Effects of aerobic and anaerobic environments on bacterial mutation rates and mutation spectra assessed by whole genome analyses

A thesis presented in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Genetics at Massey University, Palmerston North, New Zealand.

Sonal Shewaramani

2015

Abstract

For organisms that are exposed to different environments, the rates and types of spontaneous mutations that arise in each environment can vary, and potentially impact the direction of evolution as a whole. Oxidative stress is a major cause of mutation, but the effect of oxygen availability on the mutation rates and spectra of organisms grown in aerobic as compared to anaerobic environments is not well understood at the whole genome level. To investigate the mutation rates and spectra of a facultative anaerobic bacterium grown under strictly aerobic or anaerobic conditions, 24 mutation accumulation lineages, derived from *Escherichia coli* REL4536, were established and propagated through 180 and 144 single-colony population bottlenecks, respectively. Spontaneous mutation rates of 2.50×10^{-10} and 4.14×10^{-10} mutations per nucleotide per generation were obtained for aerobically and anaerobically grown cells, respectively. Mutations in the aerobic environment were significantly biased towards G → T mutations and IS*186* transposition, while C → A, T → G, A → C mutations, gross chromosomal rearrangements (GCRs) and IS*150* transposition were significantly more prevalent under anaerobic conditions. Transcriptional profiling, *via* RNAseq, of REL4536 grown under aerobic and anaerobic environments revealed that repair genes, especially those involved in the repair of GCRs, were generally up-regulated in the anaerobic environment, consistent with findings that mutation rates, especially for GCRs, are higher in the anaerobic environment.

GCRs have long been thought to play an important role in the evolutionary process, though their contributions to the process have not been specifically defined. SbcCD, an exonuclease, is involved in the repair of DNA secondary structures, and is thought to help prevent the occurrence of GCRs. Transcriptome analyses showed that in *E. coli*, *sbcC* was up-regulated during growth in an anaerobic environment, as compared to an aerobic environment. To investigate the impact of GCRs on adaptive evolution, an *E. coli* REL4536 strain with disrupted *sbcC* was constructed and evolved under anaerobic conditions for 1,000 generations in glucose-limited media in 14 parallel populations. Mutations that arose during evolution were determined by whole genome re-sequencing of selected clones, and evolved *sbcC* mutant strains displayed more GCRs and enhanced population-level fitness on average. Together, these results suggest that GCRs may play an important role in the rate of adaptation.

Acknowledgements

To begin with, I would like to offer my sincere appreciation and gratitude to my supervisors: Dr Christina Moon, Dr Sinead Leahy, Dr Rees Kassen and Prof Paul Rainey. Thank you for the support, guidance and knowledge you have given during the course of this project. In particular, I would like to thank Christina and Sinead, for their kindness, patience and encouragement over the years. It has truly meant a lot to me and I will always be grateful to both of you for being such great mentors.

I would also like to extend my gratitude towards the Marsden Fund for funding this project, towards Massey University for financial support, and finally, to AgResearch, for providing financial support and excellent research facilities.

This PhD would not have been possible without the people that were around me, with whom I have had the great pleasure of working with. I would like to acknowledge Zaneta Park and Benoit Auvray for their work on the MA simulator programme. Thank you to the Rumen Microbiology team for all the assistance and support over the last few years; it has been a great environment to work in. I would also like to express my thanks to Dr Bill Kelly, Dr Peter Janssen, Dr Ron Ronimus and Dr Dragana Gagic for always being so willing to share their expertise and wisdom. My thanks to Carrie Sang, Dong Li and Rechelle Perry, for all the reagents and technical assistance you have provided over the years. Thank you also to Paul Lovejoy, Dr Eric Altermann, Simon Guest, Paul Maclean and AgResearch Information Services for your invaluable help along the way.

To the members of the student room, both past and present, thank you so much for your friendship. I would especially like to thank Denise Martin, Eva Maier, Filomena Ng, Dr Milica Ciric, Dr Sam Noel and Renee Atua; this experience truly would not have been the same without any of you. Thank you so much for all the laughs, fun, advice and encouragement over the years. To Thomas Finn, thank you so much for being such a great colleague and friend, I truly cannot imagine anyone else who would have made this project as enjoyable as it has been.

Finally, to my family, thank you for always being there for me with all your unwavering love and support, even when you had no clue what my research was about. To my parents especially: I am well aware of all the sacrifices you have made for me over the years, and I thank you for always encouraging my curiosity. I hope you are proud.

Table of Contents

List of tables

List of figures

Non-standard abbreviations

λ-Red	Lambda Red
A	Adenine (nucleotide base)
AE	Aerobic
AHT	Anhydrotetracycline
Amp^R	Ampicillin resistant strain
AN	Anaerobic
Anc	Ancestor
Ara$^-$	Strain that cannot utilise arabinose
Ara$^+$	Strain that can utilise arabinose
ATP	Adenosine triphosphate
BER	Base excision repair
BLAST	Basic Local Alignment Sequence Tools
bp	Base pairs
BPS	Base pair substitution
C	Cytosine (nucleotide base)
cDNA	Complementary DNA
CFU	Colony forming units
CL	Confidence limits
Cm^R	Chloramphenicol resistant strain
CQ	Quantification cycle
DEPC H_2O	Diethylpyrocarbonate treated water
dH_2O	Distilled water
DM	Davis minimal
DNA	Deoxyribonucleic acid
DSB	Double-strand break
dsDNA	Double-stranded DNA
E	Amplification efficiency for qPCR reactions
G	Guanine (nucleotide base)
GCR	Gross chromosomal rearrangement
IR	Inverted repeat
IS	Insertion elements
kb	Kilobase
LB	Luria-Bertani
LTEE	Long-term experimental evolution
MA	Mutation accumulation
Mb	Megabase
MB	Mega bytes

MD	Macrodomain
MGE	Mobile genetic element
MMR	Mismtach repair
mRNA	Messenger RNA
MSS-MLE	Ma-Sandri-Sarkar Maximum Likelihood Estimator method
Nal^{R}	Nalidixic acid resistant strain
ncRNA	Non-coding RNA
NER	Nucleotide excision repair
NS	Non-structured regions
OD	Optical density
oriC	Origin of replication
PCA	Principal component analysis
PCR	Polymerase chain reaction
pol	Polymerase
ppGpp	Guanosine tetra-phosphates
pppGpp	Guanosine penta-phosphates
qPCR	Quantitative polymerase chain reaction
R^2	Correlation coefficient
Rif^{r}	Rifampicin resistant reference strain
RNA	Ribonucleic acid
ROS	Reactive oxygen species
RPKM	Reads per kilobase of gene per million reads mapped
RPM	Rotations per minute
rRNA	Ribosomal RNA
RT	Room temperature
RT-qPCR	Reverse transcription quantitative polymerase chain reaction
SCM	Small colony morphotype
SEM	Standard error of the mean
SSB	Single-strand break
ssDNA	Single-stranded DNA
T	Thymine (nucleotide base)
TCM	Typical colony morphotype
TE	Transposable element
tRNA	Transfer RNA
Ts/Tv	Transition to transversion ratio
v/v	Volume per volume
v/v/v	Volume per volume per volume
w/v	Weight per volume

Chapter One: Introduction

1.1 Mutations and evolution

Evolution is the change in gene frequency in populations of organisms over successive generations. The main processes that drive evolution include natural selection and genetic drift (1). Natural selection occurs when organisms that survive and produce more offspring than their less-adapted counterparts tend to increase in frequency, while genetic drift is when the relative frequencies of genotypes in a population change due to random sampling and chance (2). Populations of smaller sizes are more susceptible to the effects of genetic drift. For evolution to occur, natural selection and genetic drift require genetic variation upon which they can act. The primary source of genetic variation is spontaneous mutation, and processes such as gene flow, and recombination further contribute to genetic variation within populations (3, 4). Gene flow is the transfer of genes from one population to another, while recombination results in new gene combinations within organisms, which is commonly due to sexual reproduction (5). However, the ultimate source of new genetic variation is from mutation, which is the change in nucleotide sequence of an organism (5).

In evolutionary terms, mutations can generally be described by their effects on fitness, where fitness is a quantitative measure of the reproductive success of an organism (6), and therefore, its ability to contribute to the gene pool of successive generations. At the molecular level, the majority of mutations are selectively neutral, and so, have no effect on fitness (7). Some mutations are beneficial and have a positive effect on fitness, while some mutations are termed deleterious and have a negative effect on fitness. Natural selection acts to eliminate deleterious mutations from a population, while beneficial mutations are selected for, and accumulate within populations (8). Neutral mutations are subject to genetic drift and in extreme cases, may be lost, or become fixed within the populations. Generally, deleterious mutations spontaneously arise more frequently than beneficial mutations (9).

For any organism, the types of mutations that arise (section 1.3), and the spontaneous mutation rate (section 1.6) directly impact its evolutionary potential. In different environments, the types of mutations that arise, and the mutation rates of an organism, can vary, thus affecting the adaptive pathways that populations may undertake, as well as the rate of adaptation. For example, cells grown in aerobic and anaerobic environments differ in how they generate energy, resulting in differences in cellular physiological conditions and mutagenic pressures.

Escherichia coli is a facultative anaerobe; it can grow in both the presence and absence of oxygen. In the presence of oxygen, i.e. under aerobic growth conditions, *E. coli* cells generate adenosine triphosphate (ATP) by aerobic respiration. Here, oxygen is used as the terminal electron acceptor in an electron transport chain that is used to generate ATP. In contrast, in the absence of oxygen, i.e. under anaerobic growth conditions, *E. coli* cells can generate ATP by either anaerobic respiration or anaerobic fermentation. If electron acceptors, such as nitrate, nitrite, sulphate or fumarate are available, cells undergo anaerobic respiration. However, if no electron acceptors are available, then cells generate energy through a process known as anaerobic fermentation. During anaerobic fermentation, ATP yield is relatively low and thus, cells growth rate is slower (10). As a facultative anaerobe, *E. coli* is an ideal organism to use for studying the effects of aerobic and anaerobic environments on mutation rates and spectra. While genome-wide spontaneous mutation rates for *E. coli* grown under aerobic conditions have been published (11-13), similarly derived rates for *E. coli* grown under anaerobic conditions have not been published. Understanding how mutation rates and spectra vary in aerobic and anaerobic environments will provide us with insights into the process of evolution under the two environments.

1.2 Sources of mutations

Mutations can be classified into two categories: induced and spontaneous. Induced mutations occur as a result of exogenous agents, such as ionizing radiation, UV, cosmic rays, viruses and alkylating agents. These mutagens usually cause point mutations (section 1.3.1), pyrimidine dimerization or strand breaks in the DNA, amongst other mutation types (5, 14-16). In contrast, spontaneous mutations are those that occur in the absence of exogenous agents (17), and can occur due to errors in DNA replication or as a result of DNA damage occurring during normal growth (5, 17, 18). Such types of DNA damage include the oxidation, alkylation and hydrolysis (i.e. deamination, depurination, or depyrimidination) of nucleotide bases (5). During aerobic respiration, a major source of spontaneous mutations is oxidative DNA damage caused by the presence of reactive oxygen species (ROS) (19).

1.2.1 Reactive oxygen species (ROS)

ROS are normal cellular by-products of aerobic respiration and are formed when molecular oxygen (O_2) is reduced by the addition of electrons (20). ROS can also be introduced *via* environmental agents (14, 16, 21). Some of the more commonly encountered ROS are superoxide radicals (O_2^-), hydrogen peroxide (H_2O_2) and the highly reactive hydroxyl radicals (OH^-) (5, 16, 22). Though ROS play important roles in processes like homeostasis and the induction of host defence genes, they are also a major source of DNA damage and can attack cellular components. For example, they are involved in lipid peroxidation and amino acid oxidation (5, 23). ROS exposure can also result in the formation of strand breaks, damage to bases and sugar-phosphates or replication blocks (14).

More than twenty different types of oxidative stress-induced DNA lesions have been identified (24-27); see **Table A.1** for examples. In particular, guanine bases are quite vulnerable to oxidation because of their low redox potential (14). One of the most extensively studied and frequently encountered lesions is the 8-oxoG lesion, which is also commonly called the GO lesion (24). Principally, OH^- radicals react with guanine nucleotide bases to produce nucleotide bases that have the potential to be mutagenic as they can pair with both adenine (A) and cytosine (C) nucleotide bases with great affinity (19). If left unrepaired, guanine (G) to thymine (T) transversion mutations result (5, 14, 28). Additionally, the GO lesion can be further oxidised to produce more mutagenic DNA lesions such as cyanuric acid, oxaluric acid and oxazolone (19).

In organisms, the first lines of cellular defence against ROS are antioxidant molecules and systems that can neutralize the radicals *via* the actions of antioxidant enzymes, such as superoxidase dismutase, glutathione peroxidise and glutathione reductase (21, 24). In addition, organisms have multiple systems to repair DNA damage, and some of these have been identified as being involved in the repair of ROS-induced damage (discussed in section 1.4).

1.3 Classes of mutations

Mutations occur in a wide variety of types, and can be classified based on the sequence alteration they cause in DNA. Broadly, types of mutations include base pair substitutions, insertions, deletions and gross chromosomal rearrangements (5). The mutation spectrum of a DNA sequence refers to the range and relative frequencies of all the mutation types (25) relative to a reference sequence.

1.3.1 Base pair substitutions (BPSs)

Base pair substitutions (BPSs), sometimes referred to as point mutations, result from the substitution of one base pair for another (5). BPSs can be further classified into transitions and transversions (5). Transition mutations occur when a purine base (A or G) is replaced by another purine base or when a pyrimidine base (C or T) is replaced by another pyrimidine base (**Figure 1.1**). Transversion mutations occur when a purine base is substituted by a pyrimidine base or vice-versa (**Figure 1.1**). In bacteria, transition mutations are typically more common than transversion mutations (29), even though, theoretically, there are more possibilities for transversions than transitions. It is thought that this bias is due to the biochemical structure and nature of the nucleotide bases themselves (30). In essence, when a transition mutation occurs, a purine or pyrimidine nucleotide is substituted by another purine or pyrimidine nucleotide, respectively and thus, the biochemical structure and properties of the nucleotide being substituted are not changed. Another possible reason behind the prevalence of transition mutations is that transversions are more readily recognised and repaired by cellular repair systems as they cause bulky lesions, while transitions are not (30). Recently, Sakai et al. (2006) studied spontaneous mutations at one locus in *E. coli* cells grown under aerobic and anaerobic conditions and observed that the frequencies of BPSs were generally higher in aerobically grown cells (28).

BPSs can also be classified by their functional effect (5). For instance, silent or synonymous mutations are BPSs that result in the same amino acid being encoded for, and theoretically, should have no effect on protein function. Non-sense mutations are BPSs that change the amino acid codon to one of the three stop codons: TAG, TAA and TGA. The earlier termination of protein synthesis usually leads to a partial or complete loss of protein function, and therefore, such mutations are generally deleterious. Mis-sense mutations are BPSs that change the amino acid codon to that of a different amino acid. The effect of changing an amino acid can greatly vary and the resultant protein can lose, gain, change or maintain function (5).

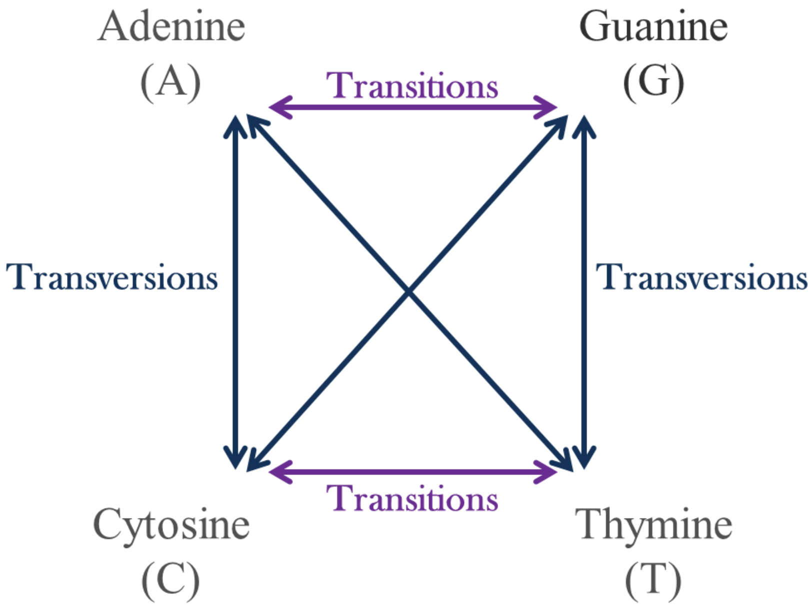

Figure 1.1. Transition versus transversion mutations. Transitions occur when a purine base (i.e. adenine or guanine) is replaced by the other purine base or when a pyrimidine base (i.e. cytosine or thymine) is replaced by the other pyrimidine base. Transversions occur when a pyrimidine base is replaced by a purine base or vice-versa.

1.3.2 Indels

Insertions or deletions, commonly also referred to as indels, involve the insertion or deletion of one or more nucleotide bases (5). Within coding regions, unless the length of the mutation is a multiple of three bases, insertions or deletions commonly result in frame-shift mutations. Frame-shifts generally result in the translation of a protein that is significantly different in sequence and usually not functional (5). Frame-shift mutations may also result in truncated proteins due to the introduction of stop codons in the gene sequence.

In this study, slippage events are also classified as indels. Replication slippage, also referred to as slipped-strand mispairing, normally occurs during DNA replication. Observed in both prokaryotic and eukaryotic cells, repetitive DNA sequences, present in the coding and non-coding regions of genomes, usually in runs of mono-, di- and trinucleotide repeats, are regions where slippage occurs. These sequences are usually unstable and additions or deletions of the repeated units can frequently occur. Viguera et al. (2001) have elucidated a model for the mechanism behind replication slippage, discussed in detail in (31).

1.3.3 Gross chromosomal rearrangements (GCRs)

Gross chromosomal rearrangements (GCRs) refer to mutations involving large-scale regions of the genome. Large-scale changes in the structure of the genome include deletions, inversions, duplications, translocations and mobile genetic element (MGE) movement (**Figure 1.2**). GCRs can generally be classified into two groups: balanced GCRs or imbalanced GCRs. Balanced GCRs are those rearrangements that change the gene order within the genome but do not remove or add any sequence (e.g. inversions and translocations). Imbalanced rearrangements are those that change the overall content of the genome (e.g. deletions and duplications) (32, 33).

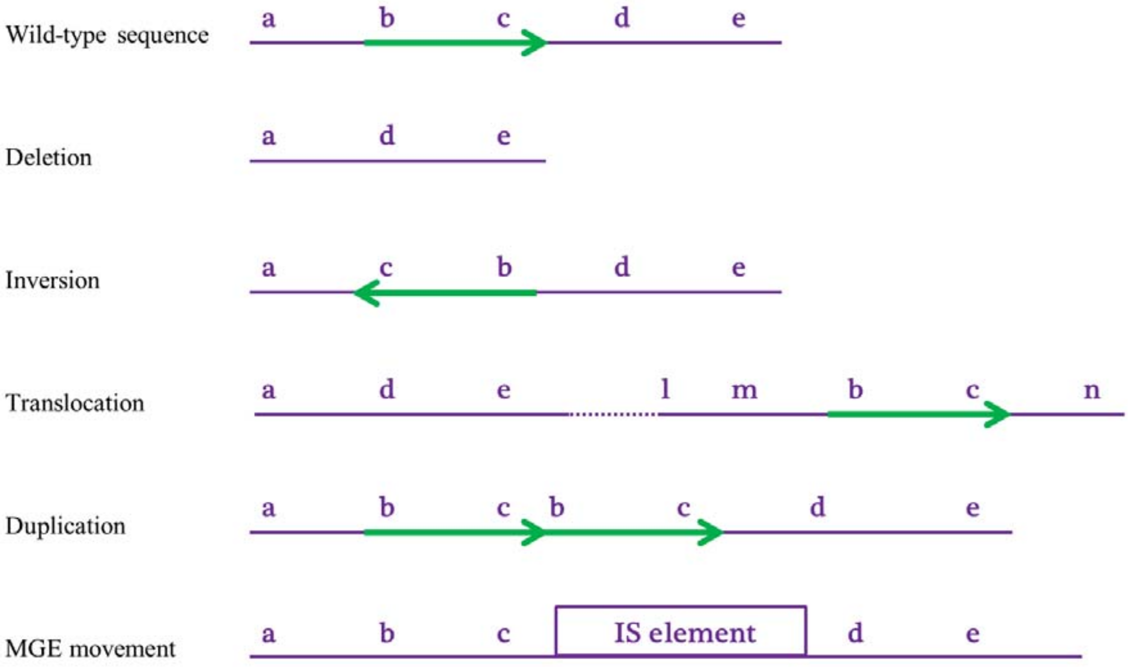

Figure 1.2. Examples of GCRs. Figure adapted from (32).

In evolution studies, GCRs are of particular interest as they can influence genome folding, gene expression, gene regulation, create new gene combinations and regulatory circuits, or even lead to the reduction in genome size by the loss of non-essential genes (34, 35). GCRs have been shown to occur during short-term adaptation to new environments in *E. coli* (35-37), *Saccharomyces cerevisiae* (38-40), and other organisms. While GCRs have been shown to confer increased fitness (41-43) in some studies, it is possible that potential fitness costs restrict the occurrences of certain GCRs in populations (33). Generally, where gene expression is affected or genome size is drastically changed as a consequence of GCRs, then the fitness of these genomes is reduced (44-46). Curiously, in the study of spontaneous mutations at one locus in *E. coli* cells grown under aerobic and anaerobic conditions by Sakai et al. (2006), GCRs were much more prevalent in anaerobically grown cells (28). However, as only a small locus was analysed in this study, it is not known if this trend generally occurs genome-wide (47, 48).

1.3.3.1 Formation of GCRs

GCRs can occur due to recombination between homologous, or partially homologous, repeated DNA sequences within the genome. Examples of such DNA sequences include ribosomal operons, tRNA genes, duplicated genes, prophages and MGEs (5). During genome replication, replication forks can stall in regions of sequence repeats, creating double-strand breaks (DSBs) in the DNA. Since these DSBs can be lethal, their repair is crucial. As recombinational repair systems (discussed in section 1.4.7) act to repair the DSBs, GCRs can be generated (33, 39).

Repeated sequences that are in the same orientation on the genome are called direct repeats (33). Recombination between direct repeats results in the duplication or deletion of the region bounded by the repeat sequences as seen in **Figure 1.3a** and **Figure 1.3b**, respectively. The deleted region can also insert itself back into the genome at another copy of the repeat, generating a translocation (33). For duplications, the duplicate regions can either be located next to each other, resulting in tandem duplications, or one of the duplicated regions can be inserted in a separate location of the genome, where the sequence is at least partially homologous. Repeated sequences that are in opposing orientations on the genome are called indirect repeats (IRs). Recombination between IRs can result in inversions of the regions in between the repeats as seen in **Figure 1.3c** (33). Thus, the occurrence of GCRs can be moderated by the number and size of repeated sequences in the genome (33). Repeated sequences most frequently involved in GCR generation are MGEs (32).

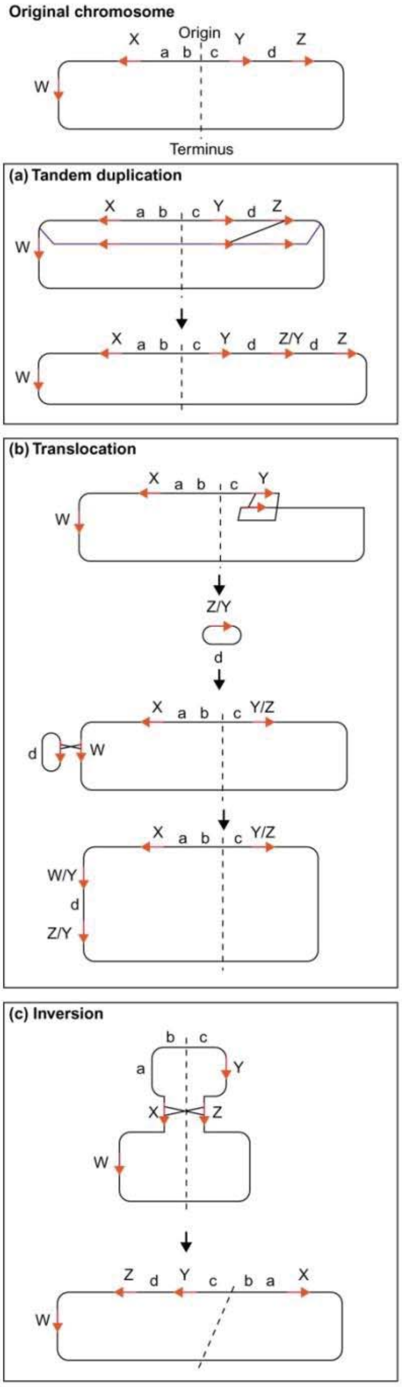

Figure 1.3. Homologous recombination of repeated sequences resulting in the formation of GCRs. The letters W, X, Y and Z represent the repeat sequences while the red arrows indicate the orientations of the sequences. The letters a, b, c and d represent regions of the genome bound by the repeat sequences. In a) recombination between repeat sequences Y and Z on a sister chromosome after replication results in a tandem duplication. In b) recombination between repeat sequences Y and Z on the same chromosome results in the excision of the DNA fragment which has gone on to recombine with another homologous sequence elsewhere in the genome, resulting in a translocation. In c) recombination between repeat sequences X and Z results in an inversion. Figure taken from (33).

1.3.3.1.1 Mobile genetic elements (MGEs)

MGEs are DNA sequences that can move within and between genomes (49) and allow the transfer of genetic information in prokaryotes and eukaryotes. MGEs have been shown to have a wide range of effects when they insert into new regions of the genome, such as adaptation to new environmental conditions and changes in genetic content, gene expression or genome structure (50). Therefore, as MGEs seem to contribute greatly to genome evolution (49, 51, 52), they are thought to be important players in changing the mode and tempo of evolution. It is noted that though there are instances where MGE movement can be beneficial to cells, such as in the adaptation to new environments, excessive MGE mobility can result in an unfavourable amount of mutagenesis (49).

Typically, MGEs encode enzymes that allow for their movement and integration into new DNA sequences, and also play a critical role in the spread of virulence factors and resistance conferring proteins amongst bacteria. Examples of MGEs include transposable elements (TEs), plasmids, bacteriophage elements and pathogenicity islands [see (53) for a complete review]. As the characteristics of many of these MGEs overlap, they can be difficult to categorize. Since TEs are the most abundant and well-studied type of MGEs, the focus of this section will be on TEs.

1.3.3.1.1.1 Transposable elements (TEs)

TEs were first described in the maize genome by McClintock in 1950 (54), and are sometimes referred to as “selfish DNA”. All TEs are flanked by a series of short repeats, either direct or indirect, which aid in their insertion and excision. The transposition of TEs around the genome is catalysed by element-specific enzymes, generally called transposases (53). In *E. coli*, two different pathways for transposition have been identified: replicative and conservative (55). Briefly, in the replicative pathway, transposition results in two copies of the TE, where one copy is at the new site even as the original site also retains a copy (56). In the conservative pathway, on the other hand, there is no replication of the TE. Instead, the TE is excised from the original site and integrated into a new site (53), in what is typically referred to as the “cut and paste mechanism”. TEs that can move by themselves are generally referred to as autonomous TEs while those that cannot mediate their own movement are referred to as non-autonomous TEs (53). Sites of insertion of TEs can generally be identified by the presence of a series of repeats, commonly of lengths ranging from 3 to 40 bp, which are generated during transposition (49).

TEs can largely be classified into two classes based on their structure and transposition mechanism: Class I and Class II (57). Class I TEs are also referred to as retrotransposons as they are first transcribed to RNA. The resultant RNA is then reverse transcribed to DNA in a step mediated by a reverse transcriptase which is often encoded for by the TE itself. The DNA copy can then be inserted into a new position in the genome (57). Conversely, Class II TEs, or DNA transposons, never use RNA intermediates and often follow a conservative mechanism of transposition. Depending on the mechanism of insertion, replication strategy and structure, TEs can then be further sub-divided within each class as detailed by Casacuberta and Gonzalez (2013) in (52). In *E. coli*, Class II TEs are more abundant than Class I TEs (57, 58). Examples of Class I elements include group II introns, retrons and diversity-generating retroelements (58) while examples of Class II elements include insertion sequence (IS) elements, Miniature inverted-repeat transposable elements (MITEs) and transposons (Tn) (53). IS elements are thought to be the most abundant bacterial TE (53).

1.3.3.1.1.1.1 IS elements

IS elements are the simplest types of TEs and range in size from 700 bp to 3,500 bp. Many different types of IS elements exist and they can be further categorised into families based on the type of encoded transposase, their overall genetic organisation and their accessory genes [discussed in detail in (59, 60)]. Most *E. coli* strains are rich in IS elements; over 40 IS elements have been identified amongst the various *E. coli* strains, with IS*3*, IS*4* and IS*5* being some of the more abundant elements. Transposition rates vary greatly between different IS elements and different *E. coli* strains, as demonstrated in recent studies by Sousa et al. (2013) (61) and Lee et al. (2014) (62).

While all IS elements typically only encode one to three proteins, a transposase is always encoded for. Thus, IS elements are autonomous TEs that are generally flanked by short sequences of IRs, with each IR being at least 3 bp long (57, 61). If IS elements are inserted into the middle of genes, the coding sequences of the genes are interrupted and gene expression is inactivated. If genes are within an operon, then IS element insertion into one gene can disrupt the expression of the other genes in the operon as well (53). IS element contributions in evolution have been well documented (60). In *E. coli*, for example, IS elements have generated a great degree of genetic diversity during laboratory evolution experiments, and aided in population adaptation to many different environments (35, 41, 50, 63-65).

1.3.3.1.1.1.2 MITEs and transposons

MITEs are non-autonomous TEs that are normally between 100 to 400 bp in length (66). MITEs, normally flanked by IRs of 10 to 40 bp on either side, are thought to have derived from IS elements (section 1.3.3.1.1.1.1) by internal deletions. MITEs normally possess the recognition sequences required for their mobility and thus, can be mobilised by the parent transposon when it is present (53, 66). While much is unknown about the mobility of these elements in prokaryotic genomes, it is thought that they are beneficial in generating new genes with new functions or in generating new regulatory signals for existing genes (66, 67).

Transposons are autonomous TEs, containing genes in addition to those required for transposition (57). These additional genes typically confer antibiotic resistance or virulence. Composite transposons are generally flanked by two IS elements and mostly utilise a conservative mode of transposition (section 1.3.3.1.1.1). The IS elements, which can be identical, have the ability to transpose independently of the transposon (57). Non-composite transposons, or complex transposons, are usually flanked by long sequences of IRs and use a replicative mode of transposition (section 1.3.3.1.1.1) (57). Over the years, many transposons have been identified in *E. coli*, examples of which include Tn3, Tn5, Tn7, Tn10 and Tn21 (53).

1.4 DNA replication and repair systems

DNA replication is typically a process of high fidelity, due to the action of accurate and efficient DNA polymerases. Some polymerases also possess proofreading ability and can hence, immediately correct mismatched nucleotide bases. Moreover, to maintain genome integrity and structure, organisms have multiple mechanisms for repairing mutated DNA and/or replicating past the damage in the sequence (5, 68, 69). DNA repair is generally defined as the cellular responses to DNA damage that result in the restoration of normal nucleotide sequence and DNA structure (5). DNA repair processes typically involve either the direct correction of the damaged bases or the excision of the damaged bases (14). While it is not always possible to reverse the damage, there are mechanisms in place to do so when applicable. In *E. coli*, for example, alkylated bases can usually be directly repaired by methyltransferases or oxidative demethylases (5, 70). Additionally, *E. coli* has many different DNA repair systems, some of which are discussed in more detail below.

1.4.1 DNA replication

In *E. coli*, DNA replication normally involves either polymerase (Pol) I or III (5, 71). DNA Pol III holoenzyme is presumed to be the main, highly accurate replicative polymerase involved in DNA replication (5, 71). Briefly, DNA replication involves unzipping the double-stranded DNA (dsDNA) and generating a forked structure that is referred to as the replication fork. As DNA synthesis can only occur in one direction (5' to 3'), the two DNA strands are differentially replicated. One strand is continuously replicated and is referred to as the leading strand while the second strand is replicated in short Okazaki fragments that are later linked together. This strand is referred to as the lagging strand. Studies have demonstrated that while Pol III can replicate both the leading and lagging DNA strands, lagging strand synthesis is more accurate (71). Reasons behind the differential fidelity of Pol III in replicating leading and lagging strands are currently unknown. Pol I, encoded by *polA*, is the most abundant DNA polymerase in *E. coli* and is involved in filling gaps between Okazaki fragments during lagging strand synthesis (5, 71). Both Pol I and Pol III has been shown to be prone to replication slippage (discussed in section 1.3.2) (72).

While DNA replication can be an error prone process itself, some DNA polymerases also possess exonuclease activity; providing them with proof-reading ability. For instance, the epsilon sub-unit of Pol III, encoded by *dnaQ*, possesses exonuclease function and mutations in *dnaQ* have been shown to impede the proofreading ability of DNA Pol III, leading to a reduced rate of growth and a strong mutator phenotype (section 1.6) (14). Pol I, in contrast, is the polymerase utilised by the various DNA repair systems. Thus, as the DNA replication process can repair damaged DNA, it also plays a crucial role in the maintenance of genomic integrity (14, 73). A list of proteins in *E. coli* known to be involved in replication is shown in **Table A.2**.

1.4.1.1 Translesion synthesis (TLS)

Under normal growth conditions, Pol III is constitutively synthesised. However, when DNA damage or a replication block is encountered, translesion synthesis (TLS) is induced. TLS allows for the replication of DNA past lesions or stalled replication forks by the use of specialized polymerases that can insert nucleotides across DNA lesions (74, 75). After the lesion has been by-passed, Pol III resumes DNA synthesis. Compared to Pol III, the TLS polymerases generally have lower fidelity and so, TLS can be an error-prone, mutagenic process itself (5).

In *E. coli*, there are three polymerases, mostly belonging to the Y-family of polymerases, that are involved in TLS; Pol II, IV and V. Pol II, while known to be involved in in re-starting stalled replication forks (76), also has proof-reading ability and has been postulated as being a backup polymerase for Pol III (71). Pol IV has been shown to bypass many lesions including alkylation and oxidatively damaged bases (5). Additionally, it has been shown to increase the cell's spontaneous mutation rate when induced (76). Pol V is the most error prone polymerase and can directly bypass most types of DNA damage (75, 77).

1.4.2 Base excision repair (BER)

The base excision repair (BER) pathway is a highly conserved DNA repair pathway that functions as the primary pathway to correct DNA damage induced by ROS (section 1.2.1) (5). Additionally, it can repair DNA damage caused by alkylation, deamination, depurination and depyrimidination (5, 15, 16, 70, 78). The BER pathway is initiated when DNA glycosylases recognise damaged bases and initiate damaged base removal *via* cleavage of the N-glycosidic bond (15, 78). Other proteins required for the repair process are an AP endonuclease or AP DNA lyase, a DNA polymerase, and a DNA ligase. Details of all genes known to be involved in this pathway in *E. coli* are provided in **Table A.2.**

At present, many different DNA glycosylase enzymes, each falling into different structural classes, are recognised (5). Each glycosylase is specialized in the detection and excision of different damaged bases. For the repair of oxidatively damaged bases, the MutM, MutY and MutT proteins are induced in response to the GO lesion, and comprise a system that is commonly referred to as the GO system (**Figure 1.4**). The formamidopyrimidine DNA glycosylase, or MutM, catalyses the removal of 8-oxoG bases paired with cytosine bases in a DNA sequence. If the lesion is not removed before replication occurs, adenine bases can be incorporated into the sequence, opposite the 8-oxoG bases (**Figure 1.4**), resulting in G:C → T:A transversions (78, 79). MutM also possesses the ability to repair alkylated purines, ring-opened purines and formamidopyrimidine lesions (78, 79). The adenine DNA glycosylase, or MutY, catalyses the removal of adenine bases paired with 8-oxoG bases in a DNA sequence (78, 79). If the lesion is not removed before replication occurs, cytosine bases can be incorporated into the sequence, opposite the 8-oxoG bases (**Figure 1.4**), resulting in A:T → C:G transversions (78, 79). Finally, MutT is involved in the hydrolysis of 8-oxo-dGTP to 8-oxo-dGMP, thus removing the oxidised guanine from the nucleotide pool (23, 80).

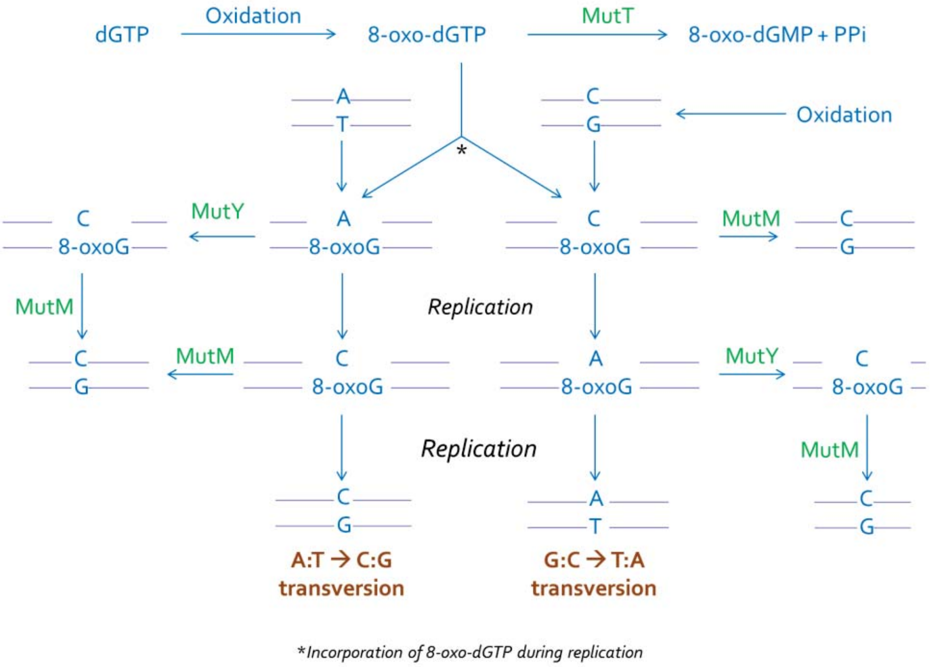

Figure 1.4. Proposed model of the GO system. If left unrepaired, the GO lesion can result in changes in the DNA sequences. The oxidised guanine base can be incorporated opposite either adenine or cytosine nucleotide bases during DNA replication. Replication of the 8-oxoG:C pair can result in the incorporation of adenine bases into the sequence, subsequently resulting in G:C → T:A transversions. Replication of the 8-oxoG:A pair can result in the incorporation of cytosine bases into the sequence, subsequently resulting in A:T → C:G transversions. MutM and MutY are involved in preventing these mutations from occurring, respectively, while MutT is involved in catalysing the removal of the 8-oxo-dGTP from the nucleotide pool. Figure adapted from (81).

As previously mentioned, the BER pathway also contains many damage-specific DNA glycosylases that can repair alkylated or deaminated bases (5, 79). Uracil, another frequently encountered DNA lesion regularly induced by the hydrolytic deamination of cytosine bases, preferentially pairs with adenine during replication. If the lesion is not removed by the uracil DNA glycosylase (encoded by *ung*) before replication occurs, C:G → T:A transitions result (5, 79). Endonuclease V, encoded by *Nfi*, recognises deaminated adenine and guanine nucleotide bases. If the lesions are not removed before replication occurs, A:T → G:C transitions result (5, 79). Additionally, ring-saturated, ring-contracted, fragmented and oxidised pyrimidines are recognised by Endonuclease III, which is encoded by *Nth* (5, 79).

1.4.3 Nucleotide excision repair (NER)

Nucleotide excision repair (NER) is the major DNA repair pathway for repairing bulky damages such as alkylated bases, UV lesions, pyrimidine dimers, helix distortions or benzoapyrene adducts (75, 82-84). These bulky lesions tend to block the progression of DNA replication forks or cause a distortion in the DNA structure, resulting in mutations. Damaged DNA nucleotides are excised *via* dual incisions made by the exinuclease (two different nucleases functioning together) (5, 82). Additionally, it is possible that the NER pathway functions as a back-up pathway for the repair of oxidative DNA damage (82, 85). Briefly, in studies conducted on human cell lines of xeroderma pigmentosum, cells unable to repair neuronal damage caused by singlet oxygen were found to have mutations in genes encoding proteins involved in the NER pathway (82, 85). Details of all genes known to be involved in this pathway in *E. coli* are provided in **Table A.2**.

1.4.4 Mismatch repair (MMR)

The mismatch repair (MMR) system is a highly conserved system involved in DNA repair. The MMR system corrects base-base mismatches that have arisen from mis-incorporation during replication as well as indels that have arisen from strand slippage. In short, the main function of MMR is to repair DNA damage that has escaped the proofreading of DNA polymerases (5). The MMR pathway prevents mismatches by detecting incorrect base-pairing between DNA strands and as it is the only system to specifically detect and correct errors on the newly synthesised strand, MMR is an essential repair system to maintain genome fidelity (83, 86-88). Details of all genes known to be involved in this pathway in *E. coli* are provided in **Table A.2**.

In *E. coli*, the methyl-directed MMR pathway can recognise and repair BPSs and indels and is comprised of proteins encoded for by the *mutS*, *mutH* and *mutL* genes (89). In methyl-directed MMR, parental DNA is methylated by the *dam* encoded DNA adenine methyltransferase while newly synthesised DNA is unmethylated. MutS and MutL bind together and activate MutH, which preferentially acts on the unmethylated strand of DNA at hemi-methylated GATC sites (89, 90). If the site of excision is 3' to the mismatch, Exonuclease I, encoded by *sbcB*, or Exonuclease X, encoded by *exoX*, are utilised. If the site of excision is 5' of the mismatch, Exonuclease VII, encoded by *xseA*, or RecJ, encoded by *recj*, are utilised (89, 90). The process is facilitated by a DNA helicase, UvrD, which is also involved in re-starting DNA replication from the lesion (89, 90).

While studies have indicated a role of the MMR pathway in the repair of oxidatively damaged bases (91), studies have also shown that the methyl-directed MMR pathway preferentially recognises transitions (90). Indeed, Lee et al. (2012) have also found a bias for A → G and T → C transitions in aerobically grown *E. coli* cells lacking MutL (13). Transitions are thought to occur due to the different pairing properties of the structural isomers of the nucleotide bases. Essentially, the four nucleotide bases of DNA are subject to spontaneous structural alterations called tautomerization. Each of the four bases exist in two forms, with each form having its own affinity; guanine and thymine bases can exist in enol or keto forms and adenine and cytosine bases can exist in amino or imino forms (5). For instance, the imino tautomers of adenine bases can pair with cytosine bases, resulting in A:T → G:C transitions whereas the enol tautomers of guanine bases can pair with thymine bases, resulting in G:C → A:T transitions (5).

The methyl-directed MMR pathway of *E. coli* can also recognise recombination heteroduplexes that contain mismatches (90). The MMR pathway has been shown to inhibit the completion of homeologous recombination (92), and has also been associated with the inhibition of homologous recombination (93). Homeologous recombination is the recombination between related but non-identical DNA sequences, and can cause substantial genome instability that can lead to GCRs. Therefore, it is not surprising that mutator strains that arise in experimental evolution studies are usually defective in MMR (9).

1.4.5 SOS response

The SOS response is a global regulatory network response, induced during conditions of extensive stress caused by exogenous and endogenous agents to enable bacterial survival (69, 76, 94). More than 40 genes are thought to be induced in the SOS response, with most of these genes encoding proteins for DNA protection, repair, replication and metabolism (75, 94). Though this mechanism is highly error prone, it can be advantageous in an evolutionary sense as cells will attempt to repair the genome and risk fixing deleterious mutations, as opposed to leaving the genome unrepaired. Moreover, the additional genetic and phenotypic variation generated may also be beneficial in helping the organism adapt to the stressful conditions. For examples of genes in *E. coli* known to be induced by this pathway, refer to **Table A.2**.

Two major proteins regulate the SOS response: RecA and LexA (69, 94). LexA is a transcriptional repressor of SOS genes. When DNA is damaged, RecA binds to single-stranded DNA (ssDNA), signalling the self-cleavage of LexA, resulting in the expression of the SOS response (69, 75, 76). Studies have shown that the presence of ROS can induce LexA self-cleavage in *E. coli* (22), thus indicating that the SOS response can be induced in the presence of oxidative DNA damage.

1.4.6 Stringent response

The stringent response is another stress response, induced in response to nutrient starvation, fatty acid limitation and heat shock (95). Many genes are thought to be induced in the stringent response, with most of these genes encoding proteins for stress-associated sigma factors, RNA degradation and amino acid biosynthesis (95, 96). For examples of genes in *E. coli* known to be induced by this pathway, refer to **Table A.2**.

The stringent response is induced by the accumulation of unusual guanosine tetra- and penta- phosphates, referred to as (p)ppGpp nucleotides, and is regulated by two enzymes*: RelA* and *SpoT* (95, 96). RelA, also known as ppGpp synthetase I, synthesises (p)ppGpp nucleotides in response to amino acid starvation while SpoT, also referred to as ppGpp synthetase II, maintains the intracellular levels of the nucleotides. The stringent response has numerous effects on cell physiology, and studies have suggested that the stringent response can be induced in the presence of oxidative DNA damage (97, 98).

1.4.7 Recombinational repair

When DNA replication is stalled, either in response to exogenous mutagenic agents or due to DNA damage, recombinational repair pathways are induced (5, 68, 69). Essentially, when the replication fork encounters unrepaired lesions or DNA strand breaks at DNA secondary structures undergoing repair, the replication complex halts. These lesions or DNA strand breaks are then repaired by enzymes involved in recombination, after which replication is re-initiated (5). While these repair mechanisms contribute greatly towards the maintenance of genome fidelity, they are also responsible for generating genome instability, as they allow for the exchange of genetic information between DNA sequences. For example, GCRs (section 1.3.3) are largely mediated through homologous recombination between repeated sequences i.e. between homologous sites. Details of all genes known to be involved in this pathway in *E. coli* are provided in **Table A.2**.

In bacteria, there are two major repair recombination mechanisms: homologous and illegitimate recombination (53, 99). While some bacteria also have a non-homologous end joining (NHEJ) pathway (99), this pathway is thought to be absent in *E. coli* (100). Illegitimate recombination requires DNA sequences that are in close proximity to each other but that are not largely homologous (53). Essentially, this mechanism relies on DNA strand annealing and is RecA independent (99). Homologous recombination, on the other hand, is the main mechanism for recombinational repair in bacteria.

1.4.7.1 Homologous recombination

Homologous recombination requires identical, or very similar, DNA sequences. It is commonly thought that a minimum of 20 bp of DNA sequence homology is required for recombination to occur (53). This mechanism can repair double-strand DNA breaks (DSBs) and single-stranded DNA breaks (SSBs) *via* the RecBCD and RecFOR (68, 101) pathways, respectively.

1.4.7.1.1 The RecBCD pathway

The RecBCD pathway is the main recombination pathway in *E. coli* (53). In this pathway, the three-enzyme RecBCD complex initiates the recombination process by binding to a blunt end of the dsDNA break (102). The DNA is then unwound and the two DNA strands are asymmetrically degraded until the RecBCD complex encounters a specific eight nucleotide sequence; the Chi site (5'-GCTGGTGG-3') (53, 102). Upon encountering a Chi site, the exonuclease activity of the complex is up-regulated and RecA proteins are loaded onto the ssDNA. This DNA-protein complex then searches out similar sequences of DNA, after which strand invasion and strand exchange follow (53, 102). Finally, the resultant crossover junctions, referred to as Holliday junctions, are resolved by some combination of RuvABC and/or RecG proteins and two recombinant, or non-recombinant, DNA molecules are produced (102). Generally, in *E. coli* genomes, there is an over-abundance of Chi sites (102, 103), indicating the importance of homologous recombination.

1.4.7.1.2 The RecFOR pathway

The RecFOR pathway can repair breaks that occur on only one DNA strand i.e. SSBs. The RecFOR pathway has also been demonstrated to repair DSBs when the RecBCD pathway is inactive (104). The RecFOR pathway follows a similar mechanism to that of the RecBCD pathway, though different proteins are involved (53, 68, 102). Essentially, the RecFOR complex loads the RecA proteins onto the ssDNA *via* the displacement of the single-strand-binding (SSB) protein. As was the case with the RecBCD pathway, the RecFOR pathway requires RecA to function.

1.4.8 Expression of repair pathways in aerobic and anaerobic environments

It is likely that different mutations will accumulate under aerobic and anaerobic environments due to the presence and action of different mutation sources (section 1.2). Thus, it is also possible that different repair systems are active under aerobic and anaerobic environments. This can be investigated by the comparative analyses of transcriptomes under both aerobic and anaerobic environments.

1.4.8.1.1 Transcriptome analysis

The transcriptome is composed of messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), and intragenic and intergenic non-coding RNA (ncRNA) (105-

108). Until very recently, hybridization-based approaches, commonly microarrays, were typically used to study the transcriptome. While these methods were reliable, rapid, and cost effective, they suffered from background noise and cross-hybridization issues. Hybridization-based approaches also required an annotated genome sequence, which was not always available. However, it is now possible to study the transcriptomes of organisms by directly sequencing mRNA, a process commonly referred to as RNAseq (105-108).

RNAseq is an efficient way to measure the transcriptome with increased dynamic range, greater sensitivity and deeper coverage than microarray technologies, providing more in-depth insight into transcriptional regulation, and hence, the biology of the cell (105, 106, 109-113). RNAseq involves either double stranded or single stranded cDNA library synthesis (113, 114). Double-stranded cDNA synthesis erases RNA strand information while single-stranded cDNA synthesis retains information about the direction of transcription (112). Various strategies for strand-specific analyses of transcriptomes have been developed as transcriptome studies of many species have revealed a previously uncharted existence of antisense transcription events. Due to its great resolution, RNAseq can also detect previously unknown genes and non-coding RNAs. In fact, it is now clear that antisense transcripts are very heavily involved in regulatory functions (108, 115-117).

Studying the transcriptomes of bacteria can be a challenging task. Bacterial mRNA can be particularly difficult to isolate as it is rather unstable and has a very short half-life, an average of 2 min (107, 110). As approximately 95% of the transcriptome is composed of rRNA and tRNA, it is ideal to enrich for mRNA and remove rRNA and tRNA from the sample for the production of high-quality transcripts. However, enrichment of bacterial mRNAs is complicated as they generally lack a 3' poly (A) tail, which facilitate enrichment in eukaryotes (109, 111). Despite these challenges, RNA sequencing has already been used to study the transcriptomes of many microbes including *Pseudomonas syringae* (108), *Streptococcus pneumoniae* (114), *Salmonella bongori* (114) and *Chlamydia trachomatis* (115). In such microbial RNAseq studies, mRNA enrichment has typically been achieved by the removal of rRNAs and other RNAs by using a hybridisation-based method, the degradation of processed RNA by a

5'–3' exonuclease that specifically degrades RNAs with a 5'-monophosphate end or by the selective polyadenylation of mRNAs by the *E. coli* poly(A) polymerase (109).

1.5 Mutation detection

Mutation detection is an area that has developed significantly in recent times; an overview of the various methods that have been used in the past are covered in this section. In many cases, molecular methods such as fluorescence *in situ* hybridization (FISH) (118), polymerase chain reaction (PCR) amplification and restriction fragment length polymorphism (RFLP) analysis of non-coding regions and microsatellite regions (119), denaturing gradient gel electrophoresis (DGGE), single-stranded conformational polymorphism (SSCP) assays (120), and most commonly, sequencing of PCR amplicons (4) have been used to detect mutations. Additionally, physical mapping by pulse-field gel electrophoresis (PFGE), PCR amplification and sequencing of rearrangement break points or comparative genome hybridisation (34) have also been used to identify GCRs (section 1.3.3). Many of the different methods listed above rely on one of three simple approaches: (i) differential hybridisation of wild-type and mutant DNA in DNA microarrays, (ii) differences in the electrophoretic properties of wild-type and mutant DNA, or (iii) the chemical/enzymatic modification of heteroduplex DNA (120). These methods generally provide a limited view of the mutational spectra by focussing on a limited number of genetic loci, or, do not provide the molecular detail of the changes in DNA sequence underlying the mutation. In recent years, however, the ease and availability of sequencing technologies has greatly advanced (section 1.5.1.1) and it is now possible to identify mutations directly across the whole genome.

1.5.1 Genome-wide mutation detection

Detecting mutations on a genome-wide basis provides a complete picture of the mutations that have arisen, allowing for a better understanding of the mutation spectrum as well as a direct molecular estimate of the mutation rate. A powerful approach to detect all mutations that have arisen within a genome is to re-sequence its genome (section 1.5.1.1). Then, by comparing the ancestral genotype with the genotypes of evolved populations, mutations can be identified across the whole genome.

1.5.1.1 Genome-wide mutation detection by genome sequencing

For many years, Sanger sequencing was the most widely used sequencing method. However, with the advent of high throughput next-generation sequencing technologies, it is now possible to sequence bacterial genomes both rapidly and affordably (4, 121, 122). These high throughput sequencing platforms are the 454 pyrosequencing, Illumina and Sequencing by Oligonucleotide Ligation and Detection (SOLiD) platforms. One of the first next-generation sequencing platforms, 454 pyrosequencing relies on the generation of a detectable light signal when a new nucleotide base is incorporated (123). The Illumina sequencing platform depends on fluorescently labelled, reversible dye-terminators to identify nucleotide bases (124) while the SOLiD sequencing platform depends on sequencing by ligation, where emulsion PCR, fluorescent probes and fluorescence imaging are used to identify nucleotide bases (125). Another commonly used sequencing technology is Ion Torrent sequencing, which utilises a principle similar to that of 454 pyrosequencing, but ultimately relies on the detection of hydrogen ions during nucleotide base incorporation (126). Additionally, when using this platform, no imaging technology is required, thus making it a very fast sequencing platform.

There are also relatively newer technologies that have the potential to overcome some of the limitations and biases of the next-generation sequencing technologies listed above (127). One such platform, the Pacific Biosciences (PacBio) platform, utilises single-molecule real-time sequencing; where phosphate-labelled nucleotide bases are incorporated into the growing DNA strands and a zero-mode waveguide detector visualizes strand synthesis (128). One very noteworthy feature of this platform is that reads of long lengths, typically between 10 to 15 kb, are generated. Other sequencing technologies that are currently not as widely used include Polony sequencing (129) and the Helicos BioSciences platform (130).

In order to detect mutations mediated by, or involving, large repetitive sequences like TEs, appropriate sequencing strategies need to be undertaken. For instance, each sequencing platform has its advantages and disadvantages (131-133), and these factors must be carefully considered when choosing a sequencing platform. Additionally, when using these technologies, single-read, paired-end and mate-pair sequencing can be performed. Single-read sequencing involves sequencing a linear nucleic acid fragment from only one end while paired-end sequencing involves sequencing the fragment from both ends (134). Paired-end sequencing, where fragments typically range in size from 100 bp to 800 bp, can greatly improve the accuracy of reference-based mapping i.e. where reads are mapped to a fully annotated reference genome sequence of a closely related organism. Mate-pair sequencing, on the other hand, refers to the method by which the sequencing library was constructed. Essentially, to create mate-pair reads, nucleic acid fragments of certain sizes are circularised before they are fragmented again (132, 134). This circularisation step allows for larger fragment sizes to be sequenced, as regions on the genome that were previously distant from one another are brought into close proximity. These larger fragments, typically ranging in size from 2 kb to 5 kb, are sequenced from both ends. Mate-pair sequencing is ideal for *de novo* genome assembly (i.e. genome assembly) without the aid of a fully annotated reference genome sequence. Additionally, mate-pair sequencing greatly aids in resolving repeated DNA sequences and thus, in elucidating GCRs (135). Alternatively, some of the newer sequencing technologies, like the PacBio platform, with the much longer reads, detect GCRs more accurately. As a result, many of the mutation rate studies published to date (13, 136, 137) have not investigated GCRs. This is discussed further in section 3.2.3.2.2.1.

Numerous studies have already used whole genome re-sequencing to identify mutations (138), generally indicating that whole genome re-sequencing greatly improves the efficiency of mutation detection and thus, provides a more comprehensive depiction of the mutation rate and spectrum. Indeed, direct estimates of the mitochondrial mutation rate and spectrum from 74 *Caenorhabditis elegans* MA lineages indicated that the mutation rate was 10-fold greater than estimated from previous indirect phylogenetic analyses (47, 48) and that the rate of insertions was higher than the rate of deletions (48). In the case of *E. coli*, Lee et al. (2012) used whole genome re-sequencing to estimate the mutation rate from their MA study where cells were grown under aerobic growth conditions (13) and obtained the most inclusive estimate of the genome-wide mutation rate to date.

1.6 Mutation rates

The rates that mutations occur at mediate the dynamics of the genetic structures of organisms, and so, play a central role in the tempo of evolution (139). Mutation rates can generally be described as either substitution rates or spontaneous mutation rates (6, 140). The substitution rate refers to the rate at which mutations accumulate in a lineage and become fixed in a population, and are generally the product of selective pressure and genetic drift (6). The spontaneous mutation rate, on the other hand, refers to the frequency of new mutations that arise per generation in an organism (6, 18). Therefore, the mutation substitution rate is generally not the same as the spontaneous mutation rate. Substitution rates are typically easier to determine, and many studies measure substitution rates (136, 141-144), however, this thesis will primarily focus on spontaneous mutation rates.

Spontaneous mutation rates have been found to vary by large orders of magnitude across different organisms (145). The highest spontaneous mutation rates measured to date have been for RNA viroids, while the lowest spontaneous mutation rates measured have been for humans (**Figure 1.5**). In general, simple organisms with relatively small genome sizes tend to have higher spontaneous mutation rates, whereas those organisms with larger genomes tend to have lower mutation rates (146). Even across bacteria, mutation rates have been shown to vary widely between strains and species (147). Additionally, mutation rates have been found to greatly vary with environment (28), stressful cellular conditions, and antibiotic exposure (138), likely due to the different mutation pressures that result from the different conditions.

The term mutator is often used to describe mutants that have an increased spontaneous mutation rate (81). Mutator strains readily arise in natural and laboratory populations of bacteria and commonly result from mutations in DNA repair genes (9, 81). While the fitness of these strains is expected to be relatively lower due to their higher mutational load, mutator strains in bacteria are also advantageous in an evolutionary sense as they have a relatively increased probability of acquiring beneficial mutations (9). Hence, mutator strains can increase the rate of evolution of bacterial populations. Generally, the evolutionary success of organisms relies on their adaptability to constantly changing environmental conditions. If mutation rates are high, it is possible that the organism can adapt more quickly to changes in the environment, even though they likely experience more harmful mutations (9). On the other hand, organisms with lower mutation rates may find it more difficult to adapt to new conditions.

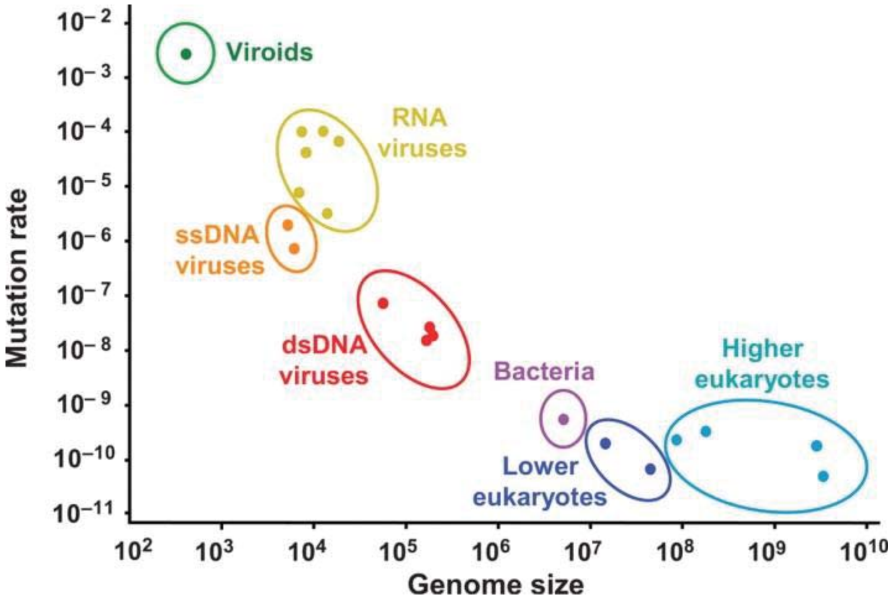

Figure 1.5. Mutation rates per nucleotide versus genome size for different organisms. Figure taken from (146).

1.6.1 Mutation rate measurement

Measuring the rate of mutation has proven to be a rather challenging task as the rate can be influenced by sequence context, complicating studies (4, 148). Nevertheless, throughout the years, methods for mutation rate measurement have developed considerably, and improved in accuracy as the availability of new genetic technologies has progressed. This section will discuss the various methods that have been used for mutation rate inference and direct measurement.

1.6.1.1 Relative rates *via* fluctuation assays

One commonly used approach to estimate mutation rates is to monitor the occurrence of phenotype mutants. Thus, mutation rates in microbes were originally studied *via* fluctuation assays. First described in 1943 by Luria and Delbrück (149), fluctuation assays involve inoculating a small number of cells into many parallel cultures, and then allowing these cultures to grow until saturation under non-selective conditions (149). The cultures are then grown on a selective, solid medium on which only spontaneous mutants that have arisen during the non-selective cultivation can grow. The distribution of the number of mutants among the parallel cultures is then used to calculate the mutation rate, with the assumption that mutants arise following a Poisson distribution. There are many other underlying assumptions associated with fluctuation assays and this has led to the design of many statistical methods for the purpose of determining better rate estimates (17, 18, 150-154). Fluctuation assays have been used to estimate mutations rates for aerobically grown *E. coli*, with estimates ranging from 4.3×10^{-10} (13) to 1.00×10^{-9} (155) mutations per locus per generation. However, as these estimates are dependent on the locus being analysed, it is possible that these rates do not generally represent the genome-wide mutation rate (47, 48). Currently, there are no published estimates of mutations rates determined *via* fluctuation assays for anaerobically grown *E. coli*.

Generally, mutation rates calculated from fluctuation assays are underestimated as mutations that do not produce a mutant phenotype do not contribute to the mutation rate calculation (4). However, these are technically simple and rapid assays to determine the relative mutation rate differences between organisms, and require no knowledge of the genetic changes underlying the phenotypic change being measured.

1.6.1.2 Inference of rates *via* phylogenetic analyses

A common molecular method to determine mutation rates in many organisms involves using phylogenetic analyses of naturally occurring populations to calculate the rate of mutation at selectively neutral sites (156). This method is especially beneficial in estimating mutation rates of extinct or fossilized specimens, where DNA from an ancestor is available and the age of the ancestor can be determined by using methods such as carbon dating. Additionally, this method has been used to measure mutation rates in laboratory evolution experiments of many organisms (157). For instance, by using this approach, Wielgoss et al. (2011) estimated a mutation rate of 4.1×10^{-4} mutations per genome per generation for populations of *E. coli* that had evolved for up to 40,000 generations under aerobic conditions (136). However, to date there are no published estimates of mutations rates determined *via* phylogenetic analyses for anaerobically grown *E. coli*.

The accuracy of this method is often debated as mutation rate estimates may be biased by the effects of natural selection acting on sequences that are assumed to be neutral but are actually not (4). For instance, Drake (2012), argues that the rate (136) calculated by Wielgoss et al. (2011) is a product of selection acting on the codon usage (158) of the populations rather than the best available *E. coli* spontaneous mutation rate (159).

Furthermore, the true number of generations separating the compared species is usually unknown and homoplasy (i.e. the similarity of traits in species from different ancestors due to convergent evolution) can also become a problem in determining true phylogeny in some cases (4, 47, 48).

1.6.1.3 Direct detection of rates *via* mutation accumulation (MA) assays

Mutation accumulation (MA) studies can be used to directly measure spontaneous mutation rates, and are performed using a population bottlenecking strategy (3). A population bottleneck is an event that drastically reduces the population size, and so, reduces the genetic diversity of the population and improves the probability of genotypes becoming fixed (6). A more accurate method to determine spontaneous mutation rates, the principle of MA experiments is to eliminate the effects of natural selection on the population, while revealing the effects of genetic drift by allowing mutations to be fixed or lost at random (3, 17, 18). In MA experiments, all new mutations (except highly deleterious ones) are allowed to accumulate within populations, irrespective of their effect on fitness, by passing the populations frequently through bottlenecks, ideally of a single organism at every generation (**Figure 1.6**). To estimate the spontaneous mutation rate in MA studies, the number of mutations present in independent populations can simply be counted (3), and expressed per unit time.

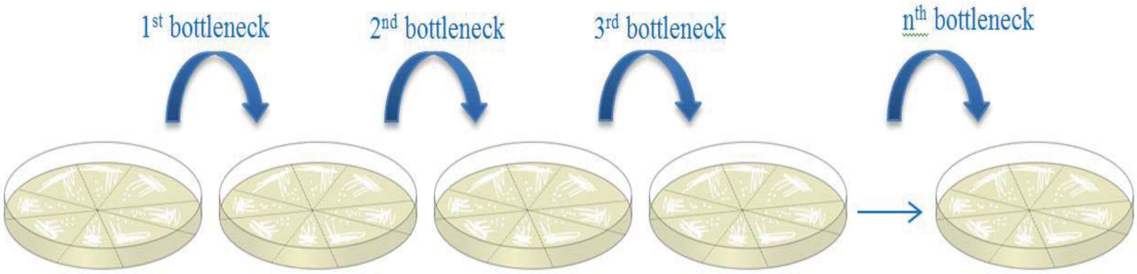

Figure 1.6. Mutation accumulation (MA) studies of bacteria. In MA studies of bacteria, frequent population bottlenecks are achieved by randomly picking single colonies that grow on agar plates to establish new populations. In this way, the effects of natural selection are reduced while the effects of genetic drift are enhanced and mutations can become fixed, irrespective of their effect on fitness.

Although it is not technically feasible to bottleneck bacteria at every generation due to their small size and rapid generation time (137, 160), MA studies have been conducted on many organisms including *E. coli* (3, 12, 13, 161-163). For *E. coli*, rates determined *via* MA assays have varied with different experimental conditions. For instance, Kibota and Lynch (1996) conducted MA assays using *E. coli* REL606 grown on Davis minimal agar plates under aerobic conditions. By measuring the fitness impact of these mutations, they reported a rate of 1.7 × 10^{-4} mutations per genome per generation (12). On the other hand, in a recent MA study conducted by Lee et al. (2012), a mutation rate of 1.0 × 10^{-3} mutations per genome per generation was determined by using whole genome re-sequencing (section 1.5.1.1) on *E. coli* MG1655 cultivated on LB agar under aerobic growth conditions (13). Not surprisingly, the relative mutation rates of *E. coli* obtained *via* MA assays differ from mutation rates of *E. coli* obtained *via* fluctuation assays (section 1.6.1.1) and phylogenetic assays (section 1.6.1.2). Apart from the different units of measurement and the varying contribution of selection in the three different methods, numerous differences in the experimental growth conditions and in the methods used to detect mutations adequately account for the different mutation rates determined from these studies. To date, there are no published estimates of mutations rates determined *via* MA assays for anaerobically grown *E. coli*.

1.7 The role of GCRs in adaptive evolution

While GCRs (section 1.3.3) have been observed in bacterial populations, little is known about their relative importance to adaptive evolution in bacteria. Adaptive evolution is the evolutionary process where organisms accumulate genetic changes that help them adapt to their environment, leading to an increase in their fitness in the same environment (6). Adaptive evolution can cause a species to gain, lose or modify a function (164) and is driven by selection for beneficial mutations while deleterious mutations are usually selected against.

As mentioned previously, Sakai et al. (2006) found that GCRs occurred twice as frequently at a small locus in anaerobically grown *E. coli* cells as opposed to aerobically grown cells (28). Thus, it is possible that GCRs could be advantageous mechanisms allowing for adaptation to anaerobic environments. Therefore, to obtain more information about the contribution of GCRs to adaptive evolution, experimental evolution (section 1.7.1) techniques can be used.

1.7.1 Experimental evolution

Experimental evolution allows for the process of evolution to be observed in the laboratory as populations, founded by the same ancestor, adapt and grow in new, controlled environmental conditions (6, 165-167). In these studies, as both the starting genotype and descending genotypes can be determined, it is possible to genetically compare descendants to their ancestor. Experimental evolution studies also allow for the investigation of the reproducibility of evolutionary outcomes, making it a very powerful tool (166). While experimental evolution has been carried out with a number of organisms to test many different hypotheses (165), bacteria remain excellent organisms to use for such studies. Advantages of using bacteria over other organisms include their growth to large population sizes, their asexual reproduction, their quick reproduction times, their small genome sizes, their ease of cultivation and their ease of storage for later examination (138, 166, 168, 169). Advances in DNA sequencing technologies (section 1.5.1.1) have now made it possible to identify whole-genome genetic changes between the ancestor and its derivatives (170) and to correlate observed phenotypes with specific genotypes (171).

MA assays (discussed in section 1.6.1.3) are examples of experimental evolution studies. Furthermore, experimental evolution experiments studying adaptive evolution are frequently conducted by periodic serial passaging of populations in flasks or multi-well plates (6, 169, 171). Briefly, serial passaging involves transferring a proportion of the culture to fresh medium for another round of growth at regular intervals. In these batch culture cultivations, the population densities vary over time and environmental factors (e.g. pH and dissolved oxygen) cannot be precisely controlled as they change as the culture ages within each batch (169, 171). Many adaptive evolution studies have been performed using bacteria, especially *E. coli* (167, 169, 171). Examples of selection pressures that have been applied to different organisms in adaptive evolution studies include temperature fluctuations (172), pH shifts (173), oxygen concentrations (174, 175), antibiotic exposure (176, 177) and nutrient limitation (38, 119, 168, 178, 179). Overall, exposure to conditions that are different to those that are normally encountered (i.e. stressful cellular conditions) has been found to increase the mutation rates of organisms (76, 147, 180). In these experiments, BPSs (section 1.3.1) are the most frequently identified mutations, possibly due to technical difficulties in detecting mutations like GCRs (138). A landmark study in experimental evolution is the ongoing

E. coli long-term evolution experiment (LTEE) started by Richard Lenski in 1988 (181).

1.7.1.1 Long-term evolution experiment (LTEE)

The LTEE is comprised of 12 *E. coli* populations, all derived from the same cell, that have been growing in flasks for over 60,000 generations in a glucose-limited environment. The evolutionary developments of these populations in this time have been closely monitored. Findings include an increase in cell size (182), the development of mutator populations (155) with biases towards A:T → C:G transversions (183), changes in the cell shape of some populations to become rounder (184) and a steady increase in the relative fitness of the populations (185). Notably, one population has even evolved the ability to utilize citrate as a source of energy, an ability wild-type *E. coli* does not possess under aerobic conditions (43, 186). Genomes of clones from these bacterial populations have been re-sequenced, such that many of the mutations acquired by the bacteria as they evolved have now been identified (187). Briefly, after 40,000 generations, the bacteria had acquired a total of 653 mutations, comprised of 627 BPSs and 26 indels. The mutations underpinning the evolution of citrate utilisation have also been identified (35, 43, 186, 188).

