EVOLUTION AT A HIGH IMPOSED MUTATION RATE

ALEXANDER WILCOX

A thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of Doctor of Philosophy

Submitted:
July 2017

COPYRIGHT STATEMENT

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Ben Dickins for his guidance and supervision throughout my PhD, and whose efforts have both made me a better scientist, and expanded my vocabulary. Never again will I think that “flense” is an obscure word limited to use by whalers and 19th century American novels. Thanks also goes to Alan McNally, Gina Manning, Oyeronke Ayansola, and all the users of the CELS micro lab at NTU.

I would also like to thank Andy Millard, Nick Waterfield, and the members of lab M116 at the University of Warwick for providing me with a stimulating environment for writing this thesis while continuing to work with bacteriophage. A special mention goes to Dan Macqueen at the University of Aberdeen for giving me the opportunity that started me off in evolutionary research.

Finally I would like to thank my family and friends for providing support or distraction as required.

TABLE OF CONTENTS

ABBREVIATIONS

ASCII	American Standard Code for Information Interchange
BER	Base-excision repair system
BAM	Binary alignment map (filetype)
bp	Base pair(s)
CDS	Coding DNA sequence
cfu	Csolony forming units
Cq	Quantification cycle
DFE	Distribution of fitness effects
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dNTPs	Deoxynucleoside triphosphates
dsDNA	Double stranded DNA
Indels	Insertion and/or deletion mutations
LB	Lysogeny broth
LPS	Lipopolysaccharide
MMR	Mismatch repair
MNNG	N-methyl-N’-nitro-N-nitrosoguanidine
MOI	Multiplicity of infection
NGS	Next-generation sequencing
OD600	Optical density at 600nm
PCR	Polymerase chain reaction
PEAR	Paired end read merger (bioinformatic tool)
pfu	Plaque forming units
pgu	Phage genome units
qPCR	Quantitative PCR
RF DNA	Replicative form DNA
RNA	Ribonucleic acid
rpm	revolutions per minute
SAM	Sequence alignment map (filetype)
sgr	Substitutions per genome per replication
SNP	Single nucleotide polymorphism
sng	Substitutions per nucleotide per generation
ssDNA	Single stranded DNA
SSM	Slipped strand mispairing
TAE	Tris-acetate-EDTA buffer
tRNA	Transfer RNA
UV	Ultraviolet
VCF	Variant call format (filetype)

ABSTRACT

Mutation is an evolutionary process that provides much of the genetic variation required for natural selection and genetic drift. The mutation rate is a major driving force behind evolution, but is also an evolved characteristic itself. Although there is much theoretical research into why it evolved to the rate it did, there is little experimental work that investigates how its alteration affects evolution.

I created a novel system in which bacteriophage ΦX174 could be evolved at a mutation rate two orders of magnitude higher than wild-type. This system used a defective proofreading gene in the host polymerase to cause mutagenesis, and did not require the use of external mutagens which often conferred a biased mutational spectrum and harmful non-mutagenic effects.

Replicate populations of ΦX174 were evolved in both wild-type and mutagenic conditions for approximately 300 generations. One mutagenic population displayed a faster rate of adaptation than the wild-type lines, acquiring many of the same adaptive mutations in a shorter time frame, and rapidly increasing in fitness . While the wild-type lines were characterised by periodic selective sweeps of individual mutations, the mutagenic conditions allowed many of these mutations to rise in frequency due to a single selective sweep. The other mutagenic population, however, evolved very differently, with an early decrease in fitness that it did not recover from. This population acquired many mutations not seen in the other lines, and lacked many of the common adaptive mutations.

The mutation rate increase was not high enough to cause extinction of the viral population. The vastly different outcomes for the two replicate populations show that while an elevated mutation rate can in turn increase the rate of adaptation, it can also prevent it altogether by altering the genetic background and "locking out" potential adaptive mutations through negative epistasis.

Chapter 1: Introduction

1.0 Introduction

"As many more individuals of each species are born than can possibly survive; and as, consequently, there is a frequently recurring struggle for existence, it follows that any being, if it vary however slightly in any manner profitable to itself, under the complex and sometimes varying conditions of life, will have a better chance of surviving, and thus be naturally selected. From the strong principle of inheritance, any selected variety will tend to propagate its new and modified form." - On the Origin of Species, 1859.

With these lines, Charles Darwin summarised the requirements for evolution: competition, variation, and heritability. The existence of these were all self-evident facts, backed up by years of observation. Yet it would be nearly a century before Watson and Crick unravelled the genetic code and the mechanisms controlling variation and heritability could be elucidated. It is now well understood how genetics fits with Darwin's original theory. DNA replication is an extremely accurate process, allowing each generation to pass on its genome to the next. But if replication fidelity were perfect, all individuals would be clones and the theory would falter. Genetic variation is required for species to adapt and change, and comes from two sources: recombination and mutation. Recombination, the joining of genetic material from multiple sources, is the defining characteristic of sexual organisms, where parental chromosomes cross over as part of meiosis. Yet for the variation between the parental genomes, and thus recombination, to be present at all, an original source is required: mutation.

Mutation is the ultimate source of variation. Without this process, DNA replication would be completely accurate and every individual would be clonal. Selection would not exist and all organisms would be genetically identical to their parents, and to their children. Mutation is therefore the fuel that drives evolution and the cause of all the diverse species on Earth today, from humans to viruses, and from plants to archaea. However, mutation is also a relatively uncommon event, and it is unclear exactly why it occurs at the rate it does. While there is a large body of theoretical work that tries to explain the causes and consequences of different mutation, there is little empirical evidence.

1.1 Mutation

While mutation refers to any change that alters the nucleotide sequence of DNA or RNA, the mechanisms responsible and the changes produced can vary greatly. This section will review the types and causes of mutations, as well as the evolution of the mutation rate.

1.1.1 Substitutions

Substitutions (figure 1.1) are the most common type of mutation, and occur when a single nucleotide is substituted for another. If a substitution occurs within a protein coding region of DNA, it will result in one codon changing into another. A missense mutation changes the codon to one with a different amino acid specificity to that of the original, potentially leading to changes in structure or function of the resultant protein. When the codon is changed to a stop, it is called a nonsense mutation, and results in early termination of protein synthesis. Finally, a synonymous (or silent) mutation occurs when the new codon has the same amino acid specificity of the original, leaving protein structure unaffected.

Substitutions can be separated into two distinct classes. In transition mutations, a purine (adenine or guanine) is substituted for the other purine, or a pyrimidine (cytosine, thymine or, in the case of RNA, uracil) for the other pyrimidine. In transversion mutations, a purine is substituted for a pyrimidine, or vice-versa. Although there are twice as many possible transversions than transitions, transitions occur at a greater frequency (undefined author and Li, 2013). Substitutions are the most common type of mutation (Chen *et al.*, 2009) and because they are typically only capable of modifying a single amino acid, are some of the most important in producing gradual evolutionary change. Yet while many point mutations are advantageous to an organism, others have been demonstrated to be strongly deleterious and the cause of many diseases, including sickle cell anaemia, cystic fibrosis, and some cancers (NCBI, 1998).

Although synonymous mutations are generally considered selectively neutral, there are circumstances in which this is not the case. For example, codon bias means some organisms preferentially produce some tRNAs for specific amino acids over others. It is thought that if a mutation switches to a codon associated with a less abundant tRNA, it may interfere with the timing of cotranslational folding causing a change in the tertiary structure of the protein, despite the amino acid sequence being unchanged (Kimchi-Sarfaty *et al.*, 2007).

Substitutions can also appear in non-coding regions of the genome. While these may be presumed to be selectively neutral, this is not always the case. Substitutions in promotor regions can affect gene expression, while substitutions in DNA binding sites can interfere with processes such as ribosome binding or mismatch repair (Zhen and Andolfatto, 2012).

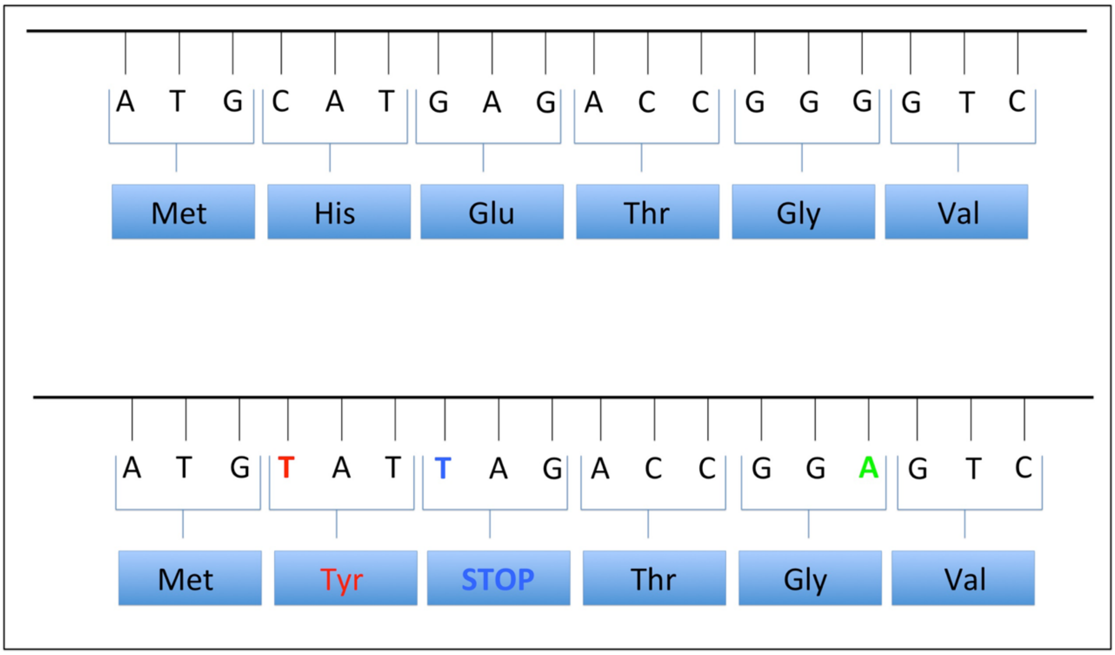

A T G C A T G A G A C C G G G G T C
Met
His
Glu
Thr
Gly
Val
A T G T A T T A G A C C G G A G T C
Met
Tyr
STOP
Thr
Gly
Val

Figure 1.1 - Substitution mutations. In this example, three substitutions have occurred after replication, denoted by different coloured bases. (i) Missense mutation (red): In the CAT codon, the cytosine has been replaced with a thymine, changing the codon from histidine to tyrosine. (ii) Nonsense mutation (blue): The first guanine in the GAG codon has been substituted for a thymine, resulting in a TAG stop codon, (iii) Synonymous mutation (green): The third guanine in the GGG codon has changed to adenine, but the new GGA codon still encodes the same amino acid as before.

1.1.2 Insertions and deletions

Insertion and deletion (indel) mutations happen when one or more nucleotides are inserted or removed from a sequence. In a coding sequence, if the number of bases of the indel is a multiple of three, this will result in the addition or removal of whole codons, as well as potentially altering adjacent codons. A frameshift mutation (figure 1.2) takes place when the indel length is not a multiple of three. All nucleotides at the 3' end of the indel are shifted along the strand, ending outside of their original reading frame. Frameshift mutations lead to a large number of codons being changed and can cause stop codons to appear earlier or later than the original. Resulting proteins and transcripts can therefore vary significantly in length and sequence to those of the original, and are usually non-functional. Frameshift mutations have been found to be strongly selected against in mutation accumulation experiments (Heilbron *et al*., 2014), and are generally considered to be deleterious.

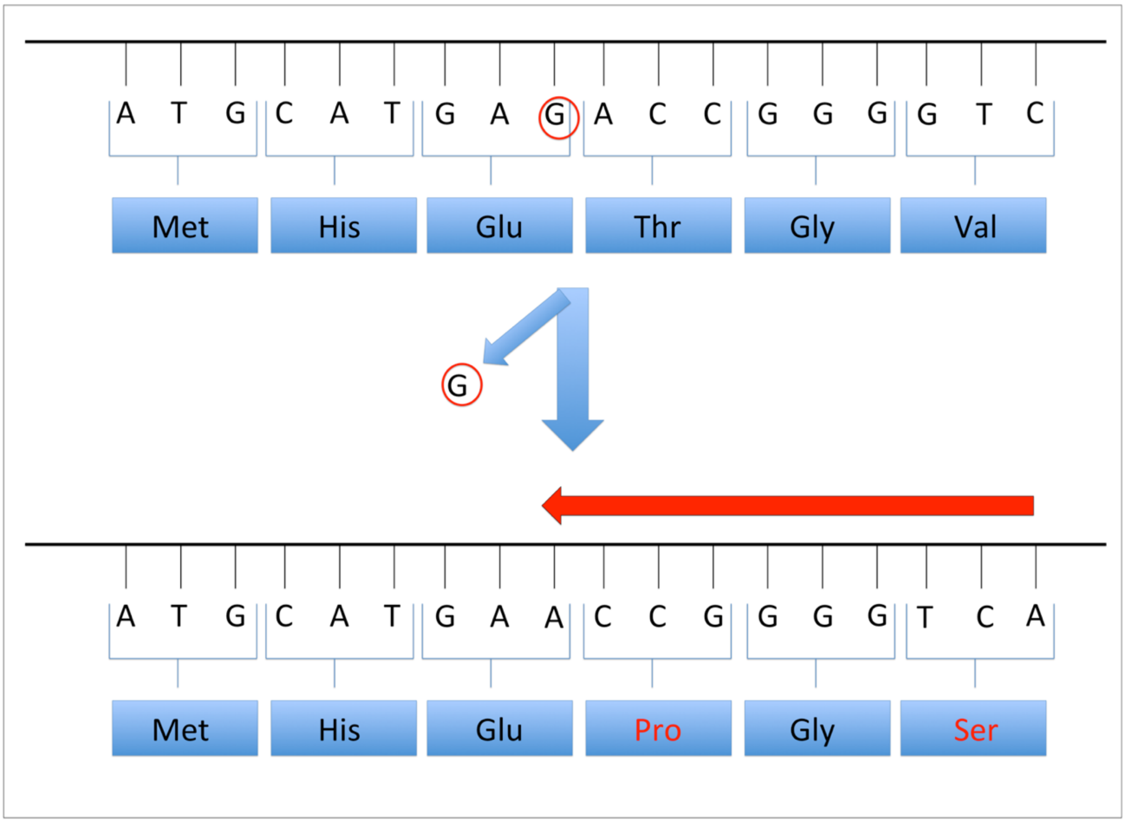

A T G C A T G A G A C C G G G G T C
Met
His
Glu
Thr
Gly
Val
G
A T G C A T G A A C C G G G G T C A
Met
His
Glu
Pro
Gly
Ser

Figure 1.2 - Frameshift mutation. During replication, a guanine has been deleted from the new strand, causing subsequent 3' nucleotides to be shifted left one position. This leads to multiple codon changes in the sequence.

1.1.3 Causes of mutation

Those mutations that arise stochastically during DNA replication are known as spontaneous mutations. These typically occur due to biochemical interactions that change the state of the nucleotides, changing the way they pair (Griffiths *et al.*, 2004).

The four nucleotides appear in tautomers, constitutional isomers which readily interconvert. In DNA, the tautomer commonly seen is the keto form although tautomers can spontaneously change to another form when a proton relocates. This affects the hydrogen bonding pattern in the nucleotide, which leads to mutation if the base is mispaired during DNA replication. Bases can also become spontaneously ionised, which can lead to mispairing.

Depurination reactions occur when a β-N-glycosidic bond is cleaved in a purine deoxyribonucleoside, releasing the adenine or guanine base and leaving an abasic site, a deoxyribose in the DNA chain that has no base. In dsDNA, abasic sites can be effectively repaired by the base-excision repair system (BER) which determines the missing base by pairing with the corresponding base on the opposite strand (Krokan and Bjørås, 2013). Because this information is not available in ssDNA, the BER will instead insert a random base, which can result in either a transversion or transition mutation.

Deamination in DNA is a hydrolysis reaction where the addition of water leads to the release of ammonia. This affects the bases differently. For example, deamination of cytosine results in uracil which, following replication, will produce a C:G --> T:A transition mutation (Coulandre *et al.*, 1978).

Slipped strand mispairing (SSM) occurs during DNA replication when the two strands denature then reanneal incorrectly, causing mispairing. This type of event results in indels and is thought to be one of the major causes of repetitive regions in DNA. Repetitive regions are self-accelerating - as their size increases so too does the likelihood of mispaired strands, which in turn leads to a larger repetitive region (Levinson and Gutman, 1987).

Induced mutations are those caused by external influences such as chemicals and radiation. Although the exact mechanisms by which this occurs vary, they typically directly affect the DNA by ionising bases or changing their configuration to a different tautomer to the usual keto form. During DNA replication, point mutations occur when these errant tautomers and ionised bases are mispaired.

1.1.4 Large-scale mutations

Whereas substitutions and indels typically affect only a small number of nucleotides, there are also mutations that can affect the genome on a larger scale. However, many of these types of mutation are more common in eukaryotes.

Gene duplication events involve the amplification of any region of DNA that contains a gene, although multiple genes could be copied together. These arise during recombination, such as when repeated sequences cause two chromosomes to become misaligned while crossing over. This unequal crossing-over can lead to a region of DNA being donated from one chromosome to the other, resulting in a duplication on one chromosome and a deletion on the other. Since gene duplications rely on recombination, they are seen more frequently in eukaryotes. Deletions can also occur on a larger scale, and can result in the removal of whole genes or chromosomal regions.

It is very unlikely that short indels are responsible for the enormous discrepancies in genome sizes between the largest eukaryotes and smallest viruses. Susumu Ohno (1970) considered gene duplication to be the single major driving force behind evolution. With an extra copy of a gene, formerly deleterious mutations in one paralogue would no longer be selected against as long as the other was still producing a functioning protein. With natural selection no longer preserving the ancestral function, the duplicated gene would be able to freely acquire more mutations and develop its own function. The power of gene duplication is demonstrated in the vertebrate calpain gene family, with phylogenetic analysis showing when duplications occurred, and molecular evolution analyses showing that selection was relaxed after many of these events (Macqueen and Wilcox, 2014).

When a gene is duplicated, the new paralogue is still a variant of the original protein, retaining its structure and function. (Ohno, 1984) proposed that frameshifts in duplicated genes allowed an unused reading frame to become active without the strong negative selection that would occur if it appeared in the only copy of an essential gene. This would lead to the synthesis of a novel polypeptide with no pre-existing biological function that could tolerate many further mutations without the constraints of selection, with the potential to eventually lead to a new, unique function. Although Ohno's example of a gene in *Flavobacterium* arising in this way was later shown to be incorrect (Negoro *et al.*, 2005), genomic analysis has demonstrated multiple instances where genes share homology with other gene families, despite being in different reading frames (Raes and Van de Peer, 2005).

Although gene duplications have been described frequently in organisms ranging from higher eukaryotes to small DNA viruses, in RNA viruses they have been rarely observed throughout evolutionary history. It is thought that this is due to strong selection against large genome sizes in RNA viruses rather than them not occurring (Simon-Loriere and Holmes, 2013).

As well as gene duplications, chromosomes or even whole genomes can become duplicated. These will not be covered in detail here because they are not applicable to the work presented in this thesis.

1.1.5 Evolution of the mutation rate

While mutation is the fuel for evolutionary change, the rate at which it happens is also an evolved characteristic. Throughout life, there are many molecular mechanisms that have arisen which work to prevent mutations from occurring and repair damage to DNA, leading to an effective decrease in the mutation rate. These mechanisms are presumably favoured by selection due to their capacity to reduce the rate of deleterious mutations (Sniegowski *et al.*, 2000).

While it may be intuitive to imagine that the mutation rate evolved to an optimal level that balances the fitness cost of deleterious mutations with the fitness gain from beneficial mutations, the physiological cost of modifying the mutation rate is also a contributing factor. This means that while the optimal mutation rate for evolution would be determined by a trade-off between minimising the number of deleterious mutations while ensuring beneficial mutations occurred at a high enough frequency, the actual mutation rate may not have evolved to this optimum level, instead being a balance between the fitness costs of having a high frequency of deleterious mutations, and the fitness costs of having and maintaining modifiers (Kimura, 1967).

Much variation has been detected in the effectiveness of these modifiers. Antimutator phenotypes have been discovered in T4 phage and *E. coli*, indicating that ever lower mutation rates are achievable (Schaaper, 1998). It is unclear whether lower rates are not usually seen in nature because they are beneath the optimum rate for generating beneficial mutations, or that the physiological costs of reducing them further are too great.

Since mutation rate modifiers act on other genes, selection will act on the mutations they generate rather than the modifiers themselves. Linkage disequilibrium means that the modifiers can still be indirectly selected for because they can hitchhike along with any beneficial mutations that they are responsible for generating. In fact, this phenomenon has been credited with the evolution of mutator phenotypes that are sometimes observed in bacteria (Raynes and Sniegowski, 2014). However, recombination disrupts the link between the modifier and the mutation, meaning indirect selection to increase the mutation rate is dependent upon the recombination rate. Since deleterious mutations are far more common than beneficial mutations and their selection is less affected by the erosion of linkage disequilibrium, selection that reduces the frequency of deleterious mutations is likely to be much stronger than selection that increases the frequency of beneficial mutations (Barton, 2010).

It is thought that elevated mutation rates will only be maintained over long periods when selection pressures change rapidly, such as antagonistic coevolution with a parasite. For example, mutator phenotypes frequently arise in populations of *Psuedomonas fluorescens* when it is experimentally evolved in the presence of a phage *in vitro* (Pál et al., 2007). However, these results could not be replicated in the bacteria's normal environment, suggesting that this phenomenon may be less common outside the laboratory (Gómez and Buckling, 2013).

It seems likely therefore that the major selective pressure for an increased mutation rate is the cost of having mutation rate modifiers, and the mutation rate evolved to a level that balances the fitness cost of modifiers with the fitness cost of deleterious mutations. A further hypothesis is that in eukaryotes the lower limit to the mutation rate is set by genetic drift (Lynch, 2010).

1.1.5.1 Drake's rule

Mutation rates can vary by orders of magnitude between different organisms, and even in different parts of the same genome. Drake (1991) compared the mutation rates of haploid DNA organisms ranging from viruses to bacteria to fungi. Despite the genome sizes varying by over 3 orders of magnitude and the mutation rate per nucleotide varying by over 4 orders of magnitude, the mutation rates per genome were all within the same order of magnitude, with the majority falling into the range 0.003-0.004. It follows from this that the mutation rate per nucleotide is inversely proportional to genome size. This has come to be known as Drake's rule, but it must be noted that since the sample size is small and methods have varied it is more of a correlation than a hard and fast law. However, further mutation rate estimates appear to support Drake’s rule (Sung *et al.*, 2012).

Drake's rule only applies to microorganisms, including dsDNA and ssDNA viruses and bacteriophages but excluding RNA viruses which have mutation rates that are on average greater by 2 orders of magnitude and do not fall into this range (Drake and Holland, 1999). Although single cellular eukaryotes appear to follow Drake's Law, higher multicellular organisms do not. It is unclear exactly why this is, but could include factors such as the removal of mutations via meiosis and recombination, the number of DNA replications that occur within a multicellular organism, and the large amount of non-coding DNA in higher organisms.

1.1.6 Consequences of evolution at a high mutation rate

Since the majority of non-neutral mutations are deleterious (Eyre-Walker and Keightley, 2007), it would be expected that at high mutation rates selection would be unable to remove them all, resulting in a decline in fitness over generations.

There are also a number of evolutionary processes that must be considered at high mutation rates, thought to affect this fitness decline either by exacerbating or offsetting it. These include mutation-selection balance, Muller's ratchet, clonal interference, mutational robustness, error catastrophe, and adaptive and compensatory evolution (Keller *et al.*, 2012).

1.1.6.1 Mutation-selection balance

Mutation and selection are two opposing forces - mutation acts on a population to create individuals that are less fit than the rest, and selection purges them to ensure that the population's fitness is maintained. How well mutation and selection are balanced determines how much of the population contains a deleterious allele, and hence its fitness.

When an organism replicates, the progeny will occasionally contain one or more mutations. The majority of these are deleterious and removed by selection. However, as mutation rate increases, it is harder for selection to remove all the deleterious mutations. Whereas usually the less-fit genotypes are outcompeted by unmutated genotypes, under mutagenic conditions they have to compete with multiple deleterious genotypes that are present at the same time. This has two immediate consequences. Firstly, the rate at which selection operates is decreased, since the difference in selection coefficients between two deleterious genotypes is smaller than that between a deleterious genotype and the wild-type. Secondly, deleterious mutations with smaller fitness costs are actually selected for when competing with deleterious mutations with large fitness costs.

If the mutation rate is increased to a level that selection cannot efficiently remove deleterious mutations, they will accumulate in the population until a new equilibrium is reached, referred to in population genetics as mutation-selection balance (Haldane, 1927). Selection presents difficulties in models, since it relies on data such as the distribution of mutational fitness effects in a genome which are too vast to measure empirically and too complex to predict. Kimura and Maruyama (1966) produced a simplified model that describes the mean fitness (*W*) of an asexual population at equilibrium (relative to 1, the fitness of the mutation-free genotype), determined only by the genome-wide deleterious mutation rate (*U*).

$$W = e^{-U}$$

Kimura and Maruyama's model says that as *U* increases, *W* will decrease. When *U* is sufficient that *W* tends towards 0, the accumulation of deleterious mutations should be too high for the population to maintain, leading to extinction. This equation has been used as a starting basis for models that attempt to describe lethal mutagenesis (Springman et al., 2010) (Bull et al., 2007). However, it treats all deleterious mutations as selectively equal, and does not take into account adaptive evolution, recombination, and epistasis, as well as many of the evolutionary processes discussed in the remainder of this introduction.

1.1.6.2 Muller's ratchet

Muller's ratchet is the name given to a model that describes the accumulation of mutations in an asexual population (Muller, 1964). This model makes two assumptions:

a) The population size is finite.

b) Deleterious mutations are irreversible (i.e. back mutations are rare enough that they are ignored)

In asexual organisms, all genomes are passed on from parent to offspring as a single unit. If a deleterious mutation appears in an individual it will invariably be present in all that individual's offspring. This means the only way for deleterious mutations to be removed is by purifying selection. However, if the mutation rate and selection coefficients are such that selection cannot remove every deleterious mutation, eventually genetic drift will lead to every member of the population acquiring at least one mutation. Once every individual contains a deleterious mutation the original unmutated genome is lost to the population. Like a ratchet with teeth that mean it can only turn forwards and never back, a further "turn" can introduce more deleterious mutations and change the genome further, but the previous fittest genome is also lost. As further deleterious mutations accumulate within the population, each fittest genotype will eventually go extinct to be replaced with the second fittest, until the fitness decline is enough to cause extinction of the entire population (figure 1.3).

Recombination allows organisms an additional way to purge deleterious mutations. If a parent has a deleterious mutation it is possible for their offspring to avoid inheriting it by combining their DNA with a second parent that does not have the same mutation. Muller proposed that asexual species were at a disadvantage to those that reproduce sexually, because of this extinction risk. Muller's ratchet has been put forward as a possible cause of the degeneration of the Y chromosome (Engelstädter, 2008), and mitochondrial DNA (Loewe, 2006). Because these are passed down through the paternal and maternal lines respectively, they do not recombine during sexual reproduction and are simply clonal copies of those from parents.

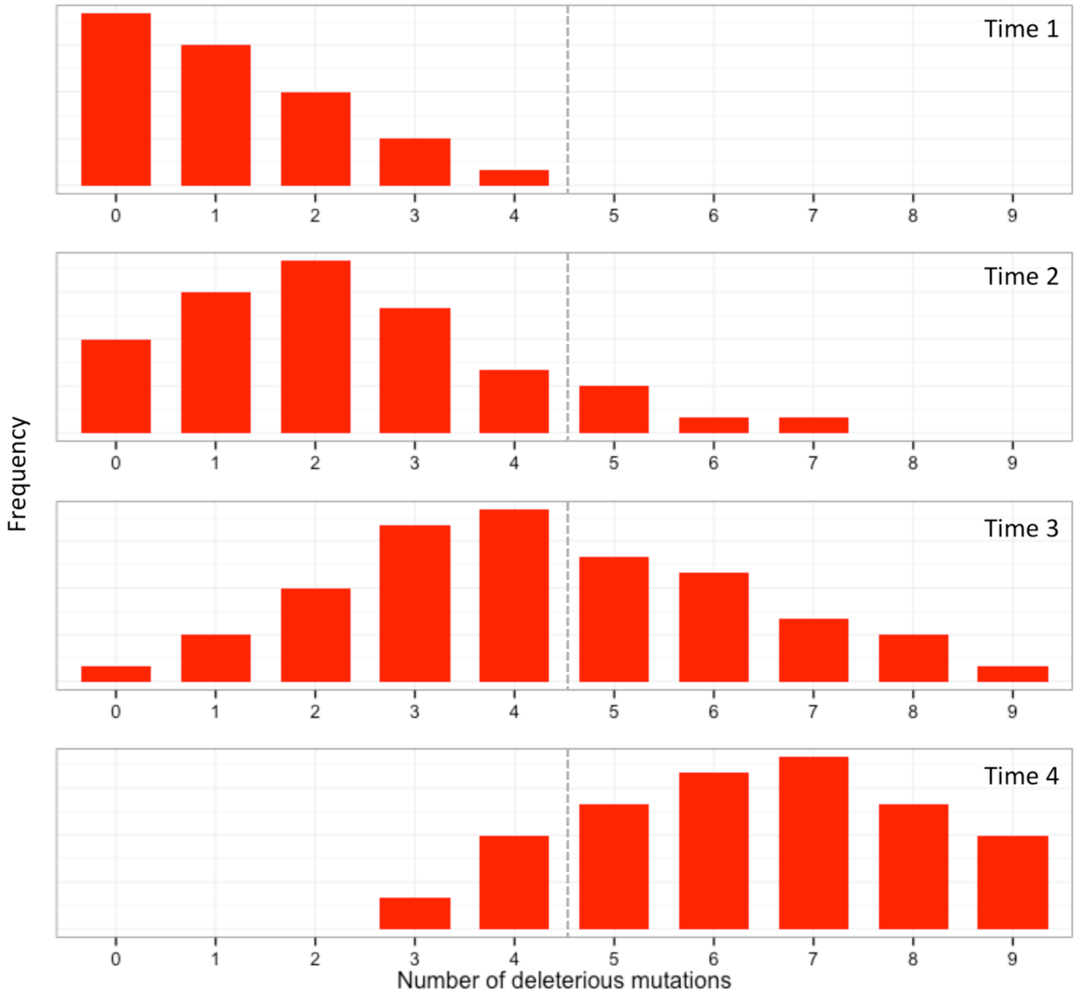


Figure 1.3 - Muller's ratchet. Distribution of mutational load in an asexual population over time. As mutations accumulate due to genetic drift, the distribution is shifted to the right. Eventually, the unmutated genotype is lost, followed by genotypes with one deleterious mutation, and so on, until the extinction threshold (dotted line) is reached.

1.1.6.3 Clonal interference

When a beneficial mutation appears in a population, positive selection can cause it to increase in frequency over subsequent generations until it becomes fixed, in a process known as a selective sweep. The traditional view is that with beneficial mutations being a rare occurrence, and beneficial mutations that are not lost to genetic drift rarer still, adaptive evolution is driven by sequential selective sweeps (ATWOOD et al., 1951).

However, with high mutation rate or population size it is possible for multiple beneficial mutations to arise simultaneously in different lineages. In a sexual population, recombination can merge these mutations into a single genotype, but in an asexual population where recombination is rare or absent, these lineages instead end up competing against each other, a phenomenon known as clonal interference.

Clonal interference has a number of effects on the dynamics of an evolving population. Beneficial mutations that would normally become fixed via a selective sweep if they occurred on their own may end up being selected against and lost if mutation with a higher selection coefficient is competing against them. The mutation that outcompetes other beneficial mutations is also affected. Because the difference in selection coefficients between two beneficial mutations is smaller than the difference between the most beneficial mutation and the wild-type, the time taken for this beneficial mutation to reach fixation is increased, slowing the rate of adaptation.

One consequence of clonal interference is that it is predicted to put a "speed limit" on evolution (Gerrish and Lenski, 1998) and may in itself be a factor in determining the evolution of the mutation rate. Consider that as the mutation rate increases, so does the frequency of beneficial mutations. It follows that frequent beneficial mutations will lead to more instances of clonal interference, which will lower the strength of selection and in turn lower the rate at which fitness improves.

1.1.6.4 Mutational robustness

While mutations in protein coding DNA sequences (CDS) often result in products with altered or impaired functions, it is also possible that the phenotype will be unaffected by these changes. This is known as mutational robustness and is the extent to which the phenotype of a population remains unaffected by changes in genotype.

Robustness allows an organism to be more tolerant of mutations. When bacteriophage T4 lysozyme mutants were created to test the effects of multiple amino acid substitutions at every position, it was found that over 55% of the substitutions did not affect the enzyme's lysis activity (Rennell *et al.*, 1991). Robustness also extends to the genomic level, with many organisms being tolerant to gene deletions or inactivations. Gene-deletion mutants for 96% of *Saccharomyces cervisae* genes were created, showing that only 18.7% of them were essential for growth under standard conditions (Giaever *et al.*, 2002).

Since robustness is intrinsically linked to mutation, it directly affects evolution. While it may be easy to think of robustness and evolvability as opposite ends of a scale, the link between the two is complex. For example, robustness can actually increase evolvability by facilitating epistasis. Robustness essentially increases the proportion of mutations that are selectively neutral. Because these are more likely to persist at low frequencies in a population due to genetic drift, the number of genetic variants in a robust population is wider. In a genetically homogeneous population, every possible mutation will have a single effect. However, in a population of genetically robust organisms, there may be numerous closely related genotypic backgrounds. This means that new mutations may have different effects depending on the organism they arise in, widening the spectrum of potential adaptations (Draghi *et al.*, 2010).

Genetic redundancy refers to genes that can be inactivated with little effect on phenotype because other genes compensate for their loss. Selection on redundant genes is often relaxed, providing another route for robustness to evolve. It is highly debated whether mutational robustness itself is a selected trait or just a by-product of other processes (Montville *et al.*, 2005).

1.1.6.5 Error catastrophe

Error catastrophe is a term that lacks a strict definition. In many studies that attempt to eradicate populations of RNA viruses by mutagenesis, error catastrophe is equated with lethal mutagenesis.

An error catastrophe occurs when mutation rate exceeds a threshold so that the fidelity of the mutation-free genotype can no longer be maintained in a population and is lost (Keller *et al.*, 2012). Although this can be thought of as similar to extinction through the accumulation of deleterious mutations, error catastrophe does not specifically refer to extinction. In fact, it is possible that this threshold is exceeded, yet the population still meets a fitness equilibrium and does not go extinct. In this way, it is possible that extinction may be avoided altogether, by shifting the genotypes of the population to those that are more robust (Bull *et al.*, 2007). When studies equate viral extinction with error catastrophe it can be misleading, since the mutation rate threshold that causes extinction will be higher than that which leads to error catastrophe and a fitness equilibrium.

Error catastrophe is intrinsically linked to the viral quasispecies model (Eigen, 1971). As opposed to a traditional population where genotypes are assumed to be homogenous, a quasispecies is a population of closely related but differing genotypes that exists at high mutation rate. This is generally used to describe RNA viruses.

In a term known informally as "survival of the flattest", the model predicts that quasispecies which occupy a low but flat position on the fitness landscape (i.e. are mutationally robust) will outcompete quasispecies on a peak (i.e. high fitness but low robustness) because they are less sensitive to deleterious mutations (Tejero *et al.*, 2011).

1.1.6.6 Adaptive and compensatory evolution

Many models of evolution at high mutation rates ignore adaptive evolution. However, when one of these models was tested empirically it was found that rather than the predicted decline being observed, a population actually increased in fitness (Springman *et al.*, 2010). Even if the assumption that a population is at optimum fitness could be enforced experimentally (by pre-adapting it to laboratory conditions, for example), it still fails to account for two other types of adaptation: reversions and compensatory mutations.

In a viral population, deleterious non-lethal mutations are constantly arising. While the majority are quickly lost through selection or genetic drift, this is not always the case. For example, if a mutation is only slightly deleterious, it might increase in frequency due to genetic drift, where it is able to persist for some time due to the weak selection acting upon it. Alternatively, if the net fitness change is still positive, deleterious mutations can increase in frequency by hitchhiking with beneficial mutations (Lang *et al.*, 2013).

Instead of being removed by selection, there are other ways in which a genotype containing a deleterious mutation can have its fitness restored. If, for example, a deleterious mutation were to revert to its ancestral state in a new mutational event, any loss of function or fitness would be reversed. Alternatively, a compensatory mutation could also occur, a second mutation that compensates for the fitness loss caused by the original by means of epistatic interaction.

Compensatory mutations appear to be common. This most likely results from there potentially being multiple compensatory mutations available for each deleterious mutation, and the fitness differences between each of these compensatory mutations and reversion. In one study, it was found that when a strain of bacteriophage ΦX174 that included deleterious mutations was grown, compensatory mutations evolved approximately twice as often as reversions (Poon, 2005). It was also estimated that on average there were at least nine potential compensatory substitutions for each deleterious mutation, although some deleterious mutations could not be compensated for at all.

At higher mutation rates where deleterious mutations are more common within a population, it follows that the frequency of compensatory mutations would also increase. Simulations have shown that compensatory mutations would be continually evolving and able to offset the fitness decline caused by the accumulation of deleterious mutations (Keller *et al.*, 2012).

A further potential consequence of frequent compensatory mutations is that they change the genetic background. Just as the compensatory mutation would not be beneficial in the absence of the deleterious mutation it compensates for, it may have potential epistatic interactions with other mutations, meaning mutations that may be beneficial or deleterious on a wild-type genome may not have the same effect in a genome containing compensatory mutations.

1.2 Bacteriophage ΦX174

1.2.1 Background

Bacteriophages (or phages) are the viruses of bacteria, the Greek word "phagein" from which their name is derived literally meaning "to devour". Much like viruses of eukaryotes, phages enter the cells of their hosts, where they replicate with the aid of intracellular host machinery.

Soon after their discovery approximately a century ago, there was much interest in the therapeutic properties of phages due to their ability to kill their bacterial hosts. However, most research into this area ceased after the discovery of antibiotics and the outbreak of war in 1939. The exception to this was the former Soviet Union where research in phage therapy continued throughout the Cold War, but political and language barriers prevented their findings from becoming known in the West. However, with antibiotic resistance a growing concern, interest in therapeutic use of bacteriophage has been renewed (Cisek *et al.*, 2016).

Bacteriophage ΦX174 will be the main organism used in the experimental portion of this thesis, so this section will discuss this phage in detail.

For such a simple organism, ΦX174 has a lot of history. In 1959, its genome became the first DNA molecule to be homogeneously purified (Sinsheimer, 1959), and in 1967 its genome was synthesised by DNA polymerase *in vitro*, opening the doors to the age of synthetic biology (M

Goulian, 1967). When pioneering the sequencing technique that bears his name, Fred Sanger chose to use ΦX174 as his subject, making it the first DNA genome ever to be completely sequenced (Sanger *et al.,* 1977, (Sanger et al., 1978). Shortly after, it was the first organism to be used in site-directed mutagenesis, for which Michael Smith received a Nobel Prize in Chemistry (Hutchison *et al*., 1978). More recently, it became the first man-made genome when Craig Venter's group assembled it *in vitro* from synthetic oligonucleotides. When introduced to host cells, this genome was capable of producing fully infectious ΦX174 virions (Smith *et al*., 2003).

ΦX174 is a tailless icosahedral bacteriophage belonging to the Microviridae family. Its capsid contains 60 copies each of two proteins that protect its genome, a circular molecule of single-stranded DNA (ssDNA) made up of 5,386 nucleotides encoding 11 genes. Its small genome is highly utilised, with 95% of nucleotides belonging to coding regions and several of its genes overlapping. Of these overlapping genes, one encodes a truncated form of a longer protein, the others in alternate reading frames encoding unique products (figure 1.5). The genes are summarised in table 1.1. During infection, ΦX174 follows the typical lytic cycle (figure 1.6).

