Adaptation of Bacteriophage to Variable Environments

Oyeronke Temidayo Ayansola

A thesis submitted in partial fulfilment of the requirements of

Nottingham Trent University for the degree of

Doctor of Philosophy

January 2020

Copyright Statement

Oyeronke Temidayo Ayansola

Acknowledgements

First and foremost, I thank God for this opportunity and granting me strength and patience to proceed successfully with this work. My sincere gratitude goes to my supervisor, Dr Benjamin Dickins, I cannot thank him enough for his efforts, invaluable guidance, advice, motivation and immense knowledge throughout the stages of this research and whose impact has made me to be a good scientist.

I express my profound gratitude to all my supervisory team both present: Prof Graham Ball, Dr Gareth McVicker, Dr Michael Loughlin, and past: Dr Alan McNally (University of Birmingham), and Dr Gina Manning (University of Wolverhampton) for their immense support and assistance in this project. Many thanks to the members of the Pathogen Research Group: Dr Jonathan Thomas – for Nanopore sequencing guidance, Dr Jody Winter, Dr Samantha McLean, Dr Naqash Masood, and Dr David Negus their advice and contributions during micro-lab meetings.

I am grateful to Nick and Susan (former NTU super lab staff) for their support and advice in designing my first chemostat. I would like to thank my friends and colleagues in CELS micro lab at NTU for their motivation and encouragement. To Dr Wilcox (University of California, Davis) thank you for help. To my friend Odette, thank you. My thanks also goes to Dr Dannise Ruiz (USA) for her advice and support along this journey.

My sincerest and deepest gratitude to my dear husband, for your encouragement, support, trust and understanding, and I know you are more proud of my achievements than I am! To my angels, thank you for your understanding and patience. I would like to thank my parents for their unending support, prayers and encouragement. My sincere gratitude also goes to my friends, families, siblings, and in-laws.

Table of Contents

Lists of abbreviations

ABI	Applied Biosystems
ATM	Amplicon Tagment Mix
bp	Base pairs
BLT	Bead-linked transposome
Core OS	Core oligosaccharide
CSV	Comma-separated values
CRISPR	Clustered regularly interspaced short palindromic repeats
dNTPs	Deoxynucleoside triphosphates
ddNTPs	Dideoxyribonucleoside triphosphates
ddPCR	Droplet digital PCR
DMSO	Dimethyl sulfoxide
DOM	Dissolved organic matter
dsDNA	double-stranded DNA
EB	Elution buffer
EPM	Enhanced PCR mix
gDNA	Genomic DNA
HBV	Hepatitis B virus
HIV	Human immunodeficiency virus
HT	Hybridization Buffer
LB	Lysogeny broth
LLB	Loading Beads
LPS	Lipopolysaccharide
LTEE	Long-term evolution experiment
MCPyV	Merkel cell polyomavirus
MOI	Multiplicity of infection
NPM	Nextera PCR Master mix
NT	Neutralised Tagment buffer
$OD600_{nm}$	Optical density
OM	Outer membrane

PCR-MM	PCR master mix
PFU	Plaque-forming units
pgu	Phage genome units
POM	Particulate organic matter
PTFE	Polytetrafluoroethylene
PyV	Polyomavirus
qPCR	Quantitative PCR
QUAST	Quality assessment tool for genome assemblies
RBS1	Nextera Resuspension Buffer
RBS	Resuspension Buffer
RF dsDNA	Replicative form double stranded DNA
SARS	Severe acute respiratory syndrome
SDM	Site directed mutagenesis
SIVs	Simians immunodeficiency viruses
s/n/r	Substitutions per nucleotide per cell infection
SPB	Sample purification beads
SPB-MM	SPB master mix
ssb	ssDNA binding protein
ssDNA	single-stranded DNA
SYBR	Safe DNA Gel Stain
TB	Tagmentation buffer
TD	Tagment DNA buffer
TMM	Tagmentation master mix
TSB	Tagment stop buffer
TSV	Tab-separated
TWB	Tagment wash buffer

Abstract

Because a virus is an obligate cellular parasite, the host is a key part of its environment. Viruses may expand their ecological niche by switching host. Successful host switching can be influenced by ecological and evolutionary factors, genetic constraints and fitness within new hosts. An outcome of host switching is reduced fitness exhibited by viruses, a phenomenon observed in the evolution of viral disease emergence and resistance. To understand the genetic basis of this cost, investigations are required at the genotypic and phenotypic level.

A host switching paradigm was developed using the model bacteriophage ϕX174 which was propagated with its laboratory bacterial host *Escherichia coli* C and with the novel host *Salmonella* enterica serovar Typhimurium, LT2 strain IJ750 or *Escherichia coli* K-12 mutant strain JWO196-2 designated as *E. coli* K-12$^{gmhB\text{-mut}}$. A chemostat was used to achieve steady-state conditions for propagation of ϕX174 and bacterial cells. Two experiments were performed using this approach. In the first, ϕX174 was cultured on *E. coli* K-12$^{gmhB\text{-mut}}$ for 3 days (~206 generations). In the second experiment, ϕX174 was cultured on *E. coli* C and *S.* Typhimurium for four consecutive periods of 10 days (~720 generations), alternating between the two hosts.

For the second chemostat experiment, the fitness and attachment rates of each viral population were measured using qPCR in liquid culture in order to identify and characterise fitness costs associated with host-switching. Deep sequencing of chemostat samples was also carried out to identify allelic changes occurring before and after host switches. Viral samples were chosen to capture substitutions associated with each host across the experiment (which might explain observed changes in fitness) and time series were picked to identify the dynamics of adaptation on a new host. Bacterial host strains were not sampled in this study.

The phenotype measures indicated the pleiotropic costs of host switching, that is a reduction in phage fitness was observed when this was tested on the host used prior to switching, and this may be explained by changes in the attachment rate. The genotype data revealed sets of changes that could be identified as signatures of adaptation to each host, although control data indicate that these may arise during DNA preparation, implicating synthesis of replicative form DNA in the host as a source of selective constraint. Some host-specific alleles and some shared alleles were identified and their fitness effects were examined in isolation after reconstruction of these alleles in the ancestor via targeted mutagenesis. The fitness effects observed for reconstructed mutants were in the direction expected although they do not fully account for the observed costs of host switching.

By analysing different phenotypes and genotypes produced during evolution, a detailed view of ϕX174's adaptation to different hosts was obtained. The results support the idea that costs associated with pathogen-host adaptation may be host-specific, associated with specific mutations, acquired early and persist. Examining these is relevant for understanding emerging infectious diseases.

Chapter One: Introduction

1.1 Bacteriophages

Bacteriophages (phages) are viruses of bacteria or archaea. In 1896, an English chemist, Ernest Hankin, provided a hint of phage-like bactericidal activity prior to its discovery over 100 years ago by Frederick William Twort (an English bacteriologist) in 1915, and, Felix d'Herelle (a French-Canadian scientist) in 1917 (reviewed in Abedon *et al*., 2011). Felix d'Herelle recognised that the particle was growing at the expense of bacteria and coined the name bacteriophages, "phage" coming from the Greek word "phagein" meaning "to eat". These discoveries gave rise to golden era of bacteriophage research from 1930 to 1970, resulting in many other discoveries including DNA as the genetic material, messenger RNA, and the determination of the genetic code. All these led to the new science of Molecular Biology. In the 1980s and 1990s phages were applied substantively in biotechnology research and industry (Abedon, 2009).

1.1.1 Importance of phages

1.1.1.1 Phages in biotechnology and health

We are in a new period of phage research due to improvements in molecular biology methods and breakthroughs in sequencing technologies. Phage and phage-derived products are important in health care research through their applications in phage therapy, an alternative to antibiotic treatment (reviewed in Górski *et al*., 2016). They are used as vaccines or as vaccine carriers (Adhya *et al*., 2014) in health care contexts also. For biotechnology applications, phages are useful in food bio-preservation and safety (Henry and Debarbieux, 2012), biofilm and bacterial growth control (Ródriguez-Rubio *et al*., 2016), plant pathogen biocontrol, as nanocages for gene delivery (Harada *et al*., 2018) and in phage display (Huang *et al*., 2012). In addition, phage-host interactions have recently influenced our genome-engineering capability owing to the discovery and dissection of the clustered regularly interspaced short palindromic repeats (CRISPR)-Cas phage resistance system (Szczepankowska 2012). While the well-studied biology of

phages has facilitated various applications of phages to industry and research, phages have also provided insights into genomic and adaptive evolution. This study fits into this latter area of interest.

1.1.1.2 Phage ecology

Phages have great ecological importance. Phages are strictly obligate parasites, require prokaryotic hosts for reproduction through different lifestyles and transmissions. Phages may develop capacity to integrate into the host genomes via lysogeny and pseudolysogeny, establishing a long-term association with their hosts and reproducing along with host cells. These types of phages are referred to as temperate (termed 'prophages' while inside the host cell) or filamentous (secreted into the environment) phages. Some phages lifestyle (lytic phages) require replicating within host cells and must kill their host via lyses to transmit to the next host cell (Clokie *et al*., 2011). The result of lytic lifestyle is catastrophic resulting in the death of bacterial host. Typically, phages replicate at a faster rate more than their hosts, the question arises how phages have not killed all host cells and drive their own existence into extinction? It is therefore evident that somehow, phages coexist with host cells.

Phage-host coexistence is an important aspect of ecological and evolutional studies (Weitz and Dushoff, 2008), focusing on the interactions between phages, hosts and environments. A number of hypotheses have been proposed to explain phage-host coexistence including arms race coevolution (section 1.3.5.1) and density dependent 'kill-the-winner' model. The model 'kill-the-winner' is explained by antagonist coexistence when phage-bacterial cells exhibit prey-predator density-dependent cycle. Phages infecting susceptible hosts are more abundance (many virions are produced per cell, refers to as virus burst size). The increase in density drives the hosts to be resistance to phage infection, thus ecosystem selects for resistance hosts. Consequently, phages reduce in abundance because they are unable to infect resistance hosts (Schwartz and Lindell, 2017). In another perspective

viral-like particles showed more consistency with temperate lifestyle than lytic (Knowles *et al*., 2016). Knowles *et al*. (2016) came up with an extension of 'Kill-the-winner' hypothesis termed 'Piggyback-the-winner' where viruses exploit hosts through temperate lifestyle rather than killing the host cells and resulting in the abundance of host cells. In the same study, using coral reefs viruses, hosts ecosystem experiments, publicly available meta-analyses and metagenomics, the high rise in contribution of temperate phages to abundance of host densities was determined. They proposed that, decrease in virus to host ratio (with increase in host density) is consistent with the reduction of lytic lifestyle and demonstrated that evidences from literature metagenomics showed relatively high temperate viruses in ecosystems with high microbial densities. The outcome of 'kill-the-winner' hypothesis is that viruses suppresses bloom of hosts, increase diversity of hosts and speeding up evolution in both viruses and hosts (Middelboe *et al*., 2009; Paterson *et al*., 2010). 'Kill-the-winner' hypothesis has been demonstrated as a mechanism that influence microbial communities, ranging from aquatic ecosystem to human body (Barr *et al*., 2013; Knowles *et al*., 2016).

Phages are found on human body niches with body surfaces that are in direct contact with the environment such as skin and mucosal surfaces harbouring most of them. The presence of phages on mucosal surfaces have been demonstrated to be even much more abundant. Many phages that colonise mucosal surfaces encodes surface proteins on the outer capsid which helps in cell surface adhesion to mucus components. This interaction maintains adherence and enhance abundance of phages in mucosal linings where they protect epithelium from bacterial infection (Barr *et al*., 2013). In this way, phages control bacterial host abundance through density dependent lytic-host predation as described in 'kill-the-winner' hypothesis.

The paradox of phage-host co-existence is more pronounced in aquatic ecosystem. In marine environments, phages were found to be numerically dominant. While bacterial cells are abundant on this planet (numbering

~10^{30}; Whitman *et al*., 1998), phages may be 10 or more times more abundant (Abedon *et al*., 2008). Since they are widespread in the aquatic ecosystem, phages contribute to ecosystem productivity by playing a vital role in the cycling of biogeochemical elements. They lyse their microbial host cells, the dominant biological biomass in the pelagic food web. Lysed hosts contribute to particulate organic matter (POM) and dissolved organic matter (DOM) in the oceans. The POM and nutrients are made available to osmotrophs (organisms which take up nutrients through their cell membranes) rather than microbes being directly consumed in the grazing food web (Abedon *et al*., 2008). In addition, virus-infected cells may sink faster in the aquatic environment compared to non-infected cells, making microbial cells available in the deep sea (Sime-Ngando, 2014), contributing to high turnover of nutrients and influencing organic-carbon release. In addition to ecological importance of phages, they play a major role in microbial evolution.

1.2 Virus evolution

1.2.1 Phage evolution and divergence

Evolution of phages is complex. Due to the complexity in their host range, lifecycle and morphology classification, there have been conflicting conclusions as regards to the origin and evolution of phages (Coetzee, 1987). This may be attributed to phages numerical dominance (section 1.1.1.2) and are virtually found everywhere bacterial hosts exist. A solution to the complexity is grouping phages according to sequence similarities and gene organisation. For better organisation and understanding of phage-hosts evolution, a model was proposed where the genetic structures and dynamics of ds (double-stranded) DNA phages with their respective host prophage genomes were mosaics via horizontal gene transfer (although not uniform for all phages due to differences in the extent of exchange of information) into a common gene pool. The result showed that phage evolution is complex and

phage genomes consist of genes that are diverse with their own evolutionary histories (Hendrix *et al*., 1999). Because phages are obligate parasites, and coexist with bacterial hosts in nature, horizontal gene transfer from phage to phage and from phage to bacteria can occur. Many studies have demonstrated horizontal gene transfer among tail fiber genes (Sandmeier, 1992; Haggard-Ljungquist *et al*., 1992). Sandmeier *et al*. (1992) compared the DNA sequences of Plasmid p15B (prophage-related found in *E. coli* 15T) and P1 phage (temperate phages of *Enterobacteria*). The study found out that p15B and P1 have undergone series of sequence acquisitions and deletions with sequence homology in; tail fibre of T4, P2, Mu and lambda phages, some sequence inversions in p15B, viral element e14 of *E. coli* K-12 and P1. The same study concluded that the evolution of phage tail fiber may be as a result of horizontal gene transfer. In a similar study, tail fiber protein homology was found in phages of different families (Mu, P1, K3, T2, lambda, TuIa, TuIb and T4) and evident from DNA sequence recombination (Haggard-Ljungquist *et al*., 1992). Through examination of sequence similarities among dsDNA phages and prophages of bacterial hosts, it is possible that dsDNA phages have a common ancestry (Hendrix *et al*., 1999), and codiverge during evolution via differences in hosts, genomes and lifecycles (Mavrich and Hatfull, 2017). Mavrich and Hatfull (2017) examined evolution driven by horizontal gene transfer between dsDNA phages and host genomes using alignment based approaches. Evidence showed that phage evolution is in two distinct modes according to lifestyles and depending on host phylum; temperate phages were distributed into high and low gene flux modes, lytic phages into lower gene flux modes. Therefore, it is possible that phage evolution and divergence is intertwined with the evolution of their host and possibly the evolution of host is also affected by evolution of their respective phages.

1.2.2 Phage coevolution

The reciprocal evolution between phages and their respective bacterial hosts is termed coevolution. Coevolution is a phenomenon seen as an important driver in maintaining diversity in microbial communities, impact community structure and ecosystem, shape the evolution of phage and bacterial strains while increasing their rate of evolution and divergence (Koskella and Brockhurst, 2014). Coevolution studies explore the interaction between phage and bacteria, in the gut (Sordi *et al.*, 2017; Sordi *et al.*, 2019)in the laboratory, and in natural communities (reviewed in Koskella and Brockhurst, 2014), giving more insight to ecology and evolution of phages.

It is possible to develop an experimental system that explore phage evolution only. While coevolution studies involve experiments in which both bacterial hosts and phages are allowed to acquire adaptations and counter-adaptations, phage evolution experimental system (phage evolution only) reduce the evolution of bacterial cells. This study explores evolution of phages only, although evolution of bacterial cells may be inevitable but kept at the very minimal (section 3.3.1.4.1).

1.2.3 Phages: models of virus evolution

The ability of phages to evolve in their host environment has become an exciting facet of phage biology. Phages have been widely used as model organisms for the study of virus evolution (reviewed in Dennehy, 2009). It is important to note that in utilising phage as a model for viral evolution, the host is unicellular therefore phage are only a partial model of metazoan infections or zoonoses. Nevertheless, phage provide a convenient model for studying many aspects of viruses and viral infection. One can study with ease large populations of phages and bacterial cells, exploring different biological scenarios such as the rate of evolution, prevalence of epistasis, the ecological bases of selection, the nature of adaptive walks, the effects of

different population sizes and the evolutionary dynamics of adaption. All these are more laborious and may be unethical for analogous studies with animals. One can also achieve control over the environment to a large extent, controlling some variables that may be difficult or impossible to control in animal or plant experiments.

It is possible to detect the evolution of phages in 'real time', tracking changes occurring per generation with experimental evolution – studying evolutionary dynamics through controlled field manipulations or laboratory experiments. Quantitative assays of phages and their hosts are rapid, convenient and accurate. Phages possess relatively small, well-understood genomes that are amenable to whole-genome sequencing as well as to molecular genetic manipulations such as site-directed mutagenesis. Because they share fundamental features with pathogenic plant, animal and human viruses (Dennehy, 2009), phages have become a highly tractable model system for viral evolution.

A pertinent feature of viral evolution is the expansion of viral host range. Ecological factors including the local ecosystem which are; biotic and abiotic factors, abundance and distribution of other organisms, microbial population, complexity and interaction of the ecosystem - host composition (densities and physiological state of host) and modes of transmission, influence viral adaptive potential. These factors may determine the ability of viruses to colonise entirely new environments in a process termed host switching or, alternatively, limit some lineages to co-divergence with existing hosts.

1.2.4 Viral co-divergence

Pathogens, including viruses, are Darwinian systems, possessing their own evolutionary capabilities, and not simply following host evolutionary history (Sabrina *et al*., 2015). However, some viruses have evolved with their hosts during macro-evolution, with only minor changes accompanying hosts' divergence into new species in a process called co-divergence (figure 1.1 A).

An example is the *Merkel cell Polyomavirus* (*MCPyV*) which infects African great apes and has been hypothesized to have co-diverged with its hosts for at least half a billion years (Buck *et al*., 2016). Madinda *et al*. (2016) were exploring the degree to which co-divergence explained patterns of diversity. The authors sought to address the question by characterising genetic diversity of *MCPyV* in seven African great ape taxa. The results showed evidence of synchronous host and *Polyomavirus* (*PyV*) divergence, suggesting co-divergence as the main process during evolution of PyV. Another example is the hepatitis B virus (HBV) which differs among non-human primates and between non-human primates and humans. For instance, the complete genome of chimpanzee HBV shares 90.3% nucleic acid with gibbon HBV, indicating a distinct species-specific variant of HBV despite overlapping ecological niche (Norder *et al*., 1996). These different, yet closely related, versions of HBV indicate a likelihood of co-divergence from a shared ancestor. It has been observed that co-divergence occurs infrequently, affecting mostly double-stranded DNA (dsDNA) viruses of *Poxviridae*, *Papillomaviridae*, *Hepadnaviridae* and *Adenoviridae* as opposed to other virus families (Geoghegan *et al*., 2017).

1.2.5 Viral host switching

In contrast to the mode of evolution described in section 1.2.4, viruses may colonise novel hosts in a process termed cross-species transmission. Cross-species transmission involves host switching by pathogens. Host switching occurs when a pathogen that colonises one host species switches to another host species (figure 1.1 B). Viral host switching may allow infection of closely related hosts or of hosts unrelated to the former one.

Host switching is common over macroevolutionary time scales relative to co-divergence, constituting a near universal feature of viruses and playing a major role in virus-host evolution (Geoghengan *et al*., 2017). Geoghegan *et al*. (2017) studied and compared the frequencies of co-divergence and host switching within viral families by analysing co-phylogenetic processes in virus

families and their hosts. The authors' analyses suggest that while co-divergence occurs among certain virus families, host switching potential is very high in all the 19 virus families studied. Over shorter time scales, host switching may still be rare because, as parasites, viruses depend on other organisms during their entire life course or for critical aspects of their life history. For example, viruses depend on host mechanisms for survival. They may adapt to overcome the host environment limitations, over longer time scales. During adaptation viruses may evolve to increase host range, or acquire mutations that enhance fitness in other hosts. It is unclear whether mutations that permit progeny viruses to cross into novel hosts are frequent compared to mutations that increase fitness on the current host (for co-divergence with original host), an area of interest in this study.

Pathogens crossing species barriers to infect novel hosts can affect agriculture, wildlife as well as human and animal health. For viruses, survival depend on the ability to infect susceptible host. Viruses therefore require the maximization of viral infectivity towards new environment encountered including in complex ecosystem. In complex systems, viruses may co-exist with other microbes to maximize viral infection (Paterson *et al*., 2010). In a complex environment, microbial diversity may favour viral host switching. As an example, in the complex environment of the intestinal microbiota, interactions between a phage (P10), a sensitive bacterial host (*E. coli* LF82) and phage-insensitive bacterial host (*E. coli* MG1655) was explored under three different conditions: *in vitro*, in mice gut with only *E. coli* LF82 and MG1655 and in the gut of conventional mice with a complete intestinal microbial flora. The P10 phage evolved to infect the resistant bacterial host *E. coli* MG1655 in the gut of the convectional host (not in other conditions), via host switching by homologous intragenomic recombination and point mutation in the phage tail fibre (De Sordi *et al*., 2017).

Emerging diseases may be defined as newly discovered diseases that are increasing in frequency from a reservoir. Re-emerging diseases are those

that, while previously declining in frequency, are now re-occurring such that they may become a significant threat to health. The frequency of host switching events is a major concern for the surveillance of emerging viral diseases, with most emerging viral diseases seemingly resulting from host switching (Geoghegan *et al*., 2017). Several emerging and re-emerging viral diseases affecting humans are associated with host switching events (Wolfe *et al*., 2007; Devaux, 2012). Often, closely related species are more vulnerable to infections via host switching (Longdon *et al*., 2018).

As an example, human immunodeficiency virus (HIV) a retrovirus, evolving at a high rate, with origin traced back to non-human primate monkeys and chimpanzees. The closest relatives of HIV, the Simian immunodeficiency viruses (SIVs; RNA lentiviruses), are also primate viruses. Primates are the natural reservoir and hosts of SIV (Sharp *et al*., 2010). Many SIVs lineage have been identified such as SIV *cpz*, and SIV *gor*. Chimpanzees are natural reservoirs of SIV *cpz* while SIV *gor* was detected in gorilla. Takehisa *et al*. (2009), suggested that SIV *gor* was acquired through a single cross-transmission event from chimpanzee to gorilla. It is unclear whether all SIV lineages are pathogenic or non-pathogenic in their natural hosts, however, studies have shown that SIV *cpz* is pathogenic in chimpanzees, affecting their reproduction, health and life span (Rudicell *et al*., 2010; Soto *et al*., 2010). Due to close proximity of human with primates in sub-Saharan Africa, SIV evolved to infect human. Chimpanzees acquired two forms of SIV, which recombined and evolved to form a unique virus of different genome structure, which in turn infected humans (Sharp *et al*., 2010).

More recently, a highly lethal virus, the Ebola virus (RNA virus) emerged. Ebola was first discovered in 1976, originally known to infect only monkeys (Miranda *et al*., 1996) and spread to other animal species most especially non-human primates (Osterholm *et al*., 2015). Through contact with infected animals, for instance, killing bats (a putative reservoir of Ebola virus) for food, animal husbandry and keeping monkeys as pet, Ebola virus disease

emerged in humans (Leroy *et al*., 2009). In 2001, human outbreaks was recorded in Gabon and Republic of Congo (Osterholm *et al*., 2015). More recently, human outbreaks was recorded in Guinea in 2013 (Baize *et al*., 2014), spreading to Liberia, Sierra Leone and parts of central Africa up until late 2018 (World Health Organisation, 2019). The Ebola virus outbreak in West Africa was unprecedented in terms of the total number of humans infected and its overall duration (from 2014 to 2019) with fatality rates varying from 25 to 90 % in past outbreaks. Most recent outbreak occurred in Democratic republic of Congo with 3,300 cases confirmed and caused more than 2,200 deaths between August 2018 to December 2019 (World Health Organisation, 2019).

Zika virus (RNA virus), was first discovered in Uganda in sentinel rhesus monkeys in 1947 and was isolated from an infected human in 1954. It has remained obscure until recent outbreaks began in 2007. Ever since then, the ongoing outbreak and has spread to more than 87 countries (World Health Organisation, 2019) around the world, and has been associated with Guillain-Barre syndrome and congenital microcephaly (in offspring of infected individuals; Weaver *et al*., 2016). Evidence of *Aedes aegypti* as a vector has been established in 61 countries. In 2018, 17 cases was recorded in Africa, 31, 587 in America and 290 cases in South-East Asia Region (World Health Organisation, 2019).

Host switching is common not only among closely related hosts, but also between hosts separated by larger phylogenetic distances. For example, the Nipah virus (RNA virus), outbreak in 1999 of severe encephalitis in humans was recorded to be pathogen of pigs, infecting people with close contact to pigs in Singapore and Malaysia (Chua *et al*., 2000). Measles virus, (RNA virus), a common infection in children, has an origin in humans that was traced back to ~1001 – 1200 AD and believed to be a result of viral evolution in an environment where humans and cattle lived in close proximity (Furuse *et al*., 2010). The severe acute respiratory syndrome (SARS) coronaviruses

(RNA virus) outbreak occurred in 2002 and 2003 where 8096 cases and 774 deaths were reported (WHO, 2014). SARS was directly linked to a member of the coronavirus group closely related to SARS-like coronaviruses isolated from bats identified as SARS reservoirs (Li *et al*., 2005). Using reverse transcription-PCR, sequencing and phylogenetic analysis, coronavirus-like were isolated in a live-animal market in China from raccoon dog, Himalayan palm civets and humans workers in the same market, suggesting host switching from bats to humans (Guan *et al*., 2003). Lau *et al*. (2005), identified coronavirus-like in Chinese horseshoe bats closely related to SARS coronavirus from human via sequencing and phylogenetic analysis.

In addition, the influenza A virus (RNA virus), a human pandemic infection, has been reported to have originated from avian hosts. The H1N1 subtype, termed 'Spanish flu', probably originated in birds in which the infection is asymptomatic (Webby and Webster, 2001), also, the H5N1 subtype, which has led to occasional outbreaks in humans, has been traced to chickens and ducks and became infectious as result of genetic re-assortment (Li *et al*., 2004). Genetic re-assortment occurs in influenza viruses when different viruses co-infect the same cell and exchange gene segments. Re-assortment may be detected via phylogenetic analysis (Steel and Lowen, 2014). It differs from co-evolution (section 1.2.2) where two or more species evolve and one species adapt to match changes in the other (and vice versa).

Aside from emerging viral diseases, host switching events have been attributed to ecological biological invasions, causing deleterious effects and seemingly resulting in declines or extinctions of species. Thus, preventing establishment of exotic species, or caused decline in established populations of exotic species (Faillace *et al*., 2017).

Viruses may invade, establish and expand in novel hosts in a few different ways including recombining and transferring their genetic content into the

host genome, exhibiting high mutation rates, or altering infectivity and virulence with relatively few mutations (Longdon *et al*., 2014). Acquisition of mutations plays an important role in emerging viral diseases, with viruses thereby acquiring the ability to infect new host populations (Nichol *et al*., 2000). Viruses rapidly mutate, enhancing their potential to adapt quickly to new hosts and providing opportunities not readily available to other parasites during novel host colonisation.

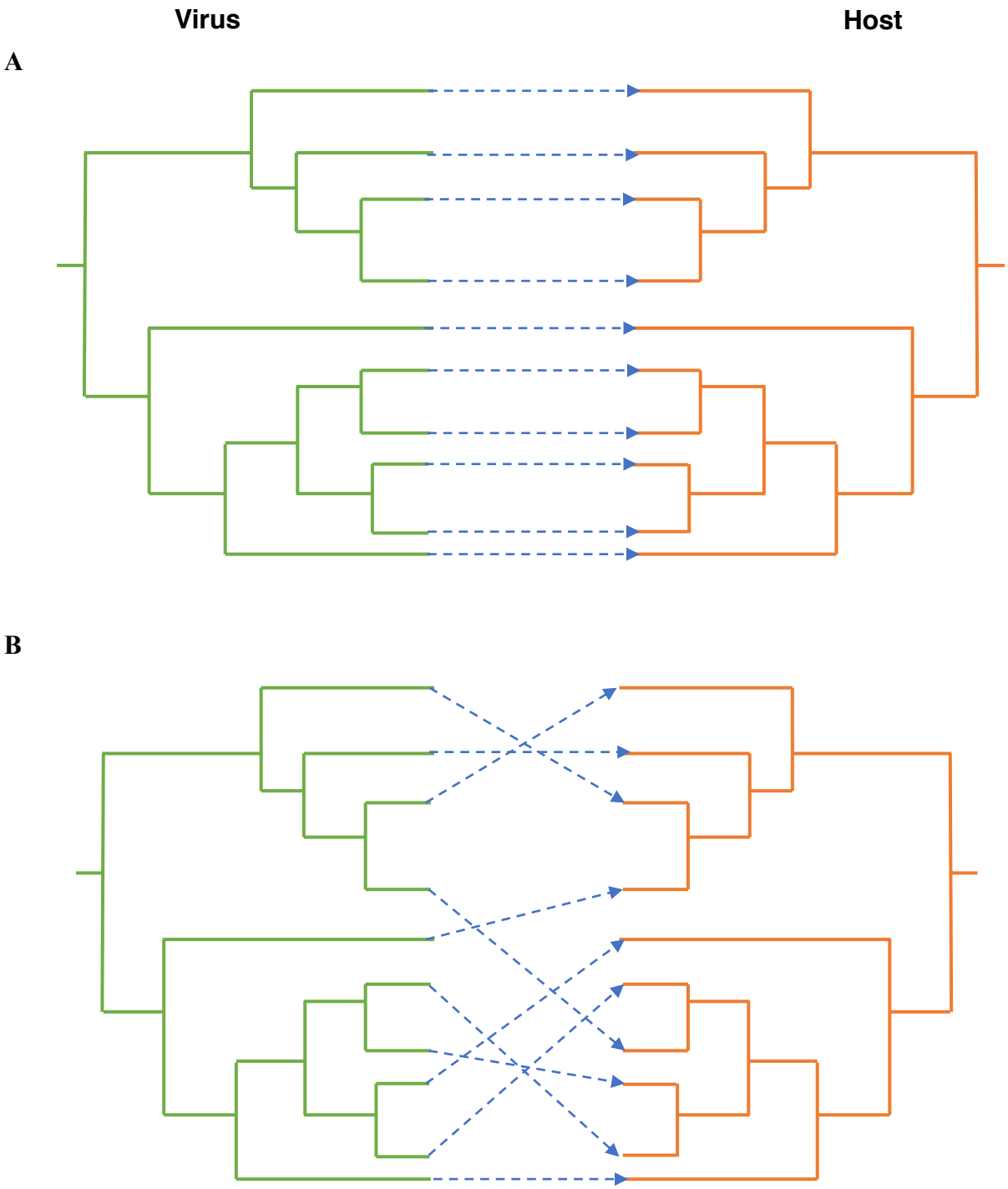


Figure 1.1: Phylogenetic patterns of host-parasite evolution. Green lines represent virus phylogeny, orange lines represent host phylogeny. Phylogenies support: A – co-divergence of virus with host, and B – multiple cross-species transmission events (re-drawn from Quantamagazine, 2019).

1.3 Mutation

A mutation is any change that alters the nucleotide sequence of the genome of an organism. Many different changes can arise as a consequence of mutation (section 1.3.4), and they either have no effect, or are beneficial or deleterious. The mechanisms by which nucleotide sequence alterations occur, and their effects on gene function, vary. While some mutations affect a large or multiple regions in the genome, others affect only one or a few nucleotides.

1.3.1 Substitutions

In cases where a single nucleotide is affected, the change is referred to as point mutation. In a population, more than one variant of a point mutation may exist in the gene pool, when only one of the variants remain it is referred to as fixation of this variant or allele. Substitution occurs when an allele becomes fixed in one population contrasted with the (inferred or observed) ancestral population. Substitutions are mutations most often necessary for successful adaptation (Shinya *et al*., 2006). There are two distinct categories of substitution: transitions and transversions. In transition, a purine base (adenine or guanine) is replaced with another purine base, or a pyrimidine base (cytosine and thymine in DNA, cytosine and uracil in RNA) is replaced with another pyrimidine. While in transversion, a purine is replaced with a pyrimidine or vice versa. Transition mutations are known to occur more than transversions (Dagan *et al*., 2002). If they occur in a coding sequence, these changes can be further categorised according to functional changes. A single base may be substituted in a manner that results in a change in polypeptide sequence and therefore protein function. Mutations with this outcome are classified as non-synonymous mutations. The common types of non-synonymous mutations are nonsense and missense mutations.

In a nonsense mutation, a change in nucleotide sequence introduces a premature stop codon in the transcribed mRNA. The resulting product is

usually an abnormally shortened and incomplete protein. Consequently, nonsense mutations most often result in the production of non-functional proteins. The functional effect depends on the position of the stop codon. In a missense mutation, a single nucleotide change may change a codon such that it is translated into a different amino acid. The new amino acid may have similar or different biochemical properties. Amino acids may be categorised according to their biochemical and physiological properties such as polarity, volume, and charges. A mutation that gives rise to an amino acid that falls within the same group as the ancestral amino acid is called a conservative mutation; substitutions between groups are referred to as radical mutations (Zhang, 2000). For instance, with respect to polarity, a substitution resulting in an amino acid change from aspartate to asparagine is called a conservative mutation because the derived and ancestral amino acid both contain carboxylic acid, but a substitution leading to a change from aspartate to threonine is a radical mutation because threonine contains an hydroxyl group. The expectation is that conservative mutations are less likely to alter protein function.

In a silent or synonymous mutation, a change in nucleotide sequence results in a codon that codes for the same amino acid as the ancestral codon, preserving the ancestral polypeptide sequence. Most often, synonymous mutations are thought to have no effect on the function and structure of proteins, however, this is not always true (Cuevas *et al*., 2012; Hunt *et al*., 2014). Nucleotide sequences in many organisms are biased towards choosing one of the several codons encoding the same amino acid over others, producing some tRNAs for a specific amino acid over another. If a mutation results in the formation of a codon associated with a less abundant tRNA, it may lead to a change in the structure of protein as a result of interfering with the timing of co-translational protein folding (Kimchi-Sarfaty *et al*., 2007). However, a non-synonymous change typically has a greater fitness effect than a synonymous change (Dagan *et al*., 2002).

1.3.2 Indels

Mutations sometimes entail insertions or deletions (indels) of one or more nucleotides from a sequence. When these occur in coding sequence and if the number of nucleotides lost or gained is three, there will either be an addition or deletion of a complete codon. However, if the indel length is not a multiple of three, a frameshift will occur. A frameshift disrupts the entire reading frame, leading to a large change in amino acid sequence downstream of the lesion, and/or a stop codon appearing later or earlier than in the ancestral sequence. Frameshifts are generally considered to be deleterious and a mutation accumulation experiment using *Pseudomonas aeruginosa* has indicated that they are strongly selected against (Heilbron *et al*., 2014).

1.3.3 Mutation in bacteriophages

Many studies have shown that accumulation of mutations enhance host switching in phages irrespective of the type of mutation in both coevolution and evolution studies (Scanlan *et al.,* 2011; De Sordi *et al*., 2017). In an experiment that investigated tripartite interactions between; phage P10, resistance bacteria *E. coli* MG1655 and sensitive bacterial strain *E. coli* LF82 in mice gut, a single point mutation on the phage tail fibre required for host specificity was shown to be involved in P10 host switching to resistance E. coli MG1655 strain (De Sordi *et al*., 2017). In a co-evolution study, Scanlan *et al*. (2011) reported that different types of mutations were observed in a coevolutionary study of 120 *Φ2* phage and 120 *Pseudomonas fluorescens* SBW25 isolates. These mutations include synonymous, non-synonymous and indels associated with tail fibre gene, first step in host adsorption, and all the mutations were involved in host range infectivity rather than general adaptation of Φ2 to *P. fluorescens*.

As discussed in section 1.3.1, codon usage depends on the availability of tRNAs and sometimes codon biases may occur in an organism, choosing some tRNAs over others. In experimental codon adaptation, phages may either have poor codon adaptation or strong codon adaptation on their hosts depending on translation efficiency; presence of phage-encoded tRNAs reducing dependency on host tRNAs (Prabhakaran *et al.*, 2014), differential codon usage in different hosts especially if phages recently switch hosts (Prabhakaran *et al.,* 2015), strand asymmetry with mutation bias most in single stranded (ss) DNA (Chithambaram *et al*., 2014). Also, Prabhakaran *et al*. (2015) suggested that phages may exhibit strong codon adaptation not because they have inefficient translation and an increase in elongation efficiency has little or no effect. The same study proposed that phage lifecycle has effect on the efficiency of translation and elongation by measuring translation initiation of Shine-Dalgarno in 24 *E. coli* lambdoid phages, 16 clades showed poor codon adaptation and these were temperate phages, while 8 clades virulent phages showed strong codon adaptation.

1.3.4 Causes of mutation

Mutations can be spontaneous or induced. Spontaneous mutations arise stochastically during or prior to DNA replication. These may occur due to errors in nucleotide pairing that change the biochemical structure of DNA during strand synthesis. Each of the nucleotide bases can appear in many forms to create tautomers, isomers which may interconvert. A keto form is readily available in DNA and may spontaneously change to another form such as imino and enol when a proton changes position. Thus, affecting the hydrogen bonding pattern in the bases, resulting in a mutation if nucleotide bases mispair. The nucleotide bases mispairing can also occur when bases become ionised (Griffiths *et al*., 2014). In addition to DNA replication errors, mutations may also arise due to spontaneous lesions, a form of DNA damage.

Spontaneous deamination in DNA is a hydrolysis reaction of bases, releasing ammonia in the process. It occurs in different bases, for example, in deamination of cytosine into uracil, uracil will pair with adenine to produce A—T, resulting in the conversion of G—C to A—T (Coulandre *et al*., 1978). Also, 5-methylcytosine deamination forms thymine, while guanine gives rise to xanthine and adenine, to hypoxanthine.

Depurination is the release of purine bases, guanine and adenine, from nucleic acids by the hydrolysis of glycosidic bonds between the base and deoxyribose, leaving deoxyribose with no base, an abasic site. The occurrence greatly depends on DNA sequences (An *et al*., 2014). An abasic site can be repaired by base-excision repair system in dsDNA. Base-excision repair initiated by DNA glycosylase, which recognises the missing base, leaving an abasic site, then pairing with a base complementary to the other strand (Krokan and Bjørås, 2013). Because, this mechanism is lacking in bacteriophage ΦX174 ssDNA, the base-excision system may insert any base randomly, resulting in either a transition or transversion mutation.

Induced mutations are caused by influences of external agents such as chemical mutagens, ultraviolet light, and ionising radiation. Chemical mutagens may mimic normal bases and are incorporated during DNA replication, some may damage bases causing mispairing or destroying pairing. Ionising radiation causes formation of excited and ionised molecules that may result in damage to DNA. Ultraviolet exposure produces a number of photoproducts which eventually cause lesions that interfere with normal base pairing (Griffiths *et al*., 2014).

Non-targeted mutations may also be generated intentionally for research purposes. For instance, to study evolutionary dynamics resulting from elevated mutation rate, Wilcox (2017) utilised a dominant negative *E. coli* DNA polymerase subunit gene to generate an accumulation of mutations in bacteriophage ΦX174 (since phage replication depends on host

polymerase). Domingo-Calap *et al*. (2009) performed a mutation accumulation experiment using ΦX174 and Qβ, exposing these phages to chemical mutagens to study the effects of random mutations acquired by the phages on fitness.

Spontaneous and induced mutations occur adventitiously, but it is possible to target a specific genetic locus through a process termed site-directed mutagenesis (SDM). SDM involves intentional and specific introduction of a mutation into a DNA sequence. The first mutagenesis experiment was done using ΦX174 (Razin *et al*., 1978). Domingo-Calap *et al*. (2009) used SDM to generate clones of ΦX174 and Qβ with designated (but randomly chosen) single-nucleotide mutations in order to examine fitness effects of mutations. Several phage studies have used SDM to study effects of substitutions (Pepin and Wichman, 2007; Holder and Bull 2001; Brown *et al*., 2010). There are several molecular methods employed for targeted mutagenesis including artificial oligonucleotide synthesis utilizing PCR and more recently CRISPR/Cas9 for genome editing (though most rely on recombination; Gupta and Musunuru, 2014).

1.3.5 Mutation in virus evolution

Viruses are characterised by high mutation rates (Sanjúan *et al*., 2010), leading to their ability to evolve rapidly and adapt to novel or rapidly changing environments (Abedon, 2009). Elevated rates of mutation may be adaptive for viruses, and can be detected via sequencing when an ancestral nucleotide sequence is compared with evolved virus isolates or with a population sample. Most often, specific mutations are required for viral adaptation to a new environment (Bull *et al*., 1997; Crill *et al*., 2000). For example, measles virus was considered to have evolved from its closest relative, rinderpest virus, a pathogen of cattle, via mutation. It was recorded to have a substitution rate of 6.0 – 6.5 x 10^{-4} substitutions/site/year (Furuse *et al*., 2010). The mutation rate of the ΦX174, a member of the *Microviridae*, was estimated to be ~1.1 X 10^{-6} substitutions *per nucleotide per cell infection*

(*s/n/r*), which is generally lower than that of RNA viruses although the dsRNA virus Φ6 comes close at ~1.4 X 10^{-6} s/n/r (collated in Sanjúan *et al*., 2010). However, it appears that ssDNA sequence evolution can proceed rapidly with *Microviridae* in the gut estimated to have substitution rates > 10^{-5} *per nucleotide per day* over a 2.5 year period in an experiment that investigated the origin and evolution of the human gut virome (Minot *et al*., 2013).

1.3.5.1 Co-evolution, co-existence and mutation in viruses

In nature and some experimental systems, virus evolution is driven by virus-host coexistence with high degree in population diversity. Since virus solely depends on host cell for infection and transmission, coevolution (section 1.2.2) is often seen as a common outcome of co-propagation of virus-host system (Koskella and Brockhurst, 2014). Coexistence may occur until the emergence of host genotype that resist viral infection evolves (Lenski and Levin, 1985). In phage-bacterial cell system, the accumulation of mutations mostly occurred on the bacterial cell surface molecules required for the attachment of phage. As the bacteria surface molecules evolve, phage must continue to evolve the capacity of specific binding to the modified version of host cell surface molecules or to an alternative receptor. Lenski and Levin (1985) argued that such bacterial-phage 'arms-race-evolution' – evolution of bacteria defences and phage counter-defences, which may later lead to mutational asymmetry. In mutation asymmetry, host resistance-mutation may occur by loss or change in gene function, causing bacterial cell to become ultimately resistant, while phage infectivity depends on specific changes in gene leading to arms-race-evolution event. Contra the mutational asymmetry hypothesis, extensive coevolutionary arms race has been shown to occur especially when there was no prior history of bacterial-phage infection (Meyer *et al*., 2012). Evidence from *E.coli* B and lambda-vir adaptation showed that lambda-vir bind to an alternative OmpF host receptor rather than LamB receptor indicating a successful counter-adaptation (Meyer *et al*.,

2012). In arms-race-coevolution, there is emergence of phage mutants that overcome host resistance. Phages accumulate mutations most often in genes encoding proteins required for host attachment, conferring broader host infectivity range (Scanlan *et al*., 2011 and Paterson *et al*., 2010).

1.4 Viruses-host genetic diversity in host switching

Genetic variability is one of the drivers of host-pathogen interactions in a population community and may enhance host range capability of pathogens. Viruses have been shown to exhibit viral receptor tropism switching via massive genetic variation of their host-binding receptor. Some group of phages (temperate phages) of *Bordetella* species (causes respiratory infections in mammals) were discovered to exhibit diversity in gene (diversity-generating retroelement, DGR) that specifies tropism switching (major tropism determinant, *mtd*) in their fibre proteins. Such modification in receptors allow broader host range abilities (Liu *et al*., 2002). The *Bordetella* phage BPP-1 displayed tropism for group of *Bordetella* hosts Bvg^+ phase; *B*. pertussis, *B*. parapertussis and *B*. bronchiseptica. Generally, BPP-1 receptor, pertacin, was known to be expressed only in BVg+, but at frequency of ~ 10^{-6}, BPP-1 gave rise to two variants that acquired tropism for; BVg^- (designated as BMP variant) and both BVg^- and BVg^+ (designated as BIP variant; Liu *et al*., 2002). The infection was suggested to occurred via template-dependent, phage encode reverse transcriptase (Brt) tropism switching, resulting from a point mutation in the variable sequence region I of *mtd* gene (Liu *et al*., 2002; Doulatov *et al*., 2004). Liu *et al*. (2004) identified two more variable regions within; *bbp36* locus, the second major redion of variability and *bbp32*, encoding a unique phage methylase with a hot-spot region for frame shift mutation. More recently, additional 92 DGRs have been identified in temperate phages integrated into bacterial phlyla Bacteroidetes, Proteobacteria and Firmicutes as prophages. The identification was done via algorithm with mapping reads of viromes from different databases including NCBI, PhagesDS, RefSeq. A novel temparate *Bacteroides dorei*

Hankyphage was discovered. Hankyphage exhibited broad host range, lysogenised 13 different species of the *genus Bacteriodes*. In the same study, five phages have hypervariable proteins similar in structure with BPP-1 tail fiber (Benler *et al*., 2018).

A large microbial population may evolve as a result of natural selection resulting in genetic diverse communities. In phage-bacteria population, phage mutants with varying degrees of host range and fitness (Schwartz and Lindell, 2017) may arouse. For instance, in tailed T7-like cyanophages, mutations occurred in tail genes (required for phage attachment). In some cases, T7-like cyanophages were able to infect both resistant and wild-type hosts with different fitness. In one case, reduction in fitness was observed on the original host and in other cases, fitness on the wild-type host increased (Schwartz and Lindell, 2017). Therefore, Schwartz and Lindell (2017) proposed that during virus evolution, phage-host genetic diversity and coexistence is driven by combination of events including; arms race, fitness costs and host range.

Sometimes, phage-bacterial interactions may drive bacteria hosts to tolerate phage infection via mutations that may have effect on host clonal population heterogeneity, generating both phage-resistance and phage-tolerant diversified population. For example, population of Mu^{L} cells, a phage-tolerant mutant isolated from *E. coli* O157:H7, stably co-exist with phage PP01 and the equilibrium population growth of Mu^{L} cells were not affected despite rapid infection of PP01 phage in a continuous culture (Fischer *et al*., 2004).

For viral macro-evolution, mutation in the nucleotide sequence of genes encoding structural proteins may influence capsid structure and can alter the host range capabilities of an organism. For example, the feline panleukopenia virus emerged in dogs as a parvovirus via a small number mutations of structural proteins (Truyen and Parrish, 1995). Mutation

appears to be essential factor for viral adaptation to a new environment, and in particular for host switching.

Following host switching, natural selection may act on traits that favour pathogen growth in the new host environment. Mutations modifying viral host cell entry, high efficiency transmission to the new host and fitness optimisation in the new host can be selected in viral populations. These mutations may enhance transmission potential, immune avoidance, virulence or facilitate the efficient use of host cellular machinery (Longdon *et al.*, 2014). The ability of pathogens to colonise and adapt to novel hosts via one or all of the pathways listed may entail fitness costs, to which we now turn.

1.5 Trade-offs

An outcome of host switching by pathogens is trade-offs. Trade-offs are events that occur when trait changes that optimise fitness come at the expense of one-another. This may occur because phenotypes provide a compromise solution (to environmental challenges) and the trade-off may manifest in a different fitness for a given genotype/phenotype when the environment changes. Adapting to a novel host may result in trade-offs, consequently affecting the chance of a host switch occurring. Trade-offs may, but do not inevitably, result in unsuccessful host switching. An unsuccessful host switch is a failed host switching attempt in which a population colonising a new host goes to extinction over several generations. For example, HIV-1 groups M and N have independently switched from chimpanzees to human (Sharp and Hahn, 2010) involving evolution of Viral protein U (Vpu). Viral particles are released from infected cells when Vpu in SIV binds and degrades CD4 receptors during infection and viral particles are released from infected cells. This function was lost in HIV-1 group N and has been suggested that it might explain why HIV-1 group N occurs rarely in Africa, while HIV-1 group M has become pandemic. Trade-offs can generate outcomes intermediate between successful and failed host switching. There

is evidence of successful host switching to humans in HIV-1 group N, though the chances were reduced as a result of trade-offs (Sauter *et al*., 2009).

As a result of fitness trade-offs, pathogens may maximise transmission and infection potential in the new environment, but with a concomitant reduction in performance on the former host. For instance, in an experiment to determine host range and fitness costs, RNA virus Φ6, a generalist phage with expanded host range was used as model system. Host range Φ6 phage mutants with ability to infect novel hosts *Pseudomonas* were identified. Each mutant had at least one of the identified nine nonsynonymous mutation in gene P3 of phage Φ6. The fitness costs associated with Φ6 mutants was investigated on the standard laboratory host and some which were costly on the original hosts (Duffy *et al*., 2006); mutations affecting gene P3 are often known to reduce growth on the original host, P3 is a host attachment gene in phage Φ6 (Ferris *et al.,* 2007). Observations of similar effects have been recorded in ΦX174 (Crill *et al*., 2000), in arthropod-borne viruses (Coffey *et al*., 2008) and in vesicular stomatitis virus in BHK-21 cell culture (Novella *et al.,* 1995).

1.6 Bacteriophage ΦX174

The bacteriophage ΦX174, in the family *Microviridae*, is an icosahedral virus, made up of a circular single-stranded (ss)DNA genome, and a protein capsid. The ΦX174 genome has a complicated architecture with a coding region of approximately 95% total length which includes several overlapping genes. Overlapping genes are constituted in one case by an alternative start codon, in others by alternative reading frames (figure 1.2). The genome is 5,386 base pairs (bp) in size and encodes 11 proteins out of which only 9 are essential in the infection process. These are A, A*, B, C, D, K, E and the structural proteins F, G, H and J (table 1.1). The icosahedral virus forms a T=1 symmetry, with the protein capsid (made up of structural proteins) consisting of 60 copies of the F protein decorated by twelve spikes on its vertices, each with five G proteins and one H protein (Sun *et al*., 2014).

Proteins A, B, C, D and E function in the production of the mature virion, while A* and K are both inessential proteins. Overlapping genes achieve an effective compression of information, such that the genome codes for more proteins than would be expected given its length.

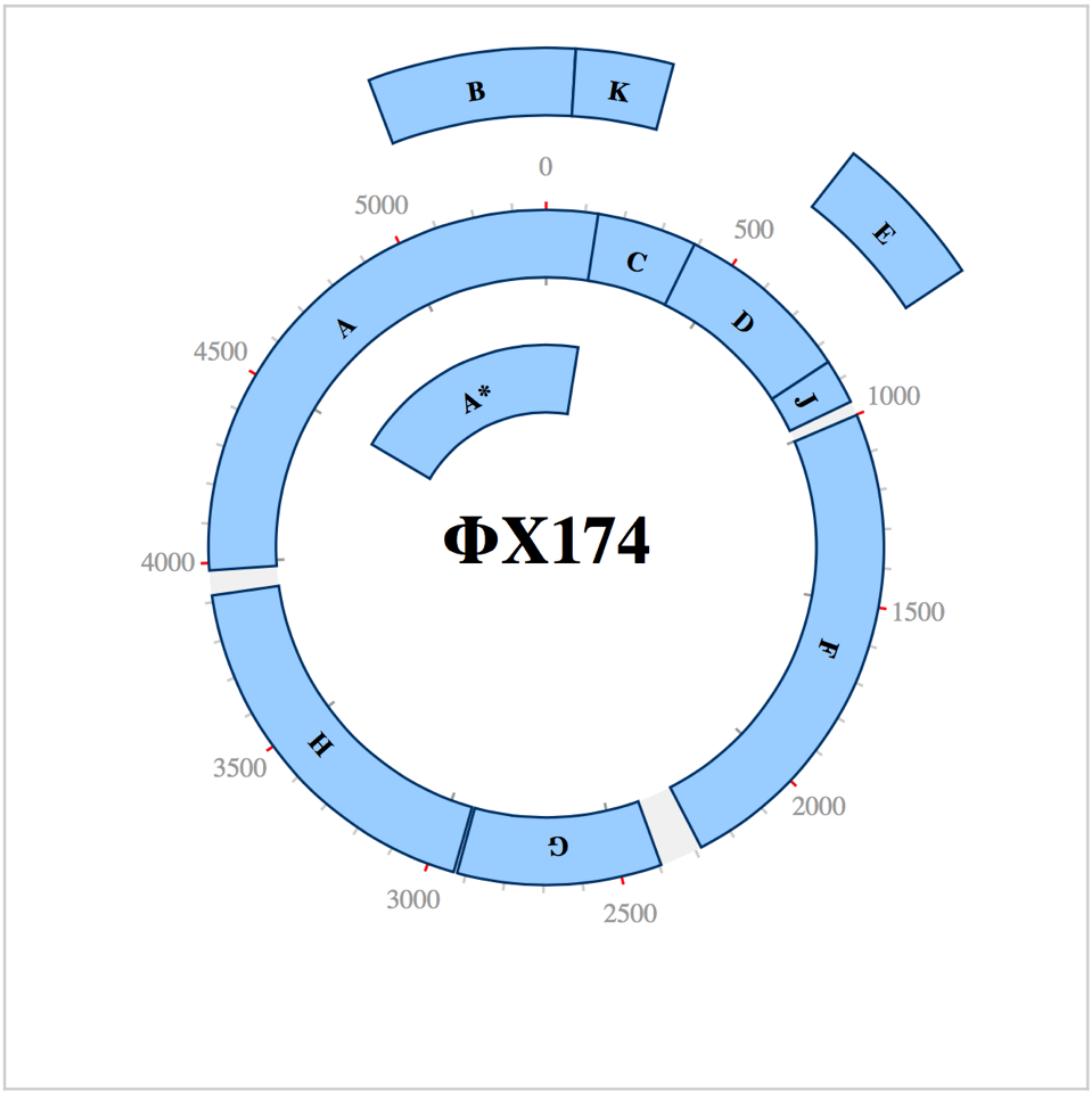


Figure 1.2: Genome map of bacteriophage ΦX174. Genes A, B, K, C, D and E overlap in different reading frames. Gene A* overlaps with A but in the same reading frame. The image was created using the Angular Plasmid library (http://angularplasmid.vixis.com).

Protein	Function	Genome location
K	Inessential. May optimise burst sizes	51-221
C	DNA replication	133-393
D	Procapsid Morphogenesis	390-848
E	Host cell lysis	568-843
J	DNA binding and packaging	848-964
F	Major coat protein	1001-2284
G	Major spike protein	2395-2922
H	Minor spike protein. DNA pilot protein	2931-3917
A	DNA replication	3981-136
A*	Unessential, may inhibits host cell replication	4497-136
B	Procapsid morphogenesis	5075-51

Table 1.1: The bacteriophage ΦX174 proteins, functions and location on the genome (Fane *et al.*, 1988; Hayashi *et al.*, 1988; Wichman *et al.*, 2005).

ΦX174 is well characterised (Hayashi *et al*., 1988) and, as a result of its small genome, is amendable to genetic manipulations. Consequently, ΦX174 was an attractive model for many researchers in molecular biology in different areas of study such as in Bull *et al*., 1997; Crill *et al*., 2000; Wichman *et al*., 2000, 2005; Holder and Bull, 2001; Poon and Chao, 2006; Pepin *et al*., 2007; Dickins and Nekrutenko, 2010; Wichman and Brown, 2010 (discussed in sections 1.7.1.1 and 5.3.8). Moreover, ΦX174 is an industrially important phage, as it can inhibit the growth of bacteria in an industry that requires *E. coli* growth (Labrie *et al*., 2014). ΦX174 has provided a method for the production of inactivated vaccines. The lysis of bacterial cells mediated by gene E (section 1.6.1.5) facilitates changes in osmotic pressure, producing empty bacterial cells lacking nucleic acids and cytoplasm which are referred to as bacterial ghosts (Yu *et al*., 2011). Bacterial ghosts have been widely used in the development of vaccines, for instance, in veterinary applications (Jalava *et al*., 2002), as carriers of nucleic acid-encoded antigens for DNA delivery to human cells (Kudela *et al*., 2005) and as recombinant vaccines (Eko *et al*., 1999).

The study of ΦX174 has a long history, beginning in 1959 as the first DNA molecule to be purified homogeneously (Sinsheimer, 1959), and the phage DNA synthesized *in vitro*, ushering in the era of synthetic biology (Gouliant, 1967). In 1977, it became the first DNA-based genome to be fully sequenced (by Fred Sanger and his team, using a sequencing technique that bears his name; Sanger *et al*., 1977), and the first to be used in SDM (Razin *et al*., 1978). It was demonstrated by Craig Venter's group that ΦX174 can be synthesised *in vitro* from synthetic oligonucleotides. ΦX174 therefore became the first human-synthesised genome capable of successful infection (Smith *et al*., 2003). More recently, it was reported that the genome of ΦX174 can be decompressed and still remain viable in host cells (Jaschke *et al*., 2012).

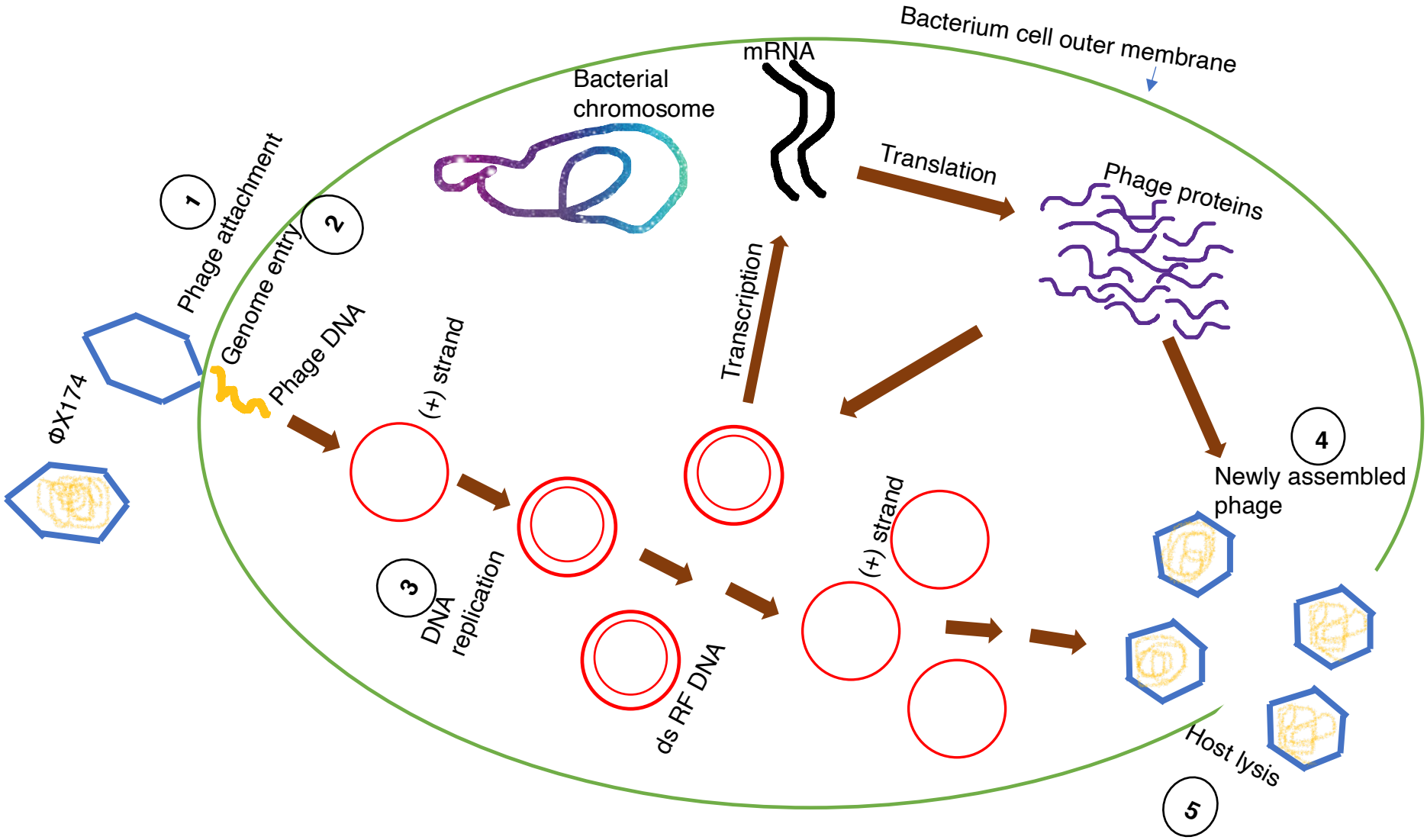


Figure 1.3: The life cycle of bacteriophage ΦX174. ΦX174 infecting a Gram negative bacterial cell outer membrane, crossing peptidoglycan layer and interacting with the cytoplasmic membrane. The five main infection stages of ΦX174 are shown: **1**- phage attachment, **2** - genome penetration, **3** - DNA replication, **4** - phage assembly and **5** - host lysis. RF DNA is Replicative form DNA.

