across assemblers A. Percent of truncated ORFs for each assembler. We calculated the fraction of truncated ORFs for each assembler as those ORFs that were 90% or less of the expected length (see **Methods**). We found that Unicycler consistently exhibited a far lower fraction of truncated ORFs. **B. Number of deletions in each assembly called via Breseq using Illumina reads alone.** We used the Breseq pipeline to call deletions using Illumina reads. Unicycler assemblies consistently had the fewest deletions; this is somewhat expected as the initial Unicycler assembly was constructed using Illumina reads only. However, all other assemblies were also polished using the same set of Illumina reads. **C. Number of SNPs in each assembly called using Illumina data alone.** We found few SNPs using Breseq for any assembler, although in a small number of cases, we observed up to 40 SNPs. These tended to be both strain and assembler-specific, rather than consistent for any one assembler or any one strain. **D. Cumulative plot of the number of Breseq-called deletions for each assembler.** The majority of long read-first assemblies have fewer than 50 deletions, far more than Unicycler. However results from ground-truth MG1655 comparison suggest that there are few false negative errors in long read-first assemblies, in contrast to Unicycler.

Nevertheless, we sought to test the possibility that false negative indels and SNPs were present in the Unicycler assemblies by comparing them to the other long read-first

assemblies. We hypothesised that if indels and SNPs in Unicycler assemblies were present primarily in repetitive regions, they should be clumped along the genome in these regions compared to other assemblers. We tested this by taking the Raven assembly for each strain as a “reference,” and aligning all other assemblies against it in a pairwise fashion. We selected the Raven assemblies as the reference because they were consistently among the most accurate. From these alignment, we could infer mismatches in the assemblies, which likely arise from errors in either the Raven assembly, or in the assembly aligned to Raven. We found four different patterns in the genome-wide error profiles. For some strains, the Unicycler assemblies consistently exhibited fewer SNPs and indels across the genome (**Fig 3.5A**). In these cases, the slope of the cumulative number of SNPs and indels in the Unicycler assemblies was approximately half the slope of the other assemblers, which is expected if the Unicycler assembly generally contains few errors - the cumulative difference in SNPs and indels between the Unicycler and Raven assemblies are errors in the Raven assembly. The slope of the cumulative error curves for all other assemblies is approximately twice the Unicycler slope because they have similar numbers of errors as the Raven assemblies spread across the genome, so the cumulative error curves reflect errors in both the Raven and the other assemblies.

The second pattern we observed was one in which similar total numbers of SNPs and indels appeared for all assemblers, but in Unicycler (and often Redbean), the majority were due to concentrated stretches of errors (**Figs. 3.5B** and **C**). In these cases, the large numbers of consecutive errors in the Unicycler assemblies tended to occur at rRNA regions, confirming that when errors occur within Unicycler assemblies, they tend to accumulate in repetitive regions.

The third pattern we observed was one in which most long read assemblers exhibited very few differences, but Unicycler and Redbean exhibited long stretches of errors, again in rRNA regions or other repetitive elements (**Figs. 3.5D** and **E**). Finally, in some cases we found that errors tended to be clumped for all assemblers, but distributed across the genome (**Fig. 3.5F**). When these errors are present in the “reference” Raven genome, these are apparent as a concerted step increase in the cumulative number of errors for all assemblers; when these are present in only one assembler, this is apparent as a step increase only for that assembler. These results show conclusively that although at first glance the Unicycler assemblies are highly accurate in terms of SNPs and indels, there are a large number of errors that are present in repetitive regions but which are not easily identified found by quantifying the fraction of truncated ORFs or by calling variants using Breseq.

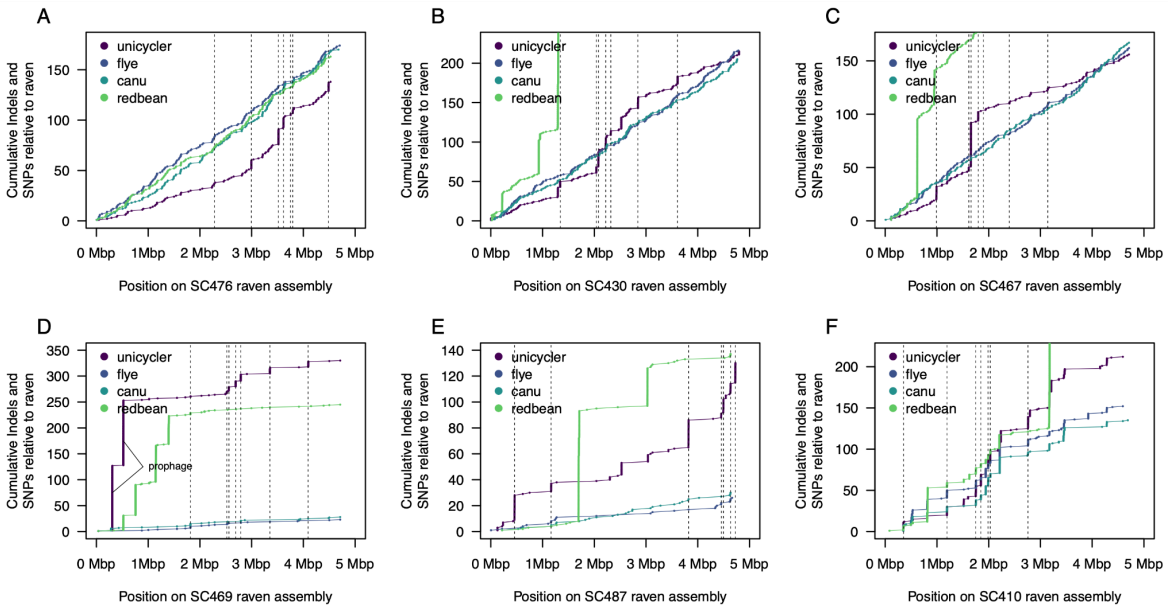


Figure 3.5. Whole genome alignments for all assemblies for six example strains. For each strain, we used the Raven assembly as a reference, aligned all other assemblies against it, and identified SNPs and indels using dnadiff. Each panel shows the cumulative number of SNPs and indels for each assembler along the genome of one strain. The location of the rRNA operons are shown as dotted vertical lines.The locations of two prophages are indicated in (**D**), where the Unicycler assembly exhibits a large number of errors. In (**B**)**,** (**C**), and (**F**), the Redbean assemblies contained a far larger number of SNPs and indels than the other assemblies, so the complete cumulative curves are not shown.

3.5.5 Effects of data quality on assembly accuracy

In addition to assembler-specific errors, we expected that differences in assembly accuracy might be dependent on the quality of the input data, which often differed between strains. Indeed, we observed that differences in accuracy were often strain-specific rather than assembler-specific. For example, some strains exhibited consistently higher discordance (**Fig. 3.3E**) or systematically higher indel numbers (**Fig. 3.4B**). These differences have two possible sources - the quality of input data, or characteristics intrinsic to each strain that preclude accurate assembly. Both the Illumina and ONT data quality differed between strains in read length and quality (**Fig. 3.2** and **Fig. S.3.2**). Variation in ONT read length was almost certainly due to subtle differences in sample treatment during genomic DNA prep. However, we did not expect substantial variation in ONT read quality, as library chemistry, flow cell chemistry, and basecalling methods were identical for all strains (9.4.1 flow cells with

RBK004 chemistry and guppy 4.2.2). Surprisingly, we observed differences in read quality even for library samples run on the same flow cell and on the same day.

We first tested for effects of input data quality on assembly structural accuracy by quantifying correlations between ONT read length (read N50) and any accuracy metric. We found no strong correlation between read length and any of the structural quality metrics **(Fig. S.3.3A-D)** for any assembler, although Redbean assemblies showed a trend for lower structural accuracy as assessed by rRNA operon orientation (**Fig. S.3.3A**, rightmost panel). We also found no correlation between read length and any sequence accuracy metric (**Fig. S.3.4**).

We also tested for correlations between ONT read quality (**Fig. 3.2B**) and assembly sequence accuracy. The result from the initial set of assemblies suggested that strains with lower read qualities exhibited both increased levels of truncated reading frames and Breseq-called indels (**Fig. 3.4**). However, this pattern was not clearcut. For example, strain SC400 exhibited very few deletions or truncated reading frames despite having relatively low quality ONT data. Interestingly, this strain is very similar to strain SC402 both phylogenetically and in genome size (**Fig. 3.1**), but strain SC402 had considerably more truncated ORFs and indels (fourth row from the bottom in the **Fig. 3.4** heatmaps) despite having very similar read quality scores. We currently do not have an explanation for this discrepancy.

To clarify the effects of ONT read quality on assembly accuracy, for six strains we collected between three and six additional sets of read data, either using different flow cells, or as independent multiplexed libraries on the same flow cells. These read sets exhibited varying levels of per read quality distributions (**Fig. 3.6A**). For each of these read sets we then performed assemblies using all five assembly methods, and calculated the sequence accuracy metrics for each (**Fig. 3.6B - D**). We observed consistent differences in assembly quality between strains, with some strains exhibiting more than 150 deletions and more than 6% truncated ORFs regardless of the ONT read set or assembly method (with the exception of Unicycler assemblies). In other cases, strains with read sets of nearly the same quality exhibited fewer than 30 deletions on average and less than 4% truncated ORFs. This strongly suggests that there are strain-specific characteristics that affect assembly accuracy. As for read length, we also checked for correlations between read quality and assembly sequence accuracy, and found no strong relationship between any metric for any assembler (**Figs. S.3.5** and **S.3.6**).

However, all of the assemblers use short read Illumina data at some stage, either for initial assembly (Unicycler) or for polishing, and in the above results, the same Illumina data were used for polishing all of the assemblies for each strain. We thus considered whether the

quality of Illumina reads affected assembly sequence accuracy, as there were clear differences in read quality (**Fig. S.3.2**). Two circumstantial pieces of evidence suggest that Illumina read quality did not have strong effects on assembly sequence accuracy. First, the two strains with the lowest average Illumina read quality, SC455 and SC400, did not exhibit substantially lower assembly metrics (in fact SC400 exhibited better assembly metrics than average). Second, the quality of the Unicycler assemblies, which depend first and foremost on the Illumina reads for assembly, did not show any clear correlation with read quality.

To directly test the effects of Illumina read quality on assembly accuracy, rather than collect additional data, we divided the MG1655 Illumina data into sets of high quality and low quality reads, such that the high quality set had mean quality scores of 38-39, and the low quality set had mean quality scores of 33 to 34 (**Fig. S.3.7**). We assembled each read set using Unicycler and compared these to the ground truth MG1655 assembly using dnadiff for genome-wide alignment. Surprisingly, we found that the low quality read set resulted in an assembly with the fewest errors (60 SNPs and indels), compared to 65 SNPs and indels for the high quality set and 71 SNPs and indels for the full read set. This suggests that Illumina read quality has little, if any, effect on assembly quality.

Overall, these results suggest that neither Illumina nor ONT read quality (as calculated by the guppy basecaller) strongly affect the resulting hybrid assemblies. Rather, for long read assemblies (but not for short-read first Unicycler assemblies), there are unidentified idiosyncratic characteristics for each strain that affect assembly accuracy, as assessed by quantifying the fraction of truncated ORFs and Breseq-called SNPs and indels. There remain two possible explanations for these differences. First, it is possible that some strains have low-accuracy assemblies because they do in fact rely on low quality ONT data, but this quality is not reflected in the quality scores assigned by the base caller. This would manifest as reads having high quality scores but low percent identity to the assembly. Alternatively, some strains may have low-accuracy assemblies because of fundamental problems in creating accurate consensus sequences.

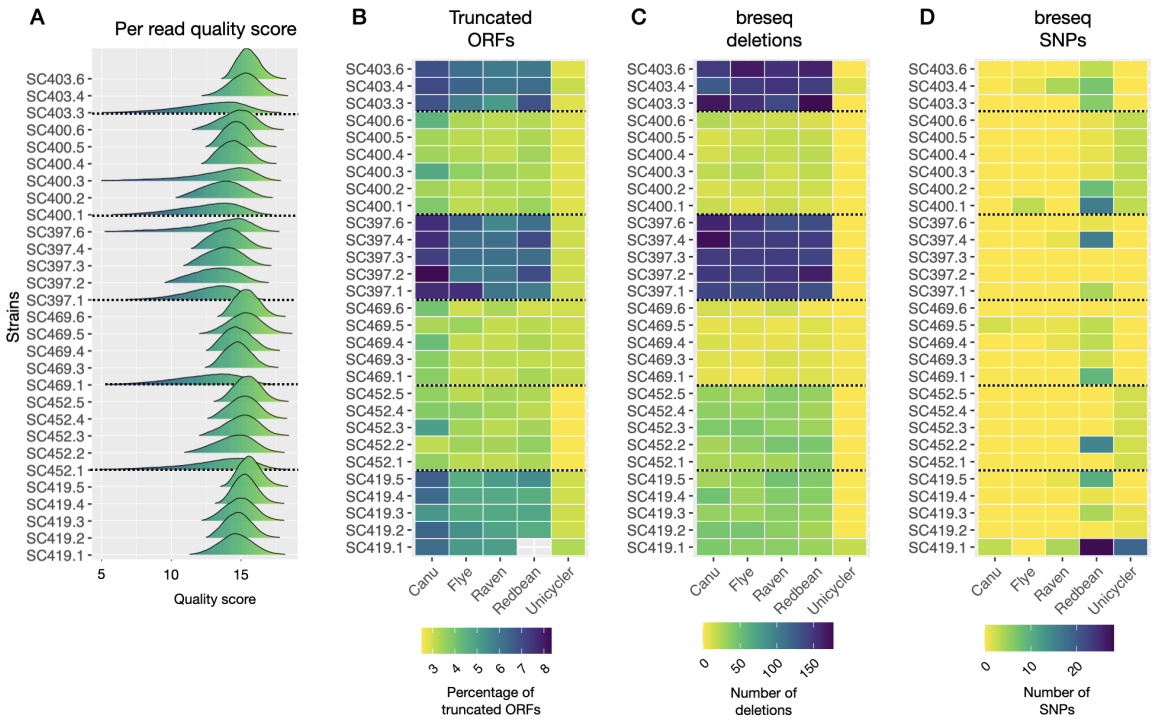


Figure 3.6. Assembly accuracy is strain dependent and does not correlate with the quality of ONT data. We collected between three and six additional ONT datasets for six strains and then performed assemblies using all five assemblers. **A.** Mean per read quality scores for each data set. The data are arranged by strain, and within strains, by the fraction of reads that are above Q15. **B.** Fraction of truncated ORFs per assembly. **C.** Number of deletions called by Breseq. **D.** Number of SNPs called by Breseq.

To distinguish between these two possibilities, we examined the relationship between read identity to the assembly and read quality score for three strains that exhibited varying levels of assembly quality (SC403, SC419, and SC452; **Fig. 3.4**). We found that all three matched almost exactly in the relationship between read identity and quality score, suggesting that the quality score assigned to reads by the guppy basecaller truly reflects the quality of the reads (**Fig. S.3.8**). These results imply that the primary limitation for accurate assembly is decreased consensus accuracy when doing long read-first assemblies. These intrinsic problems in consensus accuracy are easily diagnosed by comparing assembly metrics between the short-read first Unicycler assembly and the long-read first assemblies. When these metrics are mismatched, this suggests there are problems in consensus read accuracy for the long read-first assemblies.

3.6 Conclusion

Here we have investigated five assembly methods for their ability to generate accurate hybrid genome assemblies from Illumina and Oxford Nanopore data. We have employed three metrics of assembly quality that assess structural accuracy: assembly contiguity, discordancy of short read mapping, and the accuracy of rRNA region arrangement. We have used two metrics to assess sequence identity: the fraction of truncated reading frames observed, the number of indels and SNPs called using short read (Illumina) data.

The data here suggest that Raven and Flye are the top performing assemblers from a structural standpoint, while Unicycler is most effective at minimising indels and substitution errors outside of repetitive regions. These assemblers thus offer complementary advantages. If structural accuracy is critical for the question at hand (for example quantifying genome dynamics for rapidly evolving regions such as IS elements or plasmids), then Raven and Flye are preferable. For other applications, such as evolve and resequence experiments(Long et al. 2015), Unicycler is the preferable method. It is important to note that the metrics used here to assess assembly sequence accuracy likely overestimate the accuracy of Unicycler assemblies - when there are errors in Unicycler assemblies they are most likely in repetitive regions, such as rRNA (**Fig. 3.5**), and are not easily identified using the assembly accuracy metrics here.

We also found that there are intrinsic properties of certain bacterial strains that limit consensus accuracy. We suggest that by contrasting the qualities of short-read first Unicycler assemblies and long-read first Raven or Flye assemblies, these problems in assembly accuracy can be easily diagnosed. Finally, these results emphasise the importance of consensus assembly methods, such as that offered by tricycler (R. R. Wick et al. 2021b).

3.7 Supplementary data

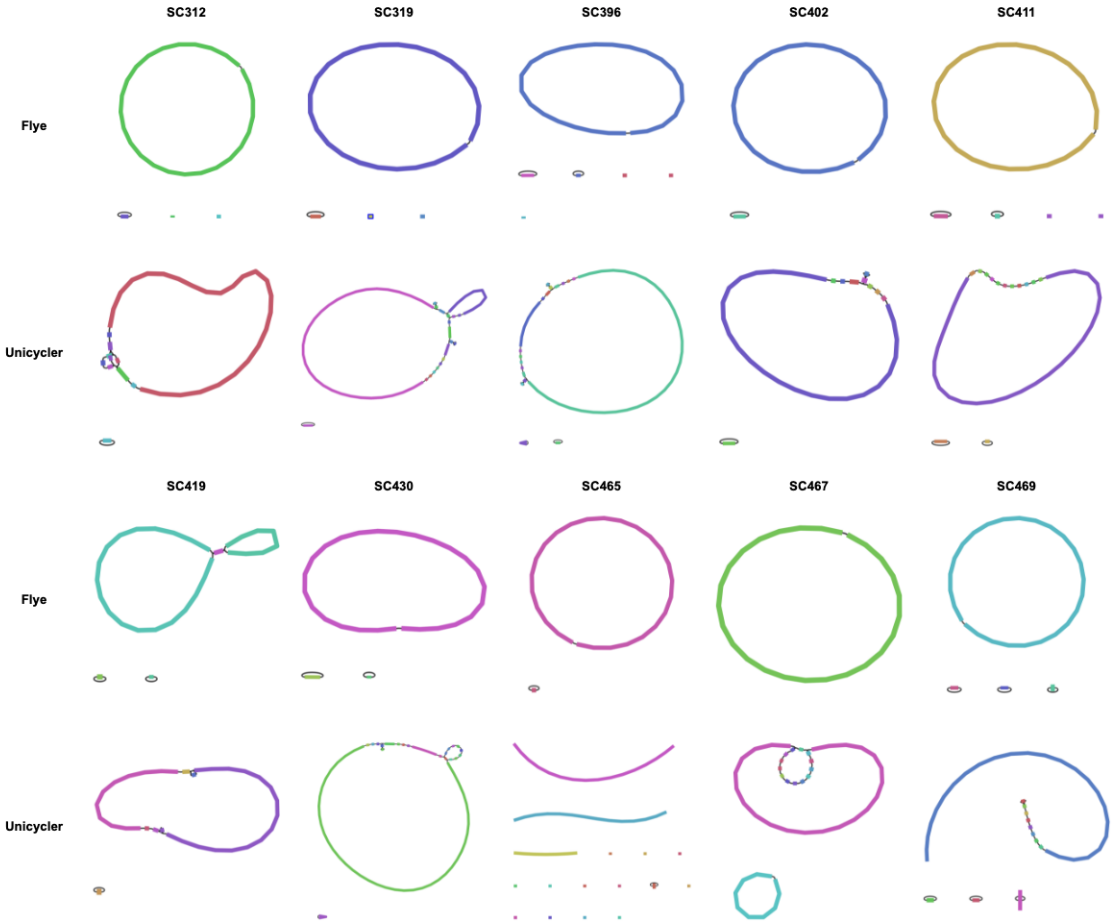


Figure S.3.1: Differences in assembly structure between assemblers. As an illustrative example, here we show bandage (R. R. Wick et al. 2015) plots of the 10 most fragmented Unicycler assemblies, coupled with the Flye assemblies for the same strain. Unicycler often produced more fragmented assemblies, while Flye was the second most contiguous assembler. We typically observed the introduction of multiple small contigs within the chromosomal contig of Unicycler assemblies.

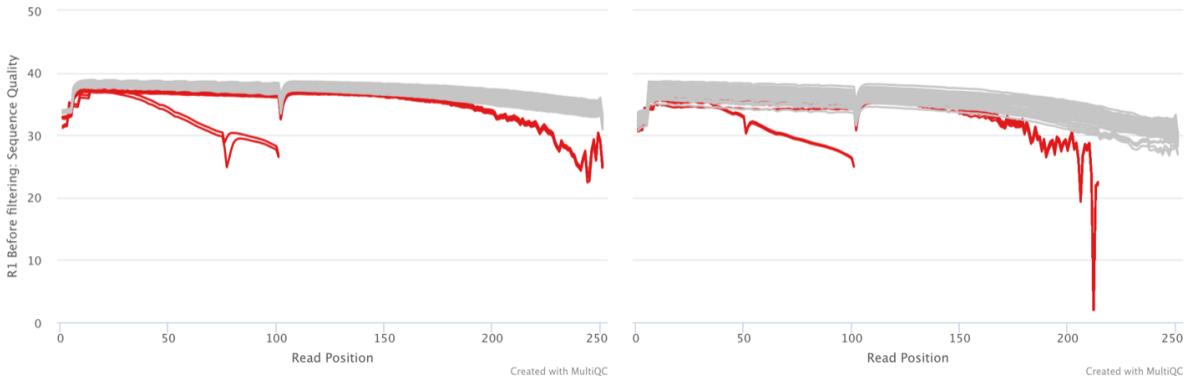


Figure S.3.2. Illumina data was systematically lower quality scores for a subset of strains. Mean quality scores at each position are indicated, calculated using fastp (S. Chen et al. 2018) and displayed using multiQC (Ewels et al. 2016). Each strain is indicated with a single line, with Read 1 shown on the left panel and Read 2 shown on the right panel. In red are the 14 strains that were sequenced on a date previous to the other 35 strains (grey lines). Two of these strains (SC455 and SC400) that were sequenced on a previous date using PE 100bp reads, in contrast to all other strains.

