Variability of rates of mutation and fitness decline during mutation accumulation in *Escherichia coli* isolates

by

Destina Mattrasingh-Williams

A thesis submitted to the Faculty of Science in partial fulfillment of the requirements for the degree of

Master of Science

in

Biology

Carleton University

Ottawa, Ontario

Abstract

During infectious disease outbreak investigations, mutation rates amongst lineages of clinical bacterial pathogens can be highly variable; what is classified as multiple outbreaks could indicate high genetic variation amongst descendants of a single outbreak event. Consequently, the best way to define the genetic boundaries of an outbreak cluster is currently unclear. Over 2720 generations of mutation accumulation on average, I explored mutation rate and fitness decline variation in nine clinical isolates of *Escherichia coli* and I found that there was high variation between, but less commonly within, genotypes. Genotypes could be generally be categorised by mutation rate and fitness decline variation between replicates as either: (1) non-mutator genotypes with low variation, non-mutator genotypes with high variation because of (2) (an) infrequent mutator replicate(s), or (3) mutator genotypes with high or low variation. My findings have important implications both for molecular epidemiology of bacterial pathogens and predicting evolution in pathogen pathways.

Key words

Single nucleotide polymorphism (SNP), fitness landscape, mutation accumulation (MA), whole-genome sequencing (WGS), epidemiological investigations, foodborne illness outbreaks, enterohemorrhagic *E. coli* (EHEC), extraintestinal *E. coli* (ExPEC)

Acknowledgements

To begin, I would like to offer my sincere appreciation and gratitude to my supervisor, Dr. Alex Wong. Thank you for the support, knowledge, and patience you have shown me during the course of this projects. You have been incredibly encouraging and intentional in all your interactions with me and all of your students. Whether I was knocking on your office door asking for troubleshooting assistance as soon as you arrived in the morning or emailing you to ask for an extension, you cared and were more than happy to help. I couldn't have asked for a better supervisor to encourage my academic growth in the field and I will always be grateful to you for being such a great mentor.

I would also like to thank my lab mates who made this learning experience what it was. Thank you, Amanda, for catching me up in numerous laboratory protocols that I performed and all the technical assistance you provided, it has been a great work environment because of you. Thank you, Kamya, Katie, Issac, and Laura, for your invaluable contributions along the way.

Thank you, Dr. Hinz for instructing me with the flow cytometer, competitive fitness assays, and creation of the reference strain, my project simply wouldn't have been possible without you. And thank you to the members of the Kassen lab, especially, Dr. Kassen, Dr. Shewaramani, and Partha for providing a workspace where I was welcome to learn and ask questions.

Thank you to the members of the Carrillo lab, Dr. Carrillo, Alexa, and Paul, for your assistance with the Nanopore platform and research directions. To Dr. Overhage and Mariam, I learned a great deal while working with you on our side project, I'm excited to see what is next for both of your research paths.

To Dr. Mullally, thank you for being a wonderful ally in the department and in my writing journey. You are an absolute inspiration to me and all of your students who continue with microbiology because of their experiences under your tutelage.

Finally, to my family and friends, thank you for being there, even when the topic seems completely foreign. To my mom, thank you for all of the sacrifices you made for us, and thank you for being our mom. To Robyn and Andre, thank you for getting me there and back, and to Alyssa for the editing help. To the Rodney's, Edward's, and A. Mattrasingh's, thank you for giving me another home. And to C, thank you for the editing, and for being my biggest fan.

Table of Contents

List of Tables

List of Figures

Non-standard Abbreviations

ω	Fitness
B	Competitor cell population
BPS	Base pair substitution
C	Total cell count
cDNA	Complementary DNA
CFIA	Canadian Food Inspection Agency
DAEC	Diffusely adherent *E. coli*
DAMP	Density-associated mutation-rate plasticity
ddNTP	Dideoxynucleotide
DGGE	Denaturing gradient gel electrophoresis
dN	Nonsynonymous mutation rate
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide
dS	Synonymous mutation rate
EHEC	Enterohemorrhagic *E. coli*
EIEC	Enteroinvasive *E. coli*
EPEC	Enteropathogenic *E. coli*
ExPEC	Extraintestinal *E. coli*
FISH	Fluorescence *in situ* hybridization
Helicos	Helicos Genetic Analysis System
HTS	High throughput sequencing
IPTG	Isopropyl β-D-1-thiogalactopyranoside
IS	Insertion sequences
LB	Lysogeny broth
LBA	Lysogeny broth agar
LFZ	Laboratory for Foodborne Zoonosis
LPS	Lipopolysaccharide

m	Total number of mutations
MA	Mutation accumulation
MGE	Mobile genetic elements
MITE	Miniature inverted repeat transposable
MMR	Mismatch repair
N	Total number of generations
NGS	Next generation sequencing
OLC	Ottawa Lab Carling
PacBio	Pacific Biosciences Real Time Sequencer
PCR	Polymerase chain reaction
PFGE	Pulse-field gel electrophoresis
PHAC	Public Health Agency of Canada
PPi	Pyrophosphate
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
s	Selection coefficient
sω	Scaled fitness
SBS	Sequencing-by-synthesis
sD	Average fitness effect of deleterious mutations
SF	Scaling factor
SMRT	Single-Molecule Real-Time
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
SSCP	Single-stranded conformational polymorphism
STEC	Shiga toxin (verotoxin)-producing *E. coli*
T	Cell population
TE	Transposable elements

Ud	Deleterious mutation rate
UPEC	Uropathogenic *E. coli*
UTI	Urinary tract infection
UTR	Untranslated region
v/v	Volume per volume
w/v	Weight per volume
WGS	Whole-genome sequencing
X-Gal	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside
YFP	Yellow-fluorescent protein

Chapter 1: Introduction

1.1 *Escherichia coli* and Foodborne Illnesses

Escherichia coli is one of the best characterized bacteria, and is largely used as a model system in bacterial genetics, molecular biology, and biotechnology (Schaechter, 2001). In Canada, the Canadian Food Inspection Agency (CFIA) and the Public Health Agency of Canada (PHAC) are part of an extensive international network of food safety research called PulseNet with an aim of achieving global standardisation for foodborne disease (Nadon *et al*., 2017). Environmentally, *E. coli* is known to inhabit both water and soil, where nutrients are limited. *E. coli* is also a minor inhabitant within the biota of the vertebrate gastrointestinal tract but may account for over 1% of the total number of bacteria in feces within the circumstances of disease caused by a pathogenic strain (Schaechter, 2001). Commensal *E. coli* strains are distinct from pathogenic strains that cause disease through the expression of virulence genes which code for cellular adherence, cell surface molecules, secretion, toxin production, invasion, and immune evasion (Finlay & Falkow, 1997).

Three *E. coli* pathotypes categories are capable of causing disease in healthy individuals: (i) enteric disease, (ii) urinary tract infections (UTIs), and (iii) sepsis and meningitis (Nataro & Kaper, 1998). Gastrointestinal (diarrhoeagenic) *E. coli* infections are further divided into six pathotypes groups based on specific interactive elements with eukaryotic cells: enteropathogenic *E. coli* (EPEC; infant diarrhea), enterohaemorrhagic *E. coli* (EHEC), enterotoxigenic *E. coli* (ETEC; traveller's diarrhea), enteroaggregative *E. coli* (EAEC), enteroinvasive *E. coli* (EIEC), and diffusely adherent *E. coli* (DAEC) (Nataro & Kaper, 1998). EPEC is the main cause of potentially fatal infant diarrhea in developing countries, and is defined by production of the outer membrane protein intimin (*eae*) which enables cellular adherence to enterocytes and lesions in the colon and subsequent colonization over normal colon biota (Donnenberg *et al*., 1993). Foodborne pathogens, EHEC, are a subgroup of Shiga toxin (verotoxin)-producing *E. coli* (STEC). EHEC virulence is partly defined by production of intimin, as well as production of the potent cytotoxins Stx1 and/or Stx2, which inhibit host cell protein synthesis. ETEC is known widely as traveller's diarrhea; these bacteria colonize the small bowel mucosa and are

partly defined by production of enterotoxins that inhibit sodium absorption and stimulate chloride secretion, leading to watery diarrhea. Alternatively, *E. coli* extraintestinal infections (ExPEC) are infections outside of the gastrointestinal tract, – including UTIs, sepsis, and meningitis. The most common ExPEC are UTIs caused by uropathogenic *E. coli* (UPEC) and associated with fimbriae and pili mediating adhesion, hemolysin toxin production, and aerobactin for iron uptake (Nataro & Kaper, 1998; Ohlsson *et al*., 2005). These pathotype groups are partially based on expression of different virulence factors, and thus a strain can belong to more than one group.

Specific *E. coli* strains associated with foodborne illness are also classified by serotype. However, this technique does not take pathogenicity into consideration (Orskov & Orskov, 1992). As explored in the modified Kauffman scheme (1944, 1947), *E. coli* serotyping is based on the surface-exposed lipopolysaccharide (LPS) and O (somatic), H (flagellar), and K (capsular) surface antigens (Edwards & Ewing, 1972; Johnson *et al*., 1996).

Within the vertebrate gut, even the large intestine, nutrients in the lumen and intestinal walls are abundant, and consequently bacteria are unlikely to starve. Strict anaerobes dominate the lower gastrointestinal tract, and this is why *E. coli*, a facultative anaerobe of the *Enterobacteriaceae* family, is more easily cultivated in the laboratory environment than the vertebrate gut. Laboratory conditions provide a (comparably) low-stress haven for *E. coli* which has evolved within the competitive environments of soil and the gut (Schaechter, 2001).

1.1.1 An Epidemiological Study of Outbreak Strains in Food Production Conditions

A simulated epidemiological study was conducted by (Markell, 2017) to provide new knowledge surrounding the effects of food production relevant field conditions on genetic divergence in *E. coli*. Field conditions provide an opportunity to observe realistic mutation rates, and the effects of bacterial selective pressures, that are comparable to those in food production environments. Additionally, the use of non-mutator lineages prevents the associated increased mutation rate and altered base substitution and indel biases (Lee

et al., 2012). Markell's use of foodborne outbreak-relevant strains - pathogenic strains with the cytotoxin genes removed - was essential in estimating the genetic divergence expected in common outbreak strains in the food production environment.

Markell (2017) inoculated field lettuce with three *E. coli* strains of serotypes relevant to foodborne illness outbreaks (Shiga-toxin removed) in two growth seasons, harvesting weekly and performing WGS for single nucleotide polymorphism (SNPs) analysis. It was hypothesized that few SNPs would accumulate throughout the field selection experiment. Interestingly, the observed 0111:NM *E. coli* strain OLC-682 showed significantly higher SNP-accumulation over the two seasons compared to the two strains inoculated in parallel, suggesting that OLC-682 may have an increased mutation rate. Notably, a federal outbreak investigation using Markell's 0111:NM strains' results would have established that the number of SNPs accumulated during the experimental timespan was attributed to two separate outbreak origins. This study demonstrated the need for further research towards our ability to cluster strains of *E. coli* as part of a single outbreak event based on SNP differences.

1.2 Mutations and Evolution

Mutations are described as any sudden change arising in genomic material of a living organism. Mutations are distinguished and characterized by their impact: effects of the mutational mechanism and "weight" of the mutation on the organism's survival. Unsurprisingly, a change to an organism's genetic code will give rise to genetic variation in the individual; this is a mutation's short-term effect. However, the culminating point of mutation lies within the long-term effects of this genetic variation on a population of organisms, where the mutation's interactions with natural selection and genetic drift brings about the evolution of a population.

Natural selection is better known as "survival of the fittest". Simply put, natural selection acts on a population through organismal survival; organisms that survive long enough to produce more offspring than their less-favourably adapted counterparts tend to grow in number (Halliburton, 2004). Genetic drift entails the impact of random sampling

and chance on the relative frequencies of genotypes in a population (Section 1.5) (Halliburton, 2004). Natural selection and genetic drift are the primary forces acting on genetic variation to drive evolution. Genetic variation can arise in multiple manners, namely, transposition, and spontaneous mutation (Section 1.4) (Kondrashov & Kondrashov, 2010a). Still, the most significant source of new genetic variation is spontaneous mutation. The types (Section 1.3) and the rate at which an organism acquires mutations (Section 1.7) directly contribute to the organism's fitness, and fitness is the ability of an organism to contribute to the gene pool of successive generations. Consequently, mutations drive evolutionary potential of an organism. Understanding, and possible manipulating, bacterial mutations provides hope for predicting evolutionary patterns that impact human health at large scales.

1.3 Sources of Mutations

Genetic changes arise from the introduction of mutations into DNA. Mutations can come about through two mechanisms: induced or spontaneous. As the name suggests, induced mutations are the result of an exogenous mutagen, either physical, biological (viruses and transposons), or chemical, which increases the likelihood of mutagenesis. Conversely, spontaneous mutations arise without exogenous mutagens, and as a result of errors or damage during conventional cellular processes, namely DNA replication and cellular growth (Rosche & Foster, 2000).

1.3.1 DNA Replication and Repair Systems

Spontaneous mutations are often the result of errors during DNA replication. A point mutation may occur during synthesis of a DNA strand if DNA polymerase mispairs, adds, or omits a nucleotide (Goodman & Tippin, 2000). In *E. coli*, Fijalkowska *et al*., (2012) found that spontaneous mutations occurred 20 times more often on the lagging strand than the leading strand. When a nucleotide addition or omission occurs, silent, missense, and nonsense mutations are potential genetic variations. Loss of normal protein functionality is possible. However, most knowledge of spontaneous mutations comes from

mutations that occur in nonessential genes during short-term evolution experiments in laboratory cultures (Drake, 1991).

Mutagenesis is greatly hindered by both the intrinsically high accuracy of DNA replication and DNA repair processes. In *E. coli*, the highly conserved biological pathway, the mismatch repair (MMR) system, is the primary corrective tool against replicative errors. MMR corrects DNA replication errors thereby preventing their permanence in future generations as well as playing a role in eliminating severely damaged cells, essential processes for cellular viability (Kunkel & Erie, 2005). Though MMR is highly conserved throughout the tree of life, MMR-defective strains are not uncommon in laboratory strains of bacteria and tend to be prominent in short- and long-term evolution experiments (LeClerc *et al*., 1996; Sniegowski *et al*., 1997; Wong *et al*., 2012).

1.4 Genetic Variation in Prokaryotes

There are a variety of mutations types which are classified based on the DNA (or RNA) sequence changes imparted. Broadly, prokaryotic genetic changes can arise from mutations, transpositional/mobile genetic elements (MGEs), and recombination– both chromosomal and genomic. In prokaryotes, the most observed types of homologous recombination events are transformation, conjugation, and transduction, each of which contributes to genetic variation (Darmon & Leach, 2014). Chromosomal mutations are changes in the number or structure of chromosomes, usually affecting blocks of genes at a time, whereas genomic mutations are changes to specific genes. Genomic and chromosomal mutations have some overlap; mutation types include base pair substitutions (BPSs)/single nucleotide polymorphism (SNPs), indels, and gene duplications.

1.4.1 Base Pair Substitutions (BPSs) - Single Nucleotide Polymorphism (SNPs)

Base pair substitutions result from the replacement of a single base pair for another (Friedberg *et al*., 2009). BPSs can occur as transversions, where a purine base (A or G) is substituted for a pyrimidine (C or T) or vice-versa, or transitions, where a purine base

replaces a purine or a pyrimidine base replaces a pyrimidine (Friedberg *et al*., 2009). Due to the similar biochemical structures of the base couples, transition mutations are more common than transversions. Bases that are structurally and biochemically similar are more easily substituted for one another and less likely to be recognised and repaired as DNA lesions.

The phenotypic effect of a BPS in a coding sequence can also be used for classification (Friedberg *et al*., 2009). A BPS is a silent or synonymous mutation when the DNA sequence change encodes the same amino acid, which theoretically would not impact protein function. Alternatively, a BPS is a missense or nonsynonymous mutation when the DNA sequence change encodes a different amino acid, the effects of which vary greatly as the resulting protein may maintain, lose, gain, or change function. Finally, a BPS that encodes an amino acid codon change to a stop codon is a nonsense mutation or stop mutation. As the name suggests, a stop mutation is generally deleterious and will produce a truncated protein with complete or partial loss of function. It is well-established that mutations which disrupt coding sequences (and nonsynonymous mutations) tend to have more deleterious effects than those that disrupt noncoding sequences (and synonymous mutations). Nonsynonymous mutations can generate: (1) non-functional proteins, with the large fitness declines, from nonsense base pair substitution mutations and coding indels, and (2) modified proteins from missense base pair substitution mutations. Synonymous mutations do not alter protein sequences and conversely rarely have a significant impact on fitness.

Single nucleotide polymorphisms (SNPs) can be localised within coding sequences, introns (in eukaryotes), or intergenic regions (Aerts *et al*., 2002). SNPs arise spontaneously and only affect a single residue unlike many indels. Retention of non-coding, synonymous, or nonsynonymous mutations with little protein impact within bacterial populations is more likely due to low selective pressure. Certain synonymous and nonsynonymous mutations may be selected for when beneficial. This is further emphasized by the, contextually small, codon biases in bacteria (Ochman, 2003). Because of this, SNP accumulation in a population is dependent upon both the strain's inherent spontaneous mutation rate in

addition to the selective pressures acting to fix or remove the particular SNP from the population.

1.4.2 Indels

Indels account for both insertions and deletions of one or more nucleotide bases; unless in a multiple of three, within coding regions indels will typically result in a frameshift mutation and subsequent non-functional or truncated protein (Friedberg *et al*., 2009). Lee *et al*. (2012) noted that small indels, ≤4 bp, were the majority of the total observed indels in a mutation accumulation (MA) experiment with both *E. coli* wildtype and MMR-deficient strains. Furthermore, the majority of indels identified were single base pair indels localized within sequence repeat "hotspots".

As summarized by Halliburton (2004), an individual gene, piece of a chromosome, or an entire chromosome can be duplicated. In gene duplication, one copy may continue to serve its original function while the other copy may produce new functions; this is thought to have been a cornerstone of bacterial evolution "leaps". Throughout evolution, though infrequently, an entire prokaryotic chromosome has been duplicated leading to new genetic material and sometimes new cellular function.

1.4.3 Transposition - Mobile Genetic Elements (MGEs)

MGEs increase genetic variation as DNA segments that encode for their ability to move and integrate elsewhere in the genome. MGEs include transposable elements (TEs) (Campbell *et al*., 1979), plasmids, and bacteriophage elements (Miller & Capy, 2004). MGEs are difficult to categorize and have a variety of potential effects upon integration into new genomic regions. Insertion can enable the spread of virulence factors and resistance amongst bacteria, but can also result in deleterious or lethal insertions into important genes. TEs are the most abundant and studied MGEs. *E. coli* contains two potential TE pathways: replicative and conservative (Derbyshire & Grindley, 1986). The replicative pathway retains one copy of the TE at the original site and another at the

destination site (Shapiro, 1979). Conversely, the conservative pathway acts as a "cut and paste" mechanism, wherein the TE is excised from one site and integrated into another (Darmon & Leach, 2014). Class II TEs, or DNA transposons, are mainly conservative, and include transposons, insertion sequences (IS), and miniature inverted repeat transposable elements (MITEs) (Darmon & Leach, 2014; Treangen *et al*., 2009).

Transposons are large autonomous TEs, > 2000 bp, containing genes to mediate their own excision and reintegration movements as well as separate genes for virulence or resistance (Darmon & Leach, 2014; Treangen *et al*., 2009). Autonomous TEs mediate their own movement whereas non-autonomous TEs cannot do so (Darmon & Leach, 2014). Like plasmids, transposons are believed to have played a crucial role in bacterial adaptation to environmental antibiotic exposure (Blot, 1994). IS elements are the most abundant and simplest form of autonomous TEs, between 700 – 3500 bp, containing only the open reading frame that encodes mobility proteins (Darmon & Leach, 2014). MITEs are nonautonomous TEs, between 100 – 400 bp long, thought to be derivatives of IS elements through internal deletions (Delihas, 2011).

Mutation accumulation studies have provided previously underestimated insight on the frequency and significance of IS elements in evolution. Lee *et al*. (2016) used a mutation accumulation- whole-genome sequencing study to demonstrate the prevalence of ISs across 520 *E. coli* lines. It was shown that the rate of IS elements insertion is relatively comparable to the rate of base pair substitutions per genome per generation in *E. coli* at 3.5×10^{-4} per genome per generation and $\sim 1 \times 10^{-3}$ per genome per generation respectively (Foster *et al*., 2015b; Lee *et al*., 2014, 2016).

1.5 Genetic Drift and Evolution

Genetic drift is random allele frequency variation between generations (Halliburton, 2004), and is an important process in evolution. Unlike natural selection, genetic drift is able to affect all genetic variations, not just adaptive or deleterious. There are four main aspects of genetic drift: (1) Genetic drift does not have a predictable direction and the allele frequency is equally likely to increase or decrease with each generation. (2)

The long-term effect of genetic drift is the reduction of genetic variation *within* a population, (3) but increased variation *among* populations. (4) The strength of genetic drift is population-size dependent.

A "sample" is the assortment of alleles that persist into the next generation. Populations are finite (random) samples of the alleles within the parent generation, and genetic drift acts through repeated random sampling of gametes from a population. Based on the principles of random sampling, large sample frequencies will more closely resemble the original population, leading to small changes over long periods of time (Wright, 1940). Across generations, allele frequency changes in large populations are difficult to observe. For example, mutations that have a minor advantage are more common but more likely to become overwhelmed before fixation. The short-term effects are generally minor unless the drift is within a very small population; smaller populations have larger allele frequency changes between generations, and genes that would otherwise be selected against in a larger population are maintained. The sample population is eventually reconstituted with offspring, matching the sample frequency, to match the original population size, and based on probability theory, the population will eventually contain one allele.

Bacterial populations are particularly useful for the study of random sampling and population sizes as mutation (and not mating of populations) is the only source of genetic variation between a population's generations. However, because bacterial populations in are relatively large, genetic drift can be difficult to experimentally observe.

1.6 Mutation Detection

Prior to the wide availability of sequencing, there were three approaches for mutation detection: (i) hybridization-based methods, (ii) enzyme-based methods, and (iii) methods based on physical properties of the DNA (Cotton, 1993). Differential hybridization of complementary DNA probes and mutant DNA is the basis of hybridization-based methodology (Rapley & Harbron, 2012). Hybridization techniques include DNA microarrays which simultaneously test SNP sites using thousands of probes (Southern, 2001), and fluorescence *in situ* hybridization (FISH) (Röscheisen *et al.*, 1994).

Hybridization-based methods are often accompanied by cross hybridization between probes (Rapley & Harbron, 2012). Enzyme-based molecular methods such as restriction fragment length polymorphism (RFLP) analysis of non-coding regions and microsatellite regions (De Visser *et al*., 2004) have predominantly been used in mutation detection. Denaturing gradient gel electrophoresis (DGGE) and single-stranded conformational polymorphism (SSCP) assays (Cotton, 1993) are mutation detection methods that use physical properties of DNA. Genomic restriction enzyme digestion and subsequent fragment length mapping by pulse-field gel electrophoresis (PFGE) is both enzyme- and physical-based; PFGE is a common strategy for comparing mutated DNA (Schwartz & Cantor, 1984).