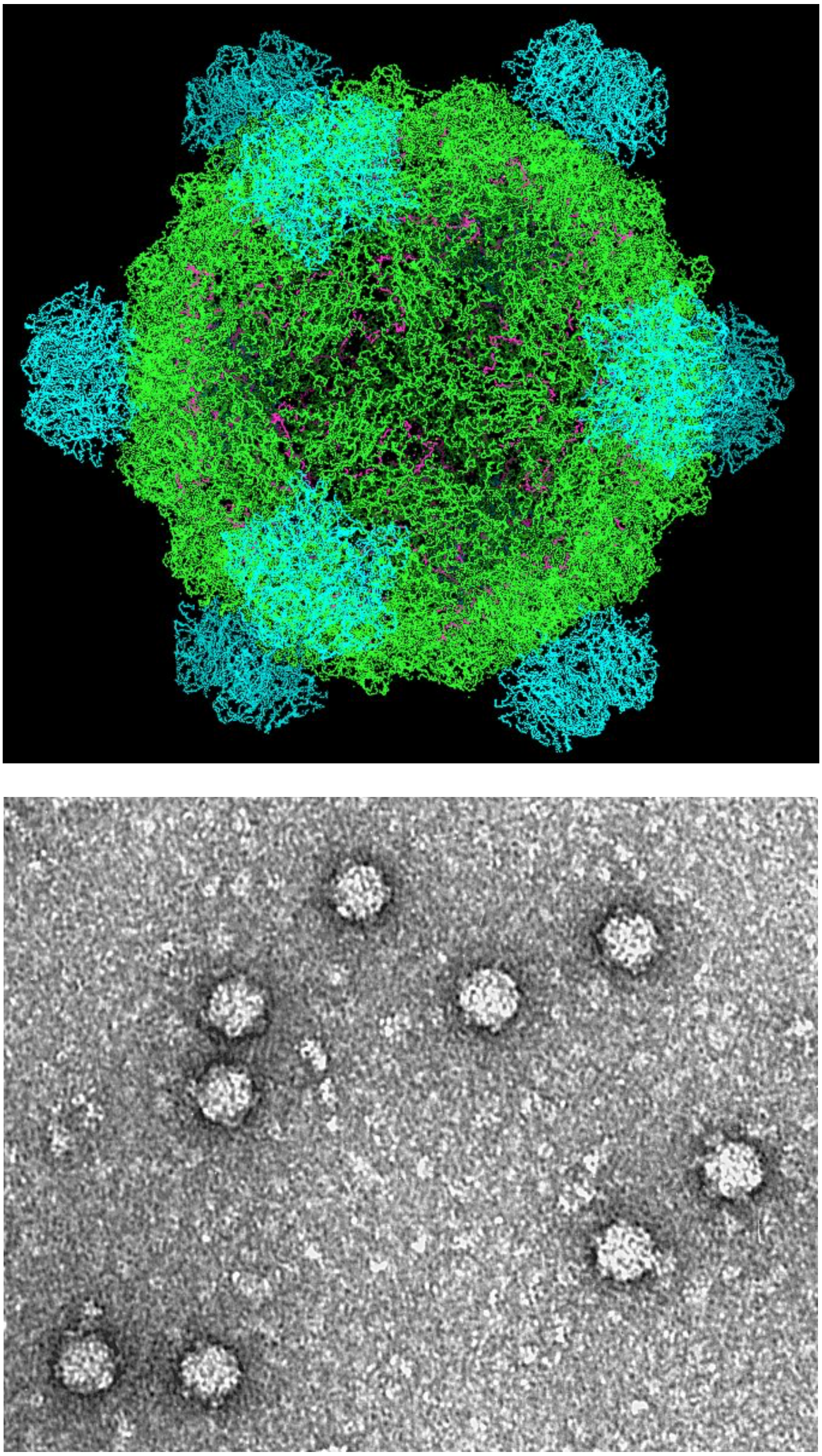


Figure 1.4 - Structure of the ΦX174 capsid (top), ΦX174 virions under electron microscopy (bottom)

Images by Fdardel (top) and ShiftFn (bottom), via Wikimedia Commons.

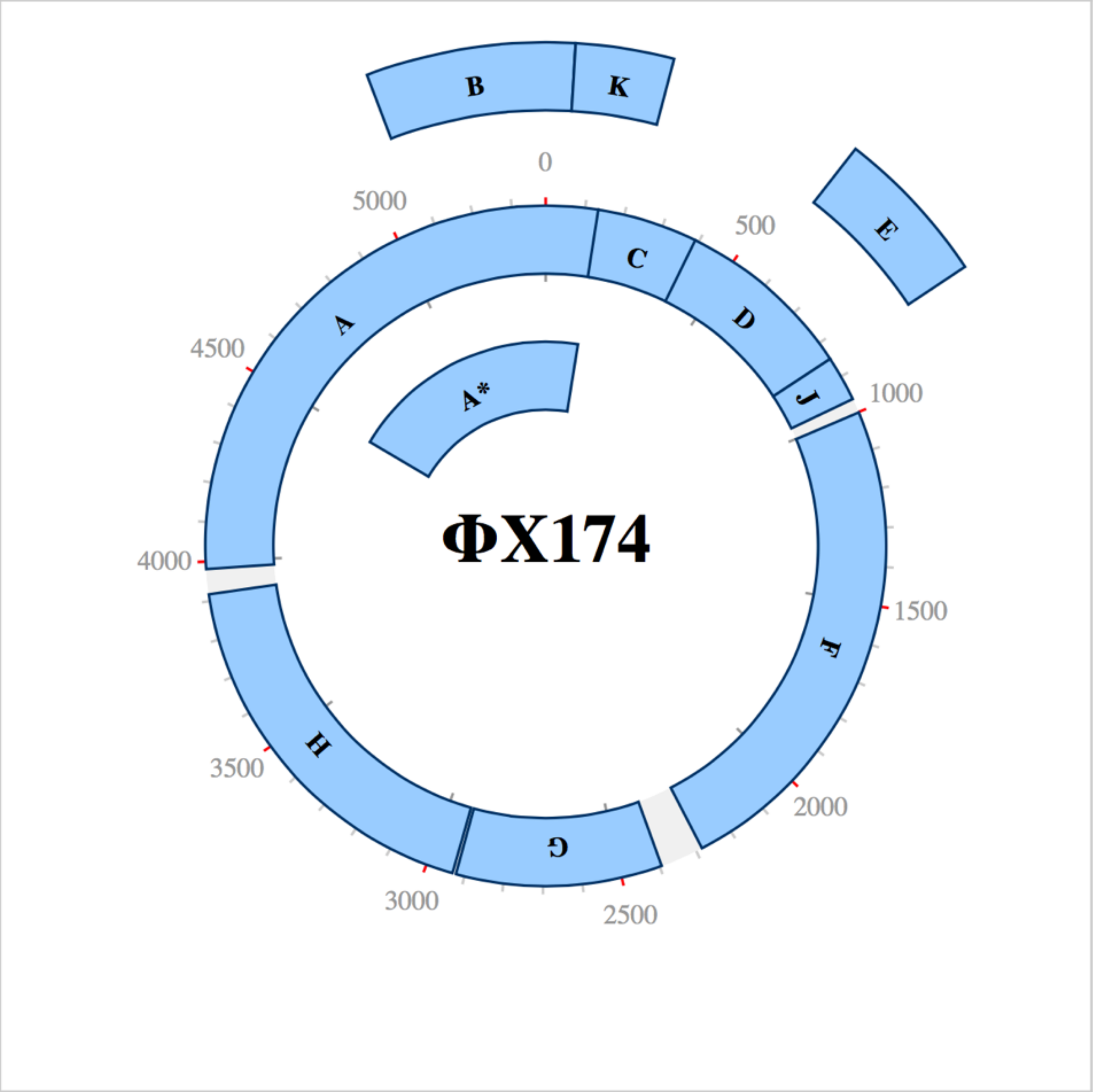


Figure 1.5 - Genome organisation of ΦX174. Genes B, K, and E overlap in different reading frames. Gene A* is a truncated form of A and shares the same reading frame. Created with AngularPlasmid.

Table 1.1 - Genes of ΦX174 and the functions of their protein products

Adapted from (Fane et al., 1988).

Gene	Function	Notes
A	DNA replication.	
A*	Inhibits host DNA replication.	Non-essential. Truncated form of protein A
B	Internal scaffolding protein used for procapsid assembly.	Overlaps gene A.
C	DNA replication	
D	External scaffolding protein used for procapsid assembly.	
E	Host cell lysis.	Overlaps gene D
F	Coat protein. 60 copies present in virion.	Green in figure 1.5
G	Spike protein. 60 copies present in virion.	Cyan in figure 1.5
H	DNA pilot protein. 10-12 copies present in virion.	
J	DNA binding protein. 60 copies present in virion.	
K	Burst size optimisation.	Unessential. Overlaps genes A and C.

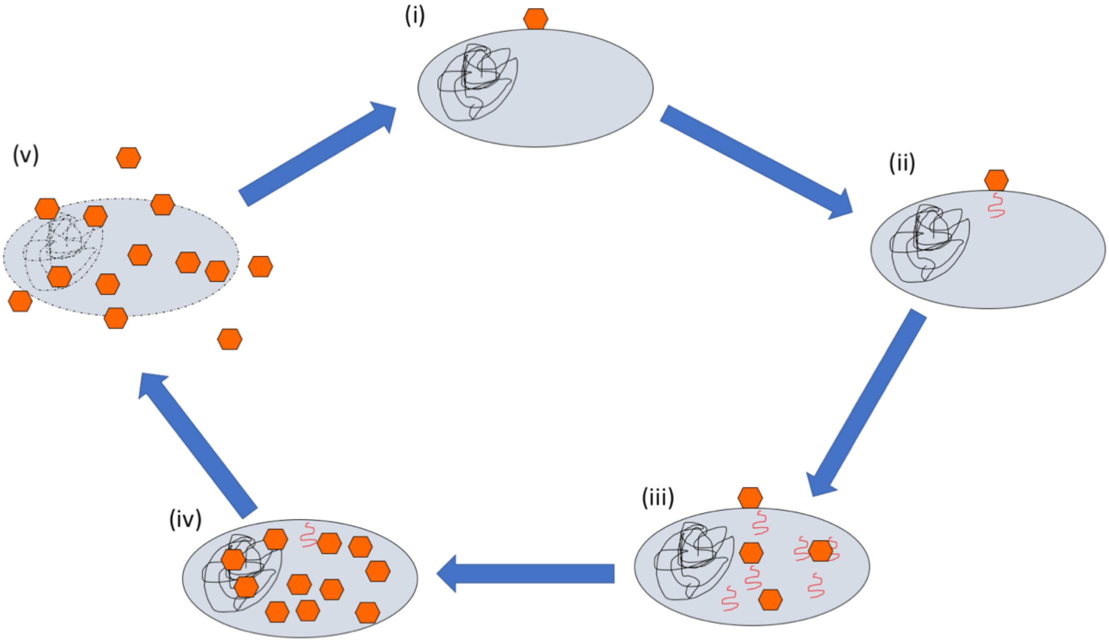


Figure 1.6 - Life cycle of ΦX174. (i) ΦX174 virion attaches to host bacterium. (ii) Genome is injected into the host cytoplasm. (iii) Replication of phage genome, synthesis of viral proteins, and assembly of progeny virions. (iv) Further replication and assembly, beginning of host lysis. (v) Lysis of host cell and release of ΦX174 virions.

1.2.2 Life cycle of ΦX174

1.2.2.1 Attachment

ΦX174 infects bacteria of the Enterobacteriaceae family, with its known hosts including strains of *Escherichia coli, Salmonella typhimirium* and *Shigella sonnei* that contain rough forms of lipopolysaccharide (LPS) in their outer membranes (Wichman and Brown, 2010).

Attachment of ΦX174 to host cells begins with a reversible reaction caused by an interaction of glucose molecules in LPS with glucose-binding residues in the coat protein F, a reaction that appears to be dependent on the presence of Ca^{2+} ions (Fane *et al.*, 1988). This is followed by an irreversible reaction that has been observed *in vitro*, but the molecular basis of which has not yet been elucidated. However, mutations in the G and H proteins have been shown to affect host range, and both proteins have demonstrated an ability to bind with LPS. One hypothesis for having an initial reversible reaction is that the phage will be able to disassociate from LPS-containing membrane fragments shed from lysed host cells (Incardona *et al.,* 1985).

1.2.2.2 Injection

While many phages use a tail to penetrate the host wall and deliver their genomic payload (Molineux and Panja, 2013), the ΦX174 virion is tailless. After attachment of the virion to LPS, ten copies of the DNA pilot protein encoded by gene H oligomerise to form a tube which extends out of one of the pentameric spikes and through the periplasm of the cell. The DNA passes out of the capsid through this tube, and into the cytoplasm of its host (Sun *et al.*, 2014).

1.2.2.3 Replication

Replication of the phage genome occurs in three stages. In stage I, the ssDNA genome is converted into a covalently closed double-stranded DNA (dsDNA) molecule called replicative form (RF) DNA. Since ΦX174 has no antisense strand, no proteins can be synthesised until this process has happened, meaning stage is carried out solely by host proteins including the DNA III polymerase holoenzyme complex (Shlomai *et al.*, 1981).

In stage II, RF DNA is amplified by a rolling circle mechanism. In addition to host proteins, this process requires the product of ΦX174 gene A, which nicks the (+) strand of RF DNA at the origin of replication before binding to the 5' end. This serves as a primer for DNA synthesis by DNA III polymerase, using the unnicked strand as a template. As replication continues along the (-) strand, the nicked strand is displaced, before being covalently closed by the host *rep* protein. The (+) strand molecule then has a (-) strand synthesised by the same process as in stage I (Eisenberg *et al.*, 1977).

In stage III replication, ssDNA genomes are synthesised and packaged in the viral procapsids. The ΦX174 C protein binds to the A protein to inhibit further replication of RF DNA. This complex binds to the procapsid and undergoes another round of rolling circle replication. From this, the new RF

DNA is released, while the (+) strand is covalently closed but remains attached to the procapsid (Aoyama and Hayashi, 1986).

1.2.2.4 Capsid assembly

Once stage I DNA replication has occurred, synthesis of phage proteins begins. As previously mentioned, the A and C proteins have functions related to stage II and III DNA replication. A further six proteins are involved in assembly of the capsid, starting with the coat and spike proteins F and G, which appear to self-assemble into pentamers (Fane *et al.*, 1988). This is followed by binding of the internal scaffolding protein B to the coat pentamer, causing a conformational change which is thought to allow the coat pentamer to associate with a spike pentamer (Ekechukwu and Fane, 1995). In the next stage of capsid morphogenesis, the pentamers are brought together with the external scaffolding protein D. It appears that there are no chemical interactions between the capsid pentamers, and that structure is maintained at this point by the external scaffolding proteins (Dokland *et al.*, 1999). Sixty copies of the DNA binding protein J bind to a ssDNA genome and enter the procapsid, displacing protein B. Each copy of J associates with a single coat protein, tethering the genome in place within the virion. This is followed by the disassociation of the external scaffolding proteins, causing a configurational change in the capsid pentamers (Hafenstein and Fane, 2002)

1.2.2.5 Lysis

Host lysis is controlled by the product of the ΦX174 E gene, a short protein that inhibits the host enzyme translocase I, involved in peptidoglycan synthesis (Bernhardt *et al.*, 2001). With host bacteria prevented from synthesising cell walls, lysis occurs when the cell attempts to divide; a similar mechanism to the β-lactam antibiotics. Time from infection to burst is asynchronous, but lasts for an average of 21 minutes (Hutchison and Sinsheimer, 1963), which supports this passive lysis mechanism being dependent on the host cell's life cycle.

1.2.3 Other genes

Two gene products from ΦX174 have not been mentioned in this section so far, A* and K. A* is a truncated form of protein A that appears to increase the efficiency of phage DNA replication by inhibiting that of the host, but is non-essential for viral propagation (Colasanti and Denhardt, 1987). A mutant with a premature stop codon in gene K was found to still be viable but had a reduced burst size compared to wild-type, indicating that the function of this protein is related to increasing burst size (Gillam *et al.*, 1985). However, the mechanism by which these two genes work is unclear.

1.2.4 Recombination

Genetic recombination can occur in ΦX174 when two molecules of RF DNA are present within the same cell. Although this process can likely occur with extra copies of RF DNA that have arisen during replication, these are all clonal and so genotype will be left unaffected. However, during coinfection of a bacterial cell with more than one phage, both can have their genomes converted to RF DNA, which can result allelic recombination.

Recombination is dependent on the host *rec*A protein, although in its absence the ΦX174 A protein can be utilised for a less effective recombination process.

Since recombination in ΦX174 takes place during DNA replication, its frequency is directly linked to the multiplicity of infection (MOI), the ratio of phages to host cells. When host cells outnumber phages, coinfection is unlikely and recombination is rare. When host cells are outnumbered by phages the opposite is true, fewer host targets mean coinfection becomes common and the probability of recombination increases.

1.3 Experimental evolution

1.3.1 History of experimental evolution

Traditionally, evolution has not been an experimental science. Much like historians, evolutionary biologists had to look to the past at evidence left behind, be that comparative studies of extant species, the fossilised remains of extinct species, or genetic data. Darwin himself only formulated the ideas in On The Origin of Species (1859) after sailing the globe and observing how different organisms had adapted to their different environments, such as the different beak shapes found in Galapagos finches or the differences between different carrier pigeons. He would have doubtless had an easier time of it if he had been able to confirm his hypotheses in the laboratory.

One of the earliest known evolution experiments was carried out in the 19th century by (Dallinger, 1888). Over several years, he grew a culture of unicellular organisms in an incubator, gradually raising the temperature from 15°C to 70°C. Although the organisms initially could not grow at 23°C or above, over the course of the experiment they adapted to the change in temperature and were eventually able to grow at 70°C. Furthermore, when the adapted organism was cultured at 15°C, it was unable to grow.

The first evolution experiments were performed in animals, such as the fruit flies *Drosophila pseudoobscura* and *Drosophila melanogaster* (Sewall Wright, 1946) (Rose, 1984), and the silver fox *Vulpes vulpes* (Belyaev, n.d.), but these were infrequent, probably due to their practical difficulties. With advances in microbiology and genetics however, experimental evolution began to arise as a field in its own right.

Microorganisms have many distinct advantages over multicellular organisms for experimental evolution. Their short reproductive times mean that experiments can run for a much larger amount of generations, their small size allows far greater populations in far smaller spaces, and they are easy to culture and grow, with conditions easier to control. These advantages also make experimental replication possible (Kawecki *et al*., 2012).

A noted evolution experiment began in 1988 at the University of Michigan, and continues to this day. Richard Lenski evolved 12 populations of *Escherichia coli* that began from a single clone. To date, the populations have been growing for over 60,000 generations and data from this experiment has been used in numerous studies. In the meantime, hundreds more evolution experiments have been performed using bacteria, viruses and yeasts (Buckling *et al*., 2009).

Rather than looking back and piecing together the evolutionary history of life on Earth with existing data, experimental evolution allows us to study dynamics in evolving populations in real time, and gain a better understanding of the mechanisms that drive evolution.

1.3.2 Serial passaging versus continuous culture

In any growth environment, the organisms must be provided with the nutrients they need to survive and reproduce (and, in the case of viruses, there must be an adequate supply of host cells and the nutrients *they* require). In addition, population sizes will quickly increase to levels that cannot be supported by the growth environment.

When studying unicellular organisms, there are two major methods used for propagation: serial passaging and chemostats. In serial passaging, a sample of a population from a culture is transferred to fresh media and allowed to grow, before the process is repeated. This allows for the periodic replenishment of nutrients, as well as the control of population sizes. By only transferring a small sample of a population, a genetic bottleneck is introduced, adding selection to the system. Typically the environment used for passaging is a culture containing growth media and/or host cells, but can include more exotic environments; one study passaged *Candida albicans* through a series of murine gastrointestinal tracts (Pavelka, 2014).

A chemostat is an enclosed growth environment to which fresh growth medium is added at a constant rate, while media in the main chamber is removed at the same rate. This allows for continual replenishment of the nutrients required for growth, as well as removal of metabolic by-products. The removed medium also contains a portion of the microorganisms being grown. A culture in a chemostat will eventually reach an equilibrium where the population growth rate is the same as the rate at which is it diluted. To achieve this, the medium in the chemostat must have an essential nutrient present in growth limiting concentration. This steady state growth can be desirable in experimental evolution because it provides for a consistent environment compared to serial passaging where conditions change over the course of each passage (Gresham and Dunham, 2014).

Although serial passaging is commonly used, experiments carried out with this method typically run for a short number of generations. A major advantage of a chemostat is that it requires far less "hands-on" time than serial passaging, because the replacement of growth medium and population control are both controlled automatically, meaning experiments can cover many more generations with little extra effort. As a comparison, one chemostat experiment with ΦX174 covered approximately 13,000 generations (Wichman *et al.*, 2005), while the longest serial passaging experiment with the same organism to date is the one presented in chapter 4 of this thesis, which reached approximately 300 generations. A notable exception is Richard Lenski's LTEE, where *E. coli* have been serially passaged for over 67,000 generations as of March 2017 (Lenski, 2017). However, this experiment has been running for over 28 years, which is not feasible for most researchers.

In the case of experimental evolution using bacteriophage, a two-chambered chemostat is required. The first chamber behaves like a typical chemostat, in which host bacteria are grown. In the second chamber, bacteriophages are grown, with a supply of host bacteria provided by a constant flow of medium from the first chamber.

One important consideration when using chemostats to grow bacteriophages is that they do not directly require nutrients; nutrients in the medium are used by the host cells, which are then utilised by the phages. Limiting nutrients would therefore limit the growth rate of the host rather than the phage. Since phages typically multiply at a much higher rate than their hosts, their growth rate is directly linked to the number of host cells within the chemostat. This means that while it is possible to control the phage population size by limiting the number of host cells that flow into the chamber, it is not possible to manipulate the MOI in a chemostat. At high MOI, the probability of a host being coinfected by multiple phages increases, increasing the probability of recombination which may not be desirable in many evolution experiments. In these circumstances, serial passaging would be preferable because population sizes can be periodically reduced to prevent the MOI increasing over time.

1.3.3 Experimental evolution of ΦX174

Due to their small genome sizes, short generation times, and ease of culture, the Microviridae, and ΦX174 in particular, have been used frequently in experimental evolution and become a model system. The first study examined convergent evolution, discovering that in nine evolving lines, over half the substitutions observed were common to multiple lineages (Bull *et al.*, 1997). It was also shown that there was not a common evolutionary trajectory shared by replicate lineages, and parallel mutations did not always appear in the same order expected by their relative fitness effects (Wichman *et al.*, 1999).

Adaptation to different hosts has also been studied. When grown in *Salmonella enterica* ΦX174's ability to grow in *E. coli* was reduced, although when adapted to *E. coli* it was still able to grow in *Salmonella*. It was determined that this was due to mutations in the coat protein, and after switching hosts these mutations swiftly reverted to restore infectivity to the phage (Crill *et al.*, 2000). It was found that the majority of mutations evolved independently in different lineages, and many of the sites where changes were seen were at sites where ΦX174 differed from S13, a closely related phage. It was concluded from this that there were limited evolutionary pathways in the *Microviridae* (Wichman *et al.*, 2000). In another study, ΦX174 was adapted to three *E. coli* mutants, each differing only in an LPS sugar group that is part of the phage receptor. Rather than repeated mutations at the same sites, high variation was observed with no common mutations shared between phages adapted to different hosts and only one mutation arising in multiple replicate lineages (Pepin *et al.*, 2008).

Adaptation to high temperature has also been investigated, showing that in a harsh environment with low starting fitness, single mutations could be responsible for fitness increases of much larger magnitude that those usually observed (Bull et al., 2000) (Holder and Bull, 2001). Another study examined compensatory evolution, where it was found that when deleterious mutations were introduced to ΦX174, they were more frequently corrected by compensatory mutations rather than reversions (Poon, 2005).

Other studies have investigated the distribution of mutational fitness effects, where it was found that the majority of random mutations tested were deleterious but not lethal. Apart from a small fraction of beneficial and lethal mutations, the remainder were effectively neutral. Fitness effects of individual mutations when the phage was grown on *S. enterica* correlated with those from *E. coli*, suggesting that most mutations have a general effect rather than one specific to the host (Domingo-Calap *et al.*, 2009; Vale *et al.*, 2012).

One study investigated clonal interference, finding that it was far less frequent in populations evolved in sub-optimal conditions. It was thought that this was due to there being more potential adaptive mutations under normal conditions, whereas mutations specific to the harsher conditions are likely to be more limited but confer the largest fitness increases (Pepin and Wichman, 2008).

The longest study with ΦX174 was carried out in a chemostat for approximately 13,000 generations. High rates of adaptive evolution were observed throughout, indicating that instead of simply adapting to a constant environment, an arms race was occurring between competing genomes within the culture. This was likely due to high levels of coinfection, an unavoidable consequence of growth within a chemostat (Wichman *et al.*, 2005).

1.3.4 Experimental evolution and mutation rate

Only one experimental evolution study could be found that specifically looked at the effects of an elevated mutation rate in phage. A single lineage of bacteriophage T7 was evolved in the presence of mutagen to test a model of lethal mutagenesis (Springman *et al.*, 2010). However, this study only looked at fitness and lethal mutagenesis, and did not examine the genetic changes or evolutionary dynamics that occurred, meaning this area is ripe for further research.

1.4 Project aims

The mutation rate is both the fuel for evolution and an evolved characteristic. Understanding how evolution is affected by an increase in this rate will help us understand why it evolved as it did. In addition to answering questions in evolutionary biology, elevated mutation rates have been considered as potential therapeutic treatments for viral infections, and a greater understanding of the processes underpinning these is required.

The main objectives of this project were:

- To investigate methods for elevating bacteriophage ΦX174 mutation rates (covered in chapter 3).
- To investigate the consequences of an elevated mutation rate on the fitness of evolving populations of ΦX174 (covered in chapter 4).
- To use next-generation sequencing to investigate evolutionary processes in these populations (covered in chapter 5).

Chapter 2: Materials and methods.

2.1 Biological strains

2.1.1 Bacteria

E. coli C1 obtained through the Yale Coli Genetics Stock Center (strain #3121) was used as the bacterial host in chapters 4 and 5 of this work. Genome sequencing and assembly was performed by Dr James Taylor of John Hopkins University.

For the work in chapter 3, *E. coli* C122 was used as the bacterial host, as well as BAF8, an amber-permissive strain of *E. coli* isogenic to C122. These were provided by Dr Bentley Fane of The University of Arizona.

2.1.2 Bacteriophage

Wild-type ΦX174 was provided by Dr Holly Wichman of The University of Idaho. The DNA sequence was obtained with GenBank ID AF176034.1 and confirmed by Illumina sequencing (appendix B.1).

In chapter 3, ΦX174 *am(E)W4,* also provided by Dr Fane, was used.

2.1.3 Plasmids

Two plasmids were used in this work. pIF2013, derived from pBR322 and containing *dnaQ926,* ampicillin, and kanamycin resistance genes, was provided by Dr Roel Schaaper of the National Institute of Environmental Health Sciences. See (Fijalkowska and Schaaper, 1996) for further details of plasmid construction. pBR322, the vector pIF2013 was derived from, was also used, and purchased from Fisher Scientific (cat # SD0041).

Primer name	**Sequence (5’ - 3')**	**Notes**
QPX-2675-F	TTGAGTCTTCTTCGGTTCCGACTA	qPCR quantification against plaque assay (taken from Vale *et al.,* 2012)
QPX-2776-R	TCACACAGTCCTTGACGGTATAAT	
QPX-590-F	ATACCCTCGCTTTCCTGCT	qPCR quantification against DNA standards (designed by PrimerDesign)
QPX-690-R	CGCCTTCCATGATGAGACA	
PHX-0001-F	GAGTTTTATCGCTTCCATG	Amplifies bases 1-2953 (amplicon 1) for Sanger sequencing (provided by Dr Holly Wichman)
PHX-2953-R	CCGCCAGCAATAGCACC	
PHX-2605-F	CAGGTTGTTTCTGTTGGTGCTG	Amplifies bases 2605-379 (amplicon 2) for Sanger sequencing (provided by Dr Holly Wichman)
PHX-0379-R	CTTGACTCATGATTTCTTACC	
PHX-0895-F	GCCGTTGCGAGGTACTAAAG	Amplifies bases 895-1500 (amplicon 3) for Sanger sequencing (provided by Dr Holly Wichman)
PHX-1500-R	TTGAGATGGCAGCAACGG	

Table 2.1 – primer sequences. This table contains the sequences of all primers used for Sanger sequencing and qPCR quantification.

2.2 Media, buffers and solutions

2.2.1 Salt solutions

CaCl$_2$ (Sigma #449709) and MgCl$_2$ (Sigma, #M8266) were dissolved in autoclaved distilled water to a concentration of 0.1M. These solutions were used to supplement media as required. During the course of this work, small amounts of precipitate were found in the stock CaCl$_2$ solution. From approximately passage 60 of the evolution experiment in chapter 4, stock solutions were instead prepared with anhydrous CaCl$_2$ (Sigma, #499609).

2.2.2 Lysogeny broth

Lysogeny broth (LB) was the media used for all liquid cultures in this work. It was prepared by adding dehydrated LB Miller Broth (Appleton Woods) to distilled water at a concentration of 25g/L (working concentrations of 10g/L tryptone, 5g/L yeast extract and 10g/L NaCl). The media was supplemented with CaCl$_2$ to a concentration of 2mM and MgCl$_2$ to a concentration of 5mM and autoclaved for 15 minutes at 121°C.

2.2.3 LB agar plates

LB agar was prepared by adding dehydrated LB to distilled water at a concentration of 25g/L. Bacto Agar powder (Appleton Woods) was added at a concentration of 15g/L, and media was autoclaved for 15 minutes at 121°C. Media was cooled in a water bath at 45°C until it was comfortable to handle. If plates were intended for use with plasmid-containing bacteria, ampicillin was added at a concentration of 100ng/µl. Agar was poured into petri dishes so that the base surface was evenly coated (approximately 25ml per dish) and left to solidify.

2.2.4 Soft LB agar

Soft LB agar was prepared by adding dehydrated LB at a concentration of 15g/L and agar at a concentration of 7g/L. The media was supplemented with $CaCl_2$ to a concentration of 2mM and $MgCl_2$ to a concentration of 5mM, and autoclaved at 121ºC for 15 minutes. Soft agar was stored at 55ºC until needed.

2.3 Bacterial overnight cultures

Overnight bacterial cultures were frequently used in this work. Unless otherwise specified, these were prepared by using a sterile plastic loop to inoculate bacteria from frozen glycerol stocks into a 50ml centrifuge tube containing 10ml of LB (supplemented with 100ng/µl ampicillin if bacteria contained a plasmid). These were incubated overnight at 37ºC, shaking at 200rpm. Overnight cultures were always prepared fresh the day before they were needed, and disposed of at the end of the day.

2.4 Bacteriophage quantification

2.4.1 Double agar overlay plaque assay

The number of bacteriophages in a sample can be quantified by growing them on a lawn of their bacterial host, and counting the number of plaques that appear.

Prior to carrying out phage overlay assays, LB agar plates were pre-warmed in an incubator at 37ºC, and molten soft agar was cooled to 50ºC in a water bath.

To ensure plates contained a number of plaques within an acceptable range (30-300), phage samples were serially diluted 10-fold in LB. In triplicate, sterile plastic bijous were prepared containing 100µl of bacterial overnight culture and 100µl of the phage dilution to be measured. To each bijou 4ml soft agar was added, and the contents mixed by replacing the lid and inverting six times. The contents were immediately poured onto an LB agar plate, and gently swirled to ensure the entire surface was coated. Plates were left at room temperature for 15 minutes to set, before being transferred to a 37ºC incubator.

Plates were removed from the incubator once plaques were large enough to be counted (usually between 4-7 hours, but occasionally overnight incubation was required). Plaques were counted manually and used to calculate the number of plaque forming units (pfu) in the original phage sample.

2.4.2 Quantitative PCR

qPCR was used as a quicker method for determining bacteriophage concentrations. By first measuring a series of samples with known concentrations, a standard curve was generated that could be cross-referenced with quantification cycle (Cq) values to determine sample concentrations. Two distinct standard curves were created, each of which required different reaction conditions.

2.4.2.1 Standard curve calibrated against plaque counts

Reactions were set up in duplicate in Lightcycler 480 96 well plates (Roche). Each qPCR reaction contained 10µl PrecisionPLUS SYBRgreen Master Mix (Primerdesign), 5µl phage sample, 0.6µl of each primer at 10mM (final concentration of 0.3µM) and 3.8µl of sterile water. Primers used were QPX-2675-F and QPX-2776-R (table 2.1). All qPCR was carried out using a Lightcycler 480 (Roche). The PCR program was:

- 95°C for 10m,
- 35 cycles of:
  - 15s at 95°C,
  - 20s at 60°C,
  - 30s at 72°C (fluorogenic data collected during this step).

To create the standard curve, a 10-fold dilution series of ΦX174 was quantified with a plaque overlay assay, and measured with qPCR. Cq values were related to pfu concentrations determined by plaque assay, and one dilution was stored at -20°C for use as a standard.

As well as phage samples, each plate contained a positive control reaction that used the standard as a template, and a negative control reaction that replaced the phage sample with sterile water. Lightcycler 480 software was used to perform an absolute quantification analysis (second derivative max method) with the resulting Cq values and the standard curve to determine phage concentrations in pfu/µl.

This method was used for all qPCR phage quantification carried out in chapters 3 and 4, other than fitness assays.

2.4.2.2 Standard curve calibrated against phage genome units

Because plaque formation is a mutable trait and could change during evolution, it was necessary to have a method of determining phage titers that did not rely on data obtained from plaque assays of ancestral phage. The method described in this section was used for the fitness assays carried out in chapter 4. A custom assay was designed by Primerdesign Ltd that included primers that amplified a region in the E gene of ΦX174 (table 2.1), and a positive control template supplied at a copy number of 2 x 10^5 per µl. As per the kit's manual, primer mix was suspended, and the positive control serially diluted and used to create a standard curve.

Reactions were set up in duplicate in Lightcycler 480 96 well plates (Roche). Each qPCR reaction contained 10µl PrecisionPLUS SYBRgreen Master Mix (Primerdesign), 5µl phage sample, 1µl primer mix (final concentration of 0.3µM for each primer) and 4µl of sterile water. Plates also contained positive and negative control reactions, replacing the phage sample with a standard or sterile water respectively. qPCR was carried out in a Lightcycler 480 (Roche).

The PCR program was:

- 95°C for 2m
- 40 cycles of:
  - 10s at 95°C

- 60s at 60°C (fluorogenic data collected during this step)

Lightcycler 480 software was used to perform an absolute quantification analysis (second derivative max method) with the resulting Cq values and the standard curve to determine phage concentrations. PCR is an exponential reaction, and under 100% efficiency the number of DNA molecules will double each cycle. However, because the genome of ΦX174 is single stranded DNA, the first cycle of the reaction will result in synthesis of the complementary strand rather than amplification. Although in subsequent cycles the product is amplified as normal, the double stranded DNA standards are subject to an extra cycle of amplification. To take this into account when determining phage titers, sample concentrations determined by the Lightcycler 480 software were multiplied by the efficiency of the reaction to give the true concentration. Titers were measured in phage genome units (pgu) rather than pfu.

2.5 Gel electrophoresis

Agarose gels were made by mixing 1-2% w/v agarose powder (Appleton Woods) in Tris-Acetate-EDTA buffer (TAE) (40mM Tris Acetate, 2mM Na_2EDTA (National Diagnostics)), depending on expected nucleic acid size. Agarose was melted using a microwave oven, and a 1:10,000 volume of SYBR Safe DNA Gel Stain (Invitrogen) added. Gels were poured into a casting tray and allowed to solidify before being placed in a gel tank (Geneflow) and submersed in TAE. DNA samples were mixed with a 1:5 volume of Blue/Orange 6X Loading Dye (Promega) and loaded into wells alongside 100bp or 1kb DNA Ladders (Promega) depending on expected fragment sizes. A

100V electrical current was passed through the gel until DNA fragments were completely separated, typically 45 minutes. Gels were visualised under UV light using an InGenius LHR gel documentation system (Syngene).

2.6 Mutation rate measurement

2.6.1 Preparation of competent BAF8 cells

Chemically competent preparations of *E. coli* BAF8 were created by treatment with calcium chloride (chapter 2.1.1). LB agar plates were streaked with the relevant bacteria from frozen glycerol stocks and incubated at 37°C until colonies were visible. A single colony was inoculated into 10ml LB and grown overnight at 37°C in a shaking incubator. 1ml of this overnight culture was added to 49ml LB and grown at 37°C in a shaking incubator until OD600 was between 0.3-0.4. This culture was divided into 25ml aliquots in chilled 50ml centrifuge tubes, and placed on ice for 10 minutes. Tubes were centrifuged for 10 minutes at 2700g in a Heraeus Megafuge 16R (Thermo Fisher) pre-chilled to 4°C. Supernatants were discarded, pellets resuspended in 5ml ice cold 0.1M $CaCl_2$, and tubes placed on ice for 30 minutes. Tubes were centrifuged as before and supernatants discarded. Pellets were resuspended in 500µl ice cold 0.1M $CaCl_2$ (supplemented with 7% DMSO) and stored at -80°C for later use.

2.6.2 Transformation of BAF8 with pIF2013

A sterile loop was used to inoculate frozen stock of *E. coli* containing the pIF2013 plasmid into a 50ml centrifuge tube containing 10ml LB (with 100ng/µl ampicillin). This was incubated overnight at 37°C, shaking at 200rpm. The plasmid was extracted using a Machary-Nagal miniprep kit following the manufacturer's instructions, and run on a 1% agarose gel to confirm it was present and of expected size.

The plasmid was transformed into BAF8 using the heat shock method. An aliquot of competent BAF8 cells was removed from the freezer and placed in an ice bucket to defrost. 100µl of competent cells were added to thin-walled PCR tube, along with 5µl of miniprep product. Tubes were placed on ice for 30 minutes, transferred to a water bath at 42°C for 45 seconds, and returned to ice for a further 2 minutes. The contents of each tube were added to a 2ml microcentrifuge tube containing 500µl SOC medium and incubated at 37°C for 45 minutes. To increase the chances of getting single colonies, four LB agar plates (with 100ng/µl ampicillin) were prepared. 500µl of the transformation reaction was spread on one, 50µl on another, and sterile loops used to streak the final two plates. Plates were incubated at 37°C until bacterial growth was visible. Single colonies were suspended in 500µl LB (with 100ng/µl ampicillin and 7% DMSO) and grown for two hours at 37°C shaking at 650rpm. These stocks, labelled BAF8 pIF2013, were stored at -80°C for future use.

2.6.3 Preparation of phage for fluctuation test

A sterile loop was used to inoculate a suspension of ΦX174 *am(E)W4* into a sterile plastic bijou containing 4ml soft agar that had cooled to approximately 45°C and 100µl of *E. coli* BAF8 overnight culture. The cap was replaced and the bijou inverted 6 times, before being decanted onto an LB agar plate. The plate was left at room temperature for 15 minutes to solidify, then incubated at 37°C until plaques were visible, approximately 4 hours. A single plaque was removed from the soft agar using a sterile 1ml pipette tip and suspended in 100µl LB in a microcentrifuge tube. This was placed in a refrigerator overnight to allow phages to diffuse out of the agar. Under a fume hood, 2 drops of chloroform were added and the tube was vortexed for 5 seconds to lyse any bacteria that had been transferred with the plaque. The tube was spun for five minutes at 13,000G in a microcentrifuge to separate cellular debris, agar and chloroform. The majority of the aqueous layer was carefully removed with a pipette and transferred to a fresh microcentrifuge tube. This phage preparation was used to initiate every culture in the subsequent fluctuation test.

2.6.4 Fluctuation test

Fluctuation tests were performed in 3 batches of 24 cultures on a Thermomixer comfort. For measuring the wild-type mutation rate, 50µl of BAF8 overnight culture was added to 500µl LB in a 2ml microcentrifuge tube. For measuring the mutation rate in the presence of *dnaQ926*, the overnight culture used was BAF8 pIF2013 and LB contained 100ng/µl ampicillin. Cultures were grown for 1 hour 30 minutes at 37°C, shaking at 650rpm, at which time approximately 400 pfu ΦX174 *am(E)W4* was added to each. After the addition of phage, cultures were grown for a further 90 minutes (wild-type) or 60 minutes (mutator). Tubes were centrifuged at 13,000G for 1 minute and the supernatant was transferred to a fresh tube. Supernatants and phage stocks used to initiate cultures were quantified with qPCR.

For each culture, supernatant containing approximately 3 x 10^3 pfu (mutator) or 2.5 x 10^4 pfu (wild-type) was added to a sterile bijou containing 4ml of molten soft agar that had cooled to approximately 45ºC and 100µl of *E. coli* C122 overnight culture. The cap was replaced and the bijou inverted 6 times, before being decanted onto an LB agar plate and lightly swirled so the entire surface was covered. Plates were left at room temperature for 15 minutes to solidify, then incubated at 37ºC overnight. After incubation, plates were checked and any plaque formation was recorded.

2.6.5 Calculating the mutation rate

The proportion of cultures where *no* mutations were observed in the amber stop codon, P_0, was determined from the proportion of plates that did not display any plaques. qPCR was used to determine N_i and N_f .

The equation:

$$m = \frac{-\ln P_0}{(N_f - N_i)}$$

was used to calculate the rate of mutation to an observable phenotype, where N_i is the number of phage plaque forming units (pfu) used to initiate the culture and N_f is the number of phage pfu in the culture after growth.

To calculate the mutation rate per nucleotide per generation μ, the equation:

$$\mu = \frac{3m}{T}$$

is used, where *T* is the number of possible mutations that result in viable plaque formation. For this calculation, it was assumed that T=8 (the number of substitutions that change the amber stop codon to a non-stop codon).

2.7 Evolution experiment

2.7.1 Experimental lines

Four lines of ΦX174 were maintained in this experiment, all originating from a single plaque. Lines A1 and A2 were grown in non-mutagenic conditions while B1 and B2 were grown under mutagenic conditions. *E. coli* C1 was transformed with pIF2013 for use as the host in lines B1 and B2, following the protocol described in 3.2.2. Because a plasmid requires antibiotic to be added to its media and uses the same DNA polymerase to replicate as the phage, its presence potentially alters selective pressures on the phage. To try and ensure conditions were similar in the non-mutagenic environment, *E. coli* C1 was transformed with pBR322, the vector that pIF2013 was derived from. This was used as the host for lines A1 and A2.