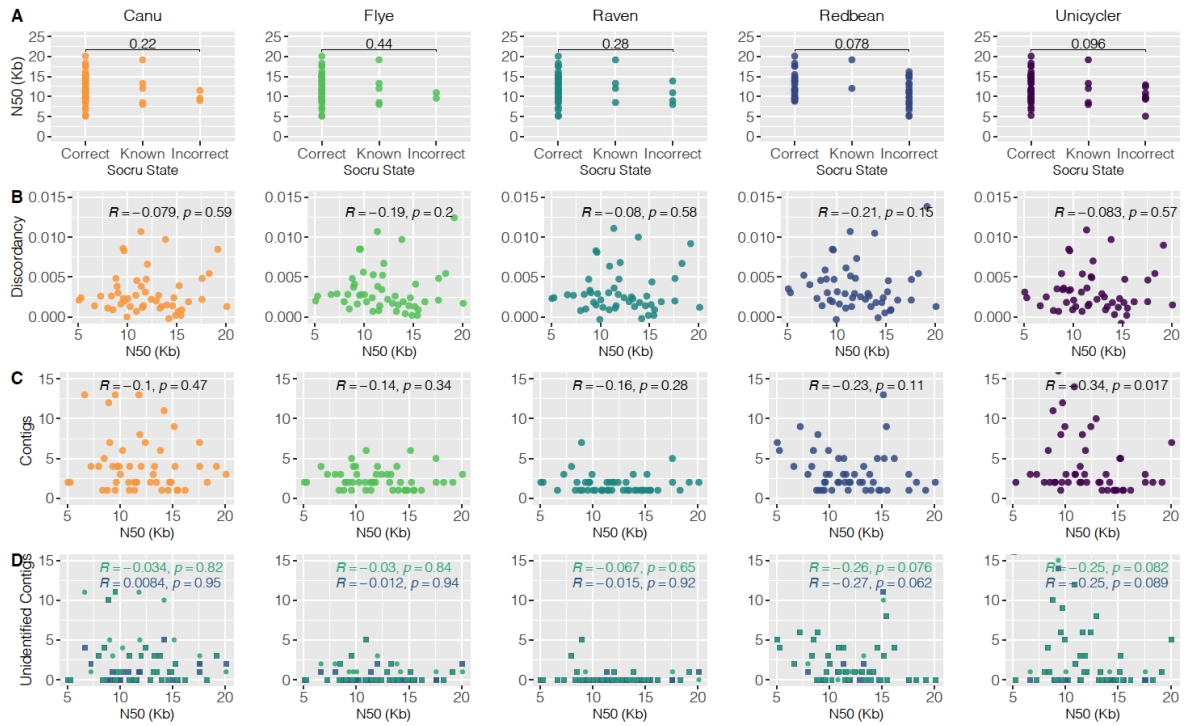


Figure S.3.3. No structural accuracy metrics significantly correlate with read length across strains. A. rRNA operon ordering vs. read N50. We tested for significant differences in the read N50 values for correctly or alternatively oriented rRNA operons and incorrectly ordered operons. For all assemblers except Redbean, we did not observe lower N50 values for assemblies with incorrectly oriented operons. **B. Read mapping discordancy vs. read N50.** We found no correlation between the fraction of discordantly mapped reads and N50. Spearman's rho and corresponding p-values (before correction for multiple comparisons) are shown in each plot. **C. Contig number vs. read N50** and **D. Unidentified contig number vs. read N50.** We found no strong correlation of either contig number or the number of extrachromosomal non-plasmid contigs ("Unidentified contigs") on read N50. As in **(B)** rho and p-values are shown in each plot.

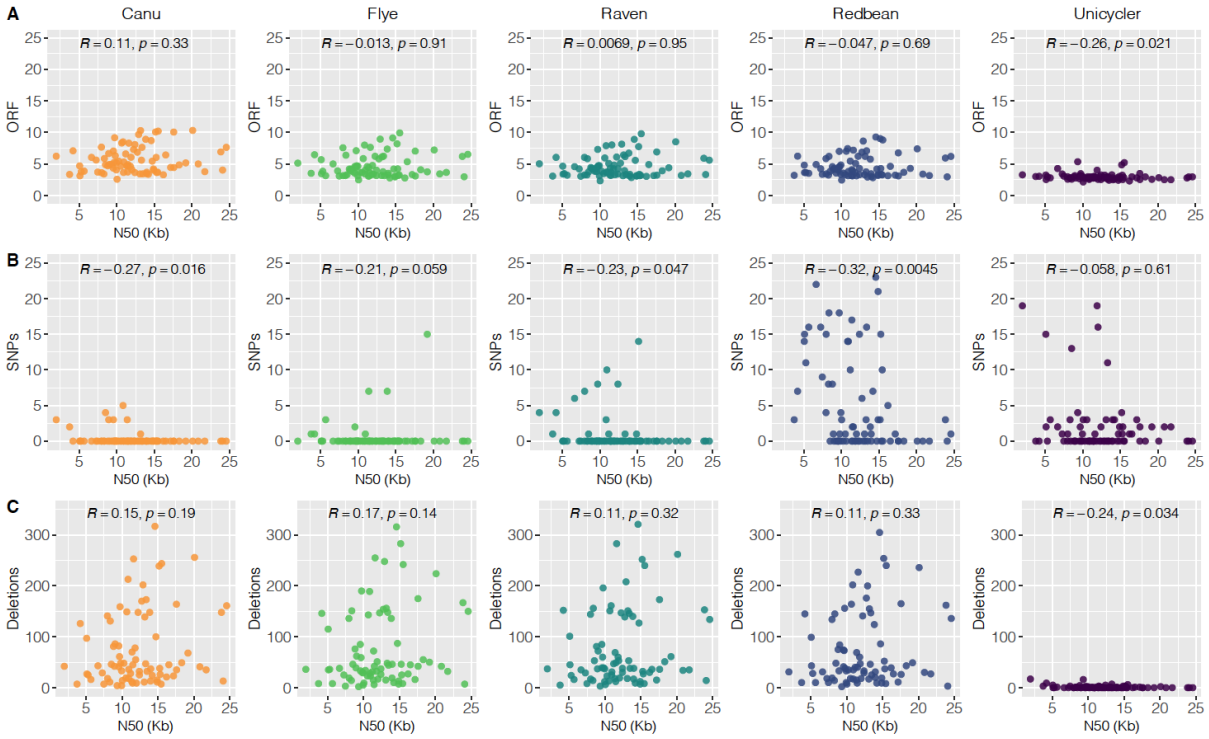


Figure S.3.4. No sequence accuracy metrics significantly correlate with read length across strains. A. Percentage of short open reading frames. B. Number of SNPs. C. Number of deletions. SNPs and deletions were called using the Breseq pipeline. Spearman's rho and the corresponding p-values before any correction for multiple comparisons are displayed on each plot.

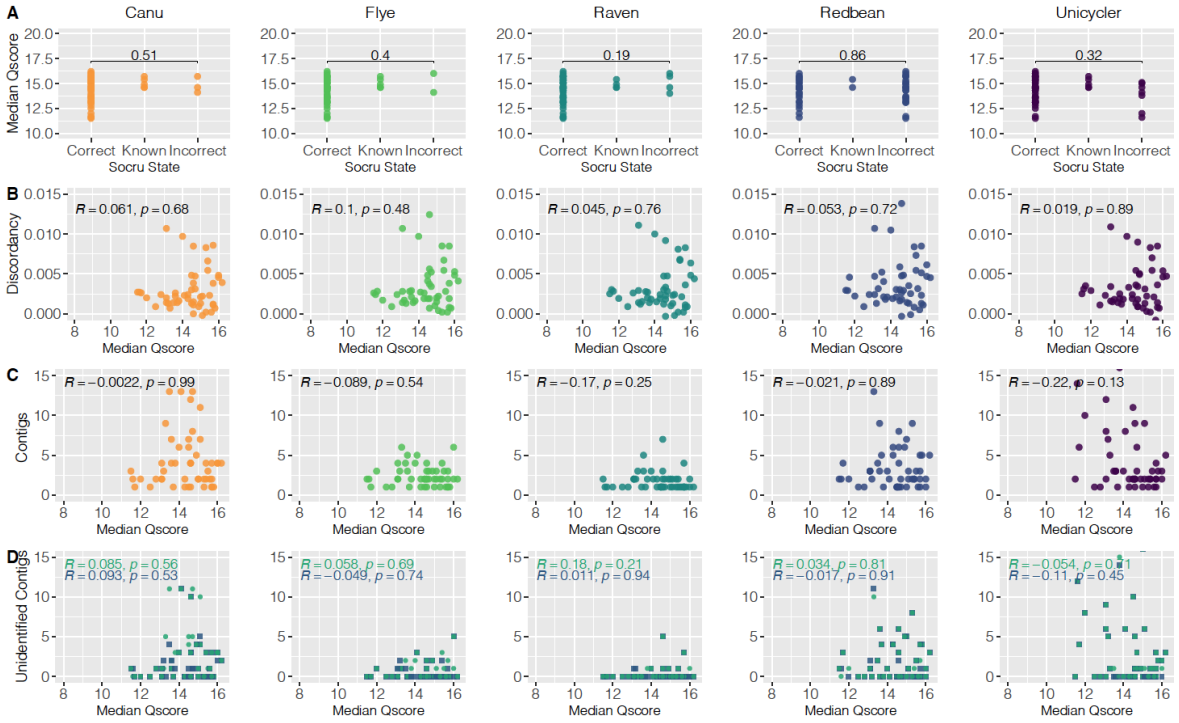


Figure S.3.5. Differences in read quality does not significantly correlate with assembly quality across structural metrics. A. rRNA operon arrangement. B. Illumina discordancy. C. Total contigs D. Unidentified contig number. No significant correlation was found across any of the assemblers with the tested metrics. Spearman's rho and corresponding p-values are displayed on each plot.

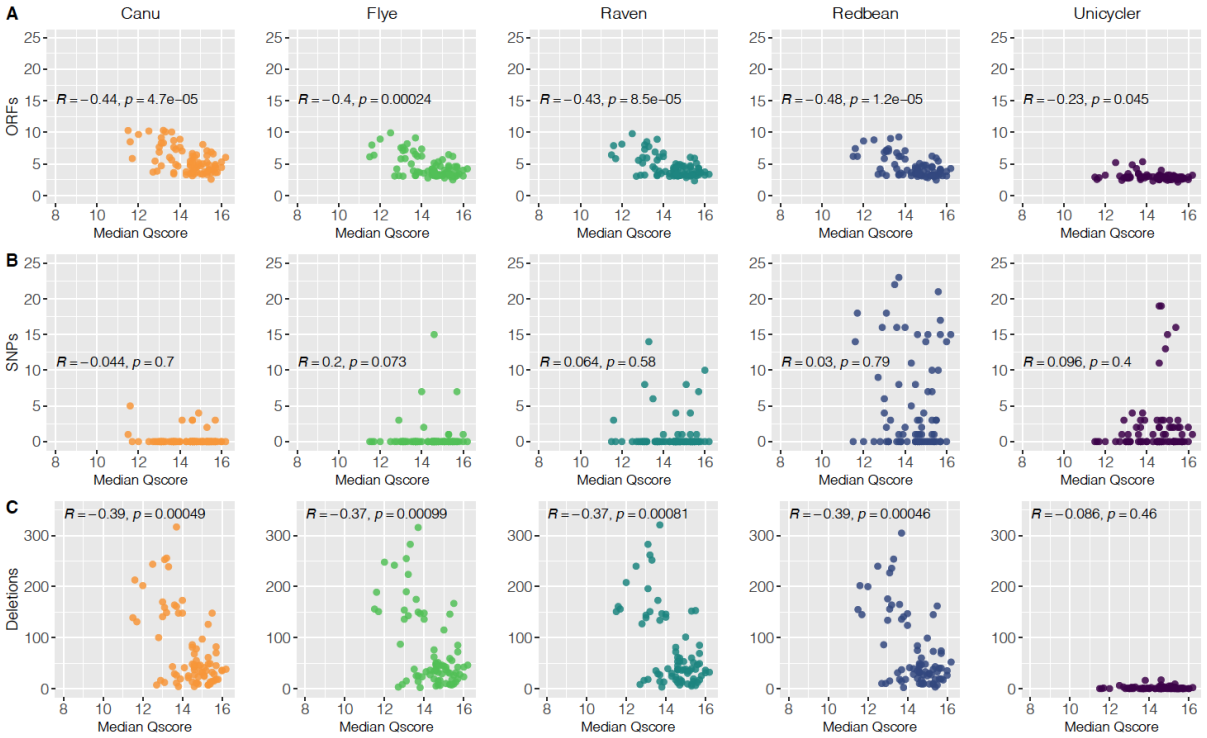


Figure S.3.6. Differences in read quality does not significantly correlate with assembly quality across sequence accuracy metrics. A. Truncated ORFs B. Breseq-called SNPs. C. Breseq-called deletions. No significant correlation was found across any of the assemblers with the tested metrics. Spearman's rho and corresponding p-values are displayed on each plot.

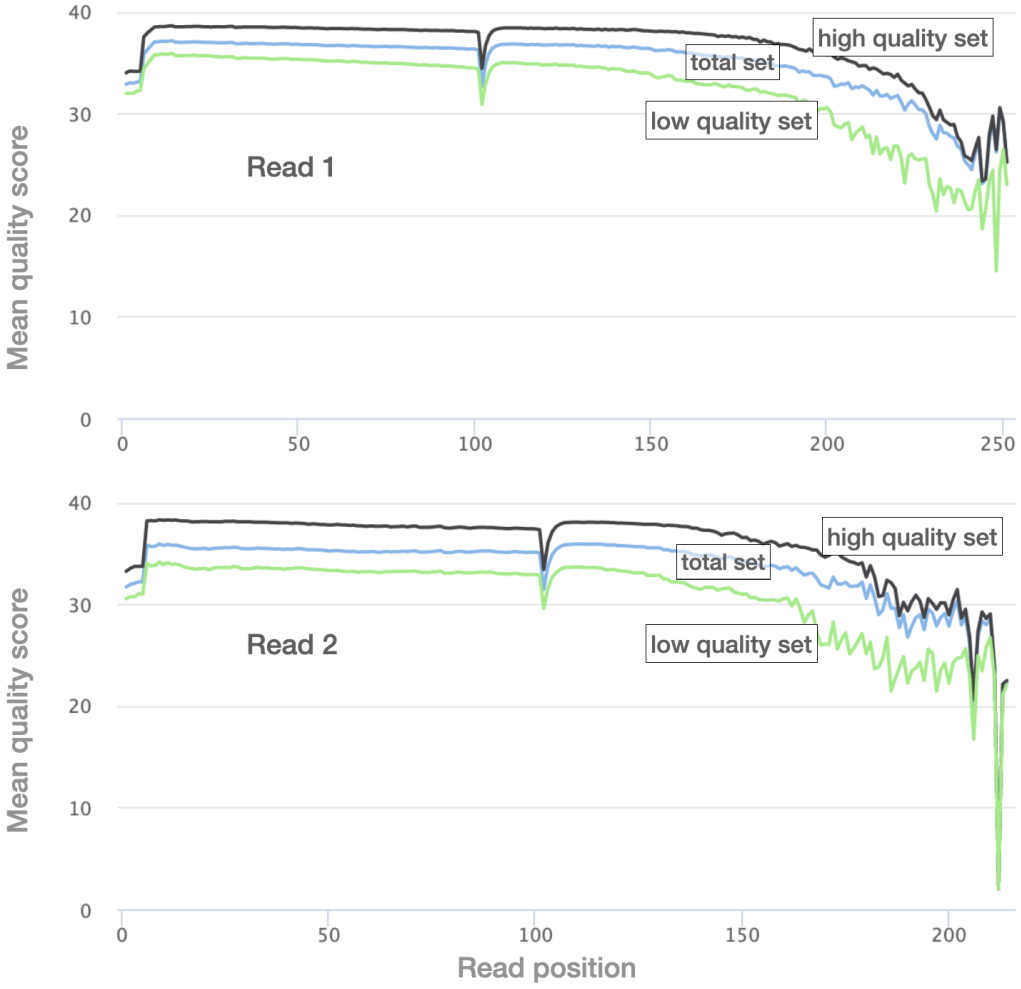


Figure S.3.7. Differences in read quality for "high quality" and "low quality" MG1655 read sets. We used filtlong to divide the MG1655 Illumina data into two read sets, one with high quality reads and the other with low quality reads. Mean quality scores at each position are indicated, calculated using fastp and displayed using multiQC. These different read sets were then used to test the effects of read quality on assembly accuracy.

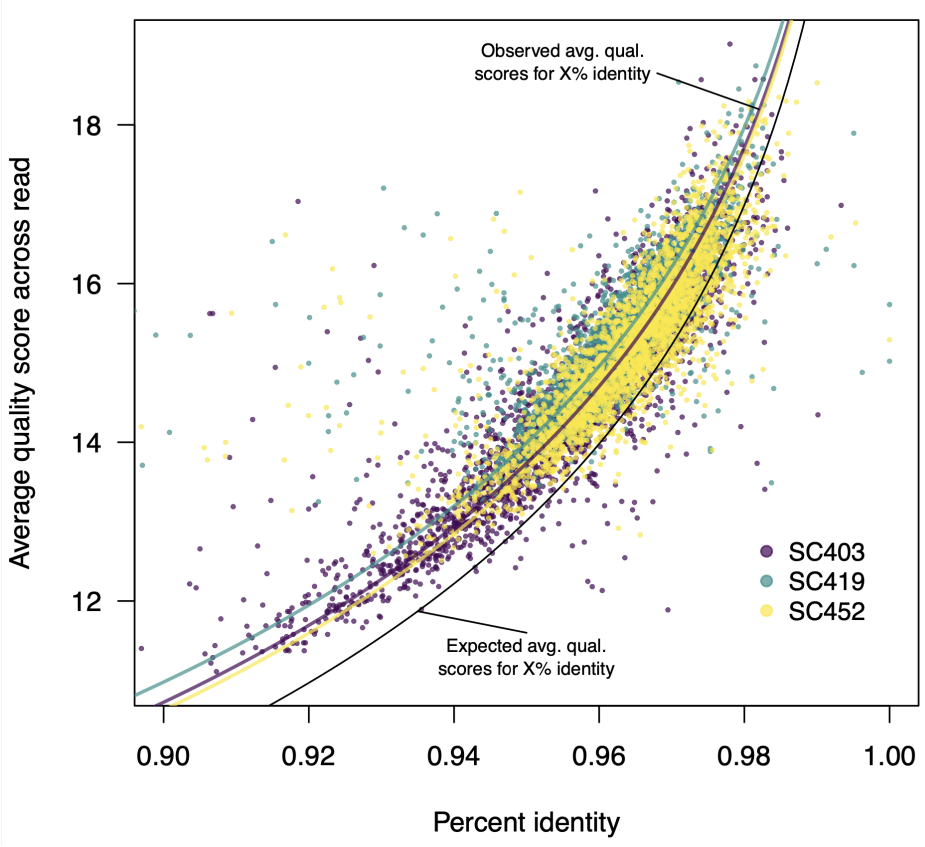


Figure S.3.8. The relationship between read identity and quality score is consistent across strains. We selected three strains (SC403, SC419, and SC452) that differed considerably in the number of truncated ORFs and indels. For each strain, we aligned all reads from a single ONT run to the assembled unicycler genome using minimap2 (H. Li 2018). We then calculated the percent identity of each read. In the figure above, each point is one read, showing the percent identity of that read and the mean quality score across the read. For clarity, only a subset of the total read numbers are plotted. The black line shows the expected relationship between read identity and average quality. For example we would expect a Q20 read to be 99% accurate and thus have 99% identity to the reference. We generally observed that the mean quality scores were systematically higher than would be expected given the percent identity. For example, a read with 94% identity should, on average, be assigned Q12.2, but we observe most reads with 94% identity assigned as having Q12.8. However, this systematic increase is the same for all strains, showing that Q scores are consistently assigned by the guppy basecaller. The purple, green, and yellow lines indicate a non-linear least squares fits for the relationship between read identity and Q score using an additional parameter that can account for systematic increases or decreases in the relationship between read identity and Q score: $Q = -\log10(1 - P)*10 + \alpha$, in which Q is the quality score, P is the percent identity, and α is a parameter allowing a systematic difference in quality score from what would be expected.

Table S.3.1. Mutations in the laboratory strain of *E. coli* K12 MG1655. We identified mutations in our laboratory isolate of MG1655 using Illumina reads and the Breseq pipeline, which finds only high quality, unambiguous SNPs, indels, and structural rearrangements. The table below shows the differences between our lab strain, ATCC 700926 (distributed by ATCC as the MG1655 genotype), and the MG1655 reference.

Coordinate	Genomic region	Type of mutation	Lab	ATCC 700926	MG1655 reference
257899	crl	INS	IS1+8	IS1+8	G
547694	ylbE	SNP	G	G	A
547832	ylbE	INS	G	G	—
1298719	oppA-ychE	INS	IS5	IS5	T
1871055	yeaJ	INS	IS1+9	—	—
2171386	gatC	INS	CC	CC	—
3558478	glpR	DEL	—	—	G
3957957	ppiC-yifO	SNP	T	T	C
4294404	gltP-yjcO	INS	GC	—	—

Table S.3.2. Summary of Oxford Nanopore sequencing data for each flow cell. All data were collected using FLO-MIN106 flowcells and either SQK-RBK004 or SQK-LSK109 library preps.

Run date	Multiplexed samples	Total Mbp	Average mbp per barcode	DNA extraction method	Library
2018.06.29	6	2048	341	Promega	SQK-RBK004
2018.08.03	5	2302	460	Promega	SQK-RBK004
2018.08.10	12	3919	327	Promega	SQK-RBK004
2018.08.20	4	6315	1579	Promega and Phenol	SQK-RBK004
2018.09.04	6	8152	1359	Phenol	SQK-RBK004
2018.09.26	10	4516	452	Phenol	SQK-RBK004
2018.09.29	8	6055	757	Phenol	SQK-RBK004

2019.02.12	10	4034	403	Phenol	SQK-RBK004
2019.02.18	11	11204	1019	Phenol	SQK-RBK004
2019.02.20	10	6077	608	Phenol	SQK-RBK004
2019.03.20	5	4145	829	Promega	SQK-RBK004
2019.03.20	8	4324	540	Promega	SQK-RBK004
2021.02.03	1	4547	378	Promega	SQK-RBK004
2021.03.12	5	14850	1350	Promega	SQK-RBK004
2021.04.07	7	12703	1154	Promega	SQK-LSK109
2021.06.04	8	9461	788	Promega	SQK-LSK109
2021.07.12	8	5562	427	Promega	SQK-RBK004

Table S.3.3. Filtered and Unfiltered ONT read quality. ONT datasets were filtered to 500mbp using filtlong.