These methods, used to detect less obvious deletion/insertion mutations, provide a limited mutational spectrum either by focussing on specific genetic loci or by showing the physical impact of mutational variation rather than the genetic changes themselves. It is now possible to directly identify mutational variation down to the nucleotide level across the whole genome.

1.6.1 Genome-wide Mutation Detection – Whole-Genome Sequencing (WGS)

Before direct sequencing was possible, mutations were indirectly inferred through phenotypically- and mathematically-based experiments. The pioneering spontaneous mutation rate experiments of Mukai (1964) and Mukai *et al*. (1972) recorded the effects of spontaneous and induced chromosomal mutations using fitness declines and lethal observations in *D. melanogaster*. Today, a "complete picture" of a genome's mutation spectrum can now be found by direct re-sequencing. When compared against the sequenced ancestral genome, the acquired mutations in an evolved population are identifiable as dissimilar. WGS allows for differentiation between evolved populations at the base pair level, making it an ideal method for SNP analysis (Bentley, 2006).

Sequencing is the ordering of nucleotides within a segment of DNA. Traditional sequencing, now referred to as first generation sequencing, was derived from the Sanger method (Sanger *et al*., 1977). In this chain-termination method, template DNA is the

scaffold for a complementary strand of DNA (cDNA). This synthetic strand is comprised of fluorescently labelled 2′-deoxynucleotides (dNTPs) and 2′, 3′-dideoxynucleotides (ddNTPs) (Sanger *et al*., 1977). This method is low throughput but highly accurate. The Maxam Gilbert method was also commercially accepted as a first-generation sequencing method, but Sanger method was more widely used. Sanger sequencing enabled scientists to determine the human genetic code. The "shotgun technique" was used to analyse longer reads This method is now automated and still essential for sequencing short DNA segments (Anderson *et al*., 1981).

The development of high throughput sequencing (HTS), which is less expensive, and less time-consuming, was the subsequent step. While the Sanger method sequences DNA fragments one-by-one and with high cost per base, HTS technologies simultaneously sequence mass fragments in parallel at substantially lower costs. The rise of next generation sequencing (NGS) platforms has led to vast opportunities in the field of HTS, both targeted (e.g., exome-wide) and genome-wide, through the production of fast, increasingly reliable, and raw genomic data. In contrast to whole-exome sequencing, WGS detects both coding and non-coding regions of the genome, as well as more efficient detection of regulatory regions (e.g., untranslated regions with potential regulatory roles such as non-coding RNAs and transcription binding sites) and copy number variation (Poduri *et al*., 2013; Stavropoulos *et al*., 2016).

NGS platforms are commonly grouped together based on method and order of invention. Although there are no universal NGS platform categories, we can consider three NGS platform methods categories: (i) second-generation, (ii) third-generation, and most (iii) most recently, fourth-generation platforms (Bentley, 2006).

Sanger, or first-generation, sequencing was the reigning gold standard of WGS for almost four decades; it sequences DNA fragments one-by-one. In opposition, second-generation sequencing allows for DNA clonal amplification which involves parallel sequencing of billions of DNA fragments. Compared to first-generation sequencing, parallel sequencing produces considerably shorter reads but provides significantly increased sequencing throughput and speed with decreased costs. Using these methods, the entire Human Genome Project, which required 13 years of international efforts and $2.7

billion, might have been sequenced in a week for a few thousand dollars (Gullapalli *et al*., 2012). Second-generation sequencing technology was first commercially introduced as Roche 454 pyrosequencing, followed by the Illumina Solexa (http://www.illumina.com/) and then the Sequencing by Oligonucleotide Ligation and Detection (SOLiD) platforms.

In 1993, Nyrén *et al*., pioneered sequencing-by-synthesis (SBS) in which DNA polymerase cleaves a detectable pyrophosphate (PPi) when a nucleotide is incorporated. Of the short-read platforms, many DNA polymerase-dependent formats of SBS exist, most notably are the 454 pyrosequencing, Illumina, and Ion Torrent (Roberts *et al*., 2013). While both use emulsion polymerase chain reaction (PCR) (Kondrashov & Kondrashov, 2010). to amplify DNA, 454 uses light detection, and Ion Torrent uses hydrogen ion detection when a new nucleotide base has been incorporated (Ronaghi *et al*., 1998; Rothberg *et al*., 2011). The HiSeq and MiSeq platforms launched by Illumina remain the most widely adopted NGS methods. Illumina's bridge amplification method creates primer hybridization clusters between identical sequences on template DNA; sequencing can then initiate from primers at both ends to form double-stranded "bridges" between primers during paired-end sequencing (Bentley *et al*., 2008; Schadt *et al*., 2010). Illumina detects nucleotide bases using fluorescently labelled, reversible dye-terminators (Bentley *et al*., 2008). The SOLiD platform detects nucleotide bases through sequencing-by-ligation (SBL), which uses PCR, fluorescent probes, and fluorescence imaging (Valouev *et al*., 2008).

Despite ongoing and widespread usage, the weakness of second-generation sequencing lays in the biased-nature of PCR amplification step. Third-generation sequencing removes the need for additional hands-on work by performing real-time DNA analysis to produce generally longer reads, a significant improvement for WGS assemblies. The first commercially released (but no longer operational) third-generation platform was the Helicos Genetic Analysis System (Helicos) by Helicos Biosciences (www.seqll.com/), and finally Single-Molecule Real-Time (SMRT) technology by Pacific Biosciences Real Time Sequencer (PacBio) (www.pacb.com/). Helicos first combined single-molecule sequencing based on fluorescent detection and SBS. SMRT sequencing was the first single-molecule sequencer with real-time detection. SMRT cells consist of millions of zero-mode

waveguides which detect (in real-time) an individual nucleotide by its a base-specific phosphate-label upon cleavage during nucleotide incorporation onto a growing DNA fragment (Churko *et al*., 2013; Eid *et al*., 2009)

Fourth-generation sequencing, like SMRT technology, is real-time single-molecule sequencing and a long-read platform with no amplification step. Its distinguishing feature is sequencing without synthesis (Mignardi & Nilsson, 2014). Though recently commercialized, fourth-generation sequencing was first conceptualized by David Deamer in 1989 (Deamer *et al*., 2016). Finally, in 2014, Oxford Nanopore (www.nanoporetech.com/) commercially released its first nanopore sequencing device, the MinION (Jain *et al*., 2015). Nanopore technology measures the change in electrical current across a nanopore (a pore of nanometer size) as a DNA molecule electrophoreses through the pore. As an exonuclease restricts the passage of nucleotides to single file, the change in current correlates to size and order of nucleotides that have traversed through the pore (Churko *et al*., 2013). MinION is both economically, portably and HTS-advantageous. The small size of the MinION device has permitted its usage in rapid surveillance of epidemics such as Ebola and Zika viruses as well as usage in outer space (Castro-Wallace *et al*., 2017).

In general, sequencing data produced from NGS platforms requires four levels of nucleotide sequence analysis (Kulski, 2016).

1. The platform's software will convert the raw signals acquired in sequencing into base calling. This outputs nucleotide sequences and associated quality scores.
2. Sequences, short and/or long, are aligned to make long continuous assemblies (contigs and scaffolds) in either of two ways: (i) reference-based mapping, or (ii) *de novo* assembly of overlapping reads. Single-read, paired-end, and/or mate-pair sequencing can be performed using the above sequencing platforms (Pushkarev *et al*., 2009). Single-read sequencing occurs at only one end of a linear nucleic acid fragment whereas paired-end sequencing occurs from both ends (Nagarajan & Pop, 2013). Paired-end sequencing is ideal in reference-based mapping and identifying large structural variations such as inversions using shorter nucleic acid fragments, typically below 800 kb (Churko *et al*.,

2013). Small fragments are typically assembled in *de novo* and whole-genome resequencing. Separately, mate-pair sequencing is a library construction method that is ideal for *de novo* genome assembly; without a fully annotated reference genome, mate-pair sequencing is often able to decrease gap regions and extend scaffold length (Reinhardt *et al*., 2008). In mate-pair sequencing larger fragments, typically between 2 kb to 5 kb, are circularized and then re-fragmented to create mate-pair reads (Nagarajan & Pop, 2013). According to Bradnam *et al*. (2013), it is recommended not to trust the results of any one assembly or associated quality scoring; a chosen assembler should be able to sufficiently provide coverage, continuity, and error-free based specific for the area of study.

3. For (optional) sequencing interpretation, the assembled sequences must then be annotated, transcribed, and translated into a user-friendly format. This includes differentiating between genomic regions – coding and noncoding (5' noncoding, and 3' terminal ends) regions, and untranslated regions (UTRs) – which allows the assembled sequences to be visualized.
4. Lastly, all the data obtained from different NGS platforms is assembled into a single, readable, bioinformatic output with tools for data representation manipulation. Bioinformatics analysis is a major bottleneck step in using sequencing data for outbreak prevention studies.

Shotgun sequencing entails cloning smaller sequences with overlapping regions and then assembling contigs based on the overlaps, essentially using short-read technologies to assemble a large genome *de novo* (Anderson *et al*., 1981). In a genome-resequencing project, reads from evolved populations can then be aligned to this reference genome in whole-genome resequencing to identify single nucleotide changes (SNPs, SNVs, and mutations) and copy number variants (Churko *et al*., 2013). Nanopore long-reads have proven to provide a strong reference framework for short read alignments (Chrystoja & Diamandis, 2014). WGS researchers still struggle to detect sequencing and amplification errors, obtain adequate coverage in regions of extreme GC/AT content, and assemble short-reads into large variants (Churko *et al*., 2013). Like increasing a camera's resolution power and taking pictures of the same object from multiple angles, by taking a

hybrid approach to sequencing (where two or more sequencing platforms are used for one assembly) and new approaches to software algorithms, high throughput WGS for food safety investigations shows great promise for calling SNPs.

1.6.2 WGS for SNPs in Food Safety Investigations

Bartels *et al*. (2014) established that WGS (using MiSeq) could successfully replace single-gene PCR amplification and Sanger sequencing in traditional *spa*-typing (sequencing of the *Staphylococcus* Protein A gene) of Methicillin Resistance *Staphylococcus aureus* (MRSA) outbreak strains; the study found a 97% agreement between *spa*-types obtained by the two methods. Furthermore, it was concluded that any *spa*-type discrepancies between the two methods would be inconsequential in an outbreak investigation because any epidemiologically linked isolates analysed under WGS would be identified as such regardless of the one gene. For traditional infection control testing of Danish patients suspected to be part of an MRSA outbreak, WGS and SNP analysis is routinely used as confirmation for MRSA isolates with related *spa*-types. WGS is more comprehensive alternative to single-gene sequencing or genotyping for a series of characterised mutations. However, at this time no standardized quality control for microbial WGS results exists.

Another prominent method, PFGE, responds more to indel variation than SNPs for profile contributions (Kudva *et al*., 2002). Because discrimination power is essential for SNP analysis, and because SNPs are less prevalent than indels, a mutation detection method that discriminates based on base pair variation is essential. Conversely, WGS is able to differentiate between microbial strains, with high resolution, down to one SNP, which is essential for MA tracking. Numerous studies have already used WGS to identify mutation accumulation (Conrad *et al*., 2011). In a wild-type *E. coli* investigation, Lee *et al*. (2012) used the combined MA-WGS strategy to estimate mutation rate and successfully obtained their most inclusive estimate of the respective genome-wide mutation rate.

WGS data can be stored in a raw form without the influence of analytical biases, but this output involves large amounts of data that require largescale storage space. This

allows for a database of raw genomic data from outbreak investigations that can be reanalyzed with improved tools and evolutionary comparisons. With constantly improving software and increasing understanding of mutation rates, re-evaluating genomic samples can help improve interpretations of the genetic boundaries surrounding an outbreak cluster.

1.7 Spontaneous Mutation Rates

Mutation rates can be measured for substitution rates or spontaneous mutation rates. Here, the substitution rate refers to the rate of mutation fixation in a population, whereas spontaneous mutation rate is the inherent rate at which new mutations arise per genome, per generation in an organism (Barrick & Lenski, 2013).

Spontaneous mutation rates are highly variable across species (Paul D. Sniegowski *et al*., 2000). RNA viroids, which have the simplest genomes, also have the highest spontaneous mutation rates, while humans and other "higher eukaryotes" have the lowest genomic spontaneous mutation rates (Gago *et al*., 2009). Generally speaking, simple organisms with small genomes tend to have higher spontaneous mutation rates and organisms with larger genomes tend to have lower spontaneous mutation rates (Gago *et al*., 2009). Spontaneous mutation rates are also not fixed for an organism and have been experimentally shown to vary greatly with environmental conditions, cellular stressors, and antibiotic exposure; this phenomenon is known as mutation rate plasticity (Conrad *et al*., 2011; Krašovec *et al*., 2017). Increased population density, for example, promotes competition for resources amongst bacteria growing in liquid, and this has been experimentally demonstrated to correlate to lower mutation rates (Krašovec *et al*., 2017). This density-associated mutation-rate plasticity (DAMP) has been observed throughout the domains of life (Krašovec *et al*., 2017). In fact, any factor that can impact the balance between mutagenesis and DNA repair can in turn impact the mutation rate (Krašovec *et al*., 2017). Still, spontaneous point mutations have been summarized to occur at a rate of 10^{-10} to 10^{-9} mutations per nucleotide per generation for various bacteria and growth conditions (Schroeder *et al*., 2018)

On the other hand, there may be individuals or lineages within a population that have a comparably increased spontaneous mutation rate; these are called mutators (Horst, 1999). Mutator phenotypes are typically the result of mutations within DNA repair genes, leaving new mutations unregulated and in disrepair; mutators naturally arise in populations both environmentally and in laboratory (Horst, 1999). The disadvantage of an increased spontaneous mutation rate is simple; unregulated mutation accumulation can disrupt essential cellular processes and typically leads to fitness decline from accumulated deleterious mutations (Section 1.9). However, an increased spontaneous mutation rate also increases the mutator strain's chances of acquiring an advantageous mutation (Section 1.8.3).

1.7.1 Scientific Contributions to the Study of Spontaneous Mutation Rates

Luria and Delbrück (1943) first attempted to experimentally separate the effects of mutation and selection through the Fluctuation Test (Section 1.7.2.1). This attempt introduced two key definitions: (1) mutation rate is the rate that mutations are passed between generations by an individual, and (2) mutation frequency is the relative proportion of mutated cells within a population. Therefore, mutation frequency depends on *when* a mutation arose during the population's growth as this affects proliferation of the mutation in subsequent mutant generations. The mutation fraction is then made up of both pre-existing mutants that exponentially proliferate, as well as the evolution of new mutants. Based on this understanding, it would be incorrect to use mutation frequency to calculate spontaneous mutation rates. As such, fluctuation tests attempt to statistically infer the mutation rate from mutation frequencies in multiple replicate populations.

One of the classic ways to estimate spontaneous mutation rates is the mutation accumulation (MA) experiment (Section 1.7.2.3); the founding MA studies were conducted in *Drosophilia melanogaster* (Bateman, 1959; Mukai, 1964; Mukai *et al*., 1972). MA experiments require maintenance of the population at an extremely low size to minimize the effects of natural selection, effectively enabling genetic drift (Section 1.5). In MA experiments, natural selection cannot effectively purge (non-lethal) deleterious mutations from the population; non-lethal mutations consequently amass genome-wide at

the rate by which they appear. Mukai's experiments found a mutation rate of one per individual per generation with a low mutational effect, <3% on average. Many population-genetic models were based upon and supported the findings of these *D. melanogaster* MA experiments (Lynch *et al*., 1999; Lynch & Walsh, 1998). Drake (1991) provided much more generalized but heavily cited mutation rate of 3-4 x 10^{-3} per genome per generation for all DNA-based microbes, concluding that an organism's cellular mutation rate is generally near-constant. However, this mutation rate was based on seven taxa, four of which were bacteriophages, and utilized reporter genes, making it a MA analysis of reporter loci rather than the genome as a whole. Since then, reporter loci have been largely discredited as being genomic representatives (Katju & Bergthorsson, 2018). The most inclusive estimate of genome-wide mutation rate following Drake's study was found in the first MA-WGS study of MMR-deficient *E. coli* by Lee *et al*. (2012). This spontaneous mutation rate was approximately threefold lower than Drake's value. Lee's combined MA-WGS method of estimating mutation rates greatly improved mutation rate and spectra accuracy; prokaryotes studied since have generated a growing database of naturally accumulated mutations with spontaneous rates of base substitution ranging between 7.9 x 10^{-10} and 2.34 x 10^{-8} per site per generation (Katju & Bergthorsson, 2018). MA-WGS experiments have also opened the door for both genetic and fitness analyses of mutations. Heilbron *et al*. (2014) were the first to use this method to also characterize fitness in evolved lineages of mutator bacteria with mutation biases. Before this, understanding of the fitness effects of spontaneous mutations was largely through indirect analysis (Eyre-walker & Keightley, 2007). The role that spontaneous mutations play in evolution can only be properly analysed using both their rate and fitness effects.

1.7.2 Measuring Spontaneous Mutation Rates

High variability across measurements of spontaneous mutation rates thus far necessitates an accurate and reproducible experimental method going forward (Williams, 2014). This may be achieved by accounting for large quantities of mutations accumulated over a large number of generations in an unbiased manner and without selective pressures.

Three well-established approaches estimating spontaneous mutation rate will be evaluated in this section.

1.7.2.1 Relative Rates – Fluctuation Assays

Luria & Delbrück (1943) pioneered the use of fluctuation assays, based on the measurable phenotype of mutants, to study mutation rates in microbes. Fluctuation assays involve inoculating and growing several parallel cultures to saturation. Cultures are initially enumerated on non-selective media to allow spontaneous mutagenesis, followed by selective solid media to detect specific mutants that arose. Statistical analysis yields an estimated mutation rate required for the observed mutants on selective media (Williams, 2014).

Survival on selective media is based on a handful of loci; thus, many locus-specific assumptions are made in this phenotypically-rooted assay (Williams, 2014). Since the origin of fluctuation assays, reporter loci have been largely discredited as being genomic representatives (Katju & Bergthorsson, 2018). Additionally, using the reversion of a mutant genotype or estimations based on antibiotic resistances is acknowledged to be highly misleading and typically an underrepresentation of genome-wide mutagenesis (Katju & Bergthorsson, 2018; Lee *et al*., 2012). However, fluctuation assays are still useful today for genotype comparisons, particularly when comparing mutation rates between samples.

1.7.2.2 Inferred Rates – Phylogenetic Analysis

Comparative genomic approaches use the differences between DNA sequences of naturally occurring populations to calculate mutation rate at selectively neutral sites (Fu, 1994). It is a particularly useful method in estimating mutation rates of specimens that cannot be laboriously measured, such as extinct specimens. However, this approach requires estimates of generation times in nature, and is not always accurate since phylogenetic assumptions are made about codon placements being neutral with respect to fitness (Dettman *et al*., 2016; Drake, 2012)

1.7.2.3 Directly Detected Rates – Mutation Accumulation (MA): Genetic Drift in a Small Population

Mutation accumulation (MA) experiments allow mutations to accumulate in a neutral and unbiased manner under conditions that minimize the effects of natural selection but reveal the effects of genetic drift; mutations are later accounted for without being specifically selected for/against (Halligan & Keightley, 2009). Reducing the effects of natural selection on new mutations promotes their retention throughout the long-term evolutionary experiment, alternatively, mutations would be selected for or against and this bias would impact the observed mutation spectra (Halligan & Keightley, 2009). Thus, most (non-lethal) mutations accumulate genome-wide and irrespective of fitness effect. After many generations, the evolved lineages can be compared to ancestral lineages for genomic sequences, fitness, and cellular viability (Halligan & Keightley, 2009). Fitness decline is almost inevitable following the accumulation of slightly deleterious mutations over thousands of generations without the purging influence of natural selection (Section 1.8.1).

Specifically, the effects of natural selection are reduced through nonselective growth conditions and maintenance of small population sizes. In a culture of microbes, population size can fluctuate by substantial numbers, making it difficult to maintain the small population sizes needed for a successful MA experiments (Halligan & Keightley, 2009). To achieve this, MA experiments passage replicate lineages through serial single-cell population bottlenecks (**Figure 1**) (Halligan & Keightley, 2009). However, it is not feasible to bottleneck bacteria at each generation because of their rapid generation time and small size (Trindade *et al*., 2010). Randomly selecting for the bottlenecked individual rescues it from potentially competing with more genetically fit population members. Although it is possible to miss rare and/or interesting mutations with this approach, it is ideal for a broad and comprehensive study of spontaneous mutation rates (Foster *et al*., 2015a).

The effects of selection will be strongest during cellular exponential growth phase. Furthermore, populations must be allowed to grow to a threshold size so that each new mutation has a continuous likelihood of occurring. Consistency is essential in MA experiments to passage populations outside of strong selective events.

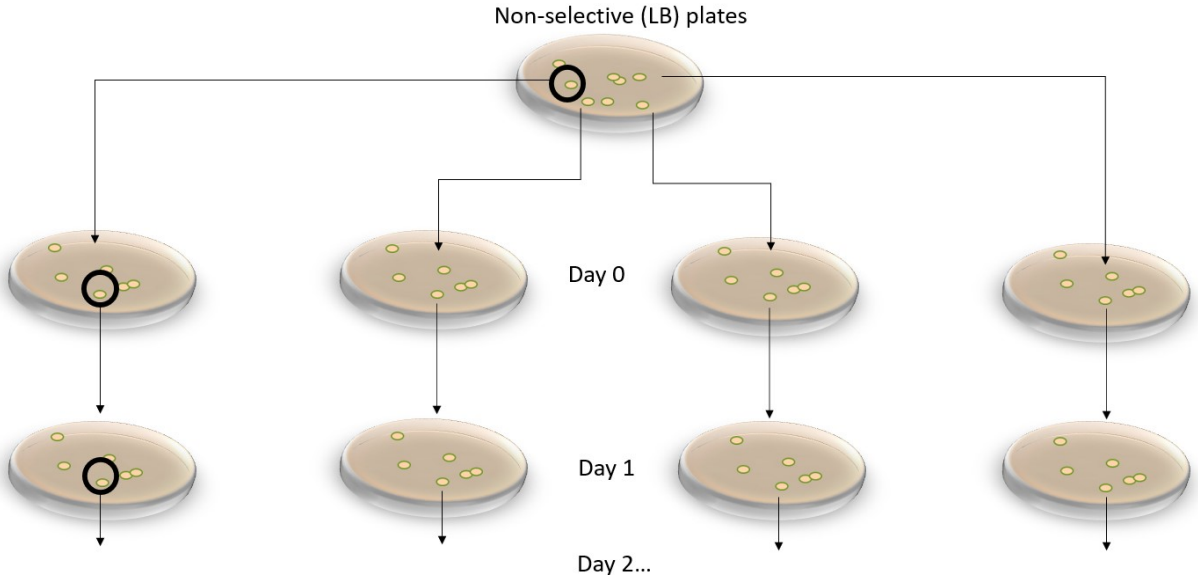


Figure 1. Mutation accumulation (MA) experimental depiction for bacteria. In MA experiments with bacterial species, a single colony of the ancestral lineage on nonselective agar is randomly selected to create MA lineage replicates (Day 0). Serial population bottlenecks are then performed by randomly passaging a single colony of the MA lineage replicate onto nonselective agar to establish new populations with genome-wide accumulated mutations (Day 1, Day 2…). This method minimizes the effects of natural selection and emphasizes the effects of genetic drift on the small population.