1.6.1 The ΦX174 lifecycle

Similar to all obligately lytic phage, ΦX174's lifecycle involves five main infection stages. A coat protein first binds to a specific receptor on the surface of a susceptible bacterial cell during the attachment processes. During the entry process, the phage injects its genomic material into the cytoplasm of the bacterium. Following this, the phage DNA is transcribed, genes are expressed and translated to functional proteins in the replication stage. Next, capsids are assembled, first as procapsids, then filled with DNA to produce a mature infectious virion. Lastly, new mature virions produced are released to the host environment as the cell burst, ready for next lytic infection cycle. The life cycle is shown in figure 1.3.

1.6.1.1 ΦX174 attachment

The bacteriophage ΦX174 infects some bacterial cell walls of the *Enterobacteriaceae* family that possess rough forms of lipopolysaccharide (LPS) in their outer membrane. These are Gram-negative bacterial strains that produce LPS containing specific terminal sugar moieties: glucose and galactose in the core oligosaccharide of rough LPS (Feige and Strim, 1976). Bacterial strains that contain such rough forms of LPS are sensitive to ΦX174 infection, these include strains of *Escherichia coli* (Michel *et al*., 2010) *S.* Typhimurium and *Shigella sonnei* (Wichman and Brown, 2010). Suzuki *et al*. (1974) discovered that replication of ΦX174 can be supported by distantly related organism *Pseudomonas aeruginosa*, but with a reduction in the production of number of phage progeny.

Attachment of ΦX174 to susceptible host cells occurs via gene F in the coat protein (table 1.1), which binds reversibly to a bacterial cell's glucose residues, a process which is dependent on the addition of calcium ions. This is followed by an irreversible reaction, the molecular basis of which remains unclear (Fane *et al*., 1988). An electron microscope study of ΦX174 by Brown *et al*. (1971) indicated that the phage attaches to the host cell wall by

one of its spikes. Through an orientation of its five-fold axis icosahedral symmetry, ΦX174 submerged about one and a half diameter of its spike into the host cell surface. Protein G functions may include recognising the host cell surface LPS residues (Inagaki *et al*., 2003), and phage bind to bacteria LPSs via one of its pentameric spikes of protein F. Previous studies demonstrated that changes in amino acids in protein F affect host range and attachment (Crill *et al*., 2000; Pepin *et al*., 2006). In addition to protein G, protein H has also been demonstrated to be involved in attachment of phage ΦX174 to the LPSs of ΦX174-sensitive strains (Suzuki *et al*., 1999; Inakagi *et al*., 2000; Kawaura *et al*., 2000).

1.6.1.2 ΦX174 penetration to host cell

Following attachment, the phage genome is injected into host bacterial cell for successful infection. Unlike tailed phages that inject their genomes into host through their tails, ΦX174 are tailless, and roll along the surface of host cells until they find a receptor for penetration (Fane *et al*., 2006). Prior to infection, pilot protein H forms a tube that translocate the genome across the host LPS. Proteins F and G at one of the five-fold vertices of interaction, recognise and interact with host LPS to initiate injection. The dissociation of the spike protein and conformational changes of ΦX174 particles during attachment to LPS bilayer membrane-like structures trigger ejection of ΦX174 genome via the protein-H tube. At 37^{o}C incubation temperature, ssDNA genome ejection proceeds, with at least one of the twelve H proteins transferred into the host cell alongside phage DNA (Sun *et al*., 2017).

1.6.1.3 ΦX174 DNA replication

The ΦX174 DNA replication phases are complex and characterised by three distinctive stages I, II and III. Stages I and II rely on recruiting host enzymes, while stage III requires nine viral proteins. In stage I, the ΦX174 ssDNA genome is converted into replicative form one (RF I) DNA, a covalently closed double-stranded (ds) DNA molecule (Fane *et al*., 1988). Since ΦX174

DNA is single-stranded, no protein can be synthesised until conversion to a dsDNA occurs. The conversion of ssDNA to RF I DNA depends entirely on 13 host proteins including ssDNA binding protein (*ssb*) and DNA III polymerase holoenzyme complex (Shlomai *et al*., 1981).

The second stage involves amplification of RF I DNA and replication of the phage (+) strand and proceeds via a rolling circle mechanism. In addition to the 13 host proteins utilized during stage I, stage II replication requires ΦX174 protein A and the host cell helicase, rep (Eisenberg *et al*., 1977). Protein A initiates stage II replication by introducing a nick on (+) strand of RF I DNA at the origin of replication. Protein A then binds covalently to the 5' end of RF II DNA and forms the RF II-gene A complex. Following this, *rep* protein binds to the RF II-gene A complex and unwinds the DNA strands, and *ssb* maintains the separated strand by binding to it. DNA polymerase III associates with RF II-gene A–*rep* protein complex and extends the viral DNA at 3' end while the 5' end is being displaced (Aoyama and Hayashi, 1986). The 5' end travels along the synthesized strand, and with the replication fork in a rolling circle. After strand synthesis completion, protein A covalently bound at the 5' end of the parental strand, acting as a nuclease, cuts the newly synthesised strand at the origin, and acting as ligase, re-joins the two ends of parental strand to form circular ΦX174 genome coated by *ssb*. Protein A is transferred to the newly synthesised strand at 5' end, severing the template for the next round of ΦX174 strand synthesis (Aoyama and Hayashi, 1986; Hayashi *et al*., 1986).

The third stage involves ssDNA synthesis and packaging into the viral procapsid. Viral DNA synthesis requires functions of host proteins, the phage procapsid and protein C (Wolfson and Eisenbrerg, 1982). ΦX174 protein C associates with the rep protein on RF II DNA, and inhibits synthesis of dsDNA. Protein C and *ssb* compete to bind to the RF II DNA-gene A-*rep* protein complex. If protein C binds first, stage III DNA proceeds, however if *ssb* binds, another round of stage II DNA synthesis ensues. Protein C will not

prevent stage II after it has begun, when protein C binds, a RF II-gene A-rep protein-protein C complex forms. This complex binds to the procapsid and a similar round of stage II (+) strand synthesis ensures, with protein A acting as ligase, producing a covalently closed circular molecule but remaining attached to the procapsid (Fane *et al.*, 1988).

1.6.1.4 ΦX174 virion assembly

A proper virion assembly requires ordered protein-protein interactions proceeding along a precise morphogenetic pathway. The assembly of an infectious virion particle requires recruiting proteins A, C, F, G, H, J and scaffolding proteins D and B which direct early assembly intermediates into larger macromolecular structures, acts as an intermediary for conformational switches. They also reduce thermodynamic barriers and ensure morphogenetic fidelity (Cherwa *et al.*, 2017). A capsid precursor, the procapsid, is the first step during virion assembly. The requirements for the procapsid begin from stage III DNA synthesis and packaging, with proteins A and C possessing functions related to stages II and III DNA replication as previously mentioned. During assembly, 60 copies of F protein, 12 H protein, 60 G protein, 60 B protein and 240 copies of D protein organises into intermediate procapsid (Hayashi *et al.*, 1988).

In the first phase, five copies of internal scaffolding protein B bind to the lower side of ΦX174 coat protein F and induce a conformational change that allows binding with one DNA pilot protein H, producing the 9S particle. This is followed by association of 6S particles (made up of pentamers of the spike protein G) to the already-formed 9S, yielding 12S intermediates. In the next phase of morphogenesis, external scaffolding protein D copies arrange twelve 12S molecules into procapsids (Cherwa *et al.*, 2011). The DNA binding protein J guides and binds to a ssDNA viral genome and enter procapsid. As the procapsid is filled, the internal scaffolding protein B is displaced. The 60 copies of protein J associate with 60 copies of coat protein F, harnessing the genome in place within the virion. After genome

packaging, the external scaffolding protein D is displaced, resulting in complete maturation of the capsid through a change in the pentamers configuration (Hafenstein and Fane, 2002).

1.6.1.5 Host lysis

ΦX174 possesses a lysis protein E, the gene for which is embedded within gene D (figure 1.2). The expression of protein E is sufficient to cause lysis of its host, forming a transmembrane tunnel, which penetrates the inner and outer membrane of the bacterial cell (Witte *et al*., 1990). The mechanism of protein E-mediated lysis is similar to the antibiotic penicillin's mechanism of action. The inhibition of MraY (encodes translocase I) activity, a membrane enzyme that catalyses the development of the first lipid-linked intermediate during peptidoglycan biosynthesis, by antibiotics lead to lysis. In the same way, protein E is a specific inhibitor of MraY, causing lysis to occur primarily due to catastrophic failure at septation (Bernhardt *et al*., 2000). Thus, protein E targets translocase I, inhibiting cell wall biosynthesis, and causing lysis of host when bacterial cell attempts to divide. The time of attachment to burst of host cell is refers to as lysis timing. Lysis timing as described by Hutchison and Sinsheimer (1963) ranges from 15 to 30 minutes but has a mean of 21 minutes at 37^{o}C, which supports the view that the lysis mechanism is dependent on the host cell's life cycle.

1.6.1.6 Non-essential proteins

The functions of the two inessential proteins A* and K have been described by researchers, although the mechanisms by which they work are unclear. Protein K was related to an increase in burst size of ΦX174. Gillam *et al*. (1985) created a mutant form of ΦX174 with a premature stop codon in protein K and found out that ΦX174 was still viable but with decrease in burst size when compared to wild-type. Protein A*, an N-terminally truncated version of protein A, has been shown to increase the efficiency of ΦX174

DNA replication by blocking replication of host cell. ΦX174 may possess greater sensitivity to cleavage by A* (Hayashi *et al*., 1988).

1.7 Experimental evolution

Traditional evolutionary studies involve piecing together past evidence from fossils of extinct species and comparative studies of extant species, an approach that depends on millions of years of evolution of the investigated organisms. Modern evolutionary studies utilise controlled field manipulations or laboratory experiments to explore evolutionary dynamics, and this is termed experimental evolution. One of the first scientists to carry out experimental evolution was William Dallinger in the 19th century (Dallinger, 1888). He cultivated unicellular organisms in an incubator for several years. At the start, the organisms normally showed signs of distressed growth at 23^{o}C. Dallinger gradually increased the temperature from 15^{o}C – 65^{o}C and at the end of his experiment the population of organisms have adapted to the environmental change and were growing at 65^{o}C. However, the adapted organisms were unable to grow at 15^{o}C.

Microorganisms are particularly useful for experimental evolutionary studies. There are several reasons for this. Microorganisms have the capacity to produce large populations with short generation times enabling evolution to be studied in real time. Microorganisms' small sizes allow large populations to be propagated in smaller spaces, while their relatively small genomes facilitate sequencing. Storage and culture is carried out in conditions that are easier to control, and the opportunities for experimental replications make microbes an excellent choice for experimental evolution (reviewed in Buckling *et al*., 2009; Kawecki *et al*., 2012). The use of microorganisms has changed the study of the mechanisms of evolutionary processes and there has been a transformation of evolutionary biology, allowing theories to be tested directly and its studied in real time (Buckling *et al*., 2009).

A prominent example of experimental evolution utilizing microbes is the *E. coli* long-term evolution experiment (LTEE). The LTEE was started in 1988 by Richard Lenski at the University of Michigan, and is still underway. Lenski and colleagues have followed the evolution of 12 populations of *E. coli*, tracking genetic changes over time in this experiment. At the time of writing, the populations have been growing for over 71,000 generations, and multiple publications have emerged from these experimental data (Barrick *et al*., 2009; Tenaillon *et al*., 2016; Good *et al*., 2017; Lenski *et al*., 2017). Since then, several evolution experiments have been performed utilizing yeasts, bacteria and viruses. Viruses, as microorganisms, share the same benefits described above and allow researchers to explore a diverse group of organisms that are extremely numerous. They play a particularly significant role in studies of host-parasite interactions. Viruses are dynamic and their properties and life cycles facilitate rapid adaptation and change, defeating the most creative expectations of many researchers (Manrubia, 2012).

Experimental evolution has led to many new discoveries, providing a wealth of information and knowledge about how evolution works, enabling evolutionary scientists to test theories directly, revealing answers to multiple long-due evolution questions. The combinations of experimental methods in evolutionary studies and high-throughput sequencing technologies has yielded many valuable approaches for evolutionary studies. The development of high-throughput sequencing and advanced methods for deep sequencing has facilitate tracking of evolutionary dynamics. The evolution tracking of genotypes within a short period, with capacity of comparisons of populations assayed from the past, contemporary, and even future predictions can be made. Thus allowing to track evolutionary processes such as elucidating adaptive mechanisms in organisms.

1.7.1 Experimental evolution of ΦX174

The bacteriophage ΦX174 has been an important model system for experimental evolution studies. This is mostly due to its rapid replication cycle, ease of cultivation (shared with other viral systems) which permit culturing large populations in a short period of time at low cost. Its common laboratory host *E. coli* C possesses a short generation time and is robust to a wide range of temperatures making it ideal for evolution experiments (Wichman and Brown, 2010). Detailed structural information is also available for ΦX174, while its particularly small genome (5,386bp) makes it amenable to genetic manipulation and facilitates accurate tracking of genetic changes that might occur during evolution and adaptation. The small target size increases expected sample coverage for a given sequencing yield. It has been employed extensively to study the patterns and processes of evolution including by cataloguing genetic variation in viral populations, tracking evolutionary dynamics across populations, elucidating mutation spectra and rates, and identifying adaptations occurring in varying environments. ΦX174 was first developed as a model organism for experimental studies by J.J. Bull in 1993 (reviewed in Wichman and Brown, 2010).

1.7.1.1 Experimental evolution studies using ΦX174

The first experimental evolution study with ΦX174 used the virus to examine the extent of parallel evolution; the rates of convergent evolution and substitution were assessed during adaptation at high temperatures in a chemostat on two different hosts (Bull *et al.*, 1997). Bull *et al.* (1997) found out that among nine lineages, half of the substitutions and one-third of nucleotide sites observed were identical in multiple lineages. In a similar parallel evolution study, Wichman *et al.* (1999) examined the molecular basis of adaptation under strong selection, discovering that although half of changes in one line appeared in the other, parallel mutations did not occur in a similar order. The authors also found out that most substitutions were adaptive in the replicate lineages but these parallel substitutions did not

show changes with largest beneficial effects or reflect common evolutionary trajectory pattern of adaptation.

Host-specific adaptation of ΦX174 has also been examined. Crill *et al*. (2000) observed adaptation of ΦX174 on *E. coli* C and *S.* Typhimurium hosts. They found that, when adapted in *S*. Typhimurium, the virus's growth rate was reduced on the *E. coli* host as a result of substitutions in the major capsid gene. When ΦX174 was forced to grow on *E. coli*, reversion mutations were observed at the same sites. In a different study ΦX174 and a closely related phage, S13, were adapted, in replicate lines, to *E. coli* C hosts. Viral samples were analysed for accumulated nucleotide changes. It was observed that changes occurred at sites where ΦX174 differed from S13, leading the authors to conclude that there were limited pathways taken by the viruses during evolution (Wichman *et al*., 2000). In another study, involving evolution of ΦX174 in three *E. coli* mutants with different LPS host receptors, the authors suggested that evolution of *Microviridae* may have relatively high levels of variation with mutations not shared between adapted phage; only one mutation occurred in multiple replicate lineages (Pepin *et al*., 2008). Based on this evidence, it is likely that evolutionary pathways depend on the starting genotype of the virus as well as on the nature of the environmental shift.

Epistatic effects in the ΦX174 genome have been studied. A study by Bull *et al*. (2000) investigated the effects of beneficial mutations that occurred when ΦX174 was grown at high temperature. It was noted that mutations occur in the genes encoding the coat and internal scaffolding proteins and coat protein fitness effects exhibited epistasis, supporting a model, diminishing returns epistasis, that beneficial mutation is scaled depending on the opportunity for fitness improvement in the genome. In diminishing returns epistasis, combinations of beneficial mutations that confer an advantage to an organism in a particular environment may reduce in benefit if introduced into a more fit environment (Chou *et al*., 2011). When fitness effects of two

single mutants and five different combinations of the corresponding double mutants in six conditions were tested, it was observed that epistatic effects differed in degree, sign and variability across host environments, even between single mutations in the same two sites (Pepin and Wichman, 2007).

Some studies investigated evolutionary dynamics using ΦX174. For instance, Dickins and Nekrutenko (2009) showed that even at positions with low-frequency genomic variation, it is possible to detect substitution dynamics occurring during adaptation using deep sequencing. Pepin and Wichman (2008) observed the evolutionary dynamics of ΦX174, while testing for beneficial mutations and clonal inference's effect on adaptation using genetic data under benign and harsh environments. They recorded that, although clonal interference may be determined by the particular beneficial mutations that arise during adaptation, its occurrence largely depends on selective conditions.

Holder and Bull (2001) investigated fitness and genetic changes during adaptation under inhibitory growth conditions. The role of mutational biases and translational efficiency of engineered ΦX174 was studied to determine evolutionary processes affecting codon compatibility between viruses and their hosts (Kula *et al*., 2018). Brown *et al*. (2013) examined the adaptive changes that occurred on ΦX174 genes when ΦX174 was grown for 50 days in a chemostat, and noted that in addition to changes in host recognition and capsid proteins, changes also occurred in genes involved in replication with host environment as a selective pressure. Also, Bull *et al*. (2006) studied the dynamics and impact of host population density on ΦX174 adaptation and the impact of adaptation on population density using chemostat. The study addressed both ecology and evolution of population density in models and concluded that in a single chambered predator-prey system (figure 1.4), virus maintained low density, while in two-chambered chemostat (figure 1.5), ΦX174 adaptation led to high viral density that favoured competition. Wichman *et al*. (2005) tracked ΦX174 evolution in a continuous culture for

180 days, the longest study carried out in a chemostat with ΦX174, and suggested that a continuous molecular evolution may ensure an indefinite arms race of the system as a consequence of co-infection which may lead to genome competing with one another in a bacterial cell.

1.8 Culturing systems

All organisms require a growth environment and a supply of nutrients to survive and reproduce. Since cellular hosts are a pre-requisite for their growth, phages are unable to reproduce without their hosts. Therefore, they must be supplied with host cells for growth, reproduction and survival. The growth medium may be a suspension or a solid matrix. There are three major methods used for propagation: plate double agar overlay assay, serial passaging and continuous culture in a chemostat. Each method provides different conditions with disadvantages and advantages for experimental study.

1.8.1 Plate double agar overlay assay

Bacterial cultures may be grown on solid agar surfaces in a Petri dish. Agar, a gel-like substance made up of agarose and agaropectin, is normally supplemented with nutrients required for bacterial growth. Phages may be grown on an agar surface alongside bacterial cells, but in a semi-solid agar. The double agar overlay is formed by pouring the semi-solid agar on a base solid agar (Clokie and Kropinski, 2009). While the base layer provides a substrate for bacterial growth where it forms a complete lawn (provided the population of bacterial cells is sufficiently large), semi-solid agar enables easy diffusion of phage in order to locate susceptible bacterial cell on incubation. Infection results in either clear (for lytic phages) or translucent (for lysogenic phages) visible zones termed plaques. This method has the disadvantage that scaling up is challenging as it involves the use of multiple Petri dishes. Nevertheless, it is a widely used method for enumerating phage

and for studying their phenotype and morphology, as well as for isolating individual plaques presumed to originate from single infectious particles.

1.8.2 Serial transfers

In serial transfer, as suggested by its name, a small inoculum from a culture is transferred to fresh media in culture flasks or tubes and incubated for a period of time, before the process in repeated several times. Populations can be serially transferred for any number of generations as desired by the experimentalist. In comparison to the plate double agar overlay assay, serial transfers are easier to handle and maintenance of many different cultures can be concurrent (Dykhuizen, 1990). The serial passaging method is typically used for experiments that run for a small number of generations in ΦX174 studies discussed in section 1.7.1.1 (Bull *et al*., 2000; Holder and Bull, 2001; Pepin and Wichman, 2008; Pepin *et al*., 2008; Baker *et al*., 2016; kula *et al*., 2018), with exception of the famous Richard Lenski's LTEE, in which *E. coli* cells have been evolved via serial transfer for more than 30 years and 71,000 generations (Lenski, 2019). The longest ΦX174 serial transfer experiment to date reached approximately 300 generations performed for about three months (Wilcox, 2017) which can be achieved in a chemostat in approximately four days, revealing an additional advantage of chemostat, viz. that it requires less "hands-on" and far less experimental time.

1.8.3 Chemostats

A chemostat is a culturing system that supports the propagation of organisms in which fresh medium is continuously added to the culture at a constant rate while at the same rate, spent medium plus cells are removed such that a constant volume is maintained (figures 1.4, 1.5). The addition of fresh medium occurs in a manner that dilution rate is less the maximum growth rate of bacterial cells. Eventually, an equilibrium is established, in

which the cells grow continuously at a constant rate and the growth rate is equal to the dilution rate (Dykhuizen 2004; Gresham and Dunham, 2014).

A one-chambered chemostat (figure 1.4) is frequently utilised for the study of evolution, selection, adaptation or fermentation processes in organisms. Usually, it uses a chemically defined medium in which a single nutrient is limiting in concentration. The concentration of the limiting nutrient defines the steady-state cell density, such that, as the concentration of the limiting nutrient increases, the steady-state cell density increases proportionally. Hence, bacterial cells grow continuously in a defined environment where all nutrients are present in excess except the limiting one (Gresham and Dunham, 2014). Although phage and bacterial cells can be confined in one chamber in a predator-prey system for the study of viral evolution, it was noted that virus adaptation maintains low density, impacting host density via feedback regulation (Bull *et al*., 2006).

A two-chambered chemostat (figure 1.5) is analogous to a one-chambered design in that fresh medium is continuously added to a bacterial vessel. However, an additional vessel for phage propagation is introduced, into which fresh, naïve bacterial cells are pumped at the same rate as the outflow. As a result, excess volume flow out as waste contains spent media, bacterial cells, and phage.

The longest chemostat experiment with ΦX174 performed by Wichman *et al*. (2005) covered 180 days (~ 13,000 ΦX174 generations). More examples of studies that utilised chemostats include: Bull *et al*. (1997) tracking substitutions and fitness rates during ΦX174 evolution, Brown *et al*. (2013) growing ΦX174 for 50 days under strong selection to determine molecular mechanisms underlying adaptation, and Wichman *et al*. (2000) adapting ΦX174 and the closely related S13 to evaluate nucleotide changes that accumulated after 10 – 11 days.

1.8.4 Serial transfers versus chemostats

Both serial transfer and chemostat culturing systems are commonly used by researchers in an experimental system. The serial transfer technique is simpler than a chemostat as it requires no special set-up other than passaging cells in culture from one flask to another, but a chemostat has several advantages making it suitable for the research described in this thesis. Chemostats are designed to provide a homogeneous and consistent environment for the growth of cells. Therefore, environment and physiological conditions are kept constant (except the change imposed by the experimentalist) and they can be automated, an advantage over serial transfer where exponential growth is re-initiated at each transfer. Furthermore, a two-chambered chemostat may reduce the chances of bacterial adaptation, depending on the interest of the experimentalist. In this study, bacterial adaption was minimised by discarding and re-inoculating with fresh naïve bacterial (further discussed in chapter 3). However, in a two-chambered chemostat, there is a possibility of between-virus competitive adaptation (Wichman *et al.*, 2005). The phage vessel has a continuous supply of healthy hosts at high density, as the input cell density is not regulated by phage density. Phages continue to replicate in the presence of a susceptible host. For the obligately lytic phage ΦX174, ~100 phage progeny are released per cell burst within 21 minutes of growth (Hutchison and Sinsheimer, 1963). Phage output per cell and burst size are high in the presence of high bacterial cell density leading to extraordinary increases in phage density. After a while, the multiplicity of infection (MOI; that is, the ratio of phages to bacteria cells) will be high, putatively resulting in selection for high competitive ability in the vessel (Bull *et al.*, 2006).

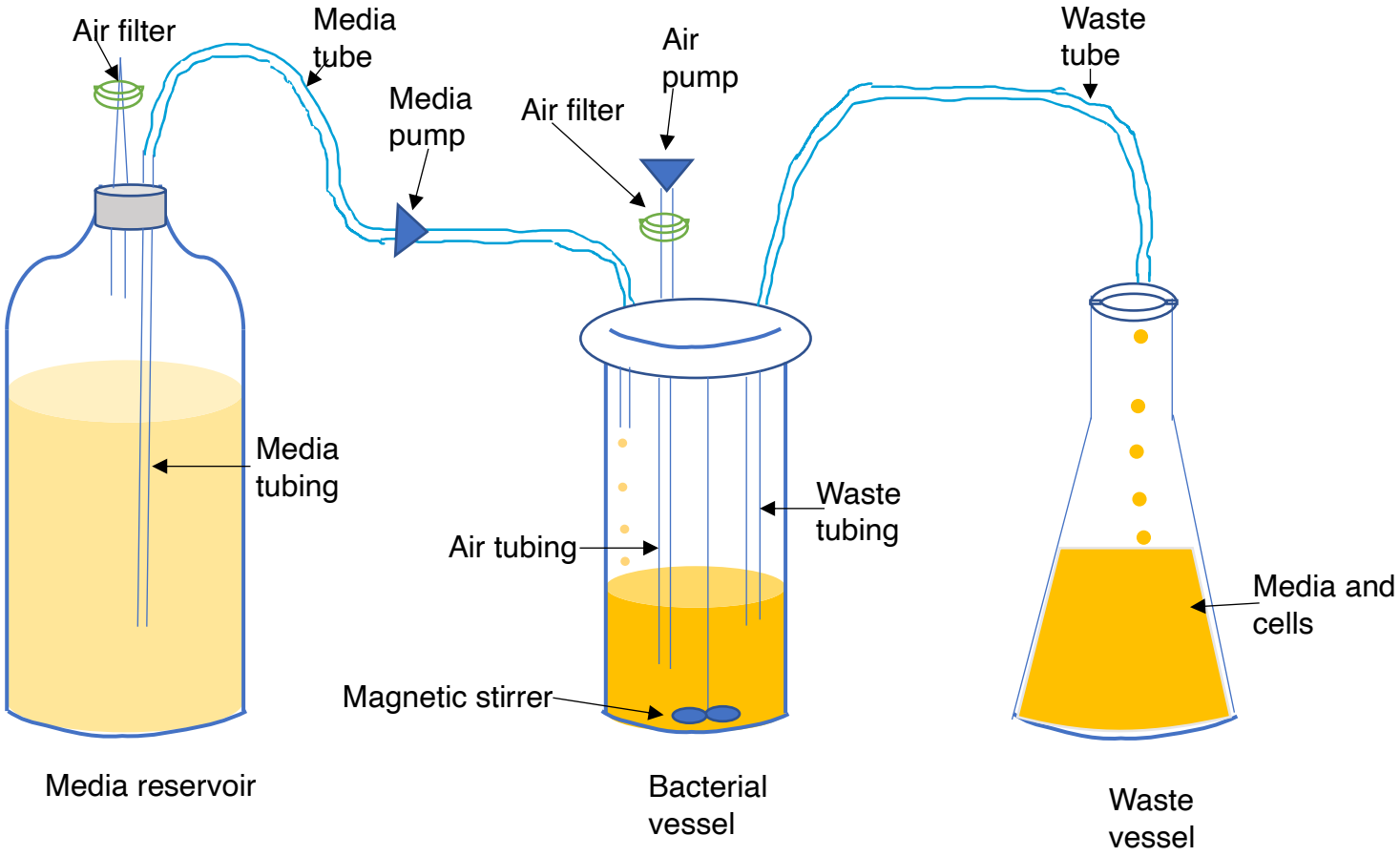


Figure 1.4: One-chambered chemostat with media reservoir (needed for growth of bacterial cells) continuously pumped into the bacterial vessel via a peristaltic pump. The culture is being continuously stirred to ensure homogeneity and aeration. Also attached to the bacteria vessel is a filter for further aeration and pressure equalisation across the system. An equal volume of spent media and cells were removed into the waste vessel.

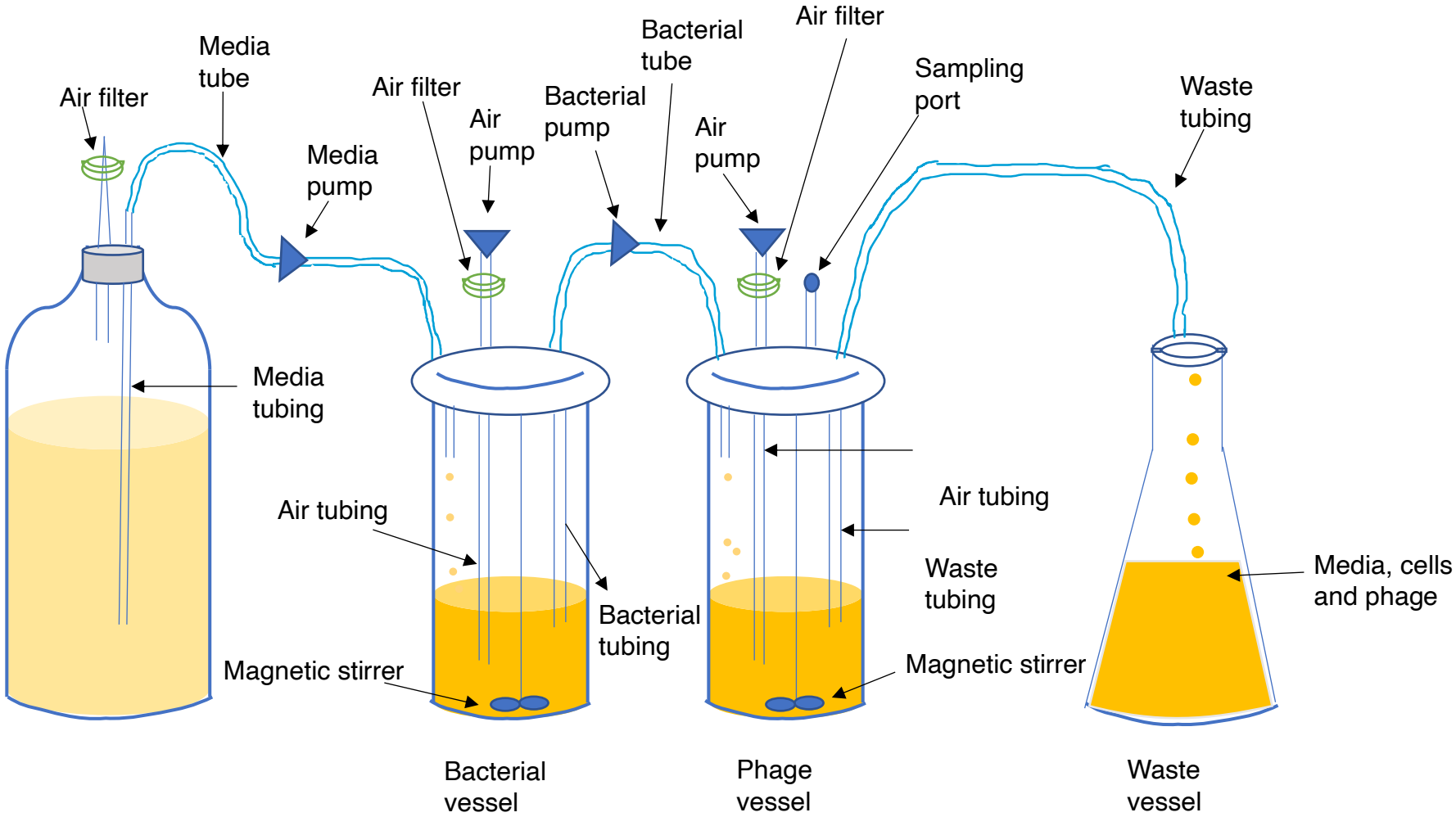


Figure 1.5: Two-chambered chemostat with media reservoir needed for bacterial cells growth continuously pumped into the bacterial vessel via a peristaltic pump. In the same way, naïve bacterial cells (that have not encountered phage) are supplied to phage vessel for propagation at the same rate to maintain constant volume. The culture and phage vessels are continuously stirred to ensure homogeneity and aeration. Attached to the bacteria and phage vessels are filters for further aeration and pressure equalisation across the system. An equal volume of spent media, phage and cells were removed as waste.

1.9 Aims and objectives

The primary aims of this study are to investigate fitness within new hosts and examine the evolutionary processes that occur during host switching, as well as ecological factors and genetic constraints that influence host-switching capability. To achieve this, a well-studied model organism ΦX174 was used, alongside the hosts, *E. coli* C, mutant strains of *S.* Typhimurium and *E. coli* K-12$^{gmhB\text{-mut}}$. Bespoke chemostat culturing system was set-up in order to hold all culture conditions constant, as far as possible, except the manipulated variable – in this case the host.

Although it has been shown that a mutant strain of *E. coli* K-12$^{gmhB\text{-mut}}$ was sensitive to ΦX174 (Ohkawa 1979; Michel *et al*., 2010), no study has examined continuous evolution of ΦX174 with *E. coli* K-12$^{gmhB\text{-mut}}$. In this thesis continuous culturing of *E. coli* K-12$^{gmhB\text{-mut}}$ was described with the aim to investigate fitness within new (non-laboratory) hosts. For the principal host-switching experiment, ΦX174 was repeatedly switched between *E. coli* C and *S.* Typhimurium. There are reports in the literature of ΦX174 adaptive evolution on *S*. Typhimurium and *E. coli* C hosts (Bull *et al*., 1997; Crill *et al*., 2000; Brown *et al*., 2013). Bull *et al*. (1997) and Crill *et al*. (2000) explored host switching between these two hosts at temperatures of 43.5^{o}C, in a continuous culturing system with different dynamics, and fitness rates were measured using the double agar overlay assay plate method (Clokie and Kropinski, 2009). Reversal in the fitness, attachment rates of ΦX174 in the hosts were recorded, and the effects on fitness and attachment rates of a particular mutation within the coat protein gene were investigated using targeted mutagenesis (Crill *et al*., 2000). However, in the study described in this thesis, ΦX174 adaptation in these hosts was allowed to proceed at a temperature of 37^{o}C (optimum growth temperature for hosts used) and the quantitative polymerase chain reaction (qPCR) method was employed for measurement of fitness and attachment rates. This study sought to investigate fitness effects and the rate of attachment of specific allelic variants associated with host switching events, and the effects of variants at

some sites have not been reported in the literature prior to this study. By utilising next- and third-generation sequencing, and qPCR, I was able to analyse different phenotypes and genotypes produced as viruses and hosts evolution occur, a situation analogous to the early stages of viral disease emergence which is a central to determinants of future viral disease emergence. Reversals in the frequencies of ancestral and derived alleles were noted when hosts were switched.

The primary objectives of this study were to:

- Develop a chemostat continuous culturing system for the propagation of ΦX174 on a mutant strain of *E. coli* K-12$^{gmhB\text{-mut}}$, the common laboratory host *E. coli* C and the novel host *S.* Typhimurium (Chapter 3).

- Evolve ΦX174 on *E. coli* K-12$^{gmhB\text{-mut}}$, *E. coli* C and *S.* Typhimurium hosts (Chapter 4).

- Switch evolving lineages of ΦX174 alternately between *E. coli* C and *S.* Typhimurium hosts (Chapter 4).

- Determine the growth and attachment rates of ΦX174 on these hosts in liquid culture and identify associated trade-offs that exists during host-switching (Chapter 4).

- Use next-generation DNA sequencing to investigate the evolutionary dynamics of ΦX174 populations in both hosts (Chapter 5).

- Establish genomic signatures of ΦX174 associated with host switching and determine specific host allele variants that either permit ΦX174 to host switch or increase fitness in the current host (Chapter 5).

- Investigate the fitness effects of some mutations through targeted mutagenesis using PCR-based site-directed mutagenesis (Chapter 6).

Chapter Two: Materials and methods

2.1 Biological strains

2.1.1 Bacteria

The *Escherichia coli* wildtype C1 obtained from Yale Coli Genetic Stock Center (strain CGSC3121) was used throughout the experiments. Genome sequencing and *de novo* assembly was performed by Dr Anton Nekrutenko of Pennsylvania State University USA. A mutant strain of *E. coli* K-12$^{gmhB\text{-mut}}$ was also obtained from the Yale Coli Genetics Stock Center (CGSC11679, JW0196-2), with knockout of all non-essential genes including a mutation on the *gmhB* gene, required for the biosynthesis of the first heptose sugar in the inner core of LPS (Baba *et al*., 2006), and used for this experiment. *S.* Typhimurium GalE^{-} , type I restrictionless (hsd), ϕX174^{s} *S*. enterica serovar Typhimurium, LT2 strain IJ750 [xyl-404 metA22 metE551 galE719 trpD2 ilv-452 hsdLT6 hsdSA29 hsdSB121 fla-66 rpsL120 H1-b H2-e nix] (provided by M. M. Susskind to I. J. Molineux as MS3849) was kindly provided by Dr Holly Wichman from the University of Idaho, USA. This strain was introduced as a novel host in the principal host switching experiment. Whole-genome sequencing with Illumina's MiSeq and Oxford Nanopore's MinION platforms was performed in house. *E.coli gro89* (*rep*$^{-}$) mutants (obtained from Bentley Fane's laboratory, University of Arizona) were used in attachment assays. The host cell *rep* protein is essential for ΦX174 replication in both stage II and III DNA synthesis (Ekechukwu *et al*., 1995). Therefore, a dysfunctional REP protein will inhibit DNA synthesis and genome packaging in ΦX174.

2.1.2 Bacteriophage

Bacteriophage ΦX174 wildtype was provided by Dr Holly Wichman of University of Idaho. The sequence of this phage was obtained from GenBank with accession AF176034.1. Deep sequencing of phage samples from the principal host switching experiment was performed in house on Illumina's MiSeq platform.

2.1.3 Plasmid

The plasmid used in this work as a sequencing spike-in was pUC18 (GenBank accession L09136), to identify the presence of contamination in library-prepared samples. dsDNA was purchased from Thermofisher Scientific (SD0051, Paisley, UK).

Host Strains	**Parent**	**Derived**	**LPS mutation pathway**	**Literature**
Escherichia coli	C-1(F–)	*E. coli*	Galactose (Gal1)	Feige, 1981
Escherichia coli	*gro89*	*E. coli*	Rep Protein	Ekechukwu *et al., 1995*
Escherichia coli	K-12	K-12	Heptose	Baba *et al*., 2006
Salmonella enterica serovar Typhimurium	LT2 (type I restriction less)	IJ750	Galactose (GalE)	Hone *et al*., 1987

Table 2.1: Hosts used in this study.

2.2 Microbiology methods

2.2.1 Culture media, buffers and solutions

2.2.1.1 Salt solutions

Salt solutions were used to supplement Lysogeny broth (LB) and agar for phage propagation and dilution. Calcium chloride dihydrate ($CaCl_2 \cdot 2H_2O$, C3306, Sigma-Aldrich, Dorset, UK) was used as a source of Ca^{2+} and magnesium chloride hexahydrate ($MgCl_2 \cdot 6H_2O$, M2670, Sigma-Aldrich), as a source of Mg^{2+}, prepared by dissolving these salts in distilled water and autoclaving. These were needed for ΦX174 recognition of bacterial cells and attachment.

2.2.1.2 Media

LB and Bacto™ agar were used. LB composition per litre (L) is: 10g Tryptone, 5g yeast extract and 10g NaCl. LB was prepared by adding dehydrated LB broth (DM370, Appleton Woods, Birmingham, UK) to distilled water at a concentration of 25g / L, supplemented with $CaCl_2 \cdot 2H_2O$ (to 2mM final concentration) and $MgCl_2 \cdot 6H_2O$ (to 10mM), before autoclaving for 15 minutes at 121°C, 15 psi for all liquid cultures in this experiment.

For Bacto agar (MN663, Appleton Woods) there were two types of agar prepared, top and bottom agar, which differ in the mass (weighed out in grams) added to LB broth. Top agar was prepared by adding dehydrated Bacto agar at a concentration of 7 g / L and LB at 25 g / L. The top agar was supplemented with $CaCl_2 \cdot 6H_2O$ and $MgCl_2 \cdot 6H_2O$. Bottom agar was prepared by adding dehydrated Bacto agar at 15 g / L concentration and LB at 25 g / L to distilled water and autoclaving for 15 minutes at 121°C. Top agar was stored at 55°C until needed while bottom agar was cooled until comfortable to handle and poured in Petri dishes (approximately 25 mL per dish as base layer), left to solidify and used as required.

2.2.2 Bacterial cell stock and overnight cultures

Aliquots of bacterial cells, for both *E. coli* C and *S.* Typhimurium were prepared from large cell cultures growing at exponential phase. An LB agar plate was streaked out with wildtype bacterial cells *E. coli* C and *S.* Typhimurium and incubated overnight at 37°C. A single isolated bacterial colony was picked with a sterile inoculating loop and transferred into a 250 mL Erlenmeyer flask containing autoclaved LB broth. The mixture was incubated overnight at 200 rpm and 37°C. Several sterile 15 mL centrifuge tubes were prepared, containing 10 mL of the overnight bacterial culture and 0.7 mL of Dimethyl sulfoxide (DMSO; D8418, Sigma-Aldrich). The mixtures were thoroughly mixed by vortexing and stored in the -80°C freezer for further use. Overnight bacterial cultures were used throughout all time points in each chemostat experiment. These were prepared by inoculating bacteria from a frozen cell stock, using a sterile inoculating loop, into 50 mL centrifuge tubes containing 15 mL of LB broth and incubated overnight at 37°C at 150 rpm before use.

2.2.3 Phage plaque assay and stock

E. coli C only was employed for the preparation of ancestral phage stock. Bacteriophage can be grown in both liquid and solid media. Prior to carrying out this experiment, phage stock was produced and stored at -80°C. A 1.5 mL micro-tube containing 0.9 mL LB and 0.1 mL bacterial cells was placed on a Thermomixer (PMHT, 13479429, Fisher Scientific, Loughborough, UK) at 37°C, shaking at 650rpm and incubated for 90 minutes. Thereafter, a sterile loop was used to inoculate ΦX174 from frozen glycerol stock into the cell culture and returned to Thermomixer for an additional one hour. Since a pure phage sample uncontaminated with bacterial cells was needed, it was necessary to kill the cells and fully separate phage from cell debris. 10% v/v chloroform ($CHCl_3$), 0.1 mL into 1 mL of phage plus bacteria (in LB media), was added to lyse bacterial cells, mixed thoroughly by a vortexing and

immediately spun down in a microcentrifuge for 5 minutes at 11,000 rpm (= bacterial cell removal procedure). Cellular debris were thereby separated from chloroform and phage. The supernatant, containing unattached phage, was carefully removed with a sterile filter tip pipette and transferred to a clean micro-tube.

In order to isolate a single plaque, a dilution series of phage was made by serially diluting 100 µl of the phage aliquot into 900 µl of LB broth. Serial dilution from 10^{-1} up to 10^{-7} was carried out in 1.5 mL micro-tubes (vortexed thoroughly between transfers). 4 mL of already-dispensed top agar kept in a molten state (55^{o}C) in a water bath was mixed with 100 µl from each dilution and 100 µl of freshly prepared overnight bacteria culture. The molten top agar (containing phage and bacterial cells) was mixed thoroughly by flipping up and down, collected at the bottom of the tube and poured on solidified thin bottom agar plates, allowed to set and incubated at 37^{o}C for 4 hours. A plate with a few, well-spread and clearly distinct plaques was chosen.

A new stock of the wildtype ΦX174 was prepared by isolating and harvesting a single clearly distinct plaque. Using a sterile plastic inoculating loop and a toothpick, the top agar portion of a chosen plaque was cut out, lifted with the toothpick and transferred into the 1.5 mL micro-tube containing 0.75 mL of LB broth and mixed for about 10 seconds on a vortex device. Following this, bacterial cell removal was carried out as described, and 0.6 mL of the phage supernatant was transferred into a new sterile 1.5 mL microtube. 1 mL of frozen bacterial cell stock was inoculated into a 50 mL centrifuge tube containing 10 mL LB broth and incubated at 37^{o}C and 200 rpm until bacteria reached exponential growth phase at an OD_{600nm} of approximately 0.4. The harvested phage stock was used to inoculate bacterial cells at MOI < 0.1 and incubated for 3 hours in order to achieve high titer phage stock. Thereafter, phages were separated from bacterial cells by adding 10% v/v chloroform as in the bacterial cell removal procedure, 8 mL of phage supernatant was stored in a fresh 50 mL centrifuge tube, ~7% v/v DMSO added, before the

mix was vortexed and the stored in the -80^{o}C freezer. This was used as ancestral phage throughout the experiment.

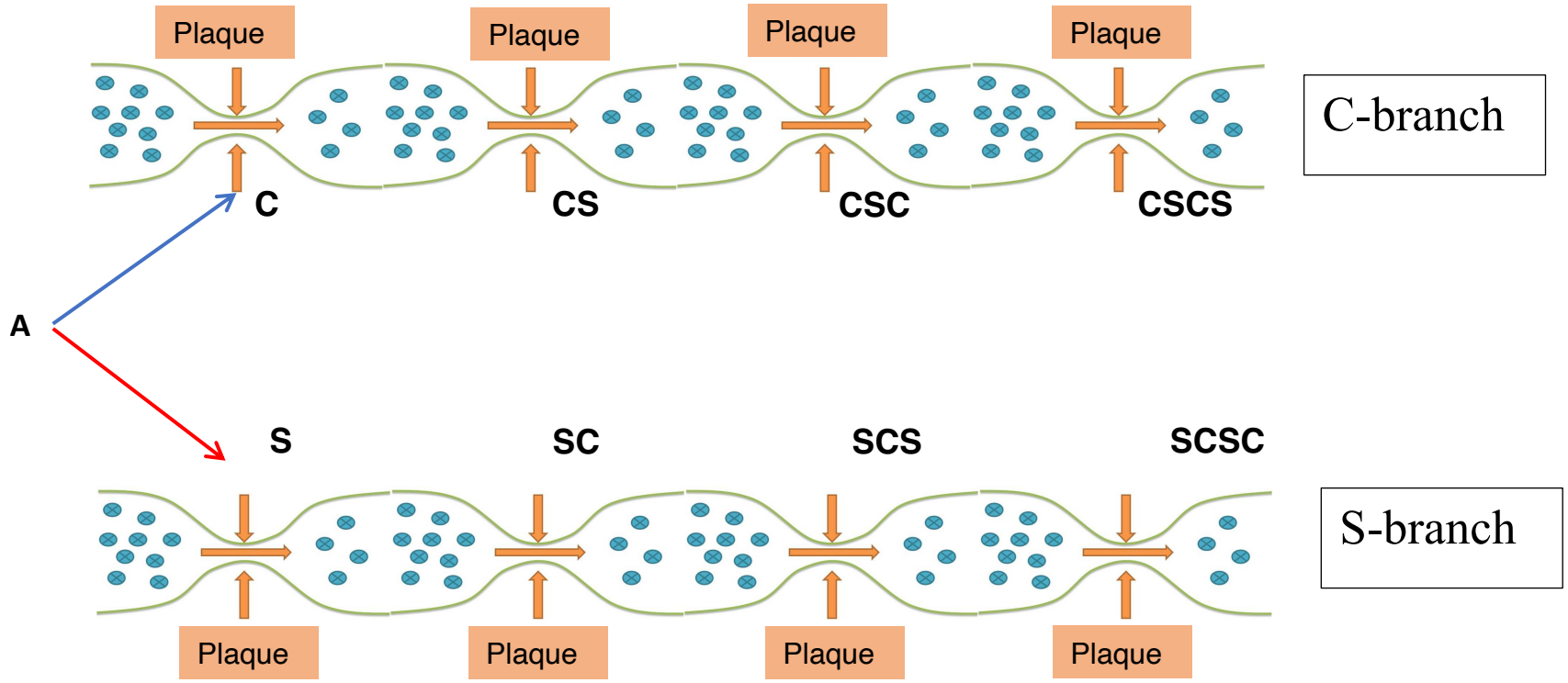


Figure 2.1: Host-switching experimental scheme: Samples are represented by letters with A indicating the ancestral ΦX174 sample. The last letters in each sample name refer to the current host exposure (viz: C**S,** SC**S**, CSC**S** – last host was *S.* Typhimurium, S**C**, CS**C**, SCS**C** – last host was *E. coli* C). The blue arrow shows the transfer that initiated C-branch, the ΦX174 lineage which first encountered *E. coli* C, while the red arrow shows the transfer that initiated S-branch populations (begun on *S.* Typhimurium). A single plaque was isolated (arrows up and down) and used to inoculate the next round of chemostat adaptation (between each host switching transfers), resulting in population bottlenecks within each branch/lineage.

2.2.4 Alternating host-switching experimental design

The key culturing system, a chemostat, used in this chapter is described in section 3.3.1. The system shown in figure 3.3 was used in continuous culture of *E. coli* K-12$^{gmhB\text{-mut}}$, while figure 2.4 shows the system used in culture of both *E. coli* C (C) and *S.* Typhimurium (S) which involved alternating switching between these two hosts.

ΦX174 that were adapted to general laboratory conditions (from Holly Wichman's laboratory, University of Idaho) were used in the host switching experiment, and grown on *S.* Typhimurium or *E. coli* C.

After the initial 10 days of growth on a given host, a single plaque was picked for chemostat inoculation for next period of selection on the alternate host which is 10 days (figure 2.1), therefore introducing a severe bottleneck. In the first instance, ΦX174 was initially grown for ~720 generations on the two hosts separately, then alternately switched between hosts for another ~720 generations for four consecutive times (figure 2.1). The last letter in each sample name refers to the most recently encountered host (viz: C**S,** SC**S**, CSC**S** – last host was *S.* Typhimurium; S**C**, CS**C**, SCS**C** – last host was *E. coli* C). ΦX174 were subjected to three consecutive periods of growth after the initial growth on each host indicated as C-adapted (alternated for three consecutive periods, viz, S**C**, CS**C**, and SCS**C**), or if on *S.* Typhimurium, S-adapted (viz, C**S**, SC**S**, CSC**S**; figure 2.1). Two experimental 'branches' were lines of independent replicates, C-branch began on *E. coli* C (viz, **C**, **C**SC, **C**SCS) and S-branch began on *S.* Typhimurium (viz, **S**, **S**C, **S**CSC). 10ml of samples were taken daily through the dedicated sampling port (figures 3.3, 3.4), and treated as described in section 3.3.1.4.1.

2.3 Molecular methods

2.3.1 Quantitative PCR for phage enumeration

Concentrations, in phage genome units (pgu) per ml, were measured using qPCR (Vale *et al*., 2012), providing an accurate method for phage growth evaluation in a liquid environment (in this respect more closely resembling the environment in which phage growth was carried out than a double agar overlay assay, section 2.2.3). A customized pair of qPCR primers (forward and reverse) targeting the genomic region 590 to 608 (a region of gene E, the gene that codes for the lysis protein, table 1.1), or 3,716 to 3,821 (a region of gene H, the gene that codes for the DNA pilot protein, table 1.1) and a positive control ΦX174 supplied at a copy number of 2 x 10^5 was used throughout the experiment (table 2.2). Lyophilised primer mix was resuspended in nuclease-free water. Following this, the positive control was diluted (via serial dilution) and used to prepare a standard curve according to the manufacturers' manual (Primerdesign Ltd, Chandler's Ford, UK). The qPCR was set up in Lightcycler 480 96 well plates (AXP480, Appleton Woods limited). Each reaction contained: 1 µl resuspended primer mix, 10 µl PrecisionPlus SYBRgreen master mix (Mini-PrecisionPLUS-SY, Primerdesign Ltd, Chandler's Ford, UK), 4 µl of nuclease-free water, and 5 µl phage lysate. The plates also contained a negative control and positive control with nuclease-free water or standard ΦX174, respectively, in place of the phage lysate. qPCR were performed with Lightcycler 480 controlled by version 1.5.0.39 of the (Roche) software, using the following conditions: 95^{o}C for 2 minutes, 40 cycles of 10 seconds at 95^{o}C, 60 seconds at 60^{o}C. All reagents (primers, SYBR green master mix and known concentration of phage lysate positive control for establishing standard curve) were from Primerdesign (Primerdesign Ltd).

An enumeration of viral particles was carried out using qPCR. A conventional method of virus quantification to calculate infectivity of bacteriophages stock involves visually counting individual plaques produced during overlay assay

of phage and bacterial infection in a semi-solid nutrient medium (section 2.2.3), resulting in an estimate of plaque-forming units (PFU) per millilitre. Healthy host cells, actively dividing in log-phase with more than 95% viability are critical for a successful plaque assay (Clokie and Kropinski, 2009). The plaque assay is considered effective for host bacteria that form a complete lawn on solid media, viruses that are capable of multiplying via infecting and lysing cells with clear plaques formed or, for those lysogenic phages that integrate with the host chromosome, cloudy plaques (Clokie and Kropinski, 2009). PFU results are affected by numerous factors, among these are plaque size, viral morphology, attachment rate, diffusion rate and change in salt concentrations, resulting in poor reproducibility of the assay (Anderson *et al*., 2011; Gallet *et al*., 2011). An important part of this work is the growth system for propagating phage in aqueous medium. Therefore, it would be ideal to enumerate phage in aqueous solutions.

With qPCR, measurement of phage genome units (pgus) takes place in a similar media composition environment (liquid form) in which the growth assays were carried out and the use of solid media (an environment that differs substantially from the experimental conditions) is avoided. Also, qPCR assay allows all the experimental replicates to be carried out simultaneously, which helped to keep experimental error low, as well saving time and space. However, qPCR has its own limitations. qPCR involves incorporation of fluorescent chemicals, dsDNA-intercalating dyes (for example, SYBR green) or hydrolysis probes (for example, Taqman), into the reaction, allowing synthesis of PCR products to be directly detected and visualized in real time, making it possible to quantify pgus in the sample. Because it is PCR-based, a set of oligonucleotide primers is needed to target specific sequence of interest. Poorly designed primers may form primer-dimers and decrease PCR efficiency, and primers that target sites with high mutation or substitution rates may give inaccurate results. For ΦX174 adaptation on *E. coli* K-12$^{gmhB\text{-mut}}$, primers targeting a section of Gene H were used (table 2.2). For the ΦX174 host-switching study in *S.* Typhimurium and *E. coli* C, primers

targeting Gene E were used (table 2.2), because it is well conserved (Bull *et al*., 1997; Crill *et al*., 2000; Poon and Chao, 2005; Pepin and Wichman 2008; Dickins and Nekrutenko, 2009; Brown *et al*., 2013) and not among the common sites of amino acid substitution during experimental evolution (Wichman *et al*., 2005; Wichman 2010). Gene E is an alternative reading frame within gene D, therefore, the sequence change is further constrained by multiple reading frames (figure 1.2). Also, there is a concern of qPCR estimating both viable and non-viable DNA phage in solution. Some researchers, even in the face these shortcomings, demonstrated that qPCR is the most precise form of bacteriophage particles enumeration when compared with other assays (Edelman and Barletta, 2003; Anderson *et al*., 2011; Klopot *et al*., 2017). Edelman and Barletta (2003) found out that qPCR has a strong correlation with the PFU measure obtained via the overlay assay. This matches the observation in this work that qPCR-based quantification is well-correlated with PFU quantification (data not shown).

2.3.2 Fitness / growth rate assays

Fitness assays were conducted over 45 minutes by examining increases in DNA concentration in liquid culture using SYBR Green-based qPCR. The phage sample to be assayed and bacterial cells were removed from -80^{0}C storage and defrosted on ice. Triplicate 2.0 mL microfuge tubes containing 0.9 mL LB medium and 0.1 mL host cells were incubated on a Thermomixer at 37^{o}C, shaking at 650rpm for 1hour 30 minutes. Host cells were grown to a density of ~1x10^8 µl. ~10^4 µl phage sample was added to individual tubes of the exponentially growing host cells and mixed briefly on a vortex device. A volume of 0.2ml of the suspension was removed immediately from the tubes, treated with chloroform (10% v/v), vortexed, centrifuged at 11,000 rpm for 4 minutes (X_0). After a further 45 minutes of growth (T), a second sample was taken (for each replicate) and treated the same way (X_i). Aliquots were quantified using qPCR to determine pgus as described in section 2.3.1. Fitness rates were estimated as:

$$Log_2 (X_i - X_0) / T$$

where T = time in hours. X_i= phage titer at 45 minutes and X_0 is phage titer at 0 minutes as in Bull *et al*. (1997). qPCR measures were triplicated for each sample drawn from a liquid culture, resulting in three technical replicates for each time point. Plotted data show biological triplicates (figures 4.3, 6.4).

2.3.3 Attachment assays

Firstly, to determine the pattern of phage concentration decline in solution, the approximate time period when most ΦX174 would have attached to host cells, several 2.0 mL microtubes were set-up on a Thermomixer containing 0.9 mL LB and 0.1 mL *gro*89 bacterial cells. The tubes were incubated at 37°C, shaking at 650 rpm for 1hour 30 minutes. Thereafter, ~10^8 of the exponential phase host cells were mixed with phage (MOI < 0.1). ΦX174 phage was sampled at different time points: T = 0, 8, 11, 14, 16 and 24 minutes using 2 ml sterile disposable syringes and immediately passed through sterile 0.22 µm cellulose acetate filters (Triple Red Laboratory Technology), retaining cells and adsorbed phage while aliquots containing unattached phage passed through the filter. Aliquots were placed on ice for further quantification using qPCR (section 2.3.1). The time period was chosen (figures 4.4, 4.5, 4.6), at T = 8 minutes, and final concentrations of unabsorbed phage in triplicates were determined from the aliquot. Attachment rate was calculated using the following equation:

$$A\mu = \frac{-In\,(N_f / N_i)}{Y \times T}$$

Aµ is the rate of attachment at a given time T, N_f the is number of unattached phages, N_i is total number of phages added, Y is the concentration of bacterial cells and T is incubation time in minutes as in Pepin *et al*. (2006). Attachment rate of phage is expressed as per cell per minute.

Primer name	Primer sequence (5’ – 3’)	Purpose	Section
QPX-590-F	ATACCCTCGCTTTCCTGCT	qPCR Quantification	2.3.1
QPX-608-R	CGCCTTCCATGATGAGACA		
QPX-3716-F	TCATCAGCAAACGCAGAATCAG		
QPX-3821-R	AATATCAACCACACCAGAAGCAG		
PHX-0001-F	GAGTTTTATCGCTTCCATG	Sanger sequencing (PCR Amplicon A)	2.5.5
PHX-2953-R	CCGCCAGCAATAGCACC		
PHX-2605-F	CAGGTTGTTTCTGTTGGTGCTG	Sanger sequencing (PCR Amplicon B)	2.5.5
PHX-0379-R	CTTGACTCATGATTTCTTACC		
PHX-2536-R	TCAAACTATCAAAATATAACGTTGACGATG	Sanger sequencing (Source Bioscience)	2.5.5
PHX-2007-R	CGGAAAACATCCTTCATAGAA		
PHX-3381-R	GATTCTCAAATCCGGCG		

Table 2.2: Primers used for qPCR and Sanger sequencing.

2.3.4 Media for isolation of mutants

The media used in this chapter were described in section 2.2.1.2, with the addition of glycerol and tetracycline hydrochloride (THA; 10460264, Fisher scientific) to both layers in the double-agar assay (top and bottom agar) to improve plaque size for easy isolation (Santos *et al.*, 2009). 5% glycerol (12144481, Fisher scientific) was added to 250 ml of both agar used prior to autoclaving. 1.5 mg / L THA was dissolved in nuclease-free water and added to freshly prepared overnight *E. coli C* bacterial cells. Super optimal broth with catabolite repression (SOC; 15544034, Fisher Scientific) medium was used for transformation.

2.3.5 Preparation of competent cells

Chemically competent *E. coli C* cells were prepared using 0.1 M $CaCl_2$ (Section 2.2.1.1) and heat shock at 42^{o}C. Bacteria cells from frozen stock were used to streak LB agar plate and incubated at 37^{o}C overnight. A single colony was inoculated into 50 mL centrifuge tubes containing 10 mL LB broth (a different batch was used for site directed mutagenesis experiment) and incubated overnight at 37^{o}C in a shaking incubator. 1 mL of the overnight cells culture was added to a 50 mL LB and incubated in a shaking incubator at 37^{o}C until $OD600_{nm}$ was ~ 0.3 during the exponential phase. The cell culture was then divided into 25 mL volumes in prechilled 50 mL centrifuge tubes, and chilled on ice for 10 minutes. Thereafter, tubes were placed in a 4^{o}C prechilled Heraeus mega-centrifuge (Model X3/X3F, Thermo-Fisher Scientific, Paisley, UK) and centrifuged for 10 minutes. Supernatants were discarded, and pellets resuspended in 5 mL ice cold 0.1M $CaCl_2$. Tubes were placed on ice for 30 minutes and immediately centrifuged and supernatant discarded. Cells pellet were resuspended in 500 µl ice cold 0.1 M $CaCl_2$, cryoprotected with 7% DMSO v/v and stored at -80^{o}C for future use.