Strain	Total reads (thousands)	Total Mbp	Read N50	Mean read quality score	Reads greater than Q15	% reads greater than Q15	Filtered mean read Q score	Filtered reads greater > Q15	Filtered % greater than Q15	Filtered read N50
SC319	108	572	10613	10.7	5479	5	11.4	6054	8.2	10836
SC307	128	753	13855	11.3	9476	7.4	13.4	10037	16.7	15159
SC312	133	770	12196	9.9	3062	2.3	12	3526	5.5	12913
SC316	112	817	14402	10.3	3798	3.4	12.6	4430	8.2	15512
SC364	280	1734	11951	10.3	9491	3.4	14.1	10740	19.7	13854
SC368	256	1637	12543	10.3	6178	2.4	13.8	7412	13.9	14585
SC375	109	1118	18820	10.2	3656	3.4	13.3	4297	10.9	20095
SC386	156	1344	16337	10.3	5363	3.4	13.7	6328	14.1	17574
SC392	96	612	14009	11.5	7596	7.9	12.8	8494	14.2	14720
SC396	245	1117	9006	10.3	5744	2.3	13.1	7367	9.2	9719
SC397	113	634	12518	10.2	2810	2.5	11.5	3168	4.7	13167
SC400	206	966	10292	10.5	6863	3.3	13.1	8452	11.4	11357
SC402	208	1208	10493	9.9	3889	1.9	13.1	5001	7.6	11601
SC403	204	762	7564	9.9	2906	1.4	11.8	4429	4.6	8354

SC406	416	1888	7970	13.4	107636	25.9	15.8	45234	84.4	10902
SC407	601	2360	7231	12.9	108432	18	15.7	41525	78	11395
SC410	225	1024	10591	12.9	39955	17.8	15.4	31742	62.8	15439
SC411	120	789	11313	13.6	29431	24.6	15	28641	52.5	12411
SC418	485	1975	7463	13.5	138437	28.5	16	48364	89.3	10935
SC419	193	735	8384	12.1	20720	10.7	14.1	21660	26.1	9556
SC423	20	155	16064	13.4	5142	26	13.7	5636	34.2	16195
SC429	240	1131	9096	13.4	66192	27.6	15.7	44740	76.5	11500
SC430	186	474	4628	13.9	58761	31.5	14.7	54388	49.3	5068
SC431	264	645	4563	12.3	44945	17	14	46603	35.7	5274
SC433	79	681	17204	13.9	28474	36.1	15.4	28004	62.3	18304
SC434	88	494	10068	13.6	25734	29.3	14	28242	39.3	10253
SC441	295	1043	6160	13.6	84888	28.8	15.7	58850	73.5	7982
SC443	173	933	10529	12.5	39064	22.5	15.4	36855	62.7	12014
SC445	27	270	19143	13.7	9314	33.9	14	9691	41.5	19177
SC446	35	261	13212	13.8	11336	32.5	14.1	12373	40.8	13288
SC452	392	2286	11753	13.6	114598	29.3	16.1	32602	94	17590
SC453	133	513	6911	12.6	21974	16.5	13.1	24264	25	7253
SC454	275	936	6855	12.2	41858	15.2	15	42007	48	8504
SC455	92	545	11361	13.4	28341	30.8	14.4	29913	43.9	11811
SC456	28	127	8608	13.7	8736	31.4	14.2	9066	43.2	8935
SC464	537	2718	9395	13.4	150196	28	16.2	36666	95.9	15242
SC465	109	549	9016	12.7	20274	18.6	13.5	23119	28.4	9315
SC467	110	519	8596	13.3	30981	28.2	13.8	33226	39.4	8808
SC469	77	423	9758	13.6	22484	29.2	14	24802	39.7	9939
SC475	211	953	8151	13.5	57510	27.2	15.5	46967	67.6	10042
SC476	365	1306	7094	13	82021	22.5	15.7	48656	73.5	9544
SC477	205	785	7747	12.1	24497	12	14.5	28549	35	9030
SC478	127	981	13971	13.9	45756	36.2	15.6	33812	70.9	14886
SC479	79	489	11705	13.9	27559	35.1	14.2	27865	43.7	11893
SC480	77	584	13064	13.8	24586	32	14.8	27378	48.7	13475
SC487	110	677	12907	13.5	32146	29.3	15.1	31947	53.8	14199
SC488	353	1243	7044	12.9	70019	19.9	15.4	41265	61.3	9652
SC489	176	530	9096	12.1	21174	12.1	12.8	24685	23	6644

SC492	52	410	13946	13.8	16665	31.9	14.1	18436	40.8	13989

Chapter 4

Growth Condition Dependent Differences in Methylation Implies Transiently Differentiated DNA Methylation States in *E. coli*

Georgia Breckell, Olin K. Silander

Article submitted for peer-review

(Microbial Genomics)

Author contributions:

Georgia Breckell: Conceptualization (equal); Methodology (lead); Investigation (lead); Visualisation (lead); Writing-original draft (lead); Writing-review & editing (equal).
Olin K. Silander: Conceptualization (equal); Methodology (supporting); Funding acquisition(lead); Resources (lead); Supervision (lead); Writing-original draft (supporting); Writing-review & editing (equal).

4.1 Preface

The following chapter has been published as a preprint on bioRxiv (DOI: https://www.biorxiv.org/content/10.1101/2022.03.24.485589v1.full) and was submitted for peer review to the journal Microbial Genomics.

DNA methylation is ubiquitous yet not well understood in bacteria. Prior to the development of long read sequencing approaches such as ONT or SMRT sequencing detection of bacterial DNA methylation was limited to lab based (i.e. restriction enzyme digests), or adaption of protocols designed for the detection of Eukaryotic methylation. Eukaryotic DNA methylation is dominated by 5mC which predominantly occurs in the context of CpG islands. In contrast, bacterial DNA methylation is most commonly 6mA, and can occur in a range of sequence motifs throughout the genome. This means that techniques such as bisulfite sequencing which detects 5mC is not able to detect the predominant bacterial methylation structure. Therefore, most bacterial methylation studies have focused on the detection of 5mC or precise roles of 6mA, with a bias towards clinically relevant pathogenic isolates. These studies have shown DNA methylation to play a role in phase variable control of gene regulation, suggesting bacterial DNA methylation may function as an epigenetic marker. In order to better understand whether DNA methylation functions as an epigenetic marker in bacterial cells, genome wide analysis of methylation patterns is required. Here we present an investigation into the genome wide patterns of DNA methylation in 3 natural isolates of *E.coli* using Nanopore sequencing data. Additionally, epigenetic systems are frequently observed to control cellular responses to environmental cues, we test our hypothesis that DNA methylation patterns respond to growth conditions by examining genome wide methylation across five growth conditions.

Olin and I conceived this project and designed the experiments. I carried out all the sequencing and data curation for this project and I performed most of the bioinformatics analysis presented in this chapter. Olin performed the analysis and produced Figure S1 and S8. This manuscript was written as a collaboration with Olin.

The formatting of this paper was modified slightly from the manuscript available at bioRxiv to ensure consistency throughout this thesis.

4.2 Abstract

DNA methylation in bacteria frequently serves as a simple immune system, allowing recognition of DNA from foreign sources, such as phages or selfish genetic elements. It is not well established whether methylation also frequently serves a more general epigenetic function, modifying bacterial phenotypes in a heritable manner. To address this question, here we use Oxford Nanopore sequencing to profile DNA modification marks in three natural isolates of *E. coli*. We first identify the DNA sequence motifs targeted by the methyltransferases in each strain. We then quantify the frequency of methylation at each of these motifs across the genome in different growth conditions. We find that motifs in specific regions of the genome consistently exhibit high or low levels of methylation. Furthermore, we show that there are replicable and consistent differences in methylated regions across different growth conditions. This suggests that during growth, *E. coli* transiently differentiates into distinct methylation states that depend on the growth state, raising the possibility that measuring DNA methylation alone can be used to infer bacterial growth states without additional information such as transcriptome or proteome data. These results provide new insights into the dynamics of methylation during bacterial growth, and provide evidence of differentiated cell states, a transient analogue to what is observed in the differentiation of cell types in multicellular organisms.

4.3 Introduction

Cellular phenotypes are determined not only by genetic and environmental factors, but also epigenetic factors (heritable changes to the phenotype which are not caused by changes to the DNA sequence). In bacteria, epigenetic inheritance of phenotypes is known to occur via a range of mechanisms, including transgenerational inheritance of transcription factors or membrane transport proteins (Lambert and Kussell 2014; Kaiser et al. 2018), protein aggregates (Govers et al. 2018), or by covalent modifications to DNA, such as methylation (Sánchez-Romero and Casadesús 2020; Hale, van der Woude, and Low 1994). There are three types of covalent DNA modifications commonly found in bacteria: C5-methyl-cytosine (5mC), C6-methyl-adenine (6mA) and N4-methyl-cytosine (4mC) (Sánchez-Romero, Cota, and Casadesús 2015; Blow et al. 2016; Oliveira 2021; John Beaulaurier, Schadt, and Fang 2018). Methylation at these sites occurs via the action of DNA methyltransferases (Heard and Martienssen 2014; Jablonka and Raz 2009; Casadesús and Low 2006), which are ubiquitous across bacteria (Oliveira and Fang 2020).

Despite the ubiquity of DNA methylation, how often it serves an epigenetic function in bacteria is not well-established. In many cases, DNA methylation does not lead to different

heritable phenotypes, and thus does not function as an epigenetic mark (Waldminghaus and Skarstad 2009; Skarstad, Boye, and Steen 1986; Collier 2009). However, a number of studies have established that DNA methylation can act to regulate cellular processes, including gene expression (D. Roberts et al. 1985; Seong, Han, and Sul 2021), sometimes in a heritable manner (D. A. Low, Weyand, and Mahan 2001; van der Woude, Hale, and Low 1998; Casadesús and Low 2006; Sánchez-Romero and Casadesús 2020). These modifications can have significant downstream phenotypic effects (Sánchez-Romero and Casadesús 2020; Park et al. 2019). Notably, in almost all well-established cases, when DNA methylation functions in an epigenetic manner, it is highly localised (e.g. at the operon-level) (Hale, van der Woude, and Low 1994), or even for a single site (Birkholz et al. 2022). One exception to this is a recent study, which suggested that genome-wide DNA methylation patterns differ between free-living and terminally differentiated bacteroids of the soil bacterium *Rhizobium leguminosarum* (Afonin et al. 2021).

To further probe possible epigenetic functions of DNA methylation in bacteria, here we characterise methylation patterns for three natural isolates of *E. coli* across a wide range of growth conditions. We profile DNA methylation using Oxford Nanopore (ONT) sequencing (Simpson et al. 2017; Rand et al. 2017), and show that by comparing samples of native methylated genomic DNA to whole genome amplified DNA it is possible to identify the expected methyltransferase binding motifs. We then use a quantitative approach to show that across the genome, methylation levels vary in a predictable fashion, and that levels of methylation differ between growth conditions. These data suggest that *E. coli* cells undergo environment-dependent transient differentiation into different methylation states during growth. These changes are not a reflection of cell cycle states, but instead are heritable changes that are gradually lost after growth ends. These results raise the possibility that in bacteria, growth states can be inferred solely by quantifying DNA methylation patterns, and that these patterns correspond to transiently differentiated epigenetic cell states.

4.4 Materials and Methods

4.4.1 Bacterial growth

We grew overnight cultures from single colonies for each natural isolate in 3mL of liquid LB media at 37°C. We then inoculated 75mL of the relevant growth media (either LB or M9 minimal media with 0.2% glucose) in a 250ml Erlenmeyer Flask with 75uL of overnight culture. We grew these at the relevant temperature (37°C, 25°C, 42°C) until an OD600 between 0.4 and 0.5 was reached, or for 24 hours or 96 hours (for WGA and late stationary phase samples). 5ml of media was removed into a 15ml falcon tube and the cells were pelleted by centrifugation at 14,000 RPM for four minutes. We removed the media and spun

the cells for an additional two minutes, after which we pipetted off any remaining media. We stored the cell pellets at -20°C until DNA extraction.

4.4.2 DNA extraction and whole genome amplification

We extracted DNA using the Promega Wizard DNA extraction kit following the gram negative bacterial extraction protocol. We performed whole genome amplification (WGA) using the Qiagen RepliG kit according to the manufacturer's protocol. We used a Qubit fluorometer to measure DNA concentration, ensuring that each sample had sufficient DNA for a ligation library prep without further concentrating the sample. We measured DNA purity with a Nanodrop. For all samples, the 260/230 and 280/230 ratios were between 1.5 and 2.3. We stored DNA at -20°C until library prep and sequencing.

4.4.3 Library preparation and DNA sequencing

We prepared ONT sequencing libraries for both the WGA and native DNA using either the SQK-LSK109 kit with barcode expansion kit EXP-NBD104 or the SQK-RBK004 kit. For the SQK-LSK109 kit we followed the manufacturer's protocol with no modifications. We modified the SQK-RBK004 protocol as follows: we eluted the samples off Agencourt AMPure XP beads using TE buffer pre-warmed to 50°C; we performed the elution itself at 50°C; and we increased the incubation time for elution to 10 minutes.

We performed ONT sequencing on a MinION Mk1B device using R9.4.1 flowcells. We used eight flowcells in total (two with SQK-RBK004 libraries and six with SQK-LSK109 libraries), with 12 samples run per flow cell. One additional flow cell was used to produce an additional 1 Gbp for a single sample that had low coverage. For each sequencing run, we demultiplexed and basecalled using Guppy v4.2.2.

For quantitative analysis of methylation, we subsampled all WGA and native sequencing reads to ensure even coverage across the genome using the following strategy: for each sample, we mapped all reads onto the relevant reference genome and determined the lowest 5th percentile of coverage over all samples, excluding the 96 hour sample, which had lower coverage for all strains (see below). For the 96 hour samples, we calculated the 5th percentile of coverage only for those samples, rather than across all samples.

We then standardised coverage across the chromosomal contig at this 5th percentile level. We first calculated the mean read length for each dataset. We then divided the genome into 10 kbp windows, and sampled an appropriate number of reads originating within each window such that the read length and the target coverage matched (e.g. if mean read length was 2 kbp and the target coverage was 100X, then we selected 500 reads originating within

the 10 Kbp window). We then mapped all reads back onto the genome to confirm that we had reached the coverage targets. If the target coverage was not achieved (for example due to irregularities in the read length distribution), the mean read length was adjusted to represent the mapped reads, and reads were resampled. We then used the ONT-fast5-api to extract the corresponding fast5 reads for each dataset (see github).

4.4.4 Identification of methyltransferases

We previously produced reference-level genomes for each strain (Breckell and Silander 2020) and annotated these using Prokka (Seemann 2014). We identified methyltransferases by using bwa mem (H. Li 2013) to map all restriction enzymes and methyltransferase enzymes in the REBASE
Gold database (R. J. Roberts et al., n.d.) to each strain. The REBASE Gold database contains only experimentally validated methyltransferase and restriction modification systems. We filtered the alignments to include only those genes which aligned for more than 97% of their length.

4.4.5 Detection of modified sites using Nanodisco

We used Nanodisco to detect DNA methylation (Tourancheau et al. 2021) with the recommended default settings. We processed fast5 reads from both WGA and native DNA samples separately with the Nanodisco preprocess command before running the Nanodisco difference command to calculate differences in the WGA and native DNA signals at each position. We used the Nanodisco merge command to create a single output file containing the native and WGA coverage for each genomic location, the mean signal difference and U and t-test p-values reporting the significance of the signal difference at each site.

4.4.6 Quantification of methylation at individual sites

The Nanodisco output includes a p-value of a two-tailed Mann-Whitney U-test for each site indicating whether or not the signal at that site differs between the modified and unmodified samples. However, this p-value is not necessarily lowest at the actual point of modification, as the nanopore detects five bases at once, and the methylation can affect the signal in unpredictable ways. For example, many bases that were identified as having signals that differed between native and WGA DNA were not highest at the expected cytosine position within GATC motifs. To ensure we identified methylated motifs, we first identified all motif locations (DCM and DAM) in the genome (CCWGG and GATC, respectively), and then identified the lowest p-value out of the focal base and either neighbouring base. We used this p-value as an indication of whether or not a CCWGG or GATC site was methylated.

To account for false positive identification of modified sites, we used the p-values from above for the DCM and DAM sites located in the first 1 Mbp of the genome. We also identified an equal number of random locations in the first 1 Mbp of the genome, and identified the lowest p-value of each random bp or either neighbouring bp. We performed this analysis only in the first 1 Mbp of the genome to minimise computational effort; it is highly unlikely that this has any effect on the results. This resulted in a set of p-values for possibly methylated sites within each target motif, and likely unmethylated random sites. We used the p-values from the random sites to establish a null distribution of p-values for unmethylated bases. We designated all DAM and DCM sites with p-values lower than the 10th percentile of the null distribution as methylated (Fig. S1). All other DAM and DCM sites we designated as unmethylated. The precise implementation of this method is available through the github repository indicated above.

4.4.7 Correlation in methylation fractions

To calculate correlations in the fraction of methylated sites, we first determined the number of DAM or DCM binding motifs within each 10 kbp window for each genome. We used this as an estimate for the number of potential DAM or DCM modifications and then calculated the fraction of DAM or DCM sites which we experimentally identified as modified in each window. We calculated the correlation between the fraction of modified sites in each window as a Pearson correlation or a partial correlation accounting for sequencing coverage, as sequencing coverage affects the likelihood that a site will be detected as modified.

4.4.8 Genome wide methylation patterns

We assessed genome wide methylation patterns by comparing the fraction of known sites vs modified sites in windows across 10 kbp windows in the genome. We discarded any regions that contained no DAM or DCM sites, as this would result in a division-by-zero problem. For the normalised data presented in Fig. S6, we simply divided the fraction of methylated sites in each window by the mean of all windows across the genome.

4.5 Results

4.5.1 Determination of methylation motifs

We first sought to determine which methyltransferases were present in each of three natural isolates of *E. coli*, denoted here as SC419, SC452, and SC469 (Ishii et al. 2006). We found the adenine methyltransferase dam (which recognizes GATC motifs) and the cytosine

methyltransferase dcm (which recognizes CCWGG motifs) in all three strains. We also found one of the adenine methyltransferases EcoKII or EcoGVI in each of the three strains. Both of these target the same motif, ATGCAT, and are present in most *E. coli* strains (Fang et al. 2012; Adzitey et al. 2020). We identified the methyltransferase EcoGIX in strains SC419 and SC469. EcoGIX is an adenine methyltransferase, with a loosely defined motif sequence (Fang et al. 2012; Forde et al. 2015). Finally, we identified EcoGVII in strain SC469, which is a close homologue of DAM (Fang et al. 2012), and recognises the same target motif.

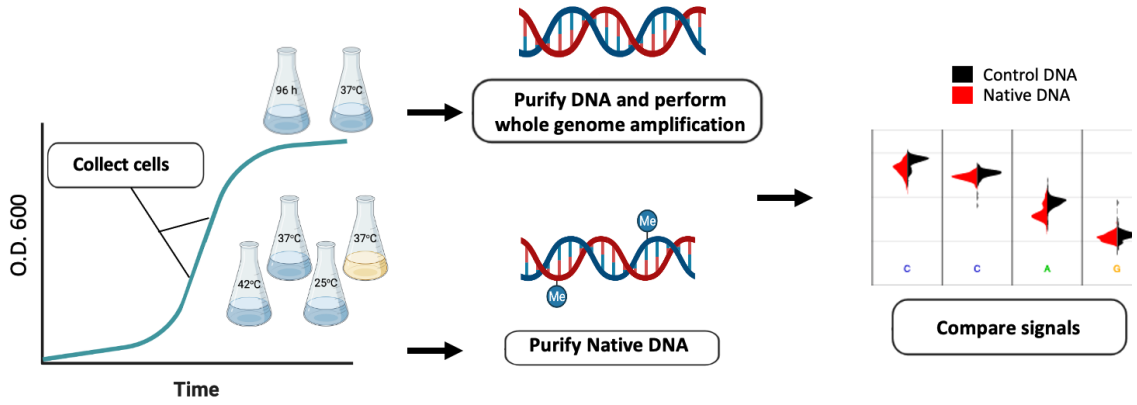


Figure 4.1. Experimental design for sampling native (possibly modified) and unmodified DNA. To sample native DNA we grew cultures until exponential phase (for the minimal M9 media, rich LB media, 42ºC and 25ºC growth conditions); or late stationary phase (for the 96 hour growth condition). For whole genome amplification, we isolated DNA from early stationary phase (24 hours of growth). After purification of genomic DNA (and whole genome amplification when necessary), we sequenced the samples using the ONT platform. To infer DNA modifications we compared the signals from native and WGA DNA using Nanodisco.

To determine whether each of these methyltransferases was active we used ONT sequencing to identify genomic sites where DNA was modified. We sequenced native DNA which may contain modified bases, and whole-genome amplified (WGA) DNA which contains few, if any, modifications. We generated at least 50-fold genomic coverage of ONT data from native DNA and at least 100-fold genomic coverage of ONT data from WGA DNA. (Methods). Note that these fold-coverage values are mean coverage values over the whole genome. To determine which genomic sites were modified we used a simple statistical approach implemented by Nanodisco (Tourancheau et al. 2021). Nanodisco uses the differences in the raw nanopore signals from each sample to assign a p-value to every position in the genome using a Mann-Whitney U-test.

We selected flanking regions from the 5,000 bases with the lowest p-values for input into MEME (Bailey et al. 2009) to identify motifs associated with modified bases. However, we found that in almost all cases, MEME identified only the cytosine methyltransferase DCM motif (CCWGG). We hypothesised that this was because methylated DCM motifs generally have smaller p-values than other motifs, due to larger signal deviations from unmethylated motifs. Because there are more than 13,000 DCM sites in each genome, the vast majority of the regions with low p-values will be DCM sites, even when considering a very large number of sites (e.g. more than 10,000). We found that using a larger number of regions for input into MEME was computationally prohibitive. We thus randomly subsampled 100,000 base pairs (and associated p-values) from the genome (representing approximately 2% of the genome). From this subsample, we selected the flanking regions for the 5,000 base pairs with the lowest p-values for input into MEME.

For all three strains, MEME identified GATC and CCWGG as significant motifs (**Table 1**). These are the canonical motifs for the DAM and DCM methyltransferases, respectively, and we had bioinformatically identified both in all three strains. As these match the DAM and DMC motifs, we assume that they contain C6-methyl-adenine (6mA) at the A position and C5-methyl-cytosine (5mC) at the second cytosine, respectively. Although we computationally identified the adenine methyltransferases EcoKII and EcoGVI in the three strains, we did not identify their target motif ATGCAT in any strains. We speculate that this is because methylated adenines are more difficult to identify (see above), and because this six-base pair motif is considerably more rare than the four-base pair motifs recognised by DAM and DCM. We also identified methyltransferase activity at two additional motifs, CCGG and GAGCC, in SC419 and SC452, respectively. Although there are no experimentally validated methyltransferases in the REBASE Gold database that are known to target these motifs, there are a number of putative type III R-M system methyltransferases that are thought to target these motifs. We mapped the sequences of each of these putative methyltransferase against each genome and identified a single genomic region in SC452 that matched all the putative GAGCC modifying methyltransferases (**Table 2**). This methyltransferase has a non-palindromic motif, and thus methylates only a single strand (Meisel et al. 1992). Surprisingly, we did not identify any CCGG-targeting methyltransferase in the SC419 genome. Finally, for the last two computationally identified methyltransferases, EcoGIX and EcoGVII, we could not confirm any activity. This is not unexpected, as the EcoGIX motif is ambiguous and the EcoGVII motif overlaps with DAM.

Table 4.1. Matches between sequence motifs identified by MEME and REBASE Gold methyltransferases. Each row indicates the top three motifs as reported by MEME.

Strain	Target motif reported by MEME	Number of motifs identified in 100 kbp	MEME p-value	Inferred REBASE Gold enzyme
SC419	CCWGG	632	3.1e-457	DCM
	GATC	625	4.9e-177	DAM
	CCGG	376	2.3e-259	unknown
SC452	CCWGG	750	2.3e-628	DCM
	GATC	681	3.3e-235	DAM
	GAGCC	111	4.1e-24	M.EcoB0880RFEP [1]
SC469	CCWGG	371	1.2e-212	DCM
	GATC	185	1.4e-30	DAM

1 This is a putative methyltransferase that is not found in the experimentally confirmed REBASE Gold database

4.5.2 Quantitative analysis of methylation levels

We next sought to determine whether there was variation in the levels of methylation across the genome, or whether all regions were equally methylated. We focused only on the most commonly methylated motifs in each genome, GATC (containing methylated adenines via DAM) and CCWGG (containing methylated cytosines via DCM). Critically, the likelihood that a site is identified as methylated depends on the coverage of that site (**Fig. S1**). Thus, to increase the likelihood that all sites across the genome had an equal probability of being identified as methylated, we subsampled each of the ONT sequencing datasets to standardise coverage across the genome (Methods).

We then used Nanodisco to compare the native and WGA datasets for all three genomes, and for each known DAM and DCM motif site took the lowest p-value from within the 3bp surrounding each motif (see Methods, Quantification of methylation at individual sites). These p-values should be indicative of the methylation status of a site, as they result from a Mann-Whitney U-test comparing the signal levels of modified and unmodified DNA. In addition, we hypothesised that sites at which all DNA molecules have a methylated nucleotide would have smaller p-values compared to sites at which only a small number of molecules are methylated, and that p-values are thus a quantitative indication of methylation status.