1.7.3 Interest in Clinical Applications of MA Experiments

When introduced into the laboratory environment, wild bacterial isolates often undergo selection for more efficient growth under new, less stressful, environmental conditions (Liu *et al*., 2017; Ronald *et al*., 2006). Newly isolated environmental *E. coli* isolates have been suggested to accumulate more mutations than laboratory isolates during the domestication process. Both, however, can acquire mutator mutations that "hitchhike"

with laboratory environment-beneficial mutations (Liu *et al*., 2017). These isolates differ genetically, and potentially phenotypically (Liu *et al*., 2017). Ronald *et al*. (2006) proved this concept in a loci-specific comparison experiment between three strains of yeast, introduced into the laboratory environment from wild strains in the 1950s, isolated from clinical origin in 1989, and isolated from environmental origin in 1996. It was shown that the newest environmental isolate demonstrated the strongest evidence of rapid protein evolution. Additionally, the laboratory strain removed from the environment longer than the other two evolved as a slow-to-intermediate rate relative to the two natural isolates. Laboratory domestication is a process of mutation and selection. This exemplifies the need for MA experiments using natural isolates as opposed to laboratory strains for the purpose of mutation control. This is the best representation of field or clinical spontaneous mutation rate data and it is necessary for reliable data applications in foodborne illness outbreak investigations.

Mutation rates can vary greatly amongst bacterial species and strains. Pathogenic bacterial strains have also been shown to have higher mutation rates than their non-pathogenic equivalents (Wirth *et al*., 2006). This makes it important for MA experiments using clinical isolates to represent pathogenic bacteria as closely as possible, for example, using pathogenic isolates lacking toxin genes.

1.8 Possible Fitness Effects of Mutations

Fitness is the measure of reproductive success of an organism. Mutations can be described by their effect on fitness: deleterious, neutral, or advantageous. It is largely suggested by evolutionary biologists that the distribution of mutation effects is bimodal between neutral and strongly deleterious (Eyre-walker & Keightley, 2007; Zeyl *et al*., 2007). A single mutation alone may have a large affect, depending on the genomic location and growth conditions, though most non-lethal spontaneous mutations have a small effect (Eyre-walker & Keightley, 2007; Halligan & Keightley, 2009; Heilbron *et al*., 2014). When more than one mutation is present the effects accumulate and the resulting fitness can surpass or reduce the fitness effect of either mutation alone. Traditionally, fitness of a population is described relative to another; this is called relative fitness mapping. Relative

fitness mapping sets the standard, or reference, fitness at ω = 1. If the marginal fitness of a mutation is less than one, the mutation worsened fitness. Conversely, if the marginal fitness of a mutation is more than one, the mutation increased fitness.

1.8.1 Deleterious Mutations

Considering most mutations have an insignificant impact on fitness, the majority of mutations that *affect* fitness are deleterious. This has been demonstrated across MA experiments in small populations (Lande, 1994; Lynch *et al*., 1995; Schwander & Crespi, 2009). This bias towards disadvantageous mutations over advantageous does suggest that a gradual "mutational meltdown" is likely inevitable for populations where the effects of natural selection are minimized, such as populations in MA experiments (Sniegowski & Lenski, 1995).

Mukai *et al*. (1972) demonstrated that minor deleterious mutations are more likely than lethal deleterious mutations through a large-scale chromosomal extraction experiment in *Drosophilia melanogaster* where the average deleterious fitness reduction from a single non-lethal (homozygous) mutation was 4-8%. Though organisms such as *C. elegans* prove to have comparably high average fitness effects from single deleterious mutations, most MA experiments have supported the general finding that most deleterious mutations impart fitness declines of only a few per cent or less (Drake *et al*., 1998; Orr, 2000; Vassilieva & Lynch, 1999). The genomic deleterious mutation rate (Ud) and average fitness effect of deleterious mutations (sD) are based on the presumption that significant fitness reductions are an accumulation of single deleterious mutations.

It is important to consider that the term "fitness" is contextually-based. For proteins, loss of function mutations are more common than gain of function; once non-functional, restorative opportunities are limited, but further opportunities for destruction are not. However, speculating that a given mutation is more likely to reduce the efficiency of a protein implies that most proteins are already operating at peak efficiency. Furthermore, peak efficiency is subject to the current environmental conditions, cellular age, and competitive interactions that a given organism experiences. Similarly, mutational effects –

deleterious, neutral, or advantageous - are very population size-dependent. Not only can mutations be eliminated by natural selection or fixed by genetic drift, but the effects of mildly deleterious mutations can be reduced to nearly neutral by the same interactions.

1.8.2 Neutral and Nearly Neutral Mutations

Neutral mutations are mutations that impart no fitness effect ($s = 1$), whereas nearly neutral mutations have negligible fitness effects ($s < 1/N$) (Section 2.6) (Ohta, 1973). Neutral and nearly neutral mutations include mutations with no effect on protein function or an effect so minimal that drift is stronger than selection. In 1993, Kimura mathematically estimated that, when all sequence changes are equally likely, 25% of all point mutations in coding sequences of a gene will be silent. Silent mutations don't alter the amino acid sequence, but may not have a neutral fitness effect when considering stability and translational efficiency for example. Furthermore, silent mutations may have nearly neutral fitness effects in large populations where natural selection is weak and alternate codon usage may be selected for or against.

Where a silent (synonymous) mutation does not change a protein's amino acid sequence and a replacement (nonsynonymous) mutation does, the strength of natural selection in a MA experiment can be estimated as a ratio of replacement to silent mutations. Based on codon availability, and assuming no codon bias, the expected ratio of replacement to silent mutations is approximately 3:1. Observed counts of nonsynonymous to synonymous mutations are typically normalized to this expectation, such that a nonsynonymous : synonymous rate ratio (dN/dS) of 1 is expected in the absence of selection. A ratio below 1 indicates the elimination of some nonsynonymous mutations by (strong) natural selection. For example, estimates of the proportion of neutral nonsynonymous mutations are generally around 0.2 – 0.3, indicating that 20 – 30% of nonsynonymous mutations were neutral, and the remainder within coding sequences were deleterious and eliminated by natural selection (Halliburton, 2004).

1.8.3 Advantageous Mutations

Advantageous mutations are the most uncommon mutations, but are responsible for adaptive evolution (Elena & Lenski, 1997; Hietpas *et al*., 2011; Lenski & Travisano, 1994; Smith & Eyre-Walker, 2002). Wiser *et al*. (2013) showed a high correlation between rapid mean fitness increase and mean cell size increase at the beginning of a MA experiment. The group suggested that this short period of rapid fitness and size increase was due to an influx of advantageous mutation accumulation followed by advantage loss due to the effects of genetic drift (Eyre-walker & Keightley, 2007). Similar to its counterpart, minor advantageous mutations are more frequent than highly advantageous mutations (Lenski *et al*., 1991).

1.9 The Fate of Mutations: Loss or Fixation

The ultimate fate of a mutation – loss or fixation – is dictated by a combination of population size and genetic drift, positive selection, and purifying selection (Halliburton, 2004). Logically, a spontaneous advantageous mutation is most likely to initially exist as a single copy in a population. As such, it is much more likely to be lost due to genetic drift within a few generations as opposed to persisting and slowly increasing in frequency. This positive selection is unlikely; until the mutant allele is found in more than a few percent of the population natural selection is too weak to have an effect and influence fixation of the advantageous mutation. Though smaller advantages are more common, for a new mutation, the smaller the fitness advantage, the less likely fixation is.

Within a population lacking recombination (i.e., in an asexual population), where offspring carry the same mutational load as their parent, Muller's ratchet describes the long-term, irreversible, mutational deterioration of a population, ultimately leading to population decline and then extinction (Muller, 1964). In such a population, individuals exist at different levels of fitness at any given point because of newly introduced deleterious mutations. Individuals at the "most fit" level – hypothetically, initially carrying no mutational load - will ultimately acquire at least one new deleterious mutation over time or leave no progeny. At this point, the "most fit" level is lost. No individuals with fewer

mutations can be expected to be found in the future. The previously "second most fit" level, then becomes the "most fit". However, it too will eventually meet the same fateful decline. According to Muller's ratchet, this cycle will continue as the population progressively declines in fitness. With an inevitably smaller population size genetic drift is more effective, accelerating the rate of mutation accumulation, which respectively accelerates the speed of Muller's ratchet. Because of Muller's ratchet, deleterious mutations are more likely to fix in smaller populations. Mutational meltdown occurs when the rate of mutation accumulation synergistically increases with the mutation load (Lynch *et al*., 1999; Sniegowski & Lenski, 1995). Beneficial mutations would need to accumulate at a rate that can sufficiently compensate for the cumulative effects of the deleterious mutations in order to escape this fate; this is unlikely. An understanding of the rate and fitness effects of spontaneous mutations can help expand knowledge in the area of mutational meltdown in populations.

Bacteria with high mutation rates are called mutator strains. Mutator strains are maintained in a natural population at low frequencies due to high mutational load from accumulated deleterious mutations (LeClerc *et al*., 1996). Dissimilarly to the research of LeClerc *et al*. (1996), Matic *et al*. (1997)detected mutator phenotypes through both MMR-deficient and all other gene-inactivating mutations. This broad detection identified low-frequency maintenance of mostly mild mutators across 504 natural *E. coli* isolates, both commensal and pathogenic (Matic *et al*., 1997). But mutator alleles have been known to "hitchhike to fixation" upon producing a highly beneficial allele; a mutator allele without a detectable deleterious load can tolerably persist within a population for longer (LeClerc *et al*., 1996).

1.10 Modelling Fitness Declines

Fitness curves associated with the phrase, "peaks and valleys" were first introduced by Armand Janet in 1895 (McCoy, 1979). However, more notably, Sewall Wright first presented the fitness landscape metaphor at the 6th International congress on genetics in 1932. Wright was asked to drastically simplify the accompanying mathematical models for this audience, and he used the simple fitness landscape model to do so. Fitness landscapes

allow the mapping of genotype or phenotype with fitness, the determinant of reproductive success. Wright had specified that fitness landscapes are inherently multidimensional to accommodate the thousands of genes that can potentially affect fitness. Because of this inherent complexity, fitness landscapes are most commonly simplified by making structural assumptions to complete a workable model that still captures the essence of static landscape (Gavrilets, 2009). The most prominent features of a fitness landscapes are the multiple peaks separated by valleys. There are two classic models of fitness landscapes: (i) the single-peak landscape, and (ii) the rugged landscape, which has multiple peaks and valleys in a 3D space (Kauffman, 1993). The vertical axis is a measurement of fitness (or adaptive value), and the horizontal axis is a measurement of time. Fitness valleys represent areas of low fitness; natural selection drives the population uphill away from the valleys and towards the fitness peaks. Because landscapes are generalized, it is important to remember that a peak is not necessarily (or even likely) the highest fitness point that individuals in the population can realize, it is more likely that a higher peak exists nearby (Kauffman, 1993).

Inferring an individual's position on the fitness landscape can be achieved by plotting various points of fitness across a long-term evolutionary experiment (ex: MA) to obtain the generalized single-peak landscape for a strain, and then plotting subsequent isolates of interest on that landscape for context. Such isolates may be cultured from clinical or environmental sources, tracking its evolutionary pathway may help achieve more accurate predictions for genetic divergence during an outbreak investigation. The shape of fitness landscapes is a fundamental parameter of much evolutionary theory, and is a topic of significant current interest (Fragata *et al*., 2018).

1.11 Intention of This Study

I evaluated variation in spontaneous mutation rates across nine *E. coli* strains, wild-type and clinically-isolated, to address the following objectives:

1. Contribution to the study of mutation rates: How much variation is there in the rate of spontaneous MA between strains, and can we use WGS to anticipate a

strain's mutation rate in comparison to another's? Mutators are known to exist at low frequencies in a population, and therefore I predicted that mutation rates will be highly variable between some genotypes.

2. Contribution to the study of food safety: Identifying potential strains with increased mutation rates is of interest in this experiment, especially with regards to 0111:H2 *E. coli* strain OLC-0682, which has demonstrated high SNV in previous research (Markell, 2017).
3. Contribution to the study of evolutionary theory: The decline of fitness relative to mutation accumulation over time and across strains will be observed to generate an overlapping fitness landscape estimate. Can we model the rise and decline of fitness? In general, I anticipate that fitness will decline across genotypes as neutral and deleterious mutation accumulation predominantly contribute to the mutational load.

Chapter 2: Materials and Methods

2.1 *Escherichia coli* Strains and Growth Conditions

Eight clinical *E. coli* strains associated with foodborne illness outbreaks or multi-drug resistant extraintestinal infections, as well as one laboratory strain, were used (**Table 1**). Three enterohemorrhagic *E. coli* strains lacking *stx1* and *stx2* genes were obtained from the Blais and Carillo laboratories at the Canadian Food Inspection Agency.

Five extraintestinal pathogenic *E. coli* strains were obtained from the Zhanel laboratory at the University of Manitoba. These isolates were collected from patients at hospitals across Canada, from a variety of non-gastrointestinal infection types (Basra *et al.*, 2018).

Table 1. Experimental *E. coli* strains

E. coli Strain	Clinical Source	Pathotype
K-12 (MG1655)	Wild-type, nonclinical laboratory strain	N/A
***OLC-0809**	Nalidixic acid-sensitive ancestor of OLC811	EHEC
****OLC-0969**	Laboratory for Foodborne Zoonosis - Guelph	EHEC
****OLC-0682**	Ottawa Lab Carling (OLC)	EHEC
PB3	UTI	ExPEC
PB4	UTI	ExPEC
PB29	Blood	ExPEC
PB33	Respiratory	ExPEC
PB35	Blood	ExPEC

(*) *E. coli* O157:H7 strain, a nalidixic acid resistant derivative of the ATCC 700728 (NCTC 12900) parent strain, OLC culture collection number 809 (Markell, 2017).

(**) CFIA's culture stock at the Ottawa Carling Laboratory (OLC) and were originally isolated by the Laboratory for Foodborne Zoonosis (LFZ) Guelph and Ottawa Lab Carling (OLC), respectively. They are OLC culture collection numbers 969 and 682 respectively (Basra *et al*., 2018).

Lysogeny Broth (LB) and Lysogeny Broth Agar (LBA) media were used for all culturing, unless otherwise noted, and were prepared by dissolving 1.0% (w/v) peptone, 0.5% (w/v) yeast extract and 0.5% (w/v) NaCl and adding 1.5% (w/v) bacteriological agar (for LBA) before autoclaving. *E. coli* cultures were streaked onto LBA using sterile inoculation loop for single colony isolation and incubated at 37ºC. All sterile and lineage-related work was performed in a B3 biohazard cabinet.

2.2 Mutation Accumulation

Cultures were stored in 50% (v/v) glycerol stock at - 80ºC. Small amounts of frozen glycerol stocks were resuscitated overnight on LBA plates using sterile inoculation loops.

Each MA line originated from a founder colony of clinical or wild-type *E. coli*. Cells were revived from frozen glycerol stock ancestor lineages. To initiate the MA lineages, each of the original nine ancestral lines were streaked for isolated colonies on LBA and incubated for 24 h at 37℃. In total, 81 well separated colonies (nine biological replicates of each of the original nine ancestral lines) were randomly selected and used to begin an independent MA line when streaked onto LBA (Cycle 0). The plates were incubated at 37℃ for 24 h. MA lineages were assigned names corresponding to their clinical or wild-type ancestral line (**Table 1**) and replicate suffixes. For example, PB3 MA lines were named "PB3-1" to PB3-9".

2.3 Maintaining MA Lines: Growth Cycle (1 – 85) Propagations

The remaining 81 MA lines were propagated every 24 h (+/- 2 hours), for 85 cycles, on LBA through a single-colony bottleneck. With a sterile inoculation loop, peripheral cells were harvested from the single colony nearest to the streak line. This unbiased harvesting approach did not consider colony morphology or size. Cells were streaked for single colonies onto LBA plates divided into three sectors to allow 3 out of 9 MA lineages of a single ancestral line to be propagated side-by-side. Plates were incubated at 37℃ for 24 hr to begin the next growth cycle with occasional growth cycles at 22℃ for 48 hr, for a total of 85 cycles.

Every 10 cycles, a sample of each MA lineage was frozen. An overnight culture was inoculated in LB and incubated at 37℃ with agitation at 150 rpm. The following day, samples were centrifuged at 10, 000 g for 5 min at 4ºC to pellet the cells. The supernatant was discarded, and the pellet was washed twice with LB. After each wash, samples were centrifuged as described. Washed cells were re-suspended in LB and mixed with an equal amount of glycerol to make glycerol stocks then stored at - 80℃ for culture rescues and fitness assay experiments.

At Cycles 30, 60, and 85, each biological replicate of each ancestral line was tested for β-galactosidase activity found in *E. coli* (and other enteric bacteria). Isopropyl β-D-1-thiogalactopyranoside (IPTG)/X-Gal plates were used to distinguish *lacZ*-containing and non *lacZ*-containing cultures. In media, IPTG, an analog of galactose, induces expression of *lacZ* (if present) by inactivating the *lac*-repressor protein, LacI; *lacZ* subsequently encodes β-galactosidase. X-Gal is hydrolyzed by β-galactosidase and the product then spontaneously dimerizes to create an insoluble blue pigment, creating distinguishable blue and white colonies for LacZ-containing and non LacZ-containing cultures, respectively. All strains used in this study were lactose-fermenting ($Lac^{+)}$, therefore, presence of Lac^{-} colonies indicates contamination. Four of the 81 MA lines, OLC809 – 4, OLC – 6, PB3 – 8, and PB3 – 9, were eliminated (Section 2.2) due to contamination during propagation.

2.4 Estimating the Number of Generations of Evolved Populations

In order to estimate the number of generations of MA experienced by each line, I measured growth from a single cell to a single colony, since one cell gives rise to 2^N cells a colony in N generations. The number of viable cells per colony was calculated by harvesting and resuspending an entire colony in 9% (w/v) saline suspension. The colony was first serially diluted in a 96-well plate, then the highest dilutions were dilution plated sequentially thrice on LBA for viable plate counts. For each of the remaining 77 MA lineages, dilution plating was done in three replicates for Cycle 0 and three replicates for Cycle 85 on one LBA plate for visual colony size comparability. This gave nine replicates of the number of viable cells per colony for Cycle 0 and Cycle 85; the replicates were averaged and the number of generations over 85 cycles was calculated as described in **Equation 1.0**.

$$\text{Number of generations} = \log_2 (N_f / N_i)$$

Equation 1.0. Estimation of the number of generations of growth in a single cycle. Number of generations is based on the number of viable cells grown from a single colony, where N_f is the final number of cells in a colony and N_i is the number of cells initially present in the colony ($N_i = 1$).

The number of generations that each MA line underwent in 24 h between cycles of propagation was then estimated as the midpoint between the number of generations in 24 h at Cycles 0 and 85.

2.5 Competitive Fitness Assays

A common reference strain was used in competitive fitness assays for comparability amongst ancestral and evolved lineages. The reference strain harboured a fluorescent tag that disrupted *lacZ* with a (nearly) neutral fitness impact in order to be considered a neutral reference strain against which all other fitness changes could be measured.

Fitness of the evolved MA lineages, and of the ancestral genotypes, was measured through competitive fitness assays. A yellow-fluorescent protein (YFP) marked ancestral

MG1655, MG1655-YFP, was used as a reference competitor to measure relative fitness (generated by Aaron Hinz, personal communication). Previous experiments indicated that the YFP tag does not incur a significant fitness cost in LB. Relative fitness was evaluated by competing MG1655-YFP against: (i) ancestral lineages (Cycle 0), and (ii) evolved MA lineages (Cycle 85) over a 24 h period. A cell type is more fit if grows to a higher cell population density than its competitor.

Ancestral lineages, evolved MA lineages, and MG1655-YFP were revived from frozen glycerol stocks by streaking for single colonies on LBA and incubating for 24 h at 37°C. The following day, a single colony of each was used to inoculate an overnight LB culture in a 24-well plate and incubated at 37°C with agitation at 150 rpm. The competition pairs were mixed in 24-well plates with equal aliquots at a 100-fold dilution with LB. Control cultures were also incubated independently; MG1655-YFP and each untagged ancestral lineage were used as the positive and negative controls respectively.

Aliquots were removed from the competition mixtures and controls before incubating the plates for 24 h at 37°C with agitation at 150 rpm. The aliquot was representative of the initial cell populations in the pre-incubation (T_0) competition. In quadruplicate, the T_0 aliquot was diluted 500-fold in buffer and the counts of YFP and total cells were measured using a Beckman Coulter Gallios flow cytometer. Rapid cell number measurements were taken for the competition pairs and controls, using the presence of fluorescence to distinguish between the cell types. Cell numbers were multiplied by the appropriate dilution factor to determine their initial and final population sizes.

Following 24 hours, aliquots of the incubated (T_{24}) competition mixture were taken, representative of the final cell populations. The T_{24} aliquot was dilute 500-fold in buffer and the two populations within were measured similarly using flow cytometry and multiplied by the appropriate dilution factor to determine their initial and final population sizes.

2.6 Modelling Relative Fitness Declines

Relative fitness was estimated from the change in population ratio between two competing strains grown together in mixed culture. Flow cytometry provided counts of fluorescent cells, and of total cells present, in a mixture. Unscaled (to the ancestral genotype) fitness was calculated as shown in **Equation 2.0 – Equation 2.2**.

$$\text{Unmarked cells} = C - YFP$$

Equation 2.0. Calculation of the unmarked cell count in a competition mixture against the reference strain at T_0 and T_{24}. Flow cytometry detected MG1655-YFP cells and the total cells in the competition population, where *C* is the total cell count and *YFP* is the reference strain cell count. The unmarked cells are what remains of the population.

$$\text{Proportion of YFP cells at } T_0 = YFP_0 / T_0$$

$$\text{Proportion of YFP cells at } T_{24} = YFP_{24} / T_{24}$$

Equation 2.1. Proportion of MG1655-YFP in the initial, T_0, population and the final, T_{24}, population. The final cell population ratio is standardized based on the initial population ratio where *T* is the total cell count and *YFP* is the reference strain cell count, subscripts *0* and *24* represent T_{24} and T_0.

$$\omega = \ln (YFP_{24} / YFP_0) / \ln (B_{24} / B_0)$$

Equation 2.2. Calculation of ancestral or MA lineage strain survivability relative to the reference strain. Relative fitness is calculated where ω is fitness, ln refers to the natural logarithm to reflect population growth (base choice is irrelevant), *YFP* is the population of the reference strain and *B* is the population of the competitor, subscripts *0* and *24* represent T_{24} and T_0.