2.3.6 PCR-based site-directed mutagenesis

Targeted mutagenesis was done using a PCR-based reaction. Here, a wildtype phage plaque (section 2.1.2) was isolated according to protocol in section 2.2.3, and diluted to a concentration of 0.5 ng / µl. PCR reactions were set up using a customised pair of primers containing desired mutation sites (table 2.3) using the Phusion mutagenesis kit (Cat. no. F541, Thermo-Fisher scientific, Paisley, UK). Each template reaction was set up in 200 *µ*l PCR tubes, containing: 32.5 µl nuclease-free water, 10 µl 5 x Phusion HF buffer, 1 µl 10mM dNTPs, 2.5 µl of each primer at 10 µM, 1 µl of 0.5 ng / µl phage template and 0.5 µl Phusion DNA polymerase. The amplification reaction was carried out in a thermal cycler with the following conditions: 98^{o}C for 30 seconds; 25 cycles of 10 seconds at 98^{o}C, 30 seconds at different annealing temperatures on table 2.3, 2 minutes 40 seconds at 72^{o}C; 72^{o}C for 10 minutes and held at 4^{o}C. A control reaction was also set up, according to manufacturer instructions. The ligation reaction was carried out at 25^{o}C, with each ligation reaction containing; 5 µl of ligation mix, 2 µl of 5X rapid ligation buffer, 2.5 µl nuclease-free water, and 0.5 µl T4 DNA ligase in a 500 µl microcentrifuge tubes, mixed thoroughly, briefly centrifuged and incubated at 25^{o}C for 5 minutes.

Primer name	Primer sequence (5' – 3')	Site	Annealing Temperature*
OYE4-MUT-F	TAACACTACT**C**GTTATATTGACCATGCC	G 1304 → C	63.2°C
OYE4-MUT-R	ACAGTCGGGAGAGGAGTG		
OYE2-MUT-F	CGATTCAATCAT**A**ACTTCGTGATAA	G 2275 → A	60.6°C
OYE2-MUT-R	CGAGTGGTCGGCAGATTG		
OYE6-MUT-F	CCCTGATGAG**T**CCGCCCCTAG	G 3129 → T	66.1°C
OYE6-MUT-R	TTAGGAACATTAGAGCCTTGAATGGC		

Table 2.3: Primers used for site-directed mutagenesis. Red nucleotides indicate the mutant sites introduced. * The annealing temperature was estimated using Thermo-scientific melting temperature calculator (https://www.thermofisher.com/uk/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/tm-calculator.html).

2.3.7 Transformation of *E. coli* C with prepared phage mutants

The ligation reactions were transformed into *E. coli* C competent cells using the heat shock method. 100 µl aliquots of competent cells were added to 200 µl thin-walled PCR tubes, 5 µl of ligation reactions were added to the competent cells, and chilled on ice for 1 hour. The PCR tubes were transferred to a heated water bath at 42^{o}C for 1 minute, and immediately returned to ice for 2 minutes. All tube contents were transferred to 1.5 mL micro-tube containing 500 µl SOC medium, incubated in a Thermomixer at 37^{o}C for 1 hour, 10% v/v chloroform added, mixed thoroughly by a vortex device and immediately centrifuged for 5 minutes at 11,000 g. Supernatant was carefully removed with sterile filter pipette tips and transferred to new autoclaved microfuge tubes.

2.3.8 Mutants phage plaque purification

To increase the plaque size, double-layer-agar plates were supplemented with glycerol and bacterial cells with THA (section 2.3.4). 500 µl of supernatant from phage mutant transformation was added to 100 µl freshly prepared overnight culture, 4 mL top agar at molten state (55^{o}C), poured on bottom agar plates, allowed to set and incubated at 37^{o}C for 4 hours. A single plaque was isolated from each plate, placed in 1.5 mL micro-tube containing 500 µl LB broth, treated with 10% v/v chloroform, mixed thoroughly and centrifuged for 5 minutes at 10,950 RCF. The supernatant was then carefully aliquoted and kept aside for phage purification. The isolated phage aliquots were purified by re-growing in bacterial cells. 1.5 mL micro-tubes tubes containing 0.9 mL LB broth and 0.1 mL bacterial cells were incubated at 37^{o}C in a thermomixer for 90 minutes. The isolated phage aliquots were added to these cultures and grown for a further 45 minutes before being treated with chloroform (as above). Aliquots were added to THA supplemented overnight cells and 4 mL top agar, incubated for 4 hours at 37^{o}C. The mutants were isolated, plaque purified via second isolation of mutants, and PCR prepared for validation with Sanger sequencing described

in section 2.11. Fitness and attachment assays were performed as described in sections 2.3.2 and 2.3.3.

2.3.9 Agarose gel electrophoresis

A 1 % w/v agarose gel was prepared by dissolving 0.5 g of agarose powder (9012-36-6, Fisher Scientific) in 50 mL Tris-Acetate-EDTA buffer (TAE: 40mM Tris Acetate, 2mM Na_2EDTA) (EC-872, National Diagnostics, Nottingham, U.K) using a microwave oven. The gel solution was cooled down to about 50^{o}C, and a 1 : 10,000 volume, 5 µl of SYBR® Safe DNA Gel Stain (S33102, Invitrogen, Thermo-fisher scientific) added. The molten gel preparation was poured into a gel casting tray and, once solidified, it was placed in an electrophoresis tank (Geneflow) and submersed in 1 x TAE buffer. Aliquots of PCR product were mixed with Blue and orange 6 x l (G1881, Promega, Southampton, U.K) in a ratio 1 Dye : 5 PCR product parts. 5 µl of the mixtures were loaded into each well along with 5 µl of appropriate molecular weight makers, 100 bp or 1 Kb DNA Ladders (G5711, Promega) depending on expected fragment sizes. Electrophoresis was performed at 100 volts for 45 minutes. The gels were visualised under Ultraviolet (UV) light to observe DNA bands, using Gel Documentation System (Syngene) and analysed with Gene System software (V1.5.5.0).

2.4 Graphics and statistical analysis

Graphics were generated using R Studio (v.3.5.5; https://www.rstudio.com/), ggplot2 (v.3.1.0) and tidyverse (v.1.2.1), with colour-blind compatible palettes generated via médialab's "i want hue" portal at http://tools.medialab.sciences-po.fr/iwanthue/ (except figures 4.4, 4.5 and 4.6; generated with Minitab v.18.0). Statistical analyses were performed for *E. coli* C and *S.* Typhimurium fitness and attachment rates in chapter 4 (sections 4.3.2 and 4.3.5 respectively) with R, p-values calculated with non-parametric ANOVA using package vegan with function adonis2 (MacArdle and Anderson, 2001) permutations set to 10,000 and the seed was set to

(124) for data reproducibility. Statistical analyses were performed for *E. coli* C and *E. coli* K-12$^{gmhB\text{-mut}}$ (chapter 4, section 4.3.1) using Mann-Whitney test with R. Also, for site directed mutagenesis fitness and attachment rates in chapter 6 (sections 6.3.2 and 6.3.3), analyses were performed with Mann-Whitney test in R. The confidence intervals (95% CI) of the observed medians recorded; F values calculated with ANOVA, and mean variance were recorded. p-values of $p > 0.05$ were interpreted to indicate no significant difference, p-values of $p < 0.05$ indicated statistically significant associations, p-values $p < 0.01$ classified as very significant associations, $p < 0.001$ were considered highly significant, and p-values $p < 0.0001$ classified as very highly statistically significant.

2.5 Sequencing methods

2.5.1 Quality and quantity assessment of nucleic acids

Genomic DNA (gDNA) extracted was assessed for both quality and quantity before library preparation (section 2.5.3.1 and 2.5.3.2) and at some stages during library preparation. Assessment of quantity involved determining the total concentration of gDNA in a given sample using a Qubit™ version 3.0 or 4.0 Fluorometer. For gDNA, the concentration of dsDNA was determined using the Qubit dsDNA HS Assay Kit (Q32854, Thermo-Fisher Scientific) if the sample was within a range of 0.2 – 100 ng / µl, or with Qubit dsDNA BR (Q32853, Thermo-Fisher Scientific) for samples within a concentration range of 2 – 1000 ng / µl.

The quality of gDNA was assessed using both a Nanodrop 2000 Spectrophotometer (ND-2000, ThermoFisher Scientific) and an Agilent 2200 TapeStation (Agilent Technologies, Cheshire, UK). The Nanodrop spectrophotometer was used for measuring purity of DNA with a ratio of absorbance at 260/280 nm and 260/230 nm. A target absorbance $A_{260/280}$ ratio of ~1.8, $A_{260/230}$ ratio of ~2.0-2.2 was generally acceptable for the nucleic acid purity. Fragment size distribution (in base pairs) was measured

for gDNA using the Agilent 2200 TapeStation. For gDNA, 2 µl of High Sensitivity D1000 sample buffer (5067-5583, Agilent Technologies) was aliquoted into Optical tube strips (401428, Agilent Technologies), followed by 2 µl of High Sensitivity D1000 ladder (5067-5583, Agilent Technologies) in the first tube in the strip and 2 µl samples in subsequent tubes. The optical tube strips were mixed by a vortex device at 2,000 rpm for 1 minute, spun down briefly and loaded into TapeStation along with loading tips and High Sensitivity D1000 ScreenTape (5067-5582, Agilent Technologies). Analysis was initiated by launching the Agilent 2200 TapeStation software and results collected after a short period of time.

2.5.2 Nucleic acid extraction

2.5.2.1 ΦX174 DNA extraction

For Illumina sequencing library preparation requires dsDNA samples/inputs. An *in vivo* DNA preparation method was used with the procedure undertaken in the last host ΦX174 was selected on for all samples, except where otherwise noted. Using a method adapted from Godson and Vapnek (1973), 3 mL overnight cultures of *E. coli* C or *S*. Typhimurium were added to 5 mL of LB in a 10 mL centrifuge tube and incubated for 2 hours at 37^{o}C, shaking at 180 rpm. 1.5 mL of the ΦX174 phage sample to be sequenced was added to each tube, and grown further for 30 minutes. Next, 30 ng / µl antibiotic chloramphenicol ready-made solution 100 mg / mL in ethanol (R4408, Sigma-Aldrich, Dorset, UK) was added to the culture. Chloramphenicol used in this study inhibits protein synthesis, allowing the continuous bacterial accumulation of Refractive Fraction (RF) dsDNA within the host cells. After 3 hours 30 minutes of growth when high ΦX174 titre has been achieved, tubes were centrifuged at 4,680 x g for 10 minutes and supernatants discarded. Phages treated with chloramphenicol accumulated dsDNA (Godson and Vapnek, 1973) and were extracted using a Qiagen miniprep kit (27104, Manchester, UK), following the manufacturer's instructions. DNA was quantified using a Qubit 3.0 Fluorometer and quality assessed with

Nanodrop 2000 Spectrophotometer at acceptable values for samples necessary for sequencing templates (section 2.5.1).

2.5.2.2 Bacterial DNA extraction

S. Typhimurium gDNA was extracted for Nanopore and Illumina sequencing using the GenElute™ Bacterial Genomic DNA Kit (NA2110, Sigma-Aldrich) following the manufacturer's instructions. 1.5 mL of overnight culture grown in LB at 37°C was centrifuged at 16,000 x g for 2 minutes. A bacterial pellet was resuspended in 180 µl Lysis Solution T, mixed with 20 µl RNase A Solution and incubated for 4 minutes at room temperature. 20 µl of Proteinase K solution was added to the sample, incubated for 30 minutes at 55°C. 200 µl of Lysis Solution C was added to the mixture and further incubated at 55°C for 10 minutes. This was followed by on-column gDNA binding. Freshly prepared 200 µl of absolute ethanol was added to the lysate, mixed thoroughly using a vortex device for 10 seconds to precipitate the DNA and achieve a homogeneous mixture. The lysate was transferred to a pre-prepared column, and centrifuged at 6,500 x g for 1 minute. After, the column was washed twice with 500 µl Wash Solution to remove contaminants, it was centrifuged at 6,500 x g for 1 minute, then at high speed (16,000 x g) for 3 minutes to dry the column. The gDNA was eluted with 60 µl of nuclease-free water by a final centrifugation at 6,500 x g for 1 minute to prevent shearing. The final gDNA was assessed for quality on Nanodrop 2000 Spectrophotometer, and good quality eluate used as template for sequencing.

2.5.3 Deep sequencing

2.5.3.1 Illumina sequencing using Nextera XT kit

DNA libraries were prepared using the Nextera XT DNA Sample preparation kit (FC-131-1024, Illumina, Cambridge, UK) and the Nextera XT DNA Index Kit (24 indices, 96 samples, FC-131-1001). Indexed and paired-end libraries

were prepared in a hard-shell skirted PCR plate for the chosen 24 samples. To control for the likelihood of cross-contamination on the PCR plate during library preparation, DNA spike-ins were employed as described by Dickins *et al*. (2015). For this purpose, a standard, readily available, high copy number cloning vector, pUC18, was utilized. The pUC18 spike-in lacks sequence homology with ΦX174 (data not shown) and are added prior to the preparation of libraries. Spike-ins were added in an alternating fashion such that samples with spike-ins were not kept in close proximity to each other and spike-in-free wells were left between spiked samples.

2.5.3.1.1 Normalisation of genomic DNA

The Nextera XT DNA library kit is sensitive to starting DNA input concentration, therefore it requires all samples to have a uniform concentration for efficiency and sequencing success. The concentrations required for samples varies with organisms. Here, selected samples were normalised to final genomic DNA inputs of 0.2 ng per µl for samples without spike-ins. Spike-in samples were normalised to 0.197 ng per µl of genomic DNA and 0.003 ng per µl of pUC18. The quantification of all genomic DNA was performed using the Qubit 3.0 Fluorometer.

2.5.3.1.2 Tagmentation of normalised genomic DNA

The normalised genomic DNA was fragmented and then tagged with adapter sequences in a single reaction utilizing the Nextera transposome. In a hard-shell PCR plate, the following were added in the order listed: 10 µl of Tagment DNA buffer (TD), 5 µl normalised genomic DNA and 5 µl of Amplicon Tagment Mix (ATM); these were mixed by pipetting. The plate was centrifuged at 280 x g at 20°C for 1 minute, sealed and run on a programmed thermal cycler (55°C for 5 minutes, held at 10°C). 5 µl of Neutralised Tagment buffer (NT) was added to each well, pipetted to mix and incubated at room temperature for 5 minutes. In this way, adapter sequences

were added to fragmented genomic DNA needed for indexing and amplification in the next step.

2.5.3.1.3 PCR amplification and indexing of library

The tagmented DNA was amplified using a PCR based program. PCR steps incorporated Nextera Index 1 (i7) 4 adapters and Index 2 (i5) 6 adapters and sequences required for cluster formation on MiSeq flow cell. The following consumables were added in the listed order into a PCR plate: 5 µl of each index 1 (i7) adapter added vertically to each well, 5 µl of each index 2 (i5) added horizontally to each well, thereby allowing for multiplexed sequencing, 15 µl Nextera PCR Master mix (NPM) to each well, pipetted to mix and centrifuged at 280g for 1 minute. The plate was firmly sealed and placed in a pre-programmed thermocycler to run following these conditions: 72^{o}C for 3 minutes; 95^{o}C for 30 seconds; 12 cycles of 95^{o}C for 10 seconds, 55^{o}C for 30 seconds, 72^{o}C for 30 seconds; 72^{o}C for 5 minutes and held at 10^{o}C.

2.5.3.1.4 PCR library clean-up and size selection

Index PCR products were purified with Agencourt AMPure XP magnetic beads (A63880, Beckman Coulter, High Wycombe, UK) and size-selected to remove short library fragments. AMPure beads (Beckman Coulter) were homogenised and brought to room temperature before use. 30 µl equivalent of 0.6x size selection were added to each well, mixed by pipetting for complete homogenisation, incubated at room temperature for 5 minutes and allowed to stand on a magnetic rack until the supernatant has cleared. The supernatant was carefully removed and discarded. Freshly prepared 80% ethanol (10644795, Fisher Scientific) was used to wash the beads on the magnetic rack twice before being air-dried for approximately 7 minutes. The washed beads were resuspended in 52.5 µl Nextera Resuspension Buffer (RBS) and 50 µl of the supernatant, excluding AMPure beads (Beckman Coulter), was transferred into a new 96-well PCR plate.

2.5.3.1.5 Final normalisation and pooling of libraries

To ensure quality results, equalise library representation and avoid issues that can affect cluster density, each library was normalised to 4 nM. Normalisation prevents overly dilute or concentrated libraries, low sequencing yield or over-clustering on the flow cell. Normalisation involves estimation of fragment size distribution in bp and quantification of the gDNA concentration in ng / µl (measured using the Qubit 3.0 Fluorometer, as described in section 2.5.1). Each library fragment size distribution was measured using Agilent 2200 TapeStation (using the High Sensitivity D1000 sample buffer and ladder as described in section 2.5.1). Depending on each library concentration and the average fragment size of all library, genomic DNA molarity was estimated using the following formula:

$$\boldsymbol{M} = \frac{\mathrm{C} \times \mathbf{10^6}}{\boldsymbol{g} \times \boldsymbol{f}}$$

where M = Molarity (nM, 4nM required), c = DNA concentration (ng / µl), g = average molecular weight of DNA (660 g/mol), f = average fragment size of all libraries. 5 µl of each normalised library was pooled into a single microtube for the subsequent step.

2.5.3.1.6 Sequencing of pooled library

Sequencing was carried out on the Illumina MiSeq platform, using a MiSeq cartridge and V3 reagent kit (MS-102-3003, 600 cycles, Illumina, Cambridge, UK). 5 µl of the pooled 4nM libraries was first denatured with freshly prepared 5 µl 0.2 N NaOH. The denaturing with NaOH ensures that the final library pool consists of single-stranded DNA after the 5-minute incubation period. Single-stranded DNA was immediately diluted with 990 µl of chilled Hybridization Buffer (HT1), resulting in a 20pM library. A sample sheet consisting of a comma-separated values (CSV) file with all the necessary information about the library, and experiment including the cartridge barcode number was created on the MiSeq platform prior to sequencing. The flow cell

was cleaned with lint free tissue and 80% freshly prepared ethanol and was loaded to MiSeq alongside required buffer. 600 µl of the 20 pM library was loaded directly unto the specified reagent cartridge and sequenced on MiSeq for 2 x 250 bp paired-end. FASTQ files were generated after the run and automatically available for download on Illumina's Basespace cloud service.

2.5.3.2 Illumina sequencing using Nextera Flex kit

In this study, several steps were introduced for internal validation during sequencing. A separate ΦX174 DNA deep-sequencing-run was carried out using the Nextera Flex kit (20018704, Illumina and Nextera DNA CD Indexes (20018704, 24 indices, 96 samples, Ilumina). ΦX174 DNA were re-extracted as described in section 2.5.2.1. DNA libraries were prepared using Nextera Flex DNA kit following manufacturer's instructions. Indexed and paired-end libraries were also prepared in a hard-shell skirted PCR plate for the chosen time series (6 samples), but DNA spike-ins were not used as described in section 2.5.3.1. The Nextera Flex kit normalisation of genomic DNA step is compatible with varying DNA inputs 1 – 500 ng. For this experiment, an average input of ~300 ng was used.

2.5.3.2.1 Tagmentation of normalised genomic DNA

The ΦX174 genomic DNA was fragmented and then tagged with adapter sequences utilising the bead-linked transposome (BLT). The BLT was mixed thoroughly with a vortex device to re-suspend. Tagmentation master mix (TMM) was prepared by combining 11 µl of the resuspended BLT with 11 µl of Tagmentation buffer 1 (TB1) and mixed vigorously. In a hard-shell PCR plate, 30 µl of genomic DNA was added, followed by 20 µl of TMM and mixed by pipetting. The plate was centrifuged at 280g at 20°C for 1 minute, sealed and run on a programmed thermal cycler (preheated lid set to 100°C, 55°C for 15 minutes, held at 10°C).

2.5.3.2.2 Tagmentation clean-up

The adapter-tagged genomic DNA was cleaned up prior to PCR amplification. 10 µl of Tagment stop buffer (TSB) was added to the tagmentation reaction mix, resuspended by slowly pipetting, and run on a programmed thermal cycler and run at 37^{o}C for 15 minutes (with the heat-lid set at 100^{o}C) and held at 10^{o}C. The library was immediately placed on a magnetic rack for aggregation of the paramagnetic beads, supernatant discarded and washed twice using 100 µl Tagment wash buffer (TWB). The TWB was left in the wells after the second wash to prevent over drying of the beads.

2.5.3.2.3 PCR amplification of tagmented DNA

The tagmented genomic DNA was amplified using a PCR-based program. PCR steps incorporated Nextera Indexes, Index 1 (i7) adapters and Index 2 (i5) adapters and sequences necessary for cluster formation on MiSeq flow cell. PCR master mix (PCR-MM) was prepared for each reaction by mixing 22 µl Enhanced PCR mix (EPM) and 22 µl Nuclease-free water. TWB supernatant (section 2.5.3.2.2) was removed and discarded. The following consumables were added immediately to the PCR plate containing the beads: 40 µl of PCR-MM, 5 µl of each index 2 (i5) adapter added horizontally to each well, 5 µl of each index 1 (i7) added vertically to each well, thereby allowing for multiplexed sequencing, pipetted to mix and centrifuged at 280g for 1 minute. The plate was firmly sealed and placed in pre-programmed thermocycler to run following these conditions: 68^{o}C for 3 minutes; 98^{o}C for 3 minutes; 5 cycles of 98^{o}C for 45 seconds, 62^{o}C for 30 seconds, 68^{o}C for 2 minutes; 68^{o}C for 1 minutes and held at 10^{o}C.

2.5.3.2.4 PCR library clean-up

Index-amplified products were purified through a bead purification procedure. Sample purification beads (SPB) were homogenised and brought to room temperature. SPB master mix (SPB-MM) was prepared for each reaction by

mixing 45 µl SPB with 40 Nuclease-free water and thoroughly mixed. 85 µl of SPB-MM was added to each PCR product, mixed by pipetting for complete homogenisation, incubated at room temperature for 5 minutes and allowed to stand on a magnetic rack until the supernatant has cleared. 125 µl of the supernatant was added to 15 µl of SPM (non-diluted) in a fresh 96-well PCR plate, mixed by pipetting, and incubated at room temperature for 5 minutes. The supernatant was removed and discarded. The beads were washed twice on the magnetic rack by adding freshly prepared 80% ethanol (10644795, Fisher Scientific) and air-dried for approximately 7 minutes. The washed beads were resuspended in 32 µl Resuspension Buffer (RBS) and 30 µl of the supernatant was transferred into a new 96-well PCR plate. The libraries were normalised, pooled and sequenced as described in sections 2.5.3.1.5 and 2.5.3.1.6.

2.5.4 Whole-genome sequencing for *S.* Typhimurium

2.5.4.1 Nanopore sequencing

The extracted *S.* Typhimurium gDNA (section 2.5.2.2) was sequenced in-house using both Illumina and Oxford Nanopore sequencing. In the latter protocol, a MinION (R9.4.1, Oxford Nanopore™ Technologies, Oxford, UK) portable sequencing device was utilised. Sequencing was performed using MinION MK1, FLO-MIN 106 flow cells, 1D PCR-based barcoding genomic DNA preparation with SQK-LSK 108 ligation kit according to manufacturer's instructions but with some modifications. Briefly, the starting mass of extracted bacterial gDNA was 1 µg, quantified with the Qubit version 3 fluorometer and adjusted to 46 µl with nuclease-free water. Prior to library preparation, the MinION and a flow cell were set up while connected to host computer, to check the number of active pores available for sequencing. 46 µl gDNA was fragmented using a Covaris g-TUBE (15240099, Fisher Scientific) by centrifuging twice at 5,000 rpm for 30 seconds. Fragmented DNA end-repair and dA-tailing was performed in a single PCR, incubated at 20°C for 5 minutes and 65°C for 5 minutes. The amplicon product was

cleaned up using 60 µl AMPure beads (A63881, Beckman Coulter) and freshly prepared 70% ethanol twice on a magnetic rack. The bead pellet was re-suspended with 31 µl nuclease-free water, incubated for 2 minutes, supernatant collected into a fresh 1.5 mL microtube and it was quantified with the Qubit fluorometer. A barcode adapter was ligated unto the end-repair product at room temperature, cleaned and size-selected using 40 µl AMPure beads (A63881, Beckman Coulter). The bead pellet was re-suspended in 25 µl nuclease-free water, incubated at room temperature for 2 minutes, then the supernatant was collected and diluted to 10 ng / µl with nuclease-free water. A barcoded PCR was set up on a thermal cycler using the following conditions: 95^{o}C for 3 minutes; 12 cycles of 95^{o}C for 15 seconds, 62^{o}C for 15 seconds, 65^{o}C for 8 minutes; 65^{o}C for 8 minutes and held at 4^{o}C. Barcoded DNA product was purified with 50 µl AMPure beads (A63881, Beckman Coulter), re-suspended in 15 µl nuclease-free water. Purified barcoded PCR amplicon was quantified with Qubit fluorometer and diluted with nuclease-free water to 1 µg in a total volume of 45 µl.

The library was adapted for Nanopore sequencing by performing end-repair and dA-tailing through a PCR-mediated reaction with the library incubated for 5 minutes at 20^{o}C and 5 minutes at 65^{o}C in a thermal cycler. The product was purified using 60 µl AMPure beads (A63881, Beckman Coulter), the bead pellet re-suspended in 31 µl of nuclease-free water and supernatant collected for the next step. Next the end-prepped DNA adapter was ligated by incubating at room temperature for 10 minutes. Finally, the product was cleaned using 50 µl AMPure beads (A63881, Beckman Coulter), resuspended in 15 µl Elution buffer (EB) incubated at room temperature for 10 minutes and supernatant collected into a fresh 1.5 mL microtube, ready for sequencing. The flow cell was primed with a freshly prepared priming mix containing 576 µl RBF (Resuspension Buffer) and 624 µl nuclease-free water. Sample was prepared for loading by mixing 35 µl RBF, 25.5 µl LLB (Loading Beads), 2.5 µl nuclease-free water and 12 µl DNA library. 75 µl of final sample was loaded to the flow cell through the SpotON sample port

drop by drop and sequenced. Files were generated and available for download and further analysis on MinKNOW experiment page.

2.5.5 Ampliconic sequencing

Cross-validation of some allelic variant sites detected during Illumina sequencing and for SDM was performed using Sanger sequencing. SDM cross-validation required amplification of phage template in a PCR-mediated reaction. Two primers (table 2.2; sequences provided by Holly Wichman) were used for amplifying ϕX174 DNA of an isolated plaque. Amplicons A and B, consisted of bases 1 – 2953 and 2605 – 379, respectively. 50 µl PCR contained: 0.5 µl Q5® High-Fidelity DNA Polymerase (E0554S, New England BioLabs, Paisley, UK), 10 µl 5X Q5 Reaction Buffer, 1 µl 10mM dNTPs, 2.5 µl of each primer at 10 µM, 1 µl of 1.0 ng / µl phage sample and 32.5 µl of nuclease-free water was added to a PCR tube. The amplification of the reaction was carried out in a thermal cycler with the following conditions: 98^{o}C for 30 seconds; 30 cycles of 10 seconds at 98^{o}C, 30 seconds at 60^{o}C, 2 minutes 30 seconds at 72^{o}C; 72^{o}C for 2 minutes and held at 4^{o}C. PCR products as well as 1kb DNA Ladders (G5711, Promega, Southampton, UK) were visualised via 1 % agarose gel electrophoresis (Section 2.3.9) to confirm the presence a single dsDNA band of the expected size. After, PCR products were purified using NucleoSpin Gel and PCR Clean-up Kit (12303368, Macherey-Nagel, Fisher Scientific), as per manufacturer's instructions. The concentration of DNA was determined using Qubit 3.0 Fluorometer and diluted to 10 ng / µl. The primers needed for sequencing were selected appropriately according to table 2.2 and diluted to 3.2 pmol / µl (as required by the sequencing company: Source Bioscience, Nottingham, UK) with nuclease-free water. Allelic variant site cross-validation from Illumina sequencing did not require PCR amplification, since these already contained high concentration dsDNA, and these were instead diluted to 100 ng / µl for sequencing. DNA samples and primers were sent to Source Biosciences for Sanger sequencing. The electropherogram results were

analysed by using 4Peaks (Griekspoor and Groothuis, 1994) and SnapGene (v.4.3.2) viewer on MacOS.

2.6 Sequencing work-flow methods

Phage dsDNA was extracted from samples stored at -80^{o}C as described in section 2.5.2. DNA extraction was performed for samples as shown in figure 2.1, sequenced using the methods described in section 2.5.3, and raw sequence data containing reads were produced. Raw sequence reads from Illumina sequencing were separated into files according to indexed samples in a process termed demultiplexing. The files produced were in the FASTQ format and were downloaded from Illumina's Basespace server. Two FASTQ files were generated, one for each end/strand, for each sample, since the sequencing run was paired end.

2.6.1 Reference genomes

The two reference genomes used in this experiment were obtained in different ways. The *E. coli* C reference genome was sequenced by Anton Nekrutenko's group (at the Pennsylvania State University, U.S.A) with input data from Oxford Nanopore and Illumina platforms. The final contigs obtained are two, one representing the entire *E.coli* C genome and the other the ΦX174 genome (putatively derived from the Illumina control track, which is frequently used as an internal control in Illumina sequencing). The ΦX174 genome was removed, leaving one contig used as the *E. coli* C reference, termed "C_reference".

2.6.1.1 *S.* Typhimurium reference genome

The *S.* Typhimurium whole genome was sequenced in-house using both the Illumina's MiSeq platform (with the Nextera XT kit and a V3 cartridge) and Oxford Nanopore's MinION as described in section 2.5.4 and is herein referred to as the "ST_assembly". The ST_assembly was constructed as

follows: after Nanopore sequencing, bases from data obtained were called using Albacore (v.2.1.10), with output in the FASTQ file format. Albacore was invoked using the command-line on a Windows 10 operating system. The command used was:

```
read_fast5_basecaller.py --flowcell FLO-MIN106 \
--kit SQK-LSK108 --output_format fast5,fastq \
--input <files_containing_reads> \
--recursive --save_path <pathway_to_data> \
--worker_threads <user_threads>
```

For quality control, adapters introduced during library preparation were removed from the output FASTQ and demultiplexed barcoded reads using Porechop (v.0.2.3). Porechop is primarily utilised for removing adapters from Nanopore reads by performing alignments to find adapters at the ends and in the middle of reads, bisecting reads when the latter are identified (https://github.com/rrwick/Porechop).

```
porechop -i  <input_reads.fastq.gz> \
-b <output_reads.fastq.gz>
```

Following this, the hybrid read sets generated by Illumina and Nanopore were assembled using Unicycler (v.0.3.0b). The Unicycler assembly workflow performs SPAdes assembly of the reads from Illumina, then scaffolds the assembly graph with long reads. It has an in-built polisher called Pilon that polishes final assembly with Illumina to minimise the rate of base-level errors (Wick *et al.*, 2017). The quality of assemblies was assessed visually, and contigs were examined to determine whether any sequences similar to the ΦX174 genome were present, using Bandage (v.0.8.1; Wick *et al.*, 2015). The command used was:

```
unicycler -1 <Illumina_reads_R1.fastq.gz> \
 -2 <Illumina_reads_R2.fastq.gz> \
-l  <Nanopore_reads.fastq .gz> -o <output>
```

To assess the quality of ST_assembly, it was compared with the NCBI reference genome. The *S.* Typhimurium LT2 chromosome (accession NC_003197.2) and its associated plasmid (NC_003277.2), were combined and are herein referred to as the "ST_reference".

The ST_reference chromosome is 4,956,769 bp in size. This was very similar to the sum of contig lengths in the ST_assembly (4,756,781 bp), including the presence of the plasmid. A program called Quality Assessment Tool for Genome Assemblies (QUAST) can be used to evaluate genome assemblies. QUAST was used to assess the similarities between ST_assembly and ST_reference. QUAST checks the quality of an assembly by calculating the length distribution of aligned blocks of contigs with respect to a reference, referred to as the NGA50, with NGA50 representing the length of derived blocks such that 50% of their total length is contained in blocks of at least that size (Mikheenko *et al*., 2018). The statistics obtained showed that the ST_assembly is 98 % identical to the ST_reference, with an NGA50 of 719,403 bp and 20.35 mismatches per 100 kbp. It appears that 12 contigs were misassembled during the Unicycler workflow. The command used for assembly comparison was:

```
quast <input_assembled.fasta> -t <user threads> -r
<input_reference.fasta> \
-o <output_file>
```

The ST_assembly contained errors, but showed a high level of similarity with the ST_reference (figure 2.4). In order to use the most accurate and contiguous/complete target for read mapping ST_reference was employed throughout the deep sequencing analysis. The absence of breaks in sequence was particularly important because the main function of the reference genome in the deep sequencing analysis (section 2.6.2) was removal of reads with ambiguous mapping.

2.6.1.2 Phage ΦX174 reference genome

The reference genomes, ΦX174 (accession AF176034) and pUC18 (L09136) were utilised during sequence data analysis. To ensure reads were captured throughout the genome and origin, 'resected' versions of ΦX174 and pUC18 (figures 2.2, 2.3) were derived with a modified coordinate system. Mapping algorithms typically treat reference genomes as linear, a problem for mapping to circular genomes. If treated as linear, there will be a break at the origin of circular genomes, meaning mapping to the reference may be inaccurate or depleted at this point. For large genomes the effect may be negligible, whereas for small genomes like ΦX174 (5,386bp) or pUC18 (2,686bp) the affected regions make up a larger proportion of the genome. The resected genomes were derived by creating FASTA files in which the second half of the genome was relocated upstream of the first half. Both original and resected genomes were used during mapping, with results merged later in the analysis (figures 2.2, 2.3).

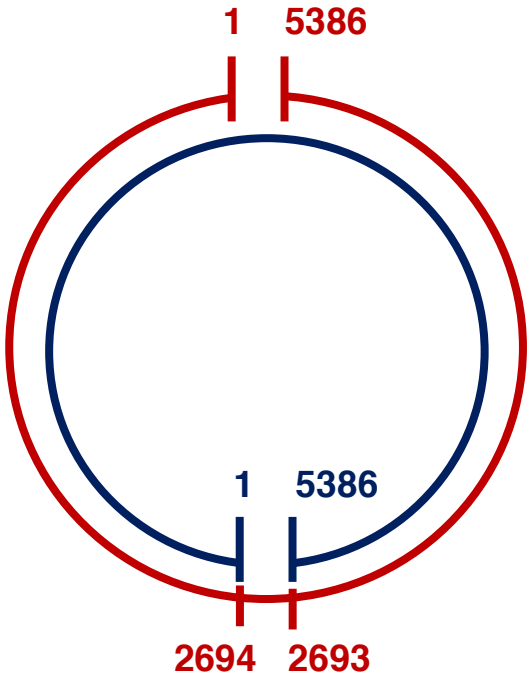


Figure 2.2: ΦX174 genome showing original and resected coordinate systems. The outer circle (in red) shows the original coordinates (starting at 1 and ending at 5386); the alternative coordinate system (inner circle in blue) begins at position 2694 (in the original coordinate system) and ends at position 2693 (in the original coordinate system).

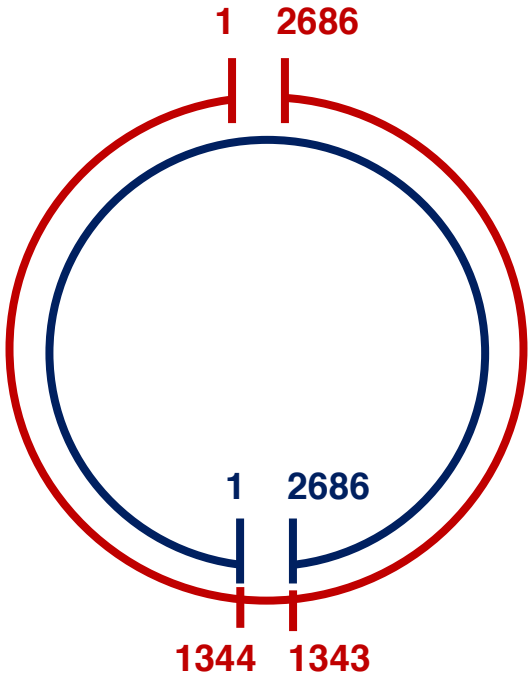


Figure 2.3: pUC18 genome showing original and resected coordinate systems. The outer circle (in red) shows the original coordinates (starting at 1 and ending at 2686); the alternative coordinate system (inner circle in blue) begins at position 1344 (in the original coordinate system) and ends at position 1343 (in the original coordinate system).

2.6.2 Sequencing analysis

The sequencing analysis was carried out using bash scripts (appendix A) that looped through the quality control (2.6.2.1) and mapping (2.6.2.2) steps for each sample. (Some loops were also employed in SNP calling). This was carried out to prevent repetition of commands. Detailed descriptions of each step follow in sections 2.6.2.1 – 2.6.2.4.

2.6.2.1 Quality control

During sequencing for short fragments, read 1 and read 2 primers from library preparation kits may be sequenced at the 3' end of reads derived from short fragments. For this reason, adapter sequences were removed by Cutadapt (v.1.14). The sequence **CTGTCTCTTATA**, the reverse complement of the last 12 nucleotides of the Nextera (Flex and XT) reads 1 and 2 transposase adapter was supplied for trimming. Cutadapt is a command-line tool that searches for adapters in list of high-throughput sequences and trims them from sequence reads in FASTQ files (Martin, 2011). In Cutadapt, quality scores corresponding to ASCII encoding were set with the argument `--quality-base=33` (this is also the default). Low quality nucleotide bases were trimmed from both 5' and 3' ends of each read in a pair prior to adapter removal; 5'-end trimming and 3'-end trimming were both supplied with a quality score cut-off of 30 using the argument `--quality-cutoff 30,30`. For trimming the arguments `-a` and `-A` were used with `-a` trimming adapter sequences from the 3' end in forward reads, and `-A` trimming 3' adapter from the reverse read in a pair. An overlap of 3 nucleotide sequences was chosen (`--overlap=3`) and an error rate of 20% (`--error-rate=0.2`), allowing a maximum of 2 nucleotide sequences of the adapter sequence to be present before trimming. The arguments `--trim-n` trimmed indeterminate (N) nucleotides on ends of the reads, `--pair-filter=any` discard or redirect pair reads if one of the reads (either forward or reverse) fulfils the filtering criterion. Output FASTQ files were generated with `--output` for forward reads and `--paired-output` for reverse reads. The numbers of available

processors was specified with the `--cores` argument. The command used was:

```
cutadapt --quality-base=33 --quality-cutoff 30,30 \
-a <adapter> -A <adapter> --error-rate=0.2 --overlap=3 \
--trim-n --pair-filter=any --minimum-length=20 \
--cores <threads> \
--output <forward_trimmed_R1.fastq.gz> \
--paired-output <reverse_trimmed_R2.fastq.gz>
```

The quality metrics of FASTQ files generated were tracked with FastQC (v.0.11.8), a tool for quality control checks on sequence data, giving an overview of data quality and comparison between trimmed and untrimmed prior to further analysis (Andrews, 2010). Moreover, a modular tool called MultiQC (v.1.7) was utilised, which summarises results from bioinformatics analyses across multiple samples and tools including FastQC into a single informative HTML report (Ewels *et al.*, 2016).

2.6.2.2 Mapping

The DNA reads after trimming were mapped to indexed reference genomes with BWA mem (Li, 2013; v.0.7.17) in two stages using the default settings, and enabling multi-threading. The command used was:

```
bwa mem -t <users_threads> <reference_genome> \
<forward_trimmed_R1.fastq.gz> <reverse_trimmed_R2.fastq.gz> \
> -o <output.sam>
```

In stage one, subtraction of the plasmid spike-in-genome was carried out to ensure pass-through read pairs generated do not map to any location of the pUC18 spike-in-genome (and to identify the presence of expected or contaminating pUC18 DNA in library-prepared samples). The trimmed FASTQ files were mapped to the pUC18 with an output in the SAM file format. Mapped reads were set aside for further analysis of potential cross-

contamination, while unmapped reads were indexed and sorted using the SAMtools (Li *et al*., 2009; v.1.9) view and sort programs, producing files in the compressed binary BAM format as output. For stage one, flag `-F 4` was utilised to select for only mapped reads, while flag `-f 4`, for unmapped reads only. The commands used were:

```
samtools view -O BAM -F 4 -o <input.bam> <output.sam> \
samtools view -O SAM -h -f 4 <input.sam> \
| samtools sort -O BAM -n -o <output.bam> -
```

BAM files generated from the unmapped reads were converted to FASTQ files using the BEDtools (v.2.27.1) bamtofastq program because BWA mem requires FASTQ inputs. This step also resulted in the discarding of unpaired singleton reads. The command used was:

```
bedtools bamtofastq -i <input.bam> \
-fq <output_R1.fastq> -fq2 <output_R2.fastq>
```

For maximal accuracy in subtracting the spike-in-genome, the unmapped reads produced using BEDtools were also mapped to the *resected* pUC18 genome with BWA mem. This was followed by the same procedure of selecting unmapped reads utilising the `-f 4` flag in SAMtools view and converting back to FASTQ files with BEDtools bamtofastq. Separately cutadapt-trimmed reads were mapped directly to the resected pUC18 genome using BWA mem and mapped reads converted to BAM files with SAMtools view for analysis of spike-in genome coverage and contamination.

In stage two, unmapped reads generated from the first stage (filtered against pUC18 and resected pUC18; section 2.6.1.2, figure 2.3) were mapped to reference genome of either C_reference or the ST_reference (depending on the last host encountered by the population from which the sample was drawn) as well as the phage ΦX174 genome. In choosing a reference genome, the `awk` command-line tool was used to select the appropriate

reference genome (for each sample) from a comma-separated values (CSV) file (prepared manually) in each iteration of the loop. The unmapped reads were mapped to the corresponding reference genome using BWA mem using the same command line as in stage one. Next, mapped reads with a mapping quality less than 20 were excluded (argument `-q 20`), prior to sorting by reference position, and indexing the output BAM files. Post-mapping steps were carried out using the SAMtools view, sort and index commands:

```
samtools view -bS -F 4 -q 20 <input.sam> | \
samtools sort -@ 3 -o <output.bam> \
samtools index <input.bam>
```

In the next step mapped reads were filtered so that all read pairs mapped exclusively to ΦX174. First, SAMtools view was used to select based on the accession ΦX174 number. This process is imperfect and leaves behind contig headers as well as cross-contig pairs (that is, pairs of reads in which each member mapped to different contigs). Second, the `awk` command line tool was used to remove cross-contig pairs. Second, the `sed` command-line tool was used to remove all contig headers except the ΦX174 genome. These filtering steps were followed by indexing with SAMtools index:

```
samtools view -O SAM -h <input.bam> AF176034.1 | \
awk '$7 == "=" || $1 ~ /^@/'| sed '/^@SQ/{/reference/!d;}'|\
samtools view -bS -o <output.bam> \
samtools index <input.bam>
```

This procedure was repeated for the resected ΦX174 genome. Indexed reads mapping to the ΦX174 genome or its resected genome in BAM file format were used for the next analysis step.

2.6.2.3 Variant calling

Freebayes (v.1.2.0) was used to call variants. Generally, Freebayes calls variants from aligned short-read data, utilising different file formats as inputs including VCF and BAM along with a reference genome, and outputs VCF files (Garrison and Martin, 2012).

Single Nucleotide Polymorphism (SNPs) were called using the following arguments: `--pooled-continuous`, outputting all alleles that pass filters (without specifying a ploidy level); `--min-mapping-quality`, excluding alignments from analysis if they possess mapping qualities less than specified (Q20 in this case); `--min-base-quality`, excluding all alleles if they have supporting base quality less than specified (Q30 in this case); `--min-alternate-fraction`, specifying the least fraction of observations (0.01 in this case) supporting an alternate allele within an individual of a population in order to assess the position; `--min-alternate-count` the smallest number of reads (1 in this case) that can support an alternate allele in a single individual so as to evaluate the position; `--bam-list`, pointing to a text file specifying the BAM files to be analysed; `--vcf`, outputting results in VCF file format:

```
freebayes --fasta-reference <ref_file_name> \
--pooled-continuous --min-alternate-fraction 0.01 \
--min-alternate-count 1 --min-mapping-quality 20 \
--min-base-quality 30 --bam-list <grouped_file.txt> \
--vcf <output.vcf>
```

Allelic variant calling was performed in two stages. In stage one, BAM files were organised into two groups according to each sample's last-encountered bacterial host; ancestral sample "A" was added with the *E. coli* C group. Lists of BAM files were created using regular expression matching (via the unix find command) with a text file as an output specifying the BAM files in each group. Freebayes was invoked for each group independently (using the `--`

`bam-list` argument) in this first stage. Additional loop iterations were also used to analyse the resected ΦX174 genome, with the corresponding references analysed (for each host). These steps lead to the production of four files:

1. C_1STCALL_phix_1.vcf
2. C_1STCALL_phix_2.vcf
3. S_1STCALL_phix_1.vcf
4. S_1STCALL_phix_2.vcf

where C and S refer to the last-encountered host (*E. coli* C and *Salmonella* Typhimurium, respectively) and 1 and 2 refer to the original and resected ΦX174 mappings.

For each mapping, results corresponding to *E. coli* C and *S.* Typhimurium were merged using bcftools (v.1.9) merge to give two files:

1. ignore_1.vcf (derived from numbers 1 and 3 above), and
2. ignore_2.vcf (derived from numbers 2 and 4 above).

In the second stage, Freebayes was again used to call variants at sites detected in any sample in either host (but still separated into two loops corresponding to conventional versus resected mappings). To achieve this an additional argument was introduced, `--variant-input`, to input the union set of variants in VCF file format. This procedure requires the listed sites to be examined in all samples. The final output from variants calling were designated as:

1. ALL_UNIONCALL_1 (for the conventional mapping)
2. ALL_UNIONCALL_2 (for the resected mapping).

```
freebayes --fasta-reference <ref_file_name> \
```

```
--pooled-continuous --min-alternate-fraction 0.01 \
--min-alternate-count 1 --min-mapping-quality 20 \
--min-base-quality 30 --variant-input <input.vcf> \
--bam-list <grouped.txt> \
--vcf <output.vcf>
```

2.6.2.4 Variant annotation

The VCF outputs generated from ΦX174-resected inputs had the coordinate system reverted to the original reference genome coordinates by employing a custom python script (appendix A.6). Following this, another python script (appendix A.7) was utilised to produce annotated tabular files in a readable, tab-separated values (TSV) format. The inputs for this script were a VCF file, the ΦX174 reference genome and a table of ΦX174 gene coordinates. The script identifies codon positions and characterises amino acid changes as synonymous, radical or conservative, taking into account the fact that some genes span the origin and detecting changes in all (including alternative) reading frames.

The results from this annotation account for amino acid changes, between-type amino acid changes, alternative reading frames, and alternative transcription start sites. Two TSV files were generated from each stage of variant calling (i.e., from both UNIONCALL inputs), one for the primary ΦX174 reference genome and the other for *corrected* resected coordinate system.

Annotation from these two files was merged using a custom R script, MergeAnnot.R (making use of tidyverse v.1.2.1) to parse the TSV files, which selected columns from each sample according to the coverage, alternative allele depth and alternative allele frequency, generating a table, and assigning values to that table, for each position and sample, from the dataset with the highest coverage. This step ensured maximal coverage for sites

close to the origin (which was the reason for using resected genome; figure 5.5 b).

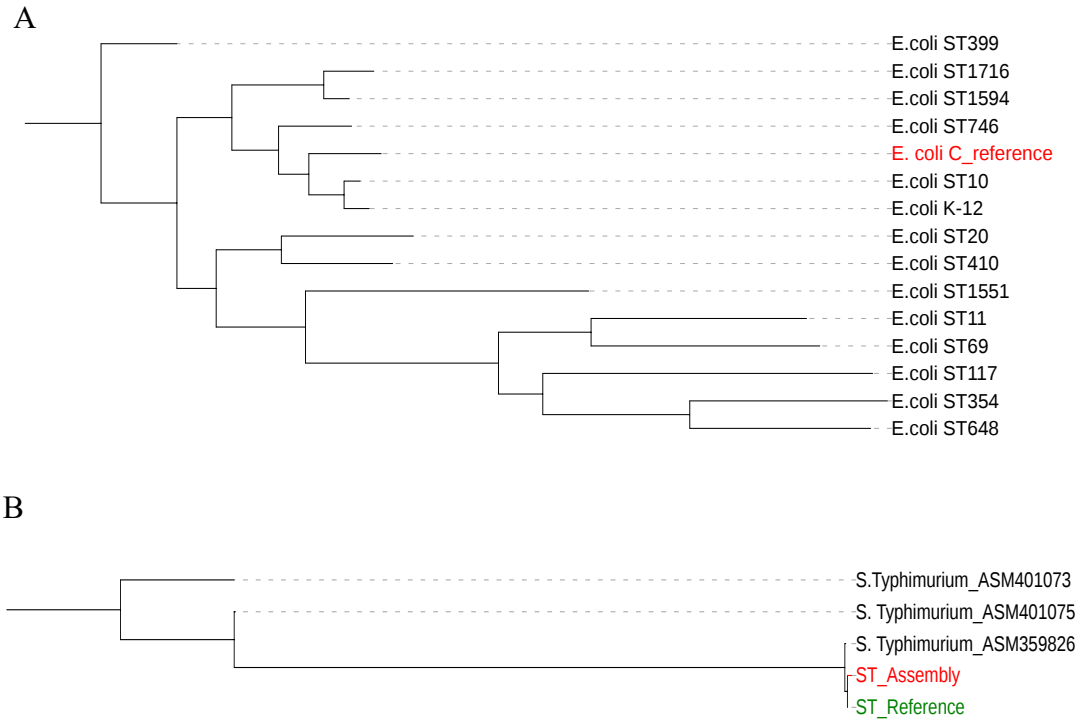


Figure 2.4: Phylogenetic tree of hosts used in this study, constructed using parsnp (v.1.2) and rendered using iTOL (https://itol.embl.de). A – *E. coli* strains from different environmental sources (Seecharran, 2018), indicating the sub-types. *E. coli* C_reference (highlighted in red) is the reference C_reference used in this study (section 2.1.1). B – ST_Assembly (highlighted in red) is the in-house sequenced strain (section 2.5.4), ST_Reference (highlighted in green) is the NCBI-obtained reference used for the analysis (accession NC_003197.2; section 2.6.1.1), other S. Typhimurium strains (highlighted in black) were obtained from NCBI (accession numbers NZ_CP034819.1, NZ_CP034831.1, NZ_PXVG01000010.1 respectively).

Chapter Three: Chemostat development and microorganism strains

3.1 Introduction

Microbial growth is regulated in response to a complex network of signals from the environment and cell genetics (Ziv *et al.*, 2013). As growth occurs, cellular processes such as metabolism, synthesis of macromolecules, cell division are coordinated. Understanding how cell growth impacts cellular processes is fundamental in microbiological, evolutionary, biotechnological, genetic and ecological studies (Bull, 2010). For several applications, evaluating and controlling the dynamics of cell growth as the cell population continuously alters the environment in which it proliferates is important to the experimentalist. A solution is the cultivation of microbial cells in a chemostat: a method that allows regulation and control of cells growth rate in a continuous, defined and invariant environments.

3.1.1 Chemostat culturing system

A chemostat system was first described in 1950s by Monod (1950) and by Novick and Szilard (1950). In a chemostat system, fresh medium is continuously added to the culture while, at the same rate, culture liquid containing microbes, metabolites and left-over nutrients are removed, therefore the culture volume is kept constant. At this point, cells are grown in a physiological state at a specific growth rate known as steady state. In a steady state, the population is presumed to reach a steady density and is maintained under constant environmental conditions in an exponential phase. For experimental evolution studies, achieving a static environment is desirable to accurately characterise biological systems in order to lower noise in quantitative phenotyping and evaluate changes in genotypes.

3.1.2 Host strains *E. coli* and *S.* Typhimurium

E. coli and *S.* Typhimurium are both gram-negative, facultative anaerobic bacteria in the family of *Enterobacteriaceae*. *E. coli* belongs to the genus *Escherichia* that is commonly found in animals' and humans' intestines and the environment after being expelled as faecal matter (Tenaillon *et al.*, 2010). *S.* Typhimurium is rod-shaped, predominantly motile with peritrichous

flagella, mainly pathogenic and widely distributed in nature (Agbaje *et al*., 2011). Both host strains are common microbiology model organisms that have been well adapted to laboratory environments. The two hosts were used in this experiment for chemostat culture as different environments for propagation ΦX174 phage.

3.2 Aims and Objectives

In experimental evolution, adaptation of microorganisms can be monitored in a laboratory-controlled conditions by the experimentalist. Different experimental designs may be used including the standard serial transfer setup and continuous culturing in experimental evolution studies. Generally, chemostats operate by continuously adding fresh media to microbial culture at a fixed dilution rate while the vessel volume remains constant. For phages, an additional vessel is require where naïve bacterial cells are continuously supply at the same rate cells and phages are remove as waste. In this study, different bespoke chemostats were designed with the main aim of evolving ΦX174 on different hosts until a presumably steady state was achieved, cells were properly mixed and biofilm formation minimised. The growth rate of ΦX174 on different hosts were measured using qPCR prior to host switching experiment in chapter 4.

The main objectives of this chapter were:

- To design low-cost chemostats.

- To build a two-chamber chemostat with uni-directional flow for the continuous culture of ΦX174 in standard conditions.

- To achieve continuous culture at steady state with an unchanging environment.

3.3 Results and Discussion

3.3.1 Chemostat development

A chemostat is a continuous flow system for growing microorganisms. It is typically made up of two parts: a nutrient reservoir and a growth chamber for bacterial growth. Through an inflow, fresh culture is continuously added to the LB medium in the growth chamber and an equal volume of spent medium containing cells is removed as waste from an outflow of the culture vessel (figures 1.4, 1.5; Gresham and Dunham, 2014). For our experiments, a population of ΦX174 was evolved in a chemostat. To achieve this, a specially designed chemostat was set up to accommodate an additional growth chamber. For convenience, the first growth chamber for bacteria (bacterial chamber in figures 3.3, 3.4) will be termed '*lagoon*' (adapted from Klavins lab, http://klavinslab.org/) and the ΦX174 chamber (phage chamber in figures 3.3, 3.4) is the '*swamp*'. Excess volume from the *lagoon* is drained continuously into the *swamp*. In this way, ΦX174 was supplied with fresh naïve bacteria (meaning they have not previously encountered the phage). In the meantime, excess suspension from the *swamp* was transferred to a waste container (figures 3.3, 3.4). All parts used for the development of the chemostats were ordered from various companies as summarized in table 3.1. The medium used for chemostat experiment was LB from the same batch while a different batch was utilised in site directed mutagenesis experiment (chapter 6). The designing of chemostats used in this study posed some unique challenges prior to the final development of the main chemostat utilised in host switching experiment. The different phases of chemostat development were divided into four main configurations with improvements due to the presence of biofilm or to avoid cross-contamination to which we now turn.

Materials name	Company	Catalog Number	Comments
Submersible magnetic stirrers – MIXdrive single	Camlab, Cambridge, UK	1169841	Used in 5L water bath, Apparatus 2 and 3
Submersible stirrers – 6 inductive magnetic stirrers	Camlab, Cambridge, UK	1169843	Used in Apparatus 4
Submersible mix control 40 for 2 MIXdrive stirrers	Camlab, Cambridge, UK	1169851	Mixdrive control used for both singles and 6 inductive in Apparatus 2, 3 and 4
PTFE Magnetic stirrer bars	Camlab, Cambridge, UK	1201475	For proper mixing in Apparatus 2, 3 and 4
PTFE Magnetic bars retriever	Camlab, Cambridge, UK	1139267	For retrieving bars from cultures in Apparat 2, 3 and 4
I150 peristaltic pump, B4R6 channel 6 roller pump head	Ipump ltd, Gloucestershire,UK	I150	Apparatus 1, 2, 3 and 4
Aquarium vacuum pump	Amazon, UK	N/A	Apparatus 1, 2, 3 and 4
Water bath 5 L, JB Nova	Appleton woods, Birmingham, UK	WA0811	Apparatus 3 water bath
Water bath 18 L, JB academy	Camlab, Cambridge, UK	1194457	Used in Apparatus 4

Thermometer	Wilko, UK	N/A	For checking water bath temperature
Autoclavable draining Tray	Cole Parmer, London, UK	WZ-06720-22	Chemostat system was placed in the tray f autoclaving
Duran bottle GL 45 thread, 3.5 L	Camlab, Cambridge, UK	1199183	Reservoir flask in Apparatus II and IV
Duran bottles GL 45 thread, 0.25 L	Camlab, Cambridge, UK		*Lagoon* and *swamp* vessels, Apparatus 1, 3 and 4
Buchner flask Pyrex, 5 L	Camlab, Cambridge, UK	1141612	For waste in Apparatus 4
Erlenmeyer flask, 2 L Pyrex	Camlab, Cambridge, UK	1142326	For waste in Apparatus 3
Screw cap GL45 3 port	Scientific Laboratories Supplies (SLS), Nottingham, UK	1129751	Caps used in Apparatus 1
Port screw cap GL 14	SLS, Nottingham, UK	1129814	For hose connection, used in Apparatus 1
Insert for port screw cap GL 14, 6.0mm	SLS, Nottingham, UK	1129818	Used for proper hose connection in Apparatus 1
Caps for samplings (NPT	Missouri, USA	990165X	Apparatus 3 and 4

port)			
Plug for ½ NPT Port	Missouri, USA	990161X	Apparatus 3 and 4
Plug for ¼ NPT Port	Missouri, USA	990160X	Apparatus 3 and 4
Caps for media and lagoon (red colour - Vaplock)	Missouri, USA	VK-205	Apparatus 2, 3 and 4
Membrane filters attached for air filtering (PTFE 0.20µm)	Appletonwoods, UK	431224	Apparatus 1, 2, 3 and 4
Rubber bungs for waste cap	Nottingham Trent University Stores, UK	N/A	Apparatus 1, 2 and 3
Silicone stoppers	Cole-Parmer, UK	WZ-06298-26	Apparatus 4
Fire lighter tips, metal inserted into rubber bunks to hold tubings	Wilko, UK	N/A	Apparatus 1, 2, 3 and 4
PTFE tubing, 1/16" ID x 1/8" OD	Cole-Parmer, London, UK	WZ06605-27	Apparatus 2, 3 and 4
Silicone tubing, 1/16" ID x 1/8" OD	Cole-Parmer, London, UK	WZ-95802-02	Apparatus 2

Ipump silicone tubing (6 x 2 mm ID x 1mm wall)	ipump	I150	Apparatus 1, 2, 3 and 4
Barbed fittings reducing connections, 3/16" x 1/16" ID, 1/32", 1-1/16", 3/8"	Cole-Parmer, London, UK	WZ-30703-46	For connecting tubings, held in place with cable ties
Cable ties, white 1000/bag	Cole-Parmer, London, UK	WZ-06830-52	Used for holding tubings in-place
Cable tie tensioning tool	Edwardes Bros (Dulwich) ltd, Kent, UK	PARTT1	For securing cable ties
Clamp Screw compressor	Amazon, UK	B01ACP95VC	Screws for clamping tubings prior to autoclave to prevent flow of media between the vessels
Silicone tubing to air pump	Cole-Parmer, UK	WZ-06516-08	Apparatus 1, 2, 3 and 4
Antifoam 204	Sigma-Aldrich ltd, Dorset, UK	A8311-50ML	Aqueous emulsion to prevent formation of foams while mixing

Table 3.1: Lists of materials used in chemostat set-up, company and catalogue number. Parts I – IV which implies stages during chemostat set-up.