To directly test this hypothesis, we subsampled reads from the WGA data (which arises from fully unmethylated reads) to reach 50x coverage across the genome. We compared this

WGA data with mixed native and WGA datasets having 50x coverage but consisting of 0%, 25%, 50%, 75% or 100% native reads. We expected that many of the native reads were fully methylated at DCM and DAM motifs. We then used Nanodisco to infer methylation status for all positions in the genome in these datasets with different ratios of WGA and native reads. We found a clear negative relationship between the fraction of native reads in the dataset and the associated p-values for each position (**Fig. 2**): as the fraction of native (possibly methylated) reads in the dataset increased, the p-values decreased. This indicates that the p-values returned by Nanodisco are correlated with the fraction of methylated molecules at a site, and may provide quantitative insight into the fraction of molecules that are methylated at any DAM or DCM position in the genome.

We then implemented a simple binary classification of DAM and DCM sites as being methylated or unmethylated (or less methylated) using a p-value cutoff (**Fig. S2 and Fig. S3**). We placed this cutoff such that 10% of non-methylated sites were inferred as being methylated, analogous to implementing a false discovery rate of 0.1 (Methods; **Fig. S4 and Fig. S5**). Although it would also be possible to implement a generative model specifying the fraction of molecules that are methylated at any one location in the genome, without a ground truth set of data for both unmethylated and methylated molecules, this is difficult. Thus, we use this simple binary classification. Importantly, this division into methylated and unmethylated status for each site does not indicate definitively that a site is methylated or unmethylated. Rather, the division establishes that specific sites are more or less methylated (**Fig. 2**). We next used this classification of sites as methylated or unmethylated to test whether there were consistent differences in methylation rates across the genome or across growth conditions.

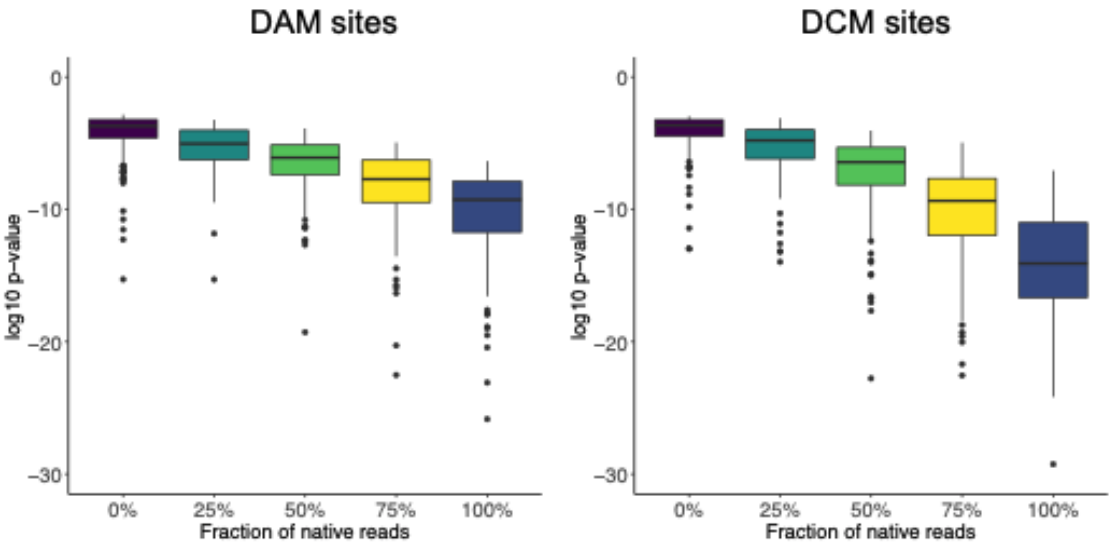


Figure 4.2. The p-values resulting from Mann-Whitney U-tests for signal deviations at DAM and DCM sites are correlated with the fraction of methylated molecules. We mixed known fractions of WGA reads (unmethylated) and native reads (possibly methylated) in silico and used Nanodisco to determine the p-value of a Mann Whitney U test at each position in the genome. We then determined the lowest p-value in a three bp window surrounding each hypothetically modified base in DAM (GATC) or DCM (GGCC) motif. For both methyltransferases, the sensitivity of the test increases as the fraction of native reads increases, with the DCM p-values decreasing to a much larger extent.

4.5.3 Identification of local and global methylation patterns

To test for differences in methylation across growth conditions, for each strain we isolated DNA from cultures grown to exponential phase in five different conditions: two replicate cultures grown at 37ºC in minimal media (M9 glucose), one grown at 37ºC in LB broth (rich media), one grown at 25ºC in minimal media (low temperature stress), one grown at 42ºC in minimal media (heat stress), and one after 96 hours of growth in minimal media (late stationary phase). For each of these growth conditions, we performed the same analyses outlined above to determine whether DAM and DCM sites were classified as methylated or unmethylated.

We then used this data to look at large scale variation in methylation marks across the genome, on the basis of both strain and growth environment. Rather than consider single sites, which exhibit considerable noise in being classified as methylated or unmethylated, we calculated the fraction of methylated sites in 10 kbp windows across the genome (approximately 500 windows total for a 5 Mbp genome; see Methods). Each of these windows contained approximately 40 DAM or DCM sites. We found that the fraction of sites

classified as methylated within each 10 kbp window varied by methyltransferase, strain, and environment (**Fig. 3**).

Overall, we inferred that a much higher fraction of DCM sites were methylated compared to DAM sites (**Fig 3.**). Part of this difference is likely due to the fact that the signal differences between methylated and unmethylated cytosines at DCM sites are much larger than between methylated and unmethylated adenines at DAM sites (**Fig. 2**). Thus, it does not reflect biological differences but differences in the sensitivity of each statistical test. Nonetheless, we observed that in some growth conditions, a strain could exhibit similar levels of methylation at DCM and DAM sites (e.g. SC452 at 42ºC) whereas another strain in the same condition could exhibit different levels of methylation (e.g. SC469 at 42ºC). This indicates that it is unlikely that the lower levels of DAM methylation are due solely to decreased sensitivity, but instead to differences in the activity of each methyltransferase. We also observed general strain-specific differences in methylation, for example, generally lower levels of both DCM and DAM methylation for SC469. However, it is difficult to determine whether this reflects real differences in methyltransferase activity between strains, or whether it is an artefact of the data analysis: for all cases, we inferred methylation status from a single unmethylated WGA dataset for each strain, and this in itself may cause differences in inferred methylation levels.

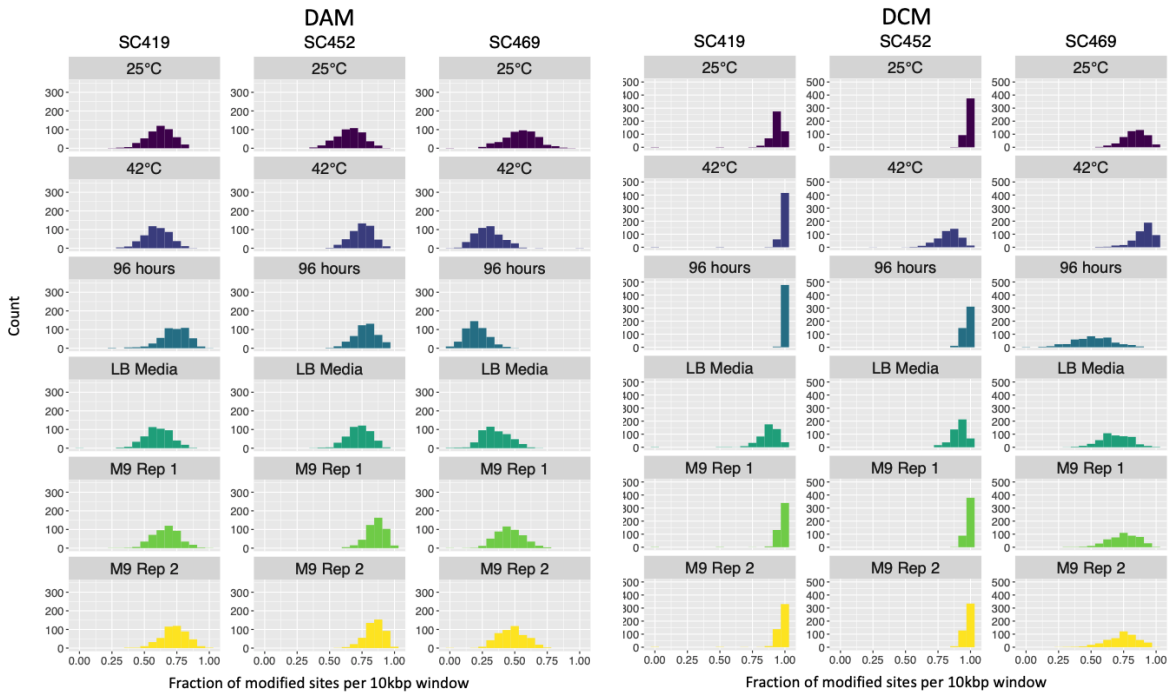


Figure 4.3. The fraction of DAM 6mA and DCM 5mC methylated sites within 10 kbp windows varies according to strain and growth condition. The histograms in each panel indicate the distribution of 10 kbp windows in which a certain fraction of sites are DAM (left panel) or DCM (right panel) methylated. This fraction ranges from almost 100% of all sites in all windows (e.g. for SC419 DCM in the 42ºC growth condition) to less than 50% of all sites in the majority of windows (e.g. for SC469 DAM in the 42ºC growth condition). With the exception of the LB rich media sample, all cultures were grown in M9 minimal glucose media.

We next considered whether there were more localised patterns of methylation across the genome. To do this, we tested for correlations in the fraction of methylated sites within the 10 kbp windows between growth conditions. Across different sets of growth conditions, we found that some 10 kbp windows consistently had the majority of sites methylated, while other windows had many fewer sites methylated (**Fig. 4A**). It is likely that some of this is due to differences in coverage, as the relationship between inferred methylation status and coverage was not totally mitigated by our subsampling scheme (Methods). To minimise this dependence, we calculated the partial correlations in methylated fractions for each 10 kbp window accounting for genome coverage (see Methods).

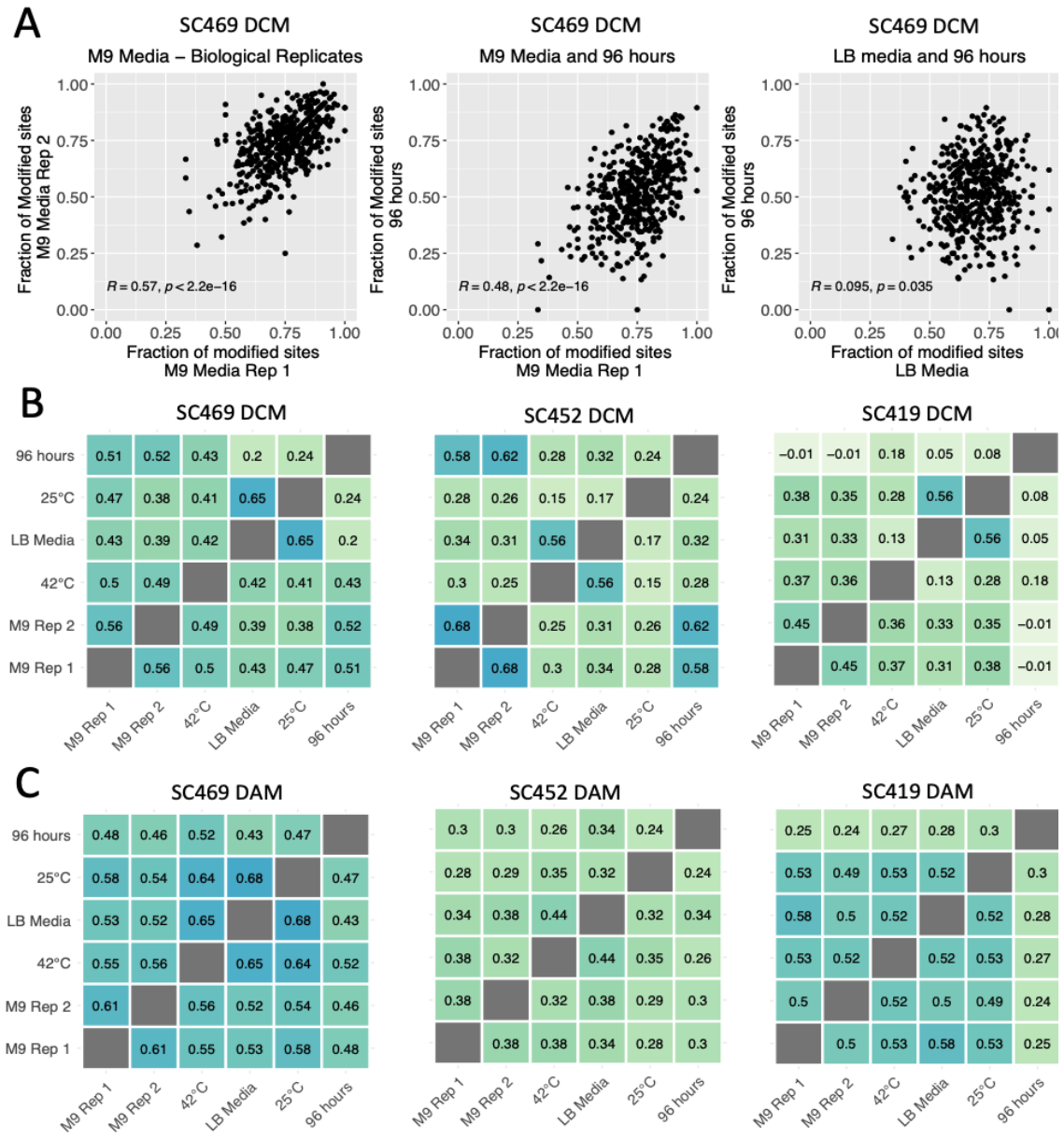


Figure 4.4. (A) The fraction of methylated sites in 10kbp windows across the genome is correlated across growth conditions. The three panels indicate the fraction of methylated DCM sites within a 10 kbp window that we inferred as methylated for strain SC469. We observed strong positive correlations in methylation patterns in replicate cultures of minimal M9 glucose media, slightly weaker correlations between M9 media and 96 hour stationary phase cultures, and almost no correlation between patterns in rich LB media and 96 hours stationary phase. Pearson partial correlations and corresponding p-values are indicated in each plot. **(B) Pairwise partial correlations in DAM and (C) DCM methylation patterns between all growth environments accounting for genome coverage.** Each panel shows all pairwise Pearson partial correlations between growth conditions in the fraction of methylated sites for all 10 kbp windows in the genome, controlling for genome and WGA coverage in each of the growth conditions.

We calculated pairwise correlations in the fraction of methylated sites in 10 kbp windows across the genome for both DAM and DCM in each strain across all pairs of growth conditions. We found replicable differences across the genome in methylation fractions (**Fig. 4**), with the correlations between some conditions being higher than others. Critically, we found that in all cases except one the replicate cultures grown in M9 minimal glucose media at 37ºC exhibited the strongest correlation with the other M9 replicate. For example, for strain SC469 DCM the partial correlation between M9 replicates 1 and 2 was 0.56. The second strongest correlations for each were with cultures at 96 hours extended stationary phase (0.51 and 0.52 for replicates 1 and 2, respectively). Similarly, for SC469 DAM, the correlation between M9 replicates was 0.61. The second strongest correlations for each replicate were with growth at 25ºC (replicate 1, 0.58) and growth at 42ºC (replicate 2, 0.56).

This pattern, in which each M9 minimal media replicate correlated most strongly with the other replicate, extended to almost all strains and methyltransferases, with the single exception of DAM in strain SC419, for which methylation patterns correlated very similarly for all pairs of conditions (**Fig. 4C, rightmost panel**). As there are a total of six independent growth conditions, there is only a one in five chance that the two M9 replicates are most highly correlated. Thus, the likelihood that they would be the most highly correlated in almost all strains for both DCM and DAM strongly suggests there are growth-condition methylation states. Furthermore, these differences exist even when growth conditions differ only subtly (e.g. growth in minimal M9 glucose media at 37ºC versus M9 at 42ºC or growth in minimal media at 37ºC versus rich media at 37ºC).

In addition to high correlations between identical growth conditions, we often found consistent correlations in methylation status between different growth conditions. For example, the methylation patterns in the rich media LB condition (grown at 37ºC) often exhibited very strong correlations with methylation patterns in the minimal media 25ºC growth condition. In three cases (SC469 DCM, SC469 DAM, and SC419 DCM), these two conditions exhibited the strongest correlation of any pair of conditions. The convergent methylation states in these two conditions may be driven by similar changes in transcriptional activity, which could have an inhibitory effect on methylation.
The lowest levels of correlation we observed were for 96 hours extended stationary phase for strain SC419 DCM (**Fig. 4B, rightmost panel**). In some cases, the partial correlations were slightly negative. However, many of the 10 kbp windows in this condition had almost 100% of all DAM sites methylated (**Fig 3, right panel**). Such low variability in methylation status means that strong correlations are difficult to obtain.

One explanation for the correlations in methylation fractions across growth conditions is that

there are consistent long-range intragenomic correlations driven by periodicity in methylation, e.g. methylation fractions are generally lower at the origin of replication and higher at the terminus, or that there is transient methylation behind the replication fork (Anton and Roberts 2021). This would be apparent as long range correlations in the fraction of methylated sites across the genome. For example, any two windows separated by a distance that is less than the periodicity should exhibit positive correlations. However, plotting the fraction of methylated sites across the genome revealed no strong long-range patterns (**Fig. 5**). To more systematically test for long-range patterns, we calculated correlations in the fraction of methylated sites within windows of increasing size, from 250 bp to 500 kbp, separated by distances of increasing size, from 0 bp to 1 Mbp. This is similar to calculating an autocorrelation function, but for almost all step sizes (Methods). Again we found no strong patterns of correlation between any windows larger than 5 kbp, nor windows separated by a distance of more than 5 kbp (**Fig. S7 and. Fig. S8**).This suggests that short-range correlations dominate, and there are few long-range correlations in the fraction of methylated sites that are driven by factors such as higher levels of methylation at the terminus.

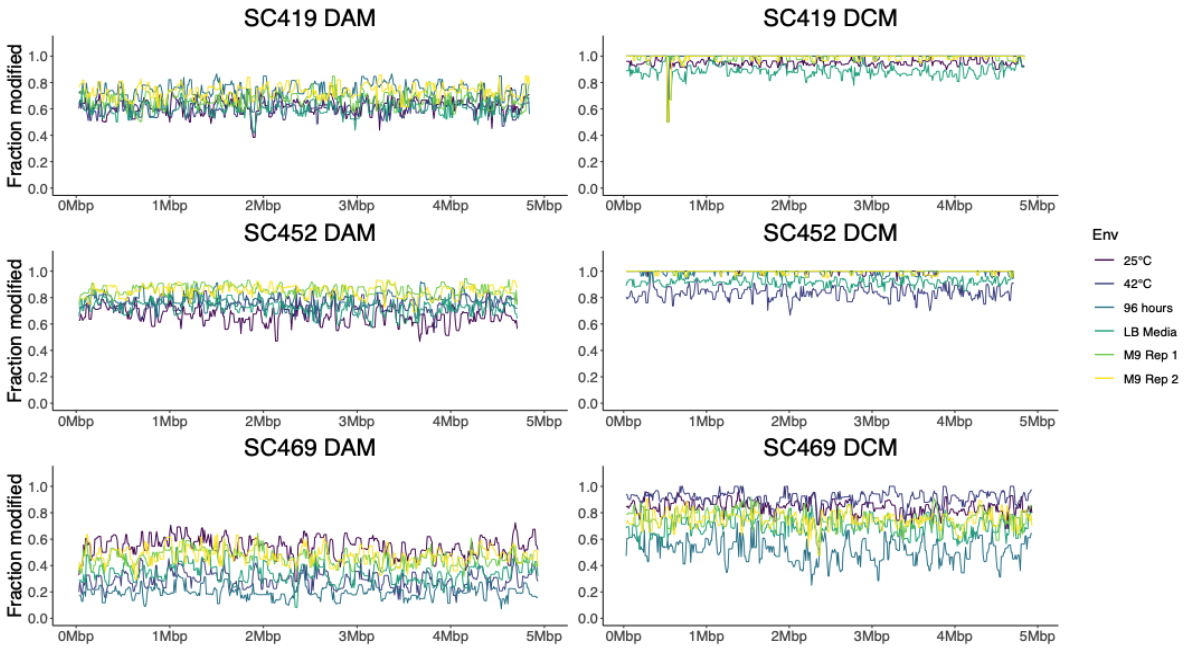


Figure 4.5. Genome wide patterns in the fraction of methylated sites. Each panel shows the fraction of methylated sites in 10 kbp windows across the entire genome, with different growth conditions indicated in different colours. No long range correlations, such as higher methylation at the replication terminus, were apparent.

4.6 Discussion

Here we have identified DNA modifications in three *E. coli* natural isolates across a range of growth conditions using ONT sequencing. We have shown that it is possible to determine the motifs at which DNA modifications occur, and that these match the motifs expected given the restriction modification systems present in each genome. However, we also found one motif (CCGG) for which we could not identify a matching RM system; this motif may be modified by a novel methyltransferase.

Furthermore, we have shown that by using a simple binary classification of sites as methylated or unmethylated, it is possible to discern replicable and consistent differences in localised methylation frequency across the genome. The methylation patterns we have observed are dependent on growth conditions, with specific localised regions (on the order of thousands of kilobases) in the genome tending to be fully methylated, while others are less methylated. These conclusions differ from some previous work. A study on diverse strains of *M. tuberculosis* showed that most differences in methylation across the genome (as determined via SMRT sequencing) are due to stochasticity in intracellular methylation, rather than consistent differences between cells in methylation rates. Consistent differences between loci in methylation (hypomethylation) were found to be exceedingly rare, on the order of 10 to 20 sites across the genome (Modlin et al. 2020). Other work has also shown that methylation remains remarkably consistent across different growth conditions, including antibiotic stress (Cohen et al. 2016) and over the growth cycle (Payelleville et al. 2018). A significant difference between these latter two studies and the data we present here is the inclusion of methylation at DCM sites (CCWGG) in addition to DAM sites (DAM). Indeed, the most notable methylation patterns that we find – although subtle – are due to differences at DCM sites (**Fig. 4B**). Differential methylation at DCM sites has been connected to major changes in ribosomal gene regulation (Militello et al. 2012).

Critical to our proposal that these methylation patterns have epigenetic effects is that DNA methylation is heritable. Sites at which both the top and bottom strand are methylated will impart hemimethylated strands to both daughter cells, which will become fully methylated by “maintenance” methyltransferases (Anton and Roberts 2021); sites that are hemimethylated will impart one hemimethylated strand to one daughter cell and one unmethylated strand,

which is more likely to remain unmethylated. This means that mother cells with methylation at a certain genomic location will have daughter cells that are also methylated at that location, but this will vary across daughter cells. Thus, if methylation affects phenotype, and methylation varies between individual cells in a population, then it acts as an epigenetic mark for the instances we have described here.

It is possible that there are unrecognised causes that drive some of the inferred differences in methylation status across the genome. For example, subtle differences in nucleotide context affects both the activity of the methyltransferase and the deviations in ONT signal. This undoubtedly influences our ability to accurately infer methylation status. However, we do not expect these differences to be dependent on growth conditions. Thus, the fact that we find both higher correlations between identical growth conditions, and consistently higher correlations between specific pairs of growth conditions (e.g. rich media (LB) at 37ºC and M9 minimal glucose media at 25ºC), suggest that nucleotide context is not the only force driving this correlation in methylation states. Additional work is required to test the repeatability of methylation patterns in different conditions, and whether other divergent growth conditions, for example antibiotic stresses or additional heat stress, lead to greater differences in methylation patterns. Similarly, methylation patterns should converge as growth conditions converge - for example we would expect more similar patterns comparing methylation during growth at 37ºC and 39ºC than to 42ºC. Again, more experimentation is needed here.