2.7.2 Growth tubes

To ensure that conditions were maintained throughout the evolution experiment, enough tubes containing hosts and growth media were prepared in advance. This method meant that each tube contained media from the same batch and host cells from the same culture, which minimised variation that could have been introduced by growing new cultures of host cells each day and preparing fresh batches of media throughout the experiment. Approximately 250 tubes were prepared for each of the hosts used during this experiment

A sterile plastic loop was used to inoculate 15ml LB (containing ampicillin to 100mg/µl) with C1 containing either pIF2013 or pBR322 plasmid from frozen stocks. This was incubated overnight at 37ºC, shaking at 200rpm. 4ml of an overnight culture and 3ml DMSO were added to 50ml falcon tubes containing 40ml LB (containing ampicillin to 100mg/µl). Tubes were briefly vortexed to mix and 550µl volumes were aliquoted into sterile 2ml microcentrifuge tubes. These "growth tubes" were stored at -20ºC for future use.

To determine host cell density, three growth tubes were placed on a Thermomixer for 90 minutes at 37ºC, shaking at 650rpm. Cells were then serially diluted tenfold in LB and 100µl of each dilution was spread in duplicate on LB agar plates. After overnight incubation at 37ºC, plates were checked for colonies. Dilutions that produced plates with approximately 30-300 colonies were selected and used to calculate cell density in colony forming units (cfu)/µl.

2.7.3 Phage preparation

1µl of ΦX174 from a glycerol stock was added to a microcentrifuge tube containing 1ml LB. From this, a tenfold dilution series of the supernatant was prepared in LB. 100µl of each dilution was added to a sterile plastic bijou containing 4ml soft agar that had cooled to approximately 45ºC and 100µl of *E. coli* C1 overnight culture. The cap was replaced and the bijou inverted 6 times, before being decanted onto an LB agar plate. The plates were left at room temperature for 15 minutes to solidify, then incubated at 37ºC until plaques were visible, approximately 4 hours.

A single plaque was removed from the soft agar using a sterile 1ml pipette tip and suspended in 100µl LB in a microcentrifuge tube. Under a fume hood, 2 drops of chloroform were added and the tube was vortexed for 5 seconds to lyse any bacteria that had been transferred with the plaque. The tube was spun for five minutes at 13,000G in a microcentrifuge to separate cellular debris, agar and chloroform. The majority of the supernatant was removed with a pipette, taking care not to take any of the non-aqueous layer, and transferred to a fresh microcentrifuge tube. Sanger sequencing was used to confirm the genotype of the plaque.

A growth tube containing C1 with plasmid pBR322 was removed from -20ºC storage, thawed at room temperature, and placed on a Thermomixer at 37ºC, shaking at 650rpm. After 90 minutes, 20µl of supernatant was added, and the tube returned to the Thermomixer for a further 2 hours. The growth tube was centrifuged at 13,000G for 30 seconds in a microcentrifuge to pellet the bacteria, and the supernatant was transferred to a fresh microcentrifuge tube and quantified using qPCR. This preparation was used to seed the first passage of all experimental lines.

2.7.4 Serial passaging

The evolution experiment was carried out by serial passaging with each line grown for 100 hours in one hour-long passages. The four lines were grown in parallel, and followed identical protocols with the exception that A1 and A2 used host tubes containing C1 with pBR322 and B1 and B2 used host tubes containing C1 with pIF2013.

Two of each growth tube were removed from -20°C storage, thawed at room temperature, and placed on a Thermomixer at 37°C, shaking at 650rpm. After 90 minutes, approximately 10^6 phage from the preparation in the last section were added to each tube, which were then returned to the Thermomixer for an hour. Tubes were then immediately centrifuged at 14,000G in a microcentrifuge to remove bacteria, and the supernatants were transferred to new microcentrifuge tubes. qPCR was used to quantify supernatants, which were then diluted to between 10^5 and 10^6 pfu/µl.

Each passage subsequent to the first followed the same protocol, but were started with between 10^6 and 10^7 phage from the previous passage instead of the initial preparation. The remaining phage were stored at -80°C for future use.

2.8 Fitness assays

The phage sample to be assayed was removed from storage at -80°C and allowed to defrost on ice. Growth tubes were incubated on a Thermomixer at 37°C, shaking at 650rpm for 90 minutes. Approximately $10^{\mathbf{5}}$ pgu phage was added to each growth tube, before being briefly vortexed and returned to the Thermomixer for a further 45 minutes. Tubes were then immediately spun for 1 minute at 14,000G on a microcentrifuge, and the supernatant was transferred to a fresh microcentrifuge tube. Supernatants and initial phage samples were quantified with qPCR to determine the initial and final titers of phage. Assays were carried out in triplicate.

Fitness was calculated in population doublings per hour by the equation

fitness = $\log_2(N_f/N_0) / t$

where t is the time in hours, N_0 is the initial titer of phage (in pfu) and N_f is the final number of phage (in pfu).

2.9 DNA sequencing

2.9.1 Sanger sequencing

Three primer pairs were used for amplifying ΦX174 DNA for sequencing (table 2.1). Amplicons 1 and 2 consisted of bases 1-2937 and 2605-379 respectively, giving complete coverage of the phage genome. The smaller amplicon 3 covered bases 895-1500, a region found to be of specific interest.

PCR reactions contained 0.5µl (1 unit) of Phusion High-Fidelity DNA Polymerase (Thermo Fisher), 10µl 5X Phusion HF buffer, 2.5µl of each primer at 10µM (0.5µM final concentration), 1µl dNTPs (each dNTP at 10mM, final concentration of 200µM) and 2µl of phage sample, made up to 50µl with sterile water.

For amplicons 1 and 2 the PCR program was:

- 98°C for 3m;
- 35 cycles of:
  - 10s at 98°C,
  - 30s at 60°C,
  - 90s at 72°C;
- 10m at 72°C,

For amplicon 3, the PCR program was:

- 98°C for 3m;
- 35 cycles of:
    - 10s at 98°C,
    - 30s at 63.9°C,
    - 15s at 72°C;
- 10m at 72°C,

PCR products were visualised alongside 100bp or 1kb DNA Ladders (Promega) on 1-2% agarose gels stained with SybrSafe to confirm that a single band of the expected size was present. PCR products were purified using the NucleoSpin Gel and PCR Clean-up kit (Machary-Nagal) following the manufacturer's protocol. DNA concentration was determined with the Qubit Fluorometer using the Qubit dsDNA BR Assay Kit (Life Technologies), and diluted to 1ng/µl per 100bp. Appropriate sequencing primers were selected from the table and diluted to 3.2pmol/µl with nuclease-free water. DNA samples and primers were sent to Source Biosciences who carried out Sanger sequencing. Electropherograms were assessed using 4Peaks (Nucleobytes) and sequence data was aligned against the reference genome using MAFFT (Katoh and Standley, 2013).

2.9.2 Illumina sequencing

2.9.2.1 dsDNA extraction

For preparation of libraries for Illumina sequencing, dsDNA samples are required instead of the ssDNA genomes of ΦX174. Using a method adapted from (Godson and Vapnek, 1973), 50µl of an overnight culture of C1 (containing no plasmid) was added to 500µl of LB in a microcentrifuge tube and grown on a Thermomixer Comfort (Eppendorf) for 2 hours at 37ºC, shaking at 650pm. Approximately 400µl of the phage sample to be sequenced was then added and the tube was returned to the Thermomixer for 30 minutes. At this point chloramphenicol was added to a concentration of 30ng/µl to inhibit protein synthesis and allow the accumulation of double stranded RF DNA within the host cells. After a further 4 hours incubation on the Thermomixer, tubes were removed. Tubes were spun in a microcentrifuge at 14,000 g for 1 minute and the supernatants were discarded. dsDNA was extracted from each pellet using a Qiagen miniprep kit, following the kit instructions. DNA was quantified using a Qubit (broad range dsDNA reagents). To confirm that the product was the expected size (of approximately 5386bp), samples were linearised with StuI (Promega) and run on a 1% agarose gel.

2.9.2.2 DNA library preparation

DNA samples were fragmented using NEBNext® dsDNA Fragmentase® (New England Biolabs). For each sample to be sequenced, reactions were set up containing 150ng of phage dsDNA, 2µl of 10X Fragmentase Reaction Buffer, 2µl of dsDNA Fragmentase and nuclease-free water added to a final volume of 20µl. Tubes were vortexed for 5 seconds then incubated at 37ºC for 20 minutes. Samples were cleaned up using a QIAquick PCR purification kit (Qiagen) following the kit instructions to remove salts, enzyme and small DNA fragments. All samples were quantified using a Qubit (Life Technologies) using the dsDNA Broad Range reagents, while 6 samples chosen at random were run on a Tapestation (Agilent Genomics) to determine the average fragment size.

Libraries were prepared using a NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® (New England Bioloabs), following the kit instructions and using the fragmented DNA as input. Libraries were quantified by Qubit (dsDNA Broad Range), diluted to identical concentrations, and pooled. To confirm the ancestral sequence, library preparation used a Nextera XT DNA Library Prep kit (Illumina).

Sequencing was carried out onsite on an Illumina MiSeq using a V2 cartridge (2 x 250bp). FASTQ files were generated by the MiSeq and automatically uploaded to Basespace.

2.10 Bioinformatic methods

2.10.1 Genome sequencing

dsDNA was extracted from frozen passage samples using the method described in chapter 2. Phage DNA from passages 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 was extracted for all lines. In addition, DNA was extracted from passages 5, 15, 25, 31, 32, 33, 34, 35, 45, 55, 65, 75, 85, and 95 in lines B1 and B2. dsDNA samples were fragmented and sequenced using the methods described in 2.9.2.

2.10.2 Preparation of FASTQ files

Raw sequence data were downloaded from Illumina's Basespace server as FASTQ files. FASTQ files are text-based and contain the determined nucleotide sequences of each read alongside corresponding quality scores (in

ASCII encoding). Since the sequencing run was paired end, two FASTQ files were produced for each sample, one for each strand.

The adaptor sequences are usually trimmed from FASTQ files provided by Illumina by default. However, the third-party library preparation kit contained unique adaptor sequences and this had to be carried out manually. In addition, because reads were short, primer sequences were sequenced at the 3' end. These were removed using Cutadapt (Martin, 2011), a command line tool that takes a list of sequences as input and trims them from reads in paired FASTQ files. Cutadapt was set to check each read twice (with argument `-n`) in case adaptors had been appended more than once, and to discard any reads that were shorter than 25 nucleotides after trimming (argument `-m`). Argument `-b` trims from both ends and was used for adaptors while `-a` trims from only the 3' end and was used for primers. The full list of adaptor and primer sequences used for trimming is in provided in table 5.1.

```
cutadapt fwd_reads.fastq rev_reads.fastq -m 25 -n 2 \
-o fwd_trimmed.fastq -p rev_trimmed.2.fastq \
-b ADAPTOR1 -b ADAPTOR2
-a PRIMER1 -a PRIMER2 -a PRIMER3
```

After trimming, all FASTQ files were analysed with FastQC (Andrews, 2010). This was used to check that no overrepresented sequences were present in the data, confirming that all adaptors had been successfully removed. FastQC also displays the distribution of base quality score at each position in the reads, allowing the overall quality of each file to be quickly determined.

FastQC also shows the distribution of read lengths in a FASTQC file. For all samples, the majority of reads were shorter than the maximum read length (250bp) (figure 5.3), indicating that the input DNA was more fragmented than intended. These small sizes resulted in multiple overlaps in the paired reads, meaning that parts of some reads would be sequenced from both directions and be overrepresented in the data.

To account for this, pairs of FASTQ files were merged using Paired End ReAd mergeR ((Zhang et al., 2014). PEAR compares each pair of reads in paired FASTQ files, and if there is sufficient overlap (10bp) combines them into a single sequence. All successfully merged pairs are written to a new FASTQ file, while reads that could not be merged are written to new paired FASTQ files. Merged reads shorter than 25 bases were discarded

```
PEAR -f fwd_trimmed.fastq -r rev_trimmed.fastq -n 25 -o merged
```

The FASTQ file pairs containing the unmerged reads were trimmed for quality at the 3' end using Sickle (Joshi & Fass, 2011). Sickle moves along reads from 5' to 3' with a sliding window. If the average quality score of the window falls beneath a certain value (set to the default of 20), the remainder of the read is trimmed. These trimmed reads are written to new FASTQ files. If a read passes the filter in one direction but not the other, reads are written to a singles file. The contents of these singles files were not used in downstream analysis. The `-n` argument was used to truncate reads if and when an N was encountered.

```
sickle pe -f merged_fwd.fastq –r merged_rev.fastq -t sanger -o
sickle_fwd.fastq -p sickle_rev.fastq -s sickle_singles_file.fastq –n
```

2.10.3 Read mapping

In order to align the reads to a reference sequence, Bowtie2 (Langmead and Salzberg, 2012) was used. The reference genome for the ancestral phage was retrieved from Genbank (AF176034.1) in FASTA format, and indexed.

```
bowtie2-build ref_genome.FASTA phix_ref
```

Following this, each FASTQ file containing a set of merged reads, as well as paired FASTQ files containing unmerged reads, were aligned against the indexed genome to produce a SAM (sequence alignment/map) file. This file contains all the data within the original FASTQ file in addition to mapping data.

```
bowtie2 -x phix_ref –U merged_reads.fastq -S mapped_singles.SAM
bowtie2 -x phix_ref -1 sickle_fwd.fastq -2 sickled_rev.fastq \
-S mapped_pairs.SAM
```

SAMtools (Li et al., 2009) was used to merge the SAM files generated from merged and paired reads, and convert the SAM files to the compressed binary BAM format, which takes up less space and can be processed more quickly. SAMtools was also used to sort and index the BAM files.

```
samtools merge mapped.SAM mapped_singles.SAM mapped_pairs.SAM
samtools view -bS mapped.SAM > mapped.BAM
samtools view -bS mapped.BAM | samtools sort - sorted.BAM
samtools index sorted.BAM
```

The LeftAlignIndels function of the Genome Analysis ToolKit (McKenna et al., 2010) was carried out on each BAM file. When indels appear in a sequence, they can often be aligned in multiple configurations (figure 5.1). It is important to align all indels to the leftmost position possible to standardise downstream processing and ensure indels are not mistaken for substitutions.

```
java -jar GenomeAnalysisTK.jar -R ref_genome.fasta \
-T LeftAlignIndels -I Sorted.BAM -o left.BAM
```

Although ΦX174 has a circular genome, the mapping algorithms treat the reference genome as linear, meaning there will be a break in the sequence (between positions 5386 and 1). If a read spans this break, it will be unable to map to the reference genome accurately, and an indel may be called erroneously. Since read lengths are short, the region of the chromosome that is affected by this is small. Assuming a maximum read length of 250bp, there are 498 positions that could be affected if they span this region. In bacterial chromosomes where genomes are typically several million bp in length (Wang et al., 2013) this is a relatively minor problem because such a small fraction of the total genome would be affected. However, the genome size of ΦX174 is only 5386bp, meaning over 9% of the total genome would be covered by this region.

To account for this, a FASTA file was created that spanned this break, running from positions 5041 - 5386 and 1-350 of the original reference genome. This was indexed in Bowtie2 as before, and FASTQ files were mapped against it. SAMtools was again used to convert to BAM, index, sort and left align indels.

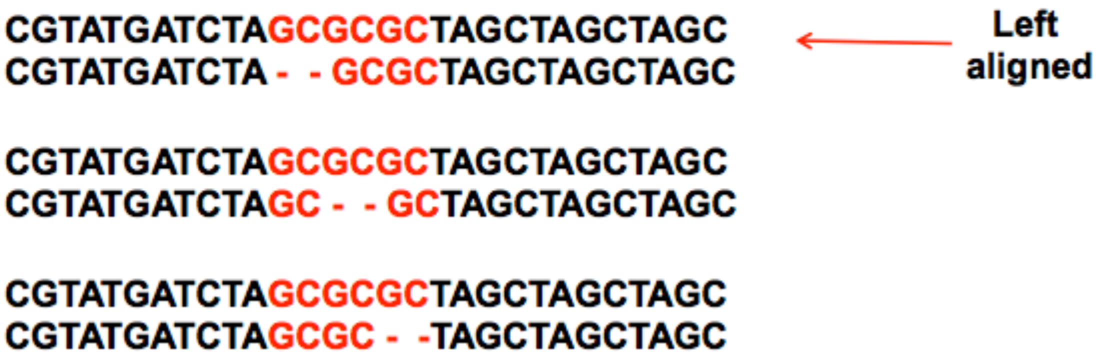

CGTATGATCTAGCGCGCTAGCTAGCTAGC
CGTATGATCTA - - GCGCTAGCTAGCTAGC
Left aligned
CGTATGATCTAGCGCGCTAGCTAGCTAGC
CGTATGATCTAGC - - GCTAGCTAGCTAGC
CGTATGATCTAGCGCGCTAGCTAGCTAGC
CGTATGATCTAGCGC - -TAGCTAGCTAGC

Figure 2.1 - Three different ways the same indel can be aligned against a reference sequence.

2.10.4 Variant calling

Quality scores for each base are given in Phred format, which is logarithmically related to the probability of the base call being erroneous. Phred score Q is given by the equation:

$$Q = -10 \log_{10} P$$

where P is the probably of an incorrect base call. A Phred score of 10 corresponds to a 90% base call accuracy, while 20 is 99%, 30 is 99.9%, and so on. Although $Q = 30$ usually gives sufficient confidence, when sequencing at high coverage a number of miscalled bases are inevitable. For example, if read depth at a particular nucleotide is 1000, and Q is uniformly 30 at that position, then it is probable that one read will be miscalled.

Since the majority of DNA fragment sizes in this experiment were smaller than the maximum read length, most pairs of reads fully or partially overlapped, and were merged with PEAR. Since each read's quality scores are calculated independently of each other, if a base is identical on both reads an updated quality score can be calculated by multiplying the individual quality scores together (figure 2.10). For example, if a base has $Q = 30$, it has a 0.1% chance of being an error. But if the same base is present on the other read with the same Q score, the probability of it being miscalled twice is 0.0001%, i.e. $Q = 60$.

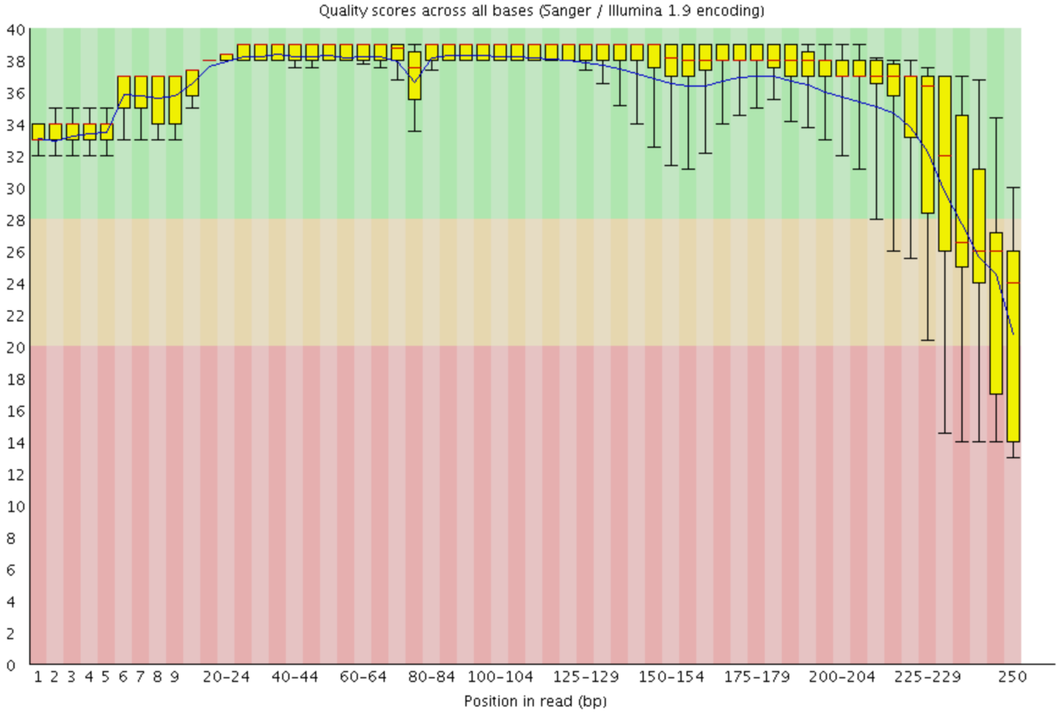
To determine mutations present in a sample, only positions with a Q of 40 or higher were considered. This allows even low frequency mutations to be identified with high certainty at the sacrifice of coverage; non-overlapping portions of reads would likely be removed by the filter. It should be noted that while a *Q* of 40 means that each base has a 0.01% chance of being miscalled, 40 is actually the maximum Phred score that the FASTQ encoding supports. In actuality, most bases that meet this quality requirement would have a much higher true score due to being a product of two quality scores.

Freebayes (Garrison and Marth, 2012) was used to call variants. To call higher frequency mutations, minimum *Q* was set to 40 (`-q 40`), and only mutations with ≥ 10% frequency (≥ 1% for indels) were returned in the output (`-F 0.1`). For each quality setting, two sets of VCF files were generated per sample, returning either SNPs or indels (using the argument `-i` to ignore indels or `-I` to ignore SNPs). The following arguments were used: `-X` (ignore multi-nucleotide polymorphisms), `-u` (ignore complex events), `-K` (output all alleles which pass input filters), `-J` (assume that samples result from pooled sequencing), and `-p 1` (no ploidy).

```
freebayes -f ref_genome.fasta  -q 30 -m 20 -F 0.1 \
-X -i -u -K -J -p 1 left.BAM > snps.vcf
```

VCF files were generated in this way for both the BAM file mapped against the reference genome, as well as the BAM files mapped against the region spanning the break of the circular phage genome. The Python script OriginPositionFixer.py (appendix X) was used to renumber the latter with genome positions that corresponded to the original reference genome. This was followed by the Python script OriginMerger.py (appendix X) that parsed the main VCF file. At each position, it checked to see if that position was present in the second VCF file, compared the coverage, selected the line with the highest depth, and wrote the line to a new file. This output of this was a merged VCF file with high coverage at the beginning and end of the genome.

Finally, the Python script VCFsimplifier.py (appendix X) was used to parse each VCF file and return a list of alleles and their frequencies in a more readable format.



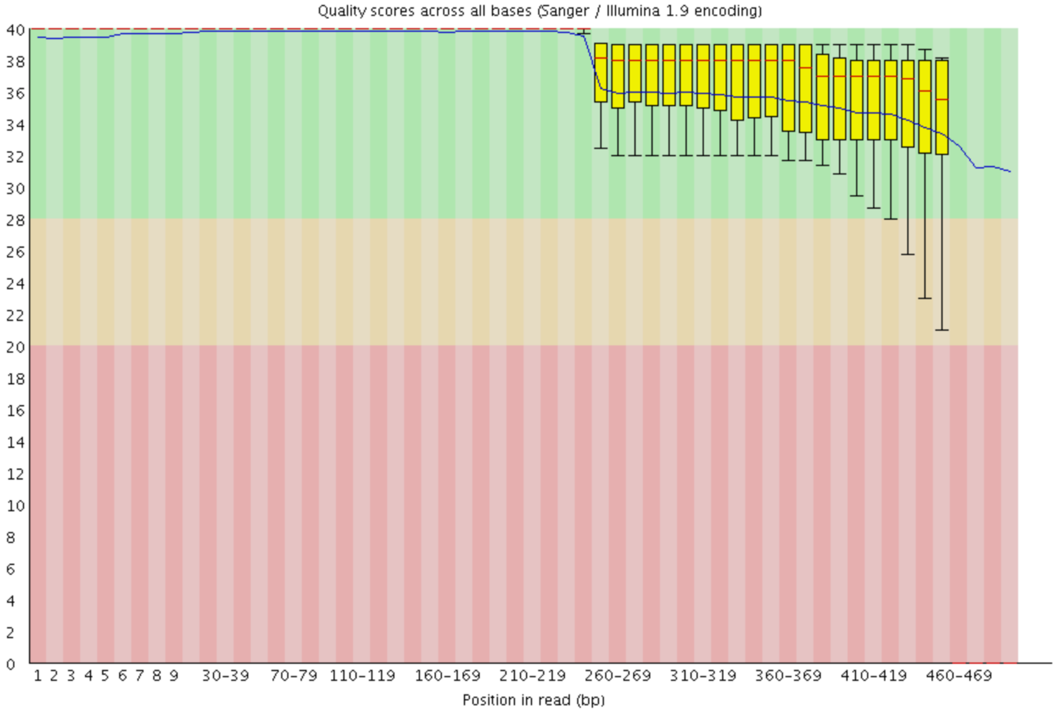


Figure 2.2 - FastQC output before (top) and after (bottom) reads were merged with PEAR. Overlapping regions from paired end reads were sequenced twice, meaning quality scores were multiplied together if the sequenced base at each position agreed. 40 is the maximum Phred value that FASTQ encoding supports, but most true quality scores will have been much higher. These data were from line A1, passage 10; but are representative of all samples.

Table 2.2 – DNA sequences for indices, primers and adaptors from the NEBNext DNA Library Prep Kit

ID	Sequence
Index 1	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
Index 2	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
Index 3	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
Index 4	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
Index 5	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
Index 6	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
Index 7	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
Index 8	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
Index 9	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
Index 10	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
Index 11	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
Index 12	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC*T
Universal primer	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC
Adaptor 1	GATCGGAAGAGCACACGTCTGAACTCCAGTC
Adaptor 2	ACACTCTTTCCCTACACGACGCTCTTCCGATC

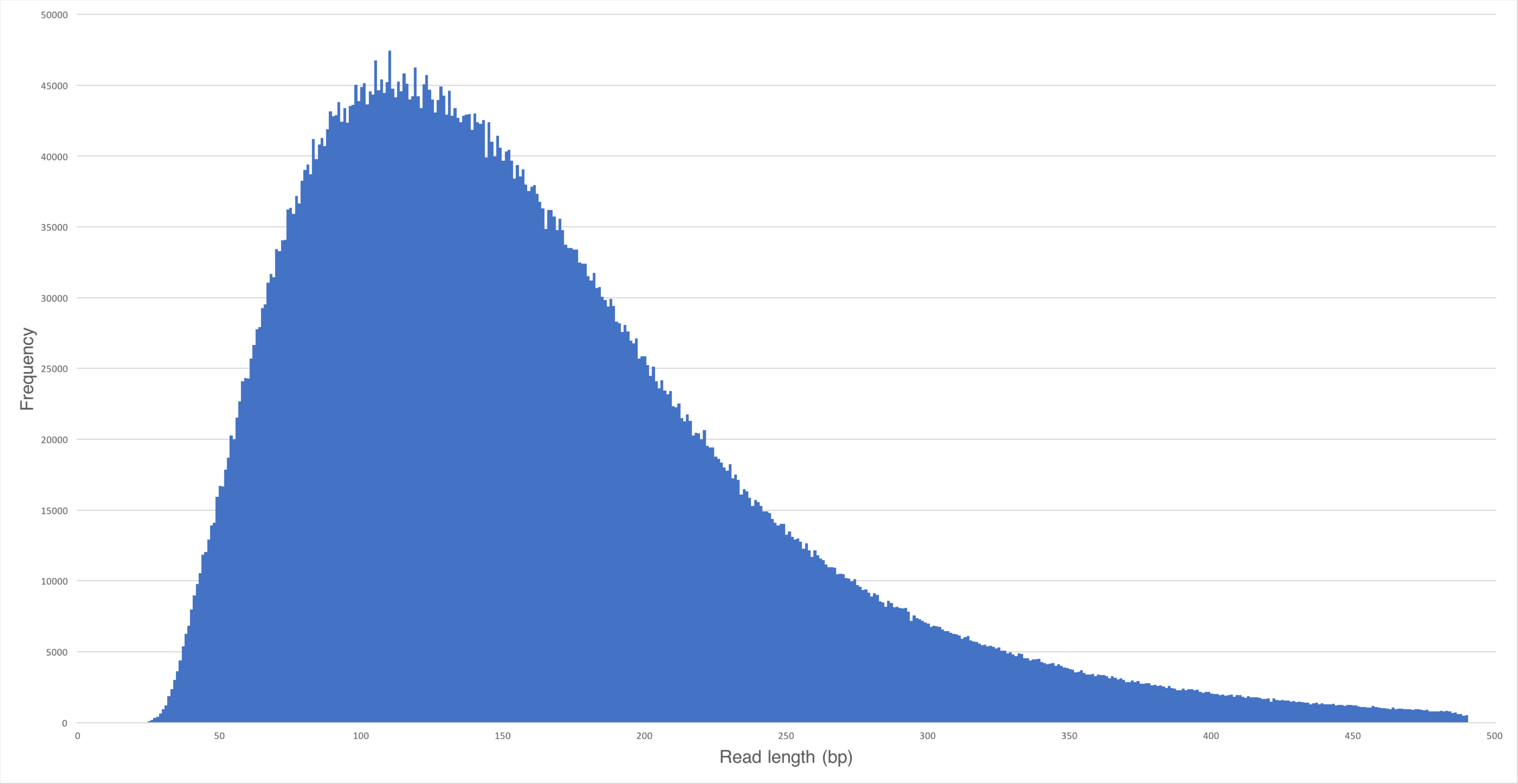


Figure 2.3 - the distribution of lengths from reads merged by PEAR. The majority of reads are 250bp or shorter, meaning paired end reads overlapped completely.

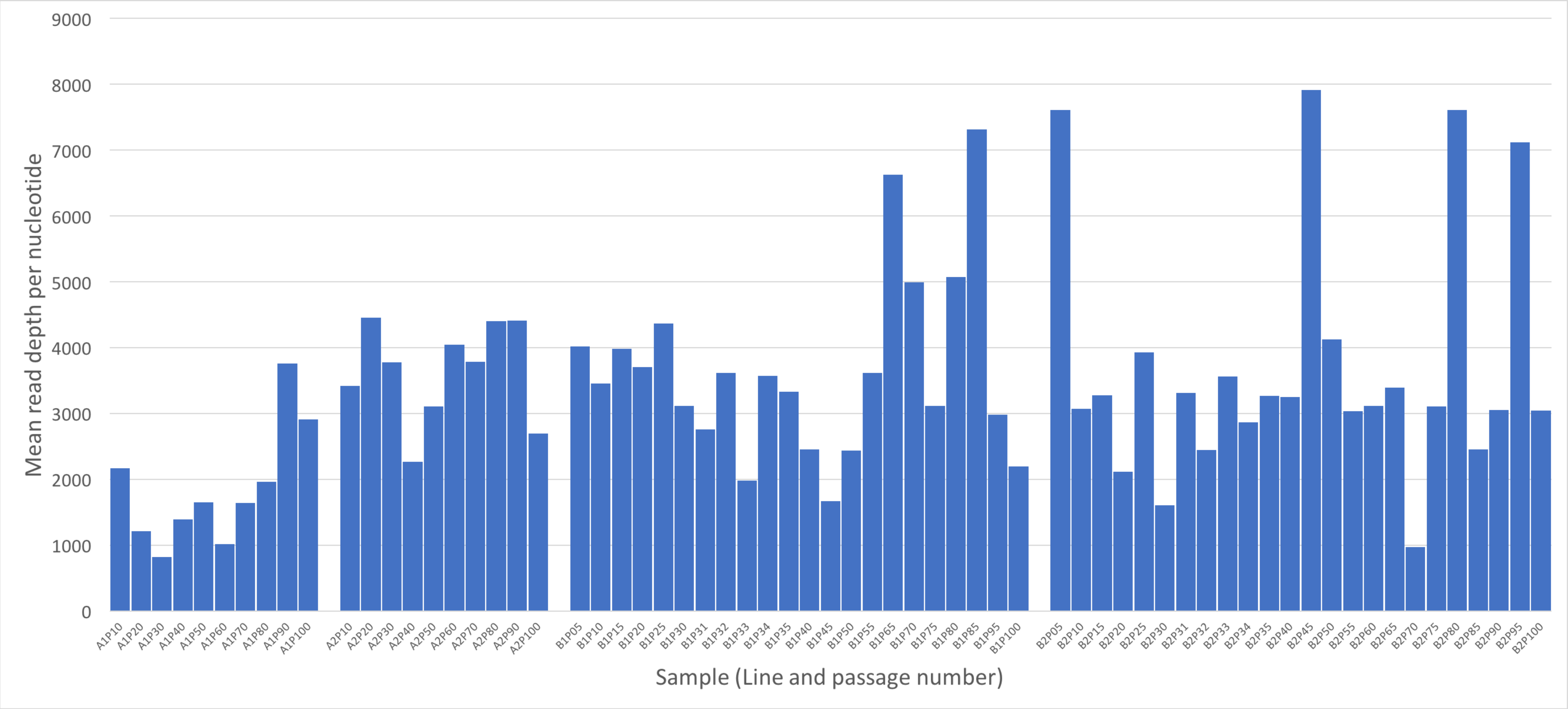


Figure 2.4 – mean number of reads per position for each sample sequenced

Chapter 3: Manipulation of the mutation rate of ΦX174

3.1 Introduction

In order to examine the effects of an elevated mutation rate on evolution, a method to manipulate mutation rate that works with the ΦX174 experimental system is required. The most suitable method would offer a broad mutational spectrum combined with limited non-mutagenic effects on the host, phage and environment.

3.1.1 Mutagens

Mutation rate of bacteriophage has been increased in the past by treatment with chemical mutagens. However, for an experimental evolution study specifically investigating the consequences of a high mutation rate, these have a number of drawbacks.

For example, in a previous experimental evolution study, Springman *et al* (2010) increased the mutation rate by adding the alkylating agent N-methyl-N'-nitro-N-nitrosoguanidine (MNNG) to passage media at a concentration of 10µg/ml, shortly before the addition of T7 bacteriophage. It was estimated that treatment with MNNG increased the T7 mutation rate by approximately three orders of magnitude.

Although MNNG is capable of methylating all oxygens and some nitrogens in DNA, its mutagenic activity primarily comes from its action on the O^6 position of guanine. This results in MNNG induced mutations being highly specific, with the G:C --> A:T transition making up 97.9% of all those observed in one study (Gordon *et al.*, 1990). Additionally, not all sequence contexts are equally susceptible to mutation, with a 5' flanking purine base increasing mutation frequency by up to nine times (Burns *et al.*, 1987).

This bias would be increased further in an experimental system using ΦX174. Because the mutagenic effect of MNNG applies directly to guanine, the lack of a complementary strand in this phage means that cytosine bases will not also be mutagenised (apart from a short period during replication when the ssDNA genome is converted to dsDNA) (Tessman *et al.*, 1965). The ΦX174 genome has only 1253 guanine residues, of which only 507 are flanked by a 5' purine base. The spectrum of possible mutations that could be induced by MNNG is therefore expected to be very limited and specific in this phage.

As well as alkylating agents, the other main class of mutagens is the nucleoside analogs, such as 2-aminopurine and 5-bromouracil. These work by substituting for specific nucleotides during DNA replication, but pairing with different amino acids during subsequent replications. While these act during DNA replication and so are not disadvantaged by the ssDNA genome of ΦX174, they still suffer from their specificity and biased mutational spectrum.

An additional drawback of mutagens is the possibility of non-mutagenic effects on host or phage physiology. MNNG, for example, triggers an adaptive response in *E. coli* that upregulates a number of genes including DNA repair enzymes, but also some unannotated transcripts (Booth *et al.*, 2013). These non-mutagenic effects would lead to a vastly different environment for the phage compared to that where mutagen was not added, potentially altering the spectrum of beneficial mutations for the phage, and would be very difficult to account for in a negative control. Nucleoside analogs can also have non-mutagenic effects, with their incorporation into genomes potentially causing transcriptional problems and sometimes inhibiting protein synthesis, or causing structural distortion and loop formation in the nucleic acids (Ojha *et al.*, 1991).

Poliovirus has been shown to become resistant to the nucleoside analog ribavirin by a single mutation in its polymerase gene (Pfeiffer and Kirkegaard, 2003). While this exact mechanism is not possible in ΦX174 (which does not possess a polymerase gene), the possibility of phage adapting to the presence of a mutagen should not be discounted.

DNA-damaging radiation such as ultraviolet (UV) light has also been used as a mutagen. Although this produces a larger mutational spectrum than the chemical mutagens, it is also more damaging to *E. coli* (Shuman and Silhavy, 2003). Additionally, constant exposure to UV would also difficult to implement in an experimental evolution system, so was also ruled out.

3.1.2 Mutator phenotypes

DNA replication fidelity is regulated in most organisms by three major factors: base insertion fidelity, proofreading, and mismatch repair (Echols and Goodman, 1991). An alternative route to increasing mutation rate would be to alter the regulation of these factors. The exact mechanisms by which these work depend on the organism, those stated in this section refer to those utilised by ΦX174, specifically those of its host *E. coli*.

Mismatch repair (MMR) is a post-replicative system that checks for and corrects mismatched base pairs (such as A-G or C-T) in a newly synthesised DNA strand. In order to determine which strand contains the erroneous nucleotide, MMR relies on the presence of the motif 5'-GATC-3', which is undermethylated for a short time after synthesis, allowing the system to distinguish between the parent and daughter strands. When in proximity to a mismatch, the MMR system nicks a nearby GATC site allowing excision and resynthesis of the strand (Fishel, 2015). ΦX174, however, contains no GATC motifs, meaning the MMR system has no effect and does not contribute to its mutation rate. When a mutant strain of ΦX174 was created with 7 GATC motifs introduced by site directed mutagenesis, it was found that the introduction of these sequences allowed MMR to act on ΦX174, decreasing mutation rate by 33 times that of the wild-type (Cuevas *et al*., 2011). Given that the mutations introducing the GATC motifs were either synonymous or intergenic and had no observable effect on growth rate, it is probable that these sites are selected against in the wild-type genome in order to keep the mutation rate high.

Base insertion and proofreading are both under the control of DNA polymerase and occur during DNA replication. As previously discussed (chapter 1.2.2.3), ΦX174 uses the DNA polymerase of its host during replication. In *E. coli*, this is the DNA III polymerase holoenzyme complex (figure 3.1). The α subunit, encoded by the *dnaE* gene is responsible for base selection (Maki and Kornberg, 1985), while the ε subunit, encoded by *dnaQ*, controls proofreading and removal of misincorporated bases. Some *E. coli* mutator phenotypes have been shown to be caused by mutations in these genes (Marinus, 2010).

A mutant of *dnaQ*, called *dnaQ926*, was created by site-directed mutagenesis that altered two amino acids (D12A; E14A). When replacing the wild-type gene on the *E. coli* chromosome, cells were inviable. However, when introduced to *E. coli* on a plasmid, a strong mutator phenotype was observed (Fijalkowska and Schaaper, 1996).

The mutational spectrum of another *dnaQ* mutant, *mutD5,* was determined with over 95% of observed mutations being single base pair substitutions. Of these, transitional mutations (A:T <--> G:C) were approximately 2-3 times as likely as transversions, with both types of transition observed in equal frequency. Although mutation bias is still present, it is less extreme than that from treatment with mutagens, and this spectrum more closely mimics that of normal non-mutagenic conditions where transitions are also more frequent than transversions (Chen *et al*., 2009).

Additionally, since *dnaQ926* is an allele of a protein already present in *E. coli*, it is unlikely that it has any harmful effects on the host or phage outside of its mutagenic ability. This makes it an ideal candidate for increasing the mutation rate in our experimental environment.

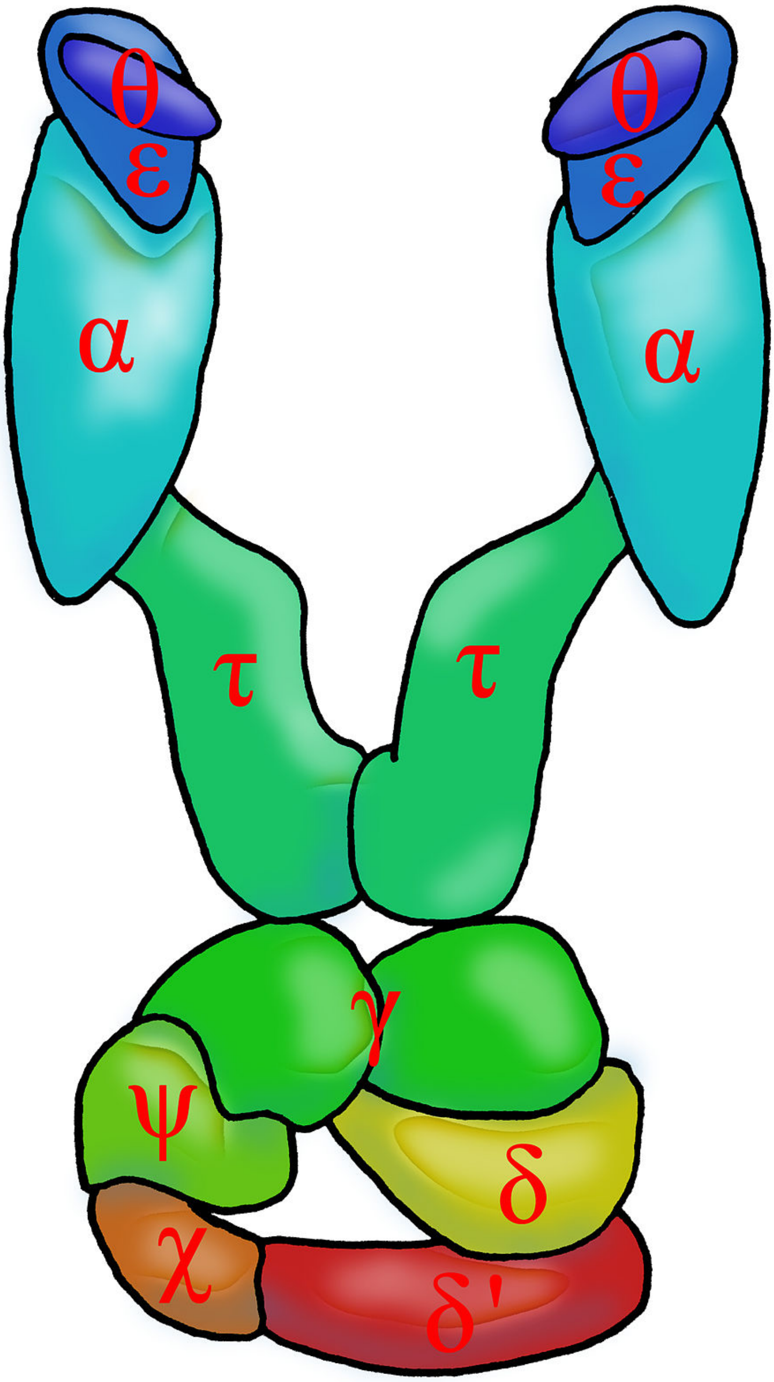


Figure 3.1 - the subunits of the *E. coli* DNA III polymerase holoenzyme complex. The ε subunit is encoded by the *dnaQ* gene and is responsible for proofreading during DNA replication.