Many IS-mediated mutations, such as inversions and IS element transposition, have also been detected in these populations (38). Some of these mutations contributed to adaptation by modifying the global regulatory network while others improved fitness in the minimal glucose medium, providing proof that IS elements can promote adaptation to the environment (63, 65). Raeside et al. (2014) recently used whole-genome sequencing, optical mapping and PCR analysis to investigate the GCRs in clones from each population after 40,000 generations (35). In total, over 110 GCRs were detected, with large IS-mediated deletions being the most frequent type of GCR.

1.7.2 The *sbcC* gene

Proteins involved in DNA replication, repair, recombination and telomere maintenance may play crucial roles in regulating occurrences of GCRs (33). Such proteins include the DNA-specific SbcCD complex in bacteria. Located within the *sbcDC* operon, expression of both the *sbcC* and *sbcD* genes in *E. coli* was found to be dependent on RpoS activity under starvation conditions (189). In the SbcCD complex, SbcD is the nuclease subunit while SbcC is the ATPase subunit that modulates the nuclease activity of SbcD (190). Essentially, the SbcCD complex has ATP-dependent dsDNA exonuclease activity and ATP-independent ssDNA endonuclease activity (189, 191, 192). SbcC, a member of the structural maintenance of chromosomes (SMC) family of proteins, has been associated with replication forks (189, 193) while overexpressed SbcD has been found in the cytoplasm of *E. coli* cells (189).

The SbcCD complex is generally thought to play an important role in maintaining genome structure (194), especially as it has been shown to enhance DNA Pol IV activity (191). During DNA replication, the SbcCD complex removes proteins bound to DNA ends to generate DSBs, which are then repaired by the recombinational pathways discussed in section 1.4.7 (195). In *Deinococcus radiodurans*, the SbcCD complex was shown to be necessary for homologous recombination, with specificity for hairpin structures (192, 194). Hairpin structures refer to the base-pairing pattern formed when IRs form secondary structures that impede the progression of polymerases in ssDNA (53). Additionally, Darmon et al. (2010) determined that in *E. coli*, in the absence of the SbcCD complex and the RecA protein, a palindromic sequence was able to initiate the formation of large duplications (196). During MGE movement, hairpin structures are generally produced (section 1.3.3.1.1). Thus, it is possible that the SbcCD complex has a role in moderating GCR formation and TE movement. Therefore, mutations in *sbcC* or *sbcD* may not only destabilize the genome through the impairment of recombinational repair (section 1.4.7), but may also possibly increase rates of TE transposition (101).

Partridge et al. (2006), in their microarray study of the transcriptional responses of anaerobic *E. coli* chemostat cultures exposed to oxygen, found that *sbcC* expression between aerobic and anaerobic environments differed and that *sbcC* expression was significantly 1.4-fold greater under aerobic conditions (10). Therefore, to determine how GCRs affect the rate of adaptation and evolution under differing environment conditions, the SbcCD complex can be used.

1.8 Project aims

In different environments, the mutation rates and types of mutations that arise in an organism can vary, thus affecting the overall direction of evolution. Aerobic and anaerobic environments differ in the presence and absence of oxygen. Oxygen availability affects the growth and metabolism of facultative anaerobes, but how this impacts the genome-wide rate of spontaneous mutations and the spectra of mutations has not been comprehensively studied. Thus, I aimed to use whole genome re-sequencing and experimental evolution techniques to determine, compare and contrast the rates and spectra of spontaneous mutations that arise in *E. coli* in aerobic and anaerobic environments. The following general research questions were investigated:

- What is the rate with which each mutation type and class occur in *E. coli* grown in aerobic and anaerobic environments?
- What types of mutations predominate in *E. coli* grown in aerobic and anaerobic environments?
- How the fidelity of the genome is maintained during growth in aerobic and anaerobic environments by understanding which DNA repair pathways are active?

Little is understood about the contribution of GCRs to adaptive evolution in *E. coli.* The SbcCD complex is involved in DNA repair and maintaining genome integrity by preventing GCRs. Furthermore, rates of GCRs have been shown to differ between aerobic and anaerobic environments (28), as have *sbcC* expression levels (10). Thus, it is of interest to determine the contribution of SbcC to the occurrence of GCRs, and the consequent impact of these GCRs on relative fitness in *E. coli*. I aimed to investigate these questions *via* the use of whole genome re-sequencing and experimental evolution techniques. The overarching goal of this research was to determine:

- How does the rate of GCR mutations impact adaptation rate?

Chapter Two: Materials and Methods

2.1 Materials

2.1.1 Bacterial strains

Bacterial strains used in this study are listed in **Table 2.1**.

Table 2.1. *E. coli* strains used in this study.

Bacterial strain	Relevant phenotype	Note	Reference
E. coli REL606	T5^{s},T6^{r}, Ara^{-}	-	(168)
E. coli REL4536	T5^{s},T6^{r}, Ara^{-}	10,000th generation descendent of REL606	(197)
E. coli REL4536 Δ*sbcC::cat*	T5^{s},T6^{r}, Ara^{-}, Cm^{r}	Modified to inactivate *sbcC* gene	This study
E. coli B113	T5^{s},T6^{r}	-	Life Technologies (USA)
E. coli DH5α	T5^{s},T6^{s}	-	Life Technologies

2.1.2 Plasmids and phage

Plasmids used in this study are listed in **Table 2.2**. Coliphage T5 and T6 stocks were provided by S. Fullard (Massey University, New Zealand).

Table 2.2. Plasmids used in this study.

Plasmid	Characteristics	Reference
pUC18 DNA	-	Thermo Fisher Scientific (USA)
pWRG99	pKD46 with I-SceI endonuclease Temperature-sensitive Amp^{r}	(198)
pWRG100	pKD3 with I-SceI recognition site Cm^{r}	(198)

2.1.3 Antibiotics

Antibiotics were purchased from Sigma-Aldrich (USA) and stock solutions were prepared as described in Sambrook and Russell (2001) (199). Antibiotic stocks solutions (**Table 2.3**) were sterilised by filtration through a sterile Millex®-GP Filter Unit (EMD Millipore, USA) containing a Millipore Express PES membrane with a 0.22 µm pore size and stored at -20ºC.

Table 2.3. Antibiotics used in this study.

Reagent	Abbreviation	Stock concentration	Working concentration	Solvent
Ampicillin	Amp	100 mg/mL	100 µg/mL	dH_2O
Carbenicillin	Cb	100 mg/mL	100 µg/mL	dH_2O
Chloramphenicol	Cm	34 mg/mL	34 µg/mL	100% ethanol
Naladixic Acid	Nal	30 mg/mL	30 µg/mL	1M NaOH
Rifampicin	Rif	50 mg/mL	100 µg/mL	100% methanol

2.1.4 Oligonucleotides

Primers (**Table 2.4**) were synthesised by Integrated DNA Technologies (Custom Science, New Zealand). Primers were re-suspended in MilliQ H_2O to a concentration of 100 µM and stored at -20ºC. For use, primers were diluted in MilliQ H_2O to a working concentration of 10 µM and stored at -20ºC.

Table 2.4. Oligonucleotides used in this study.

Name	5' → 3' sequence	PCR annealing temperature	Modifications	Application
fD1	AGA GTT TGA TCC TGG CTC AG	54ºC	-	16S rRNA gene amplification
rD1	AAG GAG GTG ATC CAG CC	54ºC	-	16S rRNA gene amplification
sbcC screening upstream	AGT GAG TAG CGG CTG GAA AAG	55ºC	-	Screening for *sbcC* gene
sbcC screening downstream	TGA AAC CCT CAG CGA ACT CAG	55ºC	-	Screening for *sbcC* gene
sbcC pWRG100 forward	ATT CAG GCT TTG TTG CTG CTG	57ºC	-	Insertion of I-SceI site for homologous recombination
sbcC pWRG100 reverse	GAG GAG TTA ACC GGC ACT GAA ATC	57ºC	-	Insertion of I-SceI site for homologous recombination
sbcC 80mer deletion	AGA TAA TTC GCC CCT CTG TAT TCA TTA TCC TGC TGA ATA GTT CAT GCT TCG TGT TCT CCG GCG AGG GTA TGC AAC GTC GT	-	5' phosphorylation HPLC purification	Second homologous recombination step
sbcC 80mer RC	ACG ACG TTG AT ACC CTC GCC GGA GAA CAC GAA GCA TGA ACT ATT CAG CAG GAT AAT GAA TAC AGA GGG GCG AAT TAT CT	-	5' phosphorylation HPLC purification	Second homologous recombination step
sbcC F	TGA AAA CCA TTT GCG AGC AGG	55ºC	-	qPCR amplification
sbcC R	GAC TGA TTA ACG CCA GGC TC	55ºC	-	qPCR amplification
yegO F	TAC TGT CGG TTG CCA TTA CCC	55ºC	-	qPCR amplification
yegO R	GAC GCC ATT GTT TCT GGT GAC	55ºC	-	qPCR amplification

Name	5' → 3' sequence	PCR annealing temperature	Modifications	Application
proC F	CCA TTC TCG GCG GTC TGA TTG	55ºC	-	qPCR amplification
proC R	ATG CCG AAC TGG TCA TGG AGG	55ºC	-	qPCR amplification
rpoD F	AGA AGA TGG CGA TGA CGA CAG	55ºC	-	qPCR amplification
rpoD R	AAT TTT TCG CGA GCC AGT TCC	55ºC	-	qPCR amplification
phoA F	GCA ACG TAC CAC GGC AAT ATC	55ºC	-	qPCR amplification
phoA R	CAT TAC GTT GCG GAT TAG GCG	55ºC	-	qPCR amplification
soxR F	GAC CAT TGG TGA AGC GTT TGG	55ºC	-	qPCR amplification
soxR R	GCA GCG CCA CTA AGG TAT GAA	55ºC	-	qPCR amplification
IS150 insertion F	CGT TGT CTC TCG TCC AGG TT	51ºC	-	Mutation screening
IS150 insertion R	AGG CGG CAA ATT TGT CTG TA	51ºC	-	Mutation screening
6 kb deletion within F	TCC GGA AGA ACT GGC TCT AA	52ºC	-	Mutation screening
6 kb deletion within R	GGA ATC AGC ACC GAC AAT TT	52ºC	-	Mutation screening
ybdk F	CCT GAA TTA ATC CCG CCA TA	52ºC	-	Mutation screening
entD R	AGC CGT GGT ATC TCG TCA AC	57ºC	-	Mutation screening
nupC F	GGA TTA TCG CAG GTG CAG TT	56ºC	-	Mutation screening
yfeaA R	GCA TTT ATT CTT GCG GTG CT	54ºC	-	Mutation screening
clpX F	ATC TGG AAT TCC GTG ACG AG	55ºC	-	Mutation screening
lon R	ATC GCG TTT ATT TTC GAA CG	52ºC	-	Mutation screening
ynhG F	GGC GAA TAC CTC ATT CAT GG	53ºC	-	Mutation screening
ydhY R	AGG TGT CCG CGG TAT AGT TG	57ºC	-	Mutation screening

Name	5' → 3' sequence	PCR annealing temperature	Modifications	Application
gltK F	TGG AGT TCC ATT GTC CCT TC	55ºC	-	Mutation screening
rihA R	GGA TCC CTG ACT GGA AGA CA	56ºC	-	Mutation screening
focA F	GCT TTC CGG CGA GTA TAT GA	55ºC	-	Mutation screening
pflA R	TGC CCA TAT CAC GGG TAA AT	54ºC	-	Mutation screening
ybdB F	CAC CAG CGA TAA CAC AAT GG	54ºC	-	Mutation screening
ybdH R	GGT GTG GAT CTA CGG CAA AC	56ºC	-	Mutation screening
iap F	AGT TTT CGA CAA AGC CGG TA	54ºC	-	Mutation screening
cysH R	GCA AAA ACA TGG CCT GAA AT	52ºC	-	Mutation screening
trkD F	GCT GGT GAT TAT GGG GCT AA	54ºC	-	Mutation screening
hdfR R	GGC CAG ATT TTC AAC AGC AT	54ºC	-	Mutation screening
yeaS F	ACG CTG GAA CTG GTG AGT TT	57ºC	-	Mutation screening
yeaR R	ACT TCA GGA GCC ACG AAG AA	56ºC	-	Mutation screening
gmhB F	GTG GGA ACA AAA GTG CTG GT	56ºC	-	Mutation screening
dkgB R	TAT GGA TGC CGT GCT GTT TA	54ºC	-	Mutation screening
kgtP F	CTC ACT ATC AGC AGG GCA CA	57ºC	-	Mutation screening
clpB R	GAA GGC ATC GCT TTC TGG TA	55ºC	-	Mutation screening
yibF F	AGT CGC CAG ATT GAC CGT AT	56ºC	-	Screening over *rhsA* gene
yibA R	CGT CTT GCC CAC CTC TTA AC	56ºC	-	Screening over *rhsA* gene
nikR F	CGT GTT GAA AGG TGA CAT GG	54ºC	-	Screening over *rhsB* gene
yhhJ R	GAA AGT ATG CCG CAG ATG GT	55ºC	-	Screening over *rhsB* gene

Name	5' → 3' sequence	PCR annealing temperature	Modifications	Application
ybfO F	CGG ATT CAG CGG ATA CTG AT	54ºC	-	Screening over *rhsC* gene
kdpA R	CGG CAG AAA GAA AGT TTT GC	53ºC	-	Screening over *rhsC* gene
ybbP F	GCT TAA CCG CGA ACT CAA TC	54ºC	-	Screening over *rhsD* gene
ylbH R	GTG ATG CGG GTT CTC TTC AT	55ºC	-	Screening over *rhsD* gene
ECB_01416 F	AGC AAC TCT GGA ATG GCT GT	57ºC	-	Screening over *rhsE* gene
ECB_01413 R	ACC GCA CCA CTG ATG TGA TA	56ºC	-	Screening over *rhsE* gene

2.1.5 Laboratory chemicals and enzymes

General laboratory chemicals were manufactured by Sigma-Aldrich, Becton-Dickinson (USA), Thermo Fisher Scientific (USA), Merck (USA) and Life Technologies (USA). BD Bacto™ Agar was manufactured by Becton-Dickinson and bacteriological agar was manufactured by Oxoid (UK). Analytical grade 100% ethanol, 96% ethanol, isopropanol and methanol were supplied by VWR International Ltd. (USA). Gas mixtures and liquid nitrogen were supplied by BOC Gases (Auckland, NZ).

Restriction endonucleases were obtained from New England Biolabs (USA). DNA polymerases were obtained from Life Technologies. Proteinase K, Lysozyme and Ribonuclease A were obtained from Sigma-Aldrich; stock solutions were prepared as described by Sambrook and Russell (2001) (199), filter sterilised through a sterile Millex®-GP Filter Unit with a 0.22 µm pore size Millipore Express PES membrane and stored at -20ºC.

2.1.6 Buffers and solutions

Standard buffers and solutions were prepared as described in Sambrook and Russell (2001) (199). Sterilisation was performed either by autoclaving at 121ºC for 20 min or by filtration through a sterile Millex®-GP Filter Unit containing a Millipore Express PES membrane with a 0.22 µm pore size. Buffers and solutions were stored at room temperature (RT) unless stated otherwise.

2.1.6.1 Tris-acetate-EDTA (TAE) buffer

TAE buffer stocks, 50×, were comprised of 2 M Tris base, 1 M glacial acetic acid and 50 mM ethylenediaminetetraacetic acid (EDTA), pH 8.0. A working solution of 1× TAE buffer was obtained by a 50-fold dilution in dH_2O.

2.1.6.2 EDTA/SDS solution

EDTA/SDS solution was comprised of 0.25 M EDTA and 20% (w/v) sodium dodecyl sulphate (SDS).

2.1.6.3 Lysis buffer

Lysis buffer was comprised of 30 mM NaCl, 50 mM Tris-HCl (pH 8.0), 5 mM EDTA (pH 8.0) and 10% (w/v) SDS.

2.1.6.4 Resazurin solution

Resazurin solution was comprised of 0.1% (w/v) resazurin.

2.1.6.5 Reducing agent

Reducing agent was comprised of 7 mM L-cysteine•HCl•H_2O and 13 mM of washed Na_2S•$9H_2O$ crystals. Reducing agent solution was sealed in nitrogen (N_2) flushed serum bottles (Wheaton Science Products, USA) with butyl rubber stoppers and solution was sterilised by autoclaving at 121°C for 15 min.

2.1.6.6 Diethylpyrocarbonate (DEPC) H_2O

DEPC H_2O was comprised of 0.1% (v/v) DEPC. Solution was sterilised by autoclaving at 121°C for 15 min.

2.1.6.7 Glycerol saline

Glycerol saline was comprised of 0.85% (w/v) NaCl and 70% (v/v) glycerol. Anaerobic 70% glycerol saline was prepared by boiling all components in a microwave oven and cooling under a stream of oxygen-free carbon dioxide (CO_2) gas.

2.1.6.8 Glucose

Anhydrous glucose, 10% (w/v), was dissolved in dH_2O and filter sterilised. To make anaerobic solution, 10% anhydrous glucose was dissolved in dH_2O and flushed with oxygen-free CO_2 gas for 30 min before filter sterilisation in an anaerobic glove box (Coy Laboratory Products Inc., USA).

2.1.6.9 Magnesium sulphate

Magnesium sulphate, 10% (w/v), was dissolved in dH_2O and filter sterilised. To make anaerobic solution, 10% magnesium sulphate was dissolved in dH_2O and flushed with oxygen-free CO_2 gas for 30 min before filter sterilisation in an anaerobic glove box.

2.1.6.10 Thiamine

Thiamine, 0.2% (w/v), was dissolved in dH_2O and filter sterilised. To make anaerobic solution, 0.2% thiamine was dissolved in dH_2O and flushed with oxygen-free CO_2 gas for 30 min before filter sterilisation in an anaerobic glove box.

2.1.7 Media

Standard media were prepared as described in Sambrook and Russell (2001) (199). Sterilisation was performed by autoclaving at 121ºC for 20 min and all media were stored at RT unless stated otherwise.

Anaerobic media was always prepared in Duran Pressure-plus bottles (Schott Duran, Germany) and bottles were sealed with butyl rubber stoppers to maintain the anaerobic conditions. Any additions to anaerobic media after sterilisation and all pouring of anaerobic solid medium into petri dishes were done under anaerobic conditions in an anaerobic glove box (Coy Laboratory Products Inc., USA).

2.1.7.1 Luria-Bertani (LB) medium

2.1.7.1.1 Aerobic preparation

Aerobic LB liquid medium was prepared by dissolving 1.0% (w/v) tryptone, 0.5% (w/v) yeast extract and 1.0% (w/v) NaCl in dH_2O and autoclaving. Solid medium was obtained by the addition of 1.6% (w/v) bacteriological agar before autoclaving.

2.1.7.1.2 Anaerobic preparation

Anaerobic LB liquid medium was prepared by dissolving 1.0% (w/v) tryptone, 0.5% (w/v) yeast extract and 1.0% (w/v) NaCl in dH_2O and boiling for one min in a microwave oven to remove oxygen. The media was then cooled to RT under a stream of oxygen-free N_2 gas. For every mL of media, 0.4 µL of 0.1% resazurin solution (section 2.1.6.4) was added and the media was autoclaved. Prior to use, for every mL of media, 20 µL of reducing agent was added *via* a sterile syringe. Once the media had changed colour from dark yellow to yellow, the media was ready for use. For solid medium, 1.6% (w/v) bacteriological agar was flushed under N_2 gas separately and mixed with liquid components before autoclaving.

2.1.7.2 Davis Minimal (DM) medium

The basic medium used for this study is Davis minimal broth (200), described in **Table 2.5**.

Table 2.5. DM medium components.

Component	Volume or weight
Salt solution	
Potassium phosphate (dibasic trihydrate)	7.0 g
Potassium phosphate (monobasic anhydrous)	2.0 g
Ammonium sulphate	1.0 g
Sodium citrate	0.5 g
dH_2O	To 1000 mL

2.1.7.2.1 DM25 Aerobic preparation

All Salt solution components listed in **Table 2.5** were dissolved in 1 L dH_2O. The medium was distributed into 250 mL aliquots and autoclaved. Before use, 62.5 µL of 10% glucose (section 2.1.6.8), 250 µL of 10% magnesium sulphate (section 2.1.6.9) and 250 µL of 0.2% thiamine (section 2.1.6.10) were added to each bottle, resulting in a glucose concentration of 25 mg/L. Unless otherwise stated, this media composition is henceforth referred to as DM25.

2.1.7.2.2 DM25 Anaerobic preparation

To make the medium anaerobic, the Salt solution components listed in **Table 2.5** were boiled for one min in a microwave oven and then cooled under a stream of oxygen-free CO_2 gas. While cooling, 400 µL of 0.1% resazurin solution (section 2.1.6.4) was added. The medium was distributed into 250 mL aliquots and autoclaved. Prior to use, 5 mL of reducing agent (section 2.1.6.5) was added to each bottle *via* a sterile 5 mL syringe. Once the media had changed colour from pale pink to colourless, 62.5 µL of 10% glucose (section 2.1.6.8), 250 µL of 10% magnesium sulphate (section 2.1.6.9) and 250 µL of 0.2% thiamine (section 2.1.6.10) were added to each bottle.

2.1.7.3 DM solid medium

2.1.7.3.1 Aerobic preparation

To make DM solid medium, 1 L dH_2O was split into three approximately equal parts. The salt solution components listed in **Table 2.5** were added to one part, 15 g Bacto Agar was added to the second part, and 0.2 g anhydrous glucose was added to the last part. The three solutions were autoclaved separately, cooled to 55°C, then combined together with 1 mL 10% (w/v) magnesium sulphate (section 2.1.6.9) and 1 mL 0.2% (w/v) thiamine (section 2.1.6.10) before pouring into petri dishes.

2.1.7.3.2 Anaerobic preparation

To make anaerobic DM solid medium, 1 L dH_2O was split into three approximately equal parts. Salt solution components listed in **Table 2.5** were added to one part, 15 g Bacto Agar was added to the second part and 0.2 g anhydrous glucose was added to the last part. The Salt solution components listed in **Table 2.5** and the dH_2O for the agar were boiled in a microwave oven for 1 min and cooled under oxygen-free N_2 gas. The bottles containing glucose and Bacto Agar were flushed with N_2 gas without boiling. The agar was mixed with the cooled dH_2O before autoclaving. After sterilisation, the three parts were combined together and 1 mL anaerobic 10% (w/v) magnesium sulphate (section 2.1.6.9) and 1 mL anaerobic 0.2% (w/v) thiamine (section 2.1.6.10) were added and the medium was poured into petri dishes.

2.1.7.4 Minimal glucose (MG) solid medium

MG solid medium was prepared in the same way as DM solid medium (section 2.1.7.3) except that 16 g Bacto Agar was added to the second part and 4.0 g anhydrous glucose was used.

2.1.7.5 Minimal arabinose (MA) solid medium

MA solid medium was prepared in the same manner as MG solid medium (section 2.1.7.4) except that 4.0 g L-arabinose was used instead of glucose.

2.2 Methods

2.2.1 Aerobic cultivation of *E. coli*

For aerobic cultivation, *E. coli* cultures were grown in serum bottles (Wheaton Science Products, USA) in media prepared aerobically, covered with AeraSeal™ Breathable Films (Raylab, New Zealand) and incubated at 37°C with an orbital shaking of 150 RPM unless otherwise stated. All sterile and lineage-related aerobic work was performed in a Class II Type a/B3 biohazard cabinet supplied by Nuaire Biological Safety Cab (USA).

2.2.2 Anaerobic cultivation of *E. coli*

All anaerobic work was carried out in an anaerobic glove box (Coy Laboratory Products Inc., USA) containing a 92% $CO_{2:}$ 8% hydrogen (H_2) atmosphere. For the anaerobic environment, *E. coli* cultures were grown in serum bottles in media prepared anaerobically, sealed with butyl rubber stoppers and incubated at 37°C with an orbital shaking of 150 RPM unless otherwise stated.

2.2.3 Culture resuscitation from frozen stocks

To resuscitate cultures, cells stored at -85°C in glycerol saline (section 2.1.6.7) were first washed in growth media to remove glycerol. Using sterile inoculation loops, small amounts of frozen culture stock were thawed in micro-centrifuge tubes and 990 µL of DM25 (section 2.1.7.2) was added to each sample. Cells were mixed and centrifuged for 3 min at 11,000 *g*. After centrifugation, 990 µL of supernatant was discarded and a wash with DM25 media was repeated a further two times. After the last wash, washed cells were used to inoculate LB (section 2.1.7.1) or DM25 media.

2.2.4 Bacterial growth courses

Growth dynamics of aerobic or anaerobic bacterial cultures in liquid media were determined by reviving cultures from frozen stocks (section 2.2.3) in the appropriate medium (DM25 or LB) and growing overnight at 37ºC with 150 RPM shaking. For 1 mL cultures, 10 µL of overnight culture was used to inoculate 990 µL of fresh media in sterile 24-well non-treated polystyrene flat-bottom tissue culture plates (Becton-Dickinson), henceforth referred to as 24-well plates. Triplicate cultures were grown for 36 h, and at regular intervals the optical densities at 600 nm (OD_{600}) were measured in an Ultraspec 1100 Pro spectrophotometer (Amersham Biosciences, Sweden). To determine colony forming units (CFU) per mL of culture, samples were serial diluted and plated on MG (section 2.1.7.4) or LB (section 2.1.7.1) agar plates as appropriate. For larger culture volumes, 500 µL of overnight culture was used to inoculate 49.5 mL of fresh media in 250 mL serum bottles in triplicate and cultures were grown for 48-72 h at 37ºC with 150 RPM shaking.

2.2.5 Gram staining

To gram stain a bacterial culture, 100 µL of culture was spread on a glass microscope slide and heat fixed. Samples were stained with 10% (w/v) crystal violet for 1 min followed by fixation of the stain with 0.3% (w/v) iodine and 0.7% (w/v) potassium iodine for 1 min. The cells were then de-colourised with 50% (v/v) acetone, and counterstained with 2.5% (w/v) safranin for 30 sec. Slides were gently rinsed with tap water between each step, and viewed using a Leica DM2500 microscope (Bio–Strategy, New Zealand).

2.2.6 Agarose gel electrophoresis

UltraPure™ Agarose (Invitrogen, Life Technologies), 1% (w/v) gels were made up in 1× TAE buffer (section 2.1.6.1) containing 1× SYBR safe nucleic acid dye (Life Technologies). Gels were cast in stands provided by Bio-Rad (USA), submerged in 1× TAE buffer in the running tank, and 5 uL of sample was mixed with 1 µL of 10× BlueJuice loading dye (Life Technologies) and loaded into each well. An appropriate molecular weight-size marker (**Table 2.6**) was also loaded in the first lane. Gels were run at 90 V for 30 min and visualised using UV trans-illumination. Gels were photographed using a Nikon D700 camera with Kodak Gel Logic 200 Imaging System (Eastman Kodak Company, USA).

Table 2.6. Molecular weight-size markers used in this study.

Molecular weight-size markers	Range	Supplier
TrackIt 100 bp DNA ladder	100 to 1,500 bp	Life Technologies
1 kb Plus DNA ladder	100 to 12,000 bp	Life Technologies
Lambda DNA/Hind III ladder	100 to 24,000 bp	Thermo Fisher Scientific
0.5-10 kb RNA ladder	500 to 10,000 bp	Life Technologies

2.2.7 Polymerase chain reaction (PCR)

PCRs were performed in 25 μL reactions unless specified otherwise. All PCR reagents were manufactured by Invitrogen and Platinum Taq or Platinum High Fidelity Taq DNA polymerases (section 2.1.5) were used. A standard reaction mix listed in **Table 2.7** was used. The amount of water added to each PCR reaction was dependent on the volume of template DNA added and therefore, it was adjusted to bring the total volume of each reaction to 25 μL. For each polymerase, the corresponding commercial PCR buffer and magnesium solution was used: 10× PCR buffer and $MgCl_2$ for Platinum Taq, and 10× High Fidelity PCR buffer and $MgSO_4$ for Platinum High Fidelity Taq.

Table 2.7. Reaction components for a standard PCR reaction.

Component	Volume per reaction	Final concentration per reaction
10× buffer	2.5 μL	1×
50 mM Mg^{+2}	1.0 μL	2 mM
10 mM dNTP mixture	0.5 μL	0.2 mM
10 μM forward primer	0.5 μL	0.2 μM
10 μM reverse primer	0.5 μL	0.2 μM
5 U/μL polymerase	0.1 μL	0.02 U/μL
Template nucleic acid	1-5 μL	1-100 ng
MilliQ H_2O	To 25 μL	N/A

All reaction components, apart from template DNA, were added together and mixed before being transferred to 0.2 mL thin-walled PCR tubes (Eppendorf, Germany). Template DNA was added and PCR amplifications were performed on proS Mastercycler machines (Eppendorf). A standard PCR programme was used: initial denaturation at 94°C for 3 min, followed by 30 cycles of 94°C for 30 sec, annealing temperature (primer dependent) for 30 sec (**Table 2.4**) and extension for 1 min per kb of product. The final elongation was for 10 min. The extension temperature was dependent on the polymerase used for the reaction: 72ºC for Platinum Taq and 68ºC for Platinum High Fidelity Taq.

2.2.7.1 PCR purifications

PCR products were purified using QIAquick PCR Purification Kits (Qiagen, Germany) as per manufacturer's instructions.

2.2.8 Quantitative PCR (qPCR)

Quantitative PCR (qPCR) was performed on DNA samples (section 2.2.10.3) using a LightCycler Fast Start DNA Master SYBR Green I (Roche Diagnostics, Switzerland) kit as per the manufacturer's instructions. qPCRs were performed on a RotorGeneQ machine (Qiagen) using a 72-well rotor disc. All reaction components, apart from template DNA, were added together and mixed before being transferred to 0.1 mL strip tubes (Qiagen). A standard reaction mix listed in **Table 2.8** was used for each reaction.

Table 2.8. Reaction components for a standard qPCR reaction.

Component	Volume per reaction
10× LightCycler Fast Start DNA Master SYBR Green I mix	2.0 µL
25 mM $MgCl_2$	1.2 µL
10 µM forward primer	1.0 µL
10 µM reverse primer	1.0 µL
Template DNA (3-5 ng)	2.0 µL
H_2O	To 20 µL

Template cDNA was added to each reaction and a standard qPCR programme was used: initial denaturation at 95°C for 10 min, followed by 40 cycles of 95°C for 10 sec, annealing at 55°C for 10 sec (see temperatures in **Table 2.4**) and extension at 72°C for 30 sec. The final elongation was at 72°C for 10 min. After the run, data was exported to LinRegPCR version 2013.0 (201) and analysed as directed in the LinRegPCR user manual.

2.2.9 Generation of *sbcC* deletion mutant

2.2.9.1 Preparation of electrocompetent cells

Electrocompetent *E. coli* cells were prepared as described in Sambrook and Russell (2001) (199). The *E. coli* strain was revived from frozen stocks (section 2.2.3) in 10 mL LB and grown overnight at 30°C. The next day, 250 mL of LB in a 1 L Erlenmeyer flask was inoculated with 1/100th volume of overnight culture. The culture was grown at 30°C with 180 RPM shaking till the culture had reached an OD_{600} of 0.4. The culture was poured into pre-chilled, sterile 250 mL centrifuge bottles and incubated on ice for 30 min. The sample was then centrifuged at 4,000 *g* for 10 min at 4°C in a pre-chilled rotor. The supernatant was discarded and the pellet was gently re-suspended in 250 mL pre-chilled 10% glycerol. The sample was centrifuged at 4,000 *g* for 10 min at 4°C and supernatant was discarded. The pellet was re-suspended in 200 mL 10% glycerol another two times. The pellet was weighed and 500 μL 10% glycerol was added to the cells. The cells were aliquoted in 50 μL volumes into pre-chilled, sterile micro-centrifuge tubes and stored at -85°C.

2.2.9.2 Transformation of electrocompetent cells

Electrocompetent cells (section 2.2.9.1) were incubated on ice for 5 min with desalted DNA and transformed by electroporation using 0.1 cm gap electroporation cuvettes and a Gene Pulser® II Electroporation System (Bio-Rad, USA). The electroporator was set at 25 μF and 1.5 kV while the pulse controller was set to 200 **Ω**. Transformed cells were transferred to a micro-centrifuge tube containing 950 μL of SOC medium and incubated for 2 h with rotary agitation at 30°C.

2.2.9.3 Lambda Red (λ-Red) recombination

A two-step recombination procedure described by Blank et al. (2011) was used to generate the *sbcC* deletion mutant (198). The process required the temperature-sensitive pWRG99 plasmid which encodes the λ-Red recombination machinery and restriction enzyme I-SceI, and the pWRG100 plasmid, which was used as a template for the recombination cassette (**Table 2.2**).

2.2.9.3.1 First homologous recombination step

2.2.9.3.1.1 Electrocompetent *E. coli* REL4536

Electrocompetent *E. coli* REL4536 cells (**Table 2.1**) were prepared as described in section 2.2.9.1 and transformed with 100 ng pWRG99 plasmid as described in section 2.2.9.2. Successful recombinant clones were identified by selective plating on LB plates supplemented with 50 µg/mL Amp (**Table 2.3**). Plates were incubated at 30°C for 24 h, and a clone was randomly selected for further use.

2.2.9.3.1.2 Transformation with pWRG100

The chloramphenicol resistance cassette with I-SceI recognition sites of plasmid pWRG100 was PCR amplified (section 2.2.7) using sbcC pWRG100 forward and sbcC pWRG100 reverse primers (**Table 2.4**). The *E. coli* REL4536 transformant containing pWRG99 (section 2.2.9.3.1.1) was grown in 200 mL LB culture at 30°C and made electrocompetent as previously described in section 2.2.9.1. To induce the λ-Red system responsible for recombination, LB was supplemented with 10 mM L-arabinose. Cells were transformed as described in section 2.2.9.2 with 200-500 ng of purified pWRG100 PCR product (section 2.2.7.1). Transformants were grown by selective plating on LB plates supplemented with 34 µg/mL Cm (**Table 2.3**) and incubation at 30°C for 24 h. Clones were re-streaked and incubated at 37°C for pWRG99 temperature-sensitive plasmid curing. The chromosomal integration of the pWRG100 chloramphenicol resistance cassette at the *sbcC* locus was verified by PCR amplification of the *sbcC* locus (sbcC screening upstream and downstream primers in **Table 2.4**) and sequencing of the amplicon (section 2.2.11). This strain is referred to as *E. coli* REL4536 Δ*sbcC::cat* (**Table 2.1**).

2.2.9.3.2 Second homologous recombination step

2.2.9.3.2.1 Transformation of *E. coli* REL4536 Δ*sbcC::cat* with pWRG99

E. coli REL4536 Δ*sbcC::cat* was made electrocompetent (section 2.2.9.1) and transformed with 100 ng pWRG99 (section 2.2.9.2). Transformants were identified by selective plating on LB plates supplemented with 50 µg/mL Amp (**Table 2.3**) and incubated at 30°C for 24 h. A randomly selected transformant was selected for further development.

2.2.9.3.2.2 Scarless deletion of *sbcC*

To create a scarless deletion of *sbcC*, a second recombination using a short DNA fragment complementary to the sequences flanking the chloramphenicol cassette was undertaken. HPLC-purified 5' phosphorylated oligonucleotides (sbcC 80mer RC and sbcC 80mer deletion primers in **Table 2.4**) at 100 µM were mixed in equal amounts and annealed together by incubating for 15 min in a water bath at 95°C, and then left to cool overnight. *E. coli* REL4536 Δ*sbcC::cat* transformed with pWRG99 (section 2.2.9.3.2.1) was grown in 200 mL LB at 30°C and made electrocompetent (section 2.2.9.1). To induce the λ-Red system, LB was supplemented with 10 mM L-arabinose. Electrocompetent cells were transformed with 1 µg of the 80-mer fragment. To induce I-SceI, and eliminate cells where homologous recombination between the 80-mer fragment and the sequences flanking the chloramphenicol resistance cassette was unsuccessful, 200 ng/mL anhydrotetracycline (AHT) was added to the SOC recovery medium. Transformations were identified by selective plating on LB plates supplemented with 50 µg/mL Cb (**Table 2.3**) and 500 ng/mL AHT. Plates were incubated at 30°C for 48 h. For pWRG99 plasmid curing, incubation at 37°C overnight was required.

2.2.10 Nucleic acid extraction

2.2.10.1 DNA extractions

DNA was extracted from cultures using a phenol based extraction method (202). For all extractions, UltraPure™ Phenol: Chloroform: Isoamyl alcohol (25:24:1, v/v/v) from Invitrogen and Chloroform: Isoamyl alcohol (24:1, v/v) from Sigma-Aldrich were used. For aerobic lineages, 100 mL LB cultures were grown for 24 h (section 2.2.1) while for anaerobic lineages; 250 mL LB cultures were grown for 36 h (section 2.2.2).

2.2.10.1.1 Cell lysis

Cultures were harvested by centrifugation at 8,000 *g* for 10 min at 4°C in 500 mL centrifuge bottles. Media was discarded and cell pellets were snap frozen in liquid nitrogen and stored at -20°C. For cell lysis, frozen pellets were re-suspended in 5 mL lysis buffer (section 2.1.6.3) and the solutions were transferred to 15 mL Falcon tubes (Becton-Dickinson, USA). Cells were once again harvested by centrifugation at 8,000 *g* for 10 min at 4°C. Supernatant was discarded and cells were re-suspended in 2 mL lysis buffer containing 20 mg/mL lysozyme and 10 µg/mL RNase A (section 2.1.5). Samples were incubated at 37ºC and orbital shaking of 150 RPM for 2 h to allow for cell lysis. EDTA/SDS solution (section 2.1.6.2), 2 mL, was added to each sample and samples were incubated for another 2 h at 65°C. Finally, 100 µL of 20 mg/mL Proteinase K was added to each sample and the mixture was incubated overnight at 65°C.

2.2.10.1.2 Extractions with Phenol: Chloroform: Isoamyl alcohol

After overnight incubation (section 2.2.10.1.1), 5 mL of Phenol: Chloroform: Isoamyl alcohol was added to each sample. Samples were inverted to mix and incubated at RT for 15 min. Samples were centrifuged at 2,500 *g* for 10 min at 4°C. The aqueous layer from each sample was transferred to a new tube, and extracted twice with an equal volume of Phenol: Chloroform: Isoamyl alcohol, and twice with Chloroform: Isoamyl alcohol, with centrifugation of the samples at 2,500 *g* for 10 min at 4°C, and transfer of the supernatant to a new tube between extractions. To precipitate the DNA, a 1/10th volume of 3 M sodium acetate (pH 5.2) and 2 volumes of 100% ethanol were added to each sample and stored at -20°C overnight. The following day, samples were centrifuged at 12,500 *g* for 10 min at 4°C to pellet the DNA. The supernatant was discarded and DNA pellets were washed twice with 5 mL 70% (v/v) ethanol. After each wash, samples were centrifuged at 12,500 *g* for 10 min at 4°C. DNA pellets were air-dried for 15 min and re-suspended in 55°C MilliQ H_2O. DNA was stored at 4°C.

2.2.10.2 RNA extractions

RNA was extracted from cultures using a hot lysis buffer and acid phenol based extraction method (202). For all extractions, Phenol: Chloroform: Isoamyl alcohol (125:24:1, v/v/v, pH 4.5) from Ambion (Life Technologies) and Chloroform: Isoamyl alcohol (24:1, v/v) from Sigma-Aldrich were used.

2.2.10.2.1 RNA handling practises

To prevent RNase contamination during RNA experiments, gloves were worn at all times and all glassware was baked in an oven at 140°C overnight. Equipment, surfaces and pipettes were wiped with Ambion RNaseZap (Life Technologies) and only RNase-free pipette tips and plastic-ware were used. RNase-free water, DEPC H_2O (section 2.1.6.6), was used to make all required solutions and buffers.

2.2.10.2.2 Growth conditions

E. coli REL4536 (**Table 2.1**) was revived from frozen stocks (section 2.2.3) in 10 mL aerobic or anaerobic DM25 (section 2.1.7.2) and grown overnight. Cultures were grown in 50 mL DM25 aliquots inoculated with 1/100th volume of overnight culture and incubated till they had just reached stationary phase. Thus, in the aerobic environment, cultures were grown for 9 h in 250 mL serum bottles (section 2.2.1) while in the anaerobic environment, cultures were grown for 16 h in 250 mL anaerobic serum bottles (section 2.2.2).

2.2.10.2.3 RNA extraction

E. coli REL4536 cultures were grown in triplicate within each environment, resulting in a total of 6 samples. Due to limitations in culture volume in serum bottles, each replicate culture was actually comprised of 12 serum bottles of 50 mL of culture, which were pooled to obtain 600 mL culture per replicate. Cells were harvested by centrifugation at 8,000 *g* for 10 min at 4°C. Media was discarded and cell pellets were snap frozen in liquid nitrogen. Samples were lysed with 5 mL of 2% SDS solution at 65°C and poured into 10 mL of acid Phenol: Chloroform: Isoamyl alcohol. The samples were incubated at 65°C for 15 min and then centrifuged at 2,500 *g* for 15 min at 4°C. The aqueous phase, containing the RNA, was transferred to a fresh tube, and 5 mL of acid Phenol: Chloroform: Isoamyl alcohol at 65°C was added. The process was repeated and the aqueous phase was transferred to a fresh tube and an equal volume of Chloroform: Isoamyl alcohol solution was added. Samples were incubated at RT for 10 min and then centrifuged at 2,500 *g* for 15 min at 4°C. The aqueous phase was transferred to a fresh tube where an equal volume of Chloroform: Isoamyl alcohol solution was added and the process was repeated. The aqueous phase was transferred to a fresh tube and an equal volume of 100% isopropanol and 1/10th volume of 3 M sodium acetate (pH 5.2) was added. After incubation at -20°C overnight, the RNA was harvested by centrifugation at 15,000 *g* for 20 min at 4°C. Supernatant was discarded and RNA pellets were washed twice with 5 mL of ice-cold 70% (v/v) ethanol. After each wash, samples were centrifuged at 12,500 *g* for 10 min at 4°C. RNA pellets were air-dried for 20 min and re-suspended in 100 µL DEPC H_2O (section 2.1.6.6). RNA was stored at -85°C.

2.2.10.2.4 DNase treatment

Following RNA extraction, Turbo DNase (Ambion) was used to treat the samples as per manufacturer's instructions. Each sample was split into five 20 µL aliquots and two rounds of DNase treatment were performed on each aliquot. DNA elimination was verified by PCR amplification of the 16S rRNA gene (**Table 2.4**).

2.2.10.2.5 RNA purifications

Following DNase treatment, RNA samples were purified using the RNeasy Mini kit (Qiagen) as per the manufacturer's instructions. For each sample, the DNase-treated aliquots were pooled together before purification. RNA was stored at -85°C.

2.2.10.2.6 RNA quality analysis

RNA quality was measured using the Agilent 2100 Bioanalyzer (Agilent Technologies, USA) with the RNA 6000 Nano Chip kit according to the manufacturer's instructions. RNA integrity was also checked by using agarose gel electrophoresis (section 2.2.6) under RNase free conditions (section 2.2.10.2.1) to detect distinct 16S and 23S ribosomal RNA (rRNA) bands.

2.2.10.3 cDNA synthesis

cDNA was synthesised from the DNase-treated RNA samples using SuperScript II Reverse Transcriptase (Invitrogen) with random hexamers (Thermo Fisher Scientific) as per manufacturer's instructions.

2.2.10.4 Nucleic acid quantity and quality analysis

Nucleic acid quantity was measured on a Qubit® 2.0 fluorometer (Life Technologies) as per the manufacturer's instructions. DNA was quantified using the Quant-iT dsDNA Broad-range (BR) Assay kit (range from 2 to 1000 ng), RNA was quantified using the Quant-iT RNA Assay kit (range from 5 to 100 ng), and cDNA was quantified using the Qubit ssDNA Assay kit (range from 1 to 200 ng) (Life Technologies).

Nucleic acid quality was measured using the NanoDrop ND-1000 UV-Vis Spectrophotometer (NanoDrop Technologies, USA). DEPC H_2O (section 2.1.6.6) was used as a blank for all samples. Nucleic acid concentration and quality were recorded using absorbance at 260 nm and 230 nm wavelengths respectively.

Agarose gel electrophoresis (section 2.2.6) was used for quantifying nucleic acid yield (using the Lambda DNA/Hind III molecular weight-size marker as described in **Table 2.6**) and checking nucleic acid integrity when required.

2.2.11 Sanger sequencing

PCR products were Sanger sequenced at the Massey Genome Service (Massey University, Palmerston North, New Zealand). Samples were fluorescence labelled, cleaned-up and sequenced on a capillary ABI3720 Genetic Analyzer (Applied Biosystems Inc., USA). Data was received as an ABI file, which is the format of the sequencing files produced by Applied BioSystems sequencing machines. ABI files, containing the sequence and associated probabilities of the four nucleotide bases, were viewed using Geneious Pro 5.6.2 (203).

2.2.12 Fluctuation assays

Classical Luria-Delbruck fluctuation tests to determine spontaneous mutation rate (149) were performed using aerobic and anaerobic DM media (section 2.1.7.2) supplemented with 1 g glucose/L (referred to as DM1000). *E. coli* REL4536 (**Table 2.1**) was revived from frozen stock (section 2.2.3) in 10 mL DM25 media and grown overnight. The following day, 100 µL of overnight culture was used to inoculate 9.9 mL DM1000 in the aerobic (section 2.2.1) and anaerobic (section 2.2.2) environment and cultures were grown for 24 h. The following day, for each environment, 30 replicate cultures were inoculated with approximately 5,000 cells/mL in 9 mL DM1000 and allowed to grow to stationary phase. After 24 h, 2 mL of each culture was concentrated to 200 µL by centrifugation at 14,000 *g* and plated on solid medium containing antibiotic (described below). Plates were incubated at 37ºC and antibiotic resistant colonies were counted after 24 h. Mutation rates were calculated using the Ma-Sandri-Sarkar Maximum Likelihood Estimator (MSS-MLE) method (150), as implemented in FALCOR (204).

Fluctuation assays were performed, assessing for nalidixic acid resistance and arabinose utilisation of spontaneous mutants. For assays assessed for nalidixic acid resistance, cultures were plated on LB agar plates (section 2.1.7.1) supplemented with 30 µg/mL nalidixic acid (n = 24) (**Table 2.3**) and on LB agar plates with no antibiotic (n = 6). For assays assessed for arabinose utilisation, cultures were plated on MA agar (section 2.1.7.5) plates (n = 24) and on MG agar (section 2.1.7.4) plates (n = 6).

2.2.13 Fitness assays

2.2.13.1 Generation of neutrally marked strain

To assess the fitness of lineages, neutrally marked reference strains of the ancestor are required. Antibiotic resistant mutant strains were generated by reviving ancestral *E. coli* REL4536 (**Table 2.1**) from frozen stocks (section 2.2.3) and growing overnight at 37°C with 150 RPM orbital shaking in 30 mL aerobic DM1000 (section 2.2.12). Overnight culture was concentrated to 1 mL and plated onto MG agar plates (section 2.1.7.4) supplemented with 100 µg/mL rifampicin (**Table 2.3**). Colonies were picked and tested for neutrality and a neutral reference competitor strain, called Rif^{r}2, was used for all competitive fitness assays (205).

2.2.13.2 Fitness assays with rifampicin resistant marker

Relative fitness was assayed by competing the reference competitor strain (Rif^{r}2) against an evolved clone or population of interest in 24-well plates in anaerobic DM media (section 2.1.7.2.2). Evolved populations were revived from frozen stocks (section 2.2.3), 10 µL of inoculum was added to 990 µL DM25 and cultures were grown overnight at 37°C with 150 RPM orbital shaking. For Rif^{r}2 cultures, inoculum was added to DM25 containing 100 µg/mL rifampicin (**Table 2.3**). The next day, for each population, five biological replicate cultures were inoculated with overnight cultures and grown overnight at 37°C with orbital shaking of 150 RPM. On the next day, referred to as T = 0, evolved population cultures and Rif^{r}2 cultures were mixed at an estimated cell ratio of 1:1 in 1 mL DM25 media. The competition cultures were incubated at 37°C with orbital shaking of 150 RPM for 24 h, referred to as T = 1.

A 10 µL aliquot of the T=0 culture was 100-fold diluted and cultures were plated on LB agar plates (section 2.1.7.1) supplemented with 100 µg/mL rifampicin (n = 2)

(**Table 2.3**) and on LB agar plates with no antibiotic (n = 2). Plates were incubated at 37ºC and antibiotic resistant colonies were counted after 24 h. A 10 µL aliquot of the T = 1 culture was 10,000-fold diluted and cultures were once again plated on LB agar plates (section 2.1.7.1) supplemented with 100 µg/mL rifampicin (n = 2) (**Table 2.3**) and on LB agar plates with no antibiotic (n = 2). Plates were incubated at 37ºC and antibiotic resistant colonies were counted after 24 h.

2.2.13.2.1 Fitness calculation

Relative fitness is the change of the ratio between two competing strains when grown together in the same environment (168). Fitness was calculated as presented in **Equation 2.1 - Equation 2.4**.

$$n_E = n_{Total} - n_R$$

Equation 2.1. Calculation for the number of evolved cells (n_E). n_{Total} is the number of cells on LB plates and n_R is the number of Rif^{r}2 cells on LB agar plates supplemented with 100 µg/mL rifampicin.

$$m_n = ln\left[\frac{n_{E,R}(1)}{n_{E,R}(0)}\right]\Big/ t$$

Equation 2.2. Calculation of Malthusian parameter (m_n) for competitors. $n_E(0)$ and $n_R(0)$ are initial densities of evolved and reference strain, respectively, $n_E(1)$ and $n_R(1)$ are final densities after one day of growth of the evolved and reference strains, respectively, and t is time.

$$D_n = ln\left[\frac{n_{E,R}(1)}{n_{E,R}(0)}\right]\Big/ ln(2)$$

Equation 2.3. Calculation of the doubling time (D_n) where the D_n ratio is a log(2) normalisation of the rate of increase.

$$\omega_{E,R} = D_E / D_R$$

Equation 2.4. Calculation of relative fitness ($\omega_{E,R}$). D_E is doubling time of the evolved cells and D_R is the doubling time of the reference strain.

2.2.14 MA assays

To determine the number and spectra of mutations accumulated in *E. coli* grown in aerobic and anaerobic environments, *E. coli* REL4536 (**Table 2.1**) was cultivated on aerobic and anaerobic DM agar plates supplemented with 0.2 g glucose/L (section 2.1.7.3).

2.2.14.1 Establishment of lineages

All MA lineages were started from single colonies of *E. coli* REL4536 (**Table 2.1**). The ancestral strain was streaked on a DM agar plate (section 2.1.7.3) and incubated for 24 h at 37ºC. To initiate the MA lineages, 100 colonies were randomly selected and each colony initiated an independent MA line by streaking on to aerobic or anaerobically prepared DM plates to establish 50 lineages in each environment.

2.2.14.2 Maintenance of lineages

The 50 MA lineages established in the aerobic environment were propagated for 180 single-colony bottlenecks and the 50 MA lineages established in the anaerobic environment were propagated for 144 single-colony bottlenecks. Colonies were transferred when they grew to a size that was readily detected by eye. Here, aerobic lineages were transferred every 23 to 25 h, while anaerobic lineages were transferred every 71 to 73 h.

All MA lineages were incubated at 37ºC in a Coy Laboratory Products Inc. (USA) forced-air incubator. Aerobic lineages were propagated in a biosafety cabinet (section 2.2.1). Anaerobic lineages were propagated in the anaerobic glove box (section 2.2.2) and placed in AnaeroJar canisters (Oxoid) to maintain an anaerobic atmosphere for cultures after removal from the chamber. Anaerobic lineages were incubated with a plate of calcium chloride pellets to remove excess moisture produced in the canister and Anaerobic Indicator (Oxoid) was placed in each canister to monitor anaerobic conditions.

2.2.14.2.1 Single-colony bottleneck procedure

For each bottleneck, the edge of a single colony was touched with a sterile inoculation loop and streaked onto a new plate to obtain single colonies. Agar plates were divided into eight sectors to allow eight separate lineages to be propagated in parallel per agar plate. Prior to streaking a new plate, spots were randomly marked on the base of the plate in each sector. Thus for the next transfer, the colony closest to the mark was selected for propagation, ensuring that there was no bias towards colonies of a particular size.

2.2.14.2.1.1 Estimation of number of generations

The number of cell divisions during the time period between bottlenecks was estimated by counting the average number of cells in 10 colonies of *E. coli* REL4536 after 24 h in the aerobic environment and the average number of cells in 10 colonies of *E. coli* REL4536 after 72 h in the anaerobic environment (section 2.2.14.2). The number of cell divisions, or generations, of colony growth were calculated as presented in **Equation 2.5**.

$$G = \frac{(logN - logNi)}{log2}$$

Equation 2.5. Estimation of the number of generations of mutation accumulation. G is the number of generations, N is the final number of cells in a colony while Ni is the number of cells initially present in the colony (Ni = 1).

2.2.14.2.2 Storage

MA lineages were frozen every 15 bottlenecks by suspending the streaked colony in 400 µL of 15% glycerol, mixing vigorously and storing at -85°C. For anaerobic lineages, anaerobic 15% glycerol (section 2.1.6.7) was used for storage.

2.2.14.2.3 Contamination test

MA lineages were routinely checked for contamination every 14 bottlenecks. Phage contamination tests (168), where lineages were examined for sensitivity to coliphages T5 and T6 (section 2.1.2), were conducted using representative colonies from each lineage. LB agar plates (section 2.1.7.1) were streaked downward with 20 µL T5 and T6 phage stock and allowed to dry. Then colonies were streaked perpendicularly across the phage streaks and plates were incubated at 37°C overnight. Sensitivity to T5 and resistance to T6 suggested that the correct strain of *E. coli* was present. *E. coli* B113 and *E. coli* DH5α were used as control strains for the T5 and T6 phage stocks (**Table 2.1**).

2.2.14.3 Mutation rate calculation

Using the ancestral genome length of 4,595,685 bp, the total number of mutations found across 24 lineages and environment-specific MA parameters, spontaneous mutation rates for aerobically and anaerobically grown $E.\ coli$ were calculated as described in **Equation 2.6**.

$$Mutation\ rate\ per\ genome\ per\ generation\ = \frac{m}{(L * N * T)}$$

$$Mutation\ rate\ per\ nucleotide\ per\ generation\ = \frac{m}{(L * N * T * G)}$$

$$Mutation\ rate\ per\ genome\ per\ day = \frac{m}{(L * D)}$$

Equation 2.6. Calculation of mutation rates from whole genome data. m is the total number of observed mutations, L is the number of lineages, N is the number of generations between bottlenecks, T is the number of bottlenecks (180 for aerobic lineages and 144 for anaerobic lineages), D is the number of days of evolution (180 for aerobic lineages and 432 for anaerobic lineages) and G is the size of the *E. coli* REL4536 genome (4,595,685 bp).

2.2.15 *sbcC* adaptive lineages

To analyse the spectra of mutations accumulated with an *E. coli sbcC* deletion mutant during adaptation to an anaerobic environment, *E. coli* REL4536 Δ*sbcC::cat* (**Table 2.1**, section 2.2.9.3.1) was cultivated for 1,000 generations in anaerobic DM25 (section 2.1.7.2.2).

2.2.15.1 Establishment of lineages

The *E. coli sbcC* mutant lineages were initiated from a single colony of *E. coli* REL4536 Δ*sbcC::cat* (**Table 2.1**, section 2.2.9.3.1.2). Cells were revived from frozen stocks (section 2.2.3) and streaked sequentially three times to a single colony on aerobic MG agar (section 2.1.7.4) and incubated for 24 h. After incubation, a well separated colony was randomly picked and used to inoculate 50 mL of aerobic DM25 in a 500 mL Erlenmeyer flask. The culture was grown overnight and used as the inoculum for all *sbcC* mutant lineages, thus the culture is henceforth referred to as the ancestral *E. coli* REL4536 Δ*sbcC::cat* culture. From this ancestral culture, 14 independent populations were created for experimental adaptive evolution under an anaerobic environment.

Establishment of the anaerobic cultures took place by inoculating 10 µL of the ancestral culture into 990 µL of anaerobic DM25 in seven out of eight wells within a 24-well plate. An un-inoculated media-only control was included alongside the group of seven cultures in each plate to assess cross-contamination within the plates. Two anaerobic plates were established in this way (14 anaerobic lineages in total) with lineages randomly numbered from 1 to 14. Aliquots of the ancestral *E. coli* REL4536 Δ*sbcC::cat* strain were stored at -85°C by mixing 500 µL of culture with 400 µL of glycerol saline (section 2.1.6.7) in a micro-centrifuge tube.

2.2.15.2 Maintenance of lineages

The *sbcC* mutant lineages were propagated for 1,000 generations in the anaerobic environment. Lineages were propagated every 24 h, as determined by growth courses to be the time taken to reach stationary phase (section 2.2.4).

Lineages were incubated at 37ºC in an Ecotron rotating incubator (Infors-HT Inc., Switzerland) with an orbital shaking of 150 RPM. Anaerobic lineages were propagated in the anaerobic glove box (section 2.2.2) and placed in Mitsubishi AnaeroPack rectangular gas boxes (Mitsubishi Gas Company Inc., Japan) to maintain an anaerobic atmosphere for cultures after removal from the chamber.

2.2.15.2.1 Daily lineage sub-culture

For each lineage, 10 µL of the previous day's culture was transferred into 990 µL fresh DM25 media every 24 h. To minimise any variation on culture growth caused by well position within the plate, the relative positions of the lineages within the blocks of eight wells in the 24-well plates were routinely rotated.

2.2.15.2.2 Storage

E. coli REL4536 Δ*sbcC::cat* anaerobic lineages were frozen every two weeks by mixing 500 µL of overnight culture with 400 µL of anaerobic glycerol saline (section 2.1.6.7) in a micro-centrifuge tube. Cultures were stored at -85°C.

2.2.15.2.3 Contamination test

sbcC mutant lineages were routinely checked for contamination every two weeks by performing phage contamination tests as previously described in section 2.2.14.2.3.

2.2.16 *In silico* analysis

2.2.16.1 Bioinformatic resources and software

All resources and software used in this study are listed in **Table 2.9**.

Table 2.9. Bioinformatic resources and software used.

Resource	Description/Application	Source	Reference
Artemis Release 16.0	Genome sequence display and analysis	https://www.sanger.ac.uk/resources/software/artemis/	(206)
Basic Local Alignment Search Tool (BLAST)	Finding regions of local similarity between sequences	http://blast.ncbi.nlm.nih.gov/Blast.cgi	(207)
Bowtie 2	Short sequence read aligner	http://bowtie-bio.sourceforge.net/index.shtml	(208)
breseq v 0.24	Computational pipeline for finding mutations in re-sequenced microbial genomes	http://barricklab.org/twiki/bin/view/Lab/ToolsBacterialGenomeResequencing	(209)
DESeq2	Analysis of differential gene expression	http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html	(210)
DNA Plotter Release 1.4	Genome visualisation	http://www.sanger.ac.uk/resources/software/dnaplotter/	(211)
Estimated Degree of Gene Expression in Prokaryotes (EDGE-pro)	Estimates gene expression levels from RNAseq data	http://ccb.jhu.edu/software/EDGE-pro/	(212)
European Molecular Biology Open Source Suite (EMBOSS)	Software analysis package for molecular biology	http://emboss.sourceforge.net/index.html	(213)
FALCOR	Fluctuation test analysis calculator	http://www.mitochondria.org/protocols/ FALCOR.html	(204)
FastQC	Quality control tool for high throughput sequence data	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/	(214)
FASTX-Toolkit v 0.0.13	Collection of tools for pre-processing manipulation of FASTA/FASTQ files	http://hannonlab.cshl.edu/fastx_toolkit/index.html	(215)

Resource	Description/Application	Source	Reference
The Gene Ontology (GO) project	A uniform controlled vocabulary of terms for describing gene products	http://geneontology.org/	(216)
Geneious Pro 5.6.2	Genome sequence alignment and analysis	http://www.geneious.com/	(203)
GenSkew	Computes genomic nucleotide skew data	http://genskew.csb.univie.ac.at/	(217)
Infernal 1.1	For finding RNA structure and sequence	http://infernal.janelia.org/	(218)
LinReg PCR 2013.0	Analysis of qPCR data	http://LinRegPCR.nl	(201)
Mauve 2.3.1	Genome alignment and comparison tool	http://gel.ahabs.wisc.edu/mauve/	(219, 220)
MUMmer 3.23	Multiple genome alignment tool	http://mummer.sourceforge.net/	(221)
Protein Analysis Through Evolutionary Relationships (PANTHER) 9.0	Protein and gene classification tool	http://www.pantherdb.org/	(222, 223)
Quality Assessment Tool (QUAST) 2.3	Quality assessment tool for genome assemblies	http://bioinf.spbau.ru/en/quast	(224)
R 3.0.0.	Software environment for statistical computing	http://www.r-project.org/	(225)
Rfam 11.0	RNA family database	http://rfam.sanger.ac.uk/	(226)
SPAdes 3.0.0	*De novo* genome assembler	http://bioinf.spbau.ru/spades	(227)
The Sequence Manipulation Suite	Web-based programs for analysing and formatting DNA and protein sequences	http://www.bioinformatics.org/SMS/	(228)
TIGRFAMS	Database of protein families	http://www.jcvi.org/cgi-bin/tigrfams/index.cgi	(229)
Trim Galore! 0.3.5	Trimming tool for sequence data	http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/	(230)

2.2.16.2 R simulator modelling

To predict the number of accumulated mutations, simulations were run on a mutation accumulation (MA) simulator programme designed for this project (unpublished data, Z.A. Park and B. Auvray). This MA simulator programme (code in **Appendix B**), run in R 3.0.0 (225), uses the number of generations between bottlenecks and an estimated mutation rate following a Poisson distribution to give an estimation of the minimum number of mutations obtained over a specified number of generations.

2.2.16.3 DNA high throughput sequencing analysis

2.2.16.3.1 Illumina sequencing

High-quality genomic DNA was sequenced on the Illumina HiSeq 2000 platform by the BGI (Shenzen, China). BGI used the Illumina Mate Pair Library Preparation Kit v2 to construct libraries with 2 kb mate-pair inserts. Paired-end sequencing with 90 bp reads was performed, such that 500-800 MB of clean data was obtained for each sample, resulting in at least a 30-fold sequencing read depth for each genome. Sequencing read depth (i.e. the average number of times any given base in the genome was sequenced) was calculated as presented in **Equation 2.7**.

$$Sequencing\ read\ depth = \frac{(L * N)}{G}$$

Equation 2.7. Calculation of sequencing read depth. L is the length of the sequenced reads (90 bp), N is the number of reads that were mapped and G is the size of the *E. coli* REL4536 genome (4,595,685 bp).

BGI provided data that had been filtered to remove reads containing ≥ 10% unreadable bases, ≥ 20% low quality (≤ Q20) bases, adapter contamination or duplicate read-pairs. FASTQ is a commonly used format for high throughput sequencing output and for each sample, two FASTQ files were provided. These files represent the forward and reverse reads and FastQC (214), a quality control tool for high throughput sequence data, was used to check the quality of the sequence data.

2.2.16.3.2 Mutation detection

The mutation detection pipeline is shown in **Figure 2.1**.

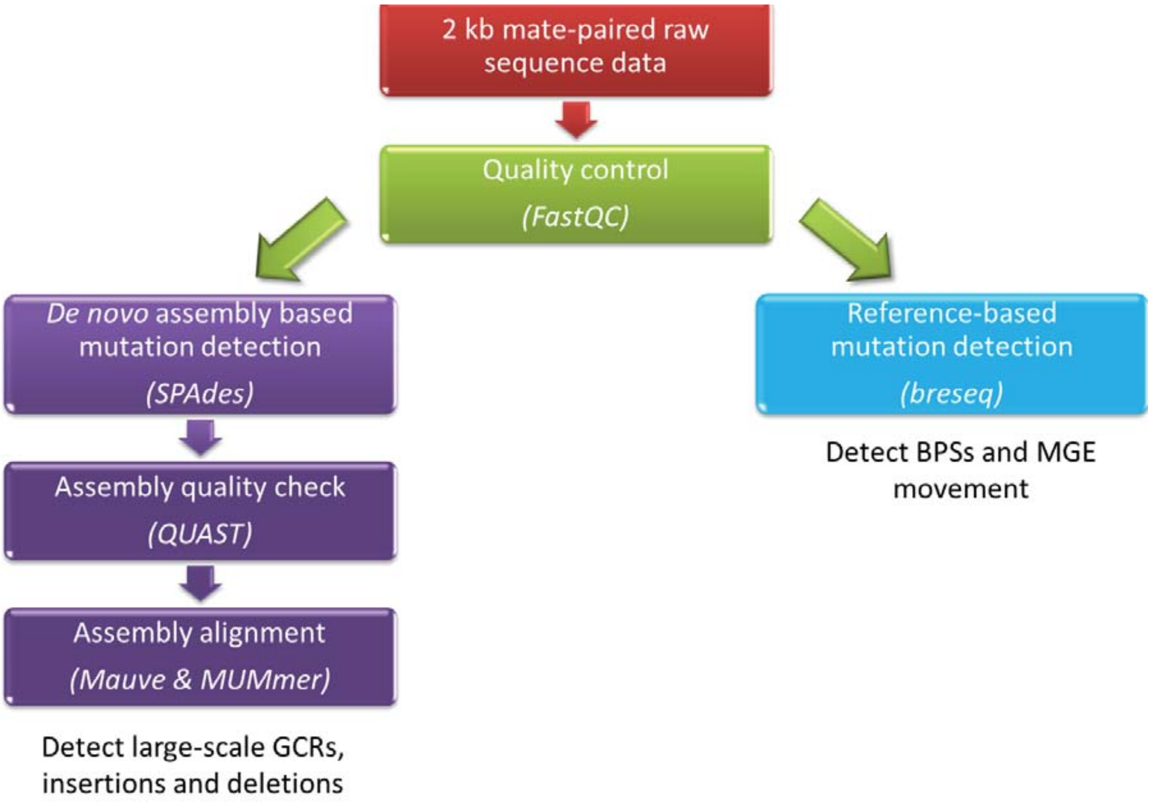
2 kb mate-paired raw sequence data
Quality control
(FastQC)
De novo assembly based mutation detection
(SPAdes)
Assembly quality check
(QUAST)
Assembly alignment
(Mauve & MUMmer)
Detect large-scale GCRs, insertions and deletions
Reference-based mutation detection
(breseq)
Detect BPSs and MGE movement

Figure 2.1. Bioinformatics pipeline developed to identify mutations within the genome sequences analysed in this study. Software programmes used are shown in brackets.

2.2.16.3.2.1 Reference-based mutation detection

To detect mutations, re-sequenced genomes were mapped to the reference *E. coli* REL4536 (**Table 2.1**) genome sequence. The reference genome was manually annotated by Finn (205) using information provided by Barrick et al. (2009) (197). Briefly, the genbank file of *E. coli* REL606 was downloaded from the NCBI database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Escherichia_coli_B_REL606_uid58803). The file was edited using Artemis (206) and European Molecular Biology Open Source Suite (EMBOSS) tools (213) to incorporate the 28 mutational differences known between the genomes of *E. coli* REL606 and *E. coli* REL4536.

BPSs (section 1.3.1) and MGE movement (section 1.3.3.1.1) were identified using breseq (209), a microbial genome re-sequencing analysis pipeline. Breseq uses reference-based alignment approaches to detect mutations relative to a reference genome. For each genome, the two FASTQ files, containing the sequencing reads, were mapped to the reference *E. coli* REL4536 genome sequence with default settings. Mutations were identified and displayed in the output directory. The quality of reads covering each BPS was subsequently manually checked to ensure that none of the detected BPSs were actually artefacts of poor quality sequencing reads.

2.2.16.3.2.2 *De novo* assembly based mutation detection

To detect insertions, deletions and large-scale GCRs (section 1.3.3), such as translocations and inversions, genomes were assembled *de novo. De novo* genome assembly was performed using the programme SPAdes 3.0.0 (227). SPAdes (227) is an assembler, with a built-in read error correction tool for Illumina reads, that can utilize mate-pair read information. Using *De Bruijn* graphs for genome assembly, a SPAdes assembly occurs in four stages (227), and even performs an error correction step devoted to repeat sequences detection (134). FASTQ files, containing the sequencing reads, were reverse complemented using the fastx reverse complement tool from the fastx toolkit (215). To assemble reads in SPAdes, the reverse-complemented reads were used as paired-end input while the sequencing reads were used as mate-pair input. The assembled scaffolds were assessed using the infoseq tool from the EMBOSS software suite (213) to determine contig length and GC percentage. *De novo* assemblies were assessed using QUality ASsessment Tool (QUAST) (224), a genome assembly evaluation tool that generates summary statistics to review the quality of assemblies. In this study, QUAST was used with the *E. coli* REL4536 genome sequence (section 2.2.16.3.2.1) serving as the reference sequence.

2.2.16.3.2.2.1 Genome alignment

MUMmer (221) is a genome alignment tool that allows for the alignment of contigs across a reference genome. To assess genome alignment, contigs obtained from the *de novo* assembly were aligned to the *E. coli* REL4536 reference genome (section 2.2.16.3.2.1) using the NUCmer pipeline on contigs that were greater than 1 kb in length. Synteny plots were generated using the ‘mummerplot’ utility. To further assess the assembled contigs, Mauve (219), a genome alignment tool, was used. Contigs obtained from the *de novo* assembly were aligned to the *E. coli* REL4536 genome using the progressiveMauve (220) algorithm with default parameters and the ‘Move Contigs’ option. Using genome alignments outputs from both MUMmer and Mauve, possible GCR sites were manually identified and experimentally verified by PCR (**Table 2.4**, section 2.2.7).