In eukaryotes, it is well-established that methylation affects gene expression (Song et al. 2005; Vanderkraats et al. 2013), and thus cell phenotypes. Here we have shown that methylation patterns are consistent and replicable in different growth conditions in *E. coli*. In addition, for identical growth conditions (in the data here, M9 minimal glucose media), there are strong correlations in which specific regions of the genome are methylated. There are two readily apparent explanations for these results. Either growth phenotypes affect patterns of methylation, or methylation patterns affect growth phenotypes (or both). We propose that it is likely that (as with eukaryotic cells) methylation affects gene expression in *E. coli* in different growth environments, although we have not established causation (L. Chen et al. 2018). This connection between methylation and transcriptional regulation has been proposed previously (J. Beaulaurier et al. 2015), and there are data that both support (Gaultney et al. 2020) and refute the connection (Mehershahi and Chen 2021). However, we note that there are a large number of other well-established instances in which this causal direction has been established (Sánchez-Romero and Casadesús 2020).

Regardless of whether methylation functions as an epigenetic mark, and regardless of its

causality, we have shown that just as bacterial cells undergo transient differentiation into different growth phenotypes, they also undergo transient differentiation into distinct methylation states. As we have not used synchronised cultures, it is unlikely that the correlated methylation is due to synchrony in the cell cycle that differs between growth conditions. This is further supported by the fact that we have shown that correlations do not arise because of short or long range correlation in methylation fractions (e.g. differences in methylation at the chromosomal replication ori or terminus). Rather, these correlations arise from localised differences across the chromosome.

This work raises the possibility of discerning bacterial growth states without measuring cell physiology or quantifying the transcriptome, similar to what can be done for differentiated eukaryotic cells. We propose that with sufficiently long reads and precise measurements, it will be possible to quantify methylation states across single molecules, and from there infer the growth state of a cell from which a particular DNA molecule has originated. In addition, with more nuanced model-based or machine learning analyses, it may be possible to more specifically assign genomic methylation patterns to specific growth states. This contrasts with more standard approaches such as single-cell transcriptome profiling, which is often of limited use in bacteria given the extremely small number of transcripts contained in most cells.

4.7 Supplementary data

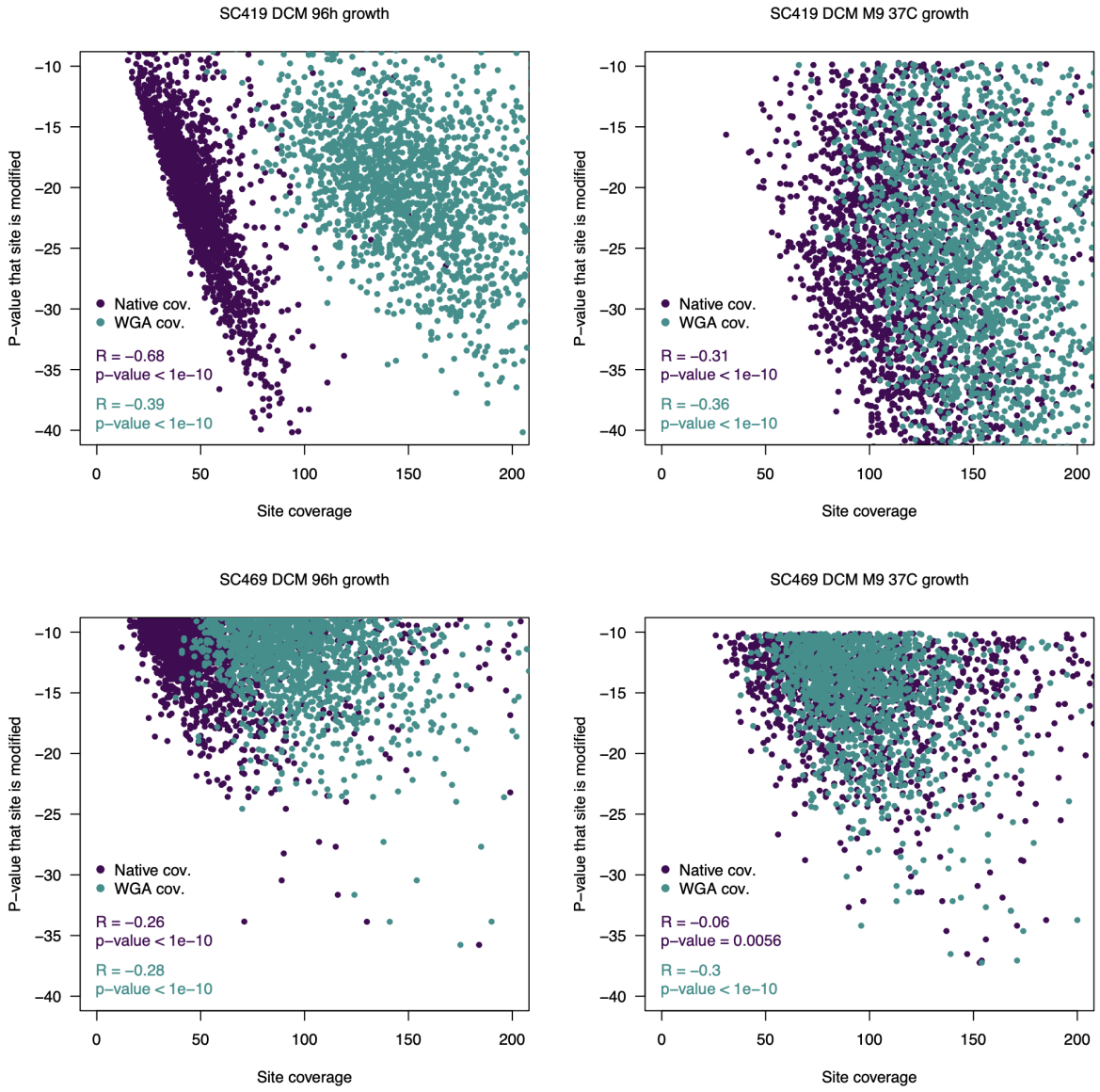


Figure S.4.1. Correlation between coverage and the Nanodisco-derived p-values. Each point indicates the coverage at individual DAM or DCM sites and the p-value of the Nanodisco Mann-Whitney U-test.. There is a clear relationship between the likelihood the p-value returned by Nanodisco (indicating a site is likely modified) and the coverage at that site, with both the coverage of the native DNA sample and the WGA sample affecting the test implemented by Nanodisco. The four examples above are all for DCM sites in two strains and two growth conditions for each. In all plots, only the sites that have p-values significantly lower than the null model background are shown. The native coverage at these sites is shown in purple; the WGA coverage at these same sites is in blue. For both native and WGA coverage, there is a strong negative correlation - sites with higher coverage have a lower p-value and a higher probability of being identified as methylated, although this differs between datasets. For example, there is only a weak relationship (R = -0.06) between native coverage and the p-value to the test in the SC469 DCM dataset.

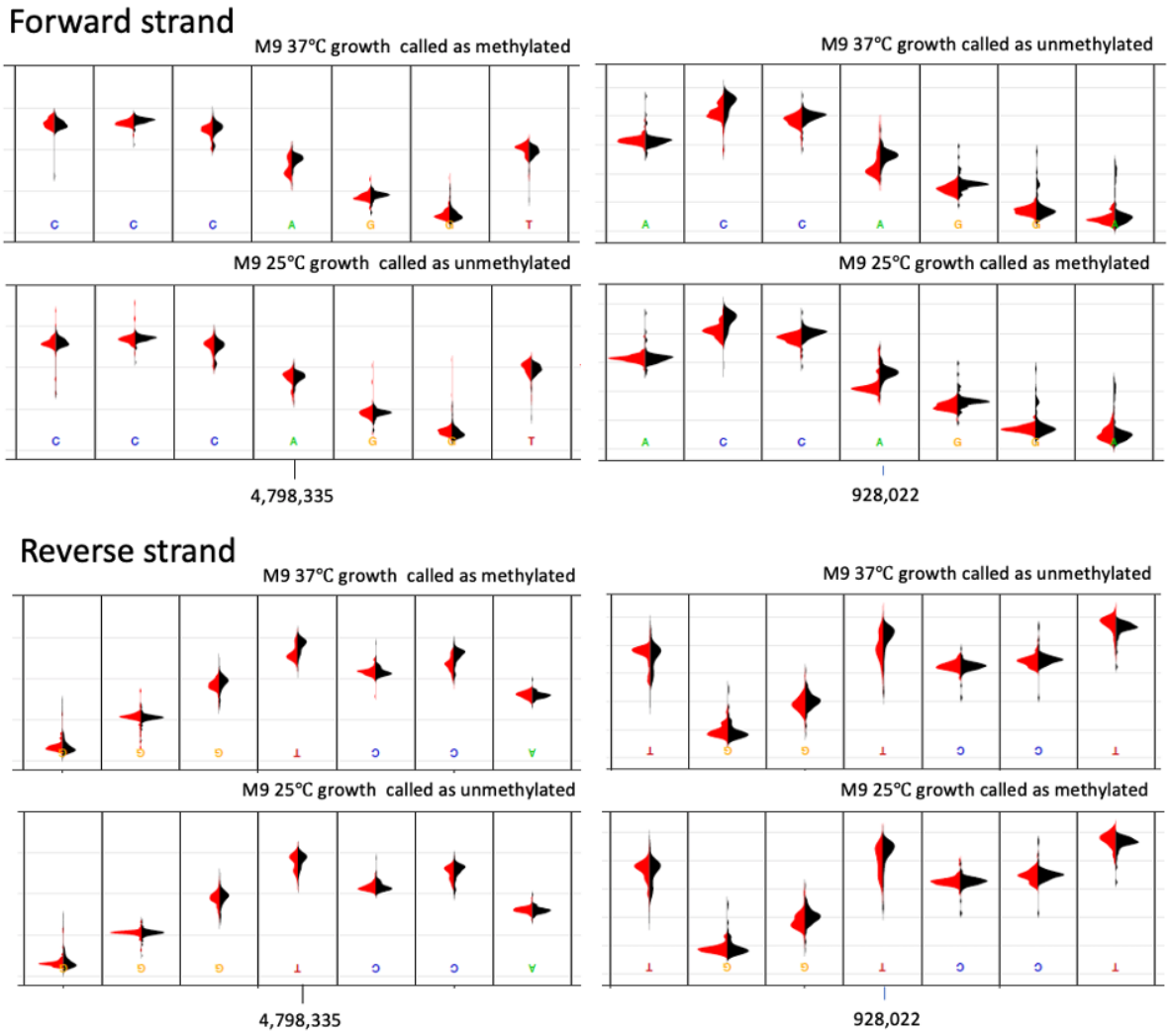


Figure S.4.2. Raw nanopore signal distributions on the forward and reverse strands at identical genomic locations of DCM sites that we inferred as methylated (top panels in each pair) or unmethylated (bottom panels in each pair). The change in the DCM CCwGG methylation status is apparent as a shift in the distribution of the red curves at the A / T position outlined with the blue box. In black are reads from the control (unmethylated whole genome amplified DNA); in red are the native DNA signals. In many cases, the shift in signal is subtle. However, the identification of these sites as methylated or unmethylated is a binary classification of a continuous state - sites that we identify as unmethylated may in fact be methylated in 40% of all cells; sites we identify as methylated may be methylated in only 60% of all cells.

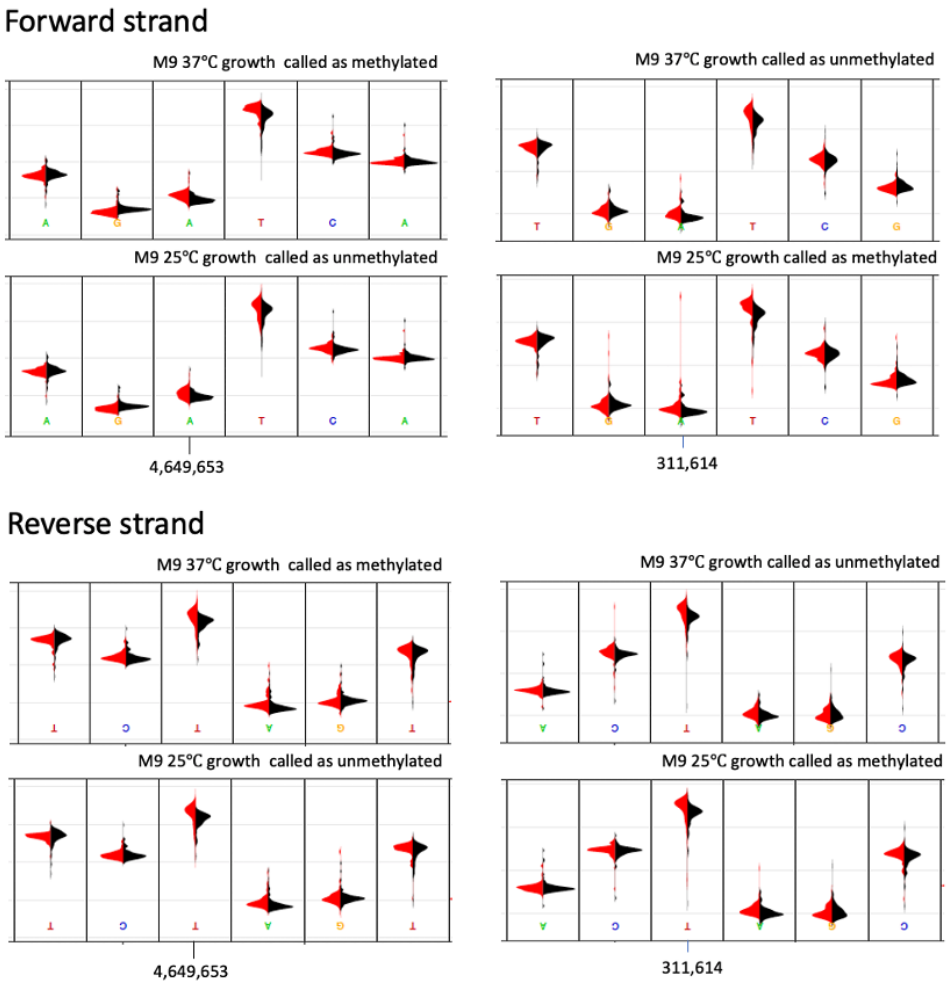


Figure S.4.3. Identical DAM sites are inferred as methylated or unmethylated across different growth conditions. The change in the DAM GATC methylation status is apparent as a shift in the distribution of the raw nanopore signal from native DNA (red curves) at the T and A positions (the A is the modified base) compared to WGA unmodified DNA (black curves). Left panels: a DAM 6mA site that we inferred as methylated in M9 37°C growth (top) but not during 25°C growth (bottom). This is most apparent as a shift in the signal at the T position, for which the overlap between red and black is less in the top panel. Right panels: a DAM GATC site that we inferred as unmethylated in M9 37°C growth (top) but methylated during 25°C growth. Again, this is most apparent as a shift in the signal at the T position, with the overlap being higher in the top panel. Note that all native DNA molecules are not necessarily methylated at positions that we call as methylated, and vice versa: at positions that we call as unmethylated, all molecules are not necessarily unmethylated.

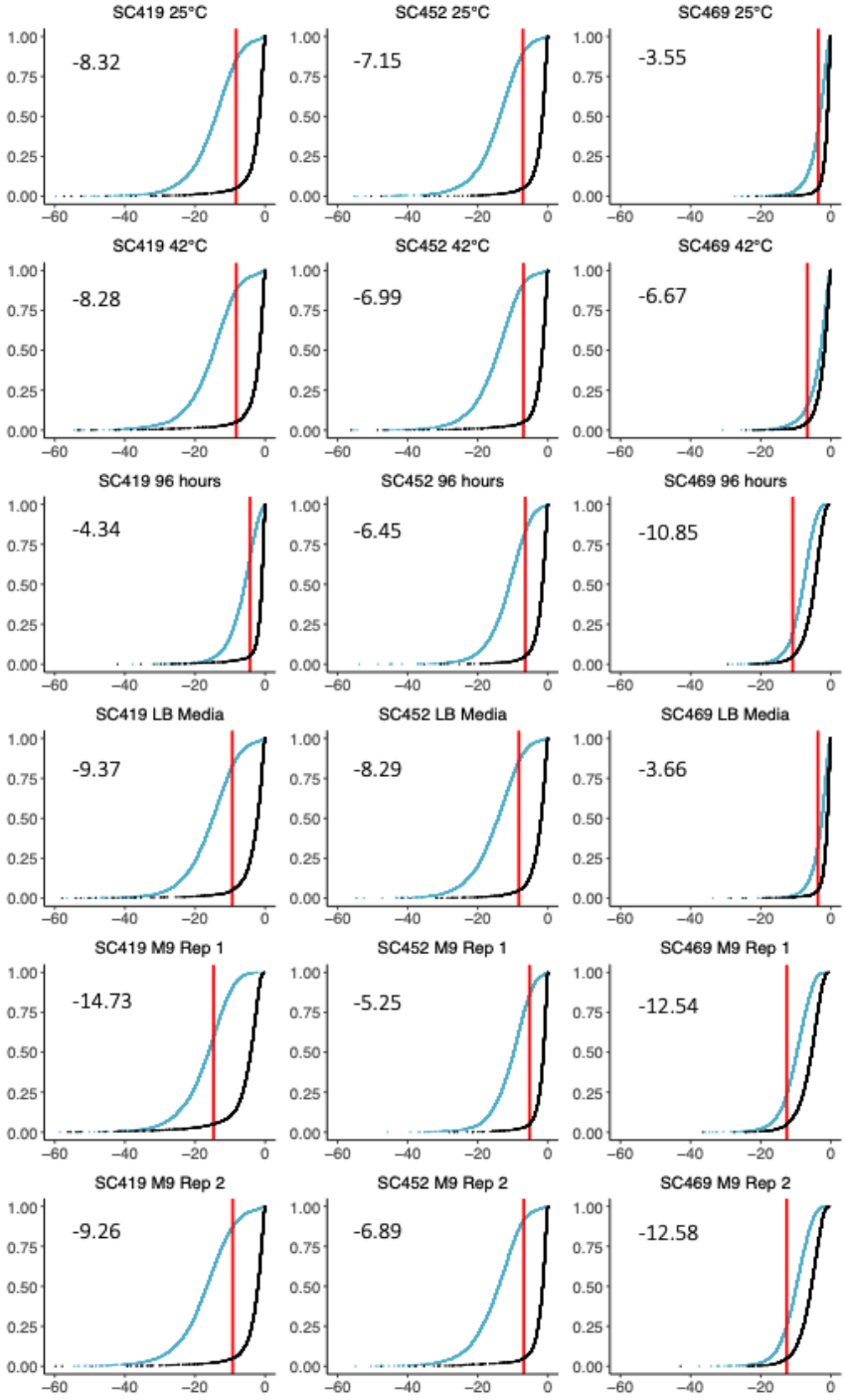


Figure S.4.4. Cumulative distributions of p-values for DAM sites relative to random (unmethylated) sites. For each combination of isolate and growth condition we used the distribution of p-values at DAM binding sites (blue) and an equal number of random sites (black) to determine a p-value cutoff. This cutoff was established such that 10% of all unmodified sites were inferred as being modified, equivalent to a 0.1 FDR. Each cutoff is shown in red, and the log10 of the p-value cutoff is noted within each plot.

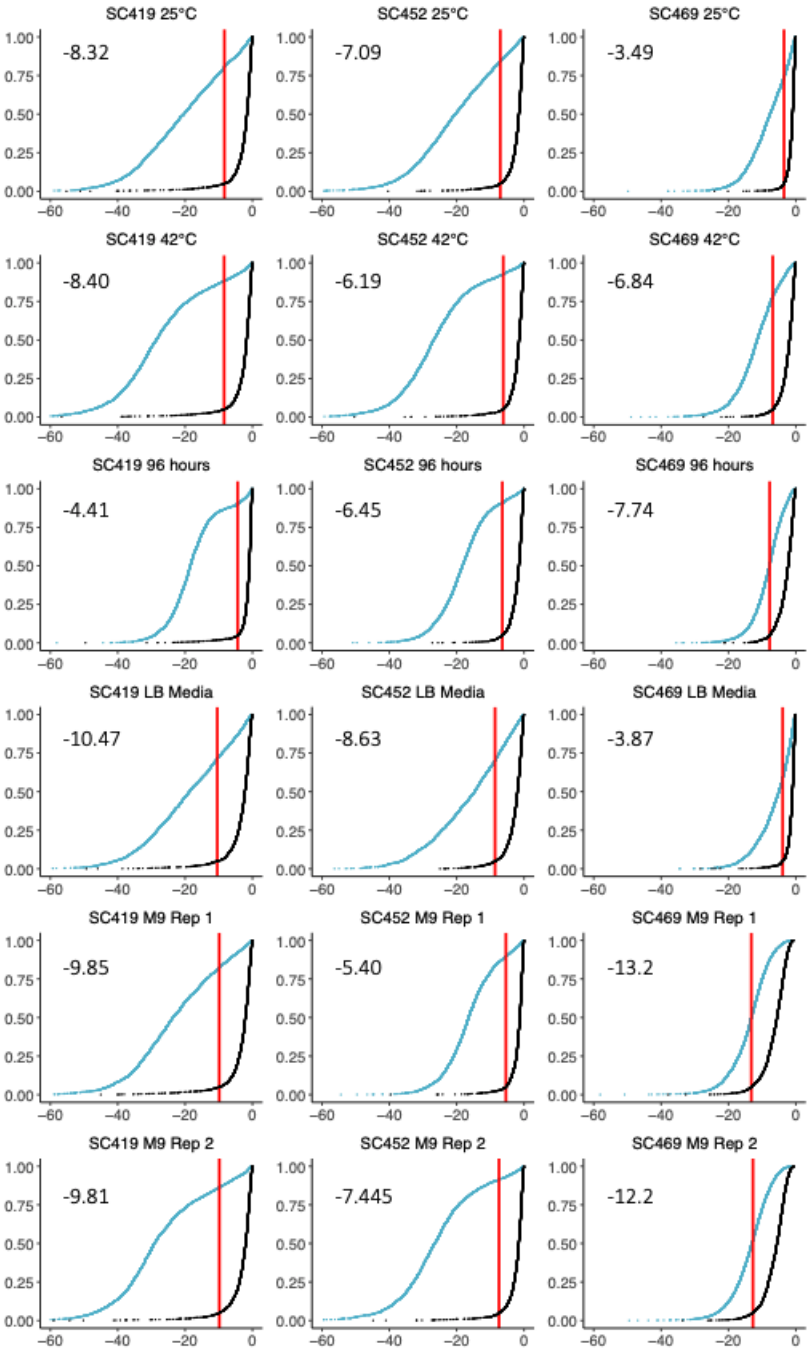


Figure S.4.5. Cumulative distribution of p-values for DCM sites relative to random sites. For each combination of isolate and growth condition we used the cumulative distribution of p-values at DCM binding sites (blue) and an equal number of random sites (black) to determine a p-value cutoff equivalent to an FDR of 0.1. Each cutoff is shown in red, and the log 10 of the p-value cutoff is noted within each plot.