Image by Alepopoli, via Wikimedia Commons.

3.1.3 Mutation rate measurements

There are two commonly used methods for estimating viral mutation rates: mutation accumulation studies and fluctuation tests (Duffy *et al.*, 2008). In mutation accumulation experiments, the frequency of new mutations and proliferation of existing mutations are tracked over time and used to calculate the mutation rate. However, a significant drawback of this method is that if a mutation has already occurred within a culture, any subsequent identical mutation events will be erroneously counted as progeny of the genome containing the original mutation (Foster, 2006). The likelihood of this occurring is increased in small genomes such as those of the Microviridae, and so this method was deemed unsuitable.

Fluctuation tests are based on the classic Luria-Delbrück experiment (Luria and Delbrück, 1943). Here, the frequency of mutation events that cause an observable phenotype is measured. If the spectrum of mutations that lead to this phenotypic change is known, the mutation rate can then be calculated.

As in mutation accumulation experiments, if a phenotype is observed more than once it is not possible to distinguish between unique mutation events and clones of a single ancestral mutation. However, if the number of mutation events in multiple replicate cultures are measured, it can be assumed that the number of mutations per culture follows a Poisson distribution (Luria and Delbrück, 1943). The proportion of these cultures that contain *no* mutations to the observable phenotype (P_0) can then be used to calculate the rate of mutation to that phenotype m with the equation:

$$m = \frac{-\ln P_0}{(N_f - N_i)}$$

Where N_i is the number of phage plaque forming units (pfu) used to initiate the culture and N_f is the number of phage pfu in the culture after growth.

For an observable phenotype, the mutant *am(E)W4* strain of ΦX174 was used, in which a single substitution in its lysis gene (E), changes tryptophan 4 (TGG) to an amber stop codon (TAG). When infecting *E. coli*, *am(E)W4* cannot synthesise its lysis protein, rendering it unable to escape the host cell, and therefore unable to form plaques on agar. A mutation in this codon will restore the ability of the phage to lyse its host and form plaques.

Phage cultures were propagated in the amber permissive BAF8 strain of *E. coli* C122. This strain contains the supF gene (Fane and Hayashi, 1991), which encodes a tyrosine transfer RNA with the anticodon CUA, complementary to the amber stop codon TAG. BAF8 suppresses the chain termination of this stop codon, allowing full synthesis of protein coding sequences containing amber mutations (Kraemer and Seidman, 1989).

3.1.4 Summary and aims

In the introduction to this chapter I have reviewed various methods that can be used to manipulate mutation rate in *E. coli*, and presumably also infecting ΦX174 bacteriophage. A plasmid containing the *dnaQ926* gene appears to offer the greatest spectrum of available mutations, as well as fewer drawbacks than the other methods. In the experimental section of this chapter, I measured the mutation rate of ΦX174 in the presence of this plasmid to confirm if the mutagenic effect was conferred to an infecting phage.

3.2 Results

The mutation rate of phage in mutagenic conditions was carried out by performing fluctuation tests in *E. coli* containing the pIF2013 plasmid, and in non-mutagenic conditions by using *E. coli* without a plasmid, using 72 replicate cultures. Phages were quantified at the beginning and end of the incubation period by qPCR. The mean number of phages in each culture, and the number of cultures where a reversion to plaque forming phenotype was *not* observed are shown in table 3.1. These data were used to calculate the rate of reversion to the plaque-forming phenotype, and the mutation rate per genome per replication, using the formulae described in 3.2.

The mutation rate of ΦX174 was calculated as (1.5 ± 0.06) x 10^{-6} substitutions per nucleotide per generation (sng) in non-mutagenic conditions, and (1.8 ± 0.12) x 10^{-4} sng in mutagenic conditions, a 120-fold increase.

	Wild-type	***dnaQ926***
Number of cultures	72	72
Cultures with no plaques	65	20
Proportion of cultures with no plaques (*P_0*)	0.901	0.278
Initial number of phages (*N_i*, pfu)	408 ± 10	388 ± 37
Final number of phages (*N_f*, pfu)	$(2.52 \pm 0.05) \times 10^4$	$(2.99 \pm 0.14) \times 10^3$
Rate of mutation to plaque-forming phenotype (*m*)	$(4.13 \pm 0.15) \times 10^{-6}$	$(4.92 \pm 0.33) \times 10^{-4}$
Substitutions per nucleotide per generation (*μ*)	$(1.5 \pm 0.06) \times 10^{-6}$	$(1.8 \pm 0.12) \times 10^{-4}$
Substitutions per genome per generation	0.01	0.97

Table 3.1 – Fluctuation test results and mutation rate calculations. N_f is a mean value averaged amount all replicates. ± denotes standard error.

3.3 Discussion

As predicted, the increase in *E. coli*'s mutation rate caused by introducing a defective proofreading gene is also conferred to bacteriophages that are using this machinery for replication.

It should be noted that a few assumptions were made when calculating the mutation rate *μ,* which are explained below. However, these assumptions are lent support by two previous studies, in which mutation rate of wild-type ΦX174 was reported as (1.0 ± 0.3) x 10^{-6} sng (Cuevas *et al.*, 2009) and (1.9 ± 1.8) x 10^{-6} sng (Raney *et al.*, 2004), measured by fluctuation test (albeit with a different phenotype to the amber reversion used here) and mutation frequencies respectively. These two independent measures are essentially the same as the wild-type mutation rate estimate of 1.5 x 10^{-6} sng presented here.

The first assumption was that *T*, the number of possible mutations that would result in a viable lysis protein, was 8. This was because of the 9 possible mutations in the amber stop codon, 8 of these resulted in a change to an amino acid codon (TAG > TAA would result in an ochre stop codon). Of these 8 substitutions, 3 have been confirmed to produce viable lysis proteins (reversion to the ancestral tryptophan and 2 mutations to tyrosine, which is on the BAF8 *E. coli* UAG tRNA), but it is still possible that some of these 5 remaining mutations may be lethal and not result in plaques. However, because this codon encodes part of the unfolded N-terminus chain before the active α-helix region (Mendel, 2006), it is unlikely that any substitutions at this position would significantly alter the structure or function of the protein. Despite this, it must be considered that *T* may be lower than 8, meaning the mutation rate could be up to 2.7 times larger than estimated here.

The effects of mutation spectrum bias on these measurements must also be considered. The calculations assume that all mutations are equally likely, but as previously mentioned, transitions occur with a greater frequency than transversions. Although sequencing plaques from the fluctuation test would allow *T* to be better estimated, the low frequency of some transversions would require an unfeasible number of individual cultures to be propagated before a mutation could be excluded with high confidence. Additionally, when the frequency of individual mutations is considered alongside their ability to restore phenotype, matters are complicated further. For example, if only the less frequent transversions resulted in viable lysis proteins, the more frequent transitions would have no effect on phenotype and so remain undetected, resulting in an underestimation of the true mutation rate. While it would be possible to modify the experiment to take these factors into account, the relative mutation rate increase between mutagenic and wild-type conditions is more important than an absolute value, so this was considered unnecessary. For simplicity's sake, it was assumed that all mutations are equally likely and the mutation rate at this position was representative of the mutation rate of entire genome.

When calculating the increase in mutation rate from the two estimates of μ, an accurate value for *T* is not necessary. Since the same phage and assay was used to measure mutation rates in wild-type and mutagenic conditions, any change in T will affect both μ values by the same magnitude, maintaining the relative increase of ~120 fold regardless.

Another consideration is that the fluctuation test only detects the substitution rate and does not detect indels. However, indels typically make up under 20% of the total mutations in most organisms (Drake, 2009), so the substitution rate should be close to the true mutation rate.

In this chapter, I investigated a variety of possible methods to increase the mutation rate of bacteriophage ΦX174 in an evolution experiment, and identified that a defective *dnaQ* gene would offer the advantage of a large mutational spectrum while lacking some of the drawbacks that would be encountered if mutagens were used. I have investigated the effects of the *dnaQ926* gene experimentally and found that it increases the ΦX174 mutation rate by two orders of magnitude, making it an ideal system to use to investigate the consequences of evolution at a high mutation rate.

Chapter 4: Evolution at a high imposed mutation rate

4.1 Introduction

4.1.1 Lethal mutagenesis

As the ultimate source of genetic variation, mutation is the fuel for evolution, giving selection the raw material on which it works. The rate at which mutations appear in a population of organisms is an important factor in their evolution; although an increase in beneficial mutations could facilitate the rate of adaptation it would be accompanied by more frequent deleterious mutations, lowering the fitness of the population as a whole. Conversely, a lower mutation rate would decrease the number of individuals that had their fitness lowered by deleterious mutations, but could lower their ability to adapt, potentially leading to extinction in times of crisis or environmental change (Agrawal and Whitlock, 2012).

As well as being the driving force behind evolution, the mutation rate is also an evolved trait. There are many molecular mechanisms in place that can prevent or repair DNA mutations, thereby lowering the mutation rate (Helleday *et al*., 2014). As discussed in section 1.1.5, this indicates that the mutation rate has evolved to lower the number of mutations with negative fitness effects.

Of all organisms, viruses have the highest mutation rates, with RNA viruses having the highest of all (Drake et al., 1998); (Sanjuán et al., 2010)). Although higher mutation rates cause beneficial mutations to arise more frequently, allowing faster adaptation to host defences, and sometimes even new hosts; it comes at the cost of a large number of deleterious mutations. These are offset by the large population sizes and high fecundity of RNA viruses, but they are close to the tolerable limit of mutation load (Arribas *et al*., 2016).

At a certain point, the mutation rate is predicted to become so high that selection is unable to counteract the accumulation of deleterious mutations. This causes a decrease in fitness over subsequent generations, leading to the eventual extinction of the population, a phenomenon known as lethal mutagenesis (Bull *et al.*, 2007). Artificially increasing the mutation rate to cause extinction is seen as a potential antiviral strategy (Domingo and Perales, 2016) and so the majority of studies on lethal mutagenesis to date have focused on it from a therapeutic perspective (Perales and Domingo, 2015). However, these only looked at the ability of the mutagen to clear the virus, meaning the genetic and evolutionary changes that underpin it remain poorly understood.

In the one study (Springman *et al.*, 2010) proposed a mathematical model for lethal mutagenesis, combining physiological parameters of the phage such as burst size, lysis time, and adsorption rate, as well as host density, the proportion of mutations with negative fitness effects and mutation rate. When they attempted to prove this empirically by evolving the dsDNA bacteriophage T7 at an elevated mutation rate, instead of the fitness decline predicted, the phage increased in fitness over the course of the experiment, rejecting the model.

Several possibilities were proposed for the failure of T7 to decline in fitness as expected (Bull *et al.*, 2013), the most important of which was adaptive evolution. The model did not take into account beneficial mutations, as well as compensatory and back mutations. However, considering the observed fitness increase, beneficial mutations and adaptive evolution undoubtedly occurred and were likely a factor in offsetting the predicted fitness decline.

4.1.2 Experimental design

In this chapter I designed and carried out an experiment to investigate the effect of an elevated mutation rate on the fitness of evolving populations of bacteriophage. ΦX174 was chosen for its lack of DNA polymerase genes, established use as a model organism in experimental evolution (Wichman and Brown, 2010) and because ssDNA viruses have intrinsically higher mutation rates than other DNA organisms. To increase the mutation rate while minimising mutational bias or non-mutagenic effects from using mutagens, the system described in the previous chapter was used.

Since mutation is a stochastic event, it was decided to carry out the experiment in duplicate to see how reproducible evolution is when increasing this element of randomness. Duplicate control lines were evolved in similar conditions but without the mutation rate increase; these would be an important comparison group to determine what would be expected to happen under non-mutagenic conditions, and what changes were products of the increase in mutation rate.

Recombination is a variable most models of lethal mutagenesis fail to take into account, so the experiment was purposely designed to minimise the likelihood of this occurring. For this reason, we elected to propagate our phages via serial passaging. While a chemostat would require less effort to maintain and could cover more evolutionary generations in the same time period, MOI is typically high in a chemostat and difficult to manipulate. Serial passaging would allow us to ensure that MOI was low at the beginning of each passage, and by transferring a small sample of phages to initiate the next passage allow us to effectively "reset" the MOI periodically as phage populations increased.

Previous work has determined that the fraction of substitutions in ΦX174 that are lethal is 0.20, and that a further 0.55 are deleterious (Domingo-Calap *et al.*, 2009). In the previous chapter, it was determined that the mutation rate in the presence of a plasmid with a defective *dnaQ* gene was such that on average each new phage genome would contain 0.97 mutations. From this, we can calculate that the number of progeny that will be viable is 78%. Of the viable progeny, we can expect 96% of them to contain a new mutation, of which 71% will be deleterious. However, it should be noted that the DFE can vary depending on genotype and environment, so there is uncertainty with the figures used for this calculation.

Since the majority of progeny will either be inviable or contain at least one deleterious mutation, fitness would be expected to decline in phages grown in the mutagenic environment compared to phages grown under normal conditions.

4.1.3 Summary and aims

While lethal mutagenesis has previously been reported in RNA viruses, these studies typically attempted to clear viral populations by treatment with mutagens and were concerned with developing treatments for viral infections rather than understanding the mechanisms at work. In this chapter an evolution experiment was carried out to investigate the consequences of viral evolution at an artificially elevated mutation rate.

The aims of this chapter were:

- To investigate if an increase in mutation rate of two orders of magnitude results in the extinction of a ssDNA virus.
- To observe how an elevated mutation rate affects the fitness of evolving populations of bacteriophages.
- To determine how reproducible evolution is at an elevated mutation rate.

4.2 Results

4.2.1 Wild-type phage

The phage used to initiate the experimental lines was isolated from a single plaque, then grown under non-mutagenic conditions until titer was sufficient (>10^5pfu/µl, approximately 2 hours). Illumina sequencing of this preparation showed that the genotype was not uniform, with a single mutation at position 1301 reaching 48.8% frequency in the population (appendix C.1).

The fitness of the ancestral phage population was found to be 3.9 ± 0.32 doublings/hour. An additional fitness assay using phages taken directly from the glycerol stock that the initial plaque was isolated from gave a value lower than zero. This suggests that the initial phage was not well adapted to our experimental conditions. Since only one mutation was detected at a high frequency and it resulted in a large fitness increase, it was considered to be an adaptation to the experimental conditions.

4.3.2 Fitness of experimental lines

Bacteriophage fitness assays are typically carried out on phage of a single genotype, isolated from a single plaque. Under normal, non-mutagenic conditions, a population of phages will generally contain little genetic variation. When a beneficial mutation does survive drift, it will quickly become fixed as a selective sweep removes competing alleles (Smith and Haigh, 2007). However, this is untrue for the populations evolved in mutagenic conditions where many co-occurring mutations at varying frequencies mean there is likely to be variation between individuals. While single plaques could be isolated and the fitness of these individuals measured, they would not be representative of the population as a whole. For this reason, fitness assays were performed on heterogeneous populations, to give an average fitness for all members.

Fitness can be measured in several ways, at its most complex based on calculations consisting of many components of phage life history such as adsorption rate, burst size, and lysis time. This could also cause difficulties in calculations when a population is heterogeneous and these components vary. For example**,** the mechanism for inducing lysis in ΦX174 is simpler than in many other phages, and results in an asynchronous lysis time; the commonly used lysis time of 21 minutes is an average (Hutchison and Sinsheimer, 1963). If the heterogeneity of the population caused lysis time to vary further between individuals, it could present difficulties in measuring this and begin to make it look like phages were being constantly released. It has previously been reported that fitness measures using multiple components underestimate true fitness in heterogeneous populations (Springman *et al.*, 2010). It was decided that in this experiment, fitness would be estimated using the average growth rate of the population as a proxy.

While it is usually desirable to carry out fitness assays in the same environment phages have evolved in, this would have presented difficulties. In the presence of *dnaQ926*, bacterial growth rates are retarded due to the constant mutagenic effects (Fijalkowska and Schaaper, 1996), which in turn is likely to affect phage growth rates. However, we are interested in measuring the fitness of the already-mutagenised phage, and by assaying it in a mutagenic environment the fitness measured would be a component of both the phage's actual fitness and the mutagenic activity occurring during the assay. This would not allow direct fitness comparisons of these lines with the ones grown under normal conditions, and for this reason it was decided to carry out fitness assays of all lines in the non-mutagenic environment. Deleterious mutations can interfere with plaque formation, so phage titers were determined during the assay using qPCR. The primers were located in a region of gene E where mutations had not been observed in previous studies with ΦX174.

Replicate lines of bacteriophage ΦX174 were evolved for 100 hours of growth (approximately 300 generations) at both its normal mutation rate of 1.5 x 10^{-6} sng and an elevated mutation rate of 1.8 x 10^{-4} sng. At the end of the experiment (passage 100), both lines evolved at the normal mutation rate (A1 and A2) had increased in fitness significantly compared to the ancestral phage confirming that the ancestral phage used was not well adapted to the experimental conditions, and beneficial mutations were available. The mutagenic lines (B1 and B2) showed very different changes in fitness. B1 had a similar fitness gain to A1 and A2. B2, however, significantly decreased in fitness over the course of the experiment (figure 4.1).

Fitness of evolutionary lines was also assayed after every 20 passages (approximately 60 generations). Both control lines A1 and A2 showed similar changes in fitness at each of these time points. After a rise in fitness at passage 20, both lines declined in fitness at passage 40 before a sharp rise at passage 80.

B1 continually increased in fitness at every time point up to passage 80. At 20 passages, the fitness was similar to A1 and A2. However, in B1 fitness increased at passages 40 and 60 over the previous time points, compared to the declines in the non-mutagenic lines.

After 20 passages, line B2 showed a negative fitness value (fewer phages were detected after growth than were present in the initial inoculum). Fitness increased to a positive value at passage 40 onwards, but remained at a similar level for the rest of the experiment.

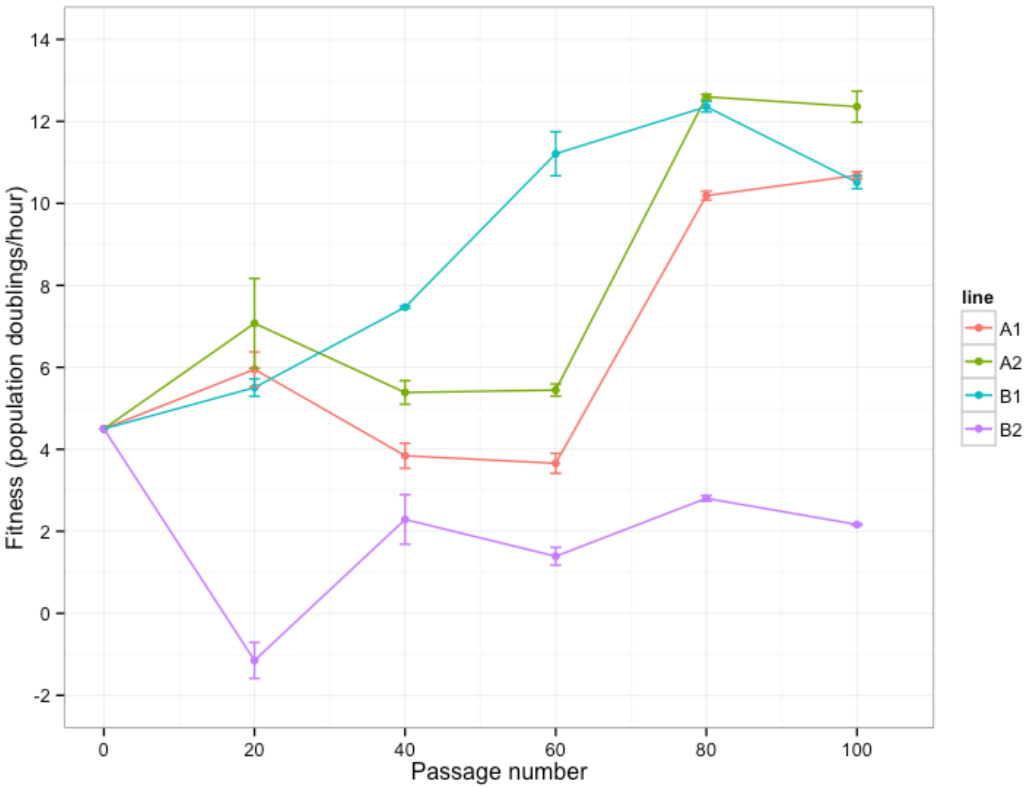


Figure 4.1 - fitness of each experimental line at intervals of 20 passages. Each time point was assayed in triplicate. Error bars indicate standard error. All lines were initiated from the same stock, the fitness of which is included here as passage 0.

	T_0	T_F			Fitness (population doublings / hour)					
		1	2	3	1	2	3	Mean	SD	SE
Anc	2.64E+05	3.78E+06	7.95E+06	4.07E+06	3.55	4.54	3.65	3.91	0.55	0.32
A1 P20	2.78E+05	4.67E+07	2.49E+07	1.21E+07	6.83	5.99	5.03	5.95	0.90	0.52
A1 P40	2.74E+05	6.73E+06	2.80E+06	6.22E+06	4.27	3.10	4.16	3.84	0.65	0.37
A1 P60	2.27E+05	2.30E+06	4.92E+06	3.88E+06	3.09	4.10	3.79	3.66	0.52	0.30
A1 P80	2.84E+05	4.91E+08	6.06E+08	6.93E+08	9.94	10.22	10.40	10.19	0.23	0.13
A1 P100	2.97E+05	9.54E+08	7.59E+08	9.93E+08	10.77	10.46	10.82	10.68	0.19	0.11
A2 P20	1.15E+05	1.68E+08	1.14E+07	6.48E+06	9.72	6.13	5.38	7.07	2.32	1.34
A2 P40	3.08E+05	2.95E+07	1.46E+07	1.25E+07	6.08	5.14	4.93	5.39	0.61	0.35
A2 P60	1.09E+05	5.48E+06	5.83E+06	8.48E+06	5.22	5.31	5.81	5.44	0.32	0.18
A2 P80	1.36E+05	1.54E+09	1.86E+09	1.79E+09	12.45	12.70	12.65	12.60	0.13	0.08
A2 P100	1.15E+05	8.69E+08	2.44E+09	8.56E+08	11.91	13.28	11.89	13.52	0.80	0.46
B1 P20	4.29E+05	1.81E+07	3.30E+07	3.17E+07	4.99	5.79	5.74	5.51	0.45	0.26
B1 P40	2.21E+05	6.04E+07	7.65E+07	6.22E+07	7.48	7.80	7.52	7.47	0.17	0.10
B1 P60	8.95E+04	2.45E+08	1.07E+09	2.45E+08	10.55	12.52	10.55	11.21	1.14	0.66
B1 P80	4.81E+04	4.32E+08	6.42E+08	4.77E+08	12.14	12.67	12.27	12.36	0.27	0.16
B1 P100	8.62E+04	1.72E+08	2.67E+08	2.61E+08	10.13	10.72	10.69	10.51	0.33	0.19
B2 P20	2.55E+06	1.08E+06	5.36E+05	2.16E+06	-1.15	-2.08	-0.22	-1.15	0.93	0.54
B2 P40	1.29E+05	2.55E+05	8.39E+05	1.72E+06	0.91	2.50	3.46	2.29	1.29	0.74
B2 P60	1.09E+05	2.11E+05	4.13E+05	3.39E+05	0.88	1.78	1.51	1.39	0.46	0.26
B2 P80	1.27E+05	1.02E+06	1.02E+06	1.17E+06	2.78	2.77	2.96	2.80	0.11	0.06
B2 P100	1.18E+05	5.90E+05	4.71E+05	7.23E+05	2.14	1.84	2.41	2.17	0.29	0.17

Table 4.1 - fitness assay measurements. T_0 is the number of phages used to initiate the assay (pgu), T_f is the number of phages at the end of the assay (pgu), with 1, 2, and 3 denoting replicate measurements. Values are provided in E-notation, where “E” represents “times 10 raised to the power of”. Fitness was calculated by the equation:

$$\text{fitness} = \log_2(N_f/N_0) / 0.75$$

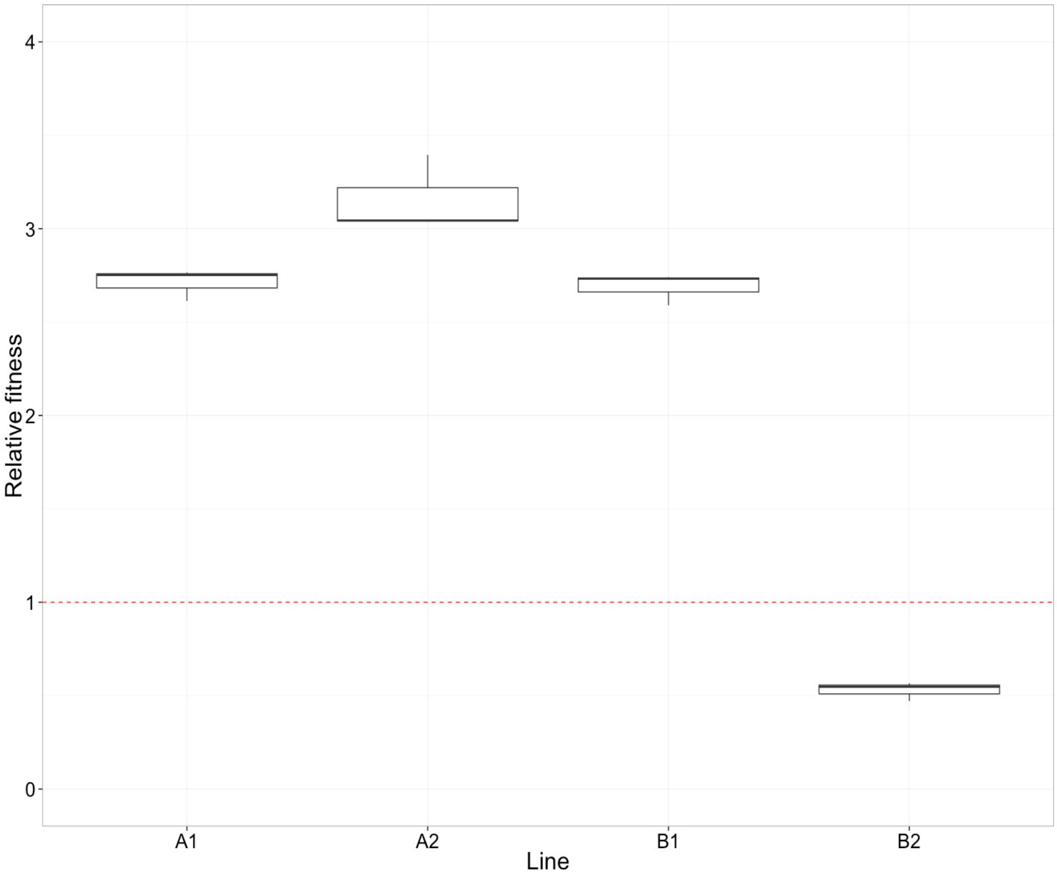


Figure 4.2 – Box and whisker plot showing fitness of experimental lines after 100 passages. Fitness is expressed relative to the fitness of the ancestral phage preparation (dotted line). Top and bottom of box indicate third and first quartiles respectively, whiskers extend to maximum and minimum.

4.2.3 Endpoint genotypes

Consensus sequences of endpoint (passage 100) genotypes are given in full in appendix B.

Over the four lines, 13 unique mutations had become the major allele (table 4.2). Of these, four were synonymous and nine were non-synonymous. Four non-synonymous substitutions appeared in multiple lines - two of them in three lines and two in all four lines. One of the synonymous mutations appeared in two lines.

Both control lines A1 and A2 contained five substitutions that were the major allele at that position (with frequencies of 89% or greater). Of the new mutations in lines A1 and A2, four were the non-synonymous mutations common to multiple lines, while each also contained a unique synonymous mutation.

Line B1 contained six mutations that were the major allele at that position (with frequencies of 69% or higher). Of these, four were the shared non-synonymous mutations, one was a unique non-synonymous mutation, and the other was a synonymous mutation shared with B2.

Line B2 contained eight mutations that were the major allele at that position (with frequencies of 70% or higher). Five of these were unique to this line (four non-synonymous, 1 synonymous), two non-synonymous mutations shared with all other lines, and one was a synonymous mutation shared with B1.

Position	Gene	Mutation	A1	A2	B1	B2	
841	D/E	V150A (GTG→GCG) /*91R (TGA→CGA)	🟥	🟥	🟥	🟩	
1301	F	T101A (ACT→GCT)	🟩	🟩	🟩	🟩	
1305	F	G102D (GGT→GAT)	🟩	🟩	🟩	🟥	
1319	F	A107T (GCC→ACC)	🟥	🟥	🟥	🟩	
1321	F	A107A (GCC→GCT)	🟥	🟩	🟥	🟥	Silent
1639	F	M213I (ATG→ATT)	🟩	🟩	🟩	🟥	
1660	F	I220I (ATT→ATC)	🟩	🟥	🟥	🟥	Silent
1968	F	N323S (AAC→AGC)	🟥	🟥	🟥	🟩	
3320	H	A130A (GCT→GCC)	🟥	🟥	🟩	🟩	Silent
3340	H	D137G (GAT→GGT)	🟩	🟩	🟩	🟩	
3423	H	I165V (ATT→GTT)	🟥	🟥	🟩	🟥	
3426	H	A166S (GCC→TCC)	🟥	🟥	🟥	🟩	
4802	A/A*	P274P (CCT→CCC)	🟥	🟥	🟥	🟩	Silent

Table 4.2 - mutations in the endpoint (passage 100) consensus genotypes compared to the ancestral sequence. Green denotes that a mutation is the major allele in that line, red denotes that the major allele at that position is unchanged from the wild-type. Note that the mutations at 841 and 4802 affect overlapping genes. 841 changes the amino acid sequences of two genes. Genes A and A* are located in the same reading frame, so the same mutation occurs in each gene (although this is located at position 102 in A*).

4.3 Discussion

The work presented in this chapter represents one of the most comprehensive experimental evolution studies into the consequences of viral evolution at an elevated mutation rate to date. Replicate populations of ΦX174 initiated from a single ancestor were evolved at a normal and elevated mutation rate. While the lines grown at a normal mutation rate showed convergent evolution, acquiring the same mutations and same fitness increase, the lines grown in mutagenic conditions resulted in dramatically different outcomes.

4.3.1 Lethal mutagenesis

These results show that lethal mutagenesis was not achieved during the course of the experiment, meaning the mutation rate was not high enough or that not enough generations had passed for the fitness to decline sufficiently. Interestingly, in line B2 the fitness appeared to be negative after 20 passages, meaning that fewer phages were detected at the end of the fitness assay than the beginning. However, all subsequent fitness assays returned positive fitness values. Presumably the population was rescued by selection of its fitter members.

While extending the duration of the experiment may eventually result in lethal mutagenesis, from a therapeutic point of view, prolonged viral infection and exposure to mutagens are undesirable to a patient, and so any further studies should focus on increasing the mutation rate further.

Although lethal mutagenesis does not occur, it was expected that the elevated mutation rate would still result in a fitness decline. This was observed in line B2, where fitness declined to approximately half that of its initial value, and then appeared to reach an equilibrium, not changing significantly from passage 40 onward. In line B1, which was evolved in identical conditions from the same ancestor, a large fitness increase was observed over the course of the experiment. The explanation for this is that adaptive evolution occurred to a much greater extent in line B1 compared to B2.

4.3.2 Parallel evolution in A1 and A2

Sequencing of endpoint genotypes showed many incidences of parallel evolution. In particular, four common substitutions were detected in both control lines. While it is probable that the mutation at 1301 arose in the ancestor, it increased to fixation in every line, strongly suggesting it is an adaptation to the experimental conditions. The remaining three mutations all arose independently in both lines and increased to high frequencies, suggesting that they too are adaptive. This is further supported by the large fitness increases observed in these lines over the course of the experiment.

A1 and A2 also contained a unique substitution each. Both of these were non-synonymous so were likely to be selectively neutral. The most plausible explanation for their fixation is that they were present on a genome on which a beneficial mutation appeared, and rose in frequency due to genetic hitchhiking. Other than these synonymous substitutions, the genomes of lines A1 and A2 are the same and produce an identical set of proteins. Both lines evolved at wild-type mutation rate showed very similar changes in fitness over the course of the experiment, suggesting that the same adaptive mutations appeared in both populations at similar time points. This example of parallel evolution demonstrates how reproducible evolution can be when an organism with a small genome and limited spectrum of beneficial mutations is grown in identical conditions.

Between passages 20 and 40, fitness of lines A1 and A2 declined somewhat, and appeared to maintain this level between passages 40 and 60. It is interesting that this occurred in both non-mutagenic lines, and suggests that there is some component of fitness outside of growth rate that our assay did not consider.

4.3.3 Adaptive evolution in B1

From the fitness increases observed in lines A1, A2 and B1, it appeared that there were a few available beneficial mutations that were of large effect. Bacteriophage with an identical genotype has previously been used in other experimental evolution studies (Schaaper, 1998) (Raynes and Sniegowski, 2014); but it is important to consider the novelty in our experimental conditions, such as the small volumes used for cultures and the presence of a plasmid in the host bacteria, may have given scope for the phage to adapt to. In fact, it is apparent that while the culture of phages was being grown to provide sufficient titer to begin the experiment, a mutation (1301) appeared and increased in frequency. A review of experimental evolution studies with ΦX174 found that this mutation was observed more frequently than any other (Wichman and Brown, 2010). This ancestral culture originated from a single, sequenced plaque where this mutation was not detected, suggesting it was adaptive and the phage population used to initiate the experimental lines was already partway through a selective sweep.

Although adaptive evolution took place in B1, it is unlikely that this was an adaptive response to the elevated mutation rate. This is because most mutations were shared with the non-mutagenic lines, and all non-synonymous mutations in this line were located in genes F and H, the coat protein and DNA pilot protein. Both these proteins have extracellular functions, and do not play any part in DNA replication.

As in the previous study where adaptive evolution appeared to more than offset the expected fitness decline (Springman *et al.*, 2010), it appears that this also happened in line B1. In that study, fitness measurements were only taken in the ancestral phage and on completion of the experiment, which suggests that fitness only increased. In this experiment, fitness was measured at 60 generation intervals, showing a steady increase in fitness over the first 240 generations before a decline in the next 60. It is possible that the same pattern occurred in that experiment - a higher fitness peak was achieved partway through the experiment before it declined to the value that was measured at the end.

If beneficial and deleterious mutations are considered separately, it could be that the expected decline in fitness through accumulation of deleterious mutations is still occurring, but is obscured in these studies by the large fitness increases from the beneficial mutations and the strong selective pressure working on them. Since every potential mutation has almost certainly occurred many times over in this experiment (box 4.1), it is unlikely that there are other beneficial mutations that have fitness effects of the magnitude already observed. If this is the case, then the fitness observed in B1 after 240 generations is probably close to the optimum fitness of the phage in this environment, meaning a continuation of the fitness decline seen in the last 60 generations would be observed if the experiment was continued.

In any similar studies in the future, phage should be adapted to the experimental system beforehand in non-mutagenic conditions. This will allow phage lines to be initiated with a starting genotype that is close to optimum fitness, and limit the influence of beneficial mutations in the experiment proper.

Box 4.1 – The coupon collector's problem

I have stated that every possible mutation has appeared many times over during this experiment, which is due to the small genome sizes, high mutation rate and large population sizes.

There are 5386 bases in the ΦX174 genome, and 3 possible mutations for each. This means there are 16,158 potential mutations that could be observed. To calculate how many mutational events are required to observe every unique mutation, we can use the coupon collector problem from probability theory, which asks the question "given *n* coupons, how many coupons do you expect you need to draw with replacement before having drawn each coupon at least once?" (Hayes and Pilling, n.d.)

There is an exact solution to this problem which returns the expected value of the number of draws needed to obtain each unique coupon. Because our *n* value of 16,158 is so large, it would require a great amount of time and computational power. However, there is an approximate solution that is accurate to one decimal place:

$$\mathrm{E}[X] \approx n(\ln n + \gamma) + \frac{1}{2}$$

where γ = 0.577216, Euler's constant. Solving for our value of *n*, this tells us that we would expect to see all 16,158 unique mutations after 165,901 mutational events.

As determined in chapter 3, at our elevated mutation rate there are an average of 0.97 mutations per phage genome. This means that 165,901 mutations would occur per 171,032 new phage. Since each passage was initiated with at least 10^6 pgu, there would be many times more than this number in each passage, even if each phage replicated only once. It must be noted that this calculation makes the false assumption that each mutation is equally likely. However, when the fecundity of the phage and the number of passages are considered, it is probable that the number of mutations calculated was exceeded by a factor of thousands.

4.3.4 Failure to adapt in B2

Unlike the other lines, B2 did not increase above the starting fitness at any of the time points assayed over the course of the experiment. It only had two mutations in common with the other lines, one of which was 1301, which was already increasing in frequency at the beginning of the experiment. Of the five substitutions unique to this line, one of them was synonymous, while the other four all result in protein sequence changes. Of particular note was the mutation at position 841. This position comprises part of the overlapping genes D and E, which are both in different reading frames. This mutation alters the protein coding sequence in both of these genes. Interestingly, it changes the stop codon in gene E to an arginine, meaning that during transcription of the gene the polypeptide chain will not terminate at the normal position and will continue to add amino acids. The next stop codon downstream in this reading frame begins at position 853. As a result, this mutation not only changes the amino acid sequence of gene D, it also causes the product of gene E to be extended by four amino acids: arginine, cysteine, asparagine and valine.

It is interesting that despite there being multiple adaptive substitutions that independently arose in all other lines, two of them failed to become fixed in B2. These mutations conferred large fitness increases in the other three lines, yet B2 only declined in fitness despite greatly beneficial mutations being available.

Since these mutations undoubtedly occurred many times over (box 4.1), the most likely explanation is that the mutations were not beneficial in B2. It is unlikely that this is a consequence of the mutagenic environment since a fitness increase still occurred in B1, meaning that it is likely caused by the different genetic background of B2. Both these mutations (at positions 1305 and 1639) appear in gene F, which encodes the coat protein. B2 also contains mutations in the external scaffolding protein and the spike protein, which interact with the coat protein during assembly and in the viral capsid respectively. It is possible that the combination of these unique mutations with 1305 and 1639 produces a negative epistatic effect, and does not confer the large fitness increase observed in the other lines. B2 also contains two unique non-synonymous mutations in gene F, so there is also the possibility that these do not merely have additive effects on fitness, but together with 1305 and 1639 adversely affect the folding of the protein or its interactions with other viral proteins or host receptors.

Although it was not observed here, the reverse may be possible too. If an epistatic interaction between multiple substitutions produces a large fitness increase but these mutations alone are negative or neutral, under normal conditions it is unlikely that they will occur in tandem. A catch-22 situation occurs: the mutations cannot occur together unless one achieves high frequency first, but they cannot increase in frequency unless they occur together. In a mutagenic environment, however, the chances of them co-occurring would be increased, leading to a spectrum of potential adaptations that would be unlikely at normal mutation rate. While this may have occurred here as well, it cannot be determined without measuring the fitness effects of the specific mutations alone and in combination.

4.3.5 Further work

There are a number of possibilities to further investigate lethal mutagenesis with experimental evolution, either by continuing this experiment, or adapting it further.

It may be that 300 generations did not provide sufficient time for a fitness decline to be observed. Since phage samples were taken and stored from every passage, the experiment carried out in this chapter could be continued and extended beyond passage 100. A less labour-intensive method that would allow for many more generations would be to conduct a similar experiment using a chemostat instead of serial passaging. However, recombination events would be more likely in this environment, something the model does not take into account.

While extending the experiment would be interesting from a theoretical perspective, increasing the mutation rate further still would be preferable and maybe even necessary to achieve lethal mutagenesis. This is harder to do experimentally without the use of mutagens, but other *dnaQ* mutants could be investigated to see if they increase mutation rate above that of *dnaQ926*.