2.2.16.3.3 Cumulative GC-skew analysis

To determine cumulative GC-skews of evolved genomes, GenSkew, a genomic nucleotide skew calculator (217), was used with default parameters.

2.2.16.3.4 Codon usage analysis

To determine the codon usage of the REL4536 genome, the ‘Codon Usage’ program of the Sequence Manipulation Suite (228), was used with default parameters and the standard genetic code.

2.2.16.4 RNA high throughput sequencing analysis

2.2.16.4.1 RNAseq data analysis

An overview of the RNAseq *in silico* analysis is presented in **Figure 2.2**.

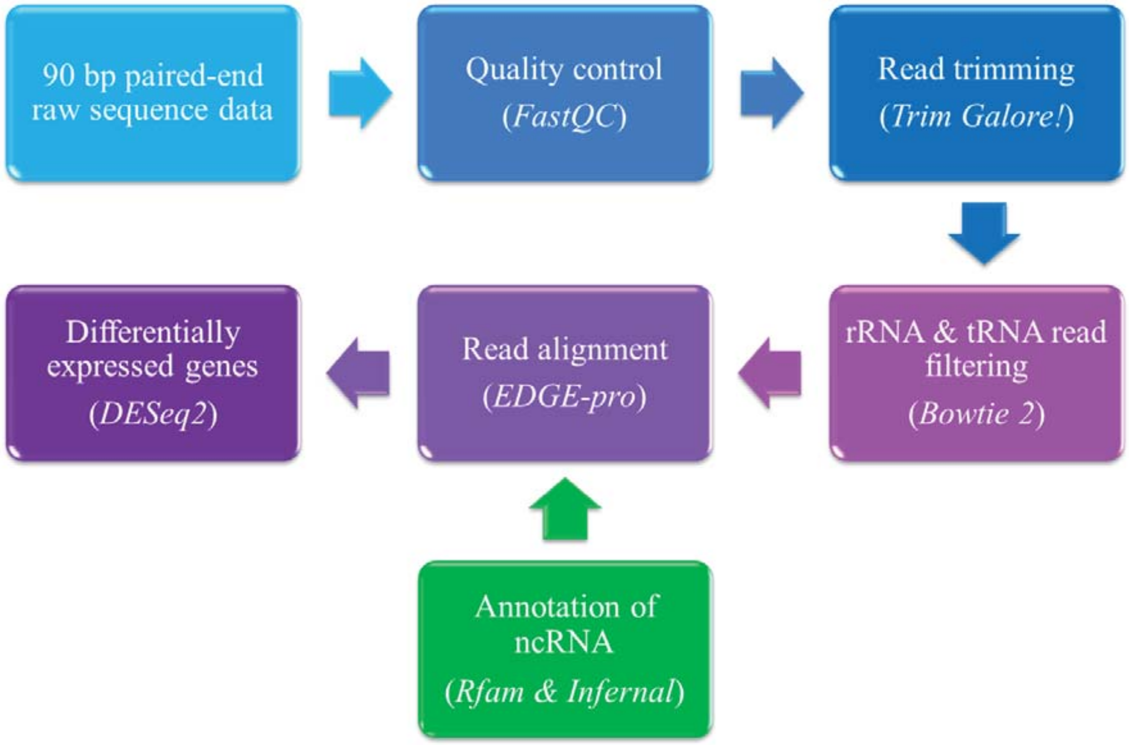
90 bp paired-end raw sequence data
Quality control (*FastQC*)
Read trimming (*Trim Galore!*)
rRNA & tRNA read filtering (*Bowtie 2*)
Read alignment (*EDGE-pro*)
Differentially expressed genes (*DESeq2*)
Annotation of ncRNA (*Rfam & Infernal*)

Figure 2.2. Overview of the workflow for RNAseq *in silico* analysis. Software programmes used are shown in brackets.

2.2.16.4.2 Illumina sequencing

High-quality total RNA was sequenced on the Illumina HiSeq 2000 platform by BGI. Messenger RNA (mRNA) was enriched for using the Ribo-Zero™ rRNA Removal Kits for bacteria (Epicentre, USA). Libraries with 200 bp inserts were constructed, paired-end sequencing was performed on each library and 90 bp reads were generated. BGI provided data that had been filtered to remove reads containing ≥ 10% unreadable bases, ≥ 20% low quality (≤ Q20) bases or adaptor contamination. For each sample, two FASTQ files were provided; these files represent the forward and reverse reads. To assess the quality of the FASTQ files, FastQC (214) was used. The first 10 bp of all reads and any reads with a quality score ≤ Q28 were trimmed using Trim Galore!, a wrapper script that can be used for quality checks, adapter trimming and trimming of paired-end and single-end data (230).

2.2.16.4.3 Removal of rRNA and tRNA reads

To identify rRNA genes, tRNA genes and non-coding RNAs (ncRNA), the Rfam database (226) with the INFERNAL (218) software package was used with default parameters. A multi-fasta file containing all rRNA, tRNA and non-coding RNA sequences present in *E. coli* REL4536 was created. Sequencing reads files were aligned to this RNA multi-fasta file using Bowtie 2 (208) with default parameters. Reads that did not align to rRNA, tRNA and ncRNA sequences were saved to a FASTQ file for further analysis.

2.2.16.4.4 Sequence read alignment

EDGE-pro (Estimated Degree of Gene Expression in PROkaryotes) is a programme specifically designed for RNAseq analysis from bacteria (212). The programme uses Bowtie 2 (208), with its default parameters, for the alignment of sequence reads in order to make an estimate of gene expression directly from alignment results. Using results from section 2.2.16.4.3, a table containing the coordinates of the tRNA and rRNA genes in the *E. coli* REL4536 genome was created. A protein translation table containing the coordinates of the protein coding genes from the *E. coli* REL4536 genome was also created. EDGE-pro, with default parameters, was run on each set of FASTQ reads. The EDGE-pro output, containing the total number of aligned reads and the reads per kilobase of gene per million reads mapped (RPKM), was used to analyse differential gene expression in DESeq2 (210).

2.2.16.4.4.1 Differential gene expression

DESeq2 is a programme which tests for differential expression based on a negative binomial distribution model (210). EDGE-pro output (section 2.2.16.4.4) was imported into DESeq2 and data was grouped according to the environment (aerobic or anaerobic). RPKM data was normalised and data grouping was assessed using a principal coordinate analysis (PCA). DESeq2 was used to identify differential expression between aerobic and anaerobic environments; a gene was identified as being differentially expressed between the two environments if it had a *p*-adj value less than 0.05 and a fold change of at least 2.

2.2.16.4.4.2 Gene ontology (GO) analysis

Gene Ontologies (GO), controlled vocabularies of defined terms, can be used to describe gene products in terms of their associated biological processes and molecular functions. Using The GO Project (216) and PANTHER (Protein ANalysis THrough Evolutionary Relationships) databases (222, 223), the significantly up-regulated genes in aerobically and anaerobically grown cells were classified into GO biological processes.

2.2.17 Statistical analysis

All statistical analyses were performed in GenStat 17th Edition (VSN international, UK) (231). To assess differences in mutation rates and spectra between aerobic and anaerobically grown cells, standard one-sided Mann-Whitney U-tests were performed. To assess differences in fluctuation assay data and in qPCR gene expression data between aerobic and anaerobically grown cells, standard two-tailed, homoscedastic Student’s *t*-tests were performed. For the analysis of competitive fitness data, standard one-tailed, heteroscedastic Student’s *t*-tests were performed. Standard one-sided Fisher’s exact tests and Pearson’s chi-squared tests were also utilised . For all tests, probability values less than 0.05 (i.e. $p < 0.05$) were considered to be significant.

Chapter Three: Mutation rate in aerobic and anaerobic environments

3.1 Introduction

Mutations are the ultimate source of genetic variation. The rates that mutations occur at can mediate the genetic organisation of an organism, and so, mutations play a central role in the evolutionary trajectory of an organism. MA studies (section 1.6.1.3) are a direct way to measure spontaneous mutation rates. In MA studies, all mutations (except highly deleterious ones) are allowed to accumulate within populations, due to the absence of a selective pressure (3). Through the combination of MA studies with whole genome re-sequencing, it is now possible to directly detect all mutations in bacterial genomes (4, 121, 122).

Although *E. coli* is one of the most widely studied microbes, estimates of its mutation rate when grown aerobically vary greatly, ranging from 1.7 × 10^{-4} to 1.0-4.0 × 10^{-3} mutations per genome per generation (11-13). In addition, mutation rate estimates based on the whole genome have not been reported for anaerobically grown *E. coli*. It was originally hypothesized that the rate at which mutations arise under aerobic conditions is greater than the rate at which they arise under anaerobic conditions, due to differences in the presence of mutagenic agents (e.g. ROS in aerobically grown cells). However, the study by Sakai et al observed a higher mutation rate in anaerobically grown *E. coli* than aerobically grown *E. coli* (28). Thus, the aim of this chapter was to use MA techniques, whole genome re-sequencing and traditional fluctuation assays to determine:

- What is the genome-wide rate at which mutations spontaneously occur in *E. coli* grown in aerobic and anaerobic environments?

3.2 Results and discussion

3.2.1 Fluctuation assays

Mutation rates for *E. coli* grown under aerobic and anaerobic conditions were first estimated using classical Luria-Delbruck fluctuation assays (section 2.2.12) (149). These frequently-used assays, where the distribution of spontaneous mutants among parallel cultures is used to calculate the mutation rate, are a rapid and simple way to estimate relative mutation rates based on a single locus. Mutation rate per locus estimates were obtained for *E. coli* REL4536 grown in DM1000 medium (section

2.1.7.2) in both the aerobic and anaerobic environment. Assays were scored for resistance to nalidixic acid (Nal^R) and for arabinose utilisation (Ara^+) as described in section 2.2.12.

Mutation rates were calculated using the MSS-MLE method (150), as implemented in FALCOR (204) (section 2.2.12). For aerobically grown cells, the mutation rates ranged from 6.50×10^{-10} to 9.60×10^{-10} mutations per generation while for anaerobically grown cells, the mutation rates were considerably higher, ranging from 4.04×10^{-9} to 4.88×10^{-9} mutations per generation (**Table 3.1**). The per locus aerobic mutation rates obtained for this study are consistent with those previously reported *via* fluctuation assays. Using the Nal^R phenotype, Lee et al. (2012) calculated a rate of 4.3×10^{-10} mutations per locus per generation for aerobically grown *E. coli* (13), which falls within the 95% confidence limits (CL) of our rate of 6.5×10^{-10} mutations per locus per generation for aerobically grown cells (**Table 3.1**). For the Ara^+ phenotype, Sniegowski et al. (1997) reported rates of approximately 1.00×10^{-9} mutations per locus per generation for aerobically grown cells (155), which falls within the 95% CL of our rate of 9.6×10^{-10} mutations per locus per generation for aerobically grown cells (**Table 3.1**).

The mutation rates discussed above are locus-based rates and as such, are relative rates. To be able to compare results of fluctuation assays to published mutation rates, it is essential to obtain a normalised rate. To obtain the mutation rate per nucleotide, the mutation rate per locus estimates were normalised to the number of nucleotides that are assumed to be responsible for conferring the mutant phenotype, as implemented by Lee et al. (2012) (13). Resistance to nalidixic acid may be conferred by 18 different point mutations in the *gyrA* gene, and 2 point mutations in the *gyrB* gene (232-234). For *E. coli* REL606, there are two possible single point mutations in the *araA* gene that confer the ability to utilize arabinose (235). It is assumed that the same two mutations can give *E. coli* REL4536 arabinose utilisation ability as well. Therefore, for the aerobic environment, the mutation rates range from 0.40×10^{-10} to 4.80×10^{-10} mutations per nucleotide per generation while for the anaerobic environment, the mutation rates range from 2.20×10^{-10} to 2.44×10^{-9} mutations per nucleotide per generation (**Table 3.1**).

In general, the mutation rates were found to be higher from growth under anaerobic conditions than under aerobic ones, though these differences were not significant by

Student's t-test ($p > 0.05$) (**Table 3.1**). These findings are consistent with Sakai et al. (2006), where a roughly two-fold higher anaerobic, as compared to aerobic, mutation rate was reported using a *rpsL*-based assay (28). In contrast, in this study, the Nal^R and Ara$^+$ scored fluctuation assays gave a six- and five- fold higher anaerobic, as compared to aerobic, mutation rate, respectively.

Table 3.1. *E. coli* mutation rates calculated from fluctuation assays.

Environment	Phenotype	Mutation rate per locus per generation (× 10^{-9})	95% CL (× 10^{-9})	Mutation rate per nucleotide per generation (× 10^{-9})	95% CL (× 10^{-9})
Aerobic					
	Nal^R	0.65	0.31 - 1.07	0.04	0.02 - 0.06
	Ara$^+$	0.96	0.52 - 1.49	0.48	0.26 - 0.75
Anaerobic					
	Nal^R	4.04	1.75 - 6.97	0.22	0.10 - 0.39
	Ara$^+$	4.88	2.51 - 7.80	2.44	1.26 - 3.90

However, as this study's estimates are dependent on the locus being analysed, they likely do not reflect genome-wide mutation rates (47, 48). Lee et al. (2012), scored for resistance to rifampicin as well as the Nal^R phenotype in their study, and after normalisation, the rates from each locus converged to rates of 3.30 × 10^{-9} and 2.10 × 10^{-9} mutations per nucleotide per generation, respectively (13). Such a phenomenon was not observed in this study. In fact, the range of mutation rates that were obtained for both aerobically and anaerobically grown cells varied greatly by locus, suggesting that the mutation rate may vary across the *E. coli* genome, with arabinose utilisation conferring a higher mutation rate than resistance to nalidixic acid. The number of mutations assumed to be responsible for each mutant phenotype may not be accurate and may explain the disagreement between the per nucleotide mutation rates observed in this study. These results point to the danger of analysing just one phenotype in mutation rate studies as estimates based on a single locus can be very biased and possibly unreliable. When studying just one locus, synonymous mutations that do not produce a mutant phenotype do not contribute to the mutation rate calculation and so, mutation rates calculated from fluctuation assays can be underestimated (4). Whole genome approaches, where these biases are eliminated, will provide more comprehensive estimates of the spontaneous mutation rate.

3.2.2 MA simulations *in silico*

A MA simulator programme (code in **Appendix B)** was developed in the statistical package R (225) to enable a prediction of how many sub-culturing events would need to be undertaken aerobically and anaerobically, under the MA experiment growth conditions, to accumulate a practical number of spontaneous mutations to analyse. The MA simulator programme is based on the accumulation of mutations following a Poisson distribution, and uses inputs of the number of generations between bottlenecks, and an estimated mutation rate, to simulate the accumulation of mutations obtained over a specified number of generations.

Using the most accurate estimate of the aerobic *E. coli* mutation rate available at the time (136, 197), the MA simulator programme was used to do an *in silico* MA study for *E. coli* REL4536 bottlenecked under aerobic and anaerobic environments (section 2.2.16.2). As Sakai et al. (2006) reported a two-fold higher anaerobic, as compared to aerobic, mutation rate (28), the anaerobic mutation rate used for the *in silico* simulations was double the estimate of the aerobic *E. coli* mutation rate. The number of generations between bottlenecks for aerobic and anaerobic lineages was estimated as described in section 2.2.14.2.1.1. The *in silico* analysis indicated that the mean number of mutations expected in any lineage after 180 rounds of bottlenecking was one to two mutations for the aerobic environment and four to five mutations for the anaerobic environment (**Figure 3.1**). As estimates of mutation rates from a few mutations are prone to sampling inaccuracies and are an improper representation of the true value (161), it was decided that 50 MA lineages would be set up in each environment to ensure the accumulation of enough mutations.

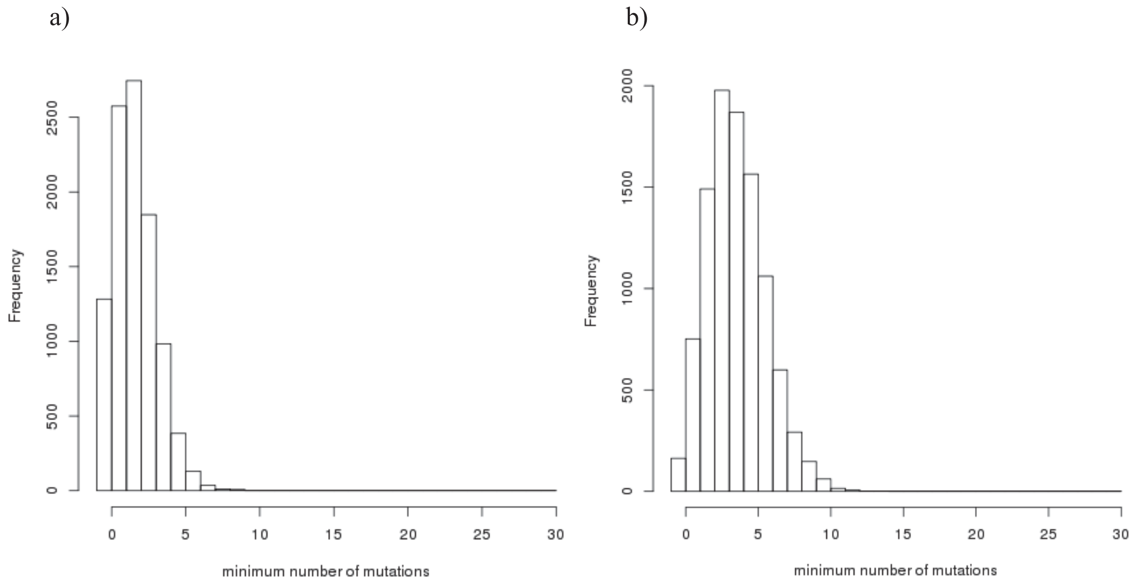

Figure 3.1. MA simulations *in silico*. 10,000 simulations were run on a programme designed to determine the minimum number of mutations expected in any lineage after 180 bottlenecks in the a) aerobic environment using a mutation rate of 4.6 × 10^{-4} per genome per generation and with bottlenecks every 25 generations and b) anaerobic environment using a mutation rate of 9.2 × 10^{-4} per genome per generation and with bottlenecks every 24 generations.

3.2.3 MA lineages

To accumulate spontaneous mutations in *E. coli* grown in aerobic and anaerobic environments, MA lineages of *E. coli* REL4536 were cultivated on DM agar plates as described in section 2.2.14. In summary, in each environment, 50 lineages were established and subjected to regular bottlenecks. Aerobic lineages were single-colony transferred daily (approx. 25 generations), for 180 bottlenecks. The anaerobic lineages were transferred every three days (approx. 24 generations) for 144 bottlenecks. After the lineages had been maintained for the desired number of bottlenecks, single colonies from 24 lineages from each environment were genome re-sequenced on an Illumina HiSeq 2000 instrument.

3.2.3.1 Next-generation sequence analysis

Whole genome re-sequencing of the genomes yielded at least 2 million raw sequencing reads for each sample (see **Table 3.2** and **Table 3.3**). To detect mutations, reference-based and *de novo* assembly based mutation detection methods were used (section 2.2.16.3.2). Summary statistics of the reference-based mapping and the QUAST (224) evaluated *de novo* assemblies are presented in **Table 3.2** and **Table 3.3**. Due to the sequencing strategy that was chosen, where multiple clones were sequenced on the same lane, the genomes were sequenced to differing depths, resulting in at least 30-fold sequencing read depth, calculated as described in **Equation 2.7**, for each genome.

3.2.3.2 Mutation identification

Using the mutation detection pipeline described previously in section 2.2.16.3.2, mutations were identified for 24 aerobic and 24 anaerobic clones. Identified mutations included BPSs, indels, and GCRs. In total, 295 mutations were identified. However, 13 mutations (12 BPSs and 2 GCRs) were excluded from the analysis as they were found in clones of lineages that had been maintained on the same agar plates and under these circumstances the possibility of cross-contamination between the lineages was higher, than if the lineages were maintained on different plates. Specifically AE-180-10, AE-180-14 and AE-180-16 had five shared mutations out of 16 mutations, AE-180-38 and AE-180-40 had one shared BPS out of a total of four, AN-144-34 and AN-144-36 had two shared mutations out of 10 total mutations while AN-144-46 and AN-144-50 had three shared BPSs out a total of 26 mutations. The relatively small number of shared mutations amongst the lineages would suggest that any cross-contamination would have occurred early on in the experiment, where there was greater chance for errors to have occurred due to technical experience in maintaining these MA lineages. That cross-contamination was the likely cause of the shared mutations, in these instances, the shared mutations were assigned to one lineage and only counted once.

3.2.3.2.1 BPS and indel detection

Using breseq (209) for reference-based mapping to the *E. coli* REL4536 reference genome sequence (section 2.2.16.3.2.1), BPSs (section 1.3.1) and indels (section 1.3.2) were detected. Indels were manually verified by analysing the alignment of the reads. In total, 147 BPSs were detected, with 74 BPSs and 73 BPSs occurring in aerobically and anaerobically grown clones, respectively. In addition, a total of 37 indels were detected, with 17 indels and 20 indels occurring in aerobically and anaerobically grown clones, respectively. These mutations will be discussed in greater depth in Chapter 4, regarding mutation spectra.

Table 3.2. Genome re-sequencing reference-based mapping and *de novo* assembly statistics for aerobic MA clones.

Lineage	Total reads	Unmapped reads	Mapped reads	Reads mapped (%)	Sequencing depth (fold coverage)	No. of contigs	No. of contigs >1000 bp	Largest contig (bp)	N50 (bp)	Genome coverage (%)
AE-180-02	8,024,374	1,402,719	6,621,655	82.52	130	73	13	11,911,991	668,405	99.24
AE-180-04	8,990,332	2,002,544	6,987,788	77.73	137	57	15	1,272,958	666,533	99.30
AE-180-06	8,383,852	1,519,626	6,864,226	81.87	134	61	16	703,912	402,575	99.28
AE-180-08	7,154,970	1,209,032	5,945,938	83.10	116	52	12	1,392,013	665,610	99.32
AE-180-10	8,095,674	1,781,202	6,314,472	78.00	124	51	12	1,389,514	665,414	99.32
AE-180-12	7,602,196	1,643,916	5,958,280	78.38	117	56	14	1,392,263	665,412	99.47
AE-180-14	9,535,352	2,211,976	7,323,376	76.80	143	58	11	1,388,925	703,705	99.27
AE-180-16	8,891,660	1,895,198	6,996,462	78.69	137	58	12	1,194,983	698,959	99.26
AE-180-20	7,770,974	1,544,146	6,226,828	80.13	122	51	12	1,195,563	703,772	99.50
AE-180-22	8,594,742	1,867,347	6,727,395	78.27	132	56	16	996,411	663,341	99.15
AE-180-24	8,744,610	1,773,339	6,971,271	79.72	137	62	14	744,448	599,521	99.32
AE-180-26	8,619,056	1,816,332	6,802,724	78.93	133	53	14	1,390,655	665,244	99.25
AE-180-28	7,999,534	1,527,627	6,471,907	80.90	127	11	11	1,388,863	665,537	98.20
AE-180-30	6,616,964	1,453,850	5,163,114	78.03	101	62	11	1,391,079	668,025	99.10
AE-180-32	8,573,024	1,739,539	6,833,485	79.71	134	68	17	698,941	501,414	99.28
AE-180-34	7,241,872	1,420,686	5,821,186	80.38	114	116	14	742,400	645,379	99.30
AE-180-36	7,901,438	1,634,714	6,266,724	79.31	123	55	13	743,533	665,061	99.45
AE-180-38	7,264,328	1,655,472	5,608,856	77.21	110	57	13	743,551	668,351	99.34
AE-180-40	8,413,810	1,654,246	6,759,564	80.34	132	53	11	1,391,505	703,427	99.41
AE-180-42	9,888,786	2,471,822	7,416,964	75.00	145	58	14	743,168	665,154	99.38
AE-180-44	7,847,260	1,794,934	6,052,326	77.13	119	48	14	742,754	645,868	99.28
AE-180-46	8,517,150	1,857,425	6,659,725	78.19	130	50	13	1,389,295	665,342	99.28
AE-180-48	8,051,738	1,813,381	6,238,357	77.48	122	56	13	789,729	668,227	99.35
AE-180-50	9,664,106	2,297,462	7,366,644	76.23	144	51	11	1,390,223	700,544	99.39

Table 3.3. Genome re-sequencing reference-based mapping and *de novo* assembly statistics for anaerobic MA clones.

Lineage	Total reads	Unmapped reads	Mapped reads	Reads mapped (%)	Sequencing depth (fold coverage)	No. of contigs	No. of contigs >1000 bp	Largest contig (bp)	N50 (bp)	Genome coverage (%)
AN-144-02	7,977,810	2,179,919	5,797,891	72.68	114	69	15	996,177	664,715	99.20
AN-144-04	8,105,756	2,113,740	5,992,016	73.92	117	51	13	1,391,268	668,078	99.26
AN-144-06	8,828,850	2,691,791	6,137,059	69.51	120	56	11	1,390,455	700,894	99.30
AN-144-08	7,069,936	2,021,098	5,048,838	71.41	99	55	12	1,394,242	665,426	99.29
AN-144-10	3,909,138	1,168,856	2,740,282	70.10	54	12	12	785,620	664,303	98.08
AN-144-12	4,438,662	1,223,138	3,215,524	72.44	63	68	16	1,307,480	579,691	99.29
AN-144-14	4,099,646	1,225,781	2,873,865	70.10	56	77	13	1,195,555	700,702	99.43
AN-144-16	3,969,008	1,251,319	2,717,689	68.47	53	52	14	1,389,012	578,673	99.51
AN-144-20	3,915,008	1,153,805	2,716,203	69.38	53	64	11	1,591,012	700,497	99.38
AN-144-22	3,599,162	1,177,897	2,421,283	67.27	47	47	11	1,388,888	667,924	99.42
AN-144-24	3,735,868	977,924	2,757,944	73.82	54	46	13	1,156,719	702,020	99.43
AN-144-26	4,118,682	1,363,821	2,754,861	66.89	54	68	13	741,258	664,713	99.33
AN-144-28	3,071,820	1,687,212	1,384,608	45.07	60	54	12	1,192,204	664,725	99.65
AN-144-30	2,854,216	917,429	1,936,787	67.86	38	47	12	744,143	665,122	99.44
AN-144-32	2,242,106	665,967	1,576,139	70.30	31	63	17	931,436	402,018	99.28
AN-144-34	2,990,526	946,894	2,043,632	68.34	40	60	12	817,571	664,850	99.19
AN-144-36	3,649,776	1,514,188	2,495,588	68.38	49	68	13	742,953	664,725	99.12
AN-144-38	4,314,884	1,479,345	2,835,539	65.72	56	50	12	936,700	664,998	99.34
AN-144-40	4,025,984	1,293,867	2,732,117	67.86	54	65	16	788,518	456,349	99.71
AN-144-42	4,217,392	1,278,102	2,939,290	69.69	58	55	13	741,332	664,715	99.35
AN-144-44	5,279,708	1,653,572	3,626,136	68.68	71	63	15	738,926	666,273	99.34
AN-144-46	6,107,634	1,907,222	4,200,412	68.77	82	50	14	815,359	641,511	99.05
AN-144-48	3,252,066	1,040,985	2,211,081	67.99	43	57	14	936,590	667,529	99.27
AN-144-50	5,491,296	1,971,580	3,519,716	64.10	69	66	16	807,135	640,201	98.67

3.2.3.2.2 GCR detection

As it had previously been implicated that GCRs (section 1.3.3) are prevalent in anaerobically grown cells (28), it was very important to be able to accurately identify these mutations within the genomes for this study. It is noted however, that many studies investigating mutation rates (13, 136, 137) do not report GCRs, as genome sequencing approaches have only recently been applied to MA studies and there is still some technical difficulty in identifying GCRs. Recently, Raeside et al. (2014) used optical mapping to identify GCRs in clones from 12 populations of Lenksi's LTEE (section 1.7.1.1) and they were able to identify 110 GCRs over 40,000 generations of evolution (35). In this study, several different software programmes were initially tested on a test sequence dataset containing GCRs, i.e. the rifampicin resistant marker described in section 2.2.13.1, until a suitable approach had been determined.

3.2.3.2.2.1 Challenges with GCR detection

The sequence reads generated by the sequencing instruments can computationally be joined together to reconstruct the target sequence in a process known as *de novo* assembly (134). GCRs are often mediated by large repetitive sequences, such as IS elements (section 1.3.3.1.1.1.1), and it is these repeat regions which often pose the major challenge to GCR detection (135) as individual sequence reads are not long enough to capture the new combinations of unique sequences that flank the repetitive sequence, as a result of repeat-sequence mediated recombination. Instead, sequence reads that map to repetitive regions can be mapped to multiple locations, making it difficult to interpret the data. In fact, many of the reference-based mapping programmes ignore any reads that map to multiple locations (62), and so, many GCRs are probably missed. Thus, if researchers restrict their analyses to reference-based mapping programmes, GCRs are probably not detected and reported. Sequence pair information is also critical for detecting GCRs, as this enables unique reads that flank repeated sequences to be linked. However, if the pair distance (library insert size) is insufficient to span the junctions of the rearrangement, GCRs cannot be detected. Nonetheless, by using *de novo* assembly and mate-pair information, GCRs with repetitive regions shorter than the library insert size can be resolved (135, 236).

In the reference *E. coli* REL4536 genome, IS elements are the most prevalent MGE, with IS element IS*150* being the largest at a length of 1446 bp (**Table 3.4**). To ensure that all IS element meditated GCR mutations could be detected, 2 kb insert libraries were used for sequencing. For this study, we hoped to find two general types of GCRs: (1) those involving the movement of MGEs or (2) those involving the rearrangement of fragments of DNA greater than 2 kb in length; the latter GCRs will collectively be referred to as large-scale GCRs throughout this thesis.

Table 3.4. Details of IS elements present in *E. coli* REL4536.

IS element	Frequency	Length (bp)
1S*1*	26	777
1S*2*	1	1,128
1S*3*	5	1,260
1S*4*	1	1,438
1S*30*	1	1,187
1S*150*	7	1,446
1S*186*	5	1,349
1S*600*	2	1,180
1S*911*	3	1,198

3.2.3.2.2.2 Large-scale GCRs

Using breseq (209) for reference-based mapping to the REL4536 reference genome sequence, GCRs involving MGEs were detected (section 2.2.16.3.2.1) and manually verified by analysing the alignment of the reads. In total, 91 GCRs were identified, with 28 and 63 events occurring in aerobically and anaerobically grown clones respectively. These mutations will be discussed in greater depth in Chapter 4, regarding mutation spectra.

Large-scale GCRs, mediated by repeat sequences greater than 2 kb in length, or where the mate-pair sequencing information was insufficient, were identified by assembling genomes *de novo* (as described in section 2.2.16.3.2.2) and aligning the assembled contigs to the REL4536 reference genome to identify differences in gene synteny (section 2.2.16.3.2.2.1). Initial analysis indicated that four aerobic clones and 10 anaerobic clones contained large-scale GCRs, with the detection of 23 events across all lineages.

Identified synteny breakpoints were validated by PCR (section 2.2.7). Overall, only seven (**Figure 3.2** and **Figure 3.3**) of the 23 large-scale GCRs identified by the *in silico* analysis were verified by PCR (**Table 3.5**). For aerobic clones, five of the six large-scale GCRs were confirmed as true GCRs while for anaerobic clones, only two of the 17 large-scale GCRs were experimentally proven. Depictions of these GCRs, as well as their sizes, can be seen in **Figure 4.8** and these mutations will be discussed in greater depth in Chapter 4, regarding mutation spectra.

Despite only 7 of the 23 GCRs predicted by assembly being experimentally verified, the 16 that were not validated experimentally were associated with large repetitive loci that were not able to be uniquely identified using the 2 kb mate-pair strategy that was used. Thus, the assembly programme consistently had trouble assembling these regions accurately and likely falsely identified GCRs for these regions. It is also possible that verification of these mutations failed due to the difficulty of PCR amplification across large repetitive regions of the genome. In bacteria, the three different rRNAs are co-transcribed from a single operon (*rrn*) which is approximately 5 kb in length. In *E. coli*, there are usually seven copies of the rRNA operon. As mentioned previously, multiple copies of repeated sequences are difficult to assemble due to ambiguous read mapping and rRNA operons are no exception.

While rRNA operons are known to be regions that assembly programmes have problems with, other troublesome loci for this *E. coli* dataset were rearrangement hot spot (*rhs*) genes. In the reference *E. coli* REL4536 genome, there are five *rhs* genes scattered around the genome – *rhsA*, *rhsB*, *rhsC*, *rhsD* and *rhsE*; with each gene being at least 4 kb long. First discovered in 1984 (237), the biological function of these genes is still not known. Each *rhs* gene has a core region, which is homologous across all five genes, and a divergent core extension sequence (238). The *rhs* core regions provide some of the most significant sequence repetition in the *E. coli* REL4536 genome and these genes have been shown to be sites where GCRs frequently occur in other studies (237, 239). Unfortunately, for this study, the 2 kb mate-pair strategy that was used for sequencing the MA clone genomes did not allow for accurate coverage of the junctions of the *rhs* genes. As such, assembly programmes were expected to be unstable in these regions due to the repetitive nature of these genes. For this dataset, only one of the verified seven large-scale GCRs involved an *rhs* gene (**Table 3.5**).

Additionally, due to the sequencing strategy that was employed, genomes of aerobically grown clones were sequenced, on average, to 100-fold coverage; almost twice the depth of anaerobically grown clones, which were sequenced, on average, to 50-fold coverage (**Table 3.2** and **Table 3.3**). Low sequencing depth makes it difficult to distinguish between real variations in the genome and assembly artefacts (240). Thus, the higher rate of false-positives observed in anaerobically grown clones could be due to the differences in sequencing depth. As it is, all 16 of the rearrangements that were not validated experimentally were in areas of around 30-fold coverage.

Lastly, the sequencing process also provides its own bias and challenges. For this study, mate-pair libraries were sequenced on an Illumina platform, generating short read lengths (90 bp) covering insert distances of 2 kb. While very effective for re-sequencing analysis, *de novo* assembly with short read lengths is not commonly performed as longer read lengths and longer insert distances are more effective in resolving gaps and areas that are difficult to assemble, such as repeated sequences (135). Furthermore, using only mate-pair data for genome assembly is challenging. The protocol used to construct the Illumina mate-pair libraries results in a non-negligible amount of read pairs with an incorrect distance or an incorrect orientation. For more accurate *de novo* assembly, assemblies are now often produced using hybrid approaches where high-depth, short-read length sequencing is complemented with lower-depth, longer read sequencing to get the best possible assembly. One such longer read sequencing technology would be the single-molecule based Pacific Biosciences RS sequencer (PacBio) system (128). PacBio sequencing generates reads of long lengths (up to about 10 kb) and the longer reads should allow for the reconstruction of entire genomes with ease. GCR detection will also improve with the increased read length, as the chances of individual sequence reads being long enough to capture the unique sequences that flank repeat sequences will be higher (240). However, PacBio sequencing data has high error rates and it is ideal to combine it with more precise, shorter read technologies. This hybrid approach has already been used for the *de novo* assembly of bacterial genomes to great effect (241-243). Unfortunately, only mate-pair reads were available for this study.

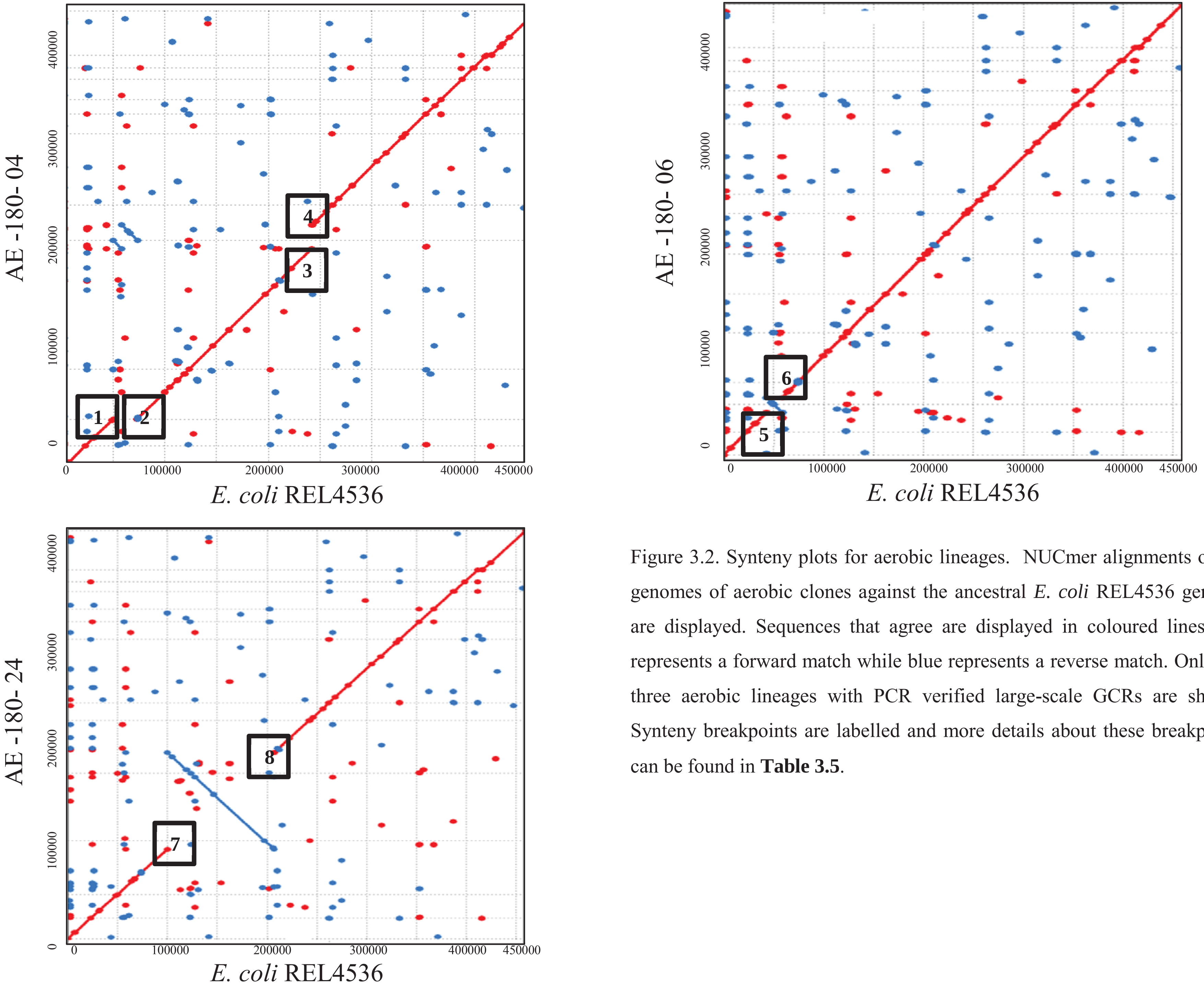

Figure 3.2. Synteny plots for aerobic lineages. NUCmer alignments of the genomes of aerobic clones against the ancestral *E. coli* REL4536 genome are displayed. Sequences that agree are displayed in coloured lines; red represents a forward match while blue represents a reverse match. Only the three aerobic lineages with PCR verified large-scale GCRs are shown. Synteny breakpoints are labelled and more details about these breakpoints can be found in **Table 3.5**.

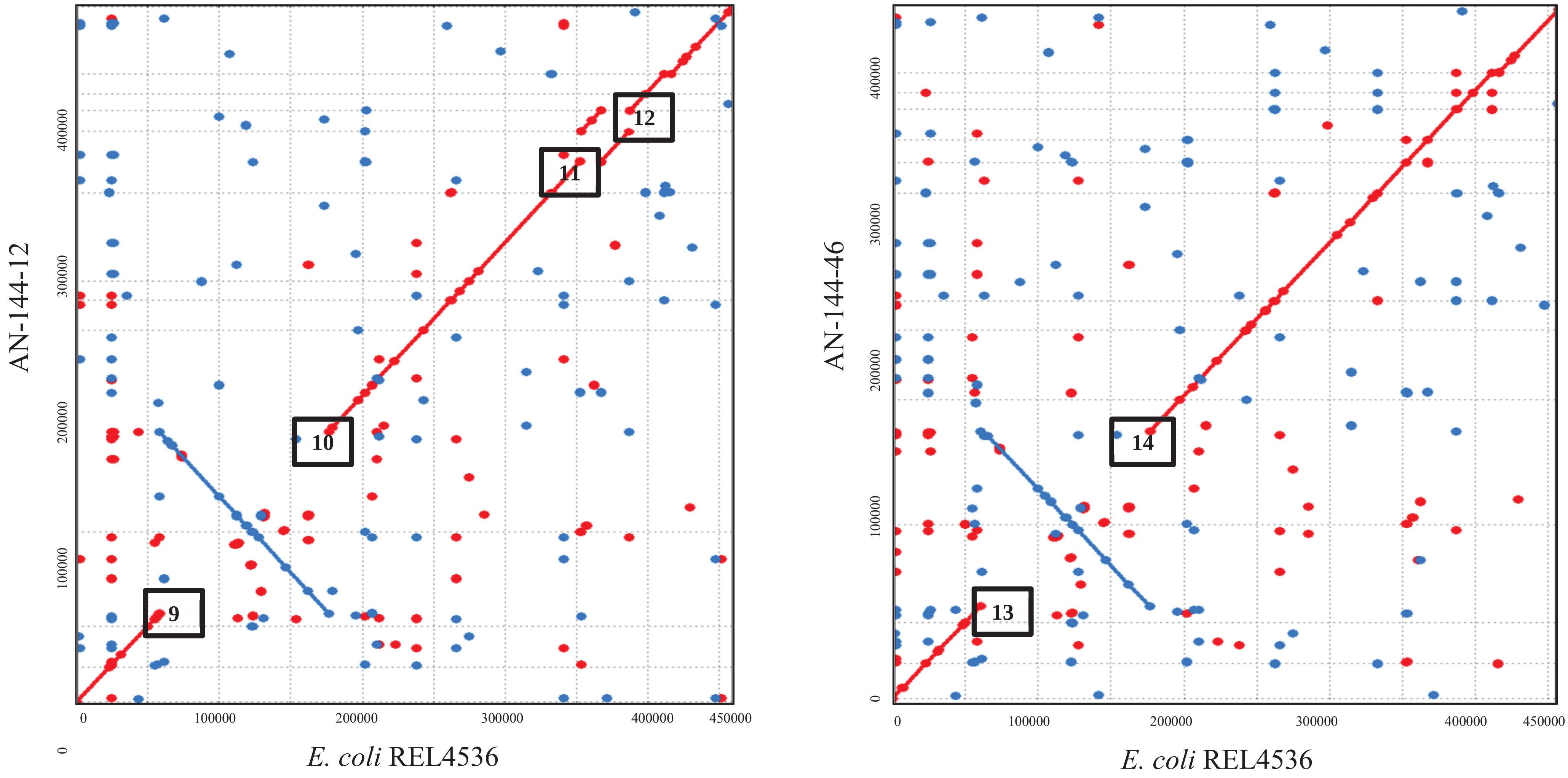

Figure 3.3. Synteny plots for anaerobic lineages. NUCmer alignments of the genomes of anaerobic clones against the ancestral *E. coli* REL4536 genome are displayed. Sequences that agree are displayed in coloured lines; red represents a forward match while blue represents a reverse match. Only the two anaerobic lineages with PCR verified large-scale GCRs are shown. Synteny breakpoints are labelled and more details about these breakpoints can be found in **Table 3.5**.

Table 3.5. Large-scale rearrangement break points identified *in silico* and verified by PCR.

Synteny breakpoint[*]	Lineage	REL4536 Reference Gene 1[†]	REL4536 Reference Gene 2[†]	PCR verified[‡]	Genomic repeat involved[§]
1	AE-180-04	*ybbP*	*ylbH*	Yes	*rhsD* gene
2	AE-180-04	*ybdK*	*entD*	Yes	IS*150* element
3	AE-180-04	*erfK*	*nac*	Yes	tRNA gene
4	AE-180-04	*nupC*	*yfeA*	Yes	-
5	AE-180-06	*clpX*	*lon*	Yes	1S*186* element
6	AE-180-06	*ybdK*	*entD*	Yes	IS*150* element
7	AE-180-24	*ynhG*	*ydhY*	Yes	IS*150* element
8	AE-180-24	*gltK*	*rihA*	Yes	IS*150* element
9	AN-144-12	*ybdK*	*entD*	Yes	IS*150* element
10	AN-144-12	*focA*	*pflA*	Yes	IS*150* element
11	AN-144-12	*nikR*	*yhhH*	No	*rhsB* gene
12	AN-144-12	*yibF*	*yibA*	No	*rhsA* gene
13	AN-144-46	*ybdB*	*ybdH*	Yes	-
14	AN-144-46	*focA*	*pflA*	Yes	IS*150* element

[*] Synteny breakpoints identified *in silico*. Numbers refer to labels shown in **Figure 3.2** and **Figure 3.3**.

[†] Genes spanning the corresponding synteny breakpoint in the reference *E. coli* REL4536 genome. Primers for PCR verification of breakpoints were targeted to these genes.

[‡] PCR was used to verify that the listed gene combinations, present in the reference genome, were not present in the evolved genome.

[§] Genomic repeats, if any, mediating the GCR

3.2.3.3 Summary of number of mutations found in lineages

A total of 124 and 158 mutations were found across the aerobic and anaerobic clones that were sequenced, respectively. A summary of the number of mutations found in each lineage is shown in **Table 3.6**. Details of all mutations found in this study are listed in **Table A.3.**

Table 3.6. Frequency of mutations found in each sequenced lineage.

Lineage	Number of mutations	Lineage	Number of mutations
AE-180-02	1	AN-144-02	0
AE-180-04	14	AN-144-04	11
AE-180-06	7	AN-144-06	16
AE-180-08	9	AN-144-08	8
AE-180-10	5	AN-144-10	11
AE-180-12	3	AN-144-12	7
AE-180-14	4	AN-144-14	8
AE-180-16	1	AN-144-16	6
AE-180-20	7	AN-144-20	4
AE-180-22	8	AN-144-22	3
AE-180-24	5	AN-144-24	12
AE-180-26	3	AN-144-26	3
AE-180-28	1	AN-144-28	2
AE-180-30	9	AN-144-30	5
AE-180-32	7	AN-144-32	3
AE-180-34	7	AN-144-34	4
AE-180-36	7	AN-144-36	4
AE-180-38	3	AN-144-38	2
AE-180-40	5	AN-144-40	15
AE-180-42	1	AN-144-42	1
AE-180-44	8	AN-144-44	8
AE-180-46	3	AN-144-46	14
AE-180-48	3	AN-144-48	2
AE-180-50	3	AN-144-50	9
Total Aerobic	*124*	*Total Anaerobic*	*158*

The mean number of mutations accumulated in aerobic lineages after 180 rounds of bottlenecking was 5.2 mutations ± 3.2 mutations per lineage. For the anaerobic lineages, the mean number of mutations after 144 rounds of bottlenecking was 6.6 mutations ± 4.6 mutations per lineage.

3.2.3.4 Estimation of mutation rates

Although more mutations were detected in anaerobically grown lineages than in aerobically grown lineages (**Table 3.6**), each set of lineages was not propagated for the same number of generations. Thus, the whole genome spontaneous mutation rate (section 1.6) for aerobically and anaerobically grown *E. coli* was calculated as described in section 2.2.14.3.

The spontaneous mutation rate of aerobically grown *E. coli* REL4536 is 2.50×10^{-10} mutations per nucleotide per generation (**Table 3.7**). For anaerobically grown *E. coli* REL4536, the mutation rate is almost double, at 4.14×10^{-10} mutations per nucleotide per generation (**Table 3.7**). Furthermore, the mutation rates of *E. coli* REL4536 grown in the two environments are significantly different (Mann-Whitney U = 196.0, $p = 0.029$). Individual mutation rates for each lineage varied greatly, ranging from 4.84×10^{-11} to 6.77×10^{-10} mutations per nucleotide per generation for aerobic clone genomes and from 0.00 to 1.01×10^{-9} mutations per nucleotide per generation for anaerobic clone genomes (**Table A.4**).

Table 3.7. Genome-wide spontaneous mutation rates for *E. coli* grown aerobically and anaerobically.

	Generations	Bottlenecks	Rate /genome /generation	SEM	Rate /nucleotide /generation	SEM
Aerobic	25	180	1.15×10^{-3}	1.46×10^{-4}	2.50×10^{-10}	3.18×10^{-11}
Anaerobic	24	144	1.90×10^{-3}	2.74×10^{-4}	4.14×10^{-10}	5.96×10^{-11}

The MA *in silico* simulations that were performed in section 3.2.2 used an aerobic mutation rate (197) that was obtained from Lenksi's LTEE (1.7.1.1). As populations in the LTEE are subjected to selection pressures, an average mutation rate of 4.6×10^{-4} mutations per genome per generation was estimated from only those mutations that were deemed to be neutral, synonymous substitutions. The mutation rate estimated from the LTEE is 2.5-fold lower than the rate estimated in this study, displaying the benefit of using whole genome sequencing and MA assays to comprehensively estimate mutation rates. In fact, Drake (2012) even argues that the mutation rate (136) calculated by Wielgoss et al. (2011) in this way is a product of selection acting on the codon usage of the populations rather than an actual estimate of the spontaneous mutation rate (159). It is important to note that selection could be a factor in this study as well. Ideally, for MA studies, a constant, effective population size of one should be maintained every generation (6). However, in this study, population sizes fluctuated as colonies grew to approximately 10^7 cells on agar plates between bottlenecking events, providing opportunities for accumulated mutations to be subject to selection. It is likely that to some degree beneficial mutations became fixed, while deleterious mutations were purged from the population (171). To determine if mutations were accumulated in a neutral manner in the MA study, and that the effects of selection were minimal, the ratio of BPSs found in protein coding regions verses non-coding regions or the ratio of non-synonymous to synonymous BPSs can be analysed. This will be discussed in further detail in section 4.2.1.1.3.

The genome-wide aerobic mutation rate obtained in this study is consistent with that determined for aerobically grown *E. coli* in a recently published MA study (13). However, Lee et al. (2012) did not take GCRs into account in their calculation of a mutation rate of 2.2×10^{-10} mutations per nucleotide per generation. If GCRs are excluded from our calculation of the mutation rate, a 1.4-fold lower rate of 1.81×10^{-10} mutations per nucleotide per generation is obtained for aerobic clone genomes. Perhaps the difference in rates between the two studies can be accounted for by the different *E. coli* strains and media used in the two studies. For their study, Lee et al. (2012) used *E. coli* MG1655 and their cultures were propagated on LB agar; a medium that is richer in carbon than the minimal DM agar used in this study (section 2.1.7.3). Furthermore, the aerobic MA lineages in their study experienced more cell divisions between bottlenecks, with 28 generations between passages (13), as opposed to the 25 generations between bottlenecks for aerobic lineages in our study; which provided a greater opportunity for mutations to occur. Generally, the difference in mutation rates between the two studies indicate the necessity of including GCRs in the estimation of spontaneous mutation rates to obtain more accurate and complete rates.

Moreover, the mutation rate calculated from anaerobic clone genomes in this study was 1.7-fold greater than the mutation rate calculated from aerobic clone genomes. Sakai et al. (2006), in their locus-based study, reported an anaerobic mutation rate that was approximately 2.4-fold higher than their aerobic mutation rate (28). Sakai et al. (2006) also used *E. coli* MG1655 and LB media for their study; factors that may perhaps account for the difference in anaerobic mutation rates between their study and our study. In this study, *E. coli* REL4536, a 10,000 generation descendent of *E. coli* REL606, was used to minimise adaptations to limited glucose media. Thus, the different mutation rates observed in the aerobically and anaerobically grown cells of this study should theoretically be in response to the different environmental conditions.

3.2.3.4.1 Mutation accumulation assays versus fluctuation assays

As it is, the mutation rates obtained from the MA assays were within the range of rates obtained from the fluctuation assays. For aerobic *E. coli* REL4536, the whole genome rate estimated from MA assays was 2.50×10^{-10} mutations per nucleotide per generation (**Table 3.7**) while the rates estimated from fluctuation assays were 4.0×10^{-11} mutations per nucleotide per generation and 4.8×10^{-10} mutations per nucleotide per generation for the Nal^{R} and Ara^{+} phenotypes, respectively (**Table 3.1**). Potential reasons for the variation amongst the loci-based mutation rates have already been discussed (section 3.2.1). Together, an average rate of 2.6×10^{-10} mutations per nucleotide per generation was obtained from fluctuation assays of aerobically grown *E. coli* REL4536. The genome-wide rate was almost six-fold greater than the rate estimated for the Nal^{R} locus but two-fold lower than the rate estimated for the Ara^{+} locus. However, the average aerobic mutation rate obtained from fluctuation assays is similar to the genome-wide rate estimated from the MA assays, suggesting that using multiple loci for fluctuation assays is preferable.

A similar observation was made for *E. coli* REL4536 mutation rates under anaerobic conditions. The whole genome rate estimated from MA assays was 4.14×10^{-10} mutations per nucleotide per generation (**Table 3.7**) while the rates estimated from fluctuation assays were 2.20×10^{-10} mutations per nucleotide per generation and 2.44×10^{-9} mutations per nucleotide per generation for the Nal^{R} and Ara^{+} phenotypes, respectively (**Table 3.1**). Together, an average rate of 1.33×10^{-9} mutations per nucleotide per generation was obtained for fluctuation assays of anaerobically grown

E. coli REL4536. The genome-wide rate was roughly two-fold greater than the rate estimated for the Nal^R locus but six-fold lower than the rate estimated for the Ara$^+$ locus. Also, the average anaerobic mutation rate obtained from fluctuation assays was nearly three-fold greater than the genome-wide rate estimated from the MA assays, once again reinforcing the benefit of using multiple loci for fluctuation assays.

3.2.3.4.2 Mutation accumulation assays versus *in silico* mutations

The MA simulations that were performed before starting the MA lineages were based on mutation rates that were approximately two-fold lower than the rates that were eventually estimated from the MA study. To determine whether our developed MA simulator programme was sufficient at predicting mutations, more simulations were run *in silico.* Using parameters and mutation rates determined from this study (**Table 3.7**), 1,000 simulations were run on the MA simulator programme (section3.2.2). The *in silico* analysis indicated that the mean number of mutations expected in any aerobic lineage after 180 rounds of bottlenecking was four to five mutations while the mean number of mutations expected in any anaerobic lineage after 144 rounds of bottlenecking was six to seven mutations (**Figure 3.4**). As can be seen in **Figure 3.4,** the observed distribution of mutations across lineages in both environments differed from theoretical predictions obtained from the MA simulator programme. It is important to note that the MA simulator programme does not consider the fitness effects of mutations, and as such, selection against highly deleterious mutations is not accounted for. In addition, the observed mutation distributions were obtained from a relatively smaller sample size (24 lineages in each environment) as opposed to the theoretical distributions, which were based on a 1,000 simulations. As such, these factors may themselves account for any differences in the observed and theoretical distribution of mutations. Nevertheless, as described in section 3.2.3.3, the mean number of mutations accumulated in aerobic and anaerobic lineages in the MA study were 5.2 mutations and 6.6 mutations, respectively. Thus, the programme, designed as part of this thesis, was suitable for its purpose of predicting the number of mutations that would accumulate in a MA study and so can be used as a useful tool in establishing future MA studies.

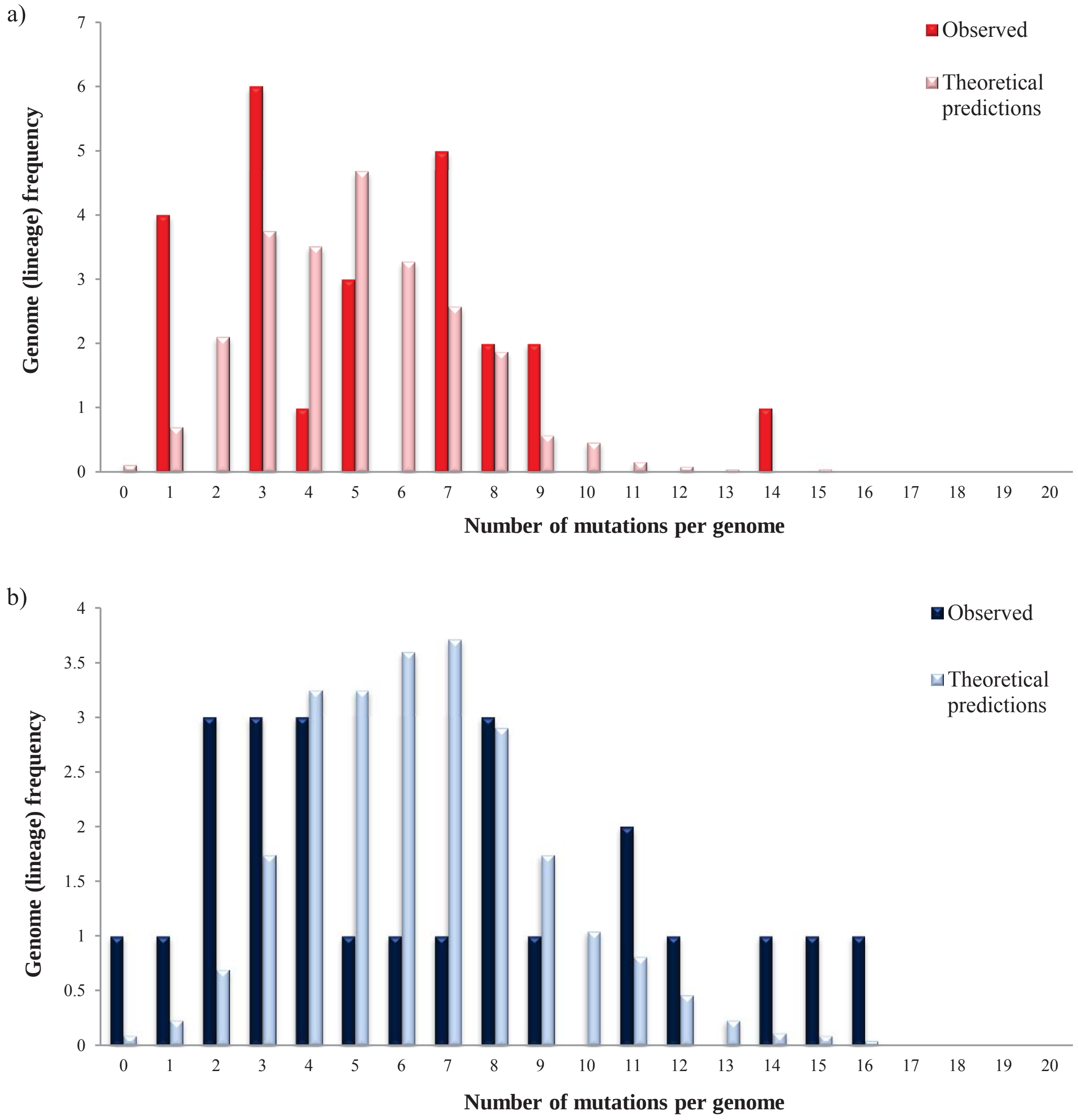

Figure 3.4. Comparison of MA mutations and *in silico* theoretical mutations. Distribution of mutations detected among a) aerobic and b) anaerobic MA lineages compared to theoretical predictions made using the MA simulator programme. 1,000 simulations were run to determine the number of mutations expected in a) aerobically grown lineages after 180 bottlenecks with a mutation rate of 1.15×10^{-3} per genome per generation and with bottlenecks every 25 generations and b) anaerobically grown lineages after 144 bottlenecks with a mutation rate of 1.90×10^{-3} per genome per generation and with bottlenecks every 24 generations.

3.3 Summary

The aim of this study was to determine the spontaneous mutation rate of *E. coli* grown in an aerobic and anaerobic environment. It had previously been reported that the anaerobic mutation rate was twice as high as the aerobic rate (28), but the study was based on just one locus of the genome. Loci-based estimates of the mutation rate tend to be limited and not necessarily applicable across the whole genome. In this study, spontaneous mutation rates were measured in two ways: by fluctuation assays and MA assays. For MA assays, whole genome re-sequencing was used to identify mutations across the entire genome.

Using both fluctuation and MA assays, it was determined that for *E. coli* REL4536, the genome-wide spontaneous mutation rate is greater in an anaerobic environment than in an aerobic environment. Our genome-wide estimates of the spontaneous mutation rate of *E. coli* REL4536 are 2.50×10^{-10} mutations per nucleotide per generation and 4.14×10^{-10} mutations per nucleotide per generation for aerobically and anaerobically grown cells respectively (**Table 3.7**). The anaerobic mutation rate is significantly 1.7-fold greater than the aerobic mutation rate and so, it is possible that particular classes of mutations are more prevalent under anaerobic growth conditions than under aerobic growth conditions. This will be investigated in the following chapter, regarding mutation spectra.

Chapter Four: Mutation spectra in aerobic and anaerobic environments

4.1 Introduction

In the previous chapter (section 3.2.3.4), the genome-wide spontaneous mutation rate for *E. coli*, determined *via* MA lineages, was shown to be greater in the anaerobic environment (4.14×10^{-10} mutations per nucleotide per generation) than in the aerobic environment (2.50×10^{-10} mutations per nucleotide per generation). To determine if there were differences in the types of mutations that accumulated in the two environments, the mutation spectra of the sequenced MA lineages was analysed. Broadly, the different classes of mutations that were studied were BPSs, indels and GCRs. The spontaneous mutation rate is the net result of mutational pressures experienced by the DNA, along with the fidelity of DNA replication and the efficiency of the various DNA repair systems. Different stages of the bacterial cell cycle experience different mutational pressures and thus, it is difficult to distinguish the relative contribution of each factor to the observed mutation rate. Therefore, to obtain a better understanding of mutation spectra, where differences in mutation spectra were present, the expressions of genes involved in maintaining genome integrity were studied to determine whether associated differences in their expression were observed. Thus, the aims of this chapter were to determine:

- What is the rate with which each mutation type and class occur in *E. coli* grown in aerobic and anaerobic environments?
- What types of mutations predominate in *E. coli* grown in aerobic and anaerobic environments?
- What DNA repair pathways are active in *E. coli* grown in aerobic and anaerobic environments?

4.2 Results and discussion

4.2.1 Spectra of mutations in aerobic and anaerobically grown *E. coli*

Among the 48 MA lineages that had been genome re-sequenced (section 2.2.16.3), a total of 282 mutations were detected, with 124 and 158 mutations accumulated in the aerobically and anaerobically grown cells, respectively. In this section, these mutations are analysed in greater detail to determine whether there are differences between the types and genomic distributions of mutations that accumulated in the two environments.

In this study, across all sequenced MA lineages, BPSs were generally the most frequently observed type of mutation, accounting for 52.1% of total mutations (**Figure 4.1a**). GCRs constituted 34.8% of all mutations while indels were less frequently observed, accounting for 13.1% of all mutations (**Figure 4.1a**). In general, mutation rates per genome per generation for BPS, indels and GCRs were greater in anaerobically grown cells than in aerobically grown cells (**Figure 4.1a**), and for GCRs, this difference was significant (Mann-Whitney U = 146.5, *p* = 0.001).

In this study, the length of time between generations for aerobically and anaerobically grown lineages differed (section 3.2.3.4). It is possible that the longer period of time between generations has enabled more mutation events to take place during growth in the anaerobic environment, contributing to the higher (per generation) rate in this environment. This would be particularly so if mutations arise independently of the genome replication phase of the cell division cycle. To investigate this further, the rate of accumulation of mutations under aerobic and anaerobic growth conditions with regard to time (mutation rates per genome per day) were calculated (**Figure 4.1b**). In contrast to the mutation rates (per generation), the rates (per day) for BPS, indels and GCRs were greater for aerobically grown cells than for anaerobically grown cells (**Figure 4.1b**), with the differences for BPS and indels being significant. It is possible that this difference in mutation rates can be attributed to the aerobically and anaerobically grown cells spending differing amounts of time in different stages of the cell cycle.

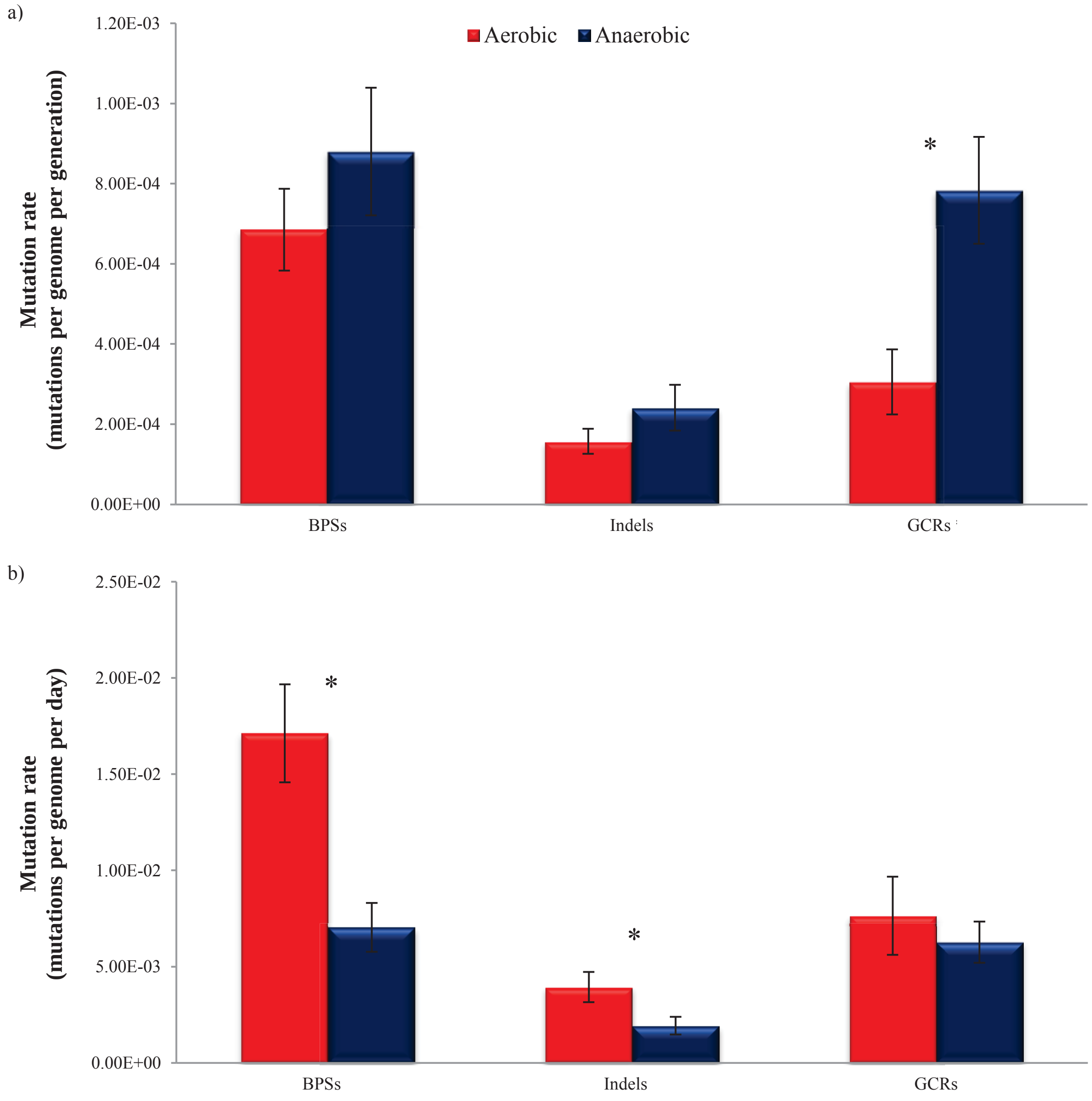

Figure 4.1. Mutation rates for different types of mutations observed in the MA study. Shown are a) mean mutation rates per genome per generation and b) mean mutation rates per genome per day of growth. Error bars represent standard error of the mean. Asterisk denotes a significant difference between the aerobic and anaerobic mutation rates ($p < 0.05$).

The bacterial cell cycle is comprised of three stages: the B period which is from the initiation of cell division to the initiation of DNA replication, the C period which is when DNA replication takes place, and finally, the D period which is from the termination of DNA replication to the completion of cell division (244). A general trend in *E. coli* has been reported, where faster growing cells spend a greater proportion of time in the C period and proportionately less time in the B period than slower growing cells, which spend a greater proportion of time in the B period and proportionately less time in the C period (245). As the aerobically grown cells in this study went through roughly 25 generations within 24 h of growth, the average time between cell division events was approximately every 60 min. Therefore, it is assumed that the aerobically grown cells spent a relatively larger proportion of time in the C period than their slower growing anaerobic counterparts. By comparison, the anaerobically grown cells went through almost 24 generations within 72 h of growth, giving an average time of 180 min per cell division. During replication, DNA is particularly susceptible to damage as it transiently exists as ssDNA; a form of DNA that has been shown to be more vulnerable to mutation (246). Thus, as aerobically grown cells spent more time in DNA replication, this is consistent with the observation that mutation rates per unit of time (as compared to per generation) were greater for aerobically grown cells than for anaerobically grown ones.

In this chapter, the mutation rates for different types of mutations for aerobically and anaerobically grown cells have been determined. However, each mutation type (e.g. BPSs, indels and GCRs, and classes therein) likely arises *via* different mutagenic agents and cellular mechanisms. For example, ROS damage to DNA will elicit different cellular responses for repair than general errors in DNA replication/repair. Furthermore, differences in cell physiology in each environment may induce different suites of mutations. Thus, the cell cycle stage that different mutation types are more likely to occur, and relative proportion of time spent in that stage of the cycle, may have significant bearing on the overall mutation rates between the two environments. Therefore, throughout this chapter, consistent with previous studies (11-13), mutation rates expressed per generation are reported. However, for comparison purposes, mutation rates per unit time (calculated as described in **Equation 2.6**) are also reported as they may provide deeper insight into the mechanisms behind differences in mutation spectra between the two environments.

4.2.1.1 BPSs

A total of 147 BPSs were detected in this study; with 74 and 73 BPSs accumulated in aerobically and anaerobically grown cells respectively. The anaerobic BPS per generation rate was 1.3-fold greater than the aerobic per generation rate (**Figure 4.1a**), though this difference was not significant (Mann-Whitney U = 254.0, *p* = 0.244). However, when considering mutation rates per genome per day, the aerobic BPS mutation rate was 2.4-fold greater than the anaerobic rate (Mann-Whitney U = 137.0, $p < 0.001$) (**Figure 4.1b**).