The fitness change in an MA lineage across 85 cycles must be measured relative to ancestral fitness rather than the fitness of a different genetic background, such as MG1655-YFP. A scaling factor was needed to consider the MG1655-YFP reference strain as neutral

fitness (ω = 1) and calculate evolved MA lineage fitness changes relative to this. Additional flow cytometer assays were used to calculate the fitness impact of the YFP tag on MG1655 and act as controls for distinguishing cell types (MG1655-YFP and ancestral lineages independently). The scaling factor is the mean of the quadruplicate competition assays between the reference strain and the genetic background's ancestor. Relative fitness is divided by the scaling factor to give scaled fitness, calculated as shown in **Equation 2.3 – Equation 2.4**. Once scaled, the mean fitness of each ancestral line will be neutral (ω = 1) and the fitness of the MA lineages will be scaled relative to its ancestral genetic background. Thus, MA lineages with ω > 1 have higher fitness than their ancestor and MA lineages with ω < 1 have lower fitness than their ancestor.

$$SF_{ancestor} = (\sum \omega_{ancestor}) / n$$

Equation 2.3. Calculation of the scaling factor for each genetic background by taking the mean of the relative ancestral fitness's for a genetic background. *SF* is the scaling factor, $\sum \omega_{ancestor}$ is the sum of ancestral relative fitness for a background, done experimentally in quadruplicate, *n*, competition assays.

$$s\omega_{ancestor} = \omega_{ancestor} / SF$$

$$= 1$$

and

$$s\omega_{MA} = \omega_{MA} / SF$$

Equation 2.4. Calculation of MA lineage fitness scaled relative to its ancestor lineage. The scaled fitness, $s\omega$, is calculated by dividing each relative fitness, ω, by the scaling factor, *SF*. Subscripts *ancestor* and *MA* represent ancestral and MA lineages, respectively. Ancestral lineage quadruplicate competition assays have sω = 1 and are the genotype's

neutral fitness. MA lineages with sω > 1 have higher fitness than their ancestor and MA lineages with sω < 1 have lower fitness than their ancestor.

2.7 DNA Sequencing

Following an 85 day Mutation Accumulation experiment, genomic DNA was isolated from *E. coli* strains PB 3, 4, 29, 33, and 35, at Cycle 0 (ancestors) and Cycle 85 (descendants) , as described (Basra *et al*., 2018), and were sequenced using the Illumina NextSeq (ancestors and descendants) and Nanopore MinION (ancestors only) machines. Ancestral genome sequences are described in Mattrasingh *et al*., (under review; Appendix I)

Ancestral and endpoint MA lines were revived from frozen glycerol stocks and grown in LB overnight at 37ºC with agitation at 150 rpm. Samples were contamination-verified using blue-white screening as previously described. A Qiagen DNeasy Blood & Tissue Kit was used to extract genomic DNA according to the bacterial DNA extraction protocol with modified enzymatic lysis pre-treatment. DNA concentration and purity were assessed using a PicoGreen dsDNA quantitation assay and agarose gel electrophoresis (using a 1 kb ladder) respectively.

2.7.1 Long-read Nanopore MinION Sequencing of MA Lines at Cycle 0

For the ancestral lines PB3, PB4, PB29, PB33, and PB35, sequencing libraries were prepared using the Genomic DNA Sequencing kit (Oxford Nanopore Technologies (ONT)) according to the manufacturer using a new FLOWMIN106.R1 Flow Cell with the SQK-RBK004 sequencing kit. The library was loaded onto the MinION Flow Cell and the Genomic DNA 48 hr sequencing protocol was run on MinKNOW software.

2.7.2 Short-read Illumina NextSeq Sequencing of MA Lines at Cycle 0 and 85

For ancestral and endpoint MA lines, sequencing libraries were constructed from genomic DNA using the Nextera XT DNA Sample Preparation Kit (Illumina Inc.) and the Nextera XT Index Kit (Illumina Inc.). Paired-end sequencing was performed with Illumina NextSeq (Health Canada in Ottawa, Ontario) using the 250 cycle NextSeq Reagent Kit v3.

2.8 DNA Sequencing Analysis

Ancestral genomes were obtained from the Carrillo lab at the CFIA (OLC809, OLC682, OLC969), downloaded from GenBank (MG1655), or assembled *de novo*. Mutations in the MA lines were detected via reference-guided assembly, as illustrated in **Figure 2**.

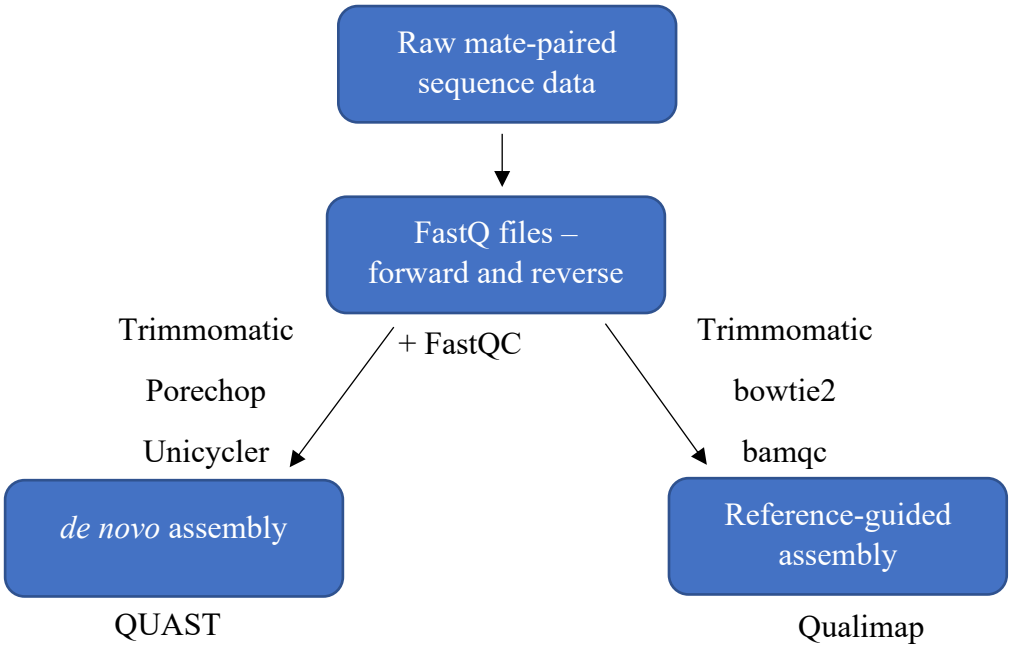


Figure 2. Bioinformatics tools pipeline used to identify SNPs within the raw short- and long-read sequencing data in this experiment. Software packages are written in black.

2.8.1 Ancestral *de novo* Assembly

Two FastQ files, forward and reverse, were outputted for each high throughput sequencing sample. FastQC analysis was performed on the raw reads for quality scoring (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Quality of the Illumina FastQC sequences was enhanced by trimming low-quality bases from the forward and reverse short sequence read ends using the Java Trimmomatic program (http://www.usadellab.org/cms/?page=trimmomatic).

Illumina trimming was performed using the following Trimmomatic parameters: LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36. FASTQC analysis was performed on the trimmed reads for quality scoring. Porechop (https://github.com/rrwick/Porechop) was used to remove the Nanopore adapter sequences from the long reads.

The hybrid assembly was made using the SPAdes-optimiser Unicycler (https://github.com/rrwick/Unicycler). Trimmed forward and reverse paired-end short reads and trimmed long reads were assembled. The reverse-complemented reads were used as paired-end input while the sequencing reads were used as mate-pair input. SPAdes made a short-read assembly graph, the graph was then bridged with long reads for contig production and the contigs were aligned to the short reads for quality improvement (Wick *et al*., 2017).

The Python tool, Quality Assessment Tool (QUAST; https://github.com/ablab/quast) was used to produce summary statistics (contig sizes, genome size, N50, etc...) for assembly quality. QUAST was used with the corresponding ancestral line as the reference sequence. Bandage (https://rrwick.github.io/Bandage/) was used to visualize the *de novo* assembly.

2.8.2 Reference-guided Assembly

For each MA line, mutations were called via reference-guided assembly against its own ancestral genome. Paired-end Fastq files were trimmed using Trimmomatic as described above, and were aligned to the relevant reference genome using bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2; Langmead & Salzberg, 2012) using default parameters. Alignment quality was assessed using bamqc (https://github.com/s-andrews/BamQC).

Mutations were called using bcftools 'call' function. Only SNP calls with a read depth of at least 20, and a quality of at least 200, were retained. SNPs common to an MA strain and its ancestor were ignored. A site in the genome of a given MA line was considered callable if both the MA assembly and the ancestral assembly both had coverage of at least 20 reads.

2.9 Calculating Mutation Rate

Using callable reads (Section 3.2) of the reference-based assembly mapped to the *de novo* assembled ancestral genome, the total number of mutations that accumulated over 85 cycles for the 77 MA lineages and the corresponding spontaneous mutation rates were calculated as described in **Equation 3.0**.

$$\text{Mutation rate per basepair per generation} = m / CN$$

Equation 3.0. Calculation of mutation rates from whole-genome sequencing data. Mutation rates for each of the nine ancestral lineages of *E. coli* are calculated where m is the number of total mutations, C is the number callable sites, N is the estimated total number of generations over 85 cycles (or bottleneck; Equation 1.0).

Chapter 3: Results

Both the rate at which new mutations accumulate, and their effects on fitness, are crucial parameters both for evolutionary theory and for molecular epidemiology (e.g. Markell, 2017). As mutations accumulate over an extended period of time, at a given mutation rate, there is a corresponding change in fitness. It is unclear, however, the degree to which rates of mutation differ between genotypes. To address these questions, I carried out MA experiments on 9 genotypes of *E. coli*, measured changes in fitness, and ascertained mutation rates by WGS.

3.1 Fitness Assessment of Evolved Lines

Fitness declines for *E. coli* propagated over 85 cycles were estimated using competition assays. Since each genotype was competed against a YFP-tagged derivative of MG1655 (reference strain MG1655-YFP), we competed the reference strain against untagged MG1655 to determine whether the YFP tag imposed a fitness cost (**Figure 3**; one-way ANOVA $P = 0.99998$). The YFP tag had a nearly neutral fitness impact, with mean $\omega = 1.00 \pm 0.00498$.

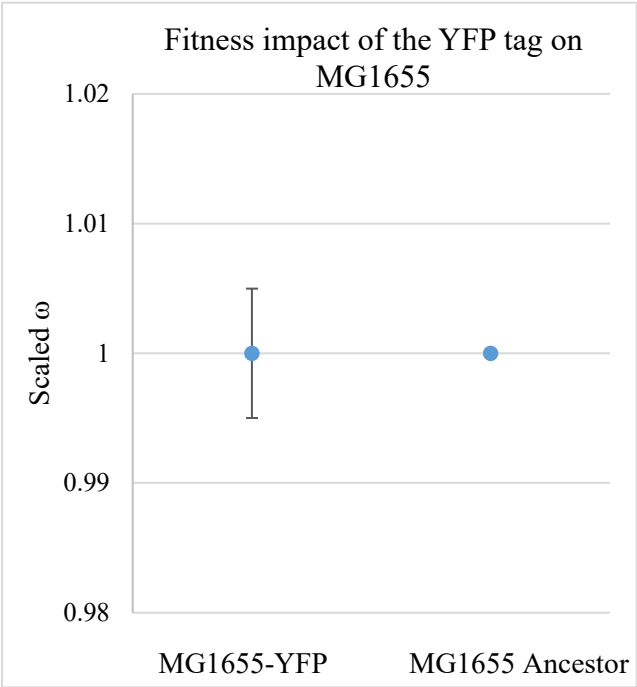


Figure 3. Fitness effect of the YFP-tag on the MG1655 ancestral lineage. The ancestral lineage has an arbitrary fitness value of 1 (ω = 1) and the effect of the YFP addition is nearly neutral.

The reference strain, MG1655-YFP, was separately competed against each ancestral and MA lineage replicate in LB over 24 h competition fitness assays. Since fitness varied between ancestors, we re-scaled the fitness of each MA line to its respective ancestor (**Equation 2.3**).

We found significant fitness variation between genotypes at Cycle 85 (**Figure 4**; one-way ANOVA P = 0.001609). Some genotypes, however, showed no loss of fitness on average (OLC809, PB3, PB29), while OLC682 showed a much larger reduction in fitness. Differences in fitness changes between genotypes could result from fewer (or more) mutations, different mutational effects, or a combination of the two; these possibilities are addressed below. Overall, the average loss in fitness across all nine genotypes was 0.034 – that is, MA lines were on average 3.4% less fit than their ancestor.

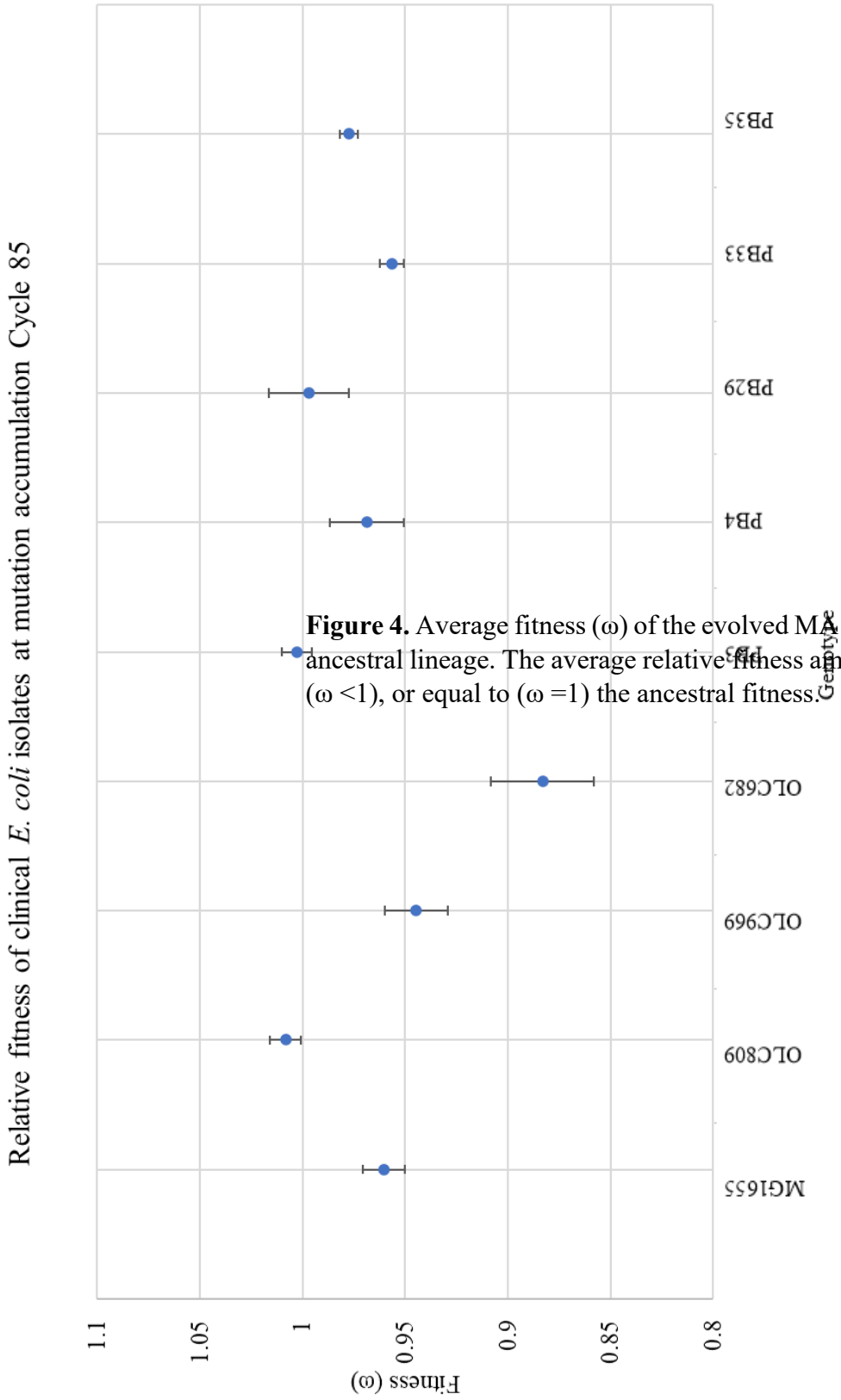
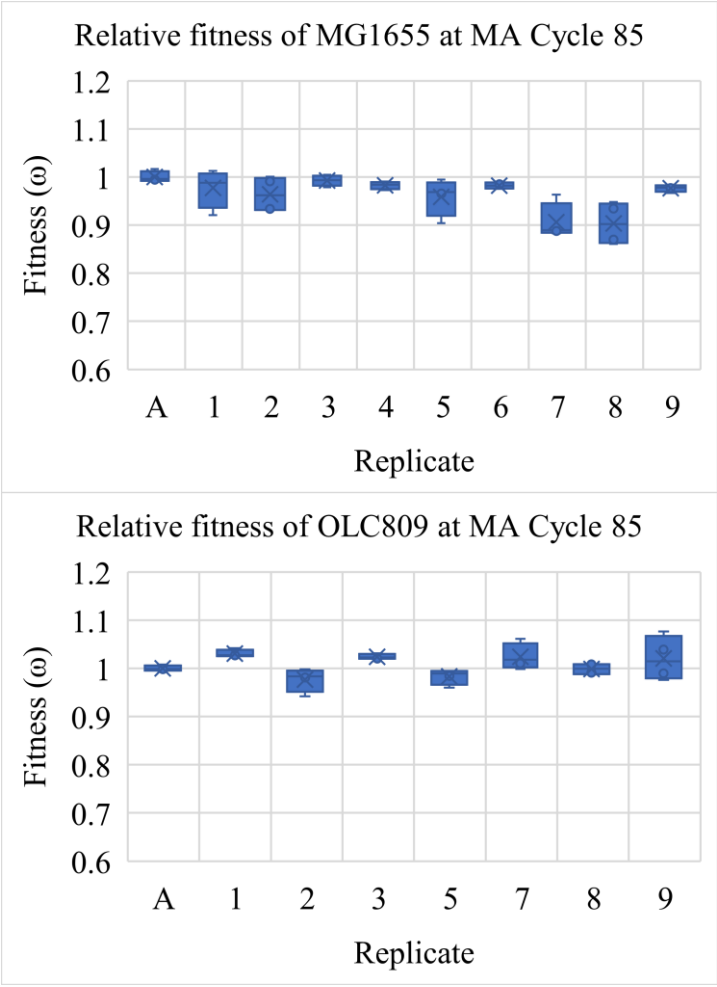
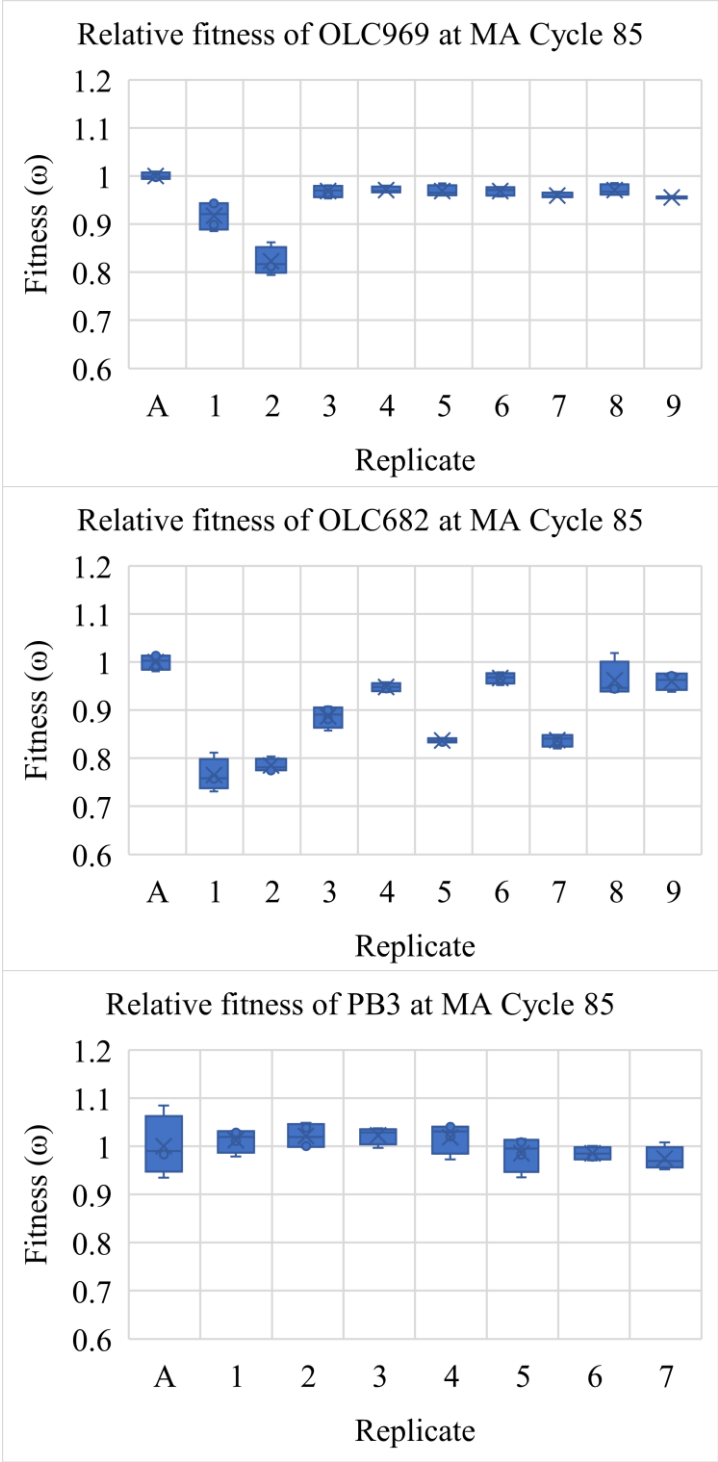
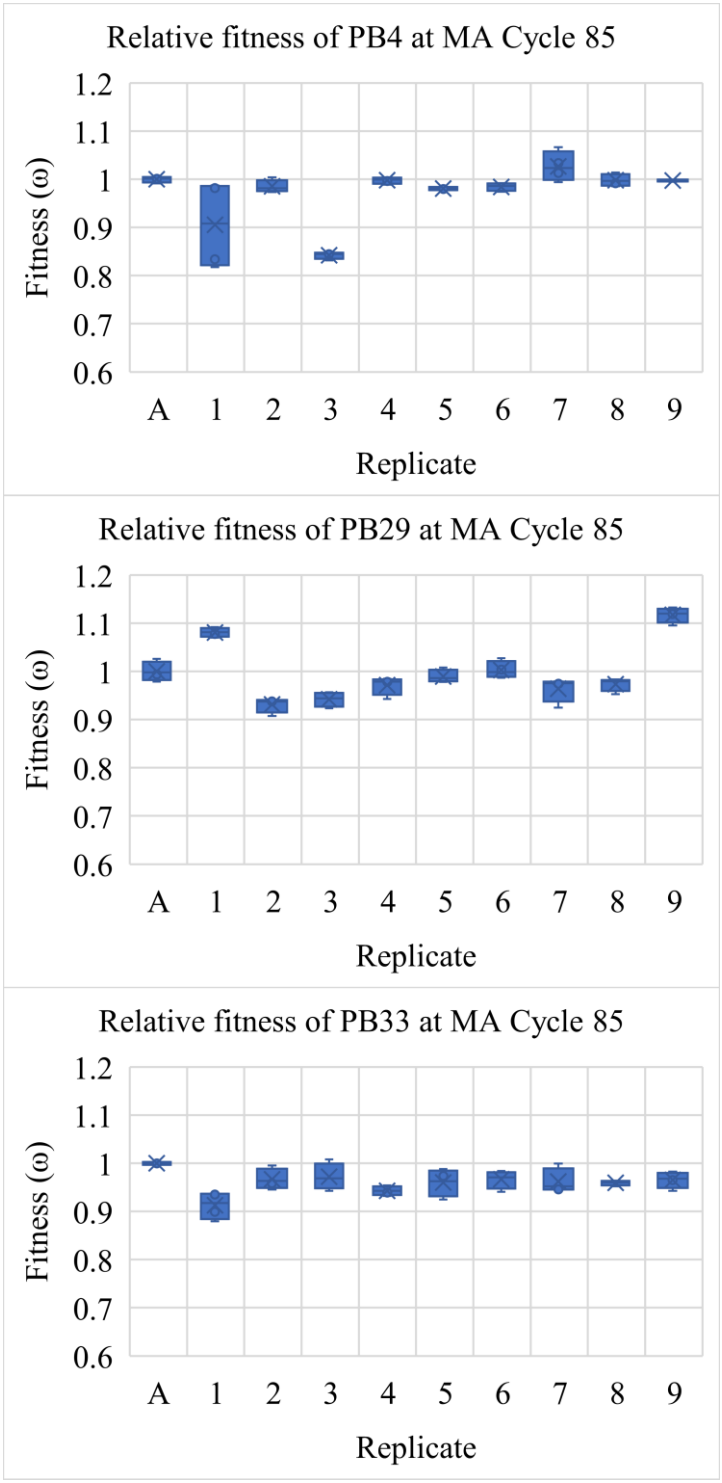


Figure 4. Average fitness (ω) of the evolved MA lineages for each genotype at
ancestral lineage. The average relative fitness among the MA lineage replicate
(ω <1), or equal to (ω =1) the ancestral fitness.