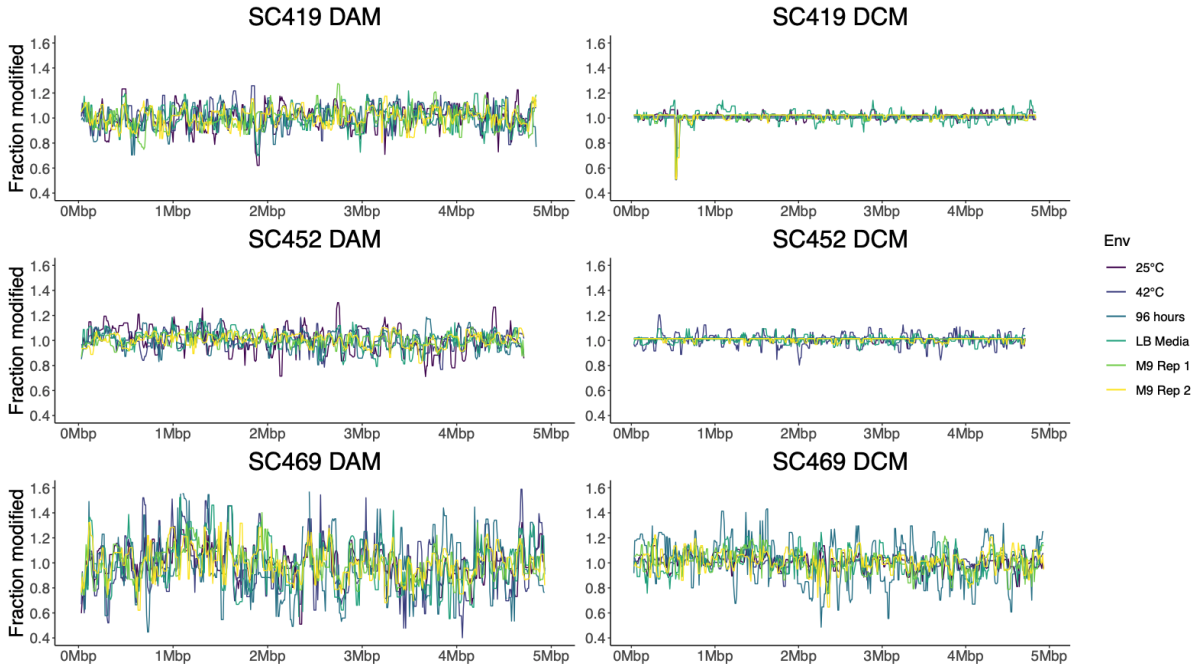


Figure S.4.6. Mean-normalised fractions of modified sites across the genome.
For each growth condition, we divided the fraction of modified sites in each window by the mean fraction of modified sites across all windows for that growth condition. This normalised fraction of modified sites are generally consistent across the genome for each methyltransferase and strain, which is clearly apparent in **Fig. 4**.

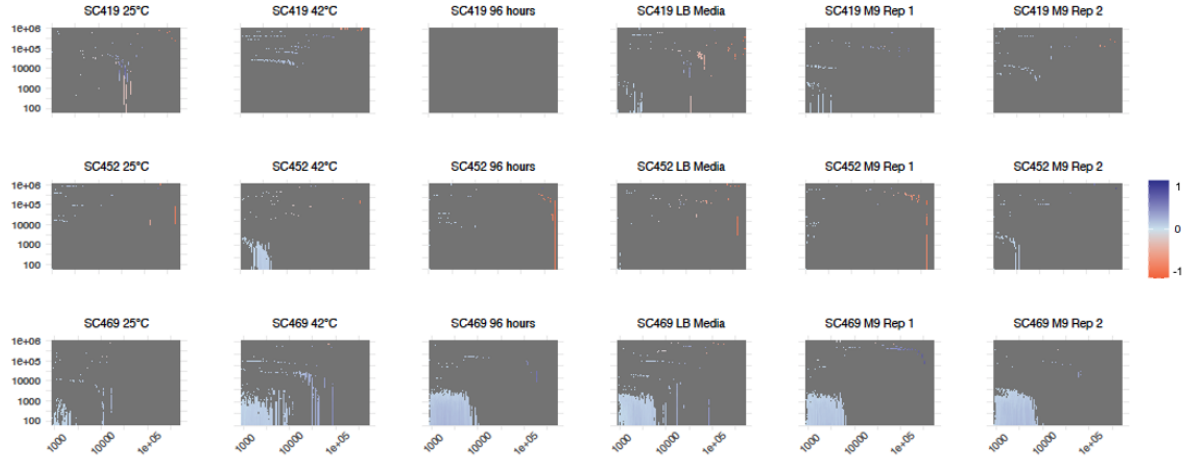


Figure S.4.7. Global autocorrelation plots for DCM methylation. Each panel is a heatmap showing the correlation for the fraction of methylated DCM sites between windows of increasing size, ranging from 250 bp to 500 kbp (different window sizes are plotted in columns), separated by increasing distances ranging from 0 (i.e. adjacent windows) to 1 Mbp (different distances are plotted in rows). Window sizes increase by a constant fraction of 4.7%; separating distances increase by a constant fraction of 9.6%. For example, the bottom left square in each heatmap shows the correlation in the fraction of methylated sites for neighbouring 250 bp windows; the middle square in each plot shows the correlation between 20.9 kbp windows separated by 7.6 kbp; the top right indicates 500 kbp windows separated by 1 Mbp. In the example here, a standard autocorrelation function (ACF) would plot the correlations between windows of a certain size separated by a specific number of windows (e.g. 10 kbp windows separated by 0 bp (neighbouring), 10 kbp (one window), 20 kbp (two windows), etc. This would be similar to several squares in the 53rd column in this plot: the squares in rows 1 (0 bp distance between windows), 56 (10 kbp distance), 64 (20 kbp distance), 68 (30.2 kbp distance), 71, 74, and 76. However, this plot shows the analogous set of correlations at almost all window sizes and distances. For clarity only correlations with $p < 0.01$ are shown. In almost all cases, the correlations are positive (i.e. windows that are close tend to have similar levels of methylation), but this correlation only exists for windows up to approximately 5-8 kbp in size and separated by a maximum of 5 kbp. This suggests that there are no long range correlations in the fraction of methylated sites. Note that the strongest correlations are observed for strain SC469, which is also the strain that exhibited the greatest variance in fraction methylated across genomic windows (**Fig. 3**). For other strains, the low level of variance in methylated fractions necessarily weakens the correlations.

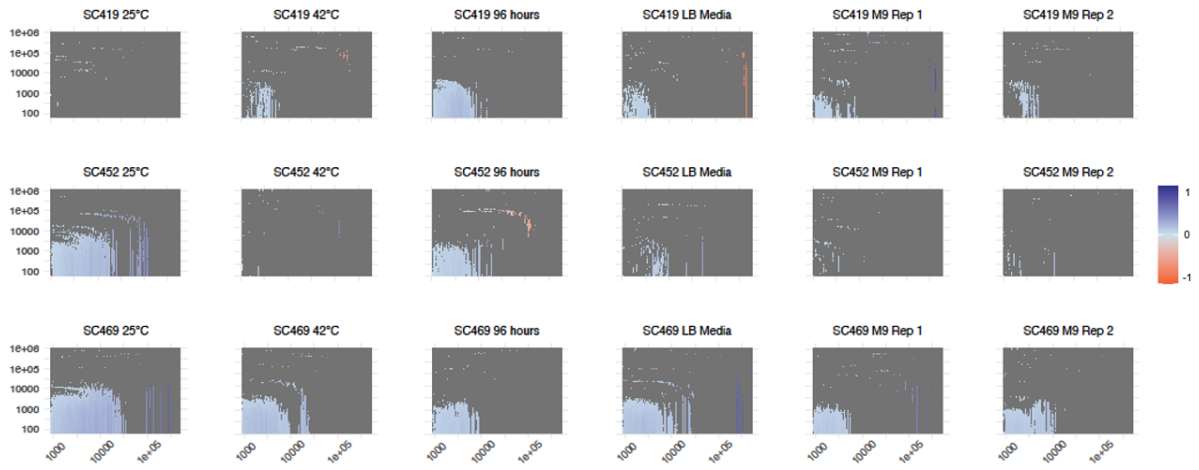


Figure S.4.8. Global autocorrelation plots for DAM methylation. The annotation and details of this plot are the same as those shown in **Supp. Fig. S7** but for DAM methylation. Again, for clarity only correlations in $p < 0.01$ are shown. The correlations here in the fraction of methylated sites in a window are in general stronger, but extend to a similar distance to those observed for DCM. Again, The strongest correlations are observed for strain SC469. However, correlations are also apparent for other strains in other conditions, also most likely due to the fact that DAM methylated fractions exhibited much greater variation than DCM (**Fig. 3**).

Chapter 5

Concluding remarks

DNA methylation is ubiquitous in bacteria. Bacterial DNA methyltransferases have been extensively studied both in the context of restriction modification systems and as orphan methyltransferases. DNA methylation has been linked with a range of cellular functions including regulating DNA repair, control of cell cycle initiation, and regulation of gene expression (Oliveira and Fang 2020; Sánchez-Romero, Cota, and Casadesús 2015; Blow et al. 2016; Oliveira 2021; John Beaulaurier, Schadt, and Fang 2018; Casadesús and Low 2006; D. Roberts et al. 1985). Control of cellular processes by DNA methylation can be broadly classified into two methods. First, the methylation state of the DNA strand can serve as a marker for a cell phase or DNA state. Secondly, DNA methylation can influence cellular processes through the competitive binding of methyltransferase with alternative regulatory proteins (Casadesús and Low 2006; D. A. Low, Weyand, and Mahan 2001; van der Woude, Hale, and Low 1998; Marinus and Casadesus 2009; Collier 2009). In both situations, DNA methylation can serve an epigenetic role and heritably influence phenotype.

Until recently, investigations of bacterial DNA methylation have relied on lab based approaches or used methods adapted from Eukaryotic studies (Y. Li and Tollefsbol 2011; Frommer et al. 1992). This has resulted in the majority of studies being focused on specific cellular processes or on virulence associated operons, in clinically relevant pathogenic isolates (P. Chen et al. 2017; Mou et al. 2015; Forde et al. 2015). DNA methylation patterns across the entire genome have not been well studied, therefore it is not well established whether DNA methylation serves an epigenetic function throughout the genome.

In this thesis, we aim to expand understanding of DNA methylation patterns in bacteria by using ONT sequencing to investigate genome wide patterns of DNA methylation in natural isolates of *E. coli*. As DNA is passed through the nanopore during sequencing, distinct squiggle patterns are recorded for canonical and non canonical bases. This allows for the detection of modified bases from native, or unaltered DNA (Simpson et al. 2017; Rand et al. 2017). In order to determine the genomic location and therefore any possible significance of modified bases, the reads must be anchored to an accurate reference genome. No reference genomes existed for our natural isolates of *E. coli,* therefore initial work in this thesis focuses on the production of reference quality de novo genome assemblies.

In chapter 2 we first use in silico experiments to investigate the structure of ONT sequencing data required to produce a complete hybrid assembly. Phillipy et al proposed the golden threshold of read length, which suggests that reads longer than the longest repeat region in a genome will lead to a complete genome assembly (Koren and Phillippy 2015). We use the *E.coli* MG1655 reference genome as a model for our natural isolates and simulate ONT reads of various lengths. We combine these reads with simulated Illumina sequence reads to investigate the effect of ONT read length on hybrid assembly completeness. We show that for *E. coli* MG1655 read lengths greater than 10 kbp result in a single contig hybrid genome assembly.

In order to generate sequencing reads longer than 10 kbp, we trial DNA extraction methods for the routine extraction of high molecular weight DNA. Using these methods we extract high molecular weight DNA for each of our 49 natural isolates. We then use this DNA to optimise sequencing library prep methods and to sequence each isolate. For each isolate, we produce ONT long read datasets containing reads that exceed our golden threshold estimate.

We use the long read assembler Flye to assemble each isolate, and polish each assembly with Illumina and Nanopore reads to produce hybrid genome assemblies. We assess each assembly for contiguity and structural accuracy through the arrangement of rRNA operons and deem 47 assemblies to be complete. This raises two key questions. First, how best to evaluate the completeness of a hybrid assembly, taking into account large scale structural accuracy and per basepair sequence accuracy. Secondly, do different assembly approaches produce the same assemblies.

In chapter 3, we aim to use the sequencing data produced for each of the 49 isolates to address the questions raised in chapter II. We quantitatively benchmark genome assemblies produced by five hybrid assembly methods across a range of metrics to determine assembly best practices. In addition to contiguity and the arrangement of rRNA operons, we use plasmid identification and the fraction of discordantly mapped Illumina reads as measures of structural accuracy. Additionally, we use the fraction of short open reading frames, and the number of variants called from short read data to determine per basepair sequence accuracy.

Each of the five assembly methods we test approaches the assembly problem uniquely. We find that assemblies built first with long reads, and polished by short reads tend to have greater structural accuracy than assemblies built from short reads first. In contrast, assemblies built from short reads first, and then scaffolded with long reads performed best on per basepair sequence accuracy metrics. However, these assemblies tended to be more

fragmented and struggled to accurately represent repetitive regions. Furthermore, we find that the distribution of read lengths and quality scores amongst Nanopore long read sequencing data was not strongly correlated with any of our metrics.

Although other studies have been conducted to assess assembly methods, the number of isolates and assembly methods used in this study exceeds other works (R. R. Wick and Holt 2019; De Maio et al. 2019). This allows for the differentiation of general and isolate specific patterns in sequencing and assembly. We found that despite having sequencing data of sufficient quality, some isolates were difficult to assemble across multiple assembly methods. We suggest that these isolates contain isolate specific idiosyncrasies which limit assembly accuracy. Despite this, the results presented in Chapter 2 highlights the intrinsic differences in assembly quality based on assembly method. These results support the recent findings of Wick et al who suggest combining multiple assembly approaches to generate a consensus assembly (R. R. Wick et al. 2021a).

In Chapter 4 we utilise the high quality reference genomes generated in Chapter 3 to investigate DNA methylation patterns in three natural isolates of *E. coli*. We first use each genome assembly to predict the presence of methyltransferases within each isolate. We then validate methyltransferase activity using comparisons between native and ground truth WGA sequencing data. Additionally, we detect the activity of unexpected methyltransferases, validating this approach as a method for methyltransferase discovery. We then use this approach to examine the genome wide patterns of DNA methylation across different environmental conditions. Our results show consistent patterns of methylation across the genome by each methyltransferase within growth conditions, as is also seen in other works (Payelleville et al. 2018). In contrast to Cohen et al (Cohen et al. 2016) who observed consistent methylation patterns in the face of antibiotic stress, we find consistent differences in methylation levels across growth conditions. Additionally, we find that these differences are not maintained into late stationary phase suggesting that during growth *E. coli* transiently differentiate into distinct methylation states.

In conclusion, this thesis presents complementary results. First, we establish quantitative quality metrics for the generation of accurate hybrid genome assemblies and investigate the effects of sequencing data quality on assembly accuracy. This work increases understanding of the limitations of hybrid assembly approaches and the quality requirements of a complete assembly. We then use these assemblies to investigate DNA methylation in natural isolates of *E. coli.* We establish methods for the detection of DNA methylation using Nanopore sequencing data, which we use to profile the frequency of DNA methylation across the genome in different growth conditions. Our results suggest genome wide DNA methylation

frequency is sensitive to growth conditions and that DNA methylation may serve as a marker of growth state. These results suggest not yet understood roles for DNA methylation in bacteria.

References

Adzitey, Frederick, Jonathan Asante, Hezekiel M. Kumalo, Rene B. Khan, Anou M. Somboro, and Daniel G. Amoako. 2020. “Genomic Investigation into the Virulome, Pathogenicity, Stress Response Factors, Clonal Lineages, and Phylogenetic Relationship of Escherichia Coli Strains Isolated from Meat Sources in Ghana.” *Genes* 11 (12). https://doi.org/10.3390/genes11121504.

Afonin, Alexey, Gribchenko Emma, Evgeny Zorin, Anton Sulima, and Vladimir Zhukov. 2021. “DNA Methylation Patterns Differ between Free-Living Rhizobium Leguminosarum RCAM1026 and Bacteroids Formed in Symbiosis with Pea (Pisum Sativum L.).” *bioRxiv*. https://doi.org/10.1101/2021.10.28.466258.

Allshire, Robin C., and Eric U. Selker. 2007. “Fungal Models for Epigenetic Research: Schizosaccharomyces Pombe and Neurospora Crassa.” *Epigenetics: Official Journal of the DNA Methylation Society*, 101–25.

Ambardar, Sheetal, Rikita Gupta, Deepika Trakroo, Rup Lal, and Jyoti Vakhlu. 2016. “High Throughput Sequencing: An Overview of Sequencing Chemistry.” *Indian Journal of Microbiology* 56 (4): 394–404.

Antipov, Dmitry, Anton Korobeynikov, Jeffrey S. McLean, and Pavel A. Pevzner. 2016. “hybridSPAdes: An Algorithm for Hybrid Assembly of Short and Long Reads.” *Bioinformatics* 32 (7): 1009–15.

Anton, Brian P., and Richard J. Roberts. 2021. “Beyond Restriction Modification: Epigenomic Roles of DNA Methylation in Prokaryotes.” *Annual Review of Microbiology* 75 (October): 129–49.

Arber, W. 2000. “Genetic Variation: Molecular Mechanisms and Impact on Microbial Evolution.” *FEMS Microbiology Reviews* 24 (1): 1–7.

Arredondo-Alonso, Sergio, Malbert R. C. Rogers, Johanna C. Braat, Tess D. Verschuuren, Janetta Top, Jukka Corander, Rob J. L. Willems, and Anita C. Schürch. 2018. “Mlplasmids: A User-Friendly Tool to Predict Plasmid- and Chromosome-Derived Sequences for Single Species.” *Microbial Genomics* 4 (11). https://doi.org/10.1099/mgen.0.000224.

Atack, John M., Aimee Tan, Lauren O. Bakaletz, Michael P. Jennings, and Kate L. Seib. 2018. “Phasevarions of Bacterial Pathogens: Methylomics Sheds New Light on Old Enemies.” *Trends in Microbiology* 26 (8): 715–26.

Auld, Josh R., Anurag A. Agrawal, and Rick A. Relyea. 2010. “Re-Evaluating the Costs and Limits of Adaptive Phenotypic Plasticity.” *Proceedings. Biological Sciences / The Royal Society* 277 (1681): 503–11.

Bailey, Timothy L., Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William S. Noble. 2009. “MEME SUITE: Tools for Motif Discovery and Searching.” *Nucleic Acids Research* 37 (Web Server issue): W202–8.

Baltrus, David A. 2013. “Exploring the Costs of Horizontal Gene Transfer.” *Trends in Ecology & Evolution* 28 (8): 489–95.

Bannister, Andrew J., and Tony Kouzarides. 2011. “Regulation of Chromatin by Histone Modifications.” *Cell Research* 21 (3): 381–95.

Bashir, Ali, Aaron A. Klammer, William P. Robins, Chen-Shan Chin, Dale Webster, Ellen Paxinos, David Hsu, et al. 2012. “A Hybrid Approach for the Automated Finishing of Bacterial Genomes.” *Nature Biotechnology* 30 (7): 701–7.

Beaulaurier, John, Eric E. Schadt, and Gang Fang. 2018. “Deciphering Bacterial Epigenomes Using Modern Sequencing Technologies.” *Nature Reviews. Genetics*, December, 1.

Beaulaurier, J., Xue-Song Zhang, Shijia Zhu, R. Sebra, C. Rosenbluh, G. Deikus, Nan Shen,

et al. 2015. “Single Molecule-Level Detection and Long Read-Based Phasing of Epigenetic Variations in Bacterial Methylomes.” *Nature Communications* 6. https://doi.org/10.1038/ncomms8438.
Bertels, Frederic, Olin K. Silander, Mikhail Pachkov, Paul B. Rainey, and Erik van Nimwegen. 2014. “Automated Reconstruction of Whole-Genome Phylogenies from Short-Sequence Reads.” *Molecular Biology and Evolution* 31 (5): 1077–88.
Bickle, T. A., and D. H. Krüger. 1993. “Biology of DNA Restriction.” *Microbiological Reviews* 57 (2): 434–50.
Bickle, Thomas A. 2004. “Restricting Restriction.” *Molecular Microbiology* 51 (1): 3–5.
Birkholz, Nils, Simon A. Jackson, Robert D. Fagerlund, and Peter C. Fineran. 2022. “A Mobile Restriction-Modification System Provides Phage Defence and Resolves an Epigenetic Conflict with an Antagonistic Endonuclease.” *Nucleic Acids Research* 50 (6): 3348–61.
Blow, Matthew J., Tyson A. Clark, Chris G. Daum, Adam M. Deutschbauer, Alexey Fomenkov, Roxanne Fries, Jeff Froula, et al. 2016. “The Epigenomic Landscape of Prokaryotes.” Edited by Gang Fang. *PLoS Genetics* 12 (2): e1005854.
Bonet, Jose, Mandi Chen, Marc Dabad, Simon Heath, Abel Gonzalez-Perez, Nuria Lopez-Bigas, and Jens Lagergren. 2021. “DeepMP: A Deep Learning Tool to Detect DNA Base Modifications on Nanopore Sequencing Data.” *bioRxiv*. https://doi.org/10.1101/2021.06.28.450135.
Breckell, Georgia, and Olin K. Silander. 2020. “Complete Genome Sequences of 47 Environmental Isolates of Escherichia Coli.” *Microbiology Resource Announcements* 9 (38). https://doi.org/10.1128/MRA.00222-20.
Buchfink, Benjamin, Klaus Reuter, and Hajk-Georg Drost. 2021. “Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND.” *Nature Methods* 18 (4): 366–68.
Bull, J. J. 1987. “Evolution of Phenotypic Variance.” *Evolution; International Journal of Organic Evolution* 41 (2): 303–15.
Camacho, Eva M., and Josep Casadesús. 2005. “Regulation of traJ Transcription in the Salmonella Virulence Plasmid by Strand-Specific DNA Adenine Hemimethylation.” *Molecular Microbiology* 57 (6): 1700–1718.
Carattoli, Alessandra, Ea Zankari, Aurora García-Fernández, Mette Voldby Larsen, Ole Lund, Laura Villa, Frank Møller Aarestrup, and Henrik Hasman. 2014. “In Silico Detection and Typing of Plasmids Using PlasmidFinder and Plasmid Multilocus Sequence Typing.” *Antimicrobial Agents and Chemotherapy* 58 (7): 3895–3903.
Caruccio, Nicholas. 2011. “Preparation of Next-Generation Sequencing Libraries Using Nextera™ Technology: Simultaneous DNA Fragmentation and Adaptor Tagging by In Vitro Transposition.” In *High-Throughput Next Generation Sequencing: Methods and Applications*, edited by Young Min Kwon and Steven C. Ricke, 241–55. Totowa, NJ: Humana Press.
Casadesús, Josep, and David Low. 2006. “Epigenetic Gene Regulation in the Bacterial World.” *Microbiology and Molecular Biology Reviews: MMBR* 70 (3): 830–56.
Casadesús, Josep, and David A. Low. 2013. “Programmed Heterogeneity: Epigenetic Mechanisms in Bacteria.” *The Journal of Biological Chemistry* 288 (20): 13929–35.
Chen, Liang, Haicheng Li, Tao Chen, Li Yu, Huixin Guo, Yuhui Chen, Mu Chen, et al. 2018. “Genome-Wide DNA Methylation and Transcriptome Changes in Mycobacterium Tuberculosis with Rifampicin and Isoniazid Resistance.” *International Journal of Clinical and Experimental Pathology* 11 (6): 3036–45.
Chen, Poyin, Henk C. den Bakker, Jonas Korlach, Nguyet Kong, Dylan B. Storey, Ellen E. Paxinos, Meredith Ashby, et al. 2017. “Comparative Genomics Reveals the Diversity of Restriction-Modification Systems and DNA Methylation Sites in Listeria Monocytogenes.” *Applied and Environmental Microbiology* 83 (3): e02091–16.
Chen, Shifu, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. “Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor.” *Bioinformatics* 34 (17): i884–90.
Chen, Ying, Fan Nie, Shang-Qian Xie, Ying-Feng Zheng, Qi Dai, Thomas Bray, Yao-Xin Wang, et al. 2021. “Efficient Assembly of Nanopore Reads via Highly Accurate and