Adaptive evolution presents a significant challenge for this model. While some adaptive mutations cannot be prevented from occurring (such as direct adaptation to the mutagenic conditions or compensatory mutations), the potential fitness gains from adaptation to the general experimental conditions could be lowered. If the experiment was initiated using phage with an optimal genotype (or close to it), many of the large fitness gains seen in line B1 would not have occurred since the mutations would already be present. This phage could be acquired by passaging the phage in the non-mutagenic experimental conditions beforehand. Phage from line A2 would be a good candidate for this, since it reached the highest fitness observed during the experiment at passage 80 and no adaptive mutation appears to have occurred in the subsequent 20 passages.

Chapter 5: Evolutionary dynamics of experimental lines

5.1 Introduction

5.1.1 The evolution experiment

In the experiment described in the previous chapter, replicate populations of bacteriophage ΦX174 were evolved at wild-type and elevated mutation rate. While both lines grown under non-mutagenic conditions demonstrated convergent evolution, the lines grown in mutagenic conditions evolved in very different ways, with one displaying parallel evolution with the non-mutagenic lines and greatly increasing in fitness, and the other acquiring novel mutations and declining in fitness.

At high mutation rates, evolutionary dynamics may be affected by many of the processes discussed in chapter 1. For example, deleterious mutations are more common at high mutation rates, so it follows that compensatory mutations would also be more frequent. If the compensatory mutation restores much of the fitness lost through the deleterious mutation, then these will be more likely to persist. Another evolutionary phenomenon that may be more readily observed at high mutation rate is clonal interference. If beneficial mutations are more likely to co-occur then it is likely that they will compete with each other, resulting in slower adaptation and the loss of one allele.

Samples of phage lysate from each passage of this experiment were stored as a "frozen fossil record". While these have previously been used to determine the changes in fitness of each line, they also serve as a permanent record of the genotypes of the bacteriophages over the course of the experiment. Using next-generation sequencing, composite genomes of populations can be obtained for multiple time points of the experiment allowing us to observed the changes in allele frequency over time and determine evolutionary trajectories.

5.1.2 Next-generation sequencing

In experimental evolution studies involving bacteriophage, sequencing is typically performed on homogenous isolates from a single plaque. However, in evolving populations it can be useful to detect substitution trajectories, which can only be inferred by conventional methods from the sequencing of numerous individual genotypes. This leads to difficulties; as well as being time- and resource-intensive, small sample sizes mean that low-frequency substitutions may not be detected. This is especially pertinent when phage is evolved at an elevated mutation rate, since this is expected to greatly increase diversity with populations.

In next-generation sequencing with platforms including Illumina^{TM}, the sample DNA is broken into multiple short fragments which are tagged with adaptors and indexes, amplified by PCR and then sequenced. Each fragment is sequenced from the 5' end on both strands for a predetermined length (which ranges from 75bp to 300bp, depending on the reagents used). These paired end reads can be reassembled by aligning them to a reference genome. Typically, multiple reads will map to the same part of the reference genome, with the average number of reads for each nucleotide position being referred to as coverage. Since each nucleotide is sequenced multiple times, when a sample consists of a single individual increased coverage makes it easier to distinguish between sequencing errors and true mutations.

Because ΦX174 has such a small genome, it is possible to get very high coverage with a small overall number of reads (e.g. as achieved in multiplex runs on the MiSeq instrument). This makes it an ideal method for detecting low frequency substitutions that may have occurred in the evolution experiment described in chapter 4. Instead of initiating sequencing with DNA from individual isolates, sample DNA can be extracted from lysate containing an entire population of phages. When sequenced, the high coverage allows the frequency of substitutions at each position to be determined (Dickins and Nekrutenko, 2009).

5.1.3 Summary and aims

When grown at an elevated mutation rate, two replicate populations of ΦX174 evolved in very different ways. Using next-generation sequencing, the substitution frequencies of these populations over the course of the experiment will be determined to investigate how these populations changed at the genetic level.

The aims of this chapter were:

- To obtain high-resolution sequence data for each evolutionary line over multiple time points.
- To use this data to infer evolutionary trajectories in evolving populations of bacteriophage.
- To identify how evolutionary processes are affected by an increased mutation rate.

5.2 Results

Phage dsDNA from passages 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 was extracted for all lines. In addition, DNA was extracted from passages 5, 15, 25, 31, 32, 33, 34, 35, 45, 55, 65, 75, 85, and 95 in lines B1 and B2. dsDNA samples were fragmented and sequenced using the Illumina platform, and output data was analysed using the workflow described in 2.10.

5.2.1 Substitutions

Over the course of the experiment, eight mutations were detected at a frequency of 10% or greater in lines A1 and A2. Of these, five were still present at high frequency after 100 passages, with four of these being non-synonymous and one being synonymous. Each line contained a non-synonymous mutation as well as two synonymous mutations, which either declined in frequency or were lost by the end of the experiment. These mutations are listed in tables 5.2 and 5.3.

In the lines evolved under mutagenic conditions, far more mutations were observed. 21 mutations in B1 and 23 in B2 were observed which reached a frequency of 10% or greater in one of the time points sequenced. It should be noted that phage from these lines were sequenced at more frequent time points (24 for B2, 22 for B1, 10 for A1 and A2). However, these high frequency mutations persisted for multiple time points and would have been observed even with less frequent sequencing.

Of the 21 mutations seen in line B1, 14 were still present at a frequency of 10% or greater by 100 passages, with six being the major allele (that with the greatest frequency) in the population. Five of these were non-synonymous mutations, with the others being synonymous. All of the seven mutations in this line which declined below 10% frequency were synonymous. Eight mutations were present in the final population at 10% frequency or higher but were not the major allele. Three of these were non-synonymous and two were synonymous. The remaining three mutations were observed in the region where the D and E genes overlap. These are in different reading frames, and so a single mutation affects two codons differently. In all three cases, this mutation was synonymous in gene D but changed the amino acid sequence in gene E. These mutations are listed in table 5.4.

12 of the 23 mutations in line B2 were lost or had declined beneath 10% frequency by passage 100. Three of these were synonymous, while another was located in an intergenic region of the genome. Two were located in the overlapping D and E genes, one affected the protein sequence of both genes, and one affecting only gene D. The remaining six were non-synonymous. Of the remaining 11 mutations that were still present at 10% or greater frequency after 100 passages, nine were the most common variants in the population. Just two of these were synonymous, with the other nine all altering protein sequences, and one of these changing sequences of both genes D and E. These mutations are listed in table 5.6.

Low frequency substitutions were detected in all lines, but were much more common in the mutagenic lines (table 5.6).

Position	Gene	Mutation	Amino acid change	First detected (passage #)	Frequency at 100 hours
1301	F	ACT→GCT	T101A	10	100%
1305	F	GGT→GAT	G102D	80	93%
1639	F	ATG→ATT	M213I	10	98%
1660	F	ATT→ATC	I220I	70	89%
3340	H	GAT→GGT	D137G	50	100%
4613	A/A*	GAG→GAA	E211E	50	1%
4622	A/A*	TAT→TAC	Y214Y	30	4%
4627	A/A*	AAT→AGT	N216S	50	0%

Table 5.1 - Every mutation in line A1 that was observed with a frequency of ≥ 10%. Table includes mutational frequency at the conclusion of the experiment (rounded to the nearest integer), and the passage number in which mutation was first detected at ≥ 10% frequency. Genes A and A* are in the same reading frame so shared mutations cause the same amino acid change in both genes. Amino acid change in table refers to gene A.

Position	Gene	Mutation	Amino acid change	First detected (passage #)	Frequency at 100 hours
1301	F	ACT→GCT	T101A	10	100%
1305	F	GGT→GAT	G102D	70	100%
1321	F	GCC→GCT	A107A	70	98%
1639	F	ATG→ATT	M213I	10	100%
3320	H	GCT→GCC	A130A	30	2%
3340	H	GAT→GGT	D137G	70	100%
4613	A/A*	GAG→GAA	E211E	60	1%
4623	A/A*	GAT→CAT	D215H	30	0%

Table 5.2 - Every mutation in line A2 that was observed with a frequency of ≥ 10%. Table includes mutational frequency at the conclusion of the experiment (rounded to the nearest integer), and the passage number in which mutation was first detected at ≥ 10% frequency. Genes A and A* are in the same reading frame so shared mutations cause the same amino acid change in both genes. Amino acid change in table refers to gene

Position	Gene	Mutation	Amino acid change	First detected (passage #)	Frequency at 100 hours
572	D E	GGT→GGC GTA→GCA	G61G V2A	95	23%
686	D E	GAA→GAG AAG→AGG	E99E K40R	95	12%
794	D E	CTT→CTC TTA→TCA	L135L L76S	95	16%
1031	F	ATG→GTG	M11V	15	4%
1301	F	ACT→GCT	T101A	5	100%
1305	F	GGT→GAT	G102D	20	100%
1307	F	TAT→CAT	Y103H	10	0%
1318	F	CAT→CAG	H106Q	15	0%
1639	F	ATG→ATT	M213I	10	100%
2277	F	ACT→ATT	T426I	10	0%
3320	H	GCT→GCC	A130A	55	92%
3339	H	GAT→AAT	D137N	15	0%
3340	H	GAT→GGT	D137G	15	100%
3389	H	GAG→GAT	E153D	65	1%
3423	H	ATT→GTT	I165V	70	70%
3426	H	GCC→TCC	A166S	65	5%
3430	H	GAG→GGG	E167G	65	23%
4614	A/A*	GCG→TCG	A212S	65	27%
4817	A/A*	GAC→GAT	D279D	75	44%
4835	A/A*	CGG→CGA	R285R	95	13%
4918	A/A*	CAG→CGG	Q313R	80	36%

Table 5.3 - Every mutation in line B1 that was observed with a frequency of ≥ 10%. Table includes mutational frequency at the conclusion of the experiment (rounded to the nearest integer), and the passage number in which mutation was first detected at ≥ 10% frequency. Genes A and A* are in the same reading frame so shared mutations cause the same amino acid change in both genes. Amino acid change in table refers to gene A. For mutations in genes D and E, top mutation refers to D, bottom to E.

Position	Gene	Mutation	Amino acid change	First detected (passage #)	Frequency at 100 hours
92	A/A*	GCT→ACT	A500T	15	0%
362	C	GCT→GTT	A77V	50	0%
756	D E	TTT→CTT CGT→CGC	F123L R63R	33	0%
781	D E	AAC→AGC ACA→GCA	N131S T132A	31	0%
841	D E	GTG→GCG TGA→CGA	V151A STOP153R	50	97%
988	Intergenic	T→C		30	0%
1301	F	ACT→GCT	T101A	5	100%
1307	F	TAT→CAT	Y103H	10	0%
1319	F	GCC→ACC	A107T	25	99%
1449	F	AAT→AGT	N150S	31	69%
1968	F	AAC→GAC	N323D	20	99%
2277	F	ACT→ATT	T426I	10	0%
2710	G	GCC→ACC	A106S	70	2%
3320	H	GCT→GCC	A130A	70	81%
3339	H	GAT→AAT	D137N	20	8%
3340	H	GAT→GGT	D137G	20	100%
3423	H	ATT→GTT	I165V	75	23%
3426	H	GCC→TCC	A166S	75	70%
3581	H	GCT→GCC	T217T	31	3%
4658	A/A*	GAT→GAC	D226D	31	3%
4760	A/A*	GGC→GGT	G260G	20	0%
4802	A/A*	CCT→CCC	P274P	55	92%
4918	A/A*	CAG→CGG	Q313R	70	39%

Table 5.4 - Every mutation in line B2 that was observed with a frequency of ≥ 10%. Table includes mutational frequency at the conclusion of the experiment (rounded to the nearest integer), and the passage number in which mutation was first detected at ≥ 10% frequency. Genes A and A* are in the same reading frame so shared mutations cause the same amino

acid change in both genes. Amino acid change in table refers to gene A. For mutations in genes D and E, top mutation refers to D, bottom to E.

	Number of unique mutations detected with a frequency of		
Line	**> 0.1%**	**> 0.5%**	**> 1%**
A1	21	8	6
A2	42	8	8
B1	171	38	26
B2	119	22	16
Ratio B:A	4.60	3.75	3.00

Table 5.5 – Low frequency mutations. The number of unique mutations detected in each line after 100 passages at three different thresholds. The bottom row shows the ratio of unique mutations (unique to the line rather that the environment) in the mutagenic lines to the control lines.

5.2.2 Indels

Indels were rare in all lines. After passage 100, the average number of indels per genome (calculated from the sum of the total indels with *Q* >= 40 divided by the average read depth) was 0.068 for A1, 0.085 for A2, 0.075 for B1 and 0.074 for B2. As discussed in section 3.1.2, the impaired error checking capability of *dnaQ926* only appears to increase the frequency of substitutions, so it is expected that indels appear at similar frequencies in mutagenic and non-mutagenic lines.

In line B1, an insertion was detected at a frequency of 1.05% at position 5350 (GC --> GTTC) in passage 65. However, no reads were found containing this insertion in the sequenced passages preceding and following it. This insertion causes a frameshift of two base pairs in genes A and A*, altering three amino acids and introducing a premature stop codon.

In line B2, an insertion at position 2372 (GTT --> GTTT) was first detected in passage 31. This increased in frequency until a peak of 4.07% in passage 45, after which it began to decline in frequency, becoming undetectable from passage 60 onwards (figure 5.5). This insertion is in an intergenic region and has no effect on coding sequences or protein structure.

No indels were detected in lines A1 and A2 at a frequency of 1% or greater.

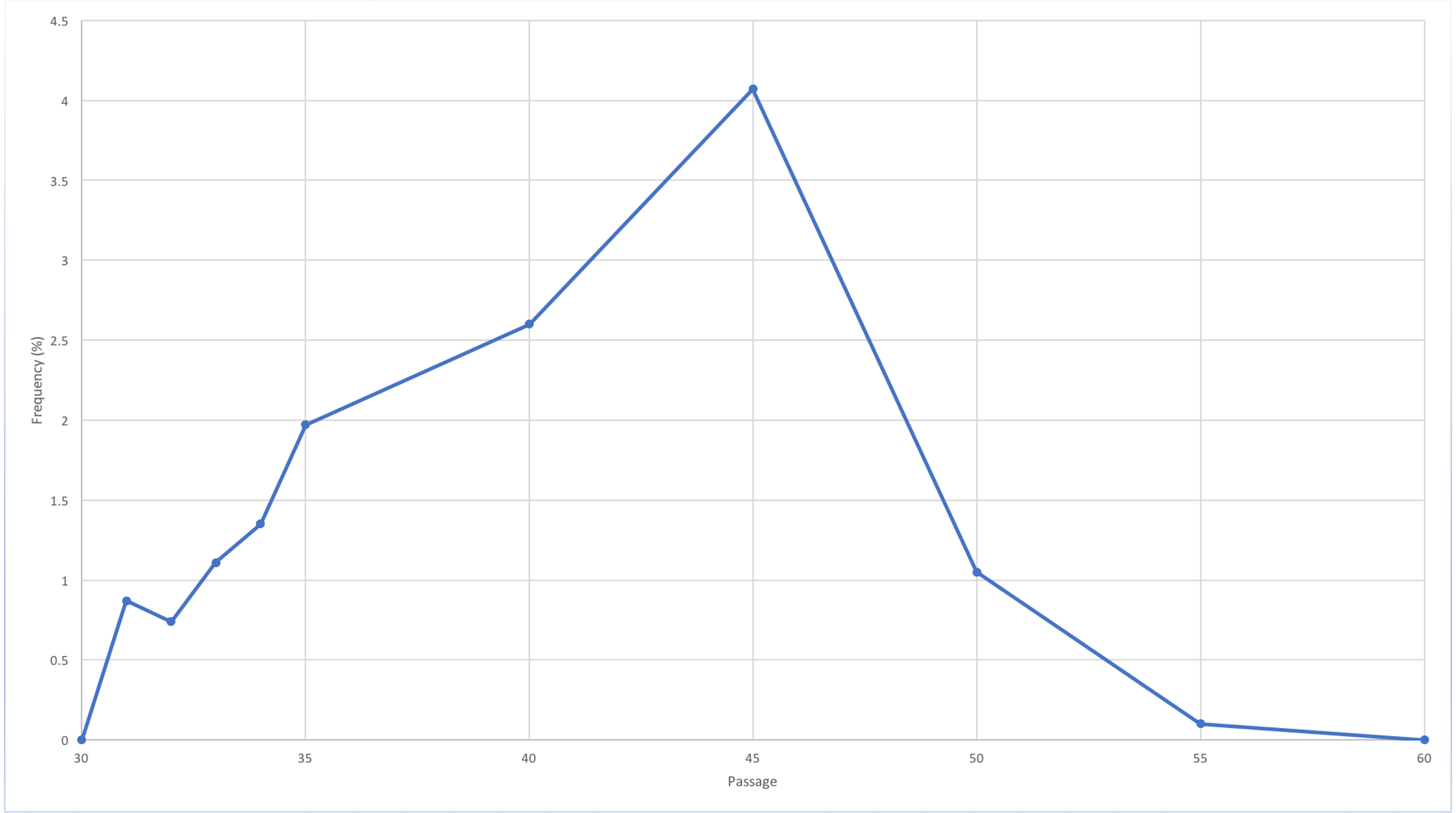


Figure 5.1 - Frequency of the insertion at position 2372 in line B2 between passages 30 and 60. This was the only indel observed in any line that persisted over multiple time points.

5.2.3 Evolutionary trajectories in lines A1 and A2

As described in chapter 4, lines A1 and A2 experienced similar changes in fitness over the course of the experiment and ended with identical genotypes other than one non-synonymous mutation each.

Both lines were characterised by sequential selective sweeps where substitutions rose to become the dominant variants in the population over a short number of passages (figures 5.6 & 5.7). All four non-synonymous substitutions that became fixed appeared in the same order in both lines, and appeared after a similar number of generations (table 5.7). These data correlate very well with the fitness measurements, giving weight to the assumption that these mutations are adaptive (figure 5.8).

Each line contained a unique synonymous mutation at high frequency. Because these mutations do not change the amino acid sequence, they are unlikely to have any significant effect on fitness and are likely to be selectively neutral. However, these synonymous mutations rose in frequency at the same rate as another mutation at position 1305. In both lines, this mutation rose in frequency from passage 60 to passage 80 in both lines. This correlates with the large gains in fitness over the same period (figure 4.1), indicating that this mutation is adaptive. Although the synonymous mutations may have had no or negligible effect on fitness, they most likely became fixed during the selective sweeps of mutation 1305 by hitchhiking on the same genome.

Although both lines were nearly identical in genotype at passage 100, over the course of the experiment there were mutations unique to both lines that rose in frequency but were ultimately lost. In line A1, mutations at positions 4622 and 4613 appeared from passage 30, and appeared to compete over the next 40 passages, reaching highs of over 90% and 40% frequency respectively, before both were ultimately lost, although there was later a slight re-emergence of 4622 (figure 5.6). Interestingly both these mutations were synonymous, yet were apparently not selectively neutral. A non-synonymous mutation at position 4627 also appeared in the population around this time, but did not increase in frequency beyond 11%, perhaps hitchhiking with a synonymous mutation. At passage 70, the adaptive mutation at 1305 was first detected and rapidly increased in frequency at the expense of these mutations.

In line A2, a synonymous mutation at position 3320 rose in frequency rapidly from passages 30 and 40 and persisted at high frequency. This mutation rapidly decreased in frequency between passages 70 and 80, which inversely correlates with an increase in frequency of the adaptive mutation at 1305 (figure 5.7). This indicates that 1305 appeared on a genotype that did not already include 3320, and hard a large selective advantage over it.

Position	A1			A2		
	Frequency		Order	Frequency		Order
	> 5%	> 90%		> 5%	> 90%	
1301	10	20	2	10	30	2
1305	70	90	4	70	90	4
1639	10	10	1	10	10	1
3340	50	60	3	70	80	3
1660	70	90	4			
1321				70	90	4

Table 5.6 - Mutations found in lines A1 and A2 at high frequency after 100 passages. Table shows the passage in which they were first detected at greater than 5% and 90% frequency, and the order in which they reached > 90% frequency.

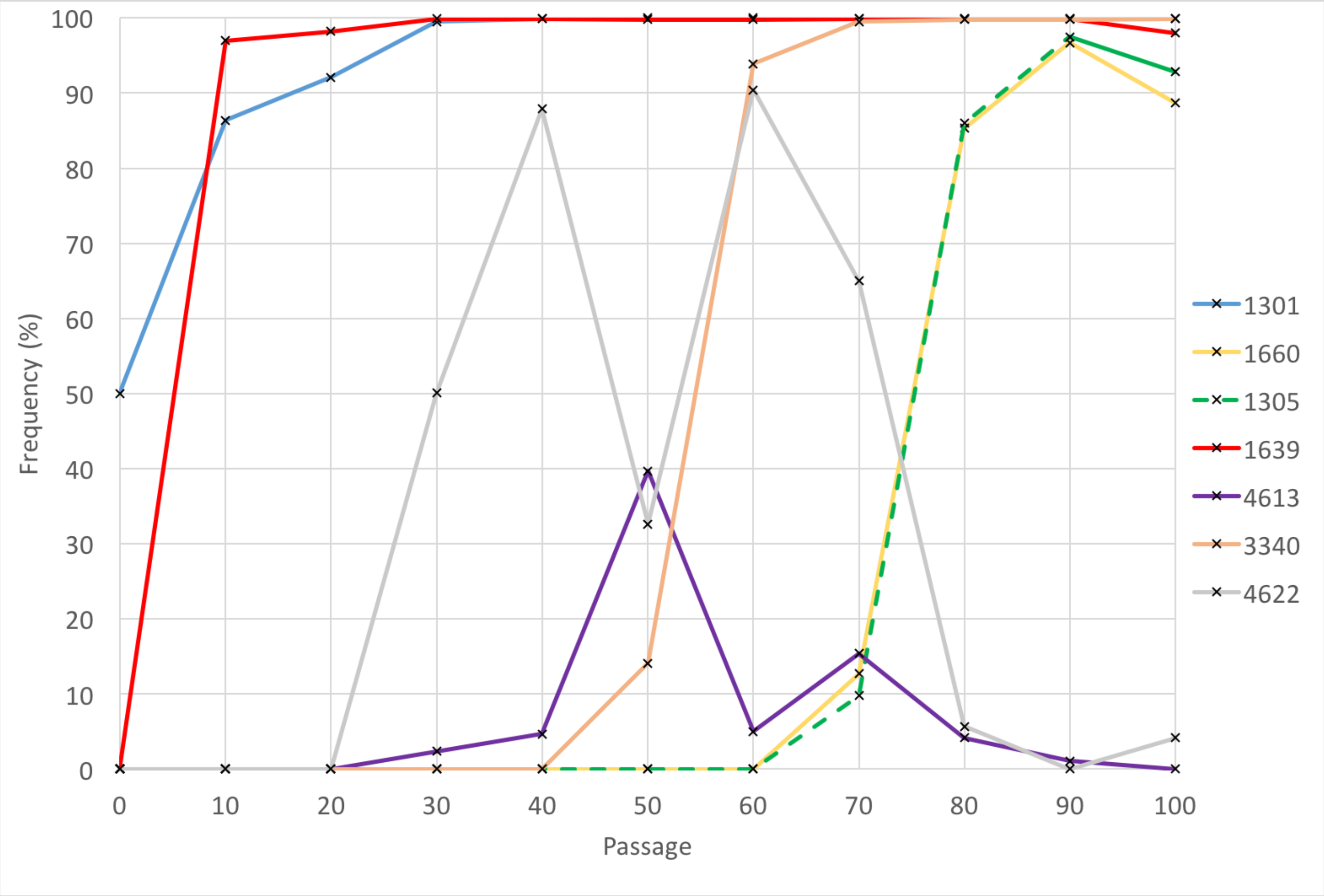


Figure 5.2 - Population dynamics of line A1 over 100 passages. All mutations that were detected at 20% frequency or higher over the course of the experiment are included. Both lines featured similar selective sweeps by mutations at 1639, 1301, 3340, and 1305. A synonymous mutation hitchhiked along with 1305. Two synonymous mutations appeared to rise in frequency after passage 20 and compete before being ultimately lost.

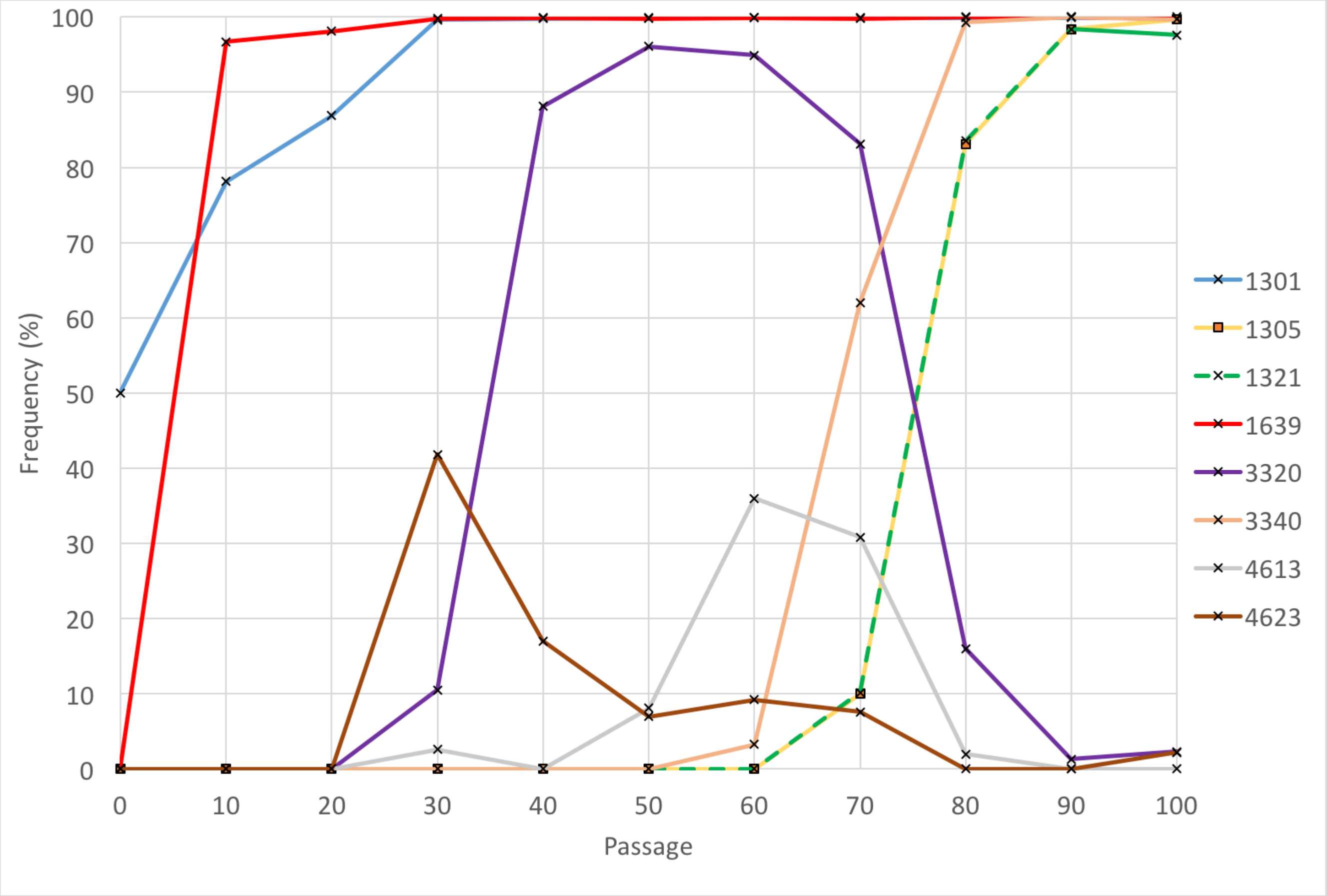


Figure 5.3 - Population dynamics of line A2 over 100 passages. All mutations that were detected at 20% frequency or higher over the course of the experiment are included. Both lines featured similar selective sweeps by mutations at 1639, 1301, 3340, and 1305. A synonymous mutation 1321 hitchhiked along with 1305. One synonymous mutation 3320 rapidly rose in frequency between passages 30 and 40, persisted for a number of passages, and then declined in frequency during the selective sweep of 3340. A non-synonymous mutation at 4623 (brown) also rose in frequency from passage 20 to 30 but did not persist.

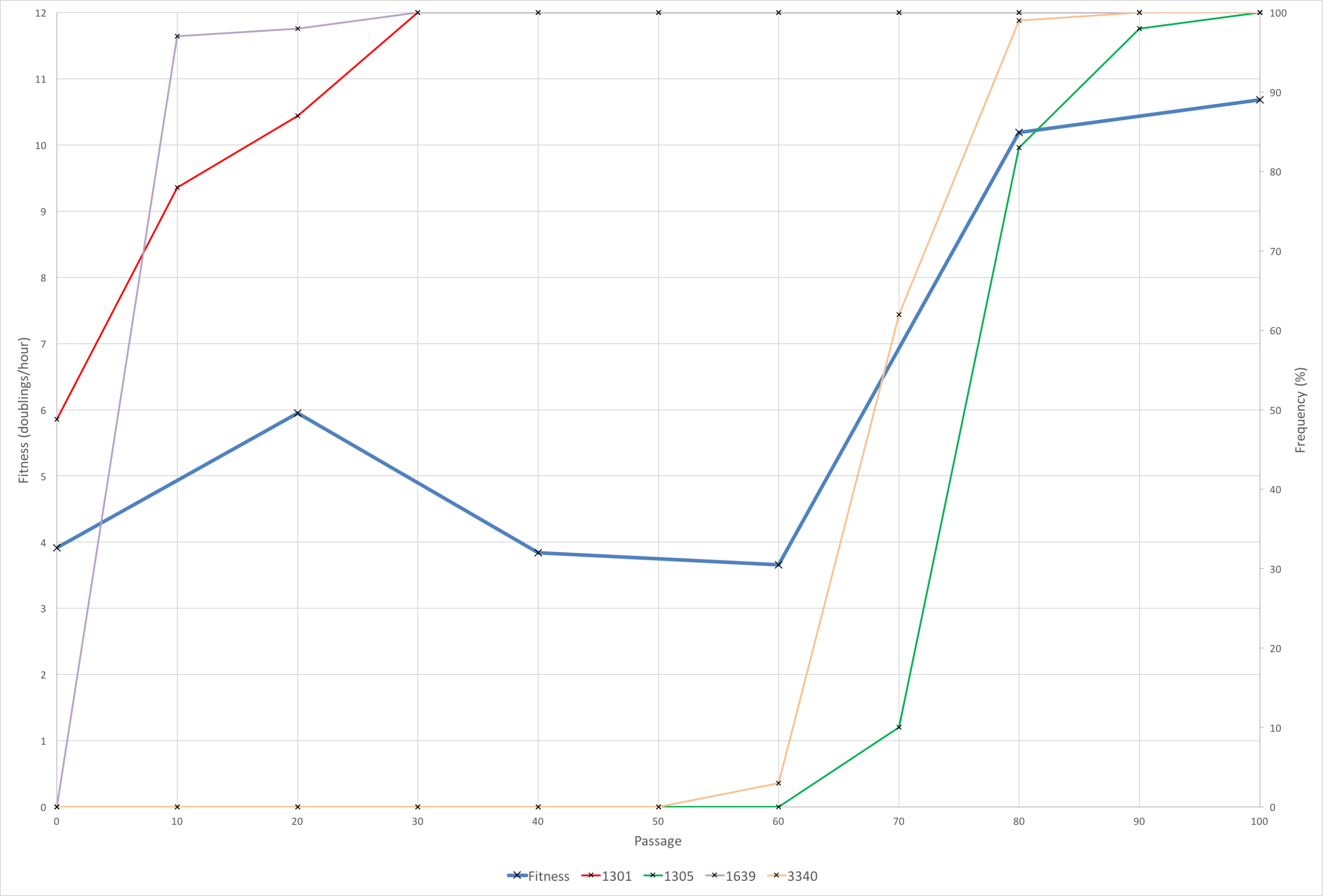


Figure 5.4 – Correlation between mutational and fitness changes in line A2. Fitness data is indicated by the blue line, while other lines represent changes in mutation frequency.

5.2.4 No mutations appeared in certain genes at high frequency

Some genes appeared to be much more tolerant of mutations than others. In the non-mutagenic lines, mutations with a frequency of 10% or greater were only observed in genes A/A*, G and F. Line B1 also contained substitutions in genes D/E (overlapping), while in line B2, substitutions were detected in genes C and G. However, these did not increase above 15% frequency and did not persist for longer than two time points.

5.2.5 Evolutionary trajectories in line B1

The evolutionary trajectories of lines A1 and A2 showed populations with little variation other than occasional selective sweeps in which beneficial mutations appeared and rose in frequency. However, the lines evolved at increased mutation rates were characterised by multiple mutations that co-occurred, many of which did not persist or become the majority allele.

During the first 30 passages of line B1, a number of mutations were present at a frequency of 10% or greater (figure 5.9). The four adaptive mutations that became fixed in lines A1 and A2 also did so in line B1. 1301, already present in the ancestral phage preparation, was the first to become fixed. However, whereas the remaining three mutations became fixed via sequential selective sweeps in the non-mutagenic lines, in B1 they increased in frequency together. It can be seen that although 1639 appeared first, mutations 3340 and 1305 appeared soon after, and almost certainly on the 1639 background. Since all these mutations have been shown to be strongly adaptive, the genotype containing all three quickly outcompeted all others and rose to 100% frequency by passage 30. This resulted in the diversity that had arisen over the first 30 passages being eliminated.

No new mutations were detected between passages 30 and 50 at a frequency of 10% or greater. From passage 55 onwards, mutations were again detected over this threshold. One of these was a synonymous mutation at 3320 which quickly rose to be the most common variant at that position. However, this mutation correlated with another large fitness increase between passages 40 and 60, despite being the only new high frequency mutation at that time (figure 5.10). This mutation also reached a frequency of over 80% in line B2, suggesting that it was in fact adaptive.

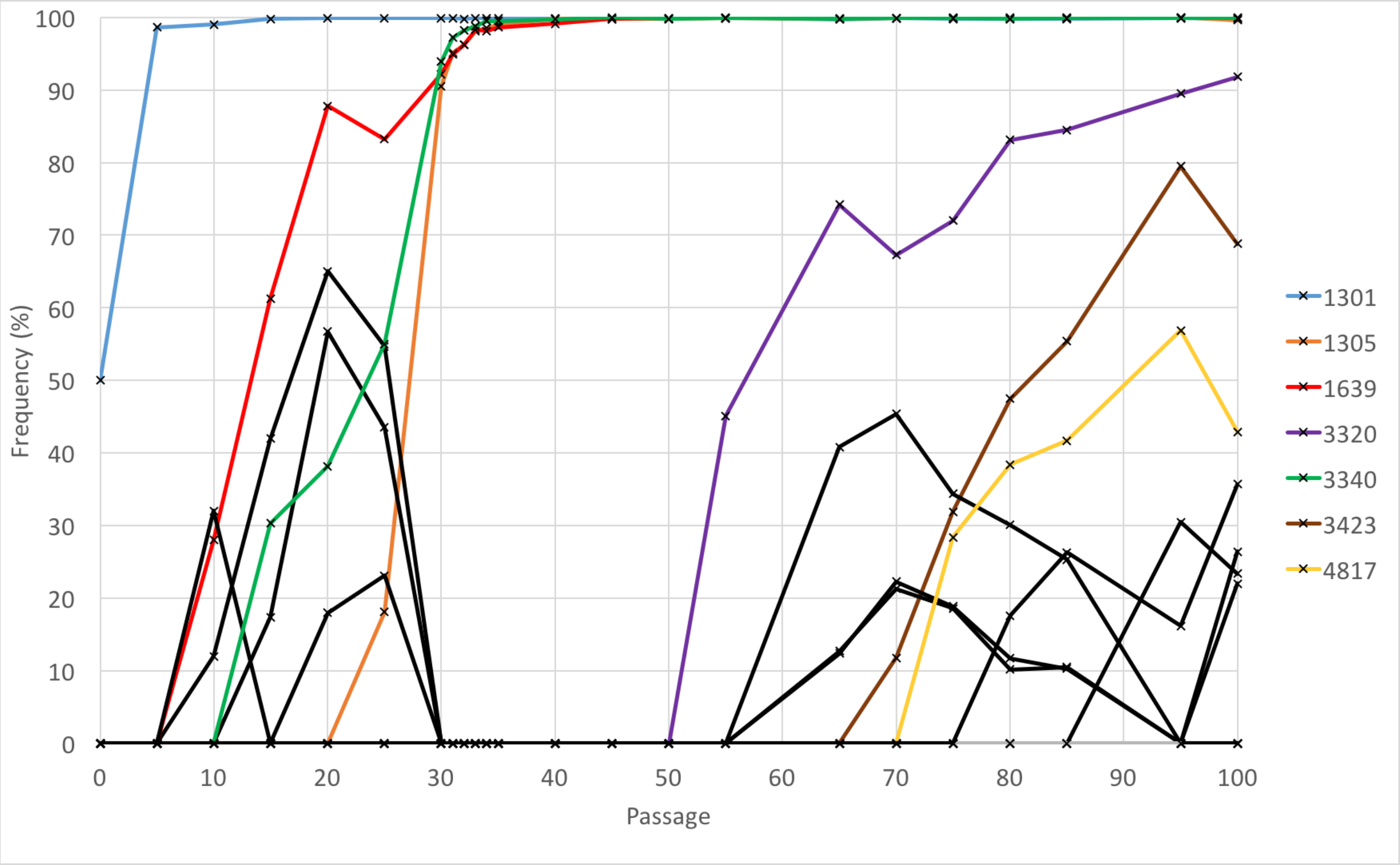


Figure 5.5 - Mutational frequencies in line B1 over 100 passages. Three adaptive mutations (1639, 3340, and 1305) that were shared with lines A1 and A2 appeared on the same genotype and became fixed by passage 30, eliminating the diversity that had previously arisen in this time. Although other mutations such as 3320 and 3423 later became the major allele, they did not rise in frequency as rapidly as the earlier adaptive mutations or those seen in lines A1 and A2. Mutations that were not discussed in detail are shown in black for clarity.

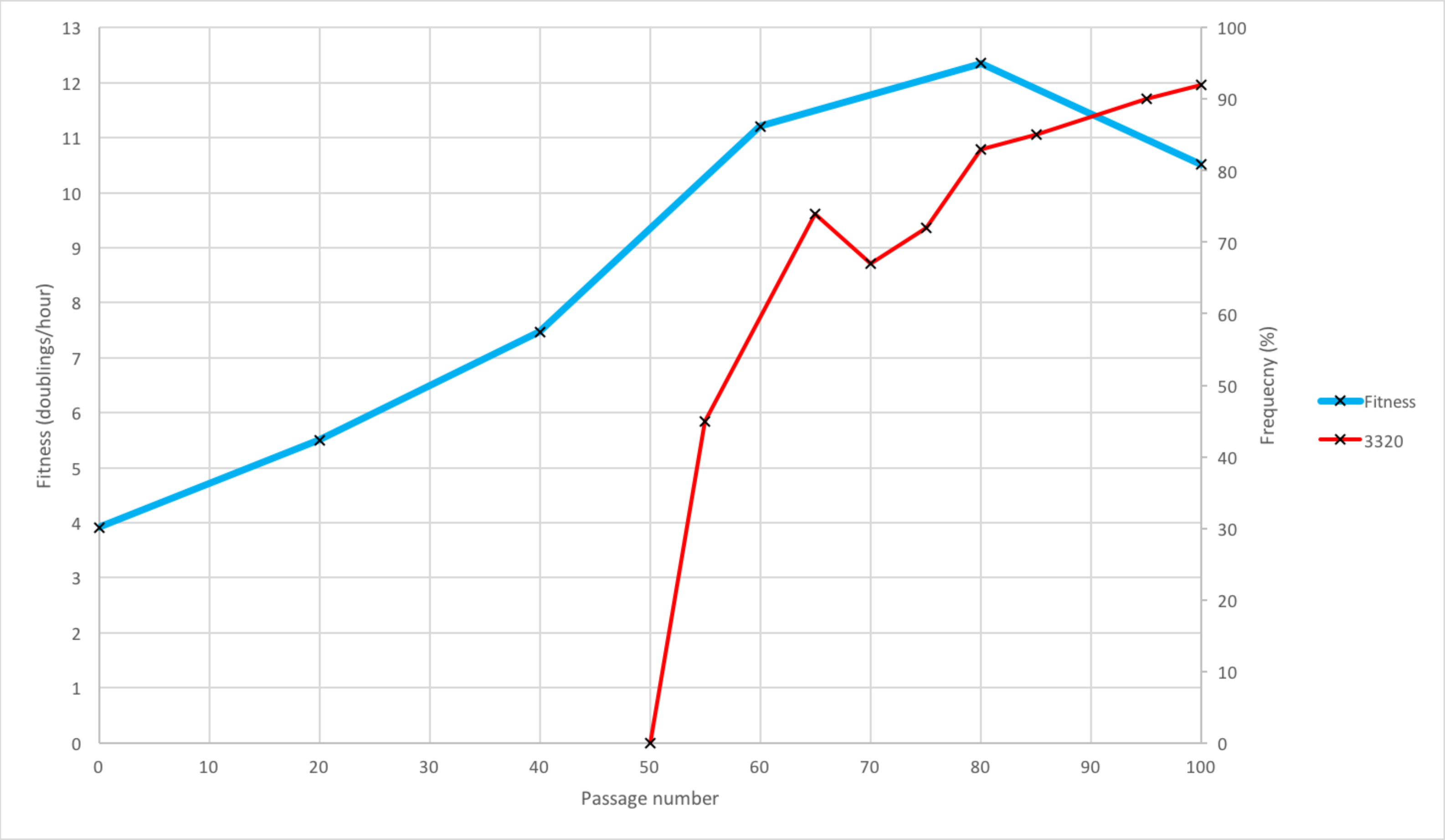


Figure 5.6 - Frequency of a silent mutation at 3320 in B1 (red) and the fitness of that line (blue). Between passages 40 and 60, this was the only mutation to significantly change in frequency, which correlates with the increase in fitness over this period.