Overall, we found less variation between MA lineages of one genotype than there was between genotypes (**Figure 5**; one-way ANOVA P = 0.001609); even MA lineages of mutator genotype PB3 (see below) followed a generally consistent fitness decline over 85 cycles. Generally speaking, MA lineages of a genotype evolved similar fitness changes, but there were exceptions to this observation. Significant fitness variation from the genotype's general trend was observed for certain replicates within PB4, PB29, and mutator genotype OLC682. Large fitness variation within some genotypes could be indicative of an early mutant clone event – a jackpot event (Luria & Delbrück, 1943), or phase variation, for example.




Relative fitness of OLC969 at MA Cycle 85
Fitness (ω)
1.2
1.1
1
0.9
0.8
0.7
0.6
A 1 2 3 4 5 6 7 8 9
Replicate
Relative fitness of OLC682 at MA Cycle 85
Fitness (ω)
1.2
1.1
1
0.9
0.8
0.7
0.6
A 1 2 3 4 5 6 7 8 9
Replicate
Relative fitness of PB3 at MA Cycle 85
Fitness (ω)
1.2
1.1
1
0.9
0.8
0.7
0.6
A 1 2 3 4 5 6 7
Replicate


Relative fitness of PB4 at MA Cycle 85
Fitness (ω)
1.2
1.1
1
0.9
0.8
0.7
0.6
A 1 2 3 4 5 6 7 8 9
Replicate
Relative fitness of PB29 at MA Cycle 85
Fitness (ω)
1.2
1.1
1
0.9
0.8
0.7
0.6
A 1 2 3 4 5 6 7 8 9
Replicate
Relative fitness of PB33 at MA Cycle 85
Fitness (ω)
1.2
1.1
1
0.9
0.8
0.7
0.6
A 1 2 3 4 5 6 7 8 9
Replicate

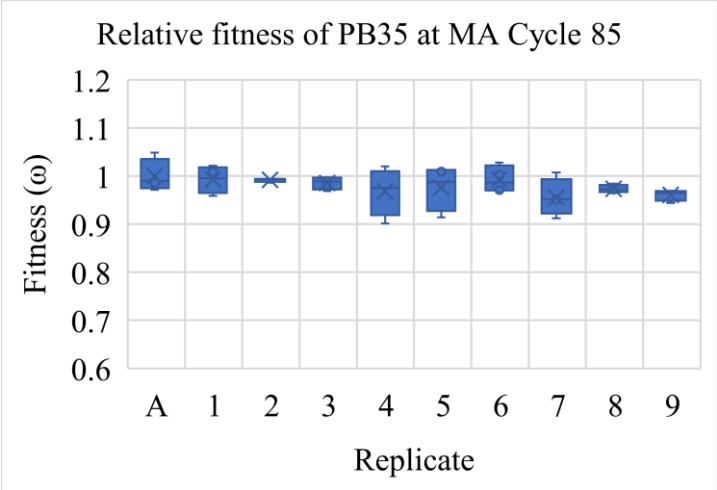
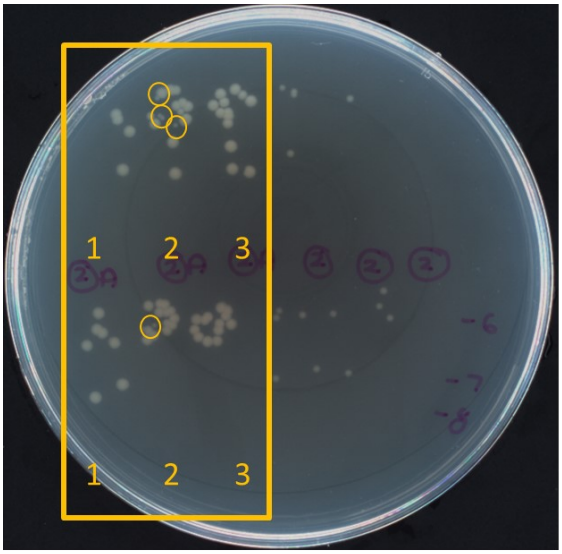


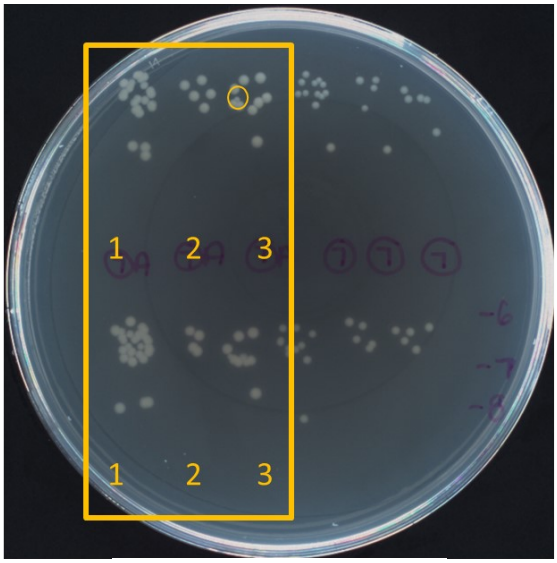
Figure 5. Fitness (ω) of the evolved MA lineages (1 – 9) at Cycle 85 and the ancestral lineage (A) at Cycle 0 scaled to the ancestral lineage. The median relative fitness (blue X) among the quadruplicate fitness assays for each replicate is either more than (ω >1), less than (ω <1), or equal to (ω =1) the ancestral fitness.

3.2 Spontaneous Mutation Rates

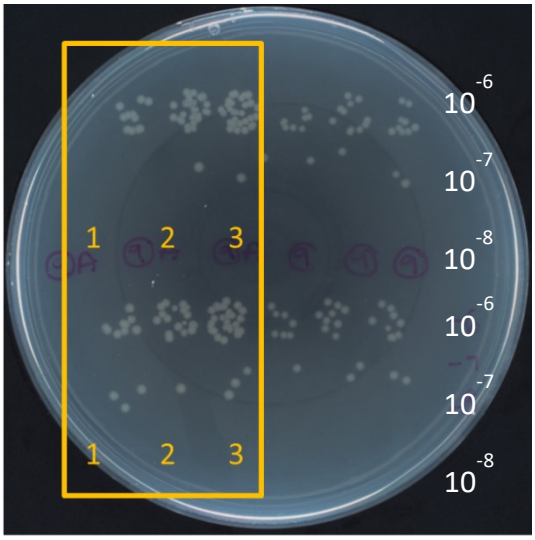
Since spontaneous mutation rates are calculated on a per generation basis, I estimated the number of generations per 24 h propagation cycle independently for each of the nine replicates per MA line. For each line, the number of cells in a colony was measured by spot-plating. This was done in triplicate by diluting and spot-plating Cycle 0 and Cycle 85. MA lineages underwent between 2592.5 and 2848.4 generations over 85 bottlenecks (**Table 3**). This method also provided a visualization of each colony's general morphology, which sometimes differed between Cycles 0 and 85. Such indicators of MA effects are shown in **Figure 6a – 6c**.



OLC682 - Replicate 2



OLC682 - Replicate 7



OLC682 - Replicate 9

Figure 6a. The effects of the 85 cycle MA on colony morphology and count in three MA lineages of OLC682. Replicate 2 (top left), Replicate 7 (top right), Replicate 9 (bottom). On each plate, the triplicate serial dilutions (10^{-6} – 10^{-8}) were performed twice (top and bottom plate), the triplicate ancestral lineage (left, yellow rectangle) is compared to the MA lineage (right). Distinctive colony size change within the ancestral lineage (yellow circle) is observable in some replicates.

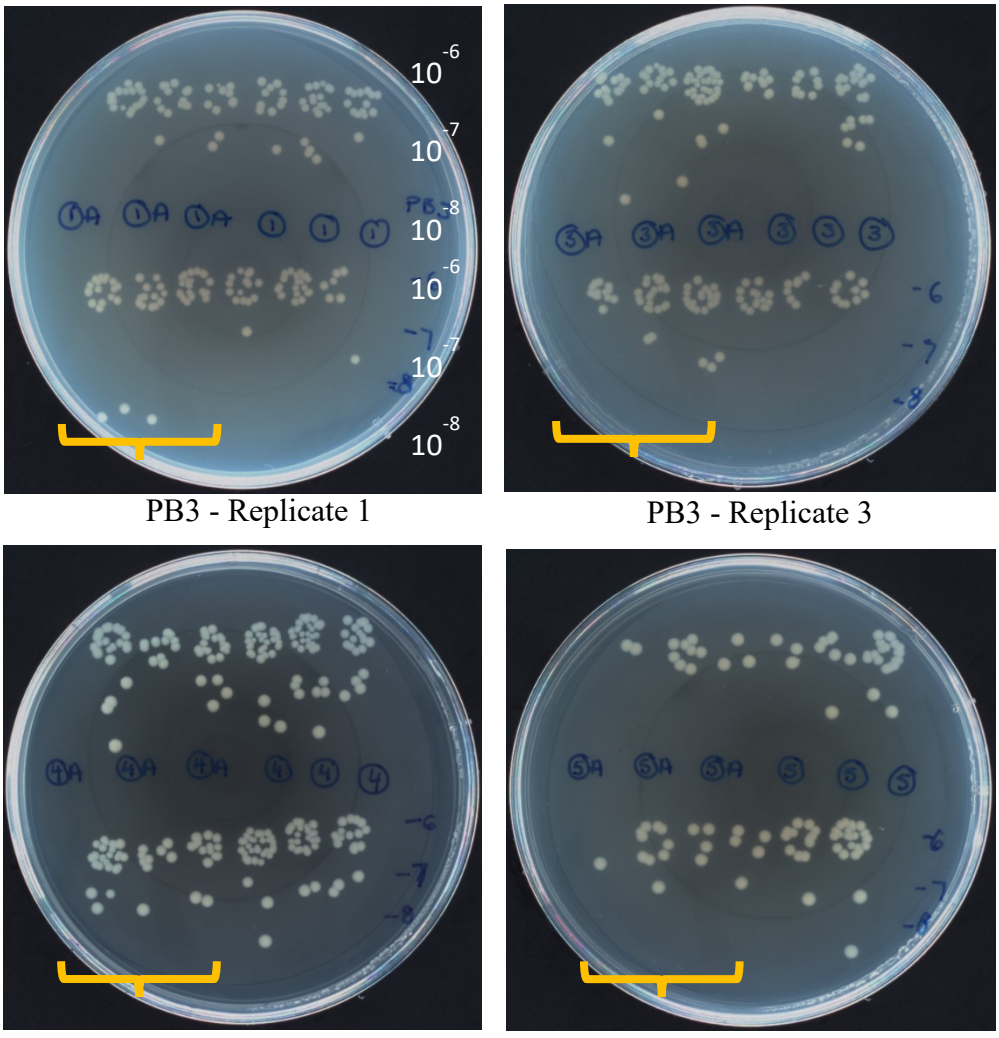
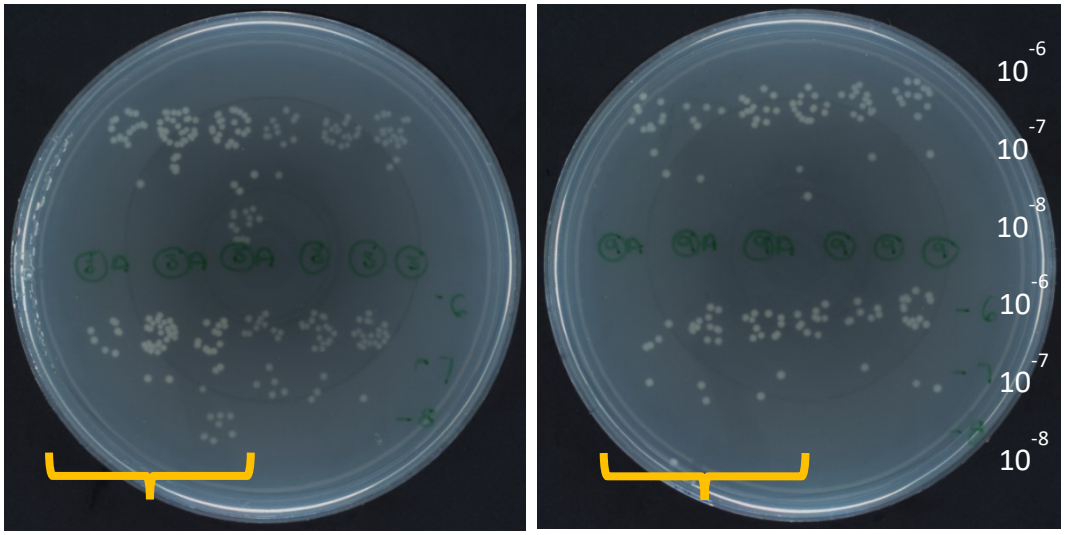


Figure 6b. The effects of the 85 cycle MA on colony morphology and count in three MA lineages of PB3. For Replicate 1 (top left), Replicate 3 (top right), Replicate 4 (bottom left), and Replicate 5 (bottom right), the triplicate serial dilutions (10^{-6} – 10^{-8}) were performed twice (top and bottom plate), the triplicate ancestral lineage (left, yellow arrow) is compared to the MA lineage (right).



PB29 - Replicate 3 PB29 - Replicate 9

Figure 6c. The effects of the 85 cycle MA on colony morphology and count in two MA lineages of PB29. For Replicate 3 (left) and Replicate 9 (right), the triplicate serial dilutions (10^{-6} – 10^{-8}) were performed twice (top and bottom plate), the triplicate ancestral lineage (left, yellow arrow) is compared to the MA lineage (right).

Mutations were called for each MA lineage via reference-guided assembly, against its own ancestral genome. We considered only mutations at callable sites, requiring 20X coverage in both the ancestral and MA lineage. As such, six lineages with a low number of callable sites compared to other genomes of the same genetic background were not included in this analysis (MG1655-2, OLC809-1, PB3-2, PB3-7, PB29-2, PB35-1). Under Qualimap analysis, many of these excluded lineages with low coverage also had low numbers of mapped reads. Furthermore, only high-quality threshold (>200) mutations at callable sites were included in this analysis, excluding ≈58000 low quality threshold mutations. Using reference-based mapping, BPs (Section 1.4.1) and small indels (Section 1.4.2) were detected. Certain lineages detected an unrealistically high number of mutations (>1500), these lineages were also removed from further analysis. In total, 2882 SNPs and small indels were detected and selected for further analysis. We found significant variation between the average mutation rates in the nine genotypes (**Tables 2**, **3**; **Figures 7**, **8**; one-way ANOVA $F = 3.700$ $P = 0.001609$).

Spontaneous mutation rates were calculated as mutations per bp per generation in the MA lineages and averaged by genotype in **Table 2**. There are three general 'trends' in spontaneous mutation rates for the nine strains: (1) most genotypes are consistently non-mutators with low mutation rate variation among replicates. (2) The PB29 genotype is an "infrequent" mutator since most MA lineages of the genotype are non-mutators, apart from a rare mutator MA lineage with a mutation rate similar to (3) mutator genotypes OLC682 and PB3. Based on mutation rate alone, the mutator and infrequent mutator genotypes all have elevated (average) spontaneous mutation rates and variation between MA lineages within the genotype, that is to say that mutator genotypes have higher mutation rates stretched across a larger range. This is possibly because mutator genotypes are more likely to be MMR-deficient; replicative error mutations have unpredictable rate and site potential.

Table 2. Average mutation rate by genotype.

Genotype	Contributing MA Lineages	Average Mutation Rate ($\times 10^{-9}$; per bp per generation)	± Standard Deviation ($\times 10^{-9}$)
MG1655	7	0.232	0.0862
OLC809	4	0.123	0.0627
OLC682	8	16.0	4.75
OLC969	8	0.217	0.129
PB3	5	5.89	1.37
PB4	7	0.182	0.165
PB29	8	8.13	21.6
PB33	9	0.643	0.305
PB35	8	0.993	0.566

The mutational potential of mutators and infrequent mutators is best demonstrated through the MA lineage assembly statistics in **Table 3**, where putative mutator lineages sustain ~100-fold more mutations than non-mutators. Conversely, there are either non-mutator genotypes such as MG1655, OLC809, OLC969, and PB4 that consistently accumulated ≤ 5 mutations over 85 cycles, or non-mutator genotypes whose MA lineages either accumulated a handfuls of mutations or no mutations as all (PB33, PB35).

Table 3. Genome re-sequencing statistics for MA lineages used to calculate *E. coli* mutation rates.