3.3.1.1 Apparatus 1

One two-chambered chemostat was set up, apparatus I (figure 3.1). Here, an LB medium vessel (the reservoir) with 2 L capacity was connected via peristaltic pump to two 0.25 L Duran bottles in series: the *lagoon* and *swamp* for bacterial cells and phage growth, respectively. A 1.5 L capacity Erlenmeyer waste flask was connected to the *swamp* vessel. The *lagoon* and *swamp* vessels were kept in a shaking water bath and incubated at 37^{o}C and mixed at 100 rpm to ensure minimal disruption to the chemostat set-up. Silicone tubing was used for all connections from reservoir – lagoon – swamp – waste vessel (inner diameter 2 mm, outer wall 1 mm for vessels and peristaltic pump tubing inner diameter 3.2mm, outer wall thickness 1.6 mm). After two days, it was noticeable that material was adhered to the tubing walls, also clumps of material were visible at the bottom of the culture vessels. Using different polymer-based surface coating material Zhang *et al*. (2006) demonstrated that cells' growth channel surfaces have effects on cell adhesion and non-specific protein adsorption. The modification of inner channel surfaces of polymethyl methacrylate and polydimethylsiloxane with polyethylene glycol and polyacrylic acid resulted in reduced wall growth and adhesion of *E. coli* (captured by optical microscope) for 7 days of cell culture in a chemostat (Zhang *et al*., 2006). However, silicone tubing is unavoidable in the chemostat, its deformability is required for peristaltic pump action to generate peristaltic waves. Silicone tubing connected to the vessels aside from that in contact with the pump can be replaced, these were replaced with polytetrafluoroethylene (PTFE) in apparatus 2. PTFE consists mainly of carbon and fluorine with high thermal stability, low frictional coefficient, used as surface coating materials and are generally non-stick and hydrophobic (Lv *et al*., 2015). PTFE are used in many industries for various applications including pans and cookware coatings, dental fillings (Sattar and Alani, 2017), and catheter coatings to reduce friction allowing other devices to pass through and prevention of bacterial cell adhesion (Cornely *et al*., 2002). The substitution of silicone tubes with PTFE coated PTFE tubing results in less fluid retention, facilitating the flow of culture and minimising the formation of

biofilm.

To further reduce the biofilm formation, it is necessary for the cell culture to be properly mixed by aeration (Miller *et al*., 2013) and/or stirring (Ziv *et al*., 2013) and therefore the latter was introduced in apparatus 2.

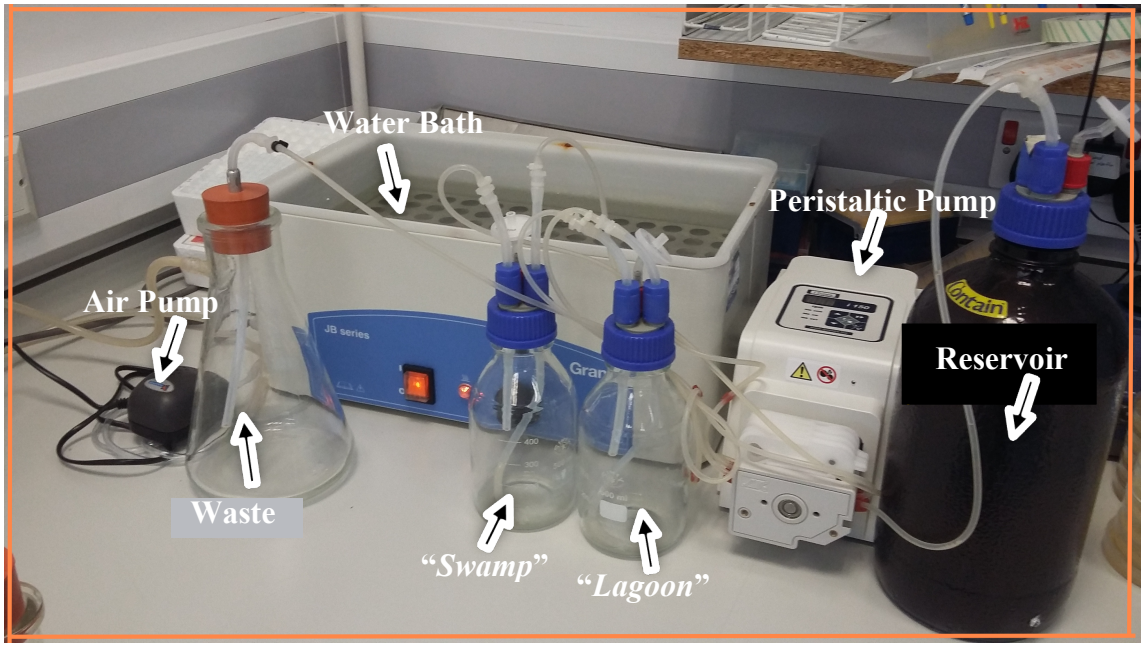


Figure 3.1: The first chemostat designed (apparatus 1) with a static water bath, silicone tubing, connected to 2L media capacity, 250 mL *lagoon* (illustrated in a generic chemostat as bacterial vessel figure 1.5) and the *swamp*, for phage growth (illustrated as phage vessel in figure 1.7) capacities and 1.5L waste bottle.

3.3.1.2 Apparatus 2

In apparatus II, caps were replaced in the swamp vessels for better sampling and to minimise contamination. The new caps were 6-ported bottle caps in which the topmost (screw-thread access) port was used for sample collection. A drip counter was introduced to help monitor the consistency of flow rate via counting the number of drops of media falling from the reservoir into the lagoon in a fixed interval. This was checked at regular intervals, ensuring the flow does not change with time. In addition, the fluid levels in the lagoon and swamp were monitored to ensure that a steady state was maintained. All electrical appliances were placed on the shelf, far from the water bath for safety. Meanwhile, a tray was used to capture any potential leakages from the old water bath. As indicated in section 3.3.1.1, stirring and PTFE tubing were deployed in apparatus 2. Stirring was achieved via submersible stirrers within the waterbath driving stirrer-bars in the lagoon and swamp (figure 3.2; table 3.1).

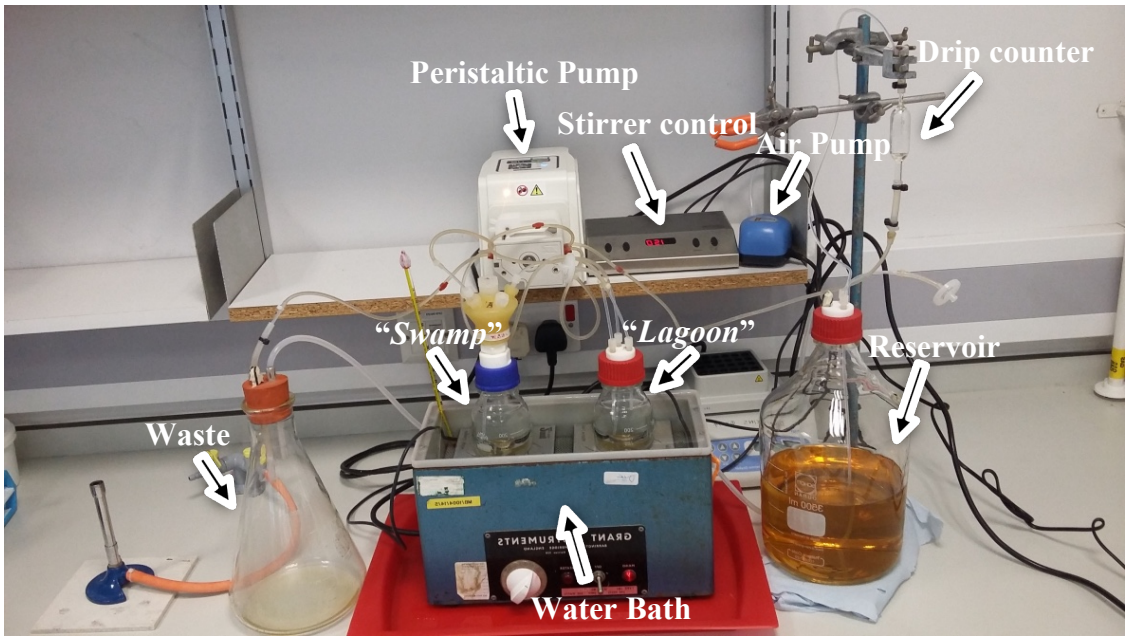


Figure 3.2: The second chemostat designed (apparatus 2) with submersible water resistance stirrers, screw cap for sampling, PTFE tubing (except the tubes connected to peristaltic pump), drip counter, tray for water leakage, 3L media capacity, 250 mL *lagoon* (illustrated in a generic chemostat as bacterial vessel figure 1.5) and the *swamp*, for phage growth (illustrated as phage vessel in figure 1.7) capacities and 1.5L waste bottle.

3.3.1.3 Apparatus 3

Apparatus 3 followed the same design considerations described in apparatus 2. The only difference between these was the replacement of water bath with a new static water bath (5L capacity; table 3.1). With the water bath replacement, no leakage of water was observed ensuring electrical safety (figure 3.3).

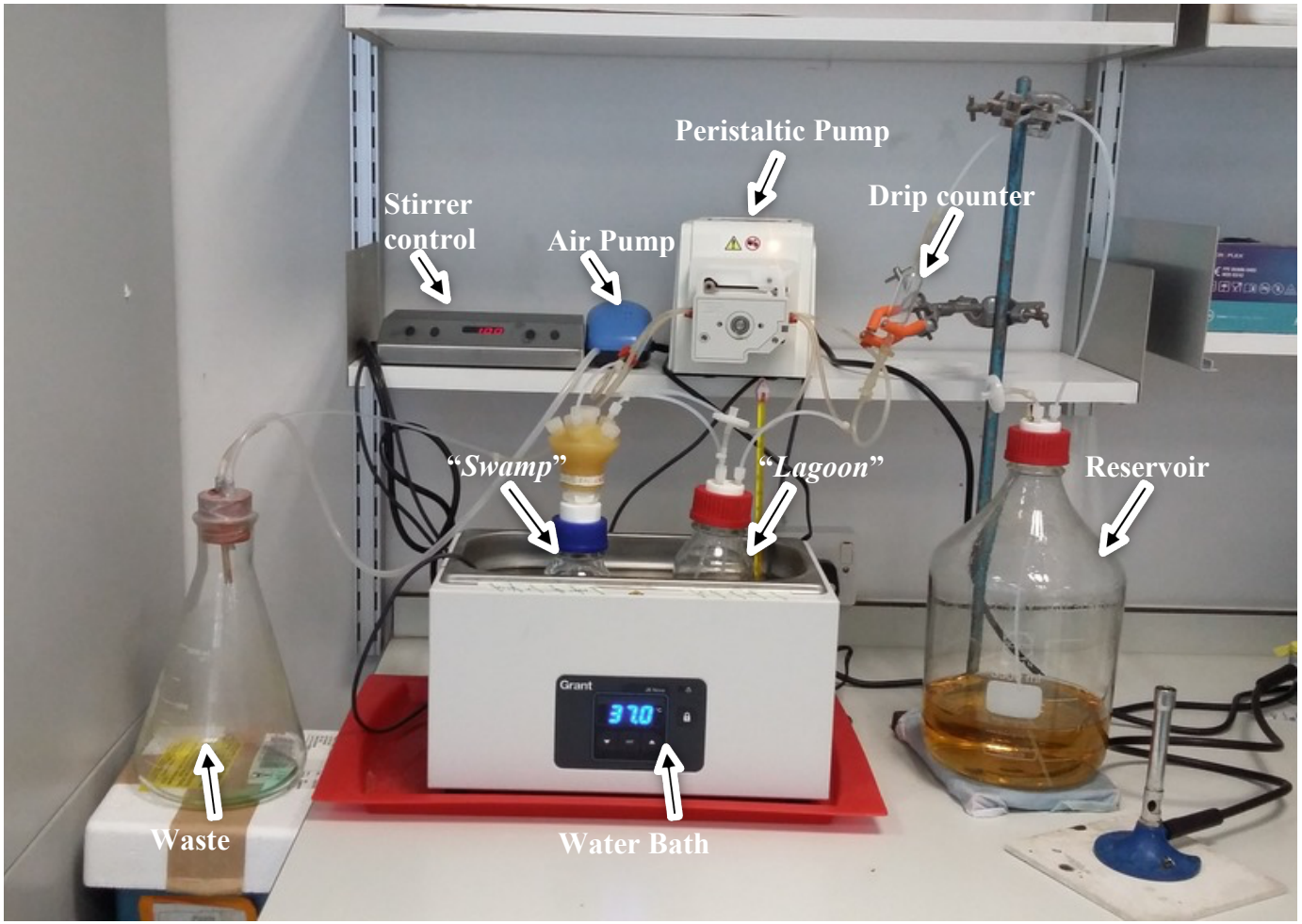


Figure 3.3: Chemostat (apparatus 3) used for continuous culture of *E. coli* K-12$^{gmhB\text{-mut}}$ with a new water bath, shorter tubing length and submersible water resistance stirrers. The *lagoon* was used for the growth of bacterial cells (illustrated in a generic chemostat as bacterial vessel figure 1.5) and the *swamp*, for phage growth (illustrated as phage vessel in figure 1.7).

3.3.1.4 Apparatus 4

The final chemostat design used for the host switching experiment (chapter 4) consisted of all the improvements made in figure 3.3 but with introduction of a new larger water bath (18L capacity; table 3.1), an extra peristaltic pump which made room for connection of additional chemostat vessels, a bigger waste vessel to accommodate the additional set-up, a larger submersible stirrer and shorter tubing length to further minimise biofilm formation (figure 3.4).

Owing to its greater capacity apparatus 4 was multiplexed: having two chemostats running in parallel. The majority of the tubing in each chemostat was PTFE, with silicone tubing used for peristalsis, waste disposal and for connecting the drip counter. In each chemostat three segments of tubing can be described:

- reservoir to *lagoon*: 30.5 cm from reservoir vessel to reservoir cap (all PTFE), 108 cm from reservoir cap to *lagoon* cap (PTFE + silicone, including short segments attaching the inline drip counter: figure 3.4), 6 cm traversing the *lagoon* cap into the *lagoon* vessel airspace (PTFE),
- *lagoon* to *swamp*: 12 cm from *lagoon* vessel (immersed) *traversing* the lagoon cap (PTFE), 57 cm for peristalsis (silicone), 6 cm traversing the *swamp* cap into the *swamp* vessel airspace (PTFE),
- *swamp* to waste: 12 cm from *swamp* vessel (immersed) traversing the *swamp* cap (PTFE), 57 cm for peristalsis (silicone), X cm to waste (material).

Peristaltic tubing was massaged to generate flow using two 4-channel peristaltic pumps. The pumps were configured as follows:

- the first pump drove fluid flow from reservoir to *lagoon* and from *lagoon* to *swamp* for both chemostats,
- the second pump drove fluid flow from *swamp* to waste for both

chemostats.

In addition to peristalsis, unidirectional flow was maintained via suction applied to the waste vessel. An aquarium pump (table 3.1) was modified by rotating the valve, to achieve suction. This suction was applied through 83 cm of vinyl tubing connected to the waste vessel.

All culture vessels were immersed in a water bath, above the internal volume line, in order to maintain a stable temperature (37^{o}C), and continuously stirred with the aid of large magnetic submersible stirrer-bars (table 3.1) to minimize biofilm formation and for aeration. Attached to the *lagoon* vessels were Duran 3-connection caps, with one port received fresh media from the reservoir, another fitted with a sterile 0.22 µm filter and the last supplying the *swamp* with naïve bacterial cells (via peristaltic pumps). The *swamp* vessels had 6-ported bottle caps, with one large NPT port (table 3.1) for easy sampling; one port received bacterial cells while another allowed spent medium, bacterial cells and phage to be transferred from the *swamp* to the waste vessel. All electrical connections were placed on a shelf above the water bath, and a drip counter connected, to measure media flow.

3.3.1.4.1 Sampling procedure

The growth system was maintained for two-day periods to minimise biofilm formation and maintain approximately neutral pH, but most importantly to minimise the adaptation of bacterial cells in the growth chambers. At the end of each period, the apparatus was sterilized and re-inoculated with fresh bacteria from frozen stock, into the *lagoon* (figure 3.4). The most recent ΦX174 sample (10 mL) was treated with chloroform (10% v/v) and centrifuged at 11,000 rpm for 4 minutes. Aliquots stored at -80^{o}C with DMSO 7% v/v were used to re-inoculate the *swamp* chamber after the chemostat has reached a steady state (after ~ 4 – 5 hours of growth) or for further analysis. By replacing bacteria and resuming phage growth in this way, ΦX174 continues to adapt (the interest of this study), while bacterial cells

were discarded to reduce host adaptation.

In a two-chambered chemostat, contamination of bacterial cells in the *lagoon* by phages is of great concern. It may occur via back-flow of bacteria + phage through the tube feeding the *swamp* with naïve bacterial cells. To control for phage contamination in the *lagoon*, the tubes delivering the bacterial cells were short and only allow for bacterial cells dripping directly into *swamp* without immersing the whole tube in the liquid phage-bacterial culture. In addition, the *lagoon* was checked and tested for phage contamination using plate overlay assay method (section 2.2.3) at regular but undefined intervals. Throughout the experiment, no contamination was observed.

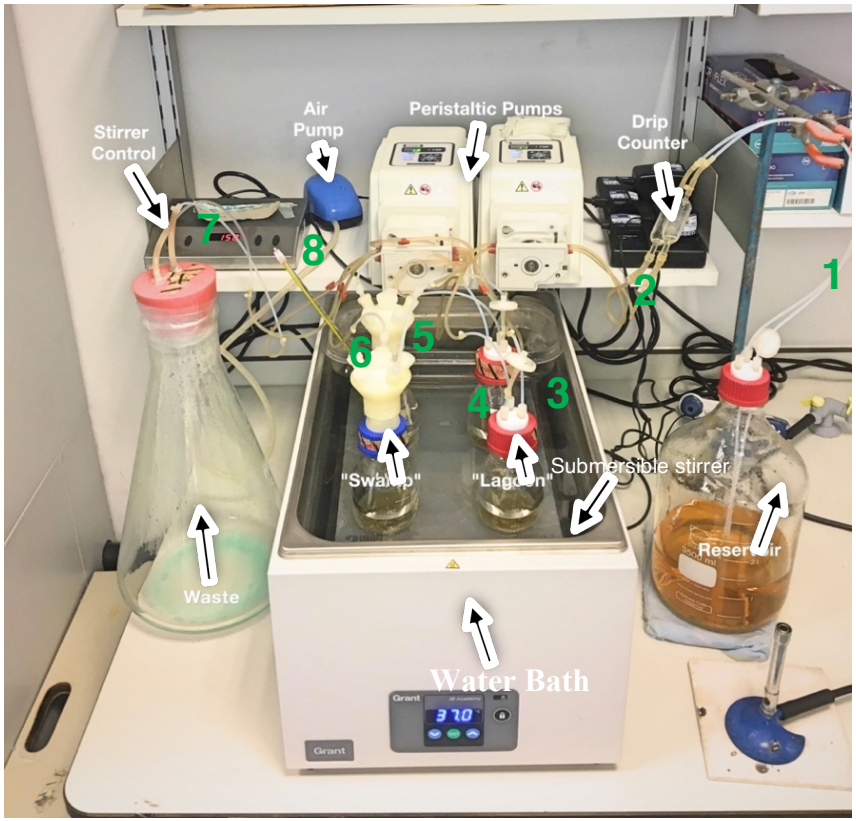


Figure 3.4: Chemostat (apparatus 4) used for continuous culture of *E. coli* C and *S*. Typhimurium with a submersible water resistance stirrer. The *lagoon* was used for the growth of bacterial cells (illustrated in a generic chemostat as bacterial vessel figure 1.5) and the *swamp*, for phage growth (illustrated as phage vessel in figure 1.7) with dilution rate of 0.38 / hour for *E. coli* C and 0.40 / hour for *S*. Typhimurium. The connections flowed from 1- Reservoir to drip counter (40 cm PTFE tubing), 2- drip counter to peristaltic tubing (5 cm silicone tubing connected to 57 cm peristaltic silicone tubing via barbed fittings), 3- peristaltic pump to lagoon (6 cm PTFE tubing), 4- lagoon (12 cm PTFE tubing) to peristaltic tubing (57 cm), 5- peristaltic tubing to swamp (6 cm PTFE), 6- swamp (12 cm PTFE) to peristaltic tubing (57 cm), 7- peristaltic tubing to waste (12 cm PTFE, fitted to 10 cm silicone tubing), 8- waste to suction aquarium pump (83 cm silicone tubing).

3.3.2 Chemostat dynamics

To evaluate how viruses adapt to different environments, ΦX174 was evolved for ~720 generations for four consecutive time periods, resulting in ~2880 generations for the whole experiment. The evolutionary study was performed in the chemostat described in section 3.3.1.4 (apparatus 4, figure 3.4.). A low-budget chemostat was designed to accommodate an additional chamber for phage growth (known in this thesis as the "*swamp*", section 3.3.1). For this experiment, we attempted to achieve a constant and fixed volume and flow rate every time the system was re-inoculated; keeping volumes constant required an approximately uniform dilution rate through each chemostat. A constant dilution rate should also result in a fixed population size although the limiting resource for the bacterial growth was unknown. The principal experimental difference from conventional chemostat studies was the use of bacteria as hosts for (and therefore as limiting resources for) phage propagation in the additional chamber (figures 3.3, 3.4).

Cultures of *E. coli* and *S.* Typhimurium were maintained in a chemostat in an approximately steady-state condition. Not all chemostat parameters describing growth efficiency could be determined, but the most important parameters, the dilution rate and the growth rate, were known. By controlling the dilution rate, the growth rate in the chemostat can also be controlled, as long as dilution rate does not exceed a critical point where cell population washes out. In steady growth state ($\boldsymbol{\mu}$), equilibrium was achieved when the rate of growth of the bacteria was equal to the dilution rate of flow $\boldsymbol{(D)\ \mu = D}$

Here, the specific growth rate was calculated using the following equation:

$$\boldsymbol{In \times = \mu t + In X_0}$$

$$\boldsymbol{\mu = \frac{In X - In X_0}{t}}$$

where X_0 is the initial cell density, X is the cell density at time t and μ is the specific growth rate. Cultures of *E. coli* and *S.* Typhimurium have a different growth rate. The *E. coli* specific growth rate μ was 0.38 cfu / hour while 0.41 cfu / hour was achieved for *S.* Typhimurium (data not shown). The dilution rate was 0.375 L/ hour (with 1.8 L of media from the reservoir spent within 24 hours) according to the following equation:

$$D = {}^{Q}/_{V}$$

Where D is the dilution rate, Q is the flow rate in L / hour and V is the volume of the culture vessel.

Doubling time (T_d), the time required for the cell concentration of a population of suspended cells to double, was calculated according to the following equation:

$$T_d = In(2) / \mu$$

For the chemostat set-up, the doubling time T_d was 1.82 hours for *E. coli* C and 1.67 hours for *S.* Typhimurium.

The mean residence time r_t is the average time bacterial cells stay in each chemostat compartment:

$$r_t = {}^{1}/_{D}$$

$$rt = 2.66\ hour$$

Overall, the dilution rate D (0.37 L / hour) is approximately equal to the chemostat growth rate μ (0.38 cfu / hour) for *E. coli* while 0.41 cfu / hour was recorded for *S.* Typhimurium. The average time bacterial cell resides in the chamber before it is diluted out is more than calculated doubling time (1.82 / hour – *E. coli* C or 1.65 / hour – *S.* Typhimurium.

The growth system was maintained for 2 days before the apparatus was sterilized and re-inoculated with fresh bacteria from frozen stock (section 2.2.2). Therefore, phage adaptation may continue while bacterial cells adapting to chemostat system are discarded every 2 days. This procedure was anticipated to limit the evolution of host cells, reduce the risk of contamination and biofilm formation.

In a chemostat, microbes are grown in growth-controlled conditions. This allows cells and phage to grow in steady-state at constant rate in an invariant environment. Although, Ferenci (2007) suggested that an ideal steady-state cannot be established in a chemostat because for limiting nutrient the residual concentration continues to decline for hundreds of hours in a chemostat. A practical solution is to remeasure the values of chemostat variables and check for likely shifts in growth rate in the population, which may be difficult because of perturbations in transcription. Evidence of perturbations in transcription levels was given by Rautio *et al*. (2006). Their study was conducted using *Trichmothecium reesei*, a filamentous fungus, where it was revealed that repeated measurements do not show a 'steady state' due to changes in transcriptional levels. In the same organism a novel transcript-based method was used to study the expression stability of marker genes and changes to the steady state were noted at the transcriptional level with genes involved in growth and macromolecule synthesis being identified as markers of disturbances in culture conditions (Rautio *et al*., 2006). To minimise perturbations in the culture conditions and limit deviation from a steady state in our experiment, the growth system was maintained for two days after which it was discarded, washed and sterilized. Despite the challenges associated with maintaining a steady state, a chemostat system has advantages over serial passaging culture methods in which the environment is continually changing as cells divide, consume nutrients and produce waste (Monod, 1950), with these changes associated with cell growth and ultimately affecting cellular physiology (Valgepea, 2013).

Continuous culture in a chemostat allows the control of growth rate where the external environment can be made static with the experimenter controlling nutrient availability. Culture cell physiology and chemical conditions including pH, biomass, growth rate, oxygen and metabolites remain stable and are approximately constant for an extended period of time (Bull, 2010).

3.3.3 Hosts used in culturing

E. coli and *S.* Typhimurium were used in this study as hosts for the propagation of phage ΦX174 (table 2.1). By comparing the amino acid sequences from 32 proteins of 72 species prokaryotes and eukaryotes, Battistuzzi *et al*. 2004 estimated the phylogenetic relationship of the genera *Escherichia* and *Salmonella*. The authors suggested divergence occurred ~102 million years ago. The genus *Escherichia* was proposed to have diverged from *Salmonella* and further split into five species; *E. albertii*, *E. fergusonii*, *E. coli*, *E. hermannii* and *E. vulneris* while *E. coli* species can be further distinguished into five subspecies: Groups A, B1, B2, D and E by examining the complete genome sequence of the genus (Meier-Kolthoff *et al*., 2014). The commonly used laboratory *E. coli* strains belong to the group A including *E. coli* utilised for experimental evolution in this study. The genus *Salmonella* are classified into two main species, *enterica* and *bongori* which are further classified into subspecies and serotypes. *S.* Typhimurium belong to the species enterica, subspecies I *enterica* and serovar Typhimurium (Agbaje *et al*., 2011). The *S.* Typhimurium and *E. coli* strains were used in this study for alternating host switching (chapter 4) and the relationship between these two strains are shown in figure 3.4 together with the *E. coli* K-12 host (also used for experimental evolution in this study but had not sustained long-term growth of ΦX174, chapter 4). The *E. coli* C strain is closely related to ST10 and *E. coli* K-12 (figure 3.4).

Enterobacteriaceae strains including *S.* Typhimurium and *E. coli* possess an outer cell membrane consisting mainly of Lipopolysaccharides (LPS) where ΦX174 bind to their host cells for infection.

3.3.4 Host cell recognition and penetration by phage ΦX174

Bacteriophage ΦX174 binds to the LPS of bacterial cell wall of Gram-negative. Specifically, the terminal galactose in the core oligosaccharide of "rough LPS" is used as the receptor in ΦX174's infection process (Feige and Stirm, 1976). The *Enterobacteriaceae* cell surface is formed by an outer membrane (OM) or cell envelope. The inner leaflet of OM comprises mainly phospholipids that are analogous to the cytoplasmic membrane. The outer leaflet is composed mainly of LPS (Letarov and Kulikov, 2017). LPS is a major component of outer membrane of the rough strains of *Enterobacteriaceae* which include *E. coli* and *S.* Typhimurium (Hayashi *et al*., 1988; Fane *et al*.,2000). The LPS molecule may be divided into two parts: lipid A, the hydrophobic membrane, and a core oligosaccharide (core OS) composed of 10 to 15 sugars molecules attached to lipid A. The core OS may be sub-divided into two regions: an inner core, often conserved among *Enterobacteriaceae*, and an outer core region, which exhibits variations in its structure. The core OS is often phosphorylated, and in this case the structure is referred to as "rough LPS". In *Enterobacteriaceae,* if rough LPS is capped by an O antigen side chain polysaccharide, this results in the formation of "smooth LPS" (Heinrichs *et al*., 1998; Yethon *et al*., 2000). Rough LPS is therefore defined by the lack of the O-antigens possessed by many pathogenic *E. coli* strains. ΦX174-sensitive bacteria appear to be exclusively rough LPS strains (Michel *et al*., 2010). The genes *waa* (previously *rfab*), *rfaE*, *lpcA* and *gmbB* encode biosynthetic enzymes for LPS production. The various *E. coli* strains exhibit five types of outer core LPS: R1, R2, R3, R4 and K-12, associated with genetic variation in the *waaQ* (heptosyltransferase responsible for addition of heptose sugar of the inner core of LPS; Yethon *et al*., 1998) operon. *E. coli* C has an R1 core type LPS as ΦX174 receptor. *E. coli* K-12 exhibits a K-12 core type with a rough

phenotype that is resistant to ΦX174. Knocking out the *rfab* and *gmhB* genes of *E. coli* K-12 has been shown to create susceptibility to ΦX174 infection (Baba *et al*., 2010). Also, *S.* Typhimurium may be sensitive to ΦX174 despite its non-R1 core LPS, since it bears different hexoses that can be modified to allow infection (Janson *et al*., 1989). These observations suggest that ΦX174 recognition and attachment depends on LPS global conformation rather than its exact (amino-acid/sugar) composition (Feige and Strim, 2006) and that ΦX174 shows some degree of host-range flexibility.

ΦX174 is dependent on calcium ions for host recognition. Upon contact and binding with the ions, the conformation of the amino acid side chain of glucose binding sites changes (Fane *et al*., 2000). This allows irreversible interaction of ΦX174 with LPS of a susceptible host. Previous studies have demonstrated that a coat protein (F), the DNA pilot protein (H; table 1.1 of chapter 1), interact specifically with Lipid A resulting in a conformational change of these proteins that functions as a trigger for phage DNA ejection (Inakagi 2000; 2005).

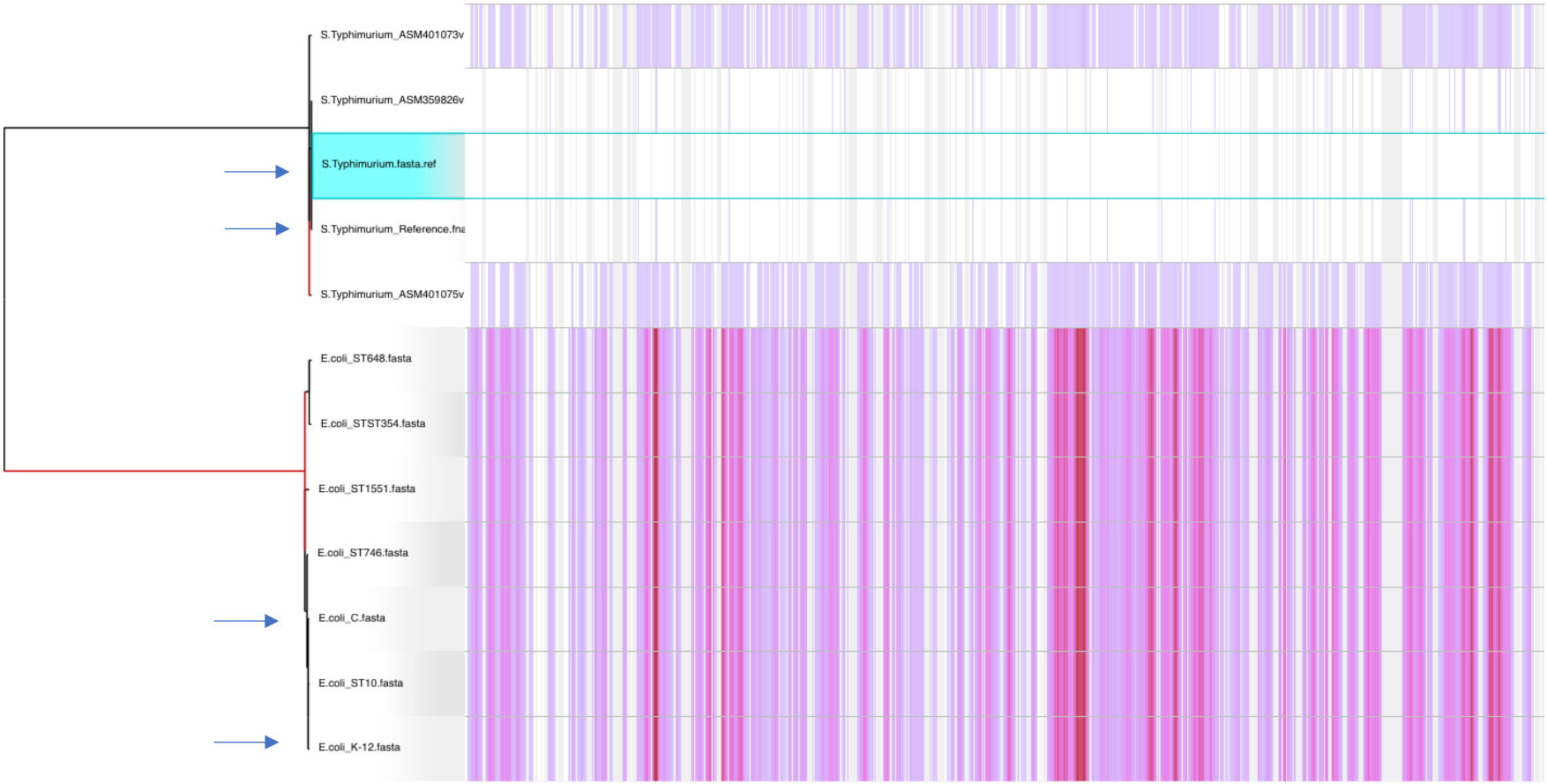


Figure 3.5: Phylogenetic tree and alignment of some Enterobacteriaceae family and hosts: *E. coli* C (sequenced by Anton Nekrutenko's group at the Pennsylvania State University, U.S.A), *E. coli* ST strains (from Seecharran, 2018), *E. coli* K-12 (obtained from NCBI with accession number NZ_CP014225.1), *S.* Typhimurium (obtained from NCBI with accession numbers NZ_PXVG01000010.1, NZ_CP034831.1, NZ_CP034819.1), S. Typhimurium_Reference (referred to as ST_reference NCBI-obtained, see section 2.6.1.1, accession number NC_003197.2), used in this study, indicated by the arrows. Tree and alignment generated by gingr (Treangen *et al.*, 2014).

3.3.5 Different studies that utilise *E. coli* and *S.* Typhimurium

Several designs of chemostat system have been reported by researchers. Because *E. coli* is well-studied organism, it is commonly used as a model organism by the developer of chemostat bioreactor systems. Schmideder *et al*. (2015) reported the design of a 48-parallel stirred-tank single-use chemostat system with 8-14 mL culture capacity fitted with fluorometric sensors for real-time dissolved oxygen, pH and temperature control and magnetically driven gas-inducing stirrers for proper oxygen transfer. Their experiment made use of *E. coli* BL21(DE3) carrying a pET28a(+) plasmid carrying a PAmCherry (27KDA) kanamycin selection marker to track the kinetic growth parameters of this strain in their chemostat. Nanchen *et al*. (2006) developed an 8-parallel bioreactor system, with a 17 mL capacity and a working volume of 10-ml cultures. In this system, *E. coli* MG1655 (in glucose-limiting M9 medium) was used and the impact of growth rate on metabolic rate was determined using ^{13}C-labelling for detecting carbon fluxes via different dilution rates. Zhang *et al*. (2006) developed the polymer-based microbioreactor system and made use of *E. coli* FB21591, a derivative of *E. coli* K-12, as model organism to determine its kinetics and growth rate. This system was additionally used to compare different structural materials suitable for minimising biofilm formation in a chemostat. Jeong *et al*., 2016 cultured two *E. coli* strains (W3110 and DST160) in a chemostat for 270 days to determine stress adaptation in the presence of high-succinate. *E. coli* W3110 showed nearly zero growth while *E. coli* DST160 exhibited unperturbed growth.

Phage ΦX174 has the ability to infect *S.* Typhimurium and *E. coli* C, therefore, various experimental evolution studies in bioreactors have explored ΦX174 as well as the bacterial hosts (section 1.8.3; summarised in table 5.4). Crill *et al*. (2000) studied the molecular and evolutionary basis of ΦX174 adaptation to both *S*. Typhimurium and *E. coli* C in a chemostat where they discovered that adaptation of ΦX174 on *S*. Typhimurium depressed its ability to grow on *E. coli* C, while ΦX174 adaptation on *E. coli*

did not affect growth on *S*. Typhimurium. In a similar study, replicate lineages of ΦX174 on either *E. coli* C or *S*. Typhimurium hosts exhibited similar substitution rates and fitness trajectories during a period of adaptation in a chemostat (Bull *et al*., 1997). Brown *et al*. (2013) designed a chemostat to determine the general patterns of genetic change underlying evolutionary adaptation in a parallel experiment with phage ΦX174 using *E. coli* C and *S.* Typhimurium as hosts for 50 days. They discovered that during adaptation a high-density chemostat environment selects for substitutions at sites associated with ΦX174 host recognition and capsid stability.

3.3.6 ΦX174 growth rate on both hosts

Experimental evolution employs various experimental designs which may influence organisms' responses to their environments. Experimental designs impose different selective pressures, most often with the aim of keeping all or some parameters constant such as nutrients, temperature, hosts, culture volume, according to the interest of the experimentalist. In propagation of microbes, evolving populations need to be supplied with fresh nutrients while end products are removed. Two major designs are usually employed, serial transfer and continuous culture (section 1.8.2 and 1.8.3). In general, populations are propagated with an ancestor that may be flourescently labelled, display a specific colour or be sequenced to determine the starting genotype. In a standard setup, a known ancestor is transferred into the desired environment and evolving populations are regularly diluted. Such an arrangement allows multiple generations of cells to undergo division and growth, enabling competition, coexistence and recombination which ultimately affect the growth rate of the evolved populations through evolution.

In this study, the growth rate of the wild-type phage (represented as A in figure 3.6) prior to chemostat adaptation was measured on *S*. Typhimurium and *E. coli* C (yellow dashed arrows 1 and 2 in figure 3.6 and horizontal dotted lines in figure 3.7). Afterwards, wild-type ΦX174 was transferred into chemostat (figure 3.6, black arrows), initially adapted on *S*. Typhimurium (S-

host; figure 3.6 and termed S-branch; figure 2.1) and parallel adaptation on *E. coli* C (C-host; figure 3.6 and termed C-branch; figure 2.1) for 10 days (figure 3.6). The growth rate after *S.* Typhimurium (days 1 and 10) and *E. coli* C adaptation (days 1 and 10) was measured on both hosts (shown as yellow dashed arrows for: day 1 - lines 3 and 5, day 10 – lines 4 and 6; figure 3.6) as absolute fitness, one of the components of growth, after 45 minutes (figure 3.7).

S. Typhimurium was introduced as a novel host. However, ΦX174 was able to infect the novel host and produce progeny, exhibiting the expected host range (Gratia, 1936). The wild-type phage growth rate was seen to be higher on *S.* Typhimurium than on *E. coli* C (figure 3.7) despite being initially isolated on *E. coli* C. Although, there was no evidence that the wild-type phage has been adapted to *E. coli* C but the initial plaque isolation was done on *E. coli* C after receiving the phage from Dr Holly Wichman's lab, University of Idaho, USA (section 2.1.2).

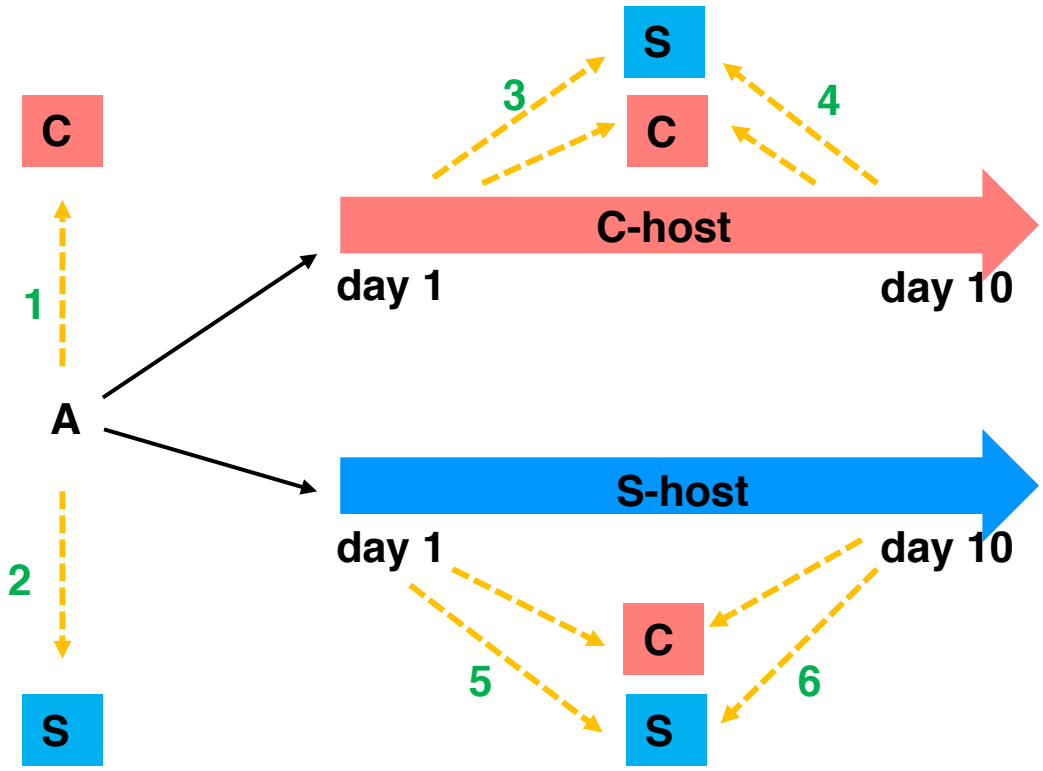


Figure 3.6: A represents the wild-type phage, C-host: adaptation of wild-type on *E. coli* C, S-host: adaptation of phages on *S.* Typhimurium. **A** was transferred (black arrows) onto C-host and S-host separately and grown for 10 days (bold red arrow and blue arrow respectively). Yellow dashed arrows indicate growth measures. Arrows 1 and 2 depict measurement of wild-type phage on *E. coli* C (C) and *S.* Typhimurium (S), respectively. Arrows 3-6 show growth rate measurements on C-host day 1 (arrow 3) and day 10 (arrow 4) and on S-host day 1 (arrow 5) and day 10 (arrow 6). Blue boxes show phage fitness on *S.* Typhimurium; red boxes show phage fitness measured on *E. coli* C.

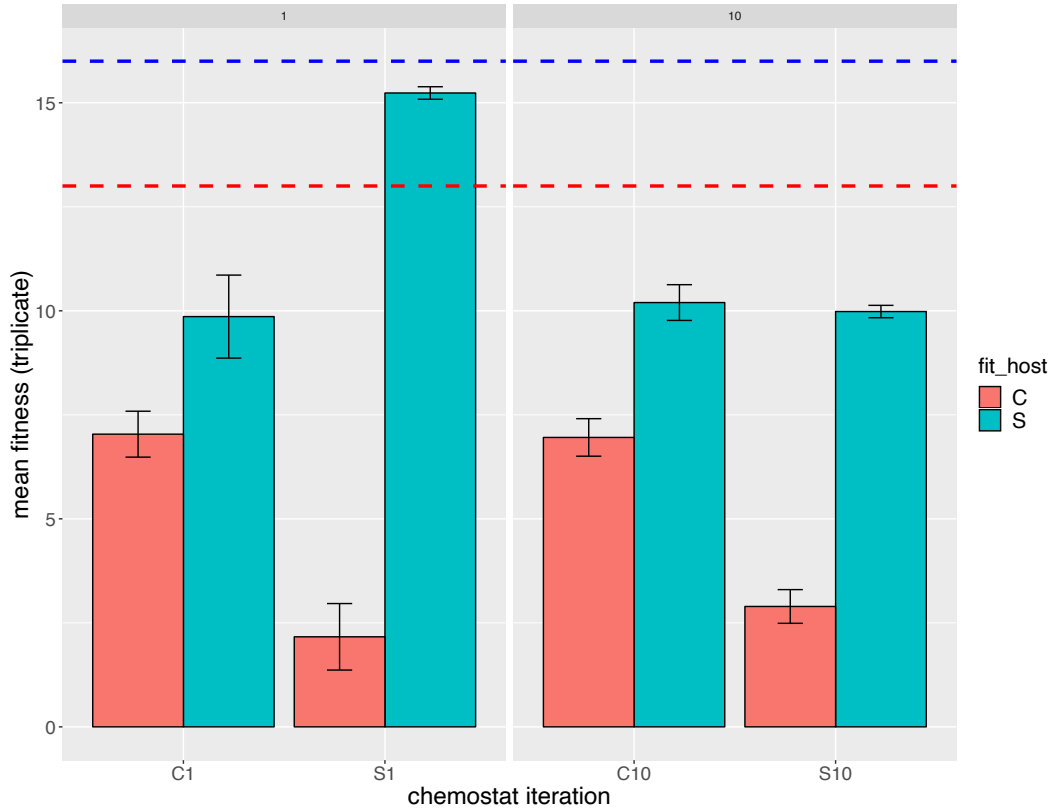


Figure 3.7: Blue bars: wild-type fitness on *S.* Typhimurium, red bars: wild-type fitness measured on *E. coli* C. Ancestral ΦX174 fitness was evaluated on *S.* Typhimurium (blue dashed line) and on *E. coli* C (red dashed line) prior to chemostat selection.

Variable	R^2	F value	p value
Days	0.211	59.289	9.999×10^{5}
Hosts	0.132	74.522	9.999×10^{5}
Fit_hosts	0.359	201.808	9.999×10^{5}
Days × Hosts	0.014	8.030	0.006
Days × Fit_hosts	0.074	20.790	9.999×10^{5}
Hosts × Fit_hosts	0.162	91.469	9.999×10^{5}
Days × Hosts × Fit_hosts	0.010	5.635	0.019

Table 3.2: Summary of non-parametric analysis of variance results for fitness assay, days 1 and 10 (represented has days) on both E. coli C and S. Typhimurium (represented has hosts). Fit_hosts implies the host on which fitness was measured. The analysis was carried out in R using the adonis2 function within vegan package. Default parameters were chosen except that 10,000 permutations were used and the seed was set to 124 before running the function (for reproducibility).

Using non-parametric ANOVA, interactions were examined for the followings; hosts (C-host and S-host, figures 3.6 and 3.7), days (day 0 for wild-type, days 1 and 10 on both hosts) and hosts on which fitness was measured (termed fit_host). In all, significant interactions were observed indicating hosts used in propagation, number of days and fit_host has effect on ΦX174 fitness (table 3.2). The interactions between all factors considered; days 1 and 10, host propagated and fit_host showed significant difference ($p < 0.01$, $F = 5.63$). There was highly significant difference between interactions of days and hosts used in propagation ($p < 0.001$, $F = 8.03$). Meanwhile, interactions between; days and fit_host, host propagated and fit_host exhibited very highly significant associations. The results indicate that ΦX174 fitness depends mostly on the hosts used and the number of adaptation days (including wild-type at day 0), table 3.2.

A chemostat system is robust for experimental evolution studies and employed in this study because any parameters may be held constant except the parameter of interest to the experimentalist. Here, presumably a steady state was achieved, the environments of ΦX174 remained invariant except the desired hosts used in propagation. The experimental design here are mainly influenced by; (a) the hosts; *E. coli* C and *S*. Typhimurium used (b) the operational settings which includes; dilution/flow rates and chemical composition of the nutrients (Joeng *et al*., 2016; Van den Bergh *et al*., 2018). Operational settings in-turns determine the population densities of hosts. The growth rates of ΦX174 measured in the two hosts on days 0, 1 and 10 differs and were high to very significant associations were observed between the hosts and days. Therefore, adaptation of ΦX174 was influenced by the chemostat environment (comparison of day 0 with days 1 and 10 in reference to fitness. In the chemostat, larger populations of and continuous selection of bacterial cells were achieved, providing more genetic diversity of the whole population which in turns may influence the fitness outcome of ΦX174.

3.4 Conclusion

In this chapter, we have shown that a continuous culture system can be made from inexpensive materials and was adequate for the growth of both bacterial cells and phages. However, more improvements and optimization may be considered for better understanding of unknown parameters such as aeration rate, cell density, automated pH test and nutrients limitation. With the advent of open source materials and inexpensive custom fabrication processes, it is possible to design a customised 3D printable system (Takahashi *et al*., 2015), software and hardware for different unique experiments (Matteau *et al*., 2015).

In the present research, nutrients and culture environments were held constant except the bacterial cells (hosts) which serve as the variable environment. For phages, the hosts in which they proliferate are the ecologically relevant source of selective pressure and rate of growth on their respective hosts is a major component of fitness and adaptive potential. The steady state environment assumed all other environmental factors constant except the hosts used. In such, the study provides an opportunity to study ΦX174 host range, costs and interactions associated with phages colonisation of a novel environment. The chemostat developed herein provides a means of controlling host selective pressure, studying growth rates while switching hosts (chapter 4) and using high throughput sequencing methods to reveal accumulation of mutants (chapter 5) in an attempt to understand the basis of adaptation during host switching.

Chapter Four: Measuring ΦX174 fitness and attachment during host switching

4.1 Introduction

4.1.1 Parasite host-switching

A pathogen depends on another organism, its host, for completion of its life cycle. Two major trends are observed in long-term virus evolution. Firstly, a virus may remain associated with the host, evolving as the host evolves by making adjustments via mutation every time with the host genetic composition changes. This pattern, in which viruses remain associated with same host over macro-evolution, is called co-divergence. For instance, humans and the closely related African great apes have slightly different versions of polyomaviruses (family *Polyomaviridae*), which were suggested to have acquired mutations that largely mirror those of their hosts, indicating that codivergence was chiefly responsible for virus diversification (Madinda *et al*. 2016). The other macro-evolutionary trend is cross-species transmission which occurs when virus populations transition into new hosts. Such viral evolution is significant and linked to emerging viral diseases like HIV, Ebola, influenza A virus subtype H5N1 and SARS (Wolfe *et al*., 2007). Host switching has been suggested to be a common occurrence over macroevolutionary time scales (Geoghengan *et al*., 2017), but occurrence frequency may be influenced by the opportunity for crossing into a new host – that is, by transmission potential.

The ability to seek and colonize new hosts may be influenced by several factors, such as transmission mechanisms, local ecosystems, competition, community composition, abiotic and biotic factors and pathogen virulence. For a pathogen to transition to an alternative host requires first, the ability to find and infect a suitable host, entailing a transmission and entry route. Second, for viable reproduction of the population, it is important for the pathogen to acquire the ability to spread, establish and adapt to the new environment. Successful transmission in the new host is often accompanied by acquisition of new genes, nucleotide substitutions, recombination/re-assortment and natural selection (Holmes *et al*., 2005).

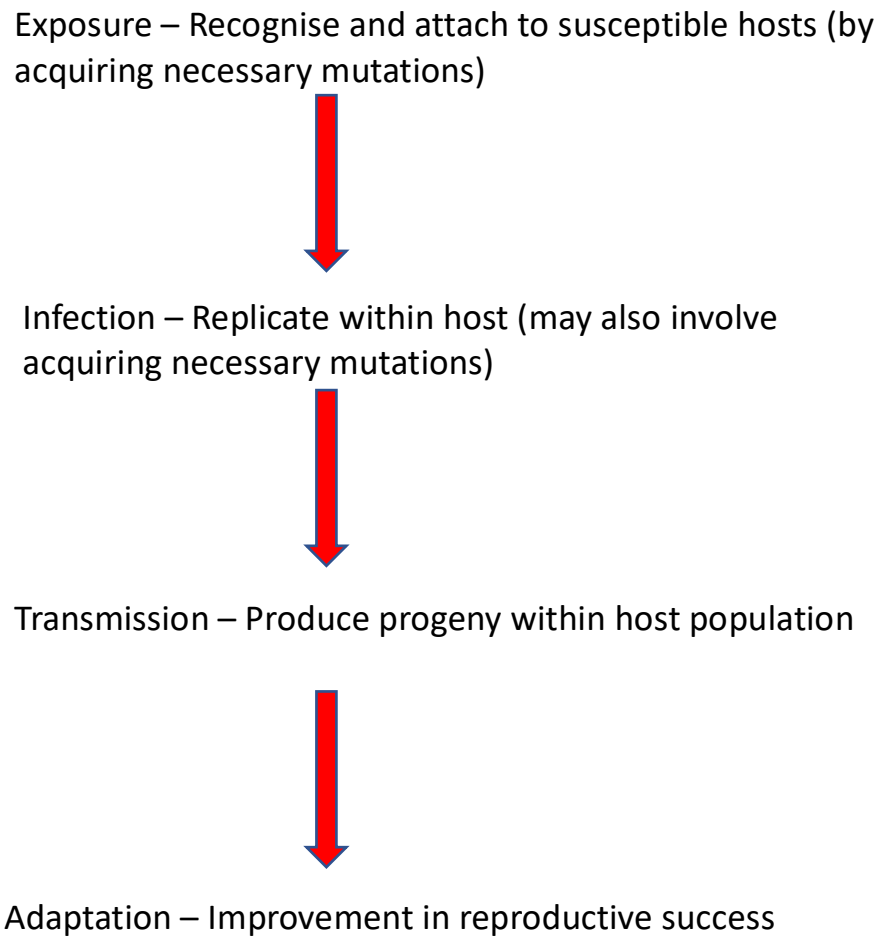


Figure 4.1: Transition pathways to successful colonisation of an alternative host during a host switching event.

Figure 4.1 illustrates likely barriers to overcome prior to successful host switching for a viral population. Such populations may maximize transmission and survival routes into the host cell regardless of variation in host physiology and the abiotic environment for successful colonisation. It may be necessary for genetic variation to be present in a population or for it to arise (via mutation or recombination) before phenotypes that can successfully survive in the new environment are produced.

A critical step in host switching success for viruses is the attachment rate. Attachment and infection of new hosts is defined by the receptor binding capability of the virus. Attachment to a new host is the first step whereby viruses interact and associate with their host cells. Receptor binding often plays a major role in host switching, for instance HIV-1 binds to CD4 host receptors, as well as to CXCR4 and CCR5 coreceptors (Philpott, 2003). Avian and mammalian influenza viruses show some host specificity by binding to different glycan linkages or sialic acids associated with a particular host (Shinya *et al*., 2006). In phages, a binding receptor is required for bacterial cell colonisation, and may involve proteinaceous targets (mainly outer member proteins), sugar moieties (within the cell wall these may be pellicles, teichoic or lipoteichoic acid) or a combination of both (Bertozzi *et al*., 2016). In all cases, attachment has been shown to involve either protruding structures or constituents of the bacterial cell wall which are exposed and easier to access (Phage Receptor Database, accessed December, 2018). Phages lose infective ability if host receptors are inaccessible or non-complementary.

4.1.2 Viral fitness trade-offs

Trade-offs occur when a trait that confers an advantage in one context of environment simultaneously confers a disadvantage in another. Genetic factors (Truyen *et al*., 1995; Diehl *et al*., 2016; Urbanowicz *et al*., 2016), environmental factors (Bull *et al*., 2000; Jessup *et al*., 2008), coexistence

(Bohannan *et al*., 2002), coevolution (Koskella and Brockhurst, 2014), competition (Huisman and Wessing, 2001; Wichman *et al.,* 2005) and environment interactions determines the existence of trade-offs.

One major barrier to host-switching capability is receptor binding, since the initial infection of novel host is an important step in host switching. Mutations may occur in viruses that enable the use of alternative receptors in a novel host (Hueffer *et al*., 2003; Weaver and Barret, 2004). While these mutations facilitate the exploitation of a new host, they may reduce viral fitness (Abedon *et al*, 2001; Duff *et al*., 2005; Wang 2006; Rodrıguez-Verdugo *et al*., 2014). Therefore, it is possible for a population to be established on and adapt to a new host, but this may entail a trade-off.

One way in which a trade-off may be revealed is that, after adaptation to a new host, a lower fitness may be observed on the original host. In ΦX174 this has been observed in the context of switching from one host to another (Crill *et al.,* 2000). If the host environment changes infrequently, evolution can lead to selection of phage specialists, able to infect a specific host. If switching between hosts occurs regularly this can impose a long-term fitness cost. In such situation, the host environments select for phage generalists - surviving under the separate host environments. Although, the phage generalists are able to infect both host environments, but in one way, may bear some costs with a lower fitness on both hosts without reaching a fitness level that would have been achieved by a specialist. In another way, host environments may select for generalists where fitness increase in one environment may results in reduction of fitness in the other environment. Both ways involve unavoidable trade-off for phage generalists in the different host environments (Remold, 2012). Microbial experimental evolution studies are well suited for exploring pathogen-host trade-offs. Using microbial model systems, it is possible to directly explore the consequences of sudden host colonization by a pathogen, determine the effects of environment on trade-off magnitude and establish the rate of pathogen-host attachment.

4.2 Aim and objectives

If adaptation to a novel host may entail fitness trade-offs, it would be beneficial for our understanding of viral evolution to investigate the costs associated with intermittent host switching and how early during adaptation this cost is likely to be observed. Trade-offs could result from fitness optimisation on the novel host or from the mutation that enables colonisation of the new host. A model organism ΦX174 was evolved continuously on three different hosts in this chapter. Fitness and attachment measurements were undertaken using qPCR to measure potential fitness costs.

The objectives of this chapter were:

- To adapt ΦX174 to different hosts: a mutant strain of *E. coli* K-12$^{gmhB\text{-}mut}$, *S.* Typhimurium and *E. coli* C in chemostat.

- To repeatedly switch ΦX174 between *E.coli* C and *S.* Typhimurium.

- To determine the costs associated with host switching via fitness and attachment measurement (using qPCR).

4.3 Results and discussion

4.3.1 *E. coli* K-12$^{gmhB\text{-mut}}$ fitness assays

The fitness component of ΦX174 growth rate was measured as doublings / hour, at a low MOI. To establish long–term reproductive success, parasites and hosts must be minimally compatible. *E. coli* K-12$^{gmhB\text{-mut}}$ was shown to be sensitive to ΦX174 infection (Ohkawa 1979; Michel *et al*., 2010), but to the best of my knowledge, an experimental evolution study utilising this host has not been described in the literature. *E. coli* K-12$^{gmhB\text{-mut}}$ utilised in this study is a mutant strain, with modification of the *gmhB* LPS gene (table 2.1), to allow attachment of ΦX174, thereby overcoming an initial level of protection of *E. coli* K-12$^{gmhB\text{-mut}}$ (host specificity and binding, an initial step during infection).

The continuous culture experiment with mutant *E. coli* K-12$^{gmhB\text{-mut}}$ shows that ΦX174 failed to survive after 2 days of continuous culture (~144 generations; figure 4.2). Successful infection was not established and this may be owing to a poor growth rate hence loss of survival, and inability of ΦX174 to successfully switch to *E. coli* K-12$^{gmhB\text{-mut}}$. According to Bohannan and Lenski (2000b), the larger the trade-off (of phage in a resistant host mutant), the higher the probability of extinction of the virulent phage. By measuring the fitness of ΦX174 on *E. coli* C (for both *E. coli* K-12$^{gmhB\text{-mut}}$ and *E. coli* C adapted), there was evidence of a substantial fitness trade-off. Since the growth rate was measured on *E. coli* C after evolving on *E. coli* K-12$^{gmhB\text{-mut}}$ for 2 days, there is possibility that ΦX174 fitness reduced on the ancestral host (*E. coli* C) over time.

Bacteriophages must be able to efficiently infect and initiate colonization of novel hosts. This includes the capacity of infection spreading between hosts as well as forming new association in the novel host. Hosts may impose barriers at different levels such as, attachment, genome entry, genome replication and gene expression, for successful phage infection (Parrish *et*

al., 2008). During adaptation to novel host, viruses may overcome these barriers, sufficiently producing transmissible progenies that are able to spread within the host population that can thrive for many generations and not go extinct. A successful host switching does not lead to population extinction, if virus in the new host environment die, regardless of the generation in which it occurs, it is a failed host switching attempt. Therefore, the ability to infect a new host does not imply sustainable long-term growth on the new host. Acquisition of mutation may allow viral adsorption to a novel host but may not be sufficient for the completion of viral life cycle in such host, further adaptation may be necessary.

The results from *E. coli* K-12$^{gmhB\text{-mut}}$ illustrate unsuccessful host colonisation. Although we have not measured fitness on the new host or in the chemostat environment a reduction in fitness was noted on the ancestral host (*E. coli* C). Although these results suggest the possibility of a fitness trade off, the evidence is insufficient to identify why colonisation was unsuccessful. Successful colonisation in nature may be driven by co-adaptation of host and pathogen, and progeny that are less fit may be lost, a case of trade-off as illustrated by Sabrina *et. al*., (2015). The production of less fit phenotypes has been speculated to enhance host range expansion (Duffy *et al*., 2005), during colonisation of new host species. In the scenario of Sabrina *et al*. (2015), trade-offs may result in reduction of host switching chances. For instance, in macro-evolution of HIV-I, it shifted from chimpanzees in the early 20th century to humans and HIV-I groups M and N have successfully shifted to humans (Sharp and Hahn 2010). The protein Vpu has evolved to antagonise tetherin restriction factors (tetherin expression block the release of HIV) in both groups (Lama *et al*., 1999). Vpu proteins bind and degrade CD4 receptors in Simian Immunodeficiency Virus (SIV) to assist viral progeny release in group M but this function is lost in group N, possibly as a result of anti-tetherin activity. It has been speculated that the loss in Vpu protein function on group N may explain why HIV group N remained a very rare pathogen in Africa, while in group M, the activity of Vpu possessing the

ability to degrade CD4 (section 1.2.5), has influenced group M prevalence in Africa (Sauter *et al*. 2009).