Recently, Lee et al. (2012) estimated a BPS mutation rate of 9.23×10^{-4} mutations per generation for aerobically grown *E. coli* (13). The value reported by Lee et al. (2012), though 1.4-fold greater, is of the same order of magnitude as the rate of 6.85×10^{-4} mutations per generation that was observed for the aerobically grown cells of this study. As Sakai et al. (2006) found BPSs to be 6.4-fold less frequent in anaerobic cells than in aerobic cells (28), the relatively higher anaerobic genome-wide BPS rate of 8.80×10^{-4} mutations per generation observed in this study was not expected. In their study, Sakai et al. (2006) used an experimental system that was based on two copies of the *rpsL* locus of the genome (28) while this study was based on the whole genome. These differences in mutation rate measurement can sufficiently explain the varying anaerobic BPS per generation mutation rates between the two studies. In addition, the two studies also utilised different *E. coli* strains and media and it is possible that these two factors also contribute to the observation of differing mutation rates; Sakai et al. (2006) used an *E. coli* MG1655 strain and LB media (section 2.1.7.1) for their study (28) while this study used *E. coli* REL4536 strains and DM media (section 2.1.7.2).

4.2.1.1.1 Nucleotide mutation rates

Single base substitutions can be classified into 12 possible substitution types. The rates for each of these were calculated for cells grown aerobically and anaerobically, with mutation rates normalized to account for the nucleotide content of the genome and the predominant mutations in both environments were G → A or C → T transitions (**Figure 4.2**). As spontaneous deamination of cytosine nucleotide bases into uracil bases, or of 5-methylcytosine nucleotide bases into thymine bases, are frequently occurring DNA lesions known to cause G → A or C → T transitions respectively, these results were expected (13, 19, 247). As can be seen in **Figure 4.2b**, G → A and C → T transitions appear to the predominant types of BPSs in aerobically grown cells per unit time. As the aerobically grown cells of this study spent more time in DNA replication, it stands to reason that the DNA of aerobically grown cells spent more time in a single-stranded state. Thus, these findings are in agreement with previous observations that ssDNA is particularly susceptible to cytosine deamination (248). On the other hand, A → G and T → C transitions, possibly caused by the spontaneous deamination of adenine nucleotide bases into hypoxanthine or by the tautomerization of thymine bases (13, 19, 247), respectively, were less common in both environments (**Figure 4.2**). Overall, these findings are consistent with mutational data from previous studies reporting a spontaneous mutation bias towards increasing the A:T content of bacterial genomes (29, 249, 250). In short, these studies imply that the nucleotide content of bacteria are not as affected by mutational biases as previously thought but rather, selection pressures that the bacteria may have been subject to over time are more likely causes for any observed variations in nucleotide content.

ROS generated during aerobic respiration, are also known to induce specific mutation types. For instance, the GO lesion is a typical DNA lesion caused by the oxidation of guanine nucleotide bases to 8-oxo-guanine nucleotide bases, and can ultimately lead to G → T and C → A transversions. In keeping with this expectation, G → T transversion mutation rates per generation were 5.4-fold greater (**Figure 4.2a**) in aerobically grown cells, as compared to anaerobically grown cells (Mann-Whitney U = 230.0, p = 0.042). Surprisingly, C → A transversion rates per generation were approximately eight-fold greater (**Figure 4.2a**) in anaerobically grown cells, as compared to aerobically grown cells (Mann-Whitney U = 225.0, p = 0.011). A similar pattern of G → T versus C → A asymmetry was also observed in mutation rates calculated per day (**Figure 4.2b**). These results were most unexpected and suggest the presence of a strand bias in the types of BPSs that arose under aerobic and anaerobic conditions and will be investigated further in section 4.2.2.1.2.

More curious, however, was the presence of G → T and C → A transversions in anaerobic lineages. These findings suggest that there may be some oxidative damage occurring during growth under the anaerobic conditions in this study and that the repair systems that are normally induced in response to this specific kind of damage are not being efficiently induced. While this may imply that the growth conditions in the anaerobic conditions are not truly anaerobic, Sakai et al. (2006) have also observed G → T and C → A transversions in the anaerobic cells of their study (28). In fact, they reported almost equivalent rates of G → T and C → A transversions in aerobically and anaerobically grown cells (28). While they did not give much thought to these mutations, they concluded that their anaerobically grown cells were not experiencing any oxidative damage as evidenced by the lack of overall G → T and C → A transversions in mutator cells (28). While ROS can also be produced as a byproduct of certain biological mechanisms, such as part of the inflammatory response or intracellular signalling (19), there is a strong likelihood that G → T and C → A transversions are spontaneously arising under anaerobic conditions, but not because of exposure to ROS. While no specific mechanism underlying this BPS spectrum is apparent, it may be associated with pH, as acids are generated as fermentation end-products during growth under anaerobic conditions. Alternatively, the slower growth rate, and the resultant physiological condition of the cells under anaerobic growth may have created specific mutagenic pressures resulting in the observed BPS spectrum.

Mutation rates per generation for T → G transversions (Mann-Whitney U = 217.5, $p = 0.032$) and A → C transversions (Mann-Whitney U = 223.0, $p = 0.021$) were at least two-fold greater in anaerobically grown *E. coli* as compared to aerobically grown *E. coli* (**Figure 4.2a**). Per day, T → G and A → C transversions occurred at rates that were not statistically significantly different between aerobically and anaerobically grown *E. coli* (**Figure 4.2b**). Therefore, it is possible that T → G and A → C transversion mutations occur independently of DNA replication. As oxidation of the guanine nucleotide can result in T → G or A → C transversions (19), these results were unexpected. Maybe these spontaneously arising transversion mutations are not being repaired efficiently under the anaerobic conditions of this study, allowing for their accumulation in anaerobic lineages. Or maybe, perhaps there is not enough oxidative stress under the aerobic growth conditions of this study, accounting for the fewer oxidative damage related mutations observed in aerobically grown cells. Or alternatively, these results raise the possibility of there being causative agents, other than ROS, that lead to T → G or A → C transversions, especially for anaerobically grown cells. To obtain a better understanding of these results, the expression values of various genes involved in repairing DNA damage under aerobic and anaerobic conditions were analysed and will be discussed in section 4.2.3.1.1.

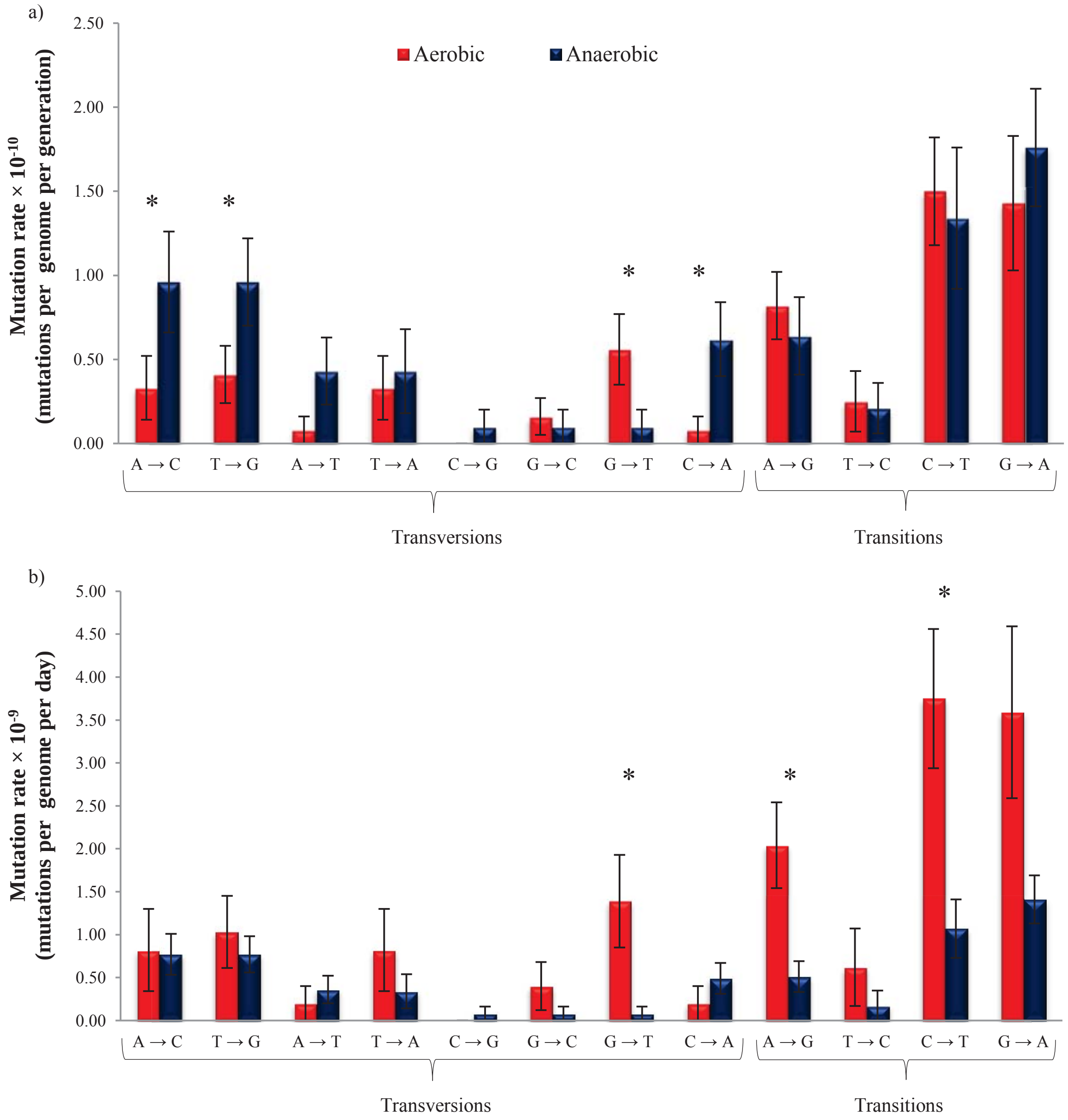

Figure 4.2. Mutation rates of single base substitutions in aerobically and anaerobically grown *E. coli*. Shown are a) mean mutation rates per genome per generation and b) mean mutation rates per genome per day of growth. Error bars represent standard error of the mean. Asterisk denotes a significant difference between the aerobic and anaerobic mutation rates ($p < 0.05$).

4.2.1.1.2 Transition bias

BPSs can be described as either being transversions or transitions. As there are four possible types of transitions and eight possible types of transversions (**Figure 1.1**), the expected ratio of transitions to transversions (Ts/Tv), or transition bias as it is more commonly known, is expected to be 0.5 if nucleotide substitutions arise completely at random (29). However, for many organisms, including *E. coli*, the Ts/Tv ratio is usually greater than 0.5; as transition mutations are more common than transversion mutations

(29). This bias is thought to be due to either the biochemical structure of the nucleotide bases themselves, or due to transitions being less likely to be repaired than transversions (30). As transitions involve the substitution of a purine or pyrimidine nucleotide with another purine or pyrimidine nucleotide, respectively, transitions are hypothesized to be easier mutations to occur as the fundamental biochemical structure and properties of the nucleotide being substituted are not changed. As transversions greatly alter the chemical structure of the DNA, the second hypothesis is that transversions are more readily recognised and repaired by cellular repair systems, allowing for the fixation of transitions in the genome.

A transition bias was also observed in this study as, across all MA lineages, more transitions than transversions were detected. For aerobically grown cells, 50 BPSs were classified as transitions and 24 were classified as transversions, resulting in a Ts/Tv value of 2.08. For anaerobically grown cells, 38 BPSs were classified as transitions and 35 were classified as transversions, resulting in a Ts/Tv ratio of 1.09. Lee at el. (2012) observed a Ts/Tv ratio of 1.28 (13) for aerobically grown *E. coli*, however, the ratio that was observed in this study for aerobically grown cells was even greater than expected. Wielgoss et al. (2011) calculated an aerobic Ts/Tv ratio of 1.99 (136) from *E. coli* populations that had been evolved for over 40,000 generations, indicating that the Ts/Tv ratio may vary according to experimental growth conditions, media and *E. coli* strains.

Per generation, transitions occurred at rates (**Figure 4.3a)** that were not statistically significantly different between the aerobically and anaerobically grown cells (Mann-Whitney U = 287.0, p = 0.494). In contrast, the anaerobic transversions per generation rate was significantly 1.7-fold greater (**Figure 4.3a**) than the aerobic per generation rate (Mann-Whitney U = 200.0, p = 0.031). However, when considering mutation rates per genome per day (**Figure 4.3b**), rates of transitions and transversions were 3.2-fold (Mann-Whitney U = 99.0, p < 0.001) and 1.7-fold (Mann-Whitney U = 252.0, p = 0.225) greater, respectively, for aerobically grown cells, as compared to anaerobically grown cells.

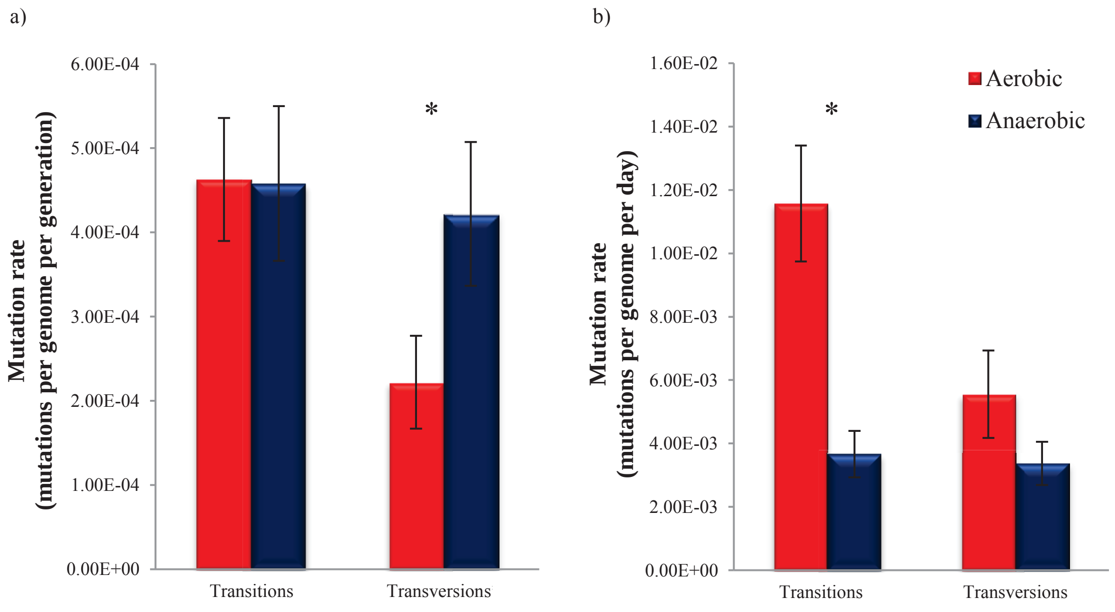

Figure 4.3. Mutation rates of transitions and transversions in aerobically and anaerobically grown *E. coli*. Shown are a) mean mutation rates per genome per generation and b) mean mutation rates per genome per day of growth. Error bars represent standard error of the mean. Asterisk denotes a significant difference between the aerobic and anaerobic mutation rates ($p < 0.05$).

Transitions and transversions occurred at similar rates in anaerobically grown cells, a phenomenon not observed for the aerobically grown cells (**Figure 4.3**). As can be seen from **Figure 4.2**, transversions, particularly T → G and A → C mutations, were frequently observed in anaerobically grown cells. While certain transversions, like G → T mutations, are known to be readily repaired in aerobically grown *E. coli* cells (19), it is possible that other transversions are less readily repaired in anaerobically grown cells, resulting in a higher mutation rate. To obtain a better understanding of these results, the expression values of various genes involved in repairing DNA damage under aerobic and anaerobic conditions were analysed and will be discussed in section 4.2.3.1.1.

4.2.1.1.3 Synonymous vs non-synonymous BPSs

DNA regions that code for proteins are referred to as protein coding regions, while regions that do not encode proteins are considered non-coding regions. Non-coding DNA includes sequences that code for mobile genetic elements, pseudogenes, tRNA and rRNA and all intergenic sequences. Across all MA lineages, 110 BPSs were found in coding regions while only 37 BPSs were detected in non-coding regions of the genome. As approximately 89% of the *E. coli* REL4536 genome is comprised of protein coding DNA, it was surprising that only 75% of all BPSs occurred in coding regions. In the REL4536 genome, protein-coding sequences are comprised of 4.08×10^6 nucleotides while the non-coding regions are comprised of 5.20×10^5 nucleotides, resulting in a coding verses non-coding ratio of 7.83. In our study, we observed mutations in coding verses non-coding regions at a ratio of 2.97 (110/37), which was significantly less than the expected ratio of 7.83 ($\chi^2 = 26.6$, df = 1, $p < 0.001$). The time colonies spend growing on agar plates allows selection to occur and as such, highly deleterious mutations will not accumulate. Thus, selection was potentially occurring in the MA lineages. Furthermore, it is possible that BPSs may have been selected for in non-coding regions, and selected against in coding regions of the genomes.

For aerobically grown genomes, 60 BPSs occurred in coding regions while only 14 BPSs were detected in non-coding regions of the genome. For anaerobically grown lineages, 50 BPSs occurred in coding regions while 23 BPSs were detected in non-coding regions of the genome. Of the 37 mutations found in non-coding regions, six occurred in tRNA sequences, one occurred in a 16S rRNA sequence, two occurred in sequences encoding MGEs and another two occurred in ncRNA sequences. The non-coding BPS in AN-144-06 occurred in the lysine riboswitch, a ncRNA that serves as a sensor for the amino acid lysine (251), while the BPS in AN-144-10 occurred in a small RNA that binds the protein Hfq.

Per generation, the rates at which mutations occurred in non-coding regions (Mann-Whitney U = 252.0, $p = 0.206$) and coding regions (Mann-Whitney U = 272.0, $p = 0.373$) were not significantly different between aerobically and anaerobically grown cells, even though the rates were greater under anaerobic conditions (**Figure 4.4a**). When considering mutation rates per genome per day (**Figure 4.4b**), rates of mutations in coding regions and non-coding regions were three-fold (Mann-Whitney U = 144.0, $p = 0.001$) and 1.5-fold (Mann-Whitney U = 263.0, $p = 0.284$) greater, respectively, for aerobically grown cells, as compared to anaerobically grown cells.

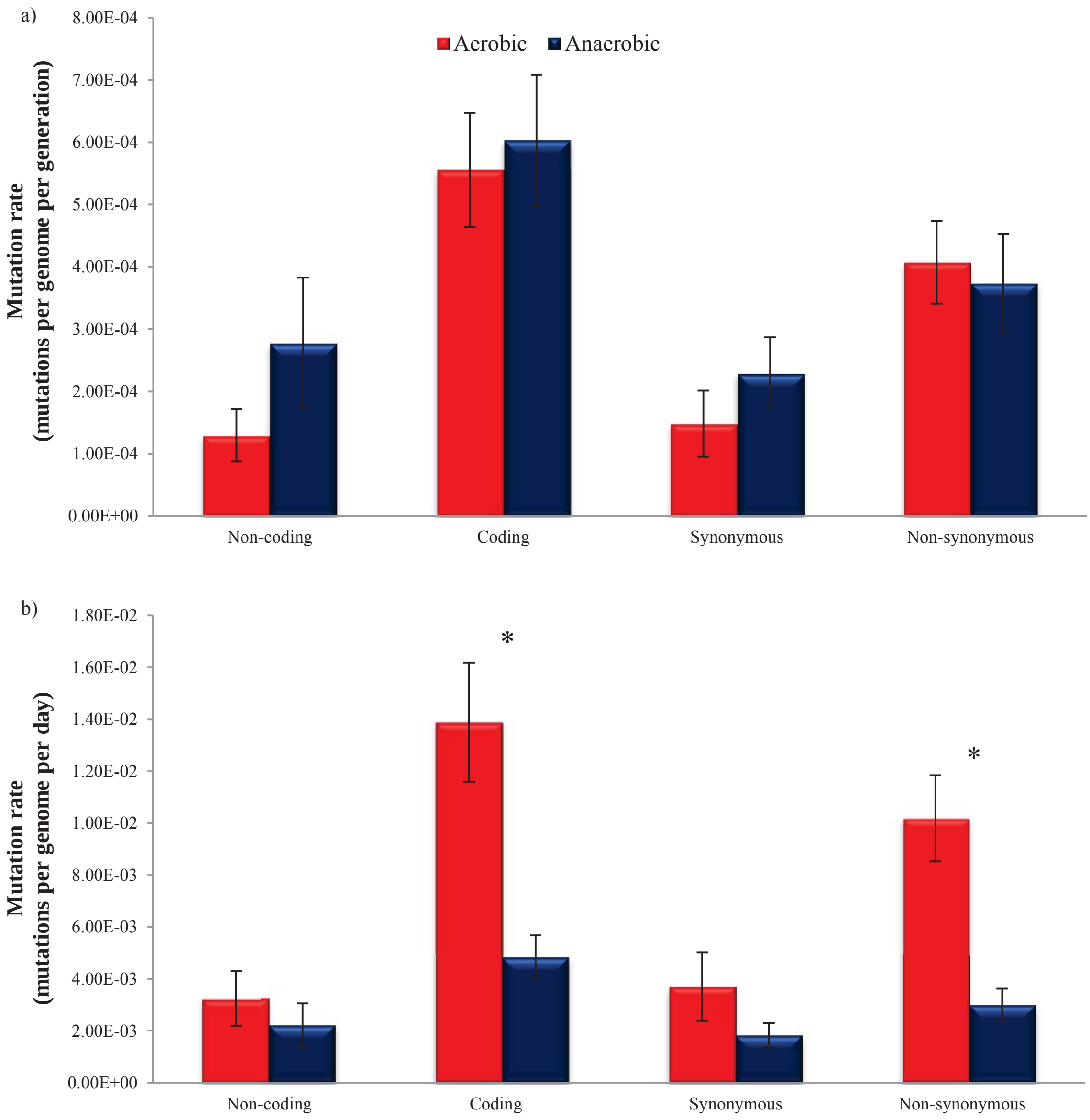

Figure 4.4. Rates of coding, non-coding, synonymous and non-synonymous mutations in aerobically and anaerobically grown *E. coli*. Shown are a) mean mutation rates per genome per generation and b) mean mutation rates per genome per day of growth. Error bars represent standard error of the mean. Asterisk denotes a significant difference between the aerobic and anaerobic mutation rates ($p < 0.05$).

Mutations in coding regions can be further categorised as being either synonymous (silent) or non-synonymous mutations depending on whether they result in a change of protein sequence. Across all MA lineages, 35 BPSs were synonymous while 75 BPSs were non-synonymous mutations. Only six of the non-synonymous mutations generated a non-sense change, with only one of those mutations occurring in anaerobically grown cells. Based on the codon usage (section 2.2.16.3.4) of REL4536 (**Table A.5**), an expected ratio of non-synonymous to synonymous mutations of 3.16 was estimated. In this study, a ratio of 2.14 was observed (75/35), which was not significantly different from expected ($\chi2 = 3.6$, df = 1, $p = 0.058$). Of the 35 synonymous mutations, based on the codon usage of the REL4536 genome, four mutations were neutral, 19 mutations resulted in a less commonly used codon while 12 mutations resulted in a more commonly used codon. Thus, there appears to have been little bias against synonymous and non-synonymous mutations in this study and as such, it can be concluded that selection was minimal in this MA study.

Lee et al. (2012) reported approximately 2.3-fold more non-synonymous than synonymous mutations in their aerobically grown *E. coli* (13). For the aerobically grown cells of this study, non-synonymous mutations were 2.8-fold more frequent, per generation, than synonymous mutations. For cells grown under anaerobic conditions in this study, however, non-synonymous mutations were only 1.6-fold more frequent per generation than synonymous mutations (**Figure 4.4a**). Therefore, in this study, the per generation rate of synonymous mutations was 1.5-fold greater in anaerobically grown cells, as compared to aerobically grown cells (Mann-Whitney U = 211.0, $p = 0.057$). Meanwhile, the per generation rate of non-synonymous mutations was not statistically significantly different between the two environments (Mann-Whitney U = 272.0, $p = 0.628$). When considering mutation rates per day (**Figure 4.4b**), rates of synonymous and non-synonymous mutations were two-fold (Mann-Whitney U = 267.0, $p = 0.316$) and 3.4-fold (Mann-Whitney U = 128.0, $p < 0.001$) greater, respectively, in aerobically grown cells, as compared to anaerobically grown cells.

4.2.1.2 Indels

For this study, indels were classified as insertions (of 1 bp or greater), deletions (of 1 bp or greater) or slippage events (section 1.3.2). A total of 37 indels were detected in this study, with 26 occurring in intergenic regions of the genome. The anaerobic indel per generation mutation rate of was 1.5-fold greater (**Figure 4.1a**) than the aerobic per generation rate, though this difference was not significant (Mann-Whitney U = 243.0, $p = 0.170$). When considering mutation rates per genome per day, the aerobic indel mutation rate was significantly two-fold greater (**Figure 4.1b**) than the anaerobic rate (Mann-Whitney U = 191.0, $p = 0.016$).

Lee et al. (2012) limited their analysis of indels to changes involving a maximum of four nucleotides and reported a rate of 8.34×10^{-5} indels per generation for their aerobically grown cells (13). In this study, an almost two-fold higher rate of 1.57×10^{-4} indels per generation was reported for the aerobically grown cells (**Figure 4.1a**). Even when the analysis was restricted to changes involving a maximum of four nucleotides, the mutation rate for the aerobically grown cells of this study, at 1.02×10^{-4} indels per generation, was 1.2-fold greater than the mutation rate reported by Lee et al. (2012). Possible reasons for the differences in mutation rates between the two studies have been discussed earlier in section 3.2.3.4. In addition, the restricted anaerobic indel per generation mutation rate was still greater (~1.7-fold) than the restricted aerobic rate, though this difference in rates was not significant (Mann-Whitney U = 245.0, $p = 0.157$). Sakai et al. (2006) only investigated single-base frameshifts in their study but did not detect any notable differences in the rates of indels between their aerobically and anaerobically grown cells (28).

In this study, slippage events were frequently observed while insertions were the least frequently observed (**Figure 4.5**). Per generation, the rate of insertions was 1.5-fold greater in aerobically grown cells, while the rate of deletions was 1.5-fold greater in anaerobically grown cells (**Figure 4.5a**). However, neither of these differences were significant ($p > 0.05$) by Mann-Whitney U-test. Per day mutation rates of insertions and deletions (**Figure 4.5b**) were 4.8- and 2.1- fold greater, respectively, in aerobically grown cells, as compared to anaerobically grown cells, though the differences in rates between the two environments were not significant by Mann-Whitney U-test ($p > 0.05$). Per generation mutation rates for slippage events were 1.8-fold greater (**Figure 4.5a**) in anaerobically grown cells, as compared to the aerobically grown cells (Mann-Whitney U = 250.0, $p = 0.055$). However, as expected, per day mutation rates of slippage events (**Figure 4.5b**) were greater in aerobically grown cells, as compared to anaerobically grown cells (Mann-Whitney U = 264.0, $p = 0.279$).

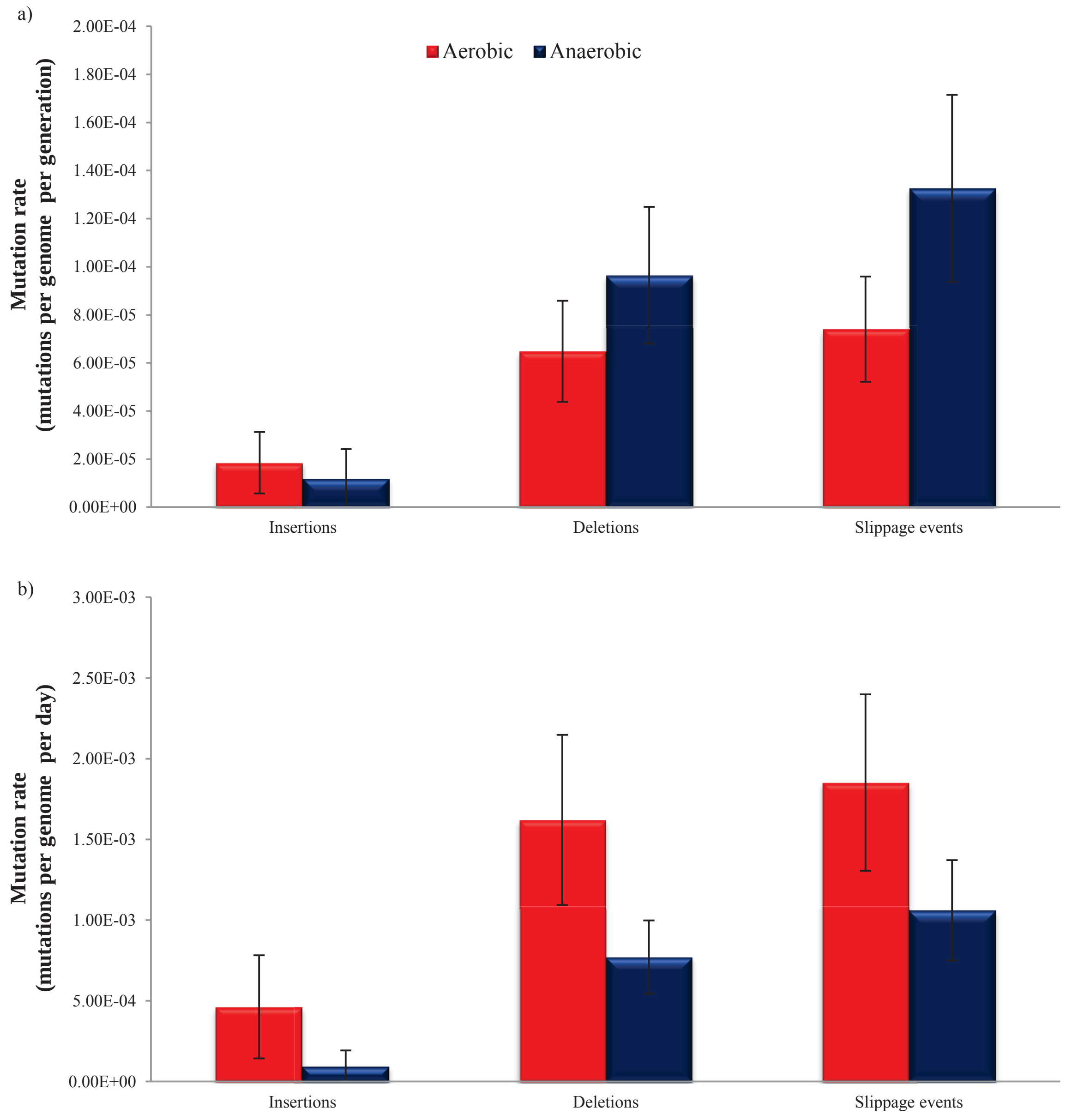

Figure 4.5. Mutation rates of indels in aerobically and anaerobically grown *E. coli*. Shown are a) mean mutation rates per genome per generation and b) mean mutation rates per genome per day of growth. Error bars represent standard error of the mean.

Slippage events are caused by short repetitive DNA sequences and in this study, all detected slippage events involved short mononucleotide repeat sequences (**Table 4.1**). Studies have shown that slippage events at mononucleotide sequences require at least two bases (252, 253) and approximately 64% of all slippage events observed in this study involved mononucleotide repeats of two bases. Furthermore, eight of the 19 slippage events were found at an intergenic, guanine mononucleotide sequence of two bases (at positions 3,866,357 to 3,866,358 in the REL4536 reference genome), suggesting that this region may be a putative slippage hotspot (section 4.2.2.2). While the majority of slippage events observed in this study occurred in intergenic regions of the genome, five slippage events were also detected in coding regions (**Table 4.1**). It is predicted that such events would lead to the disruption of gene function. However, recent reports show that genes disrupted by slippage events may still yield functional products as slippage during transcription may restore the original coding sequence (254, 255).

Table 4.1. Slippage events detected in aerobic and anaerobic MA lineages.

Lineage	Reference Position*	Mutation	5' flanking gene†	3' flanking gene†
AN-144-14	668,909	G(4) → G(3)	*yegM*	*yegL*
AE-180-30	1,043,563	G(6) → G(5)	*tppB*	*nth*
AN-144-30	1,438,008	A(5) → A(4)	*adhE*	-
AE-180-38	1,504,657	A(8) → A(9)	*umuD*	*ycgN*
AN-144-50	1,504,657	A(8) → A(9)	*umuD*	*ycgN*
AN-144-46	1,554,506	G(2) → G(3)	*ycfP*	-
AN-144-36	2,439,980	A(2) → A(1)	*yfeS*	-
AN-144-04	2,555,687	T(6) → T(7)	*iscR*	*yfhQ*
AN-144-08	2,779,247	G(2) → G(3)	*barA*	-
AE-180-48	2,802,779	A(4) → A(3)	*ECB_02652*	*fucI*
AE-180-22	3,866,357	G(2) → G(3)	*trkD*	*insJ*-5
AE-180-24	3,866,357	G(2) → G(3)	*trkD*	*insJ*-5
AE-180-26	3,866,357	G(2) → G(1)	*trkD*	*insJ*-5
AN-144-46	3,866,357	G(2) → G(3)	*trkD*	*insJ*-5
AN-144-50	3,866,357	G(2) → G(3)	*trkD*	*insJ*-5
AE-180-12	3,866,358	G(2) → G(1)	*trkD*	*insJ*-5
AE-180-44	3,866,358	G(2) → G(1)	*trkD*	*insJ*-5
AN-144-16	3,866,358	G(2) → G(1)	*trkD*	*insJ*-5
AN-144-44	4,279,523	T(2) → T(1)	*yjdB*	-

*Position of slippage event on reference *E. coli* REL4536 genome.

†Genes flanking the corresponding slippage event in the reference *E. coli* REL4536 genome. In instances where slippage events occurred in coding regions, only one gene is listed.

4.2.1.3 GCRs

In the previous chapter, large-scale GCRs were identified by using *de novo* sequencing (section 3.2.3.2.2.2). These large-scale GCRs were further analysed for this chapter and were determined to be either inversions or translocations. Therefore, for this study, GCRs were classified as IS insertions, IS deletions, partial IS insertions, IS mediated deletions, inversions, translocations or IS element independent deletions. A total of 98 GCRs were detected in this study; with 33 and 65 GCRs accumulated in aerobically and anaerobically grown cells respectively. The anaerobic GCR per generation mutation rate was significantly 2.6-fold greater (**Figure 4.1a**) than the aerobic rate per generation rate (Mann-Whitney U = 146.5, $p = 0.001$). Curiously, when considering mutation rates per genome per day (**Figure 4.1b**), the aerobic and anaerobic GCR mutation rates were not statistically significantly different (Mann-Whitney U = 266.5, $p = 0.671$). Together, these results suggest that GCRs generally occur more frequently under anaerobic growth conditions and that they may occur independently of DNA replication. These findings are largely consistent with that of Sakai et al. (2006), who reported a two-fold increase in the frequency of GCRs in anaerobically grown cells, as compared to aerobically grown cells, when using a mutation detection system based on a single gene (28).

4.2.1.3.1 Different GCR types

Across both the aerobic and anaerobically grown MA lineages, IS insertions were the most frequently observed GCR type while translocations were the least frequently observed GCR type (**Figure 4.6**). Approximately 95% of the total GCRs detected in this study involved IS elements, with 31 and 62 IS mediated GCRs occurring in aerobically and anaerobically grown cells, respectively. Briefly, IS mediated deletions and inversions were not frequently observed, and their rates of mutations between the two environments did not differ significantly (**Figure 4.6**). Translocations were only observed in aerobically grown cells while partial IS insertions were only observed in anaerobically grown cells. Rates of IS-independent deletions per generation and per day were higher in anaerobically grown cells, as compared to aerobically grown ones (**Figure 4.6**), but these differences in rates were not significant ($p > 0.05$) by Mann-Whitney U-test.

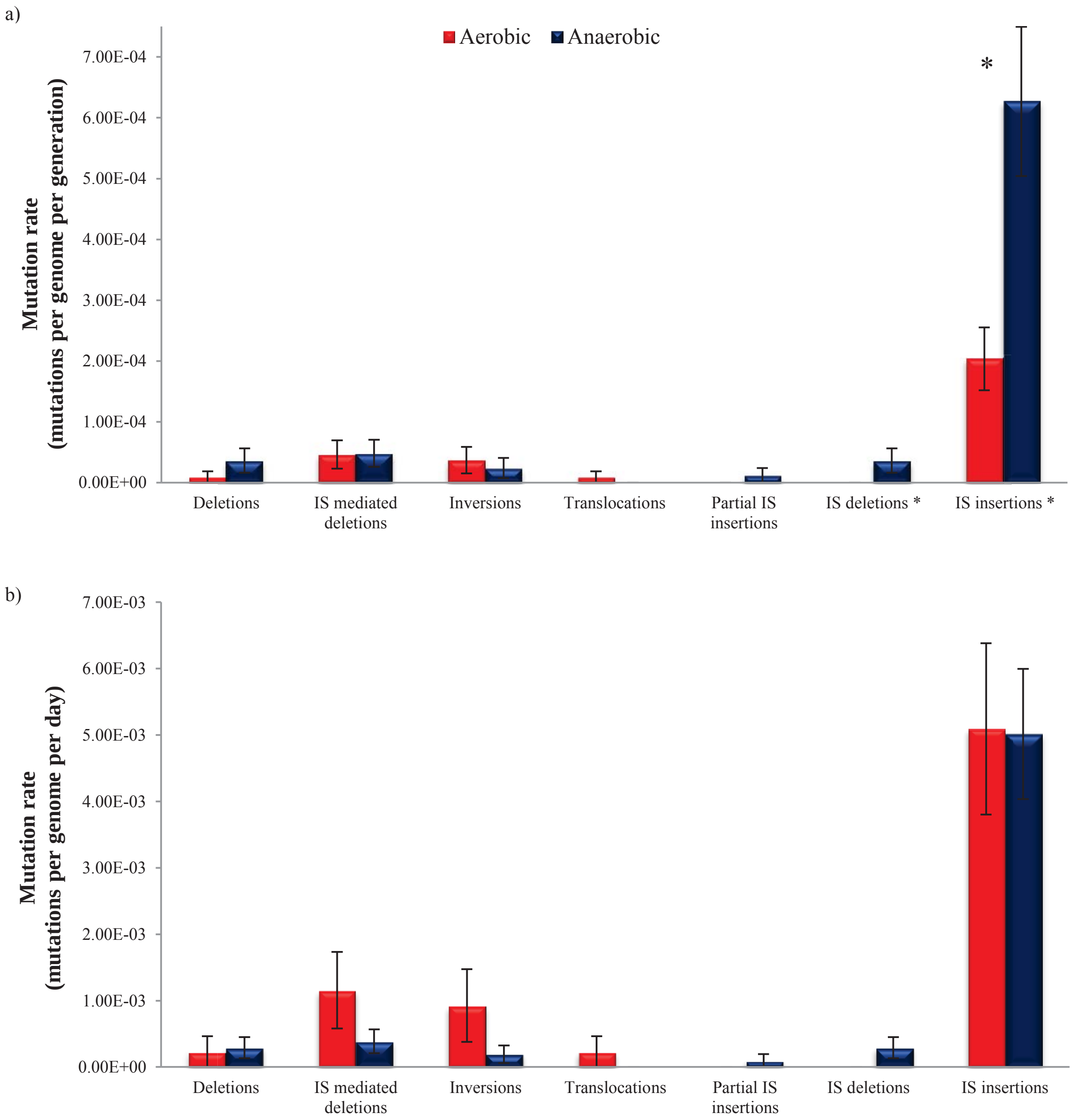

Figure 4.6. Mutation rates of different classes of GCRs in aerobically and anaerobically grown *E. coli*. Shown are a) mean mutation rates per genome per generation and b) mean mutation rates per genome per day of growth. Error bars represent standard error of the mean. Asterisk denotes a significant difference between the aerobic and anaerobic mutation rates ($p < 0.05$).

IS deletions, where only IS*150* deletion events were detected, were solely observed in anaerobically grown cells (**Figure 4.6**) but the difference in rates per generation and per day between the two environments was not significant (Mann-Whitney U = 252.0, $p >$ 0.05). In this study, an increase in IS element copy number, as well as a change in the location of an IS element, were both counted as IS insertions. Mutation rates of IS insertions per generation were three-fold greater (**Figure 4.6a**) in anaerobically grown cells, as compared to aerobically grown cells (Mann-Whitney U = 155.0, p = 0.002) while, in contrast, mutation rates of IS insertions per day (**Figure 4.6b**) between the two environments were not statistically significantly different (Mann-Whitney U = 258.0, *p* = 0.735). These results collectively indicate that IS insertions were more frequent under anaerobic growth conditions and that they may also occur independently of DNA replication.

4.2.1.3.1.1 IS element transposition

Lee et al. (2014) also investigated the movement of IS elements in their MA study, which they collectively refer to as structural variants, and observed an average rate of 2.11×10^{-4} IS insertions per generation for aerobically grown cells (62). A similar rate of 2.04×10^{-4} IS insertions per generation was observed for the aerobically grown cells of this study (**Figure 4.6a**). There are nine different IS elements present in the ancestral *E. coli* REL4536 genome (**Table 3.4**) and in this study, these IS elements differed in their contributions to mutations. Across the MA lineages, none of the 98 GCRs detected involved IS*2*, IS*30*, IS*600* or IS*911* elements. Among the IS elements that were involved in GCRs, mutation events involving IS*150* were the most abundant, while those involving IS*4* were rare (**Figure 4.7**). In some cases, IS elements are self-regulated; a higher rate of excision and a lower rate of transposition is observed for those IS elements that are present in high copy numbers in a genome (256). The most abundant IS element in the REL4536 genome is IS*1*, with 26 copies, while IS*150* is the second most abundant element, with 7 copies. As both elements, especially IS*150*, had high transposition rates (discussed below), the self-regulation theory was not supported in this study.

For IS*1* and IS*4* mediated GCRs, per generation transposition rates (**Figure 4.7a**) in the aerobically and anaerobically grown cells were not statistically significantly different (Mann-Whitney U-test, *p* > 0.05). Meanwhile, IS transposition rates per day of IS*1* and IS*4* elements were 3.2- and 2.4- fold greater (**Figure 4.7b**), respectively, in aerobically grown cells as compared to anaerobically grown cells. Per generation rates of IS*3* mediated GCRs were 3.3-fold greater (**Figure 4.7a**) in anaerobically grown cells, as compared to aerobically grown cells (Mann-Whitney U = 247.0, *p* = 0.056). When considering IS*3* transposition rates per day (**Figure 4.7b**), rates between aerobically and anaerobically grown cells were not statistically significantly different (Mann-Whitney U = 257.0, *p* = 0.793), suggesting that IS*3* transposition may be more active under anaerobic growth conditions. Rates of IS*186* mediated GCRs per generation (**Figure 4.7a**) and per day (**Figure 4.7b**) were significantly 4.6-fold (Mann-Whitney U = 241.5, p = 0.08) and 1.4-fold (Mann-Whitney U = 237.5, p = 0.025) greater, respectively, in aerobically grown cells, as compared to anaerobically grown cells, by Mann-Whitney U-test. These results imply that IS*186* transposition may not be dependent on DNA replication, and that IS*186* is more active under aerobic growth conditions. Rates of IS*150* mediated GCRs per generation were significantly 9.6-fold greater (**Figure 4.7a**) in anaerobically grown cells as compared to aerobically grown cells (Mann-Whitney U = 75.0, $p < 0.001$). When considering rates per day, IS*150* transposition rates were still significantly three-fold greater (**Figure 4.7b**) in anaerobically grown cells as compared to aerobically grown cells (Mann-Whitney U = 147.0, $p < 0.001$). These findings imply that IS*150* transposition occurs independently of DNA replication, and that, IS*150* transposition is considerably more frequent under anaerobic growth conditions.

To date, the work of Lee et al. (2014) is the only published MA study that uses wild-type *E. coli* in which IS element mobility has been examined. In their study, IS*1*, IS*2*, IS*3*, IS*4*, IS*5* and IS*186* element transposition was observed; with IS*1* and IS*5* being the most active elements (62). The mean rates per generation they reported for IS*1* (5.65×10^{-5}), IS*3* (4.27×10^{-6}), IS*4* (1.19×10^{-5}) and IS*186* (1.90×10^{-5}) are comparable to those observed in the aerobically grown cells of this study; where rates of 1.11×10^{-4}, 1.85×10^{-5}, 9.26×10^{-6} and 5.56×10^{-5} mutations per generation were observed for the IS*1,* IS*3*, IS*4* and IS*186* elements, respectively. Perhaps IS*186* transposition is specifically induced in response to ROS exposure or oxidative stress, resulting in greater mobility in aerobically grown cells, rather than anaerobically grown cells

Although, there are currently no published rates for IS*150* transposition in *E. coli* to compare results of this study to, the IS*150* element was found to be very active during Lenksi’s LTEE (35, 63-65). IS element transposition can be regulated in various ways, reviewed in (256), but it is not immediately apparent why IS*150* is so active in anaerobically grown cells. The IS*150* transposase requires programmed translational frameshifting to be induced so perhaps the increased transposition rate is in response to a specific signal under anaerobic conditions. As mentioned previously, *E. coli* can produce acetic acid, ethanol, lactic acid, formic acid and succinate acid during anaerobic fermentation (257). Perhaps the acidic environment that is created in the anaerobically grown cells of this study has led to increased IS*150* transposition. To further address these questions, the expression values of the various genes involved in IS element transposition were analysed and will be discussed in section 4.2.3.1.1.2.1.

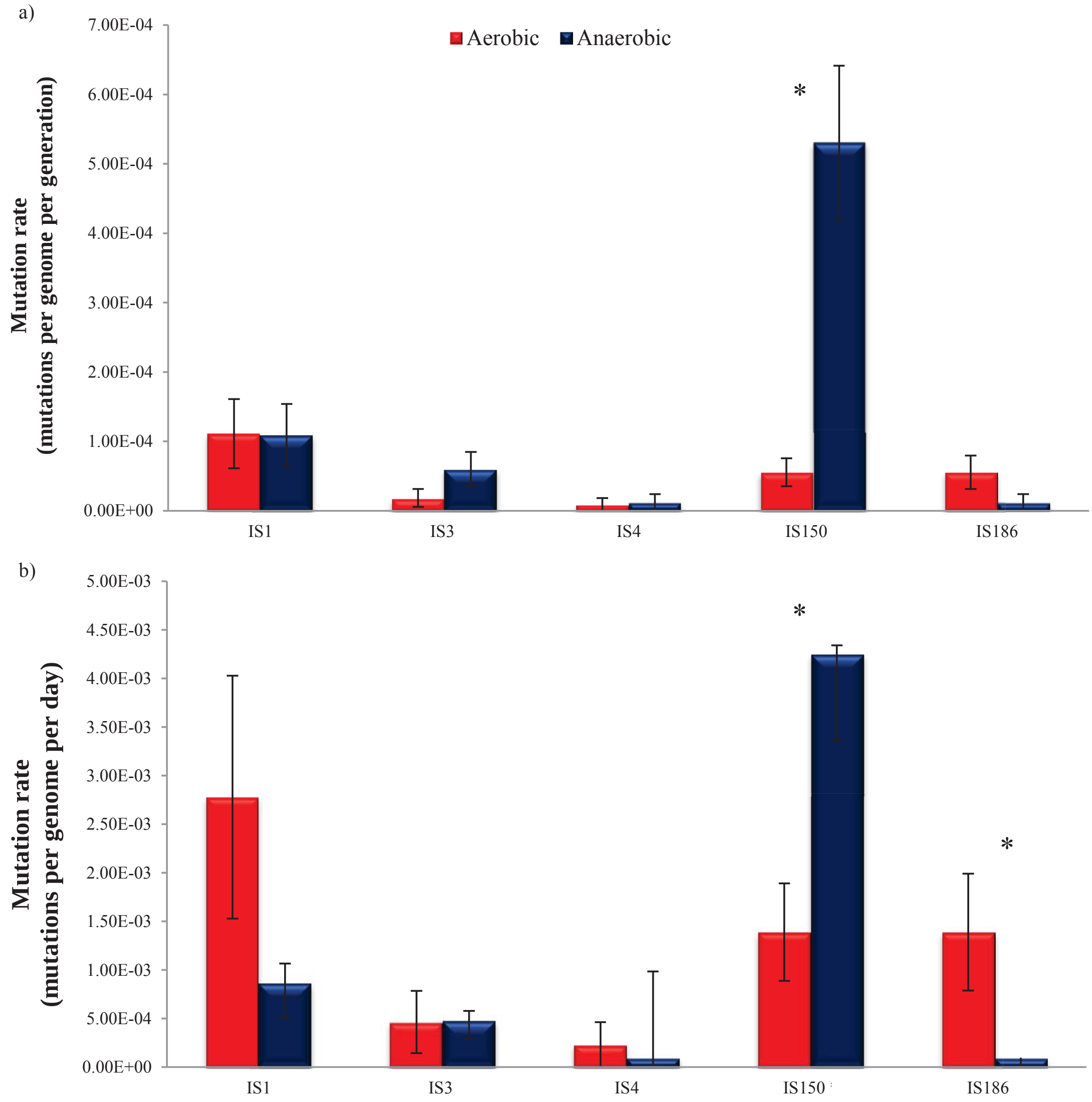

Figure 4.7. Transposition rates of different IS elements in REL4536 when grown aerobically and anaerobically. Shown are a) mean mutation rates per genome per generation and b) mean mutation rates per genome per day of growth. Error bars represent standard error of the mean. Asterisk denotes a significant difference between the aerobic and anaerobic mutation rates ($p < 0.05$).

4.2.1.3.2 Large-scale GCRs

As mentioned previously, the large-scale GCRs in the MA lineages were identified using *de novo* assembly of genome sequence data (section 3.2.3.2). In total, seven large-scale GCRs were identified, with five large-scale GCRs occurring in aerobically grown lineages and two large-scale GCRs occurring in anaerobically grown lineages. As can be seen in **Figure 4.8**, all of the large-scale GCRs that were detected involved IS-mediated inversions. Lineages AE-180-24, AN-144-12 and AN-144-46 underwent IS*150*-mediated inversions around the terminus of replication while lineage AE-180-06 underwent an IS*186*-mediated inversion of a 145,635 bp region of the genome. Lineage AE-180-04 had three large-scale GCRs; one translocation over the terminus as well as two inversions (**Figure 4.8**).

4.2.1.3.2.1 Cumulative GC-skew

Under conditions of no mutation bias, the two strands of a bacterial chromosome should be subject to the same mutation rate (247). Therefore, Chargaff's 2nd Parity Rule holds that under these conditions, equal amounts of complementary nucleotide bases (i.e. A = T and G = C) are expected in a DNA strand. However, in circular bacterial genomes, due to either mutational biases or selective pressures [reviewed in (247, 258)], the leading strand in replication is enriched in G and T while the lagging strand is enriched in A and C. The observed asymmetries in nucleotide base composition between the two strands are commonly referred to as GC- or AT- skews. As the GC-skew is usually stronger than the AT-skew, often only the GC-skew is considered (259). GC-skew, measured by plotting (G-C)/(G+C) in a sliding window along the sequence, has been correlated with the placing of the origin and terminus of replication in a genome (260)

Briefly, in a plot of the GC-skew, the measurement of strand asymmetry is positive for the leading strand, due to the excess of guanine bases over cytosine bases, and negative for the lagging strand due to the excess of cytosine bases over guanine bases. The points at which the measurements of strand asymmetry switch between positive and negative values represent either the origin or terminus of replication. However, this measurement can suffer from local fluctuations and so, cumulative GC-skews are preferred (261). Cumulative GC-skews still use a sliding window along the sequence, but involves the summation of adjacent sliding windows from an arbitrary start point to an arbitrary end

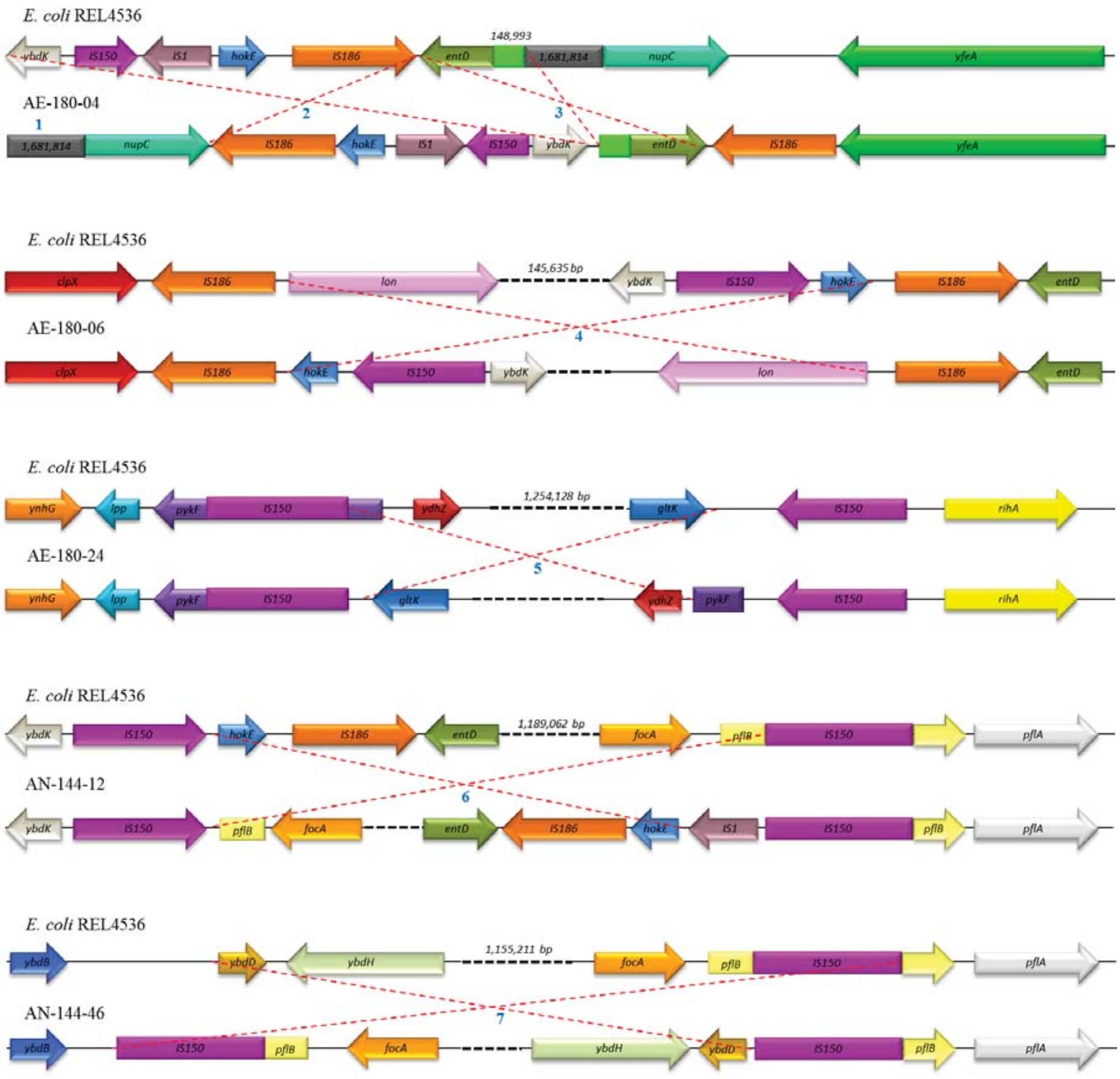

Figure 4.8. Gene diagrams of the seven large-scale GCRs that occurred in AE-180-04, AE-180-06, AE-180-24, AN-144-12 and AN-144-46. Dotted red lines indicate inversions. GCRs are labelled from 1 to 7: GCR 1 is a translocation of roughly 1.7 Mb of DNA around the terminus in AE-180-04, GCR 2 is an IS*186*-mediated inversion in AE-180-04, GCR 3 is an IS element mediated inversion in AE-180-04, GCR 4 is an IS*186*-mediated inversion in AE-180-06, GCR 5 is an IS*150*-mediated inversion of roughly 1.3 Mb of DNA around the terminus in AE-180-24, GCR 6 is an IS*150*-mediated inversion of roughly 1.2 Mb of DNA around the terminus in AN-144-12 and GCR 7 is an IS*150*-mediated inversion of roughly 1.2 Mb of DNA around the terminus in AN-144-46.

point (261). In a cumulative GC-skew plot, the peaks correspond to the switch points between the leading and lagging strands with a minimum value at the origin of replication and a maximum value at the terminus of replication. The GC-skew of *E. coli* REL4536 can be seen in **Figure 4.9**.

Chromosomal replication in circular chromosomes starts at the origin of replication and terminates at the terminus region, a region typically 180° from the origin. Chromosomal inversions around the origin or the terminus of replication generally do not change the position of genes from the origin and terminus of replication (44, 258, 262). Inversions that change the distance of a gene from either the origin or terminus are generally not preferred (44, 262) as such a modification would result in the change of a gene's expression rate (258, 263). Essentially, in what is referred to as replication-associated gene dosage, the location of a gene relative to the origin and terminus of replication is thought to influence it's copy number in the cell during replication and thus, it's expression levels (263). Genes located near the origin tend to be replicated before genes near the terminus and so tend to be more abundant in the cell during replication; selective pressures then strive to maintain the positions of genes relative to the origin (262, 264). In *E. coli* and other fast-growing bacteria, replication-associated gene dosage effects are mainly associated with the genes involved in translational and transcriptional processes (264). Changes in the symmetry of the origin and terminus of replication, such as placing the two regions closer together or further apart, can result in slower growth and complicate cell replication, termination and segregation (258). While it is generally believed that GCRs around the origin of replication rarely occur, GCRs around the terminus have been observed (44).

In all of the large-scale GCRs detected in this study, the position of the origin of the replication remained unchanged while all but one of the large-scale GCRs involved rearrangements that spanned the terminus region. As the cumulative GC-skew changes sign at the origin and terminus of replication (263, 265-267), the cumulative GC-skew of the evolved MA lineages, calculated as described in section 2.2.16.3.3 , can be used to determine if the large-scale GCRs caused any changes in the relative locations of the origin and terminus of replication. As can be seen in **Figure 4.9**, while the GCRs altered the distances between the origin of replication and terminus in almost all lineages, only two lineages (AE-180-04 and AE-180-24) were drastically affected. In AE-180-04, the origin and terminus have been brought closer together while in AE-180-24 the origin and terminus have been brought farther apart (**Figure 4.10**). Disruptions in the symmetry between the origin and terminus of replication in circular chromosomes have previously been demonstrated to be detrimental to cell growth (44). Hence, as the symmetry between the origin and termination of replication have been altered in both lineages, it is possible that the growth of these two lineages has been severely affected.

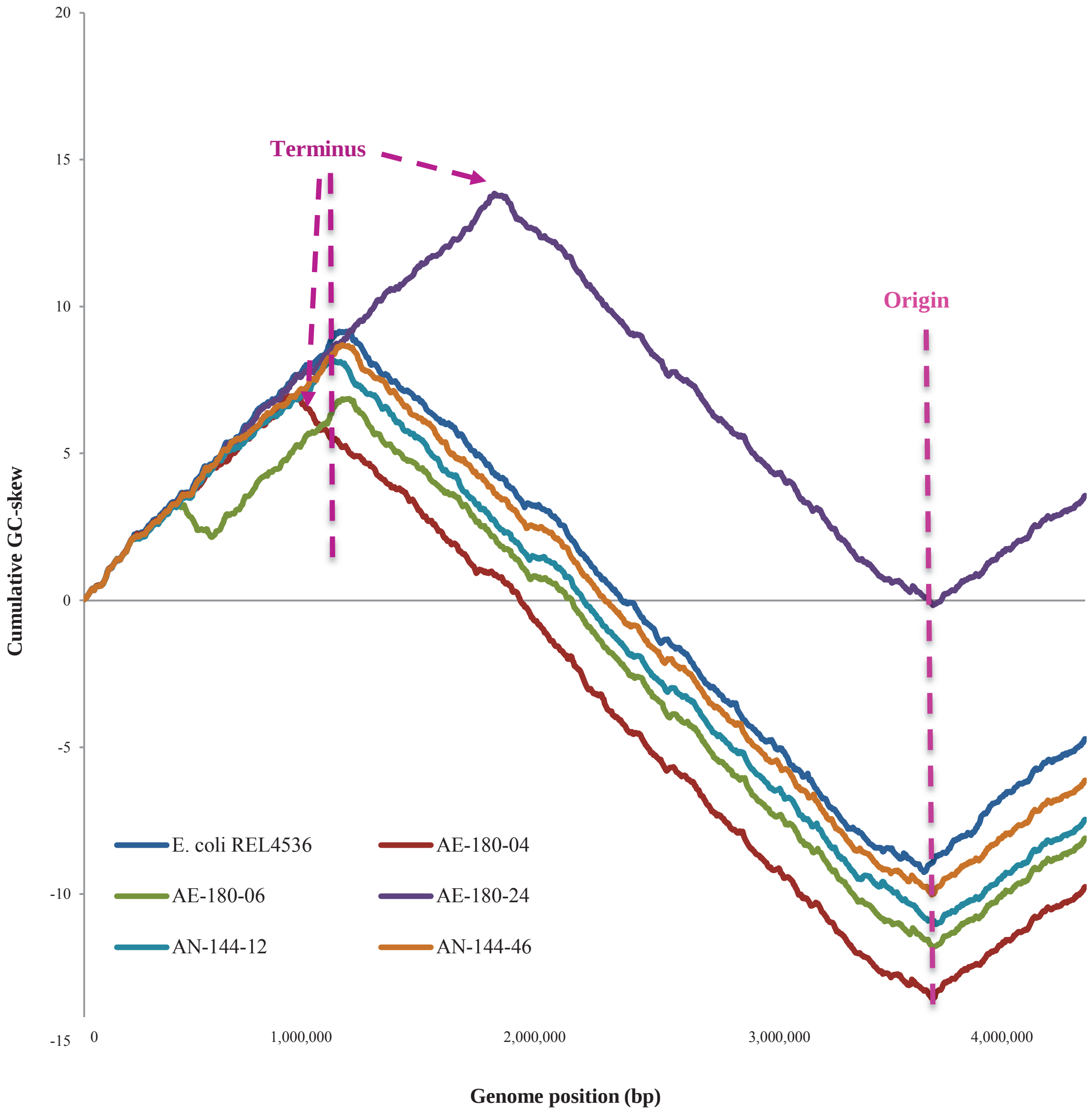

Figure 4.9. The cumulative GC-skew of *E. coli* REL4536, AE-180-04, AE-180-06, AE-180-24, AN-144-12 and AN-144-46 as a function of position in genome. The cumulative GC-skew value is calculated as (G - C)/(G + C). GC-skew has a minimum value at the origin of replication and a maximum value at the terminus of replication.

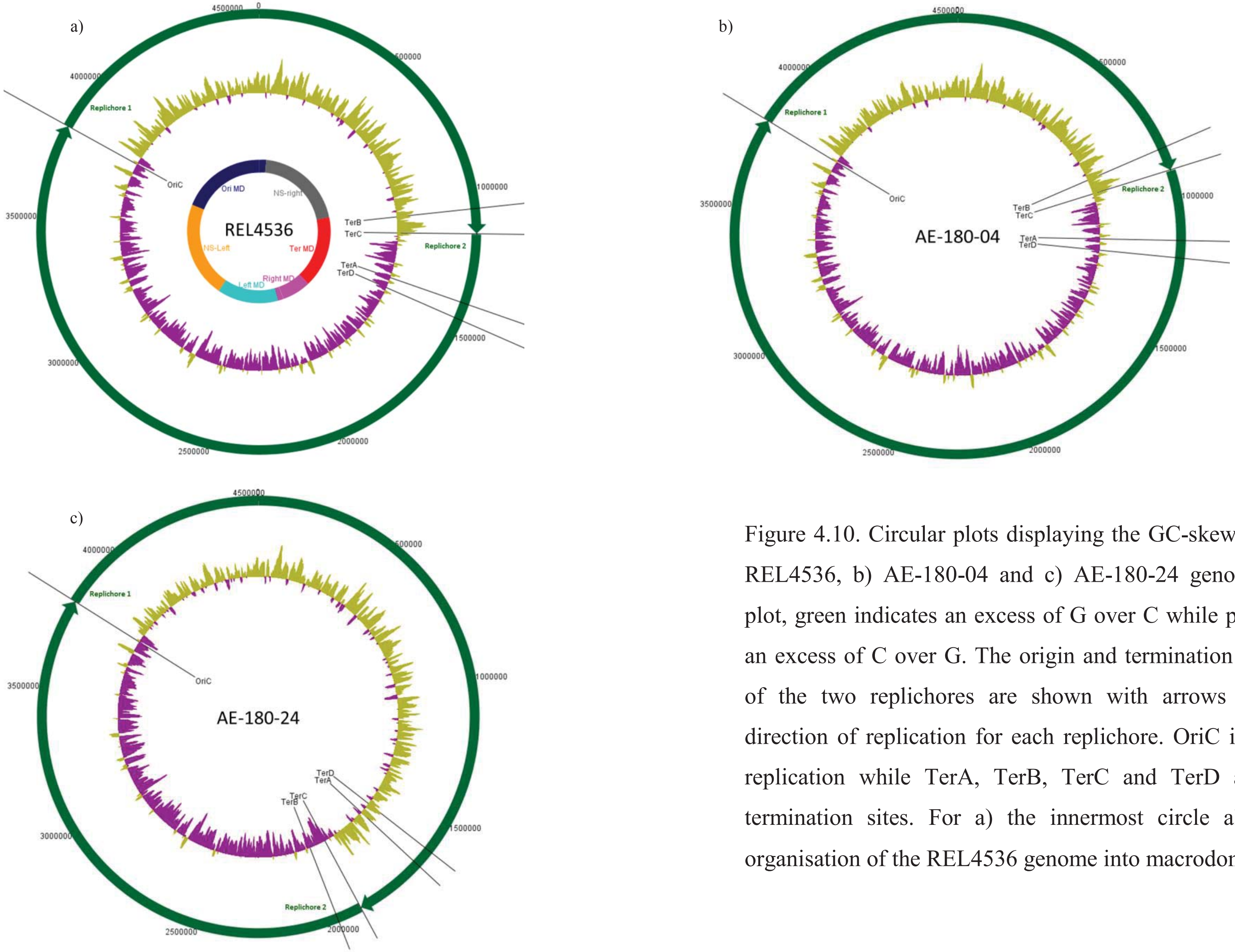

Figure 4.10. Circular plots displaying the GC-skews of: a) *E. coli* REL4536, b) AE-180-04 and c) AE-180-24 genomes. For each plot, green indicates an excess of G over C while purple indicates an excess of C over G. The origin and termination points of each of the two replichores are shown with arrows indicating the direction of replication for each replichore. OriC is the origin of replication while TerA, TerB, TerC and TerD are replication termination sites. For a) the innermost circle also shows the organisation of the REL4536 genome into macrodomains (MDs).

4.2.2 Genome-wide distribution of mutations

4.2.2.1 Genome organisation

The circular *E. coli* chromosome is organised on two different levels; into macrodomains (MDs) and replichores. MDs, roughly between 0.5 to 1 Mb in length, are discrete, structured regions of DNA to which interactions are localised (268, 269). There are four MDs in *E. coli*: the Left MD, the Right MD, the Ori MD, which contains the origin of replication (oriC), and finally the Ter MD, which contains the terminus of replication. *E. coli* also has two non-structured (NS) regions: NS^{Right} and NS^{Left}. The Ori MD is flanked by the two NS regions while the Ter MD is flanked by the Left and Right MDs (268, 269). This structure influences the mobility of chromosomal DNA; interactions between the different MDs are restricted but DNA in the NS regions can interact with any of the flanking MDs (270). Mutations that disrupt the structure of MDs, specifically the Ter MD, have been demonstrated to be detrimental to cell growth (267). The organisation of *E. coli* REL4536 into MDs can be seen in **Figure 4.10a**.

Circular chromosomes are divided into two halves, or replichores; the right arm or the first replichore and the left arm or the second replichore. The organisation of *E. coli* REL4536 into replichores can be seen in **Figure 4.10a**. For most bacterial species, the two replichores are of approximately the same length and replication occurs bi-directionally across the chromosome, with one replication fork per replichore (265, 266). In *E. coli*, the terminus region is comprised of at least 10 Ter sites dispersed in two oppositely oriented groups, with each site being A/T rich and 23 bp long. When the terminus utilisation substance, or Tus, protein binds to a Ter site, a Tus-Ter complex is created. Each Ter site binds to Tus with a specific affinity and so each replication fork can travel across five Ter sites, in what is commonly called the permissive orientation, before it encounters a Ter site in the non-permissive orientation where the DNA helicases and polymerases are stalled and replication is terminated (266). As the two replication forks are blocked in a region comprised of oppositely oriented Ter sites, this region is called the replication fork trap. The terminus region in *E. coli*, the Ter MD, also contains the dif site, a 28 bp site where chromosomal segregation during replication is initiated (267). A preference for the maintenance of the symmetry between the origin and terminus has previously been described (section 4.2.1.3.2.1).

Due to selective or mutational pressures (247, 258), the two replichores exhibit different GC-skew profiles (section 4.2.1.3.2.1). In *E. coli* genomes, the first replichore usually extends from the origin of replication, oriC, in a clockwise direction to the TerC or TerB termination sites while the second replichore extends from oriC in an anti-clockwise direction to the TerA or TerD termination sites (266). However, in the ancestral *E. coli* REL4536 genome, an inversion of approximately 1.5 Mb around the terminus by recombination occurring between two IS*1* elements has swapped the genomic locations of the Ter sites around (197), creating an asymmetrical distance between the origin and terminus of replication and a disruption in the Ter and Right MDs (**Figure 4.10a**). Nonetheless, this inversion does not impair cell growth, indicating that some level of asymmetry and imbalance in the genome can be tolerated (35).

4.2.2.1.1 Distribution of mutations within MDs

Previously studies have analysed the spatial distribution of mutations around genomes to determine if mutations were distributed at random or not (271). In particular, GCRs around the terminus have been found to comparatively preferable to those around the origin of replication, which is found in the Ori MD (44). Thus, to determine if mutations in the aerobically and anaerobically grown cells of this study displayed any trends in topology, the positions of accumulated BPSs, GCRs and indels were plotted onto the REL4536 genome (**Figure 4.11**). Under both aerobic and anaerobic conditions, BPSs and indels appear to be distributed evenly around the genome while GCRs appear to be more prevalent within the terminus containing Ter MD (section 4.2.2.1). The relative distribution of all accumulated BPSs, indels and GCRs within the macrodomains was then quantified, and a greater proportion of GCRs were found in the Ter MD for aerobically and anaerobically grown cells (**Table 4.2**). These findings are consistent with previously published reports that cells tolerate GCRs around the terminus more than they tolerate GCRs around the origin (44); possibly due to structural constraints to either maintain the symmetry between the origin and terminus of replication or to maintain genome organisation (267).