MA Lineage	Callable Sites	Number of Mutations	Total Number of Generations	Mutation Rate ($\times 10^{-9}$; per bp per generation)
MG1655-1	4,499,618	3	2731.1	0.244
MG1655-2	4,607,588	3	2796.0	0.233
MG1655-4	4,616,707	3	2654.1	0.244
MG1655-5	4,568,782	4	2639.5	0.332
MG1655-6	4,479,953	2	2688.2	0.166
MG1655-7	4,554,111	4	2740.2	0.321
MG1655-8	4,531,541	1	2652.0	0.0832
OLC809-2	5,210,939	1	2673.1	0.0718
OLC809-7	5,148,309	2	2743.7	0.142
OLC809-8	4,827,922	1	2729.8	0.759
OLC809-9	5,193,805	3	2822.1	0.205
OLC682-1	4,380,807	301	2596.9	26.5
OLC682-3	4,821,041	224	2622.8	17.7
OLC682-4	4,587,554	131	2744.9	10.4
OLC682-5	5,050,954	204	2676.7	15.1
OLC682-6	5,017,877	197	2665.0	14.7
OLC682-7	2,381,665	101	2625.5	16.2
OLC682-8	2,632,368	97	2800.1	13.2
OLC682-9	4,780,189	180	2677.3	14.1
OLC969-1	4,106,729	2	2721.4	0.179
OLC969-3	4,290,793	4	2690.0	0.347
OLC969-4	3,791,874	0	2686.1	0
OLC969-5	4,639,600	2	2741.7	0.157
OLC969-6	4,651,380	2	2665.1	0.161
OLC969-7	4,636,644	5	2719.9	0.396
OLC969-8	4,403,374	2	2635.8	0.172
OLC969-9	4,579,437	4	2711.7	0.322
PB3-1	4,934,056	87	2729.1	6.46
PB3-3	4,931,634	75	2729.6	5.57
PB3-4	4,931,841	78	2770.7	5.71
PB3-5	4,930,645	100	2616.8	7.75
PB3-6	1,567,637	17	2724.0	3.98
PB4-2	1,082,447	1	2763.4	0.334
PB4-4	4,677,808	5	2799.0	0.382
PB4-5	4,687,999	2	2771.5	0.154
PB4-6	4,634,350	1	2755.5	0.0783
PB4-7	4,362,755	0	2814.5	0
PB4-8	4,640,571	4	2592.2	0.333

PB4-9	3,913,461	0	2731.2	0
PB29-1	5,127,561	8	2640.5	0.591
PB29-3	5,052,353	878	2822.7	61.6
PB29-4	5,039,260	8	2740.0	0.579
PB29-5	5,124,515	4	2711.7	0.288
PB29-6	5,051,259	7	2705.5	0.512
PB29-7	5,130,575	11	2730.1	0.785
PB29-8	5,068,436	5	2765.2	0.357
PB29-9	5,112,857	5	2680.1	0.365
PB33-1	3,639,224	7	2744.2	0.701
PB33-2	3,627,958	6	2660.4	0.622
PB33-3	3,639,221	12	2762.0	1.19
PB33-4	3,639,224	9	2610.9	0.947
PB33-5	3,633,056	5	2735.9	0.503
PB33-6	3,617,188	8	2704.2	0.818
PB33-7	3,639,141	3	2777.2	0.297
PB33-8	3,632,806	4	2714.4	0.406
PB33-9	3,638,723	3	2724.7	0.303
PB35-2	2,130,405	6	2709.4	1.04
PB35-3	2,099,533	10	2701.3	1.76
PB35-4	2,122,814	8	2848.4	1.32
PB35-5	2,074,350	3	2814.6	0.514
PB35-6	2,126,943	9	2794.0	1.51
PB35-7	2,130,713	0	2788.5	0
PB35-8	1,940,732	4	2677.8	0.770
PB35-9	2,128,350	6	2753.8	1.02

3.3 Genome-wide Mutation Accumulation

Over 85 bottlenecks, the distribution of the number of mutations accumulated in each genotype are compared in **Figure 7a – b**. With the exception of the genotype PB29, which contains the infrequent mutator MA lineage, all non-mutator genotypes accumulated an average of ≤ 7 mutations callable site-wide

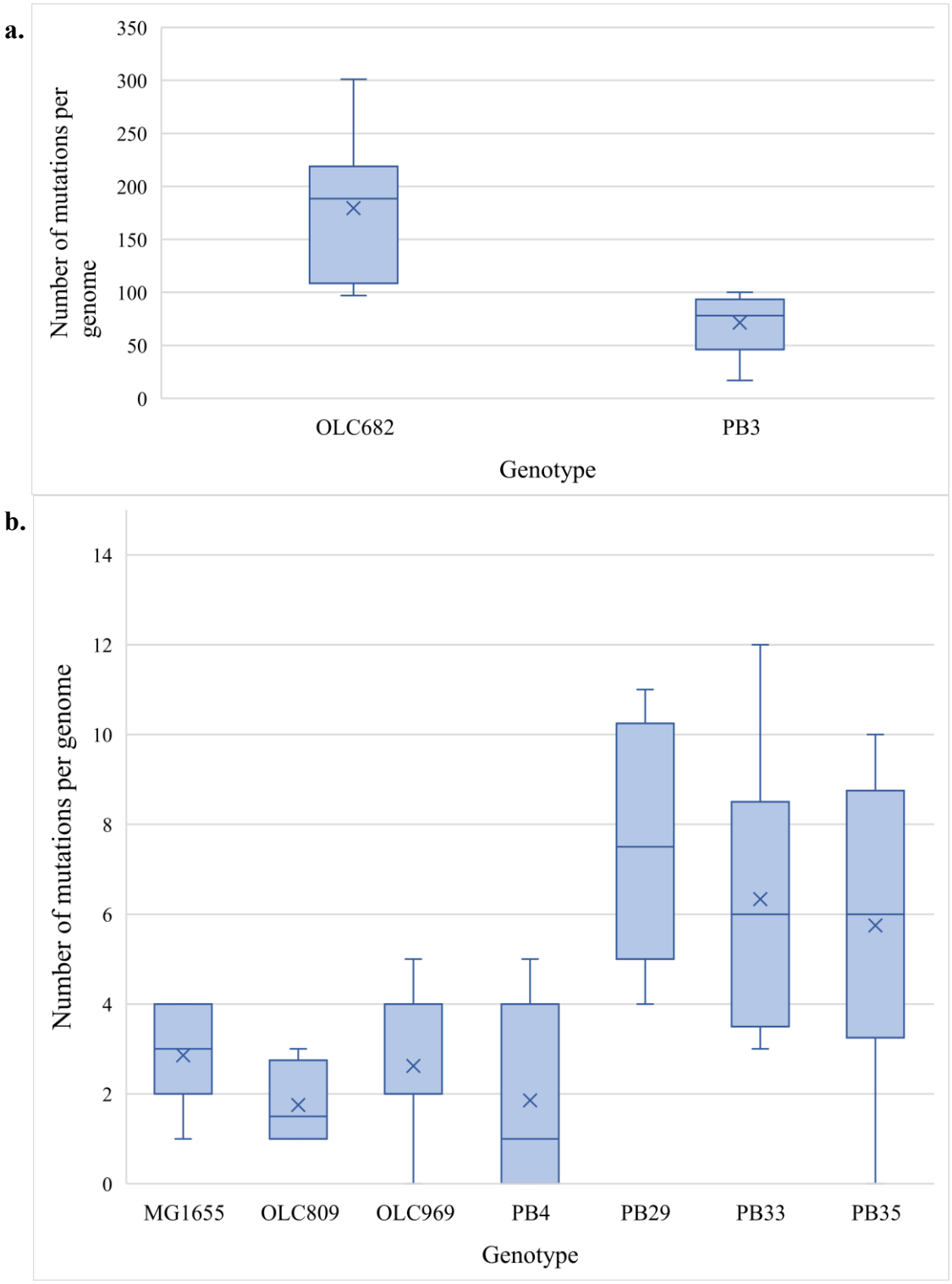


Figure 7. Comparison of the amounts of MA mutations between genotypes. A summary of the number of mutations detected within the callable bases of each genome in **a.** mutator and **b.** non-mutator MA lineages.

In a larger genome there are numerically more opportunities for mutations to accumulate. Consequently, it is possible for genomes of a certain size to be more or less predisposed to accumulate mutations; this was compared in **Figure 8a – b**. With the exception of PB33 and PB35, genotypes averaged 4.5 million callable bases or above (**Table 3**). However, 3.6 million and 2 million bp were consistently called for MA lineages of the non-mutator genotypes PB33 and PB35 respectively. In comparison to **Figure 7**, considering the number of mutations per callable base pair is largely proportionate for genotypes other than PB35. Genotype PB35 has a disproportionately elevated number of mutations accumulated per base pair than its counterparts.

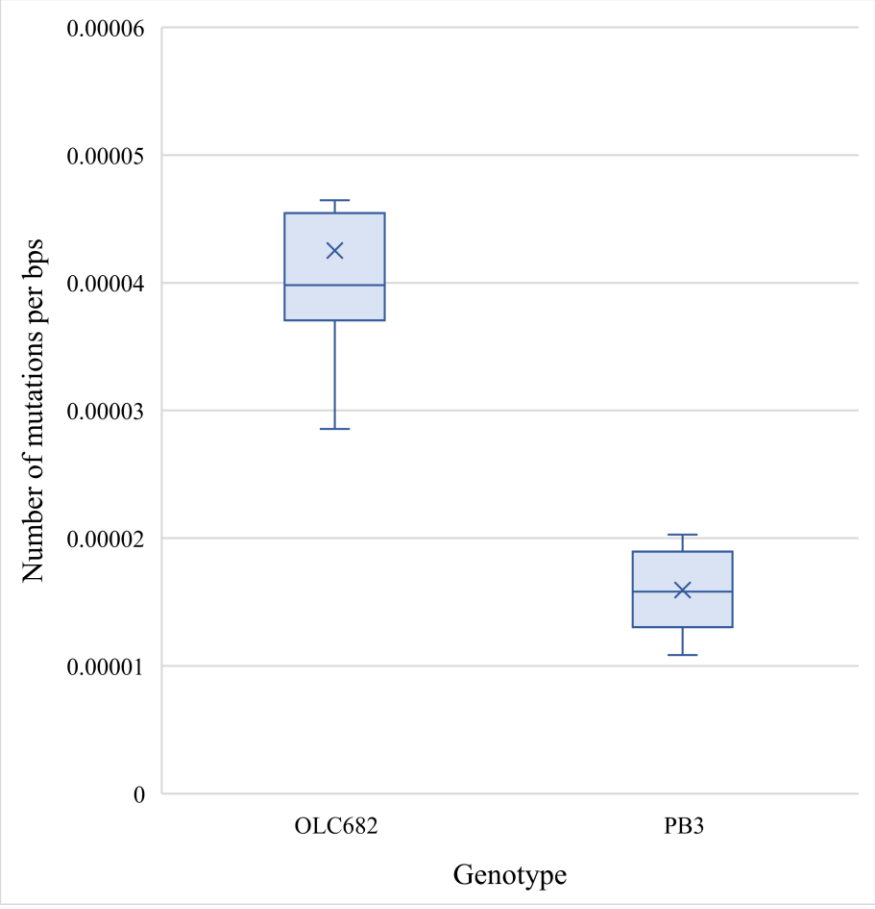


Figure 8a. Comparison of mutations per callable bps in mutator genotypes.

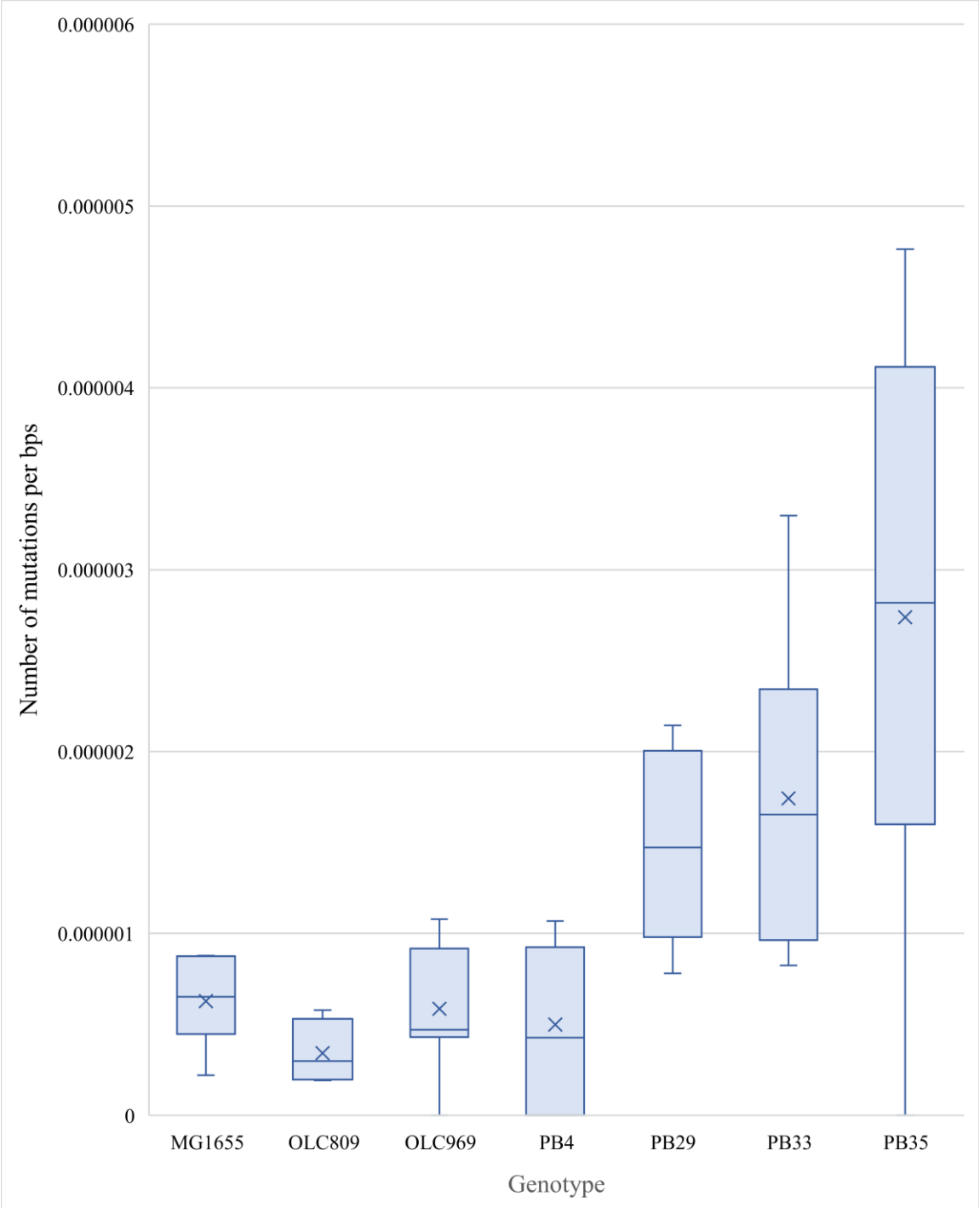


Figure 8b. Comparison of mutations per callable bps in non-mutator genotypes.

Following sequencing dataset exclusions, of 13 MA lineages, 64 MA lineages remained. Of the 2882 genome-wide mutations detected across 64 MA lineages, 28 mutations were small indels and the remaining 2774 were SNPs (Supplementary Data). There were three detected indel "hotspots" with two "hits" each within two genetic backgrounds, PB29 and OLC969.

An essential aspect of MA experiments is mutation fixation, irrespective of natural selection in order to observe unbiased long-term evolution. To do this, a small enough bottleneck must be used to restrict the efficiency of selection. Allelic variants with selection coefficients that are smaller than the power of genetic drift will evolve in an effectively neutral manner. The overall distribution of selection coefficients of our lineages is largely below one, where the accumulated mutations were fixed after genetic drift has overpowered natural selection. As such, this dataset provides a reliable perspective of genome-wide mutation accumulation without the influence of natural selection's deleterious mutation purge.

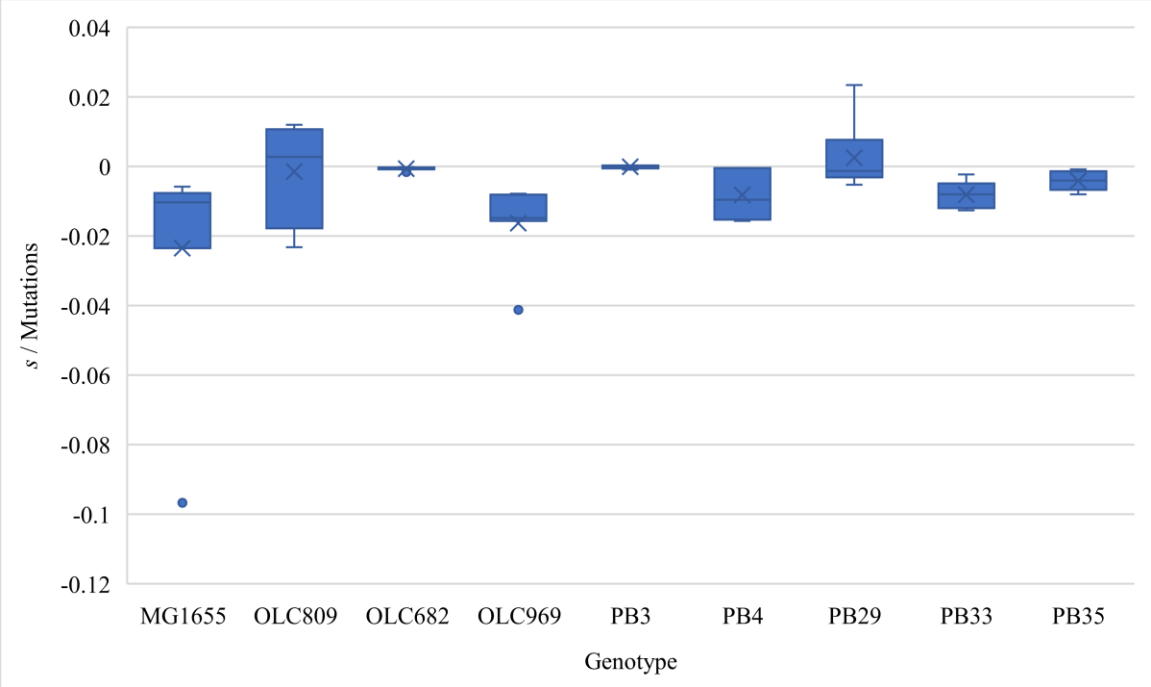


Figure 9. Distribution of the average selection coefficients per mutation of each Cycle 85 MA lineage relative to the ancestral lineage. The range of selection coefficients ($s = \omega - 1$) per mutation for each genotype identifies the accumulated mutations in MA lineages as either subject to the biases of natural selection ($s > 0$), neutral with respect to natural selection ($s = 0$), fixed irrespective of natural selection since genetic drift has overpowered natural selection ($s < 0$).

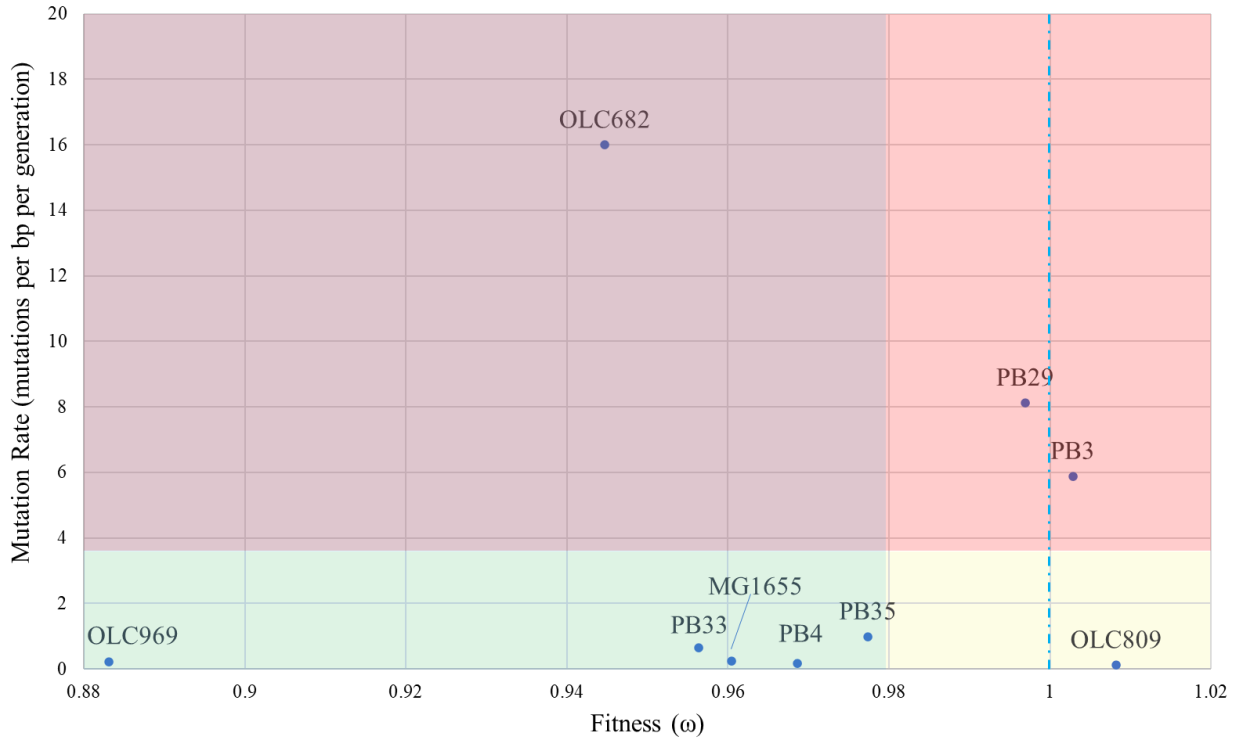


Figure 10. Distribution of the average mutation rate and fitness change of each Cycle 85 MA lineage. Vertically, lineages had a typical mutation rate (yellow box) or elevated mutation rate (red box). Horizontally, lineages had little fitness change relative to their respective ancestral lineage (blue dotted line) or more fitness change (blue box). MA lineages had either a high mutation rate with high fitness impact (top left quadrant, blue), high mutation rate with low fitness impact (top right quadrant, red), low mutation rate with high fitness impact (bottom left quadrant, green), or low mutation rate with low fitness impact (bottom right quadrant, yellow).

Chapter 4: Discussion

The aim of this study was to determine the strain-specific spontaneous mutation rates and corresponding fitness declines of eight clinical, and one laboratory, *E. coli* strains. It had previously been reported that the *E. coli* 0111:NM strain, OLC682, had a significantly elevated SNP accumulation rate compared to other EHEC strains, OLC809 and OLC969, under food production relevant conditions (Markell, 2017; McCarthy, unpublished). This report of highly variable SNP accumulation among foodborne illness-related strains prompted further research of strain-specific gauges of evolution, spontaneous mutation rate and relative fitness declines, for the purpose of more accurately clustering strains during bacterial outbreaks. By correlating the relative fitness of these MA lineages with the particular mutations that accumulated, we present a comprehensive depiction of the variable fitness effects of spontaneous mutations in clinically important pathogenic *E. coli*. Studies such as this, which focus on natural variation, are important because they provide a nearly unbiased representation of the natural mutation spectrum of the studied organism. The more we understand about *E. coli*'s mutation potential, the more readily we may predict and address sources of foodborne illness outbreaks.

A classic MA experiment was carried out for 85 cycles with 81 independent MA lineages derived from nine ancestral lineages of *E. coli*. This method founded a new population daily from a single-cell bottleneck. This reduced population size limits the efficiency of natural selection to purge deleterious mutations, and to fix beneficial mutations. Because of this, nearly all mutations become fixed by genetic drift with equal probability. Measurement of generations of growth per day were taken at Cycle 0 and Cycle 85 and averaged, resulting in an average of 2720 generations per line over the course of the MA experiment. The resulting 77 evolved MA lineages that were whole-genome sequenced were compared to their corresponding ancestral lineage to identify *de novo* mutations that accumulated over the course of the experiment. Analysis of the accumulated mutations is ongoing, and will be expanded to include the proportion of mutations that are nonsynonymous and synonymous, and the specific genes containing mutations in order to estimate the likely gene-specific fitness effects. It will also allow us to detect MMR mutations in PB29-3.

Individual lineages harbored a total mutational load of 0 - 878 spontaneous mutations. In total, 2882 SNPs and small indels were detected. Mutator strains OLC682 and PB3 had consistently elevated mutation rates compared to the remaining non-mutator genotypes, averaging 16.0 and 5.89 mutations per bp per generation respectively, although OLC682 had much more variation among its MA lineages compared to PB3. Genotype PB29 appears to be an infrequent mutator due to MA lineage PB29-3, which appears to have accrued a mutant allele that independently increased the genotype's average mutation rate and variation with >800 accumulated mutations. Infrequent mutators, such as MA lineage PB29-3 warrant further study for spontaneous mutation rate. Because there was only one replicate that presented such a distinguished mutation rate variation from the remainder of the genotype, and because the detected mutations were not found to be condensed in a specific genomic hotspot, it would be useful to identify the mutator allele that this lineage acquired and determine the frequency at which lineages within this genotype acquire the allele.