5.2.6 Evolutionary trajectories in line B2

Line B2 acquired many high frequency mutations over the course of the experiment. B2 shared two mutations with A1, A2 and B1, but also lacked two mutations that arose in all other lines. One of the mutations that did not appear in B2, at position 1639, was the first to arise and be fixed in the non-mutagenic lines, and also arose early in B1.

As in line B1, many mutations were present simultaneously in B2. However, unlike in B1 where a single genotype quickly outcompeted others and the diversity was lost, here mutations were able to accumulate throughout the experiment.

B2 also demonstrates examples of suspected clonal interference, where mutations reach high frequencies before declining when another, presumably more favourable, mutation arises and outcompetes it. Multiple occurrences of genetic hitchhiking can also be observed (seen in figure 5.11 as lines that follow nearly identical trajectories), with two pairs of mutations that ended at over 90% frequency among these. One of these pairs consisted of a mutation at 841 that affected two overlapping genes and a synonymous mutation. The other pair contained two non-synonymous mutations in gene F at positions 1319 and 1968.

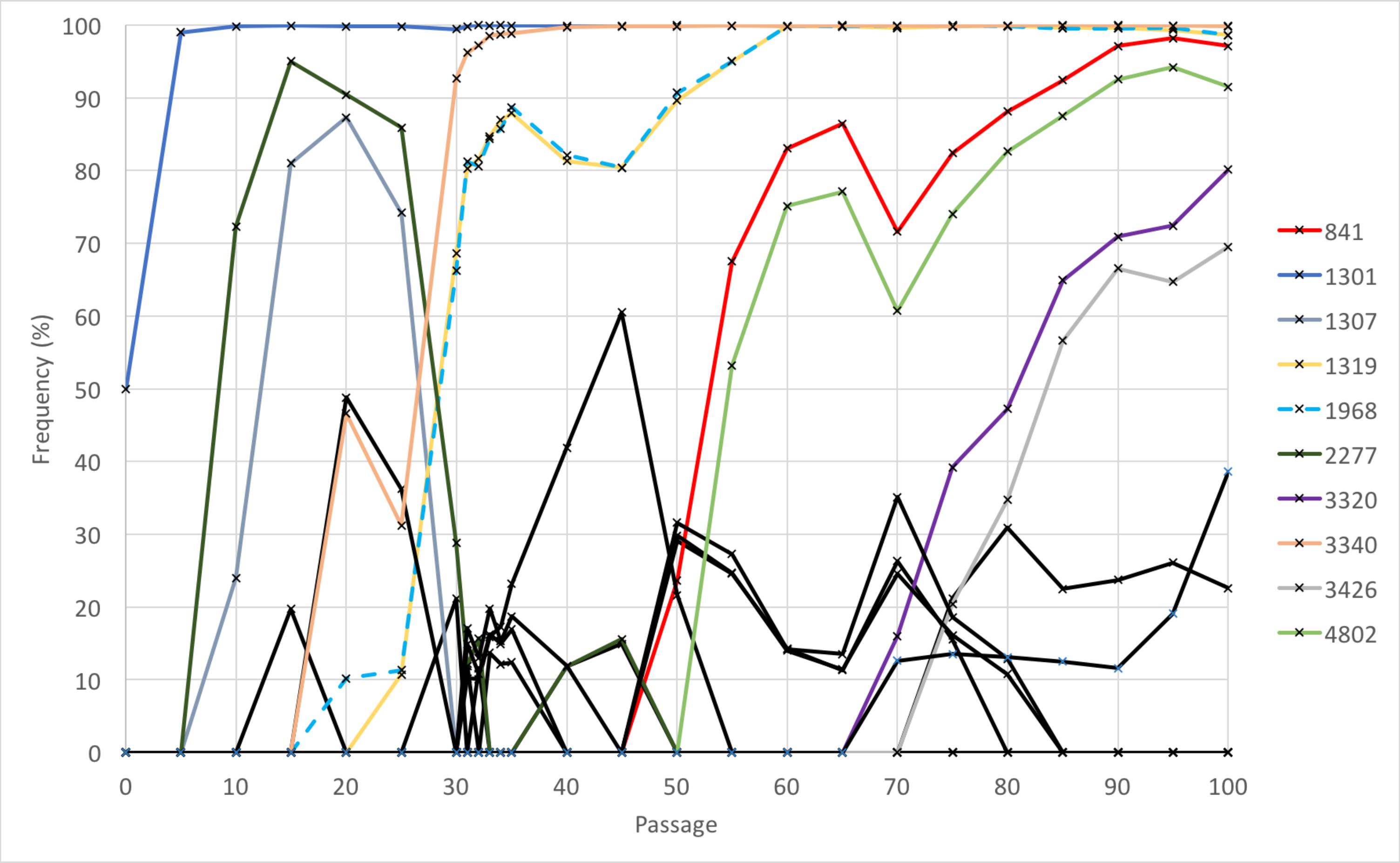


Figure 5.7 - Mutational frequencies in line B2 over 100 passages. 1301 and 3340 are the only mutations common with lines A1 and A2. This line contained many mutations not seen in the other lines, and multiple examples of genetic hitchhiking, seen when two lines have nearly identical shapes (e.g. 1319 and 1968, 841 and 4802). Mutations that were not discussed in detail are shown in black for clarity.

5.2.7 High resolution population dynamics

Although obtaining sequence data from every five passages over the experiment gave a good overview of evolutionary dynamics, it was clear that much could have happened in the short time between passages that would have remained unseen. To observe short-term evolutionary dynamics, populations from B1 and B2 were also sequenced from every passage between 30 and 35.

In line B1, this period marked the end of the selective sweep of mutations at 3340, 1639 and 1305; which eliminated most of the other mutations that had accumulated. This can be seen in figure 5.12 where these three mutations approach 100% frequency, while the frequencies of the remaining mutations simultaneous declines towards zero.

In contrast to this, so many low frequency mutations have accumulated in line B2 that they are not easily visualised on a graph (figure 5.13, figure 5.14). Most of these mutations do not appear to show any upward trend in frequency over time, suggesting that they are unlikely to be beneficial.

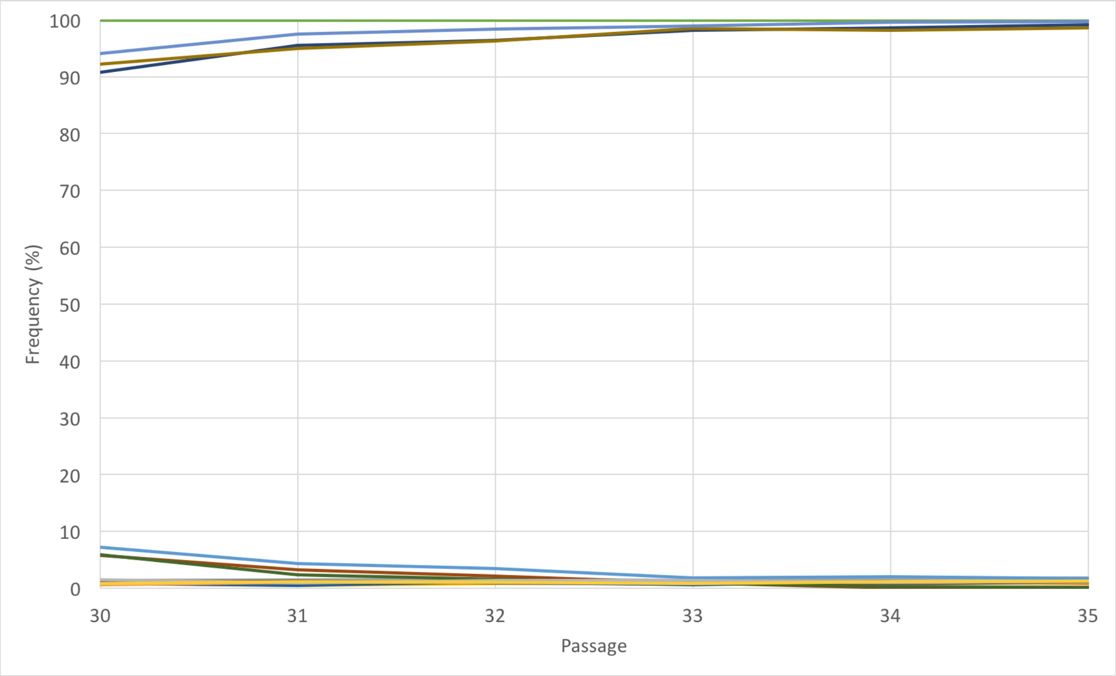


Figure 5.8 - mutation frequencies in line B1 between passages 30 and 35. Every mutation that reached a frequency of at least 1% during this time is included. This figure is only intended to show the diversity of mutations at low frequency, so individual mutations are not colour-coded or labelled. Data points have been removed for clarity.

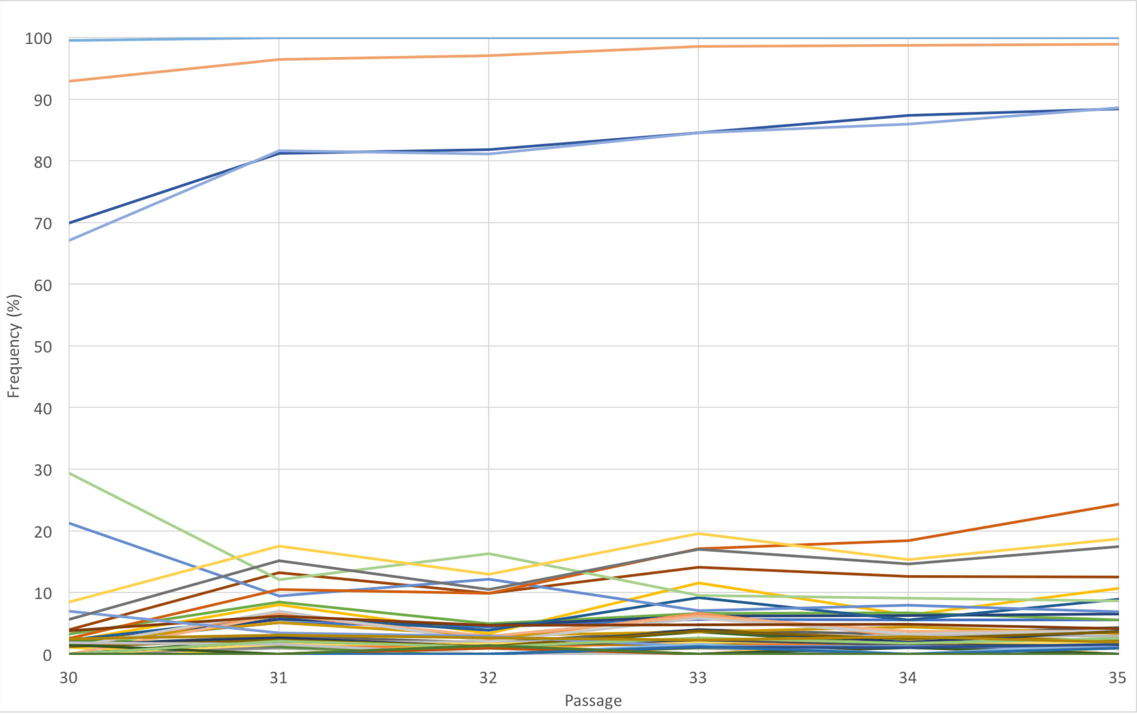


Figure 5.9 - mutation frequencies in line B2 between passages 30 and 35. Every mutation that reached a frequency of at least 1% during this time is included. This figure is only intended to show the diversity of mutations at low frequency, so individual mutations are not colour-coded or labelled. Data points have been removed for clarity.

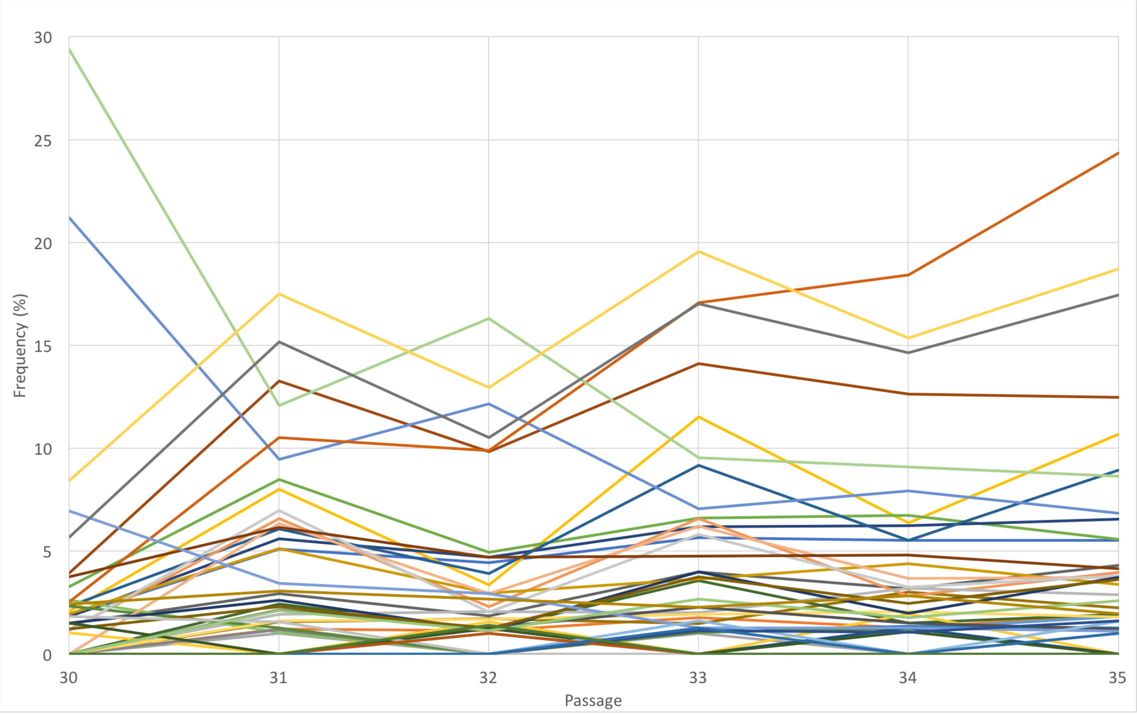


Figure 5.10 - mutation frequencies in line B2 between passages 30 and 35 with higher frequency mutations (> 30%) excluded. Every mutation that reached a frequency of at least 1% during this time is included. This figure is only intended to show the diversity of mutations at low frequency, so individual mutations are not colour-coded or labelled. Data points have been removed for clarity.

5.2.8 Competing mutations within a single codon

In both mutagenic lines, (between passages 15-25 in B1 and 20-25 in B2) mutations appear at positions 3339 and 3340. Both of these mutations are within a single codon, but result in a different amino acid; if both are present together the triplet codes for another amino acid still. Although the short read lengths of NGS mean that it is usually only possible to detect mutational frequencies in a population rather than the sequences of individual genotypes, the proximity of these two nucleotides means that they appear on the same sequencing reads. SAMtools was used to filter sequence data from these passages for reads containing both these positions, and Python was used to trim reads to just these two positions. Only four reads were found where these mutations were present together. In addition, the combined frequency of the two mutations did not exceed 99%. This means the two mutations were most likely both beneficial and arose at similar times, but had to compete against each other. On both occasions, 3340 became fixed at the expense of 3339, indicating that it had the larger selection coefficient.

In the non-mutagenic lines, 3339 and 3340 did not compete in this way, with 3340 becoming fixed in both. Since these mutations act on the same codon, they are not additive. At normal mutation rates, when the likelihood of both mutations occurring in a population at the same time is low, which mutation becomes fixed may be mainly down to luck. If 3339 had appeared first and became fixed before 3340 appeared in the population, 3340 would probably be lost even if its selection coefficient was far higher than 3339. This is because it would require a reversion of 3339 since the presence of both together would change the codon to a serine.

This demonstrates an advantage of high mutation rates, because they minimise the possibility of a beneficial mutation being excluded because a less beneficial mutation has got lucky and appeared first.

Consider an ancestral gene with a selection coefficient of 1; two mutations in that gene, *a* and *b*, with selection coefficients of 1.5 and 2 respectively; and the state *ab* where both mutations are present together with selection coefficient 0.5. In a low mutagenic environment, it is unlikely the two mutations would appear at the same time, in which case the mutation that appears first will become fixed. If this is *a*, it is almost impossible for *b* to become fixed, despite having a greater beneficial effect on fitness. As long as the selection coefficient of *ab* is lower than *a*, the *ab* state would be selected against meaning that *b* would need to appear alongside a reversion of *a* in order to be selected for.

In a high mutagenic environment, the likelihood of both *a* and *b* appearing in a population together is greater, meaning *b* would be more likely to outcompete *a* and become fixed. Even in a situation where *a* became fixed, the probability of a reversion in *a* occurring at the same time as mutation *b* would be higher in this environment, meaning the *b* state may not be permanently lost.

5.3 Discussion

5.3.1 Elevated mutation rate facilitates adaptation in line B1

Mutations 1305, 1639 and 3340 arose independently in three lines of this experiment. They were deemed to be adaptive due to this parallel evolution and their correlation with the observed fitness increases. In the non-mutagenic lines these appeared and underwent selective sweeps one by one, with both lines taking 90 passages (~270 generations) for all three to be present at over 95% frequency. In B1, however, all three substitutions reached this frequency at passage 31 (~93 generations). 1639 was the first to be detected at passage 10, followed by 3340 at passage 15 and 1305 at passage 25. All these mutations increased in frequency together, suggesting that they were not competing with each other as happens in clonal interference, but appeared on the same genome, providing a net increase in fitness each time. The most plausible explanation is that 1639 appeared first and while it was increasing in frequency, 3340 appeared on this background. This would have increased fitness further, so the 1639/3340 genotype would have started outcompeting the 1639 alone genotype. Soon after, 1305 appeared on a 1639/3340 genome, and quickly rose to become the only genotype in the population. This demonstrates one consequence of evolution at high mutation rates - beneficial mutations can occur more frequently, leading to faster adaptation than is possible under normal conditions.

Later in the experiment, mutations continued to arise in B1 that were not observed in lines A1 and A2. Because fitness continued to rise until passage

80, it is probable that these mutations were also adaptive, and had simply not had enough time to arise in the non-mutagenic lines.

5.3.2 A different evolutionary trajectory in line B2

While line B1 had many similarities to the non-mutagenic lines A1 and A2, the fate of B2 was very different. In every other line, a mutation at 1639 appeared and quickly became fixed within the populations. Because this mutation was the first to arise in both non-mutagenic lines (excluding 1301 which arose in the ancestor), increased in frequency rapidly (> 95% after 10 passages in A1 and A2), and correlates with a fitness increase in every line, it is likely to be one of the most beneficial mutations available. In line B2 this mutation had over 2% frequency after passage 10, but did not exceed 1% frequency for the remainder of the experiment.

As discussed in section 4.4.4, the failure of this mutation to persist could be attributed to negative epistasis, with it no longer being beneficial on the new genetic background that was a consequence of the elevated mutation rate. This is something that should be considered in any future models that try to account for adaptive evolution. Just because a genotype with very high fitness exists, there is no guarantee that a line evolved at high mutation rate will achieve it.

5.3.3 Lethal mutagenesis models and adaptive evolution

One major drawback of the models for describing evolution at a high mutation is ongoing adaptive evolution. These models assume starting fitness is optimal and populations can only decline in fitness. As the results presented here and in Springman *et al* (2010) show, this is untrue and adaptive evolution can still take place at high mutation rates.

Assuming a constant environment, the number of available beneficial mutations and their magnitude should both decrease as evolution proceeds. Once the potential for further adaptation is minimised, it would be expected that deleterious mutations would continue to accumulate and fitness would decline until either an equilibrium was reached or the population went extinct.

It is interesting to compare fitness in line B1 with the non-mutagenic lines. After 40 passages, the same non-synonymous mutations were fixed in B1 as in A1 and A2 after 100 passages. However figure 4.1 shows that A1 and A2 have a much higher fitness after 100 passages than B1 does after 40. Despite acquiring the adaptive mutations in approximately a third of the time taken by the lines evolved under non-mutagenic conditions, B1 failed to increase in fitness to the same extent. This discrepancy is best explained by the lower frequency mutations that were present in B1 as a result of the elevated mutation rate but that A1 and A2 lacked. This demonstrates that while mutagenic conditions can cause adaptation to occur more rapidly, it also results in a genetic background of low frequency deleterious mutations that impair the average fitness of the population.

Given more time, it would be interesting to compare B1 and B2. The models predict that fitness will eventually stop declining and equilibrate. This equilibrium fitness is calculated using the fitness of the optimal genotype. Since adaptive evolution is not considered, the optimal genotype is assumed to be that of the mutation-free wild-type. However, in line B1 fitness far exceeded that of the wild-type by a factor of nearly three. This new fitness optimum must be used to recalculate any subsequent fitness decline and equilibrium, which would increase over that calculated at the beginning. Meanwhile, line B2 did not exceed the wild-type's fitness at any point of the experiment. Despite being initiated from the same ancestor, B1 and B2 would now be expected to reach very different fitness equilibriums if the experiment was continued.

5.3.4 Rarity of indels

Indels appeared to be highly selected against. Only a single indel persisted at a frequency of 1% or greater over all lines, and this was in an intergenic region. This is in keeping with the literature, for example in Wichman’s 13,000 generation experiment (2005), only 4 indels were detected (compared to 137 substitutions), all of which were located in intergenic regions.

One possibility for this is that with 95% of the genome consisting of protein coding regions, any indels that did occur were likely to result in frameshift mutations. This would have been further exacerbated because ΦX174 contains regions were genes overlap but are in different reading frames, meaning multiple genes could be shifted out of frame by a single indel.

The physical structure of the ΦX174 genome in relation to the capsid should also be considered. In the final virion, 60 copies of protein J bind to the genome, anchoring it in place in the capsid (Hafenstein and Fane, 2002). Since the genome is small, a change in size from the insertion or deletion of nucleotides could potentially alter the configuration of the genome or the secondary structure of the ssDNA, affecting the interactions with the DNA binding protein and the viral capsid. If this is indeed the case, indels that hinder viral assembly would be selected against even if they did not affect protein structure.

5.3.5 An adaptive synonymous mutation

In line B1, a synonymous mutation at position 3320 was detected in passage 50 with a frequency of 10%. In subsequent passages, frequency rose quickly, reaching 46% by passage 55 and 74% by passage 65. No other mutations showed a similar change in frequency over this period, meaning that the rise in frequency was not a result of genetic hitchhiking. This increase in frequency also correlated with a rise in fitness that was observed between passages 40 and 60.

Unfortunately, sequence data was unavailable for passage 60 in this line. However, this was the only mutation to increase in frequency by over 10% between passages 40 and 55, suggesting it may play a role in this increase in fitness. Further evidence is provided by line B2, where this mutation arose independently and reached a frequency of 81% by the experiment's conclusion. It also reached a frequency of over 96% in line A2, but did not persist, presumably outcompeted by 1305 (figure 5.7).

Although synonymous mutations are usually assumed to be neutral, there is evidence that this is not always the case. Changing a codon to corresponding with a more or less common tRNA can affect the speed and accuracy of transcription and translation (Plotkin and Kudla, 2011), which in turn can affect co-translational folding or gene expression (Agashe *et al*., 2013). In a recent evolution experiment with *Psuedomonas fluorescens*, two synonymous mutations were detected that conferred a fitness increase similar to that of non-synonymous mutations, which they determined to be due to elevated gene expression (Bailey *et al*., 2014). There is also evidence that some other synonymous mutations in ΦX174 may be adaptive (Wichman *et al*., 2005).

The mutation detected caused GCT to change to GCC, both of which encode alanine. Data is unavailable for *E. coli* C, but in *E. coli* K12 the GCC codon is nearly thrice as abundant as GCT (Nakamura *et al.,* 2000). If this also applies to the host used here, then it is possible that the switch to a more abundant tRNA leads to an increase in expression.

This mutation is located in gene H, the DNA pilot protein. Although the tail that extends from the phage is made up of 10 copies of this protein (Sun *et al*., 2014), a ΦX174 virion contains between 10-12 copies (Burgess, 1969). While it is unclear why some virions contain extra copies of this protein and what determines this, it may be that increased expression results in more virions containing 12 copies, with these extra copies facilitating tail formation.

5.3.6 A shift to quasispecies?

The mutation rates of DNA microbes fall around an average of 0.0034 substitutions per genome per replication (sgr), (Drake *et al*., 1998), with wild-type ΦX174 being consistent with this (Cuevas *et al*., 2009). Mutation rates of RNA viruses, meanwhile, fall around an average of 0.76 sgr, approximately two orders of magnitude higher. The mutation rate of ΦX174 in mutagenic conditions was measured in chapter 3 as 0.97 sgr, which is indistinguishable from many ribovirusues.

As is most evident in figure 5.14, many low frequency mutations were present simultaneously in the population during evolution at a high mutation rate. This is characteristic of a viral quasispecies, where instead of one dominant genotype a population consists of many closely related genotypes. This is often used to describe RNA viruses, but it appears that DNA viruses may behave in a similar manner when their mutation rate is artificially elevated to the same level. In future, it would be interesting to see if the broad genetic background in line B1 facilitates adaptation to harsher conditions compared to a homogenous population grown from a single isolate.

Chapter 6: Conclusions

6.1 Summary

Mutation is the ultimate source of the genetic variation required for evolution, yet deleterious mutations vastly outnumber beneficial mutations. To compensate for this, mechanisms have evolved in all domains of life that prevent and undo mutations, lowering the mutation rate. The mutation rate appears to be a trade-off between the fitness cost of deleterious mutations, the fitness advantage of beneficial mutations, and the cost of maintaining the molecular mechanisms. The consequences of evolution at a high mutation rate are unclear. Evolutionary theory and models predict that as the mutation rate increases a population will accumulate deleterious mutations over time and decrease in fitness (Agrawal and Whitlock, 2012) (Bull et al., 2007). However, these do not account for every variable and are contradicted by empirical data (Springman *et al*., 2010). In this thesis I have used experimental evolution to investigate how populations of bacteriophage ΦX174 were affected by an elevated mutation rate.

To investigate the evolutionary consequences of an elevated mutation rate, I started **chapter three** by investigating possible ways to increase the mutation rate of bacteriophage ΦX174. My main criteria for selecting a suitable method was that mutational spectrum bias was limited, and the relative frequencies of mutations would be as close to those observed under non-mutagenic conditions as possible. Additionally, I wanted to minimise any potential non-mutagenic effects that could interfere with the host or phage during the experiment. I reviewed different chemical and physiological mechanisms that could achieve this, and selected a mutant host error checking gene as the most suitable. I created a strain of the host bacteria containing this mutant gene on a plasmid. Using an amber reversion fluctuation test, I measured the mutation rate of ΦX174 infecting this strain and a control, and found that it was two orders of magnitude higher than wild-type.

In **chapter four** I used this system to investigate how an elevated mutation rate affected bacteriophage fitness. I evolved replicate lines of ΦX174 at normal and elevated mutation rates for approximately 300 generations. In line with previous work with this organism (Bull *et al.*, 1997), the lineages of phage evolved at wild-type mutation rate underwent almost identical changes in fitness over the course of the experiment, ending with nearly identical genotypes. However, the lines grown under mutagenic conditions evolved in very different ways. Line B1 increased in fitness at a faster rate than the non-mutagenic lines, while also sharing many of the same mutations. Line B2 had a decline in fitness before reaching an equilibrium, while also acquiring many novel mutations that were not seen in the other three lines. Lethal mutagenesis was not achieved in either of the populations.

In **chapter five** I investigated evolutionary dynamics of the four lines by sequencing populations from multiple time points over the course of the experiment. The small genome of ΦX174 allowed for very high coverage, meaning the frequencies of individual mutations in the population could be determined. It was found that the rapid adaptation in line B1 was caused by three adaptive mutations that occurred at a similar time and increased in frequency together. Meanwhile, many of the mutations in B2 were unique to that line, while it failed to acquire some of the adaptive mutations shared by all other populations. Since the mutation rate was high enough that all possible mutations had undoubtedly occurred, it is likely that the adaptive mutations in other lines were not beneficial on this line's altered genetic background

6.2 Lethal mutagenesis

This work shows that while increasing mutation rate can result in a fitness decline, it can also facilitate adaptation. Since current models do not account for adaptive evolution, further research on the evolutionary dynamics of lethal mutagenesis is required.

These should be important considerations when looking at lethal mutagenesis as an antiviral strategy. If being used therapeutically, a fitness decline would not be a success unless it resulted in complete viral extinction. If a virus is mutagenised *in vivo* but not eliminated, the remaining viruses will possess a wide genetic background, which may allow a greater spectrum of beneficial mutations than were available before treatment. It is possible therefore that unsuccessful lethal mutagenesis may result in making the infection worse or harder to treat.

It is probable that a substantially higher mutation rate than the one used here would be needed to induce lethal mutagenesis in ΦX174, but it is unclear how much higher it would need to be. Given that common mutagens such as MNNG increase mutation rate in phage by a similar amount as the *dnaQ* method used here (Springman *et al.*, 2010), it seems unlikely that the larger increase required would be possible *in vivo*. This suggests that lethal mutagenesis is not a feasible antiviral strategy for DNA viruses. However, RNA viruses which have higher mutation rates by orders of magnitude and exist closer to the threshold may still be able to be treated in this way.

6.3 Reproducibility

In his book *Wonderful Life* (1989), Stephen Jay Gould hypothesised that if we were to "rewind the tape of life", evolution would play out very differently and we may end up seeing Earth populated by descendants of the extinct Cambrian fauna, unrecognisable from life as we know it. Richard Dawkins, meanwhile, said that the abundant evidence of convergent evolution in nature, such as the eye evolving independently over 40 times, means that similar traits would likely appear again (Dawkins, 2004).

The work presented in this thesis somewhat agrees with both viewpoints. The convergent evolution in lines A1, A2, and B1 is strong, with four major adaptive mutations shared by the three lines. This is probably facilitated by the small genome of ΦX174, which limits the number of potential pathways to adaptation. The drastically different outcome in line B2, however, shows that evolution is not always repeatable, and that a high mutation rate exacerbates the differences, changing the genetic background and locking out adaptive pathways.

If nothing else, this work demonstrates the importance of sufficient replication of evolution experiments. For example, in similar experiments to this, it will be important to determine the difference between lethal mutagenesis being achievable and lethal mutagenesis being inevitable.

References

Agashe, D., Martinez-Gomez, N.C., Drummond, D.A., *et al*. (2013) Good Codons, Bad Transcript: Large Reductions in Gene Expression and Fitness Arising from Synonymous Mutations in a Key Enzyme. Molecular Biology and Evolution, 30 (3): 549–560

Agrawal, A.F. and Whitlock, M.C. (2012) Mutation Load: The Fitness of Individuals in Populations Where Deleterious Alleles Are Abundant. Annual Review of Ecology, Evolution, and Systematics, 43 (1): 115–135

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

Aoyama, A. and Hayashi, M. (1986) Synthesis of bacteriophage φX174 in vitro: Mechanism of switch from DNA replication to DNA packaging. Cell, 47 (1): 99–106

Arribas, M., Cabanillas, L., Kubota, K., *et al*. (2016) Impact of increased mutagenesis on adaptation to high temperature in bacteriophage Qβ. Virology, 497 (C): 163–170

Atwood, K.C., Schneider, L.K. and Ryan, F.J. (1951) Periodic selection in Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America, 37 (3): 146–155

Bailey, S.F., Hinz, A. and Kassen, R. (2014) Adaptive synonymous mutations in an experimentally evolved Pseudomonas fluorescens population. Nature communications, 5: 4076

Barton, N.H. (2010) Genetic linkage and natural selection. Philosophical Transactions of the Royal Society B: Biological Sciences, 365 (1552): 2559–2569

Belyaev, D.K. (n.d.) Bulletin of the Moscow Society of Naturalists Biological Series

Bernhardt, T.G., Struck, D.K. and Young, R. (2001) The lysis protein E of phi X174 is a specific inhibitor of the MraY-catalyzed step in peptidoglycan synthesis. The Journal of biological chemistry, 276 (9): 6093–6097

Booth, J.A., Thomassen, G.O.S., Rowe, A.D., *et al*. (2013) Tiling array study of MNNG treated Escherichia coli reveals a widespread transcriptional response. Scientific reports, 3: 1–10

Buckling, A., Craig Maclean, R., Brockhurst, M.A., *et al.* (2009) The Beagle in a bottle. Nature, 457 (7231): 824–829

Bull, J.J., Badgett, M.R. and Wichman, H.A. (1997) Exceptional convergent evolution in a virus. …

Bull, J.J., Badgett, M.R. and Wichman, H.A. (2000) Big-Benefit Mutations in a Bacteriophage Inhibited with Heat. Molecular Biology and Evolution, 17 (6): 942–950

Bull, J.J., Joyce, P., Gladstone, E., *et al.* (2013) Empirical complexities in the genetic foundations of lethal mutagenesis. Genetics, 195 (2): 541–552

Bull, J.J., Sanjuan, R. and Wilke, C.O. (2007) Theory of lethal mutagenesis for viruses. Journal of Virology, 81 (6): 2930–2939

Burgess, A.B. (1969) Studies on the proteins of phi X174. II. The protein composition of the phi X coat. Proceedings of the National Academy of Sciences of the United States of America, 64 (2): 613–617

Burns, P.A., Gordon, A. and Glickman, B.W. (1987) Influence of neighbouring base sequence on N-methyl-N′-nitro-N-nitrosoguanidine mutagenesis in the lacI gene of Escherichia coli. Journal of Molecular Biology, 194 (3): 385–390

Chen, J.Q., Wu, Y., Yang, H., *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

Cisek, A.A., Dąbrowska, I., Gregorczyk, K.P., *et al.* (2016) Phage Therapy in Bacterial Infections Treatment: One Hundred Years After the Discovery of Bacteriophages. Current Microbiology, 74 (2): 277–283

Colasanti, J. and Denhardt, D.T. (1987) Mechanism of replication of bacteriophage ϕX174 XXII. Site-specific mutagenesis of the A∗ gene reveals that A∗ protein is not essential for ϕX174 DNA replication. Journal of Molecular Biology, 197 (1): 47–54

Coulondre, C., Miller, J.H., Farabaugh, P.J., and Gilbert, W. (1978) Molecular basis of base substitution hotspots in Escherichia coli. Nature. 274(5673): 775-780

Crill, W.D., Wichman, H.A. and Bull, J.J. (2000) Evolutionary reversals during viral adaptation to alternating hosts. Genetics

Cuevas, J.M., Duffy, S. and Sanjuán, R. (2009) Point mutation rate of bacteriophage PhiX174. Genetics, 183 (2): 747–749

Cuevas, J.M., Pereira-Gómez, M. and Sanjuán, R. (2011) Mutation rate of bacteriophage ΦX174 modified through changes in GATC sequence context. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases, 11 (7): 1820–1822

Dallinger, R. (1888) Meeting of 14th December, 1887, At King's College, Stand, WC, The President (The Rev. Dr. Dallinger, FRS) in the Chair

Darwin, C. (1859) On the Origin of Species By Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. London: John Murray

Dawkins, R. (2004) The Ancestor's Tale: A Pilgrimage to the Dawn of Life. London: Weidenfeld & Nicholson

Dickins, B. and Nekrutenko, A. (2009) High-resolution mapping of evolutionary trajectories in a phage. Genome biology and evolution, 1: 294–307

Dokland, T., Bernal, R.A., Burch, A., *et al.* (1999) The role of scaffolding proteins in the assembly of the small, single-stranded DNA virus ϕX174. Journal of molecular …, 288 (4): 595–608

Domingo, E. and Perales, C. (2016) Viral Quasispecies and Lethal Mutagenesis. European Review, 24 (01): 39–48

Domingo-Calap, P., Cuevas, J.M. and Sanjuán, R. (2009) The fitness effects of random mutations in single-stranded DNA and RNA bacteriophages. Begun, D.J. (ed.). PLoS Genetics, 5 (11): e1000742

Draghi, J.A., Parsons, T.L., Wagner, G.P., *et al.* (2010) Mutational robustness can facilitate adaptation. Nature, 463 (7279): 353–355

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

Drake, J.W. (2009) Avoiding Dangerous Missense: Thermophiles Display Especially Low Mutation Rates Casadesús, J. (ed.). PLoS Genetics, 5 (6): e1000520

Drake, J.W. and Holland, J.J. (1999) Mutation rates among RNA viruses. Proceedings of the National Academy of Sciences of the United States of America, 96 (24): 13910–13913

Drake, J.W., Charlesworth, B., Charlesworth, D., *et al.* (1998) Rates of spontaneous mutation. Genetics, 148 (4): 1667–1686

Duffy, S., Shackelton, L.A. and Holmes, E.C. (2008) Rates of evolutionary change in viruses: patterns and determinants. Nature reviews. Genetics, 9 (4): 267–276

Echols, H. and Goodman, M.F. (1991) Fidelity mechanisms in DNA replication. Annual review of biochemistry, 60 (1): 477–511

Eigen, M. (1971) Self-organization of matter and the evolution of biological macromolecules. Naturwissenschaften, 58(10):465-523.

Eisenberg, S., Griffith, J. and Kornberg, A. (1977) phiX174 cistron A protein is a multifunctional enzyme in DNA replication. Proceedings of the National Academy of Sciences of the United States of America, 74 (8): 3198–3202

Ekechukwu, M.C. and Fane, B.A. (1995) Characterization of the morphogenetic defects conferred by cold-sensitive prohead accessory and scaffolding proteins of phi X174. Journal of Bacteriology, 177 (3): 829–830

Engelstädter, J. (2008) Muller's ratchet and the degeneration of Y chromosomes: a simulation study. Genetics, 180 (2): 957–967

Eyre-Walker, A. and Keightley, P.D. (2007) The distribution of fitness effects of new mutations. Nature reviews. Genetics, 8 (8): 610–618

Fane, B.A. and Hayashi, M. (1991) Second-site suppressors of a cold-sensitive prohead accessory protein of bacteriophage phi X174. Genetics, 128 (4): 663–671

Fane, B.A., Brentlinger, K.L., Burch, A.D., *et al*. (1988) “ΦX174 *et al*., the *Microviridae*.” In Calendar, R. (ed.) The Bacteriophages. 2nd ed. New York: Plenum Press. pp. 129–148

Fijalkowska, I.J. and Schaaper, R.M. (1996) Mutants in the Exo I motif of Escherichia coli dnaQ: defective proofreading and inviability due to error catastrophe. In 1996

Fishel, R. (2015) Mismatch repair. The Journal of biological chemistry, 290 (44): 26395–26403

Foster, P.L. (2006) Methods for determining spontaneous mutation rates. Methods in enzymology, 409: 195–213

Garrison, E. and Marth, G. (2012) Haplotype-based variant detection from short-read sequencing

Gerrish, P.J. and Lenski, R.E. (1998) The fate of competing beneficial mutations in an asexual population. Genetica, 102-103 (1-6): 127–144

Giaever, G., Chu, A.M., Ni, L., *et al*. (2002) Functional profiling of the Saccharomyces cerevisiae genome. Nature, 418 (6896): 387–391

Gillam, S., Atkinson, T., Markham, A., *et al*. (1985) Gene K of bacteriophage phi X174 codes for a protein which affects the burst size of phage production. Journal of Virology, 53 (2): 708–709

Godson, G.N. and Vapnek, D. (1973) A simple method of preparing large amounts of phiX174 RF 1 supercoiled DNA. Biochimica et biophysica acta, 299 (4): 516–520

Gómez, P. and Buckling, A. (2013) Coevolution with phages does not influence the evolution of bacterial mutation rates in soil. The ISME journal, 7 (11): 2242–2244

Gordon, A.J., Burns, P.A. and Glickman, B.W. (1990) N-methyl-N'-nitro-N-nitrosoguanidine induced DNA sequence alteration; non-random components in alkylation mutagenesis. Mutation research, 233 (1-2): 95–103

Gould, S.J. (1989) Wonderful Life: The Burgess Shale and the Nature of History. New York City: W.W. Norton and Co.

Graur, D. and Li, W.-H. (2013) Fundamentals of Molecular Evolution, Second Edition. 2nd ed. Sunderland, MA: Sinauer Associates Inc.