Table 4.2. Relative distribution of mutations within the macrodomains of *E. coli*.

	Ori MD	NS^{Right}	Ter MD	Right MD	Left MD	NS^{Left}
Aerobic						
BPSs	0.26	0.18	0.19	0.07	0.15	0.16
Indels	0.23	0.00	0.23	0.23	0.08	0.23
GCRs	0.15	0.15	0.48	0.06	0.15	0.00
Anaerobic						
BPSs	0.22	0.21	0.14	0.14	0.08	0.22
Indels	0.16	0.21	0.16	0.05	0.16	0.26
GCRs	0.20	0.18	0.37	0.08	0.06	0.11
Total						
BPSs	0.24	0.19	0.16	0.10	0.12	0.19
Indels	0.19	0.13	0.19	0.13	0.13	0.25
GCRs	0.18	0.17	0.41	0.07	0.09	0.07

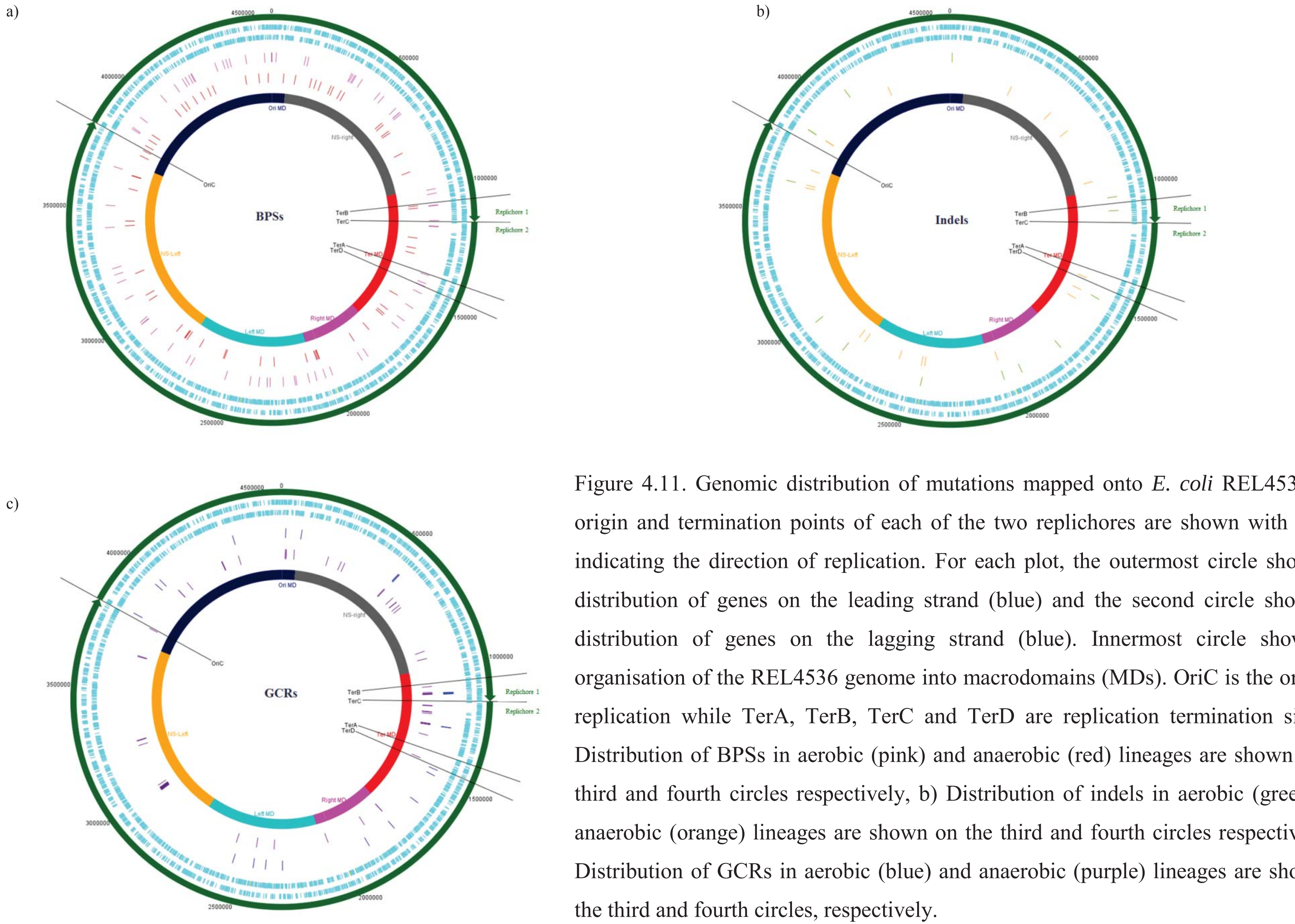

Figure 4.11. Genomic distribution of mutations mapped onto *E. coli* REL4536. The origin and termination points of each of the two replichores are shown with arrows indicating the direction of replication. For each plot, the outermost circle shows the distribution of genes on the leading strand (blue) and the second circle shows the distribution of genes on the lagging strand (blue). Innermost circle shows the organisation of the REL4536 genome into macrodomains (MDs). OriC is the origin of replication while TerA, TerB, TerC and TerD are replication termination sites. a) Distribution of BPSs in aerobic (pink) and anaerobic (red) lineages are shown on the third and fourth circles respectively, b) Distribution of indels in aerobic (green) and anaerobic (orange) lineages are shown on the third and fourth circles respectively, c) Distribution of GCRs in aerobic (blue) and anaerobic (purple) lineages are shown on the third and fourth circles, respectively.

4.2.2.1.2 First replichore versus second replichore

By using the GC-skew plot for *E. coli* REL4536, the two replichores were distinguished such that both replichores terminated around sites TerB or TerC (**Figure 4.11**). Based on this distinction, the first replichore in the REL4536 genome has a length of roughly 1.9 Mb and encodes 1,789 genes while the second replichore has a length of roughly 2.6 Mb and encodes 2,431 genes. Given the different distributions of the different mutation types (**Figure 4.11**), it was possible that the two replichores accumulated mutations at different rates. To investigate this further, the mutation rates at which mutations occurred within each replichore were calculated. For lineages with large-scale GCRs that involved inversions around the terminus region, these resulted in changes in the sizes of each replichore, and new replichore sizes were taken into account in such cases. Furthermore, inversions around the terminus region were excluded from the analysis because they were not specific to a single replichore. As the two replichores are of different sizes, mutation rates presented have been normalised and are thus, average mutation rates per nucleotide per unit time.

The per generation mutation rates for the first and second replichores were 1.5-fold (Mann-Whitney U = 248.0, p = 0.206) and 1.8-fold (Mann-Whitney U = 175.0, p = 0.009) greater, respectively, in anaerobically grown cells as compared to aerobically grown cells. Meanwhile, the per day mutation rate for the first and second replichores were 2.1-fold (Mann-Whitney U = 177.0, p = 0.010) and 1.7-fold (Mann-Whitney U = 159.0, p = 0.003) greater in aerobically grown cells as compared to anaerobically grown cells. These observations are consistent with previously reported per generation and per day mutation rates (**Figure 4.1**).

To determine if specific mutation types were responsible for the difference in mutation rates observed between the two replichores, the rates at which BPSs, indels and GCRs occurred in the two replichores under each environment were calculated (**Figure 4.12**). Per generation rates for BPSs (**Figure 4.12a)** did not appear to be biased towards a particular replichore in either environment ($p > 0.05$, Mann-Whitney U-test). However, rates of BPSs per day (**Figure 4.12b**) were 2.7-fold (Mann-Whitney U = 251.0, p = 0.007) and 2.2-fold (Mann-Whitney U = 191.5, p = 0.022) greater, in the first and second replichore, respectively, of aerobically grown cells, as compared to anaerobically grown cells. Rates of indels per generation and per day (**Figure 4.12**) did not appear to be biased towards a particular replichore in either environment ($p > 0.05$, Mann-Whitney U-test).

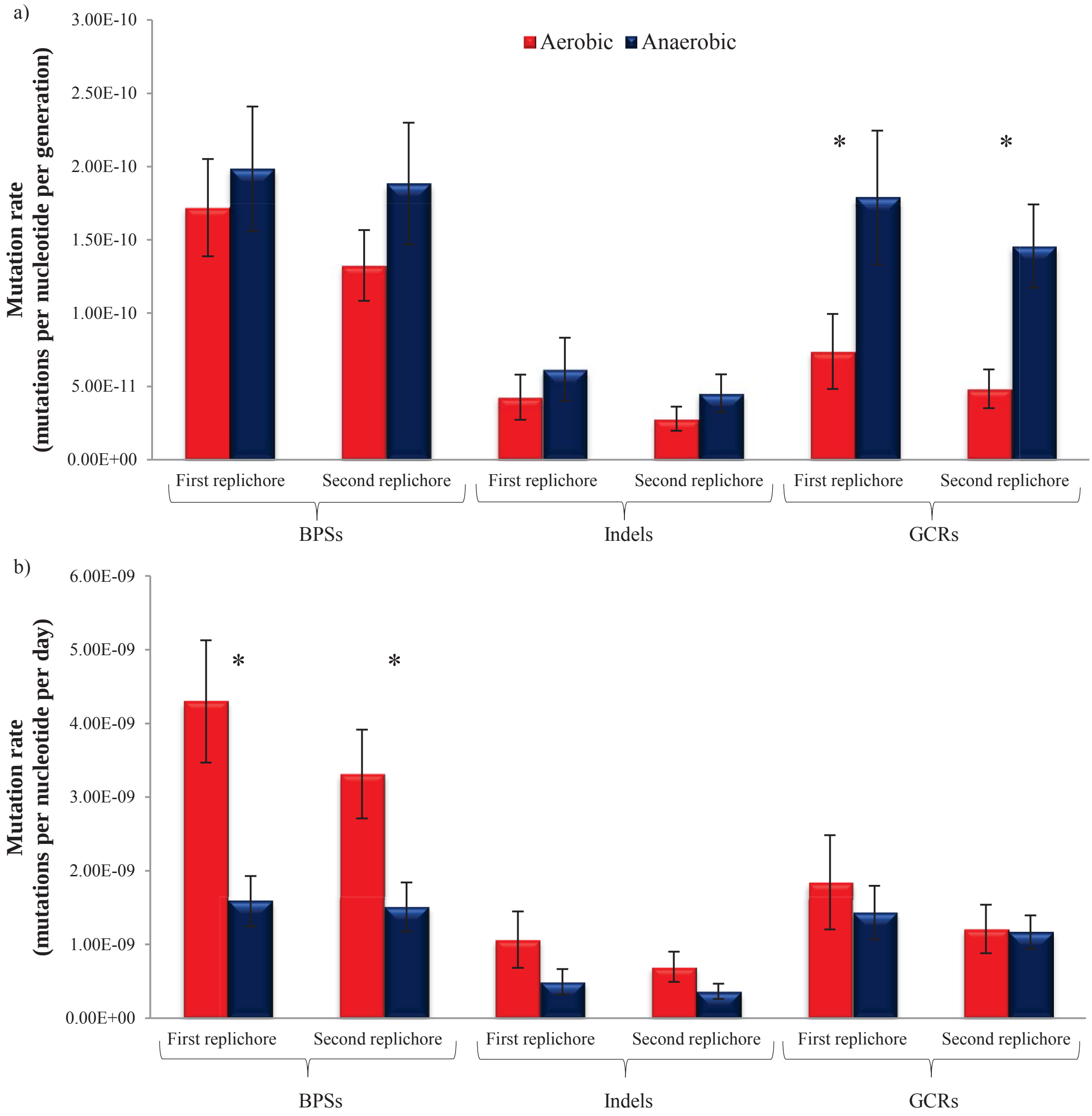

Figure 4.12. Mutation rates of different mutation types in the two replichores in aerobically and anaerobically grown *E. coli*. Shown are a) mean mutation rates per nucleotide per generation and b) mean mutation rates per nucleotide per day of growth. Error bars represent standard error of the mean. Asterisk denotes a significant difference between the aerobic and anaerobic mutation rates ($p < 0.05$).

On the other hand, rates of GCRs per generation (**Figure 4.12a**) were significantly 2.4-fold (Mann-Whitney U = 201.0, p = 0.025) and three-fold (Mann-Whitney U = 155.0, p = 0.002) greater, in the first and second replichore, respectively, of anaerobically grown cells, as compared to aerobically grown cells. Meanwhile, per day rates of GCRs (**Figure 4.12b**) did not appear to be biased towards a particular replichore in either environment ($p > 0.05$, Mann-Whitney U-test). As GCRs per generation were shown to be more prevalent in anaerobically grown cells (section 4.2.1.3), it was interesting to see that they occurred at relatively similar rates across the two differently-sized replichores and that one region of the genome in particular was not responsible for the high anaerobic GCR per generation rate.

To investigate the G → T versus C → A mutation rate asymmetry further (section 4.2.1.1.1), mutation rates for the 12 different types of BPS, normalized to account for any differences in nucleotide content per replichore, were calculated (**Figure A.1** and **Figure A.2**). For G → T transversions, mutation rates per generation were significantly greater in the first replichore under aerobic conditions (Mann-Whitney U = 228.0, $p = 0.025$). On the other hand, C → A mutation rates per generation were greater in the second replichore under anaerobic conditions (Mann-Whitney U = 250.0, $p = 0.055$). Likewise, per generation mutation rates for A → C transversions and T → G transversions were significantly greater in the second replichore under anaerobic conditions (Mann-Whitney U = 234.0, $p = 0.027$). These results suggested the presence of a strand bias in the types of BPSs that arose under aerobic and anaerobic conditions, though the cause of this is unknown.

The factors behind any asymmetric mutation pressures on the aerobically and anaerobically grown cells are not immediately apparent. It is possible that the observed BPS spectrum is a result of replication strand bias (259) where the chromosome can be clearly differentiated by GC-skew between the two replichores (258). During replication, the template leading strand is discontinuously single-stranded while the complementary Okazaki fragments are being synthesized, while the template lagging strand is maintained in a double-stranded structure during continuous leading strand synthesis (**Figure 4.13**). As ssDNA is more susceptible to DNA damage and so more prone to mutation (246, 258), the rates at which mutations arise in the template leading strand are likely to be higher than those in the template lagging strand. Moreover, as replication is slower during anaerobic growth, it seems more likely that mutation rates may be higher in the template leading strand, than for aerobically grown cells. Therefore, it is possible that this replication bias has led to the observed mutation rate asymmetry seen in **Figure 4.2**. In addition, the mutational strand bias is potentially the result of transcription bias, where the un-transcribed strand is repaired more efficiently than the transcribed strand (258, 259, 272). Alternatively, it is also possible that the observed BPS spectrum of the aerobically and anaerobically grown cells is the result of different physiological conditions generated by the cells during growth under their respective environmental conditions. Thus, to determine whether the observed BPS spectra of the aerobically and anaerobically grown cells are a result of a replication, transcriptional bias, physiology, or a combination of the three, further work will be required.

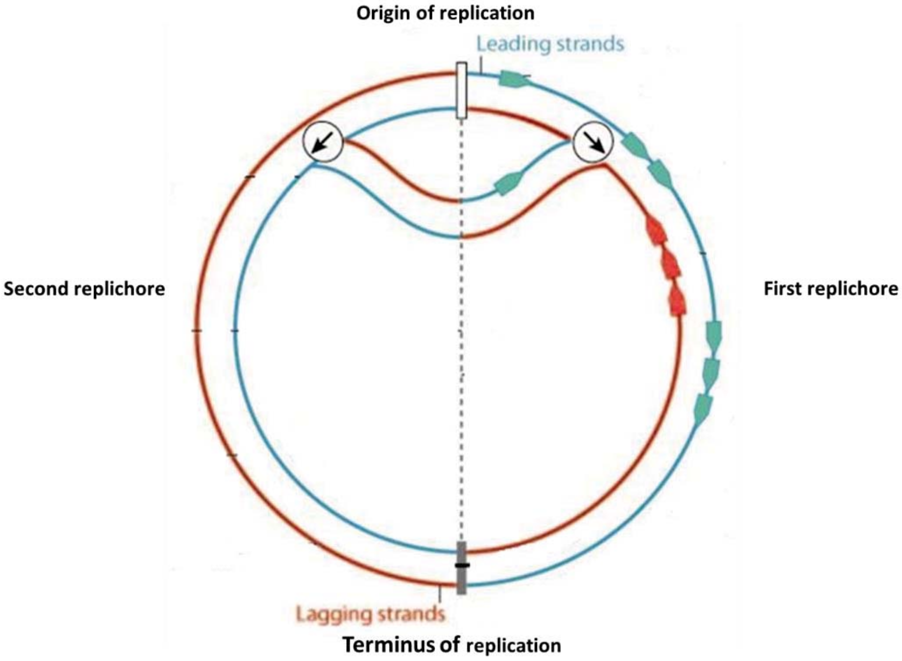

Figure 4.13. Genome organisation of *E. coli*. Due to the bi-directional nature of *E .coli* replication, each replichore has both leading (template leading strand shown in green) and lagging (template lagging strand shown in red) strands. Figure modified from (263)

4.2.2.2 Mutation hotspots

Mutation hotspots can be defined as regions of the genome where mutations are found to occur frequently (273). It is thought that hotspots are a result of some inherent signature or instability associated with the DNA sequence itself, rendering the region more susceptible to mutations. For instance, DNA sequences, such as repetitive sequences, may contain motifs that are sites for homologous recombination (274, 275) and can be putative mutation hotspots. Pathogenicity islands and prophages, *via* homologous recombination in tRNA genes, can integrate into chromosomes and are also frequently found to be mutation hotspots (276). DNA structure may also play a role, as it is possible that chromosome folding may leave some regions more exposed and thus, more prone to mutations (274).

In this study, across all 48 MA lineages, 34 genes were mutated more than once. In most of the genes, parallel mutations, that is identical mutations that have arisen independently, did not occur. However, there were some instances where identical mutations were found in independent lineages; these regions could be possible mutation hotspots. In this study, putative hotspots were identified for three BPS, two indel and 12 GCR events (**Table 4.3**).

For putative BPS hotspots, two were found solely under anaerobic growth conditions while one was identified under both aerobic and anaerobic growth conditions. Potential BPS hotspots were found in the intergenic region between *tdcA* and *tdcR*, the intergenic region between *yieP* and *rrsC* and in *citF* (**Table 4.3**), with the latter being a non-synonymous mutation found only in anaerobic lineages, suggesting that there may be some selective pressures acting on the lineages in this study. Potential indel hotspots, most likely DNA regions prone to slippage during DNA replication, were found under both aerobic and anaerobic growth conditions and included a run of eight adenine nucleotides in the intergenic region between the *umuD* and *ycgN* genes and a pair of guanine nucleotides in the intergenic region between the *trkD* and *insJ-5* genes (**Table 4.3**).

For putative GCRs hotspots, two were found solely under aerobic growth conditions, five were found only under anaerobic growth conditions and five were found in both aerobic and anaerobic growth conditions. Three of the possible hotspots were specifically for deletions (**Table 4.3**). Regions between the *ECB_01527* and *ECB_01510* genes were deleted in four separate lineages while a 27 gene deletion between the *yegR* and *yegQ* genes occurred twice in anaerobic lineages. In two separate aerobic lineages, regions between the *ybcQ* and *ompT* genes were deleted. All of the deleted regions involved either putative pathogenicity island genes or prophage genes, consistent with commonly encountered hotspots. The other eight potential GCR hotspots specifically involved IS*150* transposition (**Table 4.3**). The intergenic region between *trg* and *mokB*, in particular, was prone to IS*150* insertion, with insertion observed in 10 lineages. This region appears to have an affinity for IS*150* insertion as this mutation has been observed in other studies in *E. coli* as well (35, 188, 205). For the remaining possible hotspots, there were three parallel instances of an IS*150* insertion in the intergenic region between *mokC* and *nhaA* while *ynjII*, *pflB*, *menC*, *yfcC*, *rhaS*, *cycA* and the intergenic region between *nupC* and *yfeA* had two instances of IS*150* transposition each (**Table 4.3**).

It is possible that the multiple occurrences of certain mutations are not putative hotspots but rather an indication of selective pressures acting on the lineages in this study. However, based on the calculated ratio of synonymous to non-synonymous ratio BPSs, it would appear that selection in the MA study was minimal (section 4.2.1.1.3). Nonetheless, it would be possible to determine if a particular mutation is subject to selection by recreating the mutation in the ancestral strain and by using competitive fitness assays (section 2.2.13) to determine whether the mutation contributes to fitness. Additionally, *in silico* prediction techniques (274) could be used to search for motifs that are known to be hotspots for mutations, such as repeat regions or genomic islands (274). On the whole, these putative hotspot mutations don’t appear to be very strong as they are only found in a small number of lineages and are not localised to specific regions of the genome.

Table 4.3. Mutations identified in multiple MA lineages.

Mutation Type	REL4536 Reference Genes[†]	Reference Position 1[*]	Mutation	Lineage
BPS	*citF*	622,489	A → G	AN-144-16
BPS	*citF*	622,489	A → G	AN-144-20
BPS	*tdcA* & *tdcR*	3,172,999	C → A	AN-144-28
BPS	*tdcA* & *tdcR*	3,172,999	C → A	AN-144-38
BPS	*trkD* & *insJ-5*	3,869,337	A → G	AE-180-26
BPS	*trkD* & *insJ-5*	3,869,337	A → G	AN-144-24
Indel	*umuD* & *ycgN*	1,504,657	A(8) → A(9)	AE-180-38
Indel	*umuD* & *ycgN*	1,504,657	A(8) → A(9)	AN-144-50
Indel	*trkD* & *insJ-5*	3,866,357	G(2) → G(3)	AE-180-22
Indel	*trkD* & *insJ-5*	3,866,357	G(2) → G(1)	AE-180-26
Indel	*trkD* & *insJ-5*	3,866,357	G(2) → G(3)	AN-144-46
Indel	*trkD* & *insJ-5*	3,866,358	G(2) → G(1)	AE-180-12
Indel	*trkD* & *insJ-5*	3,866,358	G(2) → G(1)	AE-180-44
Indel	*trkD* & *insJ-5*	3,866,358	G(2) → G(1)	AN-144-16
GCR	*mokC* & *nhaA*	16,972	IS*150* insertion	AN-144-04
GCR	*mokC* & *nhaA*	16,972	IS*150* insertion	AN-144-40
GCR	*mokC* & *nhaA*	16,992	IS*150* insertion	AE-180-04
GCR	*ybcQ* & *ompT*	547,700	Deletion of 5 genes	AE-180-34
GCR	*ybcQ* & *ompT*	547,702	Deletion of 5 genes	AE-180-22
GCR	*yegR* & *yegQ*	632,692	Deletion of 27 genes	AN-144-46
GCR	*yegR* & *yegQ*	632,699	Deletion of 27 genes	AN-144-50
GCR	*ynjI*	910,345	IS*150* insertion	AN-144-24
GCR	*ynjI*	910,345	IS*150* insertion	AN-144-32
GCR	*ECB_01527* & *ECB_01510*	1,117,789	Deletion of 13 genes	AE-180-04
GCR	*ECB_01527* & *ECB_01510*	1,117,789	Deletion of 13 genes	AE-180-14
GCR	*ECB_01527* & *ECB_01510*	1,117,802	Deletion of 13 genes	AN-144-50
GCR	*ECB_01527* & *ECB_01510*	1,117,803	Deletion of 10 genes	AE-180-30
GCR	*trg* & *mokB*	1,272,399	IS*150* insertion	AN-144-50
GCR	*trg* & *mokB*	1,272,453	IS*150* insertion	AE-180-24
GCR	*trg* & *mokB*	1,272,453	IS*150* insertion	AN-144-12
GCR	*trg* & *mokB*	1,272,453	IS*150* insertion	AN-144-34
GCR	*trg* & *mokB*	1,272,455	IS*150* insertion	AN-144-46
GCR	*trg* & *mokB*	1,272,467	IS*150* insertion	AN-144-44
GCR	*trg* & *mokB*	1,272,468	IS*150* insertion	AE-180-30
GCR	*trg* & *mokB*	1,272,470	IS*150* insertion	AN-144-20
GCR	*trg* & *mokB*	1,272,470	IS*150* insertion	AN-144-30
GCR	*trg* & *mokB*	1,272,470	IS*150* insertion	AN-144-40
GCR	*pflB*	1,764,888	IS*150* deletion	AN-144-38
GCR	*pflB*	1,764,888	IS*150* deletion	AN-144-48
GCR	*menC*	2,295,162	IS*150* insertion	AE-180-08
GCR	*menC*	2,295,162	IS*150* insertion	AE-180-40
GCR	*yfcC*	2,334,210	IS*150* insertion	AN-144-04
GCR	*yfcC*	2,334,210	IS*150* insertion	AN-144-08
GCR	*nupC* & *yfeA*	2,421,315	IS*150* insertion	AN-144-08
GCR	*nupC* & *yfeA*	2,421,323	IS*150* insertion	AE-180-06
GCR	*rhaS*	4,043,793	IS*150* insertion	AN-144-14
GCR	*rhaS*	4,043,794	IS*150* insertion	AN-144-24
GCR	*cycA*	4,381,583	IS*150* insertion	AE-180-06
GCR	*cycA*	4,381,587	IS*150* insertion	AN-144-40

[*] Position of mutational event on reference *E. coli* REL4536 genome.

4.2.3 Activities of genes involved in maintaining genome integrity

The many different DNA repair pathways encoded by the *E. coli* REL4536 genome have been previously discussed. The presence of particular mutations should induce the expression of certain DNA repair pathways. As such, gene expression data may assist in understanding the mutation spectrum that was observed in the MA study for *E. coli* that was grown either aerobically or anaerobically (section 4.2.1). In particular, expression data for genes involved in the repair of the 8-oxoG lesion or recombination may provide a better comprehension of the observed BPS (**Table 4.3**) and IS element transposition (**Figure 4.7**) rates, respectively. To investigate the activities of various pathways involved in maintaining genome integrity, a transcriptome analysis using RNAseq was undertaken (section 1.4.8.1.1).

While the original aim of the transcriptome analysis was to correlate the observed mutation spectra to gene expression in the two environments, there were difficulties in obtaining sufficient biomass in actively growing (log phase) cultures to extract enough RNA for RNAseq. Thus, RNA was extracted from stationary phase REL4536 cultures grown aerobically and anaerobically (section 2.2.10.2), sequenced and analysed (section 2.2.16.4). Since the growth conditions used for obtaining the transcriptome data differed to those used for the MA lineage maintenance, there are limitations in the direct interpretation of the mutation spectra with regard to expression data. This is further discussed in section 4.2.3.1.1.

4.2.3.1 Differential gene expression

The non-stranded transcriptomes of three aerobic and two anaerobic stationary phase REL4536 cultures were sequenced (the library construction processes failed for the third anaerobic sample obtained). Paired-end sequencing yielded at least 13 million raw reads for each sample (**Table 4.4**). Before aligning the reads to the reference REL4536 genome, reads that aligned to REL4536 rRNA and tRNA genes were filtered from the dataset (section 2.2.16.4.3). Aerobic samples, had almost 12 million reads per sample for subsequent data analysis, and rRNA and tRNA read contamination ranged from 3 to 18% of total reads (**Table 4.4**). However, for the anaerobic samples, around 83% of total reads were rRNA and tRNA, leaving only approximately 2 million reads per sample for subsequent data analysis. This discrepancy in rRNA and tRNA levels was unexpected, and it is possible that the sequencing provider only used the Ribo-Zero™

rRNA Removal Kits for mRNA enrichment on the aerobic samples, and not on the anaerobic samples.

Table 4.4. Summary statistics for RNAseq data obtained from Bowtie2 and EDGE-Pro output.

	Aerobic 1	Aerobic 2	Aerobic 3	Anaerobic 1	Anaerobic 3
Total raw reads	13,257,771	13,714,491	13,189,749	13,189,749	13,714,491
Number reads after trimming	13,114,326	13,537,221	13,092,166	13,111,449	13,612,037
Number reads aligned to rRNA and tRNA genes	878,134	2,334,549	386,153	11,517,553	11,484,586
Number reads for alignment	12,236,192	11,202,672	12,706,013	1,593,896	2,127,451
Number reads uniquely aligned to genes	12,139,301	11,111,694	12,623,142	1,402,020	1,949,774
Number reads aligned more than once	28,938	30,384	22,520	134,914	107,339
Number reads not aligned	67,953	60,594	60,351	56,962	70,338
Total number reads used for RPKM calculation	12,163,225	11,130,707	12,642,386	1,401,717	1,954,087
Overall alignment rate	99.44%	99.46%	99.53%	96.43%	96.69%

Differentially expressed genes between aerobically and anaerobically grown REL4536 were determined as described in section 2.2.16.4.4.1. In aerobically grown cells, 1,273 significantly up-regulated genes (p-adj < 0.05) were identified; with 760 having at least two-fold greater expression than in anaerobically grown cells. In anaerobically grown cells, 1,260 significantly up-regulated genes were identified (p-adj < 0.05); with 680 genes having at least two-fold greater expression than in aerobically grown cells.

In aerobically grown cells, the significantly up-regulated genes, when classified into GO biological processes (section 2.2.16.4.4.2), were predominantly associated with oxidation-reduction processes, reactive nitrogen species metabolic processes and aldehyde catabolic process, which are generally consistent with cell physiology under aerobic growth conditions. In anaerobically grown cells, the significantly up-regulated genes were predominantly associated with various metabolism and biosynthesis processes. In general, most of the differentially expressed genes in aerobically and anaerobically grown *E. coli* REL4536 were associated with macromolecule biosynthesis pathways (**Figure A.3**).

The top 20 genes up-regulated in cells grown in aerobic and anaerobic conditions are listed in **Table 4.5** and **Table 4.6**, respectively. In general, in the aerobically grown cells, the most highly up-regulated genes were involved in the catabolism of propionate *via* the methylcitrate cycle, as well as the catabolism of the amino acid arginine (**Table 4.5**). In anaerobically grown cells, the most highly up-regulated genes were involved in the synthesis of arginine, the synthesis of purines and pyrimidines as well as the anaerobic catabolism of glycerol (**Table 4.6**). The up-regulation of genes involved in arginine catabolism and synthesis under aerobic and anaerobic conditions, respectively, is of interest given that arginine is vital for protein synthesis in *E. coli*. The arginine succinyltransferase (AST) pathway, encoded by the *astCADBE* operon, is the major pathway for arginine catabolism in aerobically grown *E. coli* (277). Regulated by the transcription factors RpoN, RpoS and ArgR, arginine catabolism is usually induced in response to starvation, where proteins are degraded and arginine is used as a source of both carbon and nitrogen in order to obtain glutamate for cell growth and maintenance, with the production of succinate and ammonia as by-products (277). Arginine biosynthesis in *E. coli*, on the other hand, is induced in response to nutrient limitation, sub-optimal growth conditions or cellular stress. Arginine biosynthesis involves a pathway consisting of several operons, including the *argCB* operon, under the negative control of transcription factor ArgR (278). Arginine biosynthesis has previously been shown to be stationary phase dependent (278); at stationary phase, many stationary phase proteins need to be synthesised and so intracellular levels of arginine may be depleted during protein synthesis. As a consequence, *argCBH* expression may be increased as a way to satisfy the arginine demand in cells (278). Arginine is also utilised for the biosynthesis of polyamines like putrescine and spermidine, which are required for nucleic acid and protein biosynthesis and may also prevent DNA damage in response to acidic stress (278). It is not immediately apparent why there are such differences in arginine utilisation between the two environments. However, although RNA was extracted from both aerobic and anaerobic cultures three hours after they had entered stationary phase, the growth rate in anaerobic conditions was much faster, with cultures reaching stationary phase in six hours, whereas under anaerobic conditions, it took three times as long. Thus the physiological states and intracellular stresses of aerobic and anaerobically grown cells three hours after stationary phase, were likely to have differed. Perhaps the increased utilisation of arginine under aerobic conditions is in response to starvation at stationary phase while the increased synthesis of arginine under anaerobic conditions is in response to a requirement for nucleic acid, protein or polyamine synthesis.

Table 4.5. List of the 20 most up-regulated genes in aerobically grown REL4536.

Gene	Product	Function	Fold change[†]	*p*-adj[*]
Up-regulated in aerobic cells				
osmB	Lipoprotein	Osmotically inducuble lipoprotein	44.79	3.05×10^{-236}
prpB	2-methylisocitrate lyase	Catalyses the formation of pyruvate and succinate in the methylcitrate cycle	33.58	1.49×10^{-103}
yjbJ	Putative stress-response protein	Unknown function as part of the sigma S regulon	23.86	6.17×10^{-137}
yqaE	Hypothetical protein	Potential cation transporter	19.42	1.91×10^{-67}
ytjA	Hypothetical protein	Potential membrane protein	18.24	3.89×10^{-100}
bssS	Biofilm formation protein	Regulator of biofilm formation	17.87	4.77×10^{-128}
prpC	Methylcitrate synthase	Catalyses the synthesis of 2-methylcitrate in the methylcitrate cycle	16.21	3.75×10^{-62}
astC	Bifunctional succinylornithine transaminase and acetylornithine transaminase	Catalyses the transamination of 2-N-succinylornithine and alpha-ketoglutarate in arginine degradation II (AST pathway)	16.11	4.78×10^{-64}
astB	Succinylarginine dihydrolase	Catalyses the hydrolysis of 2-N-succinylarginine in arginine degradation II (AST pathway)	16.09	7.70×10^{-58}
yejG	Hypothetical protein	Unknown function	15.54	6.34×10^{-61}
astD	Succinylglutamic semialdehyde dehydrogenase	Aldehyde dehydrogenase involved in arginine degradation II (AST pathway)	15.52	4.33×10^{-56}
hns	Global transcriptional regulator	DNA binding protein involved in stress response, gene regulation and genome organisation	15.49	1.06×10^{-70}
yeaQ	Putative inner membrane protein	Unknown function	15.39	9.79×10^{-184}
ybgS	Hypothetical protein	Unknown function	15.29	5.41×10^{-78}
ymgE	Putative inner membrane protein	Unknown function	15.13	1.73×10^{-57}
yjfY	Hypothetical protein	Unknown function	15.06	5.23×10^{-135}
prpD	2-methylcitrate dehydratase	Catalyses the dehydration of 2-methylcitrate in the methylcitrate cycle	14.28	6.28×10^{-63}
yccJ	Putative protein	Unknown function	13.87	1.00×10^{-133}
astA	Arginine succinyltransferase	Catalyses arginine hyrolysis in arginine degradation II	13.77	4.56×10^{-47}
bolA	Transcriptional regulator	Transcriptional regulator of morphogenetic pathway induced in response to oxidative stress, acid stress, heat shock, osmotic shock, and carbon-starvation stress	13.51	9.33×10^{-57}

[†] Fold change is calculated as the number of aerobic reads over anaerobic reads. [*] To identify significant expression, a *p*-adj value < 0.05 was used.

Table 4.6. List of the 20 most up-regulated genes in anaerobically grown REL4536.

Gene	Product	Function	Fold change[†]	*p*-adj[*]
Up-regulated in anaerobic cells				
yhaO	Putative transporter	Putative amino acid transporter	51.85	1.55×10^{-223}
artJ	Arginine transporter subunit	L-arginine ABC transporter	43.29	3.29×10^{-128}
zraP	Zinc resistance protein	Involved in zinc, cobalt and cadmium homeostasis	26.86	3.45×10^{-56}
allD	Ureidoglycolate dehydrogenase	Anaerobic catalysis of ureidoglycolate in allantoin degradation.	22.68	8.51×10^{-53}
yhaM	Hypothetical protein	Unknown function	20.68	3.21×10^{-140}
glpA	Sn-glycerol-3-phosphate dehydrogenase	Anaerobic dehydrogenase that catalyses the conversion of glycerol 3-phosphate to dihydroxyacetone	16.45	1.45×10^{-66}
argI	Ornithine carbamoyltransferase subunit	Catalyses the formation of L-citrulline in arginine biosynthesis pathway	15.95	1.08×10^{-60}
argB	Acetylglutamate kinase	Catalyses the phosphorylation of N-acetyl- L-glutamate in arginine biosynthesis pathway	15.31	3.44×10^{-54}
glpB	Glycerol-3-phosphate dehydrogenase	Anaerobic dehydrogenase that catalyses the formation of dihydroxyacetone from glycerol 3-phosphate	13.99	3.32×10^{-60}
argA	N-acetylglutamate synthase	Catalyses acetylglutamate in arginine biosynthesis pathway	13.36	1.82×10^{-68}
ybiA	Swarming motility protein	Deaminase in pyrimidine nucleotide biosynthetic pathway	12.31	4.36×10^{-69}
ilvM	Acetolactate synthase 2 regulatory subunit	Involved in valine biosynthesis pathway	12.12	5.04×10^{-40}
argH	Argininosuccinate lyase	Catalyses the formation of arginine from (N-L-arginino) succinate in arginine biosynthesis pathway	11.77	1.13×10^{-53}
argC	N-acetyl-gamma-glutamyl-phosphate	Catalyses the reduction of N-acetyl-5-glutamyl phosphate in arginine biosynthesis pathway	11.70	7.80×10^{-96}
yehU	Putative sensory kinase	Two component signal tranduction system	11.52	1.08×10^{-77}
nrdH	Glutaredoxin-like protein	Functions as a thioredoxin	11.46	3.67×10^{-30}
yhaL	Hypothetical protein	Unknown function	11.15	5.59×10^{-35}
purK	N5-carboxyaminoimidazole ribonucleotide synthase	Decarboxylase in purine nucleotide biosynthetic pathway	10.74	2.66×10^{-24}
ipk	4-diphosphocytidyl-2-C-methylerythritol kinase	Enzyme in the the methylerythritol phosphate pathway	10.19	6.74×10^{-58}
yjiL	Putative ATPase	Unknown function	10.18	2.42×10^{-40}

[†] Fold change is calculated as the number of anaerobic reads over aerobic reads. [*] To identify significant expression, a *p*-adj value < 0.05 was used

4.2.3.1.1 Expression of DNA repair and replication genes

Different types of DNA lesions and mutations are repaired by different repair pathways. It was observed that different mutation types arose under aerobic and anaerobic growth conditions (section 4.2.1). Thus, to determine if different repair systems were correspondingly operating under aerobic and anaerobic growth conditions, expression data for all 215 genes known to be involved in DNA repair and replication were analysed (**Table A.2**). In total, the majority (137 genes) were significant differentially expressed (*p*-adj < 0.05) between the two environments (**Figure 4.14**). Of these, 32 were up-regulated in aerobically grown cells (**Table A.2**) while 105 were up-regulated in anaerobically grown cells (**Table A.2**).

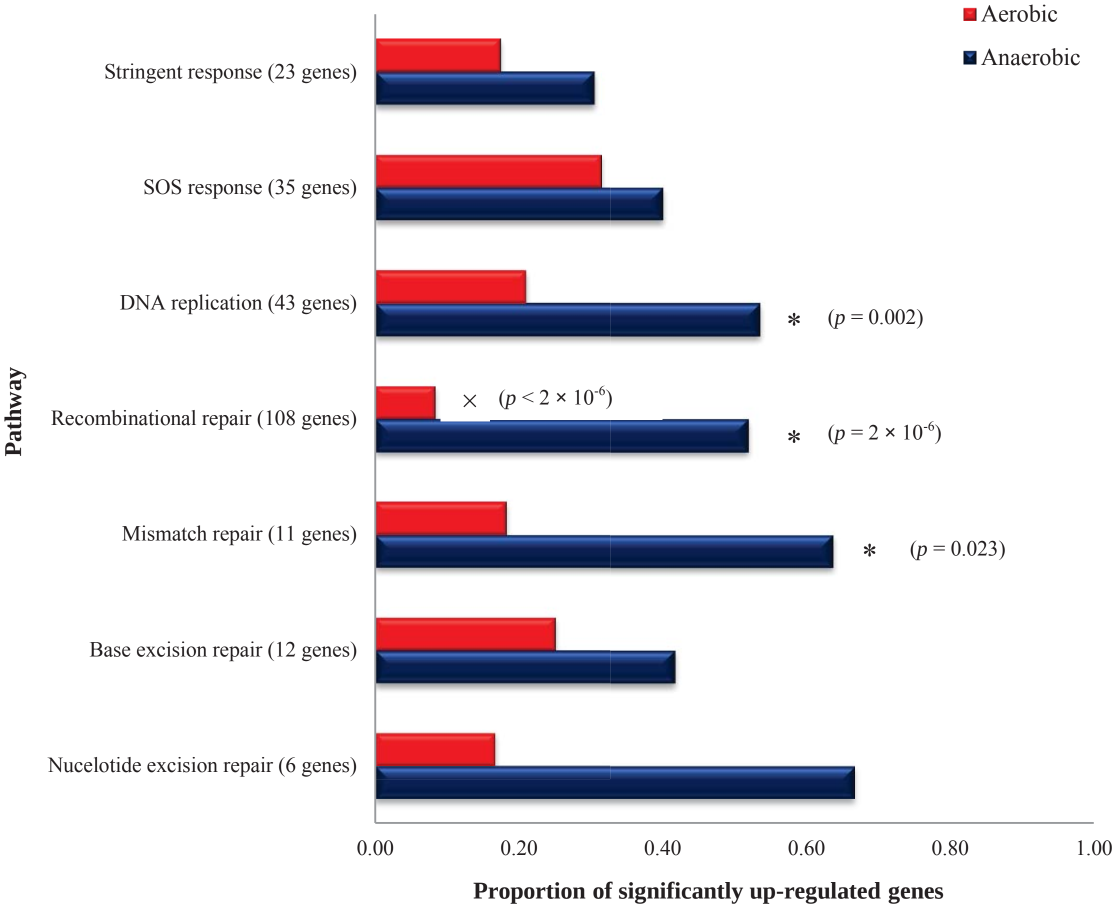

Figure 4.14. Significantly up-regulated genes, as a proportion of genes in the pathway, known to be involved in DNA repair and replication, classified by pathway. To identify significant expression, a *p*-adj value < 0.05 was used. Asterisk denotes a significant enrichment of the gene list under anaerobic conditions and × denotes a significant under-representation of the gene list under aerobic conditions by Fisher's exact test ($p < 0.05$).

While transcripts for the DNA replication and repair pathways were relatively more abundant under anaerobic conditions (**Figure 4.14**), it noted that the total mRNA content per cell was not measured. Moreover, although all samples were taken from cultures at similar ODs, the anaerobic culture RNA yields were generally much lower than those from aerobic cultures. This may be due to a technical issue, or may reflect a lower amount of RNA within the anaerobic cultures per cell. Thus, it is possible that the relatively greater abundance of DNA replication and repair transcripts under anaerobic growth conditions is an artefact of its abundance relative to total mRNA per cell.

As mentioned in section 4.2.3, differences in the growth conditions used for obtaining the transcriptome data and the mutation spectra limit the degree of interpretation of the mutation spectra with regard to expression data. In particular, the datasets were based on different phases of the growth curve (stationary phase for transcriptome data and log phase for the MA lineages) and different types of media [liquid DM media supplemented with 25 µg/mL glucose (section 2.1.7.2) for transcriptome cultures and DM agar supplemented with 200 ng/mL glucose (section 2.1.7.3) for MA lineages]. Despite these minor differences, if pathways are differentially active between the two environments, and are independent of growth phase (e.g. log vs stationary phase), it may be possible to link gene expression with the observed mutation spectra.

4.2.3.1.1.1 DNA replication

Of the 43 genes known to be involved in DNA replication (section 1.4.1), 23 were significantly up-regulated under anaerobic conditions while nine were up-regulated under aerobic conditions (**Table A.2**). It is unclear why some genes within pathways were more highly expressed under anaerobic growth conditions, while others were more highly expressed under aerobic growth conditions. For examples, these trends do not strictly correlate with operon structures. In addition, such variation in gene expression within a pathway was also commonly observed in the other DNA repair and replication pathways. It is possible that post-transcriptional or regulatory mechanisms are responsible for these observations.

As RNA was extracted from both aerobic and anaerobic cultures after they had been at stationary phase for three hours, differential expression of genes involved in replication was initially unexpected. However, during stationary phase, cells may undergo a process known as reductive division, which is where DNA replication, but not cell division, continues after entry into stationary phase (279). As cells grew more slowly under anaerobic conditions than aerobic, it is possible that the anaerobically grown cells were still undergoing reductive division after three hours, while the aerobically grown cells were not. For example, Pol III, comprised of proteins encoded by nine different genes, is presumed to be the main, highly accurate, replicative polymerase in *E. coli* (5, 71). In this study, five genes encoding Pol III subunits (*dnaX, holA, holB, holD* and *holE)* were significantly up-regulated under anaerobic conditions (**Table A.2**). The increased transcription of subunits of the replicative polymerase suggests that some of the cells in anaerobic cultures were possibly still replicating when RNA was extracted. Additionally, of the error prone polymerases involved in TLS (section 1.4.1.1) only *polB*, encoding DNA Pol II, was significantly up-regulated under anaerobic conditions (**Table A.2**).

4.2.3.1.1.2 Recombinational repair

The pathway that displayed the greatest difference in activity between the two environments was the recombinational repair pathway (section 1.4.7) (**Figure 4.14**). Apart from genes encoding IS elements (discussed later in section 4.2.3.1.1.2.1), 15 genes were significantly up-regulated under anaerobic conditions while five were up-regulated under aerobic conditions (**Table A.2**). In *E. coli*, there are two recombinational repair pathways; the RecBCD pathway, which can repair DSBs, and the RecFOR pathway, which can repair SSBs (5). The RecFOR pathway has also been demonstrated to repair DSBs when the RecBCD pathway is inactive (104). Overall, in this study, more genes involved in the recombination pathways were more highly expressed under anaerobic conditions (**Table A.2**). In the MA study, GCRs were found to be significantly more prevalent under anaerobic conditions than under aerobic ones (**Figure 4.6**). Thus, it is reasonable to assume that the increased expression of genes involved in recombinational repair under anaerobic conditions may be in response to the higher rate of GCRs.

4.2.3.1.1.2.1 IS element genes involved in transposition

As reported earlier in the chapter, IS element transposition was higher in anaerobically grown cells than in aerobically grown ones (**Figure 4.7**). To determine if increased transposition rates were due to greater activities of the IS element genes involved in transposition, the expression values of all 83 genes in REL4536 known to be located within IS elements were analysed and are listed in **Table A.2**. In total, 45 genes were significantly differentially expressed (*p-adj* < 0.05) between the two environments, with 41 genes being up-regulated in anaerobically grown cells.

The IS*150* element displayed increased transposition under anaerobic conditions (**Figure 4.7**) and all of the genes, other than *insK*-3, coding for IS*150* proteins, were significantly up-regulated in anaerobically grown cells (**Table A.2**). While IS*186* had an increased transposition rate under aerobic conditions (**Figure 4.7**), none of the genes coding for IS*186* proteins were significantly up-regulated in aerobically grown cells (**Table A.2**). Curiously, the majority of the anaerobically up-regulated genes were from IS*1* elements, which had mutation rates that were not statistically significantly different between the aerobically and anaerobically grown cells (**Figure 4.7**). As greater gene expression at stationary phase did not correlate with greater IS element transposition in the MA study, these results suggest that there may be differences between gene expression at stationary phase and the activity of IS elements at various stages of the growth cycle. Therefore, while it is possible that increased IS*150* transposition in anaerobically grown cells was due to increased expression of the genes involved in transposition, there may also be other contributing factors.

4.2.3.1.1.3 The NER pathway

Another pathway that displayed differential activity between the two environments was the NER pathway (section 1.4.3) (**Figure 4.14**). Of the six genes known to be involved in the pathway, one and four were significantly up-regulated under aerobic and anaerobic conditions, respectively (**Table A.2**). Several scenarios have been identified where the NER pathway has been implicated. These include the repair of bulky lesions that block DNA replication forks or cause a distortion in the DNA structure (84). GCRs can result from blocked replication forks and were found to be more prevalent in anaerobically grown cells (**Figure 4.6**). In addition, transversions cause changes in the structure of the DNA, and were also more prevalent in anaerobically grown cells (section 4.2.1.1.2).

Furthermore, the NER pathway may contribute in the repair of acid-induced DNA damage, as demonstrated in a study of *Streptococcus mutans* (280), and anaerobic fermentation produces an acidic environment through the production of acidic fermentation end products. Therefore, it is possible that the increased expression of genes involved in NER under anaerobic conditions may be in response to the higher rate of GCRs, transversions and anaerobic fermentation, or combinations thereof.

4.2.3.1.1.4 The MMR pathway

The MMR pathway (section 1.4.4) also displayed differential activity between the two environments (**Figure 4.14**). Of the 11 genes known to be involved in the MMR pathway, seven were significantly up-regulated under anaerobic conditions while two were up-regulated under aerobic conditions (**Table A.2**). The MMR pathway is the main pathway responsible for maintaining DNA integrity during replication, as most mismatches (i.e. incorrect base-pairing) are the result of replication errors that have not been repaired by DNA polymerases (5). Therefore, it is possible that the increased activity of the MMR pathway under anaerobic conditions in the transcriptome analysis is a consequence of reductive division still occurring under anaerobic conditions (section 4.2.3.1.1.1).

Other scenarios where the MMR pathway has been implicated have also been identified. These include the repair of slippage events (90), the recognition of recombination heteroduplexes that contain mismatches (90) or the inhibition of homologous recombination (93). Slippage events per generation were twice as frequent in anaerobically grown cells than in aerobically grown ones (**Figure 4.5a**). Additionally, GCRs, mediated *via* homologous recombination between repeated sequences, were significantly more prevalent in anaerobically grown cells (**Figure 4.6**). Therefore, it is possible that the increased expression of genes involved in MMR under anaerobic conditions may be in response to the higher rate of indels, GCRs and replication activity, or combinations thereof.

4.2.3.1.1.5 The BER pathway

The BER pathway (section 1.4.2) did not display a large degree of differential activity between the two environments (**Figure 4.14**). Of the 12 genes known to be involved in the BER pathway, five were significantly up-regulated under anaerobic conditions while three were up-regulated under aerobic conditions (**Table A.2**). As the BER pathway is considered to be the main repair pathway for ROS-induced DNA damage in *E. coli* (5), these findings were unexpected and of particular interest. One of the most extensively studied and frequently encountered ROS-induced lesions is the GO lesion (24). GO lesions typically result in higher G → T, C → A, T → G and A → C transversions. *E. coli* has three BER pathway enzymes that repair either the GO lesion itself or any GO lesion induced mutations (78, 79). Even though there are limitations in interpreting the mutation spectra in terms of gene expression data (section 4.2.3.1.1), overall, the genes associated with the prevention and repair of each mutation type did not correlate with expectations based on the observed mutation spectra. For example, formamidopyrimidine DNA glycosylase (or MutM) is responsible for preventing G → T and C → A transversions (78, 79). In the MA study, G → T transversions were more abundant in aerobically grown cells while C → A transversions were more abundant in anaerobically grown cells (**Figure 4.2**). However, *MutM* was up-regulated by more than two-fold under anaerobic conditions (**Table A.2**). Therefore, in this situation, maybe G → T transversions were efficiently repaired in anaerobically grown cells due to increased *MutM* expression, resulting in lower G → T transversion rates. Additionally, it is possible that the C → A transversions in anaerobically grown cells arose *via* different pathways and so, were not efficiently repaired. As another example, adenine DNA glycosylase (or MutY) is responsible for preventing A → C and T → G transversions (78, 79). In the MA study, A → C and T → G transversions were abundant in anaerobically grown cells (**Figure 4.2**); and *MutY* was down-regulated by almost two-fold under aerobic conditions (**Table A.2**). In this situation, maybe the A → C and T → G transversions were more efficiently repaired in aerobically grown cells due to increased *MutY* expression, resulting in lower mutation rates. Though not involved in the repair of the GO lesion, uracil-DNA glycosylase, encoded by *ung*, is another vital enzyme of the BER pathway. Involved in the removal of uracil bases from DNA, *ung* can be induced in response to cytosine deamination. In the MA study, C → T transitions and G → A transitions occurred at per generation rates that were not statistically significantly different between aerobically and anaerobically grown cells; even though per day mutation rates were greater in aerobically grown cells (**Figure 4.2**). However, *ung* was up-regulated by more than two-fold under anaerobic conditions (**Table A.2**). It is postulated that slower growth rates allow for the efficient recognition and repair of C → T and G → A transitions under anaerobic conditions, while aerobically, the faster rates of growth may result in inefficient repair of the same mutations. In this scenario, aerobically grown cells would display higher mutation rates per day. In contrast, mutation rates per generation would be similar between aerobically and anaerobically grown cells, as they spend the same proportion of time per generation as ssDNA during DNA replication, allowing for cytosine deamination. As a result, it is generally difficult to explain the increased expression of genes involved in BER under anaerobic conditions; perhaps these genes can also be simulated in response to DNA damage caused by factors other than ROS.

4.2.3.1.1.6 SOS response

The SOS response system (section 1.4.5) did not display significant differential activity between the two environments (**Figure 4.14**). Maybe at stationary phase, the SOS response is induced under both aerobic and anaerobic conditions as a direct response to the different agents of mutation operating under each environment.

4.2.3.1.1.7 The stringent response

The stringent response system (section 1.4.6) also did not display noteworthy differential activity between the two environments (**Figure 4.14**). As the stringent response is induced in response to nutrient starvation, it is likely that it is induced under both aerobic and anaerobic conditions at stationary phase.

4.2.3.1.2 Genes involved in acid response

It was postulated that fermentation during growth under the anaerobic conditions of this study have resulted in an acidic environment (257), which may have contributed to the observed mutation spectra. To further explore this possibility, gene expression data of genes previously shown to be induced in response to acidic environments were analysed. Studies suggest that while multiple metabolic processes are required for survival under acidic conditions, *E. coli* largely have four acid tolerance mechanisms [reviewed in (281)]. Each mechanism confers varying degrees of tolerance depending on the growth medium, the growth phase and studies have also shown that these mechanisms may be affected by strain-specific differences (281-283). One such mechanism is acid resistance system 1 (AR1), where ATP synthase, also known as F_0F_1 synthase, catalyses the hydrolysis of ATP in cells at stationary phase in minimal glucose medium to generate an electrochemical proton gradient (281, 282, 284). In this study, the expression of many genes encoding proteins that contribute to AR1 were significantly up-regulated under anaerobic conditions, consistent with our understanding of the mechanisms (**Table A.6**).

Studies have also shown that genes encoding proteins involved in DNA repair, cellular stress responses, membrane structure, membrane permeability, osmotic shock and ion transport are also generally induced in response to acidic cell conditions (285, 286). In this study, DNA repair genes were generally up-regulated under anaerobic conditions (section 4.2.3.1.1). Additionally, many of the genes encoding proteins previously shown to contribute to acid resistance were up-regulated under the anaerobic conditions of this study (**Table A.2**). Therefore, altogether, it seems feasible that the high level of mutations observed in the anaerobically grown MA lineages could be a result of acidic conditions generated during anaerobic fermentation.

4.3 Summary

The genome-wide spontaneous mutation rate for *E. coli* was greater in anaerobically grown MA lineages as compared to aerobically grown MA lineages. To determine if the types of mutations that prevailed in the two environments differed, the mutation spectra of 24 aerobic and 24 anaerobic MA lineage genomes were determined and compared. In general, mutation rates per generation for BPSs, indels and GCRs were higher under anaerobic growth conditions than aerobic ones (**Figure 4.1a**), while mutation rates per day for BPSs and indels were higher under aerobic growth conditions than anaerobic ones (**Figure 4.1b**). BPSs were the most prevalent mutation type in both environments with a bias towards G → T transversions in aerobically grown cells, and biases towards C → A, T → G and A → C transversions in anaerobically grown ones (**Figure 4.2**). No significant bias towards indels was detected in either environment (**Figure 4.5**) but GCRs were generally more frequent under anaerobic growth conditions (**Figure 4.6**). Generally, the frequencies of IS deletions and IS insertions underpinned the largest differences in GCR type frequency between the two environments; where there was a propensity for IS*186* and IS*150* transposition in aerobic and anaerobic cells, respectively (**Figure 4.7**). Additionally, when considering mutation rates per generation and per day together, it appears that GCRs and IS insertions, particularly IS*150* and IS*186* transposition, occur independently of DNA replication.

Gene expression data from stationary phase cultures indicates that under anaerobic conditions, there was an overall increased expression of genes involved in repair and replication (section 4.2.3.1.1). In particular, there is evidence that genes involved in GCR repair were more highly expressed under anaerobic conditions (sections 4.2.3.1.1.2 to 4.2.3.1.1.4), consistent with observations that GCRs were more prevalent in anaerobic MA lineages. Expression data for genes involved in DNA repair and acid tolerance mechanisms (section 4.2.3.1.2) suggest that acidic conditions generated during anaerobic fermentation are potentially responsible for the reported increased activity of repair genes.

Chapter Five: The contribution of gross chromosomal rearrangements to adaptive evolution

5.1 Introduction

Repeat sequences, particularly IRs, in a DNA sequence are hotspots for genome instability because of their capacity to fold into secondary structures that can interfere with molecular processes such as transcription and DNA replication (53). In ssDNA, IRs can form hairpin structures while in dsDNA, IRs can form cruciform-like structures. Both forms of secondary structures can impede the action and progression of polymerases (53). To maintain genome integrity in *E. coli*, secondary structures can be cleaved by nucleases, generating DSBs that are then repaired by homologous recombination (section 1.4.7.1) (53, 287). One such nuclease is SbcC (section 1.7.2), which forms a complex with SbcD and generates DSBs during DNA replication by hairpin cleavage. The DSBs can then be repaired by the RecBCD (section 1.4.7.1.1) or RecFOR pathways (section 1.4.7.1.2) (196, 288, 289). Thus, the SbcCD complex is thought to play an important role in maintaining genome stability (53, 104).

Improper repair of DBSs can lead to the generation of GCRs (described in section 1.3.3) (5). GCRs, such as deletions, translocations, inversions and IS element transposition, are of particular interest in evolution studies as they can modify gene expression, gene content and result in the formation of new DNA junctions (34). While GCRs have long been thought to play important roles in evolution, the impact of GCRs on population fitness and rate of adaptation in *E. coli* is not well understood (5). In our MA study of *E. coli*, GCRs were found to occur at different rates in cells grown under aerobic and anaerobic conditions, with rates being 2.6-fold greater under anaerobic conditions (section 4.2.1.3). Additionally, a microarray study of *E. coli* reported that *sbcC* was differentially expressed between aerobic and anaerobic growth conditions, with increased expression under aerobic conditions (10). Therefore, the aims of this project were to use whole genome re-sequencing and experimental evolution techniques to:

- Determine the contribution of SbcC to the occurrence of GCRs in *E. coli*.
- Determine the impact of GCRs on population fitness.

5.2 Results and discussion

5.2.1 Experimental expression of *sbcC*

Before its role in GCR occurrence could be elucidated, it was essential to determine what *sbcC* expression levels under the experimental conditions of this study were. Partridge et al. (2006) used microarrays to study the transcriptional responses of anaerobic *E. coli* chemostat cultures after exposure to oxygen. In their study, cultures of *E. coli* MG1655 were grown in Evans medium supplemented with 30 mM glucose and maintained in an environment comprised of 95% N_2 and 5% CO_2 (10). After 15 min of culture aeration (i.e. exposure to oxygen), *sbcC* expression was significantly 1.4-fold greater than it had been under anaerobic conditions. To investigate the expression of *sbcC* in *E. coli* REL4536 grown under the aerobic and anaerobic conditions of this study, a transcriptome analysis using RNAseq was conducted.

5.2.1.1 Gene expression of *sbcC* as determined using RNAseq

For transcriptome analysis, RNA was extracted from aerobically and anaerobically grown stationary phase *E. coli* REL4536 cultures (section 2.2.10.2), sequenced and analysed as described in section 2.2.16.4. As mentioned previously, RNAseq transcriptional analysis was carried out on three aerobic samples and two anaerobic samples (section 2.2.16.4.1). In total, sequencing runs generated roughly 13 million reads per samples, with approximately 12 million and 2 million reads per sample usable for aerobic and anaerobic data analysis, respectively. Summary statistics for the RNAseq data, as well as differentially expressed genes between aerobic and anaerobic REL4536, have already been discussed in section 4.2.3.1.

Biological variability of the samples was investigated by using principal component analysis (PCA) on the DESeq2-normalised RPKM values of the five samples (**Figure 5.1**). The PCA showed that samples belonging to the same environment (aerobic or anaerobic) clustered together, though there was also some variability amongst the anaerobic samples. Pearson correlation coefficients were determined for DESeq2-normalised RPKM values of the five samples. Correlation coefficients for aerobic and anaerobic samples were 0.98 and 0.98, respectively. Thus, data for all aerobic samples were combined to determine mean aerobic gene expression and the same treatment was applied to the anaerobic samples.

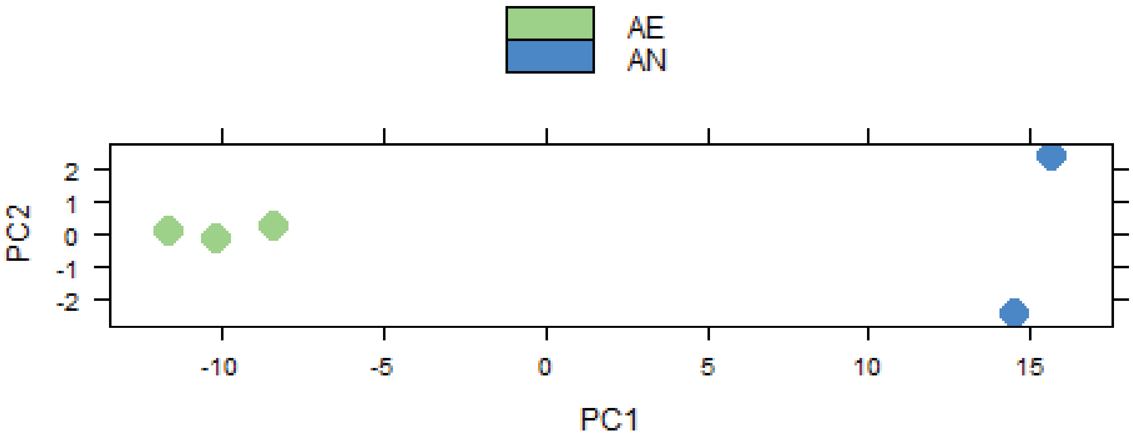

Figure 5.1. Principal component analysis (PCA) of the DESeq2-normalised RPKM values of the three aerobic and two anaerobic samples.

For this dataset, *sbcC* was found to be significantly up-regulated under anaerobic conditions with a fold change value of 2.53 (*p-adj* = 1.64 × 10^{-5}). Based on the knowledge that the SbcCD complex is involved in maintaining genome fidelity, the gene expression data for *sbcC* seems reasonable, as one may expect that there was an increase in expression of the gene under anaerobic conditions in response to the increased occurrence of GCRs under those growth conditions (section 4.2.1.3). However, these results contrast those of Partridge et al. (2006), where *sbcC* was reported to be significantly up-regulated after exposure to oxygen (10). Different *E. coli* strains, growth medium, growth stages and experimental techniques were utilised in the two studies and perhaps these factors account for the different levels of *sbcC* expression between the two studies.

5.2.1.2 Validation using reverse transcription quantitative PCR (RT-qPCR)

A second method of measuring gene expression, reverse transcription quantitative PCR (RT-qPCR), was used to verify *sbcC* expression. In this technique, a target amplicon is quantified by measuring fluorescence levels as the amplicon is amplified during PCR. A fast, sensitive and reproducible process, qPCR is commonly used to investigate gene expression (290). Many factors, including differences in the amounts and quality of starting material and the efficiencies of cDNA synthesis and amplification, can influence the accuracy and reliability of qPCR results (290). Therefore, qPCR data is frequently normalised against reference genes. As reference genes are essentially endogenous controls that display the same level of expression across all experimental conditions, the effects of any technical variation are minimised and different samples can be compared (290). It is now common practice to use at least two genes as internal controls for the relative quantification of gene expression. The $2^{-\Delta\Delta C_Q}$ method (291), first described by Livak and Schmittgen (2001), can then be used to calculate changes in gene expression. For this study, cDNA was synthesised from the RNA of the three aerobic and three anaerobic samples (section 2.2.10.3) and quantified (section 2.2.10.4). All qPCR reactions presented in this chapter were performed in triplicate and so, all values reported are mean values of the three replicates.

5.2.1.2.1 Determination of reference genes

As gene expression can differ across samples and experimental conditions, the suitability of reference genes also needed to be examined. In this study, the candidate reference genes that were chosen for investigation had either previously been used for qPCR studies in other bacteria (292, 293), or were those that were deemed to have equal expression under aerobic and anaerobic conditions *via* RNAseq. In total, five candidate reference genes were selected; the transcriptional regulator encoding *soxR*, the alkaline phosphatase encoding *phoA*, the RNA polymerase sigma factor encoding *rpoD*, the membrane protein encoding *yegO* and the pyrroline-5-carboxylate reductase encoding *proC*.

The three aerobic samples were pooled together in equal amounts, as were the anaerobic samples, and qPCR was performed on 10-fold serial dilutions of the two cDNA pools as detailed in section 2.2.8. Amplification data was analysed using LinRegPCR version 2013.0 (201). Amplification of each reference gene revealed a single peak in the melt curve analysis (data not shown), indicating that the primer pairs were specific to a single sequence. As an assumption of the $2^{-\Delta\Delta C_Q}$ method is that the amplification efficiencies (E) of the reference genes and the gene of interest should be roughly equal (291), the efficiencies of the five candidate reference genes were considered, and found to range from 2.010 to 2.070 (**Table 5.1**). These results indicate that for each gene, the qPCR amplification was 100% efficient (i.e. the PCR product was doubled with each amplification cycle). The correlation coefficients (R^2) of the five candidate reference genes ranged from 0.994 to 1.000, indicating that the standard curves of the five genes had a good fit. Quantification cycle (C_Q) values indicate the number of cycles required to reach the fluorescence threshold and give an indication of the number of gene copies present. In this study, the CQ values of the five reference genes ranged from 26 to 31 (**Table 5.1**).

An ideal reference gene should have a constant level of expression, regardless of the experimental conditions (291). This expression level is represented by N_0, which is proportional to the number of copies of the gene present in the sample. None of the five genes were significantly differentially expressed between aerobic and anaerobic samples ($p > 0.05$) as determined by Student's *t*-test (**Table 5.1**). However, the expression levels of *phoA* varied between aerobic and anaerobic samples, with eight-fold more expression in anaerobic samples (**Table 5.1**). Thus, *phoA* was not a suitable reference gene. Both *soxR* and *yegO* were also considered to not be appropriate reference genes, especially for *yegO* which had R^2 values less than the recommended threshold value of 0.999 (**Table 5.1**). Thus, *rpoD* and *proC*, with almost equivalent expression levels under both growth conditions (**Table 5.1**), and similar amplification efficiencies, were chosen as the reference genes to be utilised in the relative quantitation of *sbcC* expression.

For the five candidate reference genes, there were discrepancies between the fold changes obtained by RNAseq and RT-qPCR, with the latter technique detecting more copies of all genes across anaerobic samples (**Table 5.1**); this is discussed further in section 5.2.1.3. Potentially, the exclusion of the third anaerobic sample from RNAseq analysis, but inclusion in the RT-qPCR analysis may account for some of these differences. Nonetheless, it is difficult to explain the magnitude of the difference in expression for *phoA* by RT-qPCR (**Table 5.1**).

Table 5.1. Summary of RT-qPCR amplification of reference genes for cDNA derived from aerobic and anaerobically grown REL4536.

		soxR	*phoA*	*rpoD*	*yegO*	*proC*
RT-qPCR						
Aerobic						
	CQ	28	31	26	30	28
	N_0	5.07×10^{-9}	1.88×10^{-9}	9.69×10^{-8}	1.08×10^{-9}	2.45×10^{-8}
	N_0 SEM	1.04×10^{-9}	4.04×10^{-10}	3.02×10^{-8}	5.60×10^{-11}	5.50×10^{-9}
	R^2	0.999	0.999	1.000	0.994	0.999
Anaerobic						
	CQ	27	30	26	30	28
	N_0	9.25×10^{-9}	1.52×10^{-8}	1.15×10^{-7}	2.05×10^{-9}	3.23×10^{-8}
	N_0 SEM	2.85×10^{-9}	1.38×10^{-8}	4.51×10^{-8}	1.25×10^{-9}	1.25×10^{-8}
	R^2	0.999	0.999	0.999	0.997	1.000
Overall E ± SEM		2.07 ± 0.02	2.05 ± 0.01	2.05 ± 0.02	2.01 ± 0.02	2.03 ± 0.02
Fold change[†]		1.83	8.10	1.19	1.90	1.32
p		0.240	0.388	0.749	0.517	0.598
RNAseq						
Fold change[‡]		-1.09	1.04	-1.06	-1.05	-1.01
p-adj		0.300	0.750	0.960	0.450	0.290

[†] Fold change for RT-qPCR data is calculated as the anaerobic N_0 value over aerobic N_0 value.

[‡] For RNAseq, fold change is calculated as the anaerobic value over aerobic value and negative values indicates up-regulation in aerobic conditions.

CQ represents quantification cycle values, R^2 represents correlation coefficients and E represents amplification efficiencies with standard error of the mean (SEM) calculations. N_0 is proportional to the number of copies of the gene present in the sample with SEM calculations.

5.2.1.2.2 *sbcC* gene expression verification by RT-qPCR

To study *sbcC* expression, equal amounts of cDNA from the three aerobic and three anaerobic samples were amplified by qPCR as described in section 2.2.8 and analysed using LinRegPCR version 2013.0 (201). In order to use the $2^{-\Delta\Delta C_Q}$ method (291) for the relative quantification of *sbcC* expression, *rpoD* and *proC* were also amplified by qPCR. Amplification of *sbcC* revealed the presence of a single peak in the melt curve analysis (data not shown), indicating that *sbcC* amplification was specific. While the amplification efficiencies of all three genes fell within range of each other, the N_0 of anaerobic samples was almost two- and three- fold greater for the *rpoD* and *proC* genes, respectively, than the N_0 of aerobic samples. Average values of two independent runs are summarised in **Table 5.2**. To use the $2^{-\Delta\Delta C_Q}$ method, the reference genes need to be equally expressed under all of the experimental conditions of the assay (291). As these conditions were violated in this assay, the $2^{-\Delta\Delta C_Q}$ method was considered to not be an appropriate way to quantify the relative change in *sbcC* gene expression under aerobic and anaerobic conditions.

