Mutation rates of *Escherichia coli* with different balanced growth rates: a new fluctuation test protocol and phenotypic lag adjustments

by

Christian Terry Henderson Barna

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Mathematics
in
Applied Mathematics

Waterloo, Ontario, Canada, 2020

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Bacteria are the oldest, most abundant life form on the planet, and every other organism's livelihood is dependent on them. The bacteria *Escherichia coli* (*E. coli*) is commonly used in microbiology as a model organism to give insight into the functions of bacteria and cells in general. Of particular interest in these studies is the methods with which bacteria grow and evolve. Growth is what propagates a bacteria's species; whereas evolution is what allows them to adapt to the ever-changing world. Evolution is made possible by mutations which change a bacterium's DNA. In 1943, Luria and Delbrück developed a method, called a "fluctuation test", to estimate mutation rates from the number of mutants in a collection of parallel cultures exposed to a selecting agent after growth. The original fluctuation test methodology suffers from two major limitations. First, the bacteria are not in a reproducible, balanced state of growth throughout the test. Second, the new phenotype resulting from a mutation may not be immediately expressed (referred to as "phenotypic lag") resulting in an underestimated mutation rate. To overcome these issues, I developed a refined experimental protocol that ensures cells are in balanced growth and a suite of analysis tools that account for the effects of phenotypic lag. To test the methodology, I compared the mutation rate and phenotypic lag in fast growing *E. coli* (23 minutes per doubling) and slow growing *E. coli* (48 minutes per doubling). It is found that when not accounting for phenotypic lag, fast growing *E. coli* have a markedly lower mutation rate than slow growing *E. coli*, but when phenotypic lag is accounted for, the faster growing cells have a longer phenotypic lag, resulting in an indistinguishable mutation rate for fast and slow growing populations. The implications of mutation rate being coupled to growth rate, as well as possible explanations for why it and phenotypic lag would be growth rate dependent are discussed. Finally, possible ways to improve the experimental methodology and analysis protocols, in addition to future experiments that can be performed to further explore mutation rate - growth rate coupling are proposed.

Acknowledgements

All research for this thesis was completed on the lands of the Haudenosaunee, Anishinaabeg, and Neutral peoples [165]. I have lived in the valley of the Ose'kowáhne ("Grand River") [7] for the majority of my life, completing grades 1-12, a B.Sc., and now a Masters here. This means I have spent over 20 years reaping the benefits of being a white settler on stolen First Nation land, and for much of it I wasn't even aware. The entirety of Kitchener-Waterloo is situated on treaty land that was taken from the people it was promised to. In fact, all land within 6 miles of the "Grand River" was promised to the Haudenosaunee as part of the Haldimand proclamation in 1784 [111], but now they control less than 5% of this land due to manipulation and forceful displacement. From time immemorial the First Nations peoples have been caring for the land we call home, and for the last several hundred years they have been battling white supremacy and its destructive perception of land [39]. If we wish to have a healthy planet and just society, I believe that we have to give control of all land that is considered "North America" and beyond back to the indigenous peoples. It is possible to live sustainably on this land, but to do so we must de-colonise our minds and re-learn what our place on this planet is.

I want to thank the earth for giving me everything I need to survive and do all this science. Doing science requires resources that come from the earth, and most are extracted in an unsustainable and unethical fashion. I can confidently say that I used more plastic during this masters degree than I normally would in a decade. This has been a source for much internal conflict and has helped shape my perception of science's role in society. I think all scientists need to look at their research and seriously question if the potential results are worth the resources being consumed and the ethical implications. We desperately need a more ethical and sustainable way to do science, and if we need to take a break from actively doing science in the mean time, I think that's a fair trade.

An extra big thank you to Matt Scott. Your patience and guidance over the last 3 years is appreciated beyond belief; I truly could not have asked for a better supervisor. You have taught me so much about what it means to be a good scientist, as well as a good teacher and mentor. I always looked forward to our meetings and seeing your big smile in the halls. I sincerely hope our paths cross again in the future.

Thank you to Brian Ingalls for being much like a co-supervisor to me. Your guidance, insight, and general presence were incredibly appreciated.

Thank you Ed Vrscay for agreeing to be on my committee and making my introduction to university level mathematics nearly 10 years ago so fun.

Thank you to my parents, Lynne and Terry Barna, for your support over the last 28 years and in particular over the last 3 years. Raising me and allowing me to mooch off you for so long is so very appreciated. I couldn't have done this without you!

Thank you to Micaela Yawney, your love and support over the last year has kept me going. I'm so happy we met and have got to spend so much time getting to know each other during my masters. Here's to our future adventures together!

Thank you to my office mates. Jesse Legaspi, you are the funniest, most wonderful human who's memes and video game nights got me through. Josh Thompson, it was always nice having someone to relate to and relax with. Tim Dockhorn, thank you for lending me your brain on occasion and all the good tennis matches. Rishi Chakraborty, thank you for laughing at my jokes and always keeping an open mind. Greg Wang, thank you for your smile and giant heart.

Thank you to all my lab mates. Nate Braniff, I don't think I could ever truly communicate how much help you were to me through this degree; you provided great conversation, great ideas, great lab help, and great friendship. Nicole Wang, Linda Chan, Sarah Odinotski, Yesha Patel, Leah Fulton, Max Reed, Fatima Abrar, and Cody Receno, thank you for helping and keeping me company in the lab. Extra big shout out to everyone who helped me count plates!

Thank you to all my other friends in the Applied Math department. Lizz Webb, thank you for sharing all that hot goss and candy. Tom Bury, thank you for the conversations and racket sports games. Kat Fair, thank you for all the awesome conversations, especially about radical politics. Laura Chandler, Kate Clements, Stan Zonov, Saptarshi Pal, Sarah Walsh, Lindsey Daniels, Zahraa Abbas, Taylor Hanson, Maliha Ahmed, Alison Cheeseman, Andrew Grace, Maria Papageorgiou, Vivek Thampi, Brendon Phillips, Lauren Burnett, Giselle Sosa Jones, Luciana Chavez Rodriguez, and Ben Storer you all helped make this experience fun.

Thank you to my Outers Club friends. Tim Hill, Claire Parrott, Evan Takefman, Nicole Sos, Elif Tuzlali, and Kanishk Goomer, you all made running Outers so rewarding and fun.

Thank you to all my friends and family. Emma McKay, you are truly the best friend a person can ask for; you have taught and given me so much and for this I will forever be grateful. Shane Lawrence, your friendship and belief in me keeps me going on hard days. Seth Holland, I'm always happy when we get to chat about science and slap our hands together. Martin Bauman, I wish I got to ride on your shoulders more, but I still appreciate your friendship immensely. Rachel Barna, your support and jokes are food for my soul. Maddy MacEachern, I was gonna make an R2D2 and C3PO reference, but we're both R2D2 so that doesn't work; you get the point though. Mary Tye, you are the kindest

most wonderfulest. Alex Pearce, you are also the kindest most wonderfulest. Ilia Poichuck, you're so compassionate and fun and I am very happy we were reunited. David Paton, Will Gertler, Erin & Kyle Wagstaff, Qui Crawford, Graham Notar-Maclean, Nathan Butt, Palmer Vaughn, Saam Koukpari, Claire Leuty, Al Sachs, Ashley Beitel, Tiana Van De Veerdonk, Emily Carlson, Jackson Smith, Warren Jones, Giulia Langella, Aidan Power, Emma Raaflaub, and more, you all make being in this world so much more enjoyable.

Thank you to the D&D group that gave me something to look forward to and a world to escape to most weeks.

Finally, a big shout out to ice cream, peanut butter, and chocolate for helping me get through these wild times people call grad school.

Dedication

This thesis is dedicated to the billions and billions of bacteria who lost their lives in my selfish pursuit of knowledge.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

bp base pair 68

CDF cumulative distribution function 66

CFU colony forming units 15

CI confidence interval 63

CV coefficient of variation 104

MLE maximum likelihood estimation 63

OD optical density 51

PDF probability distribution function 21

PGF probability generating function 21

rCDF reduced cumulative distribution function 81

rCDF+K reduced cumulative distribution function and Koch 85

rCDF+K_{avg} reduced cumulative distribution function and Koch average 85

RDM rich defined media 50

SD standard deviation 69

TSS total sum of squares 66

List of Symbols

G Probability generating function (PGF): A power series built from the probabilities of a random variable with auxiliary variable z. 23

K_D Michaelis constant: A constant that is equal to the substrate concentration that gives half the maximal rate in Michaelis-Menten kinetics. 31

N_0 Initial population: The initial number of bacteria in a culture (at the time of inoculation). 7

N_f Final population: The number of bacteria in a culture at the end of an experiment. Often used interchangeably with N_t. 68

N_t Population: The total number of bacteria in a culture at time t. 7

P_0 Initial active protein number: The average amount of active protein (α-protein) that a non-mutant cell is born with. 70

S Nutrient concentration: The concentration of a growth limiting substrate. 31

η Doubling rate: The average number of doublings a bacterial culture completes in a period of time. Generally measured as doublings/hour. 7

λ Specific growth rate: The exponential growth rate of a bacterial culture. Measured as $\lambda = \eta \ln 2$. 7

$\hat{m}$ Estimated number of mutations: The estimated average number of mutations across multiple parallel cultures of bacteria after a period of growth. When presented with a subscript, the subscript denotes the fitting protocol used (i.e. $\hat{m}_{\text{MLE}}$ denotes the maximum likelihood estimated number of mutations). 62

$\hat{\mu}$ Estimated mutation rate: The estimated average spontaneous mutation rate. Also interpreted as the probability of mutation. Measured as the average number of mutations per cell per generation. When presented with a subscript, the subscript denotes the fitting protocol used (i.e. $\hat{\mu}_{\text{MLE}}$ denotes the maximum likelihood estimated mutation rate). 68

μ Mutation rate: The average spontaneous mutation rate. Also interpreted as the probability of mutation. If presented with no subscript it is measured as the average number of mutations per cell per generation. If presented with a subscript then μ_{bp} is the average number of mutations per base pair per generation, μ_{genome} is the average number of mutations per genome per generation, and μ_{cell} is the usual average number of mutations per cell per generation. 15

m Number of mutations: The average number of mutations across multiple parallel cultures of bacteria after a period of growth. Also called "mutation number". 15

$\hat{n}$ Estimated phenotypic lag length: The estimated average length of phenotypic lag in generations. When presented with a subscript, the subscript denotes the fitting protocol used (i.e. $\hat{n}_K$ denotes the Koch estimated phenotypic lag length). 77

ν Number of cultures: The number of parallel cultures used in a fluctuation test. 62

n Phenotypic lag length: The average length of phenotypic lag in generations. 70

p_r Probability of r mutants: Theoretically, the probability that a culture has r mutants. Experimentally, the proportion of cultures in a fluctuation test that have r resistant mutants. 22

ϕ Population scaling factor: A factor which scales for the difference between initial and final populations in a fluctuation test. Of the form $\phi = 1 - \frac{N_0}{N_f}$. 28

r Number of resistant bacteria: The number of bacteria expressing a specific mutant phenotype, generally antibiotic resistance, after a period of growth. 16

τ Doubling time: The average amount of time it takes for a bacteria culture to double its population. Commonly measured in minutes. 7

t Time: The amount of time since a bacterial culture has been inoculated. Commonly measured in $t = \ln(2) \cdot \text{gens}$ where "gens" represents the number of generations that the bacteria have grown on average. 7

Chapter 1

Introduction

1.1 Bacteria

Life of all forms pervades our planet, but it is bacteria which have been here the longest and are the most abundant [152, 192]. As a result, bacteria have come to play a big role in the lives of all other creatures, as well as the planet itself [192, 100]. So how did bacteria become so prevalent? Simple, they are masters of survival [123]. There are two essential aspects to survival. The first is the survival of an individual bacterium, while the second is the survival of the species through reproduction. Being masters of survival means bacteria are resilient, capable of adapting to a wide variety of environments, and are proficient at proliferation [120, 123]. These survival skills have come about through approximately 3.75 billion years of evolution [192]. In this time bacteria have come to occupy nearly every corner of the earth and grown to an estimated population of 10^{30} [192]. They have also come to have the most genetic diversity of any type of life on the planet [79], which is well communicated by the evolutionary tree in Fig. 1.1.

Bacteria are prokaryotes, meaning they are single celled organisms with all of their major components floating around together inside a single wall[1] [163]. Essentially, bacteria are tiny bags of salt water primarily filled with deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and proteins [163, 12, 152]. The "bag", often called a membrane, is made of lipids and most commonly comes in one of two forms, Gram-negative and Gram-positive [123]. The difference between these forms is that Gram-negative bacteria have two layers

[1] As opposed to eukaryotes, which have internal cellular organisation in the form of a membrane-bound nucleus and organelles, and can combine into multicellular organisms [152].

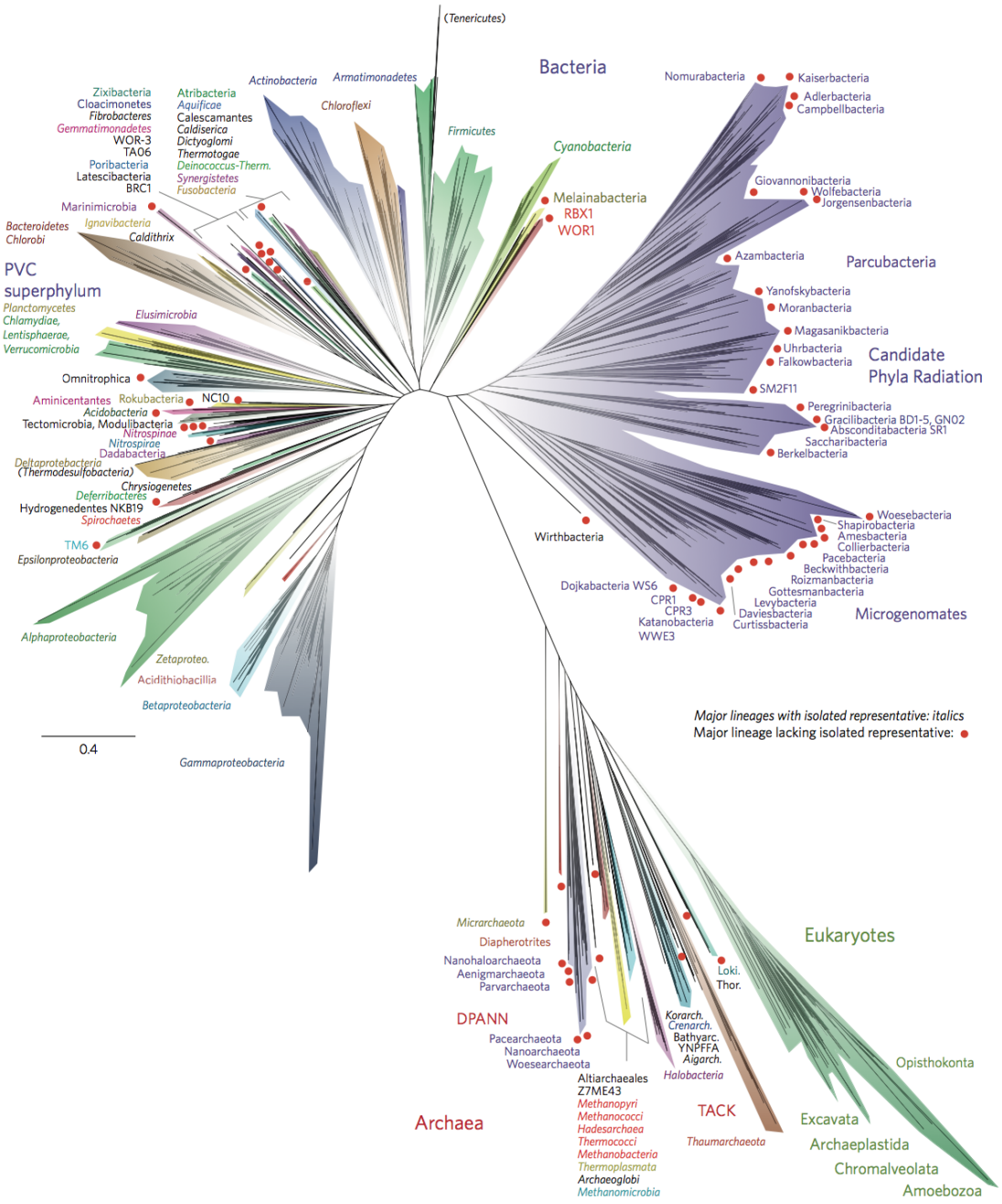

Figure 1.1: **Evolutionary tree emphasising genetic diversity.** An evolutionary tree of life which emphasises how much genetic diversity there is among bacteria versus other kingdoms. Figure 1 from Hug et al. (2016) / CC BY 4.0 [79].

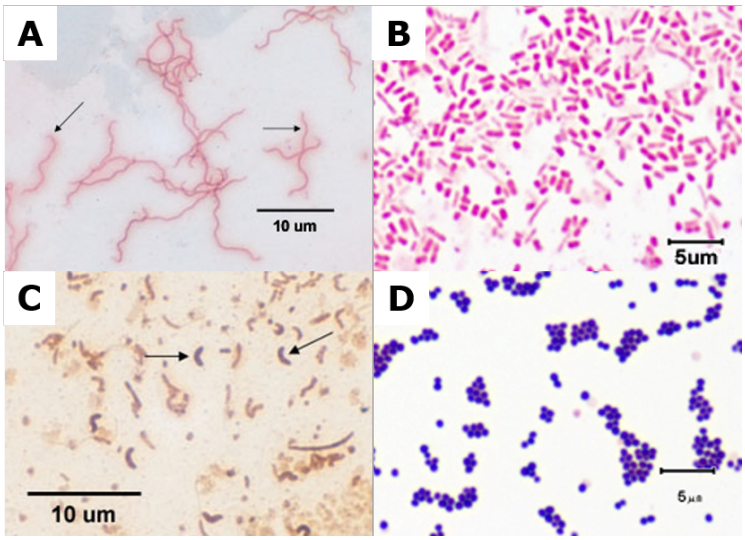

Figure 1.2: **Common bacteria shapes.** A) Spiral shaped bacteria (*Borrelia burgdorferi*). B) Gram-negative rod shaped bacilli bacteria (*Escherichia coli*). C) Kidney bean shaped vibrio bacteria (unknown *Vibrio*). D) Gram-positive spherical cocci bacteria (*Staphylococcus aureus*). (A) "Spirochete", (B) "Single Rod (bacillus)", (C) "Vibrio", and (D) "Tetrad Arrangement: Direct Stain" by Gary E. Kaiser / CC BY 3.0 [85].

of lipids while Gram-positive bacteria only have one. The name "Gram" comes from a stain which will turn Gram-positive bacteria purple and Gram-negative bacteria pink (D versus B in Fig. 1.2). Bacteria come in all sorts of shapes, sizes, and compositions. Common shapes for bacteria are rods, spheres, spirals, and something resembling a kidney bean (Fig. 1.2) [85, 163]. Size wise, bacteria have a huge range, but being smaller is beneficial because it increases their surface area to volume ratio, making diffusion of food through their membrane more efficient [192]. The smallest known bacterium, *Mycoplasma genitalium*, is about 0.2μm long, which is the size of a large virus, while the largest known bacterium, *Thiomargarita namibiensis*, can be up to 0.75mm long, which is large enough to see with the naked eye (Fig. 1.3) [192, 163].

Beyond speculation, humans knew nothing of bacteria until we invented the microscope [24]. Despite this, humans are dependent on bacteria for survival through the billions of bacteria that live in and on our bodies [161, 187]. Even though bacteria clearly live in the same ecosystems as us, due to their size, the physics they experience is foreign to us and dominated by stochastic effects [12]. This makes relating to bacteria and understanding them on an intuitive level difficult. Fortunately, one important repercussion of bacteria's

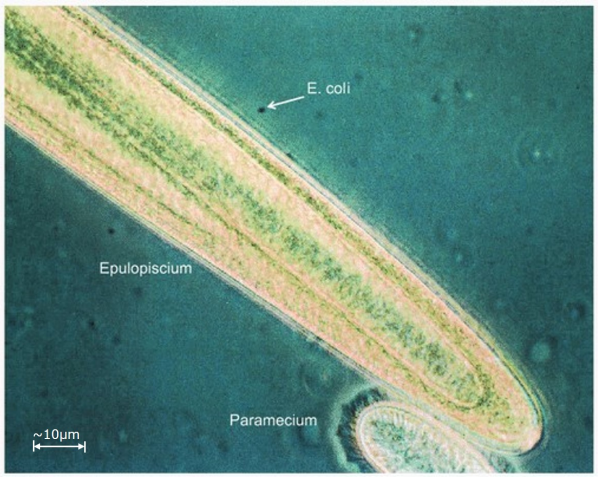

Figure 1.3: **Size comparison of different microbes.** *E coli* and *Epulopiscium* are bacteria while *Paramecium* is a eukaryotic algae. Figure 9 from Lane (2017) / CC BY 4.0 [98].

size is that they're not capable of having very many parts, at least compared to other life forms [158]. Put together with the fact that bacteria have had so long to evolve into efficient machines [123], it can be argued that they are the simplest life form on the planet. This relative simplicity facilitates the creation of meaningful quantitative descriptions for their functions and behaviours.

1.2 *Escherichia Coli*

Escherichia coli (*E. coli*) are a species of bacteria that are commonly found in the large intestines of warm-blooded creatures [180, 17] and often used for studies in the biological sciences because of their rapid growth rate and simple nutritional requirements [154]. *E. coli* are Gram-negative, rod shaped, and their size is on the order of a micrometer (see Figures 1.4 and 1.5) [154].

An *E. coli*'s ideal environment is a 37°C, pH neutral, aerated liquid full of glucose, nucleotides, amino acids, and trace amounts of a variety of elements [121]. In this ideal environment, *E. coli* are capable of doubling their population every 22 minutes [123]. In less ideal environments, this rate of population growth can decrease to the point where the bacteria stop growing and use what little resources they have to maintain themselves in the hopes that more, better resources will arrive later [123]. To grow, at a bare minimum

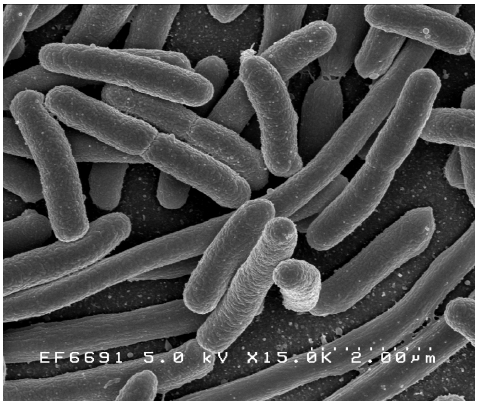

Figure 1.4: **Electron micrograph of *E. coli*.** Note the scale bar at the bottom right. "E. coli Bacteria" from the National Institute of Allergy and Infectious Diseases / CC BY 2.0 [131].

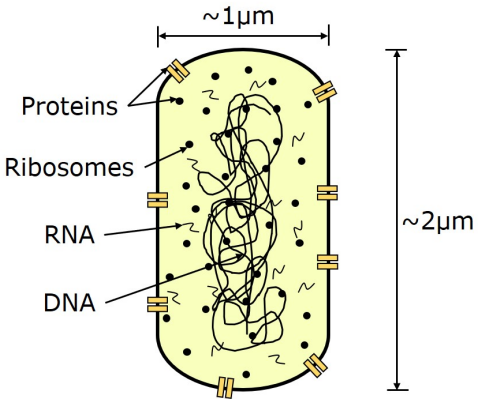

Figure 1.5: **Toy model of a bacillus bacterium such as *Escherichia coli*.**

they need a complex carbon source, nitrogen, phosphorous, sulphur and trace amounts of several different metals [121].

E. coli has been used in the lab for over a century, with its scientific journey starting when Theodor Escherich isolated it from infants with diarrhea in 1886 [3]. Scientists have studied it continuously since, resulting in us probably understanding it better than any other living thing on the planet [154, 91, 17]. So how did *E. coli* come to be studied with such vigour? In the early days of bacteriology scientists looked for easily accessible species that weren't overly virulent, grew quickly in a variety of commonly used medias, and were easily identifiable [154, 3]. *E. coli* checked all those boxes. The most common *E. coli* strains

used in the lab are ancestors of a strain called K-12. *E. coli* K-12 was first isolated in 1922 from the stool of a convalescent diphtheria patient and started being used for experiments not long after [13]. In 1947 Edward L. Tatum and Joshua Lederberg discovered that the K-12 strain was capable of conjugation [174], which was the first time the sexual transfer of genes was observed in prokaryotes [13, 3]. Since most *E. coli* strains found in nature are not capable of performing conjugation [13, 3], this boosted the popularity of K-12 and solidified *E. coli* as the bacteria of choice in microbiology [154]. Furthermore, due to the environment that *E. coli* evolved in, they are versatile creatures able to survive on many different foods and adapt quickly to changes in their environment [3]. This makes *E. coli* ideal for the lab because it allows one to easily test many different scenarios. The dedicated use of *E. coli* in scientific exploration has led to many genetic engineering protocols that uniquely work on the species, allowing for creative science that has given us things like insulin producing bacteria, which are microscopic factories that cultivate a life saving drug for humans [64].

Due to the long time continuous use of *E. coli* in the lab, we have in many ways domesticated the strains we have come to know the most about. This means popular lab strains have adapted to their comfortable lab lives and may not behave similarly to most of the *E. coli* one would find in nature [50, 105]. Furthermore, there is so much diversity among bacteria that how a specific strain works most certainly is not how another does, even within a single species. So how do we validate only studying a handful of mostly domesticated *E. coli* strains? To answer this, we must first ask the question, how do we expect to understand any life on a cellular level if we don't first understand to the best of our abilities how one specific species of bacteria behaves? The deep understanding of one species not only gives us intuition on how better to study and understand other organisms, but also gives us something to compare to. As such, all data and concepts moving forward will specifically be in the context of *E. coli* unless stated otherwise.

1.3 Bacterial Growth

Bacteria's entire existence is centred around growth. They're either growing, or they're waiting for the right conditions to grow. This is because growth *is* reproduction for them, so it's necessary for the propagation of their species [123].

E. coli reproduce through binary fission, which is a type of asexual reproduction that results in a bacterium splitting into two bacteria [152]. Accordingly, from the time a bacterium is born, it spends its whole life growing to double its original size, building up all of its components, including its genome, to double its original amount, so when it splits

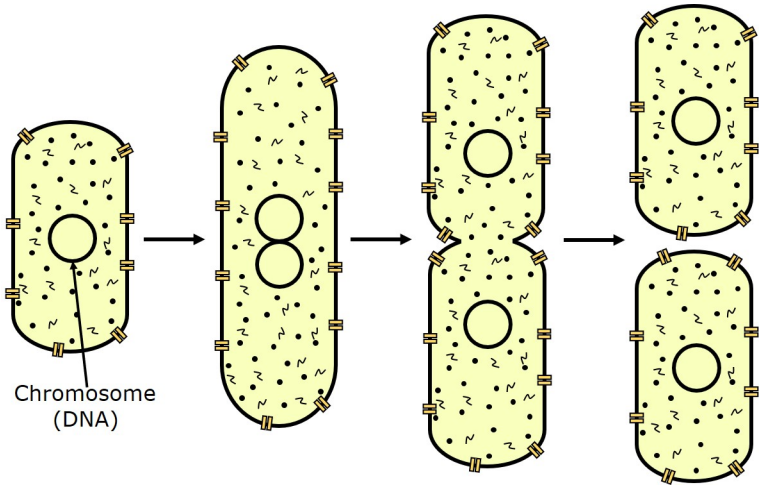

Figure 1.6: ***E. coli* replication toy model.** The bacterium lengthens and increases all of its components to double the original amount. Once the chromosome is fully replicated, the two chromosomes are moved to opposite ends and the cell septates in the middle, giving two complete bacteria.

it can create two nearly identical, fully functioning children (Fig. 1.6) [34]. The cycle is then repeated by the two children, giving four bacteria, who then double to eight bacteria, and so on. Mathematically this looks like,

$$N_t = N_0 2^{t/\tau} = N_0 2^{\eta t} \,, \tag{1.1}$$

where N_t is the total number of bacteria at time t, N_0 is the initial number of bacteria, τ is the doubling time, and η is the doubling rate [34]. Since exponentials follow a well defined set of rules, we can define a more mathematically convenient growth rate,

$$\lambda = \eta \ln 2 \,, \tag{1.2}$$

which allows the growth to be represented using the natural base e,

$$N_t = N_0 e^{\lambda t} \,, \tag{1.3}$$

where λ is called the specific growth rate [159]. When the derivative of Eq. (1.3) is taken with respect to t, we get,

$$\frac{dN_t}{dt} = \lambda N_t \,, \tag{1.4}$$

which is a famously simple differential equation. It also helps clarify why λ is considered the growth rate because the equation says that the rate of change in population over

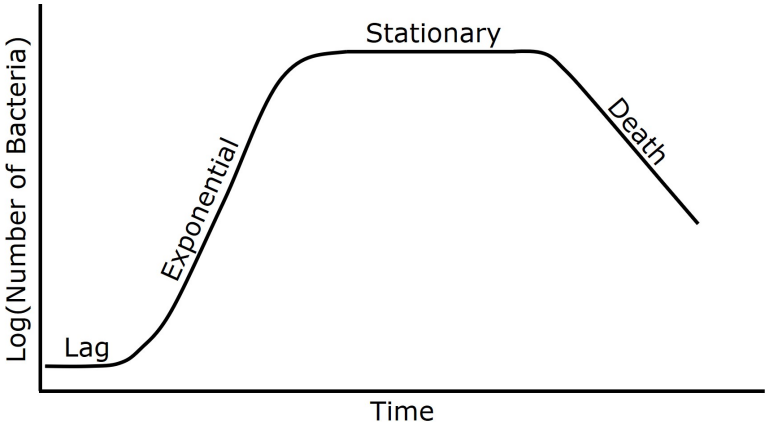

Figure 1.7: **The growth curve of a bacteria culture.** The culture in this plot is seeded from bacteria in stationary phase or from a different media. The curve represents the life cycle of a newly introduced bacteria culture in a limited medium.

time equals the current population times the growth rate. Consequently, unless stated otherwise, when people talk about bacterial growth they are generally referring to the population average growth instead of the growth dynamics of a single bacterium.

When bacteria are introduced to a new environment that contains the nutrients they need for growth, the population starts off in a state of no growth called *lag phase*, where the cells are building the necessary machinery for growth in that particular environment and increasing their size [117, 90]. The bacteria eventually start doubling and quickly accelerate the rate at which they replicate until they reach a constant exponential rate. The bacteria maintain this constant exponential rate of growth, called the log or *exponential phase*, until they start to run out of one of the necessary nutrients or the concentration of toxic waste gets too high, at which point they decelerate their net growth to zero[2]. The growthless phase is called the *stationary phase*, during which the bacteria simplify their machinery, shrink in size, and focus on maintaining homeostasis. When left in stationary phase for a long time, the bacteria eventually lose the ability to maintain homeostasis and start to die, which is called the death phase. If cells in the stationary or death phase are re-introduced to a medium sufficient for exponential growth, the process restarts and the cells enter lag phase. The pattern of growth phases, often referred to as the *growth curve*, is graphically represented in Fig. 1.7 [117]. The growth curve was considered such an essential part of the study of bacterial growth during the early years of microbiology that in a 1949 review,

[2]It is possible that not all cells stop growing, but instead some cells continue to grow more slowly, while others die, and others halt growth all together, which leads to net zero population growth.

Cornelius Bernardus van Niel stated that "nearly all that is known about the kinetics of growth of microorganisms has been learned from studies of so-called growth curves" [186].

In 1939 Alfred Hershey changed the field of microbiology by using cells that were already in exponential phase to inoculate a culture [76]. Hershey saw that the bacteria skipped the lag phase and continued to grow as they do in exponential phase when he did this, sparking a new era of research where the exponential phase of bacteria was the primary focus. Because bacteria grow at their maximal rate possible for the entirety of the exponential phase, this allows for reproducible growth [150]. In 1957, Allan Campbell communicated the importance of these attributes when he referred to the growth that takes place in the exponential phase as "balanced growth", because *all* cell constituents double at the same rate [27]. The significance of this logic is made clear by Moselio Schaechter, who said that "moving from the observation of log phase to the concept of balanced growth is like going from watching apples fall to thinking of gravity" [150]. Another term used to describe the growth that often takes place in the exponential phase is "steady state growth" because the distributions of cell attributes in a culture are time independent [134, 84]. When bacterial growth is discussed throughout this thesis, assume it is balanced growth unless stated otherwise.

1.4 Mutation

The mechanism that allows life to evolve is mutation [54, 66, 67]. Many even consider the capability of mutation a defining attribute of life, on the same level as having a cell wall [146]. With a basic understanding of what mutation is, it is easy to see that without it, life as we know it would not exist because it would not have the ability to adapt and change. So what exactly is a mutation? To answer this we must first understand what Deoxyribonucleic acid (DNA) is. DNA forms the chromosome, which codes for the nature of a living thing and is often referred to as the "blueprint of life" [152, 72]. The chemical structure of DNA is what allows it to complete this task [193, 60, 196]. DNA is a chemical with a double helix structure composed of hydrogen, oxygen, nitrogen, carbon, and phosphorous [140]. The phosphorous combines with sugars to build a "backbone" that holds the nucleobases in the centre, all together forming a nucleotide (Fig. 1.8). There are four distinct nucleobases that are used to form nucleotides, which are placed in sequences to code information into DNA. These nucleobases are adenine (A), cytosine (C), guanine (G), and thymine (T). Hydrogen bonds between opposing nucleobases are what holds the two halves of DNA together and promotes the preferred pairings of adenine with thymine, and cytosine with guanine [140]. The preferred pairings allow one to have all the information

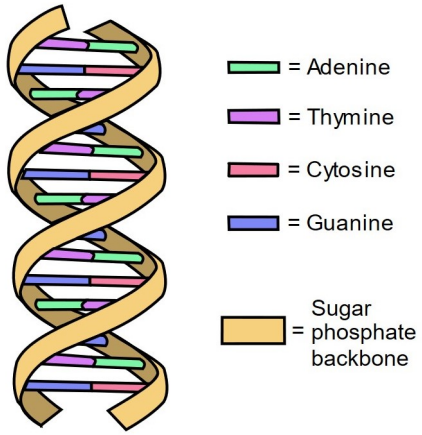

Figure 1.8: **Basic model of DNA with shape and structure.** Adenine, thymine, cytosine, and guanine are nucleobases, the order of which codes for proteins and other cell functions. "Simple diagram of double-stranded DNA" by Forluvoft [56].

held within a strand of DNA even when only in possession of one half of the strand; an important attribute for DNA replication. The ordering of the nucleobases is a base four information system reminiscent of the base two system of computers [61, 74]. A mutation is whenever this ordering changes [11].

A mutation can occur through three different types of alterations to the DNA: base substitution, insertion, and deletion [152]. A base substitution is when one or multiple of the nucleotides are traded out for nucleotides with different nucleobases, generally resulting in a less favourable hydrogen bond. When just one nucleotide is substituted, it is called a point substitution, and this is the most common mutation type [101]. An insertion is when a new nucleotide, or sequence of nucleotides are inserted into the DNA. A deletion is when a nucleotide or sequence of nucleotides are deleted from the DNA. See Fig. 1.9 for a visualisation of how these different types of mutations affect a gene.

How is the information held within the DNA's sequences of nucleobases converted into what we see as life? Bacteria's functions can largely be reduced to a complex series of chemical reactions which are catalysed by proteins made of amino acids coded for in the DNA [123]. The DNA codes for proteins by having a sequence of three base pairs uniquely represent an amino acid so that a sequence of DNA, or a gene, represents an ordered compilation of amino acids that when combined fold into proteins [152]. The DNA is converted to these amino acid compilations, called polypeptides, by the combination of two processes: *transcription* and *translation* [152]. Transcription is when a protein called RNA polymerase reads the DNA and creates *messenger RNA* (mRNA) which codes for the

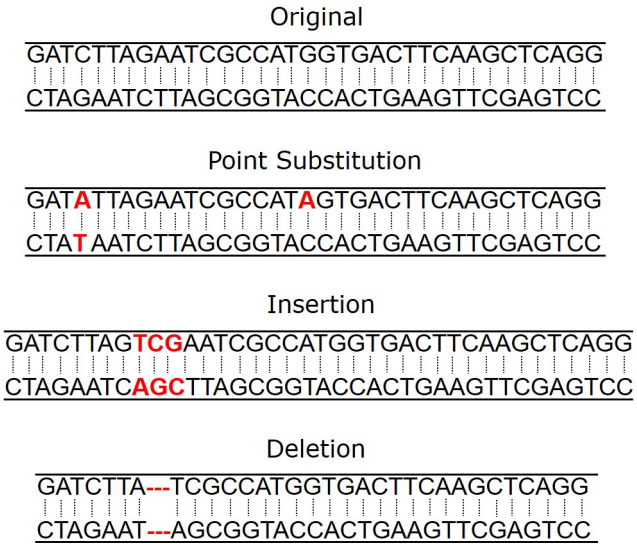

Figure 1.9: **How different mutation types affect a gene.** A, T, C, and G represent the different nucleobases. The red bases are where mutations have occurred. Two different types of point substitutions are shown; one where a pair is substituted, and one where only one base is substituted, resulting in a mismatched pair.

amino acids in a protein. Translation is when a ribosome reads the mRNA and combines the described amino acids to form a polypeptide (see Fig. 1.10) [20]. This means that different DNA results in different RNA which results in different proteins. The process is referred to as the central dogma of molecular biology and was first described by Francis Crick in 1957 [40, 41]. The implication is that when the DNA is altered through mutation, there can be changes in the proteins that are expressed. The change in protein expression can lead to new functionality for the bacteria, which is called a change in phenotype. Most commonly the change in functionality is that the protein will stop working, but occasionally there are less predictable results such as the protein becoming more efficient or gaining a whole new ability [75]. Because bacteria's phenotypes are a result of the proteins being expressed and how they work, there can be a delay between when a mutation occurs and when the new phenotype is expressed [166, 124]; this is called phenotypic lag and will be discussed in length in Section 3.2.

For bacteria to reproduce, they need to create at least two complete copies of their chromosome so that each child can inherit one. In an ideal growth environment, mutations commonly occur during this DNA replication [152, 57, 11]. For *E. coli*, the chromosome is a closed loop which is folded up to preserve space [152]. Replicating the chromosome

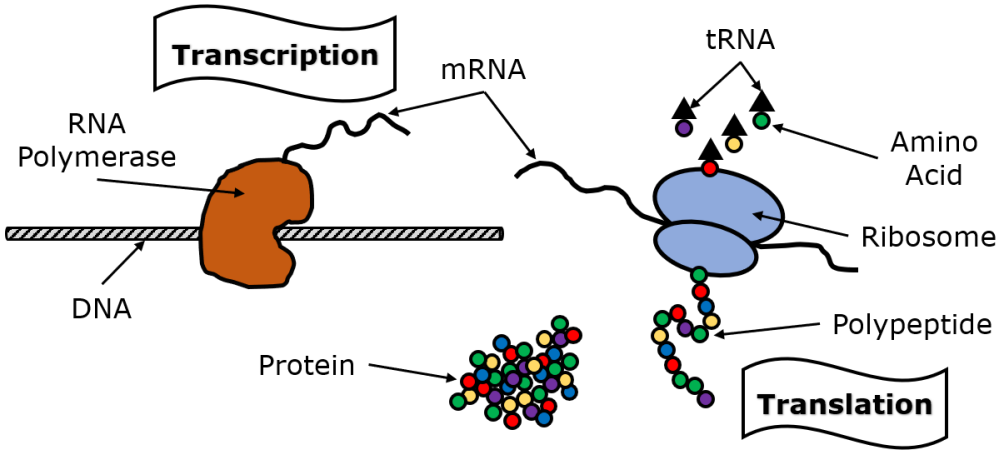

Figure 1.10: **The central dogma of molecular biology.** Transcription: RNA polymerase produces mRNA from DNA. Translation: ribosomes read mRNA and create polypeptides from amino acids which are transported by tRNA. After translation the polypeptides fold to create proteins.

requires a conglomerate of proteins. The replication starts at a point on the chromosome called the *origin* and begins with proteins "unwinding" and "unzipping" the DNA in each direction so that the two halves of the double helix are no longer connected [123]. Proteins called DNA polymerases then move along each strand of the DNA reading the nucleotides and placing the respective nucleotide pairs to make a complete strand of DNA (Fig. 1.11) [115, 123]. This process continues until the proteins reach the *terminus* on the other end of the DNA, resulting in two complete chromosomes.

Mutations primarily come about when the DNA polymerases make mistakes during DNA replication [57, 152]. There are proteins that proofread the replicated DNA and fix many of the mistakes, but they also make errors, resulting in *bona fide* mutations [152]. How many mutations come about from DNA replication is a delicate balance because if there are too many, the bacteria is likely to end up with a detrimental mutation, but if there are not enough, the bacteria's evolution could stagnate [46].

Mutations don't only come about during DNA replication. Another common source of mutations are external factors called mutagens [152]. Common mutagens are UV light and DNA-targeting antibiotics such at mitomycin-c. These particular mutagens introduce cross-links which jam DNA and RNA polymerases and cause double-strand breaks in the DNA; whereas other mutagens[3] can damage DNA in several different ways. Because the

[3]Many types of radiation, and a wide variety of chemicals and molecules can cause mutations.

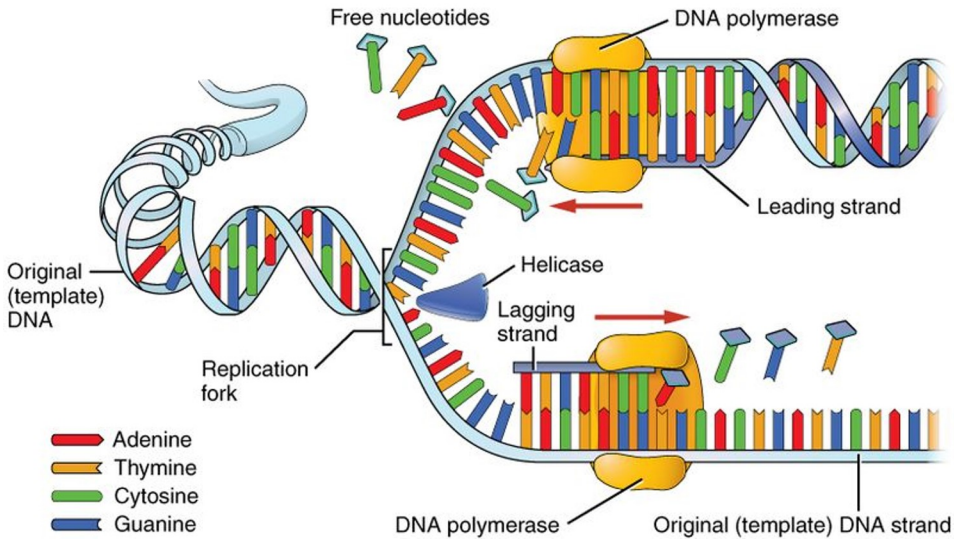

Figure 1.11: **DNA replication.** This is a simplified representation of DNA replication because many more proteins than are shown in this diagram take part in the process. The DNA polymerase is what builds the DNA and proofreads it, and as such, is the main source for mutations. "DNA Replication" by OpenStax / CC BY 4.0 / removed chromosome from original [132].

affects of mutagens can be lethal, bacteria have a system for repairing the damages. This system is called the SOS response, but the proteins that take part in the response are especially prone to errors, resulting in a high number of mutations [53].

1.5 Bacterial Growth With Mutations

1.5.1 Luria-Delbrück Fluctuation Test

In 1943, a decade before the discovery of DNA's structure, Salvador Luria and Max Delbrück set out to answer a question on many scientist's minds. The question was whether bacteria evolve in the same way as multicellular organisms [106, 205].

In Charles Darwin's seminal work "On the Origin of Species", he proposed that plants, animals, and fungi evolve by means of natural selection. This is to say that organisms

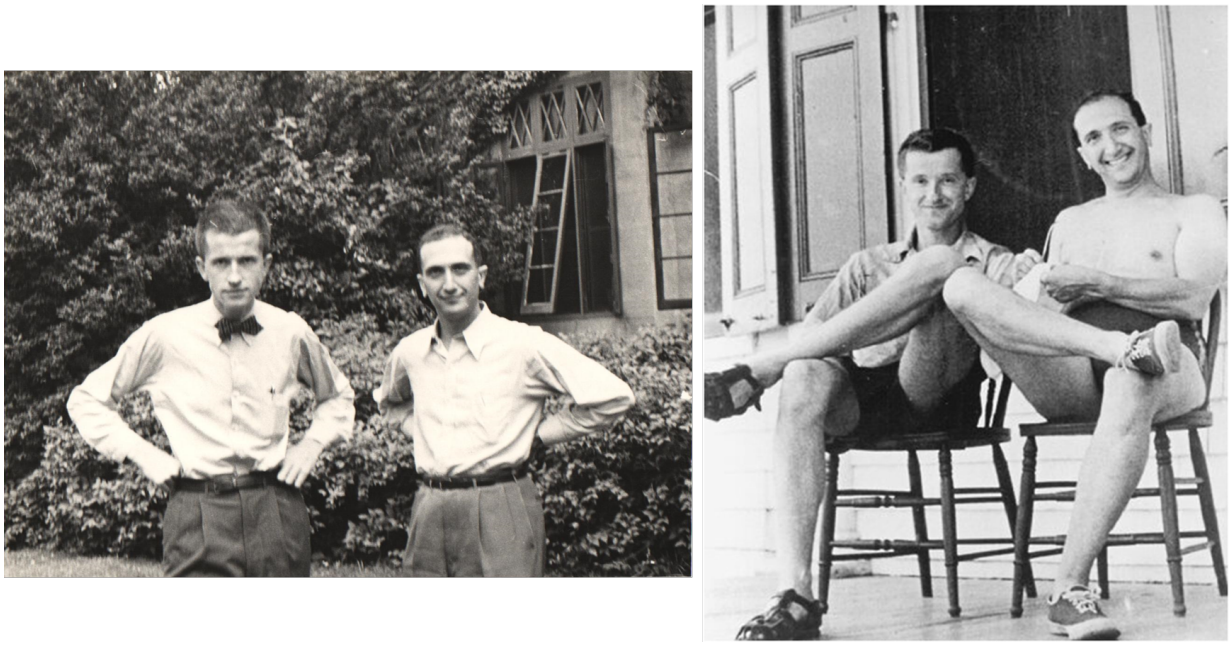

Figure 1.12: **Max Delbrück and Salvador Luria.** Left picture: Max Delbrück (left) and Salvador Luria (right) at Cold Spring Harbour Laboratory in 1946; Courtesy of the Archives at NCBS [6]. Right picture: Delbrück (left) and Luria (right) at Cold Spring Harbour Laboratory in 1953; Courtesy of Cold Spring Harbor Laboratory Archives, NY [62].

randomly gain mutations that change their phenotype, and then this phenotype is selected for through competition in nature [43]. Since bacteria appear to adapt very quickly to selective conditions, it was thought that maybe they instead mutate in *response* to the selection [149]. To test this, Luria and Delbrück designed an elaborate experiment with controlled growth and selection to probe for bacterial mutations. Luria, originally a medical doctor, came up with the idea for the experiment when he was watching a slot machine at a faculty dance at Indiana University [62, 205]. The logic went like this: since mutations are hereditary, if there is an equal probability of mutation per cell at each generation, then there is a small chance that a mutation can happen early in growth, resulting in many mutants later in time. Luria called this a *jackpot* and went on to design an experiment which would be able to identify these jackpots if they appear. The experiment came to be known as a fluctuation test because the jackpots result in large fluctuations in the number of mutant bacteria between parallel cultures after a period of growth. Delbrück, a physicist with a special interest in biology, helped with the mathematical modelling and analysis of

the experiment. Part of what makes this work so beautiful is the synergistic relationship between the experiment and the model.

A fluctuation test works by inoculating a set of tubes containing the same media with a consistent small number of cells[4]. The cells are then left to grow overnight. The next morning, each tube is plated separately with a selecting agent[5] and left to incubate. After a period of time, each plate is checked for bacterial colonies, where each colony would have been seeded by a single cell (called a colony forming unit (CFU)) that evolved a resistance to the selecting agent through mutation. The idea is that if the bacteria all have the same probability of mutation at the point of selection, called the *induced mutation* case, the distribution of cells that mutated a resistance to the selecting agent would be Poissonian (assuming a low probability of mutation), and therefore have a variance equal to its mean. Alternatively, if the cells have a constant probability of mutation per generation, called the *spontaneous mutation* case, then the variance in the number of resistant cells will be noticeably higher than the mean because each mutation will be passed onto the cell's children, resulting in the number of cells with the mutation growing exponentially and giving more plates with many mutants. See Fig. 1.13 for a comparison of how each case would play out in an experiment.

Though the induced mutation case is a straightforward Poisson process, the spontaneous mutation case required a novel mathematical analysis from Delbrück to determine estimates on the mean and variance in the number of mutants at the end of a fluctuation test. The idea is that in each time interval dt there is a probability μ that a cell will mutate and become resistant. This means that in a population of cells, you get that the total number of mutations in dt is,

$$dm = \mu N_t dt \,, \tag{1.5}$$

where m is the number of mutations[6] and N_t is the total number of bacteria. With time t measured in generations multiplied by $\ln(2)$, during exponential growth the bacterial population is governed by $N_t = N_0 e^t$. Equation (1.5) can then be integrated over the entire length of the experiment to get,

$$m = \mu(N_t - N_0) \,, \tag{1.6}$$

[4]Luria used the bacteria *Escherichia coli B* for his original experiments [106].

[5]Luria used T1 phage as his selecting agent, which is a bactriophage (virus) that will infect non-resistant *E. coli* and cause them to lyse [106, 142, 77].

[6]Note that a mutation results in a single mutant, which will then grow and produce more mutants. Consequently, a single mutation can result in many mutants. It is important that the difference between a mutation and a mutant is kept in mind throughout the thesis.

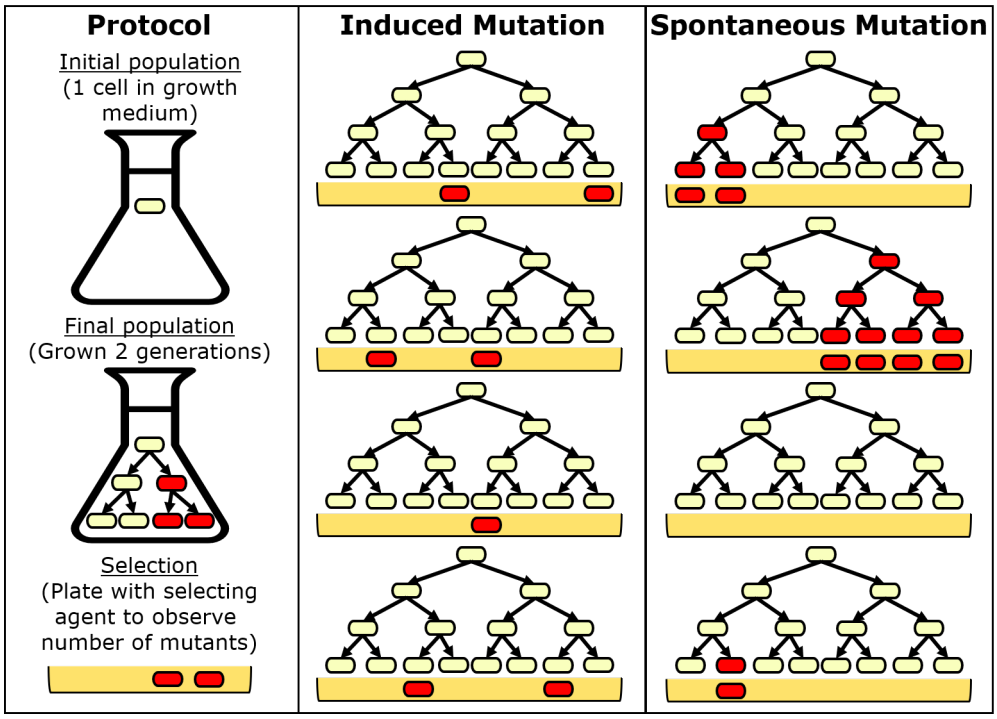

Figure 1.13: **Fluctuation tests for induced mutation and spontaneous mutation cases.** The left most column describes the general protocol for a fluctuation test while the two columns on the right show how the induced mutation and spontaneous mutation cases would appear in a fluctuation test. The "Induced mutation" represents the case where some cells mutate a resistance at the time of selection. The "Spontaneous mutation" represents the Darwinian case where some cells mutate a resistance during growth and is later selected for. When a tube containing cells is plated with a selecting agent, all non-resistant cells die, leaving only the resistant mutants to form colonies. Notice how the spontaneous mutation case has a higher variance in the number of resistant mutants amongst the different cultures.

which gives a straightforward way for going between the average number of mutations, m, and the probability of mutation (also known as the mutation rate), μ. This also gives that the units for μ are mutations per cell per generation because $(N_t - N_0)$ is the average number of cell replications that occurred in a culture. In the induced mutation case, if the selection is performed at time t, Eq. (1.6) also gives the average number of resistant mutants, which we will represent as r, and the variance in the number of mutants. Assuming $N_t >> N_0$, Eq. (1.6) can be approximated as $m = \mu N_t$, which for the induced mutation case gives

$r = \mu N_t$ and $var_r = \mu N_t$. For the spontaneous mutation case, Delbrück determined the mean and variance in the number of mutants by extending Eq. (1.5) [106]. In the spontaneous mutation case, two factors cause the number of resistant cells in a culture, r, to increase in a time interval dt: new mutants appearing through mutation and old mutants growing. The result is,

$$dr = (\mu N_t + r)dt \,, \tag{1.7}$$

where the bacterial growth rate is assumed to be the same for mutants and non-mutants. Equation (1.7) is a linear ordinary differential equation which can be integrated to find that,

$$r = t\mu N_t \,, \tag{1.8}$$

assuming there are no mutants present at the beginning of growth. Equation (1.8) gives the average number of resistant bacteria in a culture at time t, which is often taken to be the end of the experiment. To determine the variance we must consider a new timescale, $\tilde{t}$, which starts at the time of a mutation and follows the growth of the resulting lineage of mutant cells. The result is a new representation for Eq. (1.5),

$$dm = \mu N_{\tilde{t}} d\tilde{t} = \mu N_t e^{-\tilde{t}} d\tilde{t} \,, \tag{1.9}$$

because $d\tilde{t}$ starts at $(t - \tilde{t})$ so $N_{\tilde{t}} = N_0 e^{t-\tilde{t}} = N_0 e^t e^{-\tilde{t}} = N_t e^{-\tilde{t}}$. Consider that because a mutation will result in an exponentially growing lineage of mutants, the number of mutations, Eq. (1.9) can be multiplied by $e^{\tilde{t}}$ to give the number of resistant bacteria and then integrated from $\tilde{t} = 0$ to t to get Eq. (1.8). Furthermore, assuming growth is a simple birth-death process, then the variance in the number of resistant mutants will grow at approximately twice the rate of the average [88, 185], meaning we can multiply Eq. (1.9) by $e^{2\tilde{t}}$ to get,

$$var_{dr} = \mu N_t e^{\tilde{t}} d\tilde{t} \,, \tag{1.10}$$

which can then be integrated from $\tilde{t} = 0$ to t to get,

$$var_r = \mu N_t (e^t - 1) \,, \tag{1.11}$$

which is the variance in the number of mutants in a culture. Comparing Eq. (1.8) and Eq. (1.11), it is clear that the variance is indeed significantly higher than the mean in the spontaneous mutation case. With Delbrück's derived form of the mean and variance for both cases, which have been compiled in Table 1.1, the variance in the fluctuation test data could then be confidently analysed to determine if bacteria evolved by means of induced or spontaneous mutations.

	Induced Mutation	Spontaneous Mutation
Mean Number of Mutants	μN_t	$t\mu N_t$
Variance in Number of Mutants	μN_t	$\mu N_t(e^t - 1)$

Table 1.1: **Fluctuation test mean and variance in total number of mutants for induced and spontaneous mutation cases.** Assume $N_t >> N_0$. μ represents the probability of mutation per cell per generation, meaning t is the number of generations multiplied by ln(2). As determined by Delbrück in [106].

EXPERIMENT NO.	22		23	
Number of cultures	100		87	
Volume of cultures, cc	.2*		.2*	
Volume of samples, cc	.05		.2	
	Resistant bacteria	*Number of cultures*	*Resistant bacteria*	*Number of cultures*
	0	57	0	29
	1	20	1	17
	2	5	2	4
	3	2	3	3
	4	3	4	3
	5	1	5	2
	6– 10	7	6– 10	5
	11– 20	2	11– 20	6
	21– 50	2	21– 50	7
	51– 100	0	51– 100	5
	101– 200	0	101– 200	2
	201– 500	0	201– 500	4
	501–1000	1	501–1000	0
Average per sample	10.12		28.6	
Variance (corrected for sampling)	6270		6431	
Average per culture	40.48		28.6	
Bacteria per culture	2.8×10^8		2.4×10^8	
Mutation rate	2.3×10^{-8}		2.37×10^{-8}	
Standard deviation / Average (exp.)	7.8		2.8	
Standard deviation / Average (calc.)	1.5		1.5	

* Cultures in synthetic medium.

Figure 1.14: **Luria and Delbrück's fluctuation test data.** Fluctuation test data showing the "distribution of the number of resistant bacteria in a series of similar cultures". Of particular interest is the right column where they plated the entire culture instead of just a sample. Note how the variance in the number of resistant bacteria is significantly higher than the average, implying bacteria evolve by means of spontaneous mutations as opposed to induced mutations. The mutation rates in this table are calculated using Delbrück's method of the mean[7]. Table 3 from Luria and Delbrück (1943) [106].