Gresham, D. and Dunham, M.J. (2014) The enduring utility of continuous culturing in experimental evolution. Genomics, 104 (6): 1–7

Griffiths, A., Miller, J.H., Suzuki, D.T., *et al.* (2004) An Introduction to Genetic Analysis. In. 11 ed. New York: W.H.Freeman & Co Ltd

Hafenstein, S. and Fane, B.A. (2002) ϕX174 genome-capsid interactions influence the biophysical properties of the virion: evidence for a scaffolding-like function for the genome during the final stages of …. Journal of Virology

Haldane, J.B.S. (1927) A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. Mathematical Proceedings of the Cambridge Philosophical Society. 23(7): 838-844

Heilbron, K., Toll-Riera, M., Kojadinovic, M., *et al.* (2014) Fitness Is Strongly Influenced by Rare Mutations of Large Effect in a Microbial Mutation Accumulation Experiment. Genetics, 197 (3): 981–990

Helleday, T., Eshtad, S. and Nik-Zainal, S. (2014) Mechanisms underlying mutational signatures in human cancers. Nature reviews. Genetics, 15 (9): 585–598

Holder, K.K. and Bull, J.J. (2001) Profiles of Adaptation in Two Similar Viruses. Genetics, 159 (4): 1393–1404

Hutchison, C.A. and Sinsheimer, R.L. (1963) Kinetics of bacteriophage release by single cells of ϕX174-infected E. coli. Journal of molecular biology

Hutchison, C.A., Phillips, S., Edgell, M.H., *et al.* (1978) Mutagenesis at a specific position in a DNA sequence. Journal of Biological …

Incardona, N.L., Tuech, J.K. and Murti, G. (1985) Irreversible binding of phage ΦX174 to cell-bound lipopolysaccharide receptors and release of virus-receptor complexes. Biochemistry. 24: 6439-6446

Joshi NA, Fass JN. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle.

Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution, 30 (4): 772–780

Kawecki, T.J., Lenski, R.E., Ebert, D., *et al*. (2012) Experimental evolution. Trends in ecology & evolution, 27 (10): 547–560

Keller, T.E., Wilke, C.O. and Bull, J.J. (2012) Interactions between evolutionary processes at high mutation rates. Evolution, 66 (7): 2303–2314

Kimchi-Sarfaty, C., Oh, J.M., Kim, I.-W., *et al*. (2007) A “silent” polymorphism in the MDR1 gene changes substrate specificity. Science, 315 (5811): 525–528

Kimura, M. and Marayuma, T. (1966) The Mutational Load With Epistatic Gene Interactions in Fitness. Genetics. 54(6): 1337–1351

Kimura, M. (1967) On the evolutionary adjustment of spontaneous mutation rates. Genetics Research, 9 (1): 23–34

Kraemer, K.H. and Seidman, M.M. (1989) Use of supF, an Escherichia coli tyrosine suppressor tRNA gene, as a mutagenic target in shuttle-vector plasmids. Mutation research, 220 (2-3): 61–72

Krokan, H.E. and Bjørås, M. (2013) Base excision repair. Cold Spring Harbor perspectives in biology, 5 (4): a012583

Lang, G.I., Rice, D.P., Hickman, M.J., *et al*. (2013) Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. Nature, 500 (7464): 571–574

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. Nature methods, 9 (4): 357–359

Lenski, R. (2017) Lenski Lab Website. Available at: http://myxo.css.msu.edu [accessed 16 July 2017].

Levinson, G. and Gutman, G.A. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Molecular Biology and Evolution

Li, H., Handsaker, B., Wysoker, A., *et al*. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England), 25 (16): 2078–2079

Loewe, L. (2006) Quantifying the genomic decay paradox due to Muller's ratchet in human mitochondrial DNA. Genet Res, 87(2):133-59

Luria, S.E. and Delbrück, M. (1943) Mutations of Bacteria from Virus Sensitivity to Virus Resistance. Genetics, 28 (6): 491–511

Lynch, M. (2010) Evolution of the mutation rate. Trends in Genetics, 26 (8): 345–352

M Goulian, A.K.R.L.S. (1967) Enzymatic synthesis of DNA, XXIV. Synthesis of infectious phage phi-X174 DNA. Proceedings of the National Academy of Sciences, 58 (6): 2321–2328

Macqueen, D.J. and Wilcox, A.H. (2014) Characterization of the definitive classical calpain family of vertebrates using phylogenetic, evolutionary and expression analyses. Open biology

Maki, H. and Kornberg, A. (1985) The polymerase subunit of DNA polymerase III of Escherichia coli. II. Purification of the alpha subunit, devoid of nuclease activities. The Journal of biological chemistry, 260 (24): 12987–12992

Marinus, M.G. (2010) DNA methylation and mutator genes in Escherichia coli K-12. Mutation Research/Reviews in Mutation Research, 705 (2): 71–76

Martin, M. (2011) Cutadapt removes adapter sequences from highthroughput sequencing reads. EMBnet J 17: 10–12

McKenna, A., Hanna, M., Banks, E., *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20 (9): 1297–1303

Mendel, S. (2006) Interaction of the transmembrane domain of lysis protein E from bacteriophage ΦX174 with bacterial translocase MraY and peptidyl-prolyl isomerase SlyD. Microbiology, 152 (10): 2959–2967

Molineux, I.J. and Panja, D. (2013) Popping the cork: mechanisms of phage genome ejection. Nature Reviews Microbiology, 11 (3): 194–204

Montville, R., Froissart, R., Remold, S.K., *et al.* (2005) Evolution of Mutational Robustness in an RNA Virus Penny, D. (ed.). PLoS Biology, 3 (11): e381–7

Muller, H.J. (1964) The relation of recombination to mutational advance. … Research/Fundamental and Molecular Mechanisms of …

Nakamura, Y., Gojobori, T. and Ikemura, T. (2000) Codon usage tabulated from the international DNA sequence databases: status for the year 2000. Nucl. Acids Res. 28, 292.

National Center for Biotechnology Information (US). Genes and Disease [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 1998-. Anemia, sickle cell. Available from: www.ncbi.nlm.nih.gov/books/NBK22238/

National Center for Biotechnology Information (US). Genes and Disease [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 1998-. Cystic fibrosis. Available from: www.ncbi.nlm.nih.gov/books/NBK22202/

Negoro, S., Ohki, T., Shibata, N., *et al.* (2005) X-ray crystallographic analysis of 6-aminohexanoate-dimer hydrolase: molecular basis for the birth of a nylon oligomer-degrading enzyme. The Journal of biological chemistry, 280 (47): 39644–39652

Ohno, S. (1984) Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. Proceedings of the National Academy of Sciences of the United States of America, 81 (8): 2421–2425

Ojha, R.P., Roychoudhury, M. and Sanyal, N.K. (1991) Specificity of transcription and incorporation of nucleoside analogues. Journal of Molecular Structure: …, 233: 247–273

Pál, C., Maciá, M.D., Oliver, A., et al. (2007) Coevolution with viruses drives the evolution of bacterial mutation rates. Nature, 450 (7172): 1079–1081

Pavelka, N. (2014) Experimental Ecology and Evolution of Candida albicans in the mammalian gastrointestinal tract. Presented to the Experimental Approaches to Evolution and Ecology Using Yeast and Other Model Systems conference, Heidelberg, 12-15 October 2014

Pepin, K.M. and Wichman, H.A. (2008) Experimental evolution and genome sequencing reveal variation in levels of clonal interference in large populations of bacteriophage ϕX174. BMC Evol. Biol. 8 (1): 85

Pepin, K.M., Domsic, J. and McKenna, R. (2008) Genomic evolution in a virus under specific selection for host recognition. Infection, Genetics and Evolution, 8 (6): 825–834

Perales, C. and Domingo, E. (2015) Antiviral Strategies Based on Lethal Mutagenesis and Error Threshold. In. Current Topics in Microbiology and Immunology. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 1–17

Pfeiffer, J.K. and Kirkegaard, K. (2003) A single mutation in poliovirus RNA-dependent RNA polymerase confers resistance to mutagenic nucleotide analogs via increased fidelity. Proceedings of the National Academy of Sciences of the United States of America, 100 (12): 7289–7294

Plotkin, J.B. and Kudla, G. (2011) Synonymous but not the same: the causes and consequences of codon bias. Nature Reviews Genetics. 12: 32-42

Poon, A. (2005) The Rate of Compensatory Mutation in the DNA Bacteriophage X174. Genetics, 170 (3): 989–999

Raes, J. and Van de Peer, Y. (2005) Functional divergence of proteins through frameshift mutations. Trends in Genetics, 21 (8): 428–431

Raney, J.L., Delongchamp, R.R. and Valentine, C.R. (2004) Spontaneous mutant frequency and mutation spectrum for gene A of phiX174 grown in E. coli. Environmental and Molecular Mutagenesis, 44 (2): 119–127

Raynes, Y. and Sniegowski, P.D. (2014) Experimental evolution and the dynamics of genomic mutation rate modifiers. 113 (5): 375–380

Rennell, D., Bouvier, S.E. and Hardy, L.W. (1991) Systematic mutation of bacteriophage T4 lysozyme. Journal of molecular …, 222 (1): 67–88

Rose, M.R. (1984) Laboratory evolution of postponed senescence in Drosophila melanogaster. Evolution

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M. and Smith, M. (1977) Nucleotide sequence of bacteriophage ϕX174 DNA. Nature. 265: 687–695.

Sanger, F., Coulson, A.R., Friedmann, T., *et al.* (1978) The nucleotide sequence of bacteriophage ϕX174. Journal of molecular …, 125 (2): 225–246

Sanjuán, R., Nebot, M.R., Chirico, N., *et al.* (2010) Viral mutation rates. Journal of Virology, 84 (19): 9733–9748

Schaaper, R.M. (1998) Antimutator mutants in bacteriophage T4 and Escherichia coli. Genetics, 148 (4): 1579–1585

Sewall Wright, T.D. (1946) Genetics of Natural Populations. Xii. Experimental Reproduction of Some of the Changes Caused by Natural Selection in Certain Populations of Drosophila Pseudoobscura. Genetics, 31 (2): 125–156

Shlomai, J., Polder, L., Arai, K., *et al.* (1981) Replication of phi X174 dna with purified enzymes. I. Conversion of viral DNA to a supercoiled, biologically active duplex. Journal of Biological Chemistry

Shuman, H.A. and Silhavy, T.J. (2003) The art and design of genetic screens: Escherichia coli. Nature reviews. Genetics, 4 (6): 419–431

Simon-Loriere, E. and Holmes, E.C. (2013) Gene Duplication Is Infrequent in the Recent Evolutionary History of RNA Viruses. Molecular Biology and Evolution, 30 (6): 1263–1269

Sinsheimer, R.L. (1959) A single-stranded deoxyribonucleic acid from bacteriophage ϕX174. Journal of Molecular Biology

Smith, H.O., Hutchison, C.A., III, Pfannkoch, C., *et al.* (2003) Generating a synthetic genome by whole genome assembly: ϕX174 bacteriophage from synthetic oligonucleotides. Proceedings of the National Academy of Sciences of the United States of America, 100 (26): 15440–15445

Smith, J.M. and Haigh, J. (2007) The hitch-hiking effect of a favourable gene. Genetical research, 89 (5-6): 391–403

Sniegowski, P.D., Gerrish, P.J., Johnson, T., *et al.* (2000) The evolution of mutation rates: separating causes from consequences. Bioessays

Springman, R., Keller, T., Molineux, I.J., *et al.* (2010) Evolution at a High Imposed Mutation Rate: Adaptation Obscures the Load in Phage T7. Genetics, 184 (1): 221–232

Sun, L., Young, L.N., Zhang, X., *et al.* (2014) Icosahedral bacteriophage ΦX174 forms a tail for DNA transport during infection. Nature, 505 (7483): 432–435

Sung, W., Ackerman, M.S., Miller, S.F., *et al.* (2012) Drift-barrier hypothesis and mutation-rate evolution. Proceedings of the National Academy of Sciences, 109 (45): 18488–18492

Tejero, H., Marín, A. and Montero, F. (2011) The relationship between the error catastrophe, survival of the flattest, and natural selection. BMC evolutionary biology, 11 (1): 2

Tessman, I., Ishiwa, H. and Kumar, S. (1965) Mutagenic Effects of Hydroxylamine in vivo. Science, 148 (3669): 507–508

Vale, P.F., Choisy, M., Froissart, R., *et al.* (2012) The distribution of mutational fitness effects of phage ϕX174 on different hosts. Evolution, 66 (11): 3495–3507

Wang, X., Montero Llopis, P. and Rudner, D.Z. (2013) Organization and segregation of bacterial chromosomes. Nature reviews. Genetics, 14 (3): 191–203

Wichman, H.A. and Brown, C.J. (2010) Experimental evolution of viruses: Microviridae as a model system. Philosophical Transactions of the Royal Society B: Biological Sciences, 365 (1552): 2495–2501

Wichman, H.A., Badgett, M.R., Scott, L.A., *et al.* (1999) Different Trajectories of Parallel Evolution During Viral Adaptation. Science, 285 (5426): 422–424

Wichman, H.A., Millstein, J. and Bull, J.J. (2005) Adaptive molecular evolution for 13,000 phage generations: a possible arms race. Genetics, 170 (1): 19–31

Wichman, H.A., Scott, L.A., Yarber, C.D., *et al.* (2000) Experimental evolution recapitulates natural evolution. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 355 (1403): 1677–1684

Zhang, J., Kobert, K., Flouri, T., *et al.* (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics (Oxford, England), 30 (5): 614–620

Zhen, Y. and Andolfatto, P. (2012) “Methods to Detect Selection on Noncoding DNA.” In Clokie, M.R.J. and Kropinski, A.M. (eds.) Bacteriophage Plaques: Theory and Analysis. Methods in Molecular Biology. vol. Totowa, NJ: Humana Press. pp. 141–159

Appendix A – Python scripts

A.1 OriginPositionFixer.py

```
#!/usr/bin/env python3

# This script will go through the VCF files mapped against the origin region and change
# the positions to be consistent with the reference genome


f = open("origin1.vcf","r")
g = open("origin2.vcf","w")


# Run through every line, writing notes directly to new vcf:

for line in f:
        if "#" in line:
                g.write(line)

# Origin ref genome has 696 positions.
# 1-346 are equivalent to  5041 - 5386.  If position is <= 346, add 5040
# 347-696 are equivalent to 1-350. If position is >=347, subtract 346

        else:
                s = line
                a,b,c = s.split("\t",2)
                b = int(b)
                if (b <= 346):
                        b = b + 5040
                elif (b >= 347):
                        b = b - 346
                else:
                        g.write("error\n")
                b = str(b)
                g.write(a + "\t" + b + "\t" + c)

f.close()
g.close()
```

A.2 OriginMerger.py

```
# This script will parse the main VCF file.  After each line, it will search the VCF
file
# mapped against the origin region to see if that contains data for the same position.
If
# so, it will compare to see which has most reads and write that to a new output file.

# This will only work in the current directory. Will use bash to navigate through the
# directories and run this script

main = open("main.vcf","r")
origin = open("origin2.vcf","r")
merged = open("merged.vcf", "w")

# regular expression required to split with multiple delimiters
import re

# Take origin file and make a list containing each line, which will also be split into
lists

olist = []

for line in origin:
    if "#" in line:
        pass
    else:
        s = line.split()
        s[1] = int(s[1])
        olist.append(s)

# Run through every line.  Those that start with # will be written to the output:

for line in main:
    if "#" in line:
        merged.write(line)

# split the lines, and determine the position

    else:
        s = line.split()
        pos = int(s[1])

        # now see if the list of lists contains the same position

        match = 0
        for item in olist:
                if item[1] == pos:
                        match = 1
#compare the depths of the two lists and write the highest one to file

                        discard1, odepth, discard2 = re.split('DP=|;DPB', item[7],
maxsplit=2)
                        discard3, mdepth, discard4 = re.split('DP=|;DPB', s[7],
maxsplit=2)
                        try:
                                odepth = int(odepth)
                        except ValueError:
                                break
                        mdepth = int(mdepth)
                        if odepth > mdepth:
                                print (pos, odepth, mdepth)
                                x = item
                                x[1] = str(x[1])
                                y = "\t".join(x)
                                y = y + "\n"
                                merged.write(y)
                        elif odepth < mdepth:
                                merged.write(line)
                        else:
                                merged.write(line)


        # if the position was not found in both files, use the one from main
        if match == 0:
                if len(s[3]) == 1:
                        merged.write(line)
                else:
                        pass
```

```
main.close()
origin.close()
merged.close()
```

A.3 Consensus.py

```
#!/usr/bin/env python3

# This script takes a reference genome and parses the VCF file.
# At each position it will check to see if the VCF contains a new major allele
# (with frequency of > 50%).  It will write either the reference base or new allele
# to a new FASTA file.

# add reference genome as a string.  It has been prefixed with 'X'
# to correct for position (because Python strings start at position 0)
# actual reference genome not included in typed version of script
reference = “X…”

vars = open("snps40.vcf","r")
consensus = open("consensus.FASTA", "w")
consensus.write(">PhiX174_Consensus\n")

# Take vcf file and make a list containing each line, which will also be split into
lists

varlist = []
for line in vars:
    if "#" in line:
        pass
    else:
        s = line.split()
        s[1] = int(s[1])
        varlist.append(s)

# Go through genome position by position
position = 1
while position < 5387:
    base = ""

    # check each line of VCF file to see if the current position is included.
    # if so, check if it has a frequency of over 50%
    for line in varlist:
                if line[1] == position:
                        ref = line[3]
                        alt = line[4].split(",")
                        stuff = line[9].split(":")
                        depth = stuff[1]
                        freq = stuff[2].split(",")

                        if len(alt) == 3:
                                perc = round(float(freq[1])/float(depth) * 100,2)
                                if perc > 50:
                                        base = alt[0]
                                perc = round(float(freq[2])/float(depth) * 100,2)
                                if perc > 50:
                                        base = alt[1]
                                perc = round(float(freq[3])/float(depth) * 100,2)
                                if perc > 50:
                                        base = alt[2]
                        elif len(alt) == 2:
                                perc = round(float(freq[1])/float(depth) * 100,2)
                                if perc > 50:
                                        base = alt[0]
                                perc = round(float(freq[2])/float(depth) * 100,2)
                                if perc > 50:
                                        base = alt[1]
                        elif len(alt) == 1:
                                perc = round(float(freq[1])/float(depth) * 100,2)
                                if perc > 50:
                                        base = alt[0]
    # if a new major allele, write that to output.  otherwise write reference
    if base == "":
        base = reference[position]
    consensus.write(base)
    # insert newline every 70 bases
    if position % 70 == 0:
        consensus.write("\n")
    position += 1


vars.close()
consensus.close()
```

A.4 VCFsimplifier.py

```
#!/usr/bin/env python3
# This script parses vcf files and returns a text file containing
# Position, reference and allele, depth, and frequency.
# Only returns mutations with at least 10% frequency, but that number was adjusted as
required

f = open("snps.vcf","r")
g = open("snps.txt","w")



g.write("Position\tRef\tAllele\tDepth\tFreq\t%age\n")

for line in f:
    if "#" in line:
        pass
    else:
        t = line.split()
        pos = t[1]
        ref = t[3]
        alt = t[4].split(",")
        stuff = t[9].split(":")
        depth = stuff[1]
        freq = stuff[2].split(",")
        if len(alt) == 3:
                # If three different mutations at that position
                perc = round(float(freq[1])/float(depth) * 100,2)
                if perc >= 10:
                        g.write(pos + "\t\t" + ref + "\t" + alt[0] + "\t" + depth + "\t"
+ freq[1] + "\t" + str(perc) + "\n")
                perc = round(float(freq[2])/float(depth) * 100,2)
                if perc >= 10:
                        g.write(pos + "\t\t" +ref + "\t" + alt[1] + "\t" + depth + "\t" +
freq[2] + "\t" + str(perc) + "\n")
                perc = round(float(freq[3])/float(depth) * 100,2)
                if perc >= 10:
                        g.write(pos + "\t\t" +ref + "\t" + alt[2] + "\t" + depth + "\t" +
freq[3] + "\t" + str(perc) + "\n")
        elif len(alt) == 2:
                # If two different mutations at that position
                perc = round(float(freq[1])/float(depth) * 100,2)
                if perc >= 10:
                        g.write(pos + "\t\t" +ref + "\t" + alt[0] + "\t" + depth + "\t" +
freq[1] + "\t" + str(perc) + "\n")
                perc = round(float(freq[2])/float(depth) * 100,2)
                if perc >= 10:
                        g.write(pos + "\t\t" +ref + "\t" + alt[1] + "\t" + depth + "\t" +
freq[2] + "\t" + str(perc) + "\n")
        else:
                # If only one mutation at that position
                perc = round(float(freq[1])/float(depth) * 100,2)
                if perc >= 10:
                        g.write(pos + "\t\t" +ref + "\t" + alt[0] + "\t" + depth + "\t" +
freq[1] + "\t" + str(perc) + "\n")

f.close()
g.close()
```

Appendix B – Genome sequences

B.1 ΦX174 reference

The sequence of the ancestral ΦX174 genome (GenBank ID AF176034.1), confirmed by Illumina sequencing.

```
GAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTCGGATATTTCTGATGAGTCGAAAAATTATCTT
GATAAAGCAGGAATTACTACTGCTTGTTTACGAATTAAATCGAAGTGGACTGCTGGCGGAAAATGAGAAA
ATTCGACCTATCCTTGCGCAGCTCGAGAAGCTCTTACTTTGCGACCTTTCGCCATCAACTAACGATTCTG
TCAAAAACTGACGCGTTGGATGAGGAGAAGTGGCTTAATATGCTTGGCACGTTCGTCAAGGACTGGTTTA
GATATGAGTCACATTTTGTTCATGGTAGAGATTCTCTTGTTGACATTTTAAAAGAGCGTGGATTACTATC
TGAGTCCGATGCTGTTCAACCACTAATAGGTAAGAAATCATGAGTCAAGTTACTGAACAATCCGTACGTT
TCCAGACCGCTTTGGCCTCTATTAAGCTCATTCAGGCTTCTGCCGTTTTGGATTTAACCGAAGATGATTT
CGATTTTCTGACGAGTAACAAAGTTTGGATTGCTACTGACCGCTCTCGTGCTCGTCGCTGCGTTGAGGCT
TGCGTTTATGGTACGCTGGACTTTGTGGGATACCCTCGCTTTCCTGCTCCTGTTGAGTTTATTGCTGCCG
TCATTGCTTATTATGTTCATCCCGTCAACATTCAAACGGCCTGTCTCATCATGGAAGGCGCTGAATTTAC
GGAAAACATTATTAATGGCGTCGAGCGTCCGGTTAAAGCCGCTGAATTGTTCGCGTTTACCTTGCGTGTA
CGCGCAGGAAACACTGACGTTCTTACTGACGCAGAAGAAAACGTGCGTCAAAAATTACGTGCAGAAGGAG
TGATGTAATGTCTAAAGGTAAAAAACGTTCTGGCGCTCGCCCTGGTCGTCCGCAGCCGTTGCGAGGTACT
AAAGGCAAGCGTAAAGGCGCTCGTCTTTGGTATGTAGGTGGTCAACAATTTTAATTGCAGGGGCTTCGGC
CCCTTACTTGAGGATAAATTATGTCTAATATTCAAACTGGCGCCGAGCGTATGCCGCATGACCTTTCCCA
TCTTGGCTTCCTTGCTGGTCAGATTGGTCGTCTTATTACCATTTCAACTACTCCGGTTATCGCTGGCGAC
TCCTTCGAGATGGACGCCGTTGGCGCTCTCCGTCTTTCTCCATTGCGTCGTGGCCTTGCTATTGACTCTA
CTGTAGACATTTTTACTTTTTATGTCCCTCATCGTCACGTTTATGGTGAACAGTGGATTAAGTTCATGAA
GGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTACTGGTTATATTGACCATGCCGCTTTTCTT
GGCACGATTAACCCTGATACCAATAAAATCCCTAAGCATTTGTTTCAGGGTTATTTGAATATCTATAACA
ACTATTTTAAAGCGCCGTGGATGCCTGACCGTACCGAGGCTAACCCTAATGAGCTTAATCAAGATGATGC
TCGTTATGGTTTCCGTTGCTGCCATCTCAAAAACATTTGGACTGCTCCGCTTCCTCCTGAGACTGAGCTT
TCTCGCCAAATGACGACTTCTACCACATCTATTGACATTATGGGTCTGCAAGCTGCTTATGCTAATTTGC
ATACTGACCAAGAACGTGATTACTTCATGCAGCGTTACCGTGATGTTATTTCTTCATTTGGAGGTAAAAC
CTCTTATGACGCTGACAACCGTCCTTTACTTGTCATGCGCTCTAATCTCTGGGCATCTGGCTATGATGTT
GATGGAACTGACCAAACGTCGTTAGGCCAGTTTTCTGGTCGTGTTCAACAGACCTATAAACATTCTGTGC
CGCGTTTCTTTGTTCCTGAGCATGGCACTATGTTTACTCTTGCGCTTGTTCGTTTTCCGCCTACTGCGAC
TAAAGAGATTCAGTACCTTAACGCTAAAGGTGCTTTGACTTATACCGATATTGCTGGCGACCCTGTTTTG
TATGGCAACTTGCCGCCGCGTGAAATTTCTATGAAGGATGTTTTCCGTTCTGGTGATTCGTCTAAGAAGT
TTAAGATTGCTGAGGGTCAGTGGTATCGTTATGCGCCTTCGTATGTTTCTCCTGCTTATCACCTTCTTGA
AGGCTTCCCATTCATTCAGGAACCGCCTTCTGGTGATTTGCAAGAACGCGTACTTATTCGCCACCATGAT
TATGACCAGTGTTTCCAGTCCGTTCAGTTGTTGCAGTGGAATAGTCAGGTTAAATTTAATGTGACCGTTT
ATCGCAATCTGCCGACCACTCGCGATTCAATCATGACTTCGTGATAAAAGATTGAGTGTGAGGTTATAAC
GCCGAAGCGGTAAAAATTTTAATTTTTGCCGCTGAGGGGTTGACCAAGCGAAGCGCGGTAGGTTTTCTGC
TTAGGAGTTTAATCATGTTTCAGACTTTTATTTCTCGCCATAATTCAAACTTTTTTTCTGATAAGCTGGT
TCTCACTTCTGTTACTCCAGCTTCTTCGGCACCTGTTTTACAGACACCTAAAGCTACATCGTCAACGTTA
TATTTTGATAGTTTGACGGTTAATGCTGGTAATGGTGGTTTTCTTCATTGCATTCAGATGGATACATCTG
TCAACGCCGCTAATCAGGTTGTTTCTGTTGGTGCTGATATTGCTTTTGATGCCGACCCTAAATTTTTTGC
CTGTTTGGTTCGCTTTGAGTCTTCTTCGGTTCCGACTACCCTCCCGACTGCCTATGATGTTTATCCTTTG
AATGGTCGCCATGATGGTGGTTATTATACCGTCAAGGACTGTGTGACTATTGACGTCCTTCCCCGTACGC
CGGGCAATAATGTTTATGTTGGTTTCATGGTTTGGTCTAACTTTACCGCTACTAAATGCCGCGGATTGGT
TTCGCTGAATCAGGTTATTAAAGAGATTATTTGTCTCCAGCCACTTAAGTGAGGTGATTTATGTTTGGTG
CTATTGCTGGCGGTATTGCTTCTGCTCTTGCTGGTGGCGCCATGTCTAAATTGTTTGGAGGCGGTCAAAA
AGCCGCCTCCGGTGGCATTCAAGGTGATGTGCTTGCTACCGATAACAATACTGTAGGCATGGGTGATGCT
GGTATTAAATCTGCCATTCAAGGCTCTAATGTTCCTAACCCTGATGAGGCCGCCCCTAGTTTTGTTTCTG
GTGCTATGGCTAAAGCTGGTAAAGGACTTCTTGAAGGTACGTTGCAGGCTGGCACTTCTGCCGTTTCTGA
TAAGTTGCTTGATTTGGTTGGACTTGGTGGCAAGTCTGCCGCTGATAAAGGAAAGGATACTCGTGATTAT
CTTGCTGCTGCATTTCCTGAGCTTAATGCTTGGGAGCGTGCTGGTGCTGATGCTTCCTCTGCTGGTATGG
TTGACGCCGGATTTGAGAATCAAAAAGAGCTTACTAAAATGCAACTGGACAATCAGAAAGAGATTGCCGA
```

```
GATGCAAAATGAGACTCAAAAAGAGATTGCTGGCATTCAGTCGGCGACTTCACGCCAGAATACGAAAGAC
CAGGTATATGCACAAAATGAGATGCTTGCTTATCAACAGAAGGAGTCTACTGCTCGCGTTGCGTCTATTA
TGGAAAACACCAATCTTTCCAAGCAACAGCAGGTTTCCGAGATTATGCGCCAAATGCTTACTCAAGCTCA
AACGGCTGGTCAGTATTTTACCAATGACCAAATCAAAGAAATGACTCGCAAGGTTAGTGCTGAGGTTGAC
TTAGTTCATCAGCAAACGCAGAATCAGCGGTATGGCTCTTCTCATATTGGCGCTACTGCAAAGGATATTT
CTAATGTCGTCACTGATGCTGCTTCTGGTGTGGTTGATATTTTTCATGGTATTGATAAAGCTGTTGCCGA
TACTTGGAACAATTTCTGGAAAGACGGTAAAGCTGATGGTATTGGCTCTAATTTGTCTAGGAAATAACCG
TCAGGATTGACACCCTCCCAATTGTATGTTTTCATGCCTCCAAATCTTGGAGGCTTTTTTATGGTTCGTT
CTTATTACCCTTCTGAATGTCACGCTGATTATTTTGACTTTGAGCGTATCGAGGCTCTTAAACCTGCTAT
TGAGGCTTGTGGCATTTCTACTCTTTCTCAATCCCCAATGCTTGGCTTCCATAAGCAGATGGATAACCGC
ATCAAGCTCTTGGAAGAGATTCTGTCTTTTCGTATGCAGGGCGTTGAGTTCGATAATGGTGATATGTATG
TTGACGGCCATAAGGCTGCTTCTGACGTTCGTGATGAGTTTGTATCTGTTACTGAGAAGTTAATGGATGA
ATTGGCACAATGCTACAATGTGCTCCCCCAACTTGATATTAATAACACTATAGACCACCGCCCCGAAGGG
GACGAAAAATGGTTTTTAGAGAACGAGAAGACGGTTACGCAGTTTTGCCGCAAGCTGGCTGCTGAACGCC
CTCTTAAGGATATTCGCGATGAGTATAATTACCCCAAAAAGAAAGGTATTAAGGATGAGTGTTCAAGATT
GCTGGAGGCCTCCACTATGAAATCGCGTAGAGGCTTTACTATTCAGCGTTTGATGAATGCAATGCGACAG
GCTCATGCTGATGGTTGGTTTATCGTTTTTGACACTCTCACGTTGGCTGACGACCGATTAGAGGCGTTTT
ATGATAATCCCAATGCTTTGCGTGACTATTTTCGTGATATTGGTCGTATGGTTCTTGCTGCCGAGGGTCG
CAAGGCTAATGATTCACACGCCGACTGCTATCAGTATTTTTGTGTGCCTGAGTATGGTACAGCTAATGGC
CGTCTTCATTTCCATGCGGTGCATTTTATGCGGACACTTCCTACAGGTAGCGTTGACCCTAATTTTGGTC
GTCGGGTACGCAATCGCCGCCAGTTAAATAGCTTGCAAAATACGTGGCCTTATGGTTACAGTATGCCCAT
CGCAGTTCGCTACACGCAGGACGCTTTTTCACGTTCTGGTTGGTTGTGGCCTGTTGATGCTAAAGGTGAG
CCGCTTAAAGCTACCAGTTATATGGCTGTTGGTTTCTATGTGGCTAAATACGTTAACAAAAAGTCAGATA
TGGACCTTGCTGCTAAAGGTCTAGGAGCTAAAGAATGGAACAACTCACTAAAAACCAAGCTGTCGCTACT
TCCCAAGAAGCTGTTCAGAATCAGAATGAGCCGCAACTTCGGGATGAAAATGCTCACAATGACAAATCTG
TCCACGGAGTGCTTAATCCAACTTACCAAGCTGGGTTACGACGCGACGCCGTTCAACCAGATATTGAAGC
AGAACGCAAAAAGAGAGATGAGATTGAGGCTGGGAAAAGTTACTGTAGCCGACGTTTTGGCGGCGCAACC
TGTGACGACAAATCTGCTCAAATTTATGCGCGCTTCGATAAAAATGATTGGCGTATCCAACCTGCA
```

B.2 Consensus sequence of line A1 after 100 passages

```
GAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTCGGATATTTCTGATGAGTCGAAAAATTATCTT
GATAAAGCAGGAATTACTACTGCTTGTTTACGAATTAAATCGAAGTGGACTGCTGGCGGAAAATGAGAAA
ATTCGACCTATCCTTGCGCAGCTCGAGAAGCTCTTACTTTGCGACCTTTCGCCATCAACTAACGATTCTG
TCAAAAACTGACGCGTTGGATGAGGAGAAGTGGCTTAATATGCTTGGCACGTTCGTCAAGGACTGGTTTA
GATATGAGTCACATTTTGTTCATGGTAGAGATTCTCTTGTTGACATTTTAAAAGAGCGTGGATTACTATC
TGAGTCCGATGCTGTTCAACCACTAATAGGTAAGAAATCATGAGTCAAGTTACTGAACAATCCGTACGTT
TCCAGACCGCTTTGGCCTCTATTAAGCTCATTCAGGCTTCTGCCGTTTTGGATTTAACCGAAGATGATTT
CGATTTTCTGACGAGTAACAAAGTTTGGATTGCTACTGACCGCTCTCGTGCTCGTCGCTGCGTTGAGGCT
TGCGTTTATGGTACGCTGGACTTTGTGGGATACCCTCGCTTTCCTGCTCCTGTTGAGTTTATTGCTGCCG
TCATTGCTTATTATGTTCATCCCGTCAACATTCAAACGGCCTGTCTCATCATGGAAGGCGCTGAATTTAC
GGAAAACATTATTAATGGCGTCGAGCGTCCGGTTAAAGCCGCTGAATTGTTCGCGTTTACCTTGCGTGTA
CGCGCAGGAAACACTGACGTTCTTACTGACGCAGAAGAAAACGTGCGTCAAAAATTACGTGCAGAAGGAG
TGATGTAATGTCTAAAGGTAAAAAACGTTCTGGCGCTCGCCCTGGTCGTCCGCAGCCGTTGCGAGGTACT
AAAGGCAAGCGTAAAGGCGCTCGTCTTTGGTATGTAGGTGGTCAACAATTTTAATTGCAGGGGCTTCGGC
CCCTTACTTGAGGATAAATTATGTCTAATATTCAAACTGGCGCCGAGCGTATGCCGCATGACCTTTCCCA
TCTTGGCTTCCTTGCTGGTCAGATTGGTCGTCTTATTACCATTTCAACTACTCCGGTTATCGCTGGCGAC
TCCTTCGAGATGGACGCCGTTGGCGCTCTCCGTCTTTCTCCATTGCGTCGTGGCCTTGCTATTGACTCTA
CTGTAGACATTTTTACTTTTTATGTCCCTCATCGTCACGTTTATGGTGAACAGTGGATTAAGTTCATGAA
GGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTGCTGATTATATTGACCATGCCGCTTTTCTT
GGCACGATTAACCCTGATACCAATAAAATCCCTAAGCATTTGTTTCAGGGTTATTTGAATATCTATAACA
ACTATTTTAAAGCGCCGTGGATGCCTGACCGTACCGAGGCTAACCCTAATGAGCTTAATCAAGATGATGC
TCGTTATGGTTTCCGTTGCTGCCATCTCAAAAACATTTGGACTGCTCCGCTTCCTCCTGAGACTGAGCTT
TCTCGCCAAATGACGACTTCTACCACATCTATTGACATTATGGGTCTGCAAGCTGCTTATGCTAATTTGC
```

```
ATACTGACCAAGAACGTGATTACTTCATTCAGCGTTACCGTGATGTTATCTCTTCATTTGGAGGTAAAAC
CTCTTATGACGCTGACAACCGTCCTTTACTTGTCATGCGCTCTAATCTCTGGGCATCTGGCTATGATGTT
GATGGAACTGACCAAACGTCGTTAGGCCAGTTTTCTGGTCGTGTTCAACAGACCTATAAACATTCTGTGC
CGCGTTTCTTTGTTCCTGAGCATGGCACTATGTTTACTCTTGCGCTTGTTCGTTTTCCGCCTACTGCGAC
TAAAGAGATTCAGTACCTTAACGCTAAAGGTGCTTTGACTTATACCGATATTGCTGGCGACCCTGTTTTG
TATGGCAACTTGCCGCCGCGTGAAATTTCTATGAAGGATGTTTTCCGTTCTGGTGATTCGTCTAAGAAGT
TTAAGATTGCTGAGGGTCAGTGGTATCGTTATGCGCCTTCGTATGTTTCTCCTGCTTATCACCTTCTTGA
AGGCTTCCCATTCATTCAGGAACCGCCTTCTGGTGATTTGCAAGAACGCGTACTTATTCGCCACCATGAT
TATGACCAGTGTTTCCAGTCCGTTCAGTTGTTGCAGTGGAATAGTCAGGTTAAATTTAATGTGACCGTTT
ATCGCAATCTGCCGACCACTCGCGATTCAATCATGACTTCGTGATAAAAGATTGAGTGTGAGGTTATAAC
GCCGAAGCGGTAAAAATTTTAATTTTTGCCGCTGAGGGGTTGACCAAGCGAAGCGCGGTAGGTTTTCTGC
TTAGGAGTTTAATCATGTTTCAGACTTTTATTTCTCGCCATAATTCAAACTTTTTTTCTGATAAGCTGGT
TCTCACTTCTGTTACTCCAGCTTCTTCGGCACCTGTTTTACAGACACCTAAAGCTACATCGTCAACGTTA
TATTTTGATAGTTTGACGGTTAATGCTGGTAATGGTGGTTTTCTTCATTGCATTCAGATGGATACATCTG
TCAACGCCGCTAATCAGGTTGTTTCTGTTGGTGCTGATATTGCTTTTGATGCCGACCCTAAATTTTTTGC
CTGTTTGGTTCGCTTTGAGTCTTCTTCGGTTCCGACTACCCTCCCGACTGCCTATGATGTTTATCCTTTG
AATGGTCGCCATGATGGTGGTTATTATACCGTCAAGGACTGTGTGACTATTGACGTCCTTCCCCGTACGC
CGGGCAATAATGTTTATGTTGGTTTCATGGTTTGGTCTAACTTTACCGCTACTAAATGCCGCGGATTGGT
TTCGCTGAATCAGGTTATTAAAGAGATTATTTGTCTCCAGCCACTTAAGTGAGGTGATTTATGTTTGGTG
CTATTGCTGGCGGTATTGCTTCTGCTCTTGCTGGTGGCGCCATGTCTAAATTGTTTGGAGGCGGTCAAAA
AGCCGCCTCCGGTGGCATTCAAGGTGATGTGCTTGCTACCGATAACAATACTGTAGGCATGGGTGATGCT
GGTATTAAATCTGCCATTCAAGGCTCTAATGTTCCTAACCCTGATGAGGCCGCCCCTAGTTTTGTTTCTG
GTGCTATGGCTAAAGCTGGTAAAGGACTTCTTGAAGGTACGTTGCAGGCTGGCACTTCTGCCGTTTCTGA
TAAGTTGCTTGATTTGGTTGGACTTGGTGGCAAGTCTGCCGCTGATAAAGGAAAGGATACTCGTGATTAT
CTTGCTGCTGCATTTCCTGAGCTTAATGCTTGGGAGCGTGCTGGTGCTGGTGCTTCCTCTGCTGGTATGG
TTGACGCCGGATTTGAGAATCAAAAAGAGCTTACTAAAATGCAACTGGACAATCAGAAAGAGATTGCCGA
GATGCAAAATGAGACTCAAAAAGAGATTGCTGGCATTCAGTCGGCGACTTCACGCCAGAATACGAAAGAC
CAGGTATATGCACAAAATGAGATGCTTGCTTATCAACAGAAGGAGTCTACTGCTCGCGTTGCGTCTATTA
TGGAAAACACCAATCTTTCCAAGCAACAGCAGGTTTCCGAGATTATGCGCCAAATGCTTACTCAAGCTCA
AACGGCTGGTCAGTATTTTACCAATGACCAAATCAAAGAAATGACTCGCAAGGTTAGTGCTGAGGTTGAC
TTAGTTCATCAGCAAACGCAGAATCAGCGGTATGGCTCTTCTCATATTGGCGCTACTGCAAAGGATATTT
CTAATGTCGTCACTGATGCTGCTTCTGGTGTGGTTGATATTTTTCATGGTATTGATAAAGCTGTTGCCGA
TACTTGGAACAATTTCTGGAAAGACGGTAAAGCTGATGGTATTGGCTCTAATTTGTCTAGGAAATAACCG
TCAGGATTGACACCCTCCCAATTGTATGTTTTCATGCCTCCAAATCTTGGAGGCTTTTTTATGGTTCGTT
CTTATTACCCTTCTGAATGTCACGCTGATTATTTTGACTTTGAGCGTATCGAGGCTCTTAAACCTGCTAT
TGAGGCTTGTGGCATTTCTACTCTTTCTCAATCCCCAATGCTTGGCTTCCATAAGCAGATGGATAACCGC
ATCAAGCTCTTGGAAGAGATTCTGTCTTTTCGTATGCAGGGCGTTGAGTTCGATAATGGTGATATGTATG
TTGACGGCCATAAGGCTGCTTCTGACGTTCGTGATGAGTTTGTATCTGTTACTGAGAAGTTAATGGATGA
ATTGGCACAATGCTACAATGTGCTCCCCCAACTTGATATTAATAACACTATAGACCACCGCCCCGAAGGG
GACGAAAAATGGTTTTTAGAGAACGAGAAGACGGTTACGCAGTTTTGCCGCAAGCTGGCTGCTGAACGCC
CTCTTAAGGATATTCGCGATGAGTATAATTACCCCAAAAAGAAAGGTATTAAGGATGAGTGTTCAAGATT
GCTGGAGGCCTCCACTATGAAATCGCGTAGAGGCTTTACTATTCAGCGTTTGATGAATGCAATGCGACAG
GCTCATGCTGATGGTTGGTTTATCGTTTTTGACACTCTCACGTTGGCTGACGACCGATTAGAGGCGTTTT
ATGATAATCCCAATGCTTTGCGTGACTATTTTCGTGATATTGGTCGTATGGTTCTTGCTGCCGAGGGTCG
CAAGGCTAATGATTCACACGCCGACTGCTATCAGTATTTTTGTGTGCCTGAGTATGGTACAGCTAATGGC
CGTCTTCATTTCCATGCGGTGCATTTTATGCGGACACTTCCTACAGGTAGCGTTGACCCTAATTTTGGTC
GTCGGGTACGCAATCGCCGCCAGTTAAATAGCTTGCAAAATACGTGGCCTTATGGTTACAGTATGCCCAT
CGCAGTTCGCTACACGCAGGACGCTTTTTCACGTTCTGGTTGGTTGTGGCCTGTTGATGCTAAAGGTGAG
CCGCTTAAAGCTACCAGTTATATGGCTGTTGGTTTCTATGTGGCTAAATACGTTAACAAAAAGTCAGATA
TGGACCTTGCTGCTAAAGGTCTAGGAGCTAAAGAATGGAACAACTCACTAAAAACCAAGCTGTCGCTACT
TCCCAAGAAGCTGTTCAGAATCAGAATGAGCCGCAACTTCGGGATGAAAATGCTCACAATGACAAATCTG
TCCACGGAGTGCTTAATCCAACTTACCAAGCTGGGTTACGACGCGACGCCGTTCAACCAGATATTGAAGC
AGAACGCAAAAAGAGAGATGAGATTGAGGCTGGGAAAAGTTACTGTAGCCGACGTTTTGGCGGCGCAACC
TGTGACGACAAATCTGCTCAAATTTATGCGCGCTTCGATAAAAATGATTGGCGTATCCAACCTGCA
```