Table 5.2. Summary of RT-qPCR amplification of *sbcC* for cDNA derived from aerobic and anaerobically grown REL4536.

		rpoD	*proC*	*sbcC*
Aerobic				
	C_Q	22	25	28
	R^2	1.000	0.999	0.997
Anaerobic				
	C_Q	21	23	25
	R^2	0.999	0.999	0.999
E ± SEM		2.03 ± 0.01	2.00 ± 0.02	2.00 ± 0.02
N_0 fold change†		1.95	3.06	19.51

†Fold change for RT-qPCR data is calculated as the anaerobic N_0 value over aerobic N_0 value, where N_0 is proportional to the number of copies of the gene present in the sample.

C_Q represents quantification cycle values, R^2 represents correlation coefficients and E represents amplification efficiencies with standard error of the mean (SEM) calculations.

Therefore, relative quantification of *sbcC* expression was achieved by normalisation against the N_0 of both reference genes, as detailed in **Table 5.3**. Using this method, *sbcC* expression was greater in anaerobically grown samples, however this difference in expression was not significant ($p > 0.05$, Student's *t*-test).

Table 5.3. Quantification of *sbcC* expression using *rpoD* and *proC* expression values.

	Aerobic 1 N_0	Aerobic 2 N_0	Aerobic 3 N_0	Anaerobic 1 N_0	Anaerobic 2 N_0	Anaerobic 3 N_0
rpoD	5.00×10^{-6}	1.11×10^{-6}	1.38×10^{-6}	5.49×10^{-6}	2.96×10^{-6}	6.16×10^{-6}
proC	6.16×10^{-7}	1.89×10^{-7}	4.63×10^{-7}	8.58×10^{-7}	3.72×10^{-7}	2.65×10^{-6}
sbcC	6.23×10^{-8}	2.92×10^{-8}	4.51×10^{-8}	3.17×10^{-7}	5.19×10^{-8}	2.30×10^{-6}
sbcC/*rpoD*	0.012	0.026	0.033	0.058	0.018	0.373
sbcC/*proC*	0.101	0.154	0.097	0.369	0.140	0.868

	Aerobic mean N_0	Anaerobic mean N_0	Fold change[†]	*p*
sbcC/*rpoD*	0.024	0.150	6.3	0.327
sbcC/*proC*	0.118	0.459	3.9	0.189

[†] Fold change for RT-qPCR data is calculated as the anaerobic N_0 value over aerobic N_0 value, where N_0 is proportional to the number of copies of the gene present in the sample.

5.2.1.3 Comparison of gene expression data

Even though studies have shown that the two methods can be well correlated (106), in this study, there was weak correlation between the expression values determined by RNAseq and RT-qPCR (**Figure 5.2**). While there were differences in the magnitude of fold changes for the six genes, with *phoA* displaying the largest difference, all genes analysed by RT-qPCR displayed increased transcription under anaerobic conditions as opposed to values determined by RNAseq.

As both RNAseq and RT-qPCR use different approaches in assessing gene expression, it stands to reason that the different approaches contribute to some of the variation observed between the two datasets. In RNAseq analysis, reads are mapped to a reference genome or transcriptome and then normalised per million mapped reads to give the RPKM values. When using DESeq2, differences in the sequencing depths of the samples are taken into consideration, and the RPKM values are then further normalised. DESeq2 has an underlying assumption that most genes are not differentially expressed, as this allows for the calculation of the geometric mean for each gene across all samples. The geometric mean is then used to normalise the RPKM values. Results from DESeq2, such as the PCA plot (**Figure 5.1**), indicate that there were many differences between the aerobic and anaerobic samples. Thus, it is possible that normalisation may have been slightly biased; this could possibly account for the discrepancies between the RNAseq and RT-qPCR data. The usage of spike in controls in any further experiments would greatly alleviate this concern and allow for proper normalisation.

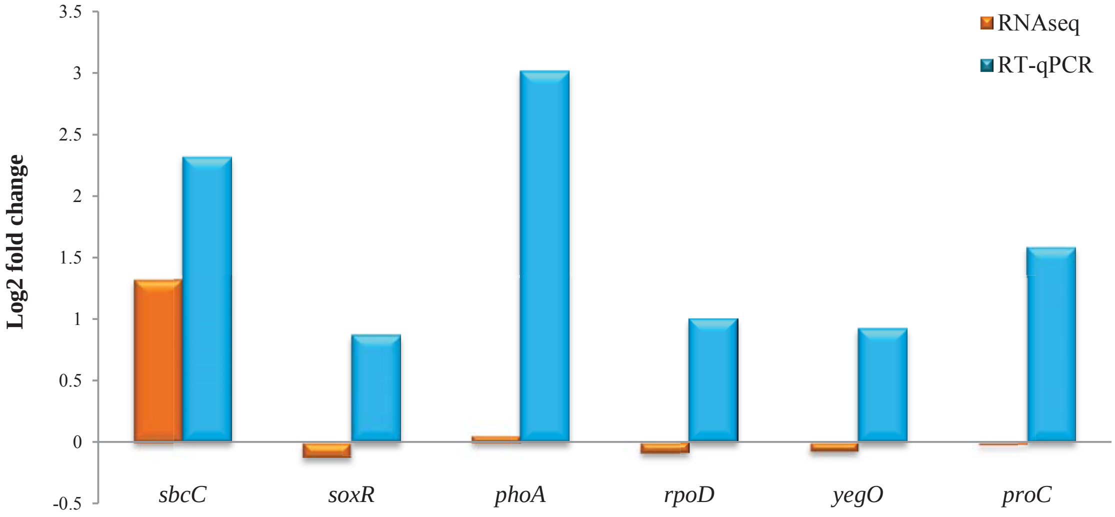

Figure 5.2. Comparison of relative log2 fold changes observed for six genes quantified by both RNAseq and RT-qPCR.

Additionally, the reference genes that were chosen for RT-qPCR analysis were ultimately not suitable. Their expression values differed between runs (**Table 5.1** and **Table 5.2**), suggesting that they were not as stable as initial experiments had indicated. While the $2^{-\Delta\Delta Cq}$ method was eventually not utilised, *sbcC* expression values were still normalised to values from both *rpoD* and *phoA* and this normalisation may have introduced some other biases. However, it is also possible that the inclusion of a third anaerobic sample (Anaerobic 2) for RT-qPCR contributes to some of the variation between RT-qPCR and RNAseq expression values. Results obtained from RT-qPCR analysis of Anaerobic 3 differ from those obtained from Anaerobic 1 and Anaerobic 2

(**Table 5.3**), suggesting that perhaps Anaerobic 3 might be an unusual sample. Certainly, more replicates for both RNAseq and RT-qPCR analysis would provide a more thorough and robust understanding of the biological relevance of differential gene expression under aerobic and anaerobic conditions. Nonetheless, most importantly, *sbcC* expression was found to be more prevalent under the anaerobic conditions of this study by both RNAseq and RT-qPCR analysis.

5.2.2 Contribution of *sbcC* to genome fidelity and fitness

Adaptive evolution (section 1.7) is the evolutionary process where organisms accumulate genetic changes that allow them to better adapt to their environment (164). In evolutionary studies, the relative fitness of evolved populations, measured by competitive fitness assays, can be used to measure adaptation. To determine the contribution of SbcC to the occurrence of GCRs, and the effects of those GCRs on population fitness in *E. coli*, experimental evolution techniques and whole genome re-sequencing were utilised. Essentially, long-term adaptive lineages were established to enable the adaptation of *E. coli* strains to anaerobic growth conditions. After lineages had undergone 1,000 generations of experimental evolution, the genomes of randomly-selected clones from the evolved populations were sequenced. Additionally, relative fitness was determined for a selection of those populations for whom clonal sequence data was available.

In order to determine the effect of *sbcC* on the occurrence of GCRs, comparison between *E. coli* REL4536 adaptive experimental evolutionary lineages containing an intact and a disrupted *sbcC* gene were employed. A parallel study examining the adaptation of *E. coli* REL4536 to anaerobic and fluctuating growth conditions for 4,000 generations by Finn (2015) had recently concluded and served as an ideal control for the *sbcC* disruption mutant, since identical operating procedures and growth conditions were used for both sets of lineages (205). Thus, it was necessary to generate an *E. coli* REL4536 strain with *sbcC* disruption.

5.2.2.1 Generation of *sbcC* mutant

An *E. coli* REL4536 Δ*sbcC::cat* strain was created (**Table 2.1**) by modifying *E. coli* REL4536 by using the λ-Red recombination procedure described in section 2.2.9. Though two homologous recombination steps were required to obtain a scarless deletion of *sbcC*, all attempts at the second homologous recombination step (section 2.2.9.3.2) were unsuccessful. Thus, in this study, the *sbcC* gene in *E. coli* REL4536 was inactivated by the insertion of roughly 1 kb of DNA sequence (the pWRG100 chloramphenicol resistance cassette) within the gene. This mutant stain was termed *E. coli* REL4536 Δ*sbcC::cat* and was used to initiate the adaptive lineages described in section 5.2.2.3.

5.2.2.2 Growth dynamics of *sbcC* mutant under anaerobic environment

Before establishing adaptive lineages of *E. coli* REL4536 Δ*sbcC::cat*, the growth dynamics of the strain under anaerobic conditions (section 2.2.2) needed to be determined. Growth courses were conducted, as described in section 2.2.4, for 1 mL cultures grown in DM25 (**Figure 5.3**). CFU per mL of culture were calculated (section 2.2.4) to determine viable cell counts of the cultures. As can be seen in **Figure 5.3**, a peak cell density 2.33 × 10^7 CFU/mL was reached after 24 h. The cell density then remained largely constant for another 24 h. Therefore, it was concluded that stationary phase in this environment was reached within 24 h and that lineages would be propagated every 24 h. In the parallel study of *E. coli* REL4536 adaptation to the same anaerobic conditions, henceforth referred to as REL4536 AN populations, a peak cell density of 3.50 × 10^7 CFU/mL was reached after 20 h (205). These lineages had also been propagated every 24 h (**Figure 5.3**).

5.2.2.3 Establishment of long-term lineages in anaerobic environment

After the growth dynamics of *E. coli* REL4536 Δ*sbcC::cat* under anaerobic conditions were confirmed, long-term adaptive lineages of *E. coli* REL4536 Δ*sbcC::cat* were established (section 2.2.15.1). Briefly, 14 independent lineages were initiated from an ancestral stock culture derived from a single colony of *E. coli* REL4536 Δ*sbcC::cat*. Ancestral culture identity was confirmed by Gram staining (section 2.2.5) and phage contamination tests (section 2.2.14.2.3). All lineages were evolved for 1,000 generations, or 152 days, with daily sub-culturing (section 2.2.15.2.1) due to time constraints. All evolved lineages were subject to fortnightly contamination tests (section 2.2.15.2.3) and fortnightly storage (section 2.2.15.2.2).

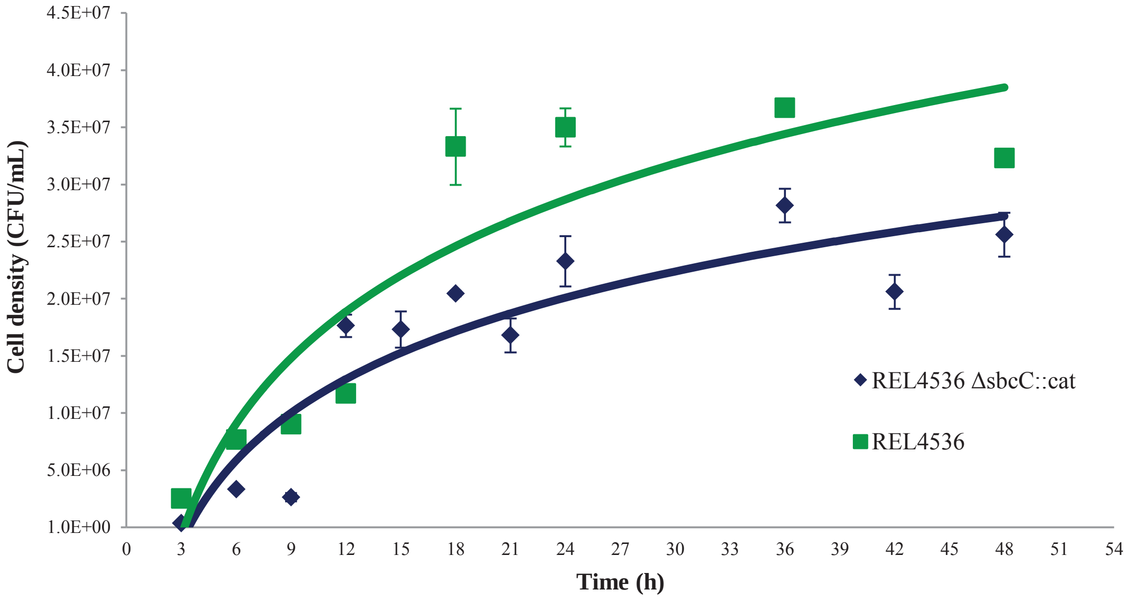

Figure 5.3. Growth dynamics of *E. coli* REL4536 and *E. coli* REL4536 Δ*sbcC::cat* under anaerobic growth conditions. Stationary phase is reached within 24 h for both strains. Data points represent the mean values from three biological replicates, with error bars representing standard error of the mean.

5.2.2.4 Mutation analysis

To identify the genetic changes that arose during the course of lineage adaptation to the anaerobic environment, whole genome re-sequencing was used. After the lineages had been propagated for 1,000 generations, each lineage was single colony streaked sequentially three times over three days to ensure purity. DNA was then extracted (section 2.2.10.1) from cultures derived from a single, randomly chosen colony from each lineage. Genomes of six *E. coli* REL4536 Δ*sbcC::cat* and four *E. coli* REL4536 AN 1,000 generation evolved clones were re-sequenced on an Illumina HiSeq 2000 instrument (section 2.2.16.3.1). To detect mutations, reference-based and *de novo* assembly based mutation detection methods were used, as detailed in section 2.2.16.3.2. Summary statistics of the reference-based mapping and the QUAST evaluated *de novo* assemblies (224) are presented in **Table 5.4**. For each sample, whole genome re-sequencing generated at least 1 million raw sequencing reads. While the genomes were sequenced to differing depths as multiple clones were sequenced on the same lane, there was at least 18-fold sequencing read depth (calculated as described in **Equation 2.7**) for each lineage clone genome.

Table 5.4. Genome re-sequencing reference-based mapping and *de novo* assembly statistics for anaerobic REL4536 *ΔsbcC::cat*-1K and REL4536 AN-1K clones.

Lineage	Total reads	Unmapped reads	Mapped reads	Reads mapped (%)	Sequencing depth (fold coverage)	No. of contigs	No. of contigs >1000 bp	Largest contig (bp)	N50 (bp)	Genome coverage (%)
REL4536 *ΔsbcC::cat*-1K-1	12,574,008	4,666,608	7,907,400	62.89%	156	59	10	1,192,370	701,142	98.66
REL4536 *ΔsbcC::cat*-1K-6	5,908,224	2,333,242	3,574,982	60.51%	70	48	10	1,388,204	667,658	99.46
REL4536 *ΔsbcC::cat*-1K-7	4,077,200	1,394,638	2,682,562	65.79%	53	47	15	1,193,857	670,931	99.23
REL4536 *ΔsbcC::cat*-1K-8	6,709,774	2,289,327	4,420,447	65.88%	87	71	14	783,995	666,064	98.92
REL4536 *ΔsbcC::cat*-1K-12	1,466,466	548,882	917,584	62.57%	18	37	22	701,988	555,278	99.24
REL4536 *ΔsbcC::cat*-1K-14	1,909,980	660,474	1,249,506	65.42%	25	69	22	666,151	273,335	99.55
REL4536 AN-1K-1	5,198,010	1,903,056	3,294,954	63.39%	65	65	12	1,193,917	703,691	99.43
REL4536 AN-1K-3	5,384,606	2,195,529	3,189,077	59.23%	63	48	10	741,515	667,786	99.41
REL4536 AN-1K-4	6,261,088	2,143,722	4,117,366	65.76%	81	52	10	1,194,274	703,911	99.48
REL4536 AN-1K-6	7,557,950	2,353,292	5,204,658	68.86%	103	66	12	742,229	668,741	99.38

Among the clones that were genome re-sequenced, a total of 57 mutations were detected. Using breseq (209) for reference-based mapping to the *E. coli* REL4536 reference genome sequence (section 2.2.16.3.2.1), 10 BPSs, 13 indels and 34 GCRs involving MGEs were detected (**Table 5.5**). Large-scale GCRs (i.e. inversions or translocations as discussed in section 3.2.3.2.2.2) were identified by assembling genomes *de novo* with PCR verification of synteny breakpoints. While two inversions, mediated by the *rhs* genes, were identified in REL4536 Δ*sbcC::cat* clones by the *in silico* analysis, both inversions were not able to be verified by PCR. Hence, these inversions were not experimentally proven, most likely due to inadequate read coverage and instability of assembly programmes over *rhs* junctions (see section 3.2.3.2.2.2 for more). Details of all mutations found in this study are listed in **Table A.7**.

Table 5.5. Frequency of mutations found in the genome of each lineage clone.

Clone	BPS	Indels	GCRs	Total
REL4536 *ΔsbcC::cat*-1K-1	0	2	3	5
REL4536 *ΔsbcC::cat*-1K-6	0	0	3	3
REL4536 *ΔsbcC::cat*-1K-7	0	0	5	5
REL4536 *ΔsbcC::cat*-1K-8	0	1	2	3
REL4536 *ΔsbcC::cat*-1K-12	3	8	6	17
REL4536 *ΔsbcC::cat*-1K-14	2	2	4	8
REL4536 AN-1K-1	0	0	2	2
REL4536 AN-1K-3	1	0	1	2
REL4536 AN-1K-4	3	0	2	5
REL4536 AN-1K-6	1	0	6	7

5.2.2.4.1 Rates of GCRs per generation per genome

The mean mutation rate per genome per generation, calculated as described in section 2.2.14.3, for the REL4536 Δ*sbcC::cat* clones was 1.4-fold greater in magnitude (Mann-Whitney U = 7.0, p = 0.186) than the mean mutation rate calculated for the REL4536 AN clones (**Figure 5.4**). Additionally, inactivation of *sbcC* in the REL4536 Δ*sbcC::cat* clones resulted in the accumulation of more GCRs, and a mean rate of 3.83×10^{-3} GCRs per genome per generation. This value was 1.4-fold greater than the mean GCR mutation rate that was estimated for the REL4536 AN clones (**Figure 5.4**). However, this difference in rates was not statistically significant (Mann-Whitney U = 6.5, p = 0.148).

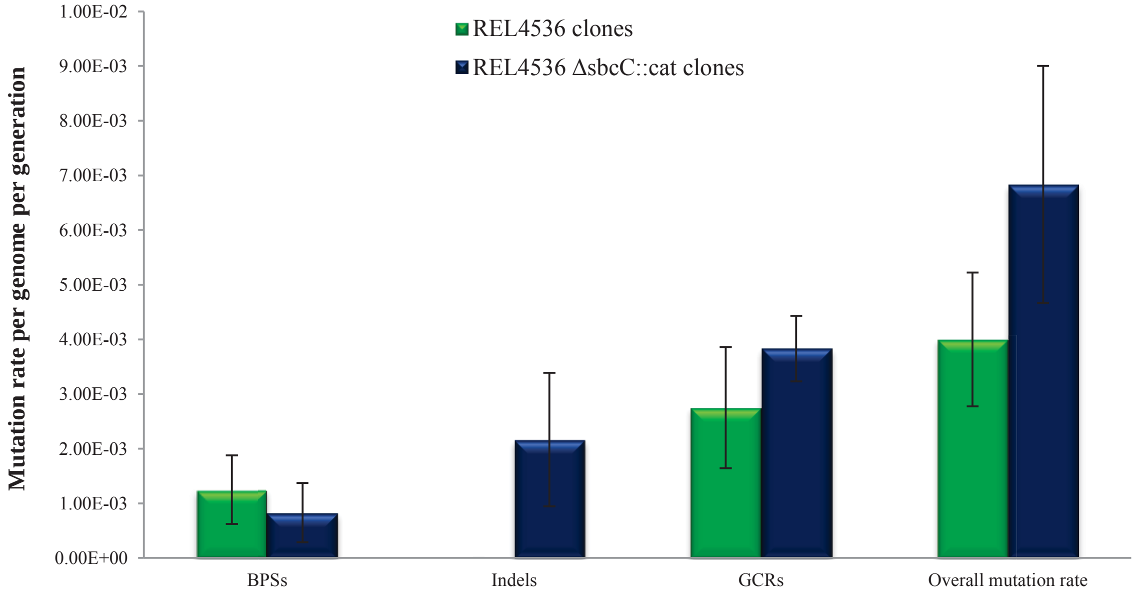

Figure 5.4. Mutation rates for different types of mutations observed amongst the REL4536 Δ*sbcC::cat*-1K and REL4536 AN-1K clones. Shown are mean mutation rates per genome per generation. Error bars represent standard error of the mean.

As can be seen from **Figure 5.4**, the most notable difference between the REL4536 Δ*sbcC::cat*-1K and REL4536 AN-1K clones was the rate at which indels occurred. The disruption of *sbcC* in the REL4536 Δ*sbcC::cat* clones resulted in a mean rate of 2.17 × 10^{-3} indels per genome per generation, while no indels were observed in REL4536 AN clones (Mann-Whitney U = 4.0, p = 0.071). As *sbcC* has demonstrated specificity for DSBs generated at palindromic and repeat sequences, these results were surprising since *sbcC* inactivation was not expected to influence the mutation rate of indels. It is important to note, however, that these results might be slightly biased due to the observation that the majority of indels amongst the REL4536 Δ*sbcC::cat*-1K clones occurring in the REL4536 Δ*sbcC::cat*-1K-12 clone specifically.

As an aside, IS*150* element transposition was particularly increased under the anaerobic conditions of this study for both the REL4536 Δ*sbcC::cat* and REL4536 AN clone genomes, with mean per generation mutation rates of 3.00 × 10^{-3} ± 6.83 × 10^{-4} and

$2.00 \times 10^{-3} \pm 1.00 \times 10^{-3}$, respectively. This phenomenon was also observed in the MA study (section 4.2.1.3.1.1), indicating that IS*150* is indeed particularly active under anaerobic conditions. Additionally, these results suggest that *sbcC* may not be involved in regulating IS*150* mediated GCRs.

5.2.2.4.2 Mutations of interest

Amongst the sequenced lineages, there were instances of the same mutations being detected in independent lineages. Such instances of parallelism suggest that either these mutations confer a selective advantage or serve as a mutation hotspot. Examples of such mutations include BPS, indels and GCRs, as described below.

5.2.2.4.2.1 REL4536 *ΔsbcC::cat*-1K-1 and REL4536 *ΔsbcC::cat*-1K-8

REL4536 *ΔsbcC::cat*-1K-1 and REL4536 *ΔsbcC::cat*-1K-8, for instance, exhibited two instances of parallelism; a roughly 33 kb deletion and a 5 bp indel was detected in clones from both lineages. As these lineages had always been maintained on separate plates during the course of evolution, we are confident that these mutations arose independently in the lineages (**Table A.7**). It seems likely that the 33 kb deletion was mediated *via* homologous recombination of IS*3* elements flanking the region. A total of 30 genes were deleted and are listed in **Table A.8**. Many of the deleted genes encode hypothetical proteins of unknown function and are likely to be associated with cryptic prophage P22 activity. Genes involved in transcription regulation were also deleted. For example, *appY*, which encodes the transcriptional regulator AppY, is induced by the global regulator ArcAB. In turn, ArcAB induces the expression of genes involved in anaerobic energy metabolism, survival at stationary phase and phosphate starvation (294). Another deleted gene of interest is *envY*, which encodes the DNA-binding transcriptional regulator EnvY. EnvY is involved in the regulation of genes encoding cellular envelope proteins that are required for survival at low temperatures and during stationary phase (295). Loss of this region of the genome has previously been reported (296) and was also observed in *E. coli* REL4536 clones evolved under anaerobic conditions for 2,000 generations (205) by Finn (2015). However, the selective advantage conferred by this deletion is not immediately apparent.

The 5 bp indel, found in both the REL4536 *ΔsbcC::cat*-1K-1 and REL4536 *ΔsbcC::cat*-1K-8 sequenced clones, restored the original coding frame of the pseudogene *dcuS*.

DcuS, along with DcuR, is part of a two component regulatory system; DcuS is a dicarboxylate-sensing histidine kinase that reports the external dicarboxylate concentration to DcuR, while DcuR regulates genes involved in the anaerobic fumarate respiratory system (297). Hence, it is possible that this mutation may have reactivated an anaerobic respiratory pathway (297) and subsequently, been selected for. The same mutation was also observed in all sequenced *E. coli* REL4536 clones evolved under anaerobic conditions for 2,000 generations by Finn (2015) (205).

5.2.2.4.2.2 The *adhE* gene

Amongst the 10 sequenced clones, four different non-synonymous mutations in *adhE* were detected (**Table 5.6**). The *adhE* gene encodes AdhE, which can function as an alcohol dehydrogenase and a coenzyme A-dependent acetaldehyde dehydrogenase. Under conditions where fermentation is possible, AdhE can catalyse the reduction of acetyl-CoA to ethanol. Alternatively, during aerobic growth, AdhE can catalyse the oxidation of acetaldehyde to acetyl-CoA (298). In summary, AdhE is involved in the maintenance of the redox balance under anaerobic conditions and previous studies have reported that *E. coli* cells were not able to grow under anaerobic growth conditions without the gene. The same study also observed that additional mutations in the *pta* gene rescued the cells and allowed them to grow (298).

Table 5.6. Details of *adhE* mutations detected in this study.

Lineage	Reference Position*	Change	Mutation type	Amino Acid Change
REL4536 *ΔsbcC::cat*-1K-12	1,439,509	A → G	Transition	Glu (568) Gly
REL4536 *ΔsbcC::cat*-1K-14	1,439,821	A → C	Transversion	Asp (672) Ala
REL4536 AN-1K-3	1,438,030	A → G	Transition	Tyr (75) Cys
REL4536 AN-1K-4	1,439,673	G → A	Transition	Ala (623) Thr
REL4536 AN-1K-6	1,438,030	A → G	Transition	Tyr (75) Cys

*Position of mutation on reference *E. coli* REL4536 genome.

As different mutations in *adhE* were found in independent lineages, with or without an active *sbcC* gene, it is likely that mutations in *adhE* generally confer a selective advantage to growth under anaerobic conditions. Lending support to this theory are findings by Finn (2015), where BPSs in *adhE* were also observed in all sequenced

E. coli REL4536 clones evolved under anaerobic conditions for 2,000 generations (205). All of the BPSs listed in **Table 5.6** resulted in the substitution of an amino acid residue for another, and so, were likely to confer a conformational change of the protein. Thus, it is possible that AdhE protein function was disrupted, which may result in the diversion of the fermentation pathway to increased succinate or acetate production to serve as additional energy sources for growth (299, 300).

5.2.2.4.2.3 IS insertions

Other instances of the same mutations being detected in independent lineages involved IS elements. For example, IS*150* insertion was detected in *cycA* at the same position (4,381,583 in the REL4536 reference genome) in both REL4536 *ΔsbcC::cat*-1K-8 and REL4536 *ΔsbcC::cat*-1K-14 clones (**Table A.7)**. Encoded by *cycA¸* the CycA transporter is involved in the uptake of glycine, serine and alanine. It has previously been demonstrated that cells with inactivated *cycA* exhibited decreased transport of the three amino acids and were resistant to d-cycloserine inhibition (301). While it is possible that these mutations could have derived from cross contamination between lineages, as both REL4536 *ΔsbcC::cat*-1K-8 and REL4536 *ΔsbcC::cat*-1K-14 lineages had been maintained on the same plates, IS*150* insertion was found at exactly the same position in sequenced REL4536 AN clones (for lineages 1, 3, 4 and 6). Therefore, it is probable that this position is a hotspot for IS*150* insertion, which may aid in adaptation to either the anaerobic environment or the 24-well growth vessels.

IS*150* insertion was also detected in *ydiQ* at the same position (974,185 in the REL4536 reference genome) in the REL4536 *ΔsbcC::cat*-1K-6, REL4536 *ΔsbcC::cat*-1K-7 and REL4536 *ΔsbcC::cat*-1K-14 clones (**Table A.7),** suggesting the possibility of this position being another hotspot for IS*150* insertion. YdiQ, encoded for by *ydiQ*, is a putative electron transport flavoprotein subunit that may be involved in the electron transport chain between the fatty acid oxidation and respiration pathways under anaerobic conditions (302). Additionally, IS*150* insertion was also detected in *ydiU* at the same position (963,716 in the REL4536 reference genome) in the REL4536 *ΔsbcC::cat*-1K-1, REL4536 *ΔsbcC::cat*-1K-6, REL4536 *ΔsbcC::cat*-1K-7 and REL4536 *ΔsbcC::cat*-1K-14 clones, suggesting this position may be another hotspot for IS*150* insertion. YdiU, encoded for by *ydiU*, is a conserved protein with unknown function. However, the advantages of inactivating both *ydiQ* and *ydiU via* insertion of IS elements remain unclear.

Another instance of parallelism was detected in clones from REL4536 *ΔsbcC::cat*-1K-7 and REL4536 *ΔsbcC::cat*-1K-14, where an IS*150* element was deleted from the *pflB* gene, resulting in the reactivation of the gene (**Table A.7)**. The *pflB* gene encodes pyruvate formate lyase I, an enzyme involved in the conversion of pyruvate to formate during anaerobic fermentation. Therefore, the presence of a functional *pflB* gene under anaerobic conditions should be beneficial and advantageous under anaerobic growth conditions. Moreover, reactivation of the *pflB* gene was also reported by Finn (2015) for some sequenced REL4536 clones evolved under anaerobic and fluctuating conditions for 2,000 generations (205). Additionally, a C → T transition in the *pflB* gene at the same position (1,766,332 in the REL4536 reference genome) was detected in the REL4536 *ΔsbcC::cat*-1K-14 and REL4536 AN-1K-4 clones. As this mutation was synonymous, and thus did not result in the substitution of amino acids, the beneficial effect of this mutation is not immediately apparent.

5.2.2.5 Fitness assessment of evolved populations.

GCRs have not been studied in exhaustive detail in most experimental evolution studies. As the aim of this study was to determine how GCRs impact adaptive evolution, determining the effect of GCRs on population-level fitness was of interest. While GCRs have long been thought to play important roles in events such as speciation and short-term adaptation to new environments (35-40), it can be difficult to predict their effects on fitness. Even though potential fitness costs can restrict the occurrences of GCRs in populations (33), many studies have found that some GCRs have conferred increased fitness (41-43).

Fitness of evolved strains can be measured by the use of competitive fitness assays. To distinguish between the evolved and ancestral strains, neutral, phenotypic markers are often used. While Lenski et al. use arabinose utilisation and growth on selective media in their fitness assays (181), Finn (2015) found that arabinose-utilising mutants of *E. coli* REL4536 were not selectively neutral under anaerobic conditions (205). Therefore, Finn (2015) isolated spontaneous mutants of REL4536 with resistance to various antibiotics and assessed these mutants for neutrality as described in (205).

Ultimately, Finn (2015) generated a rifampicin resistant spontaneous mutant, Rif^{r}2, that was neutral under both aerobic and anaerobic conditions. A single A → T transversion (at position 4,128,442 in the REL4536 reference genome) in the *rpoB* gene was found to confer rifampicin resistance in the Rif^{r}2 strain (205). This mutation resulted in the substitution of a phenylalanine for an isoleucine and thus, resistance may be due to a lower binding affinity of rifampicin to RNA polymerase. Rifampicin resistance has also been used as a marker for competitive fitness assays under oxygen-limited conditions by Puentes-Téllez et al. (2013) (303).

5.2.2.5.1 Assessment of neutrally marked strains

Finn (2015) tested the neutrality of Rif^{r}2 against the ancestral *E. coli* REL4536 strain under anaerobic conditions and reported a relative fitness value of 1.04 ± 0.05 (205). As the REL4536 *ΔsbcC::cat* strains had a disrupted *sbcC* gene, it was necessary to determine whether Rif^{r}2 could also be used for competitive fitness assays involving REL4536 *ΔsbcC::cat* evolved strains. Thus, the fitness of Rif^{r}2 against the REL4536 *ΔsbcC::cat* ancestral strain was determined as described in section 2.2.13.2. Relative fitness (***ω***) values of Rif^{r}2 against the REL4536 *ΔsbcC::cat* ancestral strain were calculated as the mean of two competitions of four biological replicates each (section 2.2.13.2.1) and a mean fitness value of 1.05 ± 0.05 was obtained. In phenotypic competitive fitness assays, an ideal marked strain should have a fitness value of 1.0. As Rif^{r}2 met these requirements, it was deemed to be an appropriate marker to use for competitive fitness assays involving both evolved REL4536 *ΔsbcC::cat* and REL4536 AN strains. Therefore, all competitive fitness assays presented in this section utilise Rif^{r}2 as the ancestral competitor strain.

5.2.2.5.2 Population fitness

Competitive fitness assays of four REL4536 *ΔsbcC::cat* populations and four REL4536 AN populations, all evolved for 1,000 generations under anaerobic growth conditions were performed as described in section 2.2.13, with an underlying assumption that the randomly-selected clones that were sequenced (section 5.2.2.4) were typical representatives of the total genetic diversity of the populations from which they were isolated. The four REL4536 *ΔsbcC::cat* lineages were chosen for fitness analysis based on the genome sequence data of representative clones (**Table A.7)**. For instance, REL4536 *ΔsbcC::cat*-1K-12 was chosen for analysis as the representative clone from this population had the most mutations and the most GCRs (**Table 5.5**). Both REL4536 *ΔsbcC::cat*-1K-6 and REL4536 *ΔsbcC::cat*-1K-7 were chosen for relative fitness assessment as the representative clones from these populations had only GCRs as mutations (**Table 5.5**). Finally, REL4536 *ΔsbcC::cat*-1K-8 was chosen as it contained only three mutations, amongst which one was the 33 kb deletion (section 5.2.2.4.2.1), and the effect of this deletion on relative fitness was of interest.

The fitness of all populations, relative to the ancestor, increased after 1,000 generations of adaptation (**Figure 5.5**). On average, the mean relative fitness of the four REL4536 *ΔsbcC::cat* populations was 1.78 ± 0.15 after the 1,000 generations. On the other hand, mean relative fitness of the four REL4536 populations was 1.64 ± 0.21. As fitness typically rises rapidly at the start of adaptation experiments (168), especially in response to novel environments, these results were consistent with expectations. Even though the assumption of there being a relationship between the number of GCRs and population fitness is rather simplistic, the association between the two was investigated next, as it is a practical starting point in understanding the dynamics between GCRs and evolution.

Populations of REL4536 *ΔsbcC::cat* displayed varying fitness dynamics that did not display a strong correlation between the number of GCRs and relative fitness (**Figure 5.5**). For example, even though the sequenced REL4536 *ΔsbcC::cat*-1K-12 clone had the most GCRs, the REL4536 *ΔsbcC::cat*-1K-12 population did not have the largest increase in fitness. In fact, the REL4536 *ΔsbcC::cat*-1K-8 population experienced the greatest increase in fitness, with a relative population fitness of 2.17 ± 0.25. This finding was of great interest, as the REL4536 *ΔsbcC::cat*-1K-8 clone had only 3 mutations; the 33 kb deletion, the 5 bp indel and an IS*150* insertion into *cycA* (section 5.2.2.4.2.3).

As fitness assays were conducted at the population level, and genome sequence data was obtained at a clonal level, it is difficult to distinguish which particular mutations were responsible for increases in fitness. Studies have previously shown that gene loss can result in increased fitness, as genes whose functions are not required under the conditions of the experiment are removed, eliminating unnecessary and costly gene expression in the process (304). However, it is unclear what factors drive the gene deletion process. Therefore, it is possible that the 33 kb deletion is the underlying factor behind the fitness increase observed for the REL4536 *ΔsbcC::cat*-1K-8 population as many genes with no assigned function, as well as the *appY* gene, were deleted in the process. Thus, it would be of interest to see if the REL4536 *ΔsbcC::cat*-1K-1 population, where the sequenced clone had the same 33 kb deletion and 5 bp indel, experienced a similar increase of relative fitness after 1,000 generations of evolution. Furthermore, in future, it would be beneficial to conduct fitness assays at the clonal level and determine if similar increases in fitness are observed. As it is possible that the sequenced clones are not typical representatives of the genotypes of the populations from which they were isolated, it would be beneficial to determine how prevalent the genotype of the REL4536 *ΔsbcC::cat*-1K-8 clone is in the population. This could be achieved by the use of qPCR to estimate gene copy numbers amongst evolved lineages. Together, these studies would indicate if the genetic basis of the fitness increase in the REL4536 *ΔsbcC::cat*-1K-8 population could be a pleiotropic effect of the three mutations.

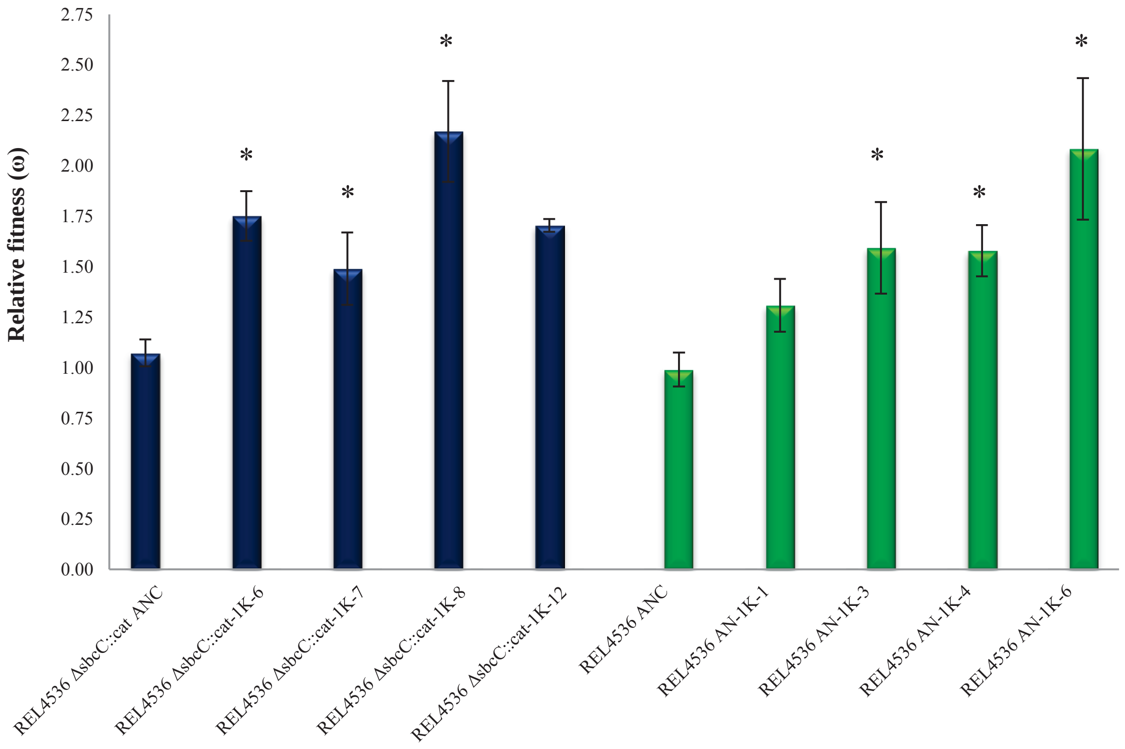

Figure 5.5. Relative fitness of anaerobic populations to an ancestral strain after 1,000 generations of evolution. ANC denotes the ancestral strain. Displayed are the mean relative fitness values of four biological replicates per lineage. Error bars represent the standard error of the mean. Asterisk denotes a significant difference between average ancestral and average lineage relative fitness by Student's *t*-test ($p < 0.05$).

Similarly, all REL4536 AN populations had increased relative fitness after 1,000 generations (**Figure 5.5**). The population displaying the greatest gain in relative fitness was REL4536 AN-1K-6 (2.08 ± 0.35), where the sequenced clone derived from this population also had the most mutations and the most GCRs of all sequenced REL4536 AN clones. Once again, as fitness assays were conducted at the population level, and genome sequence data was obtained at a clonal level, it is difficult to distinguish which particular mutations were responsible for increases in fitness. Fitness assays of the sequenced clones could potentially provide more insight. Moreover, in an independent set of fitness experiments conducted by Finn (2015), the relative fitness of REL4536 AN-1K-6 was 1.32 ± 0.04 (205), a value much lower than the one determined in this study. Thus, further experiments with this population are necessary. It is possible that the difference in relative population fitness between the two experiments is due to the use of different sub-populations of frozen stock; thus, a better understanding of the fitness dynamics of the population can be obtained with more replicates.

To further investigate potential correlation between fitness and number of GCRs, Pearson correlation coefficients between population fitness and number of detected GCRs were calculated for REL4536 *ΔsbcC::cat*-1K and REL4536 AN populations (**Table 5.7**). Correlation coefficients of -0.77 (p = 0.23) and 0.85 (p = 0.15) were obtained for the REL4536 *ΔsbcC::cat*-1K and REL4536 AN populations, respectively. Thus, there appeared to be a reasonably strong positive correlation between the number of GCRs in the sequenced clone and magnitude of fitness for the REL4536 AN lineages, but a reasonably strong negative correlation for the REL4536 *ΔsbcC::cat* lineages. When considering the REL4536 *ΔsbcC::cat*-1K and REL4536 AN populations together, a correlation coefficient of 0.24 was obtained (p = 0.57). Taken as a whole, these results suggest that there is weak correlation between the number of GCRs and fitness. These results also suggest that in the presence of functional SbcC, GCRs are well regulated and an increase in GCRs can lead to an increase in fitness. On the other hand, in the absence of functional SbcC, perhaps genome instability has drastically increased, such that population fitness does not increase in large increments due to the occurrence of too many lethal or highly deleterious GCRs.

5.2.2.6 Polymorphic evolved lineages

Finn (2015) noticed that after 2,000 generations of evolution, four of his seven anaerobic populations gave rise to colonies that were morphologically different than those typical of the ancestral population (typical colony morphotypes, TCM) when they were grown on LB agar plates (section 2.1.7.1.1) and incubated under aerobic conditions (section 2.2.1) (205). As these colonies were much smaller than TCMs, they were referred to as small colony morphotypes (SCMs). Such findings are not unusual, as many long-term adaptive lineages, including Lenski's landmark LTEE study, have reported the existence of populations with polymorphic phenotypes (300, 305, 306). Therefore, to determine if REL4536 *ΔsbcC::cat*-1K populations displayed any phenotypic polymorphism, REL4536 *ΔsbcC::cat*-1K populations, from which representative clones had been sequenced, were grown on LB agar plates, incubated under aerobic growth conditions and colony morphology was analysed. Briefly, five of the six REL4536 *ΔsbcC::cat*-1K populations were mixed populations, displaying both TCMs and SCMs (

Table A.9). Only the REL4536 *ΔsbcC::cat*-1K-6 population retained a typical colony morphology. Of note, none of the populations displaying mixed morphologies were associated with low relative fitness values (**Table 5.7**).

Table 5.7. Correlations between relative fitness and number of accumulated GCRs; and population morphology of clones under aerobic conditions.

Lineage	#GCRs	Fitness (ω)	Population colony morphology	Pearson correlation coefficients
REL4536 *ΔsbcC::cat*-1K-6	3	1.75	Typical*	-0.77 (p = 0.23)
REL4536 *ΔsbcC::cat*-1K-7	5	1.49	Mixed†	
REL4536 *ΔsbcC::cat*-1K-8	2	2.17	Mixed	
REL4536 *ΔsbcC::cat*-1K-12	6	1.70	Mixed	
REL4536 AN-1K-1	2	1.31	Typical	0.85 (p = 0.15)
REL4536 AN-1K-3	1	1.59	Typical	
REL4536 AN-1K-4	2	1.58	Typical	
REL4536 AN-1K-6	6	2.08	Typical	

*Typical morphology refers to colonies that are similar in size to the ancestor.

†Mixed populations were those with both typical and small sized colonies.

5.2.2.6.1 Genetic basis of polymorphism

Colony morphologies of the representative, sequenced clones from each population were also investigated. Only REL4536 ΔsbcC::*cat*-1K-1 and REL4536 *ΔsbcC::cat*-1K-8 clones were SCMs, as the remaining clones displayed colony sizes similar to that of the ancestral population. To determine if there was a genetic basis behind the SCMs, the whole genome sequence data (**Table A.7)** of both clones was analysed. Amongst both clones, only two mutations were shared; both clones had the 33 kb deletion and the 5 bp indel (section 5.2.2.4.2.1). Therefore, it is likely that one of the two mutations, or epistatic interactions between the two mutations, is responsible for the appearance of SCMs. Notably, neither mutation was detected in clones with typical colony morphology (**Table A.7)**. Finn (2015) also observed the 33 kb deletion and the 5 bp indel, amongst many other mutations, in the SCMs sequenced in their study (205). However, the two mutations were never reported in the same clone, indicating the likelihood of there being multiple pathways that result in a SCM phenotype in cells. Finn (2015) proposed that the TCM and SCM populations stably co-exist *via* an acetate cross-feeding mechanism, whereby SCMs arise, and are maintained, due to the secretion of acetate into the media as a by-product of anaerobic fermentation during TCM growth (205). Or alternatively, it is thought that the SCMs may have arisen due to the presence of citrate in the media.

5.3 Summary

One of the genes involved in maintaining genome integrity in *E. coli* is the *sbcC* gene, which forms a complex with the protein encoded by *sbcD* to repair DBSs. If left unrepaired, these DSBs can result in the formation of GCRs. In our MA study of *E. coli*, GCRs occurred more frequently when cells were grown under anaerobic conditions (section 4.2.1.3). Additionally, *sbcC* was shown to be more highly expressed under anaerobic growth conditions using both RNAseq and RT-qPCR (section 5.2.1.) Thus, the aims of this study were to use whole genome re-sequencing and experimental evolution techniques to determine the contribution of GCRs to adaptive evolution. To do this, the involvement of SbcC to the occurrence of GCRs in *E. coli* and the subsequent impact of those GCRs on population fitness were studied.

An *E. coli* REL4536 Δ*sbcC::cat* strain was generated (section 5.2.2.1) and used to establish 14 lineages that were sub-cultured every 24 h for 1,000 generations under anaerobic growth conditions (section 5.2.2.3), from which, selected clones were genome sequenced (section 5.2.2.4). Inactivation of *sbcC* led to an increased occurrence of GCRs, though this was not statistically significant, suggesting that proteins other than *sbcC* may be involved in GCR prevention under the anaerobic environment (section 5.2.2.4.1). Additionally, relative fitness of selected populations also increased (section 5.2.2.5.2), though, as these increases were not in accordance with the number of GCRs or mutations, the genetic basis of these increases remain unclear. Overall, a weak association between the number of GCRs in the sequenced clone and the fitness of the evolved lineages was found. Some populations displayed mixed morphology after 1,000 generations (section 5.2.2.6), growing colonies that were either similar to the ancestral strain (TCM) or much smaller (SCM) in size. While these small colonies were not associated with a loss of population fitness, it seems plausible that the 33 kb deletion, or the 5 bp indel, or more likely, both mutations, were responsible for the rise of SCMs.

In future, sequencing more clones and/or populations and measuring the relative fitness of those populations would provide a more thorough understanding of any potential trends between GCRs and fitness. Subjecting colonies to competitive fitness assays would also provide greater insight into the dynamics between GCRs and fitness. Allelic replacement studies, in which mutations are recreated in the ancestral strain and subsequently competed against the ancestor, would allow for direct measurements of fitness enhancements due to introduced mutations. Reconstructing mutants of those mutations that are of interest would also allow for the direct observation of their impact on colony morphology. Better understanding of the effects and actions of genes involved in DNA repair would also be of great value. This could be achieved through further work with the *E. coli* REL4536 Δ*sbcC::cat* strains. For instance, MA assays conducted with this strain would provide further insight on the mode of action of *sbcC*. Additionally, evolving the lineages for a longer period of time (at least for another 1,000 generations) would also be preferable, as this could increase the chances for more GCRs, which may have an impact on relative fitness, to arise.

Chapter Six: General discussion

6.1 Background

The primary aim of this thesis was to determine how aerobic and anaerobic environments affect the mutation rate and spectra of the facultative anaerobic bacterium, *E. coli* REL4536. In this study, MA lineages and whole genome re-sequencing of individual clones from the MA lineages were utilised to provide one of the most comprehensive estimates of the mutation rate and spectra in aerobic and anaerobic environments to date. The secondary aim of this thesis was to determine the impact of GCRs on the fitness of populations by determining the contribution of SbcC to the occurrence of GCRs under anaerobic growth conditions This was studied by evolving parallel populations of *sbcC* mutant strains of *E. coli* REL4536 and using whole genome re-sequencing of individual clones to identify all genetic changes that occurred within genomes as they adapted to the anaerobic environment. By using relative population fitness as a measure of adaptation, it was possible to investigate potential relationships between the number of GCRs and adaptive evolution.

6.2 General discussion of findings and future directions

Using both fluctuation and MA assays, it was determined that the genome-wide spontaneous mutation rate is significantly greater in an anaerobic environment than in an aerobic one. The average mutation rate obtained for aerobically grown *E. coli* cells *via* fluctuation assays was 2.6×10^{-10} mutations per nucleotide per generation, which is comparable to the genome-wide spontaneous mutation rate of 2.5×10^{-10} mutations per nucleotide per generation obtained *via* MA assays. The aerobic mutation rates obtained in this study are within the range of recent estimates obtained for a strain of *E. coli* K-12, though GCRs were not included in the analysis of that study. For anaerobically grown *E. coli* cells, an average mutation rate of 1.3×10^{-9} mutations per nucleotide was obtained *via* fluctuation assays. However, when using MA assays, a three-fold lower rate of 4.1×10^{-10} mutations per nucleotide per generation was obtained, revealing discrepancies that can be obtained between fluctuation assay based estimates and those based on the more comprehensive, genome-wide mutation detection.

While this study was a comprehensive analysis of the mutation rate and spectra, the mutation rates were estimated from only a small number of genomes (24 in each environment). More robust estimates could be obtained from sequencing larger sample sizes. Additionally, the mutation rates and spectra obtained from this study may not precisely match the mutation rates or spectra that are found in natural microbial populations, where populations may experience other stresses such as extreme temperatures or starvation. However, the reductionist approach undertaken within this thesis provides valuable foundation information regarding differences in mutation rate under aerobic and anaerobic conditions that would contribute to rates displayed by natural populations. Finally, it is important to note that the reported genome-wide mutation rates are still underestimations of the true spontaneous mutation rates under both aerobic and anaerobic environments. While there is negligible selection in this study (section 4.2.1.1.3), such that the substitution rate can be equated with the spontaneous mutation rate, any lethal and/or highly deleterious mutations would not have accumulated in the MA lineages and so, would not have been included in the calculations.

Another notable finding of this study was that, depending on how the mutation rate is expressed, with regard to generation or absolute time, different trends among mutation classes can be obtained. Due to the slower growth rate of anaerobically grown cells, mutation rates expressed per unit time were also calculated. Per day mutation rates for BPSs, indels and GCRs were greater in an aerobic environment, as compared to an anaerobic one, with BPSs and indels exhibiting significantly two-fold greater rates. Mutation rates calculated per generation displayed a different trend, with rates for BPSs, indels and GCRs being 1.3-, 1.5- and 2.6- fold greater, respectively, during growth in an anaerobic environment, as compared to aerobic. These differences in the per generation and the per day mutation rates can be attributed to the aerobically and anaerobically grown cells spending differing proportions of time in the different stages of the cell cycle, and indicate that some mutation types may arise independently of genome replication during cell division.

The types of mutations that prevailed in the two environments were also found to differ. In the aerobic environment, there were biases towards G → T transversions, which was expected as a result of ROS-induced DNA damage. In addition, IS*186* transposition occurred at significantly 4.6-fold greater rates in aerobically grown cells. In contrast, in the anaerobic environment, there was an unexpected propensity for C → A, T → G and

A → C transversions. These results were not consistent with our understanding of oxidative stress-induced DNA damage and repair, suggesting that other exogenous agents or cellular mechanisms present in anaerobically grown cells may be responsible for these mutations. In addition, while the general GCR mutation rate was significantly higher in anaerobically grown cells, IS element insertions displayed the greatest mutation rates. In particular, IS*150* transposition was significantly greater under anaerobic growth conditions, as compared to aerobic conditions. While IS elements are known to respond to starvation and other cellular stresses (60), the reasons for increased IS element activity under anaerobic conditions are not immediately clear.

Overall, these findings highlight the need for further studies of the mutagenic and physiological pressures associated with aerobic and anaerobic growth. In particular, determining and characterizing the agents of mutation behind the mutation biases observed in the anaerobic environment and determining the extent to which these sources contribute to the spontaneous mutation rate. Determining conditions that promote particular mutation classes (e.g. GCRs) would also greatly aid in understanding the mutational processes under both environments. Furthermore, in their study, Lee et al. (2012) demonstrated that methylated bases serve as mutational hotspots, and contributed to the prevalence of G → A and C → T transitions in aerobically grown *E. coli* (13). By analysing the sequence context of the BPSs, it would be possible to determine if anaerobically grown *E. coli* exhibit a similar trend or not. Moreover, even in the case of aerobically grown *E. coli*, there is conflicting information regarding the relationship between gene expression and mutation rates (13, 307, 308). So to determine whether the higher mutation rates can in part be explained by mutations in genes with low expression, it would be worthwhile to investigate these associations in anaerobically grown *E. coli*. Studies have also indicated that the leading and lagging strand of the *E. coli* chromosome are replicated with differential fidelity; such that the two strands differ in their susceptibility to mutations (258, 309). Further studies could involve elucidating any differences in the mutation rates of the leading and lagging strands, and to determine whether they can help explain the observed mutation spectra.

The rates at which mutations occur in populations are the combined result of the mutational pressures experienced by the DNA, accurate DNA replication and the efficiency of the pathways that find and repair DNA damage. Therefore, to determine how genome integrity is maintained during growth under aerobic and anaerobic conditions, the activities of the many DNA repair pathways in *E. coli* were investigated. Overall, expression of genes involved in repair and replication was greater under anaerobic growth conditions, than aerobic conditions. The greatest differential activity was observed for genes involved in GCR repair, consistent with findings that GCRs were more prevalent in anaerobic MA lineages. While there were some difficulties in relating gene expression of repair genes under aerobic and anaerobic conditions to the observed mutational spectra in the MA study, this study is one of few to provide insight about how genome fidelity is maintained under anaerobic conditions. As very little is known about how anaerobically grown *E. coli* maintain genome fidelity, better understanding of how the DNA repair and replication pathways function under anaerobic growth conditions will provide valuable information on how anaerobes maintain genome integrity and to establish which DNA repair and replication genes are responsible for the accumulation of certain mutations. Thus, analysing gene expression of more biological replications at different stages of the growth cycle would be beneficial. Finally, by conducting MA assays of repair-deficient strains, further insight on the mode of action of anaerobic repair genes could be gained.

Many repair and stress response genes, including those involved in acid resistance, displayed greater expression under anaerobic conditions, than under aerobic conditions. Hence, it is possible that these systems were induced by the intracellular pH generated within cells as they underwent anaerobic fermentation. In future, monitoring intracellular pH or for gas production in anaerobically grown cells will aid in better understanding the physiological conditions experienced by the cells.

We sought to understand how GCRs contribute to and impact adaptive evolution by using *sbcC*, a gene involved in maintaining genome fidelity, and experimental evolution techniques. After *E. coli* strains containing a disrupted *sbcC* had been evolved for 1,000 generations, *sbcC* mutants exhibited higher rates of mutations and GCRs, though these differences were not statistically significant. In addition, the relative population fitness of selected *sbcC* mutant populations had increased, though the genetic bases behind these increases were not able to be resolved. Moreover, there were no clear correlations between the number of GCRs detected in sequenced clones and the relative fitness of evolved populations. In fact, these results suggested that GCRs can play a significant role in adaptation under anaerobic environments. In the presence of functional SbcC, GCRs are well regulated and that an increase in GCRs can lead to an increase in fitness. On the other hand, in the absence of functional SbcC, perhaps genome instability has drastically increased, such that population fitness does not increase in large increments due to the occurrence of too many lethal or highly deleterious GCRs. Additionally, after 1,000 generations, some of the populations also displayed mixed colony morphology.

6.3 Conclusions

The evolutionary success of organisms relies on their adaptability to changing environmental conditions. As mutations are the fundamental source of genetic variation upon which natural selection and genetic drift can act, knowledge of mutation rates and the molecular spectrum of spontaneous mutations is extremely important in understanding how populations evolve over time. This thesis, using experimental evolution and whole genome re-sequencing, provides detailed insight into the mutation rates and spectra of *E. coli* grown under aerobic and anaerobic conditions. The genome-wide mutation rate of anaerobically grown cells that was reported in this study was 1.7-fold greater than the rate that was obtained for aerobically grown cells. Also, by using whole genome re-sequencing, it was possible to detect various classes of mutations including BPSs, indels and GCRs. The scale at which these mutations occurred within aerobically and anaerobically grown cells in *E. coli* was not known previously, indicating the benefit of studying mutations across the whole genome. Indeed, this is the first study to compare the genome-wide spontaneous mutation rate of *E. coli* grown aerobically and anaerobically on minimal media and provides the most comprehensive mutation rate estimate of anaerobically grown *E. coli* to date. Additionally, this study found that GCRs, a mutational class that has been rarely studied due to technical challenges in identification, were more prevalent under anaerobic growth and this was largely due to increased IS element activity.

While the findings of this study may not be relevant to complex eukaryotic organisms, due to the fact that such organisms can only metabolise energy aerobically, this PhD study can be used as a model for other facultative anaerobes with commercial or pathogenic concerns such as *Saccharomyces cerevisiae* and species belonging to the *Staphylococcus*, *Streptococcus* and *Listeria* genera. Additionally, to date, there has been limited study into the mutagenic and evolutionary processes of anaerobic organisms.

Thus, knowledge of what mutations may arise will aid in tracking the evolution of these organisms, especially in the case of pathogens as they expand and encounter new environments.

References

1. Van Ditmarsch D & Xavier JB (2014) Seeing is believing: what experiments with microbes reveal about evolution. *Trends in Microbiology* 22(1):2-4.

2. Lynch M (2010) Evolution of the mutation rate. *Trends in Genetics* 26(8):345-352.

3. Halligan DL & Keightley PD (2009) Spontaneous mutation accumulation studies in evolutionary genetics. *Annual Review of Ecology, Evolution, and Systematics* 40(1):151-172.

4. Kondrashov FA & Kondrashov AS (2010) Measurements of spontaneous rates of mutations in the recent past and the near future. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365(1544):1169-1176.

5. Friedberg EC*, et al.* (2009) *DNA Repair and Mutagenesis* (ASM Press, Washington, DC) 2nd Ed.

6. Barrick JE & Lenski RE (2013) Genome dynamics during experimental evolution. *Nature Reviews Genetics* 14(12):827-839.

7. Kimura M (1968) Evolutionary rate at the molecular level. *Nature* 217(5129):624-626.

8. Drake JW (1991) Spontaneous mutation. *Annual Review of Genetics* 25(1):125-146.

9. Denamur E & Matic I (2006) Evolution of mutation rates in bacteria. *Molecular Microbiology* 60(4):820-827.

10. Partridge JD, Scott C, Tang Y, Poole RK, & Green J (2006) *Escherichia coli* transcriptome dynamics during the transition from anaerobic to aerobic conditions. *Journal of Biological Chemistry* 281(38):27806-27815.

11. Drake JW (1991) A constant rate of spontaneous mutation in DNA-based microbes. *Proceedings of the National Academy of Sciences* 88(16):7160-7164.

12. Kibota TT & Lynch M (1996) Estimate of the genomic mutation rate deleterious to overall fitness in *Escherichia coli*. *Nature* 381(6584):694-696.

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

14. Miller JH (1996) Spontaenous mutators in bacteria: insights into pathways of mutagenesis and repair. *Annual Review of Microbiology* 50(1):625-643.

15. Friedman JI & Stivers JT (2010) Detection of damaged DNA bases by DNA glycosylase enzymes. *Biochemistry* 49(24):4957-4967.

16. Nevinsky GA (2011) Main factors providing specificity of repair enzymes. *Biochemistry (Moscow)* 76(1):94-117.

17. Rosche WA & Foster PL (2000) Determining mutation rates in bacterial populations. *Methods* 20(1):4-17.

18. Foster PL (2006) Methods for determining spontaneous mutation rates. in *Methods in Enzymology*, pp 195-213.

19. Cooke MS, Evans MD, Dizdaroglu M, & Lunec J (2003) Oxidative DNA damage: mechanisms, mutation, and disease. *The Journal of the Federation of American Societies for Experimental Biology* 17(10):1195-1214.

20. Imlay JA (2013) The molecular mechanisms and physiological consequences of oxidative stress: Lessons from a model bacterium. *Nature Reviews Microbiology* 11(7):443-454.

21. Lushchak VI (2001) Oxidative stress and mechanisms of protection against it in bacteria. *Biochemistry (Moscow)* 66(5):476-489.

22. Imlay JA (2003) Pathways of oxidative damage. *Annual Review of Microbiology* 57(1):395-418.

23. Setoyama D, Ito R, Takagi Y, & Sekiguchi M (2011) Molecular actions of Escherichia coli MutT for control of spontaneous mutagenesis. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 707(1-2):9-14.

24. Slupphaug G, Kavli B, & Krokan HE (2003) The interacting pathways for prevention and repair of oxidative DNA damage. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 531(1-2):231-251.

25. Fijal BA, Idury RM, & Witte JS (2002) Analysis of mutational spectra: Locating hotspots and clusters of mutations using recursive segmentation. *Statistics in Medicine* 21(13):1867-1885.

26. Bjelland S & Seeberg E (2003) Mutagenicity, toxicity and repair of DNA base damage induced by oxidation. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 531(1-2):37-80.

27. Skinner AM & Turker MS (2005) Oxidative mutagenesis, mismatch repair, and aging. *Science of Aging Knowledge Environment* 2005(9):re3.

28. Sakai A, Nakanishi M, Yoshiyama K, & Maki H (2006) Impact of reactive oxygen species on spontaneous mutagenesis in *Escherichia coli*. *Genes to Cells* 11(7):767-778.

29. Hershberg R & Petrov DA (2010) Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genetics* 6(9):e1001115.

30. Belle EMS, Piganeau G, Gardner M, & Eyre-Walker A (2005) An investigation of the variation in the transition bias among various animal mitochondrial DNA. *Gene* 355:58-66.

31. Viguera E, Canceill D, & Ehrlich SD (2001) Replication slippage involves DNA polymerase pausing and dissociation. *The EMBO Journal* 20(10):2587-2595.

32. Roth J, *et al.* (1996) Rearrangements of the bacterial chromosome: formation and applications. *Escherichia coli and Salmonella: cellular and molecular biology*, ed Neidhardt FC (ASM Press, Washington DC), 2nd Ed Vol 2, pp 2256-2276.

33. Hughes D (2000) Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes. *Genome Biology* 1(6):reviews0006.0001-reviews0006.0008

34. Sun S, Ke R, Hughes D, Nilsson M, & Andersson DI (2012) Genome-wide detection of spontaneous chromosomal rearrangements in bacteria. *PLoS ONE* 7(8):e42639.

35. Raeside C, *et al.* (2014) Large chromosomal rearrangements during a long-term evolution experiment with *Escherichia coli*. *mBio* 5(5):e01377-01314.

36. Bergthorsson U & Ochman H (1999) Chromosomal changes during experimental evolution in laboratory populations of *Escherichia coli*. *Journal of Bacteriology* 181(4):1360-1363.

37. Moore JM, Wimberly H, Thornton PC, Rosenberg SM, & Hastings PJ (2012) Gross chromosomal rearrangement mediated by DNA replication in stressed cells: evidence from *Escherichia coli*. *Annals of the New York Academy of Sciences* 1267(1):103-109.

38. Dunham MJ, *et al.* (2002) Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America* 99(25):16144-16149.

39. Crombach A & Hogeweg P (2007) Chromosome rearrangements and the evolution of genome structuring and adaptability. *Molecular Biology and Evolution* 24(5):1130-1139.

40. Yona AH, *et al.* (2012) Chromosomal duplication is a transient evolutionary solution to stress. *Proceedings of the National Academy of Sciences of the United States of America* 109(51):21010-21015.

41. Cooper VS, Schneider D, Blot M, & Lenski RE (2001) Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli* B. *Journal of Bacteriology* 183(9):2834-2841.

42. Maharjan RP, *et al.* (2013) A case of adaptation through a mutation in a tandem duplication during experimental evolution in *Escherichia coli*. *BMC genomics* 14(1):441-453.

43. Blount ZD, Barrick JE, Davidson CJ, & Lenski RE (2012) Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature* 489(7417):513-518.

44. Eisen JA, Heidelberg JF, White O, & Salzberg SL (2000) Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biology* 1(6):research0011.0011-research0011.0019

45. Darling AE, Miklós I, & Ragan MA (2008) Dynamics of Genome Rearrangement in Bacterial Populations. *PLoS Genetics* 4(7):e1000128.

46. Tillier ER & Collins RA (2000) Genome rearrangement by replication-directed translocation. *Nature Genetics* 26(2):195-197.

47. Denver DR, Morris K, Lynch M, Vassilieva LL, & Thomas WK (2000) High direct estimate of the mutation rate in the mitochondrial genome of *Caenorhabditis elegans*. *Science* 289(5488):2342-2344.

48. Denver DR, Morris K, Lynch M, & Thomas WK (2004) High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature* 430(7000):679-682.

49. Frost LS, Leplae R, Summers AO, & Toussaint A (2005) Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology* 3(9):722-732.

50. Gaffé J*, et al.* (2011) Insertion sequence-driven evolution of *Escherichia coli* in chemostats. *Journal of Molecular Evolution* 72(4):398-412.

51. Kazazian HH (2004) Mobile elements: drivers of genome evolution. *Science* 303(5664):1626-1632.

52. Casacuberta E & González J (2013) The impact of transposable elements in environmental adaptation. *Molecular Ecology* 22(6):1503-1517.

53. Darmon E & Leach DRF (2014) Bacterial genome instability. *Microbiology and Molecular Biology Reviews* 78(1):1-39.

54. McClintock B (1950) The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America* 36(6):344-355.

55. Derbyshire KM & Grindley NDF (1986) Replicative and conservative transposition in bacteria. *Cell* 47(3):325-327.

56. Shapiro JA (1979) Molecular model for the transposition and replication of bacteriophage Mu and other transposable elements. *Proceedings of the National Academy of Sciences of the United States of America* 76(4):1933-1937.

57. Treangen TJ, Abraham AL, Touchon M, & Rocha EPC (2009) Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiology Reviews* 33(3):539-571.

58. Simon DM & Zimmerly S (2008) A diversity of uncharacterized reverse transcriptases in bacteria. *Nucleic Acids Research* 36(22):7219-7229.

59. Mahillon J & Chandler M (1998) Insertion sequences. *Microbiology and Molecular Biology Reviews* 62(3):725-774.

60. Siguier P, Gourbeyre E, & Chandler M (2014) Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiology Reviews* 38(5):865-891.

61. Sousa A, Bourgard C, Wahl LM, & Gordo I (2013) Rates of transposition in *Escherichia coli*. *Biology Letters* 9(6).

62. Lee H, Popodi E, Foster PL, & Tang H (2014) Detection of structural variants involving repetitive regions in the reference genome. *Journal of Computational Biology* 21(3):219-233.

63. Papadopoulos D*, et al.* (1999) Genomic evolution during a 10,000-generation experiment with bacteria. *Proceedings of the National Academy of Sciences of the United States of America* 96(7):3807-3812.

64. Schneider D, Duperchy E, Coursange E, Lenski RE, & Blot M (2000) Long-term experimental evolution in *Escherichia coli*. IX. Characterization of insertion sequence-mediated mutations and rearrangements. *Genetics* 156(2):477-488.

65. Schneider D & Lenski RE (2004) Dynamics of insertion sequence elements during experimental evolution of bacteria. *Research in Microbiology* 155(5):319-327.

66. Delihas N (2008) Small mobile sequences in bacteria display diverse structure/function motifs. *Molecular Microbiology* 67(3):475-481.

67. Delihas N (2011) Impact of small repeat sequences on bacterial genome evolution. *Genome Biology and Evolution* 3(1):959-973.

68. Rocha EPC, Cornet E, & Michel B (2005) Comparative and evolutionary analysis of the bacterial homologous recombination systems. *PLoS Genetics* 1(2):e15.

69. Michel B (2005) After 30 years of study, the bacterial SOS response still surprises us. *PLoS Biology* 3(7):e255.

70. Drabløs F*, et al.* (2004) Alkylation damage in DNA and RNA repair mechanisms and medical significance. *DNA Repair* 3(11):1389-1407.

71. Fijalkowska IJ, Schaaper RM, & Jonczyk P (2012) DNA replication fidelity in *Escherichia coli*: a multi-DNA polymerase affair. *FEMS Microbiology Reviews* 36(6):1105-1121.

72. Canceill D, Viguera E, & Ehrlich SD (1999) Replication slippage of different DNA polymerases is inversely related to their strand displacement efficiency. *Journal of Biological Chemistry* 274(39):27481-27490.

73. McCulloch SD & Kunkel TA (2008) The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Research* 18(1):148-161.

74. Waters LS*, et al.* (2009) Eukaryotic translesion polymerases and their roles and regulation in DNA damage tolerance. *Microbiology and Molecular Biology Reviews* 73(1):134-154.

75. Janion C (2008) Inducible SOS response system of DNA repair and mutagenesis in *Escherichia coli*. *International Journal of Biological Sciences* 4:338-344.

76. Foster PL (2007) Stress-induced mutagenesis in bacteria. *Critical Reviews in Biochemistry and Molecular Biology* 42(5):373-397.

77. Maki H (2002) Origins of spontaneous mutations: specificity and directionality of base-substitution, frameshift, and sequence-substitution mutageneses. *Annual Review of Genetics* 36(1):279-303.

78. Robertson AB, Klungland A, Rognes T, & Leiros I (2009) Base excision repair: The long and short of it. *Cellular and Molecular Life Sciences* 66(6):981-993.

79. David SS, O'Shea VL, & Kundu S (2007) Base-excision repair of oxidative DNA damage. *Nature* 447(7147):941-950.

80. Tsuzuki T, Nakatsu Y, & Nakabeppu Y (2007) Significance of error-avoiding mechanisms for oxidative DNA damage in carcinogenesis. *Cancer Science* 98(4):465-470.