Luria and Delbrück found in their fluctuation tests that the variance in the number of resistant mutants was significantly higher than the mean number of mutants per culture, as seen in their data in Fig. 1.14. This result strongly implied that bacteria evolve through spontaneous mutations as opposed to induced mutations, meaning they obey the same rules of evolution by means of natural selection as all other life forms on the planet.

In addition to making a strong case for spontaneous mutation, Luria and Delbrück also came up with a simple, yet powerful, method for approximating the average mutation rate of the bacteria [106]. For this, one must consider that in an infinitesimal interval of time, dt, because all bacteria are independent in terms of likelihood of mutation, the number of mutations, or new mutants, will be Poisson distributed assuming μ is sufficiently small. Then note that if one were to take a large number of similar, independent cultures, the fraction of cultures with k mutations in dt will also be Poissonian,

$$P(k) = \frac{dm^k e^{-dm}}{k!}\,, \tag{1.12}$$

where dm is the average number of mutations that take place during time interval dt. If one looks at only the probability that there are no new mutants, $k = 0$, and extends the time interval dt to the entire experiment, then they have the probability that there are no mutants in a culture at the end of an experiment. Because there are no mutants in these cultures, there is no issue of inherited mutations making the distribution differ from the Poisson distribution. Accordingly, that proportion of cultures with zero mutants in a fluctuation test should approximately equal,

$$p_0 = e^{-m}\,, \tag{1.13}$$

where m is the average number of mutations across all cultures. Consequently, one can do a fluctuation test, determine the proportion of plates without any mutants, p_0, and then plug the fraction into $\mu = \frac{-\ln(p_0)}{N_t - N_0}$, which comes from combining Equations (1.6) and (1.13), to find an estimate on the average spontaneous mutation rate. The calculation requires knowing the average initial and final populations of the cultures in the fluctuation test, which can be determined experimentally.

Lastly, Delbrück loosely laid out a framework for how one could go about building a probability distribution for the number of mutants in the spontaneous mutation case, which has come to be known as the Luria-Delbrück distribution. The distribution looks a lot like a Poisson distribution, but with a thicker, longer tail (see Fig. 1.15 for comparison). Note

[7]Mutation rate estimators that use the mean number of resistant cells between parallel cultures are no longer commonly used because they rely on an assumption that no mutations will occur before a certain time, meaning jackpots will cause the estimate to be an overestimate [144].

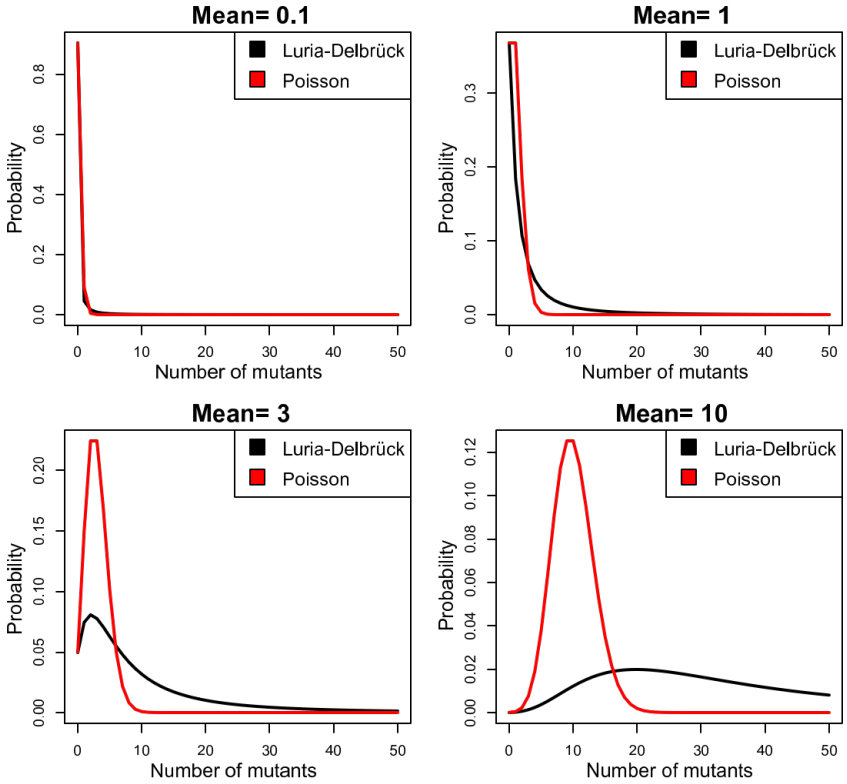

Figure 1.15: **Comparison of the probability distribution functions of the Luria-Delbrück and Poisson distributions for various different means.** The "mean" of the Luria-Delbrück distribution is the mean number of mutations per culture, but the distribution is of the number of mutants per culture. Note that the scale of the y-axis changes to help emphasise differences in shape.

that the distributions have the same p_0, or y-intercept, and that they are uniquely defined by the average number of mutations per culture, m, which is what leads to the p_0 method described above. The data from a fluctuation test can be expressed as a distribution if one has sufficiently many parallel cultures and calculates the proportion of cultures with each number of mutants (Fig. 1.16).

Luria and Delbrück's 1943 work in part earned them a Nobel Prize in 1969, but primarily for their use of bacteriophage. This fact emphasises just how monumental the work was because not only did they show the general pattern with which bacteria evolve and determined a way for approximating their mutation rates, they also helped develop the study of bacteriophage which underpins molecular biology and gives insight into how all

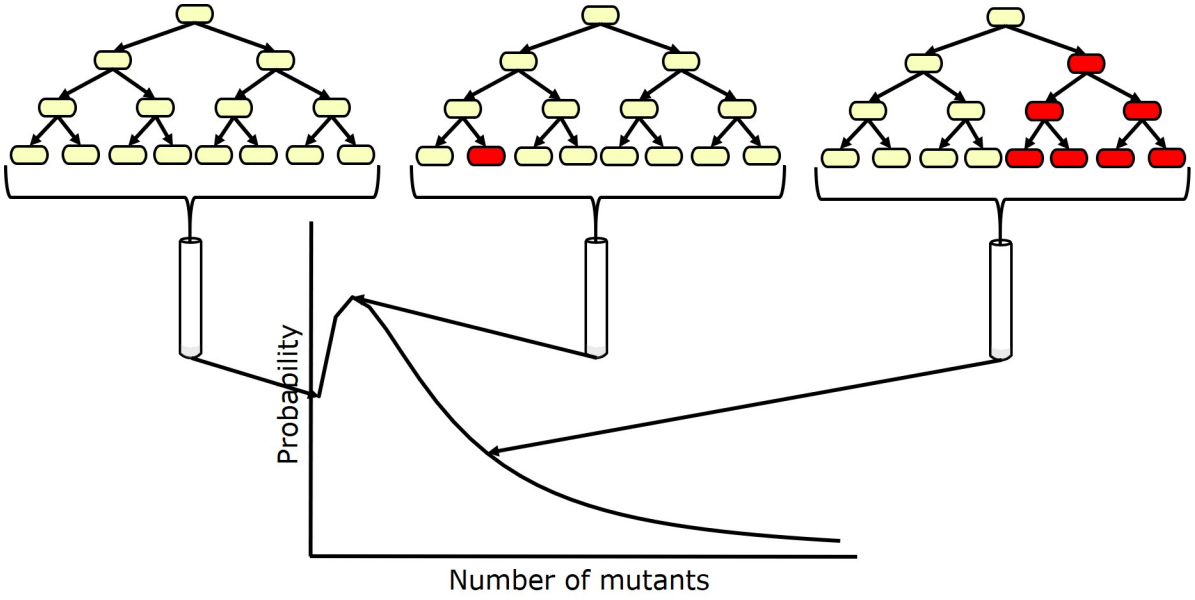

Figure 1.16: **How a fluctuation test translates to the Luria-Delbrück distribution.** Assuming a fluctuation test is performed with sufficiently many cultures, then each point on the probability distribution function is the proportion of cultures that have that many mutants.

viruses work [127, 126]. The use of fluctuation tests to approximate an average spontaneous mutation rate will be the main topic of discussion throughout this thesis.

1.5.2 Lea-Coulson Model

Luria and Delbrück's fluctuation test has been of particular interest to a number of biologists, physicists, and mathematicians since its conception [201]. Delbrück's mathematical description of the system, though clever and elegant, was not very thorough and only gave a rough approximation of the mutation rate. Since Luria and Delbrück's original paper, many have added rigour and depth to our mathematical understanding of the system, with the most powerful tool being a probability generating function (PGF) for the distribution of the number of mutants across the samples of a fluctuation test [99, 201, 4, 10]. People have of course tried to take it further, but an analytic solution for the probability distribution function (PDF) of the Luria-Delbrück distribution has eluded formulation [201].

The most common model that is used to this day was first developed by Douglas Lea

and Charles Coulson in 1947 [99]. The model was extended by Maurice Bartlett for a text in 1955 [10]. Bartlett's extension, unlike Lea and Coulson, does not rely on the simplification that the initial population is negligible compared to the final population, resulting in a more complete description of the system. The version by Bartlett will be used in this thesis, but will generally be referred to as the Lea-Coulson model to adhere to convention.

The Lea-Coulson model is built on the following assumptions [58]:

1. The cells are growing exponentially.
2. The probability of mutation is independent of previous mutations.
3. The probability of mutation is constant through a cell's lifetime.
4. The growth rates are the same for mutants and non-mutants.
5. The proportion of mutants in the total population is always small.
6. No mutants are present in the initial inoculum.
7. Reverse mutations are negligible.
8. Cell death is negligible.
9. All mutants are detected at the time of selection.
10. No mutants arise after selection.

The Lea-Coulson model is set up as a stochastic birth-death model in which mutants are born, but do not die due to assumptions 7 and 8. To build the model, imagine a culture is inoculated at time $t = 0$ with N_0 normal (non-mutant) cells. These cells then grow exponentially such that at time t the total population will be N_t. We will use r to represent the total number of mutants, or resistant cells, and p_r to represent the probability that a bacteria culture has r mutants at time t. In terms of a fluctuation test, assuming one has a sufficiently large number of parallel cultures, p_r represents the proportion of cultures with r mutants. Again, measuring time in generations, the population will grow exponentially as $\frac{dN_t}{dt} = N_t$, meaning $rdt = \frac{r}{N_t}dN_t$ represents the probability that one of the r mutants will divide during the time interval dt. Furthermore, the probability of having a cell mutate in time interval dt is μdN_t as per how μ was defined in the previous section. This means that at time $t + dt$ the probability of having r mutants, $p_r + dp_r$, is,

$$p_r + dp_r = p_{r-1}\left(\mu dN_t + \frac{r-1}{N_t}dN_t\right) + p_r\left(1 - \mu dN_t - \frac{r}{N_t}dN_t\right). \tag{1.14}$$

The first term on the right hand side represents the probability of having $(r-1)$ mutants at time t and having either a normal cell mutate or a mutant cell double. The second term represents the probability of having r mutants at time t and having none of those cells double as well as having no more normal cells mutate. We can then rewrite the left hand side as $p_r + \frac{dp_r}{dN_t} dN_t$ and rearrange to get,

$$\frac{dp_r}{dN_t} = p_{r-1}\left(\mu + \frac{r-1}{N_t}\right) - p_r\left(\mu + \frac{r}{N_t}\right) . \tag{1.15}$$

Now consider that a probability generating function, G, will have the general form,

$$G(z, N_t) = \sum_{r=0}^{\infty} p_r z^r , \tag{1.16}$$

where z is the usual auxiliary variable and the p_r's are functions dependent on N_t as described by Eq. (1.15). The first partial derivatives of the probability generating function are given by,

$$\begin{aligned} \frac{\partial G(z, N_t)}{\partial z} &= \sum_{r=0}^{\infty} r z^{r-1} p_r , \\ \frac{\partial G(z, N_t)}{\partial N_t} &= \sum_{r=0}^{\infty} z^r \frac{dp_r}{dN_t} . \end{aligned} \tag{1.17}$$

If we multiply Eq. (1.15) by z^r and sum from $r = 0$ to $r = \infty$ we get,

$$\sum_{r=0}^{\infty} z^r \frac{dp_r}{dN_t} = \sum_{r=0}^{\infty} z^r p_{r-1}\left(\mu + \frac{r-1}{N_t}\right) - \sum_{r=0}^{\infty} z^r p_r\left(\mu + \frac{r}{N_t}\right) . \tag{1.18}$$

Note that since it doesn't make sense to have negative mutants, $p_r = 0$ for all $r < 0$, meaning that the $r = 0$ case in the first sum on the right hand side equals zero, so (1.18) can be rewritten as,

$$\sum_{r=0}^{\infty} z^r \frac{dp_r}{dN_t} = \sum_{r=0}^{\infty} z^{r+1} p_r\left(\mu + \frac{r}{N_t}\right) - \sum_{r=0}^{\infty} z^r p_r\left(\mu + \frac{r}{N_t}\right) . \tag{1.19}$$

Expanding, using $\sum z^{r+1} = z \sum z^r$, and rearranging gives,

$$\sum_{r=0}^{\infty} z^r \frac{dp_r}{dN_t} = \mu(z-1) \sum_{r=0}^{\infty} z^r p_r + \frac{z-1}{N_t} \sum_{r=0}^{\infty} r z^r p_r . \tag{1.20}$$

Adding and subtracting 1 to the exponent of z in the second sum on the right hand side and once again using $\sum z^{r+1} = z \sum z^r$ then gives,

$$\sum_{r=0}^{\infty} z^r \frac{dp_r}{dN_t} = \mu(z-1) \sum_{r=0}^{\infty} z^r p_r + \frac{z(z-1)}{N_t} \sum_{r=0}^{\infty} r z^{r-1} p_r \,. \tag{1.21}$$

Substituting in the first partial derivatives from Equations (1.17) then reduces Eq. (1.21) to the first-order partial differential equation (PDE),

$$\frac{\partial G(z, N_t)}{\partial N_t} = \mu(z-1) G(z, N_t) + \frac{z(z-1)}{N_t} \frac{\partial G(z, N_t)}{\partial z} \,, \tag{1.22}$$

which is a quasi-linear PDE that can be solved using the method of characteristics. Accordingly, the characteristic equations are,

$$\begin{aligned} \frac{dN_t}{ds} &= 1 \,, \\ \frac{dz}{ds} &= \frac{-z(z-1)}{N_t} \,, \\ \frac{dG}{ds} &= \mu(z-1)G \,, \end{aligned} \tag{1.23}$$

where s is the parameterisation variable of the characteristic curves. Isolating ds in each characteristic equation and equating then gives,

$$dN_t = \frac{-N_t dz}{z(z-1)} = \frac{dG}{\mu(z-1)G} \,. \tag{1.24}$$

Equation (1.24) can now be used to determine ordinary differential equations (ODE) which can be solved to find N_t and G. Firstly, Eq. (1.24) can be rearranged to give,

$$\frac{dN_t}{dz} = \frac{N_t}{z(1-z)} \,, \tag{1.25}$$

which is an ODE for N_t with respect to z. Solving this equation by separating the variables and integrating gives,

$$N_t = \frac{Cz}{1-z} \,, \tag{1.26}$$

where C is an integration constant. By rearranging (1.24) we also find that,

$$\frac{dG}{dz} = \frac{-N_t \mu G}{z} \,, \tag{1.27}$$

which is an ODE for G with respect to z, but has a dependence on N_t. Because N_t has already been solved for, Eq. (1.27) can be combined with Eq. (1.26) to give,

$$\frac{dG}{dz} = \frac{-C\mu G}{1-z}\,, \tag{1.28}$$

which is now an ODE only dependent on G and z. Once again solving by separating the variables and integrating, we get,

$$G = D(1-z)^{C\mu}\,, \tag{1.29}$$

where D is another integration constant. Notice that from Eq. (1.26) we know that $C = \frac{N_t(1-z)}{z}$, so we only have to find D to have a complete solution for G. To do this, consider that at $t = 0$ there are only normal cells, meaning $G(z_0, N_0) = p_0 = 1$, so that,

$$1 = D(1-z_0)^{C\mu}\,. \tag{1.30}$$

Now notice that from (1.26) we also have that $z = \frac{N_t}{C+N_t}$ so that at $t = 0$ we have $z_0 = \frac{N_0}{C+N_0}$. Substituting this into (1.30) then manipulating and rearranging gives,

$$D = \left(1 + \frac{N_0}{C}\right)^{C\mu}\,. \tag{1.31}$$

We can then substitute this form for D as well as our form for C into (1.29) to get,

$$G(z, N_t) = \left(1 - z + \frac{N_0}{N_t}z\right)^{\frac{\mu N_t(1-z)}{z}}\,, \tag{1.32}$$

which is the completed probability generating function (PGF).

From the probability generating function, the mean and variance of the distribution can be calculated. To calculate the mean, the derivative of the PGF is taken with respect to z and then z is set to 1, approaching from the left. In other words, if X is a discrete random variable, then $E[X] = \lim_{z\to 1^-} \frac{\partial G(z,N_t)}{\partial z}$. When applied to the Lea-Coulson PGF this gives,

$$E[X] = t\mu N_t\,, \tag{1.33}$$

which is the same as the mean calculated by Delbrück. On the other hand, the variance is calculated with $Var[X] = \lim_{z\to 1^-}[\frac{\partial^2 G(z,N_t)}{\partial z^2} + \frac{\partial G(z,N_t)}{\partial z} - (\frac{\partial G(z,N_t)}{\partial z})^2]$ to give,

$$Var[X] = 2\mu N_t(e^t - 1) - t\mu N_t\,, \tag{1.34}$$

which differs from the variance calculated by Delbrück. The variance calculated from the Lea-Coulson model is the same as Delbrück's in the $t \to 0$ limit, but is two times larger in the $t \to \infty$ limit. The variance is larger in the Lea-Coulson formulation because the growth of the mutants is stochastic, while in the Delbrück formulation it is deterministic [10].

To date, an analytic inverse of the probability generating function (Eq. (1.16)) for the Lea-Coulson model (Eq. (1.32)) has not been found for p_r, making a closed form expression for the probability distribution function of the Luria-Delbrück distribution unattainable [201]. Fortunately, the probabilities can instead be constructed by a recursive relation found by Sarkar, Ma, and Sandri [148, 108]. To derive the recursive relation for p_r, one must first assume that $p_0 \neq 0$ so that $G(z=0) \neq 0$, causing $\ln(G(z))$ to be analytic for all z. Note that the written dependence on N_t of $G(z, N_t)$ has been dropped for readability. A new function $\zeta(z)$ can be defined such that,

$$\zeta(z) = \ln\left(\frac{G(z)}{G(0)}\right), \tag{1.35}$$

which means that $G(z) = G(0)e^{\zeta(z)}$. Taking the derivative of $G(z)$ with respect to z then gives,

$$G'(z) = G(0)e^{\zeta(z)}\zeta'(z) = G(z)\zeta'(z), \tag{1.36}$$

which can be further differentiated to give,

$$G^{(r)}(z) = \sum_{i=0}^{r-1} \binom{r-1}{i} G^{(i)}(z)\zeta^{(r-i)}(z), \tag{1.37}$$

where $G^{(r)}(z)$ represents the rth derivative of $G(z)$ with respect to z and $\binom{r-1}{i} = \frac{(r-1)!}{i!(r-i-1)!}$ are the binomial coefficients. Now note that because $\zeta(z)$ is analytic, it has a power series around $z = 0$ that can be written as,

$$\zeta(z) = \sum_{r=0}^{\infty} a_r z^r, \tag{1.38}$$

where a_r are constants. Taking the kth derivative with respect to z of the power series forms for $G(z)$ and $\zeta(z)$ (Equations (1.16) and (1.38) respectively) gives,

$$\begin{aligned} G^{(k)}(z) &= \sum_{r=k}^{\infty} \frac{r!}{(r-k)!} p_r z^{r-k}, \\ \zeta^{(k)}(z) &= \sum_{r=k}^{\infty} \frac{r!}{(r-k)!} a_r z^{r-k}, \end{aligned} \tag{1.39}$$

which when evaluated at $z = 0$ gives,

$$G^{(k)}(0) = k! p_k \,, \tag{1.40}$$

$$\zeta^{(k)}(0) = k! a_k \,, \tag{1.41}$$

because only the first term in Eqs. (1.39) are independent of z. Now evaluate Eq. (1.37) at $z = 0$,

$$G^{(r)}(0) = \sum_{i=0}^{r-1} \binom{r-1}{i} G^{(i)}(0) \zeta^{(r-i)}(0) \,, \tag{1.42}$$

then substitute in Eq. (1.40) with $k = i$, Eq. (1.41) with $k = r - i$, and the definition of the binomial coefficients to get,

$$G^{(r)}(0) = \sum_{i=0}^{r-1} \frac{(r-1)!}{i!(r-i-1)!} i! p_i (r-i)! a_{r-i} \,. \tag{1.43}$$

Rearranging and cancelling out terms then gives,

$$G^{(r)}(0) = (r-1)! \sum_{i=0}^{r-1} (r-i) a_{r-i} p_i \,. \tag{1.44}$$

Notice that the left hand side of Eq. (1.44) can be replaced with Eq. (1.40) where $k = r$,

$$r! p_r = (r-1)! \sum_{i=0}^{r-1} (r-i) a_{r-i} p_i \,, \tag{1.45}$$

which simplified is,

$$p_r = \frac{1}{r} \sum_{i=0}^{r-1} (r-i) a_{r-i} p_i \,. \tag{1.46}$$

This is a general recursive relation that can be used to find the probabilities of any discrete distribution as long as $G(0) \neq 0$ and the a_{r-i} are computable [148]. To find the probabilities of the Luria-Delbrück distribution, the a_{r-i} can be found from the Lea-Coulson probability generating function. First, we will rewrite the Lea-Coulson PGF (Eq. (1.32)) in a more amenable form,

$$G(z) = (1 - \phi z)^{\frac{m}{\phi}\left(\frac{1-z}{z}\right)} = \exp\left[\frac{m}{\phi}\left(\frac{1-z}{z}\right)\ln(1-\phi z)\right] \,, \tag{1.47}$$

where $\phi = 1 - \frac{N_0}{N_t}$ is a scaling factor which accounts for the difference between the initial and final populations[8], and $m = \mu(N_t - N_0)$ as defined by Delbrück. Using L'Hôpital's rule to compute the limit, $\lim_{z\to 0} G(z)$ gives,

$$G(0) = e^{-m}\,, \tag{1.48}$$

which agrees with Delbrück's original result. Substituting our expressions for $G(z)$ and $G(0)$ (Eqs. (1.47) and (1.48) respectively) into the definition of $\zeta(z)$ (Eq. (1.35)),

$$\zeta(z) = \ln\left(\frac{\exp\left[\frac{m}{\phi}\left(\frac{1-z}{z}\right)\ln(1-\phi z)\right]}{\exp(-m)}\right) = \frac{m}{\phi}\left[\left(\frac{1-z}{z}\right)\ln(1-\phi z) + \phi\right]. \tag{1.49}$$

The Taylor expansion of $\zeta(z)$ about $z = 0$ can then be taken to get,

$$\zeta(z) = \sum_{k=1}^{\infty} m\frac{\phi^{k-1}}{k}\left(1 - \frac{k\phi}{k+1}\right)z^k\,. \tag{1.50}$$

Comparing with the general power series form of $\zeta(z)$ from Eq. (1.38),

$$a_0 + \sum_{k=1}^{\infty} a_k z^k = \sum_{k=1}^{\infty} m\frac{\phi^{k-1}}{k}\left(1 - \frac{k\phi}{k+1}\right)z^k\,, \tag{1.51}$$

clearly gives that,

$$\begin{aligned} a_0 &= 0\,, \\ a_k &= m\frac{\phi^{k-1}}{k}\left(1 - \frac{k\phi}{k+1}\right) \qquad (k \geq 1)\,. \end{aligned} \tag{1.52}$$

Finally, substituting the Lea-Coulson form for a_k (Eq. (1.52)) with $k = r - i$ into the general expression for p_r (Eq. (1.46)),

$$p_r = \frac{m}{r}\sum_{i=0}^{r-1}\phi^{r-i-1}\left(1 - \frac{(r-i)\phi}{r-i+1}\right)p_i\,, \tag{1.53}$$

then letting $j = r - i$ and reversing the order of the summation gives the complete recursive relation,

$$\begin{aligned} p_0 &= e^{-m}\,, \\ p_r &= \frac{m}{r}\sum_{j=1}^{r}\phi^{j-1}\left(1 - \frac{j\phi}{j+1}\right)p_{r-j} \qquad (r \geq 1)\,. \end{aligned} \tag{1.54}$$

[8]Note how $\phi \approx 1$ in the $N_t >> N_0$ case.

Using the recursive relation in Eq. (1.54)[9], one can calculate the probability of having r mutants in a culture when the average number of mutations per culture is m and the growth scaling factor is $\phi = 1 - \frac{N_0}{N_t}$ where N_0 and N_t are the initial and final populations respectively. By computing the probabilities for all values of r, the probability distribution function can be determined. This recursive relation is used by the most popular analysis programs that numerically approximate the distribution and estimate a best fit for fluctuation test data [210, 112]. The numerical implementation of the relation by Zheng will be explored further in the Section 3.1.2.

Due to the idealistic assumptions of the Lea-Coulson model, extensions that loosen these assumptions and incorporate the effects of different laboratory practices have become common. The most popular extension is to assume that the normal and mutant cells have different growth rates. This was first explored by Arthur Koch [92] and has come to be considered the more general description because the equal growth rate case is simply a special case of it. Another common adjustment is to account for a non-perfect plating efficiency, or in other words, only taking a sample of the culture for the selection phase of the fluctuation test [63, 206, 208, 82]. This is of particular interest if the cells are grown to a very high final population. People have also explored the effects of phenotypic lag on the model, but the resulting extensions have been of little practical use [4, 87, 201, 10, 42, 94]. I will be using the unaltered Lea-Coulson model throughout the thesis, putting extra emphasis on careful experimental design so that I don't have to resort to using any of these model extensions.

1.5.3 Haldane Model

Another common model, popular for its understandability, is the unpublished model by J.B.S. Haldane. Haldane's model is similar in nature to the Lea-Coulson model, but has all the cells inside a culture dividing simultaneously instead of stochastically and was built from combinatorics instead of a birth-death model [147, 204]. Most importantly, it allows for figures such as Fig. 1.17, which are easy to follow and allow for the visualisation of the affects of mutants appearing at different times in growth. Haldane's model will be used throughout the thesis for more qualitative descriptions of the systems, particularly in regards to phenotypic lag and its effects.

[9] This is Zheng's form of Sarkar et al.'s recursive relation [203].

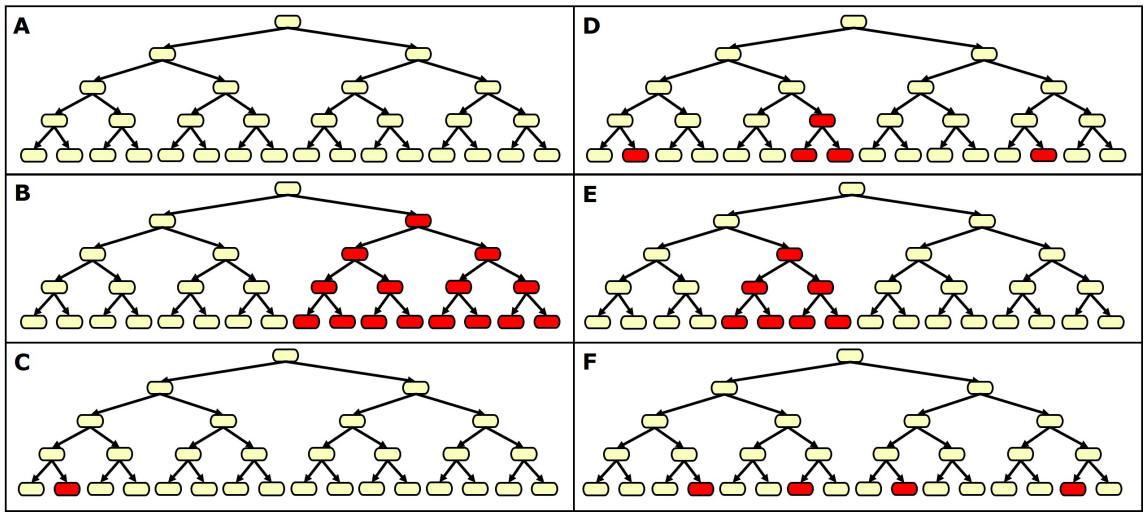

Figure 1.17: **Haldane trees describing different combinatorical ways a culture can gain mutants.** Each Haldane tree starts with one non-mutant cell that grows for 4 doublings, giving 5 generations total. In each panel, a yellow cell is a non-mutant (or normal) cell, and each red cell is a mutant cell. Panels (D), (E), and (F) show three separate methods in which a culture can have four mutant cells.

1.6 Bacterial Physiology

Physiology is simply the study of the functioning of living organisms [155]. Accordingly, bacterial physiology is the study of how bacteria function. In 1943 when Luria and Delbrück designed the fluctuation test, not much was known about bacterial physiology and as a result people did not put much importance on it. This changed in the 1950's with work led by Jacques Monod and Ole Maaløe that implied that most results pertaining to bacteria were contingent on their physiology [150]. In practice there are three "levels" of bacterial function that one can study: the intracellular level which is largely characterised by the protein dynamics within the cell, the single cell level which is characterised by the composition and growth dynamics of a single cell, and the population level which is characterised by the accumulation of the compositions and growth dynamics of all the cells within the population. As one goes from the more microscopic systems to the more macroscopic systems, the law of large numbers takes affect and their descriptions go from primarily being complex and often noisy to being simple phenomenological models [1, 145]. The relationship between the three "levels" of bacterial function and the types of models commonly used to describe each is reminiscent of the relationship between quantum

mechanics, statistical mechanics, and thermodynamics. In the context of mutations, the intracellular level is the system that causes mutations in a cell (primarily mistakes during DNA replication), the single cell level is observing which cells in the population are mutants and how many resistant offspring each will create, and the population level is what Luria and Delbrück studied by looking at how many total mutants there are in many populations. Luria and Delbrück furthered the scope of the research by using mathematics to probe for what single cell effects would lead to what they saw at the population level, giving insight into the way bacteria mutate and developing a method for calculating the average mutation rate for a single cell per generation [106]. In its essence, this is what the field commonly referred to as "bacterial physiology" is all about; quantitatively studying characteristics of bacterial populations in an attempt to illuminate the inner workings of the cells [35]. Even more specifically, "bacterial physiology" is most commonly associated with the study of the functions that allow bacteria to grow and reproduce, and the resulting dynamics [119].

1.6.1 Growth Physiology

Due to the relatively simplistic nature of bacteria and the restrictions balanced growth put on them, people realised that quantitative descriptions of their physiology could be practical. The idea of growing cells at steady state and using mathematics to describe their physiology started with Jacques Monod in 1942 [150]. Monod grew *E. coli* cells hoping to study logistic growth, but instead found what is now known as "Monod kinetics" (Fig. 1.18) [117]. He discovered that the doubling rate, η, of the cells is dependent on the concentration of a growth limiting substrate (such as a carbon source like glucose) and could be modelled using Michaelis-Menten kinetics,

$$\eta = \eta_{\max} \frac{S}{S + K_D} \,, \tag{1.55}$$

where $\eta_{\max}$ is the maximal doubling rate[10], S represents the concentration of the growth limiting substrate, and K_D is the Michaelis constant which is determined by the nature of the substrate and bacterial strain [117]. Note that because $\eta = \frac{\lambda}{\ln 2}$, both sides of Eq. (1.55) can be multiplied by $\ln 2$ and rewritten in terms of the specific growth rate, λ. Monod also found that the final yield of the cells was determined by the initial concentration of the growth limiting substrate (Fig. 1.19) [117]. The discovery of "Monod's kinetics" was a monumental step towards doing reproducible microbiology that was amenable to simple, quantitative descriptions [117]. Monod quickly moved on from the field of bacterial

[10]The growth rate when the growth limiting substrate is in saturation.

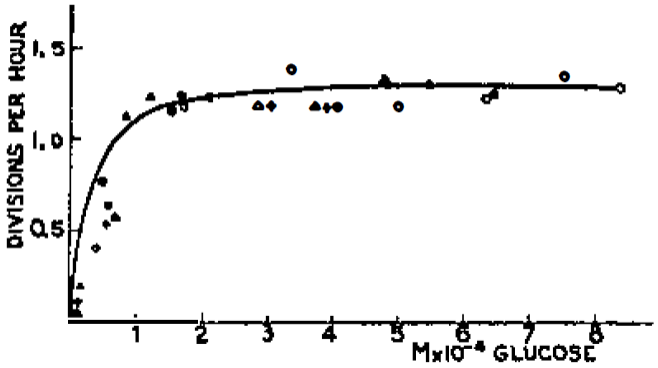

Figure 1.18: **Monod kinetics: Relationship between doubling rate and glucose concentration.** The doubling rate, η, of *E. coli* versus the concentration of glucose in the growth medium. Fitted line gives $\eta_{max} = 1.35$ doublings/hour and $K_D = 22\mu M$ for Eq. (1.55). Figure 4 in Monod (1949) [117].

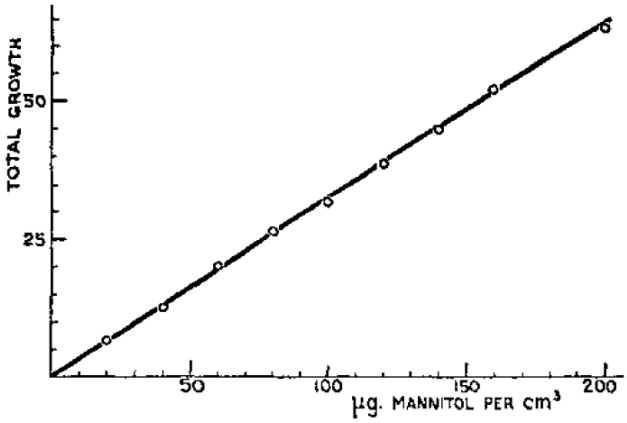

Figure 1.19: **Total bacterial growth versus carbon source concentration in *E. coli*.** The final concentration of *E. coli* is linearly proportional to the concentration of carbon source (mannitol) in the growth medium. Figure 3 from Monod (1949) [117].

physiology saying "The study of the growth of bacterial cultures does not constitute a specialised subject or branch of research: it is the basic method of Microbiology" [150]. In this comment he both outlined the importance of having a comprehensive understanding of growth, and alluded to his belief that studying growth simply for its own sake would prove sterile.

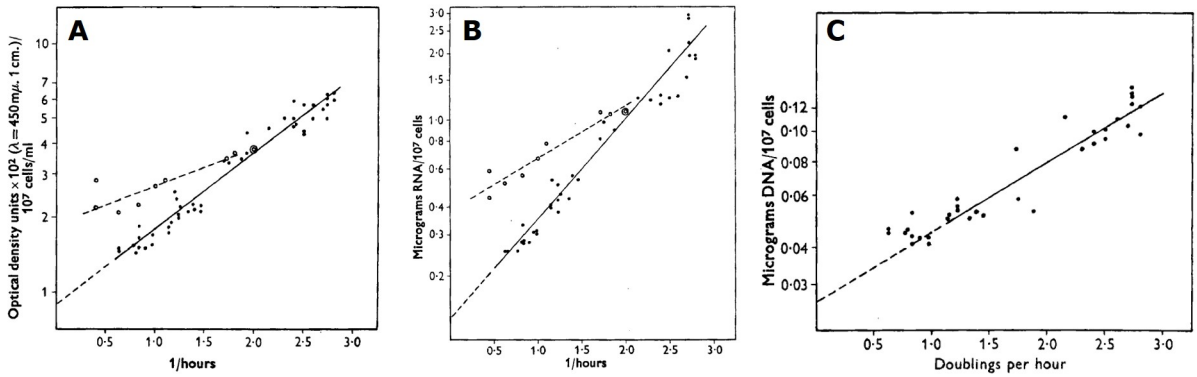

Figure 1.20: **The macromolecular composition of *Salmonella typhimurium* versus growth rate.** The points come from *Salmonella typhimurium* cells in balanced growth in several different defined media of varying quality. A) The mass/cell, M, increases approximately exponentially with doubling rate, η; $M \propto 2^{\eta}$. B) The RNA/cell, R, increases approximately exponentially with doubling rate, and faster than Mass/cell; $R \propto 2^{1.5\eta}$. C) The DNA/cell, D, increases approximately exponentially with doubling rate, and slower than Mass/cell; $D \propto 2^{0.8\eta}$. In all three panels, the units for the x-axes are doublings/hour, and the y-axes are on a log-scale. Panels (A), (B), and (C) are Figures 1, 2, and 3 respectively from Schaechter et. al (1958) [153].

Luckily, a select few did not heed Monod's warning and the field of bacterial growth physiology took off in the mid 1950's. Much of this early work came out of Ole Maaløe's lab in Copenhagen [150]. In 1958, Schaechter, Maaløe, and Kjeldgaard released a paper showing that the macromolecular composition of *Salmonella typhimurium* cells in balanced growth is primarily dependent on their growth rate, not the details of the growth media [153]. Specifically, they changed the growth rate by changing the *quality* of nutrients in the growth medium[11]. This was primarily done by using different carbon sources, but they also supplemented some media with amino acids and/or nucleotides. Surprisingly, the relations between the quantity of studied components per cell and growth rate were all approximately log linear (Fig. 1.20) [153]. The work from Maaløe's lab helped motivate the idea that growth rate is an empirically significant "state variable" (like temperature in a thermodynamic system) that can be useful in describing the state of the cell [19].

From Schaeter et. al's work, it was now known that the DNA/cell, D, increases approximately exponentially with doubling rate, $D \propto 2^{0.8\eta}$ [153]. The reason why faster

[11] Changing nutrient quality is the way growth rate is modulated throughout the thesis (unless stated otherwise).

growing cells would have more DNA was unclear though. Assuming bacteria require at least one chromosome to function and that the size of the chromosome is independent of the bacteria's physiology, then the implication of having more DNA per cell in faster growing cells would be that the cells have multiple copies of the chromosome. But why would they need multiple copies? Cooper and Helmstetter answered this question in 1968 by studying the DNA synthesis rate in *E. coli* growing at different speeds [36] . They found that the bacteria commence and end DNA synthesis in discrete patterns, and that faster growing cells have a higher rate of DNA synthesis as compared to slower growing cells, as seen in the left hand portion of Fig. 1.21. Consequently, they concluded that the fast growing bacteria must be replicating several strands of DNA in parallel (right hand side of Fig. 1.21), which would be necessary if the cells doubled in a shorter amount of time than it took a chromosome to replicate. Cooper and Helmstetter determined that replication of a single chromosome takes a fixed amount of time which is approximately 40 minutes and then the cell requires a further 20 minutes to segregate the chromosomes and divide. The result being that cells growing with a doubling time of less than 60 minutes would require parallelisation of DNA replication[12] and the cells should not be able to double faster than every 20 minutes. The parallelisation works through "forking" replications where a chromosome will start being replicated while itself is still being replicated, as seen in Fig. 1.22.

In 1996, Hans Bremer and Patrick Dennis compiled decades of work confirming and expanding on much of the previously completed work in bacterial physiology. The comprehensive review discussed and experimentally showed the composition of *E. coli* B/r cells and how they relate to growth rate [19]. The results can be found in Fig. 1.23. The compilation of all these physiological parameters was also accompanied by a compilation of simple mathematical rules that describe the origins and growth rate dependence of many of these parameters, which can be seen in Fig. 1.24. Having all of these parameters and rules comprehensively laid out for a single strain of *E. coli* facilitates further quantitative modelling of the macromolecular composition of bacterial cells and makes it easier for researchers to consider the consequences of the growth rate in their systems.

All of the bacterial physiology that has been discussed up to this point has involved cultures seeded with a large numbers of cells. A consequence is that the law of large numbers smooths over the variations in the single cell dynamics, resulting in smooth empirical rules describing the systems well [19, 145, 1]. When a culture is seeded from a small number of cells, such as in a fluctuation test, this effect weakens and more variance can be found culture to culture [2, 1]. This is especially apparent when comparing the number of cells in many parallel cultures after a long period of growth starting from small inocula. The

[12]This means fast growing cells' grandparents would have had to start building their chromosomes.

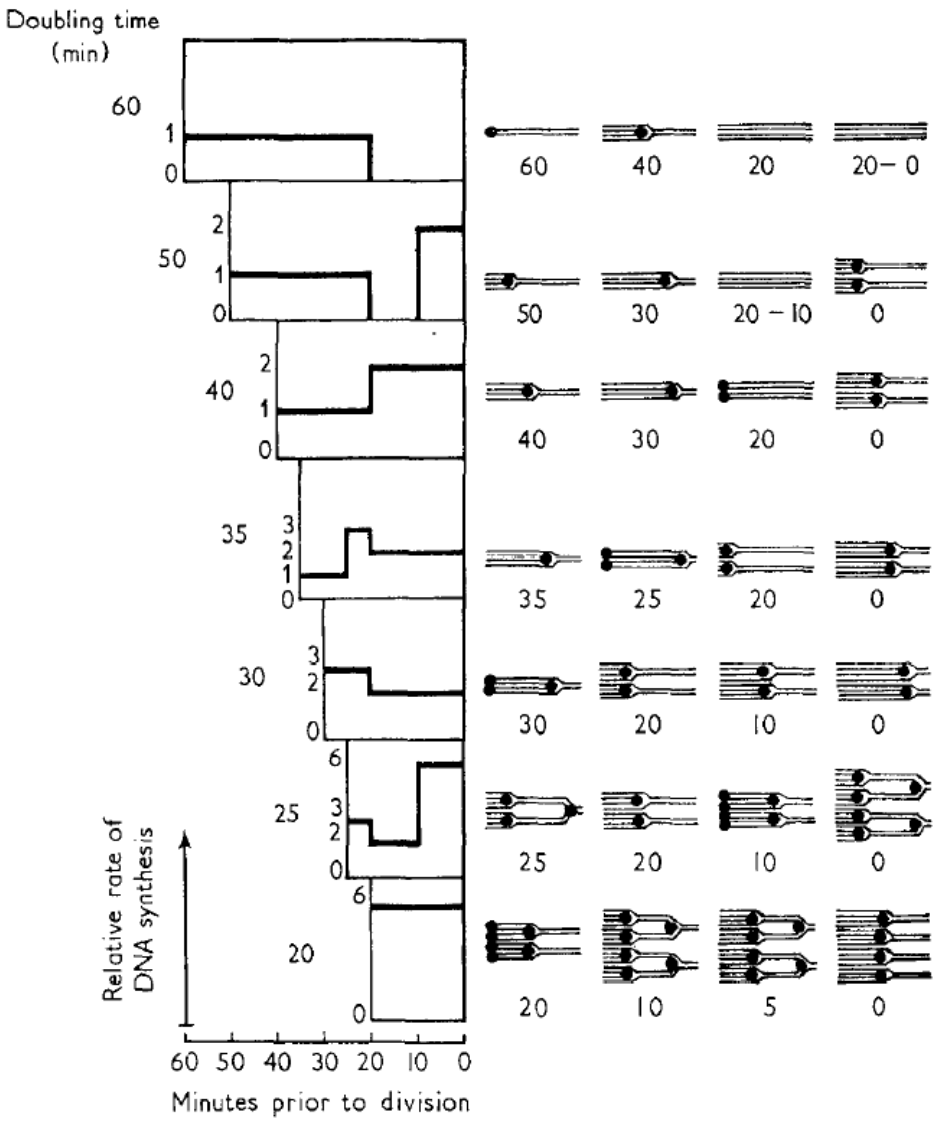

Figure 1.21: **Relative DNA synthesis rates and the consequences on DNA replication in *E. coli* with different doubling times.** The left hand portion shows the relative rate of DNA synthesis during a cell's lifetime for cells with different doubling times. The right hand portion shows the resulting DNA replication forks. Figure 1 from Cooper and Helmstetter (1968) [36].

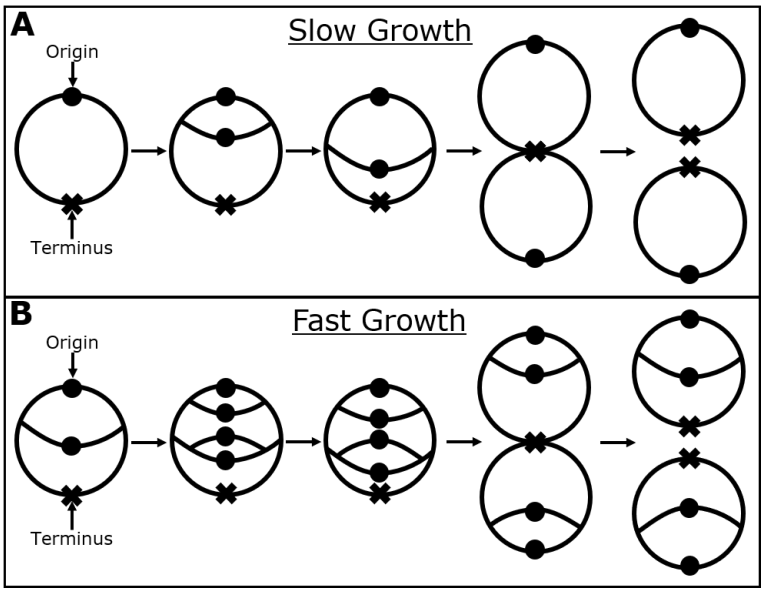

Figure 1.22: **A simplified representation of *E. coli* chromosome replication.** The "origin" is where DNA replication begins, splitting off in both directions and eventually meeting again at the "terminus" where DNA replication ends. A) Chromosome replication for slow growing *E. coli* cells where there is no parallel replication. B) Chromosome replication for fast growing cells, where there are multiple replication forks running simultaneously.

culture to culture variance is largely due to the fact that within a culture, the doubling time between cells can often vary by upwards of 10% around the mean [190]. When there are many cells, there are minimal consequences to the cell-to-cell variation, but when there are few cells, the resulting noise becomes noticeable [2, 1].

Class	No.	Parameter	Symbol	Value	Reference
I	1	Deoxyribonucleotide residues per genome	kbp/genome	4,700	4
	2	Ribonucleotide residues per rRNA precursor	nucl./prib	6,000	104
	3	Ribonucleotide residues per 70S ribosome	nucl./rib	4,566	104
	4	Amino acid residues per 70S ribosome	aa/rib	7,336	140
	5	Ribonucleotide residues per tRNA	nucl./tRNA	80	64
	6	Amino acid residues per RNA polymerase core	aa/pol	3,407	107–109
II	7	Fraction of total RNA that is stable RNA	f_s	0.98	5, 80
	8	Fraction of stable RNA that is tRNA	f_t	0.14	37, 118
	9	Fraction of active ribosomes	β_r	0.8	57
III	10	Fraction of total protein that is r-protein	α_r	0.09–0.22	Table 3
	11	Fraction of total protein that is RNA polymerase	α_p	0.009–0.01	Table 3
	12	Fraction of active RNA polymerase synthesizing rRNA and tRNA	ψ_s	0.28–0.77	Table 3
	13	Fraction of active RNA polymerase	β_p	0.15–0.32	Table 3
IV	14	Peptide chain elongation rate	c_p	12–22 aa/s	Table 3
	15	Stable RNA chain elongation rate	c_s	85 nucl./s	Table 3
	16	mRNA chain elongation rate	c_m	40–55 nucl./s	Table 3
	17	DNA chain elongation rate	c_d	500–830 nucl. bp/s	Table 3
V	18	Time to replicate the chromosome	C	40–67 min	Table 3
	19	Time between termination of replication and division	D	22–30 min	Table 3
	20	Protein per replication origin	P_O	2.5×10^8– 4×10^8 aa	Table 2

Parameter	Symbol	Units	At τ (min) and μ (doublings per h):						
			τ, 100 μ, 0.6	τ, 60 μ, 1.0	τ, 40 μ, 1.5	τ, 30 μ, 2.0	τ, 24 μ, 2.5	Observed parameter(s)	Footnote
Protein/mass	P_M	10^{17} aa/OD_{460}	6.5	5.8	5.2	5.1	5.0	P, M	*b*
RNA/mass	R_M	10^{16} nucl./OD_{460}	4.3	4.9	5.7	6.6	7.8	R, M	*c*
DNA/mass	G_M	10^8 genomes/OD_{460}	18.3	12.4	9.3	8.0	7.6	G, M	*d*
Cell no./mass	C_M	10^8 cells/OD_{460}	11.7	6.7	4.0	2.7	2.0	Cells/OD_{460}	*e*
$(P + R + G)/M$	PRD_M	μg/OD_{460}	149	137	129	131	136		*f*
Protein/genome	P_G	10^8 aa residues	3.5	4.7	5.6	6.3	6.6	P_M, G_M	
RNA/genome	R_G	10^7 nucl. residues	2.3	4.0	6.1	8.2	10.3	R_M, G_M	
Origins/genome	O_G	Dimensionless	1.25	1.32	1.44	1.58	1.73	C	*g*
Protein/origin	P_O	10^8 aa residues	2.8	3.6	3.9	4.0	3.8	P_G, O_G	*g*
Protein/cell	P_C	10^8 aa residues	5.6	8.7	13.0	18.9	25.0	P_M, C_M	
	P_C (μg)	μg/10^9 cells	100	156	234	340	450		*h*
RNA/cell	R_C	10^7 nucl. residues	3.7	7.3	14.3	24.4	39.0	R_M, C_M	
	R_C (μg)	μg/10^9 cells	20	39	77	132	211		*h*
DNA/cell	G_C	genome equiv./cell	1.6	1.8	2.3	3.0	3.8	C, D	*i*
	G_C (μg)	μg/10^9 cells	7.6	9.0	11.3	14.4	18.3		*h*
Mass/cell	M_C	OD_{460} units/10^9 cells	0.85	1.49	2.5	3.7	5.0	C_M	*j*
	M_C (μg)	μg dry weight/10^9 cells	148	258	433	641	865	μg/OD_{460}	*k*
Sum $P + R + G$	PRD_C	μg/10^9 cells	127	204	322	486	679	P_C, R_C, G_C (in μg)	*k*
Origins/cell	O_C	no./cell	1.96	2.43	3.36	4.70	6.54	C, D	*l*
Termini/cell	T_C	no./cell	1.23	1.37	1.54	1.74	1.94	D	*l*
Replication forks/cell	F_C	no./cell	1.46	2.14	3.64	5.92	9.19	C, D	*l*

Figure 1.23: **Important physiological parameters in *E. coli*.** The upper table shows values for important growth rate *independent* physiological parameters in *E. coli*. The lower table gives values for important growth rate *dependent* physiological parameters in *E. coli* growing at several different rates. Tables 1 and 2 from Bremer (1996) [19].

Parameter	Symbol	Equation	Equation no.	Reference(s)
Protein/cell	P_C	$P_C = P_O \cdot 2^{(C+D)/\tau}$	1	51
RNA/cell	R_C	$R_{C=K'}(P_{O(cp)}(1/\tau)2^{(C+D)/\tau}$, where $K' = (\text{nucl./rib}) \cdot \ln 2/[f_s \cdot (1 - f_f) \cdot \beta_r \cdot 60]$	2	27
DNA/cell	G_C	$G_C = [\tau/(C \cdot \ln 2)] \cdot [2^{(C+D)/\tau} - 2^{D/\tau}]$	3	32
Mass/cell	M_C	$M_C = k_1 \cdot P_C + k_2 \cdot R_C + k_3 \cdot G_C$, where: $k_1 = 1.35 \cdot 10^{-18}$ OD_{460} units per amino acid residue $k_2 = 4.06 \cdot 10^{-18}$ OD_{460} units per RNA nucleotide residue $k_3 = 3.01 \cdot 10^{-11}$ OD_{460} units per genome equivalent of DNA	4	27
Peptide chain elongation	c_p	$c_p = K'/[(R/P) \cdot \tau]$	5	44, 122
Ribosomal protein/total protein	α_r	$\alpha_r = (R/P) \cdot [(\text{aa/ribosome}) \cdot f_s \cdot (1 - f_f)/(\text{nucl./rib})]$	6	44, 122
Origins/cell	O_C	$O_C = 2^{(C+D)/\tau}$	7	15, 23
Termini/cell	T_C	$T_C = 2^{D/\tau}$	8	15, 23
No. of gene X/cell	X_C	$X_C = 2^{[C(1-m')+D]/\tau}$, where: m' = map location of gene X relative to location or replication origin $= (m + 16)/50$ for map locations (m) between 0 and 36 min $= (84 - m)/50$ for map locations between 36 and 84 min $= (m - 84)/50$ for map locations between 84 and 100 min	9	15, 23
Replication forks/cell	F_C	$F_C = 2 \cdot [2^{(C+D)/\tau} - 2^{D/\tau}$	10	15, 23
Origins/genome	O_G	$O_G = (C/\tau) \cdot \ln 2/(1 - 2^{-C/\tau})$	11	15, 23
No. of gene X/genome	X_G	$X_G = (O/G) \cdot 2^{-m'C/\tau}$	12	15, 23
Initiation age	a_i	$a_i = 1 + n - (C + D)/\tau$ where n is the next lower integer value of $[(C + D)/\tau]$; i.e., $n = \text{int}[(C + D)/\tau]$	13	32
Termination age	a_t	$a_t = 1 - D/\tau$	14	32
Origins per cell at initiation	O_i	$O_i = 2^n$; for a definition of n, see equation 13	15	32
Cell mass after division (a_0)	M_d	$M_d = M_C/(2 \cdot \ln 2)$	16	18
Cell mass at initiation (a_i)	M_i	$M_i = M_d \cdot 2^{a_i}$	17	18

Figure 1.24: **Simple mathematical rules of bacterial physiology.** Mathematical equations describing the composition of *E. coli* and how it relates to growth rate. Table 5 from Bremer (1996) [19].

1.6.2 Protein Partitioning Constraints

In 1960, Neidhardt and Magasanik discovered that ribosomes play a catalytic role in protein synthesis by showing that the RNA/protein ratio increases linearly with growth rate (above 0.6 doublings/hour) (Fig. 1.25) [122].

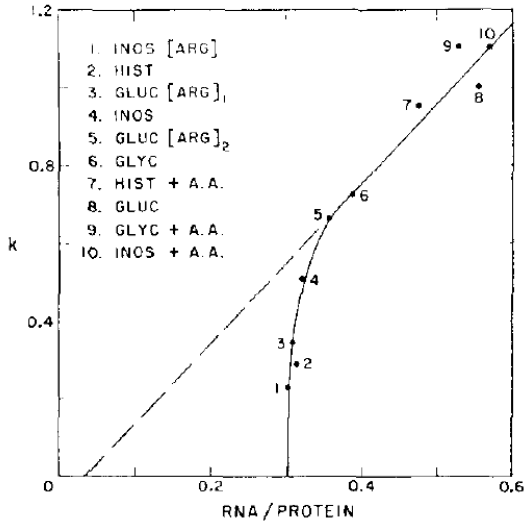

Figure 1.25: **Doubling rate versus RNA/protein ratio in *Aerobacter aerogenes*.** The y-axis "k" is doubling rate in doublings/hour while the x-axis is the ratio of RNA to protein in a culture of *Aerobacter aerogenes*. For sufficiently fast growth rates (above 0.6 doublings/hour) the RNA/protein ratio is linearly proportional to growth rate. Figure 2 from Neidhardt and Magasanik (1960) [122].

Due to the constraints of balanced growth, the total protein mass, M_P, grows exponentially with time at the same constant rate (λ) as the bacterial population, $\frac{dM_P}{dt} = \lambda M_P$. Assuming that proteins are built by N_{Rb} ribosomes translating polypeptides at a constant rate, k amino acids per ribosome per second, then the change in protein mass through time will be,

$$\lambda M_P = k N_{Rb} \,, \tag{1.56}$$

which rearranged gives,

$$\frac{N_{Rb}}{M_P} = \frac{\lambda}{k} \,. \tag{1.57}$$

Consequently, the ratio of ribosome number to protein mass is a constant. Ribosomes

are composed of ribosomal RNA (rRNA) and protein, the ratio[13] of which is growth rate independent [19]. Furthermore, a growth rate independent proportion of the total RNA in a cell is rRNA[14]. As a result, the total quantity of RNA, R, is a direct proxy for the number of ribosomes, meaning,

$$\frac{R}{M_P} \propto \frac{\lambda}{k}\,, \tag{1.58}$$

which is what Neidhardt and Magasanik observed in cells growing faster than 0.6 doublings per hour [122]. Another result of the relation in Eq. (1.57) and the fixed composition of ribosomes is that the quantity of total protein mass that is ribosomal, M_{rP}, is also growth rate dependent,

$$\frac{M_{rP}}{M_P} \propto \frac{\lambda}{k}\,. \tag{1.59}$$

Consequently, fast growing cells have more of their total protein resources allocated to ribosomes [160].

How the total protein mass[15] is allocated among different types of proteins is called the proteome. It has been shown that the proteome is split into three major categories. One category, which composes approximately half of the proteome, is a basal expression of proteins that is independent of growth rate [160, 199, 81]. The other half of the proteome is flexible and is split into two categories; one primarily includes proteins that relate to protein production, such as ribosomes, and the other includes the remaining proteins. In fast growing cells the protein producing category dominates the flexible half of the proteome while in slow growing cells it is the opposite (Fig. 1.26) [160]. The non-protein producing proteins dominate in slow growing cells because they relate to metabolism and the slow growing cells are using more complex carbon sources which require more work to turn into useful chemicals [199, 81].

[13] Approximately 2 mg rRNA to 1 mg protein [19].

[14] It has been found to be approximately 86% [19].

[15] It can be deduced from Schaechter et. al's work that the quantity of protein per cell is largely a function of growth rate alone and that it increases with growth rate.