The average spontaneous mutation rates of the nine *E. coli* genotypes varied greatly, ranging from 0.123×10^{-9} to 16.0×10^{-9} mutations per bp per generation (**Table 3**). Even so, individual mutation rates for each MA lineage varied substantially. For example, genotype PB29 does have an intermediate average mutation rate compared to the dataset across strains. However, across the eight replicates, seven individual mutation rates range from 0.228×10^{-9} to 0.785×10^{-9} mutations per bp per generation, while PB29-3 alone accounts for a large shift in the average with a mutation rate of 61.6×10^{-9} mutations per bp per generation. While this could possibly be an error in genome variation detection, the number of callable bases (**Table 2**) is comparable to all other replicates within the PB29 genotype. This large spontaneous mutation rate variation within the genotype is further supported by a substantial variation in the PB29-3's generations per day between Cycle 0 and 85; PB29-3 had 4% more generations per day than the genotype's average daily number of generations (**Table 2**). Interestingly, **Figure 6c** exemplifies phenotypic evidence to support that this was a decline in generation time for the individual. At Cycle 0, PB29-3 had 34.5 average generations per day, and at Cycle 85 it had 31.9 average generations per day, whereas the PB29 genotype without the PB29-3 outlier averaged 31.9 ± 0.169 generations per day. This places the individual's average generations per day across 85

cycles much higher than its counterparts. PB29-3 exemplifies a mutator isolate within a non-mutator genotype (seemingly unlike PB3) that has a shorter generation time than its counterparts. Changes in the gene expression of certain bacterial growth-rate dependent parameters such as gene and plasmid copy numbers, or RNA polymerase and ribosome abundance, most specifically transcription of ribosomal RNA, are the likely culprits for reduced generation time in a lineage (Bremer & Dennis, 1996; Haugen *et al*, 2008). This is potentially the result of a mutation in genes that regulate generation time, which would have been acquired prior to bottlenecking.

To enhance our understanding of the fitness effects of spontaneous mutations, fitness of the ancestral and evolved MA lineages was measured using competitive fitness assays. Using MA-WGS and flow cytometry for competition assays, it was determined that there was a wide range of relative fitness decline variation between strains of *E. coli* (**Figure 4**, **Figure 5**). Similarly to the observations of average genotypic mutations rates for mutators, OLC682 had a sizeable fitness decline with high variation between lineages, whereas PB3 had a negligible fitness change with very little fitness variation between lineages despite having a high mutation rate. Infrequent mutator, PB29-3 also had a minor fitness change despite its mutation rate of 61.6×10^{-9} mutations per bp per generation, suggesting that mutator alleles are not consistently associated with large fitness variation. In general, there was high mutation rate and fitness change variation observed between genotypes, but much less within MA lineages of a genotype. More studies using clinical isolates would be beneficial to broaden our understanding of the fitness and mutation rate variations that exist between genotypes, especially mutator and infrequent mutator genotypes.

The effects of mutations on fitness are not distributed equally across or within genotypes (**Figure 4; Figure 5**). The MA lineages were maintained for a total of 85 cycles because of the unreliable survival of replicates in the OLC682 background; multiple replicates were likely headed towards extinction with a highly deleterious mutation load (**Figure 5, Figure 6a; Replicate 2, Replicate 7**). However, as seen in **Figure 4**, it would be incorrect to categorize the general fitness changes over the mutation accumulation experiment in all nine genotypes as a "decline in fitness". Genotype OLC809 increased in

fitness over 85 bottlenecks while PB3 and PB29 (with high variation within the genotype) had little change in fitness. This demonstrates that mutator genotypes are not necessarily headed towards extinction at a faster rate than their non-mutator counterparts.

Low fitness variation within a genotype such as PB3, which accumulated 357 mutations across five replicates, may indicate accumulation of mostly neutral mutations. Otherwise, it may indicate more effective purging of deleterious mutations from the population compared to other genotypes. This particular genotype could have experienced more efficient effects of natural selection (despite the experimental design) because of increased competition within colonies. As shown in **Figure 6b**, colonies of this the PB3 genotype were consistently large in size throughout the 85 cycles, but had a relatively average and total number of generations with low lineage variability (**Table 2**). Larger cells within a set space could be indicative of quicker resource consumption and increased competition at each bottleneck, allowing for less deleterious mutation fixation. This combined with PB3's inherently high average mutation rate of 5.89×10^{-9} mutations per bp per generation (**Table 2**), could mean more mutations accumulate, largely neutral and deleterious, but deleterious mutations are eliminated very efficiently. It should be noted that PB3-6 had only 17 detected mutations, which is significantly lower than others replicates in the genotype; this is likely because of its reduced number of callable sites (**Table 3**); PB3's average mutation accumulation is likely underrepresented but can be summarized as many mutations with little impact on fitness. This may make the PB3 genotype a prime candidate for carrying large amounts of neutral and nearly neutral mutations over an extended period of time; the genome seems stable enough to allow this with little impact on fitness.

Conversely to PB3, PB29-9 demonstrates the large effect that five mutations across 5 million callable sites can have on a genome (**Table 2**; **Figure 5**). This lineage has a fitness of $\omega = 1.11$, the largest fitness increase across 85 cycles of all MA lineages owing to ≤ 5 SNPs. Four of the five detected SNPs were transition mutations. Similarly, PB29-1 has a fitness of $\omega = 1.08$ owing to only seven SNPs and one short indel, all at different sites than observed in PB29-1 but with a similar fitness effect and effect of mutations (**Figure 5**). Additionally, the PB29 genotype exemplifies substantial fitness variation with fitness

declines in six replicates, neutral/nearly neutral fitness change in two replicates, and fitness increases in two replicates (**Figure 5**). When combined, the average fitness of the evolved PB29 lineages appears to be nearly neutral compared to the ancestor. The relatively large variations in fitness observed within the PB29 genotype is important for assumptions about the genotype's average fitness effects. It should not be assumed that replicate lineages of this genotype will follow a generalised fitness trajectory because multiple replicates have strayed in a seemingly unpredictable manner.

Figure 10 summarizes the relationship between fitness change and mutation rate in each genotype. Mutator genotype OLC682 was alone in the high mutation rate/high fitness impact quadrant, this genotype had acquired many mutations which had a large cumulative impact on fitness. This supports the notion that OLC682 may be MMR-deficient, and consequently acquired large amounts of different mutations. The other two mutators, PB29 and PB3, were in the high mutation rate/low fitness impact quadrant. These mutators have an elevated mutation rate, however, mutations the cumulative fitness effects are low, making these mutators more evolutionarily predictable. The majority of (non-mutator) genotypes in this study, were in the low mutation rate/high fitness impact quadrant. OLC969 was the only genotype that had a distinguishably elevated fitness impact compared to the other genotypes in this quadrant. This means that a "high" fitness impact is more accurately a normal fitness impact, which is a slightly fitness decline across 85 cycles. Such genotypes are the best fit for current practices in food safety relatedness testing. Lastly, OLC809 was in the low mutation rate/low fitness impact quadrant, meaning its genome is highly stable compared to the other genotypes in this study.

4.1 Significance of This Study for Mutation Rate Estimations

Lee *et al*. (2012) used *E. coli* K12 in their calculation of a SNP mutation rate of 2.2×10^{-10} mutations per bp per generation. Similarly to the findings of Lee *et al*. (2012), Foster *et al*. (2015) used E. coli K12 to calculate a SNP mutation rate of 3.12×10^{-10} mutations per bp per generation. By comparison, we used clinical strains of *E. coli* with a mix of natural mutator (potentially MMR-deficient, but unconfirmed) and non-mutator genotypes to calculate an average mutation rate of 36.0×10^{-10} mutations per bp per generation, 16-

fold higher than the findings of Lee *et al*. (2012). However, our more comparable laboratory *E. coli* strain, MG1655, was estimated to have a mutation rate of 2.32×10^{-10}, which is comparable the findings of Lee *et al* (2012) and Foster *et al*. (2015) respectively. Although both publications used approximately 6000 generations in their MA experiments and I used 2700 generations, they are comparable mutation rate estimates.

Illumina sequencing is highly trusted for detecting small mutations, single-base, or multiple-base substitutions, insertions, and deletions. However, resolving longer fragments of mutations such as large indels or IS elements is more difficult. Illumina platforms output short (125 bp) paired-end read data; when mapped against the reference genome, entire affected fragment of a large rearrangement, or other sizeable variant would be fully spanned by these short sequences. In this experiment, long- and short-read analysis was performed exclusively on the ancestral strains using Nanopore MinION sequencing and Illumina NextSeq sequencing respectively. Long-read sequencing was not carried out on the MA lines due to cost and capacity restraints. Because only Illumina sequencing was performed on the MA lineages, IS element mobilization genes and other large structural variants could not be detected in this experiment. As such, the provided mutation rates for each lineage are probably an underestimate because not all mutations accumulated have been detected through short-read sequencing. In the future, large indels and structural variants may be better detected using short- and long-read sequencing on all MA lineages as was performed for the ancestral lineages.

Drake (2012) has argued that mutation rates calculated in certain long-term evolution studies is actually a product selection acting on the codon usage biases of a given population instead of genuine spontaneous mutation rates. Simply due to the experimental procedures involved in plating for MA, selection is likely a factor in this study as well. For MA studies, maintenance of an effectively low population size is essential for minimizing selections. However, when plating a bottlenecked population and allowing growth between bottlenecking cycles for mutations to accumulate, population sizes fluctuate and this enables mutations to be subject to selection. Logistically, colonies will likely experience some resource competition when they are in close physical proximity at the very least, and to an extent this would fix beneficial and purge deleterious mutations from the population.

The ratio of nonsynonymous to synonymous SNPs can be analysed in the future to determine if mutations were accumulated in a neutral manner in the MA study.

4.2 Significance of This Study for Epidemiological Food Safety

This pilot study indicated that clinical *E. coli* isolates can follow three trends of mutation accumulation: non-mutator, infrequent mutator, or mutator genotypes. The limitations of this study, however, are the use of only 2700 generations in the MA study, possible contamination of replicates, and the inconsistent sequencing of replicates with enough base calling coverage and quality to identify all mutations with accuracy. Irrespective of these limitations, mutation variation between genotypes is a valid takeaway. Epidemiology and food safety regulators should take into consideration which of the three type of mutators they have identified when performing food safety and outbreak investigations. As a hypothetical example during an outbreak investigation, if samples were taken from a mutator strain such as OLC682, one would expect to extract five isolates from a single outbreak event with > 5 SNP differences from one another in a matter of weeks. Consequently, its mutator status would need to be known, since what would normally appear to be five multiple outbreak origins is actually one. If samples were taken from an infrequent mutator genotype such as PB29, one should be wary if one of five isolates from a single outbreak has > 10 SNP differences from the rest in a matter of weeks. Its infrequent mutator status would need to be considered when isolating one sample because what would normally appear to be two outbreak origins is actually one. Lastly, if samples were taken from a non-mutator genotype, one would expect to extract five isolates from a single outbreak with ≤ 5 SNP differences from one another.

Estimating strain-specific inherent rates of spontaneous mutation was a primary goal of this study. These estimates were taken under growth conditions that are optimal for bacterial growth in order to minimize the effects of selection and allow all mutations to fix, regardless of their fitness effect, so that they could be accounted for. A logical application of this approach would be to identify how these mutation rates change for a strain when one or more selective events are applied. When nutrients are no longer ample and/or resource competition is introduced, how will the mutation rates change? The idea would

be to mimic a one aspect of a real-world food safety or bodily condition that the clinical strain might have been exposed to during an outbreak. In a future study, this could include using different media sources to perform the same MA experiment with the same strains, for example, minimal media, or media containing synthetic urine. Future experiments could also include using aerobic and anaerobic growth conditions. It is possible that the accumulated mutations in some MA lineages could produce environmental-dependent effects as seen in *Burkholderia cenocepacia* MA experiments performed by Dillon & Cooper (2016). Additionally, would there be additional or reduced mutation variation between the genotypes if grown in different environments? Because the genotypes used in this study were both ExPEC and EHEC pathogens, it is likely that they would have a different array of mutations accumulate simply because they are genetically acclimated to thrive in different environments already.

4.3 Significance of This Study for Evolutionary Theory

Generalizations are a necessity in evolutionary theory, but this study demonstrates our limited ability to accurately generalize and predict the evolutionary direction of lineages within certain genotypes. Mutation variation within lineages of a genotype was found to be predictable in most genotypes, but unpredictable in a select few. Take for example, the mutator genotypes, OLC682 and PB3. Highly variable fitness declines and spontaneous mutation rates were observed in MA lineages of OLC682. Since a variable number of bases were called with enough coverage for each lineage, it is best to focus on mutations per callable base. Between 2.85×10^{-5} and 4.65×10^{-5} mutations per callable bp were called for the eight MA lineages of OLC682. This worked out to spontaneous mutation rates between 10.4×10^{-10} and 26.5×10^{-10} mutations per bp per generation and fitness impacts between 0.76 and 0.97; no lineage in particular took a constant lead in these statistics. In comparison, PB3, which was also a mutator had very little variation between MA lineages in mutations accumulated, mutations per callable bases, spontaneous mutation rate, or fitness declines. This could possibly point to instability of the OLC682 genome; consistent with this hypothesis, phase variation which was also noticeable (e.g. **Figure 6a**). It is also possible that lineages of this genotype were at different points on a

fairly constant fitness landscape. To test this idea, more data points (than Cycle 0 and 85) could be used to create MA lineage-specific fitness landscapes for each replicate at 10-day intervals. This was the original intention but was not laboriously feasible due to the pandemic. Because of strains such as OLC682, its is perhaps possible to anticipate a genotype's mutation rate in comparison to another's with the exception of certain mutators. Because of the general consistency in fitness declines for all other genotypes, it does seem possible to model the rise and decline of fitness to generate an overlapping fitness landscape estimate per genotype.

Chapter 5: Conclusion

The evolutionary success of organisms depends upon their adaptability to changing environmental conditions. The evolution of populations as a whole relies on our understanding of mutations as the underlying source of genetic variation upon which natural selection and genetic drift can act. As such, our understanding of mutation rates, the molecular spectrum of spontaneous mutations, the fitness effects of spontaneous mutations, and most importantly, the variations of the three that exist between and within genotypes, is essential in understanding mutations that change populations. In this MA-based thesis, I used WGS to evaluate the spontaneous mutation rates and spectra of clinical *E. coli* genotypes, and competitive fitness assays to the explore fitness effects of these mutations. By using this method to detect genome-wide SNP- and small indel-accumulation in an unbiased manner, a more complete picture pertaining to the evolutionary path of eight pathogenic (and one laboratory) genotypes has been drawn. However, more analysis of the WGS data is warranted to evaluate TEs and large structural variants within the genomes.

In general, there was high mutation rate and fitness change variation observed between genotypes, but much less within replicates of a genotype. Mutator phenotypes were identified within three genotypes, but preliminary results suggest that mutator phenotypes are not consistently associated with large fitness variation. More studies using clinical isolates would be beneficial to broaden our understanding of the fitness and mutation rate variation that exist between genotypes, especially in mutator and infrequent mutator genotypes. These genotypes are of particular interest for molecular epidemiological investigations of infectious disease outbreaks. Certainly, the elevated spontaneous mutation rates that certain genotypes present under conditions with negligible selection should be considered when determining relatability of isolates in an outbreak investigation, as all strains do not acquire mutations at comparable rates. Based on the low mutation rate and fitness change variations observed within genotypes during this experiment, it appears that predicting the associated fitness landscapes for pathogenic genotypes with differing spontaneous mutation rates may be an attainable addition to outbreak investigations. With the exception of one genotype, within the replicates of a

genotype, mutator or not, there was little fitness variation between replicates. Perhaps identifying the similarities between genotypes that are an exception to this rule may be the next logical step.

Chapter 6: References

Aerts, J., Wetzels, Y., Cohen, N., & Aerssens, J. (2002). Data mining of public SNP databases for the selection of intragenic SNPs. *Human Mutation*, *20*(3), 162–173. https://doi.org/10.1002/humu.10107

Anderson, S., *et al*. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, *290*(5806). https://doi.org/10.1038/290457a0

Barrick, J. E., & Lenski, R. E. (2013). Genome dynamics during experimental evolution. *Nature Reviews Genetics*, *14*(12). https://doi.org/10.1038/nrg3564

Bartels, M. D., *et al*. (2014). Comparing whole-genome sequencing with sanger sequencing for *spa* typing of methicillin-resistant *Staphylococcus aureus*. *Journal of Clinical Microbiology*, *52*(12), 4305–4308. https://doi.org/10.1128/JCM.01979-14

Basra, P., *et al*. (2018). Fitness tradeoffs of antibiotic resistance in extraintestinal pathogenic *Escherichia coli*. *Genome Biology and Evolution*, *10*(2), 667–679. https://doi.org/10.1093/gbe/evy030

Bateman, A. J. (1959). The viability of near-normal irradiated chromosomes. *International Journal of Radiation Biology and Related Studies in Physics, Chemistry and Medicine*, *1*(2). https://doi.org/10.1080/09553005914550241

Bentley, D. R. (2006). Whole-genome re-sequencing. *Current Opinion in Genetics and Development* (Vol. 16, Issue 6, pp. 545–552). Elsevier Current Trends. https://doi.org/10.1016/j.gde.2006.10.009

Bentley, D. R., *et al*. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. https://doi.org/10.1038/nature07517

Blot, M. 1994. Transposable elements and adaptation of host bacteria. *Genetics*, 93, 5–12.

Bremer, H., & Dennis, P. (1996). Modulation of chemical composition and other parameters of the cell by growth rate.

Bradnam, K. R., *et al*. (2013). Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *GigaScience*, *2*(1). https://doi.org/10.1186/2047-217X-2-10

Campbell, A., Berg, D. E., Lederberg, E. M., Starlinger, P., Botstein, D., Novick, R. P., & Szybalski, W. (1979). Nomenclature of transposable elements in prokaryotes. *Plasmid*, *2*(3), 466–473. https://doi.org/10.1016/0147-619X(79)90030-1

Castro-Wallace, S. L., *et al*. (2017). Nanopore DNA sequencing and genome assembly on the international space station. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-18364-0

Chrystoja, C. C., & Diamandis, E. P. (2014). Whole-genome sequencing as a diagnostic test: Challenges and opportunities. In *Clinical Chemistry* (Vol. 60, Issue 5, pp. 724–733). American Association for Clinical Chemistry Inc.

https://doi.org/10.1373/clinchem.2013.209213

Churko, J. M., Mantalas, G. L., Snyder, M. P., & Wu, J. C. (2013). Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases. *Circ Res*, *112*(12). https://doi.org/10.1161/CIRCRESAHA.113.300939

Conrad, T. M., Lewis, N. E., & Palsson, B. Ø. (2011). Microbial laboratory evolution in the era of genome-scale science. *Molecular Systems Biology*, *7*(1). https://doi.org/10.1038/msb.2011.42

Cotton, R. G. H. (1993). Current methods of mutation detection. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, *285*(1), 125–144. https://doi.org/10.1016/0027-5107(93)90060-S

Darmon, E., & Leach, D. R. F. (2014). Bacterial genome instability. *Microbiology and Molecular Biology Reviews*, *78*(1), 1–39. https://doi.org/10.1128/mmbr.00035-13

De Visser, J. A. G. M., Akkermans, A. D. L., Hoekstra, R. F., & De Vos, W. M. (2004). Insertion-sequence-mediated mutations isolated during adaptation to growth and starvation in *Lactococcus lactis*. *Genetics*, *168*(3), 1145–1157. https://doi.org/10.1534/genetics.104.032136

Deamer, D., Akeson, M., & Branton, D. (2016). Three decades of nanopore sequencing. *Nature Biotechnology*, *34*(5). https://doi.org/10.1038/nbt.3423

Delihas, N. (2011). Impact of small repeat sequences on bacterial genome evolution. *Genome Biology and Evolution*, *3*(1), 959–973. https://doi.org/10.1093/gbe/evr077

Derbyshire, K. M., & Grindley, N. D. F. (1986). Replicative and conservative transposition in bacteria. *Cell* (Vol. 47, Issue 3, pp. 325–327). Elsevier. https://doi.org/10.1016/0092-8674(86)90586-6

Dettman, J. R., Sztepanacz, J. L., & Kassen, R. (2016). The properties of spontaneous mutations in the opportunistic pathogen *Pseudomonas aeruginosa*. *BMC Genomics*, 1–14. https://doi.org/10.1186/s12864-015-2244-3

Donnenberg, M. S., Tzipori, S., Mckee, M. L., O'brien, A. D., Alroy, J., & Kapert, J. B. (1993). The role of the *eae* gene of enterohemorrhagic *Escherichia coli* in intimate attachment *in vitro* and in a porcine model. *J. Clin. Invest* (Vol. 92).

Drake, J. W. (1991). A constant rate of spontaneous mutation in DNA-based microbes. *88*, 7160–7164.

Drake, J. W. (2012). Contrasting mutation rates from specific-locus and long-term mutation-accumulation procedures. *G3: Genes, Genomes, Genetics*, *2*(4), 483–485. https://doi.org/10.1534/g3.111.001842

Drake, J. W., Charlesworth, B., Charlesworth, D., & Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics*, 148, 1667–1686.

Edwards, P. R., & Ewing, W. H. (1986). *Identification of Enterobacteriaceae* (3rd ed.). London: Elsevier Applied Science.

Eid, J., *et al*. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, *323*(5910), 133–138. https://doi.org/10.1126/science.1162986

Elena, S. F., & Lenski, R. E. (1997). Test of synergistic interactions among deleterious mutations in bacteria. *Nature*, *390*(6658), 395–398. https://doi.org/10.1038/37108

Eyre-walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nat Rev Genet*, 8, 610–618. https://doi.org/10.1038/nrg2146

Fijalkowska, I. J., Schaaper, R. M., & Jonczyk, P. (2012). DNA replication fidelity in *Escherichia coli*: a multi-DNA polymerase affair. *FEMS Microbiol Rev*, *36*, 1105–1121. https://doi.org/10.1111/j.1574-6976.2012.00338.x

Finlay, B. B., & Falkow, S. (1997). Common themes in microbial pathogenicity revisited. *Microbiology and Molecular Biology Reviews*, *61*(2).

Foster, P. L., Lee, H., Popodi, E., Townes, J. P., & Tang, H. (2015b). Determinants of spontaneous mutation in the bacterium *Escherichia coli* as revealed by whole-genome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(44), E5990–E5999. https://doi.org/10.1073/pnas.1512136112

Fragata, I., Matuszewski, S., Schmitz, M. A., Bataillon, T., Jensen, J. D., & Bank, C. (2018). The fitness landscape of the codon space across environments. *Heredity*, *121*(5). https://doi.org/10.1038/s41437-018-0125-7

Friedberg E. C., *et al*. (2009) *DNA repair and mutagenesis* (2nd ed.). Washington, DC ASM Press.

Fu, Y.-X. (1994). A phylogenetic estimator of effective population size or mutation rate. *Genetics*, *136*(2), 685-692.