81. Horst JP, Wu TH, & Marinus MG (1999) *Escherichia coli* mutator genes. *Trends in Microbiology* 7(1):29-36.

82. Petit C & Sancar A (1999) Nucleotide excision repair: From *E. coli* to man. *Biochimie* 81(1-2):15-25.

83. Marti TM, Kunz C, & Fleck O (2002) DNA mismatch repair and mutation avoidance pathways. *Journal of Cellular Physiology* 191(1):28-41.

84. Reardon JT & Sancar A (2005) Nucleotide excision repair. *Progress in Nucleic Acid Research and Molecular Biology* 79:183-235.

85. Pascucci B, D'Errico M, Parlanti E, Giovannini S, & Dogliotti E (2011) Role of nucleotide excision repair proteins in oxidative DNA damage repair: An updating. *Biochemistry (Moscow)* 76(1):4-15.

86. Li GM (2008) Mechanisms and functions of DNA mismatch repair. *Cell Research* 18(1):85-98.

87. Polosina YY & Cupples CG (2010) MutL: Conducting the cell's response to mismatched and misaligned DNA. *BioEssays* 32(1):51-59.

88. Polosina YY & Cupples CG (2010) Wot the 'L-Does MutL do? *Mutation Research/Reviews in Mutation Research* 705(3):228-238.

89. Fukui K (2010) DNA mismatch repair in eukaryotes and bacteria. *Journal of Nucleic Acids* 2010:260512.

90. Harfe BD & Jinks-Robertson S (2000) DNA mismatch repair and genetic instability. *Annual Review of Genetics* 34(1):359-399.

91. Wyrzykowski J & Volkert MR (2003) The *Escherichia coli* methyl-directed mismatch repair system repairs base pairs containing oxidative lesions. *Journal of Bacteriology* 185(5):1701-1704.

92. Tham K-C, *et al.* (2013) Mismatch repair inhibits homeologous recombination *via* coordinated directional unwinding of trapped DNA structures. *Molecular Cell* 51(3):326-337.

93. Petit MA, Dimpfl J, Radman M, & Echols H (1991) Control of large chromosomal duplications in *Escherichia col*i by the mismatch repair system. *Genetics* 129(2):327-332.

94. Kelley WL (2006) Lex marks the spot: the virulent side of SOS and a closer look at the LexA regulon. *Molecular Microbiology* 62(5):1228-1238.

95. Durfee T, Hansen A-M, Zhi H, Blattner FR, & Jin DJ (2008) Transcription profiling of the stringent response in *Escherichia coli*. *Journal of Bacteriology* 190(3):1084-1096.

96. Poole K (2012) Bacterial stress responses as determinants of antimicrobial resistance. *Journal of Antimicrobial Chemotherapy* 67(9):2069-2089.

97. Farr SB & Kogoma T (1991) Oxidative stress responses in *Escherichia coli* and *Salmonella typhimurium*. *Microbiological Reviews* 55(4):561-585.

98. Vercruysse M, *et al.* (2011) Stress response regulators identified through genome-wide transcriptome analysis of the (p)ppGpp-dependent response in *Rhizobium etli*. *Genome Biology* 12(2):R17.

99. Chayot R, Montagne B, Mazel D, & Ricchetti M (2010) An end-joining repair mechanism in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* 107(5):2141-2146.

100. Bowater R & Doherty AJ (2006) Making ends meet: repairing breaks in bacterial DNA by non-homologous end-joining. *PLoS Genetics* 2(2):e8.

101. Sharples GJ (2009) For absent friends: life without recombination in mutualistic gamma-proteobacteria. *Trends in Microbiology* 17(6):233-242.

102. Cox MM (1998) A broadening view of recombinational DNA repair in bacteria. *Genes to Cells* 3(2):65-78.

103. Blattner FR*, et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* 277(5331):1453-1462.

104. Bidnenko V*, et al.* (1999) *sbcB sbcC* null mutations allow RecF-mediated repair of arrested replication forks in *rep recBC* mutants. *Molecular Microbiology* 33(4):846-857.

105. Wilhelm BT & Landry JR (2009) RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* 48(3):249-257.

106. Wang Z, Gerstein M, & Snyder M (2009) RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10(1):57-63.

107. Van Vliet AHM (2010) Next generation sequencing of microbial transcriptomes: challenges and opportunities. *FEMS Microbiology Letters* 302(1):1-7.

108. Filiatrault MJ*, et al.* (2010) Transcriptome analysis of *Pseudomonas syringae* identifies new genes, noncoding RNAs, and antisense activity. *Journal of Bacteriology* 192(9):2359-2372.

109. Sorek R & Cossart P (2010) Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nature Reviews Genetics* 11(1):9-16.

110. Skvortsov TA & Azhikina TL (2010) A review of the transcriptome analysis of bacterial pathogens in vivo: problems and solutions. *Russian Journal of Bioorganic Chemistry* 36(5):550-559.

111. Passalacqua KD*, et al.* (2009) Structure and complexity of a bacterial transcriptome. *Journal of Bacteriology* 191(10):3203-3211.

112. Ozsolak F & Milos PM (2011) RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics* 12(2):87-98.

113. Croucher NJ & Thomson NR (2010) Studying bacterial transcriptomes using RNA-seq. *Current Opinion in Microbiology* 13(5):619-624.

114. Croucher NJ*, et al.* (2009) A simple method for directional transcriptome sequencing using illumina technology. *Nucleic Acids Research* 37(22):e148.

115. Albrecht M, Sharma CM, Reinhardt R, Vogel J, & Rudel T (2010) Deep sequencing-based discovery of the *Chlamydia trachomatis* transcriptome. *Nucleic Acids Research* 38(3):868-877.

116. Costa V, Angelini C, De Feis I, & Ciccodicola A (2010) Uncovering the complexity of transcriptomes with RNA-Seq. *Journal of Biomedicine and Biotechnology* 2010:853916.

117. Faghihi MA & Wahlestedt C (2009) Regulatory roles of natural antisense transcripts. *Nature Reviews Molecular Cell Biology* 10(9):637-643.

118. Röscheisen C, Haupter S, Zechner U, & Speit G (1994) Characterization of spontaneous and induced mutations in SV40-transformed normal and ataxia telangiectasia cell lines. *Somatic Cell and Molecular Genetics* 20(6):493-504.

119. De Visser JAGM, Akkermans ADL, Hoekstra RF, & De Vos WM (2004) Insertion sequence-mediated mutations isolated during adaptation to growth and starvation in *Lactococcus lactis*. *Genetics* 168(3):1145-1157.

120. Cotton RGH (1993) Current methods of mutation detection. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 285(1):125-144.

121. Elena SF, Whittam TS, Winkworth CL, Riley MA, & Lenski RE (2005) Genomic divergence of *Escherichia coli* strains: evidence for horizontal transfer and variation in mutation rates. *International Microbiology* 8(4):271-278.

122. Bentley DR (2006) Whole-genome re-sequencing. *Current Opinion in Genetics & Development* 16(6):545-552.

123. Ronaghi M, Uhlén M, & Nyrén P (1998) A sequencing method based on real-time pyrophosphate. *Science* 281(5375):363-365.

124. Bentley DR*, et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456(7218):53-59.

125. Valouev A*, et al.* (2008) A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Research* 18(7):1051-1063.

126. Rothberg JM*, et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475(7356):348-352.

127. Schadt EE, Turner S, & Kasarskis A (2010) A window into third-generation sequencing. *Human Molecular Genetics* 19(R2):R227-240.

128. Eid J*, et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323(5910):133-138.

129. Shendure J*, et al.* (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309(5741):1728-1732.

130. Pushkarev D, Neff NF, & Quake SR (2009) Single-molecule sequencing of an individual human genome. *Nature Biotechnology* 27(9):847-850.

131. Metzker ML (2010) Sequencing technologies-the next generation. *Nature Reviews Genetics* 11(1):31-46.

132. Loman NJ, *et al.* (2012) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature Reviews Microbiology* 10(9):599-606.

133. Van Dijk EL, Auger H, Jaszczyszyn Y, & Thermes C (2014) Ten years of next-generation sequencing technology. *Trends in Genetics* 30(9):418-426.

134. Nagarajan N & Pop M (2013) Sequence assembly demystified. *Nature Reviews Genetics* 14(3):157-167.

135. Wetzel J, Kingsford C, & Pop M (2011) Assessing the benefits of using mate-pairs to resolve repeats in *de novo* short-read prokaryotic assemblies. *BMC Bioinformatics* 12(1):95.

136. Wielgoss S, *et al.* (2011) Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3: Genes|Genomes|Genetics* 1(3):183-186.

137. Trindade S, Perfeito L, & Gordo I (2010) Rate and effects of spontaneous mutations that affect fitness in mutator *Escherichia coli*. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365(1544):1177-1186.

138. Conrad TM, Lewis NE, & Palsson BO (2011) Microbial laboratory evolution in the era of genome-scale science. *Molecular Systems Biology* 7:509.

139. Saxer G, *et al.* (2012) Whole genome sequencing of mutation accumulation lines reveals a low mutation rate in the social amoeba *Dictyostelium discoideum*. *PLoS ONE* 7(10):e46759.

140. Duffy S, Shackelton LA, & Holmes EC (2008) Rates of evolutionary change in viruses: patterns and determinants. *Nature Reviews Genetics* 9(4):267-276.

141. Nachman MW & Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics* 156(1):297-304.

142. Xue Y, *et al.* (2009) Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Current Biology* 19(17):1453-1457.

143. Chen J-Q, *et al.* (2009) Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Molecular Biology and Evolution* 26(7):1523-1531.

144. Lynch M (2010) Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences* 107(3):961-968.

145. Sniegowski PD, Gerrish PJ, Johnson T, & Shaver A (2000) The evolution of mutation rates: separating causes from consequences. *BioEssays* 22(12):1057-1066.

146. Gago S, Elena SF, Flores R, & Sanjuán R (2009) Extremely high mutation rate of a hammerhead viroid. *Science* 323(5919):1308.

147. Bjedov I, *et al.* (2003) Stress-induced mutagenesis in bacteria. *Science* 300(5624):1404-1409.

148. Foster PL (1999) Sorting out mutation rates. *Proceedings of the National Academy of Sciences of the United States of America* 96(14):7617-7618.

149. Luria SE & Delbruck M (1943) Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics* 28(6):491-511.

150. Sarkar S, Ma WT, & Sandri GvH (1992) On fluctuation analysis: a new, simple and efficient method for computing the expected number of mutants. *Genetica* 85(2):173-179.

151. Bridges BA (1980) The fluctuation test. *Archives of Toxicology* 46(1):41-44.

152. Stewart FM (1994) Fluctuation tests: how reliable are the estimates of mutation rates? *Genetics* 137(4):1139-1146.

153. Gerrish P (2008) A simple formula for obtaining markedly improved mutation rate estimates. *Genetics* 180(3):1773-1778.

154. Wu X, *et al.* (2009) A robust estimator of mutation rates. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 661(1–2):101-109.

155. Sniegowski PD, Gerrish PJ, & Lenski RE (1997) Evolution of high mutation rates in experimental populations of *E. coli*. *Nature* 387(6634):703-705.

156. Fu YX (1994) A phylogenetic estimator of effective population size or mutation rate. *Genetics* 136(2):685-692.

157. Drake JW, Charlesworth B, Charlesworth D, & Crow JF (1998) Rates of spontaneous mutation. *Genetics* 148(4):1667-1686.

158. Ochman H (2003) Neutral mutations and neutral substitutions in bacterial genomes. *Molecular Biology and Evolution* 20(12):2091-2096.

159. Drake JW (2012) Contrasting mutation rates from specific-locus and long-term mutation accumulation procedures. *G3: Genes|Genomes|Genetics* 2(4):483-485.

160. Loewe L, Textor V, & Scherer S (2003) High deleterious genomic mutation rate in stationary phase of *Escherichia coli*. *Science* 302(5650):1558-1560.

161. Hall DW, Mahmoudizad R, Hurd AW, & Joseph SB (2008) Spontaneous mutations in diploid *Saccharomyces cerevisiae*: another thousand cell generations. *Genetics Research* 90(3):229-241.

162. Joseph SB & Hall DW (2004) Spontaneous mutations in diploid *Saccharomyces cerevisiae*: more beneficial than expected. *Genetics* 168(4):1817-1825.

163. Schaack S, Allen DE, Latta LC, Morgan KK, & Lynch M (2013) The effect of spontaneous mutations on competitive ability. *Journal of Evolutionary Biology* 26(2):451-456.

164. Behe MJ (2010) Experimental evolution, loss-of-function mutations, and "the first rule of adaptive evolution". *Quarterly Review of Biology* 85(4):419-445.

165. Kawecki TJ, *et al.* (2012) Experimental evolution. *Trends in Ecology & Evolution* 27(10):547-560.

166. Elena SF & Lenski RE (2003) Evolution experiments with microorganisms: The dynamics and genetic bases of adaptation. *Nature Reviews Genetics* 4(6):457-469.

167. Hindré T, Knibbe C, Beslon G, & Schneider D (2012) New insights into bacterial adaptation through *in vivo* and *in silico* experimental evolution. *Nature Reviews Microbiology* 10(5):352-365.

168. Lenski RE, Rose MR, Simpson SC, & Tadler SC (1991) Long-term experimental evolution in *Escherichia coli*. 1. Adaptation and divergence during 2,000 generations. *American Naturalist* 138(6):1315-1341.

169. Mozhayskiy V & Tagkopoulos I (2013) Microbial evolution *in vivo* and *in silico*: methods and applications. *Integrative Biology* 5(2):262-277.

170. Dettman JR, *et al.* (2012) Evolutionary insight from whole-genome sequencing of experimentally evolved microbes. *Molecular Ecology* 21(9):2058-2077.

171. Dragosits M & Mattanovich D (2013) Adaptive laboratory evolution-principles and applications for biotechnology. *Microbial Cell Factories* 12(1):64-81.

172. Kishimoto T, *et al.* (2010) Transition from positive to neutral in mutation fixation along with continuing rising fitness in thermal adaptive evolution. *PLoS Genetics* 6(10):e1001164.

173. Bennett AF & Hughes BS (2009) Microbial experimental evolution. *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology* 297(1):R17-R25.

174. Lee JY, Seo J, Kim ES, Lee HS, & Kim P (2013) Adaptive evolution of *Corynebacterium glutamicum* resistant to oxidative stress and its global gene expression profiling. *Biotechnology Letters* 35(5):709-717.

175. Puentes-Téllez PE, Kovács ÁT, Kuipers OP, & van Elsas JD (2014) Comparative genomics and transcriptomics analysis of experimentally evolved *Escherichia coli* MC1000 in complex environments. *Environmental Microbiology* 16(3):856-870.

176. Toprak E, *et al.* (2012) Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nature Genetics* 44(1):101-105.

177. Rodríguez-Verdugo A, Gaut B, & Tenaillon O (2013) Evolution of *Escherichia coli* rifampicin resistance in an antibiotic-free environment during thermal stress. *BMC Evolutionary Biology* 13(1):1-11.

178. Cooper TF & Lenski RE (2010) Experimental evolution with *E. coli* in diverse resource environments. I. Fluctuating environments promote divergence of replicate populations. *BMC Evolutionary Biology* 10(1):11-20.

179. Finkel SE & Kolter R (1999) Evolution of microbial diversity during prolonged starvation. *Proceedings of the National Academy of Sciences of the United States of America* 96(7):4023-4027.

180. Tenaillon O, Denamur E, & Matic I (2004) Evolutionary significance of stress-induced mutagenesis in bacteria. *Trends in Microbiology* 12(6):264-270.

181. Lenski RE (1998 - 2015) The *E. coli* long-term experimental evolution project site. http://myxo.css.msu.edu/ecoli (Michigan State University, East Lansing, MI).

182. Lenski RE & Travisano M (1994) Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proceedings of the National Academy of Sciences of the United States of America* 91(15):6808-6814.

183. Wielgoss S*, et al.* (2013) Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proceedings of the National Academy of Sciences of the United States of America* 110(1):222-227.

184. Philippe N, Pelosi L, Lenski RE, & Schneider D (2009) Evolution of penicillin-binding protein 2 concentration and cell shape during a long-term experiment with *Escherichia coli*. *Journal of Bacteriology* 191(3):909-921.

185. Wiser MJ, Ribeck N, & Lenski RE (2013) Long-term dynamics of adaptation in asexual populations. *Science* 342(6164):1364-1367.

186. Blount ZD, Borland CZ, & Lenski RE (2008) Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* 105(23):7899-7906.

187. Barrick JE*, et al.* (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* 461(7268):1243-1247.

188. Quandt EM, Deatherage DE, Ellington AD, Georgiou G, & Barrick JE (2014) Recursive genomewide recombination and sequencing reveals a key refinement step in the evolution of a metabolic innovation in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* 111(6):2217-2222.

189. Darmon E*, et al.* (2007) SbcCD Regulation and Localization in *Escherichia coli*. *Journal of Bacteriology* 189(18):6686-6694.

190. Connelly JC & Leach DRF (1996) The *sbcC* and *sbcD* genes of *Escherichia coli* encode a nuclease involved in palindrome inviability and genetic recombination. *Genes to Cells* 1(3):285-291.

191. Storvik KAM & Foster PL (2011) The SMC-like protein complex SbcCD enhances DNA polymerase IV-dependent spontaneous mutation in *Escherichia coli*. *Journal of Bacteriology* 193(3):660-669.

192. Liu S*, et al.* (2014) Structural basis for DNA recognition and nuclease processing by the Mre11 homologue SbcD in double-strand breaks repair. *Acta Crystallographica Section D: Biological Crystallography* 70(2):299-309.

193. Mascarenhas J*, et al.* (2006) *Bacillus subtilis* SbcC protein plays an important role in DNA inter-strand cross-link repair. *BMC Molecular Biology* 7:20.

194. Hu Y*, et al.* (2010) Characteristics of nuclease activity of the SbcCD complex from *Deinococcus radiodurans*. *Journal of Biochemistry* 147(3):307-315.

195. Connelly JC, De Leau ES, & Leach DR (2003) Nucleolytic processing of a protein-bound DNA end by the *E. coli* SbcCD (MR) complex. *DNA Repair* 2(7):795-807.

196. Darmon E*, et al.* (2010) *E. coli* SbcCD and RecA control chromosomal rearrangement induced by an interrupted palindrome. *Molecular Cell* 39(1):59-70.

197. Barrick JE & Lenski RE (2009) Genome-wide mutational diversity in an evolving population of *Escherichia coli*. *Cold Spring Harbor Symposia on Quantitative Biology* 74:119-129.

198. Blank K, Hensel M, & Gerlach RG (2011) Rapid and highly efficient method for scarless mutagenesis within the *Salmonella enterica* chromosome. *PLoS ONE* 6(1):e15763.

199. Sambrook J & Russel DW (2001) *Molecular cloning: a laboratory manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York) 3rd Ed.

200. Carlton BC & Brown BJ (1981) *Manual of methods for general bacteriology* (American Society for Microbiology, Washington, D.C.).

201. Ramakers C, Ruijter JM, Deprez RH, & Moorman AF (2003) Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neuroscience letters* 339(1):62-66.

202. Kirby KS (1965) Isolation and characterization of ribosomal ribonucleic acid. *Biochemical Journal* 96:266-269.

203. Drummond A*, et al.* (2011) Geneious http://www.geneious.com/), version 5.4.

204. Hall BM, Ma C-X, Liang P, & Singh KK (2009) Fluctuation AnaLysis CalculatOR: a web tool for the determination of mutation rate using Luria–Delbrück fluctuation analysis. *Bioinformatics* 25(12):1564-1565.

205. Finn T (2015) Understanding bacterial adaptation to aerobic and anaerobic environments through experimental evolution and whole genome sequencing. Doctor of Philosophy Thesis (Massey University).

206. Rutherford K, *et al.* (2000) Artemis: sequence visualization and annotation. *Bioinformatics* 16(10):944-945.

207. Altschul SF, Gish W, Miller W, Myers EW, & Lipman DJ (1990) Basic local alignment search tool. *Journal of molecular biology* 215(3):403-410.

208. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9(4):357-359.

209. Deatherage DE & Barrick JE (2014) Identification of mutations in laboratory evolved microbes from next-generation sequencing data using breseq. *Methods in molecular biology (Clifton, N.J.)* 1151:165-188.

210. Love M, Huber W, & Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15(12):550.

211. Carver T, Thomson N, Bleasby A, Berriman M, & Parkhill J (2009) DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics* 25(1):119-120.

212. Magoc T, Wood D, & Salzberg SL (2013) EDGE-pro: estimated degree of gene expression in prokaryotic genomes. *Evolutionary bioinformatics online* 9:127-136.

213. Rice P, Longden I, & Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* 16(6):276-277.

214. Andrews S (2010) FastQC: a quality-control tool for high-throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

215. Pearson WR, Wood T, Zhang Z, & Miller W (1997) Comparison of DNA Sequences with protein sequences. *Genomics* 46(1):24-36.

216. Ashburner M, *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics* 25(1):25-29.

217. Rattei T (2010) GenSkew: genomic nucleotide skew application http://genskew.csb.univie.ac.at/.

218. Nawrocki EP & Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29(22):2933-2935.

219. Darling ACE, Mau B, Blattner FR, & Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* 14(7):1394-1403.

220. Darling AE, Mau B, & Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* 5(6):e11147.

221. Kurtz S, *et al.* (2004) Versatile and open software for comparing large genomes. *Genome Biology* 5(2):R12.

222. Mi H, Muruganujan A, & Thomas PD (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research* 41(D1):D377-D386.

223. Mi H, Muruganujan A, Casagrande JT, & Thomas PD (2013) Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols* 8(8):1551-1566.

224. Gurevich A, Saveliev V, Vyahhi N, & Tesler G (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29(8):1072-1075.

225. R Development Core Team (2013) R: a language and environment for statistical computing. URL http://www.R-project.org/. (R Foundation for Statistical Computing, Vienna, Austria), version 3.0.0.

226. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, & Eddy SR (2003) Rfam: an RNA family database. *Nucleic Acids Research* 31(1):439-441.

227. Bankevich A*, et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19(5):455-477.

228. Stothard P (2000) The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *BioTechniques* 28(1102-1104).

229. Haft DH, Selengut JD, & White O (2003) The TIGRFAMs database of protein families. *Nucleic Acids Research* 31(1):371-373.

230. Krueger F (2013) Trim Galore! http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), version 0.3.5.

231. VSN International (2013) GenStat for Windows 17th Edition (VSN International, Hemel Hempstead, UK).

232. Yamagishi J-i, Yoshida H, Yamayoshi M, & Nakamura S (1986) Nalidixic acid-resistant mutations of the *gyrB* gene of *Escherichia coli*. *Molecular and General Genetics MGG* 204(3):367-373.

233. Yoshida H, Bogaki M, Nakamura M, & Nakamura S (1990) Quinolone resistance-determining region in the DNA gyrase *gyrA* gene of *Escherichia coli*. *Antimicrobial Agents and Chemotherapy* 34(6):1271-1272.

234. Becket E, Tse L, Yung M, Cosico A, & Miller JH (2012) Polynucleotide phosphorylase plays an important role in the generation of spontaneous mutations in *Escherichia coli*. *Journal of Bacteriology* 194(20):5613-5620.

235. Barrick JE (2015) Barrick Lab. http://barricklab.org/twiki/bin/view/Lab. (The University of Texas at Austin , Austin, Tx).

236. Miller JR, Koren S, & Sutton G (2010) Assembly algorithms for next-generation sequencing data. *Genomics* 95(6):315-327.

237. Lin R, Capage M, & Hill C (1984) A repetitive DNA sequence rhs responsible for duplications within the *Escherichia coli* K-12 chromosome. *Journal of molecular biology* 177:1 - 18.

238. Hill C (1999) Large genomic sequence repetitions in bacteria: lessons from rRNA operons and rhs elements. *Research in Microbiology* 150:665 - 674.

239. Jackson A, Thomas G, Parkhill J, & Thomson N (2009) Evolutionary diversification of an ancient gene family (rhs) through C-terminal displacement. *BMC genomics* 10(1):584.

240. Sims D, Sudbery I, Ilott NE, Heger A, & Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics* 15(2):121-132.

241. Koren S*, et al.* (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology* 30(7):693-700.

242. Bashir A*, et al.* (2012) A hybrid approach for the automated finishing of bacterial genomes. *Nature Biotechnology* 30(7):701-707.

243. Koren S*, et al.* (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing. *Genome Biology* 14(9):R101.

244. Wang JD & Levin PA (2009) Metabolism, cell growth and the bacterial cell cycle. *Nature Reviews Microbiology* 7(11):822-827.

245. Michelsen O, Teixeira de Mattos MJ, Jensen PR, & Hansen FG (2003) Precise determinations of C and D periods by flow cytometry in *Escherichia coli* K-12 and B/r. *Microbiology* 149(4):1001-1010.

246. Wright BE, Longacre A, & Reimers JM (1999) Hypermutation in derepressed operons of *Escherichia coli* K12. *Proceedings of the National Academy of Sciences of the United States of America* 96(9):5089-5094.

247. Frank AC & Lobry JR (1999) Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* 238(1):65-77.

248. Frederico LA, Kunkel TA, & Shaw BR (1990) A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry* 29(10):2532-2537.

249. Balbi KJ, Rocha EPC, & Feil EJ (2009) The temporal dynamics of slightly deleterious mutations in *Escherichia coli* and *Shigella spp*. *Molecular Biology and Evolution* 26(2):345-355.

250. Hildebrand F, Meyer A, & Eyre-Walker A (2010) Evidence of selection upon genomic GC-content in bacteria. *PLoS Genetics* 6(9):e1001107.

251. Sudarsan N, Wickiser JK, Nakamura S, Ebert MS, & Breaker RR (2003) An mRNA structure in bacteria that controls gene expression by binding lysine. *Genes & Development* 17(21):2688-2697.

252. Garcia-Diaz M & Kunkel TA (2006) Mechanism of a genetic glissando: structural biology of indel mutations. *Trends in Biochemical Sciences* 31(4):206-214.

253. Kelkar YD*, et al.* (2010) What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. *Genome Biology and Evolution* 2:620-635.

254. Tamas I*, et al.* (2008) Endosymbiont gene functions impaired and rescued by polymerase infidelity at poly(A) tracts. *Proceedings of the National Academy of Sciences* 105(39):14934-14939.

255. Wernegreen JJ, Kauppinen SN, & Degnan PH (2010) Slip into something more functional: selection maintains ancient frameshifts in homopolymeric sequences. *Molecular Biology and Evolution* 27(4):833-839.

256. Nagy Z & Chandler M (2004) Regulation of transposition in bacteria. *Research in Microbiology* 155(5):387-398.

257. Gunsalus RP & Park SJ (1994) Aerobic-anaerobic gene regulation in *Escherichia coli*: control by the ArcAB and Fnr regulons. *Research in Microbiology* 145(5–6):437-450.

258. Rocha EPC (2004) The replication-related organization of bacterial genomes. *Microbiology* 150(6):1609-1627.

259. Rocha EPC, Touchon M, & Feil EJ (2006) Similar compositional biases are caused by very different mutational effects. *Genome Research* 16(12):1537-1547.

260. Lobry JR (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Molecular Biology and Evolution* 13(5):660-665.

261. Grigoriev A (1998) Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Research* 26(10):2286-2290.

262. Mackiewicz P, Mackiewicz D, Kowalczuk M, & Cebrat S (2001) Flip-flop around the origin and terminus of replication in prokaryotic genomes. *Genome Biology* 2(12):interactions1004.1001-1004.

263. Rocha EPC (2008) The organization of the bacterial genome. *Annual Review of Genetics* 42:211-233.

264. Couturier E & Rocha EPC (2006) Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Molecular Microbiology* 59(5):1506-1518.

265. Lobry JR & Louarn J-M (2003) Polarisation of prokaryotic chromosomes. *Current Opinion in Microbiology* 6(2):101-108.

266. Duggin IG & Bell SD (2009) Termination structures in the *Escherichia coli* chromosome replication fork trap. *Journal of molecular biology* 387(3):532-539.

267. Esnault E, Valens M, Espeli O, & Boccard F (2007) Chromosome structuring limits genome plasticity in *Escherichia coli*. *PLoS Genetics* 3(12):e226.

268. Valens M, Penaud S, Rossignol M, Cornet F, & Boccard F (2004) Macrodomain organization of the *Escherichia coli* chromosome. *The EMBO Journal* 23(21):4330-4341.

269. Boccard F, Esnault E, & Valens M (2005) Spatial arrangement and macrodomain organization of bacterial chromosomes. *Molecular Microbiology* 57(1):9-16.

270. Espeli O, Mercier R, & Boccard F (2008) DNA dynamics vary according to macrodomain topography in the *E. coli* chromosome. *Molecular Microbiology* 68(6):1418-1427.

271. Foster PL, Hanson AJ, Lee H, Popodi EM, & Tang H (2013) On the mutational topology of the bacterial genome. *G3: Genes, Genomes, Genetics* 3(3):399-407.

272. Mugal CF, Von Grünberg H-H, & Peifer M (2009) Transcription-induced mutational strand bias and its effect on substitution rates in human genes. *Molecular Biology and Evolution* 26(1):131-142.

273. Yahara K, Didelot X, Ansari MA, Sheppard SK, & Falush D (2014) Efficient inference of recombination hot regions in bacterial genomes. *Molecular Biology and Evolution* 31(6):1593-1605.

274. Rogozin IB & Pavlov YI (2003) Theoretical analysis of mutation hotspots and their DNA sequence context specificity. *Mutation Research/Reviews in Mutation Research* 544(1):65-85.

275. Touchon M*, et al.* (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genetics* 5(1):e1000344.

276. Williams KP (2002) Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Research* 30(4):866-875.

277. Schneider BL, Kiupakis AK, & Reitzer LJ (1998) Arginine catabolism and the arginine succinyltransferase pathway in *Escherichia coli*. *Journal of Bacteriology* 180(16):4278-4286.

278. Weerasinghe J*, et al.* (2006) Stationary phase expression of the arginine biosynthetic operon argCBH in *Escherichia coli*. *BMC Microbiology* 6(1):14.

279. Navarro Llorens JM, Tormo A, & Martínez-García E (2010) Stationary phase in gram-negative bacteria. *FEMS Microbiology Reviews* 34(4):476-495.

280. Hanna MN, Ferguson RJ, Li Y-H, & Cvitkovitch DG (2001) *uvrA* is an acid-inducible gene involved in the adaptive response to low pH in *Streptococcus mutans*. *Journal of Bacteriology* 183(20):5964-5973.

281. Richard HT & Foster JW (2003) Acid resistance in *Escherichia coli*. *Advances in Applied Microbiology* 52:167-186.

282. Richard H & Foster JW (2004) *Escherichia coli* glutamate- and arginine-dependent acid resistance systems increase internal pH and reverse transmembrane potential. *Journal of Bacteriology* 186(18):6032-6041.

283. Castanie-Cornet M, Penfound T, Smith D, Elliott J, & Foster J (1999) Control of acid resistance in *Escherichia coli*. *Journal of Bacteriology* 181(11):3525 - 3535.

284. Trchounian A (2004) *Escherichia coli* proton-translocating F0F1-ATP synthase and its association with solute secondary transporters and/or enzymes of anaerobic oxidation–reduction under fermentation. *Biochemical and Biophysical Research Communications* 315(4):1051-1057.

285. Kannan G*, et al.* (2008) Rapid acid treatment of *Escherichia coli*: transcriptomic response and recovery. *BMC Microbiology* 8(1):37.

286. Stincone A*, et al.* (2011) A systems biology approach sheds new light on *Escherichia coli* acid resistance. *Nucleic Acids Research* 39(17):7512-7528.

287. Connelly JC, De Leau ES, & Leach DRF (1999) DNA cleavage and degradation by the SbcCD protein complex from *Escherichia coli*. *Nucleic Acids Research* 27(4):1039-1046.

288. Connelly JC, Kirkham LA, & Leach DRF (1998) The SbcCD nuclease of Escherichia coli is a structural maintenance of chromosomes (SMC) family protein that cleaves hairpin DNA. *Proceedings of the National Academy of Sciences of the United States of America* 95(14):7969-7974.

289. Eykelenboom JK, Blackwood JK, Okely E, & Leach DRF (2008) SbcCD causes a double-strand break at a DNA palindrome in the *Escherichia* coli chromosome. *Molecular Cell* 29(5):644-651.

290. Bustin SA*, et al.* (2009) The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clinical Chemistry* 55(4):611-622.

291. Livak KJ & Schmittgen TD (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* 25(4):402-408.

292. Savli H*, et al.* (2003) Expression stability of six housekeeping genes: a proposal for resistance gene quantification studies of *Pseudomonas aeruginosa* by real-time quantitative RT-PCR. *Journal of Medical Microbiology* 52(5):403-408.

293. Takle G, Toth I, & Brurberg M (2007) Evaluation of reference genes for real-time RT-PCR expression studies in the plant pathogen Pectobacterium atrosepticum. *BMC Plant Biology* 7(1):50.

294. Brøndsted L & Atlung T (1996) Effect of growth conditions on expression of the acid phosphatase (cyx-appA) operon and the appY gene, which encodes a transcriptional activator of *Escherichia coli*. *Journal of Bacteriology* 178(6):1556-1564.

295. Lundrigan MD & Earhart CF (1984) Gene envY of *Escherichia coli* K-12 affects thermoregulation of major porin expression. *Journal of Bacteriology* 157(1):262-268.

296. Le Gac M, Plucain J, Hindré T, Lenski RE, & Schneider D (2012) Ecological and evolutionary dynamics of coexisting lineages during a long-term experiment with *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America* 109(24):9487-9492.

297. Zientz E, Bongaerts J, & Unden G (1998) Fumarate regulation of gene expression in *Escherichia coli* by the DcuSR (dcuSR genes) two component regulatory system. *Journal of Bacteriology* 180(20):5421-5425.

298. Kim HJ*, et al.* (2014) Short-term differential adaptation to anaerobic stress via genomic mutations by *Escherichia coli* strains K-12 and B lacking alcohol dehydrogenase. *Frontiers in microbiology* 5:476.

299. Wolfe AJ (2005) The acetate switch. *Microbiology and Molecular Biology Reviews* 69(1):12-50.

300. Treves DS, Manning S, & Adams J (1998) Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of *Escherichia coli*. *Molecular Biology and Evolution* 15(7):789-797.

301. Baisa G, Stabo NJ, & Welch RA (2013) Characterization of *Escherichia coli* d-cycloserine transport and resistant mutants. *Journal of Bacteriology* 195(7):1389-1399.

302. Campbell JW, Morgan-Kiss RM, & E. Cronan J (2003) A new *Escherichia coli* metabolic competency: growth on fatty acids by a novel anaerobic β-oxidation pathway. *Molecular Microbiology* 47(3):793-805.

303. Puentes-Téllez PE, Hansen MA, Sørensen SJ, & van Elsas JD (2013) Adaptation and heterogeneity of *Escherichia coli* MC1000 growing in complex environments. *Applied and Environmental Microbiology* 79(3):1008-1017.

304. Koskiniemi S, Sun S, Berg OG, & Andersson DI (2012) Selection-driven gene loss in bacteria. *PLoS Genetics* 8(6):e1002787.

305. Helling RB, Vargas CN, & Adams J (1987) Evolution of *Escherichia coli* during growth in a constant environment. *Genetics* 116(3):349-358.

306. Rosenzweig RF, Sharp RR, Treves DS, & Adams J (1994) Microbial evolution in a simple unstructured environment: genetic differentiation in *Escherichia coli*. *Genetics* 137(4):903-917.

307. Lind PA & Andersson DI (2008) Whole-genome mutational biases in bacteria. *Proceedings of the National Academy of Sciences* 105(46):17878-17883.

308. Martincorena I, Seshasayee ASN, & Luscombe NM (2012) Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature* 485(7396):95-98.

309. Fijalkowska IJ, Jonczyk P, Tkaczyk MM, Bialoskorska M, & Schaaper RM (1998) Unequal fidelity of leading strand and lagging strand DNA replication on the *Escherichia coli* chromosome. *Proceedings of the National Academy of Sciences* 95(17):10020-10025.

Appendices

Appendix A

Table A.1. Examples of oxidative lesions and the mutations they produce. Table adapted from (25).

Nucleotide	DNA Lesion Produced	Mutation
Guanine		
	8-oxoguanine	G → T
	Cyanuric acid	G →T
	Oxaluric acid	G → T
	Oxazalone	G → T
	2,6-Diamino-4-hydroxy-5-formamidopyrimidine	G → T
	Spiroiminodihydantoin	G → C
	Imidazolone	G → C
Adenine		
	2-hydroxyadenine	A → T
Thymine		
	Thymine glycol	T → C
	Formylamine	T → C
	Urea	T → C
Cytosine		
	Uracil glycol	C → T
	5-hydroxyuracil	C → T

Table A.2. Expression values of all genes in REL4536 known to be involved in DNA repair and replication.

Gene	Description	Fold Change[†]	P-adj[*]
Base excision repair pathway			
alkA	3-methyl-adenine DNA glycosylase II	-1.86	4.45×10^{-8}
mug	Stationary phase uracil DNA glycosylase	-1.61	0.002
mutM	Formamidopyrimidine DNA glycosylase	2.19	9.92×10^{-7}
mutT	Nucleotide hydrolysis glycosylase	1.08	0.79
mutY	Adenine DNA glycosylase	-1.88	1.66×10^{-8}
nei	Endonuclease VIII	1.13	0.47
nfi	Endonuclease V	4.00	7.67×10^{-20}
nfo	Endonuclease IV	2.03	5.03×10^{-5}
Nth	Endonuclease III	2.56	3.83×10^{-7}
tag	3-methyl-adenine DNA glycosylase I	1.03	0.87
ung	Uracil DNA glycosylase	2.18	1.1×10^{-13}
xthA	Exonuclease III	-1.20	0.22
Mismatch repair pathway			
dam	DNA adenine methyltransferase	2.38	1.42×10^{-14}
ExoVII	Exonuclease VII	4.32	8.35×10^{-31}
exoX	Exonuclease X	2.45	6.91×10^{-19}
mutH	Endonuclease MutH	1.98	2.34×10^{-4}
mutL	Molecular matchmaker MutL	-1.00	0.98
mutS	Repair initiator MutS	5.74	2.44×10^{-30}
recJ	Exonuclease RecJ	1.32	0.013
sbcB	Exonuclease I	2.04	2.51×10^{-12}
ssb	Single-stranded DNA binding protein	-2.13	2.34×10^{-11}
xseA	Exonuclease VII	-1.72	0.001
ybcN	G:T mismatch repair protein	-1.75	0.11
Nucleotide excision repair pathway			
mfd	DNA repair endonuclease	2.86	7.92×10^{-24}
uvrA	UvrA exinuclease	1.49	0.001
uvrB	UvrB exinuclease	1.23	0.053
uvrC	UvrC exinuclease	-1.28	0.014
uvrD	Helicase II, also involved in mismatch repair	2.07	1.69×10^{-11}
ydjQ	DNA repair endonuclease also SOS inducible	2.85	1.99×10-^{24}
Recombinational repair			
helD	DNA helicase IV in RecF pathway	-3.46	6.17×10^{-17}
insA-1	IS*1* protein	1.21	0.34

Gene	Description	Fold Change†	P-adj*
insA-2	IS*1* protein	1.23	0.29
insA-3	IS*1* protein	-1.05	0.79
insA-4	IS*1* protein	1.16	0.39
insA-5	IS*1* protein	1.29	0.08
insA-6	IS*1* protein	1.29	0.15
insA-7	IS*1* protein	1.18	0.21
insA-8	IS*1* protein	1.21	0.24
insA-9	IS*1* protein	1.16	0.36
insA-10	IS*1* protein	1.08	0.68
insA-11	IS*1* protein	1.08	0.68
insA-12	IS*1* protein	1.21	0.24
insA-13	IS*1* protein	1.22	0.23
insA-15	IS*1* protein	1.19	0.36
insA-16	IS*1* protein	0.99	0.98
insA-17	IS*1* protein	1.15	0.38
insA-18	IS*1* protein	1.23	0.20
insA-19	IS*1* protein	1.19	0.32
insA-20	IS*1* protein	3.72	3.43×10^{-20}
insA-21	IS*1* protein	1.18	0.33
insA-22	IS*1* protein	1.20	0.31
insA-23	IS*1* protein	1.22	0.24
insA-24	IS*1* protein	1.02	0.89
insA-25	IS*1* protein	6.14	6.19×10^{-34}
insA-26	IS*1* protein	1.22	0.27
insA-27	IS*1* protein	1.21	0.27
insA-28	IS*1* protein	1.28	0.13
insB-1	IS*1* protein	1.49	0.11
insB-2	IS*1* protein	1.92	0.01
insB-3	IS*1* protein	1.83	0.008
insB-4	IS*1* protein	3.17	8.24×10^{-14}
insB-5	IS*1* protein	2.74	7.06×10^{-17}
insB-6	IS*1* protein	2.68	4.94×10^{-8}
insB-7	IS*1* protein	2.62	3.77×10^{-7}
insB-8	IS*1* protein	3.15	2.02×10^{-14}
insB-9	IS*1* protein	2.45	8.33×10^{-11}
insB-10	IS*1* protein	1.83	0.001
insB-11	IS*1* protein	2.82	3.84×10^{-9}
insB-12	IS*1* protein	2.81	8.06×10^{-9}
insB-13	IS*1* protein	2.71	5.42×10^{-8}
insB-15	IS*1* protein	3.20	2.57×10^{-14}

Gene	Description	Fold Change†	P-adj*
insB-16	IS*1* protein	1.66	0.005
insB-17	IS*1* protein	2.90	3.76×10^{-9}
insB-18	IS*1* protein	2.14	6.97×10^{-5}
insB-19	IS*1* protein	2.41	3.17×10^{-5}
insB-20	IS*1* protein	2.07	0.005
insB-21	IS*1* protein	1.84	0.001
insB-22	IS*1* protein	1.76	0.003
insB-23	IS*1* protein	2.20	4.69×10^{-5}
insB-24	IS*1* protein	1.50	0.04
insB-26	IS*1* protein	1.88	8.54×10^{-4}
insB-27	IS*1* protein	1.86	0.001
insB-28	IS*1* protein	1.68	0.02
insD	IS*2* protein	1.33	0.13
insE-1	IS*3* protein	-1.44	0.06
insE-2	IS*3* protein	-1.34	0.13
insE-3	IS*3* protein	-1.74	0.003
insE-4	IS*3* protein	-1.49	0.05
insE-5	IS*3* protein	-1.96	1.19×10^{-4}
insF-1	IS*3* protein	1.41	0.02
insF-2	IS*3* protein	1.50	0.003
insF-3	IS*3* protein	1.39	0.02
insF-4	IS*3* protein	1.03	0.86
insF-5	IS*3* protein	1.35	0.02
insG	*IS4* protein	-1.73	3.76×10^{-5}
insI	IS*30* protein	1.50	0.02
insJ-1	IS*150* protein	1.69	0.01
insJ-2	IS*150* protein	1.83	0.006
insJ-3	IS*150* protein	1.93	0.004
insJ-4	IS*150* protein	1.89	0.006
insJ-5	IS*150* protein	1.91	0.004
insK-2	IS*150* protein	1.49	0.02
insK-3	IS*150* protein	1.11	0.61
insK-4	IS*150* protein	1.45	0.04
insK-5	IS*150* protein	2.24	2.32×10^{-7}
insL-1	IS*186* protein	1.04	0.84
insL-2	IS*186* protein	-1.13	0.44
insL-3	IS*186* protein	-1.03	0.82
insL-4	IS*186* protein	-1.13	0.44
insL-5	IS*186* protein	-1.07	0.65
yis1	IS*600* protein	1.19	0.66

Gene	Description	Fold Change[†]	P-adj[*]
yis2	IS*600* protein	1.09	0.55
insN	IS*911* protein	1.90	4.92×10^{-4}
intR	Integrase	-2.86	6.12×10^{-15}
pepA	Aminopeptidase A/I	-2.07	1.31×10^{-14}
recA	DNA recombination protein	-1.04	0.81
recB	Component of the RecBCD complex	1.98	1.01×10^{-8}
recC	Component of the RecBCD complex	1.58	7.03×10^{-6}
recD	Component of the RecBCD complex	1.41	0.039
recF	DNA helicase	-1.19	0.23
recG	DNA helicase for double stranded DNA repair	-1.41	0.045
recN	Protein involved in DSB repair	2.26	6.16×10^{-8}
recO	Component of RecFOR complex	1.39	0.038
recQ	DNA helicase involved in RecF recombination	2.47	2.88×10^{-9}
recR	Component of RecFOR complex	-2.79	3.01×10^{-11}
recT	Recombinase in RecE recombination pathway	2.87	1.21×10^{-5}
rus	Holliday junction endonuclease	1.34	0.52
ruvA	Component of RuvAB complex	2.44	5.63×10^{-10}
ruvB	Component of RuvAB complex	1.26	0.055
ruvC	Holliday junction nuclease	1.50	0.004
sbcC	Double stranded DNA exonuclease	2.51	3.72×10^{-12}
sbcD	Double stranded DNA exonuclease	2.76	2.63×10^{-12}
topB	DNA topoisomerase III	1.72	1.69×10^{-5}
xerC	Recombinase protein XerC	3.58	2.88×10^{-15}
xerD	Recombinase protein XerD	-1.24	0.080
yigN	Putative recombination limiting protein	5.48	9.53×10^{-57}
yqgF	Predicted transcription antitermination factor	2.53	1.05×10^{-7}
DNA replication			
dinB	DNA polymerase IV for translesion synthsis	-1.04	0.77
dinG	Helicase, SOS inducible	1.53	0.002
dnaA	Replication initiation protein	1.22	0.077
dnaB	DNA helicase	6.03	1.16×10^{-56}
dnaC	DNA replication initiation protein	-1.68	2.64×10^{-5}
dnaE	DNA polymerase III	-1.28	0.051
dnaG	DNA primase	-2.46	1.33×10^{-19}
dnaJ	DNA chaperone protein	1.62	8.81×10^{-27}
dnaK	DNA chaperone protein	2.38	1.53×10^{-77}
dnaN	DNA polymerase III beta subunit	-1.00	0.98
dnaQ	DNA polymerase III epsilon subunit	-1.55	2.22×10^{-5}
dnaT	Replication protein, SOS inducible	-1.93	6.66×10^{-8}
dnaX	DNA polymerase III gamma&tau subunit	3.23	9.14×10^{-26}

Gene	Description	Fold Change[†]	P-adj[*]
hda	Replication initiation factor	2.28	6.11×10^{-5}
holA	DNA polymerase III subunit	1.46	0.033
holB	DNA polymerase III delta subunit	1.99	3.96×10^{-7}
holC	DNA polymerase III subunit	-1.09	0.53
holD	DNA polymerase III subunit	4.83	5.84×10^{-25}
holE	DNA polymerase III theta subunit	3.39	6.41×10^{-11}
ligA	DNA ligase	-2.47	2.65×10^{-20}
ligB	DNA ligase	2.17	7.12×10^{-9}
mukB	Cell division protein involved in partitioning	1.08	0.52
mukE	Cell division protein involved in partitioning	3.93	4.22×10^{-27}
mukF	Cell division protein involved in partitioning	2.42	1.17×10^{-8}
nrdA	Ribonucleoside diphosphate reductase 1	1.76	2.70×10^{-4}
nrdB	Ribonucleoside diphosphate reductase 1	1.48	0.004
nrdD	Ribonucleoside-triphosphate	2.23	1.08×10^{-5}
nrdE	Ribonucleoside-diphosphate reductase 2	2.42	2.92×10^{-15}
nrdF	Ribonucleoside-diphosphate reductase 2	2.00	4.45×10^{-7}
pioO	Calcium-binding protein	1.38	0.078
polA	DNA polymerase I	-1.48	0.002
polB	DNA polymerase II for translesion synthsis	2.39	1.02×10^{-13}
priA	Primosomal replication factor	1.40	0.026
priB	Primosomal replication factor	-1.67	0.045
priC	Primosomal replication factor	-1.11	0.49
rep	Helicase that prevents DSBs	4.11	1.06×10^{-15}
seqA	Replication regulator	5.44	1.93×10^{-38}
tdk	Thymidine kinase/deoxyuridine kinase	-2.41	2.99×10^{-7}
tus	Replication terminator	-3.69	3.58×10^{-36}
umuC	DNA polymerase V subunit	1.28	0.16
umuD	DNA polymerase V subunit	1.29	0.075
ycaJ	Protein involved in recombination	1.34	0.008
yraO	DnaA initiator-associating factor	1.22	0.14
SOS response			
dinD	DNA damage inducible protein	3.35	4.03×10^{-14}
dinF	Member of the MATE family of multidrug efflux transporters	1.29	0.22
dinI	DNA damage-inducible protein I	-1.26	0.30
ftsK	Cell division protein	-1.12	0.19
lexA	Transcription repressor LexA	-2.11	3.07×10^{-5}
recX	Inhibitor of RecA	-1.07	0.80
rimK	ribosomal protein S6 modification protein	-2.36	1.17×10^{-7}
sulA	Cell division inhibitor SulA	-3.88	8.56×10^{-32}

Gene	Description	Fold Change[†]	P-adj[*]
yafP	Predicted acyltransferase with acyl-CoA N-acyltransferase domain	1.53	0.008
ydjM	Predicted inner membrane protein,	-1.54	0.004
yebG	DNA damage-inducible protein	1.21	0.22
yjiW	Toxin-like protein of the SOS response	3.02	3.67×10^{-16}
Stringent response			
ahpC	Alkyl hydroperoxide reductase	1.75	0.004
appA	Acid phosphatase	1.68	0.009
appY	DNA-binding transcriptional activator	-1.05	1.00
glnG	NtrC transcriptional dual regulator	-2.17	1.91×10^{-4}
leuB	3-isopropylmalate dehydrogenase	1.28	0.65
mazG	Nucleoside triphosphate pyrophosphohydrolase	2.20	1.19×10^{-4}
phoQ	Sensory histidine kinase	-1.34	0.18
phoR	Sensory histidine kinase	-1.28	0.29
phoU	Negative regulator of the *Pho* regulon	1.17	0.58
rplK	50S ribosomal subunit protein L11	2.17	6.13×10^{-5}
ssuE	NAD(P)H-dependent FMN reductase	1.66	0.30
tas	Putative NAD(P)-linked reductase that acts in starvation-associated mutation	3.73	0.006
tauA	Taurine ABC transporter - periplasmic binding protein	1.21	0.48
tauB	Taurine ABC transporter - ATP binding subunit	1.47	0.20
tauC	Taurine ABC transporter - membrane subunit	-1.46	0.25
ycgW	Sigma factor inhibitor	2.10	0.25
yhgI	Iron-sulfur cluster scaffold protein	1.73	0.09
yibD	UDP-glucuronate:LPS(HepIII) glycosyltransferase	-1.53	0.15
yjbA	Predicted phosphate starvation-inducible protein	2.37	5.64×10^{-4}
yjiY	Inner membrane protein	2.04	3.87×10^{-4}
yodA	Cadmium-induced cadmium binding protein	-2.54	0.007

[†] Fold change is calculated as the number of anaerobic reads over aerobic reads, and negative values indicate up-regulation in aerobic conditions

[*] To identify significant expression, a *p*-adj value < 0.05 was used

Appendix B

```
#MA simulator programme code

nmuts.max  <-  30 ##  maximum  number  of  mutations  you're
tracking in a single bacterium

nsims <- 10000 ## number of simulation runs

ngens <- 24  ## number of generations per bottleneck event
(excluding first generation)

mutrate <- 0.00046  ## mutation rate

nbottlenecks <- 180  ## number of bottleneck events

#### functions

nextgen <- function(pop, mutrate, nmuts.lim = 100, largeN =
FALSE) {

#### pop: vector of counts of bacteria with respectively 0,
1, 2... mutations

#### mutrate: mutation rate

#### nmuts.lim: absolute maximum number of mutations

#### largeN: logical, should be TRUE if the population size
goes over 2 ^ 14, because

####  rmultinom crashes if its argument "size" is too large

nmuts <- length(pop) - 1   ## maximum number of mutations
you're tracking in a single bacterium

p <- dpois(0 : nmuts, mutrate) ## probability of observing
0, 1, 2, ... new mutations for a bacterium (from Poisson
distribution)

pop.new <- rep(0, nmuts.lim)   ## new mutation counts after
reproduction AND mutation!

pop  <-  2  *  pop  ##  population  doubles,  all  counts  are
doubled

for  (i  in  1  :  (nmuts  +  1))  {    ##  loop  through  bacteria
with increasing number of mutations

        if (pop[i] > 0) {
            if (largeN) {          ## the code below draws
new counts based on a multinomial distribution (with
probabilities of each

                N <- pop[i]        ## number of mutations
coming from the Poisson distribution) and adds the counts
in pop.new

                x <- vector(length = nmuts + 1)

                for (j in 1 : (nmuts + 1)) {

                    if (N > 0) {

                        if (j > 1) {

                            px <- p[j] / (1 - sum(p[1 : (j
- 1)]))

                        } else {

                            px <- p[1]

                        }

                        x[j] <- rbinom(1, N, px)

                        N <- N - x[j]

                    }

                }

            } else {

                x <- rmultinom(1, pop[i], dpois(0 : nmuts,
mutrate))

            }

            pop.new[i : (i + nmuts)] <- pop.new[i : (i +
nmuts)] + x

        }

    }

      pop.new[nmuts + 1] <- pop.new[nmuts + 1] +
sum(pop.new[(nmuts + 2) : nmuts.lim]) ## counts of bacteria
with more than
```

```
        pop.new[(nmuts + 2) : nmuts.lim] <-
0                                        ## the number
of mutations tracked are collapsed and added to the

              pop.new[1 : (nmuts +
1)]
 ## last mutation count tracked

}

pop <- array(dim = c(ngens + 1, nmuts.max + 1, nbottlenecks
+ 1, nsims)) ## array of counts, for each gen, each number
of mutation tracked (0 to nmuts.max),

## each bottleneck and each simulation run

for (sim in 1 : nsims) {                         ## looping
through simulation runs

if (sim %% 100 == 0) print(sim)

    pop[1, , 1, sim] <- c(1, rep(0, nmuts.max))

     for (gen in 2 : (ngens + 1)) {           ## looping
through generations

         pop[gen, , 1, sim] <- nextgen(pop = pop[gen - 1, ,
1, sim], mutrate = mutrate, largeN = TRUE)       ##
generating new counts

    }

     for (b in 2 : (nbottlenecks + 1)) {      ## looping
through bottlenecks

         selected <- sample(1 : (nmuts.max + 1), 1, prob =
pop[ngens + 1, , b - 1, sim])  ## selecting the bacterium
to start after bottleneck

        pop[1, , b, sim] <- rep(0, nmuts.max + 1)

        pop[1, selected, b, sim] <- 1

         for (gen in 2 : (ngens + 1)) {       ## looping
through generations

             pop[gen, , b, sim] <- nextgen(pop = pop[gen -
1, , b, sim], mutrate = mutrate, largeN = TRUE) ##
generating new counts

         }

    }
```

```
}

save.image("pop120_10000dataan.RData")

min.mut <- function(x) { min(which(x>0)) - 1 }



apply(pop[ngens+1, , nbottlenecks+1, ],2,min.mut)

median(apply(pop[ngens+1, , nbottlenecks+1,
],2,min.mut)) # Minimum number of mutations per
simulation at nbottlenecks and ngenserroreree.

png(filename = "Rplot%03ddata.png")

hist(apply(pop[ngens+1, , nbottlenecks+1, ],2,min.mut),
xlab = "minimum number of mutations",

main = paste("Number of mutations after", nbottlenecks,
"bottlenecks,

total bacteria =",sum(pop[ngens+1,,100+1,nsims]), ", ngens
= ", ngens, "

nsims = ", nsims), breaks = -1:30 ) # Or do the breaks
until nmuts.max

dev.off()
```

Appendix C

Table A.3. List of all mutations detected in aerobic and anaerobic MA clone genomes.

AE-180-02

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
Indel	Deletion	1,774,108	1,774,152	44	-	-	*ycaC*	Putative hydrolase	*dmsC*	Dimethyl sulfoxide reductase, anaerobic, sub-unit C	-

AE-180-04

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	516,399	516,399	1	C → T	Ser 345 Asn	*allC*	Allantoate amido-hydrolase	-	-	-
SNP	Transition	1,436,448	1,436,448	1	A → G	-	*oppA*	Oligopeptide transporter subunit	*ychE*	Hypothetical protein	Intergenic
SNP	Transversion	3,162,607	3,162,607	1	T → G	Lys 399 Thr	*yhaO*	Putative transporter	-	-	-
SNP	Transversion	3,511,575	3,511,575	1	T → A	Leu 544 Lys	*zntA*	Zinc/cadmium/mercury/lead transporting ATPase	-	-	-
SNP	Transition	4,062,788	4,062,788	1	C → T	Gly 220 Asp	*glpF*	Glycerol facilitator	-	-	-
Indel	Deletion	10,311	10,312	1	-C	-	*yaaH*	Hypothetical protein	-	-	-
Indel	Insertion	4,100,614	4,100,649	35	-	-	*argE*	Acetylornithine deacetylase	*argC*	N-acetyl-gamma glutamyl-phosphate reductase	Intergenic
GCR	IS insertion	16,992	16,994	1,446	-	-	*mokC*	Regulatory protein for HokC	*nhaA*	pH-dependent proton antiporter	Intergenic IS*150* (+) + 3 bp insertion
GCR	IS mediated deletion	1,117,789	1,125,298	7,509	-	-	*ECB_01527*	Outer membrane lipoprotein Blc	*ECB_01510*	Putative tail component of prophage	IS*3* mediated 13 genes deleted[†]
GCR	IS insertion	3,988,830	3,988,838	777	-	-	*yihF*	Hypothetical protein	-	-	IS*1* (-) + 9 bp insertion

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
GCR	IS insertion	2,421,291	2,421,294	1,349	-	-	*nupC*	Nucleoside transporter	*yfeA*	Putative diguanylate cyclase	Intergenic IS*186* (-) + 6 bp insertion
GCR	Inversion‡	578,932	584,572	~5.6 kb	-	-	*ybdK*	Carboxylate-amine ligase	*insL-3*	IS*186* hypothetical protein	IS*186* mediated
GCR	Inversion‡	583,943	731,438	~150 kb	-	-	*entD*	Phospho-pantetheinyl-transferase component of enterobactin synthase multienzyme complex	*erfK*	Hypothetical protein	-
GCR	Translocation‡	736,256	2,420,089	~1.7 Mb	-	-	*yeeN*	Hypothetical protein	*nupC*	Nucleoside transporter	-

†Deleted genes encode for hypothetical proteins, a putative pathogenicity island protein, a putative AraC-type regulatory protein, a putative S lysis protein, a putative lysozyme, cold shock proteins and an IS3 element.

‡Depicted in **Figure 4.8**.

AE-180-06

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	1,483,117	1,483,117	1	G → A	Thr 130 Met	*dhaR*	DNA-binding transcriptional regulator	-	-	-
SNP	Transition	2,951,909	2,951,909	1	G → A	Glu 134 Lys	*yqgE*	Hypothetical protein	-	-	-
SNP	Transition	3,411,614	3,411,614	1	G → A	Ser 284 Phe	*yhfX*	Putative amino acid racemase	-	-	-
SNP	Transition	4,224,073	4,224,073	1	G → A	Ala 263 Ala	*yjcD*	Putative permease	-	-	-
GCR	Inversion[†]	430,943	582,472	~145 kb	-	-	*lon*	DNA-binding ATP-dependent protease La	*hokE*	Small toxic polypeptide	IS*186* mediated inversion
GCR	IS insertion	2,421,323	2,421,326	1,347	-	-	*nupC*	Nucleoside (except guanosine) transporter	*yfeA*	Putative diguanylate cyclase	Intergenic IS*186* (-) + 4 bp insertion
GCR	IS insertion	4,381,583	4,381,585	1,472	-	-	*cycA*	D-alanine/ D-serine/ glycine permease	-	-	IS*150* (-) + 29 bp insertion

[†] Depicted in **Figure 4.8**.

AE-180-08

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	382,017	382,017	1	C → T	Arg 720 Arg	*sbcC*	Exonuclease subunit SbcC	-	-	-
SNP	Transversion	1,085,806	1,085,806	1	A → C	Ala 192 Ala	*ynfM*	Putative transporter	-	-	-
SNP	Transition	1,179,190	1,179,190	1	C → T	Gly 219 Gly	*ydeM*	Hypothetical protein	-	-	-
SNP	Transversion	1,398,532	1,398,532	1	G → T	Arg 663 Arg	*acnA*	Aconitate hydratase	-	-	-
SNP	Transition	4,227,885	4,227,885	1	A → G	-	*yjcF*	Hypothetical protein	*actP*	Acetate permease	Intergenic
SNP	Transversion	4,340,894	4,340,894	1	A → C	Ile 268 Leu	*amiB*	N-acetyl-muramoyl- L alanine amidase II	-	-	-
SNP	Transition	4,406,122	4,406,122	1	G → A	Gln 405 Gln	*mpl*	UDP-N-acetyl muramate: L-alanyl-gamma-D-glutamyl-mesodiamino-pimelate ligase	-	-	-
GCR	IS insertion	1,774,157	1,774,162	1,349	-	-	*ycaC*	Putative hydrolase	*dmsC*	Dimethyl sulfoxide reductase subunit C	Intergenic IS*186* (-) + 6 bp insertion
GCR	IS insertion	2,295,162	2,295,167	1,349	-	-	*menC*	O-succinyl-benzoate synthase	-	-	IS*186* (-) + 6 bp insertion

AE-180-10

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	2,096,853	2,096,853	1	C → T	Trp 57*	*pagP*	Palmitoyl transferase	-	-	*Protein truncation
SNP	Transition	2,130,355	2,130,355	1	C → T	Gln 215*	*yehM*	Hypothetical protein	-	-	*Protein truncation
SNP	Transition	3,082,154	3,082,154	1	G → A	His 79 His	*parE*	DNA topoisomerase IV subunit B	-	-	-
SNP	Transition	3,875,202	3,875,202	1	G → A	Thr 147 Ile	*hdfR*	Transcriptional regulator	-	-	-
Indel	Insertion	3,625,341	3,625,342	1	+C	-	*insK-4*	IS*150* protein	*glyS*	Glycyl-tRNA synthetase subunit beta	Intergenic C(1) → C(2)

AE-180-12

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	2,693,118	2,693,118	1	A → G	Val 47 Ala	*mltB*	Murein hydrolase B	-	-	-
Indel	Deletion	2,972,159	2,972,172	13	-	-	*ECB_ 02798*	Hypothetical protein	*flu*	Antigen 43 (Ag43) phase variable biofilm formation auto-transporter	Intergenic
Indel	Slippage	3,866,358	3,866,359	1	-G	-	*trkD*	Potassium transport protein Kup	*insJ-5*	IS*150* protein	Intergenic G(2) → G(1)

AE-180-14

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transversion	1,625,077	1,625,077	1	G → C	Pro 162 Ala	*ycdT*	Putative diguanylate cyclase	-	-	-
SNP	Transition	4,391,221	4,391,221	1	A → G	Tyr 196 His	*msrA*	Methionine sulfoxide reductase	-	-	-
SNP	Transition	4,562,374	4,562,374	1	C → T	Gly 145 Asp	*rsmC*	16S ribosomal RNA methyl-transferase	-	-	-
GCR	IS mediated deletion	1,117,789	1,125,298	7,509	-	-	*ECB_01527*	Outer membrane lipoprotein Blc	*ECB_01510*	Putative tail component of prophage	IS*3* mediated 13 genes deleted[†]

[†] Deleted genes encode for hypothetical proteins, a putative pathogenicity island protein, a putative AraC-type regulatory protein, a putative S lysis protein, a putative lysozyme, cold shock proteins and an IS3 element.

AE-180-16

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
Indel	Deletion	2,426,705	2,426,871	182	-	-	*valU*	tRNA-Val	*valY*	tRNA-Val	*valX* encoding tRNA-Val deleted

AE-180-20

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transversion	150,907	150,907	1	T → A	Gln 231 Leu	*panC*	Pantoate-beta-alanine ligase	-	-	-
SNP	Transition	639,501	639,501	1	C → T	Cys 257 Cys	*ECB_02003*	Hypothetical protein	-	-	-
SNP	Transversion	647,341	647,341	1	G → T	Ser 78 Ile	*ECB_01994*	Hypothetical protein	-	-	-
SNP	Transversion	2,187,475	2,187,475	1	G → C	Pro 11 Ala	*fruB*	Bifunctional PTS system fructose specific transporter subunit IIA/HPr protein	-	-	-
SNP	Transversion	2,318,314	2,318,314	1	G → T	Pro 51 Gln	*nuoE*	NADH de-hydrogenase subunit E	-	-	-
SNP	Transversion	2,670,786	2,670,786	1	A → C	Tyr 622 Ser	*nrdE*	Ribonucleotide diphosphate reductase subunit alpha	-	-	-
Indel	Deletion	1,103,198	1,103,199	1	-G	-	*rspB*	Putative dehydrogenase	*ECB_01546*	Excisionase	Intergenic

AE-180-22

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	387,797	387,797	1	A → G	-	*phoR*	Phosphate regulon sensor protein	*brnQ*	Putative branched chain amino acid transporter	Intergenic
SNP	Transversion	4,181,661	4,181,661	1	T → G	Tyr 199 Asp	*yjbF*	Putative lipoprotein	-	-	-
Indel	Deletion	1,973,806	1,973,933	127	-	-	*lysQ*	tRNA-Lys	*lysZ*	tRNA-Lys	Intergenic
Indel	Slippage	3,866,357	3,866,358	1	+G	-	*trkD*	Potassium transport protein Kup	*insJ-5*	IS*150* protein	Intergenic G(2) → G(3)
GCR	IS mediated deletion	547,702	551,543	3,841	-	-	*ybcQ*	Putative anti-termination protein	*ompT*	Outer membrane protease	IS*1* mediated 5 genes deleted†
GCR	IS insertion	551,542	-	777	-	-	*appY*	DNA-binding transcriptional activator	-	-	IS*1* insertion required to mediate deletion
GCR	IS insertion	1,272,948	-	777	-	-	*cybB*	Cytochrome b561 peptide	-	-	IS*1* insertion required to mediate deletion
GCR	IS mediated deletion	1,272,948	1,273,794	846	-	-	*ydcA*	Hypothetical protein	*gapC*	Glycer-aldehyde3-phosphate dehydrogenase	IS*1* mediated *cybB* encoding for cytochrome b561 peptide deleted

†Deleted genes encode for hypothetical proteins and a DNA-binding transcriptional activator.

AE-180-24

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transversion	3,199,393	3,199,393	1	A → C	-	*yraJ*	Putative outer membrane protein	*yraK*	Putative fimbrial-like adhesin protein	Intergenic
SNP	Transition	4,594,928	4,594,928	1	A → G	-	*yjjY*	Hypothetical protein	*lasT*	Putative RNA methyl-transferase	Intergenic
Indel	Slippage	3,866,357	3,866,358	1	+G	-	*trkD*	Potassium transport protein Kup	*insJ-5*	IS*150* protein	Intergenic G(2) → G(3)
GCR	IS insertion	1,272,453	1,272,455	1,473	-	-	*trg*	Chemotaxis protein III, ribose and galactose sensor receptor	*mokB*	Regulatory peptide	Intergenic IS*150* (-) + 30 bp insertion
GCR	Inversion[†]	997,452	2,069,680	~1.3 Mb	-	-	*pykF*	Pyruvate kinase	*gltK*	Glutamate and aspartate transporter subunit	IS*150* mediated inversion around the terminus

[†] Depicted in **Figure 4.8**.

AE-180-26

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	3,869,337	3,869,337	1	A → G	-	*yieP*	Predicted transcriptional regulator	*rrsC*	16S ribosomal RNA	Intergenic
Indel	Slippage	3,866,357	3,866,358	1	-G	-	*trkD*	Potassium transport protein Kup	*insJ-5*	IS*150* protein	Intergenic G(2) → G(1)
GCR	IS insertion	2,510,962	2,510,973	1,438	-	-	*yfgC*	Putative peptidase	-	-	IS*4* (+) + 12 bp insertion

AE-180-28

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	4,479,286	4,479,286	1	C → T	Arg 170 Arg	*yjhQ*	Putative acetyl-transferase	-	-	-

AE-180-30

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	34	34	1	C → T	-	-	-	*thrL*	Thr operon leader peptide	Intergenic
SNP	Transition	19,326	19,326	1	C → T	Arg 205 Cys	*nhaR*	Transcriptional activator	-	-	-
SNP	Transition	638,827	638,827	1	A → G	Ile 33 Val	*ECB_02003*	Hypothetical protein	-	-	-
SNP	Transition	1,152,029	1,152,029	1	C → T	Gly 167 Gly	*yneI*	Putative succinate semialdehyde dehydrogenase	-	-	-
SNP	Transition	3,684,416	3,684,416	1	G → A	Gly 85 Ser	*yibL*	Hypothetical protein	-	-	-
SNP	Transversion	4,375,263	4,375,263	1	A → T	Asp 135 Glu	*ECB_04073*	Putative acetyl-CoA: acetoacetyl-CoA transferase	-	-	-
Indel	Deletion	1,043,563	1,043,564	1	-G	-	*tppB*	Putative tripeptide transporter permease	*nth*	Endonuclease III	Intergenic G(6) → G(5)
GCR	Deletion	1,117,803	1,124,384	6,581	-	-	*ECB_01527*	Outer membrane lipoprotein Blc	*ECB_01510*	Putative tail component of prophage	10 genes deleted†
GCR	IS insertion	1,272,468	1,272,470	1,146	-	-	*trg*	Chemotaxis protein III, ribose and galactose sensor receptor	*mokB*	Regulatory peptide	Intergenic IS*150* (+) + 3 bp insertion

†Deleted genes encode for hypothetical proteins, a putative pathogenicity island protein, a putative AraC-type regulatory protein, a putative S lysis protein, a putative lysozyme and cold shock proteins.