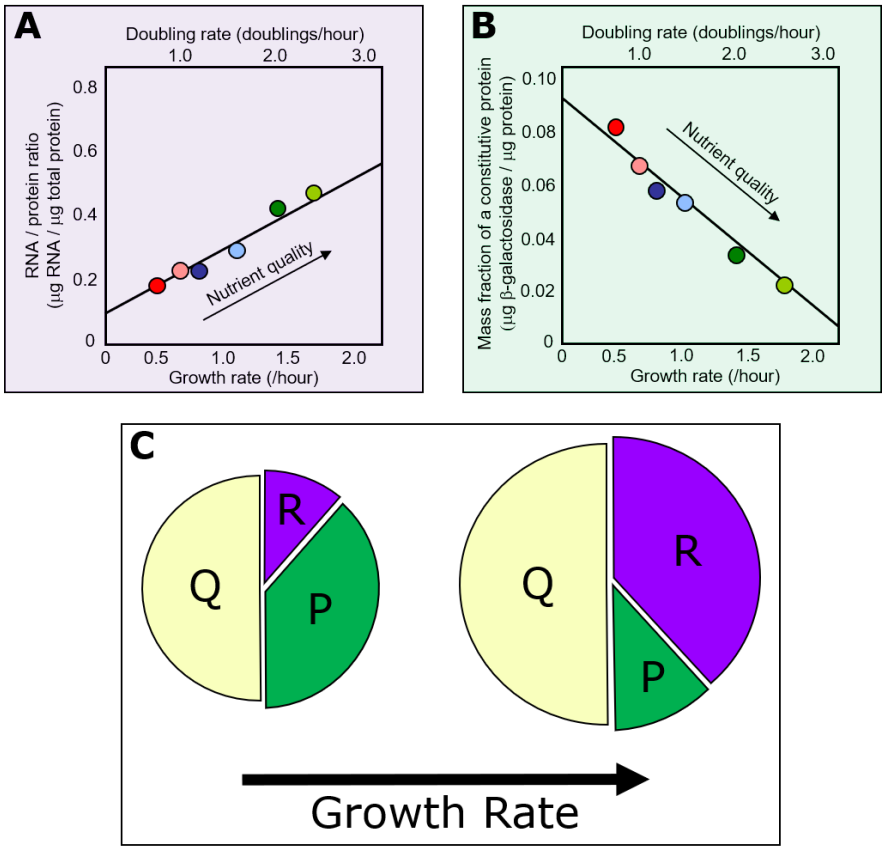

Figure 1.26: **Protein partitioning in slow and fast growing *E. coli* cells.** A) The RNA/protein ratio, which is a direct proxy for the amount of ribosomal protein (represented by "R" in (C)), with different nutrient-modulated growth rates. B) The mass fraction of a constitutive protein (β-galactosidase under the control of a synthetic TetO1 promoter)(part of "P" in (C)) with different nutrient-modulated growth rates. C) A pie chart representation of the proteome at different growth rates. The proteome is larger for fast growing cells because the protein/cell is positively correlated with growth rate. The partitioning is different due to the positive correlation of ribosomal protein with growth rate, meaning the other proteins are necessarily anti-correlated (as seen in panel (B)). The "Q" portion represents a basal protein expression that is growth rate independent, "R" represents ribosomal proteins which are proteins that participate in protein production, and "P" represents the remaining proteins. Panels (A) and (B) are Figures 2A and 2C respectively from Scott et al. (2010) [160].

1.7 Growth Rate - Mutation Rate Coupling

Like most science, my journey started with a question: Is a bacteria's mutation rate dependent on their balanced growth rate? I believe both potential answers to this question have interesting implications. If no, this means bacterial mutation is either decoupled from the rest of its physiology, or there are at least two cellular mechanisms that are cancelling each other out in order to maintain a constant mutation rate across all exponential growth rates. If yes, this means the genotype of the cell is intricately connected to the bacteria's environment. A coupling between the mutation rate and the growth rate also implies that the cells have developed an evolutionary feedback loop based on how well they are growing.

In a simple minded sense, it is reasonable to expect some coupling between the mutation rate and the growth rate simply because most of a cell's physiological attributes are coupled to the growth rate. A more convincing argument would lay out the mechanisms which could cause the coupling. The most obvious framework for this argument would be based on the proteome's growth rate dependence. It has already been discussed that DNA replication is performed by proteins, so the next question would be which category of the proteome do these proteins reside? Of particular interest are the DNA repair proteins that correct the errors that the primary DNA polymerases make. To further this argument, one can employ the fact that both the quantity of DNA/cell and protein/cell increases with growth rate, but the increase is slower in the DNA/cell [19]. In addition, the amount of DNA synthesis being performed in a cell during its life increases with growth rate. How do the cells then manage their proteins that tend to the DNA as the growth rate increases? Are the DNA repair proteins increased in a way that is exactly proportional to the increase in the amount of DNA replication being performed or otherwise, and how would either affect the mutation rate? Despite there being one primary cause of mutations in the absence of mutagens, there are many routes towards mutation rate - growth rate coupling, all of which can be combined in difficult-to-predict ways.

With all this in mind, it is of no surprise that people have questioned if mutation rate is growth rate dependent in the past [106, 124, 4, 87, 201, 10, 42, 94, 92]. In fact, even Luria and Delbrück mentioned it in their original 1943 paper by saying "the chance (of mutation) may vary in some manner during the life cycle of each bacterium and may also vary when the physiological conditions of the culture vary...it seems reasonable to assume that the chance (of mutation) is proportional to the growth rate of the bacteria"[16] [106].

Another argument to motivate the question of growth rate - mutation rate coupling

[16]The particular quote comes from a point in the paper when developing the mathematical theory and arguing for the choice of the doubling time as the base time unit, but its implications are far reaching.

comes in the context of stress induced mutation. It is known that the SOS response results in a higher expression of a number of proteins that perform the task of DNA repair, often in an error prone manner [53]. These proteins are not co-regulated with ribosomal proteins, and so they are anti-correlated with the growth rate (Fig. 1.27). Consequently, it seems reasonable to hypothesise that there would be some form of growth rate dependence for mutation rate in the presence of a mutagen that activates the SOS response.

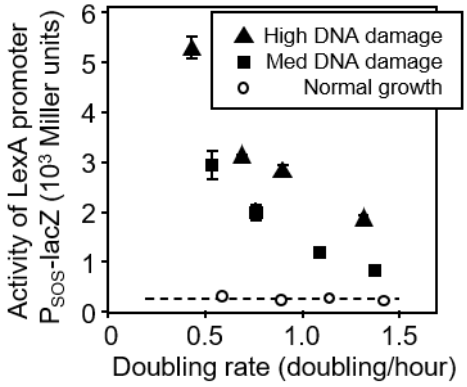

Figure 1.27: **Growth rate dependence of SOS response proteins.** The SOS response is regulated by the protein LexA which acts to repress the SOS response until it senses DNA damage at a stalled replication fork [49, 25]. The DNA damaging antibiotic Mitomycin C (MMC) is introduced at two different concentrations. At higher concentrations of MMC, the activity of the LexA promoter increases, implying the induction of the SOS response. On the graph, the open circle relates to 0 mg/mL MMC, the squares relate to 0.02 mg/mL MMC, and the triangles relate to 0.2 mg/mL MMC. This figure is of unpublished preliminary data from Matthew Scott.

Only a few other groups have attempted to experimentally probe for coupling between the spontaneous mutation rate and the bacteria's environment to my knowledge. One of these was performed by Ram Maharjan and Thomas Ferenci in 2018 when they performed fluctuation tests by growing cells in a turbidostat[17] with different media [110]. They modified the media in two ways, by changing the concentration of glucose, and by changing if the culture had access to oxygen. Of particular interest is the change in concentration of glucose, as it changes the growth rate in much the same way as that discussed throughout Section 1.6.1 [117]. Maharjan and Ferenci found that, in the presence of oxygen, the

[17] A turbidostat is a machine which produces continuous cultures. See Appendix A.3.2 for more information.

mutation rate increases in slower growing cells [110]. Though this is a step in the right direction towards addressing the question about whether the mutation rate and growth rate are coupled, their methodology is problematic because fluctuation tests were not designed to be performed in continuous cultures, and it doesn't explore the modulation of growth rate through varying the carbon source itself.

The purpose of this thesis is to attempt to observe the relationship between the spontaneous mutation rate and the exponential growth rate of *E. coli* by performing fluctuation tests that, as closely as possible, follow the assumptions of the Lea-Coulson model while varying growth rate by changing the nutrient quality, as is traditional to studies of bacterial physiology. The experimental methodology I developed to do this is detailed in Chapter 2. In Chapter 3, the analysis tools developed and used to analyse the results of the experiments are presented and explored. Next, the results of the experiments are provided in Chapter 4. Finally, in Chapter 5, I summarise the results, discuss their implications, and detail potential future work on the subject.

Chapter 2

Experimental Methods

To test the coupling between a bacteria's growth physiology and mutation rate requires a careful growth protocol. In this chapter, I discuss issues with the traditional fluctuation test methodology and motivate a new approach. I then detail the experimental methodology of this new approach, including data from control experiments and an in depth discussion on the selection agent chosen for my fluctuation tests.

2.1 Steady State Fluctuation Tests in Different Growth Media

Many fluctuation tests have been performed since Luria and Delbrück's seminal work, but few have strayed much from the original methodology [137, 169, 45, 124, 26, 51, 101, 18]. This is surprising considering how much has been discovered about bacterial physiology since the original fluctuation tests, as laid out in Section 1.6. Because the primary goal of any quantitative experiment is to have it obey the model assumptions as closely as possible, in the case of a fluctuation test, the goal is to have the experiment obey the Lea-Coulson assumptions in Section 1.5.2. One important assumption, which is indirectly held in assumptions 1 and 3, is that the cells have the same physiology throughout growth. The issue with Luria and Delbrück's original methodology is that the bacteria are left to grow overnight so that they reach saturation [106]. Consequently, the cells transition from exponential phase to stationary phase at some point. Through this transition, the physiology of the bacteria changes, which could have unexpected affects on the mutation

rate, possibly breaking the 3rd assumption of the Lea-Coulson model[1] [59, 170, 106, 124, 18]. Because the population of bacteria is close to its maximum, but still growing during the physiological change from exponential to stationary phase, based off Delbrück's definition of m (Eq. (1.5): $dm = \mu N_t dt$), many new mutations are likely to occur at this time, as seen in Fig. 2.1. If there is indeed any change in the mutation rate during this period, the effect will then be magnified by the large number of doublings occurring. The ironic part is it seems many biologists take great care to seed their experiments with exponentially growing cells [137] even though at the beginning of the experiment when the population is small, the probability of a mutation occurring is very low. Furthermore, assumption 1 is compromised by the fact that cells spend a period of time decelerating and in stationary phase, and therefore not growing exponentially. If mutations continue to occur during stationary phase, which there is evidence they do[2] [59, 23, 113, 136], this can be especially problematic because the model will not account for these mutations and categorically result in an overestimate on the mutation rate. Fortunately, if the cells require a period of growth before expressing the mutant phenotype (i.e. phenotypic lag), then it is likely that the majority of these mutants that arise during the stationary phase will not survive the selection phase of a fluctuation test, assuming the selection is immediate and does not also require a period of growth to take effect. If selection is not immediate though, difficult-to-predict dynamics could occur, diminishing the quantitative power of the fluctuation test. Without a lot of extra analysis and control experiments, the issues caused by the deceleration and stationary phases can not be accounted for in any reasonable way.

Fluctuation tests are labour intensive at the start and end of the experiment, but during the long period of growth in the middle they require no input. Consequently, it is appealing to grow the cells overnight. Furthermore, it is ideal to have some samples with zero mutants as well as having as many countable selection plates as possible, which puts a practical limit on how long the cells can be grown for. To keep the final populations at a desirable number during overnight growth requires the limitation of a nutrient so that growth halts, as described by Monod [117] (Fig. 1.19). Although this makes for easily controllable and

[1] This problem and its consequences are even discussed by Luria and Delbrück (1943) [106] and expanded upon by Newcombe (1948) [124]. Newcombe attempted to run a control in which they grew cells purely in exponential phase to see if it made a difference, but they did so on plates, making it incomparable to the batch culture fluctuation test results. Stewart et al. (1990) [170] appears to be the first to try to adjust the model by introducing nutrition depletion and assumptions on how nutrition would affect the mutation rate to derive altered expressions for the expected number of mutants. Despite the problem being addressed on a number of occasions, no one appears to have attempted to explicitly account for it in their experiments since Newcombe.

[2] There is even suggestions that the mutation rate could increase during stationary phase due to stress induced mutagenesis [59].

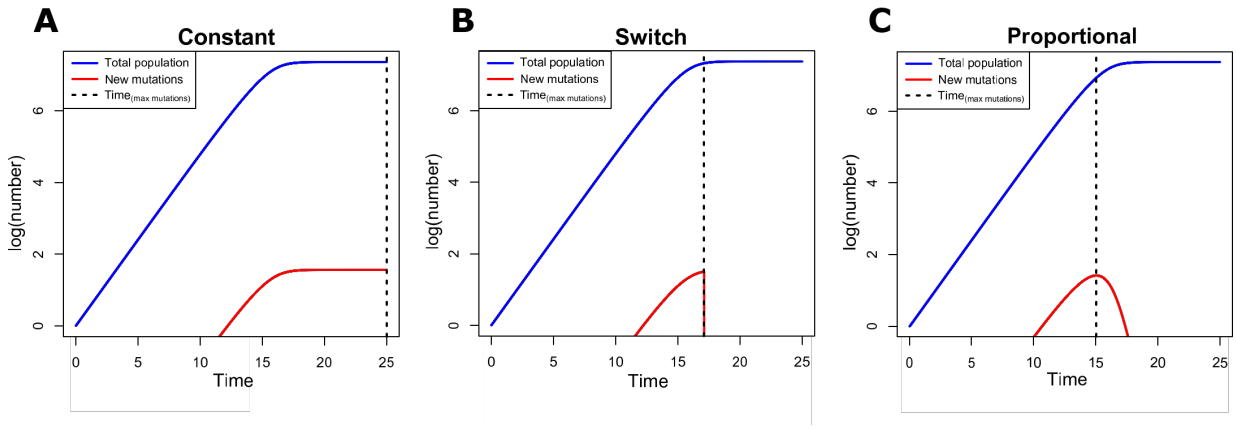

Figure 2.1: **Total bacterial population and expected number of new mutations in a culture during late exponential phase into stationary phase.** The total bacterial population (blue line), N_t, is modelled using Monod kinetics and the expected number of new mutations (red line), m_t, is calculated using $m_t = \mu N_t$ where the mutation rate, μ, is made dependent on growth rate in three different ways. A) μ remains constant regardless of growth rate. B) μ goes to zero when the growth rate falls below 10% of its maximal value. C) μ is directly proportional to growth rate with proportionality constant $\frac{3}{\lambda_{\max}}$. The time when the maximum number of new mutations are expected to appear is shown with the dotted black line; notice how in (A) it is in stationary phase, in (B) it is near the end of the transition from exponential phase to stationary phase, and in (C) it is near the beginning of the transition from exponential phase to stationary phase. For all three cases, a significant number of new mutations are arising either during the transition from exponential to stationary phase, or in stationary phase itself. Simulated growth with Monod kinetics by numerically solving the coupled ordinary differential equations: $\frac{dN}{dt} = \lambda N$ and $\frac{dS}{dt} = -\alpha\lambda N$ where $\lambda = \lambda_{\max}\frac{S}{K_D+S}$, $\lambda_{\max} = \ln 2$, $K_D = 2.2 \cdot 10^{-5}$, $\alpha = 10^{-7.5}$, $N_0 = 1$, $S_0 = 5 \cdot 10^{-5}$, and $\mu_0 = 3 \cdot 10^{-3}$.

consistent final populations, it results in the issues caused by cells transitioning to stationary phase. One popular way researchers have gotten around these problems, consciously or not, is by studying mutation rates with continuous cultures instead of batch cultures [129, 96, 110] (see Appendix A.3 for descriptions of batch and continuous cultures). The method used to determine mutation rates from continuous cultures is different than a fluctuation test in that it is based on mutant accumulation [58]. In order for this method to work, there must be a significant number of mutants, which requires a long period of growth. Unfortunately, continuous cultures have a number of problems, most of which are

exacerbated by long growth. One problem with continuous cultures is that the machines preferentially select mutations that promote surface growth because these mutants will not get removed during dilution, introducing a selection bias [194]. Another problem is that it is easier for fast growing mutants to take over continuous cultures, which could result in a bias in which phenotypes make it to the end of the experiment [194]. Both of these problems show that the machines which maintain continuous cultures do not perform a neutral selection, which imposes a practical limitation on how long a continuous culture experiment can be run. Moreover, the fluctuations that are inherent to Luria and Delbrück's system can be detrimental to the mutant accumulation method [58, 8] Consequently, using continuous cultures potentially introduces just as many issues as it resolves [52].

The lack of care for physiology during fluctuation tests has led to an uncertainty in the results of the past century, especially regarding their reproducability. Their significance can also be brought into question, though I don't believe it necessary to discard all earlier results. The problem is simply that earlier results aren't describing what they claim to be describing. Instead, the experiments are determining the average spontaneous mutation rate over several generations of exponential growth, the deceleration phase, and potentially an unknown amount of time in stationary phase. This is likely just as indicative of what cells experience in nature [91], but it is difficult to determine how the specific attributes of the media, as well as the amount of time spent in stationary phase, affected the mutation rate, taking away any potential for a deeper quantitative understanding of the process.

With all of the issues of growing bacteria to saturation and using continuous cultures in mind, I set out to do fluctuation tests in a different way, which would keep the cells in exponential phase for the entirety of their growth[3] while in batch culture. In addition, my fluctuation tests would be done in different growth media, allowing for different exponential growth rates. With the cells being in balanced growth throughout the experiment, they will have reproducibly different physiologies in different mediums as described in Section 1.6. Consequently, direct analysis of whether the bacteria's physiology, and by extension environment, affects the mutation rate of bacteria in a quantitatively verifiable and reproducible way can be performed.

[3]With the exception of potentially at the very beginning, but as mentioned earlier, this should have minimal consequences.

2.2 Strain and Media

The experiments discussed in this thesis are completed using a wild-type *Escherichia coli* strain called NCM3722 which, like many strains used in the lab, is a K12 derivative. NCM3722 was chosen over MG1655, a particularly popular strain, because NCM3722 appears to be genetically closer to the original *E. coli* K12 [167]. The hope is that since K12 was directly cultivated from nature, then NCM3722 should be a better proxy for naturally occurring *E. coli*, as well as being less likely to have genetic adaptations to lab growth, which could mean it has a wider variety of robust physiological phenotypes [107]. NCM3722 has a $4.7 \cdot 10^6$ base pair chromosome and $6.7 \cdot 10^4$ base pair F-like plasmid that have been fully sequenced [21]. The cells lack flagella, meaning they have no control over movement, which should make it easier to get homogeneous solutions of cells in culture. Figure 2.2 shows microscope images of *E. coli* NCM3722 cells growing in differing qualities of growth medium and therefore expressing different physiologies, as apparent from their differing sizes.

There are two types of growth media that can be used in the lab: defined media and undefined media. Undefined media are ill-defined chemical solutions that allow for cell growth. Lysogeny broth (LB), an undefined medium, is the most popular medium used in microbiology because it is cheap and gives incredibly fast growth. It is made from tryptone, sodium chloride, and yeast extract [14]. The problem is that yeast extract is produced by growing and lysing yeast with little care for the specifics of how they are grown, so the details of its contents are mostly unknown [93, 141]. Furthermore, none of the nutrients in yeast extract are present in saturating amounts, meaning the bacteria undergo many minor nutrient shifts, resulting in the cells never truly being in balanced growth [125, 162]. Another consequence is that the inconsistencies in yeast extract are likely to result in chemical differences between different batches of LB which can be a large hindrance on reproducability [188]. As such, when a quantitative approach that requires reproducible growth is being employed in the lab, undefined media are best not used, and instead defined media should be used. Defined media are media in which the concentrations of all the components are fully understood, allowing for much greater reproducibility and confidence in the cells' physiology experiment to experiment. For the experiments discussed in this thesis, defined media were used. The media which were used during the growth phase of the fluctuation test use 3-(N-morpholino)propanesulfonic acid (MOPS) as a buffer, with variable carbon sources. The MOPS based media were developed in the 1970's by Frederick Neidhardt specifically for use in quantitative studies of bacterial physiology [121]. The complete MOPS media recipes can be found in Appendix B.1.

The carbon source used in the MOPS medium can be changed to give a different growth

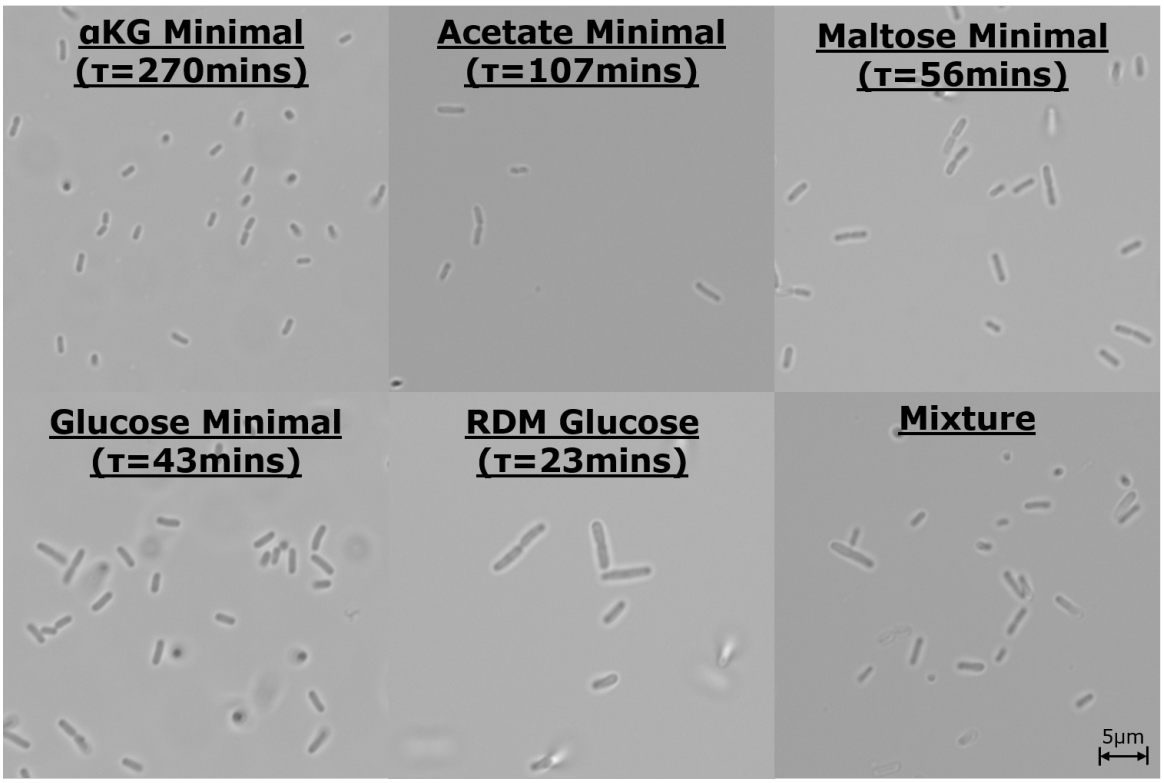

Figure 2.2: ***E. coli* NCM3722 microscope images.** Microscope images of *E. coli* NCM3722 taken from exponential growth in different medias. The doubling time for each media is given below the medium name. Note how cells with shorter doubling times, and therefore faster growth rates, are larger.

rate while having every other attribute of the medium stay the same. In this thesis we report full data from experiments using glucose and maltose as the carbon source, as well as some preliminary data using α-ketoglutarate (αKG) and acetate as carbon sources. The main growth media used in this study were minimal maltose medium (0.2% (w/v) maltose in MOPS minimal buffer; recipe in Appendix B.1.5.), and rich defined glucose medium (0.2% (w/v) glucose in MOPS minimal buffer supplemented with nucleotides (ACGU) and amino acids (EZ); recipes in Appendices B.1.4, B.1.2, and B.1.3). The minimal maltose medium will be colloquially referred to as "maltose minimal" and the rich defined glucose medium will be referred to as "RDM glucose". Acetate and α-ketoglutarate were both used to form a MOPS based minimal medium like maltose minimal.

The exponential growth rates of the cells in all relevant media were determined using

the turbidity of the culture (i.e. light scattering at 600nm, called the optical density (OD) or more specifically OD_{600}) and the viable cell count (called colony forming units (CFU))[4]. Although turbidity should be proportional to the cell number density [168], the growth rate as determined by CFU was categorically faster than the growth rate determined by OD when CFU growth is started from a small number of cells and less than fifteen doublings were observed[5]. The growth rates as determined by OD and CFU for NCM3722 in RDM glucose, maltose minimal, and αKG minimal can be found in Table 2.1. These growth rates and their standard deviations were determined by averaging and comparing the slopes of the lines of best fit for several independent log-linear growth curves such as those shown in Fig. 2.3.

Media	OD λ (/hr)	OD τ (min)	CFU λ (/hr)	CFU τ (min)
RDM glucose	1.727 $\pm$ 0.067	24.10 $\pm$ 0.93	1.796 $\pm$ 0.037	23.16 $\pm$ 0.49
Maltose minimal	0.741 $\pm$ 0.018	56.14 $\pm$ 1.41	0.877 $\pm$ 0.093	47.78 $\pm$ 4.89
αKG minimal	0.150 $\pm$ 0.004	277.08 $\pm$ 6.88	0.393 $\pm$ 0.033	106.29 $\pm$ 8.70

Table 2.1: ***E. coli* NCM3772 growth rates and doubling times.** Specific growth rates (λ) and doubling times (τ) plus or minus one standard deviation for *E. coli* NCM3772 in RDM glucose, maltose minimal, and αKG minimal MOPS based media as determined using both OD_{600} and CFU. Means and standard deviations are from comparing several independent growth curves. See Fig. 2.3 for respective R^2 values.

OD works as a relative measure when determining growth rate because it is a proxy for cell mass[6], but without calibration it cannot give a reliable indication of the number cells that are in a sample due to the growth dependence in the average cell size, as seen in Figures 1.20 and 2.2 [168, 153]. Because fluctuation tests require a consistent seed of a small number of cells, it is imperative that a cell concentration can be determined from a measured OD. As such, several growth curves were completed where the OD and CFU were measured simultaneously at each time point. This allows one to create a plot for CFU vs. OD where the equation of the line of best fit can be used to determine the cell concentration at any OD as seen in Table 2.2 and Fig. 2.4.

[4]See Appendix A for more info on OD and CFU.
[5]This apparent inconsistency is discussed in more detail in Section 4.4.
[6]See Appendix A.1 for details.

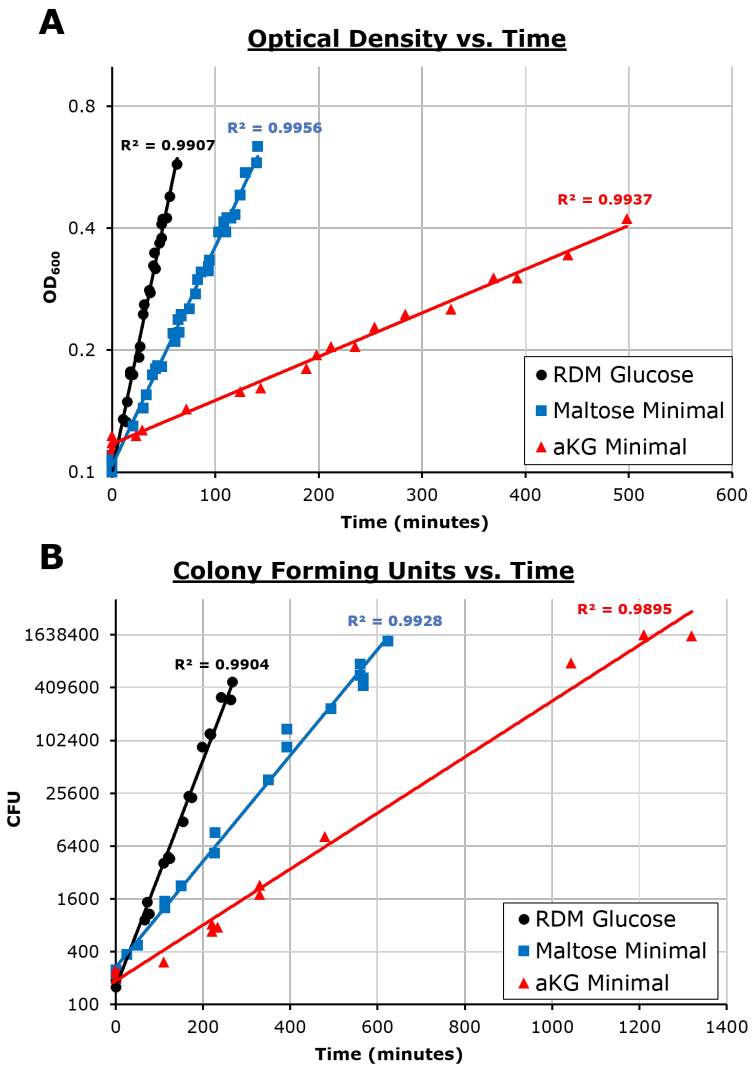

Figure 2.3: ***E. coli* NCM3722 OD_{600} and CFU growth curves.** A) The 600nm optical density (OD_{600}) of a sample of culture growing in exponential phase at different times is plotted on a semi-log plot for RDM glucose, maltose minimal, and αKG minimal. B) The number of colony forming units (CFU) in a sample of culture at different times during exponential growth is plotted on a semi-log plot for RDM glucose, maltose minimal, and αKG minimal. In both plots the slope of the line of best fit is the growth rate and the coefficient of determination, R^2, is provided for each line. The data is a compilation of several independent growth experiments.

Medium	OD_{600} to Cell Concentration
RDM glucose	$CFU/mL = 5.67 \cdot 10^8 \times OD_{600}$ - $7.50 \cdot 10^6$
Maltose minimal	$CFU/mL = 1.09 \cdot 10^9 \times OD_{600} + 5.18 \cdot 10^6$
αKG minimal	$CFU/mL = 3.54 \cdot 10^9 \times OD_{600}$ - $4.86 \cdot 10^7$

Table 2.2: ***E. coli* NCM3772 OD_{600} to cell concentration.** Equations for determining the number of viable cells (CFU) per millilitre from the optical density of a culture at 600nm (OD_{600}). Note that the intercept is at least two orders of magnitude smaller than the slope in all three cases, meaning it can reasonably be ignored in most cases. See Fig. 2.4 for respective R^2 values.

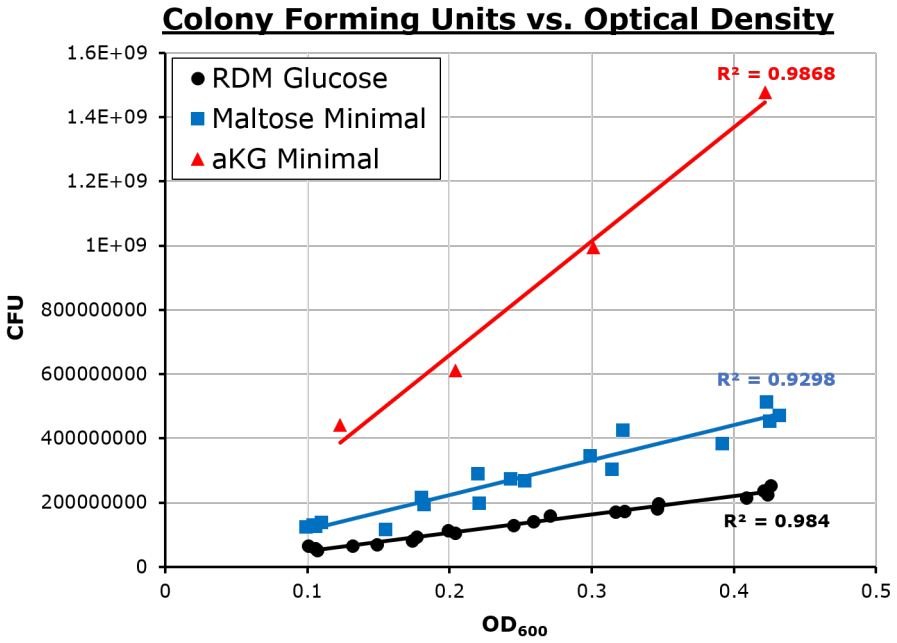

Figure 2.4: ***E. coli* NCM3772 viable cell counts (CFU) versus optical density (OD_{600}).** The number of colony forming units (CFU) in a sample of culture is plotted versus its 600nm optical density (OD_{600}) for RDM glucose, maltose minimal, and αKG minimal. The equation of the line of best fit for each medium allows for the calculation of the cell concentration from the OD_{600}. The difference in slopes is due to fast growing, larger cells (black) being more efficient light scatterers than the slow growing, smaller cells (red). The coefficient of determination, R^2, is provided for each line.

2.3 Mutant Selection

Potentially the most important experimental choice when designing a fluctuation test is how mutants will be distinguished from the wild type bacteria. The selecting agent is what probes for the mutation and allows one to make this distinction. Consequently, it is imperative that a good selecting agent which will accurately select a mutation in a well defined gene, is chosen. In other words, a good selecting agent should only have one path towards selection, otherwise it would probe for a variety of mutations in different genes, which would be very difficult to account for [51]. Also, the mutation which allows the mutants to survive the selection should ideally not affect the growth rate of the cells [51, 99, 106]. If this cannot be achieved, it can be mathematically accounted for in the models [92, 201, 210], but one will have to do many control experiments to determine the average growth rate of the mutants. Lastly, the mutations that are selected for should be of any type (point mutations, insertions, deletions) [51]. A lot of commonly used selection agents don't satisfy all these conditions [51, 48].

One selecting agent that has all of the attributes of a good selector is D-cycloserine (often referred to as just cycloserine or cyc) [51]. Cycloserine is an antibiotic that is an analogue of the amino acid D-alanine [55, 9]. It is cyclic in shape (Fig. 2.5) and is transported into the cell by the same permease protein that transports D-alanine, L-alanine, D-serine, and glycine into the cell [37, 143, 191]. Once in the cell, cycloserine inhibits D-alanyl-D-alanine ligases A and B, and alanine racemase activities, which interrupts cell wall creation [97, 128]. The permease protein that transports cycloserine resides on the outer membrane of *E. coli* K12 and is coded for by a single copy 1413-bp gene called *cycA* [51]. If *cycA* is disabled through mutation, the cells stop making these permease proteins (referred to as CycA), and cycloserine can no longer enter the cell; this is a simple

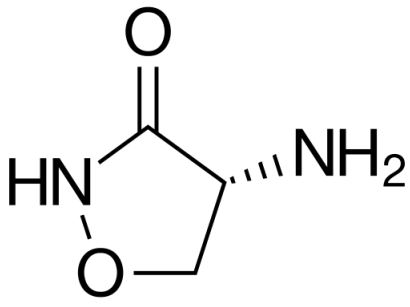

Figure 2.5: **D-cycloserine chemical structure.** Note the cyclic shape of the chemical. "Cycloserine" by Yikrazuul [184].

genotype-phenotype relationship which is ideal. It has also been suggested that *cycA* can be disabled through any type of mutation on any part of the gene[7], and the mutation does not change the growth rate in minimal media [51]. When growing *E. coli* in media with amino acids, one may expect some sort of growth inhibition when *cycA* is switched off, but since it would only partially inhibit the uptake of three amino acids and most mutations happen near the end of the experiment, the resulting effects should be negligible. The bigger issue with cycloserine as a selection agent comes from the fact that it selects for the absence of a permease protein. As a result, the CycA proteins present at the time of mutation must be diluted out through growth in order for the cell to become resistant to cycloserine, resulting in phenotypic lag (details in Section 3.2) [29]. Phenotypic lag can be adjusted for during model fitting (details in Section 3.3), but this is certainly not ideal. Regardless of the issue of phenotypic lag, cycloserine still makes for arguably the best selecting agent in fluctuation tests because its advantages over other popular agents are clear and abundant, and the other commonly used selecting agents also likely suffer from phenotypic lag [124, 18, 170, 101, 29]. As such, cycloserine is used as the selecting agent in the experiments discussed throughout this thesis.

In order to use a selecting agent, one must know at which concentrations it works on the cells being studied. Accordingly, cycloserine inhibition curves were experimentally determined for all relevant media. This was done by growing the cells in the medium of interest with several different concentrations of cycloserine. First they were grown overnight so the bacteria could adapt their physiology to the media with antibiotic. The cells were then diluted into the same medium and concentration of cycloserine they adapted to, and growth rates were measured using OD once they reached exponential phase. The susceptibility to the antibiotic was quantified by the concentration of antibiotic which causes the growth rate to be half of what it is in the absence of the antibiotic (called the half-inhibition concentration (IC_{50})). The inhibition curves for RDM glucose, glucose minimal, maltose minimal, and acetate minimal can be seen in Fig. 2.6 while the IC_{50}'s can be seen in Fig. 2.7, from which it is clear that slower growing cells are more susceptible to cycloserine. It was found that a cycloserine concentration of 100μM consistently results in zero growth, so this concentration was chosen to be used for all media during the selection phase of the fluctuation tests.

[7]It has been found that a wide variety of mutation types and locations resulted in a defective *cycA* gene, which was taken as evidence that "any" mutation will disable the gene [51].

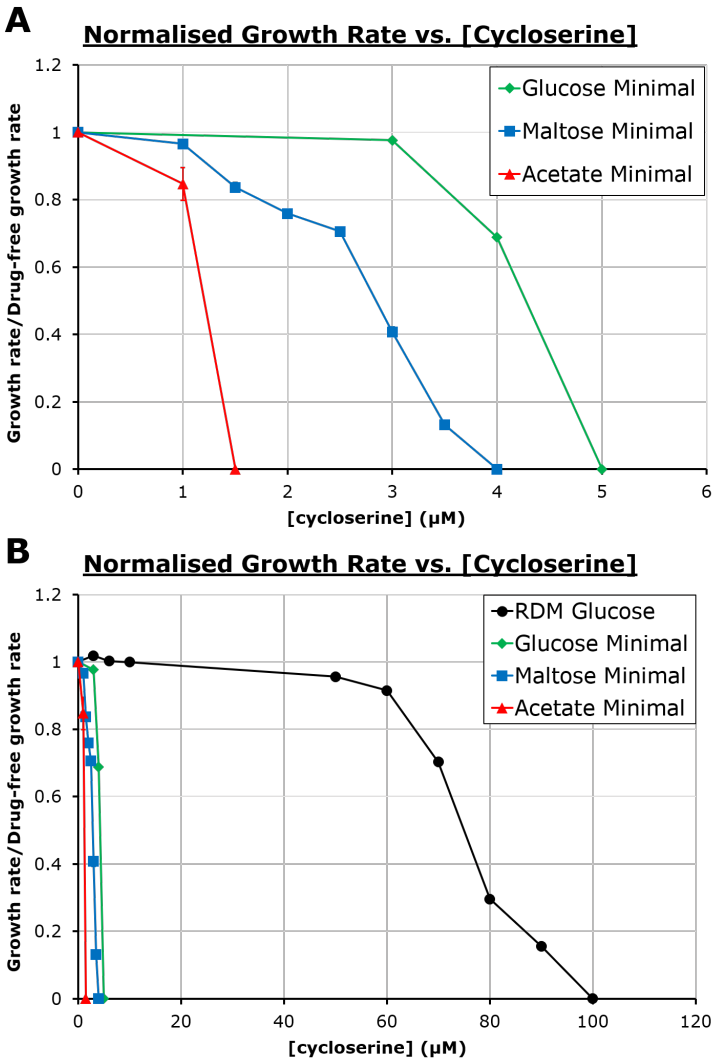

Figure 2.6: ***E. coli* NCM3722 cycloserine inhibition curves.** *E. coli* are grown in different concentrations of cycloserine for which their exponential growth rates are determined then divided by the drug-free growth rate. These normalised growth rates are plotted versus their respective cycloserine concentrations. A) The inhibition curves for *E. coli* NCM3722 in glucose minimal, maltose minimal, and acetate minimal. B) The inhibition curves from (A) plus RDM glucose, which requires much more cycloserine to be inhibited. For both plots error bars are plus or minus one standard deviation where applicable.

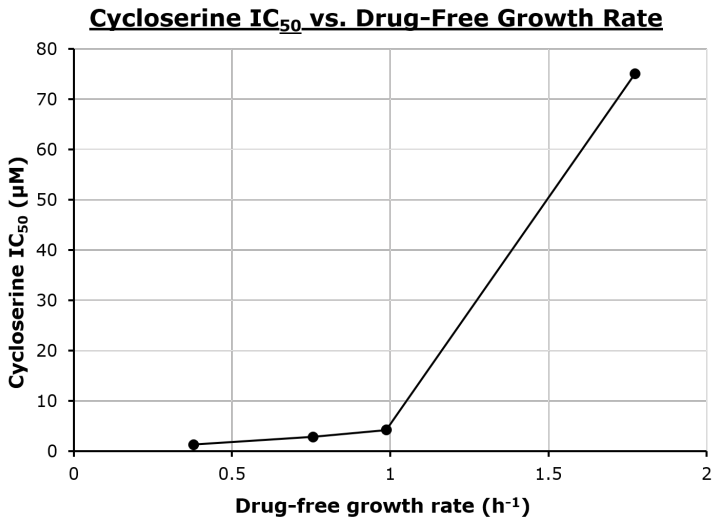

Figure 2.7: ***E. coli* NCM3722 cycloserine half-inhibition concentration versus drug-free growth rate.** The concentrations of cycloserine at which cultures grow at half their uninhibited rate (IC_{50}), as determined by Fig. 2.6, are plotted versus their uninhibited growth rate.

2.4 Experimental Outline

A fresh stock plate was prepared monthly by streaking a -80°C glycerol stock of *E. coli* NCM3722 onto a sterile 1.5% agar + LB plate, which was then incubated for 12-18 hours, sealed with parafilm, and stored in a 4°C fridge. A single colony was taken from the stock plate, inoculated into a test tube with LB, and then incubated in a 37°C water bath, shaking at 250 rotations per minute (rpm). Once the tube of LB and cells was noticeably turbid, 10μL was transferred to a tube with 1mL of the medium of interest (RDM glucose, maltose minimal, or αKG minimal). This tube was then put in the shaker bath overnight to allow for the bacteria to fully adapt to the new medium. The next morning the OD was measured in order to calculate the necessary dilution for the cells to reach an OD of approximately 0.3 after one to two hours of growth, which is sufficient time for the cells to reach exponential phase. The calculated dilution was then performed in a tube with 1mL of medium and put in the shaker bath to grow. When the culture was at an OD of 0.3-0.4, the cell concentration was calculated using the data from Table 2.2. The cells were then serial diluted in 4°C MOPS buffer solution and a 14.2mL "master mix" was prepared at a concentration of 1000 cells per 200μL of buffer. For RDM glucose experiments this master mix was then left to sit for 30 minutes[8][9], while in the maltose minimal and αKG minimal experiments the next phase was immediately commenced. 60 tubes were prepared beforehand with 300μL of buffer plus sufficient carbon source (and supplements for RDM) for a 500μL final volume and put in the shaker bath for approximately 30 minutes to warm to 37°C. 200μL of master mix was then put into all 60 tubes at intervals of 30 seconds, retrieving the tube from the shaker bath at the beginning of the 30 seconds, and returning it at the end. After the 200μL of master mix was added to each tube, the tube was swirled by hand in order to mix the cells with the carbon solution. After every fifteen tubes were inoculated, a 5 minute and 30 second break was taken, during which 50μL of the master mix was pour plated[10] onto one or two LB + agar plates in order to determine the initial inoculum size and how it varies over the inoculation period. Initial inoculum plates were also made at the beginning and end of the inoculation period.

The tubes were then left to grow in the shaker bath to reach a final population of

[8]When exponentially growing bacteria's carbon sources are removed, they continue growing for a period of time. This effect is more noticeable in fast growing cells, so in order to make sure that the variance in number of cells per inoculum between the beginning and end of the inoculation period are comparable, one needs a settling period. Data for how much the cells continue to grow after being diluted into cold buffer can be found in Appendix C.1.

[9]The calculation of the number of cells in a sample from the OD must be adjusted for the extra growth during the settling period in the buffer.

[10]See Appendix A.2 for information about the plating protocol

approximately $4.5{\cdot}10^5$ cells[11]. During their growth, a fresh solution of 10mM cycloserine was made from powder and filter sterilised. In addition, 60 plates with a base layer of 1.5% agar plus M9 minimal[12] with glucose ("M9 glucose") were prepared for selection and cell counting, 50 of which contained cycloserine at a concentration of 100μM. Once growth was complete, 1mL of ice cold MOPS buffer solution was added to each tube at the same 30 second intervals as the inoculation period. The buffer added to 50 of the tubes had cycloserine in it for a final concentration of 100μM in the 1.5mL solution. The buffer added to tubes 13, 14, 15, 29, 30, 44, 45, 58, 59, and 60 did not have cycloserine; these tubes were serial diluted and plated during the 5 minute and 30 second breaks in order to determine the final cell population.

After adding the buffer, the outside of the remaining tubes were cleaned with ethanol and the entire contents of the tubes were plated on the plates containing cycloserine by adding 3mL of 1% agar + M9 glucose + 100μM cycloserine to the tube, swirling, and pouring directly onto the plate. All the plates were then placed in a 37°C air incubator. The initial inoculum plates were left to incubate overnight, while the population plates were left to incubate until colonies were reasonably sized for counting (generally 18-36 hours). All initial and final population plates were counted by hand. The selection plates were left to incubate for approximately 48 hours in the RDM glucose experiment or 60 hours in the maltose minimal experiment. The selection plates were counted and objects of unclear origin were marked. The selection plates were then incubated for another 12 hours and counted again, with special focus on checking the unclear objects for further growth to see if they were colonies. The data was then recorded and analysed using the methods described in Chapter 3. Figure 2.8 outlines the experiment's key steps in the form of a flow chart while Table 2.3 has links to videos of me performing the experiment.

Length	Link
1 minute	https://youtu.be/B5FfjP6Vm9w
3 minutes	https://youtu.be/gct4ji-V0yA
5 minutes	https://youtu.be/Z1Ft_cqK_oU

Table 2.3: **Links to videos of a fluctuation test being performed.** An 8 hour fluctuation test with *E. coli* NCM3722 in RDM glucose was recorded and sped up to be 1 minute, 3 minutes, and 5 minutes.

[11] 3 hours and 33/34 minutes for RDM glucose; 7 hours and 6 minutes for maltose minimal; 18 hours and 45 minutes for αKG minimal.

[12] See Appendix B.2 for recipe.

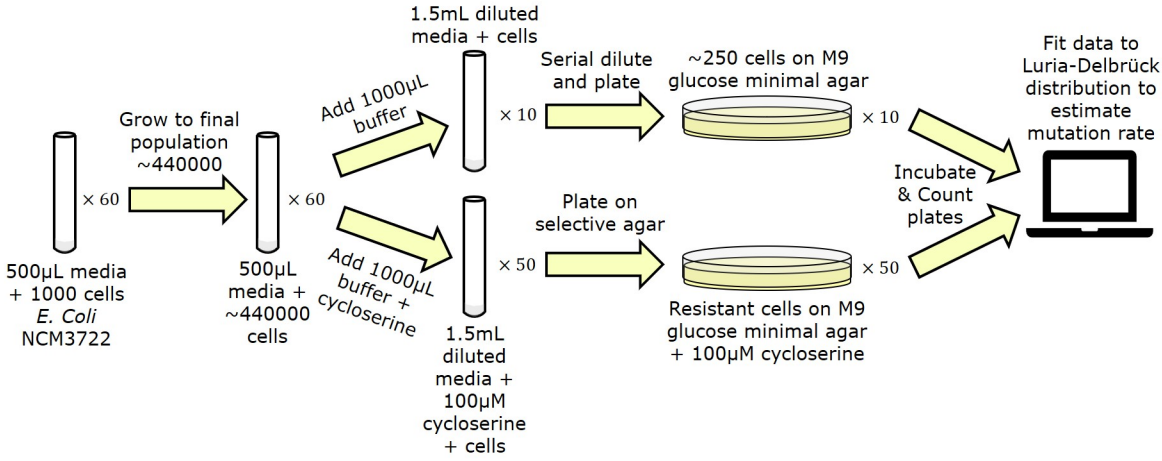

Figure 2.8: **Fluctuation test experimental procedure flowchart.** A flowchart outlining the key steps in the fluctuation test procedure developed and used for this thesis.

Chapter 3

Analysis

To gain meaning from fluctuation test data an analysis must be performed. In this chapter, I will introduce some of the most popular analysis methods used. Then I will introduce the concept and repercussions of a delayed phenotypic expression after mutation, and detail a few protocols that attempt to adjust for these effects, which will be tested through application to simulated and historical data. Finally, how one would go about comparing results from different fluctuation tests is discussed in preparation for determining if the growth rate affects the mutation rate.

3.1 Mutation Rate Estimation

Besides designing and performing the experiment to obey as many of the model assumptions as possible, the most important aspect of a fluctuation test is the fitting of the data to the model in order to determine the average number of mutations. There are several ways to do this and several tools have been developed to make the process easy and accurate for experimentalists. Historically scalar estimators were used, but in modern times the most commonly used tools are maximum likelihood estimators. In this section two scalar estimators, an implementation of a maximum likelihood estimator, and a more flexible total sum of squares fitting mechanism will be detailed. Additionally, how to convert the average number of mutations to the mutation rate and the associated error is discussed.

3.1.1 Scalar Estimators

When the fluctuation test was first described by Luria and Delbrück in 1943, Delbrück proposed a method for approximating the average number of mutations, m, called the p_0 method [106]. The p_0 method uses the fact that the zero point (or probability that no events occurred) is the same for both the Poisson distribution and the Luria-Delbrück distribution. Noting that the zero point of the Poisson distribution with mean m is,

$$p_0 = e^{-m} \,, \tag{3.1}$$

then an estimate on the average number of mutations per culture in a fluctuation test, $\hat{m}$, can be found with,

$$\hat{m} = -\ln p_0 \,, \tag{3.2}$$

where p_0 is the proportion of cultures that don't have any mutants [106]. The p_0 method works particularly well when estimating data with $0.1 \leq p_0 \leq 0.7$, which corresponds to $0.3 \leq m \leq 2.3$ [58]. To determine the error on an estimate from the p_0 method, the probability of having zero mutants in a culture, p_0, is considered binomial [99], giving a variance of,

$$\sigma^2_{p_0} = \frac{p_0(1-p_0)}{\nu} \,, \tag{3.3}$$

where ν is the number of parallel cultures that were selected for mutants. Plugging Eq. (3.1) into Eq. (3.3) and noting that the variance in m will be e^{2m} times bigger than the variance in p_0 gives [99],

$$\sigma^2_{\hat{m}} = \frac{e^{\hat{m}} - 1}{\nu} \,. \tag{3.4}$$

To calculate the 95% confidence interval, the error is assumed to be normal distributed, giving the interval,

$$\left(\hat{m} - 1.96\sqrt{\frac{e^{\hat{m}} - 1}{\nu}}, \hat{m} + 1.96\sqrt{\frac{e^{\hat{m}} - 1}{\nu}} \right) . \tag{3.5}$$

The p_0 estimate will be used as an initial guess during the maximum likelihood estimate as well as a simple estimator when analysing data in the next chapter, primarily for historical reasons.

Another popular scalar estimate is a median estimator developed by Jones et al. [83]. The estimate is of the form,

$$\hat{m} = \frac{r_m - \ln 2}{\ln r_m - \ln(\ln 2)} \,, \tag{3.6}$$

where r_m is the number of mutants in the median culture. The estimate is derived by deducing a dilution which is likely to halve the number of plates with mutants. The Jones median estimator works well when estimating data in the range $3 \leq r_m \leq 40$, which corresponds to $1.5 \leq m \leq 10$ [58]. Errors for this method are not explored because the method is only used as a potential initial guess in the maximum likelihood estimator to follow [203].

3.1.2 Maximum Likelihood Estimation

A commonly used package for analysing fluctuation tests is called rSalvador and was created by Qi Zheng in the programming language R [210]. The package is quite comprehensive and incorporates much of the work Zheng has done on the study of the Luria-Delbrück distribution in the last two decades, meaning rSalvador is capable of many different tasks; regardless, for my purposes only the base analysis tools were used. The most useful of these tools fits fluctuation test data to the Lea-Coulson model, and determines a confidence interval (CI) for the estimated average number of mutations, $\hat{m}$.

The following derivations follow Qi Zheng's work in [203]. The fitting tool determines an optimal $\hat{m}$ with a maximum likelihood estimation (MLE) method that uses the log-likelihood function,

$$l(X, m) = \sum_{i=1}^{\nu} \log p(X_i, m) \,, \tag{3.7}$$

where $X = (X_1, X_2, ..., X_\nu)$ is the experimental data from a fluctuation test ran with ν samples, so each X_i is the number of resistant mutants found in one experimental sample and is necessarily an integer. The goal of the fitting method is to find the value of m which makes the derivative with respect to m of the log-likelihood function, also known as the *score*,

$$U(X, m) = \frac{\partial l(X, m)}{\partial m} = \sum_{i=1}^{\nu} \frac{\frac{\partial p(X_i, m)}{\partial m}}{p(X_i, m)} \,, \tag{3.8}$$

go to zero. In order to do this, the probabilities and their derivatives with respect to m must be found. First, the probabilities of the Luria-Delbrück distribution are determined using the recursive method described by Sarkar et al. [148, 203]. The relation is of the

form[1],

$$
\begin{aligned}
p_0 &= e^{-m}\,, \\
p_r &= \frac{m}{r}\sum_{j=1}^{r}\phi^{j-1}(1-\frac{j\phi}{j+1})p_{r-j} \qquad (r\geq 1)\,,
\end{aligned}
\tag{3.9}
$$

where p_r is the probability that a culture has r mutants, m is the average number of mutations per culture, and $\phi = 1-\frac{N_0}{N_t}$ is a known constant that scales for total growth (N_0 and N_t are the initial and final populations respectively). Now to determine the derivative of the probabilities with respect to m, the i^{th} derivative of the probability generating function (Eq. (1.47)) with respect to m is taken,

$$
\frac{\partial^i G}{\partial m^i} = \left[\frac{1}{\phi}\left(\frac{1-z}{z}\right)\ln(1-\phi z)\right]^i \exp\left[\frac{m}{\phi}\left(\frac{1-z}{z}\right)\ln(1-\phi z)\right]\,,
\tag{3.10}
$$

where G is the probability generation function (PGF) and z in the usual auxiliary variable. Equation (3.10) can then be written in terms of a power series by using the definition of the PGF (Eq. (1.16)), taking the Taylor expansion around z=0 of the coefficient in front of the exponential, and noting that the exponential term is the Lea-Coulson PGF (Eq. (1.47)),

$$
\frac{\partial^i p_0}{\partial m^i} + \sum_{r=1}^{\infty}\frac{\partial^i p_r}{\partial m^i}z^r = \left[-1+\sum_{r=1}^{\infty}\frac{\phi^{r-1}}{r}\left(1-\frac{r\phi}{r+1}\right)z^r\right]^i\left(p_0+\sum_{r=1}^{\infty}p_r z^r\right)\,.
\tag{3.11}
$$

Equating the coefficients of z^r in Eq. (3.11) then gives,

$$
\begin{aligned}
p_r^{(1)} &= -p_r + \sum_{k=1}^{r}\frac{\phi^{k-1}}{k}\left(1-\frac{k\phi}{k+1}\right)p_{r-k}\,, \\
p_r^{(2)} &= -p_r^{(1)} + \sum_{k=1}^{r}\frac{\phi^{k-1}}{k}\left(1-\frac{k\phi}{k+1}\right)p_{r-k}^{(1)}\,,
\end{aligned}
\tag{3.12}
$$

where $p_r^{(i)} = \frac{\partial^i p_r}{\partial m^i}$.

Newton's method will now be used to find the root of the score function on the m axis, meaning the derivative with respect to m of the score function must also be determined.

[1] See Section 1.5.2 for the derivation of this relation.

The negative derivative of the score function, which is often referred to as the Fisher information, has the form,

$$J(X,m) = -\frac{\partial^2 l(X,m)}{\partial m^2} = \sum_{i=1}^{\nu} \left[\left(\frac{p^{(1)}(X_i,m)}{p(X_i,m)} \right)^2 - \frac{p^{(2)}(X_i,m)}{p(X_i,m)} \right] . \tag{3.13}$$

The Newton's method algorithm can then be written as,

$$\tilde{m}_{j+1} = \tilde{m}_j + \frac{U(X,\tilde{m}_j)}{J(X,\tilde{m}_j)} , \tag{3.14}$$

where $\tilde{m}$ is an estimate on m, the score function $U(X,m)$ is as defined in Eq. (3.8), and the Fisher information $J(X,m)$ is as defined in Eq. (3.13). For the initial guess $\tilde{m}_0$ the Jones median estimate [83] or the p_0 estimate [106] is used. The root that Eq. (3.14) finds is the optimal estimate on the average number of mutations per culture, $\hat{m}$.

The tool that calculates the confidence interval on the estimate $\hat{m}$ works by first recognising that $2(l(X,\hat{m}) - l(X,m_0))$ asymptotically has a chi-squared distribution with one degree of freedom, where $l(X,m)$ is again the log-likelihood function (Eq. (3.7)). Assuming a large sample size, this leads to,

$$l(X,m) = l(X,\hat{m}) - \frac{1}{2}\chi^2_{\alpha,1} , \tag{3.15}$$

where $\chi^2_{\alpha,1}$ is the $(1-\alpha)^{\text{th}}$ quantile of the chi-squared distribution with one degree of freedom. The log-likelihood, $l(X,m)$, is then assumed to have its only maximum at $\hat{m}$, meaning two points, $\hat{m}^-$ and $\hat{m}^+$, will satisfy Equation (3.15) when $0 < \alpha < 1$. Newton's method can be used to determine $\hat{m}^-$ and $\hat{m}^+$ using the algorithm,

$$\tilde{m}^{\pm}_{j+1} = \tilde{m}^{\pm}_j - \frac{l(X,\tilde{m}^{\pm}_j) - l(X,\hat{m}) + \frac{1}{2}\chi^2_{\alpha,1}}{U(X,\tilde{m}^{\pm}_j)} . \tag{3.16}$$

For the initial guess, $\tilde{m}_0$, the log-likelihood function is assumed the be quadratic in m so that it has two easily found roots. These two roots, $\tilde{m}^-_0$ and $\tilde{m}^+_0$, are then used as the initial guesses for $\hat{m}^-$ and $\hat{m}^+$ respectively,

$$\tilde{m}^{\pm}_0 = \hat{m} \pm \frac{1}{2}\sqrt{\frac{\chi^2_{\alpha,1}}{J(X,\hat{m}_j)}} . \tag{3.17}$$

One can determine the confidence interval, $(\hat{m}^-, \hat{m}^+)$, by choosing the desired value for α. For a 95% confidence interval ($\text{CI}_{95\%}$), $\alpha = 0.05$, and for a 84% confidence interval ($\text{CI}_{84\%}$), $\alpha = 0.16$, which are the two intervals that will be looked at in this thesis.