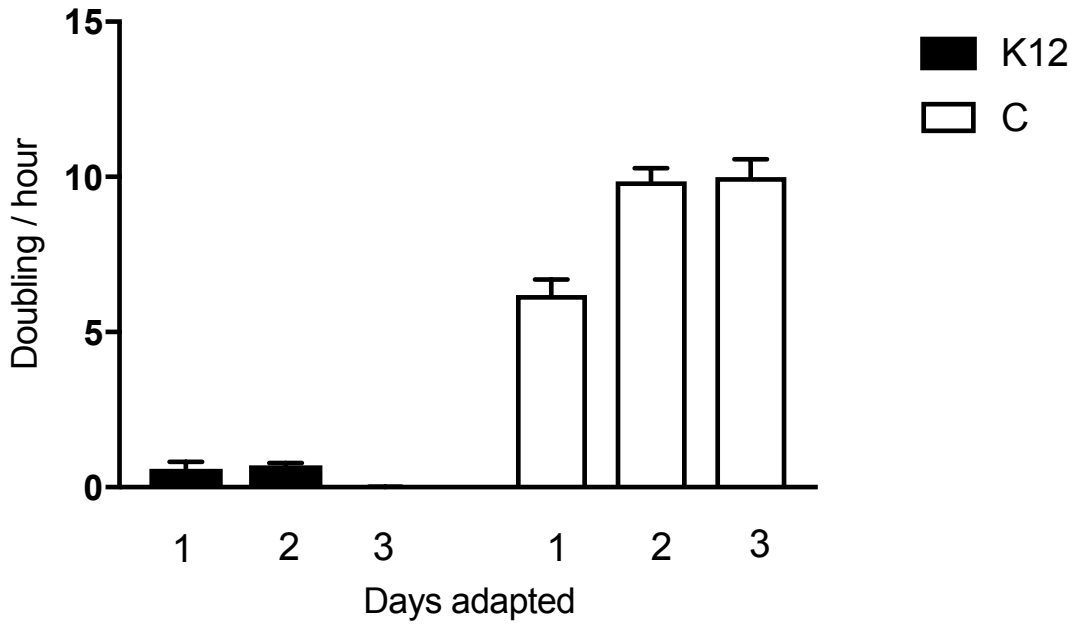


Figure 4.2: Fitness of ΦX174, measured in *E. coli* C, for populations adapted in *E. coli* K-12$^{gmhB\text{-mut}}$ or *E. coli* C in the chemostat for 3 days. After two days in *E. coli* K-12$^{gmhB\text{-mut}}$, fitness on *E. coli* C declined to zero. Fitness differences between K-12$^{gmhB\text{-mut}}$ - and C-adapted phages are very significant ($p = 0.003$) with Mann-Whitney test (section 2.4). Triplicate biological replicates (each derived from triplicate technical replicates) are plotted together with 95% confidence intervals of the means.

4.3.2 Trade-offs during alternating host switching

One consequence of host switching is that it may come at a cost for the pathogen. Phage ΦX174 was alternately switched between two hosts more distantly related than the two *E. coli* strains. *S.* Typhimurium has been used in ΦX174 studies (Bull *et al*., 1997; Crill *et al*., 2000; Brown *et al*., 2013), so, while it was a more distant host, it was expected to support ΦX174 growth (figure 4.3).

Alternating switching was carried out to determine if infection would be successful, whether adaptations can be reversed by returning the population to previous host and the costs associated with host switching. The existence of a trade-off was inferred by comparing fitness differences of ΦX174 on *E. coli C*, its normal laboratory host, and *S.* Typhimurium, a novel host with modified receptor necessary for ΦX174 attachment. ΦX174 successfully attached to the modified receptor on *S.* Typhimurium, infected, established, and, we may infer, was able to undergo multiple compete life cycles throughout the experiment.

In general adaptation of ΦX174 to *E. coli* C was accompanied by a reduction in fitness measured on *S*. Typhimurium (that is fitness on the *Salmonella* host was decreased when viruses adapted to a non-*Salmonella* host; compare blue bars in figure 4.3). The converse pattern is observed for ΦX174 adapted to *S*. Typhimurium, with *E. coli*-measured fitness decreased when viruses adapted to a non-*E. coli* host (compare red bars in figure 4.3). These patterns were more marked in later iterations, but there was no significant effect of day of adaptation (day 1 versus day 10) throughout the experiment (ANOVA, $p = 0.250$, table 4.2). For successful infection and in response to adaptation to different environmental factors, trade-offs are often assumed to play a role (section 4.1.2). The last host for ΦX174 selection has a very highly significant effect (ANOVA, $p<0.0001$) and the presence of a very highly significant interaction between this and the host on which fitness was measured provides evidence for a trade-off effect (ANOVA, $p<0.0001$).

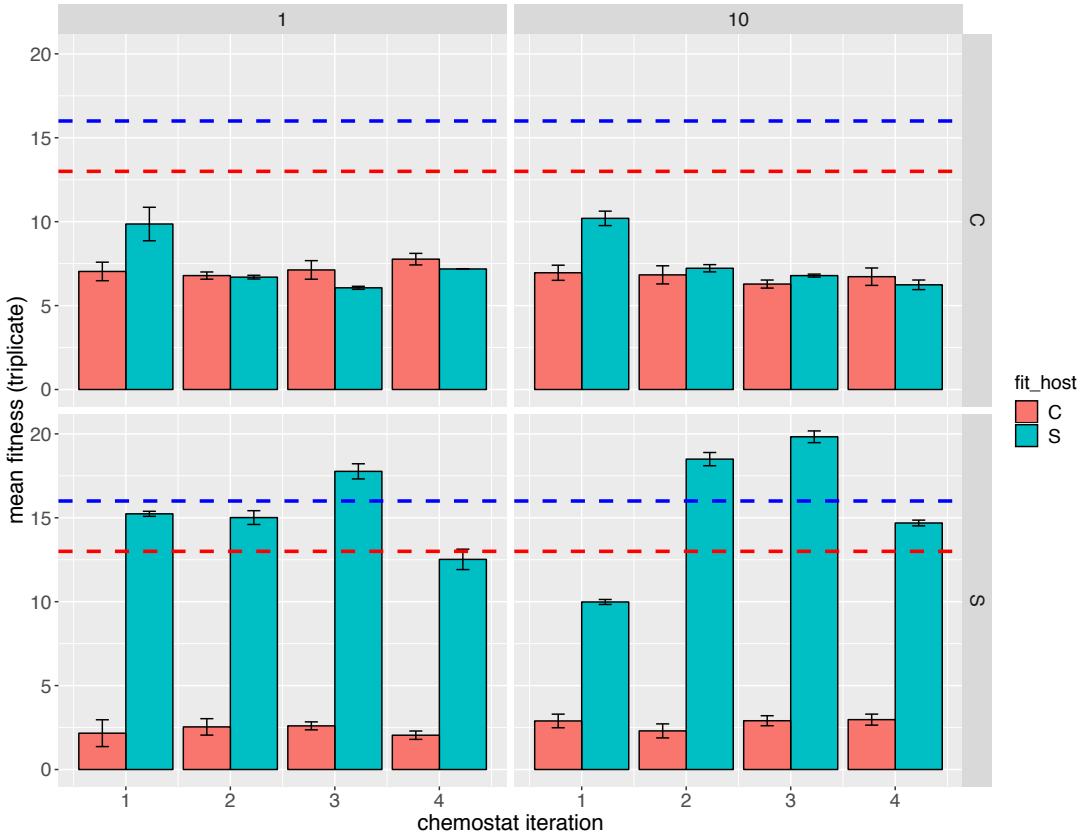


Figure 4.3: Fitness (doublings/hour, number of replicates = 3) of ΦX174 on *E. coli* C and *S.* Typhimurium (indicated as fit_host) after 45 minutes of incubation for S-adapted and C-adapted phage populations. Iterations at days 1 and 10 as indicated by facet headings implies fitness evaluated on day 1 and day 10. Chemostat iterations numbers 1 – 4 implies: C branch (C, CS, CSC, CSCS on *E. coli* C) or S branch (S, SC, SCS, SCSC on *S.* Typhimurium) see section 2.2.4. Ancestral ΦX174 fitness was evaluated on *S.* Typhimurium (blue dashed line) and on *E. coli* C (red dashed line) prior to chemostat selection. Triplicate biological replicates (each derived from triplicate technical replicates) are plotted together with 95% confidence intervals of the means.

Variables	R^2	F value	p value
Days	0.001	1.317	0.250
Last_hosts	0.151	155.849	9.999×10^5
Fit_hosts	0.409	421.324	9.999×10^5
Num_seen	0.001	1.384	0.230
Branch	0.001	1.173	0.280
Days × last_hosts	0.002	1.541	0.211
Days × fit_hosts	0.002	1.688	0.180
Last_host × fit_hosts	0.347	357.076	9.999×10^5
Days × last_host × fit_hosts	0.003	2.670	0.0838

Table 4.2: Summary of non-parametric analysis of variance results. Variables are defined as follows: days are fitness on both hosts on day 1 or 10, last_hosts means the last host of ΦX174 adaptation prior to fitness measurement, fit_host is the host in which fitness was measured, num_seen implies the number of times ΦX174 has been on *E. coli* C or *S.* Typhimurium, branch is the two parallel lines run independent of each other (see section 2.2.4). The analysis was carried out in R using the adonis2 function within vegan package. Default parameters were chosen except that 10,000 permutations were used and the seed was set to 124 before running the function (for reproducibility).

An effect was also seen in relation to the chemostat environment (discussed in the section 4.3.4). Fitness was measured in liquid culture but not in the chemostat environment. The initial ancestral ΦX174 fitness measures on *E. coli* C prior to chemostat selection on *E. coli* C were reduced by almost one half after adaptation to *E. coli* C in the chemostat (~13 to 7 doublings/hour). For *S.* Typhimurium-adapted isolates, initial ancestral fitness measured on *S.* Typhimurium was ~16 doublings/hour, close in value to day 1 of chemostat selection (~15 doublings/hour; figure 4.3).

The results obtained in this study showed that viral adaptation can be reversed by the population when it is switched to a previous environment. Host switching was carried out four consecutive times (twice in each host branch) and each time, the fitness measured on both hosts exhibit a similar pattern. Such reversals in fitness during experimental evolution have been observed in both; laboratory populations with ϕ6 virus (Burch and Chao 1999) and ΦX174 (Crill *et al*. 2000), natural populations in boid snakes (Lynch and Wagner 2010) and frogs (Wiens 2011).

Overall there are consequences of growth on *E. coli* C for chemostat-adapted population across multiple host switches and higher absolute growth rates were achieved on *S.* Typhimurium host. Although there was loss of fitness in *E. coli* C when ancestral phage, A, was propagated in the chemostat, successful reproduction was achieved and viruses persisted throughout the 40 days of adaptation (figure 4.3).

In this study, we have been able to demonstrate that during adaptation to the novel host *S.* Typhimurium, a trade-off is shown where performance on the original host *E. coli* C was reduced in a DNA phage. Observations of similar effects have been seen in some other studies, including with ssDNA viruses. Crill *et al.* (2000) observed a similar trade-off effect for ΦX174 adaptation in *S*. Typhimurium and *E. coli* C hosts. Several studies examined life history trade-offs (for fitness components) in different phages, utilising various environmental stressors (Dessau *et al*., 2012; García-Villada and Drake

2013). In a short-term evolution study that examined the occurrence of trade-off in fecundity and the ability of Qß (an ssRNA phage) to remain viable outside the host, it was revealed that costs occurred in fecundity at the expense of production of viable phage virions, thereby reducing the amount of resources required to produce virions (García-Villada and Drake 2013). A study by Dessau *et al*. (2012) investigated fitness trade-offs of pleiotropic mutations and protein structure utilising phage Φ6 (RNA virus) under an environmental stressor, heat shock, that imposed extreme virus mortality. The study showed that a single amino acid mutation in the viral protein improved stability but at cost of the viral reproduction rate.

Trade-offs in extracellular organisms generally have been investigated (Goldhil and Turner, 2014). West Nile virus, a mosquito-borne virus, responsible for several illnesses including meningitis and encephalitis, was stressed by varying pH, required for conformational rearrangements of glycoprotein during its infection. The result demonstrated that acquisition of a substitution increased acid resistance in the virus; however, a reduction in viral fitness was also observed (MartÍn-Acebes and Saiz, 2010). In another study, amino acid replacement of foot-and-mouth disease virus was shown to increase acid resistance in cells with reduction in plaque size (Vazquez-Calvo *et al*., 2014).

In addition, trade-offs associated with host switching have been demonstrated in macro-evolutionary studies (Truyen *et al*., 1996; Duffy *et al*., 2005). A successful infective virus will likely maximise reproductive ability, and establish long-term reproductive success. Maximising reproductive ability may result in reduction in performance on the original host, and may result in host expansion. In this study, a reduction in growth rate occurred in *S.* Typhimurium (S-adapted, figure 4.3) when fitness was measured on *E. coli* C. The *S*. Typhimurium-measured fitness was higher when other populations (either from another linage or, in some cases, from ancestral or descendent populations) were recently adapted on *S*. Typhimurium. The

same reciprocal pattern of fitness changes was observed with *E. coli* C-measured fitness which was highest in populations recently adapted on *E. coli* C and lower in related or un-related populations adapted on *S.* Typhimurium. The reduction in fitness did not result in reproductive failure (as observed in *E. coli* K-12$^{gmhB\text{-mut}}$). It has been suggested that infectivity loss or reduction in original host may enable virus host range expansion, broadly exploiting and efficiently utilizing host resources to accomplish the goal of evolving as a generalist with clear costs, rather than a specialist (Duffy *et al.* 2005). In a microevolution study, Truyen *et al.* (1996) demonstrated host switching of canine parvovirus from cats to dogs, where CPV-2 virus responsible for the initial observed outbreak in dogs lost the ability to infect the original cat host. Pathogens may accomplish a long-term success in adaptation to a suitable host, even if the performance on the original host would be lost, considering part of the completion of the pathogen life cycle do not depend on the original host. For example, in macroevolution of SIV virus, the virus evolved the ability to establish a viable population in humans to become HIV-I, reduction in performance or inability to infect original host was not important. However, there may be a trade-off between transmission potential and virulence for pathogen that kills the host cells, for instance HIV and even lytic phage.

4.3.3 Phage-host coevolution and phage evolution

For an obligate parasite like phage that the whole or part of its lifecycle depends on the bacterial host it infects, phage interact and evolve in response to its host. In the same way, bacterial host evolve in response to phage infection in a reciprocal coevolution event. In such an experimental system, selection favours emergence of resistance host and phage capable of infecting resistance host via accumulation of mutations. This eventually leads to mutation and counter-mutation cycles arms-race (section 1.3.5.1) where selective sweep may act and resistant hosts and resistance-infective phages replace their wild-type. As coexistence of phage-host continues, resistance breaking phage mutants accumulate facilitated by an arms race.

It is possible to separate the evolution studies of phage from coevolving host by transferring phage to fresh population of bacterial cells, allowing bacterial cells to be held evolutionary constant while phage continues to evolve as described in this study (section 3.3.1.4.1). Although, evolution of bacteria hosts may occur but in the present study, efforts were made to minimise host evolution as much as possible. In experimental designs, results obtained from coevolution studies are usually different from an evolution studies with respect to phages. As an example, Paterson *et al*. (2010) propagated phage Φ2 and its laboratory host *P. fluorescens* under two separate conditions; bacterial host evolution was held constant while Φ2 adaptation continues and coevolution where both *P. fluorescens and* Φ2 were allowed to evolve. The results obtained in Paterson *et al*. (2010) study showed the difference in genetic diversity of Φ2 evolved isolates; coevolved Φ2 populations exhibited larger genetic distance, more mutation sites and greater genetic divergence when compared with Φ2 ancestor than evolved isolates.

Because recognising a susceptible host and adsorption is the initial process of phage infection (figure 4.1), most common mode of phage infectivity is by mutations in genes required for host adsorption (Meyer *et al*., 2012; Scanlan *et al*., 2011 and Paterson *et al*., 2010). Most often, these phage mutants may retain their ability to infect their original host but in some cases may result in fitness cost on the original host (Crill *et al*., 2000). As such, high degree of diversity may arise in the population with phage mutants possessing varying degrees of growth rates. For instance, in Schwartz and Lindell (2011) study, T7-like cyanophages (tailed-lytic phages) overcame resistance in *Proclorococcus* (Cyanobacteria) via mutation in tail genes required for attachment. This is further discussed in chapter 5.

4.3.4 A potential effect of the chemostat environment

The chemostat culturing system was developed to create an environment with a continuous supply of host cells for ΦX174. The environment was designed to be stable and unchanging across the experiment except for the

host being used, therefore, adaptation to the host was imposed change in this study. Ancestral ΦX174 fitness was calculated before evolution in chemostat culturing system. Results show that ΦX174 fitness (measured on *E. coli* C) was lower than ancestral fitness (measured on *E. coli* C) during the experiment. This was also true for fitness measured on *S.* Typhimurium (except CS10, SCS1, and SCS10). One environment for ancestral ΦX174 was the previous environment prior to chemostat transfer (from Wichman's laboratory, section 2. 1.2) and another environment was our chemostat system. Ancestral ΦX174 fitness measured on *E.coli* C prior to chemostat transfer was different from fitness after day 1 measured on the same host (*E. coli* C). It may be deduced that ΦX174's environment is affected not only by the type of host but also by culture conditions. In a population, both environmental and ecological factors (including culture conditions, not evaluated in this study) influence phage adaptation. In chemostat environment, phage-host may co-infect with co-infection selecting for the ability to compete with different viral genomes (Koskell and Brockhurst, 2014). Other evolutionary processes that may influence phage evolution are coexistence and mutation discussed in sections 1.3.5, 1.4 and 4.3.3. So overall, viral infection and evolution is complex and a function of biotic and abiotic environmental factors (although the limiting resource in the chemostat was unknown).

Successful adaptation may be determined by properties of the viral population and how these interact with the immediate environment. If a successful adaptation is driven by co-adaptation initiated by the fitness costs of phenotypes (Sabrina *et. al.*, 2015), and successful adaptation, in turn, depends on the environment, then, the extent of trade-off may vary with the culture environment (such as chemostat versus serial transfer). As an example, a study by Bohannan *et al*. (2002) examined trade-offs and coexistence in *E. coli* cells and T4-phage infection and found that the magnitude of trade-offs is determined by genetic and environmental factors. The authors compared trade-offs between competitive and resistance abilities in a chemostat culture and in batch culture for T4 phage in *E. coli*

cells (environmental change inclusive of glucose utilisation of the bacterial cells), and found that fitness was higher in batch culture than in the chemostat system. This is consistent with the observation described here of an initial high fitness of ancestral ΦX174 prior to chemostat adaptation.

The data in this study do not capture the pattern expected in an adaptive walk in a constant environment. It is expected that fitness should increase rapidly early in adaptation process and then plateau (Gerrish and Lenski, 1998; Good *et al*., 2017). Comparing ΦX174 fitness between days 1 and 10 on both hosts throughout the alternating switching period, there was no substantial difference (ANOVA, $p = 0.250$, table 4.2) in phage growth rate (section 4.3.2).

If assumed that the chemostat design does not select for faster phage growth, could the system be driven by competitive interactions? Competitive interactions can occur internally, via viral genome competition within host cells or externally, via access to host cells (Dennehy, 2014). In the *E. coli* C-adapted populations, competition may have occurred, with co-infection selecting for the ability to compete with different viral genomes. Such selection may profoundly affect growth dynamics, causing viruses to evolve a reduced potential to grow when hosts are abundant, resulting in decreased fitness (Turner and Chao, 1998). Although, decline in fitness was observed in the chemostat, the competing viral progeny may stably coexist. On the other hand, in *S.* Typhimurium-adapted populations the decrease in fitness between ancestral fitness and chemostat fitness was less marked and, in some cases, the fitness of chemostat-adapted viruses was higher than ancestral fitness. This does not imply the absence of competitive interactions. It may be that the most rapidly growing viral progeny out-compete slower growing populations, resulting in population with a high growth rate. Viral competition and co-infection is often shaped by the environment bacteria inhabit, available resources, and interactions among surrounding viruses (Kneitel, 2009; Diaz-Munoz, 2017), which in turns affect fitness in the global environment. Overall, within-host viral genotypes and

intraspecific competition are also expected to drive growth rate dynamics in different environments.

Another factor for consideration in describing overall growth dynamics in the chemostat system was the bottlenecks imposed. These were derived from two process: the re-inoculation of each chemostat after 2 days of growth (within each 10-day growth period), and the single plaque transfer from the population on one host to the next period of selection on the alternative host (section 2.2.4). The plaque transfer bottleneck is the more severe of these two bottlenecks. A bottleneck is also an intrinsic feature of the fitness measurement protocol where a small fraction of the population was transferred to fresh bacterial cells for evaluating fitness (section 2.3.2). Population bottlenecks, an inherent feature in the life cycle of many pathogens, are constituted by substantial reductions in the size of a population. These occur in nature and have been shown to have implications for adaptation. This is particularly the case for extreme bottlenecks occurring during pathogen transmission from host to host (Abel *et al*., 2015; Leonard *et al*., 2017; LeClair and Wahl, 2018). Adaptive walks during host switching depend on the severity of the bottleneck imposed (LeClair and Wahl, 2018), and, when extremely severe, a bottleneck may have negative effects on adaptation. In some cases adaptation may no longer be possible, leading to extinction. Population bottlenecks may contribute to virus extinction during *E. coli* K-12$^{gmhB\text{-mut}}$ adaptation. In this experiment, a bottleneck was imposed artificially by selecting a representative plaque at random. It is possible that a low fitness phage was selected, subsequently affecting the reproductive success of the following generations. Novella *et al*. (1995) demonstrated that selecting for lower fitness genomes causes gradual virus elimination. In addition, bottlenecks have an effect on overall fitness of organisms, driven by random sampling and selection. Duarte *et al*. (1992) showed a reduction in fitness following bottlenecked passages of vesicular stomatitis virus, the most severe fitness loss associated with passages on a new host. They also recorded no fitness change, or statistically insignificant changes, in some passaged clones. During host switching adaptation in this study, there were

no observed significant differences between ΦX174 fitness changes for days 1 and 10 (section 4.3.2; figures 4.3). As earlier mentioned (section 2.2.4), bottleneck was artificially imposed within-host cultures (after 2 days of growth) leading to small reductions in population size during the 10 days adaptation period. Bottleneck in-between hosts (alternate host-switching) introduces initial ΦX174 low population. These effects (bottlenecks; within-host and in-between hosts) may reduce the rate of adaptation.

The conclusion from the results for growth rates in this study are similar to a study conducted by Crill *et al.* (2000) on evolutionary reversal with ΦX174 phage. However Crill *et al.* (2000) observed higher growth rates on *E. coli* C than on *S.* Typhimurium and near zero (some at zero) growth rates observed after selection on *S.* Typhimurium for the *E. coli* C host. This difference may arise from chemostat dynamics, flow rate and dilution rate, which affect the physiological state of host in the chambers. The chemostat used for Crill *et al.* (2000) study had a lower mean residence time and a higher flow rate (Bull *et al.*, 1997) compared to this study design. This may cause differences in the hosts' intrinsic abilities to support phage growth.

4.3.5 Attachment rate of evolved populations

During infection ΦX174 follows a typical lytic life cycle, finding a suitable host receptor, attaching to the host, injecting its genome into the host's cytoplasm, synthesizing its proteins, assembling progeny virions and finally lysing host cells leading to the release of mature viruses. Of great importance in the ΦX174 life cycle is finding and attaching to a host, which is a pre-requisite for successful infection.

In this study, a trade-off exists when the fitness of ΦX174 populations, recently selected on the novel host *S.* Typhimurium, was measured on *E. coli* C. We can ask whether the rate at which ΦX174 adsorbs to *S.* Typhimurium explains these observations. To investigate this, the attachment rate was quantified in solution using qPCR on both hosts. Attachment assays were initially performed to determine the appropriate time period after which most

ΦX174 would have been adsorbed onto bacterial cells. As shown in figure 4.4, the maximum attachment rate of phage was between 8 and 10 minutes after incubation, measured as cell per minute. Hence, rates of attachment were calculated after 8 minutes in exponentially growing bacterial hosts, measured as per cell per minute. To further clarify that ~ 8 minutes was the optimal time period, where ~99% of ΦX174 have attached to host cells, a mutant strain of *E.coli* C (*gro89*: a mutation in the rep protein) was employed. The *gro89 E. coli* allow attachment of ΦX174 but exclusively inhibit stage III DNA synthesis, resulting in an unassembled genome and no progeny. This enables the determination of ΦX174 PFU counts without these being conflated with progeny that might have emerged from early-lysing cells (figure 4.5; note that negative "replication rates" on wildtype *E. coli* C in this figure are presumed to arise from viruses emerging from early-lysing cells). This procedure was used to identify the time period of maximal phage attachment. At 8 minutes, 98.56% phage were adsorbed and by 26 minutes 100% of ΦX174 in the solution had been adsorbed (figure 4.5).

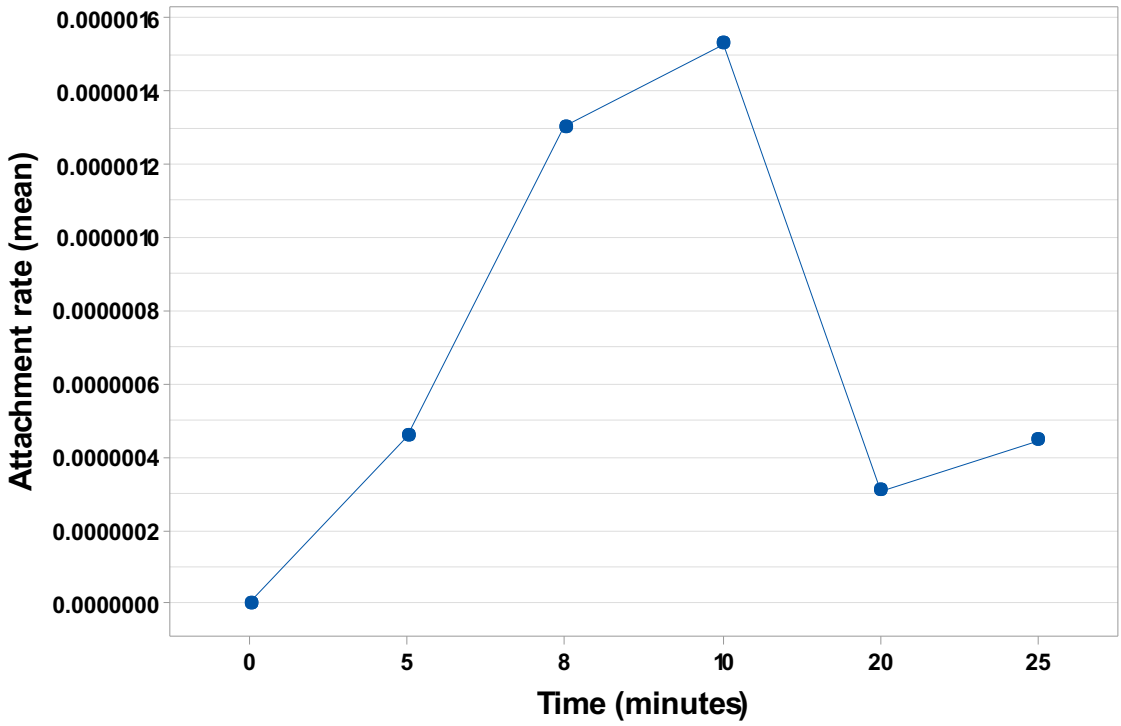


Figure 4.4: Preliminary attachment time course on *E. coli* C, with maximal attachment observed between 8 and 10 minutes (MOI < 0.1, number of replicates = 3). Attachment rate of phage expressed as per cell per minute.

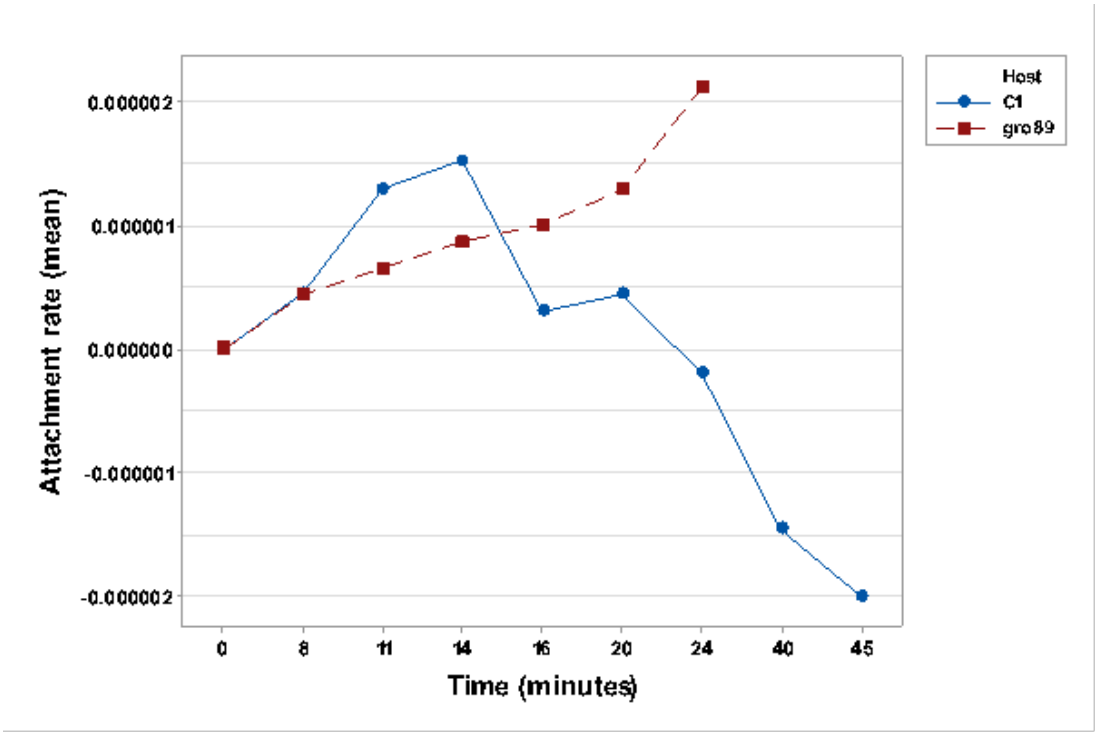


Figure 4.5: Attachment rate (mean of 3 technical replicates) on both *E. coli* C and *gro89*, showing optimal attachment rate at ~ 8 minutes (MOI < 0.1). 100% ΦX174 were adsorbed by ~ 26 minutes, may be an indication of one ΦX174 generation. Attachment rate of phage expressed as per cell per minute.

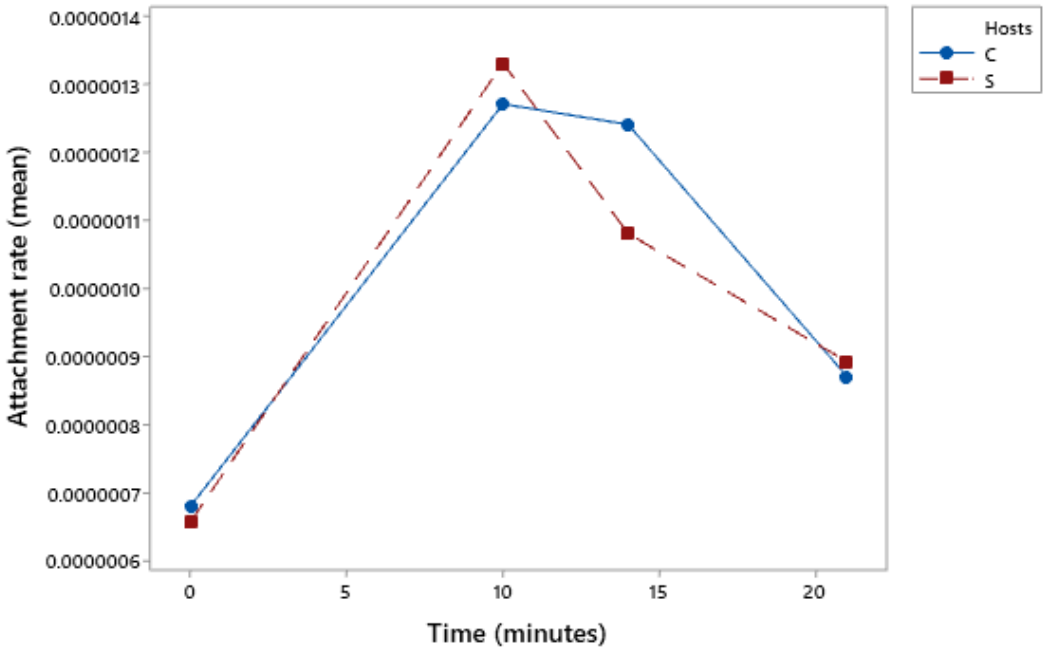


Figure 4.6: Preliminary attachment (mean of 3 technical replicates) time course on *E. coli* C (blue line) and *S.* typhimurium (red line) with maximal attachment observed between 8 and 10 minutes (MOI < 0.1). Attachment rate of phage expressed as per cell per minute. Data was obtained from a different experiment.

The attachment rate experiment was performed for chemostat-adapted phages at days 1 and 10 on both recent and previously encountered host strains. The last host encountered has no significant effect on attachment rate (ANOVA, $p = 0.062$), and on significant interactions between days and last host of adaptation was identified (ANOVA, $p = 0.164$). significant effect was observed on days (1 vs 10) of adaptation (ANOVA, $p = 0.002$). The attachment host for ΦX174 selection has a very highly significant effect (ANOVA, $p<0.0001$, table 4.3)

As with the fitness study, a statistically significant interaction was observed between the effect of the last host encountered and the effect of the host on which attachment was measured (ANOVA, $p=0.005$). However, this interaction was not as marked as that observed for fitness measures (ANOVA, $p < 0.0001$ for fitness, contrast figures 4.3 and 4.7). Although, for example, higher rates of attachment (measured on *S.* Typhimurium) were observed for ΦX174 recently selected on *S.* Typhimurium, it may be that changes in attachment rate between hosts do not fully account for fitness changes.

Overall attachment rates for *E. coli* C-adapted and *S*. Typhimurium-adapted ΦX174 measured on both hosts appeared to exhibit similar patterns as observed in their respective fitness graphs. There is possibility that rates of attachment were more biologically significant than observed or that, as one of the factors determining trade-offs in the system, it could be measured on a more sensitive scale beyond the experimental assay design. It is possible that the higher the rate of attachment, the higher the chances of replication, hence the higher the fitness. However, changes in attachment rate may not be the only factor determining increases or decreases in fitness. Productivity, size and density of both hosts and phage are also plausible determining factors (Roychoudhury *et al*., 2013). High attachment rate may even be a detrimental in certain circumstance, for example Gallet *et al*. (2009) showed that high attachment rate of lambda phage in a biofilm environment was deleterious to the phage fitness.

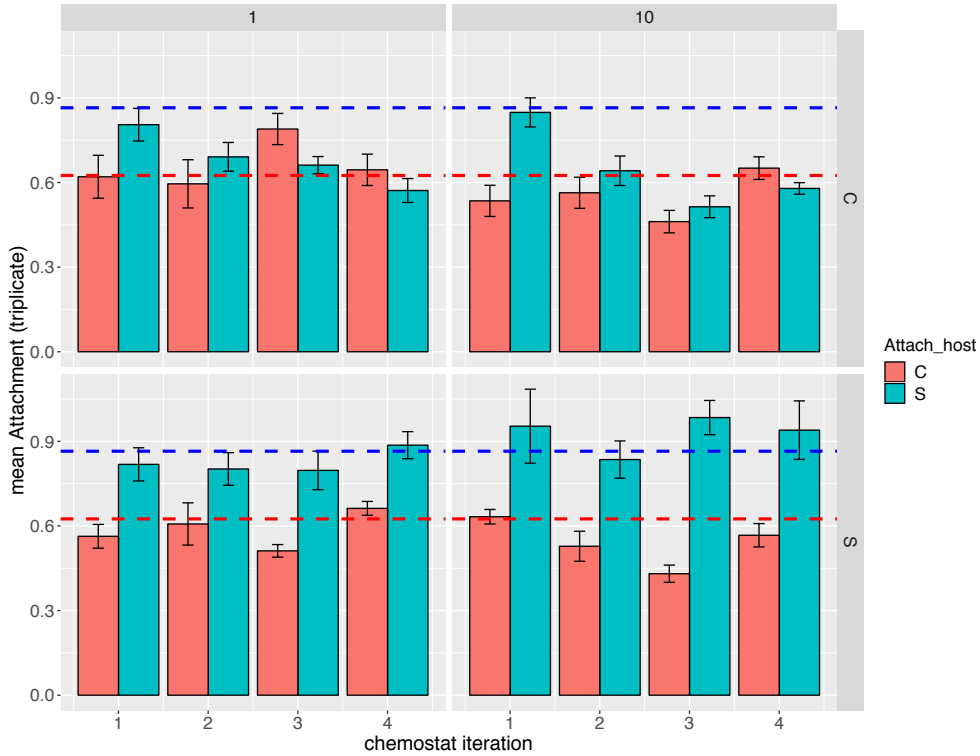


Figure 4.7: Attachment rate of ΦX174 (/cell/minute, mean of 3 replicates) on *E. coli* C and *S.* Typhimurium (indicated as fit_host) after 8 minutes incubation for S-adapted and C-adapted phage populations. Iterations at days 1 and 10 as indicated by facet headings implies fitness evaluated on day 1 and day 10. Chemostat iterations numbers 1 – 4 implies: C branch (C, CS, CSC, CSCS on *E. coli* C) or S branch (S, SC, SCS, SCSC on *S.* Typhimurium) see section 2.2.4. Ancestral ΦX174 fitness was evaluated on *S.* Typhimurium (blue dashed line) and on *E. coli* C (red dashed line) prior to chemostat selection. Triplicate biological replicates (each derived from triplicate technical replicates) are plotted together with 95% confidence intervals of the means.

Variable	R^2	F value	p value
Days	0.059	7.256	0.002
Last_host	0.023	2.830	0.062
Attach_hosts	0.102	12.572	9.999×10^{5}
Num_seen	0.014	1.773	0.185
Branch	0.011	1.375	0.293
Days × last_host	0.015	1.906	0.164
Days × attach_hosts	0.013	1.654	0.211
Last_host × attach_hosts	0.044	5.378	0.005
Days × last_host × attach_hosts	0.021	2.615	0.076

Table 4.3: Summary of non-parametric analysis of variance results. Variables are defined as follows: days are attachment on both hosts on day 1 or 10, last_hosts means the last host of ΦX174 adaptation prior to attachment measurement, fit_host is the host in which attachment rate was measured, num_seen implies the number of times ΦX174 has been on *E. coli* C or *S.* Typhimurium, branch is the two parallel lines run independent of each other (see section 2.2.4). The analysis was carried out in R using the adonis2 function within vegan package. Default parameters were chosen except that 10,000 permutations were used and the seed was set to 124 before running the function (for reproducibility).

4.4 Conclusion

In conclusion, like all viruses, bacteriophages must encounter and infect suitable host cells for successful reproduction. The ability to infect large numbers of hosts increases the likelihood that phage will encounter a permissive host. However this must be married to an ability to productively infect a host and attain reproduction rate sufficient for persistence in that host. As such, the breadth of host range may be the most important factor for long-term persistence. Selection for the fittest phenotypes may come at a cost for phage requiring trade-offs in less fit phenotypes as discussed (sections 4.3.2). A similar experimental approach to that used here may uncover insights into new emerging diseases based on host shifts. This information may also be helpful for understanding drug resistance mechanisms, for exploring ways to prevent future development of drug resistance and, more importantly, for assessing the impact of trade-offs that occur on acquisition of resistance to drugs. Adaptation to host environment is not only governed by the host alone, but also by the global culturing environment and its dynamics including phage-phage competition.

The results of the attachment rate assay cannot conclusively show that the observed trade-offs were a consequence of variation in attachment rates and the last host of adaptation, therefore it would be helpful to identify genotypic factors that may be responsible for the phenotype changes observed. It is possible to examine the influence of genotype on the magnitude of trade-off during host-switching, and also consequences of environmental trade-offs on genotype.

The results show that fitness reversals occurred every time ΦX174 was switched back to the alternative host. Studies have confirmed that whether switching occurs instantly, with adaptations taking years (Teotonio and Rose, 2000), months (Lenski 1988), or days (Burch and Chao, 1999; Crill *et al.*, 2000) or if environmental switching occurs slowly (Tan and Gore, 2012),

evolutionary reversals do eventually occur. However, it would be of interest to understand genetic factors that determine this reversibility. This could be carried out by identifying reproducible changes in allele frequency in ΦX174 populations that correlate with the current host to which a population is exposed.

Chapter Five: Deep sequencing of ΦX174

5.1 Introduction

5.1.1 Mutating to adapt to host environments

Organisms must cope with fluctuating environments including that of the host on which they proliferate. When organisms are introduced to novel environments or hosts, the ability to grow rapidly and survive depends on their adaptive potential. Adaptation to novel hosts is a typical example of the evolutionary phenomenon of adaptation and invasion of a new ecological niche. A new host may create challenges for viral entry, replication or lysis. Overcoming these barriers depends on the likelihood of viruses acquiring the necessary set of mutations for infection in the new host.

While adaptation to novel hosts can entail fixation of several mutations, a single mutation may allow switching to a new host and enhance adaptation in it. For example, in the 1990s, Venezuelan equine encephalitis virus switched from rodents to horses (with efficient replication in the latter host) through a single mutation (Anishchenko *et al*., 2006). Likewise, a single mutation in an RNA virus enhanced host range abilities and ease of host switching (Duffy *et al*., 2007).

In contrast, multiple mutations may be required for new host colonisation. For instance, five mutations were shown to cause influenza A virus, subtype H5N1, transmission between ferrets (Linster *et al*., 2014). Several mutations on DNA phage (SBW25Φ2) of *Pseudomonas fluorescens* allowed rapid infection of some hosts (Hall *et al*., 2011). In situations where multiple mutations are required for adaptation, identifying the types of mutations that arise, and the ability of beneficial mutations to be fixed are important to understand the evolutionary dynamics of colonisation. Some questions that can be addressed include, whether mutations are adaptive in all genetic backgrounds or whether mutations fix simultaneously or in a particular order (Longdon *et al*., 2014). Understanding the nature of mutations, identifying the phenotypic effect and dynamics of accumulations of adaptive mutations are

paramount as well (Lang and Desai, 2014). Answers to these questions may improve our understanding of the molecular basis of adaptation in host environments. Whole-genome sequencing and deep sequencing methods have played a key role in identifying the genetic basis of adaptation.

5.1.2 Theories of adaptation

Adaptation differs from natural selection. Natural selection arises from differences in the survival and reproduction capability of individual organisms with differences in phenotypes attributable to heritable nucleotide changes. Adaptation represents the change in composition of a population of organisms towards phenotypes that best fit the immediate environment (Orr, 2005a). Theorists have tried to model adaptation through various phenotype- and genotype-based models. In Fisher's geometric model, optimum fitness arises from the best combination of traits values in the particular environment. As a result of a recent change in the environment for a population, mean trait values in the population may be far from the optimum. Adaptation may result in a population moving towards the new optimum as much as possible. This can be achieved through the acquisition of random mutations some of which lie closer to the new trait value optimum. Mutations that bring trait values closer to the optimum are to be favoured over mutations that resulted in movement further away. Equivalently, mutations that contribute to adaptation must be favourable (but they must also escape accidental loss when rare, which will be the case for newly arising mutations).

Fisher's model indicates that mutations can lead to different phenotypic effect sizes (with fitness effects emerging from size and direction in phenotypic space) and the model suggests that mutations of small effect are most likely to contribute to adaptation (Orr, 2005b). However Motoo Kimura's model, which incorporates time, shows that intermediate size mutations probably also play a role in adaptation. John Gillespie examined evolution in

protein sequences in his landscape model and observed that an adaptive walk towards an optimum is exponential in terms of the distribution of fitness effects of fixed mutations. However, biological scenarios involving moving optima were not considered (and may be captured in fitness landscape models). Impediments to adaptation include clonal interference, epistatic and pleotropic interactions, competition and recombination. The relative importance of these factors can be addressed through experimental work, leading to a situation in which theoretical work on adaptation lags behind the empirical (Orr, 2005b).

Through experimental evolution, the biological scenarios listed may be revealed. Advances in molecular methods for experimental evolution, varieties of model systems and high-throughput next-generation sequencing technologies make it possible to identify, track and quantify mutations in several replicate populations over many generations.

5.1.3 Sequencing technologies

Since the development of the ground-breaking sequencing method by Fred Sanger and his colleagues (Sanger *et al.*, 1977; Sanger *et al.*, 1978), over forty years ago, there has been a rapid improvement, triggered by improvements in methods. This has provided great opportunities for fast, robust and low-cost DNA sequencing.

Through first-generation Sanger sequencing, it is possible to sequence a short stretch of DNA using chain termination based on random incorporation of dideoxyribonucleoside triphosphates (ddNTPs; Metzker, 2005). However, this is time consuming and the original method entailed risk caused by radioisotopes used for ddNTP labelling. After some methodological improvements, an automated DNA sequencer, the Prism from Applied Biosystems (ABI), was introduced in 1996. Subsequent improvements in molecular biology methods and new developments in almost all branches of

science gave rise to the next phase of DNA sequencing (Ari and Arikan, 2016).

DNA sequencing underwent drastic changes in the 2000s with the introduction of next-generation sequencing, sometimes referred to as second-generation sequencing. These sequencing methods used either sequencing by ligation or sequencing by synthesis. Several sequencing platforms are available (at the time of writing) including prominent platforms from Illumina such as MiSeq, Nextseq 500, Hiseq 2500 with differences in the spectrum of applications, output yields and the read lengths delivered. Illumina sequencing platforms utilise the sequencing-by-synthesis method. The procedure begins with DNA extraction and moves on to library preparation. Library preparation entails fragmentation of input DNA into multiple short fragments that are tagged with specific adapters and indexes. Following this, tagged DNA fragments are then amplified, loaded and sequenced on a flow cell (containing immobilised primer sequences complementary to adapters in the DNA library). An *in situ* amplification process generates clusters on the flow cell via bridge amplification. Clusters are switched over the other end to achieve opposite strand amplification. Thereafter, sequencing by synthesis occurs during which fluorescently tagged deoxynucleoside triphosphates (dNTPs) are added to the DNA strand and at the same time, the platform records which base was added through the unique emissions of each of the four bases (on excitation). Each time a fluorophore-labelled nucleotide is added it results in chain termination, but after detection, a chemical reversing agent is added. DNA fragments are therefore sequenced in multiple cycles from the 5' end of each strand prior to switching over, to a given length, ranging from 75 cycles or bases to 300 cycles, depending on the reagents employed.

After sequencing, demultiplexing, the process of dividing the sequence reads into corresponding files for each tagged sample, is carried out, generating FASTQ files for read pairs. The paired reads generated can then be checked

for quality in order to select the highest quality reads. Following quality control, reads may either be assembled (to reconstruct genomes *de novo*) or aligned to a reference genome (for re-sequencing or deep sequencing). In deep sequencing the number of reads mapped to the same part of the reference genome is critical and this is referred to as the coverage.

In comparison with Sanger sequencing, Illumina sequencing and similar second-generation techniques increase DNA sequencing throughput, save time and, in the case of Illumina sequencing, are reported to have a consensus accuracy of 99.9 % (Morey *et al*., 2013). This technique comes with disadvantages, including biased GC representation introduced during sequencing that may lead to uneven coverage (Chen *et al*., 2013). substitution errors may occur as a result of noise background in each sequencing cycle (Hutchison, 2007) and sequence biases (that is errors that are sequence-specific) may result from library preparation steps including the PCR amplification procedure (Pienaar et al., 2006).

A third generation of sequencing technologies has arisen more recently which addresses some of the limitations of second-generation sequencing systems, although accuracy is still a challenge for these systems. Third-generation technologies are mostly based on direct detection of nucleotide sequences without amplification thereby reducing the complexity of library preparation. These sequencing technologies include Nanopore sequencing (section 2.5.4.1). Nanopore technology sequences DNA based on translocation through nanometre-sized pores propelled by a potential difference across the membrane in which the pores are embedded. Sequences are inferred from changes in current induced during this process (Ambarda *et al*., 2016; Rusk, 2014). Advantages of this system include, its small size, ease of access (ideal for in-field use as well as in the laboratory), generation of (ultra-)long reads, relative affordability and reduced biases from DNA library preparation. However, the error rates (after base calling) are quite high (38.2 %; Laver *et al*., 2015).

Second-generation sequencing technologies are currently the most commonly used sequencing platforms. They have many applications for addressing varied scientific questions. These platforms' greater accuracy makes them the currently best choice for deep sequencing and therefore of use for tracking variation in populations during experimental evolution.

5.2 Aims and objectives

When ΦX174 was host-switched intermittently four consecutive times, it was discovered that fitness trade-offs were repeatedly manifested (section 4.3.2). In order to determine the genotypes responsible for the fitness costs observed, this chapter uses a deep-sequencing approach. The aim is to use the detection of nucleotide changes to identify mutations that may underlie the observed phenotypic and fitness changes. By analysing different genotypes produced as evolution occurs, we can provide a detailed view of the changes that occur when a parasite adapts to a new host. Using deep sequencing, the allelic variants along with their nucleotide frequencies in each population sample were determined and related to host-switching events. Deep sequencing does not just determine a nucleotide sequence change but reconstructs a population's sequence variation (Goldman and Domschke, 2014). The challenge for this method is in error-rate determination, since per-read error rates are higher than consensus error rates (which are reduced by averaging over reads with unbiased errors). There may also be biases introduced during library preparation and concerns related to whether DNA prepared from a population is representative.

The main objectives of this chapter were to:

- Obtain high-resolution deep sequence data of ΦX174 populations evolved through alternating host switching at multiple time points.

- Assess repeatability of experimental measures at the level of DNA extraction and sequencing.

- Detect genotype changes associated with adaptation to new hosts.

- Determine the likelihood of host-specific mutations associated with the two hosts, *E.coli* C and *S*. Typhimurium.
- Track genetic changes through time series during adaption of ΦX174 in the two hosts utilised.

5.3 Results and discussions

5.3.1 Viral samples

Ancestral phage ΦX174 used to initiate this experiment was isolated from a single plaque, then cultivated in the *E. coli* C host strain until the titer was sufficiently high (~ 10^6 pfu/ul, for 3 hours; section 2.2.3). Therefore, starting phage was pre-adapted in *E. coli* C. Illumina sequencing of the cultivated ancestral phage was done to ascertain variation putatively present (or arising) in the wildtype phage. The results showed that mutations were acquired by the starting phage (compared with GenBAnk sequence of ΦX174 accession number AF176034) as shown in table 5.1. The pre-adapted ancestral phage was adapted in either *S.* Typhimurium or *E. coli* C for 10 days for four consecutive times as described in chapter 2 (section 2.2.4).

Site	Nucleotide change	Protein(s)	Amino acid change	Radicality	Approximate allele frequencies
648	C – G	D(E)	H86D(F26L)	RAD(CON)	1%
944	G – T	J	V32L	CON	2%
1307	T – C	F	Y102H	RAD	1%
1460	C – A	F	Q153K	RAD	57%
1956	T – G	F	V318G	CON	1%
2275	G – A	F	M424I	CON	93%
2321	T – C	INT	INT	INT	2%
2971	C – T	H	A13V	CON	1%
3339	G – A	H	D136N	RAD	96%
5360	A – C	A/A*(B)	I459I/I287I(K95Q)	SYN/SYN(RAD)	1%

Table 5.1: Allelic variants in the ancestral wild-type ΦX174 used to initiate the experiment, showing alleles with frequencies ≥ 1 % in comparison with GenBank sequence AF176034. Amino acid changes are categorised as CON: conservative, RAD: radical or SYN: synonymous. Mutations 648G and 5360C affect overlapping genes (with the alternative reading frame indicated in parentheses). The mutation at 5360 occurs in genes A and A* (located in the same reading frame), as well as in the alternative reading frame of gene B.

5.3.1.1 Nomenclature of viral samples

Two lines were maintained in evolution of ΦX174 on *E. coli* C and *S.* Typhimurium (figure 2.1). For clarity in describing results the history of host exposures of each lineage or branch of the experiment is described.

C-branch was exposed to:

1. *E. coli* C (= first iteration on C), LABEL: C,
2. switched over to *S.* Typhimurium (= first iteration on S), LABEL: CS,
3. returned to *E. coli* C (= second iteration on C), LABEL: CSC,
4. returned to *S.* Typhimurium (= second iteration on S), LABEL: CSCS.

S-branch was exposed to:

1. *S.* Typhimurium (= first iteration on S), LABEL: S,
2. switched over to *E. coli* C (= first iteration on C), LABEL: SC,
3. returned to *S.* Typhimurium (= second iteration on S), LABEL: SCS,
4. returned to *E. coli* C (= second iteration on C), LABEL: SCSC.

Two sequencing runs were performed with two different kits: Nextera XT and Nextera Flex (section 2.5.3.1 and section 2.5.3.2, respectively, table 5.2). Phage dsDNA was extracted from samples from days 1 and 10 (for every adaptation period listed above) and for days 2, 3 and 8 for both S and SCSC periods from S-branch. DNA extraction was carried out on the host most recently encountered by the population/sample.

Additional dsDNA was extracted for samples SCSC day 10 providing technical duplicates within the Nextera XT run. These were labelled SCSC10**A** and SCSC10**B**. Additional extractions/technical duplicates were also undertaken within the Nextera Flex run for SCSC day 2. Two of these Nextera Flex technical duplicates were labelled 2NDSCSC2**A** and 2NDSCSC2**B**. A third Nextera Flex replicate, from SCSC day 2, was extracted on the other host, *S.* Typhimurium and was labelled 2NDSCSC2**S** (cross-host DNA preparation control).

One sample within the Nextera XT run (viz., S1) showed results incongruous with other *S.* Typhimurium-adapted phage, so this sample was extracted again (via a new DNA preparation) and sequenced on the Nextera Flex run (2NDS1; with the original S1 excluded from the analyses presented in this thesis). An additional two cross-run repeats were carried out with samples S2 and SCSC3 (labelled 2NDS2 and 2NDSCSC3, respectively), in the Nextera Flex run. Sample 2NDSCSC3 failed to provide data and is excluded for this reason. dsDNA was sequenced using the Illumina platform, and output data was analysed as described in section 2.6.2.

A

A	C 1	C 10	SC 1	SC 10	CSC 1	CSC 10	SCSC 1	SCSC 2	SCSC 3	SCSC 8	SCSC 10	S 1	S 2	S 3	S 8	S 10	CS 1	CS 10	SCS 1	SCS 10	CSCS 1	CSCS 10

B

2ND SCSC 2A	2ND SCSC 2B	2ND SCSC 2(S)	2ND SCSC 3	2ND S 1	2ND S 2

Table 5.2: Lists of samples sequenced with **A**- Nextera XT kit and **B**- Nextera Flex (represented as 2ND). SCSC 10A, SCSC 10B indicate repeated samples sequenced with Nextera XT. SCSC2A and SCSC2B are repeated samples sequenced with Nextera Flex. The repeated samples sequenced were treated differently during DNA extraction. SCSC2S represent SCSC2 sample with DNA extraction performed in *S*. Typhimurium.

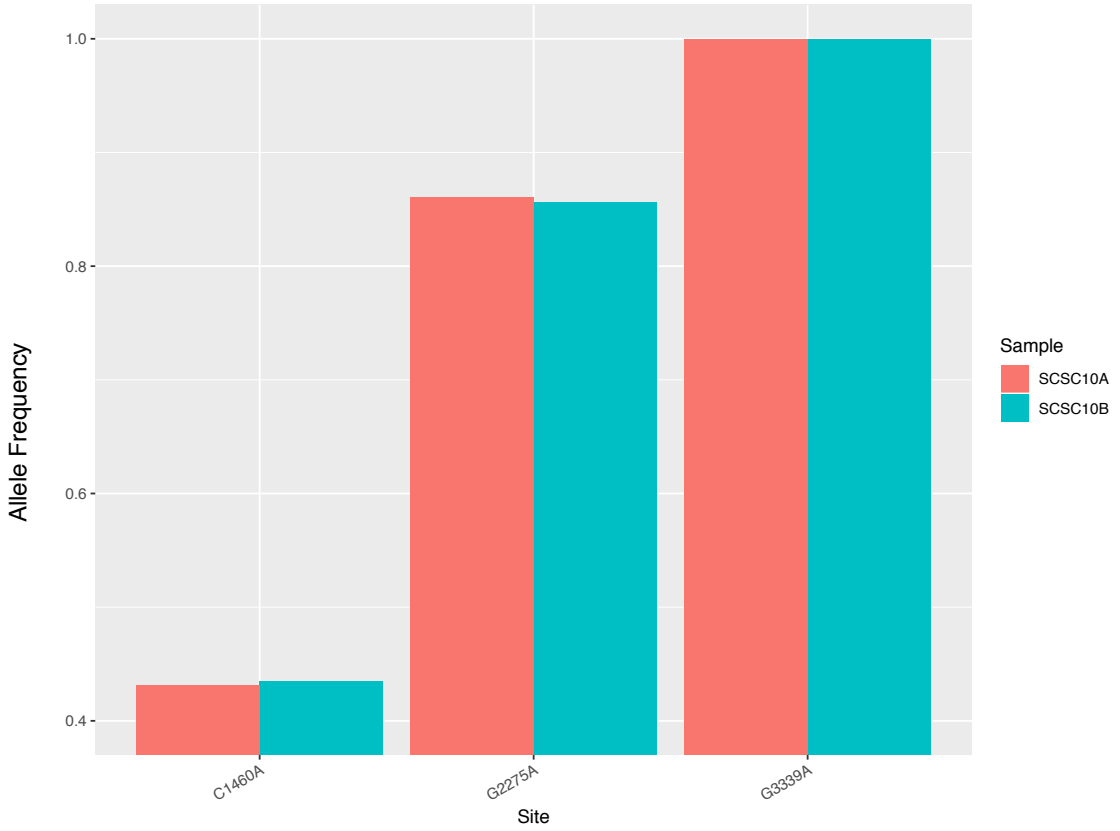
5.3.2 Controlling for sequencing biases

The advent of affordable and easy-to-use NGS technologies has contributed to the recent growth of population genomics (Andrews and Luikart, 2014). Whilst these technologies allow understanding of biological diversity, revealing the genetic bases of rare genetic disorders, they come with their own sources of method-specific error (e.g., PCR biases in amplification steps) and human error (e.g., cross-contamination, DNA preparation errors or misinterpretation of data). These errors may impair medical and scientific applications of the technology. The most commonly used NGS platforms from Illumina includes DNA preparation, library preparation, amplification and sequencing steps (section 5.1.3). In this experiment, several steps were introduced to control for sequencing biases as much as possible.

5.3.2.1 ΦX174 genome dsDNA preparation controls

The dsDNA preparation controls were included because the ΦX174 genome consists of ssDNA and Illumina sequencing technology requires dsDNA. ΦX174 ssDNA was converted to dsDNA using a method adapted from Godson and Vapnek (1973). During ΦX174 DNA replication, ssDNA is converted into dsDNA RF (section 1.6.1.3) utilising host proteins, prior to protein synthesis (Fane *et al*., 1988). The ß-lactam antibiotic chloramphenicol inhibits protein synthesis, therefore allowing the accumulation of RF dsDNA during the DNA replication stage of its lifecycle as described in section 2.5.2.1 (Godson and Vapnek, 1973). To test whether some alleles amplified more than others during accumulation of RF dsDNA, within-run replicates for DNA preparation were set up. This involves repeated sample DNA preparation labelled with suffix A and B for sample individuation, for example SCSC10A and SCSC10B (section 5.3.1.1) with prepared DNA entering the same multiplexed library preparation procedure. The results from the first Nextera XT sequencing run for these samples

showed that allele frequency estimates were very similar (figure 5.1) with the largest absolute difference in allele frequency being ~1 % (appendix B).



Figures 5.1: Allele frequency at high-frequency sites (sites with frequencies greater than 40 %) compared between within-run replicates from the Nextera XT run (repeated from the DNA preparation stage). The identity of substitutions is given on the x axis and allele frequencies are shown on the y axis (note that the y axis begins at a frequency of 40%).