B.3 Consensus sequence of line A2 after 100 passages

```
GAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTCGGATATTTCTGATGAGTCGAAAAATTATCTT
GATAAAGCAGGAATTACTACTGCTTGTTTACGAATTAAATCGAAGTGGACTGCTGGCGGAAAATGAGAAA
ATTCGACCTATCCTTGCGCAGCTCGAGAAGCTCTTACTTTGCGACCTTTCGCCATCAACTAACGATTCTG
TCAAAAACTGACGCGTTGGATGAGGAGAAGTGGCTTAATATGCTTGGCACGTTCGTCAAGGACTGGTTTA
GATATGAGTCACATTTTGTTCATGGTAGAGATTCTCTTGTTGACATTTTAAAAGAGCGTGGATTACTATC
TGAGTCCGATGCTGTTCAACCACTAATAGGTAAGAAATCATGAGTCAAGTTACTGAACAATCCGTACGTT
TCCAGACCGCTTTGGCCTCTATTAAGCTCATTCAGGCTTCTGCCGTTTTGGATTTAACCGAAGATGATTT
CGATTTTCTGACGAGTAACAAAGTTTGGATTGCTACTGACCGCTCTCGTGCTCGTCGCTGCGTTGAGGCT
TGCGTTTATGGTACGCTGGACTTTGTGGGATACCCTCGCTTTCCTGCTCCTGTTGAGTTTATTGCTGCCG
TCATTGCTTATTATGTTCATCCCGTCAACATTCAAACGGCCTGTCTCATCATGGAAGGCGCTGAATTTAC
GGAAAACATTATTAATGGCGTCGAGCGTCCGGTTAAAGCCGCTGAATTGTTCGCGTTTACCTTGCGTGTA
CGCGCAGGAAACACTGACGTTCTTACTGACGCAGAAGAAAACGTGCGTCAAAAATTACGTGCAGAAGGAG
TGATGTAATGTCTAAAGGTAAAAAACGTTCTGGCGCTCGCCCTGGTCGTCCGCAGCCGTTGCGAGGTACT
AAAGGCAAGCGTAAAGGCGCTCGTCTTTGGTATGTAGGTGGTCAACAATTTTAATTGCAGGGGCTTCGGC
CCCTTACTTGAGGATAAATTATGTCTAATATTCAAACTGGCGCCGAGCGTATGCCGCATGACCTTTCCCA
TCTTGGCTTCCTTGCTGGTCAGATTGGTCGTCTTATTACCATTTCAACTACTCCGGTTATCGCTGGCGAC
TCCTTCGAGATGGACGCCGTTGGCGCTCTCCGTCTTTCTCCATTGCGTCGTGGCCTTGCTATTGACTCTA
CTGTAGACATTTTTACTTTTTATGTCCCTCATCGTCACGTTTATGGTGAACAGTGGATTAAGTTCATGAA
GGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTGCTGATTATATTGACCATGCTGCTTTTCTT
GGCACGATTAACCCTGATACCAATAAAATCCCTAAGCATTTGTTTCAGGGTTATTTGAATATCTATAACA
ACTATTTTAAAGCGCCGTGGATGCCTGACCGTACCGAGGCTAACCCTAATGAGCTTAATCAAGATGATGC
TCGTTATGGTTTCCGTTGCTGCCATCTCAAAAACATTTGGACTGCTCCGCTTCCTCCTGAGACTGAGCTT
TCTCGCCAAATGACGACTTCTACCACATCTATTGACATTATGGGTCTGCAAGCTGCTTATGCTAATTTGC
ATACTGACCAAGAACGTGATTACTTCATTCAGCGTTACCGTGATGTTATTTCTTCATTTGGAGGTAAAAC
CTCTTATGACGCTGACAACCGTCCTTTACTTGTCATGCGCTCTAATCTCTGGGCATCTGGCTATGATGTT
GATGGAACTGACCAAACGTCGTTAGGCCAGTTTTCTGGTCGTGTTCAACAGACCTATAAACATTCTGTGC
CGCGTTTCTTTGTTCCTGAGCATGGCACTATGTTTACTCTTGCGCTTGTTCGTTTTCCGCCTACTGCGAC
TAAAGAGATTCAGTACCTTAACGCTAAAGGTGCTTTGACTTATACCGATATTGCTGGCGACCCTGTTTTG
TATGGCAACTTGCCGCCGCGTGAAATTTCTATGAAGGATGTTTTCCGTTCTGGTGATTCGTCTAAGAAGT
TTAAGATTGCTGAGGGTCAGTGGTATCGTTATGCGCCTTCGTATGTTTCTCCTGCTTATCACCTTCTTGA
AGGCTTCCCATTCATTCAGGAACCGCCTTCTGGTGATTTGCAAGAACGCGTACTTATTCGCCACCATGAT
TATGACCAGTGTTTCCAGTCCGTTCAGTTGTTGCAGTGGAATAGTCAGGTTAAATTTAATGTGACCGTTT
ATCGCAATCTGCCGACCACTCGCGATTCAATCATGACTTCGTGATAAAAGATTGAGTGTGAGGTTATAAC
GCCGAAGCGGTAAAAATTTTAATTTTTGCCGCTGAGGGGTTGACCAAGCGAAGCGCGGTAGGTTTTCTGC
TTAGGAGTTTAATCATGTTTCAGACTTTTATTTCTCGCCATAATTCAAACTTTTTTTCTGATAAGCTGGT
TCTCACTTCTGTTACTCCAGCTTCTTCGGCACCTGTTTTACAGACACCTAAAGCTACATCGTCAACGTTA
TATTTTGATAGTTTGACGGTTAATGCTGGTAATGGTGGTTTTCTTCATTGCATTCAGATGGATACATCTG
TCAACGCCGCTAATCAGGTTGTTTCTGTTGGTGCTGATATTGCTTTTGATGCCGACCCTAAATTTTTTGC
CTGTTTGGTTCGCTTTGAGTCTTCTTCGGTTCCGACTACCCTCCCGACTGCCTATGATGTTTATCCTTTG
AATGGTCGCCATGATGGTGGTTATTATACCGTCAAGGACTGTGTGACTATTGACGTCCTTCCCCGTACGC
CGGGCAATAATGTTTATGTTGGTTTCATGGTTTGGTCTAACTTTACCGCTACTAAATGCCGCGGATTGGT
TTCGCTGAATCAGGTTATTAAAGAGATTATTTGTCTCCAGCCACTTAAGTGAGGTGATTTATGTTTGGTG
CTATTGCTGGCGGTATTGCTTCTGCTCTTGCTGGTGGCGCCATGTCTAAATTGTTTGGAGGCGGTCAAAA
AGCCGCCTCCGGTGGCATTCAAGGTGATGTGCTTGCTACCGATAACAATACTGTAGGCATGGGTGATGCT
GGTATTAAATCTGCCATTCAAGGCTCTAATGTTCCTAACCCTGATGAGGCCGCCCCTAGTTTTGTTTCTG
GTGCTATGGCTAAAGCTGGTAAAGGACTTCTTGAAGGTACGTTGCAGGCTGGCACTTCTGCCGTTTCTGA
TAAGTTGCTTGATTTGGTTGGACTTGGTGGCAAGTCTGCCGCTGATAAAGGAAAGGATACTCGTGATTAT
CTTGCTGCTGCATTTCCTGAGCTTAATGCTTGGGAGCGTGCTGGTGCTGGTGCTTCCTCTGCTGGTATGG
TTGACGCCGGATTTGAGAATCAAAAAGAGCTTACTAAAATGCAACTGGACAATCAGAAAGAGATTGCCGA
GATGCAAAATGAGACTCAAAAAGAGATTGCTGGCATTCAGTCGGCGACTTCACGCCAGAATACGAAAGAC
CAGGTATATGCACAAAATGAGATGCTTGCTTATCAACAGAAGGAGTCTACTGCTCGCGTTGCGTCTATTA
TGGAAAACACCAATCTTTCCAAGCAACAGCAGGTTTCCGAGATTATGCGCCAAATGCTTACTCAAGCTCA
```

```
AACGGCTGGTCAGTATTTTACCAATGACCAAATCAAAGAAATGACTCGCAAGGTTAGTGCTGAGGTTGAC
TTAGTTCATCAGCAAACGCAGAATCAGCGGTATGGCTCTTCTCATATTGGCGCTACTGCAAAGGATATTT
CTAATGTCGTCACTGATGCTGCTTCTGGTGTGGTTGATATTTTTCATGGTATTGATAAAGCTGTTGCCGA
TACTTGGAACAATTTCTGGAAAGACGGTAAAGCTGATGGTATTGGCTCTAATTTGTCTAGGAAATAACCG
TCAGGATTGACACCCTCCCAATTGTATGTTTTCATGCCTCCAAATCTTGGAGGCTTTTTTATGGTTCGTT
CTTATTACCCTTCTGAATGTCACGCTGATTATTTTGACTTTGAGCGTATCGAGGCTCTTAAACCTGCTAT
TGAGGCTTGTGGCATTTCTACTCTTTCTCAATCCCCAATGCTTGGCTTCCATAAGCAGATGGATAACCGC
ATCAAGCTCTTGGAAGAGATTCTGTCTTTTCGTATGCAGGGCGTTGAGTTCGATAATGGTGATATGTATG
TTGACGGCCATAAGGCTGCTTCTGACGTTCGTGATGAGTTTGTATCTGTTACTGAGAAGTTAATGGATGA
ATTGGCACAATGCTACAATGTGCTCCCCCAACTTGATATTAATAACACTATAGACCACCGCCCCGAAGGG
GACGAAAAATGGTTTTTAGAGAACGAGAAGACGGTTACGCAGTTTTGCCGCAAGCTGGCTGCTGAACGCC
CTCTTAAGGATATTCGCGATGAGTATAATTACCCCAAAAAGAAAGGTATTAAGGATGAGTGTTCAAGATT
GCTGGAGGCCTCCACTATGAAATCGCGTAGAGGCTTTACTATTCAGCGTTTGATGAATGCAATGCGACAG
GCTCATGCTGATGGTTGGTTTATCGTTTTTGACACTCTCACGTTGGCTGACGACCGATTAGAGGCGTTTT
ATGATAATCCCAATGCTTTGCGTGACTATTTTCGTGATATTGGTCGTATGGTTCTTGCTGCCGAGGGTCG
CAAGGCTAATGATTCACACGCCGACTGCTATCAGTATTTTTGTGTGCCTGAGTATGGTACAGCTAATGGC
CGTCTTCATTTCCATGCGGTGCATTTTATGCGGACACTTCCTACAGGTAGCGTTGACCCTAATTTTGGTC
GTCGGGTACGCAATCGCCGCCAGTTAAATAGCTTGCAAAATACGTGGCCTTATGGTTACAGTATGCCCAT
CGCAGTTCGCTACACGCAGGACGCTTTTTCACGTTCTGGTTGGTTGTGGCCTGTTGATGCTAAAGGTGAG
CCGCTTAAAGCTACCAGTTATATGGCTGTTGGTTTCTATGTGGCTAAATACGTTAACAAAAAGTCAGATA
TGGACCTTGCTGCTAAAGGTCTAGGAGCTAAAGAATGGAACAACTCACTAAAAACCAAGCTGTCGCTACT
TCCCAAGAAGCTGTTCAGAATCAGAATGAGCCGCAACTTCGGGATGAAAATGCTCACAATGACAAATCTG
TCCACGGAGTGCTTAATCCAACTTACCAAGCTGGGTTACGACGCGACGCCGTTCAACCAGATATTGAAGC
AGAACGCAAAAAGAGAGATGAGATTGAGGCTGGGAAAAGTTACTGTAGCCGACGTTTTGGCGGCGCAACC
TGTGACGACAAATCTGCTCAAATTTATGCGCGCTTCGATAAAAATGATTGGCGTATCCAACCTGTA
```

B.4 Consensus sequence of line B1 after 100 passages

```
GAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTCGGATATTTCTGATGAGTCGAAAAATTATCTT
GATAAAGCAGGAATTACTACTGCTTGTTTACGAATTAAATCGAAGTGGACTGCTGGCGGAAAATGAGAAA
ATTCGACCTATCCTTGCGCAGCTCGAGAAGCTCTTACTTTGCGACCTTTCGCCATCAACTAACGATTCTG
TCAAAAACTGACGCGTTGGATGAGGAGAAGTGGCTTAATATGCTTGGCACGTTCGTCAAGGACTGGTTTA
GATATGAGTCACATTTTGTTCATGGTAGAGATTCTCTTGTTGACATTTTAAAAGAGCGTGGATTACTATC
TGAGTCCGATGCTGTTCAACCACTAATAGGTAAGAAATCATGAGTCAAGTTACTGAACAATCCGTACGTT
TCCAGACCGCTTTGGCCTCTATTAAGCTCATTCAGGCTTCTGCCGTTTTGGATTTAACCGAAGATGATTT
CGATTTTCTGACGAGTAACAAAGTTTGGATTGCTACTGACCGCTCTCGTGCTCGTCGCTGCGTTGAGGCT
TGCGTTTATGGTACGCTGGACTTTGTGGGATACCCTCGCTTTCCTGCTCCTGTTGAGTTTATTGCTGCCG
TCATTGCTTATTATGTTCATCCCGTCAACATTCAAACGGCCTGTCTCATCATGGAAGGCGCTGAATTTAC
GGAAAACATTATTAATGGCGTCGAGCGTCCGGTTAAAGCCGCTGAATTGTTCGCGTTTACCTTGCGTGTA
CGCGCAGGAAACACTGACGTTCTTACTGACGCAGAAGAAAACGTGCGTCAAAAATTACGTGCAGAAGGAG
TGATGTAATGTCTAAAGGTAAAAAACGTTCTGGCGCTCGCCCTGGTCGTCCGCAGCCGTTGCGAGGTACT
AAAGGCAAGCGTAAAGGCGCTCGTCTTTGGTATGTAGGTGGTCAACAATTTTAATTGCAGGGGCTTCGGC
CCCTTACTTGAGGATAAATTATGTCTAATATTCAAACTGGCGCCGAGCGTATGCCGCATGACCTTTCCCA
TCTTGGCTTCCTTGCTGGTCAGATTGGTCGTCTTATTACCATTTCAACTACTCCGGTTATCGCTGGCGAC
TCCTTCGAGATGGACGCCGTTGGCGCTCTCCGTCTTTCTCCATTGCGTCGTGGCCTTGCTATTGACTCTA
CTGTAGACATTTTTACTTTTTATGTCCCTCATCGTCACGTTTATGGTGAACAGTGGATTAAGTTCATGAA
GGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTGCTGATTATATTGACCATGCCGCTTTTCTT
GGCACGATTAACCCTGATACCAATAAAATCCCTAAGCATTTGTTTCAGGGTTATTTGAATATCTATAACA
ACTATTTTAAAGCGCCGTGGATGCCTGACCGTACCGAGGCTAACCCTAATGAGCTTAATCAAGATGATGC
TCGTTATGGTTTCCGTTGCTGCCATCTCAAAAACATTTGGACTGCTCCGCTTCCTCCTGAGACTGAGCTT
TCTCGCCAAATGACGACTTCTACCACATCTATTGACATTATGGGTCTGCAAGCTGCTTATGCTAATTTGC
ATACTGACCAAGAACGTGATTACTTCATTCAGCGTTACCGTGATGTTATTTCTTCATTTGGAGGTAAAAC
CTCTTATGACGCTGACAACCGTCCTTTACTTGTCATGCGCTCTAATCTCTGGGCATCTGGCTATGATGTT
GATGGAACTGACCAAACGTCGTTAGGCCAGTTTTCTGGTCGTGTTCAACAGACCTATAAACATTCTGTGC
```

```
CGCGTTTCTTTGTTCCTGAGCATGGCACTATGTTTACTCTTGCGCTTGTTCGTTTTCCGCCTACTGCGAC
TAAAGAGATTCAGTACCTTAACGCTAAAGGTGCTTTGACTTATACCGATATTGCTGGCGACCCTGTTTTG
TATGGCAACTTGCCGCCGCGTGAAATTTCTATGAAGGATGTTTTCCGTTCTGGTGATTCGTCTAAGAAGT
TTAAGATTGCTGAGGGTCAGTGGTATCGTTATGCGCCTTCGTATGTTTCTCCTGCTTATCACCTTCTTGA
AGGCTTCCCATTCATTCAGGAACCGCCTTCTGGTGATTTGCAAGAACGCGTACTTATTCGCCACCATGAT
TATGACCAGTGTTTCCAGTCCGTTCAGTTGTTGCAGTGGAATAGTCAGGTTAAATTTAATGTGACCGTTT
ATCGCAATCTGCCGACCACTCGCGATTCAATCATGACTTCGTGATAAAAGATTGAGTGTGAGGTTATAAC
GCCGAAGCGGTAAAAATTTTAATTTTTGCCGCTGAGGGGTTGACCAAGCGAAGCGCGGTAGGTTTTCTGC
TTAGGAGTTTAATCATGTTTCAGACTTTTATTTCTCGCCATAATTCAAACTTTTTTTCTGATAAGCTGGT
TCTCACTTCTGTTACTCCAGCTTCTTCGGCACCTGTTTTACAGACACCTAAAGCTACATCGTCAACGTTA
TATTTTGATAGTTTGACGGTTAATGCTGGTAATGGTGGTTTTCTTCATTGCATTCAGATGGATACATCTG
TCAACGCCGCTAATCAGGTTGTTTCTGTTGGTGCTGATATTGCTTTTGATGCCGACCCTAAATTTTTTGC
CTGTTTGGTTCGCTTTGAGTCTTCTTCGGTTCCGACTACCCTCCCGACTGCCTATGATGTTTATCCTTTG
AATGGTCGCCATGATGGTGGTTATTATACCGTCAAGGACTGTGTGACTATTGACGTCCTTCCCCGTACGC
CGGGCAATAATGTTTATGTTGGTTTCATGGTTTGGTCTAACTTTACCGCTACTAAATGCCGCGGATTGGT
TTCGCTGAATCAGGTTATTAAAGAGATTATTTGTCTCCAGCCACTTAAGTGAGGTGATTTATGTTTGGTG
CTATTGCTGGCGGTATTGCTTCTGCTCTTGCTGGTGGCGCCATGTCTAAATTGTTTGGAGGCGGTCAAAA
AGCCGCCTCCGGTGGCATTCAAGGTGATGTGCTTGCTACCGATAACAATACTGTAGGCATGGGTGATGCT
GGTATTAAATCTGCCATTCAAGGCTCTAATGTTCCTAACCCTGATGAGGCCGCCCCTAGTTTTGTTTCTG
GTGCTATGGCTAAAGCTGGTAAAGGACTTCTTGAAGGTACGTTGCAGGCTGGCACTTCTGCCGTTTCTGA
TAAGTTGCTTGATTTGGTTGGACTTGGTGGCAAGTCTGCCGCTGATAAAGGAAAGGATACTCGTGATTAT
CTTGCTGCTGCATTTCCTGAGCTTAATGCCTGGGAGCGTGCTGGTGCTGGTGCTTCCTCTGCTGGTATGG
TTGACGCCGGATTTGAGAATCAAAAAGAGCTTACTAAAATGCAACTGGACAATCAGAAAGAGGTTGCCGA
GATGCAAAATGAGACTCAAAAAGAGATTGCTGGCATTCAGTCGGCGACTTCACGCCAGAATACGAAAGAC
CAGGTATATGCACAAAATGAGATGCTTGCTTATCAACAGAAGGAGTCTACTGCTCGCGTTGCGTCTATTA
TGGAAAACACCAATCTTTCCAAGCAACAGCAGGTTTCCGAGATTATGCGCCAAATGCTTACTCAAGCTCA
AACGGCTGGTCAGTATTTTACCAATGACCAAATCAAAGAAATGACTCGCAAGGTTAGTGCTGAGGTTGAC
TTAGTTCATCAGCAAACGCAGAATCAGCGGTATGGCTCTTCTCATATTGGCGCTACTGCAAAGGATATTT
CTAATGTCGTCACTGATGCTGCTTCTGGTGTGGTTGATATTTTTCATGGTATTGATAAAGCTGTTGCCGA
TACTTGGAACAATTTCTGGAAAGACGGTAAAGCTGATGGTATTGGCTCTAATTTGTCTAGGAAATAACCG
TCAGGATTGACACCCTCCCAATTGTATGTTTTCATGCCTCCAAATCTTGGAGGCTTTTTTATGGTTCGTT
CTTATTACCCTTCTGAATGTCACGCTGATTATTTTGACTTTGAGCGTATCGAGGCTCTTAAACCTGCTAT
TGAGGCTTGTGGCATTTCTACTCTTTCTCAATCCCCAATGCTTGGCTTCCATAAGCAGATGGATAACCGC
ATCAAGCTCTTGGAAGAGATTCTGTCTTTTCGTATGCAGGGCGTTGAGTTCGATAATGGTGATATGTATG
TTGACGGCCATAAGGCTGCTTCTGACGTTCGTGATGAGTTTGTATCTGTTACTGAGAAGTTAATGGATGA
ATTGGCACAATGCTACAATGTGCTCCCCCAACTTGATATTAATAACACTATAGACCACCGCCCCGAAGGG
GACGAAAAATGGTTTTTAGAGAACGAGAAGACGGTTACGCAGTTTTGCCGCAAGCTGGCTGCTGAACGCC
CTCTTAAGGATATTCGCGATGAGTATAATTACCCCAAAAAGAAAGGTATTAAGGATGAGTGTTCAAGATT
GCTGGAGGCCTCCACTATGAAATCGCGTAGAGGCTTTACTATTCAGCGTTTGATGAATGCAATGCGACAG
GCTCATGCTGATGGTTGGTTTATCGTTTTTGACACTCTCACGTTGGCTGACGACCGATTAGAGGCGTTTT
ATGATAATCCCAATGCTTTGCGTGACTATTTTCGTGATATTGGTCGTATGGTTCTTGCTGCCGAGGGTCG
CAAGGCTAATGATTCACACGCCGACTGCTATCAGTATTTTTGTGTGCCTGAGTATGGTACAGCTAATGGC
CGTCTTCATTTCCATGCGGTGCATTTTATGCGGACACTTCCTACAGGTAGCGTTGACCCTAATTTTGGTC
GTCGGGTACGCAATCGCCGCCAGTTAAATAGCTTGCAAAATACGTGGCCTTATGGTTACAGTATGCCCAT
CGCAGTTCGCTACACGCAGGACGCTTTTTCACGTTCTGGTTGGTTGTGGCCTGTTGATGCTAAAGGTGAG
CCGCTTAAAGCTACCAGTTATATGGCTGTTGGTTTCTATGTGGCTAAATACGTTAACAAAAAGTCAGATA
TGGACCTTGCTGCTAAAGGTCTAGGAGCTAAAGAATGGAACAACTCACTAAAAACCAAGCTGTCGCTACT
TCCCAAGAAGCTGTTCAGAATCAGAATGAGCCGCAACTTCGGGATGAAAATGCTCACAATGACAAATCTG
TCCACGGAGTGCTTAATCCAACTTACCAAGCTGGGTTACGACGCGACGCCGTTCAACCAGATATTGAAGC
AGAACGCAAAAAGAGAGATGAGATTGAGGCTGGGAAAAGTTACTGTAGCCGACGTTTTGGCGGCGCAACC
TGTGACGACAAATCTGCTCAAATTTATGCGCGCTTCGATAAAAATGATTGGCGTATCCAACCTGCA
```

B.5 Consensus sequence of line B2 after 100 passages

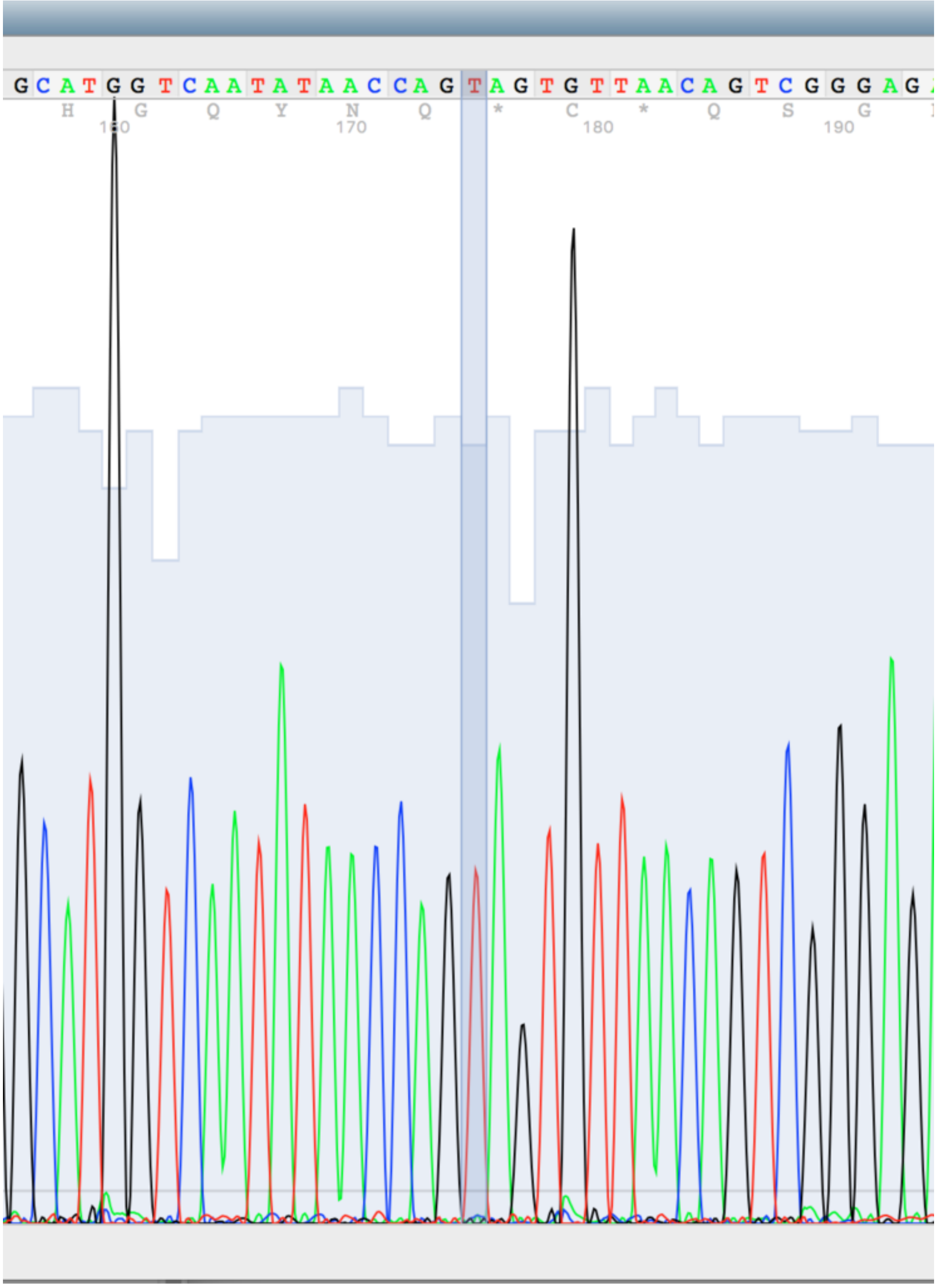
```
GAGTTTTATCGCTTCCATGACGCAGAAGTTAACACTTTCGGATATTTCTGATGAGTCGAAAAATTATCTT
GATAAAGCAGGAATTACTACTGCTTGTTTACGAATTAAATCGAAGTGGACTGCTGGCGGAAAATGAGAAA
ATTCGACCTATCCTTGCGCAGCTCGAGAAGCTCTTACTTTGCGACCTTTCGCCATCAACTAACGATTCTG
TCAAAAACTGACGCGTTGGATGAGGAGAAGTGGCTTAATATGCTTGGCACGTTCGTCAAGGACTGGTTTA
GATATGAGTCACATTTTGTTCATGGTAGAGATTCTCTTGTTGACATTTTAAAAGAGCGTGGATTACTATC
TGAGTCCGATGCTGTTCAACCACTAATAGGTAAGAAATCATGAGTCAAGTTACTGAACAATCCGTACGTT
TCCAGACCGCTTTGGCCTCTATTAAGCTCATTCAGGCTTCTGCCGTTTTGGATTTAACCGAAGATGATTT
CGATTTTCTGACGAGTAACAAAGTTTGGATTGCTACTGACCGCTCTCGTGCTCGTCGCTGCGTTGAGGCT
TGCGTTTATGGTACGCTGGACTTTGTGGGATACCCTCGCTTTCCTGCTCCTGTTGAGTTTATTGCTGCCG
TCATTGCTTATTATGTTCATCCCGTCAACATTCAAACGGCCTGTCTCATCATGGAAGGCGCTGAATTTAC
GGAAAACATTATTAATGGCGTCGAGCGTCCGGTTAAAGCCGCTGAATTGTTCGCGTTTACCTTGCGTGTA
CGCGCAGGAAACACTGACGTTCTTACTGACGCAGAAGAAAACGTGCGTCAAAAATTACGTGCAGAAGGAG
CGATGTAATGTCTAAAGGTAAAAAACGTTCTGGCGCTCGCCCTGGTCGTCCGCAGCCGTTGCGAGGTACT
AAAGGCAAGCGTAAAGGCGCTCGTCTTTGGTATGTAGGTGGTCAACAATTTTAATTGCAGGGGCTTCGGC
CCCTTACTTGAGGATAAATTATGTCTAATATTCAAACTGGCGCCGAGCGTATGCCGCATGACCTTTCCCA
TCTTGGCTTCCTTGCTGGTCAGATTGGTCGTCTTATTACCATTTCAACTACTCCGGTTATCGCTGGCGAC
TCCTTCGAGATGGACGCCGTTGGCGCTCTCCGTCTTTCTCCATTGCGTCGTGGCCTTGCTATTGACTCTA
CTGTAGACATTTTTACTTTTTATGTCCCTCATCGTCACGTTTATGGTGAACAGTGGATTAAGTTCATGAA
GGATGGTGTTAATGCCACTCCTCTCCCGACTGTTAACACTGCTGGTTATATTGACCATACCGCTTTTCTT
GGCACGATTAACCCTGATACCAATAAAATCCCTAAGCATTTGTTTCAGGGTTATTTGAATATCTATAACA
ACTATTTTAAAGCGCCGTGGATGCCTGACCGTACCGAGGCTAACCCTAATGAGCTTAATCAAGATGATGC
TCGTTATGGTTTCCGTTGCTGCCATCTCAAAAACATTTGGACTGCTCCGCTTCCTCCTGAGACTGAGCTT
TCTCGCCAAATGACGACTTCTACCACATCTATTGACATTATGGGTCTGCAAGCTGCTTATGCTAATTTGC
ATACTGACCAAGAACGTGATTACTTCATGCAGCGTTACCGTGATGTTATTTCTTCATTTGGAGGTAAAAC
CTCTTATGACGCTGACAACCGTCCTTTACTTGTCATGCGCTCTAATCTCTGGGCATCTGGCTATGATGTT
GATGGAACTGACCAAACGTCGTTAGGCCAGTTTTCTGGTCGTGTTCAACAGACCTATAAACATTCTGTGC
CGCGTTTCTTTGTTCCTGAGCATGGCACTATGTTTACTCTTGCGCTTGTTCGTTTTCCGCCTACTGCGAC
TAAAGAGATTCAGTACCTTAACGCTAAAGGTGCTTTGACTTATACCGATATTGCTGGCGACCCTGTTTTG
TATGGCAGCTTGCCGCCGCGTGAAATTTCTATGAAGGATGTTTTCCGTTCTGGTGATTCGTCTAAGAAGT
TTAAGATTGCTGAGGGTCAGTGGTATCGTTATGCGCCTTCGTATGTTTCTCCTGCTTATCACCTTCTTGA
AGGCTTCCCATTCATTCAGGAACCGCCTTCTGGTGATTTGCAAGAACGCGTACTTATTCGCCACCATGAT
TATGACCAGTGTTTCCAGTCCGTTCAGTTGTTGCAGTGGAATAGTCAGGTTAAATTTAATGTGACCGTTT
ATCGCAATCTGCCGACCACTCGCGATTCAATCATGACTTCGTGATAAAAGATTGAGTGTGAGGTTATAAC
GCCGAAGCGGTAAAAATTTTAATTTTTGCCGCTGAGGGGTTGACCAAGCGAAGCGCGGTAGGTTTTCTGC
TTAGGAGTTTAATCATGTTTCAGACTTTTATTTCTCGCCATAATTCAAACTTTTTTTCTGATAAGCTGGT
TCTCACTTCTGTTACTCCAGCTTCTTCGGCACCTGTTTTACAGACACCTAAAGCTACATCGTCAACGTTA
TATTTTGATAGTTTGACGGTTAATGCTGGTAATGGTGGTTTTCTTCATTGCATTCAGATGGATACATCTG
TCAACGCCGCTAATCAGGTTGTTTCTGTTGGTGCTGATATTGCTTTTGATGCCGACCCTAAATTTTTTGC
CTGTTTGGTTCGCTTTGAGTCTTCTTCGGTTCCGACTACCCTCCCGACTGCCTATGATGTTTATCCTTTG
AATGGTCGCCATGATGGTGGTTATTATACCGTCAAGGACTGTGTGACTATTGACGTCCTTCCCCGTACGC
CGGGCAATAATGTTTATGTTGGTTTCATGGTTTGGTCTAACTTTACCGCTACTAAATGCCGCGGATTGGT
TTCGCTGAATCAGGTTATTAAAGAGATTATTTGTCTCCAGCCACTTAAGTGAGGTGATTTATGTTTGGTG
CTATTGCTGGCGGTATTGCTTCTGCTCTTGCTGGTGGCGCCATGTCTAAATTGTTTGGAGGCGGTCAAAA
AGCCGCCTCCGGTGGCATTCAAGGTGATGTGCTTGCTACCGATAACAATACTGTAGGCATGGGTGATGCT
GGTATTAAATCTGCCATTCAAGGCTCTAATGTTCCTAACCCTGATGAGGCCGCCCCTAGTTTTGTTTCTG
GTGCTATGGCTAAAGCTGGTAAAGGACTTCTTGAAGGTACGTTGCAGGCTGGCACTTCTGCCGTTTCTGA
TAAGTTGCTTGATTTGGTTGGACTTGGTGGCAAGTCTGCCGCTGATAAAGGAAAGGATACTCGTGATTAT
CTTGCTGCTGCATTTCCTGAGCTTAATGCCTGGGAGCGTGCTGGTGCTGGTGCTTCCTCTGCTGGTATGG
TTGACGCCGGATTTGAGAATCAAAAAGAGCTTACTAAAATGCAACTGGACAATCAGAAAGAGATTTCCGA
GATGCAAAATGAGACTCAAAAAGAGATTGCTGGCATTCAGTCGGCGACTTCACGCCAGAATACGAAAGAC
CAGGTATATGCACAAAATGAGATGCTTGCTTATCAACAGAAGGAGTCTACTGCTCGCGTTGCGTCTATTA
TGGAAAACACCAATCTTTCCAAGCAACAGCAGGTTTCCGAGATTATGCGCCAAATGCTTACTCAAGCTCA
AACGGCTGGTCAGTATTTTACCAATGACCAAATCAAAGAAATGACTCGCAAGGTTAGTGCTGAGGTTGAC
```

```
TTAGTTCATCAGCAAACGCAGAATCAGCGGTATGGCTCTTCTCATATTGGCGCTACTGCAAAGGATATTT
CTAATGTCGTCACTGATGCTGCTTCTGGTGTGGTTGATATTTTTCATGGTATTGATAAAGCTGTTGCCGA
TACTTGGAACAATTTCTGGAAAGACGGTAAAGCTGATGGTATTGGCTCTAATTTGTCTAGGAAATAACCG
TCAGGATTGACACCCTCCCAATTGTATGTTTTCATGCCTCCAAATCTTGGAGGCTTTTTTATGGTTCGTT
CTTATTACCCTTCTGAATGTCACGCTGATTATTTTGACTTTGAGCGTATCGAGGCTCTTAAACCTGCTAT
TGAGGCTTGTGGCATTTCTACTCTTTCTCAATCCCCAATGCTTGGCTTCCATAAGCAGATGGATAACCGC
ATCAAGCTCTTGGAAGAGATTCTGTCTTTTCGTATGCAGGGCGTTGAGTTCGATAATGGTGATATGTATG
TTGACGGCCATAAGGCTGCTTCTGACGTTCGTGATGAGTTTGTATCTGTTACTGAGAAGTTAATGGATGA
ATTGGCACAATGCTACAATGTGCTCCCCCAACTTGATATTAATAACACTATAGACCACCGCCCCGAAGGG
GACGAAAAATGGTTTTTAGAGAACGAGAAGACGGTTACGCAGTTTTGCCGCAAGCTGGCTGCTGAACGCC
CTCTTAAGGATATTCGCGATGAGTATAATTACCCCAAAAAGAAAGGTATTAAGGATGAGTGTTCAAGATT
GCTGGAGGCCTCCACTATGAAATCGCGTAGAGGCTTTACTATTCAGCGTTTGATGAATGCAATGCGACAG
GCTCATGCTGATGGTTGGTTTATCGTTTTTGACACTCTCACGTTGGCTGACGACCGATTAGAGGCGTTTT
ATGATAATCCCAATGCTTTGCGTGACTATTTTCGTGATATTGGTCGTATGGTTCTTGCTGCCGAGGGTCG
CAAGGCTAATGATTCACACGCCGACTGCTATCAGTATTTTTGTGTGCCTGAGTATGGTACAGCTAATGGC
CGTCTTCATTTCCATGCGGTGCATTTTATGCGGACACTTCCCACAGGTAGCGTTGACCCTAATTTTGGTC
GTCGGGTACGCAATCGCCGCCAGTTAAATAGCTTGCAAAATACGTGGCCTTATGGTTACAGTATGCCCAT
CGCAGTTCGCTACACGCAGGACGCTTTTTCACGTTCTGGTTGGTTGTGGCCTGTTGATGCTAAAGGTGAG
CCGCTTAAAGCTACCAGTTATATGGCTGTTGGTTTCTATGTGGCTAAATACGTTAACAAAAAGTCAGATA
TGGACCTTGCTGCTAAAGGTCTAGGAGCTAAAGAATGGAACAACTCACTAAAAACCAAGCTGTCGCTACT
TCCCAAGAAGCTGTTCAGAATCAGAATGAGCCGCAACTTCGGGATGAAAATGCTCACAATGACAAATCTG
TCCACGGAGTGCTTAATCCAACTTACCAAGCTGGGTTACGACGCGACGCCGTTCAACCAGATATTGAAGC
AGAACGCAAAAAGAGAGATGAGATTGAGGCTGGGAAAAGTTACTGTAGCCGACGTTTTGGCGGCGCAACC
TGTGACGACAAATCTGCTCAAATTTATGCGCGCTTCGATAAAAATGATTGGCGTATCCAACCTGTA
```

Appendix C – Miscellaneous

C.1 Initial phage preparation

The initial phage preparation was derived from a single plaque, then amplified by growing in media with its host (chapter 4.2.3). Illumina sequencing found this preparation contained a mutation at position 1301 with 48.8% frequency (chapter 4.3.1). Sanger sequencing was performed on the original plaque. This screenshot of the chromatogram (from 4peaks) shows that this mutation was not present in the original plaque (highlighted base), so probably arose during the amplification step.


GCATGGTCAATATAACCAGTAGTGTTAACAGTCGGGAG
H G Q Y N Q * C * Q S G