3.1.3 Total Sum of Squares Fitting

A fitting method that does not rely on the assumption that the data is solely composed of integers and is also capable of fitting only specific portions of the data, in order to accommodate for phenotypic lag, which will be discussed later in this chapter. Consequently, I developed a simple least squares fitting method that relies on the total sum of squares (TSS) of the difference between the data and the theoretical distribution (see Appendix D.2 for the R code). It works by first turning the data into a cumulative distribution function (CDF) and then looping through a sequence of theoretical average number of mutations, $\tilde{m}$, at steps of 0.001. For each $\tilde{m}$, a theoretical CDF is produced using rSalvador and the square difference between every point in the experimental CDF and its theoretical counterpart is found and summed, giving a fitting measure (see Fig. 3.1). The $\tilde{m}$ which gives the smallest sum of squared differences is considered the optimal fit, $\hat{m}$. This fitting method exhibits good agreement with the MLE fitting, especially for high sample numbers, n (Fig 3.2).

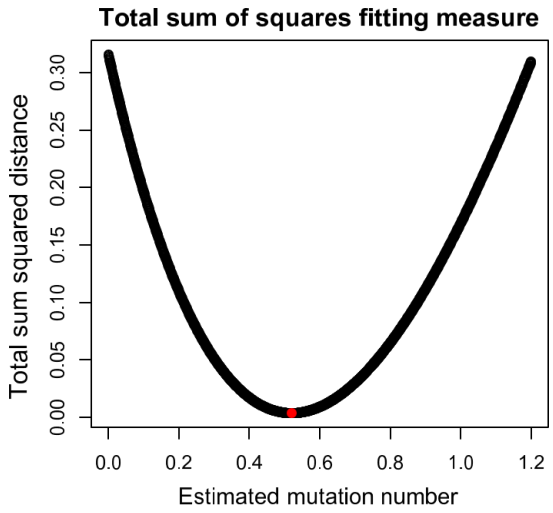

Figure 3.1: **The total sum of squares fitting measure for a sequence of estimated mutation numbers.** The total sum of squares (TSS) fitting measure is the sum of the square distances between all the points in an experimental CDF and the corresponding points in a theoretical CDF built with rSalvador. The TSS fit is found for a set of 100 simulated samples with an average number of mutations, $m = 0.5$. A sequence of guesses ($0 \leq \tilde{m} \leq 1.2$) for m is made, and for each $\tilde{m}$ the fitting measure is calculated. These fitting measures are plotted versus their corresponding guess, $\tilde{m}$. Note how there is a clear minimum (represented by a solid red point), which is chosen as the optimal estimate, $\hat{m}$.

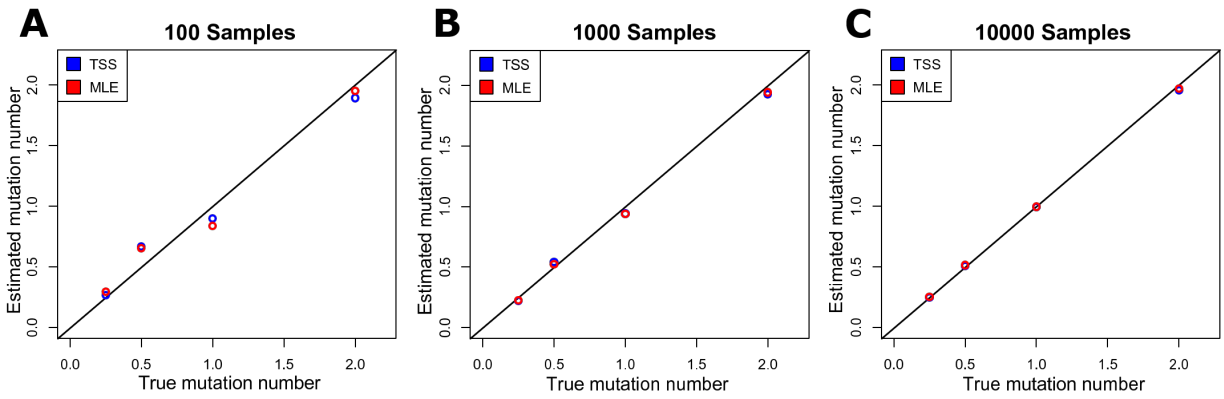

Figure 3.2: **Comparison of the rSalvador maximum likelihood (MLE) and the total sum of squares (TSS) estimates.** Data with three different sample sizes (ν = 100, 1000, and 10000) was simulated using rSalvador. For each sample size, four different average mutation numbers (m = 0.25, 0.5, 1, and 2) are simulated. For each data set the rSalvador maximum likelihood estimator and my total sum of squares estimator are used to estimate the average number of mutations. A) A comparison of the MLE and TSS estimates for data simulated with 100 samples. B) The same as (A), but with 1000 simulated samples. C) The same as A and B, but with 10000 simulated samples. The black line, $y = x$, is provided in order to easily gauge how far from the true value the estimates are.

The developed TSS fitting method does not lead to a straight-forward way for determining the confidence intervals such as the one used by rSalvador. Therefore, a bootstrapping method is used to determine the confidence intervals on the TSS fit [133]. The bootstrapping method used is a generic nonparametric bootstrap implemented with the "boot" package in R [28]. While using bootstrapping to determine the confidence intervals on the experimental data, 10000 replicates were used to find a bias-corrected and accelerated bootstrap interval [47] with the "boot" package.

3.1.4 Determining the Mutation Rate

All three types of analysis methods detailed in this section find an estimate for the average number of mutations per culture. The average number of mutations is not a particularly meaningful parameter though, because the more doublings in the culture, the more chances for mutation, the higher the average number of mutations. In other words, the number of mutations is dependent on the amount of growth. As a result, the average number of

mutations per culture must be converted to a parameter that is independent of experimental variables if one wishes to easily discuss and compare how prone bacteria are to mutation. The most common parameter of choice is the mutation rate, which gives the average number of mutations per cell per generation. To get the mutation rate, μ, the average number of mutations, m, is divided by the total change in population during the experiment, $(N_f - N_0)$, which is equivalent to the average number of doublings performed in each culture because if a culture starts with N_0 cells, ends with N_f cells, and bacteria grow by doubling, then the only way to get that growth is by having $N_f - N_0$ cells double or grow a generation. The result,

$$\mu = \frac{m}{(N_f - N_0)}\,, \tag{3.18}$$

can be interpreted as a rate of mutation or a probability of mutation. The mutation rate can be further generalised by accounting for the size of the gene that results in resistance when mutated. This is done by dividing the per cell mutation rate in Eq. (3.18) by the length of the target gene in base pairs (bp), giving a mutation rate with units of mutations per base pair per generation, μ_{bp}. Mutation rates are often given in this form because it is the most general. The ability to give a per base pair mutation rate is dependent on knowing which gene is being mutated, which is difficult if there are multiple pathways to resistance. Another form for mutation rate that is sometimes discussed is the per genome mutation rate, μ_{genome}. To calculate the per genome mutation rate, the per base pair mutation rate is multiplied by the length of the organism's genome.

Of course, when the conversion from the average number of mutations to the mutation rate is done for estimates from experimental data, there are errors involved. The error on the estimate of the average number of mutations, $\hat{m}$, comes in the form of a confidence interval either found from the log-likelihood for the MLE estimate or bootstrapping for the TSS estimate. When the estimate of the mutation rate, $\hat{\mu}$, is calculated, the error on $\hat{m}$ will propagate through and be combined with the error in the initial and final populations (σ_{N_0} and σ_{N_f} respectively) which are experimentally determined standard deviations. To do this, the confidence interval for $\hat{m}$ must first be converted into a standard deviation. Because the confidence intervals are not symmetric, the portion on each side of the mean must be considered separately and the maximum and minimum of the interval will be denoted C^+ and C^- respectively. To convert the confidence interval into a standard deviation, the error $\hat{m}$ is assumed to be normally distributed around the mean [135], giving,

$$C^{\pm}_{\hat{m}} = \hat{m} \pm Z\sigma^{\pm}_{\hat{m}}\,, \tag{3.19}$$

where Z is a statistical number related to the normal distribution that equals 1.96 for 95% confidence intervals and 1.4 for 84% confidence intervals (i.e. $Z_{95\%} = 1.96$ and $Z_{84\%} = 1.40$)

[135], which are the two confidence intervals that will be discussed in this thesis (see Section 3.4). Rearranging Eq. (3.19) for the upper and lower standard deviations (SD) gives,

$$\sigma_{\hat{m}}^{\pm} = \frac{1}{Z}(\pm C_{\hat{m}}^{\pm} \mp \hat{m}) \,. \tag{3.20}$$

Now using error propagation rules [195] and assuming there is no covariance between $\hat{m}$ and $(N_f - N_0)$, the right and left standard deviations of $\hat{\mu}$ are found to be,

$$\sigma_{\hat{\mu}}^{\pm} = \hat{\mu}\sqrt{\left(\frac{\sigma_{\hat{m}}^{\pm}}{\hat{m}}\right)^2 + \left(\frac{\sqrt{\sigma_{N_0}^2 + \sigma_{N_f}^2}}{N_f - N_0}\right)^2} \,. \tag{3.21}$$

The standard deviations of $\hat{\mu}$ can then be converted to confidence intervals by again assuming that the error in $\hat{\mu}$ is normally distributed, giving,

$$C_{\hat{\mu}}^{\pm} = \hat{\mu} \pm Z\sigma_{\hat{\mu}}^{\pm} \,, \tag{3.22}$$

which can be used to compare mutation rates from different experiments.

All discussed estimators for the average number of mutations, and by extension the mutation rate, are designed under the assumptions of the Lea and Coulson model found in Section 1.5.2 [58]. The experiments performed for this thesis were designed to obey as many of these assumptions as possible, but there are physical limitations that make assumption 9, that all mutants are detected at the time of selection, nearly impossible to obey when selecting for the absence or presence of an active protein [29]. In other words, once a mutation happens, the cell that first inherits the mutant chromosome and all of its subsequent children are considered mutants, but they may not be selected for because their mutant phenotype takes time to manifest due to the time it takes to dilute or accumulate proteins [29]. This period of time between when the mutation occurs and when the mutant phenotype is expressed is called phenotypic lag, and it results in fewer mutants appearing in fluctuation tests causing underestimates on the average number of mutations [10, 4, 170, 29, 18]. Methods for how to adjust for the bias caused by phenotypic lag are proposed in this chapter, but to understand them phenotypic lag and its consequences must first be discussed in detail.

3.2 Phenotypic Lag

Phenotypic lag is how long it takes for a mutant phenotype to be expressed after a mutation occurs [124, 4]. One of the most common and consequential forms of phenotypic lag is a repercussion of proteins needing to be accumulated or diluted out in order for the mutant cells to express the mutant phenotype [29]. To better understand the details and consequences of phenotypic lag, I will focus on protein dilution leading to resistance, but many of the arguments to follow can also be used to describe protein accumulation with minimal adjustments[2]. Imagine that the gene that codes for a protein, which we will refer to as the α-protein, gains a mutation during replication and one of the children inherit it. The mutation will likely affect the expression of the α-protein in some way. The most common effect will be that the cell stops making functional α-proteins, which for the sake of this argument is the same as producing no α-proteins; we will employ this simplification moving forward. At birth, the new mutant cell will inherit a portion of its non-mutated parent's α-proteins. What portion of the α-proteins that each progeny will inherit is dependent on how the cell manages this protein, but each child will likely inherit on average half of the α-proteins, subject to some partition error [80, 29]. Assuming there are P_0 α-proteins independently distributed around the parent cell at the time of division, the number of α-proteins inherited by one child will be a binomial random number of probability 0.5 and P_0 trials, while the other child will inherit the remainder [80, 29]. The new mutant cell then starts growing, doubling all of its components except the α-proteins. When this cell eventually splits, each of its children will inherit approximately half of all its proteins, giving two cells with on average one half the normal number of α-proteins at birth. The doubling process continues in this fashion, giving the pattern seen in Fig. 3.3. As growth continues, the number of α-proteins in the mutant children decays exponentially[3]. The question now is: at what point do the mutant cells start expressing the mutant phenotype? The answer is dependent on the details of the protein in question and how the mutant phenotype is being selected for. Regardless, there will be a length of time that passes called phenotypic lag. If P_0 is the average number of α-proteins in a normal cell at the time of birth, and P_r is the maximum number of α-proteins that a cell can have while expressing the mutant phenotype, then the phenotypic lag length in generations, n, is determined by $n = \log_2(\frac{P_0}{P_r})$. In the experimental system used in this thesis, the α-protein is a permease protein coded by the gene *cycA* and the mutant phenotype is expressed as resistance to the antibiotic cycloserine [51]. This means the mutant cells become resistant

[2]For protein accumulation one simply assumes that mutant cells produce a certain number of proteins each generation, and once they have enough, they become resistant [29].

[3]Due to the degradation of proteins, the cells will likely reach zero sooner than predicted by this model.

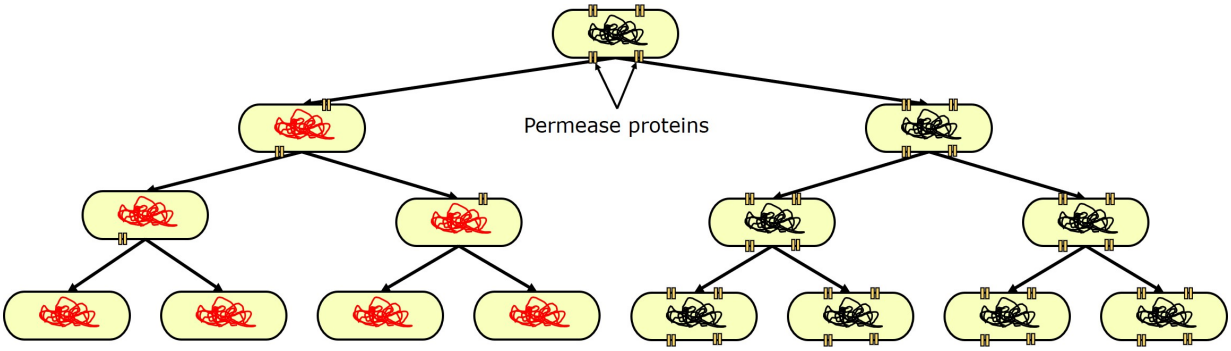

Figure 3.3: **Permease dilution causing phenotypic lag.** A Haldane tree in which a normal black genome becomes red when mutated, resulting in the cell no longer producing permease proteins. When a mutant doubles, each child inherits half of its parent's permease proteins. This assumes no partition error and no partial proteins.

when so many of these permease proteins have been diluted out that the cell is no longer able to uptake a fatal amount of cycloserine.

In a deterministic system, if there are n generations of phenotypic lag, then no mutants will appear until n generations after a mutation occurs. In this time the mutant continues to grow, meaning when the mutant phenotype does get expressed, there are now 2^n resistant cells, as apparent in Fig. 3.4. Consequently, in a fluctuation test, the samples with a small number of mutants ($r < 2^n$) will not appear in the CDF, except to artificially increase the number of apparent samples with zero mutants (see Fig. 3.5). It also means that in cultures that have had multiple mutation events, the mutant lineages that are less than n generations old will not appear, resulting in some of the samples with greater than 2^n mutants appearing with artificially low mutant counts that are still greater than or equal to 2^n. The consequence of these two affects is a discontinuous CDF with an artificially high y-intercept and a slightly flatter shape. In reality, partition error will mean that there is a chance cultures with greater than zero, but less than 2^n, mutants will appear. The quantity of these cultures compared to the no lag case will be significantly reduced though, meaning the y-intercept will still be artificially high and the CDF will still appear more flat. See Fig. 3.5 for a comparison of simulated fluctuation test data with phenotypic lag both accounting and not accounting for partition error.

Another common form of phenotypic lag is a repercussion of having multiple copies of the chromosome (generally due to fast growth) which is called effective polyploidy [171]. Phenotypic lag as a consequence of effective polyploidy can have different effects depending on the type of mutation. The two types of mutation are recessive and dominant, where

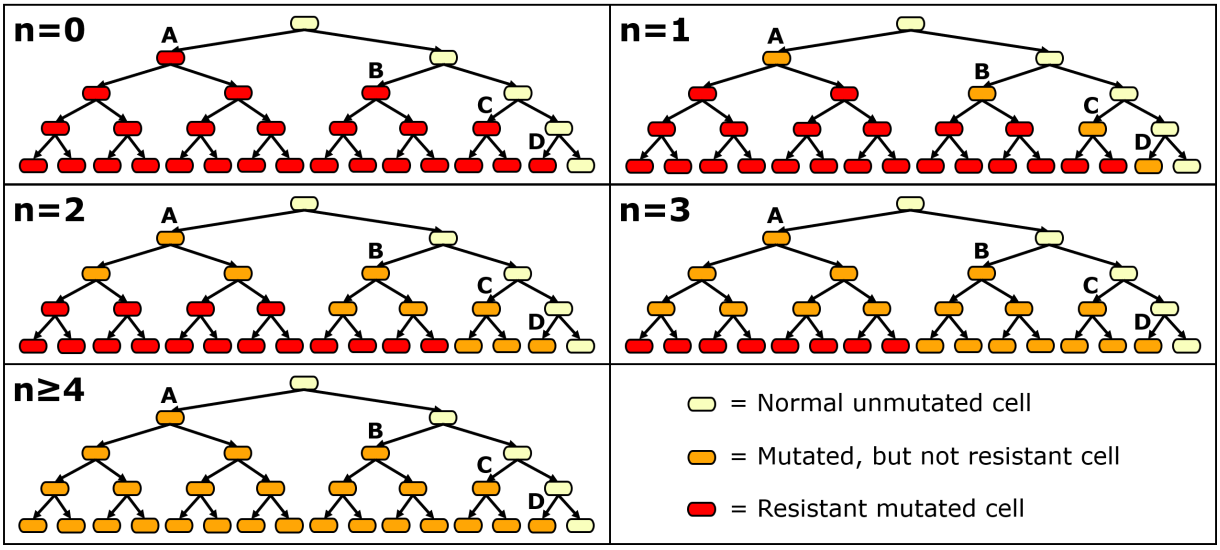

Figure 3.4: **Haldane trees for different phenotypic lag lengths.** The length of phenotypic lag in generations is represented by n. In each case, the culture starts from 1 normal cell and is grown for 4 generations, with one new mutant appearing in each generation. After 4 generations, a selector is introduced to the culture so that only the cells that are expressing the mutant phenotype survive. For $n = 0$, all mutants appear, giving 15 resistant cells. For $n = 1$, cells must grow for at least 1 generation to become resistant, meaning the D lineage will not become resistant, resulting in 14 resistant cells. For $n = 2$, the C and D lineages do not grow for long enough, giving 12 resistant cells. For $n = 3$, only the A lineage grows for enough time, giving 8 resistant cells. For $n \geq 4$, no lineage has sufficient time to grow and hence there are no resistant cells. This is all assuming no partition error.

a recessive mutation means that the mutant phenotype is fully expressed after all the chromosomes in a cell have the mutation, while a dominant mutation requires the cell to only have one mutant chromosome to express the phenotype [171]; if the selecting agent chemically combines with its target, as is often the case with antibiotics, then the mutation which grants resistance is recessive [29]. In the context of a fluctuation test, polyploidy with a dominant mutation will cause a different type of phenotypic lag than described above, in that after a mutation occurs a mutant will immediately appear, but only one of its children will be a mutant for several generations until the mutant chromosome becomes homozygous (i.e. there will be only one mutant for several generations before the mutants start doubling as described in the Luria-Delbrück model) [171]. In combination with the higher number of

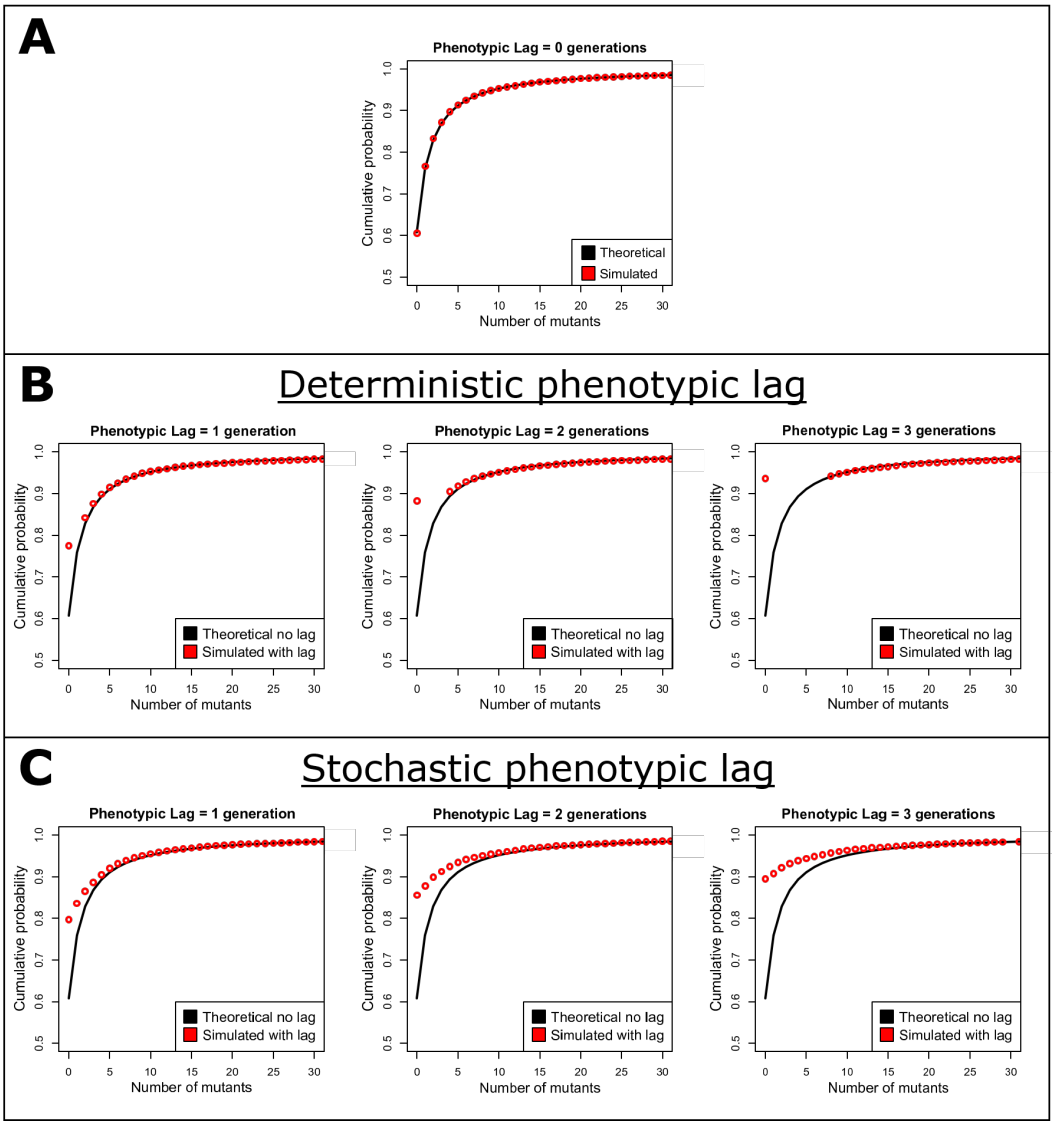

Figure 3.5: **Simulated fluctuation test data with phenotypic lag.** The cumulative distributions of 10000 simulated cultures with initial population $N_0 = 1 \cdot 10^3$, final population $N_f = 5 \cdot 10^5$, mutation rate $\mu = 1 \cdot 10^{-6}$, and phenotypic lag lengths $n = 0$, 1, 2, 3. A) No phenotypic lag (n=0). B) Deterministic phenotypic lag: no partition error resulting in a discontinuous CDF. C) Stochastic phenotypic lag: protein dilution with binomial partition error and an initial number of active proteins, $P_0 = 100$.

chromosomes giving more opportunities for mutational events, the result is a larger number of measured mutations and therefore a higher measured mutation rate. On the other hand, polyploidy with a recessive mutation will have no affect on fluctuation test data because the effects of there being more chromosomes to mutate cancel out the effects of the phenotypic lag [171, 29]. But if polyploidy with a recessive mutation is combined with protein dilution, it can have an effect on fluctuation test data that is different than in the presence of just protein dilution [29]. When a cell has multiple copies of a gene, it will use all of them to build proteins [29]. Consequently, if some of the genes have a mutation which causes the α-protein to stop being produced, but the cell keeps sending resources to them in an attempt to make the protein, then the cell will start producing less α-proteins. The result would be a gradual decrease in α-protein numbers in the parent cells even before the mutant chromosome monopolises the cells, giving a shorter time needed to become resistant once the cell contains only mutated chromosomes. Because polyploidy is a growth dependent physiological parameter as described in Section 1.6.1, the phenotypic lag length could become dependent on growth rate when protein dilution is required. Furthermore, due to the constraints of the proteome described in Section 1.6.2, there is reason to believe that most α-protein concentrations would be dependent on growth rate, adding another avenue for which phenotypic lag length can couple to growth rate. Finally, because the cell size, and by extension surface area, is growth rate dependent, if chemical diffusion is dependent on the surface area and density of permease proteins (as one may expect), then there is further potential for phenotypic lag length to be coupled to growth rate, especially when the α-protein is a permease. There being three separate routes to phenotypic lag length coupling to the cells' physiology, all of which likely present and potentially combining in non-linear ways, makes inferring the exact effect very difficult. Regardless, it is beneficial to be aware of this potential coupling when designing experiments and simulations.

3.2.1 Simulating Phenotypic Lag

In order to study the effects of phenotypic lag on fluctuation test data, and by extension, the estimated average number of mutations, a method for simulating data with phenotypic lag was developed. The simulation uses the code from Sun et. al [171, 172] as a base, in which they use the ideas of Zheng [202], and Hamon and Ycart [68]. For the simulation, the initial population N_0, final population N_f, and mutation rate μ are input and the number of mutations that will appear during growth is drawn from a Poisson distribution with probability $\mu(N_f - N_0)$ (cf. Eq. (1.6)). The times at which each of these mutants appear are then drawn from an exponential distribution with rate $\lambda = \ln(2)$, so that the time is in generations, as in the derivations by Delbrück, and Lea and Coulson (see Sections

1.5.1 and 1.5.2).. The number that is drawn from the exponential distribution gives how long prior to the end of the experiment the mutant appears (i.e. $t_{\text{birth}} = t_{\text{final}} - t_{\text{exp}}$). The mutant is then grown by drawing replication times from the same exponential distribution for each mutant and doubling them at these drawn times until the end of the experiment is reached. The method allows for the tracking of all the mutant lineages (which start from a single mutant). At the end of the "experiment" when all the mutant lineages have been realised, the number of mutants which are alive (i.e. have not yet doubled) are counted to get a total number of mutants in the culture. The mutant number from this "experiment" constitutes one point in a set of fluctuation test data, so to create meaningful simulated data, this protocol must be repeated many times.

To include the effects of phenotypic lag in the simulation required adding my own elements to the code and inputting an average initial protein number, P_0, and average phenotypic lag length in generations, n. For deterministic phenotypic lag, at the time of adding up all the mutants in a culture, all the lineages that are less than n generations old, meaning they are composed of less than 2^n mutant cells, are discarded. For stochastic phenotypic lag, whenever a new mutant lineage appears, the first mutant is born with a number of proteins, $\tilde{P}$, which is equal to a binomial random number with probability 0.5 and $2P_0$ trials. When this cell doubles, one child inherits a number of proteins, $\tilde{\tilde{P}}$, equal to a binomial random number with probability 0.5 and $\tilde{P}$ trials, while the other child inherits the remaining $\tilde{P} - \tilde{\tilde{P}}$ proteins. This process repeats until the cells no longer have time to double and the original $\tilde{P}$ proteins are distributed between all live mutants in the lineage. To include phenotypic lag, at the end of the "experiment" if a mutant has less than $\frac{P_0}{2^n}$ proteins then it is counted. The effects of different starting protein numbers are shown in Fig. 3.6. Based off measured protein numbers of the CycA permease protein [189, 102] (which will be the active protein during selection in the experiments throughout this thesis), and motivated by the fact that the simulated distribution does not appear to have a strong dependence on the number of proteins, for future simulations the average starting protein number will be chosen as $P_0 = 100$.

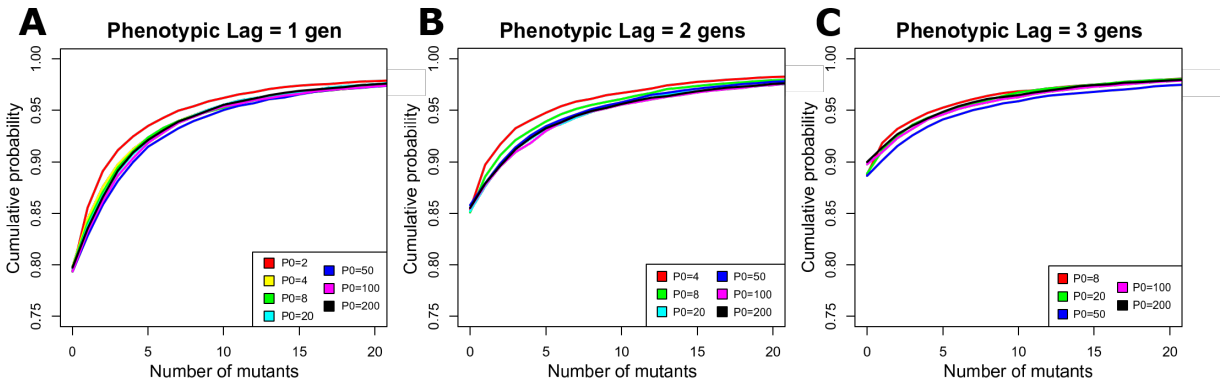

Figure 3.6: **Simulated fluctuation test data with average phenotypic lag of 1, 2, and 3 generations and varying starting protein amounts.** Cumulative distributions for 10000 simulated cultures with initial population $N_0 = 1000$, final population $N_f = 500000$, and mutation rate $\mu = 1 \cdot 10^{-6}$. A) Simulated phenotypic lag length of 1 generation and starting protein numbers $P_0 = 2$, 4, 8, 20, 100, and 200. B) Simulated phenotypic lag length of 2 generations and starting protein numbers $P_0 = 4$, 8, 20, 100, and 200. C) Simulated phenotypic lag length of 3 generations and starting protein numbers $P_0 = 8$, 20, 100, and 200. The protein number P_0 is never allowed to be less than 2^n where n is the average length of phenotypic lag in generations. Note how the shape of the distribution becomes more dependent on P_0 with increasing phenotypic lag, but in all cases the differences are not large for $P_0 > 2^n$.

3.3 Adjusting Fit for Phenotypic Lag

Adjustments to model fitting are a necessity when aspects of an experiment fail to satisfy all the assumptions of the model. Many such adjustments exist for the Luria-Delbrück system, but few have investigated and implemented a system for adjusting for phenotypic lag. In the case of phenotypic lag, assumption 9 of the Lea-Coulson model (Section 1.5.2), which is that all mutants are detected at the time of selection, fails. Instead, the young mutant lineages go undetected, as explained in Section 3.2. The result is that in the presence of phenotypic lag, the usual fitting mechanisms described in Section 3.1 categorically give underestimates for the average number of mutations, m.

3.3.1 Koch Adjustment

In 1981, Arthur Koch laid out a basic protocol to adjust for phenotypic lag [92]. His protocol was to divide the number of observed mutants[4] by 2^n (where n is the number of generations of phenotypic lag), fit a Luria-Delbrück distribution to the adjusted data, and then multiply the fitted $\hat{m}'$ by 2^n to get a more accurate estimate, $\hat{m}$ [5]. By dividing the original distribution by 2^n, one is essentially looking at what the distribution looked like n generations earlier if there was no phenotypic lag. Fitting to this adjusted distribution then gives an estimate, $\hat{m}'$, on the average number of mutations n generations earlier in a hypothetical system free of phenotypic lag. Multiplying this fitted $\hat{m}'$ by 2^n then adjusts for the growth of all the mutants during the n generations of lag that were removed by the original division by 2^n. The concept of this adjustment is straight-forward and logical, but it is not without flaws. The main flaw is that the protocol on average gives over-estimates for $\hat{m}$ when applied to simulated data. I believe the primary reason for this is that the adjusted data still includes mutants that would not have been present n generations earlier, resulting in a higher estimate for $\hat{m}'$. In particular, some of the cultures with many mutants could be from several younger lineages of mutants that would not have existed n generations earlier, but the adjustment protocol instead treats them like a single, old, large lineage.

To implement Koch's adjustment protocol, I used the programming language R as well as tools from rSalvador [210]. The protocol allows for a prediction of both the length of phenotypic lag, $\hat{n}$, and an associated average number of mutations, $\hat{m}$. First, the fluctuation test data, which is a set of numbers specifying the number of resistant mutants in each sample, is turned into a cumulative distribution with all samples that have greater than 300 mutants binned[6] at 300. A sequence of guesses on the length of phenotypic lag, $\tilde{n}$, is then cycled through and for each $\tilde{n}$ the number of mutants, which is the x-axis of the CDF, is divided by $2^{\tilde{n}}$. The TSS fitting mechanism is then employed to get an associated estimate on the mutation number, $\tilde{m}$, and the associated error, which is the total sum squared difference. The TSS errors for each $\tilde{n}$ are then compared and the minimum is chosen to give a prediction for phenotypic lag length, $\hat{n}$. The associated optimal average number of mutations $\hat{m}'$ is finally multiplied by $2^{\hat{n}}$ to get an adjusted estimate, $\hat{m}$. See Fig. 3.7 for an infographic of the protocol being applied to a set of simulated data.

[4] Koch worked with quartiles due to computational limitations, but the concept remains the same when applied to the entire cumulative distribution.

[5] In Koch's original paper he multiplies by $2^n - 1$ at this stage, but this doesn't make sense because it fails in the $n = 0$ case.

[6] Foster claims that high mutant counts can be truncated at 150 with "little loss of precision" as long as there aren't too many outliers [58], while the program flan bins mutants at $r = 1024$ [112]. I took a sort of middle ground approach, leaning more towards the side of Foster to reduce computational costs.

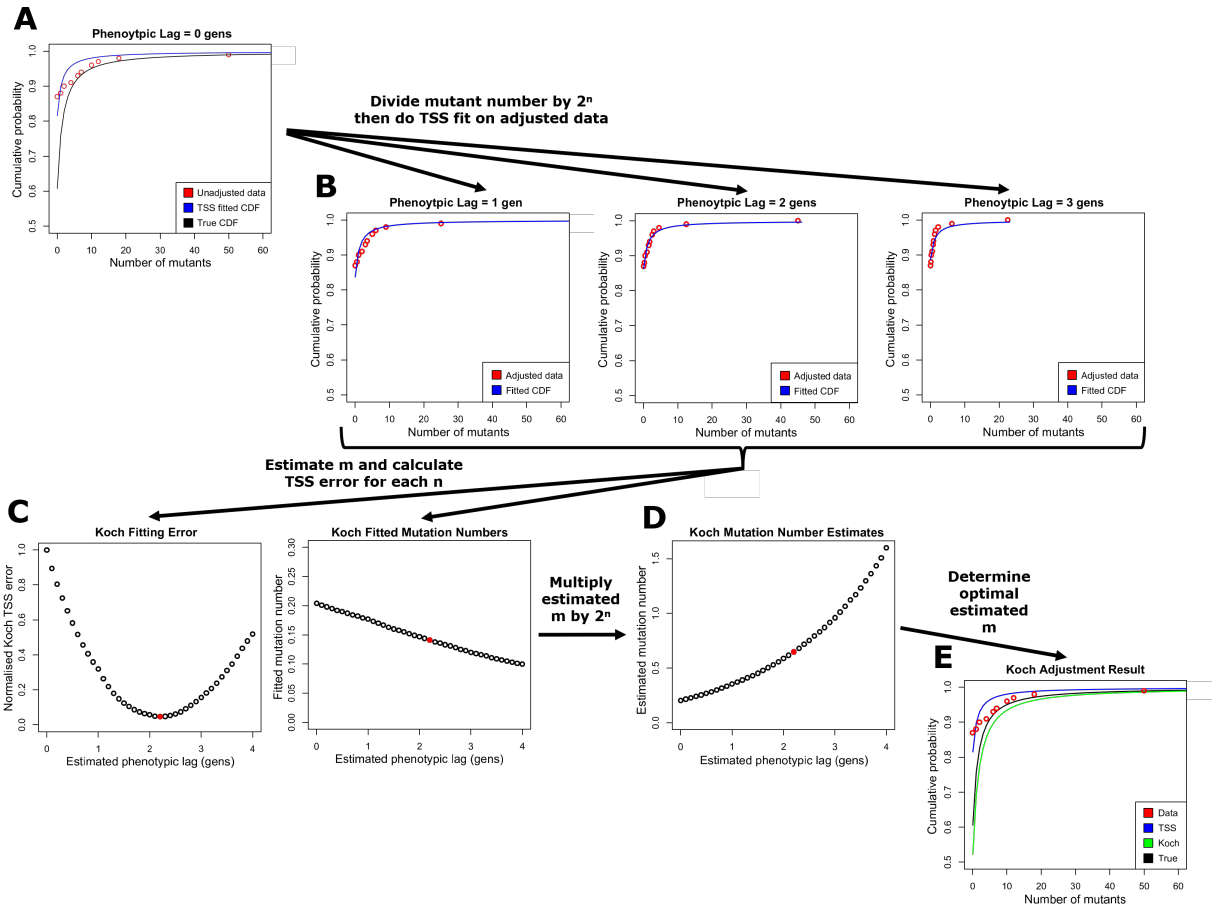

Figure 3.7: **Infographic for the Koch adjustment protocol.** The Koch adjustment system is applied to simulated fluctuation test data with 100 samples, initial population $N_0 = 1000$, final population $N_f = 500000$, mutation rate $\mu = 1 \cdot 10^{-6}$, average number of mutations $m = 0.499$, and phenotypic lag length $n = 2$ generations. A) The data's cumulative distribution with the CDF's for the TSS estimate $\hat{m}_{\text{TSS}}$ and the true m. B) The number of mutants in each tube divided by $2^{\tilde{n}}$ where $\tilde{n} = 1$, 2, and 3 generations, and the CDF for the TSS estimate $\tilde{m}_{\text{TSS}}$. C) The TSS fitting error and the estimated average number of mutations, $\tilde{m}$, for each $\tilde{n}$; the solid red points are the optimal estimates, $\hat{n}$ and $\hat{m}'$. D) The estimated $\tilde{m}$ from (C) converted to average number of mutations estimates by multiplying by $2^{\tilde{n}}$; the solid red point is the optimal estimate, $\hat{m}_K$. E) The unadjusted data with the CDF's corresponding to the Koch estimate $\hat{m}_K$, the TSS estimate $\hat{m}_{\text{TSS}}$, and the true m.

To study the efficacy of the Koch adjustment method, it was implemented on three sets of simulated fluctuation test data with stochastic phenotypic lag. One set is 100 simulations of 100 samples each, one is 10 simulations of 1000 samples each, and one is 1 simulation of 10000 samples. All simulations have the parameters: initial population $N_0 = 1 \cdot 10^3$ cells, final population $N_f = 5 \cdot 10^5$ cells, mutation rate $\mu = 1 \cdot 10^{-6}$ mutations per cell per generation, and initial protein number $P_0 = 100$ proteins, which corresponds to an average number of mutations per culture $m = 0.499$. These values are chosen such that the simulations have similar parameters to my experiments. Furthermore, each set of simulations is done for several phenotypic leg lengths, $n = 0$, 1, $\log_2(3)$, 2, $\log_2(5)$, $\log_2(6)$, $\log_2(7)$, and 3 generations[7]. Results from the Koch adjustment protocol being applied to each simulation set can be found in Fig. 3.8.

When the Koch protocol is applied to the simulations, as already mentioned, overestimates on the average number of mutations, m, are given on average, and this appears to be independent of sample size. Conversely, the estimates on the phenotypic lag length, n, are underestimates when long phenotypic lag ($n \geq 2$) is present, and slight overestimates for short lag ($n < 2$). In particular, the system is quite good at telling when there is no phenotypic lag present. Regardless of the estimates of m on average being overestimates, the average adjusted estimate on m is still much closer to the true value inputted into the simulations than the unadjusted MLE fits. Furthermore, the fact that the system can tell there is phenotypic lag present and can generally predict if there's more lag in one set of data than another is beneficial, even if the magnitude is not correct on average. One especially nice attribute of the Koch protocol is that the distribution of TSS errors always creates a smooth quadratic-like shape with a clear minimum, making the choice for $\hat{n}$ very straightforward (Panel (C) in Fig. 3.7). Arguably the biggest problem with the protocol is the magnitude of standard deviation in its estimates of m across the 100 simulations with 100 samples, though the edges of the error bars still lay closer to the true m than the MLE fit. In addition, the standard deviations in the predictions on n are also fairly substantial. Despite the apparent issues with the Koch adjustment protocol, as made clear through its use on simulations, it remains a clearly beneficial tool for making better predictions on the average number of mutations than the traditional maximum likelihood estimate when phenotypic lag is present.

[7] These numbers will be referred to as $\log_2$ integers.

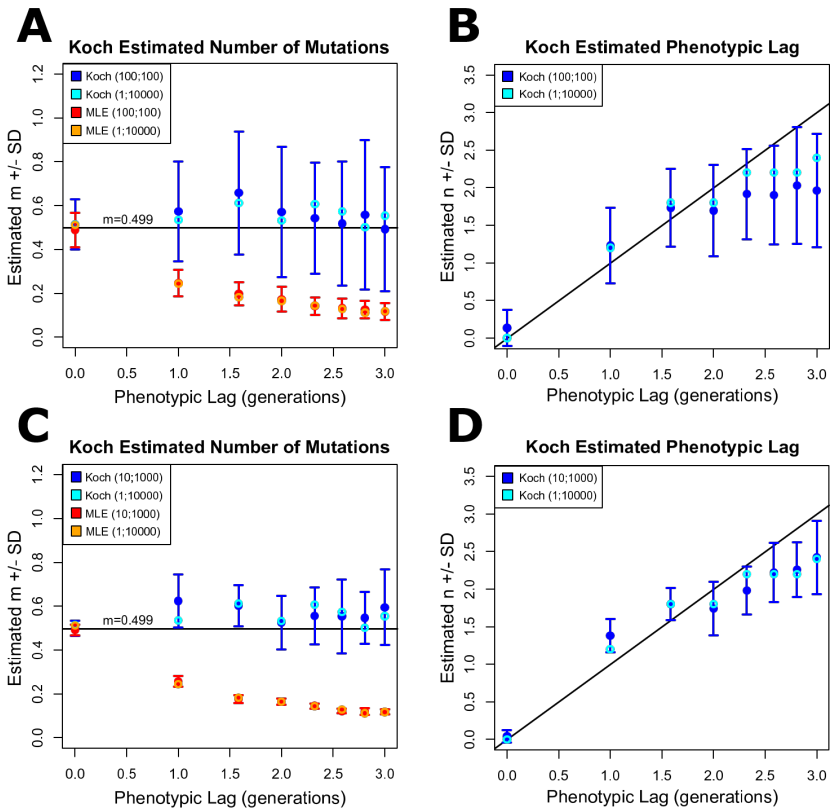

Figure 3.8: **Koch adjustment protocol applied to simulated fluctuation test data with phenotypic lag.** Fluctuation test data is simulated with an average number of mutations $m = 0.499$ and $\log_2$ integer phenotypic lag lengths (n). Three sets of data are simulated, one with 100 simulates of 100 samples (100;100), one with 10 simulates of 1000 samples (10;1000), and one with 1 simulate of 10000 samples (1;10000). A) The MLE fitting method and the Koch phenotypic lag adjustment protocol are applied to the 100;100 and 1;10000 data sets, and the average estimated number of mutations ($\langle \hat{m} \rangle$) $\pm$ one standard deviation (SD) is plotted for each simulated phenotypic lag. B) The Koch protocol is applied to the 100;100 and 1;10000 data sets, and the average estimated phenotypic lag length ($\langle \hat{n} \rangle$) $\pm$ SD is plotted for each simulated phenotypic lag. C) The MLE fitting method and the Koch protocol are applied to the 10;1000 and 1;10000 data sets, and $\langle \hat{m} \rangle \pm$ SD is plotted for each simulated n. D) The Koch protocol is applied to the 10;1000 and 1;10000 data sets, and $\langle \hat{n} \rangle \pm$ SD is plotted for each simulated n. In (B) and (D) the black line, $y = x$, is provided in order to easily gauge how far from the true value the estimates are.

3.3.2 Reduced CDF Adjustment

Another adjustment method, which has been alluded to by Foster and others [58, 4], has been extended and implemented by me. In this method, only the latter part of the experimental cumulative distribution (CDF) is considered while fitting. Accordingly, I will refer to the method as the "reduced CDF (rCDF)" adjustment protocol. With phenotypic lag present, the left portion of the CDF experiences the most noise from partition errors because the cultures with many mutants have a smaller proportion of mutants not being selected due to a significant number of these mutants coming from mutations early in growth. Consequently, the latter part of the experimental CDF matches the true CDF much closer, as is made clear by Fig. 3.5. As a result, if one only fits to the later portion of the CDF, one should theoretically get a more accurate estimate of the average number of mutations, m. The main downfall of this adjustment method is that by ignoring early points of the CDF, one then has to fit to a smaller number of points. If there are many samples, this is not a problem, but when there aren't, the error in the fit has the potential to be troublesome. Furthermore, the reduced CDF fitting procedure on average gives an underestimate for the average number of mutations per culture, m. An underestimate on m is given because, as seen in Fig. 3.5, the cumulative distribution for simulated data with phenotypic lag is flatter than the Luria-Delbrück distribution, and a more flat curve leads to a smaller fitted mutation rate. Though removing the beginning of the distribution allows one to ignore the most dominant effects of phenotypic lag, it does not entirely rid the data of the effects. The underestimate of m is a consequence of missing young mutant lineages in cultures with multiple mutant lineages, as well as partition error.

To implement the reduced CDF adjustment protocol, the computer program R is once again used. Like the Koch protocol, the data is first turned into a CDF with all cultures with mutant numbers greater than or equal to 300 being binned at 300. Then a sequence of phenotypic lag guesses $\tilde{n}$, which are all equal to $\log_2$ integers, is cycled through. For each $\tilde{n}$, all points with mutant number less than $2^{\tilde{n}}$ are removed. The reduced CDF is then fitted using the TSS fitting protocol to find an associate estimate on the mutation number, $\tilde{m}$. Also for each $\tilde{n}$, the theoretical distribution is adjusted to mimic deterministic phenotypic lag by moving all points with $x < 2^{\tilde{n}}$ to zero, making the y-intercept artificially high. Next, the sum squared distance between the zero mutant point of the experimental data and the adjusted theoretical data is calculated because the zero point is the point which is most sensitive to phenotypic lag. Comparing the zero point of the theoretical CDF adjusted to mimic deterministic phenotypic lag and the experimental zero point can give insight into the length of phenotypic lag. The TSS fitting error for the reduced CDF and the zero point error are then added together to get a total fitting error. Finally, the $\tilde{n}$ which gives

the minimum total fitting error is used as the optimal estimate on the phenotypic lag, $\hat{n}$, and the associated $\tilde{m}$ is chosen as the optimal average number of mutations estimate, $\hat{m}$. The algorithm is laid out in the form of an infographic in Fig. 3.9.

The reduced CDF adjustement protocol was applied to the same simulation data as the Koch fitting protocol to study its efficacy. The results of this study can be found in Fig. 3.10.

It can be seen from the results of the application of the reduced CDF adjustment protocol to simulated data that the protocol on average gives an underestimate for the average number of mutations as predicted. The protocol also gives an underestimate on the length of phenotypic lag for all lags greater than or equal to two generations. On the other hand, the protocol does a good job of predicting phenotypic lags of length less than 2 generations. Furthermore, the standard deviation in the estimates of both mutation number and phenotypic lag length have a noticeably lower standard deviation than the Koch protocol. Finally, though not made clear through the simulation data[8], the reduced CDF adjustment protocol fails when there are no samples with zero mutants because it uses the difference between the experimental zero and theoretical zero to estimate the length of phenotypic lag.

[8]This was found by applying the rCDF method to several sets of historical data.

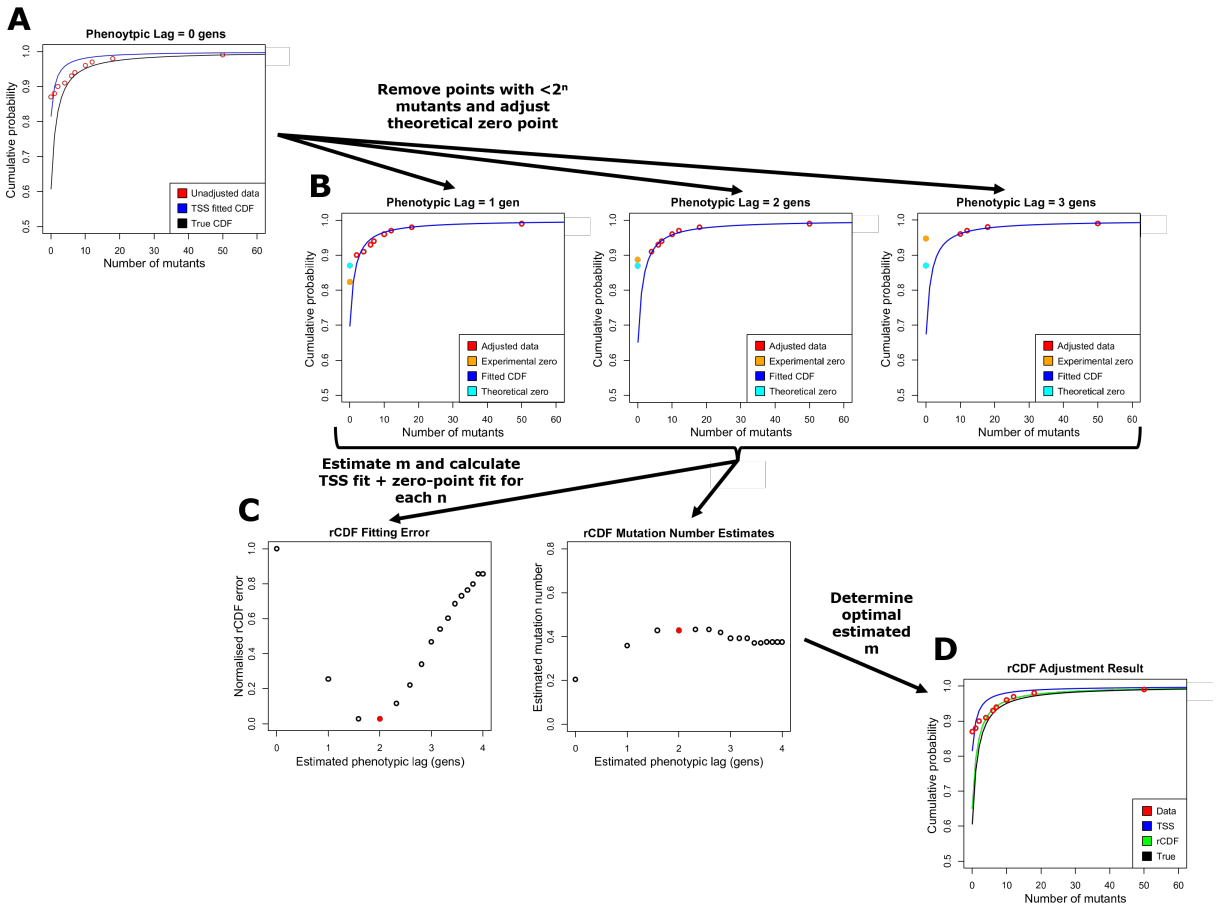

Figure 3.9: **Infographic for the reduced CDF adjustment protocol.** The reduced CDF (rCDF) adjustment system is applied to simulated fluctuation test data with 100 samples, initial population $N_0 = 1000$, final population $N_f = 500000$, mutation rate $\mu = 1 \cdot 10^{-6}$,, average number of mutations $m = 0.499$, and phenotypic lag length $n = 2$ generations. A) The data's cumulative distribution with the CDF's for the TSS estimate $\hat{m}_{\text{TSS}}$ and the true m. B) All data points with number of mutants less than $2^{\tilde{n}}$ where $\tilde{n} = 1$, 2, and 3 generations are removed, and the reduced CDF is fitted with TSS; the zero point of the fitted CDF is adjusted to include all points $< 2^{\tilde{n}}$ to mimic deterministic phenotypic lag. C) The TSS fitting error and the estimated average number of mutations, $\tilde{m}$, for each $\tilde{n}$; the solid red points are the optimal estimates, $\hat{n}$ and $\hat{m}$. D) The unadjusted data with the CDF's corresponding to the rCDF estimate $\hat{m}_{rCDF}$, the TSS estimate, $\hat{m}_{\text{TSS}}$, and the true m.

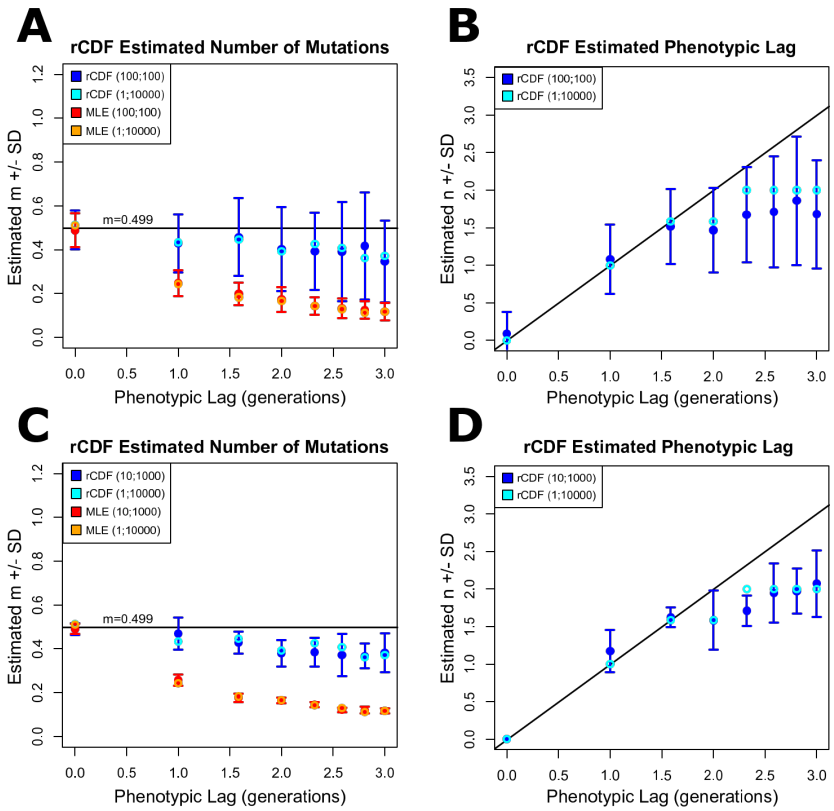

Figure 3.10: **Reduced CDF adjustment protocol applied to simulated fluctuation test data with phenotypic lag.** Fluctuation test data is simulated with and average number of mutations $m = 0.499$ and $\log_2$ integer phenotypic lag lengths (n). Three sets of data are simulated, one with 100 simulates of 100 samples (100;100), one with 10 simulates of 1000 samples (10;1000), and one with 1 simulate of 10000 samples (1;10000). A) The MLE fitting method and the reduced CDF (rCDF) phenotypic lag adjustment protocol are applied to the 100;100 and 1;10000 data sets, and the average estimated number of mutations ($\langle \hat{m} \rangle$) $\pm$ one standard deviation (SD) is plotted for each simulated phenotypic lag. B) The rCDF protocol is applied to the 100;100 and 1;10000 data sets, and the average estimated phenotypic lag length ($\langle \hat{n} \rangle$) $\pm$ SD is plotted for each simulated phenotypic lag. C) The MLE fitting method and the rCDF protocol are applied to the 10;1000 and 1;10000 data sets, and $\langle \hat{m} \rangle \pm$ SD is plotted for each simulated n. D) The rCDF protocol is applied to the 10;1000 and 1;10000 data sets, and $\langle \hat{n} \rangle \pm$ SD is plotted for each simulated n. In (B) and (D) the black line, $y = x$, is provided for easy comparison.

3.3.3 Combination Reduced CDF & Koch Adjustment

It is clear from the results of applying the Koch and reduced CDF protocols to simulated data that there are flaws in each system, but is it possible to use the best parts of each to get an even better adjustment protocol? First off, the Koch system is the more obvious choice for predicting the length of the phenotypic lag, n, primarily due to the smooth fitting landscape leading to an obvious choice for $\hat{n}$ on every occasion regardless of sample size. On the other hand, it appears that the reduced CDF method is a better choice for predicting the average number of mutations, m, because the Koch adjusted estimate $\hat{m}_K$ increases exponentially with phenotypic lag length, meaning variability in phenotypic lag estimates, $\hat{n}$, can result in larger variability in $\hat{m}_K$[9]. Consequently, a potentially good protocol would be to use the Koch system to determine the phenotypic lag length, $\hat{n}$, and then the reduced CDF system to determine the associated average number of mutations, $\hat{m}$. This method will be referred to as the "reduced CDF + Koch (rCDF+K)" protocol.