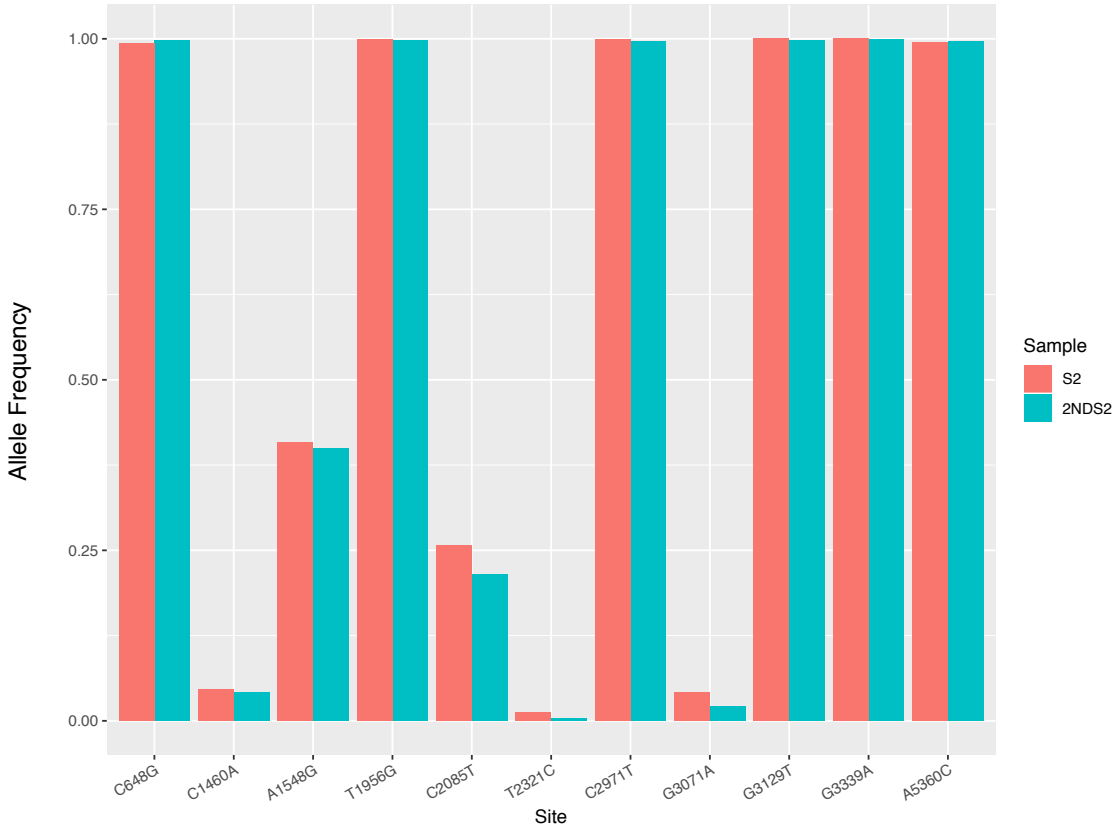


Figure 5.2: Allele frequency at sites for S2 and 2NDS2, representing between-run replications of both DNA preparation and library preparation. S2 was sequenced with Nextera XT, while 2NDS2 was sequenced with Nextera Flex (see section 5.3.1.1 for sample nomenclature). The identity of substitutions is given on the x axis and allele frequencies are shown on the y axis.

5.3.2.2 Nextera XT and Flex sequencing run

In order to check for sequencing repeatability, an additional sequencing run was carried out. One run was prepared using a Nextera XT kit, while the second run was prepared using a Nextera Flex kit. Both runs were sequenced with MiSeq V3 cartridges on the MiSeq Illumina platform, but at different times (sections 2.5.3.1, 2.5.3.2). In the Nextera XT DNA run, all samples were captured and sequenced (including duplicates for one sample for the DNA and library preparation procedures; section 5.3.2.1; figure 5.1). For the Nextera Flex DNA run, 6 isolates were re-sequenced delivering between-run replicates (viz, S2 and 2NDS2), more technical replicates (2NDSCSC2A and 2NDSCSC2B), and a cross-host DNA preparation control (2NDSCSC2S; sections 2.5.3.2 and 5.3.1.1 for methods and nomenclature, respectively). For S2 and 2NDS2 (figure 5.2), similar allele frequencies and allele identities were observed. The samples 2NDSCSC2A and 2NDSCSC2B also exhibited similar allele frequencies, but substantial differences between these and SCSC2 (from the Nextera XT run) were noted. While most identified sites' allele frequencies for 2NDSCSC2A and 2NDSCSC2B stayed at below 3 %, SCSC2 identified sites showed an average of ~25 %, except for sites 2275A and 3339A in both sets (figure 5.3). This may be as result of PCR bias, arising during PCR amplification steps, during library preparation. However, the numbers and types of variant alleles observed remain the same.

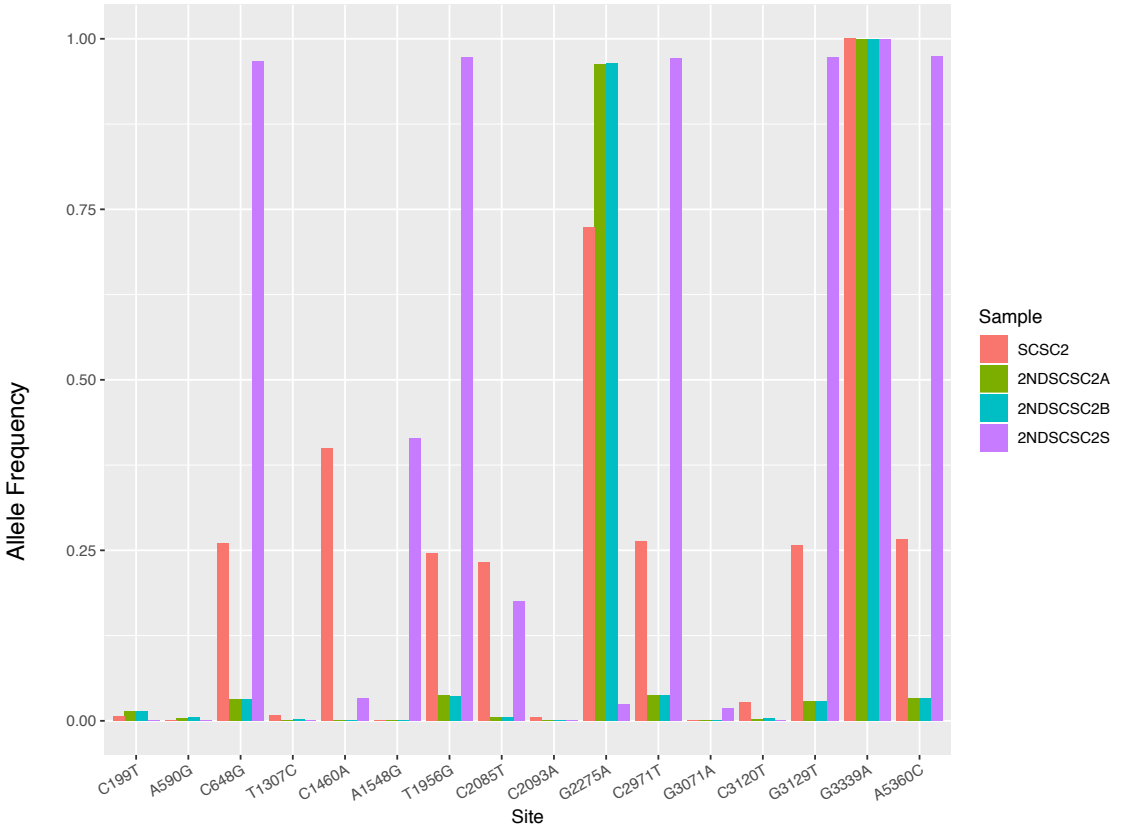


Figure 5.3: Allele frequencies of nucleotide changes for SCSC2, 2NDSCSC2A, 2NDSCSC2B and 2SCSC2S (section 5.3.2.1). SCSC2 was sequenced with Nextera XT, while 2NDSCSC2A, 2NDSCSC2B and 2NDSCSC2S were sequenced with Nextera Flex. DNA for 2NDSCSC2S was prepared in a host environment (S) different from the last host encountered by the population (C; see section 5.3.1.1 for sample nomenclature). The identity of substitutions is given on the x axis and allele frequencies are shown on the y axis.

5.3.2.3 Cross-host DNA preparation controls

The bacteriophage ΦX174 was cultivated in two different hosts, since the RF dsDNA DNA replication stage utilises host proteins, ΦX174 DNA preparation was carried out in the last host it was propagated on. For instance, SCSC RF dsDNA was prepared in the *E. coli* C host. To account for changes that may occur during RF dsDNA preparation, DNA preparation of the sample 2NDSCSC2**S** for which last host was *E. coli* C, was carried out in the *S.* Typhimurium host (hence the suffix: S).

The result shows that both allele frequencies and the mutations acquired differed substantially from the other SCSC series samples (figure 5.3). Instead the results for 2NDSCSC2S mostly mirror (except for some alleles such as 2085T) the pattern seen in samples for which the last host was *S.* Typhimurium rather than *E. coli* C (figure 5.11). This indicates that mutation can accumulate within the period of RF dsDNA replication which was carried out for 3 hours 30 minutes. Alternatively the pattern of nucleotide changes in 2NDSCSC2S may arise as a result of selection in a single round of phage attachment. If the selection coefficient for a minority of phages with different capsid residues is very large and positive (that is, selection is very strong because these viruses are much more efficient at attachment) then alleles responsible for these differences could increase in frequency even in a single round of selection. The library preparation-related changes (in 2NDSCSC2S) are occurring faster than originally expected (Bull *et al.*, 1997; Wichman *et al.*, 1999; 2000). However, the use of a *S.* Typhimurium host may impose stronger than expected selection in a short time period.

When assessing the significance of rapid changes, it is also of note that ancestral ΦX174, sample A, was cultivated in the *E. coli* C host strain for 3 hours in order to increase its concentration (section 5.3.2.1). Table 5.1 listed the variants acquired in sample A in comparison to reference ΦX174. 10 allele variants were observed (table 5.3) after initial growth in *E. coli* C, with 2

near fixation. This suggests that some nucleotide changes can occur within a few hours of growth. The early onset of nucleotide changes does not necessarily imply simplicity of ΦX174 evolution (dominated by short-time-scale changes) since complex allele frequency dynamics were observed (over a period of 10 days) within the SCSC time series (discussed in section 5.3.6.1). However, the findings in 2NDSCSC2S do complicate interpretation of the main effect of last-encountered host (discussed in section 5.3.3). In brief, this is because changes that are replicated between iterations of host switching and between branches may be arising during DNA preparation.

Despite the differences observed in 2NDSCSC2S, most of the changes that occurred were found in similar studies (table 5.4), some using the same RF dsDNA accumulation method (Godson and Vapnek, 1973) and others using PCR-based methods (which may also introduce biases during preparation). Therefore, there is a need for improved DNA preparation procedures.

5.3.2.4 pUC18 spike-in control and coverage

A pUC18 spike-in was introduced to check for cross-contamination between samples in the Nextera XT run (section 2.5.3.1; after method in Dickins *et al*., 2014). The spike-ins were added in an alternating fashion such that samples with spike-ins were not kept in close proximity to each other on the plate used during library preparation and spike-in-free wells were left between spiked samples (figure 5.4A). During analysis, cross-contamination for 'jumped' spike-ins in between wells were checked (section 2.6.2.2). There was very low to no coverage in samples that were not spiked indicating the absence of carryover into these samples from spiked samples (figure 5.4B).

5.3.2.5 Bioinformatics quality control

More common biases introduced during sequencing are from data analysis which includes library constructions and coverage (Ross *et al*., 2013). The

ΦX174 genome is small (5,386bp) therefore the yield obtained from a single multiplex run was sufficient for drawing inferences about allele frequency in the population, with coverage between 700 and 15,000, reaching as high as 35,000 in some genome regions of some samples (figure 5.5). The bioinformatics workflow used included steps for removal of adapters introduced during library preparation and removal of low-quality bases. The workflow also made allowance for read pairs mapping across the origin by mapping to an alternative coordinate system and merging the results at the end (section 2.6.2.2, figures 2.2 and 2.3). A comprehensive awareness of DNA and library preparation procedures, as well as sequencing biases can drive both wet and dry laboratory improvements for deep sequencing analysis.

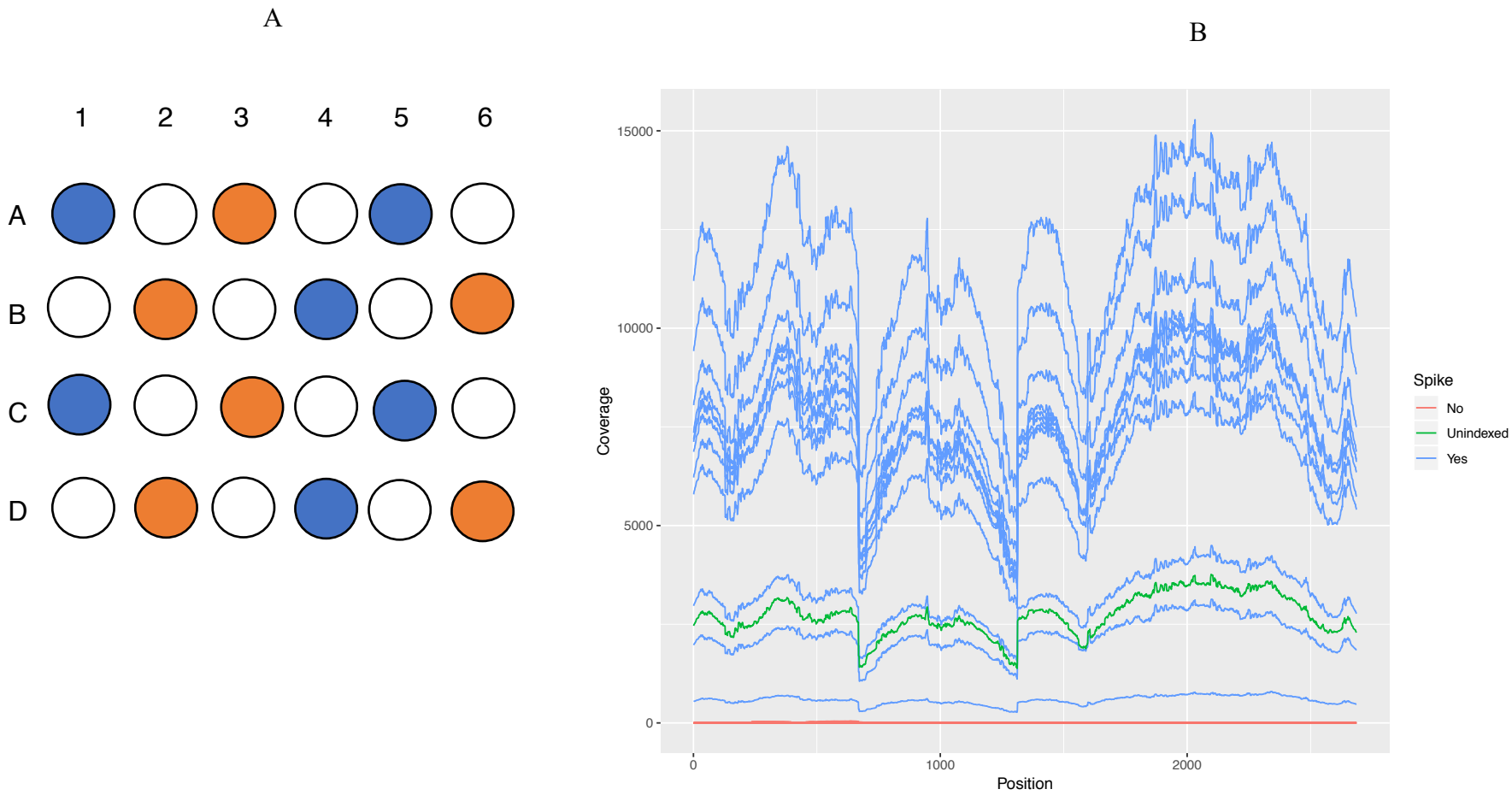


Figure 5.4: The 96-well plate layout and pUC18 spike-in coverage. (A) the 96-well plate layout used during the DNA library preparation, prior to sequencing. Blue-coloured wells represent sample DNA + pUC18 spike-in, orange wells represent sample DNA only, while non-coloured wells are blanks left between DNA wells (appendix C). (B) coverage of pUC18 spike-in in samples after sequencing, The x axis shows genome positions (pUC18 is 2,686 base pairs in size); the y axis is the coverage at each position.

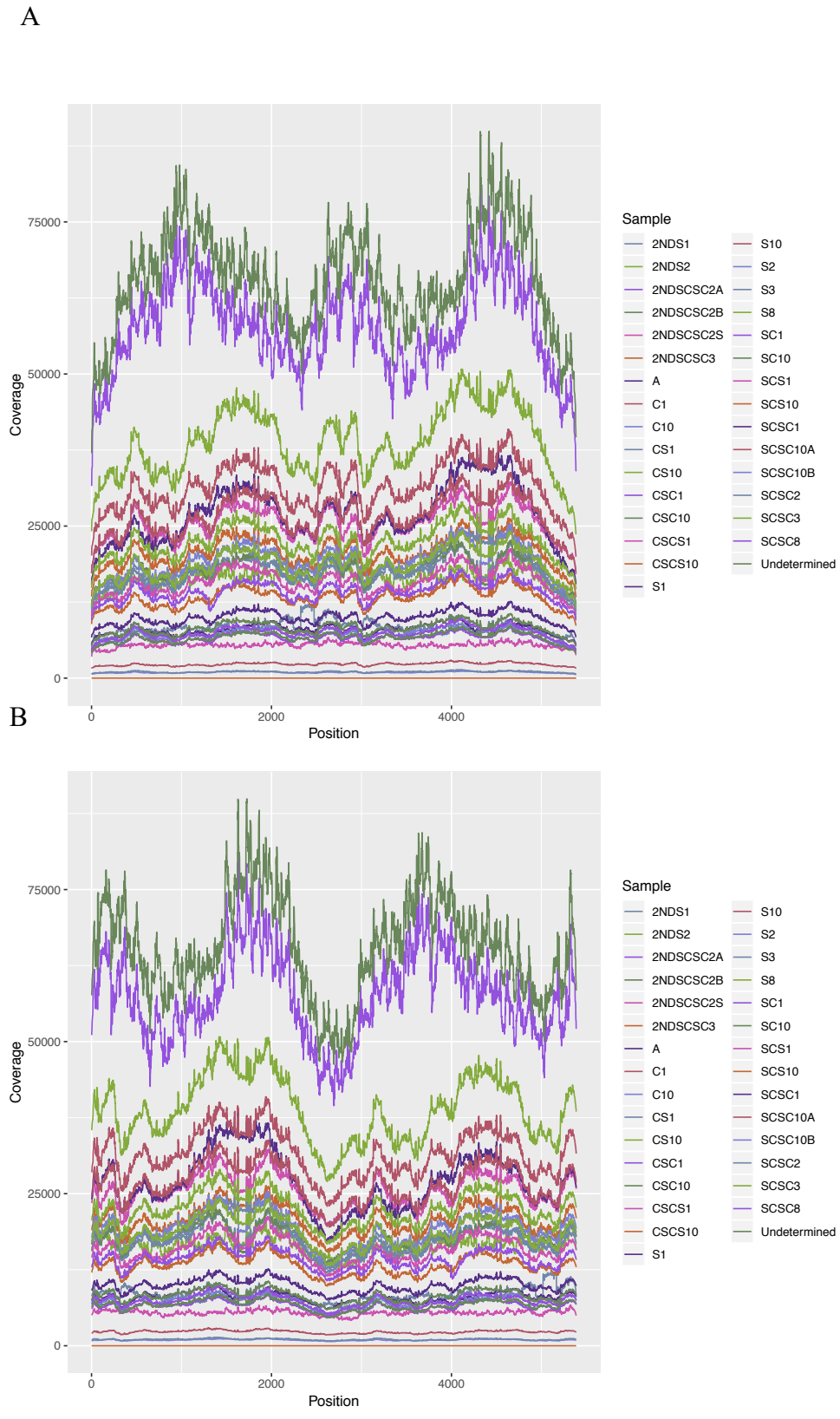


Figure 5.5: Coverage of the ΦX174 genome in (A) the reference coordinate system and (B) the resected coordinate system. The x axis shows genome positions (ΦX174 is 5,386 base pairs in size); the y axis is the coverage at each position.

5.3.3 Derived allele distribution across the ΦX174 genome

After analysis, the complete genome sequence of the ancestral and all evolved samples revealed that 35 nucleotide changes rose to detectable levels in both hosts' cultures, of which 31 were found in *E. coli* C-adapted samples, and 35 were found in *S.* Typhimurium-adapted samples (making 4 sites unique to *S.* Typhimurium-adapted samples). Two sets of multi-nucleotide events were observed (figure 5.9; appendix B) and no insertions or deletions relative to the reference were identified.

The nucleotide changes that were detected in both host environments were in genes associated with (table 1.1):

1. genome replication and packaging: replication initiation gene A, overlapping A*),
2. procapsid morphogenesis: gene B (overlapping A/A*) and D (overlapping with gene E),
3. DNA pilot: gene H,
4. the major coat: gene F,
5. DNA binding and packaging: gene J,
6. the DNA maturation: gene C ,
7. host cell lysis: gene E
8. non-essential: gene K (overlapping C): and in one intergenic region. (as summarised in appendix C)

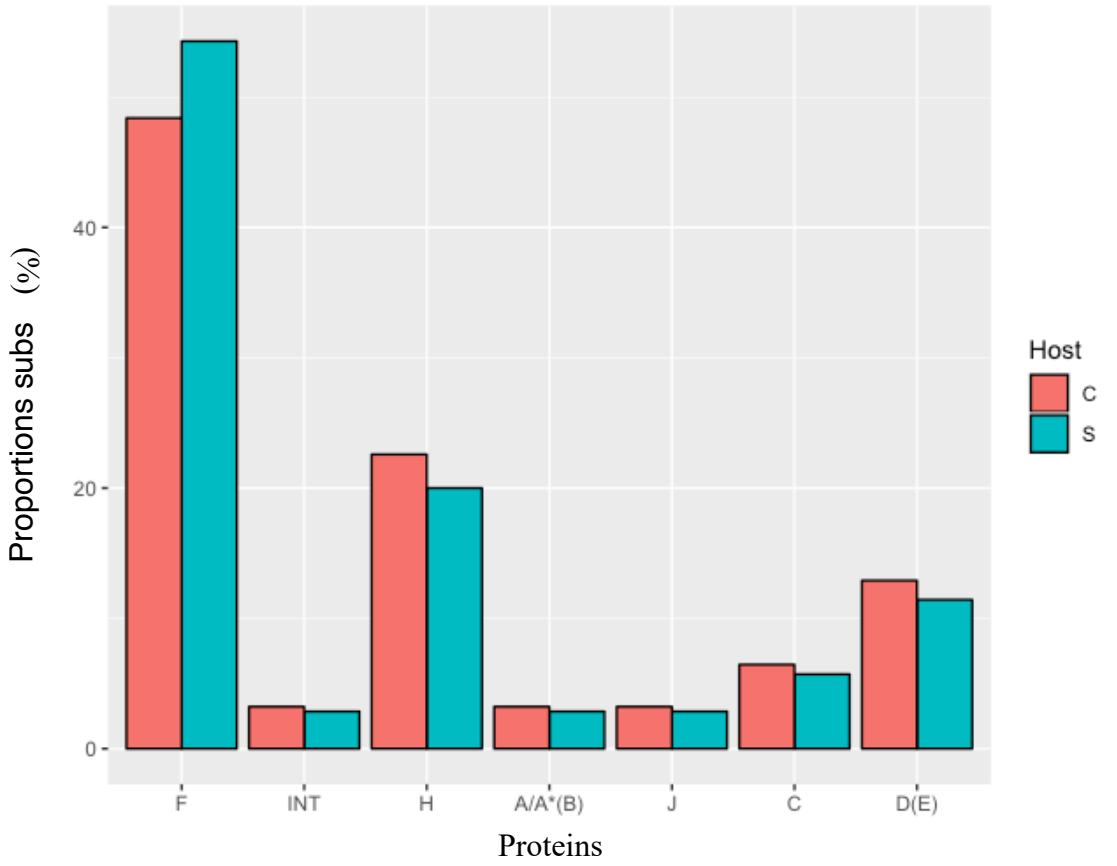


Figure 5.6: The distribution of substitutions identified in each gene in samples most recently adapted to *E.coli* C and *S.* Typhimurium. The x axis identifies genome blocks in ΦX174; the y axis indicates the proportions of substitutions (%) in each region observed in either *E. coli* C (C; red bar) or *S.* Typhimurium (S; green bar).

5.3.4 Distribution of nucleotide changes across ΦX174 genes and frequency classifications

In this experiment, nucleotide changes occurred in seven blocks. The distribution of substitutions observed across ΦX174 genes was not significantly different than expected given their lengths in nucleotide sequences (chi-squared test $x^2 = 28$, df = 24, p = 0.26). For clarity of interpretation, I will classify allele frequencies according to their frequency. Alleles with frequencies less than or equal to 1 % are herein referred to, putatively, as *rare variants*. Alleles with frequencies between 1 and 5 % are classified as *low-frequency variants*, and those greater than 5 % are *frequent variants*. Two genome regions manifested multi-nucleotide substitution events in some samples. These are referred to as *complex variants*.

5.3.4.1 Derived alleles in host-recognition genes

Protein H is involved in binding to the LPS receptor of the host bacterial cell and injecting DNA through spike protein H (Inagaki *et al.*, 2000). In this experiment ~22 % of the total substitutions occurred in gene H for *E. coli* C and ~20% (7 / 35 x 100) for *S.* Typhimurium. All nucleotide changes in gene H were found in both hosts, although with different allele frequencies. Shared alleles were found at nucleotide positions 2971(T), 3071(A), 3111(A), 3120(T), 3129(T), 3132(A), and 3339(A) (with derived alleles given in parentheses). Additional shared alleles that were newly acquired, and not found in the ancestral phage, were 3071A, 3111A, and 3120T. Alleles 3129T and 3132A were *rare variants* in ancestral ΦX174 (~10^4, not shown in table 5.1 because this was less than 1%), while 3339A was a *frequent variant* in all samples, having a frequency near fixation and maintaining it across all samples in both hosts with exception of C10 (~28 %) as shown in figures 5.10 and 5.11. The substitution 3071A is absent in the ΦX174 experimental evolution literature listed in table 5.4. Allele 3071A was a *rare variant* in *E. coli* C while being a *low-frequency variant* in *S.* Typhimurium. Although the difference in allele frequencies was not wide (~1 % in *E. coli* C, ~3 % in *S.*

Typhimurium), 3071A may contribute to differences in cell receptor binding in *S*. Typhimurium. Meanwhile, 2971T showed a substantial difference in allele frequency between the two hosts, being a frequent variant of ~100% in *S*. Typhimurium, but < 2% in *E. coli* C-adapted samples. This substitution may be involved in cell receptor binding, which differs in structure and composition between the two hosts (Inagaki *et al*., 2003), contributing to attachment and fitness trade-offs discussed in chapter 4. Pepin and Wichman (2008) suggested that 2971T may be an adaptive mutation in *S*. Typhimurium, owing to its arising in ΦX174 grown in harsh conditions. A similar pattern of allele frequency was observed for 3129T, almost fixed (~99 %) for *S*. Typhimurium, low (~25 %) for *E. coli* C, and also found to be adaptive in Pepin and Wichman (2008) study.

Substantial allele frequency changes were observed over the course of the experiment at some sites in gene F, ~48 % in *E. coli* and ~58 % in *S*. Typhimurium were found in this gene, affecting the major capsid protein, which is involved in recognition and attachment to LPS. Gene F substitutions have all been observed in the literature (table 5.4) except 1148A. 1148A is a *rare variant* across all samples except C10.

Complex variants were detected at sites 1301 and 1304 (1301-1304) as well as at sites 1346 and 1347 (1346-1347). These alleles are excluded in figures 5.10 and 5.11 but shown in figure 5.9 SCSC time series and table 5.3. These exhibit complex changes in one set of samples (described in section 5.3.6.2). A relatively high frequency at 1304 (1304G ancestral and 1304C derived) is found in the *S*. Typhimurium host (appendix C). The substitutions at 1301, 1304, 1346, and 1347 in gene F persisted throughout the duration of *S*. Typhimurium adaptation. Some of these sites have been observed in the literature and known to be important in host recognition (Crill *et al*., 2000; Pepin *et al*., 2007; Brown *et al*., 2013). Variants at sites 1301, 1346, and 1347 may be classified as *rare variants*, and it seems likely (due to low allele frequency) that these allelic variants offer smaller fitness advantages, but

because they are likely *S*. Typhimurium-specific, they may have positive selective advantages in this host.

Of particular interest is position 1304, for which the wildtype allele is G. The derived allele 1304C is a *frequent variant*, being nearly fixed (~100 %) across *S*. Typhimurium-adapted phage samples, but is largely absent in *E. coli* C-adapted samples (with the exception of samples in the SCSC series; section 5.3.6.1). This suggests that 1304C may be an allele required for adaptation to *S*. Typhimurium. This allele may partly be responsible for the large fitness increment observed in *S*. Typhimurium-adapted populations over the course of the experiment (chapter 4). The fitness effects of 1304C alleles in isolation are reported in chapter 6. In a study reported by Crill *et al*. (2000), a 1305G mutation occurs consistently (affecting the same codon as 1304C) when phage ΦX174 was grown on *S*. Typhimurium. In their study the substitution at site 1305G produced major changes in growth rate on *E. coli* C. The derived amino acid is different, being arginine (R) in this study and aspartic acid (D) in Crill *et al*. (2000).

Several sites show similar patterns with allele frequencies correlating with the recently-encountered host. More than half of the substitutions that track host switches occurred in genes F and H, a pattern (figure 5.6) that is consistent with the literature. Proteins F and H had been shown to have interactions with host LPS (Mckenna *et al*., 1994; Hafenstein and Fane, 2002), and the two hosts utilised possess distinct LPS structures. As the first point of contact with the host cell membrane, it is expected that ΦX174 may acquire mutations to enable successful attachment, a necessary first step in the infection process. Also, gene F covers a large percentage of the genome map, approximately 24 % (although mutations do not appear to be over-represented in this region: section 5.3.4). Some studies have demonstrated that mutations in the viral pilot protein (Cherwa *et al*., 2011; Young *et al*., 2014) as well as the capsid protein (Crill *et al*., 2000) result in defects in adsorption (further discussed in chapter 6, fitness effects of mutations).

5.3.4.2 Derived alleles in replication and packaging genes

The proteins associated with ΦX174 replication and packaging are A, H and C. Protein H is a multi-functional protein, which functions include host recognition, ejection of viral genome into host and piloting the genetic content to DNA replication of ssDNA to dsDNA sites (section 1.6.1.3). The protein has been discussed in section 5.3.4.1, the focus here will be on genes A and C, including associated overlapping genes.
Protein A acts in *cis* to initiate rolling circle replication of ΦX174, accomplished by binding of the protein to the formed dsDNA at the site of initiation (Hayashi *et al*., 1986). Genes A*, B and the 5' region of K overlap with gene A. The A* protein is encoded in the same reading frame as protein A and shares functions with A, except that it lacks the ability to package ΦX174 DNA into the capsid and cannot initiate rolling circle replication (Van der Ende *et al*., 1982).

A mutation occurred at site 5360 (absent in similar ΦX174 experiments, table 5.4), within the region of overlap between genes A, A* and B. 5360C is a *frequent variant* and this synonymous mutation possessed near fixed frequencies on all *S*. Typhimurium samples (figures 5.10, 5.11). Because protein A functions in DNA binding, unwinding and affinity for host helicase, a silent nucleotide change may result in quicker production of new viral stands (Eisenberg, 1980), an advantage for accumulation of viral progeny, high population size and probably higher fitness.

Protein C is involved in DNA replication and packaging the viral genome into the procapsid (section 1.6.1.4). In addition, it down-regulates transcription from one of ΦX174's three promoters and has been seen to be adaptive at high temperature (Brown *et al*., 2010). Substitutions appeared in sites 199(T) and 323(G) of protein C. The 323G allele has arisen in many experiments in both hosts, but not in experiments involving serial passaging (table 5.4). There is a possibility that site 323G is an adaptive mutation for ΦX174 in the

presence of large host populations and constant temperature in the chemostat environment.

5.3.4.3 Derived alleles in the non-coding region

Three non-coding regions exist in the ΦX174 genome, they are located between genes J and F, F and G, H and A. Intergenic regions between genes H and A, and genes F and G have been suggested to function during production of dsDNA from ssDNA (Arai and Kornberg, 1981). One allele arose in the F and G intergenic region 2321C, not seen in the literature (table 5.4). The mutation at 2321C is within the primosome recognition and binding site (Van Der Avoort *et al*., 1984). Although the allele is a *rare variant* on both hosts, the presence was higher (~1 % more) in the *S*. Typhimurium host. We may therefore speculate that a nucleotide change in this region improves ΦX174 interactions with the *S*. Typhimurium primosome.

5.3.4.4 Derived alleles in procapsid assembly genes

Several proteins function in viral procapsid assembly. This involved process entails interactions of proteins F, B, D, G, J and H (Fane *et al*., 2006; Cherwa *et al*., 2011). There were no nucleotide changes observed in gene G, one in gene J (944) and one in B (overlapping A and A*). The substitutions observed in genes F and H have been discussed in section 5.3.4.1.

The external scaffolding protein, gene D, overlaps with gene E (encoding the host cell lysis protein). Protein D organises assembly intermediates of the procapsid. Some alleles arose in gene D at sites 530(T), 572(C), 590(G), and 648(G), with three alleles (530T, 590G, and 648G) present in this study only (table 5.4). Allele frequencies at sites 530(T) and 590(G) were low making these alleles *rare variants* in both hosts, while 648G was a rare variant in *E. coli* C and nearly fixed in *S*. Typhimurium. Protein D has been shown to control fidelity in capsid formation and a mutation in gene D led to an increase in virion fitness (Cherwa *et al*., 2017). Therefore, the 648G

substitution may contribute to the increase in *S*. Typhimurium fitness discussed in chapter 4.

5.3.5 Mutation spectrum of nucleotide changes in samples

ΦX174 was initially isolated on an *E. coli* host (Sinsheimer, 1959), is typically grown on *E. coli* C host in the laboratory, and can infect rough strains of *Salmonella* (Hayashi *et al*., 1988). The wild-type ΦX174 for this experiment has most recently been propagated in *E. coli* C, therefore, *S*. Typhimurium may be considered a novel host environment for the virus. Likewise, the specific growth environment in this study, chemostat, represents a novel abiotic environment. Thus, ΦX174 adaptation involves two novel factors for ΦX174 populations adapted on *S*. Typhimurium (S, CS, SCS, and CSCS, section 5.3.1.1) and one novel factor for those propagated on the typical laboratory *E. coli* C host (C, SC, CSC, and SCSC, section 5.3.1.1). For *S*. Typhimurium samples, genetic changes that occurred on day 1 for all samples (10 or 11 allelic variants) remained almost the same up till day 10 (table 5.3). In *E. coli* C samples, SC and SCSC acquired more alleles in day 10 than at day 1, C decreased (9 - 6), while CSC remained unchanged (7). There is the possibility that the numbers of adaptive environmental factors affect the numbers of nucleotide changes expected to accumulate – at least ordinally (in numbers as shown in table 5.3). Since *S*. Typhimurium combines both novel host and chemostat culturing environment (two adaptive environmental factors – novel host and chemostat environment) in comparison with *E. coli* C (one adaptive environmental factor – chemostat environment), hence may need to acquire adaptive alleles with stable pattern.

In tracking the pattern of nucleotide changes from S1 – S10, most mutations were either at frequency of ~6% and below or ~97 % and above (6 near fixation, table 5.3), with only two *frequent variants* at ~37 % (1458A) and ~25 % (2085T) (figure 5.7). In some instances, it is possible that mutations could

attain near fixation by hitchhiking with large-effect beneficial mutations (Barton, 2000), the nucleotide substitution pattern of *S*. Typhimurium samples indicate that hitchhiking does not explain the observed high frequency. Hitchhiking occurs when an allele changes frequency because it is near a beneficial mutation undergoing selective sweeps and this may manifest in linked allele frequency changes. Genetic tracking of *S*. Typhimurium samples in figure 5.7 did not exhibit such a pattern.

Nucleotide changes may occur during the DNA preparation procedure (section 5.3.2.1), which could also explain the apparent persistence of some polymorphic sites. Although confounding *E. coli* C and *S*. Typhimurium comparisons, the DNA preparation was a constant factor within *E. coli* C or within *S*. Typhimurium. Therefore, the changes in observed allele frequency remain informative and the existence of a main effect attributable to *E. coli* C and *S*. Typhimurium is still evident.

The near fixation alleles (approximately half of cumulative polymorphisms) remained the same in number (6, table 5.3) and pattern (figure 5.11) for *S*. Typhimurium, suggesting most alleles acquired may be adaptive. *E. coli* C on average possesses ~2 fixed alleles, 4 fewer than *S*. Typhimurium. A closer look at the SCSC time series (figure 5.8), tracked from day 1 to 10 shows a complex pattern of nucleotide changes (figure 5.10) with clonal interference apparently acting on the alleles (discussed in section 5.3.6.2).

Samples Identities	Variants (polymorphic) (sampled days)					Variants (fixed) (sampled days)				
	1	2	3	8	10	1	2	3	8	10
A	10					2				
C	9				6	2				0
SC	7				12	1				2
CSC	7				7	2				2
SCSC	7	8	11	14	16	2	0	1	1	1
2NDSCSC2S		11					6			
S	11	11	10	10	11	6	6	6	6	6
CS	10				10	6				6
SCS	11				10	6				6
CSCS	11				11	6				6

Table 5.3: A summary of ΦX174 allelic (non-reference) variants in both hosts. The first column indicates sample identities including original starting ancestral sample A and the sample extracted in its non-recent host, 2NDSCSC2S (section 5.3.1.1). The second column is the number of polymorphic variations (*low-frequency and common variants*, ≥ 1 %). The third column is the number of substitutions (>95%).

5.3.6 Allelic variants through time series

Genomes were sequenced at multiple points during growth on a novel host. This was done to capture any possible changes in allele frequency that occurred in the population time series that may indicate ongoing adaptation. Figure 5.7 shows a clear, but unchanging allele frequency pattern for S samples time series (S1, S2, S3, S8, S10), some mutations were rare, some low and others common, and the pattern is approximately fixed from days 1 till 10. This suggests that after rapid fixation of alleles, further changes did not occur in the period.

5.3.6.1 Haplotypes in ΦX174 time series

In the SCSC time series there seem to be haplotypes that possess linked dynamics. The substitutions at sites 648, 1956, 2085, 2971, 3129 and 5360 (set X) had allele frequencies of near zero in day 1, increasing to ~25 % by day 2, 70 % in day 3, before falling to ~35 % in day 8 (except 2085T which was at ~5 %), and to ~14 % at day 10. All through the time series dynamics, these mutations' allele frequencies travelled together. Another set of substitutions at sites 2275 and 1460 (set Y) followed a similar pattern but in opposite direction (figure 5.8). In the first instance, set Y appeared to have high allele frequencies, suggesting they may be beneficial, but the trend dropped down by day 3. Meanwhile set X that was at zero frequency in day 1, rising to 70 % by day 3, coinciding with a drop in set Y. The trend displayed by these haplotypes suggest the presence of clonal interference.

5.3.6.2 Clonal interference in ΦX174 time series

Clonal inference is a phenomenon where multiple clones interfere and compete against each other in a population. When beneficial mutations arise in a population, positive selection causes a rise in frequency culminating in the fixation of beneficial alleles and loss of competing neutral or deleterious alleles, a process termed a selective sweep. However, in a large population

and at high mutation rates, it is likely that beneficial alleles will arise in different lineages at the same time (although sometimes alleviated in large populations and at high mutation rates; Bollback and Huelsenbeck, 2007). For a sexual population, these mutations can merge together, forming a single lineage allowing alleles to be fixed together as a result of recombination (Hill and Robertson, 1966), but in an asexual population (or in a population with limited recombination) these lineages may compete with each other. This is the phenomenon referred to as clonal interference. Clonal interference is often thought to slow down adaptation by delaying the rate at which beneficial mutations are fixed, since mutations with larger selection coefficients out-compete less beneficial counterparts, affecting the pattern of genetic variation (Gerrish and Lenski, 1998; Wahl and Krakauer, 2000). If beneficial mutations arise rarely, it is possible for them to become fixed prior to the appearance of competitive mutations. However, if another set of mutations with higher selection coefficient arise, they may end up being extinct, a possible example is set X in SCSC, appearing to decrease in frequency by day 10. The mutation with higher coefficient may also be affected. This is because the difference in coefficients between two sets of beneficial mutation appears to be smaller than between wildtype and the most beneficial mutation, as seen in set Y. There appears to be some evidence for competing haplotypes in the SCSC time series, the sets of haplotypes may have similar fitness (so that their frequency is changing arbitrarily). It is also possible that there are direct competition interactions in the two sets X and Y.

Complex variants illustrated in figure 5.9 exhibited evidence of clonal interference. For set 1301 -1304, the wild-type allele (ACTG) showed a frequency at near fixation (~100%) in SCSC day 1, ~77 % in day 2, decreasing further to ~35 % in day 3, and rising again in day 8 to ~78 %. Meanwhile, a change at 1304(C; yielding ACTC) was near zero in day 1, increasing to ~23 % in day 2, in day 3 rose to ~63 % at the time ACTG was decreasing. Finally the ACTC allele declined again to ~6 % as the wild-type

ACTG increased. In summary, each time the ACTG allele frequency decreased, the *complex variant* ACTC increases in frequency. This suggests that there was competition between the two sets of alleles. A similar effect was shown in the second set of multi-nucleotide event observed in the same time series. The 1346 and 1347 wild-type (GA) allele frequency was ~100 %, while its *complex variants* AG were *rare variants* near (~0 %) from days 1 – 3. As GA allele frequency decreased to ~81 % in day 8, the AG allele rose from ~0 to ~15 %. In day 8, GA increased to ~97 % and AG decreased to ~2%.

Overall, there was evidence of clonal interference occurring in the population of the typical laboratory host, *E. coli* C, where sets of putatively beneficial mutations appeared contemporaneously, however, this pattern was not observed in the novel host, *S*. Typhimurium. The competing alleles were unable to occupy the same position on the genome. Clonal interference forced the independent beneficial mutations to compete rather than combining, thus impeding allele fixation (Abedon, 2008). This observation is similar to Pepin and Wichman's, (2008) ΦX174 study, where evidence of clonal interference was detected more in a benign environment than in a harsh one and suggested the existence of a higher number of potential adaptive mechanisms in the benign environment in comparison to harsh environment, which decreases the waiting period for accumulation of beneficial mutations accumulation.

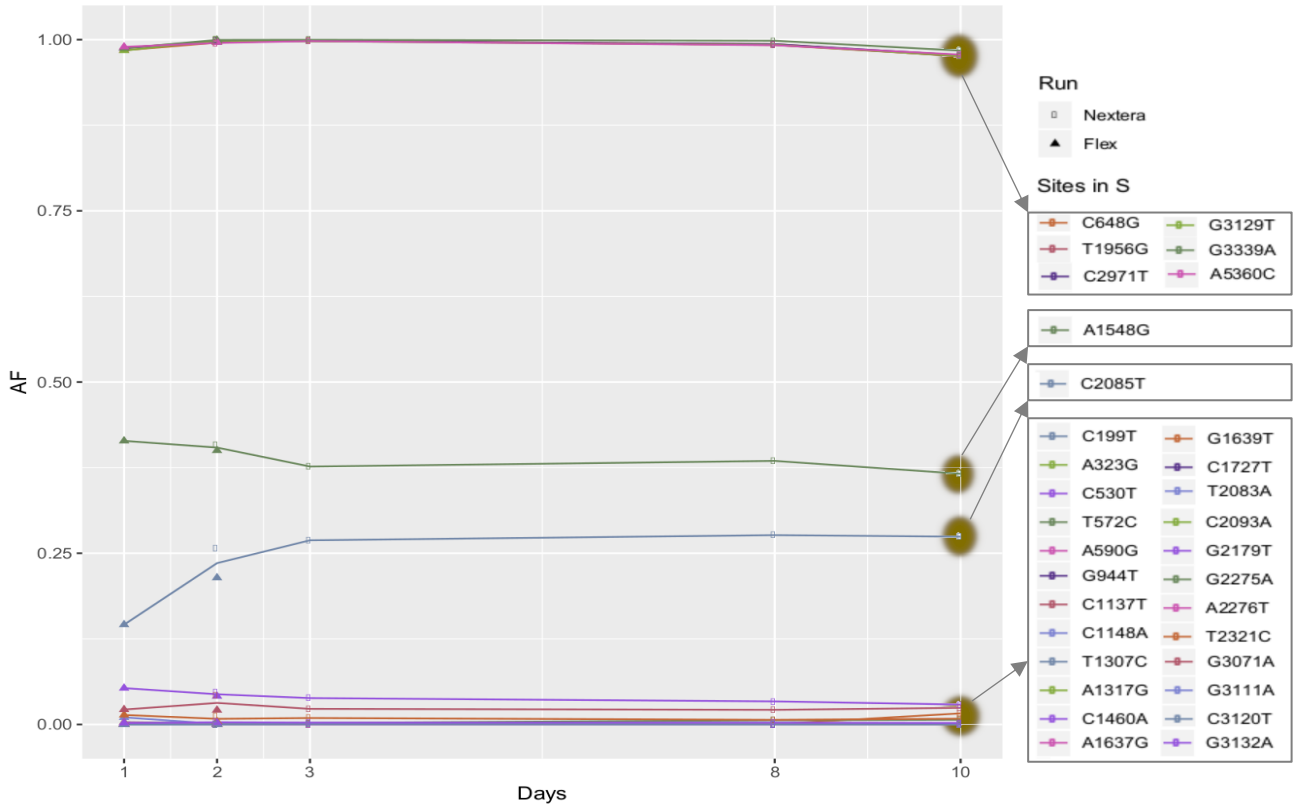


Figure 5.7: Allele frequency during the S time series, with allele frequency (AF) on the y-axis versus days of growth on *S*. Typhimurium on the x axis. Data is presented for samples S 1, 2, 3, 8, 10 and 2NDS2. Where replicate data points are present (S2 and 2NDS2), line segments are interpolated. Samples extractions were sequenced with both Nextera XT kit (circles) or Nextera Flex kit (triangles). Colours differentiate allele changes given in the key in the format "Ancestral_Location_Derived" (e.g., A1548G means a change from A to G at position 1548).

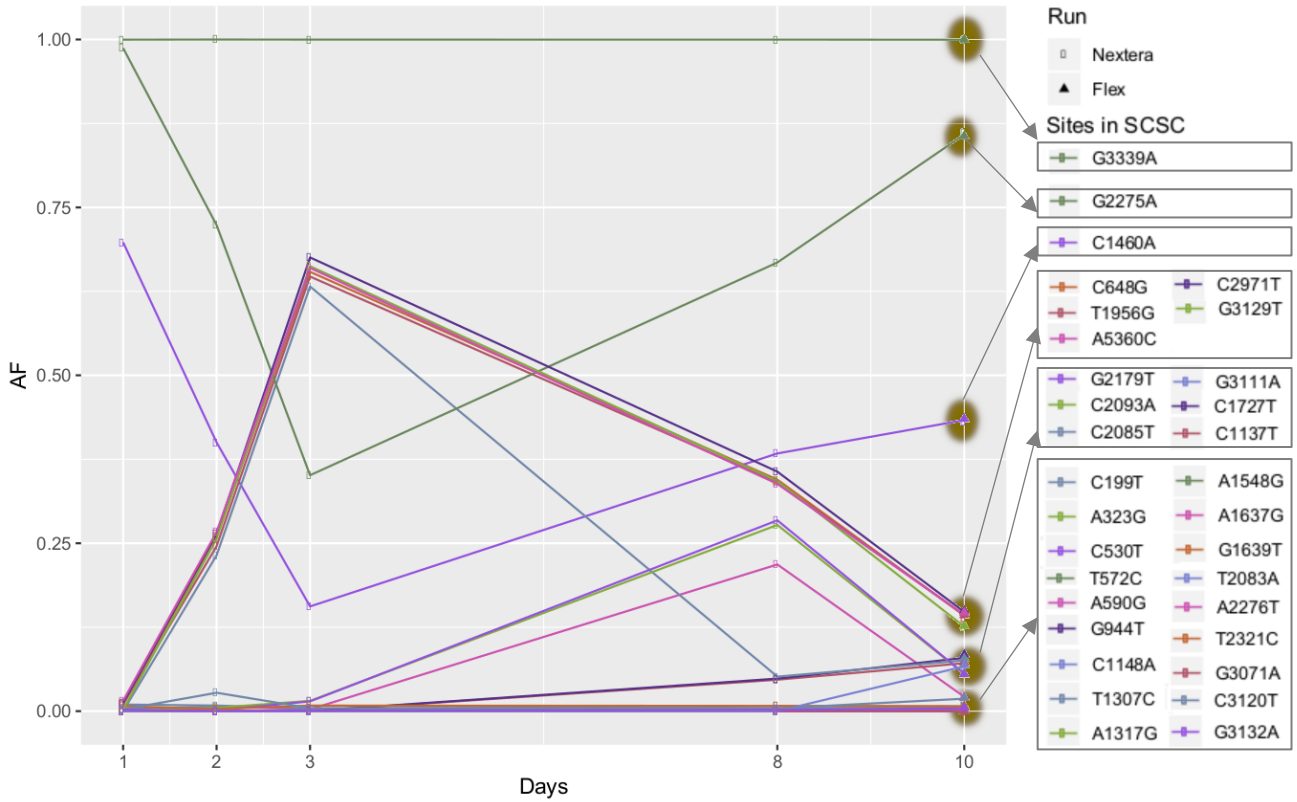


Figure 5.8: Allele frequency during the SCSC time series, with allele frequency (AF) on the y axis, versus days of growth on the *E. coli* C host on the x axis. Data is presented for samples SCSC 1, 2, 3, 8, 10A and B (the latter two being DNA preparation replicates for day 10). The plotted lines are mean values, and all samples were sequenced in the first, Nextera XT, run. Alternative symbols indicate replicates and colours differentiate allele changes given in the key in the format “Ancestral_Location_Derived” (e.g., A1548G means a change from A to G at position 1548).

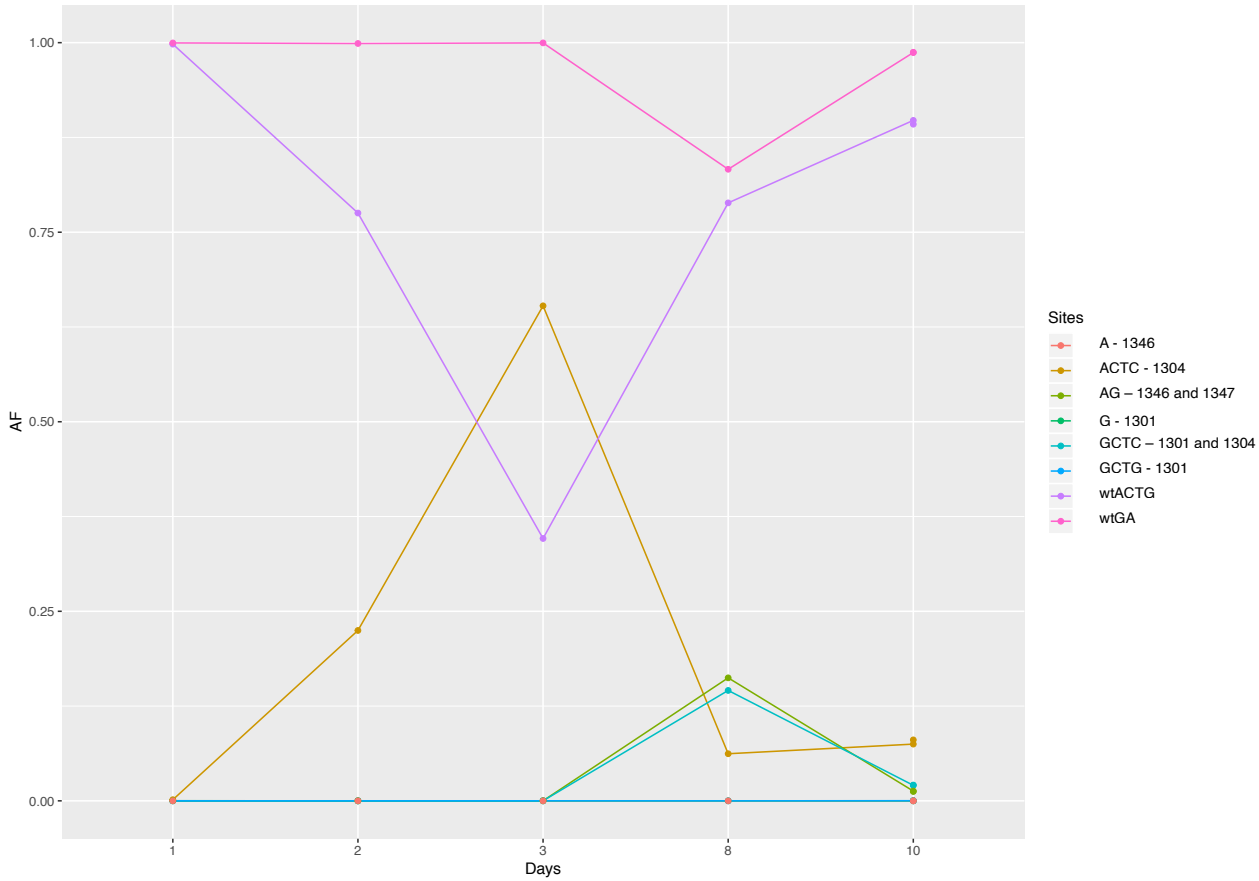


Figure 5.9: Multi-nucleotide events in the SCSC time series. The y axis shows allele frequency (AF), while the x axis represents sampling days 1, 2, 3, 8 and 10. Overlapping points at day 10 are derived from samples SCSC10A and B replicate samples (from DNA preparation stage). Two sets of multi-nucleotide events are represented between sites 1301 and 1304, inclusive. wtACTG is ΦX174 wild-type, ACTC - change to a C at position 1304, GCTG – change to a G at position 1301, GCTC – changes at both 1301 and 1304 (G and C, respectively), and G – a change in 1301G. 1346 – 1347; wtGA is ΦX174 wild-type, AG – changes in 1346 and 1347, A – change in 1346A.

5.3.7 Frequency reversals at variant sites

The main pattern of interest in the genetic data is the recurrent gain and loss of the same set of alleles in samples exposed to a presumably identical host environment (figures 5.10, 5.11 for C-branch and S-branch, respectively). This study examines mutations in populations evolved on either *E. coli* C or *S.* Typhimurium in the chemostat environment, and nucleotide changes are followed as switching to the alternate host occurs. This approach identifies a set of nucleotide changes that are likely important for host-specific adaptations. Some sites with a host-specific pattern manifested intermediate allele frequencies in one host and close to zero frequencies in the other host. Other sites with a host-specific pattern were approximately fixed in one host with close to zero frequencies in the other. The key feature of the observed nucleotide changes was that they exhibited similar patterns of changes (gain and loss) in each host for each (alternating) host switch. The recurrence of these patterns across branches of the experiment (figures 5.10, 5.11) is similar to parallel evolution. Parallel evolution is a term used to describe the acquisition of traits in independent lineages. Numerous examples of parallel evolution have been described in the literature (Schluter *et al.*, 2004), including in ΦX174 studies (Bull *et al.*, 1997; Wichman *et al.*, 1999; Wichman *et al.*, 2000; Holder and Bull, 2001; Brown *et al.*, 2013). Parallel evolution provides evidence that nucleotide changes are adaptive and occurred in response to natural selection (Bollback and Huelsenbeck, 2009; Longdon *et al.*, 2018). The presence of parallel changes in this experiment therefore also suggests that allelic variants associated with each host are adaptive. Alternatively, the evidence from the host crossover control sample (2NDSCSC2S; discussed in section 5.3.2.3) indicates that observed alleles may arise during DNA preparation. However, it is still possible to argue that these changes are parallel since they would then presumably arise independently each time the preparation procedure is repeated. We now turn to the identity of some of these putatively adaptive, host-specific alleles.

The 1460A allele can be categorised as a *frequent variant* for populations adapted in *E. coli* C (~60 % allele frequency). In contrast, populations grown in *S*. Typhimurium possessed this allele in the form of a *low-frequency variant* (~ 3%). The pattern was repeated during four consecutive periods of selection on alternate hosts (or during repeated DNA preparation procedures on respective hosts). Therefore, the 1460A allele showed host specificity for *E. coli* C throughout the experiment, despite being polymorphic in the population. Changes at this site have also been observed in evolutionary studies that involve adaptation of ΦX174 to *E. coli* C in continuous culture (Wichman *et al*., 1999; Wichman *et al*., 2000; Crill *et al*., 2000). A phylogeny study observed that sites 1460(A) and 2085(T), overlapping with this study, lie on the outside of coat protein (Redondo *et al*., 2017). These sites may be necessary for LPS recognition of bacterial host cell. The 2275A allele followed a similar pattern (being a *rare variant* in *S*. Typhimurium and a *frequent variant* in *E. coli* C), a pattern also observed (at the same site or at a site on the same codon) in ΦX174 evolution studies in the *E. coli* C host in a chemostat environment (Brown *et al*., 2013; Dickins and Nekrutenko, 2010) as well as during serial transfer (Wilcox, 2017).

Meanwhile, a number of mutations have been identified as likely host-specific for *S*. Typhimurium. *E. coli* C-adapted populations (section 5.3.1.1) showed a clear pattern of *rare variants* at these sites while the same alleles were *frequent variants* in *S*. Typhimurium (figure 5.11), but some variants in *E.coli C* for S-adapted populations had allele frequencies of ~20 % for SCSC samples on days 2 through 10 (figure 5.11; discussed in section 5.3.1.1). The SCSC chemostat iteration was the last period of ΦX174 adaptation on the *E. coli* C host for the S-branch lineage, following host switching back and forth. It is plausible, if host switching continues, that ΦX174 may acquire adaptive mutations that will enhance host-range capabilities. Substitutions at sites 648(G), 1956(G), 2971(T), 3129(T) and 5360(C) exhibited allele frequencies near fixation, while 2085T and 1548G were ~50 %.

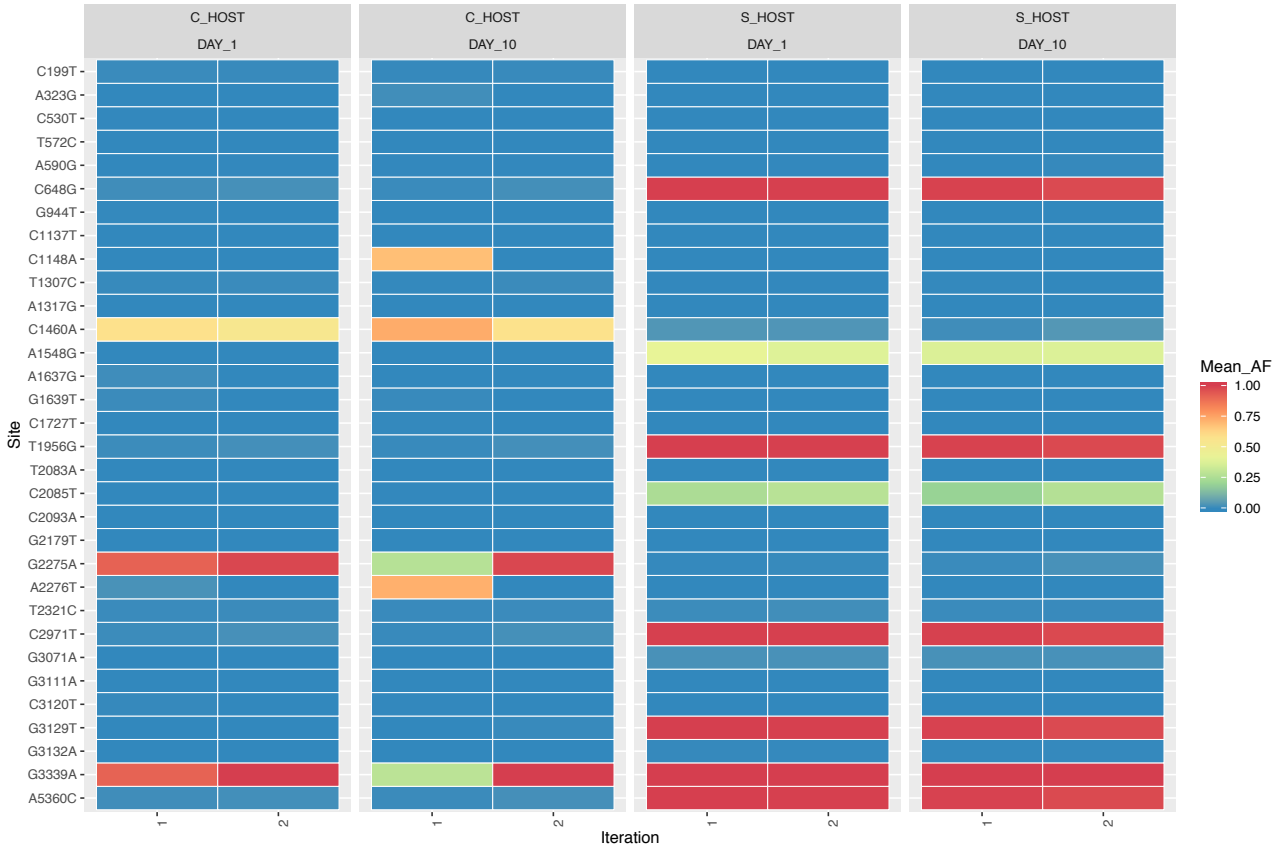


Figure 5.10: Allelic variation patterns for the C-branch experiment. Iteration (on the x axes) denotes the number of times each lineage of ΦX174 has encountered a particular host. Colours indicated as Mean_AF (mean allele frequency) in the key represent the allele frequency of for each allele/sample which varies from 0 to 1 (if DNA preparation replicates are present the mean value is displayed). The header for each heatmap shows the present host for the sample (C_HOST for *E. coli* C and S_HOST for *S.* Typhimurium) as well as days 1 (DAY_1) and 10 (DAY_10) indicating the duration of adaptation on the present host. Please note that all alleles lying within multi-nucleotide events in at least one sample are excluded from this figure, viz, alleles at sites 1301, 1304, 1307, 1346, and 1347.