Intact Error Correction.” *Nature Communications* 12 (1): 60.
Clark, Tyson A., Iain A. Murray, Richard D. Morgan, Andrey O. Kislyuk, Kristi E. Spittle, Matthew Boitano, Alexey Fomenkov, Richard J. Roberts, and Jonas Korlach. 2012. “Characterization of DNA Methyltransferase Specificities Using Single-Molecule, Real-Time DNA Sequencing.” *Nucleic Acids Research* 40 (4): e29.
Cohen, Nadia R., Christian A. Ross, Saloni R. Jain, R. Shapiro, A. Gutierrez, Peter Belenky, Hu Li, and J. J. Collins. 2016. “A Role for the Bacterial GATC Methylome in Antibiotic Stress Survival.” *Nature Genetics* 48: 581–86.
Collier, Justine. 2009. “Epigenetic Regulation of the Bacterial Cell Cycle.” *Current Opinion in Microbiology* 12 (6): 722–29.
Deatherage, Daniel E., and Jeffrey E. Barrick. 2014. “Identification of Mutations in Laboratory-Evolved Microbes from next-Generation Sequencing Data Using Breseq.” *Methods in Molecular Biology* 1151: 165–88.
De Maio, Nicola, Liam P. Shaw, Alasdair Hubbard, Sophie George, Nicholas D. Sanderson, Jeremy Swann, Ryan Wick, et al. 2019. “Comparison of Long-Read Sequencing Technologies in the Hybrid Assembly of Complex Bacterial Genomes.” *Microbial Genomics* 5 (9). https://doi.org/10.1099/mgen.0.000294.
Dewitt, T. J., A. Sih, and D. S. Wilson. 1998. “Costs and Limits of Phenotypic Plasticity.” *Trends in Ecology & Evolution* 13 (2): 77–81.
Dimitriu, Tatiana, Andrew Matthews, and Angus Buckling. n.d. “Increased Copy Number Couples the Evolution of Plasmid Horizontal Transmission and Antibiotic Resistance.” https://doi.org/10.1101/2020.08.12.248336.
Dohm, Juliane C., Philipp Peters, Nancy Stralis-Pavese, and Heinz Himmelbauer. 2020. “Benchmarking of Long-Read Correction Methods.” *NAR Genomics and Bioinformatics* 2 (2): lqaa037.
Earl, Dent, Keith Bradnam, John St John, Aaron Darling, Dawei Lin, Joseph Fass, Hung On Ken Yu, et al. 2011. “Assemblathon 1: A Competitive Assessment of de Novo Short Read Assembly Methods.” *Genome Research* 21 (12): 2224–41.
“Escherichia Coli Str. K-12 Substr. MG1655, Complete Genome - Nucleotide - NCBI.” n.d. Accessed May 9, 2022. https://www.ncbi.nlm.nih.gov/nuccore/556503834.
Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. “MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report.” *Bioinformatics* 32 (19): 3047–48.
Fang, Gang, Diana Munera, David I. Friedman, Anjali Mandlik, Michael C. Chao, Onureena Banerjee, Zhixing Feng, et al. 2012. “Genome-Wide Mapping of Methylated Adenine Residues in Pathogenic Escherichia Coli Using Single-Molecule Real-Time Sequencing.” *Nature Biotechnology* 30 (12): 1232–39.
Farine, Damien R., Pierre-Olivier Montiglio, and Orr Spiegel. 2015. “From Individuals to Groups and Back: The Evolutionary Implications of Group Phenotypic Composition.” *Trends in Ecology & Evolution* 30 (10): 609–21.
Feng, Suhua, Zhenhui Zhong, Ming Wang, and Steven E. Jacobsen. 2020. “Efficient and Accurate Determination of Genome-Wide DNA Methylation Patterns in Arabidopsis Thaliana with Enzymatic Methyl Sequencing.” *Epigenetics & Chromatin* 13 (1): 42.
Flusberg, Benjamin A., Dale R. Webster, Jessica H. Lee, Kevin J. Travers, Eric C. Olivares, Tyson A. Clark, Jonas Korlach, and Stephen W. Turner. 2010. “Direct Detection of DNA Methylation during Single-Molecule, Real-Time Sequencing.” *Nature Methods* 7 (6): 461–65.
Forde, Brian M., Minh-Duy Phan, Jayde A. Gawthorne, Melinda M. Ashcroft, Mitchell Stanton-Cook, Sohinee Sarkar, Kate M. Peters, et al. 2015. “Lineage-Specific Methyltransferases Define the Methylome of the Globally Disseminated Escherichia Coli ST131 Clone.” *mBio* 6 (6): e01602–15.
Freddolino, Peter L., Sasan Amini, and Saeed Tavazoie. 2012. “Newly Identified Genetic Variations in Common Escherichia Coli MG1655 Stock Cultures.” *Journal of Bacteriology* 194 (2): 303–6.
Frommer, M., L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy,

and C. L. Paul. 1992. “A Genomic Sequencing Protocol That Yields a Positive Display of 5-Methylcytosine Residues in Individual DNA Strands.” *Proceedings of the National Academy of Sciences of the United States of America* 89 (5): 1827–31.
García-Del Portillo, F., M. G. Pucciarelli, and J. Casadesús. 1999. “DNA Adenine Methylase Mutants of Salmonella Typhimurium Show Defects in Protein Secretion, Cell Invasion, and M Cell Cytotoxicity.” *Proceedings of the National Academy of Sciences of the United States of America* 96 (20): 11578–83.
Gaultney, Robert A., Antony T. Vincent, Céline Lorioux, Jean-Yves Coppée, Odile Sismeiro, Hugo Varet, Rachel Legendre, Charlotte A. Cockram, Frédéric J. Veyrier, and Mathieu Picardeau. 2020. “4-Methylcytosine DNA Modification Is Critical for Global Epigenetic Regulation and Virulence in the Human Pathogen Leptospira Interrogans.” *Nucleic Acids Research* 48 (21): 12102–15.
Genereux, Diane P., Brooks E. Miner, Carl T. Bergstrom, and Charles D. Laird. 2005. “A Population-Epigenetic Model to Infer Site-Specific Methylation Rates from Double-Stranded DNA Methylation Patterns.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (16): 5802–7.
Ghodsi, Mohammadreza, Christopher M. Hill, Irina Astrovskaya, Henry Lin, Dan D. Sommer, Sergey Koren, and Mihai Pop. 2013. “De Novo Likelihood-Based Measures for Comparing Genome Assemblies.” *BMC Research Notes* 6 (1): 334.
Goldstein, Sarah, Lidia Beka, Joerg Graf, and Jonathan L. Klassen. 2019. “Evaluation of Strategies for the Assembly of Diverse Bacterial Genomes Using MinION Long-Read Sequencing.” *BMC Genomics* 20 (1): 23.
Goodwin, Sara, James Gurtowski, Scott Ethe-Sayers, Panchajanya Deshpande, Michael C. Schatz, and W. Richard McCombie. 2015. “Oxford Nanopore Sequencing, Hybrid Error Correction, and de Novo Assembly of a Eukaryotic Genome.” *Genome Research* 25 (11): 1750–56.
Gormley, Niall A., Mark A. Watson, Stephen E. Halford, Niall A. Gormley, Mark A. Watson, and Stephen E. Halford. 2005. “Bacterial Restriction-Modification Systems.” In *Encyclopedia of Life Sciences*. Chichester, UK: John Wiley & Sons, Ltd.
Govers, Sander K., Julien Mortier, Antoine Adam, and Abram Aertsen. 2018. “Protein Aggregates Encode Epigenetic Memory of Stressful Encounters in Individual Escherichia Coli Cells.” *PLoS Biology* 16 (8): e2003853.
Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. “QUAST: Quality Assessment Tool for Genome Assemblies.” *Bioinformatics* 29 (8): 1072–75.
Hale, W. B., M. W. van der Woude, and D. A. Low. 1994. “Analysis of Nonmethylated GATC Sites in the Escherichia Coli Chromosome and Identification of Sites That Are Differentially Methylated in Response to Environmental Stimuli.” *Journal of Bacteriology* 176 (11): 3438–41.
Heard, Edith, and Robert A. Martienssen. 2014. “Transgenerational Epigenetic Inheritance: Myths and Mechanisms.” *Cell* 157 (1): 95–109.
Henderson, Ian R., and Steven E. Jacobsen. 2007. “Epigenetic Inheritance in Plants.” *Nature* 447 (7143): 418–24.
Hernandez-Beeftink, Tamara, Hector Rodriguez-Perez, Ana Diaz-de Usera, Rafaela Gonzalez-Montelongo, Jose M. Lorenzo-Salazar, Fabian Lorenzo-Diaz, and Carlos Flores. 2018. “Shallow MinION Sequencing to Assist de Novo Assembly of the Streptococcus Agalactiae Genome.” *bioRxiv*. https://doi.org/10.1101/485029.
Hernday, Aaron, Margareta Krabbe, Bruce Braaten, and David Low. 2002. “Self-Perpetuating Epigenetic Pili Switches in Bacteria.” *Proceedings of the National Academy of Sciences of the United States of America* 99 Suppl 4 (December): 16470–76.
Holliday, Robin. 2006. “Epigenetics: A Historical Overview.” *Epigenetics: Official Journal of the DNA Methylation Society* 1 (2): 76–80.
Homer, Nils. n.d. *DWGSIM: Whole Genome Simulator for Next-Generation Sequencing*. Github. Accessed December 13, 2021. https://github.com/nh13/DWGSIM.
Horesh, Gal, Grace A. Blackwell, Gerry Tonkin-Hill, Jukka Corander, Eva Heinz, and Nicholas R. Thomson. 2021. “A Comprehensive and High-Quality Collection of

Escherichia Coli Genomes and Their Genes." *Microbial Genomics* 7 (2). https://doi.org/10.1099/mgen.0.000499.
Hunt, Martin, Taisei Kikuchi, Mandy Sanders, Chris Newbold, Matthew Berriman, and Thomas D. Otto. 2013. "REAPR: A Universal Tool for Genome Assembly Evaluation." *Genome Biology* 14 (5): R47.
Hyatt, Doug, Gwo-Liang Chen, Philip F. Locascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification." *BMC Bioinformatics* 11 (March): 119.
Ishii, Satoshi, Winfried B. Ksoll, Randall E. Hicks, and Michael J. Sadowsky. 2006. "Presence and Growth of Naturalized Escherichia Coli in Temperate Soils from Lake Superior Watersheds." *Applied and Environmental Microbiology* 72 (1): 612–21.
Jablonka, Eva, and Marion J. Lamb. 2006. "The Changing Concept of Epigenetics." *Annals of the New York Academy of Sciences* 981 (1): 82–96.
Jablonka, Eva, and Gal Raz. 2009. "Transgenerational Epigenetic Inheritance: Prevalence, Mechanisms, and Implications for the Study of Heredity and Evolution." *The Quarterly Review of Biology* 84 (2): 131–76.
Jaenisch, Rudolf, and Adrian Bird. 2003. "Epigenetic Regulation of Gene Expression: How the Genome Integrates Intrinsic and Environmental Signals." *Nature Genetics* 33 Suppl (March): 245–54.
Jain, Miten, Hugh E. Olsen, Benedict Paten, and Mark Akeson. 2016. "The Oxford Nanopore MinION : Delivery of Nanopore Sequencing to the Genomics Community." *Genome Biology*, 1–11.
Jan Grimbergen, Ard, Jeroen Siebring, Ana Solopova, and Oscar P. Kuipers. 2015. "Microbial Bet-Hedging: The Power of Being Different." *Current Opinion in Microbiology* 25: 67–72.
Jeltsch, Albert. 2002. "Beyond Watson and Crick: DNA Methylation and Molecular Enzymology of DNA Methyltransferases." *Chembiochem: A European Journal of Chemical Biology* 3 (4): 274–93.
Johnston, Calum, Bernard Martin, Gwennaele Fichant, Patrice Polard, and Jean-Pierre Claverys. 2014. "Bacterial Transformation: Distribution, Shared Mechanisms and Divergent Control." *Nature Reviews. Microbiology* 12 (3): 181–96.
Jong, Imke G. de, Patsy Haccou, and Oscar P. Kuipers. 2011. "Bet Hedging or Not? A Guide to Proper Classification of Microbial Survival Strategies." *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 33 (3): 215–23.
Julio, Steven M., Douglas M. Heithoff, Daniele Provenzano, Karl E. Klose, Robert L. Sinsheimer, David A. Low, and Michael J. Mahan. 2001. "DNA Adenine Methylase Is Essential for Viability and Plays a Role in the Pathogenesis of Yersinia Pseudotuberculosis andVibrio Cholerae." *Infection and Immunity* 69 (12): 7610–15.
Kahng, L. S., and L. Shapiro. 2001. "The CcrM DNA Methyltransferase of Agrobacterium Tumefaciens Is Essential, and Its Activity Is Cell Cycle Regulated." *Journal of Bacteriology* 183 (10): 3065–75.
Kaiser, Matthias, Florian Jug, Thomas Julou, Siddharth Deshpande, Thomas Pfohl, Olin K. Silander, Gene Myers, and Erik van Nimwegen. 2018. "Monitoring Single-Cell Gene Regulation under Dynamically Controllable Conditions with Integrated Microfluidics and Software." *Nature Communications* 9 (1): 212.
Kaufman, James H., Ignacio Terrizzano, Gowri Nayar, Ed Seabolt, Akshay Agarwal, Ilya B. Slizovskiy, and Noelle Noyes. n.d. "Integrative and Conjugative Elements (ICE) and Associated Cargo Genes within and across Hundreds of Bacterial Genera." https://doi.org/10.1101/2020.04.07.030320.
Kearns, Daniel B. 2008. "Division of Labour during Bacillus Subtilis Biofilm Formation." *Molecular Microbiology*.
Klassen, Jonathan L., and Cameron R. Currie. 2012. "Gene Fragmentation in Bacterial Draft Genomes: Extent, Consequences and Mitigation." *BMC Genomics* 13 (January): 14.
Knierim, Ellen, Barbara Lucke, Jana Marie Schwarz, Markus Schuelke, and Dominik Seelow. 2011. "Systematic Comparison of Three Methods for Fragmentation of Long-Range

PCR Products for next Generation Sequencing." *PloS One* 6 (11): e28240.
Kolmogorov, Mikhail, Jeffrey Yuan, Yu Lin, and Pavel A. Pevzner. 2019. "Assembly of Long, Error-Prone Reads Using Repeat Graphs." *Nature Biotechnology* 37 (5): 540–46.
Kong, Huimin, Lee-Fong Lin, Nicole Porter, Shawn Stickel, Devon Byrd, Janos Posfai, and Richard J. Roberts. 2000. "Functional Analysis of Putative Restriction–modification System Genes in the Helicobacter Pylori J99 Genome." *Nucleic Acids Research* 28 (17): 3216–23.
Koren, Sergey, and Adam M. Phillippy. 2015. "ScienceDirect One Chromosome , One Contig : Complete Microbial Genomes from Long-Read Sequencing and Assembly." *Current Opinion in Microbiology* 23: 110–20.
Koren, Sergey, Michael C. Schatz, Brian P. Walenz, Jeffrey Martin, Jason T. Howard, Ganeshkumar Ganapathy, Zhong Wang, et al. 2012. "Hybrid Error Correction and de Novo Assembly of Single-Molecule Sequencing Reads." *Nature Biotechnology* 30 (7): 693–700.
Koren, Sergey, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. 2017. "Canu: Scalable and Accurate Long-Read Assembly via Adaptive *k* -Mer Weighting and Repeat Separation." *Genome Research*, 1–35.
Krueger, Felix, Benjamin Kreck, Andre Franke, and Simon R. Andrews. 2012. "DNA Methylome Analysis Using Short Bisulfite Sequencing Data." *Nature Methods* 9 (2): 145–51.
Kurtz, Stefan, Adam Phillippy, Arthur L. Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, and Steven L. Salzberg. 2004. "Versatile and Open Software for Comparing Large Genomes." *Genome Biology* 5 (2): R12.
Kussell, Edo, and Stanislas Leibler. 2005. "Phenotypic Diversity, Population Growth, and Information in Fluctuating Environments." *Science* 309 (5743): 2075–78.
Lambert, Guillaume, and Edo Kussell. 2014. "Memory and Fitness Optimization of Bacteria under Fluctuating Environments." *PLoS Genetics* 10 (9): e1004556.
Lande, Russell. 1976. "Natural Selection and Random Genetic Drift in Phenotypic Evolution." *Evolution; International Journal of Organic Evolution* 30 (2): 314–34.
Lannoy, Carlos de, Dick de Ridder, and Judith Risse. 2017. "The Long Reads Ahead: De Novo Genome Assembly Using the MinION." *F1000Research* 6 (December): 1083.
Lee, Heewook, Thomas G. Doak, Ellen Popodi, Patricia L. Foster, and Haixu Tang. 2016. "Insertion Sequence-Caused Large-Scale Rearrangements in the Genome of Escherichia Coli." *Nucleic Acids Research* 44 (15): 7109–19.
Liao, Yu-Chieh, Shu-Hung Lin, and Hsin-Hung Lin. 2015. "Completing Bacterial Genome Assemblies: Strategy and Performance Comparisons." *Scientific Reports* 5 (March): 8747.
Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM," March. http://arxiv.org/abs/1303.3997.
———. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." *Bioinformatics* 34 (18): 3094–3100.
Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79.
Liu, Qian, Li Fang, Guoliang Yu, Depeng Wang, Chuan-Le Xiao, and Kai Wang. 2019. "Detection of DNA Base Modifications by Deep Recurrent Neural Network on Oxford Nanopore Sequencing Data." *Nature Communications* 10 (1): 2449.
Li, Yuanyuan, and Trygve O. Tollefsbol. 2011. "DNA Methylation Detection: Bisulfite Genomic Sequencing Analysis." *Methods in Molecular Biology* 791: 11–21.
Løbner-Olesen, Anders, Ole Skovgaard, and Martin G. Marinus. 2005. "Dam Methylation: Coordinating Cellular Processes." *Current Opinion in Microbiology* 8 (2): 154–60.
Loman, Nicholas J., Chrystala Constantinidou, Jacqueline Z. M. Chan, Mihail Halachev, Martin Sergeant, Charles W. Penn, Esther R. Robinson, and Mark J. Pallen. 2012. "High-Throughput Bacterial Genome Sequencing: An Embarrassment of Choice, a

World of Opportunity." *Nature Reviews. Microbiology* 10 (9): 599–606.
Long, Anthony, Gianni Liti, Andrej Luptak, and Olivier Tenaillon. 2015. "Elucidating the Molecular Architecture of Adaptation via Evolve and Resequence Experiments." *Nature Reviews. Genetics* 16 (10): 567–82.
Low, David A., and Josep Casadesús. 2008. "Clocks and Switches: Bacterial Gene Regulation by DNA Adenine Methylation." *Current Opinion in Microbiology* 11 (2): 106–12.
Low, D. A., N. J. Weyand, and M. J. Mahan. 2001. "Roles of DNA Adenine Methylation in Regulating Bacterial Gene Expression and Virulence." *Infection and Immunity* 69 (12): 7197–7204.
Lu, Hengyun, Francesca Giordano, and Zemin Ning. 2016. "Oxford Nanopore MinION Sequencing and Genome Assembly." *Genomics, Proteomics & Bioinformatics* 14 (5): 265–79.
Marçais, Guillaume, James A. Yorke, and Aleksey Zimin. 2015. "QuorUM: An Error Corrector for Illumina Reads." *PloS One* 10 (6): e0130821.
Marinus, Martin G., and Josep Casadesus. 2009. "Roles of DNA Adenine Methylation in Host–pathogen Interactions: Mismatch Repair, Transcriptional Regulation, and More." *FEMS Microbiology Reviews* 33 (3): 488–503.
"Medaka — Medaka Documentation." n.d. Accessed November 7, 2021. https://nanoporetech.github.io/medaka/.
"Megalodon." n.d. Accessed May 18, 2022. https://nanoporetech.github.io/megalodon/.
Mehershahi, Kurosh S., and S. Chen. 2021. "Methylation by Multiple Type I Restriction Modification Systems Avoids Influencing Gene Regulation in Uropathogenic Escherichia Coli." *bioRxiv*. https://doi.org/10.1101/2021.01.08.425850.
Meisel, A., T. A. Bickle, D. H. Krüger, and C. Schroeder. 1992. "Type III Restriction Enzymes Need Two Inversely Oriented Recognition Sites for DNA Cleavage." *Nature* 355 (6359): 467–69.
Militello, Kevin T., Robert D. Simon, Mehr Qureshi, Robert Maines, Michelle L. VanHorne, Stacy M. Hennick, Sangeeta K. Jayakar, and Sarah Pounder. 2012. "Conservation of Dcm-Mediated Cytosine DNA Methylation in Escherichia Coli." *FEMS Microbiology Letters* 328 (1): 78–85.
Modlin, Samuel J., Derek Conkle-Gutierrez, Calvin Kim, Scott N. Mitchell, Christopher Morrissey, Brian C. Weinrick, William R. Jacobs, Sarah M. Ramirez-Busby, Sven E. Hoffner, and Faramarz Valafar. 2020. "Drivers and Sites of Diversity in the DNA Adenine Methylomes of 93 Mycobacterium Tuberculosis Complex Clinical Isolates." *eLife* 9 (October). https://doi.org/10.7554/eLife.58542.
Mou, Kathy T., Usha K. Muppirala, Andrew J. Severin, Tyson A. Clark, Matthew Boitano, and Paul J. Plummer. 2015. "A Comparative Analysis of Methylome Profiles of Campylobacter Jejuni Sheep Abortion Isolate and Gastroenteric Strains Using PacBio Data." *Frontiers in Microbiology* 5 (January): 782.
Murigneux, Valentine, Leah W. Roberts, Brian M. Forde, Minh-Duy Phan, Nguyen Thi Khanh Nhu, Adam D. Irwin, Patrick N. A. Harris, et al. 2021. "MicroPIPE: An End-to-End Solution for High-Quality Complete Bacterial Genome Construction." *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2021.02.02.429319.
Nanney, D. L. 1958. "EPIGENETIC CONTROL SYSTEMS." *Proceedings of the National Academy of Sciences of the United States of America* 44 (7): 712–17.
Ni, Peng, Neng Huang, Feng Luo, and Jianxin Wang. 2018. "DeepSignal: Detecting DNA Methylation State from Nanopore Sequencing Reads Using Deep-Learning." *bioRxiv*, August, 385849.
Norman, Anders, Lars H. Hansen, and Søren J. Sørensen. 2009. "Conjugative Plasmids: Vessels of the Communal Gene Pool." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 364 (1527): 2275–89.
Oliveira, Pedro H. 2021. "Bacterial Epigenomics: Coming of Age." *mSystems* 6 (4): e0074721.
Oliveira, Pedro H., and Gang Fang. 2020. "Conserved DNA Methyltransferases: A Window