Another protocol which is worth exploring involves also using Koch to estimate the phenotypic lag length, but then taking the average of the associated estimates on the mutation numbers found by the Koch and reduced CDF protocols. Interestingly, the Koch estimate $\hat{m}_K$ appears to often be equally an overestimate as the reduced CDF estimate $\hat{m}_{\text{rCDF}}$ is an underestimate, meaning if one takes the average of the two, they get an estimate which is on average very close to the true m. This method will be referred to as the "reduced CDF + Koch average (rCDF+K_{avg})" protocol.

For both the rCDF+K and rCDF+K_{avg} protocols to work, only $\log_2$ integers can be estimated for $\hat{n}$ so that the reduced CDF fitting method works properly, meaning the predicted $\hat{n}$ may not be as accurate as in the Koch protocol. Both of the combination protocols discussed were applied to the same simulated data as the Koch and reduced CDF protocols in order to study their efficacy. The results can be found in Figures 3.11 and 3.12.

For both the combination protocols, the estimates of the phenotypic lag length are nearly indistinguishable from the estimates from the Koch protocol. When the rCDF+K protocol is applied to the simulated data, its estimates of the average number of mutations are generally underestimates, getting more drastic at higher phenotypic lags. The main benefit to the rCDF+K protocol is that it has the smallest standard deviation in $\hat{m}$ among all the discussed protocols when applied to the 100 simulations of 100 samples. The rCDF+K_{avg} protocol on the other hand does the best job at predicting the average number

[9]This is a potential cause of the high standard deviation in the estimates of the number of mutations when Koch is applied to simulated data (Fig. 3.8).

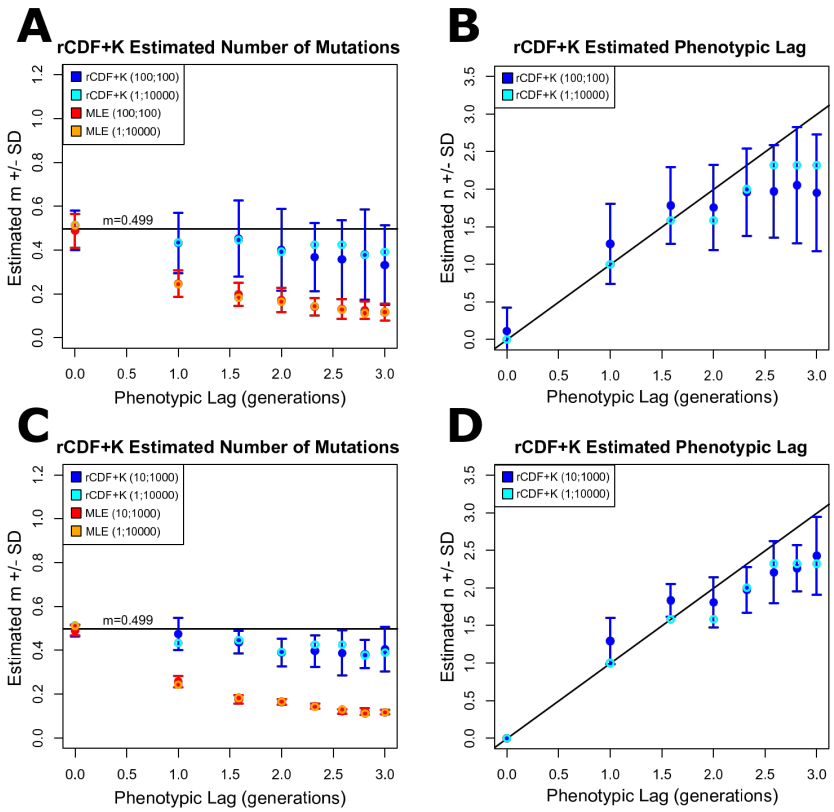

Figure 3.11: **rCDF+Koch adjustment protocol applied to simulated fluctuation test data with phenotypic lag.** Fluctuation test data is simulated with an average number of mutations $m = 0.499$ and $\log_2$ integer phenotypic lag lengths (n). Three sets of data are simulated, one with 100 simulates of 100 samples (100;100), one with 10 simulates of 1000 samples (10;1000), and one with 1 simulate of 10000 samples (1;10000). A) The MLE fitting method and the rCDF+Koch (rCDF+K) phenotypic lag adjustment protocol are applied to the 100;100 and 1;10000 data sets, and the average estimated number of mutations ($\langle \hat{m} \rangle$) $\pm$ one standard deviation (SD) is plotted for each simulated phenotypic lag. B) The rCDF+K protocol is applied to the 100;100 and 1;10000 data sets, and the average estimated phenotypic lag length ($\langle \hat{n} \rangle$) $\pm$ SD is plotted for each simulated phenotypic lag. C) The MLE fitting method and the rCDF+K protocol are applied to the 10;1000 and 1;10000 data sets, and $\langle \hat{m} \rangle$ $\pm$ SD is plotted for each simulated n. D) The rCDF+K protocol is applied to the 10;1000 and 1;10000 data sets, and $\langle \hat{n} \rangle$ $\pm$ SD is plotted for each simulated n. The black line, $y = x$, is provided for easy comparison.

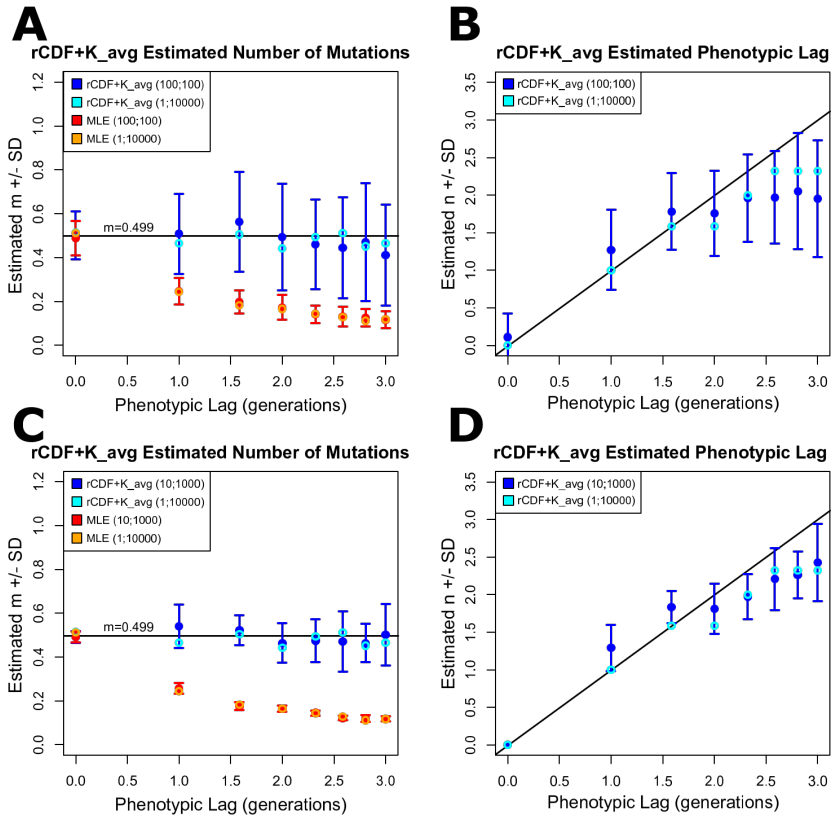

Figure 3.12: **rCDF+Koch average adjustment protocol applied to simulated fluctuation test data with phenotypic lag.** Fluctuation test data is simulated with an average number of mutations $m = 0.499$ and $\log_2$ integer phenotypic lag lengths (n). Three sets of data are simulated, one with 100 simulates of 100 samples (100;100), one with 10 simulates of 1000 samples (10;1000), and one with 1 simulate of 10000 samples (1;10000). A) The MLE fitting method and the rCDF+Koch average (rCDF+K_{avg}) phenotypic lag adjustment protocol are applied to the 100;100 and 1;10000 data sets, and the average estimated number of mutations ($\langle \hat{m} \rangle$) $\pm$ one standard deviation (SD) is plotted for each simulated phenotypic lag. B) The rCDF+K_{avg} protocol is applied to the 100;100 and 1;10000 data sets, and the average estimated phenotypic lag length ($\langle \hat{n} \rangle$) $\pm$ SD is plotted for each simulated phenotypic lag. C) The MLE fitting method and the rCDF+K_{avg} protocol are applied to the 10;1000 and 1;10000 data sets, and $\langle \hat{m} \rangle$ $\pm$ SD is plotted for each simulated n. D) The rCDF+K_{avg} protocol is applied to the 10;1000 and 1;10000 data sets, and $\langle \hat{n} \rangle$ $\pm$ SD is plotted for each simulated n. In (B) and (D) the black line is $y = x$.

of mutations, with the average $\hat{m}$ very close to the m inputted to the simulations. Also, the standard deviation in $\hat{m}$ for rCDF+K_{avg} among the 100 simulations of 100 samples is smaller than that of Koch and only a bit larger than that of the estimates from the rCDF+K protocol, which is promising.

3.3.4 Error in Phenotypic Lag Adjusted Estimates

As with all model fitting and parameter estimating, finding errors on the predictions is critical for determining the significance of the results. How to find confidence intervals for the estimated average number of mutations and mutation rates for the fitting methods which do not account for phenotypic lag were discussed throughout Section 3.1, but how to find confidence intervals on the phenotypic lag adjusted estimates has not been touched upon. Instead, the adjustment protocols were ran on many simulated data sets to get an idea of what level of variance one may expect from the mechanism, but this is of no help when applying the methods to experimental data. Because all four phenotypic lag adjustment protocols use the total sum of squares fitting to fit the adjusted data and the distribution being fitted to is unknown, the obvious choice is to use bootstrapping to calculate confidence intervals. Unfortunately, bootstrapping is a very computationally expensive practice, and when combined with how computationally inefficient my adjustment systems are, the computational cost can become prohibitive. This can be worked around slightly by assuming the predicted phenotypic lag length of the original data is the true lag, and then plugging this lag in and only fitting for the average number of mutations. If one wishes to estimate the phenotypic lag length as well, the run times can be forbidding and the errors can be difficult to interpret. As such, errors on the phenotypic lag adjusted estimates will only be provided for my experimental data and what is provided will be calculated with the phenotypic lag fixed at the initial estimate. Furthermore, the provided error bars will be bias-corrected and accelerated confidence intervals [47] from 10000 bootstrap replicates as calculated by the "boot" package in R [28].

3.3.5 Application of Phenotypic Lag Adjustments to Historical Data

In Luria and Delbrück's original 1943 paper they mention phenotypic lag as a possible reason for the discrepancy between the theoretical variance in the number of mutants and their measured variance, but quickly dismiss the idea due to the abundance of samples with

only a single mutant[10] [106]. Since then, many fluctuation tests have been performed and many have mentioned phenotypic lag in passing, but few have attempted to account for the lag in any significant manner. One exception to this is Newcombe's 1948 paper, "Delayed Phenotypic Expression of Spontaneous Mutations in *Escherichia Coli*" [124], which was written with the core purpose of quantifiably addressing and accounting for phenotypic lag in fluctuation tests. Unfortunately, the computational limitations of the time meant that their adjustments, though similar in theory to the reduced CDF protocol, are rudimentary. I will now apply my analysis tools to some historically significant fluctuation test data.

Newcombe Data Fitting

Newcombe (1948) [124], in addition to openly discussing the probable presence of phenotypic lag, provides one of the few published fluctuation tests with a statistically significant number of samples (200 samples total from 8 experiments with 25 samples, each with similar final populations) and reported raw data. Newcombe used *Escherichia coli* B/r which was grown in an undefined medium broth, presumably[11] giving a doubling time of less than 30 minutes. In four of the experiments, the cultures were inoculated with approximately 10 cells, while in the other four, they were inoculated with 10^4 cells[12]. In all 8 experiments, the cultures were grown overnight to an average final population of $3.5 \cdot 10^8$ cells with a standard deviation between experiments of $0.7 \cdot 10^8$, which gives a coefficient of variation of 20%. The selecting agent used by Newcombe to isolate mutants was T1 bacteriophage. There are several pathways towards resistance to T1 phage for *E coli*, but the most common are mutations to the *fhuA* or *tonB* genes [69, 103, 70], which are 2241 base pairs [38] and 717 base pairs [138] respectively [32]. The existence of several pathways towards resistance poses a problem for determining mutant fitness and the per base pair mutation rate. Another problem with Newcombe's experiment is that even though they attempted to address the potential physiological consequences of the different growth phases through control experiments, insufficient care was taken to ensure balanced growth, as made clear through the use of broth as the growth medium and allowing cells to reach

[10] We now know that stochastic effects can cause there to be samples with single mutants, as discussed in Section 3.2. In addition, it is unclear what effects growing to saturation will have on phenotypic expression in the presence of lag.

[11] The growth rate was not provided in the paper, but growth in broth generally results in a near-maximal growth rate.

[12] Because I wish to combine the data from all 8 experiments, when calculating mutation rates I will use the average, which is $5 \cdot 10^3$ cells.

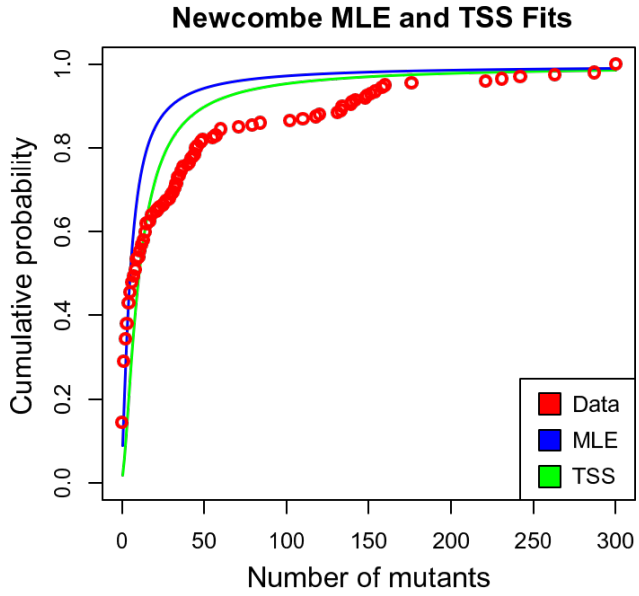

Figure 3.13: **Newcombe *E. coli* B/r in broth fluctuation test data with MLE and TSS fit.** The compiled data from the fluctuation tests performed by Newcombe (1948) [124] with *E. coli* B/r in broth are plotted as a cumulative distribution with the CDF of best fit as determined by rSalvador's maximum likelihood estimator (MLE) and the CDF of best fit as determined from the total sum of squares (TSS) method.

saturation during the fluctuation tests[13]. Newcombe determined from their analysis that there is likely between 2 and 6 generations of phenotypic lag on average and their adjusted mutation rate[14] is $3.17 \cdot 10^{-8}$ mutations per generation per cell [124]. Armitage also analysed Newcombe's data and estimated a phenotypic lag of 4 generations and adjusted mutation rate[15] of $2.7 \cdot 10^{-8}$ mutations per generation per cell [5]. For the results of my analysis protocols applied to Newcombe's data, see Figures 3.13, 3.14, 3.15, and 3.16 and Table 3.1.

[13] As with Luria and Delbrück, this work was prior to Schaechter et al.'s seminal work [153] so it is understandable that minimal physiological care was taken.

[14] Newcombe adjusted for phenotypic lag by developing a method to use the cultures with a large number of mutants to calculate the mutation rate, which has the form $\mu = \frac{r_{\max} - \langle r \rangle}{\nu N_f}$ where $r_{\max}$ is the maximum number of resistant cells in a single culture and $\langle r \rangle$ is the average number of resistant cells across all cultures.

[15] Armitage adjusted for phenotypic lag by only looking at an upper quartile of the data distribution.

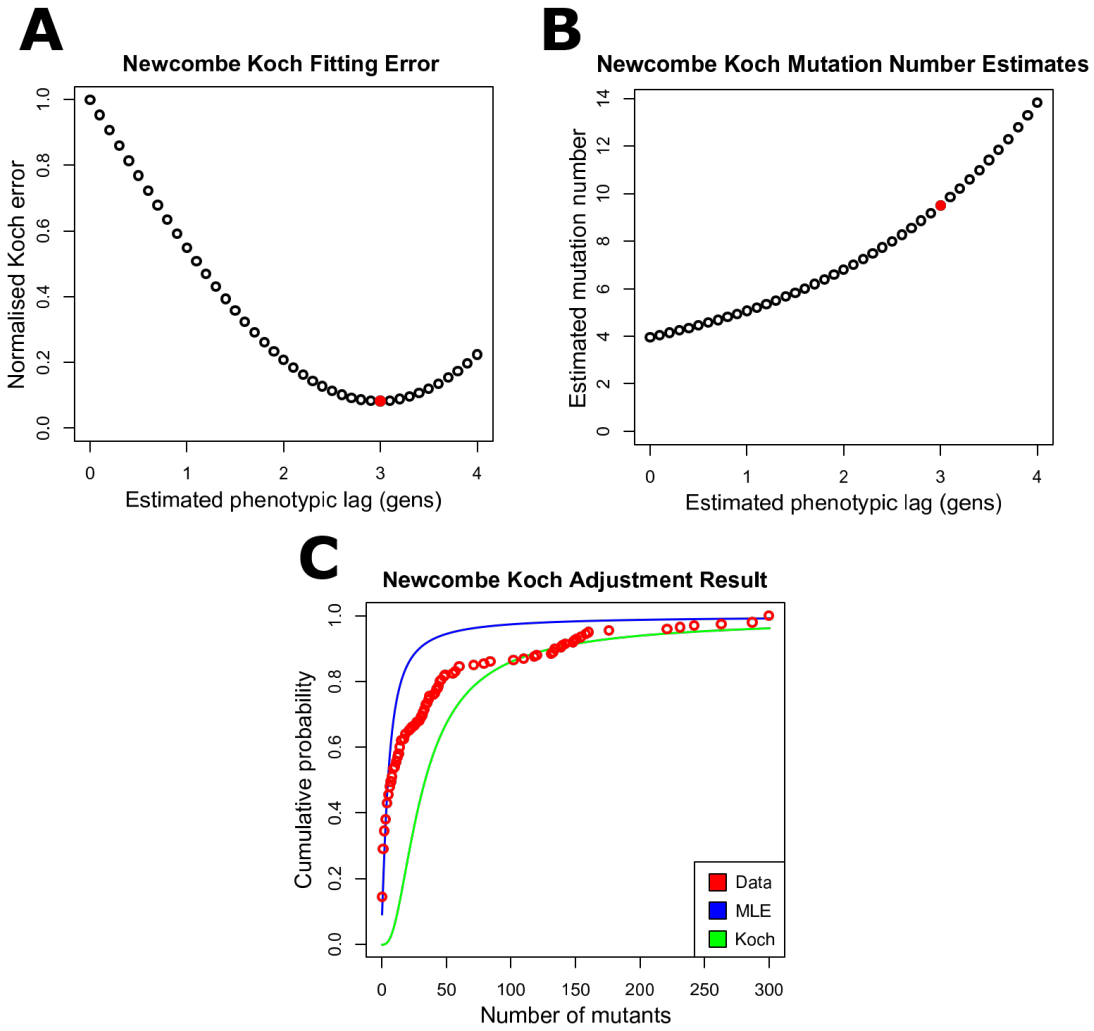

Figure 3.14: **Newcombe fluctuation test data: Koch adjusted fit.** The Koch adjustment protocol (see Section 3.3.1) applied to a set of fluctuation test data from Newcombe (1948) [124]. A) The Koch fitting error for a sequence of guessed phenotypic lags; the red point is the minimum. B) The Koch estimated average number of mutations for a sequence of guessed phenotypic lags; the red point is the chosen estimate. C) The Newcombe fluctuation test data plotted as a cumulative distribution with the CDF's from the MLE and Koch estimated mutation numbers.

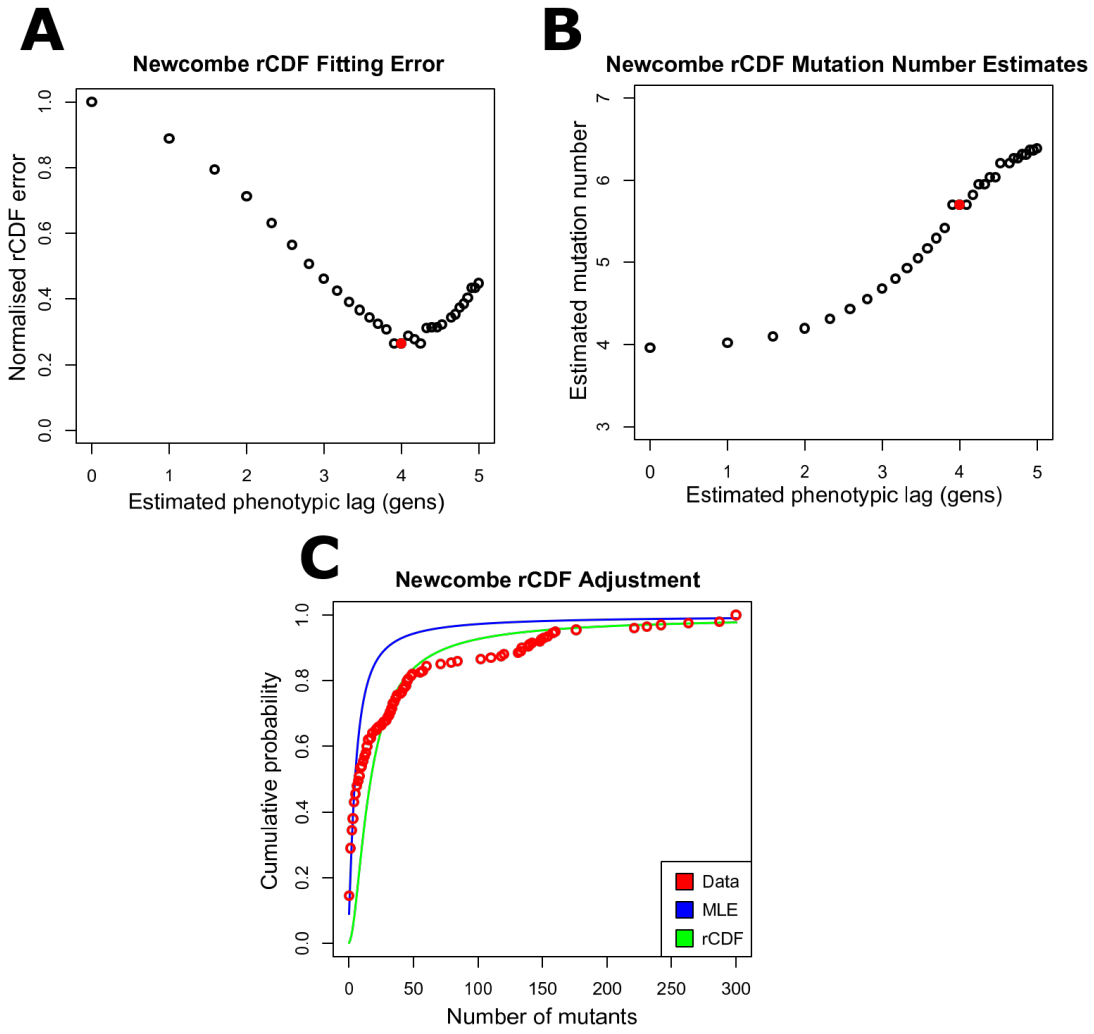

Figure 3.15: **Newcombe fluctuation test data: reduced CDF adjusted fit.** The reduced CDF (rCDF) adjustment protocol (see Section 3.3.2) applied to a set of fluctuation test data from Newcombe (1948) [124]. A) The rCDF fitting error for a sequence of guessed phenotypic lags; the red point is the minimum. B) The rCDF estimated average number of mutations for a sequence of guessed phenotypic lags; the red point is the chosen estimate. C) The Newcombe fluctuation test data plotted as a cumulative distribution with the CDF's from the MLE and rCDF estimated mutation numbers.

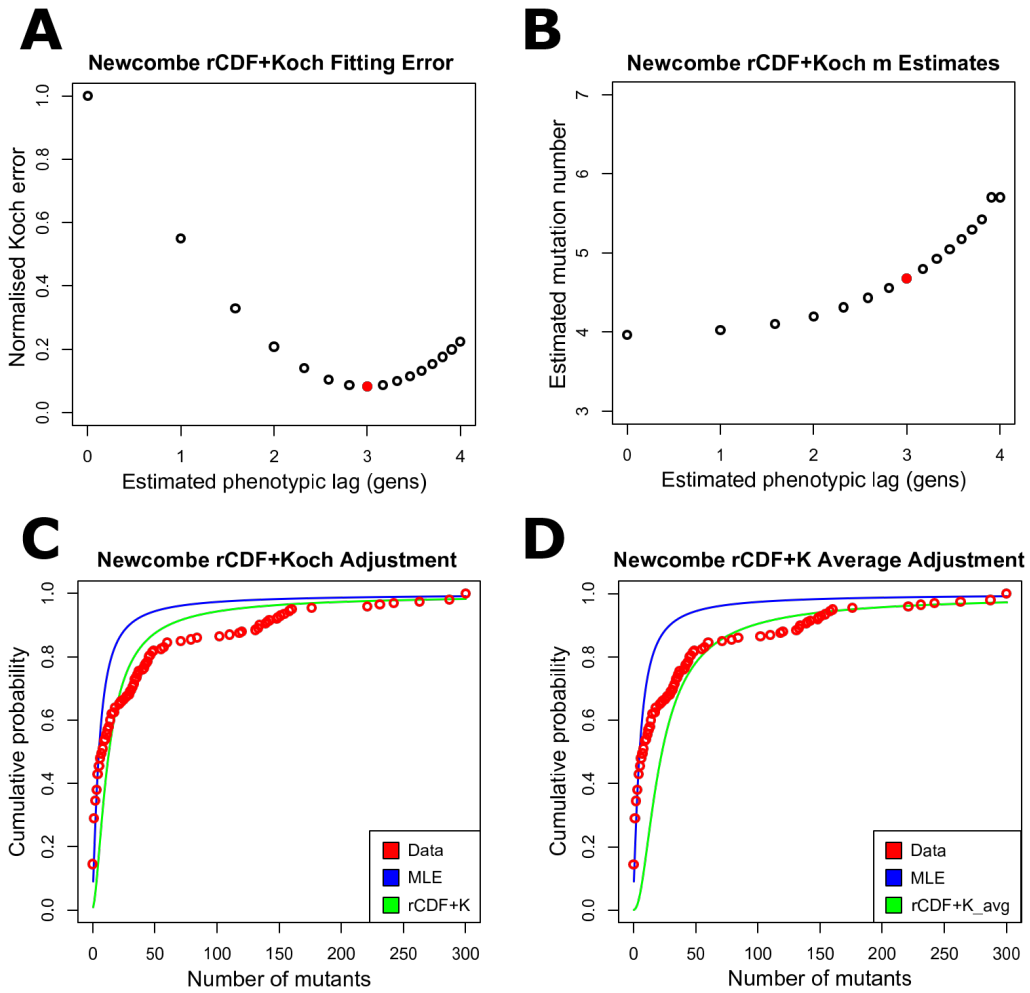

Figure 3.16: **Newcombe fluctuation test data: reduced CDF + Koch adjusted fits.** The hybrid rCDF + Koch and rCDF + Koch average adjustment protocols (see Section 3.3.3) applied to a set of fluctuation test data from Newcombe (1948) [124]. A) The rCDF + Koch fitting error for a sequence of guessed phenotypic lags; the red point is the minimum. B) The rCDF + Koch estimated average number of mutations for a sequence of guessed phenotypic lags; the red point is the chosen estimate. C) The Newcombe fluctuation test data plotted as a cumulative distribution with the CDF's from the MLE and rCDF + Koch estimated mutation numbers. D) The Newcombe fluctuation test data plotted as a cumulative distribution with the CDF's from the MLE and rCDF + Koch average estimated mutation numbers.

Newcombe mutation rates of *E. coli* B/r in broth				
Analysis protocol	Phenotypic lag length in generations (n)	Average number of mutations per culture (m)	Average number of mutations per cell per generation (μ_{cell}) ($\times 10^{-8}$)	Average number of mutations per base pair per generation (μ_{bp}) ($\times 10^{-12}$)
p_0	N/A	1.93	0.554	2.47
MLE	N/A	2.40	0.688	3.07
TSS	N/A	3.96	1.14	5.07
Koch	3	9.50	2.72	12.2
rCDF	4	5.70	1.63	7.30
rCDF+K	3	4.68	1.34	5.99
rCDF+K_{avg}	3	7.09	2.03	9.07

Table 3.1: **Newcombe *E. coli* B/r mutation rates in broth.** Mutation rates of *E. coli* B/r grown in broth (likely doubling time < 30 minutes) as determined by several different analysis methods. Data from Newcombe (1948) [124] where 200 cultures were inoculated with an average of 5005 cells and grown to saturation with an average final population of $(3.5 \pm 0.7) \cdot 10^8$ cells. T1 bacteriophage resistant mutants are selected and counted. To determine the per base pair mutation rate, the per cell mutation rate is divided by the size of the *fhuA* gene which is 2241 base pairs. No errors provided due to computational limitations.

Note that the mutation rate increases upwards of 4-fold when the data is adjusted for phenotypic lag. Also, the Koch adjustment, which gives the highest estimate and may be the most reliable because the reduced CDF protocol does not function optimally when there are very few zeros, gives an almost identical estimate to Armitage. Finally, the per base pair mutation rate appears to be about two orders of magnitude lower than the commonly recorded 10^{-10} mutations per base pair per generation [101, 197, 116]. This underestimate may be from dividing by too large a gene size, which would be the case if the mutations must take place in a specific portion of the gene for resistance to be gained. The underestimation may also be a repercussion of the experimental procedures used, especially since the experiment was performed in 1948 with the very original protocol that did not take into account bacterial growth physiology.

Boe et al. Data Fitting

Another fluctuation test with many samples and reported raw data is Boe et al. (1994) [18]. Boe et al. famously performed 23 fluctuation tests with 48 cultures each, all with a similar final population, giving a total of 1104 samples. Moreover, Boe et al. also addresses the likely presence of phenotypic lag and claims it to be a potential explanation for why their data differs from the theoretical Luria-Delbrück distribution. They grew *Escherichia coli* MG1655 in defined AB minimal medium with limiting 0.05% (w/v) glucose, which gave a doubling time of 72 ± 5 minutes. The cultures were inoculated with approximately $1.2 \cdot 10^4$ cells and grown overnight to a final population of approximately $1.2 \cdot 10^9$ cells[16]. The selecting agent used was nalidixic acid, which is an antibiotic that affects chromosomal DNA replication [65]. There are two main issues with Boe et al.'s experiment. The first issue is that they grew the cells to saturation (see Section 2.1 for a detailed discussion on the potential problems with this). The second issue is that nalidixic acid resistance can be achieved through mutations to several different genes [78], making it difficult to confidently say if a mutant has a change in fitness, as well as making it difficult to determine the per base pair mutation rate. Fortunately, it has been shown that the majority of mutants that are resistant to high levels[17] of nalidixic acid have mutations in one of two regions in the *gyrA* gene which code for an amino acid in DNA gyrase subunit A [104, 15]. Consequently, it seems reasonable to determine the per base pair mutation rate by simply dividing the per cell mutation rate by 3 or 6 to account for the commonly mutated nucleotide triplets that code for one or both of the amino acids associated with resistance. A downside to the majority of nalidixic acid resistance being from these particular mutations is that the

[16] The initial and final population numbers were not provided in detail within the paper [18].

[17] Boe et al. used a relatively high concentration of 100 μg/mL.

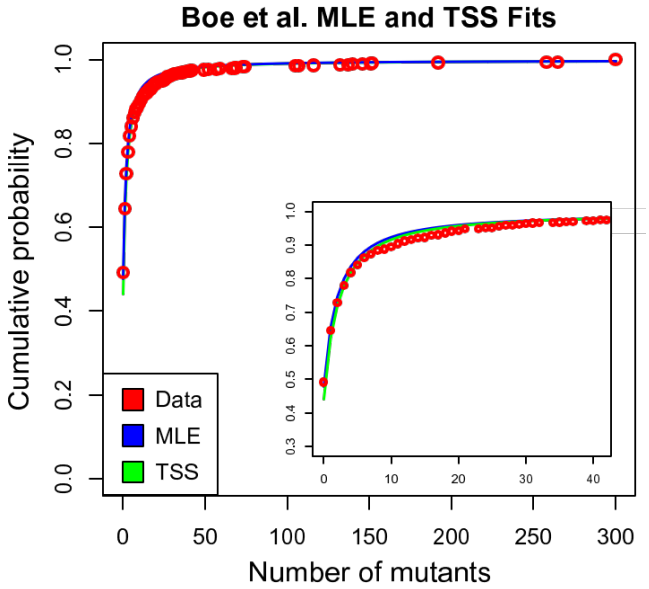

Figure 3.17: **Boe et al. *E. coli* MG1655 in AB minimal with glucose fluctuation test data with MLE and TSS fits.** The compiled data from the fluctuation tests performed by Boe et al. (1994) [18] with *E. coli* MG1655 in AB minimal medium with limiting glucose are plotted as a cumulative distribution with the CDF of best fit as determined by rSalvador's maximum likelihood estimator (MLE) and the CDF of best fit as determined from the total sum of squares (TSS) method. Plot with smaller domain and range inlaid in plot covering full domain and range.

system favours point substitution mutations, meaning the mutation rate found is more representative of the point substitution mutation rate than the overall mutation rate. Regardless, the data from Boe et al. was studied by applying my phenotypic lag adjustment systems to it. See Figures 3.17, 3.18, 3.19, and 3.20 and Table 3.2 for the results.

Note that a very small amount of lag is predicted by the Koch protocol, which agrees with Boe et al.'s prediction that if there is phenotypic lag, it is short [18]. The prediction of lag also agrees with Carballo-Pacheco et al.'s prediction that phenotypic lag is present in the Boe et al. system [29], though they give no indication of how long the lag is. In addition, the unadjusted per cell[18] mutation rate is only slightly higher than that of Lee et al. (2012) [101] who also use *E. coli* MG1655 and nalidixic acid, but grow the cells in LB. On the other hand, the adjusted per cell mutation rate is upwards of 2 times as large

[18]Lee et al. divide the per cell mutation rate by a larger gene size than me to calculate the per base pair mutation rate.

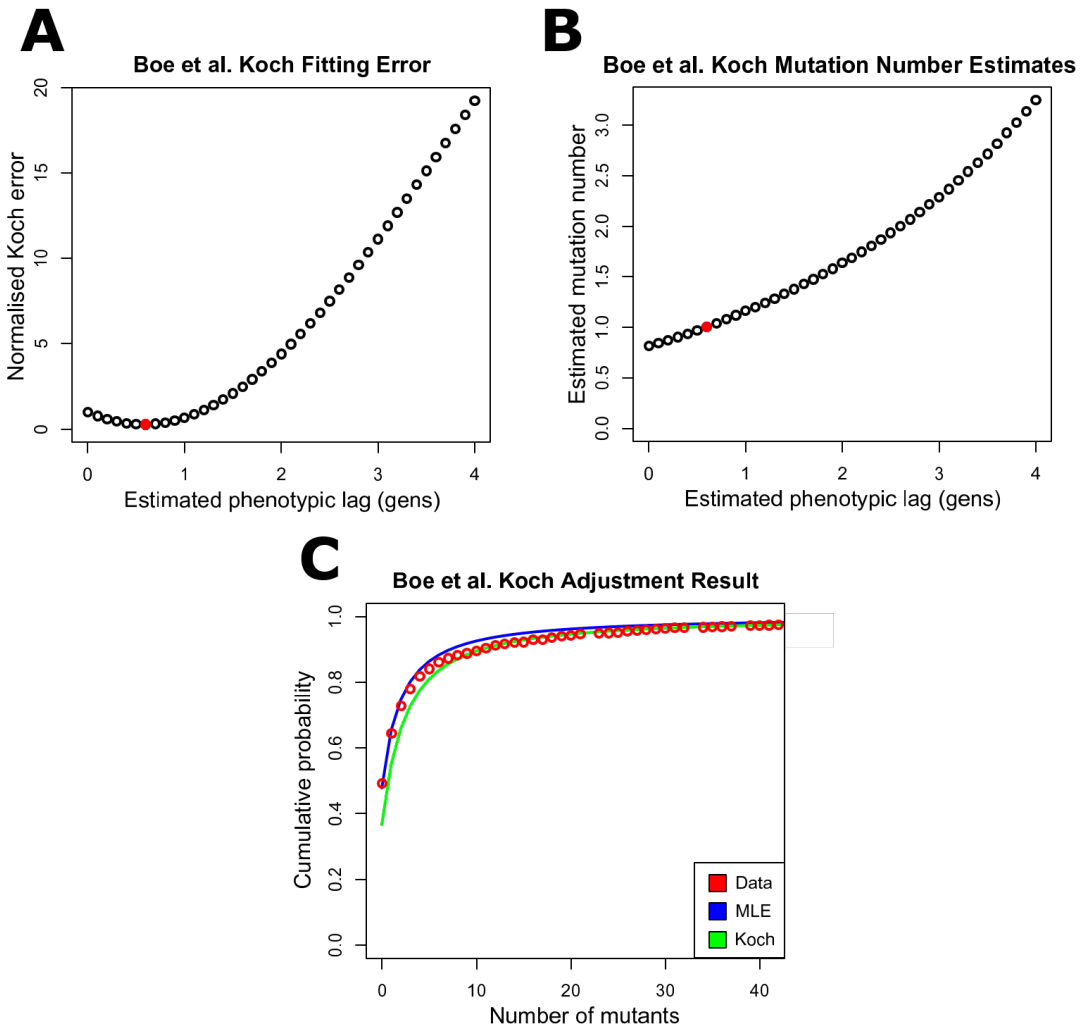

Figure 3.18: **Boe et al. fluctuation test data: Koch adjusted fit.** The Koch adjustment protocol (see Section 3.3.1) applied to a set of fluctuation test data from Boe et al. (1994) [18]. A) The Koch fitting error for a sequence of guessed phenotypic lags; the red point is the minimum. B) The Koch estimated average number of mutations for a sequence of guessed phenotypic lags; the red point is the chosen estimate. C) The Boe et al. fluctuation test data plotted as a cumulative distribution with the CDF's from the MLE and Koch estimated mutation numbers.

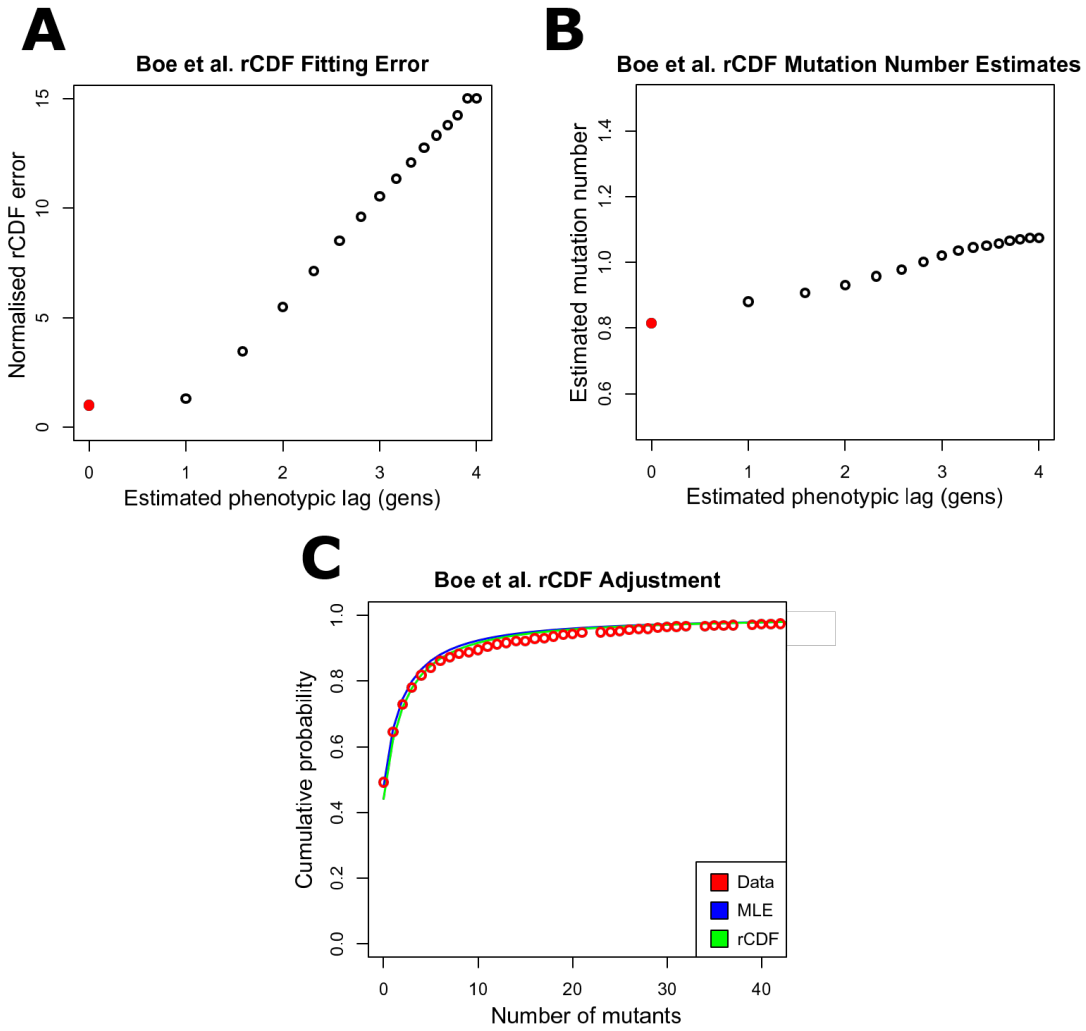

Figure 3.19: **Boe et al. fluctuation test data: reduced CDF adjusted fit.** The reduced CDF (rCDF) adjustment protocol (see Section 3.3.2) applied to a set of fluctuation test data from Boe et al. (1994) [18]. A) The rCDF fitting error for a sequence of guessed phenotypic lags; the red point is the minimum. B) The rCDF estimated average number of mutations for a sequence of guessed phenotypic lags; the red point is the chosen estimate. C) Boe et al. fluctuation test data plotted as a cumulative distribution with the CDF's from the MLE and rCDF estimated mutation numbers.

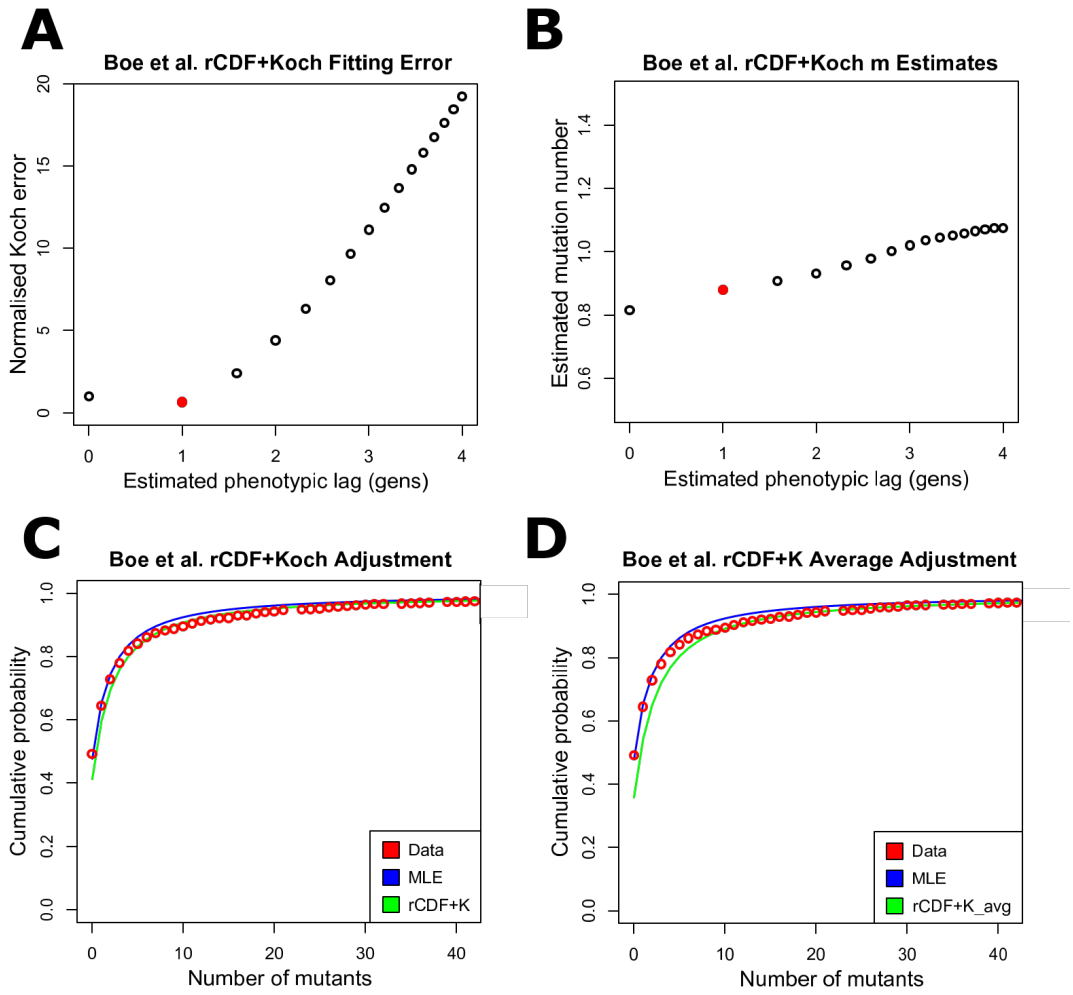

Figure 3.20: **Boe et al. fluctuation test data: reduced CDF + Koch adjusted fits.** The hybrid rCDF + Koch and rCDF + Koch average adjustment protocols (see Section 3.3.3) applied to a set of fluctuation test data from Boe et al. (1994) [18]. A) The rCDF + Koch fitting error for a sequence of guessed phenotypic lags; the red point is the minimum. B) The rCDF + Koch estimated average number of mutations for a sequence of guessed phenotypic lags; the red point is the chosen estimate. C) The Boe et al. fluctuation test data plotted as a cumulative distribution with the CDF's from the MLE and rCDF + Koch estimated mutation numbers. D) The Boe et al. fluctuation test data plotted as a cumulative distribution with the CDF's from the MLE and rCDF + Koch average estimated mutation numbers.

Boe et al. mutation rates of *E. coli* MG1655 in AB minimal with glucose				
Analysis protocol	Phenotypic lag length in generations (n)	Average number of mutations per culture (m)	Average number of mutations per cell per generation (μ_{cell}) ($\times 10^{-10}$)	Average number of mutations per base pair per generation (μ_{bp}) ($\times 10^{-10}$)
p_0	N/A	0.710	5.91	1.97
MLE	N/A	0.737	6.14	2.05
TSS	N/A	0.815	6.79	2.26
Koch	0.6	1.00	8.37	2.79
rCDF	0	0.815	6.79	2.26
rCDF+K	1	0.88	7.33	2.44
rCDF+K_{avg}	1	1.021	8.51	2.84

Table 3.2: **Boe et al. *E. coli* MG1655 mutation rates in AB minimal medium with glucose.** Mutation rates of *E. coli* MG1655 grown in AB minimal media with limiting 0.05% glucose carbon source (doubling time = 72 ± 5 minutes) as determined by several different analysis methods. Data from Boe et al. (1994) [18] where 1104 cultures were inoculated with $1.2 \cdot 10^4$ cells on average and grown to saturation with final population of $1.2 \cdot 10^9$ cells on average. Nalidixic acid resistant mutants are selected and counted. The per base pair mutation rates are calculated by dividing the per cell mutation rate by 3 base pairs because the resistance is generally achieved through mutation of one of two specific nucleotide triplets that code for an amino acid. No errors provided due to computational limitations.

as Lee et al. when phenotypic lag is accounted for. Finally, my estimates for the per base pair mutation rate from Boe et al.'s system agrees well with the per base pair mutation rate Lee et al. calculate with whole-genome sequencing [101].

3.4 Comparison Between Fluctuation Tests

When comparing data from different experiments, one often employs both qualitative and quantitative methods. The qualitative methods help give intuition for the differences between the results while the quantitative methods determine just how real and significant those differences are. The most obvious and powerful qualitative comparison one can perform between data from different fluctuation tests is through plotting the cumulative distribution functions (CDF) and/or probability distribution functions (PDF) from each experiment on the same plot. One can only do this if the cells grew the same amount in each experiment (same final populations with sufficiently small initial populations), meaning the only variable that could change the shape of the data is the mutation rate, μ [207]. When the CDFs are plotted together, one can compare the medians (0.5 on the y-axis) to get a rough estimate on the relative mutation rates between tests. How the PDF and CDF look for a wide array of average mutation numbers is shown in Fig. 3.21.

Methods for quantitatively comparing mutation rates from separate experiments are not abundant. The most obvious it to compare the 95% confidence intervals ($CI_{95\%}$) of each experiment and see if there is overlap [156]. Though a good starting technique, it is not definitive considering 95% confidence intervals can overlap even if the two variables have significant difference [95]. As a result, it has been suggested that the comparison of the 84% confidence intervals instead be conducted [109, 207]. The 84% confidence interval is chosen because when these intervals do not overlap, it mimics a statistical test with p-value 0.05 [109]. This test can be used to compare the average number of mutations, m, or the mutation rates, μ, from different experiments and has showed success at distinguishing between data simulated with different mutation rates [207]. The confidence interval test is particularly powerful when it is used with mutation rates instead of the number of mutations because then it no longer relies on each experiment having the same amount of growth (or final population) [207]. The downside to using mutation rates is that the variance in the initial and final populations will inevitably add to the uncertainty in the mutation rate[19]. On the other hand, it has be shown that having variance in the final

[19]Most investigators seem to ignore this phenomena [207], with Foster arguing that "if the denominator is larger than the numerator, the variance of the ratio will be smaller than the variance of the numerator, and thus no great harm should be done by ignoring the variance of the denominator" [58], but this appears

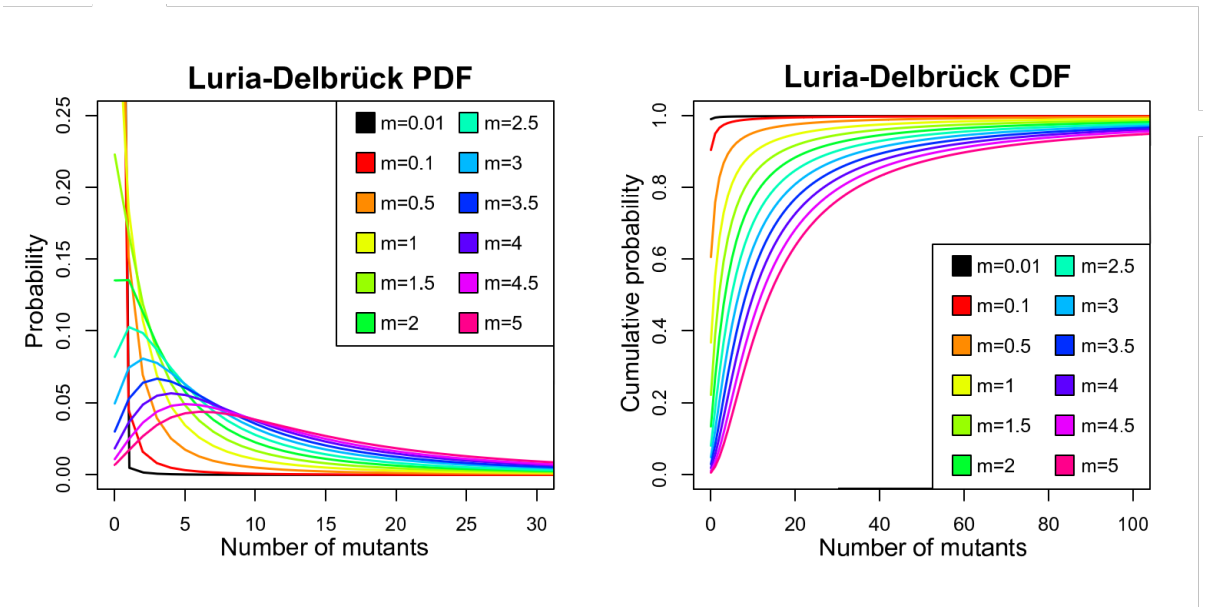

Figure 3.21: **Luria-Delbrück PDF and CDF for several different average mutation numbers.** Each probability distribution function (PDF) is created using rSalvador and each cumulative distribution function (CDF) is built by taking the cumulative sum of the the associated PDF.

population will affect the estimate on the average number of mutations [198, 209] so if this variance is not taken into account when estimating the number of mutations, then it may be beneficial to account for it during the calculation of the mutation rate.

Another test, which was developed by Zheng, is a likelihood ratio test [207]. The test requires that the final populations of the experiments being compared are the same and it calculates a p-value by comparing the log-likelihood that the experiments have the same number of mutations, m, to the log-likelihood of them having different numbers of mutations. In the next chapter we will use all three methods (95% confidence, 84% confidence, and likelihood ratio) to compare the mutation rates from multiple experiments performed in the same media as well as comparing experiments performed in different growth media in order to investigate if bacterial mutation rates are dependent on their exponential growth rates.

to disregard the rules of error propagation.

Chapter 4

Experimental Results

Fluctuation tests as described in Section 2.4 were performed by me in two different media, a rich defined medium with glucose (RDM glucose) with a doubling time of 23 minutes (Table 2.1), and a minimal medium with maltose (maltose minimal) with a doubling time of 48 minutes (Table 2.1). In each case, several different fitting protocols were applied to the data to estimate the average number of mutations per culture, m, as described in Section 3.1. Adjustments for phenotypic lag were then applied in order to estimate the length of phenotypic lag, n, and the corresponding adjusted average number of mutations. The corresponding mutation rate, μ, which gives the average number of mutations per cell[1] per generation is then calculated for all estimates of m as described in Section 3.1.4. The mutation rates are also given in more general units by assuming that mutations are randomly distributed around the genome and adjusting for the size of the genome and gene of interest. Because mutations were measured by selecting for cycloserine resistance in my experiments, we can assume that every observed mutation took place in the *cycA* gene (Section 2.3) which is 1413 base pairs (bp) long [51] (i.e. the per cell mutation rate is effectively a per gene mutation rate). Accordingly, to get a per base pair mutation rate, μ_{bp}, the per cell mutation rate must be divided by the length of the *cycA* gene ($\mu_{\text{bp}} = \frac{\mu}{1413}$ mutations per base pair per generation). Furthermore, to get the per genome mutation rate, μ_{genome}, one must multiply the per base pair mutation rate by the chromosome size, which is 4,678,045 base pairs for *E. coli* NCM3722 ($\mu_{\text{bp}} = \mu \cdot 4.68 \cdot 10^6$ mutations per genome per generation). Errors on all estimates of m and μ are provided as described throughout Chapter 3. The results for each medium are compared using the methods described in Section 3.4 in order to determine if there is a significant difference in mutation rate when

[1] Later in the chapter this mutation rate will be represented with μ_{cell} in order to emphasise the units.

the exponential growth rate of the bacteria is altered through changes in the quality of the nutrients they have access to. Finally, some difficulties faced during experimentation and preliminary results for a third medium, α-ketoglutarate minimal, are discussed.

4.1 RDM Glucose

Two fluctuation tests, performed as described in Section 2.4, were done in a MOPS based rich defined media (RDM) using glucose as the carbon source. Each fluctuation test produced 50 samples which were selected for mutants, combining for a total of 100 points. Additionally, 10 independent samples from which to determine the final population were grown in each experiment, combining for 20 population points. The RDM glucose medium used gives a very fast CFU-based specific growth rate of 1.80 ± 0.04 /hr, which corresponds to a doubling time of 23.2 ± 0.5 minutes. The cells were grown from an average initial population of 975 cells with a standard deviation of 99 cells, giving a coefficient of variation (CV) of 10.1%. The growth period was 3 hours and 33 minutes in one experiment and 3 hours and 34 minutes in the other. The result was an average final population of $(4.2 \pm 0.5) \cdot 10^5$ cells, giving a CV of 11.4%. Calculating the growth rate from the difference between initial and final populations along with the growth time gives a doubling time of 24.3 minutes, which agrees well with the independently determined growth rate.

With the fluctuation test data coming from the combination of two separate runs, it is important to confirm that they are sufficiently similar to be combined. It can be seen that the 84% confidence intervals for their MLE estimated average number of mutations, $\hat{m}_{\text{MLE}}$, are $(0.122, 0.313)$ and $(0.270, 0.546)$ which comfortably overlap. Furthermore, from the likelihood ratio test the p-value for the results being the same is 0.105. These two statistical measures support the decision to combine the two experiments into a single data set. The combined raw data for all population plates as well as the mutant plates is included in Table 4.1.