Figure 5.11: Allelic variation patterns for the S-branch experiment. Iteration (on the x axes) denotes the number of times each lineage of ΦX174 has encountered a particular host. Colours indicated as Mean_AF (mean allele frequency) in the key represent the allele frequency of for each allele/sample which varies from 0 to 1 (if DNA preparation replicates are present the mean value is displayed). The header for each heatmap shows the present host for the sample (C_HOST for *E. coli* C and S_HOST for *S.* Typhimurium) as well as days 1 (DAY_1) and 10 (DAY_10) indicating the duration of adaptation on the present host. Please note that all alleles lying within multi-nucleotide events in at least one sample are excluded from this figure, viz, alleles at sites 1301, 1304, 1307, 1346, and 1347.

5.3.8 Common nucleotide changes from ΦX174 studies

About 75 % of mutations associated with host adaptation in this study have been observed by researchers in similar bacterial host experiments. These sites are summarised in table 5.4. Some experiments were performed in a continuous culture system, while others through serial transfers, with different dsDNA preparation (RF DNA accumulation or PCR based) and sequencing (Sanger or Illumina sequencing) methods employed. The hosts utilised in this literature are either *E. coli* C, *S*. Typhimurium or both. Brown *et al*. (2013), Bull *et al*. (1997), and Crill *et al*. (2000) utilised both hosts in their studies together with a chemostat environment, but introduced additional temperature selective pressure (43^{o}C) for *S*. Typhimurium populations. Meanwhile, Pepin and Wichman (2008) provided alternative selective pressures by adapting ΦX174 in different levels of calcium chloride ($CaCl_2$) to create harsh and benign environments (associated with deficient or excess calcium ions respectively). Despite the differences in selective pressures, most of the mutations observed in those studies and here overlapped, with most in protein F and H, responsible for host recognition. For example, mutations at sites 1460 and 2093 were noted by Crill *et al*. (2000) as substitutions found in *S*. Typhimurium adaptation, appearing only once throughout the 55 days of adaptation. However, here, though shared between both hosts, the allele frequency in *S.* Typhimurium is at ~3 % at the maximum while stable around ~ 60 % in *E. coli* C for 1460A, and the 2093A mutation is found in both hosts at a very low frequency of ~ 10^4 except in SCSC8 which shows a ~27 % allele frequency.

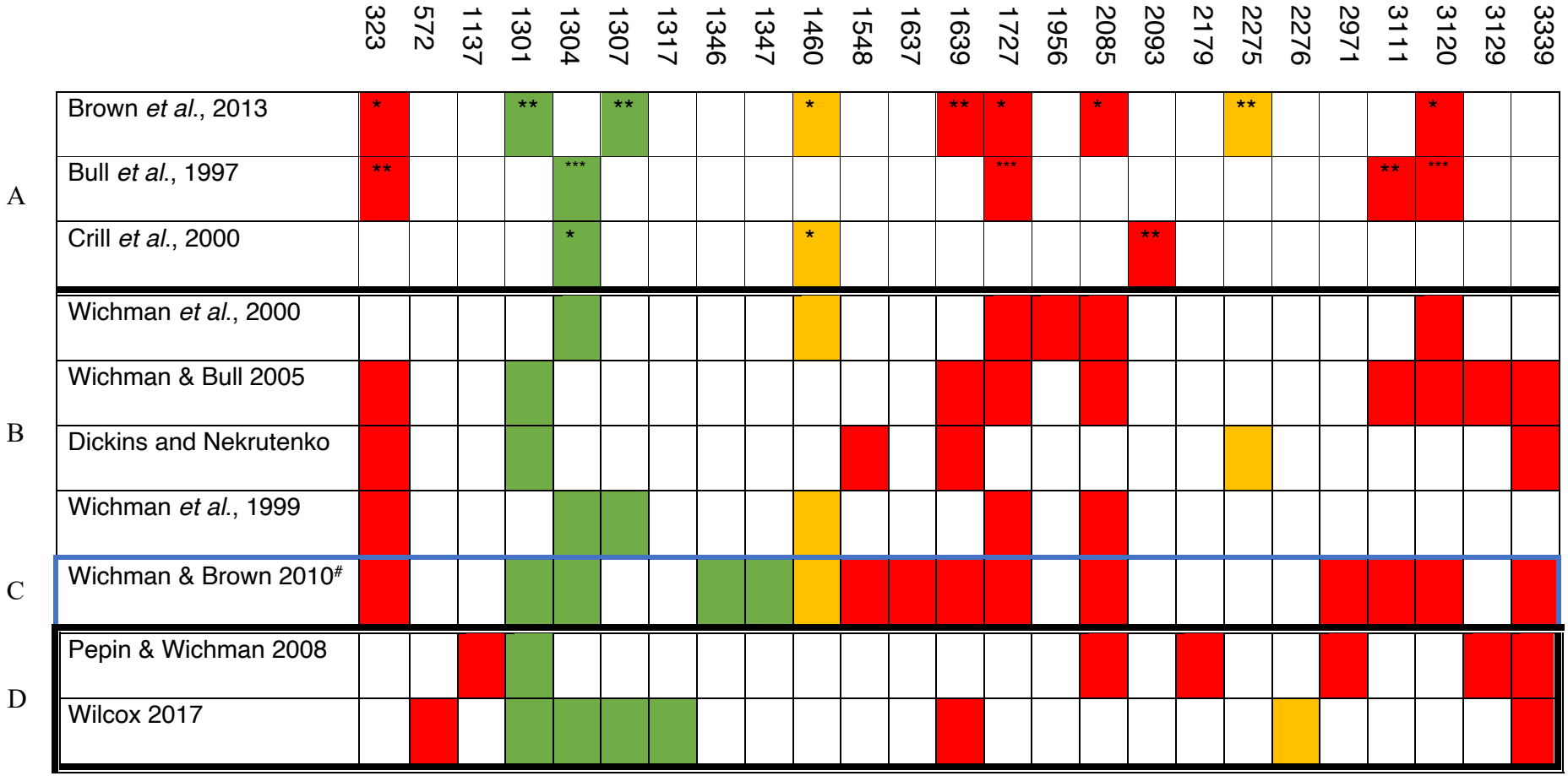


Table 5.4: ΦX174 alleles in this study compared with the literature. Set A (separated with first border line) studies that utilised both hosts, ** appears in *S.* Typhimurium only, * in *E. coli* only, *** in both E coli and *S.* Typhimurium. Set B (with border line) studies that used *E. coli* C only. Set C (blue-coloured border) a review that compiled changes seen in different studies. Set D (thick border line) studies that utilised serial passaging. Green boxes are changes in this study showing host specificity with *S.* Typhimurium, yellow boxes indicate host specificity for *E. coli* C and red boxes are changes that occurred in both host.

5.4 Conclusion

The approach taken in this study has been to use experimental evolution of phage ΦX174 as a model system to study host switching events. In addition to phenotypic analyses discussed in chapter 4, this chapter has probed genetic variation associated with viral host switching using a deep-sequencing approach. A number of control procedures were included, which suggest possible sources of error affecting inferred or actual allele frequencies, during the different stages of DNA and library preparation. No cross-contamination was detected using spike-in controls during library preparation and within-run controls (from the DNA preparation stage on) indicated within-run consistency. However, significant deviations in allele frequency seem most likely to have arisen during DNA preparation (section 5.3.2.3) although differences in the library preparation procedure (most likely during PCR-steps) are also implicated in some variation (section 5.3.2.2).

ΦX174 was placed in novel laboratory conditions, in a custom-designed chemostat with specified bacterial host cells, and genetic changes were tracked through time. Despite the likely errors from DNA and/or library preparation, the results obtained do provide information on the genetics of adaptation during and after host switching events. The study shows that mutations that enable adaptation to a novel host environment are observed reproducibly (either owing to adaptation in the chemostat or arising from DNA preparation on the corresponding host), and it provides an opportunity to observe changes over time series (for which DNA preparation procedures are constant). The interpretation of one of these time series (SCSC series) hinted at the presence of antagonistic evolutionary dynamics, suggesting that putative haplotypes may have competed with one another during adaptation to *E. coli* C.

The emergence of new alleles occurred throughout the experiment. Some of these nucleotide changes appeared in both hosts while some are largely

host-specific. Also, it was observed that evolutionary reversals occurred, meaning that the same set of alleles seem to appear every time ΦX174 was evolved in a given host or every time DNA are prepared on that host. Parallel evolution of host-specific changes (even if occurring during library preparation) suggests that the identified alleles constitute a signature of host-specific adaptation. Comparison of host-specific sites with the experimental evolution literature also supports the view that at least some of the alleles observed in this study are adaptive.

The majority of the reversion events are host-specific mutations occurring in gene F; this can be explained partly by gene F's large relative size as a target for mutations (compared with other genes in the genome). However, protein F, a capsid protein has been described to have important interactions with host LPS (Hafenstein and Fane, 2002). Since *E. coli* C and *S.* Typhimurium differ in LPS structure, individual effects of some of the mutations may result in differences in the growth pattern of ΦX174 on these hosts. Differences in growth rate have been observed and discussed in chapter 4. Examining the effects of host-specific alleles in isolation may contribute to uncovering mutations associated with an increase in viral growth rate. The next chapter examines the fitness effects of some of these putatively host-specific mutations through targeted mutagenesis.

Chapter Six: Fitness effects of reconstructed alleles

6.1 Introduction

6.1.1 The fitness effects of mutations

Mutations are the ultimate source of population genetic variation. Without the variation created by mutation, populations cannot adapt to novel environments and natural selection cannot operate. Natural selection is a key mechanism of evolution that determines the change in heritable traits of populations over generations. The outcome of adaptation by natural selection is a change in the phenotypes produced within a population. The phenotypic variation produced is underlain by genetic changes. If genetic changes affect phenotypes that influences growth rate or survival, then the mutation is presumed to possess a fitness effect (Gordo *et al.*, 2011).

Fitness is a measure of natural selection, representing the ability of an organism to survive and reproduce in its environment (Orr, 2009). The effect of mutations on fitness contributes to the direction and strength of natural selection (Eyre-Walker and Keightley, 2007; Keightley and Eyre-Walker, 2010; Gerrish and Hengartner, 2017). A population with a broad host range may experience a range of selection coefficients at a given locus/for a given allele. For populations adapting to a particular novel host environment, differentially selected alleles with lower selection coefficients may arise for effective reproduction and adaptation to the new environment.

Mutations possess fitness effects and these depend on the environment as well as between and within host species/genome in such an environment. The study of fitness effects of mutations is central to experimental evolution and crucial in understanding the outcome of adaptation at the molecular level. Understanding the fitness effects of mutations is therefore critical for evaluating the mechanisms responsible for the emergence of pathogens, as well as for understanding resistance mechanisms and adaptation to novel hosts. The probability of pathogen emergence in novel hosts and survival in different host environments will depend on the range and distribution of

fitness effects of mutations (Pepin *et al.*, 2006), and whether the mutations are differentially selected in a given host, which may result in host specificity. Host-specific fitness variation is partly a consequence of accumulation of acquired mutations. The measurement of fitness variation is used to predict and test the response to adaptation through quantitative genetic methods, with fitness components taken to be indicators of fitness.

Adaptation in a novel host environment is an outcome of the nucleotide changes. Adaptation frequently entails the acquisition/fixation of multiple mutations in response to environmental change. Therefore, the individual contribution of mutations must be disentangled from the population and environmental interactions for a better understanding of the fitness effect of the acquired mutations. An explicit understanding of mutational fitness effects can be achieved by constructing genotypes differing at amino acid sites (Sanjúan, 2010). A suitable method to study the fitness effects of mutation is through introduction of the acquired mutations separately into an organism via site-directed mutagenesis and testing of fitness in different environments. This may be achieved relatively straightforwardly in a well-studied organism like ΦX174 with known structure and functional proteins.

6.1.2 Site-directed mutagenesis

Targeted mutagenesis has brought substantial benefits to molecular genetic work since its inception in 1983 when site-directed mutagenesis was used to study the lipoprotein signal peptide of *E. coli* (reviewed in Inouye, 2016). Since then, it has contributed to our understanding of the structures and functions of proteins and nucleic acids, as well as being used in protein engineering and gene editing. The technology is mostly used to introduce mutations at desired sites of target genes so that effects on protein function may be evaluated during the life cycle of organisms.

Several methods are available to effect targeted mutagenesis, these includes cassette mutagenesis, Kunkel's method, whole-plasmid mutagenesis, the electroporation method and PCR-based SDM (Sambrooks and Russel, 2001). PCR-based SDM requires the design and synthesis of a DNA primer (table 2.3) that contains the desired mutation, but is otherwise complementary to the template DNA around the allele site. This allows the primer to hybridize with DNA in the region of interest and with nucleotides upstream and downstream of the target site/locus. During mutagenesis extension of single-stranded primer is accomplished by a DNA polymerase. This creates copies of the target region containing the mutated site and the amplicons produced may be transformed into competent bacteria (Chapnik *et al*., 2008). To achieve this in ΦX174 or other DNA phages, the template should be isolated from plaque-purified virus to increase the homogeneity of sequences (Sanjúan, 2010). Finally, mutants are selected and cross-validated by DNA sequencing to check for the desired allele.

6.1.3 Host-recognition proteins

As discussed in (section 1.6.1), a lytic viral lifecycle includes recognition of and attachment to susceptible hosts, transporting of the viral genome into the host, replication, utilizing host machinery, and production of viable progeny through bursting of host cell. Fundamental steps in a bacteriophage lifecycle are recognising and infection of a susceptible host. The infection step includes translocation of genetic material into target host cell, essential for viral survival and propagation. Some viruses directly encode their genome in a nucleocapsid or capsid. For some the genome is protected and enclosed in icosahedral capsid (Liu *et al*., 2010; Olia *et al*., 2011; Hu *et al*., 2013; Peralta *et al*., 2013), where genome delivery may be via the tail, or through tail-less mechanisms relying on host organelles, or a pilot protein (Molineux and Panja, 2013; Peralta *et al*., 2013; Sun *et al*., 2013). Bacteriophage ΦX174 is a tail-less icosahedral virus made up of ssDNA and four capsid proteins F, G, H and J (Hayashi *et al*., 1988). The ΦX174 icosahedron possesses twelve

vertices with spikes on each, and each constituted of protein H and G (McKenna *et al*., 1992). The spike protein was reported to recognise and adsorb to lipopolysaccharide receptors of susceptible bacterial cells (Brown *et al*., 1971; Rakhuba *et al*., 2010). The major capsid F protein consist of 5-fold vertex with eight β strands and seven loops insertion. The F protein interacts and binds with G protein through two of the insertion loops EF, lying between His73 and Pro234, and FG, found between Trp243 and Gly262, most mutations affecting assembly are found in these two loops (Ilag and Incardona, 1993). In the infection process, G and F proteins have been demonstrated to interact with lipopolysaccharide to produce a channel for injection of ΦX174 ssDNA into bacterial host cell while H protein recognises and penetrates host lipopolysaccharide membrane alongside with the ssDNA (Jazwinski *et al*., 1975; McKenna and Rossmann, 1994; Suzuki *et al*., 1999; Inagaki *et al*., 2000 and 2003; Sun *et al*., 2017).

6.2 Aim and objectives

In order to identify the genetic changes that might underlie the fitness costs and changes in attachment rate observed in chapter 4, deep sequencing of populations was carried out in chapter 5 and ΦX174 allelic variants specific to each bacterial host were identified. The aim here is to address to what extent some of these allelic variants contribute to fitness costs and differences in attachment rates as a follow up to the observations in chapter 5. This was achieved using site-directed mutagenesis and fitness measures of the mutants produced. Some studies previously measured distributions of fitness effects of single mutations in ΦX174, however, these engineered mutations were chosen randomly along the genome of ΦX174 (Vale, 2012) or as part of stand-alone mutation studies to determine the effect of a specific allele on protein function (Ruboyianes *et al*., 2009; Young *et al*., 2014). In this chapter, fitness effects of mutations were studied on the two susceptible bacterial hosts, *S.* Typhimurium and *E. coli* C. Here, the measured the attachment rate and fitness effects of mutations were measured at the following host-specific sites: 2275 (protein F; *E. coli* C), 1304 (protein F; *S.* Typhimurium) and 3129 (protein H; a shared site between the two host populations). Most allelic variants observed in chapter 5 occurred in genes F and H. Therefore, this chapter focused on evaluating the fitness effects of mutations in these proteins.

The objectives of this chapter were:

- To introduce individual mutations onto a wild-type ΦX174 background through targeted mutagenesis.

- To determine the fitness effects of host-specific (and shared) mutations on *E. coli* C and *S*. Typhimurium via qPCR.

- To measure the rate of attachment of host-specific mutations and shared mutations using qPCR.

6.3 Results and Discussion

6.3.1 Proteins F and H: structures and interactions

Deep sequencing of phage isolates revealed allelic variants that are shared between phage populations adapted to *S*. Typhimurium and *E. coli* C, as well as other variants that were specific to *E. coli* C or to *S.* Typhimurium (summarised in figure 6.2). Here, fitness effects of three allelic variants sites were studied to infer and detect whether these sites contribute to the trade-offs identified in chapter 4, and determine the effect of a single point mutation on protein function. Through PCR-based site directed mutagenesis described in section 2.3.6, three alleles sites were chosen: 1304C – a polymorphic allele found in *S.* Typhimurium-adapted populations, 2275A – a fixed (~96%) allele in most *E. coli* C adapted populations and 3129T – a fixed allele in *S*. Typhimurium, between ~2% and ~25% in *E. coli* C. Both positions 1304 and 2275 are located in gene F, while 3129 is in gene H (figure 6.3).

A mature infectious virion particle consists of a protein coat which encloses a core that contains ssDNA and protein. The virion core contains 60 copies of protein J and one molecule of ssDNA, and the protein coat of ΦX174 contains protein H (12 molecules) and proteins F and G (60 molecules each). Multiple protein-protein and DNA-protein interactions exist in ΦX174. The illustration in figure 6.1C shows the interactions between proteins F, G and J and the DNA – protein backbone, embedded in a complex network. A significant change in DNA or protein sequences may influence function. For instance, a mutation in protein H may influence DNA ejection mechanisms (Marchler-Bauler *et al*., 2017). Many prokaryotic viruses use the tail for genome delivery across host cell walls, some tail-less viruses rely on host organelles, whereas ΦX174 genome transport requires DNA pilot protein H (Molineux and Panja, 2013; Peralta *et al*., 2013 and Sun *et al*., 2013). Protein H of ΦX174 is a pilot protein with one single molecule found on each 12 spikes on the phage capsid. It functions in the ejection of ssDNA viral

genome into host cell's cytoplasm by forming a tube through oligomerisation of H molecules (figures 6.1A, 6.1B). The oligomerised tube does not affect virus assembly but is essential for infectivity with inner diameter that allows its ssDNA chain to be injected (Sun *et al*., 2013). Part of protein H lies outside the capsid and function in host lipopolysaccharide recognition during infection process.

During ΦX174 genome assembly, while forming an icosahedral capsid, the F protein first forms an aggregate of 9S, the G protein is found as a 6S aggregate decorating each of the 12 pentagonal vertices, and both 9S and 6S form a 12S complex structure, then assemble into protein shell of the virion (Fujisawa and Hayashi, 1975). The J protein binds to the complex and packages the ssDNA viral genome, acting as a linker between the icosahedral F proteins (Bernal *et al*., 2004). A complex network of protein - protein and DNA – protein interactions are formed as shown in figure 6.1C. Protein F is the major capsid protein with 60 copies in the virion, and function in host attachment. Also, the complex structure forms a well-conserved hydrophilic channel, required for DNA ejection (McKenna *et al*., 1994).

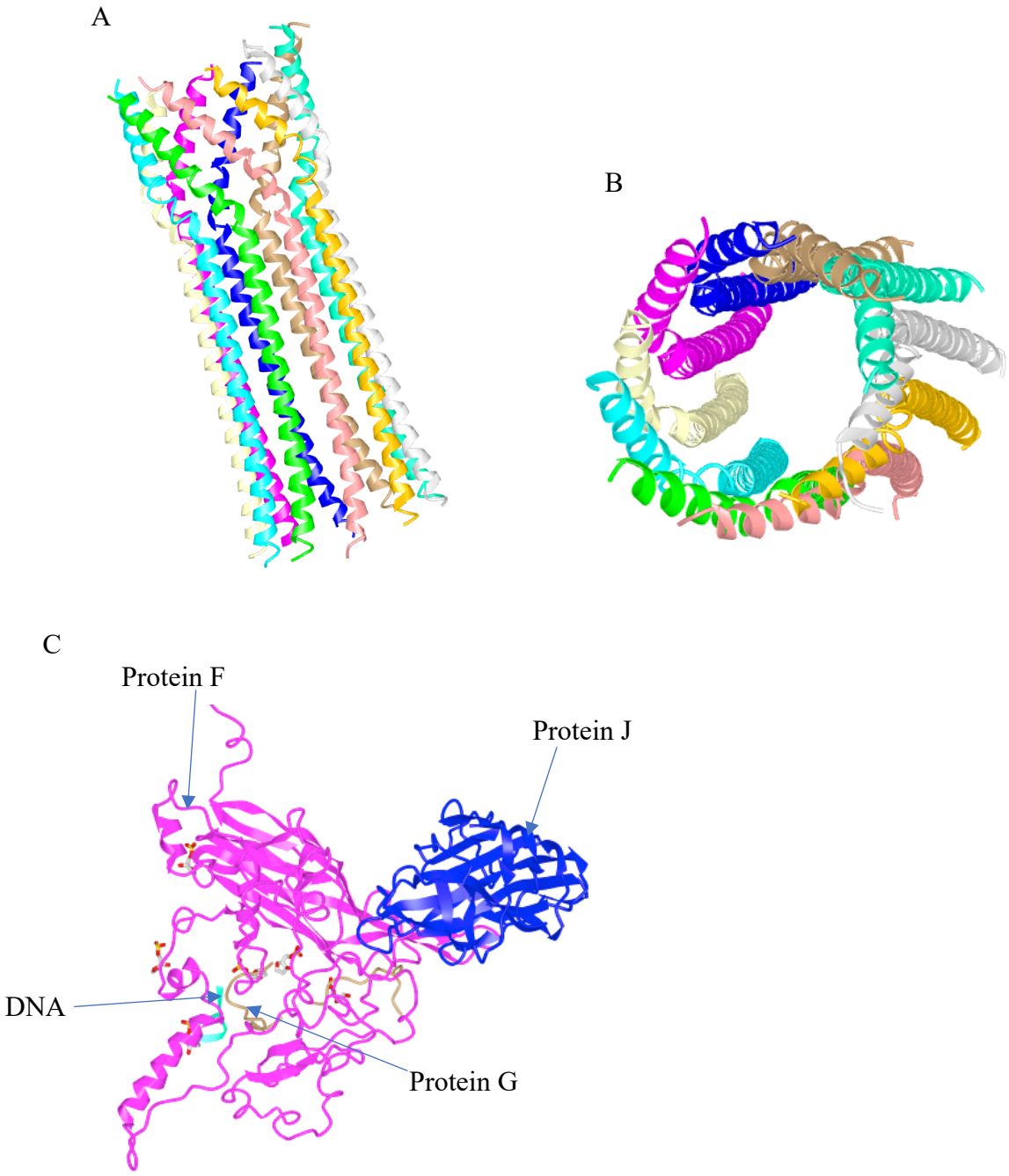
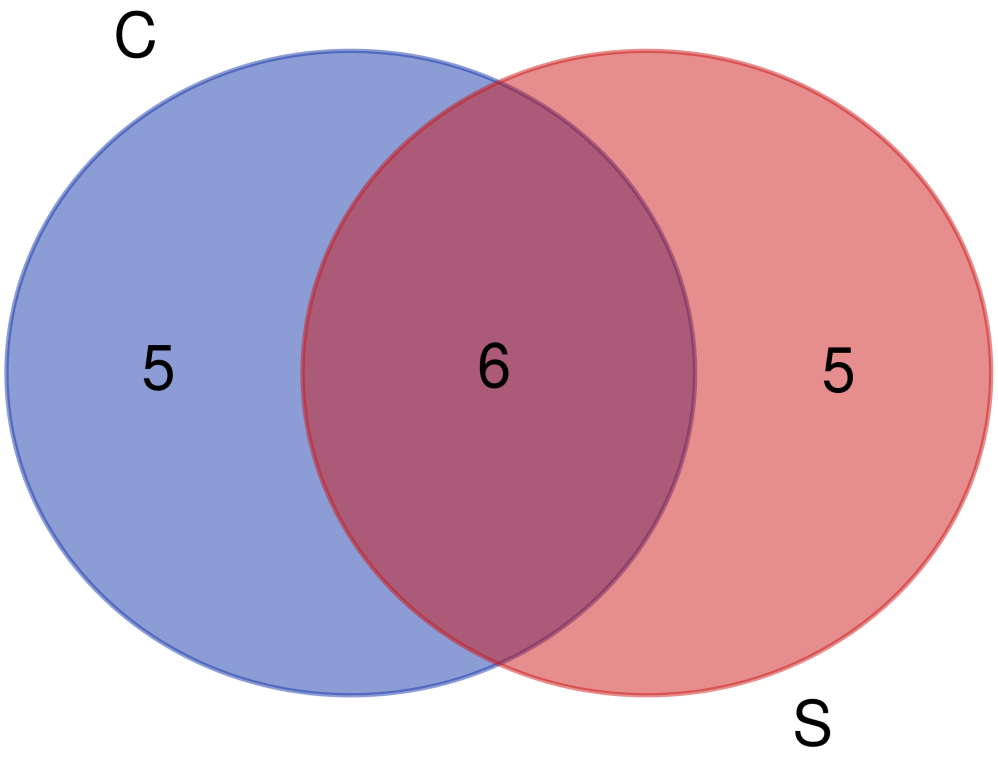


Figure 6.1: H protein oligomerised tube (A) and cross section of the tube (B) showing super-helical coiled-coil of ten alpha-helices. Protein-protein and DNA-protein interactions of F, J and G (C). Images based on Sun *et al.* (2014) for A and B and Bernal *et al.* (2004) for C, visualised and adapted with iCn3D (NCBI).

Two distantly related bacterial hosts (figure 3.5) were used as different environments for ΦX174 adaptation, both differing in lipopolysaccharide structure. The H protein has been well-characterised and studied. Inagaki *et al*., (2003) demonstrated that a prompt decrease in affinity of H protein occurred when residue changes were introduced to the outer R-core of lipopolysaccharide. Site 3129 was chosen as an interesting site for examining the fitness effect and attachment rate changes caused by mutation because it occurred repeatedly on ΦX174 adapted to both hosts despite the difference in liposaccharide (one of the 6 alleles shared between both hosts; figure 6.2) and it has almost fixed allele frequency of ~98% on both hosts.

A



B

C → 1137 1727 2093 2179 2275

S → 648 1304 1548 1956 3071

CS → 1460 2085 2971 3129 3339 5360

Figure 6.2 : A - the total number of shared and likely host-specific alleles observed during deep-sequencing of ΦX174 on bacterial hosts *E. coli* C (C) and *S.* Typhimurium (S). Alleles with frequency ≤ 2 % are excluded. B – Lists of allelic sites: C, seen only in *E. coli* C; S, seen only in *S.* Typhimurium; CS, occurring in both hosts. The circled sites (2275, 1304 and 3129) are those evaluated for fitness effects. Venn diagram produced using software available here: http://bioinformatics.psb.ugent.be/webtools/Venn/.

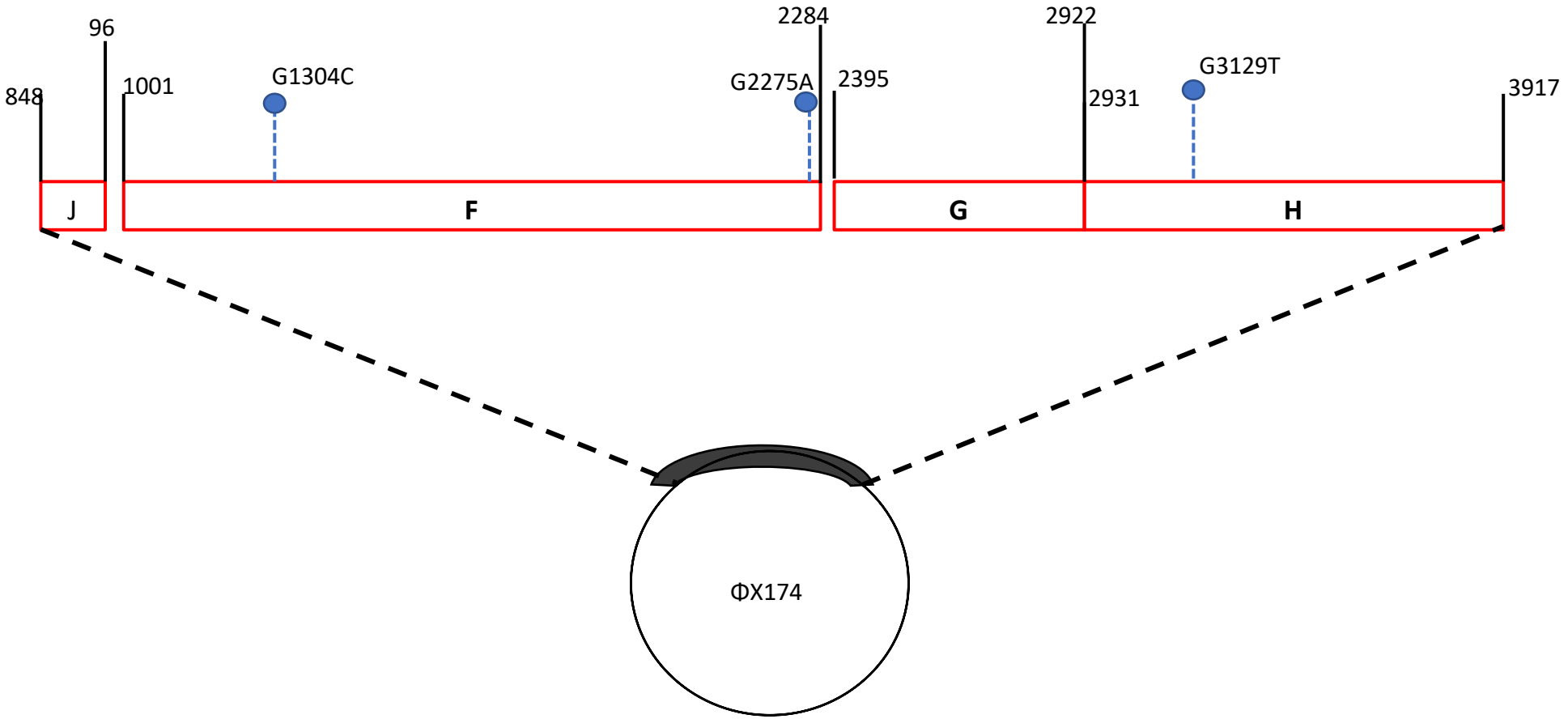


Figure 6.3: Protein F, G and H location in ΦX174 genome. Blue dashed-arrows are the position of mutations introduced through site-directed mutagenesis.

6.3.2 Alleles in gene H: fitness and attachment effects

Previous studies have shown that mutations in ΦX174 of DNA pilot H protein result in defects in host attachment and assembly (Cherwa *et al*., 2010; Young *et al*., 2014). A defect in H protein can mask the functions of other proteins since optimal synthesis of other proteins requires *de novo* biosynthesis of H protein (Ruboyianes *et al*., 2009). It is therefore possible for mutations in H to indirectly influence the entire viral lifecycle, including overall fitness in host cells. Because the H protein recognises and penetrates the host lipopolysaccharide membrane (Jazwinski *et al*., 1975; McKenna and Rossmann, 1994; Suzuki *et al*., 1999; Inagaki *et al*., 2000 and 2003), there is a possibility that it recognises different host lipopolysaccharide in species-specific manner, affecting attachment and infectivity. The relative fitness and attachment rate of 3129T were measured on *S*. Typhimurium and *E. coli* C hosts and the statistical interactions determined using Mann-Whitney test in R (section 2.4). The mean fitness relative to ancestral ΦX174 was not statistically significantly on *S.* Typhimurium than *E. coli* C (ANOVA, p= 0.1; figure 6.4). Moreover, no significant difference (ANOVA, p = 0.1) exists in attachment rate of 3129T (figure 6.5). Cherwa *et al*., (2010) describe interactions between B-H protein, and showed that over-expression of B protein restores assembly but the virions produced are still significantly less infectious. The conclusion from this study may explain why the significant differences in mean fitness (especially on *E. coli* C) did not exhibit the same magnitude differences in attachment rate for 3129T.

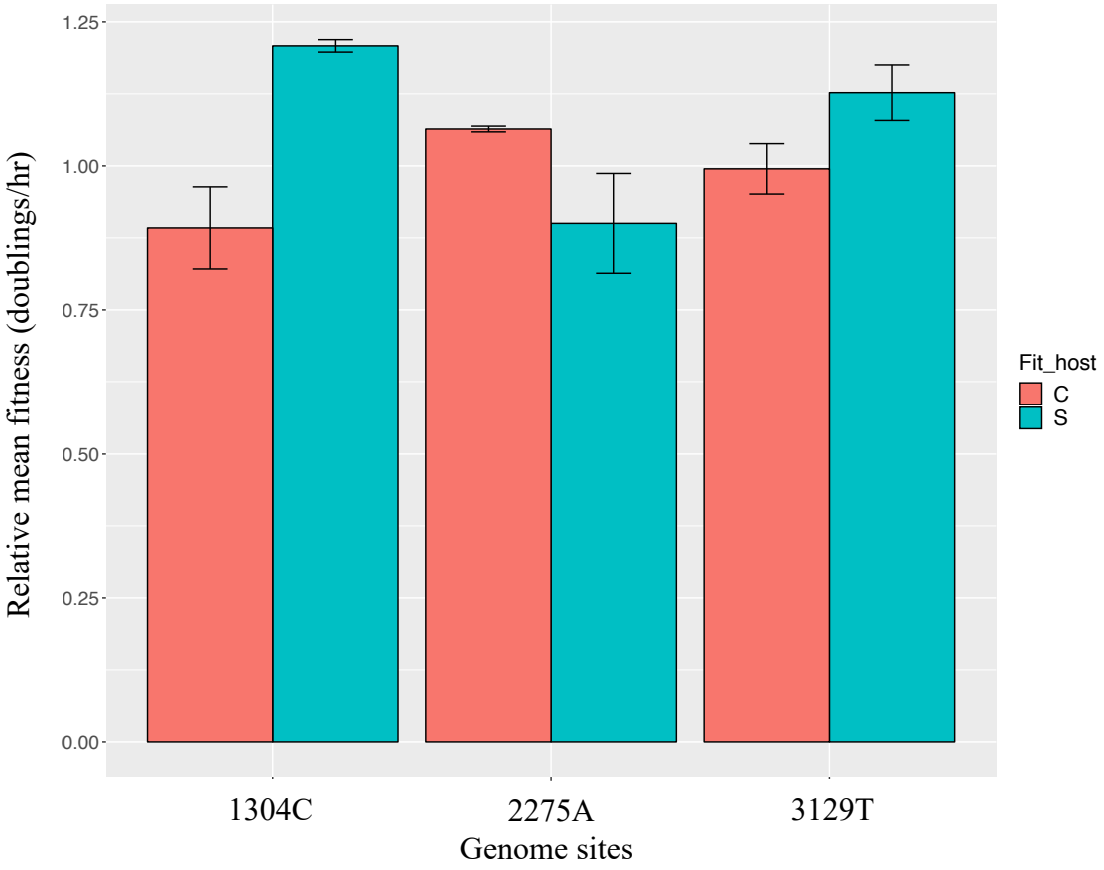


Figure 6.4: Relative fitness of mutant ΦX174 phage. The y axis shows mean fitness (doublings/hour), relative to ancestral ΦX174, measured on *E. coli* C (C; red bars) and *S.* Typhimurium (S; blue bars) for SDM-engineered mutant alleles (identified on the x axis). Site 1304 was changed from G to C, site 2275 from G to A, and site 3129 from G to T. Fitness was measured after 45 minutes of incubation. Triplicate biological replicates (each derived from triplicate technical replicates) are plotted together with 95% confidence intervals of the means.

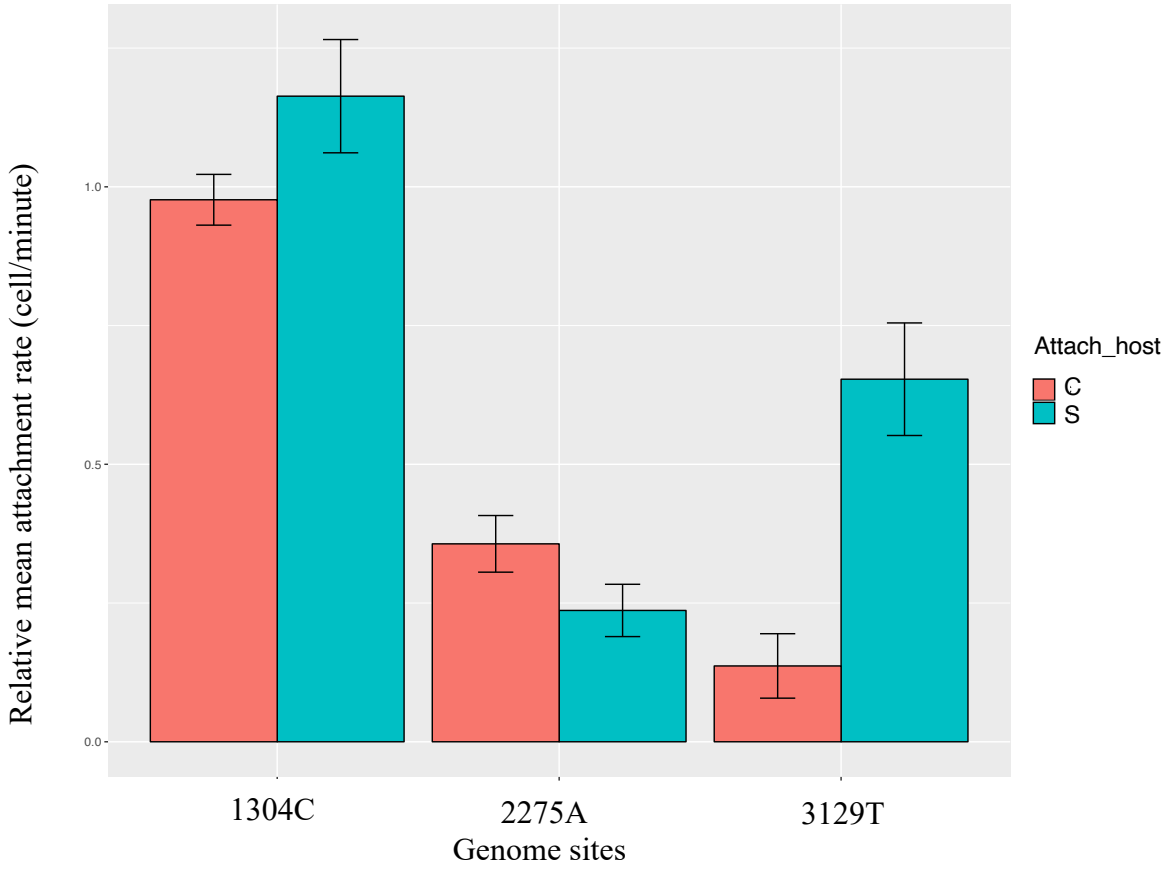


Figure 6.5: Relative attachment rate of mutant ΦX174 phage. The y axis shows mean attachment relative to ancestral ΦX174 measured on *E. coli* C (C; red bars) and *S.* Typhimurium (S; blue bars) for SDM-engineered mutant alleles (identified on the x axis). Site 1304 was changed from G to C, site 2275 from G to A, and site 3129 from G to T. Attachment rate was measured after 8 minutes of incubation. Triplicate biological replicates (each derived from triplicate technical replicates) are plotted together with 95% confidence intervals of the means.

6.3.3 Alleles in gene F: fitness and attachment effects

The major functions of the F protein are host attachment and ejection of viral DNA, hence allelic variation in the F protein is expected to affect attachment rate and possibly fitness. To address this, the fitness effects and attachment rates of likely host-specific mutations 1304C (specific to *S*. Typhimurium) and 2275A (specific to *E. coli* C) both located on protein F were examined using Mann-Whitney test in R (section 2.4). The polymorphic site 1304C may putatively reflect responses to selection pressures within the chemostat. For 2275A, there was no significant difference in attachment rate (relative to ancestral ΦX174) on both hosts (ANOVA, $p = 0.1$), while a statistically significant association was recorded for relative fitness (ANOVA, $p = 0.046$), with higher fitness on *E. coli* C. For the 1304C allele, host-specific in *S*. Typhimurium populations, there is a significant difference for attachment rates measure (ANOVA, $p = 0.046$). Likewise, there was no significant difference in relative fitness of 1304C (ANOVA, $p = 0.10$) on both hosts.

In section 4.3.2, it was demonstrated that trade-off exists with significant performance reduction on *E. coli* C for populations adapted to *S*. Typhimurium. Since, the 1304C allele was probably host-specific to *S*. Typhimurium, it was expected to contribute significantly to the increase in fitness on this host. As expected, higher relative fitness was observed in comparison to *E. coli* C but not comparable to the difference in fitness magnitude recorded in chapter 4 (figure 4.3). Therefore, a single mutation at 1304(C) was not solely responsible for the trade-off magnitude demonstrated in chapter 4 of this research.

6.3.4 Epistatic effects

Epistasis may be defined as a deviation from the sum of independent effects of mutations in a double or multiple mutant. In epistasis, the effect of one gene is influenced by the collective interaction with other genes. Different forms of epistasis have been described depending on the combination of

mutational effects in comparison to an expected outcome calculated based on the separate individual effects (Van den Bergh *et al*., 2018). In general, epistatic interactions may be divided into negative or positive depending on whether the effect decreases or increases fitness, respectively, over the additive expectation. If a mutant has two beneficial mutations, the combination of these mutations may improve fitness more than the sum of the individual fitness effects would predict, likewise, the combination of two deleterious mutations in a mutant may result in more negative effects than expected, such epistasis effects is known as synergistic. In antagonistic epistasis, the combination of two or more mutations may reduce magnitude of change in absolute fitness than expected. For instance, the sum of two beneficial mutations may not enhance fitness as much as the sum of their individual fitness effects, or two deleterious mutations may have less negative effects than expected. In sign epistasis, a beneficial mutation satisfies conditional beneficial mutation where a mutation may be beneficial on some genetic background and deleterious on some (Weinriech *et al*., 2005). Beneficial mutation may acquire a deleterious fitness effect, or a deleterious mutation may lead to a marginal fitness increase on an alternative background or environment (Van den Bergh *et al*., 2018). In a rare form of epistasis, multiple mutations combine to abruptly increase fitness. If a single mutation is missing, the sum of the others will have no effect on fitness, such epistasis effect is termed all-or-none (Van den Bergh *et al*., 2018).

Epistasis may be a good explanation of the difference in magnitude of fitness effects measured for independent alleles versus at population-level, providing a partial explanation for why natural selection does not always give rise to individuals that possess the most-optimal phenotypes. Epistasis of higher order fitness resulting from complex interactions of an organism's genetic context (de Visser *et al*., 2011), may have profound effects on the evolutionary process. In higher-order epistasis, epistasis interactions vary with organism genomic background which has effect on fitness landscape

(Weinriech *et al*., 2018). Fitness landscape is an evolutionary concept that describes the relationship between genotypes and fitness by mapping a set of genotypes to fitness and organising the sets of genotype according to changes in mutation from one form to another (MaCandlish, 2011), a phenomenon that is essential for understanding evolutionary dynamics (Poelwijk *et al*., 2007).

In general, if a population is carrying more than one high frequency mutation, it likely consists of individuals carrying more than one derived/mutant allele. For such individual, beneficial mutations may arise at the same time. Rokyta *et al*. (2011), utilising ssDNA ID11 phage (a wildtype relative of ΦX174), provided evidence that the fitness effects of beneficial double mutations were less than their component beneficial single mutations, resulting in antagonistic mutation effects. In the experiment described in this thesis, six near fixation mutations arose across *S*. Typhimurium-adapted populations (table 5.3). There is a possibility that the combinations of the observed mutations conferred a much lower fitness when measured on *E. coli* C (figure 4.3), relative to the effect of a single near fixed mutation (1304C). In another way, high fitness observed when measured on *S*. Typhimurium (for *S*. Typhimurium-adapted populations) may be a consequence of positive epistasis (synergistic epistasis between beneficial alleles). In synergistic epistasis a combination of mutants confer high fitness, with beneficial mutation having a positive effect. The acquisition of new mutations, complex interactions within/between the species and genomes, the requirement for multiple mutations and the order in which the mutations occur exhibit complex interplay in evolution studies (Meyer *et al*., 2012). The complexity of epistasis effects have been examined by combining both evolutionary and structural analyses. Redondo *et al*. (2016) studied ΦX174 capsid proteins substitutions along q phylogeny (reconstructed using Bayesian phylogenetic methods) and showed that mutations on the capsid did not significantly effect fitness, probably because they are largely neutral. Also, stably folded proteins are very rare in most cases except in an environment

that support the growth of large populations of phages (Wylie and Shakhnovich, 2012). This results in a complicated relationship and difficulty in evaluating correlations between genotype, phenotype, fitness components and/or fitness effects.

Bacteriophage ΦX174 is a good model organism for studying epistatic effects. Its capsids are structurally constrained by amino acids of their constituent proteins and by DNA sequence interactions. Some studies have highlighted epistatic interactions between mutations and with the environment (at the level of fitness and attachment rates) in ΦX174 (Crill *et al*., 2000; Pepin *et al*., 2006 and Pepin and Wichman, 2007). These studies indicate that epistatic effects differed in degree, sign and variability (across environments) which influences evolutionary processes. Crill *et al*. (2000) and Pepin *et al*. (2006) showed that mutations in coat protein F (both studies examined fitness effects resulting from mutations at site 1305, which lies on the same codon as 1304) affected fitness, attachment rate and unidentified phenotypes, and suggested that mutations on protein F have epistatic effects on fitness. The results from the research presented in this thesis also suggest the possibility of epistatic effects on fitness in the context of mutations arising during adaptation. Evidence from figure 6.4 shows a reduction in fitness for ΦX174 population on the *S.* typhimurium-adapted line (1304C), but not as much as fitness trade-offs observed in figure 4.3 of chapter 4. Although more mutagenesis work with combinations of mutations would be required to establish this robustly.

6.4 Conclusion

Bacteriophage ΦX174 is a tail-less small icosahedral with a capsid containing an ssDNA genome that encodes four structural proteins F, G, H and J. These structural proteins function in infection processes including host attachment, DNA ejection and packaging (McKenna *et al*., 1994). The initial steps of viral infection involve recognising a suitable host lipopolysaccharide, adsorption and viral DNA injection, which are all major functions of the structural proteins. It is likely that mutations affecting the initial steps of host LPS recognition may contribute to the ability of ΦX174 to shift host environments. Studies have shown that mutations affecting host cell recognition and kinetics of ΦX174 DNA ejection were identified and isolated from genes F, G and H (Incardona, 1974; Bull *et al*., 1997; Young *et al*., 2014). The majority of allelic variants from deep-sequencing analysis (section 5.3.3, figure 5.6) were found on these structural proteins.

In this chapter, the fitness effects of allelic variants were measured at three sites, one allele identified in populations adapting to *E. coli* C and *S.* Typhimurium in protein H, and other two alleles in protein F, specific to and likely differentially selected for on one host or the other. One major aim of this chapter is determine whether allelic variants contribute to the magnitude of trade-off observed in chapter 4 (figure 4.3). Due to the complexity of genome structural protein and DNA interactions, probably leading to ΦX174 structural epistasis, a single point mutation did not appreciably contribute to mean fitness as observed in figure 4.3 (ANOVA, $p<0.0001$).

However, there was a significant difference in attachment rate most especially for the mutation in protein H. The significant difference on attachment rate is evidence of the protein's function in host attachment. It can also be deduced that recognition and attachment of protein H to host receptor may be species-specific (considering the percentage of allele frequency in *S*. Typhimurium – near fixation; figures 5.10 and 5.11 of chapter

5). In the same way that attachment to host LPS is species-specific, other stages of the infection process may also differ between hosts, leading to selection at the other stages of the life cycle. The possibility could be investigated with whole-transcriptome-based comparative analyses of phage infection *E. coli* C and *S*. Typhimurium populations, to determine different gene expression levels of phage infection in both bacterial host.

Chapter Seven: Conclusions and future directions

In a changing environment, occasional switches in host infection may occur as response to environmental change, (promoting host switching) a process referred to as emergence. Viruses may overcome the limitation of a host environment by switching over to novel hosts. Although it is possible for a population to be established on and adapt to a new host, this often entails a trade-off (Abedon *et al*, 2001; Duff *et al*., 2005; Rodrıguez-Verdugo *et al*., 2014; Wang 2006). In particular, adaptation to a new host can result in lower fitness on the original host (Crill *et al.,* 2000). If switching between hosts occurs regularly this may impose fitness cost (McMullen *et al*., 2017). Host switching has implications for drug resistance mechanisms, vaccine efficacy, pathogenesis and the threat from emerging viral diseases. Working with a host-parasite system is expected to be relevant to understanding these implications.

Although the dynamics of virus evolution can be complex, with recent advances in deep sequencing technologies and molecular methods it is possible to seek answers to evolutionary questions by studying genetic changes occurring over time and in different conditions. The availability of diverse sequencing technologies, and standardised bioinformatics formats and tool kits for handling sequencing data has facilitated the study of viral evolution through deep sequencing.

This project was designed to understand the evolutionary processes occurring in large populations of phage ΦX174 exposed to novel or alternating host environments. As the first DNA genome to be sequenced and artificially synthesised (Sanger *et al*., 1977; Sanger *et al*.,1978; Smith *et al*., 2003), and as a model organism used by many researchers (Bull *et al*., 1997; Crill *et al*., 2000; Wichman *et al*., 2000, 2005; Holder and Bull 2001; Poon and Chao, 2006; Pepin *et al*., 2007; Dickins and Nekrutenko, 2010; Wichman and Brown, 2010; Brown *et al*., 2014; Wilcox, 2017; Redondo *et al*., 2017), this well-understood organism has benefitted the current study. The usual laboratory host *E. coli* C and novel hosts *E. coli* K-12$^{gmhB\text{-mut}}$ or *S.* Typhimurium were utilised in two separate experiments. First, ΦX174 was

grown for 3 days in an *E. coli* K-12$^{gmhB\text{-mut}}$ mutant strain (carrying a mutation affecting biosynthesis of the 1st heptose sugar). Second, ΦX174 was adapted to *E. coli* C and *S.* Typhimurium (carrying a mutation GalE^{-} affecting LPS structure) in an alternating fashion for four consecutive periods, with each period lasting for 10 days on each host (figure 2.1).

Phage ΦX174 was propagated in the different hosts (in both experiments) utilising the continuous culture environment of a bespoke chemostat. An important feature of a chemostat is that microbes can be grown in a steady state under constant environmental conditions. The apparatus also allows manipulation of environmental conditions. In the chemostat designed for this study, an additional chamber (the swamp) was introduced for propagation of ΦX174. Every day samples were taken from the swamp, with ΦX174 separated from bacterial cells debris, and cryoprotected for phenotype and genotype analysis.

The principal aim of this study was to identify a signature of adaptation to novel hosts and detect evidence of adaptive change over time for days 1 and 10, including tracking time series for days 2, 3 and 8. The first step taken towards this was analysing the phenotypes produced by measuring the growth rates (a proxy for fitness) and attachment rates. For both sets of experiments, the rates were measured with qPCR in liquid culture.

In the first experiment, evolution of ΦX174 in the *E. coli* K-12$^{gmhB\text{-mut}}$ host had a profound effect on its fitness, with a large fitness cost when measured on the original host. Eventually, an attempt to experimentally evolve ΦX174 in the host failed after 3 days (~206 generations) even when ΦX174 possess ability to grow in *E. coli* K-12$^{gmhB\text{-mut}}$ (Michel *et al*., 2010) and closely related to *E. coli* C (figure 3.5), presumably reflecting a low fitness in the new environment. The failure of this population to persist may be as a result of chemostat environment used. In a chemostat, bacterial cells are gradually expelled as waste, in addition to phages. If the chemostat flow rate far exceeds reproduction rate of ΦX174 (evidenced from the reduced fitness

observed in the first day), then in the long run ΦX174 will go to extinction and the few progeny produced would have been washed away.

In the second experiment, sustained ΦX174 infection in *S*. Typhimurium was demonstrated through multiple 10-day growth periods alternating between *E. coli* C and *S.* Typhimurium. The infection of ΦX174 in *S*. Typhimurium was successful despite the typical laboratory host *E. coli* C being distantly related to *S*. Typhimurium (figure 4.3). Fitness and attachment assays were carried out to measure phenotypes produced. Here, both assays were carried out on the hosts that a lineage was most recently exposed to as well as on the alternative host. ΦX174 adapted to the *S*. Typhimurium host showed fitness costs when propagated on *E. coli* C. However, the fitness rates measured on *S*. Typhimurium does not decrease on adaptation to *E. coli* C (relative to the result when phages were adapted to *S*. Typhimurium). Interestingly, host switching across the replicate populations consistently exhibited the same pattern. According to Woolhouse *et al* (2005), the probability of successful adaptation depends on primary infections, initial transmissibility of infection in the new host population, the number of mutations or genetic changes required to colonise the new host, as well as the probability of these genetic changes occurring. Since adaptation in *S*. Typhimurium occurred repeatedly but came at a cost, further investigation was undertaken in an attempt to unveil the genetic changes associated with the trade-offs observed.

Deep sequencing was utilised to determine the genetic changes in ΦX174 populations before and after host switching as well as to capture adaptation to the *S*. Typhimurium host through two time series. This approach included internal validation procedures and a range of controls. These include repeated sample DNA preparations (entailing the accumulation of RF I DNA and, in one case, including a crossover host for preparation), addition of plasmid DNA spike-ins in alternate samples to check for cross contamination during library preparation, and a repeat sequencing run, carried out at a different time entirely (also including between- and within-run sample DNA preparations). Using a work-flow that including mapping against spike-in

DNA, the results revealed no evidence for cross-contamination and matching patterns of substitutions in repeat preparation samples. However, disparity in allele frequency was observed when ΦX174 DNA preparation was performed in a crossover host (representing the previous host), with the corresponding host-specific set of mutations observed. In addition, differences in allele frequency were observed between two different sequencing runs. Despite these differences, ~75 % of substitutions recorded were similar to many studies utilising the same set of hosts. Relatively few mutations (1639T, 2085T, 3071A and 1304C) were observed (not present in the ancestor) for ΦX174 adaptations in the new host. When larger numbers of mutations are required to infect a new host, the likelihood of infection may be greatly reduced (Hall *et al*., 2011), this may explain why infection was successful. Moreover, evidence from the literature and substitutions observed in the time series suggest that most changes may be adaptive. The recurrence of the pattern of allelic variants (in two independent lines and across two alternations in each, or resulting from repeated independent preparations in each host) indicates parallel evolution and also provides evidence that the changes may be adaptive.

Some alleles, that were not present in the ancestor, were shared between hosts, while others alleles were present in one host only (and were reversed in the alternative host). It is probable that the latter set of host-specific alleles contributed to fitness trade-offs seen in *E. coli* C *S*. Typhimurium. To investigate further, a subset of these likely host-specific substitutions or shared alleles were examined for their fitness effects on the ancestral background through targeted mutagenesis. The single point mutations did not appreciably contribute to the magnitude of fitness costs in *S*. Typhimurium. This may be as a result of complexity of protein-protein and DNA-protein interactions via epistasis.

Ecological and environmental factors also contribute to viral evolution complexity. Viruses are obligate parasites, the whole or part of their lifecycle

depend on the host they infect. In evolution, viruses interact and evolve in response to their host. In the same way, bacterial hosts evolve in response to phages infection in reciprocal coevolution events. Therefore, in a population, viruses and hosts may co-evolve and co-exist. In the present study, the focus was on phage and the evolution of hosts was not evaluated. We evaluated phage evolution only. Every two days, the chemostat apparatus was discarded and re-inoculated with fresh naïve bacterial cells from stock. This was done in an attempt to minimise host evolution and biofilm formation. Due to the complexity of virus evolution, the evolution of bacteria hosts may occur in this study. For better understanding of ecological and environmental influences on virus-host-switch study, future work may explore the evolution of both host and phage, effects of different flow rates and mixed host environment.

In summary, an important survival challenge of an evolving viral population is the potential to respond to changes in the environment; an instance of this is provided by host switching. Viruses adapting to a novel host may experience fitness costs on the original host. Trade-offs are complex evolutionary scenarios that requires a tight link between phenotypic costs and genotypic changes during pathogen-host adaptation. Examining these is relevant for understanding of emerging infectious diseases, and is expected to contribute to a better knowledge of general constraints, and the costs and benefits, for evolving parasite populations, of adapting to new host environments. The work detailed in this study was achieved by developing a controlled and crossover experimental design to explore viral host switching. It was the first study to utilise *E. coli* K-12$^{gmhB\text{-mut}}$ in an adaptation experiment, examine nucleotide changes occurring in a time series on a novel host, utilise internal controls for sample sequencing preparation in experimental evolution study, and investigate the fitness effects of some host-specific sites (2275A and 3129T).

This study did not conclusively identify the genetic basis of the trade-off observed in switching from *E. coli* C to *S*. Typhimurium. Although, it is likely that the host-specific mutations detected contribute to the costs, gene regulation can also play a central role in ΦX174 adaptation to the *S*. Typhimurium environment. Whole-transcriptome sequencing may reveal more about the genetic basis and evolutionary forces that influence viral adaptation. With whole-transcriptome sequencing, alterations in transcription marking life history transitions can be examined. For instance, the accumulation of transcripts for different gene products involved in DNA replication and capsid morphogenesis may be detected alongside changes in bacterial host transcription.

To further emphasise controls for biases in DNA and library preparations, explicit comparisons of different methods required for ΦX174 DNA may be explored (including PCR-based methods). In this study, I highlight the importance of monitoring the effects of DNA preparation procedure and this warrants further investigation using alternative procedures. As an additional control, droplet digital PCR (ddPCR), a tool for detecting and quantifying nucleic acids with high specificity and sensitivity (Mazaika and Homsy, 2015), may be employed to validate the presence of mutations, as well as confirm and compare the allele frequency with estimates obtained from bioinformatic analysis. Employing ddPCR may allow the measurement of more components of fitness (which entails life cycle history) such as latent period, substantiating measurements of fitness rates and better inferring the associated costs. ddPCR may also be used to explore phage-phage competition (Morella *et al*., 2018) in the chemostat system.

References

Abedon, S. T. (2008). Bacteriophage Ecology: Population growth, Evolution, and impact of Bacterial Viruses. *Advances in Molecular and Cellular Microbiology*. Cambridge University Press.

Abedon, S. T. (2009). Chapter 1 Phage Evolution and Ecology. *Advances in Applied Microbiology, 67,* 1-45.

Abedon, S. T., Herschler, T. D., & Stopar, D. (2001). Bacteriophage latent-period evolution as a response to resource availability. *Applied Environment Microbiology*, *67* (0099-2240), 4233-4241.