AE-180-32

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	164,340	164,340	1	C → T	-	*sfsA*	Sugar fermentation stimulation protein A	*ligT*	2'-5' RNA ligase	Intergenic
SNP	Transversion	1,083,277	1,083,277	1	G → T	-	*ydgE*	Multidrug efflux system protein MdtI	*ydgD*	Putative peptidase	Intergenic
SNP	Transition	1,449,253	1,449,253	1	G → A	Ala 60 Ala	*narJ*	Molybdenum co-factor assembly chaperone subunit	-	-	-
SNP	Transition	1,847,760	1,847,760	1	T → C	-	*ECB_ 00826*	Hypothetical protein	*ECB_ 00825*	Putative replication protein for prophage	Intergenic
SNP	Transition	3,228,132	3,228,132	1	G → A	Ala 513 Val	*yhbX*	Putative inner membrane hydrolase	-	-	-
SNP	Transition	4,195,683	4,195,683	1	A → G	-	*malM*	Maltose regulon periplasmic protein	*yjbI*	Hypothetical protein	Intergenic
GCR	IS insertion	1,464,063	1,464,065	1,446	-	-	*chaA*	Calcium/ sodium antiporter	*IdrC*	Small toxic polypeptide	Intergenic IS*150* (+) + 3 bp insertion

AE-180-34

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	418,865	418,865	1	G → A	Arg 167 Cys	*cyoC*	Cytochrome O ubiquinol oxidase	-	-	-
SNP	Transversion	2,305,613	2,305,613	1	C → T	Gln 43*	*ECB_ 02200*	YadA protein	-	-	*Protein truncation
SNP	Transition	2,964,028	2,964,028	1	C → A	-	*mltC*	Murein trans-glycosylase C	*nupG*	Nucleoside transporter	Intergenic
SNP	Transversion	3,930,625	3,930,625	1	T → G	Leu 199 Arg	*corA*	Magnesium/ nickel/ cobalt transporter	-	-	-
GCR	IS mediated deletion	547,700	551,936	4,236	-	-	*ybcQ*	Putative anti-termination protein	*envY*	DNA binding transcriptional activator of porin biosynthesis	IS*1* mediated 5 genes deleted[†]
GCR	IS insertion	551,935	-	777	-	-	*ompT*	Outer membrane protease	-	-	IS*1* insertion required to mediate deletion
GCR	IS insertion	3,835,180	3,835,190	777	-	-	*yieC*	Carbohydrate specific outer membrane porin	-	-	IS*1* (+) + 9 bp insertion

[†] Deleted genes encode for hypothetical proteins and a DNA-binding transcriptional activator.

AE-180-36

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transversion	613,960	613,960	1	T → A	Leu 127 His	*ahpF*	Alkyl hydroperoxide reductase	-	-	-
SNP	Transition	1,558,121	1,558,121	1	C → T	Pro 49 Pro	*ycfF*	Purine nucleoside phospho-ramidase	-	-	-
SNP	Transition	2,354,692	2,354,692	1	G → A	Gly 110 Glu	*flk*	Flagella bio-synthesis regulator	-	-	-
SNP	Transversion	2,394,749	2,394,749	1	T → A	Phe 1109 Leu	*evgS*	Hybrid sensory histidine kinase in two component regulatory system	-	-	-
SNP	Transition	4,212,193	4,212,193	1	A → G	Lys 92 Arg	*tyrB*	Aromatic amino acid amino transferase	-	-	-
GCR	IS insertion	1,625,833	1,625,842	778	-	-	*ycdT*	Putative diguanylate cyclase	*pgaA*	Outer membrane protein PgaA	Intergenic IS*1* (+) + 10 bp insertion
GCR	IS insertion	2,367,200	2,367,202	777	-	-	*yfcQ*	Putative fimbrial like adhesin protein	-	-	IS*1* (+) + 9 bp insertion

AE-180-38

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transversion	1,176,674	1,176,674	1	T → G	-	*ydeO*	Transcriptional regulator	*ydeN*	Hypothetical protein	Intergenic
SNP	Transition	2,055,723	2,055,723	1	G → A	Gly 277 Ser	*asnB*	Asparagine synthetase B	-	-	-
Indel	Slippage	1,504,657	1,504,658	1	+A	-	*umuD*	DNA polymerase V subunit UmuD	*ycgN*	Hypothetical protein	Intergenic A(8) → A(9)

AE-180-40

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	2,194,503	2,194,503	1	C → T	Pro 172 Leu	*spr*	Putative outer membrane lipoprotein	-	-	-
GCR	IS insertion	86,845	86,853	777	-	-	*leuL*	Leu operon leader peptide	*leuO*	Leucine transcriptional activator	Intergenic IS*1* (-) + 9 bp insertion
GCR	IS insertion	1,489,865	1,489,873	777	-	-	*ymgE*	Putative inner membrane protein	*ycgR*	Flagellar motility regulator	Intergenic IS*1* (-) + 9 bp insertion
GCR	IS insertion	2,295,162	2,295,167	1,349	-	-	*menC*	O-succinyl-benzoate synthase	-	-	IS*186* (-) + 6 bp insertion
GCR	IS insertion	4,015,454	4,015,456	1,446	-	-	*yihS*	Putative glucosamine isomerase	-	-	IS*150* (+) + 3 bp insertion

AE-180-42

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transversion	1,030,806	1,030,806	1	G → T	Asp 120 Tyr	*ydhF*	Putative oxidoreductase	-	-	-

AE-180-44

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transversion	366,206	366,206	1	G → T	Arg 230 Leu	*yaiW*	Putative DNA binding transcriptional regulator	-	-	-
SNP	Transversion	1,010,594	1,010,594	1	T → G	Gly 238 Gly	*ydhQ*	Hypothetical protein	-	-	-
SNP	Transition	2,245,987	2,245,987	1	T → C	Ser 119 Ser	*atoB*	Acetyl-CoA acetyl-transferase	-	-	-
SNP	Transition	2,507,214	2,507,214	1	G → A	Trp 556*	*hyfR*	DNA binding transcriptional activator	-	-	*Protein truncation
SNP	Transition	2,789,800	2,789,800	1	T → C	-	*ygdH*	Conserved protein of unknown function	*sdaC*	Serine/proton symporter	Intergenic
SNP	Transition	2,893,210	2,893,210	1	G → A	Ser 180 Ser	*lysS*	Lysyl-tRNA synthetase	-	-	-
SNP	Transition	2,953,571	2,953,571	1	C → T	Trp 28*	*yggR*	Putative transporter	-	-	*Protein truncation
Indel	Slippage	3,866,358	3,866,359	1	-G	-	*trkD*	Potassium transport protein Kup	*insJ-5*	IS*150* protein	Intergenic G(2) → G(1)

AE-180-46

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	539,113	539,113	1	G → A	Ala 46 Val	*ECB_00489*	Hypothetical protein	-	-	-
SNP	Transition	2,023,482	2,023,482	1	G → A	Ala 277 Val	*rhsC*	rhsC element core protein	-	-	-
SNP	Transversion	4,138,247	4,138,247	1	G → T	Leu 25 Leu	*thiS*	Sulfur carrier protein	-	-	-

AE-180-48

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	1,665,297	1,665,297	1	C → T	Pro 77 Ser	*yccM*	Putative 4Fe-4S membrane protein	-	-	-
SNP	Transition	1,878,302	1,878,302	1	G → A	Gly 351 Asp	*moeA*	Molybdopterin biosynthesis protein	-	-	-
Indel	Slippage	2,802,779	2,802,780	1	-A	-	*ECB_ 02652*	Hypothetical protein	*fucI*	L-fucose isomerase	Intergenic A(4) → A(3)

AE-180-50

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	372,876	372,876	1	C → T	Val 139 Ile	*proC*	Pyrroline-5-carboxylate reductase	-	-	-
Indel	Deletion	2,057,486	2,057,603	117	-	-	*glnV*	tRNA-Gln	*glnX*	tRNA-Gln	Intergenic
GCR	IS insertion	1,904,314	1,904,317	1349	-	-	*ybiL*	Catecholate siderophore receptor Fiu	*ybiX*	Putative hydroxylase	Intergenic IS186 (-) + 6 bp insertion

AN-144-02

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
-	-	-	-	-	-	-	-	-	-	-	-

AN-144-04

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transversion	2,355,170	2,355,170	1	T → G	Ser 269 Ser	*flk*	Flagella biosynthesis regulator	-	-	-
SNP	Transition	4,109,399	4,109,399	1	G → A	Thr 9 Thr	*btuB*	Vitamin B12 transporter	-	-	-
SNP	Transition	4,131,602	4,131,602	1	G → A	Gly 257 Ser	*rpoC*	DNA-directed RNA polymerase subunit beta	-	-	-
Indel	Deletion	312,641	312,642	1	-C	-	*yahG*	Hypothetical protein	*yahI*	Putative carbamate kinase	Intergenic
Indel	Slippage	2,555,687	2,555,688	1	+T	-	*iscR*	DNA-binding transcriptional regulator	*yfhQ*	Putative methyl-transferase	Intergenic G(6) → G(7)
GCR	IS insertion	16,972	16,974	1,446	-	-	*mokC*	Regulatory protein for HokC	*nhaA*	pH-dependent sodium/proton antiporter	Intergenic IS*150* (-) + 3 bp insertion
GCR	IS insertion	21,796	-	777	-	-	*ECB_00021*	Putative usher protein	-	-	IS*1* insertion required to mediate deletion
GCR	IS mediated deletion	21,797	23,292	1,495	-	-	*ECB_00021*	Putative usher protein	*ECB_00025*	Hypothetical protein	IS*1* mediated *ECB_00022* encoding for hypothetical protein deleted
GCR	IS insertion	316,559	316,560	777	-	-	*yahK*	Putative oxidoreductase	-	-	IS*1* (+) + 9 bp insertion
GCR	IS insertion	632,696	632,698	1,446	-	-	*ECB_02013*	Hypothetical protein	-	-	IS*150* (-) + 3 bp insertion
GCR	IS insertion	2,334,207	2,334,210	1,446	-	-	*yfcC*	Hypothetical protein	-	-	IS*150* (+) + 3 bp insertion

AN-144-06

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transversion	2,057,271	2,057,271	1	T → A	-	*glnW*	tRNA-Gln	-	-	-
SNP	Transition	2,057,274	2,057,274	1	T → C	-	*glnW*	tRNA-Gln	-	-	-
SNP	Transversion	2,057,279	2,057,279	1	A → T	-	*glnW*	tRNA-Gln	-	-	-
SNP	Transversion	2,057,290	2,057,290	1	C → G	-	*glnW*	tRNA-Gln	-	-	-
SNP	Transversion	2,057,291	2,057,291	1	T → A	-	*glnW*	tRNA-Gln	-	-	-
SNP	Transition	2,285,153	2,285,153	1	G → A	-	*ais*	Hypothetical protein	*yfbE*	UDP-4-amino-4-deoxy-L arabinose-oxoglutarate amino-transferase	Intergenic
SNP	Transversion	2,765,259	2,765,259	1	G → T	-	*ygcE*	Putative kinase	*ECB_ 02621*	Hypothetical protein	Intergenic
SNP	Transition	3,778,383	3,778,383	1	C → T	Pro 561 Leu	*ade*	Cryptic adenine deaminase	-	-	-
SNP	Transition	4,067,306	4,067,306	1	C → T	Val 113 Val	*hslV*	ATP-dependent protease peptidase	-	-	-
SNP	Transition	4,178,271	4,178,271	1	C → T	-	*lysC*	Aspartate kinase III	*pgi*	Glucose-6-phosphate isomerase	Intergenic
SNP	Transition	4,530,135	4,530,135	1	G → A	Arg 159*	*hsdR*	Type I restriction enzyme subunit R	-	-	*Protein truncation

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
Indel	Deletion	2,057,314	2,057,440	126	-	-	*glnW*	tRNA-Gln	*glnU*	tRNA-Gln	*metU* encoding for tRNA-Met deleted
GCR	IS insertion	1,219,034	1,219,036	1,446	-	-	*narZ*	Nitrate reductase 2 (NRZ), subunit alpha	-	-	IS*150* (-) + 3 bp insertion
GCR	IS insertion	4,296,002	4,299,101	1,446	-	-	*dcuR*	DNA binding transcriptional activator	*yjdI*	Hypothetical protein	IS*150* (+) + 3 bp insertion
GCR	IS mediated deletion	4,296,003	4,299,101	3,098	-	-	*dcuR*	DNA binding transcriptional activator	*yidL*	Putative transporter	1S*150* mediated 4 genes deleted[†]
GCR	IS insertion	4,299,101	-	1,446	-	-	*lysU*	Lysyl-tRNA synthetase	*yidL*	Putative transporter	IS*150* insertion required to mediate deletion

[†]Deleted genes encode for hypothetical proteins, a putative acyltransferase and a lysyl-tRNA synthetase.

AN-144-08

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	3,834,647	3,834,647	1	G → A	Thr 250 Ile	*yieC*	Carbohydrate-specific outer membrane porin	-	-	-
Indel	Deletion	545,010	545,012	2	-AT	-	*ybcO*	Hypothetical protein	-	-	-
Indel	Slippage	2,779,247	2,779,248	1	+G	-	*barA*	Hybrid sensory histidine kinase	-	-	G(2) → G(3)
GCR	IS insertion	946,957	946,959	1,446	-	-	*ydiZ*	Hypothetical protein	-	-	IS*150* (+) + 3 bp insertion
GCR	IS insertion	1,440,512	1,440,523	1,438	-	-	*adhE*	Bifunctional acetaldehyde-CoA; alcohol dehydrogenase	*tdk*	Thymidine kinase	Intergenic IS*4* (-) + 12 bp insertion
GCR	IS insertion	2,334,207	2,334,210	1,446	-	-	*yfcC*	Hypothetical protein	-	-	IS*150* (+) + 3 bp insertion
GCR	IS insertion	2,421,315	2,421,320	1,349	-	-	*nupC*	Nucleoside (except guanosine) transport	*yfeA*	Putative diguanylate cyclase	Intergenic IS*186* (+) + 6 bp insertion
GCR	IS insertion	3,658,970	3,658,972	1,446	-	-	*yiaT*	Hypothetical protein	-	-	IS*150* (+) + 3 bp insertion

AN-144-10

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	202,529	202,529	1	C → T	Pro 587 Leu	*yaeT*	Outer membrane protein	-	-	-
SNP	Transition	361,170	361,170	1	C → T	Tyr 671 Tyr	*yaiT*	Hypothetical protein	-	-	-
SNP	Transversion	2,815,104	2,815,104	1	C → A	-	*metV*	tRNA-Met	-	-	-
SNP	Transversion	3,669,885	3,669,885	1	G → C	Gly 68 Ala	*rhsA*	rhsA element core protein	-	-	-
SNP	Transition	3,671,097	3,671,097	1	G → A	Ser 472 Asn	*rhsA*	rhsA element core protein	-	-	-
SNP	Transversion	3,671,135	3,671,135	1	A → T	Ile 484 Leu	*rhsA*	rhsA element core protein	-	-	-
SNP	Transition	3,671,137	3,671,137	1	A → G	Ile 485 Met	*rhsA*	rhsA element core protein	-	-	-
SNP	Transversion	4,222,812	4,222,812	1	T → G	-	*soxR*	DNA-binding transcriptional dual regulator	*yjcD*	Putative permease	Intergenic
GCR	IS insertion	297,069	297,071	777	-	-	*ykgG*	Putative transporter	-	-	IS*1* (-) + 9 bp insertion
GCR	IS insertion	3,240,880	3,240,882	1,446	-	-	*yhbE*	Inner membrane protein	*rpmA*	50S ribosomal protein L27	Intergenic IS*150* (+) + 3 bp insertion
GCR	IS insertion	3,663,953	3,663,955	1,446	-	-	*aldB*	Aldehyde dehydrogenase B	-	-	IS*150* (+) + 3 bp insertion

AN-144-12

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transversion	2,748,853	2,748,853	1	A → C	-	*insJ-3*	IS*150* protein	*cysH*	Phospho-adenosine phosphosulfate reductase	Intergenic
SNP	Transversion	2,070,044	2,070,044	1	A → C	-	*gltK*	Glutamate and aspartate transporter subunit	*insK-2*	IS*150* putative transposase	Intergenic
SNP	Transversion	2,070,045	2,070,045	1	C → A	-	*gltK*	Glutamate and aspartate transporter subunit	*insK-2*	IS*150* putative transposase	Intergenic
GCR	IS insertion	546,043	546,045	777	-	-	*ybcQ*	Putative antitermination protein	-	-	IS*1* (-) + 9 bp insertion
GCR	Inversion[†]	582,320	1,764,886	~1.2 Mb	-	-	*hokE*	Small, toxic polypeptide	*pflB*	Pyruvate formate lyase I	IS*150* mediated inversion around the terminus
GCR	IS insertion	1,272,453	1,272,455	1,446	-	-	*trg*	Chemotaxis protein III, ribose and galactose sensor receptor	*mokB*	Regulatory peptide	Intergenic IS*150* (-) + 3 bp insertion
GCR	IS insertion	1,764,019	1,764,886	777	-	-	*focA*	Formate transporter	*pflB*	Pyruvate formate lyase I	Intergenic IS*1* (-) + 9 bp insertion

[†] Depicted in **Figure 4.8**.

AN-144-14

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	119,955	119,955	1	C → T	Gln 133 Gln	*ppdD*	Putative major pilin subunit	-	-	-
SNP	Transversion	227,057	227,057	1	T → G	-	*rrsH*	16S ribosomal RNA	-	-	-
SNP	Transition	381,185	381,185	1	G → A	Leu 998 Phe	*sbcC*	Exonuclease subunit	-	-	-
SNP	Transversion	1,630,940	1,630,940	1	A → C	Lys 117 Asn	*ycdQ*	N-Glycosyl-transferase	-	-	-
SNP	Transition	4,577,364	4,577,364	1	C → T	Val 281 Met	*lplA*	Lipoate-protein ligase A	-	-	-
Indel	Slippage	668,909	668,910	1	-G	-	*yegM*	Multidrug efflux system subunit MdtA	*yegL*	Hypothetical protein	Intergenic G(4) → G(3)
GCR	IS insertion	1,872,676	1,872,678	777	-	-	*yliB*	Putative peptide transporter subunit	-	-	IS*1* (-) + 9 bp insertion
GCR	IS insertion	4,043,793	4,043,797	1,446	-	-	*rhaS*	Transcriptional activator	-	-	IS*150* (-) + 3 bp insertion

AN-144-16

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	61,021	61,021	1	G → A	Pro 95 Leu	*imp*	Organic solvent tolerance protein	-	-	-
SNP	Transition	622,489	622,489	1	A → G	Leu 304 Pro	*citF*	Citrate lyase, citrate ACP transferase subunit alpha	-	-	-
SNP	Transversion	1,090,701	1,090,701	1	A → C	Phe 22 Cys	*ynfJ*	Putative voltage gated ClC-type chloride channel ClcB	-	-	-
Indel	Deletion	3,611,726	3,611,886	160	-	-	*dppA*	Dipeptide transporter	*proK*	tRNA-Pro	Intergenic
Indel	Slippage	3,866,358	3,866,359	1	-G	-	*trkD*	Potassium transport protein	*insJ-5*	IS*150* protein	Intergenic G(2) → G(1)
GCR	IS insertion	230,817	230,819	1,446	-	-	*rrlH*	23S ribosomal RNA	-	-	IS*150* (-) + 3 bp insertion

AN-144-20

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	622,489	622,489	1	A → G	Leu 304 Pro	*citF*	Citrate lyase, citrate ACP transferase subunit alpha	-	-	-
SNP	Transversion	1,240,503	1,240,503	1	T → G	Lys 268 Gln	*yncE*	Hypothetical protein	-	-	-
GCR	IS insertion	668,193	668,196	1,447	-	-	*yegM*	Multidrug efflux system subunit MdtA	*yegL*	Hypothetical protein	Intergenic IS*150* (+) + 4 bp insertion
GCR	IS insertion	1,272,470	1,272,472	1,446	-	-	*trg*	Chemotaxis protein III, ribose and galactose sensor receptor	*mokB*	Regulatory peptide	Intergenic IS*150* (+) + 3 bp insertion

AN-144-22

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transversion	1,656,135	1,656,135	1	T → A	Pro 166 Pro	*torD*	Chaperone protein	-	-	-
SNP	Transition	3,021,648	3,021,648	1	G → A	His 624 His	*glcB*	Malate synthase G	-	-	-
SNP	Transversion	4,462,188	4,462,188	1	C → A	Val 59 Leu	*fecE*	Iron-dicitrate transporter ATP binding subunit	-	-	-

AN-144-24

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transversion	305,643	305,643	1	A → T	Ala 208 Ala	*yahA*	Putative DNA binding transcriptional regulator	-	-	-
SNP	Transversion	634,226	634,226	1	C → A	Arg 17 Ile	*ECB_ 02011*	Bacteriophage P2 C-like protein	-	-	-
SNP	Transition	684,565	684,565	1	C → T	-	*yegH*	Hypothetical protein	*wza*	Lipoprotein for capsular polysaccharide translocation	Intergenic
SNP	Transversion	1,860,047	1,860,047	1	T → G	-	*cmr*	Multidrug efflux system protein	*ybjG*	Undecaprenyl pyrophosphate phosphatase	Intergenic
SNP	Transition	3,869,337	3,869,337	1	A → G	-	*yieP*	Putative transcriptional regulator	*rrsC*	16S ribosomal RNA	Intergenic
Indel	Deletion	2,423,812	2,423,843	32	-	-	*alaX*	tRNA-Ala	*alaW*	tRNA-Ala	Intergenic
GCR	IS insertion	61,532	61,534	1,446	-	-	*imp*	Organic solvent tolerance protein	*djlA*	DnaJ like membrane chaperone protein	Intergenic IS*150* (-) + 3 bp insertion
GCR	IS insertion	910,345	910,346	1,446	-	-	*ynjI*	Hypothetical protein	-	-	IS*150* (-) + 3 bp insertion
GCR	Partial IS insertion	2,070,036	2,070,038	935	-	-	*gltK*	Glutamate and aspartate transporter subunit	*insK-2*	IS*150* putative transposase	Intergenic IS*150* (+) insertion
GCR	IS insertion	3,243,267	3,243,269	1,446	-	-	*nlp*	DNA-binding transcriptional regulator	*murA*	UDP-N-acetyl-glucosamine L-carboxyvinyl transferase	Intergenic IS*150* (+) + 3 bp insertion

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
GCR	IS insertion	3,808,732	3,808,734	1,446	-	-	*yidX*	Hypothetical protein	-	-	IS*150* (-) + 3 bp insertion
GCR	IS insertion	4,043,794	4,043,795	1,472	-	-	*rhaS*	Transcriptional activator	-	-	IS*150* (+) + 29 bp insertion

AN-144-26

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	3,440,980	3,440,980	1	G → A	Lys 701 Lys	*yhgF*	Putative transcriptional accessory protein	-	-	-
GCR	IS insertion	1,123,004	1,123,008	1,261	-	-	*ynfN*	Hypothetical protein	-	-	IS*3* (-) + 6 bp insertion
GCR	IS deletion	1,764,874	1,766,323	1,449	-	-	*pflB*	Pyruvate formate lyase I	-	-	IS*150* deletion

AN-144-28

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	1,766,332	1,766,332	1	C → T	Asn 753 Asn	*pflB*	Pyruvate formate lyase I	-	-	-
SNP	Transversion	3,172,999	3,172,999	1	C → A	-	*tdcA*	DNA-binding transcriptional activator	*tdcR*	DNA-binding transcriptional activator TdcR	Intergenic

AN-144-30

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	1,544,112	1,544,112	1	G → A	Val 196 Val	*ycfU*	Outer membrane lipoprotein transporter subunit	-	-	-
SNP	Transition	4,262,523	4,262,523	1	A → G	Val 23 Ala	*phnL*	Carbon-phosphorus lyase complex subunit	-	-	-
Indel	Slippage	1,438,008	1,438,009	1	-A	-	*adhE*	Bifunctional acetaldehyde-CoA/alcohol dehydrogenase	-	-	A(5) → A(4)
Indel	Deletion	3,702,798	3,702,937	139	-	-	*yibD*	Putative glycosyl transferase	-	-	-
GCR	IS insertion	1,272,470	1,272,472	1,446	-	-	*trg*	Chemotaxis protein III, ribose and galactose sensor receptor	*mokB*	Regulatory peptide	Intergenic IS*150* (+) + 3 bp insertion

AN-144-32

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	905,014	905,014	1	C → T	Ala 251 Thr	*sppA*	Protease 4	-	-	-
SNP	Transition	2,675,708	2,675,708	1	G → A	-	*proX*	Glycine betaine transporter periplasmic subunit	*ygaXY*	Putative transporter	Intergenic
GCR	IS insertion	910,345	910,346	1,446	-	-	*ynjI*	Hypothetical protein	-	-	IS*150* (-) + 3 bp insertion

AN-144-34

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	3,898,622	3,898,622	1	G → A	Glu 272 Glu	*wecB*	UDP-N-acetyl glucosamine-2-epimerase	-	-	-
Indel	Insertion	3,625,341	3,625,342	1	+C	-	*insK-4*	IS*150* putative transposase	*glyS*	Glycyl-tRNA synthetase subunit beta	Intergenic C(1) → C(2)
GCR	IS insertion	1,272,453	1,272,454	1,446	-	-	*trg*	Chemotaxis protein III, ribose and galactose sensor receptor	*mokB*	Regulatory peptide	Intergenic IS*150* (-) + 3 bp insertion
GCR	IS insertion	2,521,102	2,521,103	777	-	-	*yfgF*	Putative inner membrane protein	-	-	IS*1* (+) + 9 bp insertion

AN-144-36

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transversion	2,761,200	2,761,200	1	T → G	-	*ygcW*	Putative deoxy-gluconate dehydrogenase	-	-	-
SNP	Transversion	2,761,201	2,761,201	1	A → C	-	*ygcW*	Putative deoxy-gluconate dehydrogenase	-	-	-
Indel	Slippage	2,439,980	2,439,981	1	-A	-	*yfeS*	Hypothetical protein	-	-	A(2) → A(1)
GCR	Deletion	2,973,150	2,990,149	~17 kb	-	-	*ECB_ 02798*	Hypothetical protein	*ECB_ 02819*	KpsS protein	19 genes deleted†

†Deleted genes encode for hypothetical proteins, an antigen 43 phase-variable biofilm formation autotransporter, a putative DNA repair protein, the YeeV-YeeU toxin-antitoxin system, capsule polysaccharide export proteins and 3-deoxy-manno-octulosonate cytidylyltransferase.

AN-144-38

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transversion	3,172,999	3,172,999	1	C → A	-	*tdcA*	DNA-binding transcriptional activator	*tdcR*	DNA-binding transcriptional activator TdcR	Intergenic
GCR	IS deletion	1,764,888	1,766,334	1,446	-	-	*pflB*	Pyruvate formate lyase I	-	-	IS*150* deletion

AN-144-40

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transversion	672,941	672,941	1	A → T	Leu 441 Gln	*yegD*	Putative chaperone	-	-	-
SNP	Transition	1,078,845	1,078,845	1	C → T	Gly 223 Gly	*pntA*	NAD (P) trans-hydrogenase subunit alpha	-	-	-
SNP	Transition	2,118,729	2,118,729	1	G → A	Leu 137 Leu	*metG*	Methionyl-tRNA synthetase	-	-	-
SNP	Transversion	2,539,230	2,539,230	1	T → G	Ser 579 Ser	*pbpC*	Penicillin-binding protein	-	-	-
SNP	Transition	2,544,706	2,544,706	1	A → G	Leu 408 Ser	*yfhM*	Hypothetical protein	-	-	-
SNP	Transition	3,416,952	3,416,952	1	G → A	His 150 His	*dam*	DNA adenine methylase	-	-	-
SNP	Transversion	4,290,429	4,290,429	1	T → A	Ser 496 Cys	*fumB*	Anaerobic class I fumarate hydratase	-	-	-
GCR	IS insertion	16,972	16,974	1,446	-	-	*hokC*	Small toxic membrane polypeptide	*nhaA*	pH-dependent proton antiporter	Intergenic IS*150* (-) + 3 bp insertion
GCR	IS insertion	549,471	549,477	1,260	-	-	*ECB_ 00515*	Hypothetical protein	-	-	IS*3* (-) + 5 bp insertion
GCR	IS insertion	1,113,441	1,113,442	1,473	-	-	*ECB_ 01533*	Hypothetical protein	*hokD*	Small toxic polypeptide	Intergenic IS*150* (+) + 3 bp insertion
GCR	IS insertion	1,181,523	1,181,525	1,446	-	-	*yddA*	Multidrug ABC transporter membrane	-	-	IS*150* (+) + 3 bp insertion

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
GCR	IS insertion	1,272,470	1,272,472	1,446	-	-	*trg*	Chemotaxis protein III, ribose and galactose sensor receptor	*mokB*	Regulatory peptide	Intergenic IS*150* (+) + 3 bp insertion
GCR	IS insertion	3,217,019	3,217,021	1,446	-	-	*nlpI*	Lipoprotein NlpI	*pnp*	Polynucleotide phosphorylase/ polyadenylase	Intergenic IS*150* (-) + 3 bp insertion
GCR	IS insertion	4,099,134	4,099,136	1,446	-	-	*ppc*	Phosphoenol-pyruvate carboxylase	*argE*	Acetylornithine deacetylase	Intergenic IS*150* (-) + 3 bp insertion
GCR	IS insertion	4,381,587	4,381,589	1,446	-	-	*cycA*	D-alanine/D-serine/glycine permease	-	-	IS*150* (-) + 3 bp insertion

AN-144-42

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
GCR	IS insertion	3,000,325	3,000,330	1,260	-	-	*ECB_ 02827*	Polysialic acid transport ATP binding KpsT	-	-	IS3 (+) + 5 bp insertion

AN-144-44

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transversion	1,290,872	1,290,872	1	T → G	Lys 285 Gln	*insF-2*	IS*3* protein	-	-	-
SNP	Transversion	1,317,845	1,317,845	1	A → C	Ile 12 Ser	*ECB_ 01335*	Hypothetical protein	-	-	-
Indel	Deletion	951,549	951,560	11	-	-	*ECB_ 01690*	Hypothetical protein	*thrS*	Threonyl-tRN A synthetase	Intergenic
Indel	Slippage	4,279,523	4,279,524	1	-T	-	*yjdB*	Putative cell division protein	-	-	T(2) → T(1)
GCR	IS insertion	1,268,600	-	1,446	-	-	*ydcJ*	Hypothetical protein	-	-	IS*150* insertion required to mediate deletion
GCR	IS mediated deletion	1,268,600	1,272,468	3,868	-	-	*mdoD*	Glucan biosynthesis protein D	*mokB*	Regulatory peptide	IS*150* mediated 3 genes deleted[†]
GCR	IS insertion	1,272,467	1,272,468	1,446	-	-	*trg*	Chemotaxis protein III, ribose and galactose sensor receptor	*mokB*	Regulatory peptide	Intergenic IS*150* (+) + 3 bp insertion
GCR	IS insertion	1,325,141	1,325,143	1,260	-	-	*intR*	Integrase	-	-	IS*3* (-) + 5 bp insertion

[†] Deleted genes encode for a hypothetical protein, a putative DNA-binding transcriptional regulator and a methyl-accepting chemotaxis protein III.

AN-144-46

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	252,026	252,026	1	T → C	Ile 248 Ile	*yafL*	Putative lipoprotein and C40 family peptidase	-	-	-
SNP	Transition	319,176	319,176	1	G → A	Ala 15 Thr	*yahO*	Hypothetical protein	-	-	-
SNP	Transition	804,570	804,570	1	C → T	Thr 48 Thr	*ntpA*	dATP pyro-phospho-hydrolase	-	-	-
SNP	Transversion	1,290,872	1,290,872	1	T → G	Lys 285 Gln	*insF-2*	IS3 protein	-	-	-
SNP	Transversion	3,051,763	3,051,763	1	A → C	Val 96 Gly	*hybA*	Hydrogenase-2 protein	-	-	-
SNP	Transversion	3,603,182	3,603,182	1	A → C	-	*ldrE*	Small, toxic polypeptide	*ldrF*	Small, toxic polypeptide	Intergenic
Indel	Slippage	1,554,506	1,554,506	1	+G	-	*ycfP*	Hypothetical protein	-	-	G(2) → G(3)
Indel	Slippage	3,866,357	3,866,358	1	+G	-	*trkD*	Potassium transport protein	*insJ-5*	IS*150* protein	Intergenic G(2) → G(3)
GCR	IS insertion	603,036	603,038	1,443	-	-	*entA*	2,3-di-hydroxy-benzoate-2,3-de-hydrogenase	*ybdB*	Hypothetical protein	Intergenic IS*150* (-) + 3 bp insertion
GCR	IS insertion	605,950	605,952	1,446	-	-	*ybdB*	Hypothetical protein	*ybdD*	Hypothetical protein	Intergenic IS*150* (+) + 3 bp insertion
GCR	Inversion[‡]	606,656	1,764,886	~1.1 Mb	-	-	*ybdD*	Hypothetical protein	*pflB*	Pyruvate formate lyase I	IS*150* mediated inversion around the terminus

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
GCR	Deletion	632,692	654,838	~22 kb	-	-	*yegR*	Hypothetical protein	*yegQ*	Putative peptidase	27 genes deleted[†]
GCR	IS insertion	1,085,303	1,085,305	1,446	-	-	*ynfM*	Putative transporter	-	-	IS*150* (-) + 3 bp insertion
GCR	IS insertion	1,272,455	1,272,457	1,446	-	-	*trg*	Chemotaxis protein III, ribose and galactose sensor receptor	*mokB*	Regulatory peptide	Intergenic IS*150* (-) + 3 bp insertion

[†] Deleted genes encode for hypothetical proteins, an integrase, a bacteriophage P2 Cox-like protein, a bacteriophage P2 C-like protein, a replication gene A protein, a bacteriophage capsid portal protein, baseplate assembly proteins J and W, a putative phage protein and a DNA-binding transcriptional regulator.

[‡] Depicted in **Figure 4.8**.

AN-144-48

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transversion	2,063,688	2,063,688	1	A → C	Ile 166 Ile	*lnt*	Apolipoprotein N-acyltransferase	-	-	-
GCR	IS deletion	1,764,888	1,766,334	1,446	-	-	*pflB*	Pyruvate formate lyase I	-	-	IS*150* deletion

AN-144-50

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	1,667,983	1,667,983	1	C → T	Thr 33 Met	*ymcD*	Hypothetical protein	-	-	-
SNP	Transition	3,808,711	3,808,711	1	G → A	Gly 36 Asp	*yidX*	Hypothetical protein	-	-	-
Indel	Slippage	1,504,657	1,504,658	1	+A	-	*umuD*	DNA polymerase V subunit	*ycgN*	Hypothetical protein	Intergenic A(8) → A(9)
Indel	Slippage	3,866,357	3,866,358	1	+G	-	*trkD*	Potassium transport protein	*insJ-5*	IS*150* protein	Intergenic G(2) → G(3)
Indel	Deletion	2,745,500	2,745,671	171	-	-	*iap*	Alkaline phosphatase isozyme conversion amino-peptidase	*insL-5*	Putative transposase for IS*186*	Intergenic
GCR	Deletion	632,699	654,814	~22 kb	-	-	*yegR*	Hypothetical protein	*yegQ*	Putative peptidase	27 genes deleted[†]
GCR	IS mediated deletion	1,117,802	1,125,343	7,541	-	-	*ECB_01527*	Outer membrane lipoprotein Blc	*ECB_01510*	Putative tail component of prophage	IS*3* mediated 13 genes deleted[‡]
GCR	IS insertion	1,179,527	1,179,528	777	-	-	*ydeM*	Hypothetical protein	-	-	IS*1* (-) + 9 bp insertion
GCR	IS insertion	1,272,399	1,272,401	1,446	-	-	*trg*	Chemotaxis protein III, ribose and galactose sensor receptor	*mokB*	Regulatory peptide	Intergenic IS*150* (+) + 3 bp insertion

[†] Deleted genes encode for hypothetical proteins, an integrase, a bacteriophage P2 Cox-like protein, a bacteriophage P2 C-like protein, a replication gene A protein, a bacteriophage portal protein, baseplate assembly proteins J and W, a putative phage protein and a DNA-binding transcriptional regulator.

[‡] Deleted genes encode for hypothetical proteins, a putative pathogenicity island protein, a putative AraC-type regulatory protein, a putative S lysis protein, a putative lysozyme, cold shock proteins and an IS3 element.

Table A.4. Mutation rates calculated from individual lineages.

Lineage	No. of BPSs	No. of Indels	No. of GCRs	Mutation rate per genome per generation	Mutation rate per nucleotide per generation	Lineage	No. of BPSs	No. of Indels	No. of GCRs	Mutation rate per genome per generation	Mutation rate per nucleotide per generation
AE-180-02	0	1	0	2.22×10^{-4}	4.84×10^{-11}	AN-144-02	0	0	0	0.00	0.00
AE-180-04	5	2	7	3.11×10^{-3}	6.77×10^{-10}	AN-144-04	3	2	6	3.18×10^{-3}	6.93×10^{-10}
AE-180-06	4	0	3	1.56×10^{-3}	3.38×10^{-10}	AN-144-06	11	1	4	4.63×10^{-3}	1.01×10^{-9}
AE-180-08	7	0	2	2.00×10^{-3}	4.35×10^{-10}	AN-144-08	1	2	5	2.31×10^{-3}	5.04×10^{-10}
AE-180-10	4	1	0	1.11×10^{-3}	2.42×10^{-10}	AN-144-10	8	0	3	3.18×10^{-3}	6.93×10^{-10}
AE-180-12	1	2	0	6.67×10^{-4}	1.45×10^{-10}	AN-144-12	3	0	4	2.03×10^{-3}	4.41×10^{-10}
AE-180-14	3	0	1	8.89×10^{-4}	1.93×10^{-10}	AN-144-14	5	1	2	2.31×10^{-3}	5.04×10^{-10}
AE-180-16	0	1	0	2.22×10^{-4}	4.84×10^{-11}	AN-144-16	3	2	1	1.74×10^{-3}	3.78×10^{-10}
AE-180-20	6	1	0	1.56×10^{-3}	3.38×10^{-10}	AN-144-20	2	0	2	1.16×10^{-3}	2.52×10^{-10}
AE-180-22	2	2	4	1.78×10^{-3}	3.87×10^{-10}	AN-144-22	3	0	0	8.68×10^{-4}	1.89×10^{-10}
AE-180-24	2	1	2	1.11×10^{-3}	2.42×10^{-10}	AN-144-24	5	1	6	3.47×10^{-3}	7.56×10^{-10}
AE-180-26	1	1	1	6.67×10^{-4}	1.45×10^{-10}	AN-144-26	1	0	2	8.68×10^{-4}	1.89×10^{-10}
AE-180-28	1	0	0	2.22×10^{-4}	4.84×10^{-11}	AN-144-28	2	0	0	5.79×10^{-4}	1.26×10^{-10}
AE-180-30	6	1	2	2.00×10^{-3}	4.35×10^{-10}	AN-144-30	2	2	1	1.45×10^{-3}	3.15×10^{-10}
AE-180-32	6	0	1	1.56×10^{-3}	3.38×10^{-10}	AN-144-32	2	0	1	8.68×10^{-4}	1.89×10^{-10}
AE-180-34	4	0	3	1.56×10^{-3}	3.38×10^{-10}	AN-144-34	1	1	2	1.16×10^{-3}	2.52×10^{-10}
AE-180-36	5	0	2	1.56×10^{-3}	3.38×10^{-10}	AN-144-36	2	1	1	1.16×10^{-3}	2.52×10^{-10}
AE-180-38	2	1	0	6.67×10^{-4}	1.45×10^{-10}	AN-144-38	1	0	1	5.79×10^{-4}	1.26×10^{-10}
AE-180-40	1	0	4	1.11×10^{-3}	2.42×10^{-10}	AN-144-40	7	0	8	4.34×10^{-3}	9.44×10^{-10}
AE-180-42	1	0	0	2.22×10^{-4}	4.84×10^{-11}	AN-144-42	0	0	1	2.89×10^{-4}	6.30×10^{-11}
AE-180-44	7	1	0	1.78×10^{-3}	3.87×10^{-10}	AN-144-44	2	2	4	2.31×10^{-3}	5.04×10^{-10}
AE-180-44	3	0	0	6.67×10^{-4}	1.45×10^{-10}	AN-144-46	6	2	6	4.05×10^{-3}	8.81×10^{-10}
AE-180-48	2	1	0	6.67×10^{-4}	1.45×10^{-10}	AN-144-48	1	0	1	5.79×10^{-4}	1.26×10^{-10}
AE-180-50	1	1	1	6.67×10^{-4}	1.45×10^{-10}	AN-144-50	2	3	4	2.60×10^{-3}	5.67×10^{-10}

Table A.5. Codon usage of *E. coli* REL4536 genome.

Amino Acid	Codon	Number	Average frequency per 1000 codons	Fraction
Glycine	GGG	15,682	10.24	0.18
Glycine	GGA	18,541	12.10	0.21
Glycine	GGT	24,335	15.89	0.27
Glycine	GGC	29,995	19.58	0.34
Glutamic acid	GAG	13,767	8.99	0.33
Glutamic acid	GAA	27,654	18.05	0.67
Aspartic acid	GAT	28,345	18.50	0.61
Aspartic acid	GAC	17,873	11.67	0.39
Valine	GTG	21,647	14.13	0.26
Valine	GTA	17,217	11.24	0.21
Valine	GTT	26,854	17.53	0.32
Valine	GTC	17,660	11.53	0.21
Alanine	GCG	37,872	24.72	0.30
Alanine	GCA	31,691	20.69	0.25
Alanine	GCT	26,585	17.35	0.21
Alanine	GCC	30,418	19.86	0.24
Arginine	AGG	16,818	10.98	0.11
Arginine	AGA	18,542	12.10	0.12
Serine	AGT	16,355	10.68	0.12
Serine	AGC	26,620	17.38	0.20
Lysine	AAG	21,034	13.73	0.37
Lysine	AAA	35,903	23.44	0.63
Asparagine	AAT	27,131	17.71	0.50
Asparagine	AAC	27,220	17.77	0.50
Methionine	ATG	25,269	16.50	1.00
Isoleucine	ATA	20,888	13.64	0.27
Isoleucine	ATT	27,476	17.94	0.36
Isoleucine	ATC	28,381	18.53	0.37
Threonine	ACG	24,632	16.08	0.29
Threonine	ACA	19,304	12.60	0.23
Threonine	ACT	16,504	10.77	0.19
Threonine	ACC	24,658	16.10	0.29
Tryptophan	TGG	27,568	18.00	1.00
Stop	TGA	27,166	17.73	0.46
Cysteine	TGT	19,349	12.63	0.38
Cysteine	TGC	31,238	20.39	0.62
Stop	TAG	8,920	5.82	0.15
Stop	TAA	22,844	14.91	0.39
Tyrosine	TAT	20,900	13.64	0.55
Tyrosine	TAC	17,292	11.29	0.45

Amino Acid	Codon	Number	Average frequency per 1000 codons	Fraction
Leucine	TTG	25,204	16.45	0.20
Leucine	TTA	22,902	14.95	0.18
Phenylalanine	TTT	36,083	23.55	0.57
Phenylalanine	TTC	27,622	18.03	0.43
Serine	TCG	23,579	15.39	0.18
Serine	TCA	28,378	18.52	0.21
Serine	TCT	18,637	12.17	0.14
Serine	TCC	18,661	12.18	0.14
Arginine	CGG	28,589	18.66	0.19
Arginine	CGA	23,545	15.37	0.16
Arginine	CGT	24,399	15.93	0.16
Arginine	CGC	38,047	24.84	0.25
Glutamine	CAG	34,462	22.50	0.57
Glutamine	CAA	25,837	16.87	0.43
Histidine	CAT	25,332	16.54	0.53
Histidine	CAC	22,275	14.54	0.47
Leucine	CTG	34,439	22.48	0.27
Leucine	CTA	8,987	5.87	0.07
Leucine	CTT	21,163	13.81	0.17
Leucine	CTC	13,963	9.11	0.11
Proline	CCG	29,387	19.18	0.32
Proline	CCA	29,297	19.12	0.32
Proline	CCT	17,050	11.13	0.19
Proline	CCC	15,909	10.39	0.17

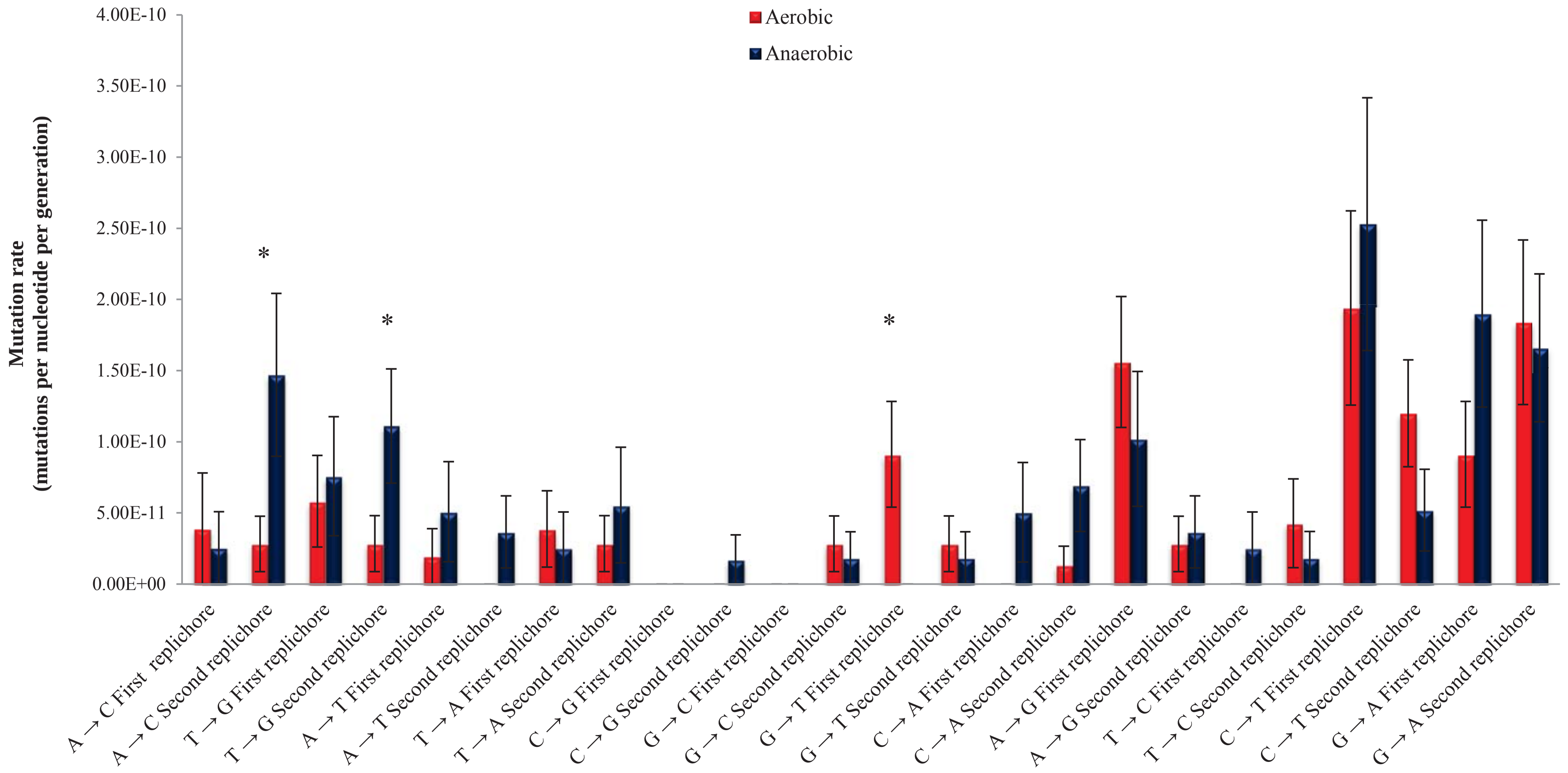

Figure A.1. Mutation rates of different BPSs in the two replichores in aerobically and anaerobically grown *E. coli*. Shown are mean mutation rates per nucleotide per generation. Error bars represent standard error of the mean. Asterisk denotes a significant difference between the aerobic and anaerobic mutation rates ($p < 0.05$).

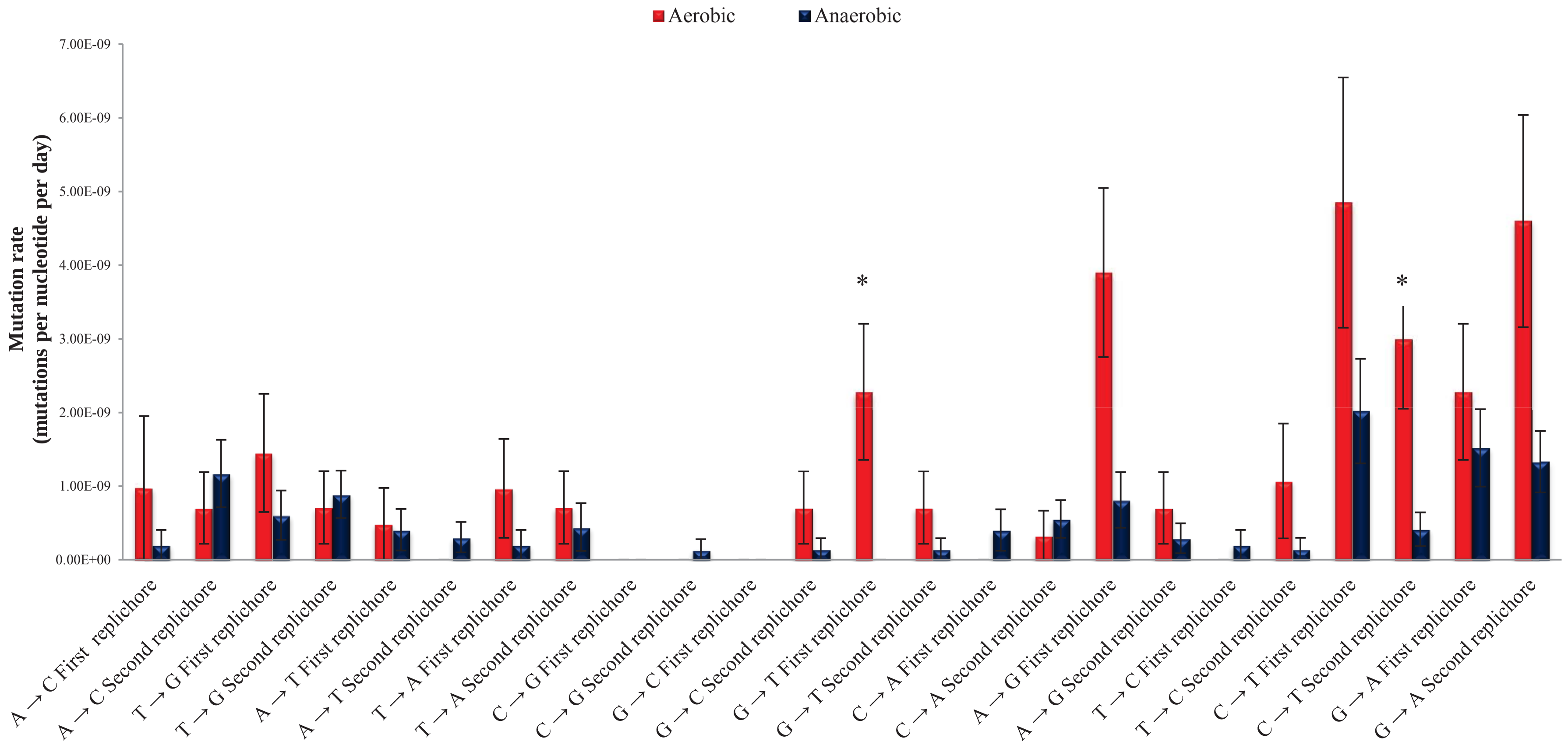

Figure A.2. Mutation rates of different BPSs in the two replichores in aerobically and anaerobically grown *E. coli* per day of growth. Error bars represent standard error of the mean. Asterisk denotes a significant difference between the aerobic and anaerobic mutation rates ($p < 0.05$).

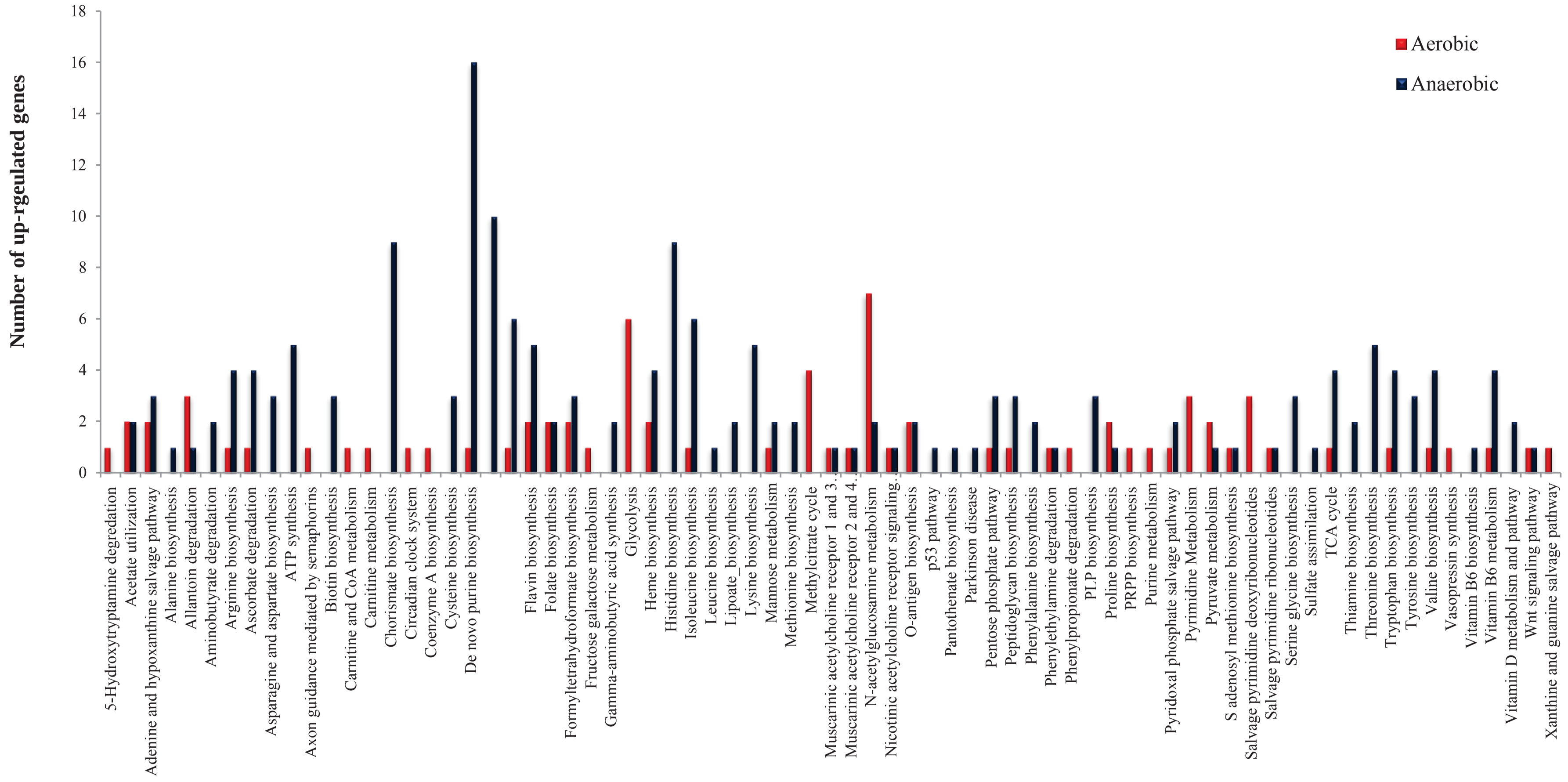

Figure A.3. Functional category enrichment analysis of differentially expressed genes. The x-axis shows GO pathways terms while the y-axis shows the number of significantly enriched genes (*p-adj* <0.05) for each environment.

Table A.6. Genes induced in response to acidic stress.

Gene	Description	Fold change[$]	*p*-adj value[*]
aceA	Isocitrate lyase	7.95	1.83×10^{-28}
aceB	Malate synthase A	8.48	3.08×10^{-31}
adiA	Arginine decarboxylase	1.07	0.62
adiY	DNA-binding transcriptional activator	1.56	0.007
atpA	ATP synthase subunit	2.05	2.39×10^{-14}
atpB	ATP synthase subunit	1.89	1.87×10^{-9}
atpC	ATP synthase subunit	-1.71	1.27×10^{-5}
atpD	ATP synthase subunit	1.02	0.91
atpE	ATP synthase subunit	1.62	0.08
atpF	ATP synthase subunit	3.14	2.32×10^{-19}
atpG	ATP synthase subunit	1.88	1.05×10^{-6}
atpH	ATP synthase subunit	2.76	2.38×10^{-16}
cadA	Lysine decarboxylase 1	-1.03	0.90
cadB	Lysine:cadaverine antiporter	-1.04	0.93
cadC	DNA-binding transcriptional activator	-2.27	4.80×10^{-9}
dppA	Dipeptide ABC transporter subunit	1.84	2.21×10^{-13}
dppB	Dipeptide ABC transporter subunit	2.92	5.98×10^{-13}
dppC	Dipeptide ABC transporter subunit	2.51	3.71×10^{-10}
dppD	Dipeptide ABC transporter subunit	5.82	5.36×10^{-62}
dppF	Dipeptide ABC transporter subunit	6.62	3.28×10^{-61}
gadA	Glutamate decarboxylase A	-3.40	2.28×10^{-18}
gadB	Glutamate decarboxylase B	-8.92	2.37×10^{-39}
nrdH	Glutaredoxin-like protein	11.46	3.67×10^{-30}
nrdI	Flavodoxin	5.45	1.03×10^{-12}
oppA	Peptide ABC transporter	1.61	7.69×10^{-7}
sdhA	Succinate dehydrogenase subunit	4.22	5.02×10^{-36}
sdhB	Succinate dehydrogenase subunit	5.78	6.98×10^{-76}
sdhC	Succinate dehydrogenase subunit	2.81	1.19×10^{-19}
sdhD	Succinate dehydrogenase subunit	3.69	5.27×10^{-25}
speA	Arginine decarboxylase	-5.71	1.36×10^{-46}
speB	Agmatinase	-2.24	2.12×10^{-8}
speC	Ornithine decarboxylase	1.46	0.003
speD	Adenosylmethionine decarboxylase	2.11	4.00×10^{-12}
speE	Spermidine synthase	2.26	1.95×10^{-7}
yagU	Inner-membrane protein	1.91	2.04×10^{-6}
yfiD	Stress-induced pyruvate formate-lyase	1.60	0.002
ygaE	DNA-binding transcriptional repressor	-1.13	0.30
yjdE	Arginine:agmatine antiporter	-1.20	0.22

Table A.7. List of all mutations found in REL4536 Δ*sbcC::cat*-1K and REL4536 AN-1K clone genomes.

REL4536 *ΔsbcC::cat*-1K-1

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
Indel	Deletion	2,997,535	2,997,895	360	-	-	*ECB_02825*	Pyro-phosphorylase	-	-	-
Indel	Deletion	4,295,377	4,295,382	5	-	-	*dcuR*	DNA binding transcriptional activator DcuR	*yjdI*	Hypothetical protein	-
GCR	IS mediated deletion	546,987	582,320	~33 kb	30 genes[†]	-	*ybcQ*	Putative anti-termination protein	*hokE*	Small, toxic polypeptide	IS*1* mediated
GCR	IS insertion	963,716	963,718	1,473	-	-	*ydiU*	Hypothetical protein	-	-	IS*150* (+) + 30 bp
GCR	IS deletion	1,764,886	7,764,889	1,446	-	-	*pflB*	Pyruvate formate lyase 1	-	-	1S*150* deletion

[†] Deleted genes listed in **Table A.8**.

REL4536 *ΔsbcC::cat*-1K-6

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
GCR	IS insertion	624,989	624,992	1,260	-	-	*citC*	Citrate lyase synthetase	-	-	1S*3* + 5 bp insertion
GCR	IS insertion	963,716	963,718	1,446	-	-	*ydiU*	Hypothetical protein	-	-	IS*150* + 3 bp insertion
GCR	IS insertion	974,185	974,188	1,446	-	-	*ydiQ*	Putative electron transfer flavoprotein	-	-	IS*150* (+) + 3 bp insertion

REL4536 *ΔsbcC::cat*-1K-7

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
GCR	IS insertion	963,716	963,718	1,446	-	-	*ydiU*	Hypothetical protein	-	-	IS*150* + 3 bp insertion
GCR	IS insertion	974,185	974,188	1,446	-	-	*ydiQ*	Putative electron transfer flavoprotein	-	-	IS*150* (+) + 3 bp insertion
GCR	IS insertion	1,089,037	1,089,040	1,438	-	-	*ynfK*	Putative dithiobiotin synthetase	-	-	IS*4* + 12 bp insertion
GCR	IS deletion	1,764,886	1,766,323	1,446	-	-	*pflB*	Pyruvate formate lyase I	-	-	1S*150* deletion
GCR	IS insertion	4,303,100	4,303,102	1,260	-	-	*cadA*	Lysine de-carboxylase I	-	-	IS*3* + 5 bp insertion

REL4536 *ΔsbcC::cat*-1K-8

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
Indel	Deletion	4,295,377	4,295,382	5	-	-	*dcuR*	DNA binding transcriptional activator DcuR	*yjdI*	Hypothetical protein	Intergenic
GCR	IS mediated deletion	546,987	582,320	~33 kb	30 genes[†]	-	*ybcQ*	Putative anti-termination protein	*hokE*	Small, toxic polypeptide	IS*1* mediated
GCR	IS insertion	4,381,583	4,381,585	1,446	-	-	*cycA*	D-alanine/ D-serine/ glycine permease	-	-	IS*150* insertion

[†] Deleted genes listed in **Table A.8**.

REL4536 *ΔsbcC::cat*-1K-12

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transversion	500,208	500,208	1	A → C	His 31 Pro	*ylbH*	Hypothetical protein	-	-	-
SNP	Transition	1,439,509	1,439,509	1	A → G	Glu 568 Gly	*adhE*	Bifunctional acetaldehyde-CoA; alcohol dehydrogenase	-	-	-
SNP	Transition	1,448,670	1,448,670	1	G → A	Ala 18 Val	*narI*	Nitrate reductase 1	-	-	-
Indel	Deletion	642,689	642,717	29	-	-	*ECB_ 02001*	Hypothetical protein	*ECB_ 02000*	Bacteriophage capsid portal protein	
Indel	Deletion	933,388	933,401	14	-	-	*celB*	PTS system N,N'-diacetylc-hitobiose-specific transporter subunit IIC	-	-	-
Indel	Deletion	1,518,864	1,518,891	28	-	-	*ycgZ*	Hypothetical protein	*ycgF*	Putative FAD-binding phosphor-diesterase	
Indel	Deletion	1,578,069	1,578,091	23	-	-	*rne*	Ribonuclease E	-	-	-
Indel	Deletion	1,780,947	1,780,979	33	-	-	*ycaJ*	Recombination factor protein RarA	-	-	-
Indel	Deletion	1,804,234	1,804,246	13	-	-	*aqpZ*	Aquaporin Z	*ybjE*	Putative transporter	-
Indel	Deletion	2,666,207	2,666,272	66	-	-	*stpA*	DNA binding protein	*ygaW*	Putative inner membrane protein	

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
Indel	Deletion	2,768,001	2,768,026	26	-	-	*ygcG*	Hypothetical protein	-	-	-
GCR	IS insertion	1,272,468	1,272,470	1,446	-	-	*trg*	Chemotaxis protein III, ribose and galactose sensor receptor	*mokB*	Regulatory peptide	IS*150* insertion
GCR	IS insertion	1,578,092	1,578,094	1,446	-	-	*rne*	Ribonuclease E	-	-	IS*150* + 3 bp insertion
GCR	IS insertion	1,882,390	1,882,392	1,446	-	-	*ybiW*	Putative pyruvate formate lyase	-	-	IS*150* + 3 bp insertion
GCR	IS insertion	2,169,478	2,169,479	1,446	-	-	*cirA*	Colicin I receptor	-	-	IS*150* + 3 bp insertion
GCR	IS insertion	4,095,663	4,095,666	1,446	-	-	*yijP*	Hypothetical protein	-	-	IS*150* (+) + 3 bp insertion
GCR	IS insertion	4,207,648	4,207,650	1,446	-	-	*pspG*	Phage shock protein G	-	-	IS*150* + 3 bp insertion

REL4536 *ΔsbcC::cat*-1K-14

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transversion	1,439,821	1,439,821	1	A → C	Asp 672 Ala	*adhE*	Bifunctional acetaldehyde-CoA/alcohol dehydrogenase	-	-	-
SNP	Transition	1,766,332	1,766,332	1	C → T	Asn 75 Asn	*pflB*	Pyruvate formate lyase I	-	-	-
Indel	Deletion	2,322,072	2,322,092	21	-	-	*nuoA*	NADH dehydrogenase subunit A	*lrhA*	DNA-binding transcriptional repressor protein	-
Indel	Deletion	4,595,674	4,595,685	12	-	-	*lasT*	Putative RNA methyl-transferase	-	-	-
GCR	IS insertion	963,716	963,718	1,446	-	-	*ydiU*	Hypothetical protein	-	-	IS*150* + 3 bp insertion
GCR	IS insertion	974,185	974,188	1,446	-	-	*ydiQ*	Putative electron transfer flavoprotein YdiQ	-	-	IS*150* (+) + 3 bp insertion
GCR	IS insertion	3,244,960	3,244,962	1,446	-	-	*yrbA*	Putative DNA-binding transcriptional regulator	*yrbB*	Hypothetical protein	IS*150* + 3 bp insertion
GCR	IS insertion	4,381,583	4,381,585	1,446	-	-	*cycA*	D-alanine/ D-serine/ glycine permease	-	-	IS*150* + 3 bp insertion

REL4536 AN-1K-1

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
GCR	IS insertion	749,665	749,667	1,255	-	-	*yeeI*	Hypothetical protein	-	-	IS*3* + 5 bp insertion
GCR	IS insertion	4,381,583	4,381,588	1,472	-	-	*cycA*	D-alanine/ D-serine/ glycine permease	-	-	IS*150* + 3 bp insertion

REL4536 AN-1K-3

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	1,438,030	1,438,030	1	A → G	Tyr 75 Cys	*adhE*	Alcohol dehydrogenase	-	-	-
GCR	IS insertion	4,381,583	4,381,588	1,472	-	-	*cycA*	D-alanine/ D-serine/ glycine permease	-	-	IS*150* + 3 bp insertion

REL4536 AN-1K-4

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	1,439,673	1,439,673	1	G → A	Ala 623 Thr	*adhE*	Alcohol dehydrogenase	-	-	-
SNP	Transition	1,706,026	1,706,026	1	G → A	Ala 547 Ala	*uup*	ABC transporter ATPase component	-	-	-
SNP	Transition	1,766,332	1,766,332	1	C → T	Asn 753 Asn	*pflB*	Pyruvate formate lyase I	-	-	-
GCR	IS insertion	4,381,583	4,381,588	1,472	-	-	*cycA*	D-alanine/ D-serine/ glycine permease	-	-	IS*150* + 3 bp insertion
GCR	IS insertion	4,490,400	4,490,402	1,349	-	-	*fimA*	Major type 1 subunit fimbrin (pilin)	-	-	IS*186* + 6 bp insertion

REL4536 AN-1K-6

Mutation Type	Mutation Class	Reference Position 1	Reference Position 2	Size (bp)	Change	Amino Acid Change	Reference Gene 1	Description Gene 1	Reference Gene 2	Description Gene 2	Notes
SNP	Transition	1,438,030	1,438,030	1	A → G	Tyr 75 Cys	*adhE*	Alcohol dehydrogenase	-	-	-
GCR	IS insertion	224,516	224,518	1,446	-	-	*rrsH*	16S ribosomal RNA	-	-	IS*150* + 3 bp insertion
GCR	IS insertion	2,872,437	2,872,439	1,446	-	-	*yqeB*	Hypothetical protein	-	-	IS*150* + 3 bp insertion
GCR	IS insertion	2,992,382	2,992,385	777	-	-	*kpsS*	KpsS protein capsular biosynthesis	-	-	IS1 + 9 bp insertion
GCR	IS insertion	3,637,233	3,637,236	1,446	-	-	*xylH*	D-xylose transporter subunit	-	-	IS*150* + 3 bp insertion
GCR	IS insertion	4,381,584	4,381,586	1,446	-	-	*cycA*	D-alanine/ D-serine/ glycine permease	-	-	IS*150* + 3 bp insertion
GCR	IS insertion	4,581,548	4,581,551	1,446	-	-	*nadR*	Nicotinamide-nucleotide adenylyl-transferase	-	-	IS*150* + 3 bp insertion

Table A.8. Genes deleted in the 33 kb deletion event observed in REL4536 *ΔsbcC::cat*-1K-1 and REL4536 *ΔsbcC::cat*-1K-8 clones.

Gene	Description
insB-6	IS*1* protein
insA-6	IS*1* protein
ECB_00514	Hypothetical protein
ECB_00515	Hypothetical protein
ECB_00516	Hypothetical protein
ECB_00517	Hypothetical protein
appY	DNA-binding transcriptional activator
ompT	Outer membrane protease
envY	DNA-binding transcriptional activator of porin
ybcH	Hypothetical protein
nfrA	Outer membrane bacteriophage N4 receptor
ECB_00524	Inner membrane bacteriophage N4 receptor
yhhI-2	Putative transposase
ECB_00526	Hypothetical protein
ECB_00527	Hypothetical protein
ECB_00528	Hypothetical protein
ECB_00529	Hypothetical protein
ECB_00530	Hypothetical protein
cusS	Sensor kinase
cusR	DNA-binding transcriptional activator
cusC	Copper/silver efflux system outer membrane protein
ylcC	Periplasmic copper-binding protein
cusB	Copper/silver efflux system membrane fusion protein
cusA	Copper/silver efflux system, membrane component
pheP	Phenylalanine transporter
ybdG	Putative mechanosensitive channel
nfnB	Dihydropteridine reductase
ybdF	Hypothetical protein
ybdJ	Putative inner membrane protein
ybdK	Carboxylate-amine ligase

Table A.9. Colony morphotypes of REL4536 Δ*sbcC::cat*-1K populations and clones after 1,000 generations.

Lineage	Population morphology	Sequenced clone morphology
REL4536 *ΔsbcC::cat*-1K-1	Mixed*	Small
REL4536 *ΔsbcC::cat*-1K-6	Typical†	Typical
REL4536 *ΔsbcC::cat*-1K-7	Mixed	Typical
REL4536 *ΔsbcC::cat*-1K-8	Mixed	Small
REL4536 *ΔsbcC::cat*-1K-12	Mixed	Typical
REL4536 *ΔsbcC::cat*-1K-14	Mixed	Typical

*Mixed populations were those with both typical and small sized colonies.

†Typical morphology refers to colonies that are similar in size to the ancestor.