The first method used to estimate the average number of mutations, m, from the data was Delbrück's p_0 method, and it gave an average number of mutations, $\hat{m}_{p_0} = 0.288$, with a 95% confidence interval of $(0.175, 0.401)$ which corresponds to a mutation rate of $\hat{\mu}_{p_0} = 6.79 \cdot 10^{-7}$ mutations per cell per generation with 95% confidence interval $(3.73 \cdot 10^{-7}, 9.87 \cdot 10^{-7})$. When the data is fit using rSalvador's MLE system, $\hat{m}_{\text{MLE}} = 0.294$ with a 95% confidence interval of $(0.193, 0.424)$ is found, which corresponds to a mutation rate of $\hat{\mu}_{\text{MLE}} = 6.93 \cdot 10^{-7}$ with $\text{CI}_{95\%} = (4.10 \cdot 10^{-7}, 1.04 \cdot 10^{-6})$. Impressively, the p_0 estimate agrees very closely with the MLE estimate despite the difference in sophistication. Finally, the TSS fitting method gives $\hat{m}_{\text{TSS}} = 0.329$ with $\text{CI}_{95\%} = (0.217, 0.521)$ and

Initial Population	Final Population	Number of Mutants				
	$4.2 \cdot 10^5$	0	0	0	0	0
	$4.2 \cdot 10^5$	0	0	0	0	1
960	$4.2 \cdot 10^5$	0	0	25	0	28
967	$3.0 \cdot 10^5$	0	0	0	0	0
852	$4.8 \cdot 10^5$	0	0	0	0	0
$1.05 \cdot 10^3$	$4.6 \cdot 10^5$	54	0	0	21	1
$1.07 \cdot 10^3$	$4.0 \cdot 10^5$	2	0	0	0	1
$1.07 \cdot 10^3$	$4.5 \cdot 10^5$	14	1	0	0	0
$1.06 \cdot 10^3$	$4.9 \cdot 10^5$	0	0	0	0	0
$1.05 \cdot 10^3$	$4.9 \cdot 10^5$	0	0	0	12	0
928	$4.4 \cdot 10^5$	0	0	0	0	1
804	$4.1 \cdot 10^5$	0	1	1	0	0
792	$3.6 \cdot 10^5$	3	1	0	0	0
$1.09 \cdot 10^3$	$3.6 \cdot 10^5$	0	0	1	0	0
900	$3.7 \cdot 10^5$	0	0	0	0	0
984	$3.9 \cdot 10^5$	0	0	9	0	8
$1.07 \cdot 10^3$	$4.5 \cdot 10^5$	0	0	10	0	0
944	$4.6 \cdot 10^5$	0	0	1	0	0
	$4.4 \cdot 10^5$	0	0	0	0	177
	$4.5 \cdot 10^5$	0	2	1	9	0
Mean ± SD	**Mean ± SD**	**Doubling Time**				
975 ± 99	$(4.2 \pm 0.5) \cdot 10^5$	24.35 minutes				

Table 4.1: **RDM glucose fluctuation test data.** Combined data from two fluctuation tests with *E. coli* NCM3722 in a MOPS based rich defined medium with glucose as the carbon source and D-cycloserine as the selecting agent.

$\hat{\mu}_{\text{TSS}} = 7.77 \cdot 10^{-7}$ with $\text{CI}_{95\%} = (4.60 \cdot 10^{-7}, 1.26 \cdot 10^{-6})$, which also agrees well with the MLE estimate. Plots of the experimental CDF along with the CDF's for $\hat{m}_{\text{MLE}}$ and $\hat{m}_{\text{TSS}}$ are shown in Fig. 4.1.

Attempts to account for potential phenotypic lag with the goal of getting a better fit on the data were pursued and all four adjustment protocols discussed in Section 3.3 were applied. The Koch protocol predicts the length of phenotypic lag to be $\hat{n}_K = 1.6$ generations and gives an estimate of the average number of mutations of $\hat{m}_K = 0.724$ with $\text{CI}_{95\%} = (0.464, 1.14)$ (Fig. 4.2). The reduced CDF protocol gives an estimate on the phenotypic lag of $\hat{n}_{\text{rCDF}} = 2$ generations and on mutation number of $\hat{m}_{\text{rCDF}} = 0.730$

with $\text{CI}_{95\%} = (0.377, 1.30)$ (Fig. 4.3). The Koch and reduced CDF estimates of m agree very well, which is especially surprising when considering that the average estimates of each when ran on simulated data are noticeably different. The Koch estimate of $\hat{m}_K = 0.724$ and the rCDF estimate of $\hat{m}_{\text{rCDF}} = 0.730$ are both approximately a 2.5 fold increase from the MLE estimated $\hat{m}_{\text{MLE}} = 0.294$. The hybrid Koch and rCDF adjustment protocols give an estimate on n of $\hat{n}_{\text{rCDF+K}} = \log_2(3) \approx 1.58$ while the two estimates of the average number of mutation are $\hat{m}_{\text{rCDF+K}} = 0.559$ with $\text{CI}_{95\%} = (0.255, 1.09)$ and $\hat{m}_{\text{rCDF+K}_{\text{avg}}} = 0.638$ with $\text{CI}_{95\%} = (0.363, 1.08)$ (Fig. 4.4). The two hybrid estimates don't agree with the reduced CDF and Koch estimates quite as well as the rCDF and Koch estimates agree with each other, but they still increase the estimated average number of mutations as desired from a phenotypic lag adjustment. All estimates are compiled in Table 4.2 and plotted together as a histogram with errors in Fig. 4.5.

For no particularly good reason beyond the fact that the Koch adjusted and rCDF adjusted fits are nearly indistinguishable, and the error in the Koch estimate is the smallest among the phenotypic lag adjustment protocols, I will primarily focus on the Koch adjusted estimate during future discussions.

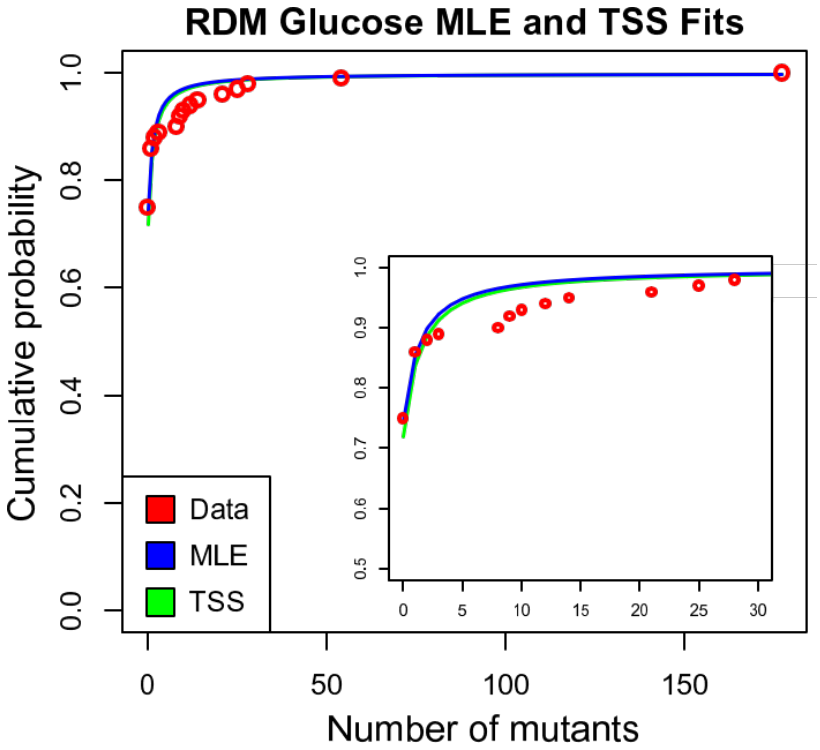

Figure 4.1: **RDM glucose fluctuation test data with MLE and TSS fit.** The compiled data from the fluctuation tests performed with *E. coli* NCM3722 in RDM glucose are plotted as a cumulative distribution with the CDF of best fit as determined by rSalvador's maximum likelihood estimator (MLE) and the CDF of best fit as determined from the total sum of squares (TSS) method. Plot with smaller domain and range inlaid in plot covering full domain and range.

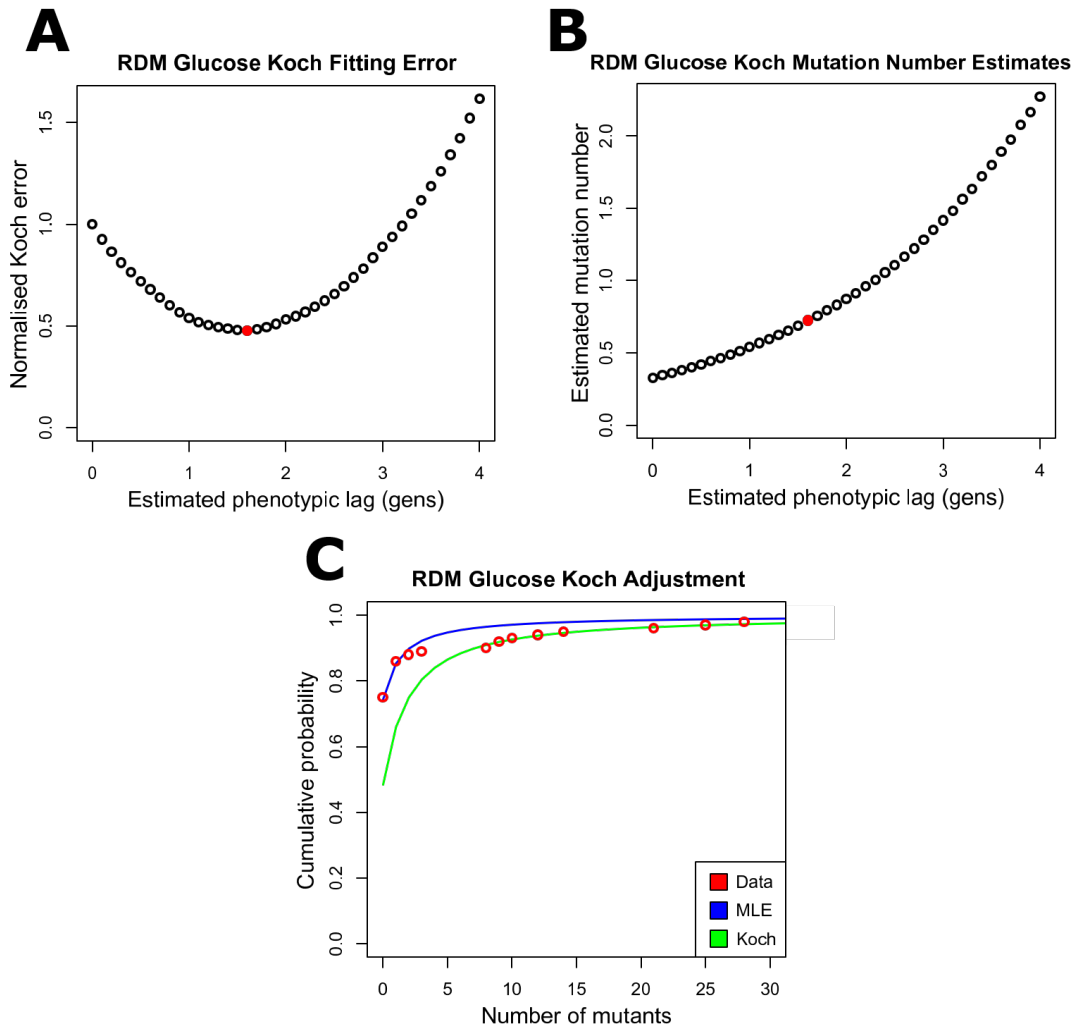

Figure 4.2: **RDM glucose fluctuation test data Koch adjusted fit.** The Koch adjustment protocol (see Section 3.3.1) applied to a set of fluctuation test data from *E. coli* NCM3722 grown in RDM glucose. A) The Koch fitting error for a sequence of guessed phenotypic lags; the red point is the minimum. B) The Koch estimated average number of mutations for a sequence of guessed phenotypic lags; the red point is the chosen estimate. C) The RDM glucose fluctuation test data plotted as a cumulative distribution with the CDF's from the MLE and Koch estimated mutation numbers.

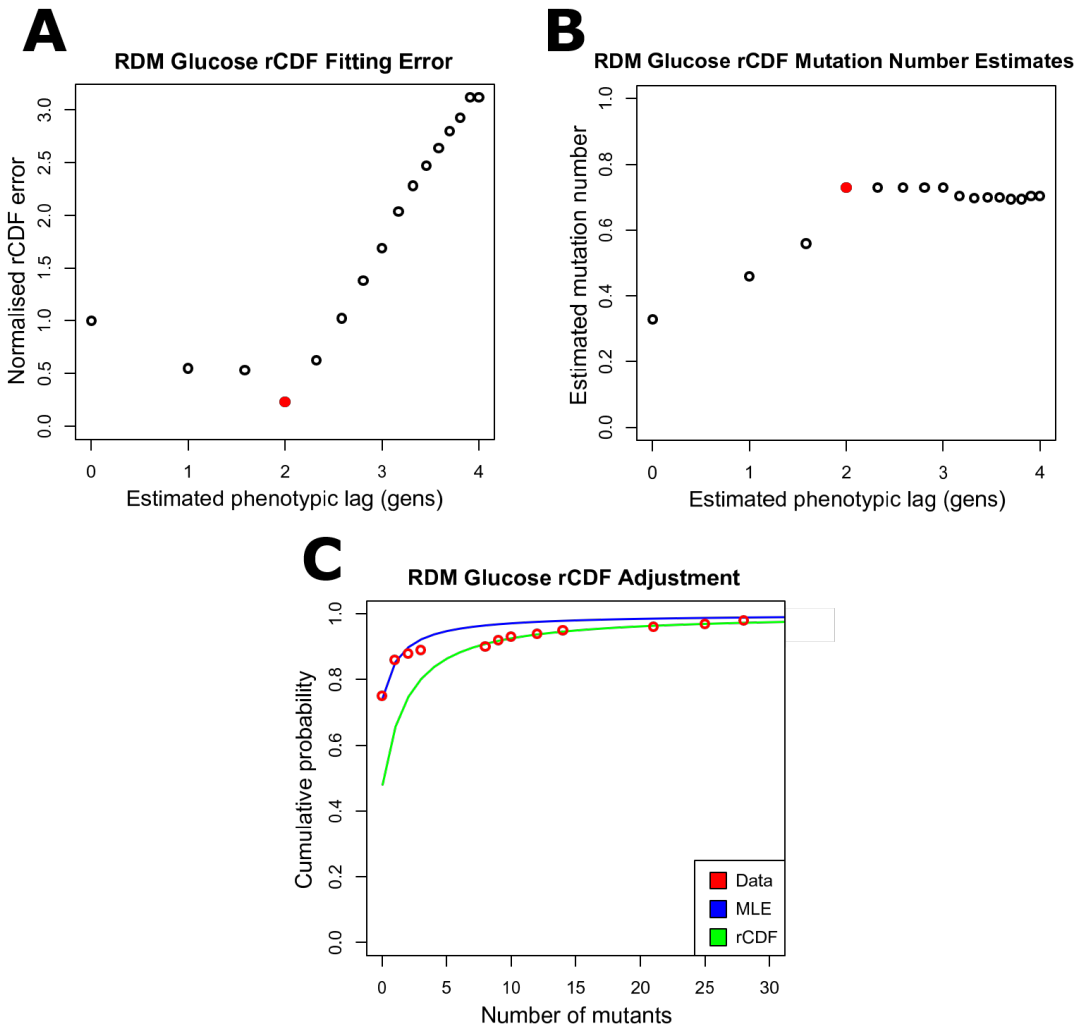

Figure 4.3: **RDM glucose fluctuation test data reduced CDF adjusted fit.** The reduced CDF (rCDF) adjustment protocol (see Section 3.3.2) applied to a set of fluctuation test data from *E. coli* NCM3722 grown in RDM glucose. A) The rCDF fitting error for a sequence of guessed phenotypic lags; the red point is the minimum. B) The rCDF estimated average number of mutations for a sequence of guessed phenotypic lags; the red point is the chosen estimate. C) The RDM glucose fluctuation test data plotted as a cumulative distribution with the CDF's from the MLE and rCDF estimated mutation numbers.

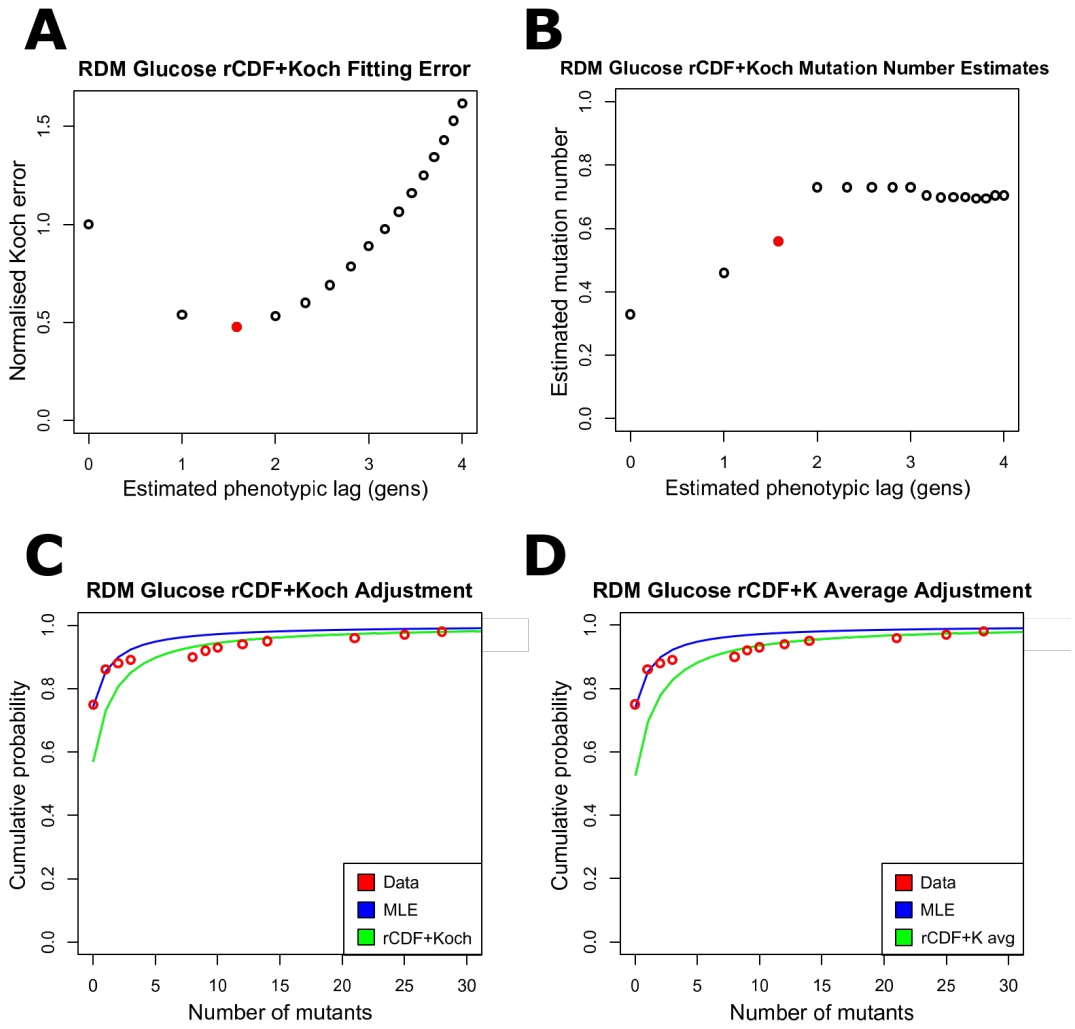

Figure 4.4: **RDM glucose fluctuation test data reduced CDF + Koch adjusted fits.** The hybrid rCDF + Koch and rCDF + Koch average adjustment protocols (see Section 3.3.3) applied to a set of fluctuation test data from *E. coli* NCM3722 grown in RDM glucose. A) The rCDF + Koch fitting error for a sequence of guessed phenotypic lags; the red point is the minimum. B) The rCDF + Koch estimated average number of mutations for a sequence of guessed phenotypic lags; the red point is the chosen estimate. C) The RDM glucose fluctuation test data plotted as a cumulative distribution with the CDF's from the MLE and rCDF + Koch estimated mutation numbers. D) The RDM glucose fluctuation test data plotted as a cumulative distribution with the CDF's from the MLE and rCDF + Koch average estimated mutation numbers.

Mutation rates of *E. coli* NCM3722 in RDM glucose					
Analysis protocol	Phenotypic lag length in generations (n)	Average number of mutations per culture (m) with 95% confidence interval	Average number of mutations per cell per generation (μ_{cell}) with 95% confidence interval ($\times 10^{-7}$)	Average number of mutations per base pair per generation (μ_{bp}) with 95% confidence interval ($\times 10^{-10}$)	Average number of mutations per genome per generation (μ_{genome}) with 95% confidence interval ($\times 10^{-3}$)
p_0	N/A	0.288 (0.175, 0.401)	6.79 (3.73, 9.87)	4.80 (2.64, 6.98)	2.25 (1.23, 3.27)
MLE	N/A	0.294 (0.193, 0.424)	6.93 (4.10, 10.4)	4.90 (2.90, 7.36)	2.29 (1.35, 3.44)
TSS	N/A	0.329 (0.217, 0.521)	7.77 (4.60, 12.6)	5.50 (3.26, 8.93)	2.57 (1.52, 4.18)
Koch	1.6	0.724 (0.464, 1.14)	17.1 (9.86, 27.6)	12.1 (6.98, 19.5)	5.66 (3.27, 9.15)
rCDF	2	0.730 (0.377, 1.30)	17.2 (8.05, 31.2)	12.2 (5.70, 22.1)	5.70 (2.67, 10.3)
rCDF+K	1.58	0.559 (0.255, 1.09)	13.2 (5.44, 26.1)	9.34 (3.85, 18.4)	4.37 (1.80, 8.63)
rCDF+K_{avg}	1.58	0.638 (0.363, 1.08)	15.1 (7.75, 26.0)	10.7 (5.49, 18.4)	4.99 (2.57, 8.61)

Table 4.2: ***E. coli* NCM3772 mutation rates in RDM glucose.** Mutation rates of *E. coli* NCM3722 grown in MOPS based rich defined media with glucose carbon source (doubling time = 23.2 ± 0.5 minutes) as determined by the different analysis methods described in Chapter 3. The estimated phenotypic lag length from the rCDF+K and rCDF+K_{avg} protocols is $\log_2(3) \approx 1.58$. Errors in the phenotypic lag adjustment protocols do not account for the error in the estimated phenotypic lag. The per base pair mutation rate was calculated by dividing the per cell mutation rate by 1413 base pairs. The per genome mutation rate was calculated by multiplying the per base pair mutation rate by $4.68 \cdot 10^6$ base pairs.

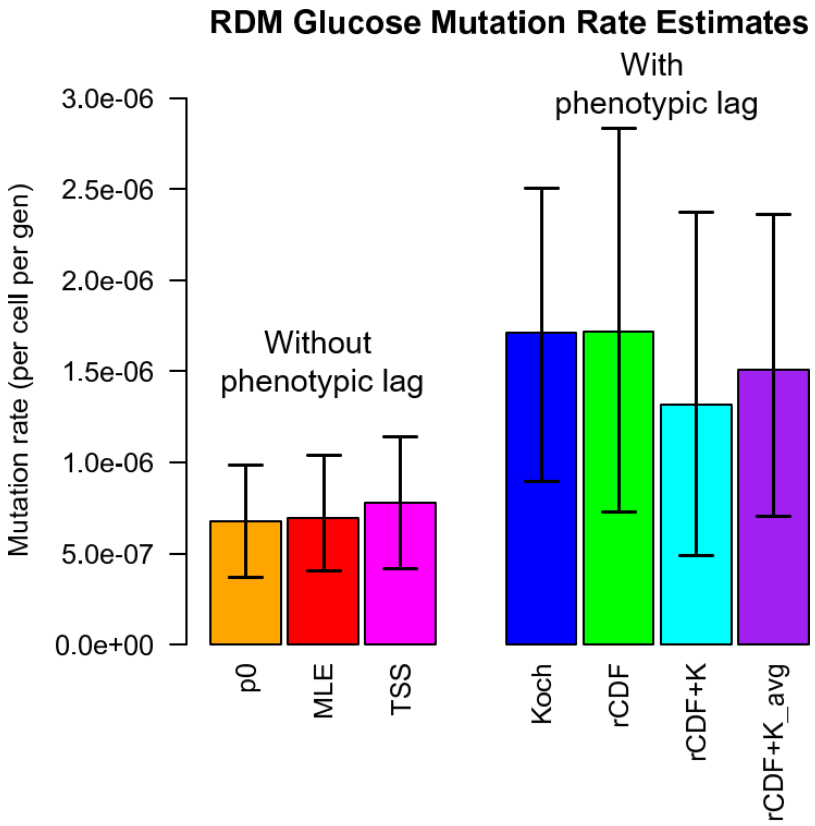

Figure 4.5: **Histogram comparing mutation rates calculated from different fitting protocols for *E. coli* NCM3772 in RDM glucose.** The per cell mutation rates with their 95% confidence intervals are plotted for all fitting methods described in Chapter 3 and compiled in Table 4.4. No error provided for rCDF, rCDF+K, and rCDF+K_{avg} due to computational limitations.

4.2 Maltose Minimal

MOPS based minimal media using maltose as the carbon source was used as the growth medium for two fluctuation tests. Each experiment was done with 50 samples to be selected for mutants and 10 samples to be used for final population counts, giving a total of 100 mutant plates and 20 population plates. The specific growth rate produced by maltose minimal medium as measured using CFU's is 0.88 ± 0.09 /hr, which corresponds to a doubling time of 47.8 ± 4.9 minutes. Each culture was seeded with an initial population of $(1.23 \pm 0.16) \cdot 10^3$ cells, which has a CV of 12.9%. The cells were grown for 7 hours and 6 minutes in both experiments, leading to an average final population of $4.5 \cdot 10^5$ cells with a standard deviation of $1.2 \cdot 10^5$ cells, giving a CV of 25.6%. Calculating the growth rate from the initial and final populations, along with the growth time, gives a doubling time of 50 minutes, which is within the standard deviation of the growth rate which was independently determined in control experiments.

The fluctuation test data coming from two separate runs means it is important to confirm that they are sufficiently similar to be combined. The 84% confidence intervals for their MLE estimated average number of mutations, $\hat{m}_{\text{MLE}}$, are $(0.586, 0.973)$ and $(0.845, 1.32)$ which overlap. Also, the p-value from the likelihood ratio test is 0.162. These two statistical measures support the decision to combine the two experiments into a single data set. The combined raw data for all population plates as well as the mutant plates is included in Table 4.3.

Delbrück's p_0 method gives an average number of mutations of $\hat{m}_{p_0} = 1.02$ with a 95% confidence interval of $(0.760, 1.28)$, which corresponds to a mutation rate of $\hat{\mu}_{p_0} = 2.25 \cdot 10^{-6}$ mutations per cell per generation with 95% confidence interval $\text{CI}_{95\%} = (9.79 \cdot 10^{-7}, 3.52 \cdot 10^{-6})$. When the data is fit using rSalvador's MLE system, an average mutation number of $\hat{m}_{\text{MLE}} = 0.908$ with a 95% confidence interval of $(0.710, 1.14)$ is found, which gives a mutation rate of $\hat{\mu}_{\text{MLE}} = 2.00 \cdot 10^{-6}$ with $\text{CI}_{95\%} = (9.02 \cdot 10^{-7}, 3.13 \cdot 10^{-6})$. The TSS fitting method gives $\hat{m}_{\text{TSS}} = 0.813$ with $\text{CI}_{95\%} = (0.641, 1.02)$ and $\hat{\mu}_{\text{TSS}} = 1.79 \cdot 10^{-6}$ with $\text{CI}_{95\%} = (8.12 \cdot 10^{-7}, 2.80 \cdot 10^{-6})$, which again agrees well with the MLE estimate. Plots of the experimental CDF along with the CDF's for $\hat{m}_{\text{MLE}}$ and $\hat{m}_{\text{TSS}}$ are shown in Fig. 4.6.

When all four phenotypic lag adjustment protocols described in Section 3.3 were applied to the maltose minimal data, they predicted that there was no phenotypic lag ($\hat{n} = 0$). See Fig. 4.7 for the fitting errors and estimated average number of mutations for the Koch and reduced CDF methods. All estimates are compiled in Table 4.4 and plotted together as a histogram with errors in Fig. 4.8.

Physically, the prediction of no phenotypic lag in maltose minimal is surprising consider-

Initial Population	Final Population	Number of Mutants				
	$4.7 \cdot 10^5$	2	0	0	2	4
	$4.3 \cdot 10^5$	6	0	0	87	3
$1.33 \cdot 10^3$	$4.5 \cdot 10^5$	2	0	2	1	0
$1.17 \cdot 10^3$	$3.6 \cdot 10^5$	4	0	1	5	0
$1.24 \cdot 10^3$	$3.9 \cdot 10^5$	0	0	0	2	292
$1.29 \cdot 10^3$	$3.9 \cdot 10^5$	1	4	0	1	3
$1.20 \cdot 10^3$	$5.4 \cdot 10^5$	1	0	3	4	0
$1.40 \cdot 10^3$	$6.2 \cdot 10^5$	11	0	1	4	2
$1.38 \cdot 10^3$	$6.5 \cdot 10^5$	1	2	0	1	2
$1.42 \cdot 10^3$	$5.5 \cdot 10^5$	0	0	1	1	8
972	$2.7 \cdot 10^5$	139	1	0	0	0
$1.06 \cdot 10^3$	$4.9 \cdot 10^5$	2	1	0	2	1
$1.02 \cdot 10^3$	$7.0 \cdot 10^5$	0	2	1	2	2
$1.02 \cdot 10^3$	$3.3 \cdot 10^5$	1	1	1	0	1
$1.14 \cdot 10^3$	$3.2 \cdot 10^5$	1	2	103	1	1
$1.22 \cdot 10^3$	$3.6 \cdot 10^5$	19	55	2	0	0
$1.32 \cdot 10^3$	$4.4 \cdot 10^5$	0	3	0	1	11
$1.49 \cdot 10^3$	$5.3 \cdot 10^5$	0	0	2	4	4
	$3.6 \cdot 10^5$	1	1	4	0	0
	$4.6 \cdot 10^5$	0	0	51	0	0
Mean ± SD	**Mean ± SD**	**Doubling Time**				
$(1.23 \pm 0.16) \cdot 10^3$	$(4.5 \pm 1.2) \cdot 10^5$	49.93 minutes				

Table 4.3: **Maltose minimal fluctuation test data.** Combined data from two fluctuation tests with *E. coli* NCM3722 in MOPS based minimal media with maltose as the carbon source and D-cycloserine as the selecting agent.

ing the nature of the system. I believe the prediction is potentially a result of experimental error because I had not quite perfected my plate counting protocol at the time of collecting the maltose minimal data, which may have resulted in some false positives. Fortunately, the presence of a small number of false positives would most affect the shape of the cdf at low mutant numbers, meaning the reduced CDF method would ignore the portion of the data with strong effects from this experimental error. Looking at the rCDF estimated average number of mutations, $\tilde{m}_{\text{rCDF}}$, for different estimated phenotypic lag lengths ((D) in Fig. 4.7) shows that for phenotypic lag of up to $\log_2(11) \approx 3.46$ generations, the re-

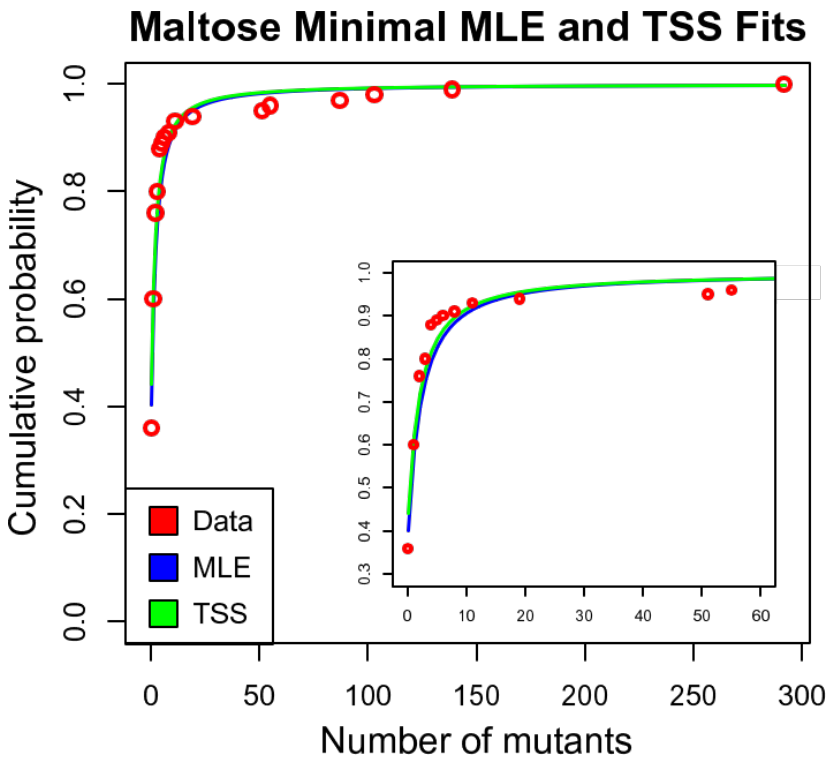

Figure 4.6: **Maltose minimal fluctuation test data with MLE and TSS fit.** The compiled data from the fluctuation tests performed with *E. coli* NCM3722 in maltose minimal media are plotted as a cumulative distribution with the CDF of best fit determined by rSalvador's maximum likelihood estimator (MLE) and the CDF of best fit as determined from the total sum of squares (TSS) method. Plot with smaller domain and range inlaid in plot covering full domain and range.

sulting estimate on m is within the 95% confidence interval[2] of the TSS fitted m (i.e. $0.641 \leq \tilde{m}_{\text{rCDF}} \leq 1.02$ for $0 \leq \tilde{n} \leq 3.46$). Consequently, it is likely reasonable to propose that the estimates given for the average number of mutants with no phenotypic lag adjustment are representative of the true number of mutants regardless of potential false positives in the data. Unfortunately a downside to this error is that it makes it impossible to say with any confidence what the phenotypic lag may be.

[2]The estimate for m with phenotypic lag length of $n = 2$ generations is outside of the confidence interval by 0.003, which is insignificant enough to consider it within the interval for the sake of the argument.

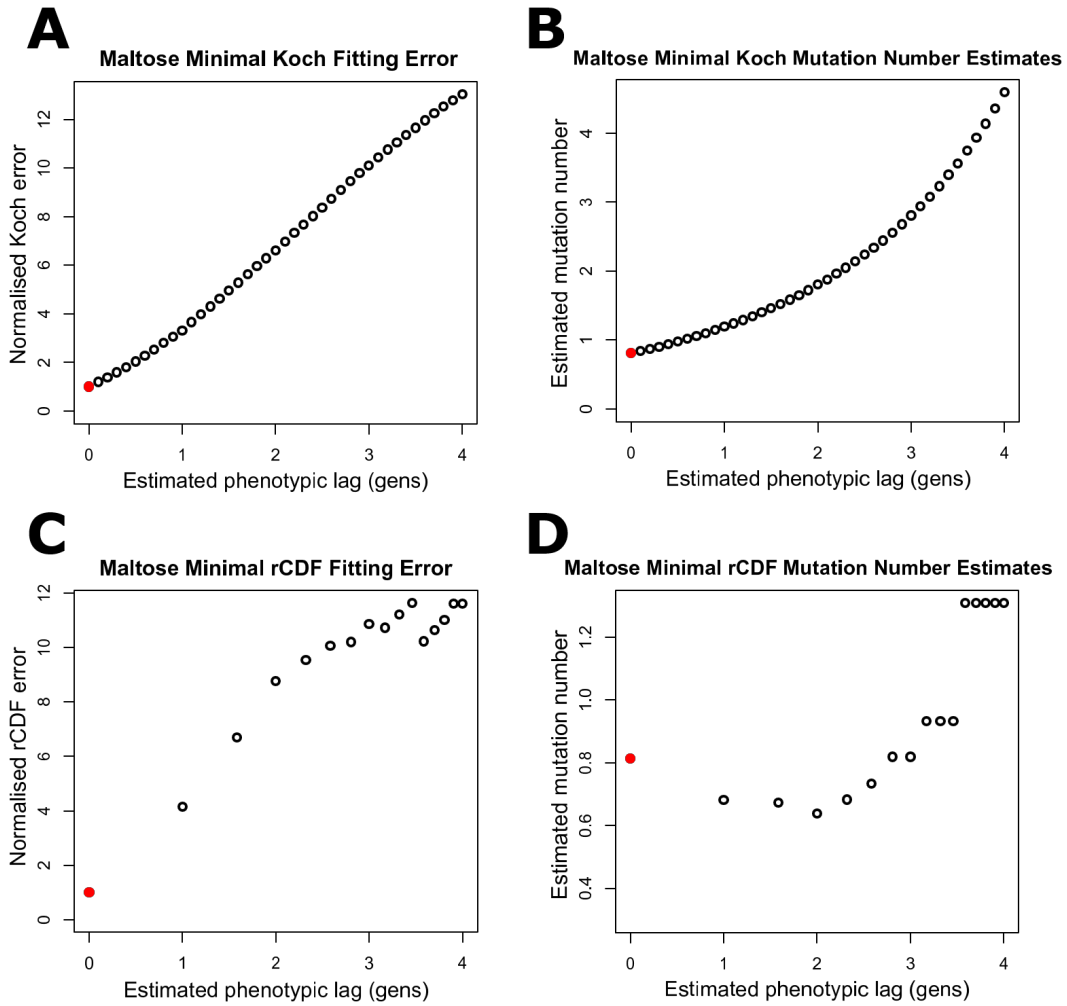

Figure 4.7: **Maltose minimal fluctuation test data Koch and reduced CDF fit details.** The Koch adjustment protocol (see Section 3.3.1) and reduced CDF (rCDF) adjustment protocol (see Section 3.3.2) applied to a set of fluctuation test data from *E. coli* grown in maltose minimal media. A) The Koch fitting error for a sequence of guessed phenotypic lags; the red point is the minimum. B) The Koch estimated average number of mutations for a sequence of guessed phenotypic lags; the red point is the chosen estimate. C) The rCDF fitting error for a sequence of guessed phenotypic lags; the red point is the minimum. B) The rCDF estimated average number of mutations for a sequence of guessed phenotypic lags; the red point is the chosen estimate.

Mutation rates of *E. coli* NCM3722 in maltose minimal					
Analysis protocol	Phenotypic lag length in generations (n)	Average number of mutations per culture (m) with 95% confidence interval	Average number of mutations per cell per generation (μ_{cell}) with 95% confidence interval ($\times 10^{-7}$)	Average number of mutations per base pair per generation (μ_{bp}) with 95% confidence interval ($\times 10^{-10}$)	Average number of mutations per genome per generation (μ_{genome}) with 95% confidence interval ($\times 10^{-3}$)
p_0	N/A	1.02 (0.760, 1.28)	22.5 (9.79, 35.2)	15.9 (6.93, 24.9)	7.45 (3.24, 11.6)
MLE	N/A	0.908 (0.710, 1.14)	20.0 (9.02, 31.3)	14.1 (6.38, 22.1)	6.62 (2.99, 10.4)
TSS	N/A	0.813 (0.641, 1.02)	17.9 (8.12, 28.0)	12.7 (5.75, 19.8)	5.93 (2.69, 9.28)
Koch	0	0.813 (0.641, 1.02)	17.9 (8.12, 28.0)	12.7 (5.75, 19.8)	5.93 (2.69, 9.28)
rCDF	0	0.813 (0.641, 1.02)	17.9 (8.12, 28.0)	12.7 (5.75, 19.8)	5.93 (2.69, 9.28)
rCDF+K	0	0.813 (0.641, 1.02)	17.9 (8.12, 28.0)	12.7 (5.75, 19.8)	5.93 (2.69, 9.28)
rCDF+K_{avg}	0	0.813 (0.641, 1.02)	17.9 (8.12, 28.0)	12.7 (5.75, 19.8)	5.93 (2.69, 9.28)

Table 4.4: ***E. coli* NCM3772 mutation rates in maltose minimal.** Mutation rates of *E. coli* NCM3722 grown in MOPS based minimal media with maltose carbon source (doubling time = 47.8 ± 4.9 minutes) as determined by the different analysis methods described in Chapter 3. Errors in the phenotypic lag adjustment protocols do not account for error in the estimated phenotypic lag length. The per base pair mutation rate was calculated by dividing the per cell mutation rate by 1413 base pairs. The per genome mutation rate was calculated by multiplying the per base pair mutation rate by $4.68 \cdot 10^6$ base pairs.

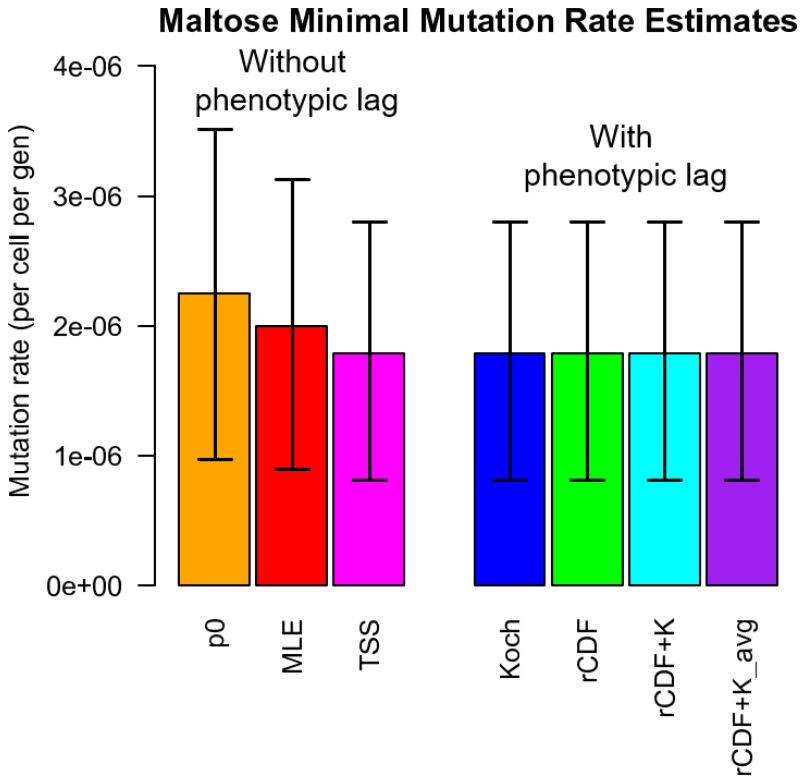

Figure 4.8: **Histogram comparing mutation rates calculated from different fitting protocols for *E. coli* NCM3772 in maltose minimal.** The per cell mutation rates with their 95% confidence intervals are plotted for all fitting methods described in Chapter 3 and compiled in Table 4.4.

4.3 Comparison

The primary objective of the research performed in this thesis was to determine if the spontaneous mutation rate in bacteria is growth rate dependent. This was addressed by performing fluctuation tests in different quality growth media while keeping the cells in the physiologically consistent state of exponential growth. In order to confidently compare the data between the fluctuation tests, the same initial and final populations were required. The average initial population in RDM glucose was 975 ± 99 while in maltose minimal it was $(1.2 \pm 0.2) \cdot 10^3$, which when compared have intersecting standard deviations, meaning they can be considered the same for our purposes. The average final population in RDM glucose was $(4.2 \pm 0.5) \cdot 10^5$ while in maltose minimal it was $(4.5 \pm 1.2) \cdot 10^5$, which also have intersecting standard deviations, meaning comparison of fluctuation test data between media can be done with confidence. Both quantitative comparison methods described in Section 3.4 will be used for the data not adjusted for phenotypic lag, while the confidence interval method will be used to compare estimates of the average number of mutations that account for phenotypic lag.

When the data is not adjusted for phenotypic lag, the likelihood ratio test developed by Zheng for rSalvador can easily be used. When applied to the RDM glucose and maltose minimal data, it gives a p-value of $2.28 \cdot 10^{-7}$, meaning it is unlikely that the mutation rates for each set of data are actually the same. In addition, the 84% confidence intervals of the maximum likelihood estimates (MLE) of the average number of mutations are $(0.219, 0.384)$ for RDM glucose and $(0.763, 1.07)$ for maltose minimal which clearly do not overlap, providing further evidence that the unadjusted mutation rate is likely to be different in each growth medium. The same holds true for the p_0 and total sum of squares estimates. Finally, from looking at Fig. 4.9 it appears that the distributions for each data set are describing different processes..

When the data is adjusted for phenotypic lag, it is not as easy to compare the results from the two different mediums. The only way is to observe the confidence intervals as determined through bootstrapping, but as explained in Section 3.3.4, the errors on adjusted data are difficult to find. Regardless, the errors on the fitted number of mutations are calculated and it is found that the 84% confidence intervals for the Koch estimated average number of mutations are $(0.531, 1.02)$ for RDM glucose and $(0.686, 0.955)$ for maltose minimal which are comfortably overlapping. Consequently, it is possible that the two sets of data have the same average number of mutations and therefore the bacteria may have the same mutation rate in both media when phenotypic lag is considered. In addition, RDM glucose is estimated to have 1.6 generations of phenotypic lag while maltose minimal is estimated to have no phenotypic lag. The implication is that when phenotypic lag is

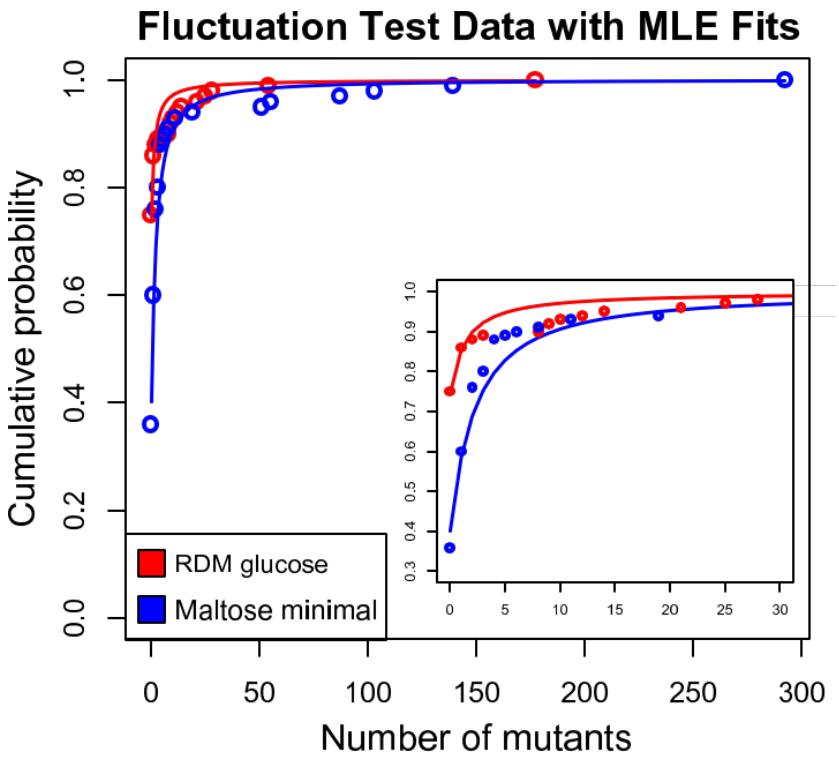

Figure 4.9: **RDM glucose and maltose minimal fluctuation test data with CDF's from the rSalvador MLE fits.** Fluctuation test data for *E. coli* NCM3772 grown in MOPS based rich defined media with glucose (doubling time = 23 minutes) and MOPS based minimal media with maltose (doubling time = 48 minutes). The CDF corresponding to the maximum likelihood estimate (MLE) of the average number of mutations per culture for each data set are displayed in the same colour as the respective data. Plot with smaller domain and range inlaid in plot covering full domain and range.

accounted for, the growth rate dependency shifts from the mutation rate to the phenotypic lag length. Finally, once again looking at the cumulative distribution for the data, but now with the CDF's for the adjusted fit, it appears that when adjusted for phenotypic lag, the processes being described are similar (Fig. 4.10).

In conclusion, if phenotypic lag is not accounted for, it appears that mutation rate is growth rate dependent and slower growing *E. coli* have higher mutation rates. But if phenotypic lag is accounted for, the mutation rate appears to be growth rate independent and instead the phenotypic lag is growth rate dependent. See Table 4.5 for a compilation of the results for each medium and Fig. 4.11 for a histogram comparing the mutation

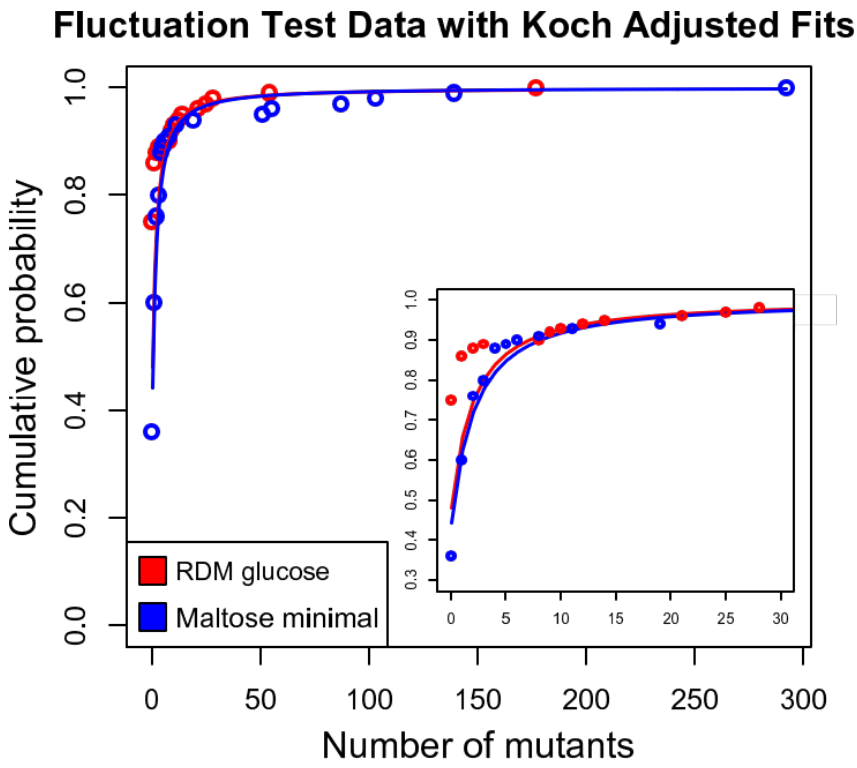

Figure 4.10: **RDM glucose and maltose minimal fluctuation test data with CDF's from the Koch phenotypic lag adjusted fits.** Fluctuation test data for *E. coli* NCM3772 grown in MOPS based rich defined media with glucose (doubling time = 23 minutes) and MOPS based minimal media with maltose (doubling time = 48 minutes). The CDF corresponding to the Koch phenotypic lag adjustment estimate of the average number of mutations per culture for each data set are displayed in the same colour as the respective data. For RDM glucose there is an an estimated phenotypic lag length of $n = 1.6$ generations while for maltose minimal data there is no estimated phenotypic lag. Plot with smaller domain and range inlaid in plot covering full domain and range.

rates. Without further experimentation it is not possible to distinguish between these two scenarios. Possible experiments that could illuminate which is the more likely of the two scenarios are discussed in Section 5.2.3.

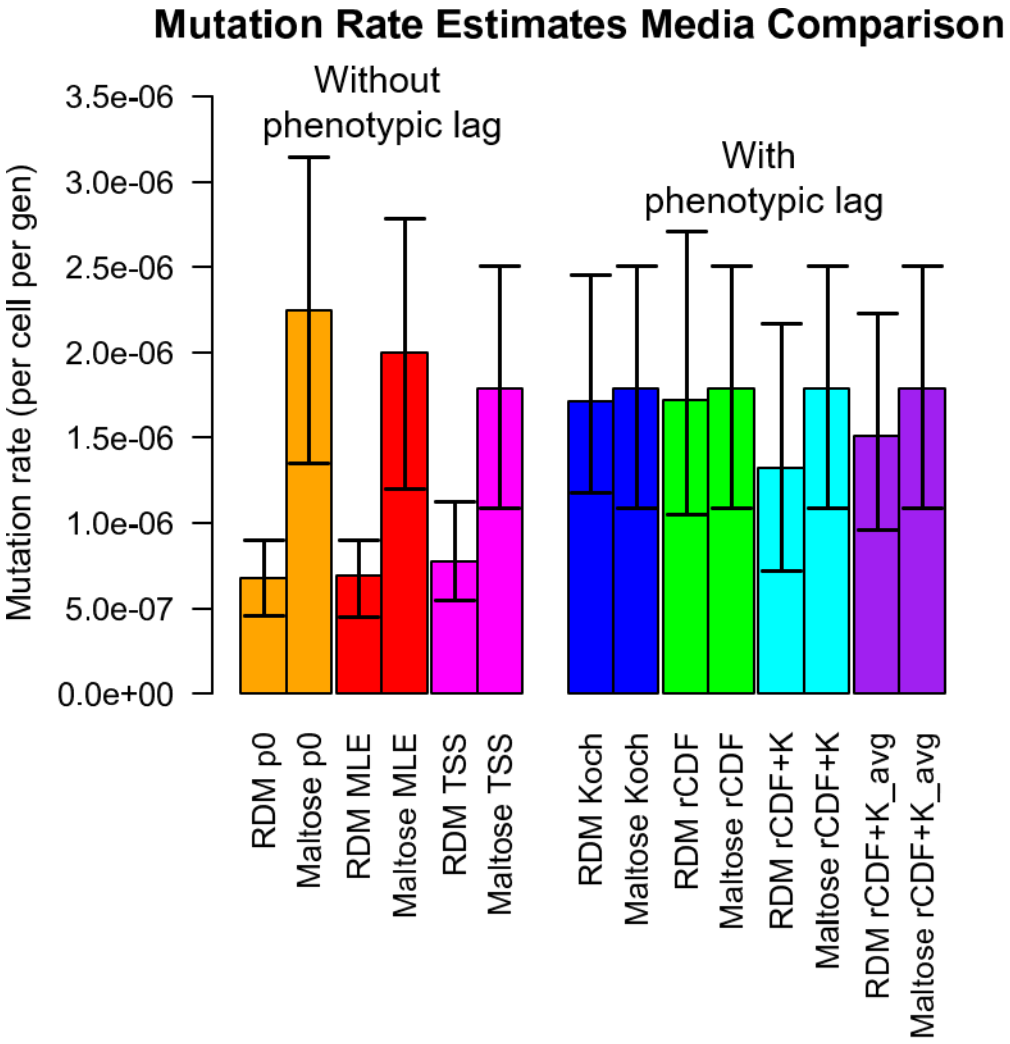

Figure 4.11: **Histogram comparing mutation rates calculated from different fitting protocols for *E. coli* NCM3772 in RDM glucose and maltose minimal.** The per cell mutation rates with their 84% confidence intervals are plotted for all fitting methods described in Chapter 3 and compiled in Table 4.5.

Mutation rates of *E. coli* NCM3722						
	RDM glucose (doubling time = 23.2 ± 0.5 minutes)			Maltose minimal (doubling time = 47.8 ± 4.9 minutes)		
Analysis protocol	Phenotypic lag length in generations (n)	Average number of mutations per culture (m) with 84% confidence interval	Average number of mutations per cell per generation (μ_{cell}) with 84% confidence interval ($\times 10^{-7}$)	Phenotypic lag length in generations (n)	Average number of mutations per culture (m) with 84% confidence interval	Average number of mutations per cell per generation (μ_{cell}) with 84% confidence interval ($\times 10^{-7}$)
p_0	N/A	0.288 (0.207, 0.369)	6.79 (4.60, 9.00)	N/A	1.02 (0.834, 1.21)	22.5 (13.4, 31.5)
MLE	N/A	0.294 (0.219, 0.384)	6.93 (4.54, 9.03)	N/A	0.908 (0.763, 1.07)	20.0 (12.0, 27.9)
TSS	N/A	0.329 (0.248, 469)	7.77 (5.49, 11.3)	N/A	0.813 (0.686, 0.955)	17.9 (10.9, 25.1)
Koch	1.6	0.724 (0.531, 1.02)	17.1 (11.8, 24.6)	0	0.813 (0.686, 0.955)	17.9 (10.9, 25.1)
rCDF	2	0.730 (0.468, 1.13)	17.2 (10.5, 27.1)	0	0.813 (0.686, 0.955)	17.9 (10.9, 25.1)
rCDF+K	1.58	0.559 (0.322, 0.910)	13.2 (7.22, 21.7)	0	0.813 (0.686, 0.955)	17.9 (10.9, 25.1)
rCDF+K_{avg}	1.58	0.638 (0.432, 0.928)	15.1 (9.64, 22.3)	0	0.813 (0.686, 0.955)	17.9 (10.9, 25.1)

Table 4.5: ***E. coli* NCM3772 mutation rates in RDM glucose and maltose minimal.** Mutation rates of *E. coli* NCM3722 grown in MOPS based rich defined media with glucose (RDM glucose) and minimal media with maltose (maltose minimal) as determined by the different analysis methods described in Chapter 3. 84% confidence intervals are given for easy comparison between media. Errors in the phenotypic lag adjustment protocols do not account for error in the estimated phenotypic lag length. The per base pair mutation rate was calculated by dividing the per cell mutation rate by 1413 base pairs. The per genome mutation rate was calculated by multiplying the per base pair mutation rate by $4.68 \cdot 10^6$ base pairs.

4.4 Difficulties

There were two major difficulties in the experimental set-up. The first difficulty was determining the growth rates of the cells. In particular, it was found that the growth rates as determined by colony forming units (CFU) at low density were systematically higher than those found with optical density (OD) at high density. Which one to use in order to determine how long to grow the cells for in a fluctuation test was an obvious choice because how the growth rate was determined using CFU's mimicked the growth in a fluctuation test, but there remained the nagging question of why the discrepancy? It was found that the growth rates that were determined from CFU's were similar to those determined from OD's while measuring the same high density cultures (found while collecting data for Fig. 2.4). Accordingly, considering the cultures were at a much lower density for the independent CFU growth curves (Fig. 2.3), then this suggests that the difference in cell densities is a likely cause for the discrepancy in growth rates. At higher densities, it is possible that there are crowding effects that either reduce the growth rate or increase the death rate[3] of the cells [33, 173]. Both these effects would give the appearance of a decreased growth rate, making high density growth appear slower. The difference in growth rates was most significant in slow growing media such as αKG minimal, complicating the search for an optimal slow growth medium. I suspect the greater effect in slow growing cells is a repercussion of the cell density being much higher in a slow growing culture of the same OD as a fast growing culture (Fig. 2.4), meaning the crowding effects are likely to be more pronounced at lower OD's.