into Fundamental Mechanisms of Epigenetic Regulation in Bacteria." *Trends in Microbiology*, May. https://doi.org/10.1016/j.tim.2020.04.007.
"ONT Accuracy." n.d. Oxford Nanopore Technologies. Accessed April 13, 2022. https://nanoporetech.com/accuracy.
Oxford Nanopore Technologies. n.d. "Rapid Sequencing gDNA - Barcoding (SQK-RBK004)."
Page, Andrew J., Emma V. Ainsworth, and Gemma C. Langridge. 2020. "Socru: Typing of Genome-Level Order and Orientation around Ribosomal Operons in Bacteria." *Microbial Genomics* 6 (7). https://doi.org/10.1099/mgen.0.000396.
Park, Hye-Jee, Boknam Jung, Jungkwan Lee, and Sang-Wook Han. 2019. "Functional Characterization of a Putative DNA Methyltransferase, EadM, in Xanthomonas Axonopodis Pv. Glycines by Proteomic and Phenotypic Analyses." *Scientific Reports* 9 (1): 2446.
Partridge, Sally R., Stephen M. Kwong, Neville Firth, and Slade O. Jensen. 2018. "Mobile Genetic Elements Associated with Antimicrobial Resistance." *Clinical Microbiology Reviews* 31 (4). https://doi.org/10.1128/CMR.00088-17.
Paul Coupland, Liz Sheridan. n.d. "Needle Shearing DNA for PacBio >20 Kb Libraries." http://www.2einteractive.com/pacbio/Needle-Shearing-DNA-for-PacBio-20kb-Libraries-Sanger.pdf.
Payelleville, Amaury, Ludovic Legrand, J. Ogier, Céline Roques, A. Roulet, O. Bouchez, Annabelle Mouammine, A. Givaudan, and Julien Brillard. 2018. "The Complete Methylome of an Entomopathogenic Bacterium Reveals the Existence of Loci with Unmethylated Adenines." *Scientific Reports* 8. https://doi.org/10.1038/s41598-018-30620-5.
Payne, Alexander, Nadine Holmes, Vardhman Rakyan, and Matthew Loose. 2019. "BulkVis: A Graphical Viewer for Oxford Nanopore Bulk FAST5 Files." *Bioinformatics* 35 (13): 2193–98.
Peterson, Stacey N., and Norbert O. Reich. 2006. "GATC Flanking Sequences Regulate Dam Activity: Evidence for How Dam Specificity May Influence Pap Expression." *Journal of Molecular Biology* 355 (3): 459–72.
———. 2008. "Competitive Lrp and Dam Assembly at the Pap Regulatory Region: Implications for Mechanisms of Epigenetic Regulation." *Journal of Molecular Biology* 383 (1): 92–105.
Pevzner, Pavel A., Paul A. Pevzner, Haixu Tang, and Glenn Tesler. 2004. "De Novo Repeat Classification and Fragment Assembly." *Genome Research* 14 (9): 1786–96.
Phillippy, Adam M., Michael C. Schatz, and Mihai Pop. 2008. "Genome Assembly Forensics: Finding the Elusive Mis-Assembly." *Genome Biology* 9 (3): R55.
Pigliucci, Massimo. 2001. *Phenotypic Plasticity: Beyond Nature and Nurture*. JHU Press.
Pingoud, A., M. Fuxreiter, V. Pingoud, and W. Wende. 2005. "Type II Restriction Endonucleases: Structure and Mechanism." *Cellular and Molecular Life Sciences: CMLS* 62 (6): 685–707.
Poptsova, Maria S., Irina A. Il'icheva, Dmitry Yu Nechipurenko, Larisa A. Panchenko, Mingian V. Khodikov, Nina Y. Oparina, Robert V. Polozov, Yury D. Nechipurenko, and Sergei L. Grokhovsky. 2014. "Non-Random DNA Fragmentation in next-Generation Sequencing." *Scientific Reports* 4 (March): 4532.
Quick, Josh. 2018. "Ultra-Long Read Sequencing Protocol for RAD004 Protocol by Josh Quick." January 22, 2018. https://doi.org/10.17504/protocols.io.mrxc57n.
Rainey, Paul B., Hubertus J. E. Beaumont, Gayle C. Ferguson, Jenna Gallie, Christian Kost, Eric Libby, and Xue-Xian Zhang. 2011. "The Evolutionary Emergence of Stochastic Phenotype Switching in Bacteria." *Microbial Cell Factories* 10 Suppl 1 (August): S14.
Rand, Arthur C., Miten Jain, Jordan M. Eizenga, Audrey Musselman-Brown, Hugh E. Olsen, Mark Akeson, and Benedict Paten. 2017. "Mapping DNA Methylation with High-Throughput Nanopore Sequencing." *Nature Methods* 14 (4): 411–13.
Rang, Franka J., Wigard P. Kloosterman, and Jeroen de Ridder. 2018. "From Squiggle to Basepair: Computational Approaches for Improving Nanopore Sequencing Read Accuracy." *Genome Biology*. https://doi.org/10.1186/s13059-018-1462-9.

Redondo-Salvo, Santiago, Raúl Fernández-López, Raúl Ruiz, Luis Vielva, María de Toro, Eduardo P. C. Rocha, M. Pilar Garcillán-Barcia, and Fernando de la Cruz. 2020. “Pathways for Horizontal Gene Transfer in Bacteria Revealed by a Global Map of Their Plasmids.” *Nature Communications* 11 (1): 3602.
Reisenauer, A., L. S. Kahng, S. McCollum, and L. Shapiro. 1999. “Bacterial DNA Methylation: A Cell Cycle Regulator?” *Journal of Bacteriology* 181 (17): 5135–39.
Rhoads, Anthony, and Kin Fai Au. 2015. “PacBio Sequencing and Its Applications.” *Genomics, Proteomics & Bioinformatics* 13 (5): 278–89.
Roberts, D., B. C. Hoopes, W. R. McClure, and N. Kleckner. 1985. “IS10 Transposition Is Regulated by DNA Adenine Methylation.” *Cell* 43 (1): 117–30.
Roberts, Richard J., Marlene Belfort, Timothy Bestor, Ashok S. Bhagwat, Thomas A. Bickle, Jurate Bitinaite, Robert M. Blumenthal, et al. 2003. “A Nomenclature for Restriction Enzymes, DNA Methyltransferases, Homing Endonucleases and Their Genes.” *Nucleic Acids Research* 31 (7): 1805–12.
Roberts, Richard J., Tamas Vincze, Janos Posfai, and Dana Macelis. n.d. “REBASE—a Database for DNA Restriction and Modification: Enzymes, Genes and Genomes.” https://watermark.silverchair.com/gku1046.pdf?token=AQECAHi208BE49Ooan9kkhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAAj4wggI6BgkqhkiG9w0BBwagggIrMIICJwIBADCCAiAGCSqGSIb3DQEHATAeBglghkgBZQMEAS4wEQQM8KgS53oFJCO8PfHKAgEQgIIB8dHXz5gEfRShLOhm_N-k0fphNJLSjFndecQCMtuPxKtl6u0.
Ruan, Jue, and Heng Li. 2019. “Fast and Accurate Long-Read Assembly with wtdbg2.” *bioRxiv*, January, 530972.
Sánchez-Romero, María A., and Josep Casadesús. 2020. “The Bacterial Epigenome.” *Nature Reviews. Microbiology* 18 (1): 7–20.
Sánchez-Romero, María A., Ignacio Cota, and Josep Casadesús. 2015. “DNA Methylation in Bacteria: From the Methyl Group to the Methylome.” *Current Opinion in Microbiology* 25 (June): 9–16.
Schalamun, Miriam, and Benjamin Schwessinger. 2017. “DNA Size Selection (>1kb) and Clean up Using an Optimized SPRI Beads Mixture.” protocols.io. July 1, 2017. https://www.protocols.io/view/dna-size-selection-1kb-and-clean-up-using-an-optim-idmca46.
Schlichting, Carl, and Massimo Pigliucci. 1998. *Phenotypic Evolution: A Reaction Norm Perspective*. Sinauer.
Seemann, Torsten. 2014. “Prokka: Rapid Prokaryotic Genome Annotation.” *Bioinformatics* 30 (14): 2068–69.
Seong, Hoon Je, Sang-Wook Han, and Woo Jun Sul. 2021. “Prokaryotic DNA Methylation and Its Functional Roles.” *Journal of Microbiology* 59 (3): 242–48.
Shen, Wei, Shuai Le, Yan Li, and Fuquan Hu. 2016. “SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation.” *PloS One* 11 (10): e0163962.
Simpson, Jared T., Rachael E. Workman, P. C. Zuzarte, Matei David, L. J. Dursi, and Winston Timp. 2017. “Detecting DNA Cytosine Methylation Using Nanopore Sequencing.” *Nature Methods* 14 (4): 407–10.
Skarstad, K., E. Boye, and H. B. Steen. 1986. “Timing of Initiation of Chromosome Replication in Individual Escherichia Coli Cells.” *The EMBO Journal* 5 (7): 1711–17.
Smits, Wiep Klaas, Oscar P. Kuipers, and Jan-Willem Veening. 2006. “Phenotypic Variation in Bacteria: The Role of Feedback Regulation.” *Nature Reviews. Microbiology* 4 (4): 259–71.
Sommer, Ralf J. 2020. “Phenotypic Plasticity: From Theory and Genetics to Current and Future Challenges.” *Genetics* 215 (1): 1–13.
Song, Fei, Joseph F. Smith, Makoto T. Kimura, Arlene D. Morrow, Tomoki Matsuyama, Hiroki Nagase, and William A. Held. 2005. “Association of Tissue-Specific Differentially Methylated Regions (TDMs) with Differential Gene Expression.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (9): 3336–41.
Soucy, Shannon M., Jinling Huang, and Johann Peter Gogarten. 2015. “Horizontal Gene Transfer: Building the Web of Life.” *Nature Reviews. Genetics* 16 (8): 472–82.

Sović, Ivan, Krešimir Križanović, Karolj Skala, and Mile Šikić. 2016. “Evaluation of Hybrid and Non-Hybrid Methods for de Novo Assembly of Nanopore Reads.” *Bioinformatics* 32 (17): 2582–89.
Srikhanta, Yogitha N., Stefanie J. Dowideit, Jennifer L. Edwards, Megan L. Falsetta, Hsing-Ju Wu, Odile B. Harrison, Kate L. Fox, et al. 2009. “Phasevarions Mediate Random Switching of Gene Expression in Pathogenic Neisseria.” *PLoS Pathogens* 5 (4): e1000400.
Stamatakis, Alexandros. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.” *Bioinformatics* 30 (9): 1312–13.
Stoiber, Marcus H., Joshua Quick, Rob Egan, Ji Eun Lee, Susan E. Celniker, Robert Neely, Nicholas Loman, Len Pennacchio, and James B. Brown. 2016. “De Novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing.” *bioRxiv*, 094672.
Timp, Winston. Letter to Ubioinfo Slack Channel. 2018, November 2, 2018.
“Tombo.” n.d. Accessed September 6, 2021. https://nanoporetech.github.io/tombo/.
Tørresen, Ole K., Bastiaan Star, Pablo Mier, Miguel A. Andrade-Navarro, Alex Bateman, Patryk Jarnot, Aleksandra Gruca, et al. 2019. “Tandem Repeats Lead to Sequence Assembly Errors and Impose Multi-Level Challenges for Genome and Protein Databases.” *Nucleic Acids Research* 47 (21): 10994–6.
Tóth, Adrienn Gréta, István Csabai, Maura Fiona Judge, Gergely Maróti, Ágnes Becsei, Sándor Spisák, and Norbert Solymosi. 2021. “Mobile Antimicrobial Resistance Genes in Probiotics.” *bioRxiv*. https://doi.org/10.1101/2021.05.04.442546.
Totsika, Makrina, Scott A. Beatson, Nicola Holden, and David L. Gally. 2008. “Regulatory Interplay between Pap Operons in Uropathogenic Escherichia Coli.” *Molecular Microbiology* 67 (5): 996–1011.
Touchon, Marie, Claire Hoede, Olivier Tenaillon, Valérie Barbe, Simon Baeriswyl, Philippe Bidet, Edouard Bingen, et al. 2009. “Organised Genome Dynamics in the Escherichia Coli Species Results in Highly Diverse Adaptive Paths.” *PLoS Genetics* 5 (1). https://doi.org/10.1371/journal.pgen.1000344.
Tourancheau, Alan, Edward A. Mead, Xue-Song Zhang, and Gang Fang. 2021. “Discovering Multiple Types of DNA Methylation from Bacteria and Microbiome Using Nanopore Sequencing.” *Nature Methods* 18 (5): 491–98.
Traxler, Matthew F., and Daniel E. Rozen. 2022. “Ecological Drivers of Division of Labour in Streptomyces.” *Current Opinion in Microbiology* 67 (April): 102148.
Treangen, Todd J., and Steven L. Salzberg. 2011. “Repetitive DNA and next-Generation Sequencing: Computational Challenges and Solutions.” *Nature Reviews. Genetics* 13 (1): 36–46.
Utturkar, Sagar M., Dawn M. Klingeman, Miriam L. Land, Christopher W. Schadt, Mitchel J. Doktycz, Dale A. Pelletier, and Steven D. Brown. 2014. “Evaluation and Validation of de Novo and Hybrid Assembly Techniques to Derive High-Quality Genome Sequences.” *Bioinformatics* 30 (19): 2709–16.
Vanderkraats, Nathan D., Jeffrey F. Hiken, Keith F. Decker, and John R. Edwards. 2013. “Discovering High-Resolution Patterns of Differential DNA Methylation That Correlate with Gene Expression Changes.” *Nucleic Acids Research* 41 (14): 6816–27.
Vaser, Robert, and Mile Šikić. 2021. “Time- and Memory-Efficient Genome Assembly with Raven.” *Nature Computational Science* 1 (5): 332–36.
———. n.d. “Raven: A de Novo Genome Assembler for Long Reads.” https://doi.org/10.1101/2020.08.07.242461.
Vaser, Robert, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. 2017. “Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads.” *Genome Research* 27 (5): 737–46.
Vasudevan, Karthick, Naveen Kumar Devanga Ragupathi, Jobin John Jacob, and Balaji Veeraraghavan. 2019. “Highly Accurate-Single Chromosomal Complete Genomes Using IonTorrent and MinION Sequencing of Clinical Pathogens.” *Genomics*, April. https://doi.org/10.1016/J.YGENO.2019.04.006.

Vasu, Kommireddy, and Valakunja Nagaraja. 2013. “Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense.” *Microbiology and Molecular Biology Reviews: MMBR* 77 (1): 53–72.
Veening, Jan-Willem, Wiep Klaas Smits, and Oscar P. Kuipers. 2008. “Bistability, Epigenetics, and Bet-Hedging in Bacteria.” *Annual Review of Microbiology* 62 (1): 193–210.
Veening, Jan-Willem, Eric J. Stewart, Thomas W. Berngruber, François Taddei, Oscar P. Kuipers, and Leendert W. Hamoen. 2008. “Bet-Hedging and Epigenetic Inheritance in Bacterial Cell Development.”
Waddington, C. H. 1942. “CANALIZATION OF DEVELOPMENT AND THE INHERITANCE OF ACQUIRED CHARACTERS.” *Nature* 150 (November): 563.
Waldminghaus, Torsten, and Kirsten Skarstad. 2009. “The Escherichia Coli SeqA Protein.” *Plasmid* 61 (3): 141–50.
Walker, Bruce J., Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, et al. 2014. “Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement.” Edited by Junwen Wang. *PloS One* 9 (11): e112963.
Watson, Michael E., Jr, Justin Jarisch, and Arnold L. Smith. 2004. “Inactivation of Deoxyadenosine Methyltransferase (dam) Attenuates Haemophilus Influenzae Virulence.” *Molecular Microbiology* 53 (2): 651–64.
Watson, Mick. n.d. *Ideel*. Github. Accessed November 9, 2018. https://github.com/mw55309/ideel.
Wenger, Aaron M., Paul Peluso, William J. Rowell, Pi-Chuan Chang, Richard J. Hall, Gregory T. Concepcion, Jana Ebler, et al. 2019. “Accurate Circular Consensus Long-Read Sequencing Improves Variant Detection and Assembly of a Human Genome.” *Nature Biotechnology* 37 (10): 1155–62.
West-Eberhard, M. J. 1989. “Phenotypic Plasticity and the Origins of Diversity.” *Annual Review of Ecology and Systematics* 20 (1): 249–78.
Weyand, Nathan J., Bruce A. Braaten, Marjan Van Der Woude, Julie Tucker, and David A. Low. 2001. “The Essential Role of the Promoter-Proximal Subunit of CAP in Pap Phase Variation: Lrp-and Helical Phase-Dependent Activation of papBA Transcription by CAP from- 215.” *Molecular Microbiology* 39 (6): 1504–22.
Whitman, Douglas W., Taracad Narayanan Ananthakrishnan, and Others. 2009. *Phenotypic Plasticity of Insects: Mechanisms and Consequences*. Science Publishers, Inc.
Wick, Ryan. n.d. *Filtlong: Quality Filtering Tool for Long Reads*. Github. Accessed July 27, 2021a. https://github.com/rrwick/Filtlong.
———. n.d. *Porechop: Adapter Trimmer for Oxford Nanopore Reads*. Github. Accessed December 15, 2021b. https://github.com/rrwick/Porechop.
Wick, Ryan R., and Kathryn E. Holt. 2019. “Benchmarking of Long-Read Assemblers for Prokaryote Whole Genome Sequencing.” *F1000Research* 8 (December): 2138.
———. 2021. “Polypolish: Short-Read Polishing of Long-Read Bacterial Genome Assemblies.” *bioRxiv*. https://doi.org/10.1101/2021.10.14.464465.
Wick, Ryan R., Louise M. Judd, Louise T. Cerdeira, Jane Hawkey, Guillaume Méric, Ben Vezina, Kelly L. Wyres, and Kathryn E. Holt. 2021a. “Trycycler: Consensus Long-Read Assemblies for Bacterial Genomes.” *Genome Biology* 22 (1): 266.
———. 2021b. “Trycycler: Consensus Long-Read Assemblies for Bacterial Genomes.” *Genome Biology* 22 (1): 266.
Wick, Ryan R., Louise M. Judd, Claire L. Gorrie, and Kathryn E. Holt. 2017a. “Unicycler: Resolving Bacterial Genome Assemblies from Short and Long Sequencing Reads.” *PLoS Computational Biology* 13 (6): e1005595.
———. 2017b. “Completing Bacterial Genome Assemblies with Multiplex MinION Sequencing.” *Microbial Genomics* 3 (10): e000132.
Wick, Ryan R., Mark B. Schultz, Justin Zobel, and Kathryn E. Holt. 2015. “Bandage: Interactive Visualization of de Novo Genome Assemblies.” *Bioinformatics* 31 (20): 3350–52.

Wilson, Geoffrey G., and Noreen E. Murray. 1991. “RESTRICTION AND MODIFICATION SYSTEMS.” *Annual Review of Genetics* 25: 585–627.
Wion, Didier, and Josep Casadesús. 2006. “N6-Methyl-Adenine: An Epigenetic Signal for DNA-Protein Interactions.” *Nature Reviews. Microbiology* 4 (3): 183–92.
Woude, M. van der, B. Braaten, and D. Low. 1996. “Epigenetic Phase Variation of the Pap Operon in Escherichia Coli.” *Trends in Microbiology* 4 (1): 5–9.
Woude, M. van der, W. B. Hale, and D. A. Low. 1998. “Formation of DNA Methylation Patterns: Nonmethylated GATC Sequences in Gut and Pap Operons.” *Journal of Bacteriology* 180 (22): 5913–20.
Wright, R., C. Stephens, and L. Shapiro. 1997. “The CcrM DNA Methyltransferase Is Widespread in the Alpha Subdivision of Proteobacteria, and Its Essential Functions Are Conserved in Rhizobium Meliloti and Caulobacter Crescentus.” *Journal of Bacteriology* 179 (18): 5869–77.
Wu, C. T., and J. R. Morris. 2001. “Genes, Genetics, and Epigenetics: A Correspondence.” *Science* 293 (5532): 1103–5.
Wu, Xiaolin, Bo Cao, Patricia Aquino, Tsu-Pei Chiu, Chao Chen, Susu Jiang, Zixin Deng, et al. 2020. “Epigenetic Competition Reveals Density-Dependent Regulation and Target Site Plasticity of Phosphorothioate Epigenetics in Bacteria.” *Proceedings of the National Academy of Sciences of the United States of America*, June. https://doi.org/10.1073/pnas.2002933117.
Xue, Bingkan, and Stanislas Leibler. 2018. “Benefits of Phenotypic Plasticity for Population Growth in Varying Environments.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (50): 12745–50.
Yang, Chen, Justin Chu, René L. Warren, and Inanç Birol. 2017. “NanoSim: Nanopore Sequence Read Simulator Based on Statistical Characterization.” *GigaScience* 6 (4): 1–6.
Yang, Yu, and Jun Hang. 2013. “Fragmentation of Genomic DNA Using Microwave Irradiation.” *Journal of Biomolecular Techniques: JBT* 24 (2): 98–103.

Appendices

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Georgia Breckell
Name/title of Primary Supervisor:	Dr Olin Silander
In which chapter is the manuscript /published work:	2

Please select one of the following three options:

☑ The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Breckell G, Silander OK. Complete Genome Sequences of 47 Environmental Isolates of Escherichia coli. Microbiol Resour Announc. 2020;9(38):e00222-20. Published 2020 Sep 17. doi:10.1128/ MRA.00222-20

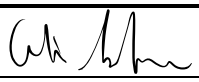
☐ The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate:

- Describe the contribution that the candidate has made to the manuscript/published work:

☐ It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Georgia Breckell
Date:	23/5/22
Primary Supervisor's Signature:	
Date:	23/5/22

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Georgia Breckell
Name/title of Primary Supervisor:	Dr Olin Silander
In which chapter is the manuscript /published work:	3

Please select one of the following three options:

☐ The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

☑ The manuscript is currently under review for publication – please indicate:

- The name of the journal:
  Microbial Genomics

- The percentage of the manuscript/published work that was contributed by the candidate: 90

- Describe the contribution that the candidate has made to the manuscript/published work:
  She concieved the project, designed the experiments and analyses in collaboration with her supervisor Olin Silander. She performed most experiments and carried out the majority of the bioinformatic analysis and data curation with inputs from her supervisor Olin Silander. She wrote this manuscript in collaboration with her supervisor Olin Silander. Analysis and generation of Figures 5, S2, S7 and S8 was

☐ It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Georgia Breckell
Date:	23/5/22
Primary Supervisor's Signature:	
Date:	23/5/22

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Georgia Breckell
Name/title of Primary Supervisor:	Dr Olin Silander
In which chapter is the manuscript /published work:	4

Please select one of the following three options:

☐ The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

☑ The manuscript is currently under review for publication – please indicate:

- The name of the journal:
  Microbial Genomics

- The percentage of the manuscript/published work that was contributed by the candidate: 90

- Describe the contribution that the candidate has made to the manuscript/published work:
  She concieved the project, designed the experiments and analyses in collaboration with her supervisor Olin Silander. She performed most experiments and carried out the majority of the bioinformatic analysis and data curation with inputs from her supervisor Olin Silander. She wrote this manuscript in collaboration with her supervisor Olin Silander. Analysis and generation of Figure S1 and S8 was

☐ It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Georgia Breckell
Date:	23/5/22
Primary Supervisor's Signature:	
Date:	23/5/22

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.