Gago, S., Elena, S. F., Flores, R., & Sanjuan, R. (2009). Extremely high mutation rate of a hammerhead viroid. *Science*, *323*(5919). https://doi.org/10.1126/science.1169202

Goodman, M. F., & Tippin, B. (2000). The expanding polymerase universe. *Nature Reviews Molecular Cell Biology*, *1*(2). https://doi.org/10.1038/35040051

Gullapalli, R. R., Desai, K. V, Santana-Santos, L., Kant, J. A., & Becich, M. J. (2012). Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. *Journal of Pathology Informatics*, *3*(1), 40. https://doi.org/10.4103/2153-3539.103013

Halliburton, R. (2004). *Introduction to Population Genetics*. Pearson. https://www.pearson.com/us/higher-education/program/Halliburton-Introduction-to-Population-Genetics/PGM246740.html

Halligan, D. L., & Keightley, P. D. (2009). Spontaneous mutation accumulation studies in evolutionary genetics. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 151–172. https://doi.org/10.1146/annurev.ecolsys.39.110707.173437

Heilbron, K., Toll-Riera, M., Kojadinovic, M., & MacLean, R. C. (2014). Fitness is strongly influenced by rare mutations of large effect in a microbial mutation

accumulation experiment. *Genetics*, *197*(3). https://doi.org/10.1534/genetics.114.163147

Hietpas, R. T., Jensen, J. D., & Bolon, D. N. A. (2011). Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(19), 7896–7901. https://doi.org/10.1073/pnas.1016024108

Horst, J. (1999). *Escherichia coli* mutator genes. *Trends in Microbiology*, *7*(1). https://doi.org/10.1016/S0966-842X(98)01424-3

Haugen, S., Ross, W. & Gourse, R. (2008). Advances in bacterial promoter recognition and its control by factors that do not bind DNA. *Nature Reviews Microbiology*, 6, 507–519. https://doi.org/10.1038/nrmicro1912

Jain, M., Fiddes, I. T., Miga, K. H., Olsen, H. E., Paten, B., & Akeson, M. (2015). Improved data analysis for the MinION nanopore sequencer. *Nature Methods*, *12*(4). https://doi.org/10.1038/nmeth.3290

Johnson, R. P., *et al*. (1996). Growing concerns and recent outbreaks involving non-O157:H7 serotypes of verotoxigenic *Escherichia coli*. *Journal of Food Protection*, *59*(10), 1112–1122. https://doi.org/10.4315/0362-028X-59.10.1112

Katju, V., & Bergthorsson, U. (2018). Old Trade, New tricks : Insights into the spontaneous mutation process from the partnering of classical mutation. *Genome Biol. Evol.*, *11*(1), 136–165. https://doi.org/10.1093/gbe/evy252

Kauffman, S. A. (1993). *The Origins of Order: Self-organization and Selection in Evolution*. Oxford University. https://books.google.ca/books?hl=en&lr=&id=lZcSpRJz0dgC&oi=fnd&pg=PR13&dq=(Kauffman,+1993)+the+origins+of+order&ots=9-AOfYbMQw&sig=dOK41u2C2pUsAJ8CEkeoh5dm33o#v=onepage&q=(Kauffman%2C1993) the origins of order&f=false

Kauffmann, F. (1947). The serology of the *coli* group. *Acta Pathologica Microbiologica Scandinavica*, *21*(1), 20–45. https://doi.org/10.1111/j.1699-0463.1944.tb00031.x

Kondrashov, F. A., & Kondrashov, A. S. (2010). Measurements of spontaneous rates of mutations in the recent past and the near future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1544), 1169–1176. https://doi.org/10.1098/rstb.2009.0286

Krašovec, R., Richards, H., Gifford, D. R., Hatcher, C., Faulkner, K. J., Belavkin, R. V., Channon, A., Aston, E., McBain, A. J., & Knight, C. G. (2017). Spontaneous mutation rate is a plastic trait associated with population density across domains of life. *PLoS Biology*, *15*(8), 1–19. https://doi.org/10.1371/journal.pbio.2002731

Kudva, I. T., *et al*. (2002). Strains of *Escherichia coli* O157:H7 differ primarily by insertions or deletions, not single-nucleotide polymorphisms. *Journal of Bacteriology*, *184*(7). https://doi.org/10.1128/JB.184.7.1873-1879.2002

Kulski, J. K. (2016). Next-generation sequencing — An overview of the history, tools, and "omic" applications. *Next Generation Sequencing - Advances, Applications and Challenges*. InTech. https://doi.org/10.5772/61964

Kunkel, T. A., & Erie, D. A. (2005). DNA mismatch repair. *Annual Review of Biochemistry*, *74*(1), 681–710. https://doi.org/10.1146/annurev.biochem.74.082803.133243

Lande, R. (1994). Risk of population extinction from fixation of new deleterious mutations. *Evolution*, *48*(5). https://doi.org/10.1111/j.1558-5646.1994.tb02188.x

LeClerc, J. E., Li, B., Payne, W. L., & Cebula, T. A. (1996). High mutation frequencies among *Escherichia coli* and *Salmonella* pathogens. *Science*, *274*(5290), 1208–1211. https://doi.org/10.1126/science.274.5290.1208

Lee, H., Doak, T. G., Popodi, E., Foster, P. L., & Tang, H. (2016). Insertion sequence-caused large-scale rearrangements in the genome of *Escherichia coli*. *Nucleic Acids Research*, *44*(15), 7109–7119. https://doi.org/10.1093/nar/gkw647

Lee, H., Popodi, E., Foster, P. L., & Tang, H. (2014). Detection of structural variants involving repetitive regions in the reference genome. *Journal of Computational Biology*, *21*(3), 219–233. https://doi.org/10.1089/cmb.2013.0129

Lee, H., Popodi, E., Tang, H., & Foster, P. L. (2012). Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(41), E2774–E2783. https://doi.org/10.1073/pnas.1210309109

Lenski, R. E., Rose, M. R., Simpson, S. C., & Tadler, S. C. (1991). Long-term experimental evolution in *Escherichia coli* - Adaptation and divergence during 2,000 generations. *The American Naturalist*, *138*(6). https://doi.org/10.1086/285289

Lenski, R. E., & Travisano, M. (1994). Dynamics of adaptation and diversification: A 10,000-generation experiment with bacterial populations. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(15), 6808–6814. https://doi.org/10.1073/pnas.91.15.6808

Liu, B., Eydallin, G., Maharjan, R. P., Feng, L., Wang, L., & Ferenci, T. (2017). Natural *Escherichia coli* isolates rapidly acquire genetic changes upon laboratory domestication. *Microbiology (United Kingdom)*, *163*(1), 22–30. https://doi.org/10.1099/mic.0.000405

Luria, S. E., & Delbrück, M. (1943). Mutations of bacteria from virus sensitivity to virus resistance. *Genetics, 28*(6), 491–511.

Lynch, M., *et al*. (1999). Perspective: Spontaneous deleterious mutation. *Evolution*, *53*(3). https://doi.org/10.1111/j.1558-5646.1999.tb05361.x

Lynch, M., Conery, J., & Burger, R. (1995). Mutation accumulation and the extinction of small populations. *The American Naturalist*, *146*(4). https://doi.org/10.1086/285812

Lynch, M., & Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer.

Markell, J. A. (2017). Accumulation of Single Nucleotide Polymorphism (SNP) Mutations in *Escherichia coli* Grown Under Food Production Conditions and Their Importance in Outbreak Strain Epidemiology.

Matic, I., *et al*. (1997). Highly variable mutation rates in commensal and pathogenic *Escherichia coli*. *Science*, *277*(5333), 1833–1834. https://doi.org/10.1126/science.277.5333.1833

McCoy, J. W. (1979). The origin of the "adaptive landscape" concept. *The American Naturalist*, *113*(4). https://doi.org/10.1086/283418

Mignardi, M., & Nilsson, M. (2014). Fourth-generation sequencing in the cell and the clinic. *Genome Medicine*, *6*(4), 1–4. https://doi.org/10.1186/gm548

Miller, W. J., & Capy, P. (2004). Mobile genetic elements as natural tools for genome evolution. *Methods in molecular biology* (260), 1–20. Humana Press. https://doi.org/10.1385/1-59259-755-6:001

Mukai, T. (1964). The genetic structure of natural populations of *Drosophila melanogaster* spontaneous mutation rate of polygenes controlling viability. *Genetics*, *50*(1), 1–19. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1210633/

Mukai, T., Chigusa, S. I., Mettler, L. E., & Crow, J. F. (1972). Mutation rate and dominance of genes affecting viability in *Drosophila melanogaster*. *Genetics*, *72*(2), 335–355. http://www.ncbi.nlm.nih.gov/pubmed/4630587

Nadon C., *et al*. (2017). PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. *Euro Surveill*, *22*(23), 305-344. https://doi.org/10.2807/1560-7917.ES.2017.22.23.30544

Nagarajan, N., & Pop, M. (2013). Sequence assembly demystified. *Nature Reviews Genetics*, *14*(3). https://doi.org/10.1038/nrg3367

Nataro, J. P., & Kaper, J. B. (1998). Diarrheagenic *Escherichia coli*. *Clinical Microbiology Reviews*, *11*(1), 142–201. American Society for Microbiology. https://doi.org/10.1128/cmr.11.1.142

Nyrén, P., Pettersson, B., & Uhlén, M. (1993). Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Analytical Biochemistry*, *208*(1), 171–175. https://doi.org/10.1006/abio.1993.1024

Ochman, H. (2003). *Neutral Mutations and Neutral Substitutions in Bacterial Genomes*. https://doi.org/10.1093/molbev/msg229

Ohlsson, J., *et al*. (2005). Structure-activity relationships of galabioside derivatives as inhibitors of *E. coli* and *S. suis* adhesins: Nanomolar inhibitors of *S. suis* adhesins. *Organic and Biomolecular Chemistry*, *3*(5), 886–900. https://doi.org/10.1039/b416878j

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature*. https://www.nature.com/articles/246096a0.pdf?origin=ppub

Orr, H. A. (2000). Adaptation and the cost of complexity. *Evolution*, *54*(1). https://doi.org/10.1111/j.0014-3820.2000.tb00002.x

Orskov, F., & Orskov, I. (1992). *Escherichia coli* serotyping and disease in man and animals. *Canadian Journal of Microbiology*, *38*(7), 699-704.

https://doi.org/10.1139/m92-115

Poduri, A., Evrony, G. D., Cai, X., & Walsh, C. A. (2013). Somatic mutation, genomic variation, and neurological disease. *Science*, 341(6141). American Association for the Advancement of Science. https://doi.org/10.1126/science.1237758

Pushkarev, D., Neff, N. F., & Quake, S. R. (2009). Single-molecule sequencing of an individual human genome. *Nature Biotechnology*, *27*(9). https://doi.org/10.1038/nbt.1561

Rapley, R., & Stuart Harbron. (2012). *Molecular Analysis and Genome Discovery Second edition* (2nd ed.). Wiley-Blackwell. www.wiley.com/wiley-blackwell.

Reinhardt, J. A., Baltrus, D. A., Nishimura, M. T., Jeck, W. R., Jones, C. D., & Dangl, J. L. (2008). *De novo* assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae*. *Genome Research*, *19*(2). https://doi.org/10.1101/gr.083311.108

Roberts, R. J., Carneiro, M. O., & Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome Biology*, *14*(7), 405. https://doi.org/10.1186/gb-2013-14-7-405

Ronaghi, M., Uhlén, M., & Nyrén, P. (1998). DNA sequencing: A sequencing method based on real-time pyrophosphate. *Science*, *17*(5375), 363–365. https://doi.org/10.1126/science.281.5375.363

Ronald, J., Tang, H., & Brem, R. B. (2006). Genomewide evolutionary rates in laboratory and wild yeast. *Genetics*, *174*(1), 541–544. https://doi.org/10.1534/genetics.106.060863

Rosche, W. A., & Foster, P. L. (2000). Determining mutation rates in bacterial populations. *Methods*, *20*(1), 4–17. https://doi.org/10.1006/meth.1999.0901

Röscheisen, C., Haupter, S., Zechner, U., & Speit, G. (1994). Characterization of spontaneous and induced mutations in SV40-transformed normal and Ataxia Telangiectasia cell lines. *Somatic Cell and Molecular Genetics*, *20*(6), 493–504. https://doi.org/10.1007/BF02255840

Rothberg, J. M., *et al*. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, *475*(7356), 348–352. https://doi.org/10.1038/nature10242

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12). https://doi.org/10.1073/pnas.74.12.5463

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, *19*(R2), R227–R240. https://doi.org/10.1093/hmg/ddq416

Schaechter, M. (2001). *Escherichia coli* and *Salmonella* 2000: The view from here. *Microbiology and Molecular Biology Reviews*, *65*(1), 119–130. https://doi.org/10.1128/mmbr.65.1.119-130.2001

Schroeder, J. W., Yeesin, P., Simmons, L. A., & Wang, J. D. (2018). Sources of spontaneous mutagenesis in bacteria. *Critical Reviews in Biochemistry and Molecular*

Biology, *53*(1), 29–48. Taylor and Francis Ltd. https://doi.org/10.1080/10409238.2017.1394262

Schwander, T., & Crespi, B. J. (2009). Twigs on the tree of life? Neutral and selective models for integrating macroevolutionary patterns with microevolutionary processes in the analysis of asexuality. *Molecular Ecology*, *18*(1). https://doi.org/10.1111/j.1365-294X.2008.03992.x

Schwartz, D. C., & Cantor, C. R. (1984). Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell*, *37*(1), 67–75. https://doi.org/10.1016/0092-8674(84)90301-5

Shapiro, J. A. (1979). Molecular model for the transposition and replication of bacteriophage Mu and other transposable elements. *Proceedings of the National Academy of Sciences of the United States of America*, *76*(4), 1933–1937. https://doi.org/10.1073/pnas.76.4.1933

Smith, N. G. C., & Eyre-Walker, A. (2002). Adaptive protein evolution in *Drosophila*. *Nature*, *415*(6875), 1022–1024. https://doi.org/10.1038/4151022a

Sniegowski, P D, & Lenski, R. E. (1995). Mutation and adaptation: The directed mutation controversy in evolutionary perspective. *Annual Review of Ecology and Systematics*, *26*(1). https://doi.org/10.1146/annurev.es.26.110195.003005

Sniegowski, Paul D., Gerrish, P. J., Johnson, T., & Shaver, A. (2000). The evolution of mutation rates: Separating causes from consequences. *BioEssays*, *22*(12). https://doi.org/10.1002/1521-1878(200012)22:12<1057::AID-BIES3>3.0.CO;2-W

Sniegowski, Paul D., Gerrish, P. J., & Lenski, R. E. (1997). Evolution of high mutation rates in experimental populations of *E. coli*. *Nature*, *387*(6634), 703–705. https://doi.org/10.1038/42701

Southern, E. M. (2001). DNA microarrays. History and overview. *Methods in molecular biology*, *170*, 1–15. Humana Press. https://doi.org/10.1385/1-59259-234-1:1

Stavropoulos, D. J., *et al*. (2016). Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine. *Npj Genomic Medicine*, *1*(1), 1–9. https://doi.org/10.1038/npjgenmed.2015.12

Treangen, T. J., Abraham, A.-L., Touchon, M., & Rocha, E. P. C. (2009). Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiology Reviews*, *33*(3), 539-571. https://doi.org/10.1111/j.1574-6976.2009.00169.x

Trindade, S., Perfeito, L., & Gordo, I. (2010). Rate and effects of spontaneous mutations that affect fitness in mutator *Escherichia coli*. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1544), 1177–1186. https://doi.org/10.1098/rstb.2009.0287

Valouev, A., *et al*. (2008). A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Research*, *18*(7), 1051–1063. https://doi.org/10.1101/gr.076463.108

Vassilieva, L. L., & Lynch, M. (1999). The rate of spontaneous mutation for life-history traits in *Caenorhabditis elegans*. *Genetics*, *151*(1), 119-129.

Williams, A. B. (2014). Spontaneous mutation rates come into focus in *Escherichia coli*. *DNA Repair*, *24*, 73–79. https://doi.org/10.1016/j.dnarep.2014.09.009

Wirth, T., *et al*. (2006). Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Molecular Microbiology*, *60*(5). https://doi.org/10.1111/j.1365-2958.2006.05172.x

Wiser, M. J., Ribeck, N., & Lenski, R. E. (2013). Long-term dynamics of adaptation in asexual populations. *Science*, *342*(6164). https://doi.org/10.1126/science.1243357

Wong, A., Rodrigue, N., & Kassen, R. (2012). Genomics of adaptation during experimental evolution of the opportunistic pathogen *Pseudomonas aeruginosa*. *PLoS Genetics*, *8*(9), e1002928. https://doi.org/10.1371/journal.pgen.1002928

Wright, S. (1940). Breeding structure of populations in relation to speciation. *The American Naturalist*, *74*(752), 232–248. https://doi.org/10.1086/280891

Zeyl, C., Mizesko, M., & De Visser, J. A. G. M. (2007). Mutational meltdown in laboratory yeast populations. *Evolution*, *55*(5). https://doi.org/10.1111/j.0014-3820.2001.tb00608.x

Appendix

2
3
4
5
6 **Hybrid Nanopore-Illumina assemblies for five extra-intestinal pathogenic**
7 ***Escherichia coli* isolates**
8
9 Mattrasingh D[1], Hinz A[1], Phillips L[1], Carroll AC[1], and Wong A[1*]
10
11 [1]Department of Biology, Carleton University, Ottawa, Canada
12 *Corresponding author: alex.wong@carleton.ca
13
14
2
15 **Abstract**
16
17 Extra-intestinal pathogenic *Escherichia coli* (ExPEC) are important sources of multi-drug
18 resistant infections, particularly in hospitals. We report hybrid Nanopore-Illumina
19 assemblies for 5 ExPEC isolates with varying drug resistance profiles.
20
21 **Announcement**
22
23 Extra-intestinal pathogenic *Escherichia coli* (ExPEC) cause serious illnesses, including blood
24 and urinary tract infections (UTIs), and have a wide variety of antibiotic resistance profiles
25 (1). Five ExPEC isolates with varying degrees of documented antibiotic resistance were
26 obtained from the Zhanel laboratory, and were collected as part of the CANWARD survey of
27 antibiotic resistant pathogens in Canada (2, 3). Two isolates were derived from UTIs (PB3
28 and PB4), two from blood infections (PB29 and PB35), and one from a respiratory infection
29 (PB33).
30 DNA was extracted from overnight cultures of each strain using the One-4-All Genomic
31 DNA Miniprep kit (BioBasic, Markham). Short reads were generated on the Illumina
32 NextSeq platform using 150bp paired-end reads with Nextera XT library preparation,
33 generating a total of 162,925,724 clusters passing filter. Long reads were generated on the

Nanopore MinION platform using the rapid barcoding kit, generating a total of 697,371 reads. Quality scoring before and after trimming was performed using FASTQC v0.11.7 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and Illumina reads were trimmed using Trimmomatic v.0.38 (4) with parameters LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36. Porechop v0.2.4 (https://github.com/rrwick/Porechop) was used to remove adapter sequences from the Nanopore reads. Hybrid assemblies were generated using the SPAdes-optimiser Unicycler v0.4.8 (5). Quast v4.6 (6) was used to produce summary statistics for the assembly. Genome annotations were carried out with the NCBI Prokaryotic Genome Annotation Pipeline v4.12 (7), serotypes were predicted using SeroTypeFinder v2.0.1 (8), and AMR profiles were predicted with ResFinder v3.1 (9). Genomes varied between 4,880,873bp (PB4) and 5,309,474bp (PB29) in length, and maintained a %GC between 50.6 and 50.85 (Table 1). Strains harboured similar numbers of CDS, rRNA, and tRNA (Table 2). PB4 and PB29 carry two predicted CRISPR sequences each, while the others have none. Strains carried between 0 and 6 chromosomal resistance mutations, and between 1 and 14 resistance genes (Table 2). PB33 and PB35 both belong to serotype O25:H4-ST131, a major epidemic clone of ExPEC (10, 11).

3

We have provided *de novo* hybrid genome assemblies of five extra-intestinal pathogenic *E. coli* isolates. Three of the isolates exhibit a genomic signature of multidrug resistance characterized by 14 or more resistance mutations/genes (PB29, PB33, and PB35). The remaining two had relatively fewer resistance determinants (PB3 and PB4). These genomes will support future investigations in *E. coli* pathogenesis and resistance evolution.

Data Availability

This sequencing project has been deposited in GenBank under BioProject PRJNA648312, with accession numbers JACFYA000000000 - JACFYE000000000. The versions described

in this paper are the first versions, JACFYA010000000 - JACFYE010000000.

References

1. Dale AP, Woodford N. 2015. Extra-intestinal pathogenic Escherichia coli (ExPEC): Disease, carriage and clones. J Infect 71:615-26.

2. Zhanel GG, Adam HJ, Baxter MR, Fuller J, Nichol KA, Denisuik AJ, Lagace-Wiens PR, Walkty A, Karlowsky JA, Schweizer F, Hoban DJ. 2013. Antimicrobial susceptibility of 22746 pathogens from Canadian hospitals: results of the CANWARD 2007-11 study. J Antimicrob Chemother 68 Suppl 1:i7-22.

3. Basra P, Alsaadi A, Bernal-Astrain G, O'Sullivan ML, Hazlett B, Clarke LM, Schoenrock A, Pitre S, Wong A. 2018. Fitness tradeoffs of antibiotic resistance in extra-intestinal pathogenic Escherichia coli. Genome Biol Evol 10:667-679.

4. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-20.

5. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595.

6. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072-5.

7. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res 44:6614-24.

8. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. 2015. Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data. J Clin Microbiol 53:2410-26.

9. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 67:2640-4.

10. Coque TM, Novais A, Carattoli A, Poirel L, Pitout J, Peixe L, Baquero F, Canton R, Nordmann P. 2008. Dissemination of clonally related Escherichia coli strains expressing extended-spectrum beta-lactamase CTX-M-15. Emerg Infect Dis 14:195-200.

89 11. Nicolas-Chanoine MH, Blanco J, Leflon-Guibout V, Demarty R, Alonso MP, Canica MM,
90 Park YJ, Lavigne JP, Pitout J, Johnson JR. 2008. Intercontinental emergence of 4
Escherichia coli clone O25:H4-91 ST131 producing CTX-M-15. J Antimicrob Chemother
92 61:273-81.
93

Table 1. Assembly characteristics for 5 ExPEC isolates.

Strain	Total length	Length in contigs >500bp	#contigs	N50	Largest contig
PB3	5,224,675	5,220,584	48	4,043,831	4,043,831
PB4	4,880,873	4,880,309	23	688,147	1,168,174
PB29	5,309,474	5,298,387	55	473,125	1,155,588
PB33	5,168,956	5,168,377	27	710,413	1,099,130
PB35	5,236,721	5,236,154	41	497,725	871,437

Table 2. Annotation of 5 ExPEC genomes.

Strain	#protein CDS[1]	#rRNA (complete)	#tRNA	#CRISPR	#resistance genes	#resistance mutations	ESBL[2]	Serotype
PB3	4,793	22	89	0	2	0	None	O6:H1
PB4	4,414	20	84	2	1	1	None	O-:H9
PB29	4,909	14	90	2	13	0	CTX-M-15	O-:H9
PB33	4,772	21	83	0	8	6	CTX-M-15	O25:H4
PB35	4,792	21	83	0	14	5	CTX-M-15	O25:H4

[1]Predicted protein coding sequences, excluding pseudogenes

[2]ESBL: Type (if any) of extended-spectrum beta-lactamase gene present in genome.