The greatest difficulty faced during the fluctuation tests was making sure that the variance in the final population wasn't exorbitant. The problem primarily comes from the fact that there is variance in cell-to-cell growth rates and there is variance in the initial populations, which is then extenuated by exponential growth. If one lets the cell-to-cell coefficient of variation (CV) in the growth rate be σ_λ and the CV in the initial population be σ_{N_0}, then one will end up with a final population CV of,

$$\sigma_{N_f} = \sqrt{\sigma_{N_0}^2 + (t\langle\lambda\rangle\sigma_\lambda)^2}\,, \tag{4.1}$$

where $\langle\lambda\rangle$ is the average specific growth rate and t is how long the culture is grown for [195]. Assuming a σ_λ of 10% [190] and plugging in my parameters from the fluctuation

[3]Simply having dead cells present, which are measured by the OD but not the CFU, will not result in the growth rate discrepancy, instead there needs to be some death rate that is effectively decreasing the measured growth rate (i.e. $N_t = N_0 e^{(\lambda-\delta)t}$ where δ is a death rate and $(\lambda-\delta)$ is now the slope of a growth curve on a log-linear plot).

test, for RDM glucose one gets $\sigma_{N_f} = 64.7\%$ and for maltose minimal $\sigma_{N_f} = 63.6\%$. These two coefficients of variance should be treated as upper bounds on the amount of error one can expect in an experiment. The actual measured final population coefficients of variance being 11.4% and 25.6% for RDM glucose and maltose minimal respectively gives confidence that the growth was well controlled in the experiment. In fact, by rearranging Eq. (4.1) and plugging in the measured values for σ_{N_f}, it can be estimated that the cell-to-cell coefficient of variances were 0.8% for RDM glucose and 3.5% for maltose minimal. Regardless of this success, fluctuation tests are traditionally done by allowing the cells to exhaust the carbon source, resulting in a low variance in the final population [106, 209]. Because of this, the variance in my maltose minimal final populations can appear quite high for a fluctuation test. Furthermore, high variance in the final population can potentially cause problems for fitting because when compiling the data from each sample in a fluctuation test, one is implicitly assuming that they all have the same final population meaning they all had the same number of doublings or opportunities for a mutation to occur. Zheng has explored the effects of final population variance, noting that a CV below 20% should not lead to many issues [209]. To account for data with high variance in the final population, Zheng developed a distribution and fitting method that claims to account for the effects, called the B_0 method [210, 209]. When the B_0 method is applied to my RDM glucose and maltose minimal data, one gets $m = 0.294$ and $m = 0.919$ respectively, which are very close to what is given by the normal MLE fitting ($m = 0.294$ and $m = 0.908$). Considering it has been shown that the normal MLE methods give underestimates on m when there is variance in the final populations [198] and Zheng's B_0 method has been shown to successfully adjust for this [209], it can be argued that the fact that the B_0 and MLE estimates agree so well suggests that the normal MLE method is not giving an underestimate in this particular case. Zheng argued that increasing the initial population will lead to less variance at the end of the experiment [209]. I took this into consideration and increased the initial population from approximately 200 cells in my early trials to approximately 1000 cells in my final experiments. In addition, I increased the volume of media during growth from 200μL in my early trials to 500μL in my final trials. The increase in media volume was to combat the potential evaporation of media during growth which would lead to variance and faulty dilution calculations. The combination of these two efforts helped reduce the variance in my final populations in RDM glucose and maltose minimal. Unfortunately, the large variance in my acetate minimal and α-ketoglutarate minimal trials remained. Why exactly the variance in these slow growing media was particularly high is unclear, but I suspect it is primarily due to near unavoidable evaporation during an approximately 19 hour growth period, and the elevated opportunity for a fast growing mutant to take over the population.

4.4.1 α-Ketoglutarate Minimal

The original goal was to perform fluctuation tests in three different growth media with three noticeably different growth rates. At first, for the slowest growth medium a MOPS minimal media with acetate as the carbon source was explored, but was quickly traded out for α-ketoglutarate (α-KG) as the carbon source. Acetate was abandoned because there were noticeable growth rate changes during exponential growth with *E. coli* MG1655 and large variances in the final population counts with *E. coli* NCM3722. These phenomena were thought to be a repercussion of the idea that some *E. coli* strains have a propensity to gain a mutation that makes them better at growing in acetate [157, 183], possibly similar to that found with glycerol [200].

In terms of growth rates determined using CFU, α-KG minimal media had what I was aiming for with a doubling time greater than 100 minutes. Unfortunately, how drastically different the doubling times were when measured by CFU vs OD (106 minutes vs 277 minutes) raised suspicions, although a large variation between the CFU and OD measured growth rates appeared to be an inherent feature of slow growth media. A long doubling time was desired because it guaranteed a physiology drastically different than that of RDM glucose and maltose minimal. Ultimately, the same difficulties of high variance in the final populations that plagued the preliminary acetate experiments ended up plaguing the α-KG minimal experiments, as seen in Table 4.6. Furthermore, the maximum coefficient of variance in the final populations according to Eq. (4.1) would be 73.8% for α-KG minimal, while the measured CV was 61.5%. All things considered, this isn't bad, but when the data is being used for a fluctuation test and the CV's in the final populations of RDM glucose and maltose minimal are significantly lower than the estimated error, this result becomes insufficient for its purposes. I believe that the most promising approach towards fixing the issue of variance in the final populations in slow growth media is to increase the volume of growth medium per sample even further. The downside to this is that it would raise the costs, difficulty, and resource use of each experiment.

Initial Population		**Final Population**	
		472388	268657
948	932	432090	490299
732	896	454478	705224
744	1028	360448	326866
764	988	387313	317910
852	940	389552	355970
764	912	539552	436567
864	1140	622388	526119
808	1048	647015	355970
		550746	463433
Mean ± SD		**Mean ± SD**	
898 ± 118		$(2.2 \pm 1.3) \cdot 10^5$	
CV = 13.2%		**CV** = 61.5%	

Table 4.6: **α-ketoglutarate minimal fluctuation test population data.** Data is from two separate attempts at a fluctuation test in MOPS based minimal media with α-ketoglutarate as the carbon source. The average growth time across the two experiments was 18 hours and 28 minutes, giving an average doubling time of 139 minutes.

Chapter 5

Conclusion

Evolution is the theory which underpins all of modern biology and mutations are the mechanism that drives evolution [67]. Bacteria are the oldest, most abundant life form on the planet, from which every other living thing descends [192]. Accordingly, the study of bacterial evolution is important both for understanding the fundamental principles of bacteria and potentially gaining insight into the workings of all organisms. Throughout history, the study of evolution has often been done with a quantitative approach due to the nature of genetics [114, 67, 54]. The study of bacteria has also lent itself to a quantitative approach due to bacteria's relative simplicity, especially in their behaviour during balanced growth [150]. Luria and Delbrück had the insight to combine these two fields and design a quantitatively inspired experiment to determine the method and rate with which bacteria mutate [106]. Their experiment, the fluctuation test, relies on several simplifying assumptions, one of which being that the cells are growing exponentially and by extension have consistent physiology. Unfortunately, in their experiments they failed to implement the special care needed for consistent physiology[1], and this oversight has continued up to present. My research set out to remedy this problem by growing cells in the exponential phase for the entirety of a fluctuation test. Doing so results in mutation rates consistent with the traditional order of magnitude of 10^{-10} to 10^{-9} mutations per base pair per generation [101, 197, 116], but with a new level of confidence that the determined mutation rate is specific to the exponential phase of growth. I used this methodology to also observe if the mutation rate exhibits any intrinsic growth rate dependence. From these studies it was found that when one does not consider the effects of phenotypic lag, the mutation rate appears to be anticorrelated with the growth rate (i.e. the mutation rate

[1] To be fair to them, bacterial physiology was not a well developed field at the time of their first study.

is higher for slower growing cells). But when the effects of phenotypic lag are considered, the growth dependence in the mutation rate is lost, and instead the phenotypic lag appears to be directly correlated with the growth rate (i.e. the phenotypic lag is longer in faster growing cells). See Table 5.1 for a summary of the results.

Mutation rates of *E. coli* NCM3722				
Growth medium	Specific growth rate (λ) in per hour and doubling time (τ) in minutes $\pm$ standard deviation	Average number of mutations per base pair per generation without phenotypic lag (MLE) (μ_{bp}) with 84% & 95% confidence intervals ($\times 10^{-10}$)	Phenotypic lag length in generations (range from different protocols) (n)	Average number of mutations per base pair per generation with phenotypic lag (Koch) (μ_{bp}) with 84% & 95% confidence interval ($\times 10^{-10}$)
RDM glucose	$\lambda = 1.80 \pm 0.04$ $\tau = 23.2 \pm 0.5$	4.90 $(3.21, 6.39)_{84\%}$ $(2.90, 7.36)_{95\%}$	$1.58 - 2$	12.1 $(8.35, 17.4)_{84\%}$ $(6.98, 19.5)_{95\%}$
Maltose minimal	$\lambda = 0.88 \pm 0.09$ $\tau = 47.8 \pm 4.9$	14.1 $(8.49, 19.7)_{84\%}$ $(6.38, 22.1)_{95\%}$	0	12.7 $(7.71, 17.8)_{84\%}$ $(5.75, 19.8)_{95\%}$

Table 5.1: **Summary of experimental results from fluctuation tests with *E. coli* NCM3722 in RDM glucose and maltose minimal.** Summary of the growth rates, the maximum likelihood estimated (MLE) per base pair mutation rate without accounting for phenotypic lag, the estimated phenotypic lag length, and the Koch estimated per base pair mutation rate with accounting for phenotypic lag. For the mutation rates, the 84% and 95% confidence interval are provided for easy comparison between fitting protocols and media. The fluctuation tests were performed in two experiments with 50 cultures each, giving a total of 100 samples in each medium; D-cycloserine was used as the selecting agent.

Within the two media used for growth during the fluctuation tests in this thesis (RDM glucose and maltose minimal), there are clear differences in the physiologies of the bacteria grown. The high growth rate of RDM glucose results in a physiology unique to especially fast growing cells, which can only be achieved in optimal conditions. Conversely, the lower growth rate of maltose minimal can be achieved through a variety of conditions [153]. One characteristic of the physiology at both growth rates is the presence of multiple DNA replication forks in the cell at some time during growth [36]. For cells growing exponentially in RDM glucose, the number of forks is greater than one for the entirety of the cell's life cycle, even reaching upwards of four forks late in life. On the other hand, in maltose minimal there is a short period of time when there is no DNA replication being performed and the number of forks never exceeds two (See Fig. 1.21). A consequence of this is that there will be more chromosomes in cells growing in RDM glucose than in maltose minimal. Accordingly, effective polyploidy will play a role in both media, but will likely be more substantial in RDM glucose. Furthermore, having more or less DNA replication being performed in a cell at any given time could have consequences if all the proteins that participate in the action are not proportionally scaled. Since most mutations arise during DNA replication [152], one may extrapolate the potential for a correlation between growth rate and mutation rate.

Physiological differences not only have the potential to directly affect the mutation rate, but to also affect the way in which we measure it. For fluctuation tests, one must always choose a specific mutant phenotype to select for, and the specific characteristics of these phenotypes are also often coupled to the cell's physiology. For the experiments in this thesis, it is cycloserine resistance that is selected for, but there are a number of ways this resistance can couple to physiology, causing variable phenotypic lag. If one assumes that the CycA permease proteins are an unregulated protein, making it part of the "P" class in Fig. 1.26, then the concentration of these proteins will be higher in slow growing cells than in fast growing cells [160]. Also, bacteria cells growing at different growth rates have different sizes, as made clear by Figures 1.20 and 1.4. Combining the ideas of the growth rate dependent CycA permease protein concentration and cell size, it is possible that the density of CycA permease proteins on the surface of the cells will be significantly lower in fast growing cells. This could be an explanation for why the cycloserine half inhibition concentration is so much higher in RDM glucose than in the slower growing cells, as seen in Fig. 2.7. The difference in CycA permease protein density could also have a significant affect on phenotypic lag.

5.1 Implications

On first inspection, it appears that slow growing bacteria may have higher mutations rates than fast growing bacteria. Further exploration is certainly required to say this definitively, but it is interesting to explore the consequences regardless.

Mutations can be beneficial, deleterious, or neutral. Most commonly, mutations are neutral [71, 89], meaning they have very minimal affect on the functioning of the cell. On the other hand, when beneficial or deleterious mutations that change the fitness or growth rate of the cell arise, there can be drastic consequences. If there is a relationship between growth rate and mutation rate, mutations that cause changes in growth rate could result in a feedback loop. If growth rate and mutation rate are positively correlated, a positive feedback loop would be established. The result would be faster growing cells having a greater ability to explore mutant phenotypes due to more mutations per generation and more generations per time. This has the potential to allow for a cell to go from slow growth to fast growth very quickly, but it also has the potential to result in successful lineages dying off shortly after being established. It would also mean that slower growing cells would have lower mutation rates, essentially trapping them in a less competitive state with fewer opportunities to seek improvement. These combined facts would lead one to believe that a positive correlation would not be a particularly beneficial or stable system. Alternatively, a negative correlation between mutation rate and growth rate would result in a negative feedback loop which has the potential to be more stable. If slower growing cells have a higher mutation rate, it allows these cells to take more risks in their search for improved fitness. Conversely, by having a lower mutation rate during fast growth, the cells are more likely to maintain their success once they get it. Combining these facts makes a negative correlation more appealing in terms of maximising fitness within a population of bacteria. Finally, not having any correlation between mutation rate and growth rate would likely be the most stable system. Additionally, it is often said that organisms need the perfect balance between mutation and DNA repair in order to propagate into the future [46], which would be easiest to manage if the mutation rate was decoupled from growth rate.

Bacteria's growth rates are primarily determined by their environment. Consequently, if the mutation rate is growth rate dependent, then a change in environment can change the bacteria's mutation rate. Furthermore, mutations are what cause change in an organism's genome, and it is through the accumulation and selection of these changes that living things evolve and diversify [54, 66, 67]. As a result, one may expect that different environments

can promote a quicker genetic diversification of bacteria than others[2]. Over large time scales, such as are commonly studied in evolutionary biology, this coupling of environment to mutation rate has the potential to have major consequences.

It is well known that certain strains of bacteria can be very dangerous to humans [187]. To combat the dangers of bacteria, humans have developed many antibiotics which can treat dangerous infections [86]. Unfortunately, the selective pressure that the abundant use of antibiotics puts on bacteria has led to a popular rise of mutated strains that are resistant to common treatments [139, 44]. As my research specifically pertains to the characteristics of bacterial evolution towards antibiotic resistance, it seems reasonable to suggest that there are potential applications to the work. The applications are unlikely to be direct though, considering bacterial growth within an animal is far more complicated than in a test tube [91, 164], especially when so much care is taken to keep the cells in balanced growth. Regardless, different infections can grow at different rates [73, 164], so with simplifying assumptions and an understanding of how mutation rates relate to growth rates, one may be able to infer how likely an infection is to develop a resistance to an antibiotic.

5.2 Future Work

5.2.1 Changes to Experiment Protocol

The study of bacterial mutation rates through fluctuation tests is predicated on a number of assumptions, as with any research that uses a model. For fluctuation tests, it is the assumptions laid out in Section 1.5.2 that are generally used to build the model that allows for a quantitative analysis of the experimental results. Though extra care was taken to satisfy more of these assumptions than is traditional, there are still inevitably discrepancies between the experimental conditions and the model. It is likely that the most dominant effect is that of phenotypic lag, which has been discussed in detail and attempted to be accounted for during the analysis of the data. Other possible discrepancies could be caused by cell death, imperfect plating efficiency, unknown effects during the potential short lag phase at the beginning of growth, and mutations either causing changes to fitness or mutability. Of particular interest is the potential for the *cycA* mutation to cause a fitness change in media with amino acid supplements, as alluded to in Section 2.3. I attempted to account for these effects, but a more careful job could likely have been done to better

[2]In many ways this is what led me to this research in the first place.

ensure no effects on growth rate by removing alanine and glycine in addition to serine from the rich defined growth medium. Furthermore, control experiments where selected mutants are cultured and their growth rates measured in different media could illuminate how significant the effects of the mutation are, especially in the presence of alanine and glycine. Fortunately, the presence of alanine and glycine in the growth medium should not have any major affects on selection because the mutants are grown on M9 minimal with glucose plates during the selection phase.

In experimentation, it is a scientists job to do everything possible to remove unwanted interactions between the environment and the system of study in order to make the experiment as significant and easily reproduced as possible. In the case of my fluctuation test methodology, by keeping the bacteria purely in exponential phase, I in theory make more easily modelled and reproduced results[3] because the cells' physiology is then primarily a product of the growth rate instead of the details of its growth medium (see Section 1.6.1). That being said, there are still several places in which the methodology can be improved and further validated. The most clear place for improvement is in the final population numbers. If a way could be found to reduce the variance in these numbers, especially in slow growing cells, one would be able to better explore the relationship between the growth rate and the mutation rate by performing fluctuation tests at a wide range of growth rates. The most obvious first step in trying to reduce this variance is by further increasing the initial population number [209]. This comes with the risk of potentially having a mutant in the inoculum of some of the cultures though, so finding the right balance of large initial inoculum and low probability of introducing a mutant is essential. Another way final population numbers could be altered in an attempt to improve fluctuation test results is by increasing the final population, and by extension the expected number of mutations, m. Many analysis techniques work best for higher values of m than found in this thesis [58]. Furthermore, having a higher average number of mutations would result in more data of high mutant numbers which would be beneficial for the reduced CDF phenotypic lag adjustment protocol because there would be more data to fit the reduced CDF to. Everything considered, I think exploring more final population numbers in order to find an optimal number which gives at least some zeroes and not too many jackpots in all media could be very beneficial.

The next largest problem with my methodology lays in human error. These are long, labour intensive experiments that can lead to mistakes and back pain. Having more samples would make fitting better and likely also decrease variance in the final population, but it comes at a cost. In this regard, it would be beneficial if there was a way to do my version

[3]Unfortunately my methodology is far more difficult than the traditional fluctuation test, so this effect may be cancelled out by the effects of human error.

of fluctuation tests, but with a larger number of samples while not increasing the work load. To increase sample size and decrease labour, people often use multi-well plates, but this makes it very difficult to maintain physiological consistency because the wells are so small, making oxidation difficult, and they must be warmed in an air incubator, which are known to be inconsistent [168]. Using a turbidostat is also a common way to decrease labour, though in this case I don't think it would decrease it significantly and it would be subject to the issues discussed in Section 2.1. A popular tool that is very versatile is using fluorescent proteins to probe the inner workings of a cell [31]. In the case of a fluctuation test, I suspect one could design a cell which has a gene in its chromosome that when mutated, the cell makes a fluorescent protein. Flow cytometry could then be used to determine the number of cells in a culture expressing the proteins, and by extension have a mutation. The most obvious way this could be done would be by having a broken fluorescent protein gene that "turns on" when fixed through mutation [30, 8], but this would most likely be biased towards point mutations and, depending on the location and nature of the "broken" portion of the gene, could cause a growth defect. This difference in growth rate between normal and mutant cells can be adjusted for by either determining the average mutant's growth rate experimentally, or by fitting for a growth rate difference while fitting for mutation rate [92, 201, 210], but the bias towards a specific type of mutation is much more difficult to address[4]. For these reasons, one should be careful exploring this path. As in all experimental biology, one must find the right balance between use of advanced technology and reducing the amount of unknown mechanical error.

5.2.2 Improved Analysis

In this thesis I explored two novel implementations of protocols that adjust for phenotypic lag and developed two novel hybrids of the two methods.. That being said, all four protocols have clear issues, the most blaring of which being that there is no easy way to determine a fitting error on the estimate of the phenotypic lag length, and by extension the associated average number of mutations. Consequently, it is difficult to know how confident one can be in the results of the adjustment when applied to experimental data as well as making it difficult to compare the results between fluctuation tests. If a method could be developed for approximating these errors, the adjustment protocols would immediately become much more useful. Furthermore, the standard deviation in the adjusted fitting results when

[4]Potentially a much more clever way to have mutations of a specific gene cause the expression of fluorescent proteins could be found, but none immediately come to mind.

applied to simulated data of sample sizes similar to that pursued by researchers[5] points towards potential instabilities in the algorithm. It would be beneficial to explore the stability of the algorithms and try to find ways to decrease the variability in their estimates. Finally, it is not clear how well the phenotypic lag adjustments work for data with different average numbers of mutations. Exploring if the system works best for specific ranges of mutation number could help influence experimental design and analysis protocols while also giving further insight into potential pitfalls in the adjustment algorithms.

A number of data analysis techniques were discussed in this thesis, but there are other potential approaches that can be explored. Some possible techniques are more advanced hybridisations of the Koch and reduced CDF protocols, the most obvious of which is where one would use rCDF to predict phenotypic lag if it is less than two generations and Koch if it is greater than or equal to two generations because from Figures 3.8 and 3.10 it is clear that rCDF is slightly better at predicting short lag and Koch is slightly better at predicting long lag. Once a better prediction for phenotypic lag length is made, the most obvious choice to estimate the average number of mutants would be with the rCDF+K average technique due to its apparent accuracy seen in Fig. 3.12. There are possibly more creative combinations and alterations to the already developed analysis techniques that can also be done. In addition, there has been a decent amount of mathematics developed to describe the actions and consequences of phenotypic lag [201, 4, 170]. It would be great if any of these models, most of which are extensions of the Lea-Coulson model, could be used to develop a fitting technique. Another notable method for fitting fluctuation test data while accounting for phenotypic lag has been indirectly alluded to recently by Carballo-Pacheco et al. (2020) [29]. In their paper they run a simulation with parameters mimicking that of a specific experiment and show that a discrepancy between mutation rates from fluctuation tests and sequencing data can be explained by the presence of phenotypic lag as a consequence of protein dilution. They also develop a highly successful statistical method for determining if an experiment has phenotypic lag present by comparing it to simulations. I believe it is reasonable to imagine taking this one step further and fitting data directly to large sample size simulations in order to get estimates on parameters. This does come with downsides though, the first of which is how computationally expensive it would be to run so many simulations, especially with the quantity of points needed for them to be considered a satisfactory representation of the theoretical case. Second is that it would add another variable to fit for, in the form of the protein number. Third is that it hard-wires the assumptions of the simulation into the fitting mechanism, which is generally considered less convincing than fitting to a theoretical model like that of Lea and Coulson. Arguably

[5] Actually even 100 samples is significantly more than most people perform; more commonly people do 20-30 samples [137]

the most debilitating of these issues is the first one, but if one has the computational power and time, then it may have the potential to be a very advantageous methodology.

5.2.3 Future Experiments

One experiment that can be done to test the significance of my results is to do some fluctuation tests using the traditional methodology of growing to saturation, but with all the same conditions (strain, selecting agent, inoculum size, media, final population) as my experiments. This would allow one to test the assumptions of Section 2.1 and better determine the potential inconsistencies of previous fluctuation tests. It could also allow for a rough prediction on how much the mutation rate changes during deceleration and stationary phase.

Another experiment that could be done immediately to test and likely add to the significance of my results is to experimentally determine the length of phenotypic lag in my system. This could be done in a number of ways, the simplest of which is growing cells in exponential phase and introducing a mutagen (such as UV light or some type of DNA targeting antibiotic like mitomycin C) to the environment at a defined time [29]. Then at discrete time intervals, samples of the culture are introduced to cycloserine and plated. Accounting for growth, one can then look for a spike in the number of mutants a period of time after the mutagen was introduced, giving an indication of how long the phenotypic lag was. The biggest advantage to this experiment is it would lead to the determination of whether phenotypic lag is present in my system and if it is growth rate dependent, allowing for the distinction between the scenarios of "mutation rate is growth rate dependent" and "mutation rate is growth rate independent, but phenotypic lag is growth rate dependent" (as mentioned in Section 4.3). Another benefit to this experiment is that the phenotypic lag adjustment protocols would no longer have to fit for phenotypic lag, as it can be manually input. The effects of knowing the length of phenotypic lag prior to fitting were studied on the simulated data used throughout Section 3.3 and gave promising results (Fig. 5.1).

A practical study that could shed further light on the relationship between bacterial physiology and mutation rate would be to perform the same fluctuation tests as done in this thesis, but while the cells are under stress so that the SOS response is turned on. The SOS response could be turned on by including a small amount of mitomycin C (MMC), which is a DNA targeting antibiotic, to the growth media. The goal would be to add just enough MMC to induce the SOS response, but not enough to affect the growth rate. Because the SOS response is mutagenic in nature (due to its up regulation of error prone DNA polymerases), the result would be an increased mutation rate which would give insight

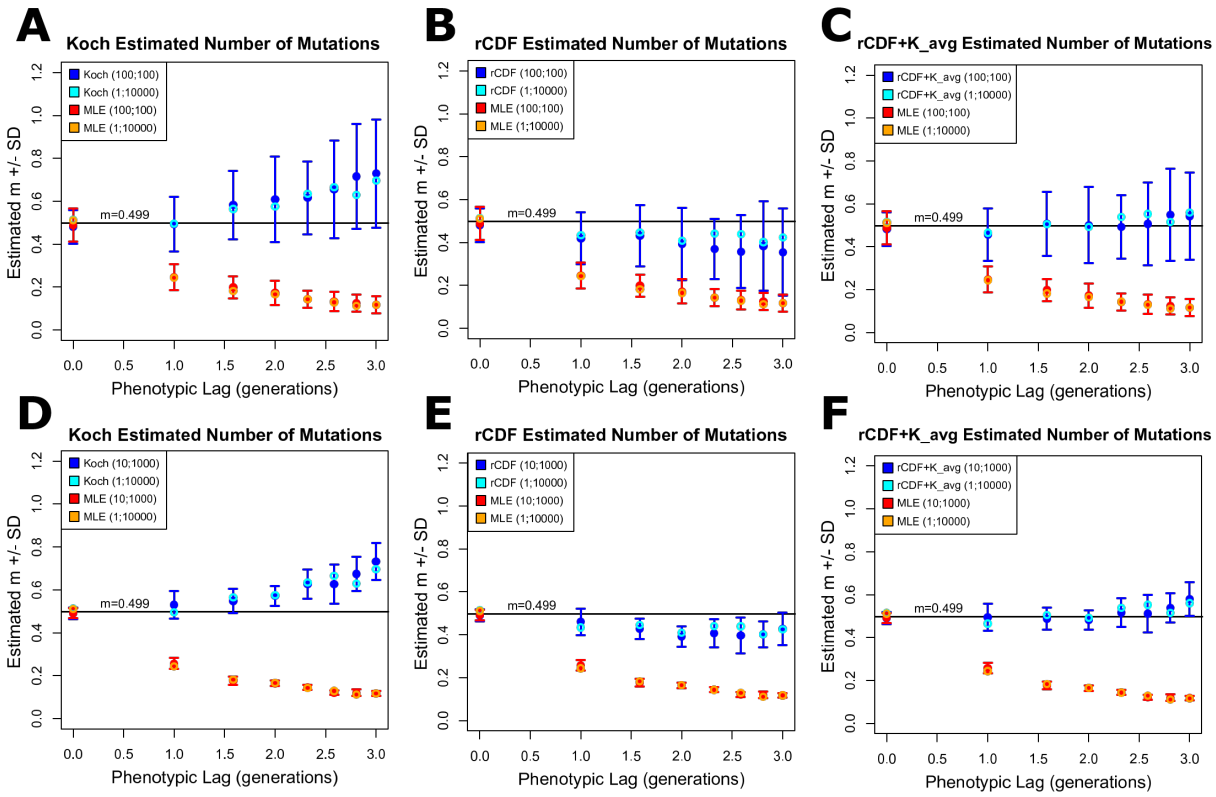

Figure 5.1: **Estimations of the average number of mutations in simulated data from the Koch, reduced CDF and rCDF+K_{avg} protocols when the phenotypic lag length is known.** All data simulated with an average number of mutations $m = 0.499$. The points on the plots are the average estimated number of mutations $\pm$ one standard deviation for several simulated phenotypic lag lengths. Every plot includes the maximum likelihood estimates (MLE), which clearly decrease with increasing phenotypic lag length. A) Koch adjusted fit applied to 100 simulations of 100 cultures (100;100) and 1 simulation of 10000 cultures (1;10000). B) rCDF applied to 100;100 and 1;10000 simulations. C) rCDF+K_{avg} applied to 100;100 and 1;10000 simulations. D) Koch adjusted fit applied to 10 simulations of 1000 cultures (10;1000) and 1;10000. E) rCDF applied to 10;1000 and 1;10000 simulations. F) rCDF+K_{avg} applied to 10;1000 and 1;10000 simulations.

into the magnitude of the mutational consequences of the SOS response. The SOS response being turned on would also likely result in a stronger coupling between the mutation rate and the growth rate because the proteins used in the SOS response are unregulated and therefore anti-correlated with growth rate (See Fig. 1.27). One could use this experiment to further study how bacteria behave in times of stress, which can be common in nature [91].

Finally, experiments which probe for mutation rate - growth rate coupling can be performed by determining the mutation rate through sequencing the DNA of a sample of cells that are in balanced growth. The concept is similar to a fluctuation test in that the cells must be grown for a defined amount of time in order for mutations to accumulate and a rate determined, but the DNA will be directly observed so all mutations which occurred will be known. As a result, the effects of phenotypic lag will not matter because mutations will be measured directly by reading the DNA and do not need to have a phenotypic expression. Another consequence is that all mutation types can be observed and the likelihood of each can be determined. One way to perform this experiment while being able to monitor the physiology and single cell dynamics is by growing the bacteria in a mother machine [190] and taking a sample of the outflow at a specified time to be sequenced. Accordingly, one can know the approximate physiological state of the cells being sequenced.

The study of nature through science is a never ending pursuit, with knowledge constantly being stacked upon itself to build stronger and stronger theories. Even the most insignificant appearing results in a field can open doors to studies with significant consequences. Here's to hoping that what was done in this thesis can help open one of those doors, or at least a window.

Bibliography

[1] Rosalind J. Allen and Bartlomiej Waclaw. Bacterial growth: A statistical physicist's guide. *Reports on Progress in Physics*, 82(1):016601, 2018.

[2] Antonio A. Alonso, Ignacio Molina, and Constantinos Theodoropoulos. Modeling bacterial population growth from stochastic single-cell dynamics. *Applied and environmental microbiology*, 80(17):5241–5253, 2014.

[3] Brooke Anderson. Standing on the shoulders of a tiny giant. *Small Things Considered: https://schaechter.asmblog.org/schaechter/2016/08/standing-on-the-shoulders-of-a-tiny-giant.html*, August 21, 2016.

[4] P. Armitage. The statistical theory of bacterial populations subject to mutation. *Journal of the Royal Statistical Society. Series B, Methodological*, 14(1):1–40, 1952.

[5] P. Armitage. Statistical concepts in the theory of bacterial mutation. *Epidemiology & Infection*, 51(2):162–184, 1953.

[6] Archives at NCBS. Max delbrück and salvador luria, ms-001_7_1_58_3_p_0023. *http://archives.ncbs.res.in/node/728*, 1946 (accessed September 24, 2020).

[7] The Decolonial Atlas. Haudenosaunee country in mohawk. *https://decolonialatlas.wordpress.com/2015/02/04/haudenosaunee-country-in-mohawk-2/*, February 4, 2015 (accessed September 22, 2020).

[8] Jürgen Bachl, Mark Dessing, Carina Olsson, R. C. von Borstel, and Charles Steinberg. An experimental solution for the luria–delbrück fluctuation problem in measuring hypermutation rates. *Proceedings of the National Academy of Sciences*, 96(12):6847–6849, 1999.

[9] Gary Baisa, Nicholas J. Stabo, and Rodney A. Welch. Characterization of escherichia coli d-cycloserine transport and resistant mutants. *Journal of bacteriology*, 195(7):1389–1399, 2013.

[10] Maurice Stevenson Bartlett. *An introduction to stochastic processes: with special reference to methods and applications*. CUP Archive, 1978.

[11] John R. Battista and Ashlee M. Earl. Mutagenesis and dna repair: the consequences of error and mechanisms for remaining the same. In Robert V. Miller and Martin J. Day, editors, *Microbial evolution : gene establishment, survival, and exchange*, chapter 1. ASM Press, Washington, D.C, 2004.

[12] Howard C. Berg. Motile behavior of bacteria. *Physics Today*, 53(1), 2000.

[13] Mehmet Berkmen and Paul Riggs. How did e. coli get named k-12? *Small Things Considered:* https://schaechter.asmblog.org/schaechter/2016/01/how-did-e-coli-get-named-k-12.html, January 20, 2016.

[14] Giuseppe Bertani. Studies on lysogenesis i.: the mode of phage liberation by lysogenic escherichia coli1. *Journal of bacteriology*, 62(3):293, 1951.

[15] Kamya Bhatnagar and Alex Wong. The mutational landscape of quinolone resistance in escherichia coli. *PLOS ONE*, 14(11):1–18, 11 2019.

[16] I. K. Blaby, V. de Crécy-Lagard, and T.J. Lyons. 1.22 - modes of culture/microbial. In Murray Moo-Young, editor, *Comprehensive Biotechnology (Second Edition)*, pages 303 – 314. Academic Press, Burlington, second edition edition, 2011.

[17] Zachary D. Blount. The natural history of model organisms: The unexhausted potential of *E. coli*. *eLife*, 4:e05826, mar 2015.

[18] Lars Boe, Tim Tolker-Nielsen, Karen-Margrethe Eegholm, Henrik Spliid, and Astrid Vrang. Fluctuation analysis of mutations to nalidixic acid resistance in escherichia coli. *Journal of bacteriology*, 176(10):2781–2787, 1994.

[19] Hans Bremer, Patrick P Dennis, et al. Modulation of chemical composition and other parameters of the cell by growth rate. *Escherichia coli and Salmonella: cellular and molecular biology*, 2(2):1553–1569, 1996.

[20] Sydney Brenner, François Jacob, and Matthew Meselson. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, 190(4776):576–581, 1961.

[21] Steven D. Brown and Suckjoon Jun. Complete genome sequence of escherichia coli ncm3722. *Genome Announcements*, 3(4), 2015.

[22] Vernon Bryson and Waclaw Szybalski. Microbial selection. *Science*, 116(3003):45–51, 1952.

[23] Harold J. Bull, Mary-Jane Lombardo, and Susan M. Rosenberg. Stationary-phase mutation in the bacterial chromosome: Recombination protein and dna polymerase iv dependence. *Proceedings of the National Academy of Sciences*, 98(15):8334–8341, 2001.

[24] William Bulloch. *The history of bacteriology*. Heath Clark lectures; 1936. Dover Publications, New York, 1979.

[25] Matej Butala, Darja Žgur-Bertok, and Steve JW Busby. The bacterial lexa transcriptional repressor. *Cellular and Molecular Life Sciences*, 66(1):82, 2009.

[26] J. Cairns and P. L. Foster. Adaptive reversion of a frameshift mutation in escherichia coli. *Genetics*, 128(4):695–701, 1991.

[27] Allan Campbell. Synchronization of cell division. *Bacteriological reviews*, 21(4):263, 1957.

[28] Angelo Canty. Resampling methods in r: The boot package. *R News*, 2:2–7, 01 2002.

[29] Martín Carballo-Pacheco, Michael D. Nicholson, Elin E. Lilja, Rosalind J. Allen, and Bartlomiej Waclaw. Phenotypic delay in the evolution of bacterial antibiotic resistance: Mechanistic models and their implications. *PLoS computational biology*, 16(5):e1007930–, 2020.

[30] Neal F. Cariello, Sabrina Narayanan, Puntipa Kwanyuen, Heidi Muth, and Warren M. Casey. A novel bacterial reversion and forward mutation assay based on green fluorescent protein. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, 414(1-3):95–105, 1998.

[31] M. Chalfie, Y. Tu, G. Euskirchen, W. W. Ward, and D. C. Prasher. Green fluorescent protein as a marker for gene expression. *Science*, 263(5148):802–805, 1994.

[32] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 11 2018.

[33] D. E. Contois. Kinetics of bacterial growth: Relationship between population density and specific growth rate of continuous cultures. *Microbiology*, 21(1):40–50, 1959.

[34] Stephen Cooper. *Bacterial growth and division : biochemistry and regulation of prokaryotic and eukaryotic division cycles*. Academic Press, San Diego, 1991.

[35] Stephen Cooper. The origins and meaning of the schaechter-maaloe-kjeldgaard experiments. *Microbiology-sgm*, 139:1117–1124, 06 1993.

[36] Stephen Cooper and Charles E. Helmstetter. Chromosome replication and the division cycle of escherichia coli br. *Journal of molecular biology*, 31(3):519–540, 1968.

[37] S. D. Cosloy. D-serine transport system in escherichia coli k-12. *Journal of bacteriology.*, 114(2):679–684, 1973.

[38] James W. Coulton, Patrizia Mason, D. R. Cameron, Gilles Carmel, Richard Jean, and H. N. Rode. Protein fusions of beta-galactosidase to the ferrichrome-iron receptor of escherichia coli k-12. *Journal of bacteriology*, 165(1):181–192, 1986.

[39] Six Nations Council. *Land Rights: A Global Solution for the Six Nations of the Grand River*. Six Nations Lands & Resources Department, 2019.

[40] Francis H. C. Crick. On protein synthesis. In F. K. Sanders, editor, *Symposia of the Society for Experimental Biology*, volume 12, pages 138–63. Cambridge University Press, 1958.

[41] Francis H. C. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[42] Kenny S. Crump and David G. Hoel. Mathematical models for estimating mutation rates in cell populations. *Biometrika*, 61(2):237–252, 1974.

[43] Charles Darwin. *On the origin of species*. John Murray, London, 1859.

[44] Julian Davies and Dorothy Davies. Origins and evolution of antibiotic resistance. *Microbiology and molecular biology reviews*, 74(3):417–433, 2010.

[45] M. Demerec. Production of staphylococcus strains resistant to various concentrations of penicillin. *Proceedings of the National Academy of Sciences - PNAS*, 31(1):16–24, 1945.

[46] John W. Drake. A constant rate of spontaneous mutation in dna-based microbes. *Proceedings of the National Academy of Sciences*, 88(16):7160–7164, 1991.

[47] Bradley Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987.

[48] V. I. Enne, A. A. Delsol, J. M. Roe, and P. M. Bennett. Rifampicin resistance and its fitness cost in enterococcus faecium. *Journal of antimicrobial chemotherapy.*, 53(2):203–207, 2004.

[49] Ivan Erill, Susana Campoy, and Jordi Barbé. Aeons of distress: an evolutionary perspective on the bacterial sos response. *FEMS microbiology reviews*, 31(6):637–656, 2007.

[50] Gustavo Eydallin, Ben Ryall, Ram Maharjan, and Thomas Ferenci. The nature of laboratory domestication changes in freshly isolated escherichia coli strains. *Environmental Microbiology*, 16(3):813–828, 2014.

[51] Tamás Fehér, Botond Cseh, Kinga Umenhoffer, Ildikó Karcagi, and György Pósfai. Characterization of cyca mutants of escherichia coli. an assay for measuring in vivo mutation rates. *Mutation research*, 595(1-2):184–190, 2006.

[52] Thomas Ferenci. Bacterial physiology, regulation and mutational adaptation in a chemostat environment. In Robert K. Poole, editor, *Advances in Microbial Physiology*, volume 53 of *Advances in Microbial Physiology*, pages 169 – 315. Academic Press, 2007.

[53] I. J. Fijalkowska, R. L. Dunn, and R. M. Schaaper. Genetic requirements and mutational specificity of the escherichia coli sos mutator activity. *Journal of Bacteriology*, 179(23):7435–7445, 1997.

[54] R. A. Fisher. *The Genetical Theory of Natural Selection*. Clarendon Press, 1930.

[55] National Center for Biotechnology Information. Pubchem compound summary for cid 6234, cycloserine. *Pubchem: https://pubchem.ncbi.nlm.nih.gov/compound/Cycloserine*, (accessed September 5, 2020).

[56] Forluvoft. Simple diagram of double-stranded dna. *Wikimedia Commons: https://commons.wikimedia.org/wiki/File:DNA_simple2.svg*, January 16, 2008.

[57] Patricia L. Foster. Directed mutation in escherichia coli: Theory and mechanisms. In Alfred I. Tauber, editor, *Organism and the Origins of Self*, pages 213–234. Springer Netherlands, Dordrecht, 1991.

[58] Patricia L. Foster. Methods for determining spontaneous mutation rates. *Methods in Enzymology*, 409:195–213, 2006.

[59] Patricia L. Foster. Stress-induced mutagenesis in bacteria. *Critical Reviews in Biochemistry and Molecular Biology*, 42(5):373–397, 2007.

[60] Rosalind E. Franklin and R. G. Gosling. Molecular configuration in sodium thymonucleate. *Nature (London)*, 171(4356):740–741, 1953.

[61] George Gamow. Possible relation between deoxyribonucleic acid and protein structures. *Nature*, 173(4398):318–318, 1954.

[62] Cristy Gelling. Luria & delbrück: Jackpots and epiphanies. *Genes to Genomes: a blog from the Genetics Society of America: http://genestogenomes.org/luria-delbruck-jackpots-and-epiphanies/*, March 29, 2016.

[63] Philip Gerrish. A simple formula for obtaining markedly improved mutation rate estimates. *Genetics (Austin)*, 180(3):1773–1778, 2008.

[64] D. V. Goeddel, D. G. Kleid, F. Bolivar, H. L. Heyneker, D. G. Yansura, R. Crea, T. Hirose, A. Kraszewski, K. Itakura, and A. D. Riggs. Expression in escherichia coli of chemically synthesized genes for human insulin. *Proceedings of the National Academy of Sciences*, 76(1):106–110, 1979.

[65] William A. Goss, William H. Deitz, and Thomas M. Cook. Mechanism of action of nalidixic acid on escherichia coli ii. inhibition of deoxyribonucleic acid synthesis. *Journal of Bacteriology*, 89(4):1068–1074, 1965.

[66] J. B. S. Haldane. A mathematical theory of natural and artificial selection, part v: Selection and mutation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 23(7):838–844, 1927.

[67] J. B. S. Haldane. *The Causes of Evolution*. Princeton Science Library. Princeton University Press, 1990.

[68] Agnes Hamon, Bernard Ycart, et al. Statistics for the luria-delbrück distribution. *Electronic journal of statistics*, 6:1251–1272, 2012.

[69] K. Hantke and V. Braun. Functional interaction of the tona/tonb receptor system in escherichia coli. *Journal of bacteriology*, 135(1):190–197, 1978.

[70] Klaus Hantke. Compilation of Escherichia coli K-12 outer membrane phage receptors – their function and some historical remarks. *FEMS Microbiology Letters*, 367(2), 02 2020. fnaa013.

[71] Daniel L. Hartl, Daniel E. Dykhuizen, and Antony M. Dean. Limits of adaptation: the evolution of selective neutrality. *Genetics*, 111(3):655–674, 1985.

[72] D. Harvey. *The Secret Life of Genes: Decoding the Blueprint of Life*. Firefly Books, 2019.

[73] Maria Schei Haugan, Frederik Boëtius Hertz, Godefroid Charbon, Berivan Sahin, Anders Løbner-Olesen, and Niels Frimodt-Møller. Growth rate of escherichia coli during human urinary tract infection: Implications for antibiotic effect. *Antibiotics*, 8(3):92, 2019.

[74] Brian Hayes. Computing science: The invention of the genetic code. *American Scientist*, 86(1):8–14, 1998.

[75] Christopher D. Herring, Anu Raghunathan, Christiane Honisch, Trina Patel, M. Kenyon Applebee, Andrew R. Joyce, Thomas J. Albert, Frederick R. Blattner, Dirk Van den Boom, Charles R. Cantor, et al. Comparative genome sequencing of escherichia coli allows observation of bacterial evolution on a laboratory timescale. *Nature genetics*, 38(12):1406–1412, 2006.

[76] A. D. Hershey. Factors limiting bacterial growth. *Journal of Bacteriology*, 37(3):285–299, 1939.

[77] A. D. Hershey and Martha Chase. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of general physiology*, 36(1):39–56, 1952.

[78] David C. Hooper. Mechanisms of fluoroquinolone resistance. *Drug Resistance Updates*, 2(1):38 – 55, 1999.

[79] Laura A. Hug, Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, Alex W. Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A. Relman, Kari M. Finstad, Ronald Amundson, Brian C. Thomas, and Jillian F. Banfield. A new view of the tree of life. *Nature microbiology*, 1(16048), 2016.

[80] Dann Huh and Johan Paulsson. Random partitioning of molecules at cell division. *Proceedings of the National Academy of Sciences*, 108(36):15004–15009, 2011.

[81] Sheng Hui, Josh M Silverman, Stephen S. Chen, David W Erickson, Markus Basan, Jilong Wang, Terence Hwa, and James R. Williamson. Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria. *Molecular systems biology*, 11(2):784, 2015.

[82] M. E. Jones. Accounting for plating efficiency when estimating spontaneous mutation rates. *Mutation research*, 292(2):187–189, 1993.

[83] M. E. Jones, S. M. Thomas, and A. Rogers. Luria-delbrück fluctuation experiments: Design and analysis. *Genetics*, 136(3):1209–1216, 1994.

[84] Suckjoon Jun, Fangwei Si, Rami Pugatch, and Matthew Scott. Fundamental principles in bacterial physiology—history, recent progress, and the future with focus on cell size control: a review. *Reports on progress in physics*, 81(5):056601–056601, 2018.

[85] Gary E. Kaiser. Microbiology laboratory manual: Lab 1. *http://faculty.ccbcmd.edu/courses/bio141/labmanua/lab1/lab1.html*, Aug 2017.

[86] Garima Kapoor, Saurabh Saigal, and Ashok Elongavan. Action and resistance mechanisms of antibiotics: A guide for clinicians. *Journal of anaesthesiology, clinical pharmacology*, 33(3):300, 2017.

[87] W. S. Kendal and P. Frost. Pitfalls and practice of luria-delbrück fluctuation analysis: a review. *Cancer research (Chicago, Ill.)*, 48(5):1060–1065, 1988.

[88] David G. Kendall. Stochastic processes and population growth. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(2):230–282, 1949.

[89] Motoo Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, 1983.

[90] N. O. Kjeldgaard, O. Maaløe, and M. Schaechter. The transition between different physiological states during balanced growth of salmonella typhimurium. *Journal of general microbiology*, 19(3):607–616, 1958.

[91] Arthur L. Koch. The adaptive responses of escherichia coli to a feast and famine existence. In A.H. Rose and J.F. Wilkinson, editors, *Advances in Microbial Physiology*, volume 6, pages 147 – 217. Academic Press, 1971.

[92] Arthur L. Koch. Mutation and growth rates from luria-delbrück fluctuation tests. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 95(2):129–143, 1982.

[93] U. P. Kokko. On the suitability of yeast extract for bacteriological culture media. *Acta Pathologica Microbiologica Scandinavica*, 23(6):528–535, 1946.

[94] Sohei Kondo. A theoretical study on spontaneous mutation rate. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 14(4):365–374, 1972.

[95] Martin Krzywinski and Naomi Altman. Points of significance: error bars. *Nature methods*, 10(921), 2013.

[96] H. E. Kubitschek and H. E. Bendigkeit. Mutation in continuous cultures. i. dependence of mutational response upon growth-limiting factors. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 1(2):113–120, 1964.

[97] M. P. Lambert and F. C. Neuhaus. Mechanism of d-cycloserine action: alanine racemase from escherichia coli w. *Journal of bacteriology.*, 110(3):978–987, 1972.

[98] Nick Lane. Origin of the eukaryotic cell. *Molecular Frontiers Journal*, 1:1–13, 10 2017.

[99] D. E. Lea and C. A. Coulson. The distribution of the numbers of mutants in bacterial populations. *Journal of genetics*, 49(3):264–285, 1949.

[100] Edward R. Leadbetter and Jeanne S. Poindexter, editors. *Bacteria in nature*, volume 1. Plenum Press, New York, 1985.

[101] Heewook Lee, Ellen Popodi, Haixu Tang, and Patricia L. Foster. 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*, 109(41):16416–16417, 2012.

[102] Gene-Wei Li, David Burkhardt, Carol Gross, and Jonathan S. Weissman. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, 157(3):624–635, 2014.

[103] Ping Li, Hong Lin, Zhiqiang Mi, Shaozhen Xing, Yigang Tong, and Jingxue Wang. Screening of polyvalent phage-resistant escherichia coli strains based on phage receptor analysis. *Frontiers in microbiology*, 10:850, 2019.

[104] Patricia Komp Lindgren, Åsa Karlsson, and Diarmaid Hughes. Mutation rate and evolution of fluoroquinolone resistance in escherichia coli isolates from patients with urinary tract infections. *Antimicrobial agents and chemotherapy*, 47(10):3222–3232, 2003.

[105] Bin Liu, Gustavo Eydallin, Ram P. Maharjan, Lu Feng, Lei Wang, and Thomas Ferenci. Natural escherichia coli isolates rapidly acquire genetic changes upon laboratory domestication. *Microbiology*, 163(1):22–30, 2017.

[106] Salvador E. Luria and Max Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28(6):491, 1943.

[107] Eric Lyons, Michael Freeling, Sydney Kustu, and William Inwood. Using genomic sequencing for classical genetics in e. coli k12. *PLOS ONE*, 6(2):1–16, 02 2011.

[108] W. T. Ma, G. vH. Sandri, and S. Sarkar. Analysis of the luria–delbrück distribution using discrete convolution powers. *Journal of applied probability*, 29(2):255–267, 1992.

[109] Ian MacGregor-Fors and Mark E. Payton. Contrasting diversity values: statistical inferences based on overlapping confidence intervals. *PLoS One*, 8(2):1–4, 2013.

[110] Ram P. Maharjan and Thomas Ferenci. The impact of growth rate and environmental factors on mutation rates and spectra in escherichia coli. *Environmental microbiology reports*, 10(6):626–633, 2018.

[111] R. Matthews. Why you *Must* plot your growth data on semi-log graph paper. *Six Nations Lands and Resources:* http://www.sixnations.ca/LandsResources/HaldProc.htm, 2008 (accessed September 22, 2020).

[112] Adrien Mazoyer, Rémy Drouilhet, Stéphane Despréaux, and Bernard Ycart. flan: An r package for inference on mutation models. *R Journal*, 9:334–351, 06 2017.

[113] Gregory J. McKenzie, Reuben S. Harris, Peter L. Lee, and Susan M. Rosenberg. The sos response regulates adaptive mutation. *Proceedings of the National Academy of Sciences*, 97(12):6646–6651, 2000.

[114] Gregor Mendel. Versuche uber pflanzen-hybriden. *Verhandlungen des naturforschenden Vereins in Brunn fur*, 4:3–47, 1866.

[115] Matthew Meselson and Franklin W. Stahl. The replication of dna in escherichia coli. *Proceedings of the national academy of sciences*, 44(7):671–682, 1958.

[116] R. Milo and R. Phillips. *Cell Biology by the Numbers*. CRC Press, 2015.

[117] J. Monod. The growth of bacterial cultures. *Annual review of microbiology*, 3(1):371–394, 1949.

[118] J. Monod. La technique de culture continue. théorie et applications. *Ann. Inst. Pasteur*, 79(4):390–410, 1950.

[119] Nature. Bacterial physiology. https://www.nature.com/subjects/bacterial-physiology, (accessed August 31, 2020).

[120] Frederick C. Neidhardt. Bacterial growth: Constant obsession withdn/dt. *Journal of Bacteriology*, 181(24):7405–7408, 1999.

[121] Frederick C. Neidhardt, Philip L. Bloch, and David F. Smith. Culture medium for enterobacteria. *Journal of bacteriology*, 119(3):736–747, 1974.

[122] Frederick C. Neidhardt and Boris Magasanik. Studies on the role of ribonucleic acid in the growth of bacteria. *Biochimica et biophysica acta*, 42:99–116, 1960.

[123] Frederick C. Neihardt, John L. Ingraham, and Moselio Schaechter. *Physiology of the bacterial cell : a molecular approach*. Sinauer Associates, Sunderland, Mass, 1990.

[124] H. B. Newcombe. Delayed phenotypic expression of spontaneous mutations in escherichia coli. *Genetics (Austin)*, 33(5):447–476, 1948.

[125] Hiroshi Nikaido. The limitations of lb medium. *Small Things Considered:* https://schaechter.asmblog.org/schaechter/2009/11/the-limitations-of-lb-medium.html, November 9, 2009.

[126] NobelPrize.org. The nobel prize in physiology or medicine 1969. *Nobel Media AB 2020:* https://www.nobelprize.org/prizes/medicine/1969/summary/, (accessed August 31, 2020).

[127] NobelPrize.org. Press release. *Nobel Media AB 2020:* https://www.nobelprize.org/prizes/medicine/1969/press-release/, (accessed August 31, 2020).

[128] Masafumi Noda, Yumi Kawahara, Azusa Ichikawa, Yasuyuki Matoba, Hiroaki Matsuo, Dong-Geun Lee, Takanori Kumagai, and Masanori Sugiyama. Self-protection mechanism in d-cycloserine-producing streptomyces lavendulae. gene cloning, characterization, and kinetics of its alanine racemase and d-alanyl-d-alanine ligase, which are target enzymes of d-cycloserine. *The Journal of biological chemistry.*, 279(44):46143–46152, 2004.

[129] A. Novick and L. Szilard. Experiments with the chemostat on spontaneous mutations of bacteria. *Proceedings of the National Academy of Sciences - PNAS*, 36(12):708–719, 1950.

[130] Aaron Novick and Leo Szilard. Description of the chemostat. *Science*, 112(2920):715–716, 1950.

[131] National Institute of Allergy and Infectious Diseases. Scanning electron micrograph of escherichia coli, grown in culture and adhered to a cover slip. *https://www.flickr.com/photos/niaid/7316101966*, November 14, 2002.

[132] OpenStax. The action of dna polymerase during replication. *Wikimedia Commons: https://commons.wikimedia.org/wiki/File:0323_DNA_Replication.jpg*, July 3, 2016.

[133] Jeremy Orloff and Jonathan Bloom. Bootstrap confidence intervals. *MIT OpenCourseWare (CC BY-NC-SA 4.0). 18.05 Introduction to Probability and Statistics (Spring 2014): https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading24.pdf*, 2014 (accessed October 18, 2020).

[134] P. R. Painter and A. G. Marr. Mathematics of microbial populations. *Annual Review of Microbiology*, 22(1):519–548, 1968. PMID: 4879521.

[135] Rod Pierce. Confidence intervals. *Math Is Fun: http://www.mathsisfun.com/data/confidence-interval.html*, July 20, 2020 (accessed September 7, 2020).

[136] P. Pletnev, I. Osterman, P. Sergiev, A. Bogdanov, and O. Dontsova. Survival guide: Escherichia coli in the stationary phase. *Actanaturae*, 7(4):22–33, 2015.

[137] Cassie F. Pope, Denise M. O'Sullivan, Timothy D. McHugh, and Stephen H. Gillespie. A practical guide to measuring mutation rates in antibiotic resistance. *Antimicrobial agents and chemotherapy*, 52(4):1209–1214, 2008.

[138] Kathleen Postle and Rhonda F. Good. Dna sequence of the escherichia coli tonb gene. *Proceedings of the National Academy of Sciences*, 80(17):5235–5239, 1983.

[139] Leslie Pray. Antibiotic resistance, mutation rates and mrsa. *Nature Education*, 1(1):30, 2008.

[140] Leslie Pray. Discovery of dna structure and function: Watson and crick. *Nature Education*, 1(1), 2008.

[141] Lucas Proust, Alain Sourabié, Martin Pedersen, Iris Besançon, Eloi Haudebourg, Véronique Monnet, and Vincent Juillard. Insights into the complexity of yeast extract peptides and their utilization by streptococcus thermophilus. *Frontiers in Microbiology*, 10:906, 2019.

[142] Mark Ptashne. *A genetic switch : phage lambda revisited*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y, 3rd ed. edition, 2004.

[143] J. C. Robbins and D. L. Oxender. Transport systems for alanine, serine, and glycine in escherichia coli k-12. *Journal of bacteriology.*, 116(1):12–18, 1973.

[144] William A. Rosche and Patricia L. Foster. Determining mutation rates in bacterial populations. *Methods*, 20(1):4 – 17, 2000.

[145] Richard Routledge. Law of large numbers. *Encyclopædia Britannica: https://www.britannica.com/science/law-of-large-numbers*, (accessed September 3, 2020).

[146] Carl Sagan. Definitions of life. In Mark A. Bedau and Carol E. Cleland, editors, *The Nature of Life: Classical and Contemporary Perspectives from Philosophy and Science*, chapter 3, pages 303–306. Cambridge University Press, 2010.

[147] S. Sarkar. Haldane's solution of the luria-delbrück distribution. *Genetics (Austin)*, 127(2):257–261, 1991.

[148] S. Sarkar, W. T. Ma, and G. vH. Sandri. On fluctuation analysis: a new, simple and efficient method for computing the expected number of mutants. *Genetica*, 85(2):173–179, 1992.

[149] Sahotra Sarkar. Lamarck contre darwin, reduction versus statistics: Conceptual issues in the controversy over directed mutagenesis in bacteria. In Alfred I. Tauber, editor, *Organism and the Origins of Self*, pages 235–271. Springer Netherlands, Dordrecht, 1991.

[150] Moselio Schaechter. A brief history of bacterial growth physiology. *Frontiers in Microbiology*, 6:289, 2015.

[151] Moselio Schaechter. Why you *Must* plot your growth data on semi-log graph paper. *Small Things Considered: https://schaechter.asmblog.org/schaechter/2018/07/why-you-must-plot-your-growth-data-on-semi-log-graph-paper.html*, July 26, 2018.

[152] Moselio Schaechter, John L. Ingraham, and Frederick C. Neidhardt. *Microbe*. American Society for Microbiology, Washington, D.C, 2006.

[153] Moselio Schaechter, Ole Maaløe, and Niels O. Kjeldgaard. Dependency on medium and temperature of cell size and chemical composition during balanced growth of salmonella typhimurium. *Microbiology*, 19(3):592–606, 1958.

[154] Moselio Schaechter and Frederick C. Neidhardt. Introduction. In F.C. Neidhardt, J.L. Ingraham, and R. Curtiss, editors, *Escherichia coli and Salmonella typhimurium : cellular and molecular biology*, volume 1, chapter 1. American Society for Microbiology, Washington, D.C, 1987.

[155] Bradley Titus Scheer. Physiology. *https://www.britannica.com/science/physiology*, November 22, 2018.

[156] Nathaniel Schenker and Jane F. Gentleman. On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, 55(3):182–186, 2001.

[157] Dominique Schneider and Richard E. Lenski. Dynamics of insertion sequence elements during experimental evolution of bacteria. *Research in microbiology.*, 155(5):319–327, 2004.

[158] Erwin Schrödinger. *What is Life? The Physical Aspect of the Living Cell*. Cambridge University Press, 1944.

[159] Matthew Scott. *Quantitative Methods in Bacterial Physiology*. Waterloo, Ontario, Canada, 2017.

[160] Matthew Scott, Carl W. Gunderson, Eduard M. Mateescu, Zhongge Zhang, and Terence Hwa. Interdependence of cell growth and gene expression: origins and consequences. *Science*, 330(6007):1099–1102, 2010.

[161] Ron Sender, Shai Fuchs, and Ron Milo. Revised estimates for the number of human and bacteria cells in the body. *PLoS biology*, 14(8):e1002533, 2016.

[162] Guennadi Sezonov, Danièle Joseleau-Petit, and Richard d'Ari. Escherichia coli physiology in luria-bertani broth. *Journal of bacteriology*, 189(23):8746–8749, 2007.

[163] Paul Singleton and Diana Sainsbury. *Introduction to bacteria : for students in the biological sciences*. Wiley, Chichester, 1981.

[164] Harry Smith. Pathogenicity and the microbe in vivo: The 1989 fred griffith review lecture. *Microbiology*, 136(3):377–383, 1990.

[165] Jackson Smith, Cassandra Puckett, and Wendy Simon. Know the land territories campaign; indigenous allyship: An overview. *http://www.lspirg.org/knowtheland*, 2016 (accessed September 22, 2020).

[166] T. M. Sonneborn and Ruth Stocking Lynch. Hybridization and segregation in paramecium aurelia. *Journal of Experimental Zoology*, 67(1):1–72, 1934.

[167] Eric Soupene, Wally C. van Heeswijk, Jacqueline Plumbridge, Valley Stewart, Daniel Bertenthal, Haidy Lee, Gyaneshwar Prasad, Oleg Paliy, Parinya Charernnoppakul, and Sydney Kustu. Physiological studies of escherichia coli strain mg1655: Growth defects and apparent cross-regulation of gene expression. *Journal of Bacteriology*, 185(18):5611–5626, 2003.

[168] Keiran Stevenson, Alexander F. McVey, Ivan B. N. Clark, Peter S. Swain, and Teuta Pilizota. General calibration of microbial growth in microplate readers. *Scientific reports*, 6(1):1–7, 2016.

[169] F. M. Stewart. Fluctuation tests: How reliable are the estimates of mutation rates? *Genetics*, 137(4):1139–1146, 1994.

[170] F. M. Stewart, D. M. Gordon, and B. R. Levin. Fluctuation analysis: The probability distribution of the number of mutants under different conditions. *Genetics*, 124(1):175–185, 1990.

[171] Lei Sun, Helen K. Alexander, Balazs Bogos, Daniel J. Kiviet, Martin Ackermann, and Sebastian Bonhoeffer. Effective polyploidy causes phenotypic delay and influences bacterial evolvability. *PLOS Biology*, 16(2):1–24, 02 2018.

[172] Lei Sun, Helen K. Alexander, Balazs Bogos, Daniel J. Kiviet, Martin Ackermann, and Sebastian Bonhoeffer. Data from: Effective polyploidy causes phenotypic delay and influences bacterial evolvability. *Dryad*, February 19, 2019.

[173] Sotaro Takano, Bogna J. Pawlowska, Ivana Gudelj, Tetsuya Yomo, and Saburo Tsuru. Density-dependent recycling promotes the long-term survival of bacterial populations during periods of starvation. *MBio*, 8(1), 2017.

[174] E. L. Tatum and Joshua Lederberg. Gene recombination in the bacterium escherichia coli. *The Journal of Bacteriology*, 53(6):673–684, 1947-06-01.

[175] Teknova. 10x acgu solution. https://www.teknova.com/acgu-solution-10x.html, (accessed September 9, 2020).

[176] Teknova. 10x mops buffer (used in ez rich defined medium kit). https://www.teknova.com/mops-10x-for-ez-rich-defined-medium-kit-m2105.html, (accessed September 9, 2020).

[177] Teknova. Mops ez rich defined medium kit. https://www.teknova.com/mops-ez-rich-defined-medium-kit.html, (accessed September 9, 2020).

[178] Teknova. Mops minimal media kit. https://www.teknova.com/mops-ez-rich-minimal-media-kit.html, (accessed September 9, 2020).

[179] Teknova. Supplement ez 5x. https://www.teknova.com/supplement-ez-5x.html, (accessed September 9, 2020).

[180] Olivier Tenaillon, David Skurnik, Bertrand Picard, and Erick Denamur. The population genetics of commensal escherichia coli. *Nature Reviews Microbiology*, 8(3), March 2010.

[181] theLabRat.com. M9 minimal media recipe (1000 ml). http://www.thelabrat.com/protocols/m9minimal.shtml, 2005 (accessed September 9, 2020).

[182] P. Thomas, A. C. Sekhar, and M. M. Mujawar. Nonrecovery of varying proportions of viable bacteria during spread plating governed by the extent of spreader usage and proposal for an alternate spotting-spreading approach to maximize the cfu. *Journal of applied microbiology*, 113(2):339–350, 2012.

[183] David S. Treves, Shannon Manning, and Julian Adams. Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of escherichia coli. *Molecular biology and evolution*, 15(7):789–797, 1998.

[184] Vaccinationist. Skeletal formula of d-cycloserine. *Wikimedia Commons:* https://commons.wikimedia.org/wiki/File:Cycloserine.svg, June 13, 2015.

[185] N.G. Van Kampen. *Stochastic Processes in Physics and Chemistry*. North-Holland Personal Library. Elsevier Science, 1992.

[186] C. B. Van Niel. The kinetics of growth of micro-organisms. In Arthur K. Parpart, editor, *The chemistry and physiology of growth.*, chapter 5. New Jersey. Princeton University Press., 1949.

[187] T. Venkova, C. C. Yeo, and M. Espinosa. *The Good, The Bad and The Ugly: Multiple Roles of Bacteria in Human Life*. Frontiers Research Topics. Frontiers Media SA, 2018.

[188] C. H. Wang and A. L. Koch. Constancy of growth on simple and complex media. *Journal of Bacteriology*, 136(3):969–975, 1978.

[189] Mingcong Wang, Christina J. Herrmann, Milan Simonovic, Damian Szklarczyk, and Christian von Mering. Version 4.0 of paxdb: protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*, 15(18):3163–3168, 2015.

[190] Ping Wang, Lydia Robert, James Pelletier, Wei Lien Dang, Francois Taddei, Andrew Wright, and Suckjoon Jun. Robust growth of escherichia coli. *Current biology*, 20(12):1099–1103, 2010.

[191] R. J. Wargel, C. A. Shadur, and F. C. Neuhaus. Mechanism of d-cycloserine action: transport systems for d-alanine, d-cycloserine, l-alanine, and glycine. *Journal of bacteriology.*, 103(3):778–788, 1970.

[192] Trudy M. Wassenaar. *Bacteria : the benign, the bad, and the beautiful*. Wiley-Blackwell, Hoboken, N. J., 2012.

[193] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature (London)*, 171(4356):737–738, 1953.

[194] Aryeh Wides and Ron Milo. Understanding the dynamics and optimizing the performance of chemostat selection experiments. *arXiv preprint arXiv:1806.00272*, 2018.

[195] The Free Encyclopedia Wikipedia. Propogation of uncertainty. *https://en.wikipedia.org/wiki/Propagation_of_uncertainty*, (accessed September 7, 2020).

[196] M. H. F. Wilkins, A. R. Stokes, and H. R Wilson. Molecular structure of nucleic acids: Molecular structure of deoxypentose nucleic acids. *Nature (London)*, 171(4356):738–740, 1953.

[197] Ashley B. Williams. Spontaneous mutation rates come into focus in escherichia coli. *DNA Repair*, 24:73 – 79, 2014.

[198] Bernard Ycart and Nicolas Veziris. Unbiased estimation of mutation rates under fluctuating final counts. *PLOS ONE*, 9(7):1–10, 07 2014.

[199] Conghui You, Hiroyuki Okano, Sheng Hui, Zhongge Zhang, Minsu Kim, Carl W. Gunderson, Yi-Ping Wang, Peter Lenz, Dalai Yan, and Terence Hwa. Coordination of bacterial proteome with metabolism by cyclic amp signalling. *Nature*, 500(7462):301–306, 2013.

[200] Zhongge Zhang and Milton H. Saier, Jr. A novel mechanism of transposon-mediated gene activation. *PLOS Genetics*, 5(10):1–9, 10 2009.

[201] Qi Zheng. Progress of a half century in the study of the luria–delbrück distribution. *Mathematical Biosciences*, 162(1):1–32, 1999.

[202] Qi Zheng. Statistical and algorithmic methods for fluctuation analysis with salvador as an implementation. *Mathematical Biosciences*, 176(2):237–252, 2002.

[203] Qi Zheng. New algorithms for luria–delbrück fluctuation analysis. *Mathematical Biosciences*, 196(2):198–214, 2005.

[204] Qi Zheng. On haldane's formulation of luria and delbrück's mutation model. *Mathematical Biosciences*, 209(2):500–513, 2007.

[205] Qi Zheng. The luria-delbrück distribution: early statistical thinking about evolution. *Chance*, 23(2):15–18, 2010.

[206] Qi Zheng. A bayesian approach for correcting for partial plating in fluctuation experiments. *Genetics research*, 93(5):351–356, 2011.

[207] Qi Zheng. Methods for comparing mutation rates using fluctuation assay data. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 777:20–22, 2015.

[208] Qi Zheng. A new practical guide to the luria–delbrück protocol. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 781:7–13, 2015.

[209] Qi Zheng. A second look at the final number of cells in a fluctuation experiment. *Journal of theoretical biology*, 401:54–63, 2016.

[210] Qi Zheng. rsalvador: An r package for the fluctuation experiment. *G3 (Bethesda, Md.)*, 7(12):3849–3856, 2017.

APPENDICES

Appendix A

Lab Practices

In order to study bacteria in the laboratory, techniques to measure their attributes had to be developed. Most importantly for my purposes, techniques for measuring the exact quantity of bacteria in a culture as well as their growth rate were necessary.

A.1 Optical Density

One technique that has been used since the early days of bacterial growth physiology and continues to be used today is optical density (OD), which measures the turbidity of a culture. The optical density of a bacterial culture is measured with a spectrophotometer which shines light of a specific wavelength (commonly 600nm for *E. coli*) through a small liquid sample and measures how much light comes out the other side (see Fig. A.1). Because bacteria scatter the light, by measuring how much of the light is lost when travelling through the culture one indirectly measures how much cell volume or mass is in the sample [117, 168, 153]. Therefore, one can measure the growth of a bacteria culture by observing the change in optical density of a culture through time when the cells' characteristics are constant. Due to the growth dependent nature of cell size (see Section 1.6.1) the measure of OD does not give a measure of the number of cells without some sort of conversion factor [168]. To find this conversion factor one plates cultures and counts the number of colony forming units.

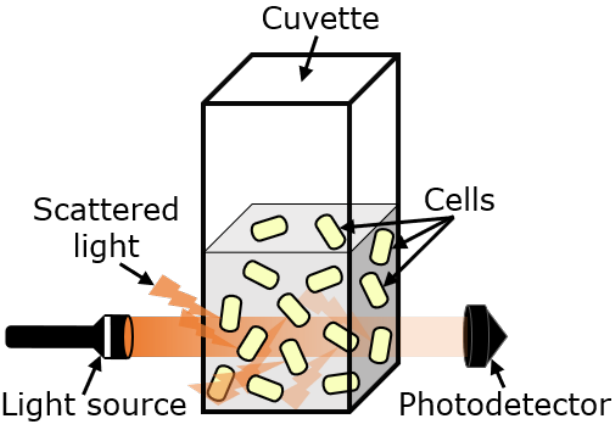

Figure A.1: **Spectrophotometer diagram.** Light is shone through a cuvette filled with culture and the bacteria scatter some of the light so that the light exiting the other side of the cuvette is lower in intensity. The change in intensity between the light on either side of the cuvette is a proxy for the number of cells. Cells not to scale.

A.2 Colony Forming Units

Colony forming units (CFU) are the bacteria in a culture which when placed on a surface with nutrients will grow to form a colony of bacteria[1] The goal when placing the bacteria on the surface is to have all the colony forming units sufficiently dispersed so that each formed colony was seeded by a single bacteria. If achieved, the number of colonies on the surface will tell you how many viable bacteria were in the placed sample (also called the viable cell count). To do this, culture samples are generally placed on small dishes/plates that contain nutrients and agar, which works as a sort of scaffolding. The samples must be sufficiently dilute so that there are not too many cells (generally 100-300 is ideal), and the samples are spread evenly around the dish in the hopes that cells won't overlap. These dishes are then incubated at the ideal growing temperature for the bacteria. In practice, the plated sample generally has to be serial diluted from culture in order to reach an ideal concentration for plating (if the culture has $10^8 \frac{\text{cells}}{\text{mL}}$, then to have 200 cells on the plate, a $5 \cdot 10^5\times$ dilution must be performed, which is usually done as a $100\times$ then $100\times$ then $50\times$ dilution). Once the sample is ready, there are several different plating techniques that can be performed, but in the experiments outlined in this thesis, pour plating is the main technique practised. Pour plating is done by first adding a base layer of high percentage

[1] This colony is observed as a single goopy dot of bacteria on a surface.

agar (1-1.5%) with nutrients and letting it solidify. Then the sample is mixed with a low percentage agar (0.5-0.8%) with nutrients either directly on the plate, or in a tube then poured onto the plate[2] (see Fig. A.2 for the plating procedures). Once the sample is added to the plate, it must be swirled around to assure even distribution of the cells. Other plating techniques require direct contact with the plate and cells with tools, which can damage or remove cells, giving less accurate counts [182]. After the plates have incubated and the colonies counted, the number is then multiplied by the dilution amount to get the number of cells per millilitre in the culture. To find the relation between the OD and CFU, one must determine the CFU for a culture with different optical densities. The result will be a linear relation which can be used to calculate the number of cells in a sample from the OD and vice versa (see Fig. 2.4).

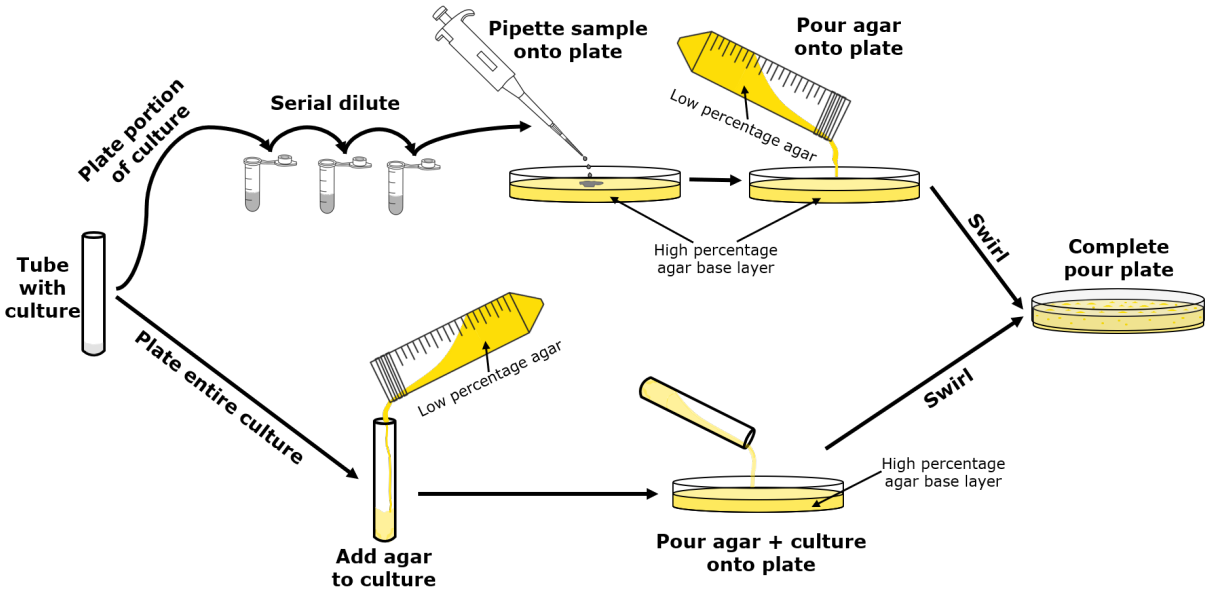

Figure A.2: **Pour plating protocol infographic.** Two different pour plating protocols are outlined; one for plating a portion/sample of the culture, and one for plating the entire culture.

[2]When plating only a portion of the culture, adding the sample directly to the plate with a pipette then adding the agar and swirling is the better technique, but when plating the entire culture (such as in the selection phase of a fluctuation test), combining the sample with the agar in a tube and then pouring it onto the plate is better (see Appendix C.2 for validation).

A.3 Measuring Bacterial Growth

Measuring the rate at which bacteria grow is essential to understanding a bacteria's physiology (see Section 1.6.1). To measure growth rate, bacteria need an environment in which to grow and we need a way to measure said growth over time. The environments commonly used to grow bacteria in lab settings come in two forms: batch culture and continuous culture [16]. Batch culture is when a population of bacteria is grown in a container with finite resources for a period of time. Continuous cultures are grown in machines which remove cells throughout growth in order to keep the bacterial population in a culture constant. Both environments lend themselves to a straight forward way of determining growth rate.

A.3.1 Batch Culture

In batch culture growth, a container (commonly a test tube) will be filled with growth medium and a small inoculum of bacteria. To measure the bacterial grow that ensues, either the optical density (OD) or viable cell count (CFU) is measured periodically. If bacteria are in constant conditions and have adapted their physiology to said conditions, they will grow exponentially. In particular, the bacteria need consistent access to food in non-limiting amounts, which is achieved by providing a saturating amount of nutrients in the growth medium. During exponential growth, the natural logarithm of the OD or CFU can be plotted versus time to get a straight line[3], of which the slope is the specific growth rate, λ [151]. Eventually the bacteria will consume so much of the nutrients that they will no longer be in saturating amounts and the cells will stop growing exponentially [117]. Consequently, if one wants to observe bacterial growth for a long period of time in batch culture, they must re-dilute the cells into a tube with the same conditions as the previous tube before the nutrients reach non-saturating levels. The resulting growth plot is a sawtooth wave, which can be seen in Fig. A.3.

Generally, one can get sufficient information to determine the growth rate of a bacteria by measuring only a couple doublings during the exponential phase. When I measure growth rates, I inoculate 1-3mL of growth medium with a sample of culture that was grown in the same growth medium overnight. I choose the dilution such that after a few doublings the OD will be at approximately 0.1, at which point I start taking a measurement

[3]The coefficient of determination (R^2) for this line can give insight into how exponential the data is, and therefore how confident one can be that the cells are in balanced growth ($R^2 > 0.99$ is good and common).

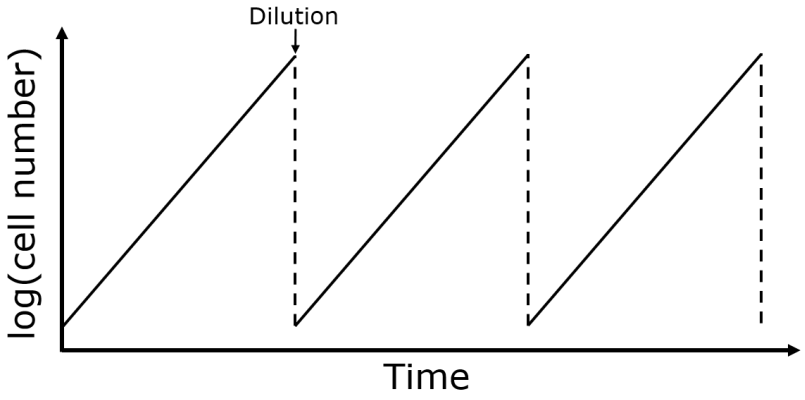

Figure A.3: **Batch culture growth curve.** When bacteria are grown in batch culture for long periods of time they will run out of nutrients, so to keep them in exponential phase they must be periodically re-diluted.

at approximately every half doubling for 2-3 doublings (OD $\approx$ 0.1, 0.15, 0.2, 0.3, 0.4)[4] The same can be down by plating samples of the culture and counting the number of colonies, which can be done with cells at any concentration (though it is much more time and resource consuming). In theory, using optical density and viable cell counts to determine growth rate should be equivalent, though in practice I have found this not to be the case at slow growth (see Section 4.4).

A.3.2 Continuous Culture

Due to the necessity of actively making measurements during batch growth, it can be a lot of work to grow cultures in this fashion, especially for long periods of time. To combat this, continuous culture methods were developed. Continuous culture growth is a method of growth which keeps the population of the growing culture on average constant by periodically removing culture and adding fresh growth medium [118]. This is usually done by one of two machines: a turbidostat or a chemostat (see Fig. A.4) [16]. A turbidostat measures the optical density (or turbidity) of the culture and keeps it constant through periodic dilutions [22]. A chemostat keeps the concentration of a nutrient constant in the culture through a constant dilution [130]. During balanced growth when bacteria are

[4]Every spectrophotometer has an optimal range below which it can't properly measure the OD and above which the cells are too dense or have begun reaching saturation [168]. This range must be considered when planning a growth experiment.

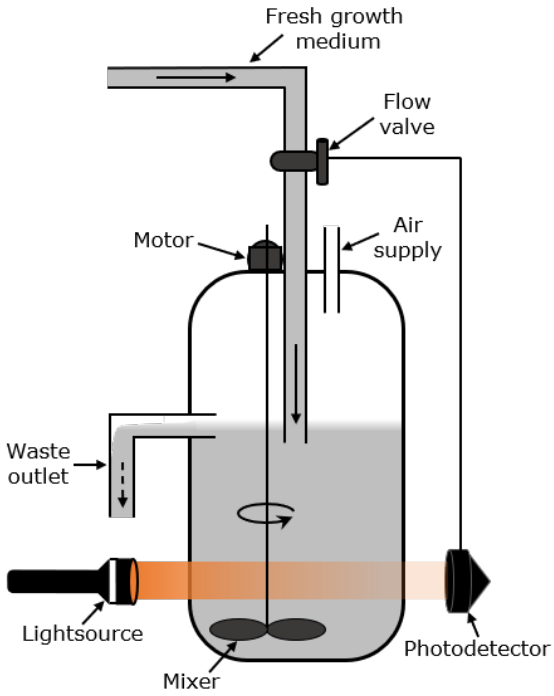

Figure A.4: **Turbidostat diagram.** Diagram of a turbidostat, which maintains a constant turbidity (OD) of a culture through dilution. A chemostat is simply a turbidostat, but with the light source and photodetector removed so that the dilution rate is constant.

consuming nutrients at a constant rate and the concentration of cells are increasing in the culture at a constant rate, the turbidostat and chemostat are equivalent. To determine the growth rate of the culture, one looks at the dilution rate of the machine. The faster the cells are growing, the faster the culture must be diluted to maintain a constant population, and vice versa. Upon further inspection, it can be shown that the growth rate of the culture is exactly equivalent to the dilution rate, which can be determined from the machine [16].

Though continuous cultures require little active input, they do require a larger amount of time to set up and calibrate. Furthermore, they impose a selection bias on the culture by constantly removing cells [194]. In particular, they favour cells that stick to the walls of the growth container and allow for fast growing mutants to completely take over the population (unlike in batch culture where a fast growing mutant can only compose a portion of the total population). Consequently, even though continuous cultures seem ideal for long growth experiments, there are practical restrictions on how long they are useful before unaccountable effects dominate.

Appendix B

Media Recipes

B.1 MOPS Based Media

B.1.1 10× MOPS Buffer

Chemical	Concentration	Mass in 100mL H_2O
3-(N-morpholino)propanesulfonic acid (MOPS)	400mM	8.372g
Tricine	40mM	716.8mg
Iron Sulfate	0.10mM	2.78mg
Ammonium Chloride	95mM	508.2mg
Potassium Sulfate	2.76mM	48.1mg
Calcium Chloride	5μM	73.5μg
Magnesium Chloride	5.25mM	106.8mg
Sodium Chloride	500mM	2.922g
Ammonium Molybdate	$2.92{\cdot}10^{-6}$mM	3.6μg
Boric Acid	$4.0{\cdot}10^{-4}$mM	24.74μg
Cobalt Chloride	$3.02{\cdot}10^{-5}$mM	7.18μg
Copper Sulfate	$9.62{\cdot}10^{-6}$mM	2.4μg
Manganese Chloride	$8.08{\cdot}10^{-5}$mM	16μg
Zinc Sulfate	$9.74{\cdot}10^{-6}$mM	2.8μg

Table B.1: **10× MOPS buffer recipe.** Supplied by Teknova [176].

B.1.2 5× EZ Supplement

Chemical	Concentration	Mass in 100mL H_2O
L-Alanine	4.0mM	35.6mg
L-Arginine HCl	26mM	548.6mg
L-Asparagine	2.0mM	26.4mg
L-Aspartic Acid, Potassium Salt	2.0mM	26.6mg
L-Glutamic Acid, Potassium Salt	3.0mM	44.1mg
L-Glutamine	3.0mM	43.8mg
L-Glycine	4.0mM	30.0mg
L-Histidine HCl H_2O	1.0mM	21.0mg
L-Isoleucine	2.0mM	26.2mg
L-Proline	2.0mM	23.0mg
L-Threonine	2.0mM	23.8mg
L-Tryptophan	0.5mM	10.2mg
L-Valine	3.0mM	35.2mg
L-Leucine	4.0mM	52.5mg
L-Lysine HCl	2.0mM	36.5mg
L-Methionine	1.0mM	14.9mg
L-Phenylalanine	2.0mM	33.0mg
L-Cysteine HCl	0.5mM	7.9mg
L-Tyrosine	1.0mM	18.1mg
Thiamine HCl	0.05mM	1.3mg
Calcium Pantothenate	0.05mM	1.2mg
para-Hydroxy Benzoic Acid	0.05mM	0.7mg

Table B.2: **5× EZ supplement recipe.** Adapted from Teknova's recipe [179] to not include L-Serine (Teknova's normal 5× EZ supplement has a L-Serine concentration of 50mM).

B.1.3 10× ACGU Supplement

Chemical	Concentration	Mass in 100mL H_2O
Guanine	1.99mM	30.1mg
Adenine	1.99mM	26.9mg
Cytosine	1.99mM	22.1mg
Uracil	1.99mM	22.3mg

Table B.3: **5× ACGU supplement recipe.** Supplied by Teknova [175].

B.1.4 Rich Defined Media (MOPS)

Solution	Volume in 100mL H_2O
10× MOPS buffer	10mL
0.132M K_2HPO_4	1mL
20% (w/v) glucose	1mL
5× EZ supplement	20mL
10× ACGU supplement	10mL

Table B.4: **MOPS based rich defined media with glucose (RDM glucose) recipe.** Recipe from Teknova [177].

To use the rich defined medium with glucose, first the 10× MOPS buffer, 0.132M K_2HPO_4, and 20% (w/v) glucose are combined at the volumes described in Table B.4 then adjusted to 70mL with sterile Milli-Q H_2O (which can be stored at 4°C for extended periods of time). Then in each individual tube to be used for bacterial growth, the above buffer+glucose solution is added to be 70% of the total volume, 5× EZ is added at 20%, and 10× ACGU is added at 10% (i.e. for 3mL of growth medium, one combines 2.1mL buffer+glucose, 600μL 5× EZ, and 300μL 10× ACGU).

B.1.5 Minimal Defined Media (MOPS)

Solution	Volume in 100mL H_2O
10× MOPS buffer	10mL
0.132M K_2HPO_4	1mL
20% (w/v) maltose	
or	
2M α-ketoglutaric acid	1mL
or	
4M acetate	

Table B.5: **MOPS based minimal media recipe.** Three possible carbon sources are provided; adding maltose gives the maltose minimal media predominantly used throughout the thesis. Recipe from Teknova [178].

B.2 M9 Based Media

B.2.1 M9 Salts

Chemical	Mass in 1L H_2O
Na_2HPO_4-$7H_2O$	64g
KH_2PO_4	15g
NaCl	2.5g
NH_4Cl	5.0g

Table B.6: **M9 salts recipe.** M9 salts used for M9 minimal media. Recipe from the-LabRat.com [181].

B.2.2 M9 Minimal Media

1. Make M9 salts and sterilize by autoclaving.
2. Measure approximately 700mL of sterile distilled H_2O.
3. Add 200mL of M9 salts.
4. Add 2mL of sterile 1M $MgSO_4$.
5. Add 20mL of sterile 20% (w/v) glucose (or other carbon source).
6. Add 100μL of 1M $CaCl_2$.
7. Adjust to 1000mL with sterile distilled H_2O.

Recipe from theLabRat.com [181]

Appendix C

Control Experiments

C.1 Continued Growth After Dilution

The inoculation period of my fluctuation test is lengthy at 55 minutes, during which the cells are sitting in buffer. This means that the cells in the last inoculum will have been sitting in buffer for approximately 50 minutes longer than the cells in the first inoculum. If the cells continue to grow during this time in the buffer, the variance between inocula could end up being substantial. Accordingly, a control experiment was performed to observe how much growth there is after the cells are diluted into buffer from each relevant growth medium. *Escherichia coli* NCM3722 cells were grown in three different MOPS based media (RDM glucose, maltose minimal, and acetate minimal) to exponential phase after an adaption period and then serial diluted in cold buffer composed of MOPS and K_2HPO_4 (Recipe from Appendix B.1.5 excluding carbon source). Cells were left in the buffer and samples of the same quantity were plated periodically to track growth. See Table C.1 for RDM glucose results, Table C.2 for maltose minimal results, and Table C.3 for acetate minimal results.

From the results of the observation of growth in buffer, it appears that cells from RDM glucose continue growing for approximately 30 minutes and then settle. On the other hand, cells from maltose minimal and acetate minimal appear to be stable for approximately 60 minutes and then begin to show growth. Because the cells sit in buffer for the entire 55 minutes of my inoculation period, it appears the best protocol for reducing the variance in the initial inoculua is to let cells from RDM glucose sit in the buffer for 30 minutes before commencing inoculation. Conversely, for cells from minimal media, it appears best to begin the inoculations immediately.

Time in buffer	5 mins	15 mins	30 mins	45 mins	60 mins
Colonies ± SD	337 ± 21	396 ± 44	450 ± 27	455 ± 21	450 ± 39

Table C.1: ***E. coli* NCM3722 growth in buffer after dilution from RDM glucose balanced growth.** *E. coli* NCM3722 grown in MOPS based rich defined medium with glucose until well established in balanced growth and then serial diluted in cold MOPS based buffer and let sit. Three plates are made from a constant sample of the buffer+culture periodically to track growth. At each time point the average number of colonies on the plate ± one standard deviation are presented.

Time in buffer	6 mins	22 mins	62 mins	123 mins
Colonies ± SD	263 ± 25	281 ± 5	342 ± 18	437 ± 19

Table C.2: ***E. coli* NCM3722 growth in buffer after dilution from maltose minimal balanced growth.** *E. coli* NCM3722 grown in MOPS based minimal medium with maltose until well established in balanced growth and then serial diluted in cold MOPS based buffer and let sit. Three plates are made from a constant sample of the buffer+culture periodically to track growth. At each time point the average number of colonies on the plate ± one standard deviation are presented.

Time in buffer	5 mins	20 mins	50 mins	100 mins	150 mins
Colonies	840	837	827	979	1177

Table C.3: ***E. coli* NCM3722 growth in buffer after dilution from acetate minimal balanced growth.** *E. coli* NCM3722 grown in MOPS based minimal medium with acetate until well established in balanced growth and then serial diluted in cold MOPS based buffer and let sit. One plate is made from a constant sample of the buffer+culture periodically to track growth.

C.2 Comparison of Pour Plating Techniques

Escherichia coli NCM3722 cells were grown in MOPS based rich defined medium to exponential phase after an adaption period and then serial diluted and distributed evenly among 10 test tubes. Each test tube was levelled to 1mL of culture with buffer. Five tubes were poured directly onto a petri dish with a base of 1% agar with LB, and then approximately 3mL of 0.7% agar with LB was poured onto the plate and swirled to mix the culture and agar (this method will be referred to as the "plate then agar" method). The other five tubes had 4mL of 0.7% agar with LB added to them, were swirled by hand, and then were poured onto a petri dish with a base of 1% agar with LB and swirled (this method will be referred to as the "agar then plate" method). The "plate then agar" method resulted in an average colony count plus or minus one standard deviation of 407 ± 35 cells while the "agar then plate" method resulted in 426 ± 26 cells. Notice that the "agar then plate" method produces both a marginally higher count (meaning there are likely fewer cells lost in the process) and a marginally lower variance, making it a more efficient and reliable plating technique. Consequently, the "agar then plate" method was used while plating the selection plates during the fluctuation test (the cultures with selecting agent used to determine mutant numbers).

Appendix D

Code

D.1 Convert Data to a Cumulative Distribution

```
# data is fluctuation test data as a vector
# bin all samples with more than "top" mutants
top <- 300
for (i in 1:length(data)){
  if (data[i] >= top){
    data[i] <- top
  }
}

# make function take takes in data and returns a cdf
cdf.data <- function(data){
    # sort data by number of mutants
    nums <- sort(unique(data))
    # count how many samples have each number of mutants
    counts <- integer(length(nums))
    for (i in 1:length(data)){
      for (j in 1:length(nums)){
        if (data[i] == nums[j]){
          counts[j] <- counts[j]+1
        }
      }
```

```
    }

    # turn data into a probability distribution function
    pdf <- counts/(sum(counts))
    # turn into a cumulative distribution function
    cdf <- cumsum(pdf)

    # compile mutant numbers (x-axis) and
    #cumulative probabilities (y-axis) in a matrix
    C <- matrix(nrow=2,ncol=length(nums))
    C[1,] <- nums
    C[2,] <- cdf

    return(C)
}
```

D.2 Total Sum of Squares Fitting

```
# load necessary packages
library("rsalvador")

# create function which estimates m by fitting data to a Luria-Delbruck
#CDF through minimisation of the total sum of squares distance
# m0 is the rsalvador MLE estimate of m
# cdf is a vector of the experimental cumulative probabilities
# nums is a vector of the number of mutants corresponding to cdf
TSS.fit.LD <- function(m0,cdf,nums){

    l <- 0
    fit_TSS <- c()

    # loop through sequence of guesses on m and
    #calculate TSS error for each
    m <- seq(0,m0*2,by=0.001)
    for (b in m){
        l <- l+1
```

```
        # build theoretical pdf with rSalvador then turn into cdf
        k <- 0:(floor(max(nums))+1)
        pdf_t <- prob.LD(b,k=(floor(max(nums))+1))
        cdf_t <- cumsum(pdf_t)

        # reduce & interpolate theoretical cdf to only include same points
        #as experimental cdf
        intercdf <- c()
        for (a in 1:length(nums)){
            interp  <- approx(k,cdf_t,nums[a],method="linear")
            intercdf[a] <- interp[2]
        }
        intercdf <- as.vector(intercdf,mode="numeric")

        # calculate total sum of squares distance
        f <- sum((intercdf-cdf)^2)
        fit_TSS[l] <- f
        }

        # find which m gives best TSS fit
        for (t in 1:length(fit_TSS)){
            if (fit_TSS[t] == min(fit_TSS)){
                m_best_TSS <- m[t]
        }
    }

    # compile best m and its associated fit
    m_and_fit <- c(m_best_TSS,min(fit_TSS))

    return(m_and_fit)
}
```

D.3 Fluctuation Test Simulation

The following code is adapted from Sun et al. (2018) [171, 172], which is in the public domain.

```
# load necessary packages
library(data.table)

# simulate growth of one culture
# treats normal cells deterministically and mutant lineages
#stochastically like the Lea-Coulson formulation

# INPUTS:
# N0: initial number of cells
# Nf: final number of cells after growth
# divdist: name of the distribution of interdivision times;
#can be 'exp' or 'const'
# mu: probability of mutation per division
# pheno: length of phenotypic lag in generations
# protein: number of selected proteins a normal cell is
#born with on average

# OUTPUT:
# nmuts: total number of mutants
# nmuts_pheno: number of resistant bacteria

simculture.lag.partition <- function(N0,Nf,divdist,mu,pheno,protein){

  # determine number of generations of growth
  growthgens <- log2(Nf/N0)

  # calculate final number of normal cells in culture
  # number of mutants assumed negligible in comparison
  Nf <- N0*2^growthgens

  # define exponential growth rate of non-mutant cells
  beta <- log(2)
```

```
# calculate mean number of mutations that occur during culture growth
mmut <- mu*(Nf-N0)

# draw actual number of mutations from Poisson distribution
#with mean=nmut
# gives number of mutant lineages to be simulated
numdnmut <- rpois(1,mmut)

# if no mutations occur, return this and stop simulation
if(numdnmut==0){
  nmuts <- 0
  nmuts_pheno <- nmuts
  NMUTS <- c(nmuts,nmuts_pheno)
  return(NMUTS)
}

# draw "developing time" of each mutant lineage from
#exponential distribution
Tclones <- rexp(numdnmut,rate=beta)

# define vectors needed later
muts <- c()
muts_pheno <- c()

# analyse each mutant lineage
for(clone in 1:numdnmut){

  Tclone <- Tclones[clone]

  # initialize data table to keep track of all cells, at first
  #containing only the progenitor
  # columns: phyloID is a string of L's and R's uniquely specifying
  #line of descent from the progenitor
  # tb is cell's birth time, measured from time lineage is initiated
  # td is cell's division time, measured from time lineage is initiated
  # proteins is the number of selected protein in the cell
```

```
switch(divdist$name,
       exp = {newtd <- rexp(1,rate=beta)},
       const = {newtd <- 1},
)
protein_0<-rbinom(1,2*protein,0.5)
newgen <- data.table(phyloID="",tb=0, td=newtd,proteins=protein_0)
cells <- newgen
setkey(newgen,td)

while(newgen[1,td] <= Tclone){
# iterate until there is no mother cell left that still divides
#before end of clone's growth time

  mothers <- newgen[td<=Tclone]

  # produce two daughter cells for each mother cell in current gen
  # append "L" and "R" to each mother's ID
  newIDs <- c(paste(mothers[,phyloID],"L",sep=""),
              paste(mothers[,phyloID],"R",sep=""))
  # daughters' birth times are their mothers' division times
  newtb <- rep(mothers[,td],2)
  # daughters' division times are their birth times plus
  #independent random number drawn from interdivision time distribution
  switch(divdist$name,
         exp = {newtd <- newtb + rexp(length(newtb),rate=beta)},
         const = {newtd <- newtb + 1},
  )
  # split proteins in mother between children binomially with p=0.5
  nprot1 <- rbinom(length(mothers[,proteins]),mothers[,proteins],0.5)
  nprot1 <- ifelse(nprot1<0,0,nprot1)
  nprot2 <- mothers[,proteins]-nprot1
  nprot2 <- ifelse(nprot2<0,0,nprot2)
  newprotein <- c(nprot1,nprot2)

  # build information table for daughters
  newgen <- data.table(phyloID=newIDs, tb=newtb, td=newtd,
                       proteins=newprotein)
  cells <- rbindlist(list(cells,newgen))
```

```
      setkey(newgen,td)
    }

    # cells "alive" at end of clone's growth time: those that did not
    #yet divide by end of clone growth time
    setkey(cells,td)
    livecells <- cells[td>Tclone]
    muts <- c(muts,dim(livecells)[1])

    # determine which cells have diluted out sufficient protein
    #to be resistant
    muts_pheno_0 <- c()
    prot_live <- livecells[,proteins]
    for (i in 1:length(prot_live)){
      if (prot_live[i] < (protein/(2^pheno))){
        muts_pheno_0 <- c(muts_pheno_0,1)
      }
      else if (prot_live[i] >= (protein/(2^pheno))){
        muts_pheno_0 <- c(muts_pheno_0,0)
      }
    }
    muts_pheno <- c(muts_pheno,sum(muts_pheno_0))
  }

  # count number of total mutants
  nmuts <- sum(muts)

  # count number of resistant mutants
  nmuts_pheno <- sum(muts_pheno)

  # compile into a vector
  NMUTS <- c(nmuts,nmuts_pheno)

  return(NMUTS)
}
```

D.4 Adjusting Fit for Phenotypic Lag

All proceeding procedures require loading the rSalvador package as well as the "Convert Data to a Cumulative Distribution" function and the "Total Sum of Squares Fitting" function. The following procedures also require the input of fluctuation test data:

```
# compile data
data <- #insert fluctuation test data as a vector here
popmean <- #insert average final population per culture here
pop0 <-  #insert average initial population per culture here
phi <- 1-(pop0/popmean)
```

D.4.1 Koch Adjustment

```
# prep vectors
mvec_K <- c()
fitpheno_K <- c()

# bin all samples with more than "top" mutants
top <- 300
for (i in 1:length(data)){
    if (data[i] >= top){
        data[i] <- top
    }
}

# turn fluctuation test data into a cdf
C <- cdf.data(data)
nums <- C[1,]
cdf <- C[2,]

# fit data using rSalvador MLE
m0 <- newton.LD(data)
#build theoretical pdf for MLE fit
k0 <- 0:max(data)
p0 <-prob.LD(m0,k=max(data))
#turn into cdf
```

```
c0 <- cumsum(p0)

# loop through potential phenotypic lags (in generations)
pheno_tot_K <- seq(0,4,0.1)
w <- 0
for (p in pheno_tot_K){
    w <- w+1

    # adjust data for phenotypic lag
    nums <- nums/(2^p)

    # find optimal m using TSS fitting and its associated error
    m_and_error <- TSS.fit.LD(m0*2,cdf,nums)
    mlag <- m_and_error[1]
    # adjust estimated m for lag and save it
    mvec_K[w] <- mlag*(2^p)

    # record error for best TTS m estimate
    fitpheno_K[w] <- m_and_error[2]

    # compute theoretical pdf for optimal m
    k <- 0:(max(nums)+1)
    plag <- prob.LD(mlag,k=(max(nums)+1))
    #turn into cdf
    clag <- cumsum(plag)

    # plot experimental & theoretical cdfs
    plot(mutnum,cdf,col="red",xlab="Number of mutants",
         ylab="Cumulative probability",ylim=c(0,1),xlim=c(0,50))
    lines(k,clag,col="blue")
    legend("bottomright",c("Adjusted data","Fitted CDF"),
           fill=c("red","blue"))
}

# find lag which gives best fit
for (t in 1:length(fitpheno_K)){
    if (fitpheno_K[t] == min(fitpheno_K)){
        pheno_best_K <- pheno_tot_K[t]
```

```
        t_best <- t
    }
}

# save optimal lag length and m estimate for Koch adjustment
lag_K <- pheno_best_K
mlag_K <- mvec_K[t_best]

# plot how the fitting error changes with phenotypic lag length
plot(pheno_tot_K,fitpheno_K/fitpheno_K[1],col="black",
     xlab="Estimated phenotypic lag (gens)",
     ylab="Normalised Koch TSS error",main="Koch Fitting Error",
     ylim=c(0,max(fitpheno_K/fitpheno_K[1])))
points(pch=19,lag_K,fitpheno_K[t_best]/fitpheno_K[1],col="red")

# plot how m estimate changes with phenotypic lag length
plot(pheno_tot_K,mvec_K,col="black",
     xlab="Estimated phenotypic lag (gens)",
     ylab="Estimated mutation number",
     main="Koch Mutation Number Estimates",ylim=c(0,max(mvec_K)))
points(pch=19,lag_K,mlag_K,col="red")

# build cdf for theoretical distribution with Koch estimate of m
pdf_K <- prob.LD(mlag_K,k=max(data))
cdf_K <- cumsum(pdf_K)

# plot the cdf of the data, Koch estimate CDF, and MLE estimate CDF
plot(nums,cdf,col="red",main="Koch Adjustment Result",
     xlab="Number of mutants",ylab="Cumulative probability",
     ylim=c(0,1),xlim=c(0,50))
lines(k0,c0,col="blue")
lines(0:(max(data)),cdf_K,col="green")
legend("bottomright",c("Data","MLE","Koch"),fill=c("red","blue","green"))
```

D.4.2 Reduced CDF Adjustment

```
# prep vectors
```

```
mvec_rCDF<-c()
fitpheno_rCDF<-c()

# bin all samples with more than "top" mutants
top <- 300
for (i in 1:length(data)){
    if (data[i] >= top){
        data[i] <- top
    }
}

# turn fluctuation test data into a cdf
C <- cdf.data(data)
nums <- C[1,]
cdf <- C[2,]

# fit data using rSalvador MLE
m0 <- newton.LD(data)
#build theoretical pdf for MLE fit
k0 <- 0:max(data)
p0 <-prob.LD(m0,k=max(data))
#turn into cdf
c0 <- cumsum(p0)

# loop through potential phenotypic lags (in generations)
pheno_tot_rCDF <- c(0,1,log2(3),2,log2(5),log2(6),log2(7),3,log2(9),log2(10),
                    log2(11),log2(12),log2(13),log2(14),log2(15),4)
w <- 0
for (p in pheno_tot_rCDF){
    w <- w+1

    # reduce cdf to only include points with
    #greater than or equal to 2^p mutants
    cdf_r <- c()
    nums_r <- c()
    z <- 1
    if (p == 0){
        cdf_r <- cdf
```

```
    nums_r <- nums
} else {
    for (i in 1:length(cdf)){
        if (nums[i] >= (2^p)){
            cdf_r[z] <- cdf[i]
            nums_r[z] <- nums[i]
            z <- z+1
        }
    }
}

# find optimal m using TSS fitting and its associated error
m_and_error <- TSS.fit.LD(m0*2,cdf_r,nums_r)
mlag <- m_and_error[1]
# save optimal rCDF m estimate
mvec_rCDF[w] <- mlag

# compute theoretical pdf for optimal m
k <- 0:(max(nums)+1)
plag <- prob.LD(mlag,k=(max(nums)+1))
#turn into cdf
clag <- cumsum(plag)

# error from TSS fitting of reduced cdf
d <- m_and_error[2]

# note the zero point of the experimental data
cdf_0 <- cdf[1]

# compress all theoretical points < 2^p into 0
i_best<-0
for (i in 1:min(length(k),ceiling(2^(max(pheno_tot_rCDF)+1)))){
    if (k[i] < 2^p){
        i_best <- i_best + 1
    }
}
clag_0 <- clag[i_best]
```

```
    # calculate difference between theoretical and experimental zeros
    d0 <- (clag_0-cdf_0)^2

    # calculate rCDF fitting error
    fitpheno_rCDF[w] <- d0+d

    # another choice is:
    #fitpheno_rCDF[w] <- (((sum(counts_0)-sum(counts))*d0)+
                          ((sum(counts))*d))/(sum(counts_0))
    #which more explicitly accounts for lost information when removing
    #points through a taking a weighted average

    # plot experimental & theoretical cdfs
    plot(mutnum,cdf,col="red",xlab="Number of mutants",
         ylab="Cumulative probability",ylim=c(0,1),xlim=c(0,50))
    lines(k,clag,col="blue")
    points(pch=19,0,cdf_0,col="orange")
    points(pch=19,0,clag_0,col="cyan")
    legend("bottomright",c("Adjusted data", "Experimental zero",
                           "Fitted CDF", "Theoretical zero"),
                               fill=c("red","orange","blue","cyan"))
}


# find lag which gives best fit
for (t in 1:length(fitpheno_rCDF)){
    if (fitpheno_rCDF[t] == min(fitpheno_rCDF)){
        pheno_best_rCDF <- pheno_tot_rCDF[t]
        t_best <- t
    }
}

# save optimal lag length and m estimate for Koch adjustment
lag_rCDF <- pheno_best_rCDF
mlag_rCDF <- mvec_rCDF[t_best]

# plot how the fitting error changes with phenotypic lag length
plot(pheno_tot_rCDF,fitpheno_rCDF/fitpheno_rCDF[1],col="black",
```

```
     xlab="Estimated phenotypic lag (gens)",
     ylab="Normalised rCDF error",main="rCDF Fitting Error",
     ylim=c(0,max(fitpheno_rCDF/fitpheno_rCDF[1])))
points(pch=19,lag_rCDF,fitpheno_rCDF[t_best]/fitpheno_rCDF[1],col="red")

# plot how m estimate changes with phenotypic lag length
plot(pheno_tot_rCDF,mvec_rCDF,col="black",
     xlab="Estimated phenotypic lag (gens)",
     ylab="Estimated mutation number",
     main="rCDF Mutation Number Estimates",ylim=c(0,max(mvec_rCDF)))
points(pch=19,lag_rCDF,mlag_rCDF,col="red")

# build cdf for theoretical distribution with rCDF estimate of m
pdf_rCDF <- prob.LD(mlag_rCDF,k=max(data))
cdf_rCDF <- cumsum(pdf_rCDF)

# plot the cdf of the data, rCDF estimate CDF, and MLE estimate CDF
plot(nums,cdf,col="red",main="rCDF Adjustment Result",
     xlab="Number of mutants",ylab="Cumulative probability",
     ylim=c(0,1),xlim=c(0,50))
lines(k0,c0,col="blue")
lines(0:(max(data)),cdf_rCDF,col="green")
legend("bottomright",c("Data","MLE","rCDF"),fill=c("red","blue","green"))
```

D.4.3 rCDF & Koch Hybrid Adjustments

```
# prep vectors
mvec_rCDFK_K<-c()
fitpheno_rCDFK<-c()

# bin all samples with more than "top" mutants
top <- 300
for (i in 1:length(data)){
    if (data[i] >= top){
        data[i] <- top
    }
```

```
}

# turn fluctuation test data into a cdf
C <- cdf.data(data)
nums <- C[1,]
cdf <- C[2,]

# fit data using rSalvador MLE
m0 <- newton.LD(data)
#build theoretical pdf for MLE fit
k0 <- 0:max(data)
p0 <-prob.LD(m0,k=max(data))
#turn into cdf
c0 <- cumsum(p0)

# loop through potential phenotypic lags (in generations)
pheno_tot_rCDF <- c(0,1,log2(3),2,log2(5),log2(6),log2(7),3,log2(9),log2(10),
                    log2(11),log2(12),log2(13),log2(14),log2(15),4)
w <- 0
for (p in pheno_tot_rCDF){
    w <- w+1

    # adjust data for phenotypic lag
    nums <- nums/(2^p)

    # find optimal m using TSS fitting and its associated error
    m_and_error <- TSS.fit.LD(m0*2,cdf,nums)
    mlag <- m_and_error[1]
    # adjust estimated m for lag and save it
    mvec_rCDFK_K[w] <- mlag*(2^p)

    # record error for best TTS m estimate
    fitpheno_rCDFK[w] <- m_and_error[2]
}


# find lag which gives best fit
for (t in 1:length(fitpheno_rCDFK)){
```

```
    if (fitpheno_rCDFK[t] == min(fitpheno_rCDFK)){
        pheno_best_rCDFK <- pheno_tot_rCDFK[t]
        t_best <- t
    }
}

# save optimal lag length
lag_rCDFK <- pheno_best_rCDFK
# determine optimal rCDF estimate of m
#(use calculated mvec_rCDF from the rCDF algorithm)
mlag_rCDFK <- mvec_rCDF[t_best]
# save Koch estimate for the same lag
mlag_rCDFK_K <- mvec_rCDFK_K[t_best]

# plot how the fitting error changes with phenotypic lag length
plot(pheno_tot_rCDF,fitpheno_rCDFK/fitpheno_rCDFK[1],col="black",
     xlab="Estimated phenotypic lag (gens)",
     ylab="Normalised Koch error",main="rCDF+Koch Fitting Error",
     ylim=c(0,max(fitpheno_rCDFK/fitpheno_rCDFK[1])))
points(pch=19,lag_rCDFK,fitpheno_rCDFK[t_best]/fitpheno_rCDFK[1],col="red")

# plot how m estimate changes with phenotypic lag length
plot(pheno_tot_rCDF,mvec_rCDFK,col="black",
     xlab="Estimated phenotypic lag (gens)",
     ylab="Estimated mutation number",
     main="rCDF+Koch Mutation Number Estimates",ylim=c(0,max(mvec_rCDFK)))
points(pch=19,lag_rCDFK,mlag_rCDFK,col="red")

# build cdf for theoretical distribution with rCDF+K estimate of m
pdf_rCDFK <- prob.LD(mlag_rCDFK,k=max(data))
cdf_rCDFK <- cumsum(pdf_rCDFK)

# plot the cdf of the data, rCDF+K estimate CDF, and MLE estimate CDF
plot(nums,cdf,col="red",main="rCDF+Koch Adjustment Result",
     xlab="Number of mutants",ylab="Cumulative probability",
     ylim=c(0,1),xlim=c(0,50))
lines(k0,c0,col="blue")
lines(0:(max(data)),cdf_rCDFK,col="green")
```

```
legend("bottomright",c("Data","MLE","rCDF+K"),fill=c("red","blue","green"))

# take the average of the rCDF and Koch estimated m with determined lag
mlag_rCDFKavg <- (mlag_rCDFK+mlag_rCDFK_K)/2

# build cdf for theoretical distribution with rCDF+K_avg estimate of m
pdf_rCDFKavg <- prob.LD(mlag_rCDFKavg,k=max(data))
cdf_rCDFKavg <- cumsum(pdf_rCDFKavg)

# plot the cdf of the data, rCDF+K_avg estimate CDF, and MLE estimate CDF
plot(nums,cdf,col="red",main="rCDF+K_avg Adjustment Result",
     xlab="Number of mutants",ylab="Cumulative probability",
     ylim=c(0,1),xlim=c(0,50))
lines(k0,c0,col="blue")
lines(0:(max(data)),cdf_rCDFKavg,col="green")
legend("bottomright",c("Data","MLE","rCDF+K_avg"),
       fill=c("red","blue","green"))
```

Glossary

amino acids Organic compounds composed of an amino group, carboxyl group, and a side chain which is unique to each different amino acid. Amino acids are the building blocks of proteins. 4

bacteria A single-celled organism that constitutes one of the kingdoms of life. 1

bacterial physiology The study of the functions which allow bacteria to grow and reproduce, and the resulting dynamics. Historically done by observing population dynamics and using mathematics to infer cellular behaviour. 30

chromosome A DNA molecule which is a single, long, tangled circle of DNA in bacteria; often equivalent to the genome in bacteria. 9

coefficient of variation The standard deviation of a variable divided by its mean. 104

colony forming units A measure of the number of viable cells in a culture found from plating a sample of the culture and counting the number of colonies, each of which was established by a single cell. 51

conjugation The direct transfer of DNA between two proteins through cell to cell contact. 6

cumulative distribution function A function which gives the probability that a random variable will be less than or equal to a value. 66

DNA polymerases Proteins which move along the DNA, replicating the nucleotide sequence and correcting errors. Come in five different types. 12

doubling rate How many doublings a bacterial culture completes in a period of time. Generally measured as doublings/hour. 7

doubling time How long it takes a bacterial culture to double its population on average. Generally measured in minutes. 7

eukaryote An organism whose cells have internal membrane-bound organisation; of particular importance is the nucleus, which holds the cell's DNA. Eukaryotes can form multicellular or single celled organisms. 1

fluctuation test An experiment which observes the number of mutants in many parallel cultures after a period of growth and uses the data to determine attributes of the mutation process. First developed by Salvador Luria and Max Delbrück in 1943. 14

gene A portion of the genome which codes for the synthesis of material, commonly protein. 10

genome The total genetic information of a living organism. 6

half-inhibition concentration (IC_{50}) The concentration of antibiotic which causes the balanced growth rate of a bacteria to be half of what it is when there is no antibiotic present. 55

inoculate To introduce cells to a new medium. 9

mutagens External factors that cause mutations such as chemicals or light. 12

mutation A change in the sequence of nucleotides in a strand of DNA. 9

nucleotides A molecule composed of phosphorous, sugar, and nucleobases; nucleotides are stacked together to form DNA. 4

optical density A measure of the amount of light that passes through a liquid. Commonly performed using 600nm light, denoted OD_{600}. When used to measure bacterial cultures in liquid medium, the optical density is proportional to the total cell mass in the culture. 51

phenotype The observable physical properties of an organism; these include the organism's appearance, development, and behaviour. 11

probability distribution function A function which gives the probability that a random variable will equal a value. 21

probability generating function A representation of the probability distribution function in power series form. 21

prokaryote A single celled organism with no internal membrane-bound organisation. 1

proteome How the total amount of protein in a cell is allocated amongst the different types of proteins. 40

ribosome The molecule which makes proteins. 11

selecting agent An agent that selects for a specific phenotype. Commonly an antibiotic or phage. 15, 54

SOS response A stress response system in which bacteria up-regulate error prone DNA polymerases; commonly used to quickly repair DNA damage caused by mutagens. 13, 43

turbidostat A machine which periodically measures the optical density of a growing bacteria culture and dilutes the culture with fresh media. These are used to maintain exponential growth and population density. 43