Abedon, S. T., Thomas-Abedon, C., Thomas, A., & Mazure, H. (2011). Bacteriophage prehistory: Is or is not Hankin, 1896, a phage reference? *Bacteriophage*, *1*(3),174-178.

Abel, S., Abel zur Wiesch, P., Davis, B. M., & Waldor, M. K. (2015). Analysis of Bottlenecks in Experimental Models of Infection. *PLOS Pathogens*, *11*(6), e1004823.

Agbaje, M., Begum, R. H., Oyekunle, M. A., Ojo, O. E., & Adenubi, O. T. (2011). Evolution of Salmonella nomenclature: A critical note. *Folia Microbiologica*, *56*(6), 497–503.

Anderson, B., Dean, T., Pasternack, G., Rashid, M. H., Carter, C., Senecal, A., Rajanna, C., Revazishvili, T., & Sulakvelidze, A. (2011). Enumeration of bacteriophage particles. *Bacteriophage*, *1*(2), 86–93.

Adhya, S. Merril, C. R., & Biswas, B. (2014). Therapeutic and prophylactic applications of bacteriophage components in modern medicine. *Cold Spring Harbor perspectives in medicine*, *4*(1), a012518.

Andrews, K. R., & Luikart, G. (2014). Recent novel approaches for population genomics data analysis. *Molecular Ecology*, *23*(7), 1661-7.

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

An, R., Jia, Y., Wan, B., Zhang, Y., Dong, P., Li, J., & Liang, X. (2014). Non-Enzymatic Depurination of Nucleic Acids: Factors and Mechanisms. *PLoS ONE*, *9*(12), e115950.

Ambardar, S., Trakroo, D., Lal, R., Gupta, R., & Vakhlu, J., (2016). High Throughput Sequencing: An Overview of Sequencing Chemistry. *Indian Journal of Microbiology*, *56*(4), 394–404.

Anishchenko, M., Weaver, S. C., Paessler, S., Austgen, L., Greene, I. P., & Bowen, R. A. (2006). Venezuelan encephalitis emergence mediated by a phylogenetically predicted viral mutation. *Proceedings of the National Academy of Sciences*, *103*(13), 4994–4999.

Aoyama, A., & Hayashi, M. (1986). Synthesis of bacteriophage φ{symbol}X174 in vitro: Mechanism of switch from DNA replication to DNA packaging. *Cell*, *47*(1), 99–106.

Arai, K., & Kornberg, A. (1981). Unique primed start of phage phi X174 DNA replication and mobility of the primosome in a direction opposite chain synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, *78*(1), 69-73.

Araujo, S. B. L., Braga, M. P., Brooks, D. R., Agosta, S. J., Hoberg, E. P., Von Hartenthal, F. W., & Boeger, W. A. (2015). Understanding host-switching by ecological fitting. *PLoS ONE*, *10*(10), 1-17.

Ari, S. & Arikan, M. (2016). Next Generation Sequencing:Advantages, Disadvantages and Future. *In "Plant Omics-Trends and Applications*, 109 – 136.

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Takai, Y., Datsenko, K. A., Tomita, M., Wanner, B. L., & Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: The Keio collection. *Molecular Systems Biology*, *2*, 1-11.

Baize, S., Pannetier, D., Oestereich, L., Rieger, T., Koivogui, L., Magassouba, N'Faly, Soropogui, B., Sow, M. S., Keïta, S., De Clerck, H., Tiffany, A., Dominguez, G., Loua, M., Traoré, A., Kolié, M., Malano, E. R., Heleze, E., Bocquin, A., Mély, S., Raoul, H., Caro, V., Cadar, D., Gabriel, M., Formenty, P., Kolié, M., Van Herp, M., Tappe, D., Cadar, D., Gabriel, M., Schmidt-Chanasit, J., Impouma, B., Diallo, A. K., Van Herp, M., & Günther, S. (2014). Emergence of Zaire Ebola Virus Disease in Guinea. *New England Journal of Medicine*, *371*(15), 1418–1425.

Baker, C. W., Thaweethai, T., Baker, M. H., Yuan, J., Joyce, P., Miller, C. R., & Weinreich, D. M. (2016). Genetically Determined Variation in Lysis Time Variance in the Bacteriophage ϕX174. *GenesIGenomesIGenetics*, *6*(4), 939–955.

Benler, S., Cobián-Güemes, A. G., McNair, K., Hung, S. H., Levi, K., Edwards, R., & Rohwer, F. (2018). A diversity-generating retroelement encoded by a globally ubiquitous Bacteroides phage. *Microbiome*, *6*(1), 191.

Battistuzzi, F. U., Feijao, A., & Hedges, S. B. (2004). A genomic timescale of prokaryote evolution: Insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evolutionary Biology*, *4*.

Barr, J. J., Auro, R., Furlan, M., Whiteson, K. L., Erb, M. L., Pogliano, J., Stotland, A., Wolkowicz, R., Cutting, A. S., Doran, K. S., Salamon, P., Youle, M., & Rohwer, F. (2013). Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proceedings of the National Academy of Sciences*, *110*(26), 10771–10776.

Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., Schneider, D., & Lenski, R. E. (2009). Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature, 461*, 1243–1247.

Barton N. H. (2000). Genetic hitchhiking. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *355*(1403), 1553-62.

Bernal, R. A., Hafenstein, S., Esmeralda, R., Fane, B. A., & Rossmann, M. G. (2004). The ϕX174 protein J mediates DNA packaging and viral attachment to host cells. *Journal of Molecular Biology*, *337*(5), 1109–1122.

Bernhardt, T. G., Roof, W. D., & Young, R. (2002). Genetic evidence that the bacteriophage phi X174 lysis protein inhibits cell wall synthesis. *Proceedings of the National Academy of Sciences*, *97*(8), 4297–4302.

Bertozzi, S. J., Storms, Z., & Sauvageau, D. (2016). Host receptors for bacteriophage adsorption. *FEMS Microbiology Letters, 363*(4), 1-11.

Bohannan, B. J. M., Kerr, B., Jessup, C. M., Hughes, J. B., & Sandvik, G. (2002). Trade-offs and coexistence in microbial microcosms. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, *81*(1–4), 107–115.

Bollback, J. P., & Huelsenbeck, J. P. (2007). Clonal interference is alleviated by high mutation rates in large populations. *Molecular Biology and Evolution*, *24*(6), 1397–1406.

Bollback, J. P., & Huelsenbeck, J. P. (2009). Parallel genetic evolution within and between bacteriophage species of varying degrees of divergence. *Genetics*, *181*(1), 225–234.

Brown, D. T., Mackenzie, J. M., & Bayer, M. E. (1971). Mode of Host Cell Penetration by Bacteriophage ϕX174. *Journal of Virology*, *7*(6), 836-846.

Brown, C. J., Millstein, J., Williams, C. J., & Wichman, H. A. (2013). Selection Affects Genes Involved in Replication during Long-Term Evolution in Experimental Populations of the Bacteriophage ϕX174. *PLoS ONE*, *8*(3), e60401.

Brown, C. J., Zhao, L., Evans, K. J., Ally, D., & Stancik, A. D. (2010). Positive selection at high temperature reduces gene transcription in the bacteriophage X174. *BMC Evolutionary Biology*, *10*(378), 1-12.

Buck, C. B., Van Doorslaer, K., Peretti, A., Geoghegan, E. M., Tisza, M. J., An, P., Katz, J. P., Pipas, J. M., McBride, A. A., Camus, A. C., McDermott, A. J., Dill, J. A., Delwart, E., Ng, T. F. F., Farkas, K., Austin, C., Kragerger, S., Davison, W., Pastrana, D. V., & Varsani, A. (2016). The Ancient Evolutionary History of Polyomaviruses. *PLOS Pathogens*, *12*(4), 1-26.

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

Bull, A. T. (2010). The renaissance of continuous culture in the post-genomics age. *Journal of Industrial Microbiology and Biotechnology*.

Bull, J. J., Badgett, M. R., & Wichman, H. A. (2000). Big-benefit mutations in a bacteriophage inhibited with heat. *Molecular Biology and Evolution*, *17*(6), 942–950.

Bull, J. J., Badgett, M. R., Wichman, H. A., Huelsenbeck, J. P., Hillis, D. M., Gulati, A., Ho, C., & Molineux, I. J. (1997). Exceptional convergent evolution in a virus. *Genetics*, *147*(4), 1497-507.

Bull, J. J., Millstein, J., Orcutt, J., & Wichman, H. A. (2006). Evolutionary Feedback Mediated through Population Density, Illustrated with Viruses in Chemostats. *The American Naturalist*, *167*(2), E39–E51.

Burch, C. L., & Chao, L. (1999). Evolution by small steps and rugged landscapes in the RNA virus ϕ6. *Genetics 151,* 921–927.

Chapnik, N., Sherman, H., & Froy, O. (2007). A one-tube site-directed mutagenesis method using PCR and primer extension. *Analytical Biochemistry*, *372*(2), 255-257.

Chen, Y.-C., Yu, C.-H., Chiang, T.-Y., Liu, T., &. Hwang, C.-C. (2013). Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. *PLoS ONE*, *8*(4), e62856.

Cherwa, J. E., Organtini, L. J., Ashley, R. E., Hafenstein, S. L., & Fane, B. A. (2011). In vitro assembly of the øx174 procapsid from external scaffolding protein oligomers and early pentameric assembly intermediates. *Journal of Molecular Biology*, *412*(3), 387–396.

Cherwa, J. E., Tyson, J., Prevelige, P. E., Bedwell, G. J., Edwards, A. G., Dokland, T., Brooke, D. & Fane, B. A. (2016). ϕX174 Procapsid Assembly:

Effects of an Inhibitory External Scaffolding Protein and Resistant Coat Proteins In Vitro. *Journal of Virology*, *91*(1), e01878-16.

Cherwa, J. E., Young, L. N., & Fane, B. A. (2011). Uncoupling the functions of a multifunctional protein: The isolation of a DNA pilot protein mutant that affects particle morphogenesis. *Virology*, *411*(1), 9–14.

Prabhakaran, R., Chithambaram, S., & Xia, X. (2014). The effect of mutation and selection on codon adaptation in Escherichia coli bacteriophage. *Genetics*, *197*(1), 301–315.

Chou, H. H., Chiu, H. C., Delaney, N. F., Segrè, D. & Marx, C. J. (2011). Diminishing Returns Epistasis Among Beneficial Mutations Decelerates Adaptation. *Science, 332*, 1190-1192.

Chua, K.B. (2000). Nipah virus: A recently emergent deadly paramyxovirus. *Science*, *288*(5470), 1432-1435.

Clokie, M. R.J. & Kropinski, A. (2009). Methods and Protocols, Volume 1: Isolation, Characterization, and Interactions. “Methods in molecular biology”. *Humana press*, 69 – 81.

Clokie, M. R. J., Millard, A. D., Letarov, A. V., & Heaphy, S. (2011). Phages in nature. *Bacteriophage*, *1*(1), 31–45.

Coetzee J. In: Phage Ecology. Goval S M, Gerba C, Bitton G, editors. New York: Wiley; 1987. pp. 45–85.

Coffey, L. L., Vasilakis, N., Brault, A. C., Powers, A. M., Tripet, F., & Weaver, S. C. (2008). Arbovirus evolution *in vivo* is constrained by host alternation. *Proceedings of the National Academy of Sciences USA*, *105*(19), 6970–6975.

Cornely, O. A., Bethe, U., Pauls, R., & Waldschmidt, D. (2002). Peripheral Teflon Catheters: Factors Determining Incidence of Phlebitis and Duration of Cannulation. *Infection Control & Hospital Epidemiology*, *23*(5), 249–253.

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

Crill, W. D., Wichman, H. A., & Bull, J. J. (2000). Evolutionary reversals during viral adaptation to alternating hosts. *Genetics*, *154*(1), 27-37.

Cuevas, J. M., Domingo-Calap, P., & Sanjuán, R. (2012). The fitness effects of synonymous mutations in DNA and RNA viruses. *Molecular Biology and Evolution*, *29*(1), 17–20.

Dagan, T., Talmor, Y., & Graur, D. (2002). Ratios of radical to conservative amino acid replacement are affected by mutational and compositional factors and may not be indicative of positive Darwinian selection. *Molecular Biology and Evolution*, *19*(7), 1022–1025.

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

Dennehy, J. J. (2014). What Ecologists Can Tell Virologists. *Annual Review of Microbiology*, *68*(1), 117–135.

Dennehy, J. J. (2009). Bacteriophages as model organisms for virus emergence research. *Trends in Microbiology*, *17*(10), 450-457.

Dessau, M., Goldhill, D., McBride, R. L., Turner, P. E., & Modis, Y. (2012). Selective Pressure Causes an RNA Virus to Trade Reproductive Fitness for Increased Structural and Thermal Stability of a Viral Enzyme. *PLoS Genetics*, *8*(11).

De Sordi, L., Khanna, V., & Debarbieux, L. (2017). The Gut Microbiota Facilitates Drifts in the Genetic Diversity and Infectivity of Bacterial Viruses. *Cell Host and Microbe*, *22*(6), 801-808.e3.

De Sordi, L., Lourenço, M. & Debarbieux, L. (2019). "I will survive": A tale of bacteriophage-bacteria coevolution in the gut. *Gut microbes*, *10*(1), 92-99.

Devaux, C. A. (2012). Emerging and re-emerging viruses: A global challenge illustrated by Chikungunya virus outbreaks. *World Journal of Virology*, *1*(1), 11-22.

de Visser, J.A.G.M., Cooper, T.F., & Elena, S.F. (2011). The causes of epistasis. *Proceedings of the Royal Society B: Biological Sciences, 278*, 3617–3624.

Díaz-Muñoz, S. L. (2017). Viral coinfection is shaped by host ecology and virus–virus interactions across diverse microbial taxa and environments. *Virus Evolution*, *3*(1), vex011.

Dickins, B., & Nekrutenko, A. (2009). High-Resolution Mapping of Evolutionary Trajectories in a Phage. *Genome Biology and Evolution*, *1*, 294–307.

Dickins, B., Rebolledo-Jaramillo, B., Su, M. S. W., Paul, I. M., Blankenberg, D., Stoler, N., Ne Stoler, N., Makova, K. D., & krutenko, A. (2014). Controlling for contamination in re-sequencing studies with a reproducible web-based phylogenetic approach. *BioTechniques*, *56*(3), 134–141.

Diehl, W. E., Lin, A. E., Grubaugh, N. D., Carvalho, L. M., Kim, K., Kyawe, P. P., McCauley, S. M., Donnard, E., Kucukural, A., McDonel, P., Schaffner, S. F., Garber, M., Rambaut, A., Andersen, K. G., Sabeti, P. C., & Luban, J. (2016). Ebola Virus Glycoprotein with Increased Infectivity Dominated the 2013–2016 Epidemic. *Cell*, *167*(4), 1088-1098.

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

Doulatov, S., Hodes, A., Dai, L. Mandhana, N., Liu, M., Deora, R., Simons, R. W., Zimmerly, S. & Miller, J. F. (2004). Tropism switching in *Bordetella* bacteriophage defines a family of diversity-generating retroelements. *Nature* 431**,** 476–481.

Dykhuizen D. E., & Dean A. M. (2004), Evolution of specialists in an experimental microcosm. *Genetics*, *167*, 2015–2026.

Duarte, E., Clarke, D., Moyat, A., Domingo, E., & Holland, J. (1992). Rapid fitness losses in mammalian RNA virus clones due to Muller's ratchet (RNA virus mutation/virus populations/replicative competition of virus clones/vesicular stomatitis virus). *Genetics*, *89*, 6015–6019.

Duffy, S., Burch, C. L., & Turner, P. E. (2007). Evolution of host specificity drives reproductive isolation among RNA viruses. *Evolution*, *61*(11), 2614–2622.

Duffy, S., Turner, P. E., & Burch, C. L. (2006). Pleiotropic costs of niche expansion in the RNA bacteriophage Φ6. *Genetics*, *172*(2), 751–757.

Dykhuizen, D. E. (1990). Experimental studies of natural-selection in bacteria. *Annual Review of Ecology and Systematics*, *21*, 373-398.

Edelman, D. C. & Barletta, J. (2003). Real-time PCR provides improved detection and titer determination of bacteriophage. *BioTechniques*, *35*(2), 368-375.

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.

Eisenberg, S. (1980). The role of gene A protein and *E. coli rep* protein in the replication of ϕX 174 replicative form DNA. *Proceedings of the Royal Society of London. Series B. Biological Sciences, 210*(1180).

Ekechukwu, M. C., Oberste, D. J., & Fane, B. A. (1995). Host and PhiX174 Mutations Affecting the Morphogenesis or Stabilization of the 50s Complex, a Single-Stranded. *Genetics*, *140*, 1167–1174.

Eko, F. O., Witte, A., Huter, V., Kuen, B., Fürst-Ladani, S., Haslberger, A., Katinger, A., Hensel, A., Szostak, M. P., Resch, S., Mader, H., Raza, P., Brand, E., Marchart, J., Jechlinger, W., Haidinger, W., & Lubitz, W. (1999). New strategies for combination vaccines based on the extended recombinant bacterial ghost system. In *Vaccine*, 17, 1643–1649.

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048.

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

Faillace, C. A., Lorusso, N. S., & Duffy, S. (2017). Overlooking the smallest matter: viruses impact biological invasions. *Ecology Letters*, *20*(4), 524-538.

ΦX174 chapter - Fane (2006).pdf. (n.d.).

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*. 129–148.

Feige, U. (1981). The lipopolysaccharide of Escherichia coli C studies on the anomeric configurations of the hexoses in the R1 core. *Zentralblatt fur Bakteriologie, Mikrobiologie und Hygiene*, 250(1-2), 52-62.

Feige, U., & Strim, S. (1976). On the structure of the Escherichia coli C cell wall lipopolysaccharide core and on its fX174 receptor region. *Biochem. Biophys. Res. Commun*. *71*, 566–573.

Ferenci, T. (2007). Bacterial Physiology, Regulation and Mutational Adaptation in a Chemostat Environment. *Advances in Microbial Physiology*. Academic Press.

Ferris, M. T., Joyce, P., & Burch, C. L. (2007). High frequency of mutations that expand the host range of an RNA virus. *Genetics*, *176*(2), 1013–1022.

Fischer, C. R., Yoichi, M., Unno, H., & Tanji, Y. (2004). The coexistence of Escherichia coli serotype O157:H7 and its specific bacteriophage in continuous culture. *FEMS Microbiology Letters*, *241*(2), 171–177.

Fujisawa, H., & Hayashi, M. (1976). Viral DNA-Synthesising Intermediate Complex Isolated During Assembly of Bacteriophage ΦX174. *Journal of Virology*, *19*(2), 409–415.

Furuse, Y., Suzuki, A., & Oshitani, H. (2010). Origin of measles virus: Divergence from rinderpest virus between the 11th and 12th centuries. *Virology Journal*, *7*(52), 1-4.
Gallet, R., Lenormand, T., & Wang, I. N. (2012). Phenotypic Stochasticity Protects Lytic Bacteriophage Populations From Extinction During The Bacterial Stationary Phase. *Evolution*, *66*(11), 3485–3494.

Gallet, R., Shao, Y., & Wang, I. N. (2009). High adsorption rate is detrimental to bacteriophage fitness in a biofilm-like environment. *BMC Evolutionary Biology*, *9*(241).

García-Villada, L., & Drake, J. W. (2013). Experimental selection reveals a trade-off between fecundity and lifespan in the coliphage Qß. *Open Biology*, *3*(6), e130043.

Garrison, E. & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. Available at http://arxiv.org/abs/1207.3907, [accessed 11, March 2019].

Gerrish P.J., & Hengartner N. (2017). Inferring the Distribution of Fitness Effects (DFE) of Newly-Arising Mutations Using Samples Taken from Evolving Populations in Real Time. *Algorithms for Computational Biology*, *10252*, 103-114.

Gerrish, P.J. & Lenski, R.E. (1998). The fate of competing beneficial mutations in an asexual population. *Genetica, 102-103*(1-6), 127-44.

Geoghegan, J. L., Duchêne, S., & Holmes, E. C. (2017). Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLoS Pathogens*, *13*(2), 1006215.

Gillam, S., Atkinson, T., Markham, A., & Smith, M. (1985). Gene K of bacteriophage phi X174 codes for a protein which affects the burst size of phage production. *Journal of Virology*, *53*(2), 708–9.

Good, B. H., McDonald, M. J., Barrick, J. E., Lenski, R. E., & Desai, M. M. (2017). The dynamics of molecular evolution over 60,000 generations. *Nature*, *551*(7678), 45–50.

Godson, G. N., & Vapnek, D. (1973). A simple method of preparing large amounts of ΦX174 RF I supercoiled DNA. *BBA Section Nucleic Acids And Protein Synthesis*, *299*(4), 516–520.

Goldhill, D.H., and Turner, P.E. (2014). The evolution of life history trade-offs in viruses. *Current Opinion in Virology 8*, 79–84.

Goldman, D., & Domschke, K. (2014). Making sense of deep sequencing. *International Journal of Neuropsychopharmacology*, *17*(10), 1717-25.

Gordo, I., Perfeito, L., Sousa, A. (2011). Fitness Effects of Mutations in Bacteria. *J Molecular Microbiology Biotechnology*, *21*, 20-35.

Górski, A., Międzybrodzki, R., Weber-Dąbrowska, B., Fortuna, W., Letkiewicz, S., Rogóż, P., Jończyk-Matysiak, E., Dąbrowska, K., Majewska, J., & Borysowski, J. (2016). Phage Therapy: Combating Infections with Potential for Evolving from Merely a Treatment for Complications to Targeting Diseases. *Frontiers in Microbiology*, *7*(1515).

Goulian, M., & Kornberg, A. (1967). Enzymatic synthesis of DNA XXIV Synthesis of infectious phage Φx174 DNA. *Proceedings of the Natural Adcademy Science United State of America, 58*(6), 2321-2328.

Gratia, A., 1936. Des relations numeriques entre bacteries lysogenes et particules de bacteriophage. *Ann. Inst. Pasteur*, *57*, pp.652-676.

Gresham, D., & Dunham, M. J. (2014). The enduring utility of continuous culturing in experimental evolution. *Genomics*, *104*(6), 399–405.

Griffiths AJF, Miller JH, Suzuki DT., Lewontin R.C., and Gelbart W. M. (2000) An Introduction to Genetic Analysis. 7th edition. *New York*.

Guan, Y., Zheng, B. J., He, Y. Q., Liu, X. L., Zhuang, Z. X., Cheung, C. L., Luo, S. W., Li, P. H., Zhang, L. J., Guan, Y. J., Butt, K. M., Wong, K. L., Chan, K. W., Lim, W., Shortridge, K. F., Yuen, K. Y., Peiris, J. S. & Poon, L. L. (2003). *Science, 302*(5643), 276-8.

Gupta, R. M., & Musunuru, K. (2014). Expanding the genetic editing tool kit: ZFNs, TALENs, and CRISPR-Cas9.(zinc finger nucleases, transcription activator-like effector nucleases, and clustered regularly interspaced short palinromic repeats-associated systems)(Report). *Journal of Clinical Investigation*, *124*(10), 4154.

Hafenstein, S., & Fane, B. A. (2002). phi X174 genome-capsid interactions influence the biophysical properties of the virion: evidence for a scaffolding-like function for the genome during the final stages of morphogenesis. *Journal of virology*, *76*(11), 5350-6.

Haggard-Ljungquist, E., Halling, C., & Calendar, R. (1992). DNA sequences of the tail fiber genes of bacteriophage P2: Evidence for horizontal transfer of tail fiber genes among unrelated bacteriophages. *Journal of Bacteriology*, *174*(5), 1462–1477.

Hall, A. R., Scanlan, P. D., & Buckling, A. (2011). Bacteria-Phage Coevolution and the Emergence of Generalist Pathogens. *The American Naturalist*, *177*(1), 44–53.

Harada, L. K., Silva, E. C., Campos, W. F., Del Fiol, F. S., Vila, M., Dąbrowska, K., Krylov, V. N., & Balcão, V. M. (2018). Biotechnological applications of bacteriophages: State of the art. *Microbiological Research*, *212-213*, 38-58.

Hayashi, M., Aoyama, A., Richardson, D. L., &. Hayashi M. N.(1988). Biology of the bacteriophage phiX174, in The Bacteriophages, Vol. 2, edited by R. Calendar. Plenum, New York, 1-71.

Heilbron, K., Toll-Riera, M., Kojadinovic, M., & MacLean, R. C. (2014). Fitness is strongly influenced by rare mutations of large effect in a microbial mutation accumulation experiment. *Genetics*, *197*(3), 981–990.

Heinrichs, D. E., Monteiro, M. A., Perry, M. B., & Whitfield, C. (1998). The assembly system for the lipopolysaccharide R2 core-type of Escherichia coli is a hybrid of those found in Escherichia coli K-12 and Salmonella enterica. Structure and function of the R2 WaaK and WaaL homologs. *Journal of Biological Chemistry*, *273*(15), 8849–8859.

Hendrix, R. W., Smith, M. C. M., Burns, R. N., Ford, M. E., & Hatfull, G. F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(5), 2192–2197.

Henry, M. & Debarbieux, L. (2012). Tools from viruses: Bacteriophage successes and beyond. *Virology*, *434*(2), 151-161.

Holder, K. K. & Bull, J. J. (2001). Profiles of adaptation in two similar viruses. *Genetics*, *159*(4), 1393-1404.

Hill, W.G., and Robertson, A. (2008). The effect of linkage on limits to artificial selection. *Genetics Research 89*, 311–336.

Holder, K. K. & Bull, J. J. (2001). Profiles of adaptation in two similar viruses. *Genetics*, *159*(4), 1393-1404.

Holmes, E. C., Ghedin, E., Miller, N., Taylor, J., Bao, Y., St George, K., Grenfell, B. T., Salzberg, S. L., Fraser, C. M., Lipman, D. J., & Tauben-berger J. K. (2005). Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *PLoS Biology*. *3*, e300.

Hone, D., Morona, R., Attridge, S., & Hackett, J. (1987). Construction of defined galE mutants of Salmonella for use as vaccines. *The Journal of infectious diseases*, *156*(1), 167-174.

Huang, J. X., Bishop-Hurley, S. L., & Cooper, M. A. (2012). Development of anti-infectives using phage display: biological agents against bacteria, viruses, and parasites. *Antimicrobial agents and chemotherapy*, *56*(9), 4569-4582.

Hueffer, K., Parker, J. S. L., Weichert, W. S., Geisel, R. E., Sgro, J. Y., & Parrish, C. R. (2003). The natural host range shift and subsequent evolution of canine par- ovirus resulted from virus-specific binding to the canine transferrin receptor. *Journal of Virology*. *77*, 1718–1726.

Huisman, J. & Wessing, F. J. (2001). Biological conditions for oscilla- tions and chaos generated by multispecies competition. *Ecology*, *82*, 2682–2695.

Hunt, R. C., Simhadri, V. L., Iandoli, M., Sauna, Z. E., & Kimchi-Sarfaty, C. (2014). Exposing synonymous mutations. *Trends in Genetics*, *30*(7), 308-21.

Hutchison, C. A. (2007). DNA sequencing: Bench to bedside and beyond. *Nucleic Acids Research*, *35*(18), 6227–6237.

Hutchison, C.A., Phillips, S., Edgell, M.H., Gillam S., Jahnke P., & Smith M. (1978). Mutagenesis at a specific position in a DNA sequence. *Journal of Biological*, *253*(18), 6551-60.

Hutchison, C.A. & Sinsheimer, R.L. (1963) Kinetics of bacteriophage release by single cells of ϕX174-infected *E. coli*. *Journal of molecular biology*, *7*(2), 206-208.

Hu, B., Margolin, W., Molineux, I. J., & Liu, J. (2013). The bacteriophage t7 virion undergoes extensive structural remodeling during infection. *Science (New York, N.Y.)*, *339*(6119), 576-579.

Ilag, L. L., & Incardona, N. L. (1993). Structural Basis for Bacteriophage ΦX174 Assembly and Eclipse as Defined by Temperature-Sensitive Mutations. *Virology 196*(2), 758-768.

Inagaki, M., Kawaura, T., Wakashima, H., Kato, M., Nishikawa, S., & Kashimura, N. (2003). Different contributions of the outer and inner R-core residues of lipopolysaccharide to the recognition by spike H and G proteins of bacteriophage ϕX174. *FEMS Microbiology Letters*, *226*(2), 221–227.

Inagaki, M., Tanaka, A., Suzuki, R., Wakashima, H., Kawaura, T., Karita, S., Nishikawa, S., & Kashimura, N. (2000). Characterization of the Binding of Spike H Protein of Bacteriophage ΦX174 with Receptor Lipopolysaccharides, *The Journal of Biochemistry*, 127(4), 577–583.

Inouye, M. (2016). The first application of site-directed mutagenesis using oligonucleotides for studying the function of a protein. *Gene*, *593*, 342-343.

Jalava, K., Hensel, A., Szostak, M., Resch, S., and Lubitz, W. (2002). Bacterial ghosts as vaccine candidates for veterinary applications. *Journal of Controlled Release, 85*, 17–25.

Jansson, P. -E, Wollin, R., Bruse, G. W., & Lindberg, A. A. (1989). The conformation of core oligosaccharides from Escherichia coli and Salmonella typhimurium lipopolysaccharides as predicted by semi-empirical calculations. *Journal of Molecular Recognition*, *2*(1), 25–36.

Jaschke, P. R., Lieberman, E. K., Rodriguez, J., Sierra, A., & Endy, D. (2012). A fully decompressed synthetic bacteriophage øX174 genome assembled and archived in yeast. *Virology*, *434*(2), 278–284.

Jazwinski, S. M., Lindberg, A. A., & Kornberg, A. (1975). The gene H spike protein of bacteriophages ØX174 and S13: I. Functions in phage-receptor recognition and in transfection. *Virology*, *66*(1), 283-293.

Jeong, H., Lee, S. J., & Kim, P. (2016). Procedure for adaptive laboratory evolution of microorganisms using a chemostat. *Journal of Visualized Experiments*, *2016*(115).

Kula, A., Saelens, J., Cox, J., Schubert, A. M., Travisano, M., & Putonti, C. (2018). The Evolution of Molecular Compatibility between Bacteriophage Φx174 and its Host. *Scientific Reports*, *8*(1), 8350.

Kawaura, T., Inagaki, M., Kato, M., Karita, S., Nishikawa, S. & Kashimura, N. (2005). Recognition of Receptor Lipopolysaccharides by Spike G Protein of Bacteriophage ϕX174. *Bioscience, Biotechnology, and Biochemistry*, *64*(9), 1993–1997.

Kawecki, T. J., Lenski, R. E., Ebert, D., Hollis, B., Olivieri, I., & Whitlock, M. C. (2012). Experimental evolution. *Trends in Ecology and Evolution*, *27*(10), 547-60.

Keightley, P. D. & Eyre-Walker, A. (2010). What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? *Philosophical Transactions of the Royal Society B: Biological Sciences, 365*(1544), 1187-1193.

Kimchi-Sarfaty, C., Sauna, Z. E., Kim, I. W., Ambudkar, S. V., Oh, J. M., Calcagno, A. M., & Gottesman, M. M. (2006). A “Silent” Polymorphism in the MDR1 Gene Changes Substrate Specificity. *Science, 315*, 525–528.

Kneitel, J. (2009). Gause's Competitive Exclusion Principle. *In Encyclopedia of Ecology*, *3*, 1731–1734.

Knowles, B., Silveira, C. B., Bailey, B. A., Barott, K., Cantu, V. A., Cobian-Guëmes, A. G., … Rohwer, F. (2016). Lytic to temperate switching of viral communities. *Nature*, *531*(7595), 466–470.

Koskella, B., & Brockhurst, M. A. (2014). Bacteria-phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiology Reviews*, *38*(5), 916–931.

Krokan, H. E. & Bjørås, M. (2013). Base excision repair. *Cold Spring Harbor Perspectives in Biology*, *5*(4), 1–22.

Kudela, P., Paukner, S., Mayr, U. B., Cholujova, D., Schwarczova, Z., Sedlak, J., Bizik, J., & Lubitz, W. (2005). Bacterial ghosts as novel efficient targeting vehicles for DNA delivery to the human monocyte-derived dendritic cells. *Journal of Immunotherapy*, *28*(2), 136–143.

Labrie, S. J., Dupuis, M. È., Tremblay, D. M., Plante, P. L., Corbeil, J., & Moineaua, S. (2014). A new Microviridae phage isolated from a failed biotechnological process driven by Escherichia coli. *Applied and Environmental Microbiology*, *80*(22), 6992–7000.

Lama, J., Mangasarian, A., & Trono, D. (1999). Cell-surface expression of CD4 reduces HIV-1 infectivity by blocking Env incorporation in a Nef- and Vpu- inhibitable manner. *Current Biology*, *9*, 622–631.

Lang, G. I., & Desai, M. M. (2014). The spectrum of adaptive mutations in experimental evolution. *Genomics*, *104*(6), 412–416.
Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*, *3*, 1–8.

Lau, S. K., Woo, P. C., Li, K. S., Huang, Y., Tsoi, H. W., Wong, B. H., Wong, S.S., Leung, S. Y., Chan, K. H & Yuen, K. Y. (2005). Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(39), 14040–14045.

Leclair, J. S. & Wahl, L. M. (2018). The Impact of Population Bottlenecks on Microbial Adaptation. *Journal of Statistical Physics*, *172*(1), 114-125

Lenski, R. E. (1988). experimental studies of pleiotropy and epistasis in escherichia coli. Ii. Compensation for maladaptive effects associated with resistance to virus t4
Richard. *Evolution*, *42*(3), 433-440.

Lenski, R.E. (2017). Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. *International Society for Microbial Ecology Journal 11*, 2181–2194.

Lenski, R. (2019) Lenski Lab Website. Available at: http://myxo.css.msu.edu [accessed 10 January 2019].

Lenski, R. E., & Levin, B. R. (1985). Constraints on the coevolution of bacteria and virulent phage: a model, some experiments, and predictions for natural communities. *American Naturalist*, *125*(4), 585–602.

Leonard, S. A., Weissman, D. B., Koelle, K., Greenbaum, B., & Ghedin, E. (2017). Transmission Bottleneck Size Estimation from Pathogen Deep-Sequencing Data, with an Application to Human Influenza A Virus. *Journal of Virology*, *91*(14), e00171-17.

Leroy, E. M., Epelboin, A., Mondonge, V., Pourrut, X., Gonzalez, J. P., Muyembe-Tamfum, J. J. & Formenty, P. (2009). Human Ebola outbreak resulting from direct exposure to fruit bats in Luebo, Democratic Republic of the Congo, 2007. *Vector Borne Zoonotic Dis 9*, 723–728.

Letarov, A. V., & Kulikov, E. E. (2018). Adsorption of bacteriophages on bacterial cells. *Biochemistry (Moscow)*, *82*(13), 1632–1658.

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

Li, K. S., Guan, Y., Wang, J., Smith, G. J. D., Xu, K. M., Duan, L., Rahardjo, A. P., Puthavathana, P., Buranathai, C., Nguyen, T. D., Estoepangestie, A. T. S., Chaisingh, A., Auewarakul, P., Long, H. T., Hanh, N. T. H., Webby, R. J., Poon, L. L. M., Chen, H., Shortridge, K. F., Yuen, K. Y., Webster, R. G.Peiris, J. S. M. (2004). Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. *Nature*, *430*(6996), 209–213.

Linster, M., van Riel, D., van Boheemen, S., de Graaf, M.,Osterhaus, A.D.M.E., Rimmelzwaan, G.F., Schrauwen, E.J.A., Fouchier, R.A.M., Lexmond, P., Bestebroer, T.M., Matrosovich, M., Mänz, B., Baumann, J., & Herfst, S. (2014). Identification, Characterization, and Natural Selection of Mutations Driving Airborne Transmission of A/H5N1 Virus. *Cell*, *157*, 329–339.

Liu, M., Deora, R., Doulatov, S. R., Gingery, M., Eiserling, F. A., Preston, A., Duncan, J., Simons, R. W., Cotter, P. A., Parkhill, J. & Miller J. F. (2002). Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage. *Science, 295*, 2091-2094.

Liu, M., Gingery, M., Doulatov, S. R., Liu, Y., Hodes, A., Baker, S., Davis, P., Simmonds, M., Churcher, C., Mungall, K., Quail, M. A., Preston, A., Harvill, E. T., Maskell, D. J., Eiserling, F. A., Parkhill, J. & Miller, J. F. (2004). Genomic and genetic analysis of Bordetella bacteriophages encoding reverse transcriptase-mediated tropism-switching cassettes. *J Bacteriol*. *186*(5), 1503-17.

Liu, X., Zhang, Q., Murata, K., Baker, M. L., Sullivan, M. B., Fu, C., Dougherty, M. T., Schmid, M. F., Osburne, M. S., Chisholm, S. W., & Chiu, W. (2010). Structural changes in a marine podovirus associated with release of its genome into Prochlorococcus. *Nature Structural & Molecular Biology*, *17*(7), 830-836.

Li, W., Shi, Z., Yu, M., Ren, W., Smith, C., Epstein, J. H., Wang, H., Crameri, G., Hu, Z., Zhang, H., Zhang, J., McEachern, J., Field, H., Daszak, P., Eaton, B. T., Zhang, S., Wang, L. F. (2005). Bats are natural reservoirs of SARS-like coronaviruses. *Science*, *310*(5748), 676–679.

Longdon, B., Brockhurst, M. A., Russell, C. A., Welch, J. J., & Jiggins, F. M. (2014). The Evolution and Genetics of Virus Host Shifts. *PLoS Pathogens*, *10*(11), e1004395.

Longdon, B., Day, J. P., Alves, J. M., Smith, S. C. L., Houslay, T. M., McGonigle, J. E., Tagliaferri, L., & Jiggins, F. M. (2018). Host shifts result in parallel genetic changes when viruses evolve in closely related species. *PLoS Pathogens*, *14*(4), e1006951.

Lv, M., Zheng, F., Wang, Q., Wang, T., & Liang, Y. (2015). Surface structural changes, surface energy and antiwear properties of polytetrafluoroethylene induced by proton irradiation. *Materials and Design*, *85*, 162–168.

Lynch, V. J., & G. P. Wagner. (2010). Did egg-laying boas break Dollo's Law? Phylogenetic evidence for reversal to oviparity in sand boas (Eryx: Boidae). *Evolution, 64*, 207–216.

McArdle, B.H. & Anderson, M. J. (2001). Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology*, *82*, 290 - 297.

Madinda, N. F., Ehlers, B., Wertheim, J. O., Akoua-Koffi, C., Bergl, R. A., Boesch, C., Akonkwa, D. B. M., Eckardt, W., Fruth, B., Gillespie, T. R., Gray, M., Hohmann, G., Karhemere, S., Kujirakwinja, D., Langergraber, K., Muyembe, J., Nishuli, R., Pauly, M., Petrzelkova, K. J., Robbins, M. M., Todd, A., Schubert, G., Stoinski, T. S., Wittig, R. M., Zuberbühler, K., Peeters, M., Leendertz, F. H., & Calvignac-Spencer, S. (2016). Assessing Host-Virus Codivergence for Close Relatives of Merkel Cell Polyomavirus Infecting African Great Apes. *Journal of virology*, *90* (19), 8531-8541.

Manrubia, S. C. (2012). Modelling viral evolution and adaptation: challenges and rewards. *Current Opinion Virology, 2*(5):531–537.

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C.J., Lu, S., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D. I., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., Zheng, C., Geer, L. Y.,& Bryant, S. H. (2017). CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research, 45*, D200–D203.

Martín-Acebes, M.A., and Saiz, J.C. (2011). A West Nile virus mutant with increased resistance to acid-induced inactivation. *Journal of General Virology, 92*, 831–840.

Martin, M. (2011). Cutadapt removes adapter sequences from highthroughput sequencing reads. *European Molecular Biology network J*, *17*, 10–12.

Matteau, D., Baby, V., Pelletier, S., & Rodrigue, S. (2015). A small-volume, low-cost, and versatile continuous culture device. *PLoS ONE*, *10*(7).

Mavrich, T. N., & Hatfull, G. F. (2017). Bacteriophage evolution differs by host, lifestyle and genome. *Nature Microbiology*, *2*.

McArdle, B.H. & Anderson, M.J. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* 82, 290–297.

McCandlish, D. M. (2011). Visualizing fitness landscapes. *Evolution*, *65*(6), 1544-1558.

McKenna, R., Ilag, L. L., & Rossmann, M. G. (1994). Analysis of the Single-stranded DNA Bacteriophage ϕX174, Refined at a Resolution of 3·0 Å. *Journal of Molecular Biology*, *237*(5), 517–543.

Meier-Kolthoff, J., Hahnke, R., Petersen, J., Scheuner, C., Michael, V., Fiebig, A., Rohde, C., Rohde, M., Fartmann, B., Goodwin, L., Chertkov, O., Reddy, T., Pati, A., Ivanova, N., Markowitz, V., Kyrpides, N., Woyke, T., Göker, M. & Klenk, H. (2014). Complete genome sequence of DSM 30083T, the type strain (U5/41T) of Escherichia coli, and a proposal for delineating subspecies in microbial taxonomy. *Standards in Genomic Sciences*, 9(1).

Metzker, M. L. (2005). Emerging technologies in DNA sequencing. *Genome Research, 15*(12), 1767-76.

Meyer, J. R., Dobias, D. T., Weitz, J. S., Barrick, J. E., Quick, R. T., & Lenski, R. E. (2012). Repeatability and contingency in the evolution of a key innovation in phage lambda. *Science*, *335*(6067), 428–432.

Michel, A., Clermont, O., Denamur, E., & Tenaillon, O. (2010). Bacteriophage PhiX174's ecological niche and the flexibility of its escherichia coli lipopolysaccharide receptor. *Applied and Environmental Microbiology*, *76*(21), 7310–7313.

Middelboe, M., Holmfeldt, K., Riemann, L., Nybroe, O., & Haaber, J. (2009). Bacteriophages drive strain diversification in a marine Flavobacterium: Implications for phage resistance and physiological properties. *Environmental Microbiology*, *11*(8), 1971–1982.

Mikheenko, A., Prjibelski, A., Antipov, D., Saveliev, V., &. Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*, *34*(13), i142–i150.

Miller, A. W., Befort, C., Kerr, E. O., & Dunham, M. J. (2013). Design and use of multiplexed chemostat arrays. *Journal of Visualized Experiments*, (72).

Minot, S., Bryson, A., Chehoud, C., Wu, G. D., Lewis, J. D., & Bushman, F. D. (2013). Rapid evolution of the human gut virome. *Proceedings of the National Academy of Sciences*, *110*(30), 12450–12455.

Miranda, M. E., Ksiazek, T. G., Retuya, T. J., Khan, A. S., Sanchez, A., Fulhorst, C. F., Rollin, P. E., Calaor, A. B., Manalo, D. L., Roces, M. C., Dayrit, M. M., & Peters, C. J. (1999). Epidemiology of Ebola (Subtype Reston) Virus in the Philippines, 1996. *The Journal of Infectious Diseases*, *179*(1), 115-119.

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

Monod, J. (1950). Technique, Theory and Applications of Continuous Culture. *Ann. Inst. Pasteur*, *79*(4), 390–410.

Morella, N. M., Yang, S. C., Hernandez, C. A., & Koskella, B. (2018). Rapid quantification of bacteriophages and their bacterial hosts in vitro and in vivo using droplet digital PCR. *Journal of Virological Methods*, *259*, 18–24.

Morey, M., Fraga, J.M., Fernández-Marmiesse, A., Cocho, J.A., Castiñeiras, D., and Couce, M.L. (2013). A glimpse into past, present, and future DNA sequencing. *Molecular Genetics and Metabolism, 110*, 3–24.

Nanchen, A., Schicker, A., & Sauer, U. (2006). Nonlinear dependency of intracellular fluxes on growth rate in miniaturized continuous cultures of

Escherichia coli. *Applied and Environmental Microbiology*, *72*(2), 1164–1172.

Nichol, S. T., Arikawa, J., & Kawaoka, Y. (2000). Emerging viral diseases. *PNAS*, *97*(23), 12411-12412.

Norder, H., Ebert, J. W., Fields, H. A., Mushahwar, I. K., & Magnius, Lars O. (1996). Complete Sequencing of a Gibbon Hepatitis B Virus Genome Reveals a Unique Genotype Distantly Related to the Chimpanzee Hepatitis B Virus. *Virology*, *218*(1), 214-223.

Novella, I. S., Clarke, D. K., Quer, J., Duarte, E. A., Lee, C. H., Weaver, S. C., Elena, S. F., Moya, A., Domingo, E., & Holland, J. J. (1995). Extreme fitness differences in mammalian and insect hosts after continuous replication of vesicular stomatitis virus in sandfly cells. *Journal of virology*, *69*(11), 6805-6809.

Novella, I. S., Elena, S. F., Moya, A., Domingo, E., & Holland, John J. (1995). Size of Genetic Bottlenecks Leading to Virus Fitness Loss Is Determined by Mean Initial Population Fitness. *Journal of Virology*, *69*(5) 2869-2872.

Novick, A., & Szilard, L. (1950). Description of the chemostat. *Science*, *112*(2920), 715–716.

Ohkawa, T. (1979). Ter mutation and susceptibility to ϕx174 phage in E. coli K12. *Biochemical and Biophysical Research Communications*, *91*(3), 1051–1056.

Olia, A. S., Prevelige, P. E., Johnson, J. E., & Cingolani, G. (2011). Three-dimensional structure of a viral genome-delivery portal vertex. *Nature Structural & Molecular Biology*, *18*(5), 597-603.

Orr, H. A. (2005a). The genetic theory of adaptation: A brief history. *Nature Reviews Genetics 6*(2), 119-27.

Orr, H. A. (2005b). Theories of adaptation : what they do and don ' t say Edited by Foxit Reader. *Genetica*, *123*, 3–13.

Orr, H. A. (2009). Fitness and its role in evolutionary genetics. *Nature Reviews Genetics*, *10*, 531–539.

Osterholm, M. T., Moore, K. A., Kelley, N. S., Brosseau, L. M., Wong, G., Murphy, F. A., Peters, C. J., LeDuc, J. W., Russell, P. K., Van Herp, M., Kapetshi, J., Muyembe, J. J., Ilunga, B. K., Strong, J. E., Grolla, A., Wolz, A., Kargbo, B., Kargbo, D. K., Sanders, D. A. & Kobinger, G. P. (2015). Transmission of Ebola viruses: what we know and what we do not know. *mBio*, *6*(2), e00137.

Parrish CR, Holmes EC, Morens DM, Park EC, Burke DS, Calisher CH, Laughlin,Ca. A., Saif, L. J., & Daszak, P. (2008). Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiology and Molecular Biology Reviews*, 72(3), 457–70.

Paterson, S., Vogwill, T., Buckling, A., Benmayor, R., Spiers, A. J., Thomson, N. R., … Brockhurst, M. A. (2010). Antagonistic coevolution accelerates molecular evolution. *Nature*, *464*(7286), 275–278.

Pepin, K. M., Samuel, M. A., & Wichman, H. A. (2006). Variable pleiotropic effects from mutations at the same locus hamper prediction of fitness from a fitness component. *Genetics*, *172*(4), 2047–2056.

Pepin, K. M., & Wichman, H. A. (2007). Variable epistatic effects between mutations at host recognition sites in ϕX174 bacteriophage. *Evolution*, *61*(7), 1710–1724.

Pepin, K. M., & Wichman, H. A. (2008). Experimental evolution and genome sequencing reveal variation in levels of clonal interference in large populations of bacteriophage ϕX174. *BMC Evolutionary Biology*, *8*(85).
Peralta, B., Gil-Carton, D., Castaño-Díez, D., Bertin, A., Boulogne, C., Oksanen, H. M., Bamford, D. H., & Abrescia, N. G. A. (2013). Mechanism of Membranous Tunnelling Nanotube Formation in Viral Genome Delivery. *PLoS Biology*, *11*(9), e1001667.

Phage receptor database (2019). Phage receptor database Website. Available at: https://phred.herokuapp.com [accessed 27 October, 2018].

Philpott, S. M. 2003. HIV-1 coreceptor usage, transmission, and disease progression. *Current HIV Research*, *1*, 217–227.

Pienaar, E., Theron, M., Nelson, M., & Viljoen, H. J. (2006). A quantitative model of error accumulation during PCR amplification. *Computational biology and chemistry*, *30*(2), 102–111.

Plank, L. D., & Harvey, J. D. (2009). Generation Time Statistics of Escherichia coli B Measured by Synchronous Culture Techniques. *Journal of General Microbiology*, *115*(1), 69–77.

Poelwijk, F., Kiviet, D.J., Weinreich, D.M. & Tans, S.J. (2007). Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, *445*, 383–386.

Poon, A. F. Y., & Chao, L. (2006). Functional origins of fitness effect-sizes of compensatory mutations in the DNA bacteriophage ϕX174. *Evolution*, *60*(10), 2032-43.

Prabhakaran, R., Chithambaram, S., & Xia, X. (2014). Aeromonas phages encode tRNAs for their overused codons. In *International Journal of Computational Biology and Drug Design*, 7, 168–182.

Prabhakaran, R., Chithambaram, S., & Xia, X. (2015). Escherichia coli and Staphylococcus phages: Effect of translation initiation efficiency on differential codon adaptation mediated by virulent and temperate lifestyles. *Journal of General Virology*, *96*(5), 1169–1179.

Quantamagazine (2019), Quantamagazine Website. Available at: https://www.quantamagazine.org/viruses-would-rather-jump-to-new-hosts-than-evolve-with-them-20170913/ [accessed 13 January 2019].

Rakhuba, D.V., Kolomiets, E.I., Dey, E. S., &2 and Novik, G.I. (2010). Bacteriophage Receptors, Mechanisms of Phage Adsorption and Penetration into Host Cell. *Polish Journal of Microbiology*, *59*(3), 145-155.

Rautio, J. J., Smit, B. A., Wiebe, M., Pentillä, M., & Saloheimo, M. (2006). Transcriptional monitoring of steady state and effects of anaerobic phases in chemostat cultures of the filamentous fungus Trichoderma reesei. *BMC Genomics*, *7*.

Razin, A., Hirose, T., & Riggs, A. D. (1978). Efficient correction of a mutation by use of chemically synthesized DNA. *Proceedings of the National Academy of Sciences*, *75*(9), 4268–4270.

Remold, S. (2012). Understanding specialism when the jack of all trades can be the master of all. *Proceedings of the Royal Society B: Biological Sciences*. Royal Society.

Redondo, R. A. F., de Vladar, H. P., Włodarski, T., Bollback, J. P (2017). Evolutionary interplay between structure, energy and epistasis in the coat protein of the ϕX174 phage family. *Journal of the Royal Society, Interface*, *14*(126), 20160139.

Rodriguez-Verdugo, A., Gonzalez-Gonzalez, A., Carrillo-Cisneros, D., Bennett, A. F., & Gaut, B. S. (2014). Different tradeoffs result from alternate genetic adaptations to a common environment. *Proceedings of the National Academy of Sciences*, *111*(33), 12121–12126.

Rodríguez-Rubio, L., Gutiérrez, D., Donovan, D. M., Martínez, B., Rodríguez, A., & García, P. (2016). Phage lytic proteins: biotechnological applications beyond clinical antimicrobials. *Critical Reviews in Biotechnology*, *36*(3), 542-52.

Roychoudhury, P., Shrestha, N., Wiss, V. R., & Krone, S. M. (2013). Fitness benefits of low infectivity in a spatially structured population of

bacteriophages. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1774), e20132563.

Rokyta, D. R., Joyce, P., Caudle, S. B., Miller, C., Beisel, C. J., & Wichman, H. A. (2011). Epistasis between beneficial mutations and the phenotype-to-fitness map for a ssDNA virus. *PLoS Genetics*, *7*(6), e1002075.

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biology*, *14,* 51-71.

Ruboyianes, M. V., Chen, M., Dubrava, M. S., Cherwa, J. E., & Fane, B. A. (2009). The Expression of N-Terminal Deletion DNA Pilot Proteins Inhibits the Early Stages of X174 Replication. *Journal of Virology*, *83*(19), 9952–9956.

Rudicell, R. S., Holland, J. J., Wroblewski, E. E., Learn, G. H., Li, Y., Robertson, J. D., Greengrass, E., Grossmann, F., Kamenya, S., Pintea, L.… (2010). Impact of simian immunodeficiency virus infection on chimpanzee population dynamics. *PLoS Pathog*, *6*, e1001116.

Rusk, N. (2014). Genomics: Nanopores read long genomic DNA. *Nature Methods*, *11*(9), 887.

Sambrook, J., E. F. Fritsch and T. Maniatis, chapter 3 (2001). Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sandmeier, H., Iida, S., & Arber, W. (1992). DNA inversion regions Min of plasmid p15B and Cin of bacteriophage P1: Evolution of bacteriophage tail fiber genes. *Journal of Bacteriology*, *174*(12), 3936–3944.

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, (2019), Sanjuán Lab Website. Available at: https://www.uv.es/rsanjuan/Viral_mutation_rates_snr.htm [accessed 15 March, 2019].

Sanjuán, R. (2010). Mutational fitness effects in RNA and single-stranded DNA viruses: common patterns revealed by site-directed mutagenesis studies. Philosophical transactions of the Royal Society of London. *Series B, Biological sciences*, *365*(1548), 1975-82.

Santos, S. B., Carvalho, C. M., Sillankorva, S., Nicolau, A., Ferreira, E.C., & Azeredo, J. (2009). The use of antibiotics to improve phage detection and enumeration by the double-layer agar technique. *BMC Microbiology*, *9*, 1-10.

Sattar, M. M., Patel, M., & Alani, A. (2017). Clinical applications of polytetrafluoroethylene (PTFE) tape in restorative dentistry. *British Dental Journal*. Nature Publishing Group.

Sauter, D., Schindler, M., Specht, A., Landford, W. N., Münch, J., Kim, K. A., … Kirchhoff, F. (2009). Tetherin-Driven Adaptation of Vpu and Nef Function and the Evolution of Pandemic and Nonpandemic HIV-1 Strains. *Cell Host and Microbe*, *6*(5), 409–421.

Scanlan, P. D., Hall, A. R., Lopez-Pascua, L. D. C., & Buckling, A. (2011). Genetic basis of infectivity evolution in a bacteriophage. *Molecular Ecology*, *20*(5), 981–989.

Schluter, D., Clifford, E. A., Nemethy, M., & McKinnon, J. S. (2004). Parallel Evolution and Inheritance of Quantitative Traits. *The American Naturalist*, *163*(6), 809–822.

Schmideder, A., Severin, T. S., Cremer, J. H., & Weuster-Botz, D. (2015). A novel milliliter-scale chemostat system for parallel cultivation of microorganisms in stirred-tank bioreactors. *Journal of Biotechnology*, *210*, 19–24.

Schmitt, M. W., Fox, E. J., Prindle, M. J., Reid-Bayliss, K. S., True, L. D., Radich, J. P., & Loeb, L. A. (2015). Sequencing small genomic targets with high efficiency and extreme accuracy. *Nature Methods*, *12*(5), 423–425.

Schwartz, D. A., & Lindell, D. (2017). Genetic hurdles limit the arms race between Prochlorococcus and the T7-like podoviruses infecting them. *ISME Journal*, *11*(8).

Seecharran (2018). Elucidating the unknown ecology of bacterial pathogens from genomic data (Doctoral dissertation). Nottingham Trent University, Nottingham, United Kingdom.

Sharp, P.M., & Hahn, B.H. (2010). The evolution of HIV-1 and the origin of AIDS. *Philosophical Transactions of the Royal Society B: Biological Sciences, 365*, 2487–2494.

Shinya, K., Ebina, M., Yamada, S., Ono, M., Kasai, N., & Kawaoka, Y. (2006). Avian flu: influenza virus receptors in the human airway. *Nature*, *440*, 435–436.

Shlomai, J., Polder, L., Arai, K., & Komberg, A. (1981). Replication of phi X174 DNA with purified enzymes. I. Conversion of viral DNA to a

supercoiled, biologically active duplex. *Journal of Biological Chemistry, 256*(10), 5233-8.

Sime-Ngando, T. (2014). Environmental bacteriophages: viruses of microbes in aquatic ecosystems. *Frontiers in microbiology*, *5*, 355.

Sinsheimer, R.L. (1959) A single-stranded deoxyribonucleic acid from bacteriophage ϕX174. *Journal of Molecular Biology, 1*(1), 43-53.

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.

Soto, P. C., Stein, L. L., Hurtado-Ziola, N., Hedrick, S. M. & Varki, A. (2010). Relative over-reactivity of human versus chimpanzee lymphocytes: Implications for the human diseases associated with immune activation. *J Immunol. 184*, 4185–4195.

Steel J., Lowen A.C. (2014). Influenza A Virus Reassortment. *Influenza Pathogenesis and Control, 385*(1), 377-401.

Sun, L., Molineux, I. J., Fane, B. A., Rossmann, M. G., Zhang, X., Young, L. N., Zbornik, E. (2014). Icosahedral bacteriophage ΦX174 forms a tail for DNA transport during infection. *Nature*, *505*(7483), 432–435.

Sun, Y., Roznowski, A. P., Pollack, L., Fane, B. A., Klose, T., Mauney, A., Pollack, L., Fane, B. A., Rossmann, M. G. (2017). Structural changes of tailless bacteriophage ΦX174 during penetration of bacterial cell walls. *Proceedings of the National Academy of Sciences*, *114*(52), 13708–13713.

Suzuki, R., Inagaki, M., Karita, S., Kawaura, T., Kato, M., Nishikawa, S., Nishikawa S., & Morita, J. (1999). Specific interaction of fused H protein of bacteriophage ϕX174 with receptor lipopolysaccharides. *Virus Research*, *60*(1), 95–99.

Suzuki, M., Kaneko-Tanaka, Y., & Azegami, M. (1974). Transfection of non-host bacterial spheroplasts with bacteriophage ΦX174 DNA. *Nature*, *252*(5481), 319-321.

Szczepankowska, A. (2012). Role of CRISPR/cas System in the Development of Bacteriophage Resistance. *Advances in Virus Research*, *82*, 289-338.

Takahashi, C. N., Miller, A. W., Ekness, F., Dunham, M. J., & Klavins, E. (2015). A low cost, customizable turbidostat for use in synthetic circuit characterization. *ACS Synthetic Biology*, *4*(1), 32–38.

Takehisa, J., Kraus, M. H., Ayouba, A., Bailes, E., Van Heuverswyn, F., Decker, J. M., Li, Y., Rudicell, R. S., Learn, G. H., Neel, C., Ngole, E. M., Shaw, G. M., Peeters, M., Sharp, P. M. & Hahn B. H. (2009). Origin and biology of simian immunodeficiency virus in wild-living western gorillas. *J Virol., 83*, 1635–1648.

Tan, L., & Gore, J. (2012). Slowly Switching Between Environments Facilitates Reverse Evolution in Small Populations. *Evolution*, *66*(10), 3144-3154.

Tenaillon, O., Barrick, J. E., Ribeck, N., Deatherage, D. E., Blanchard, J. L., Dasgupta, A., Wu, G. C., Wielgoss, S., Cruveiler, S., Medigue, C., Schneider., & Lenski, R. E. (2016). Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature, 536*, 165–170.

Tenaillon, O., Skurnik, D., Picard, B., & Denamur, E. (2010, March). The population genetics of commensal Escherichia coli. *Nature Reviews Microbiology*.

Teotonio, H., & Rose, M. R. (2000). Variation in the reversibility of evolution. *Nature, 408*, 463–466.

Turner, P. E., & L. Chao. (1998). Sex and the evolution ofintrahost competition in RNA virus F6. *Genetics, 150*, 523–532.

Treangen, T.J., Ondov, B.D., Koren, S., and Phillippy, A.M. (2014). The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology 15*(11), 524-539.

Truyen, U., Evermann, J. F., Vieler, E., & Parrish, C. R. (1996). Evolution of canine parvovirus involved loss and gain of feline host range. *Virology*, *215*(2), 186–189.

Truyen, U., & Parrish, C. R. (1995). The evolution and control of parvovirus host ranges. *Seminars in Virology*, *6*(5), 311-317.

Urbanowicz, R. A., McClure, C. P., Sakuntabhai, A., Sall, A.A., Kobinger, G., Müller, M. A., Sall, A. A., McClure, C. P., Holmes, E. C., Rey, F. A., Simon-Loriere, E., Ball, J. K. (2016). Human Adaptation of Ebola Virus during the West African Outbreak. *Cell*, *167*(4), 1079-1087.

Vale, P. F., Choisy, M., Froissart, R., Sanjuán, R., & Gandon, S. (2012). The Distribution Of Mutational Fitness Effects Of Phage ΦX174 On Different Hosts. *Evolution*, *66*(11), 3495–3507.

Valgepea, K., Adamberg, K., Seiman, A., & Vilu, R. (2013). Escherichia coli achieves faster growth by increasing catalytic and translation rates of proteins. *Molecular BioSystems*, *9*(9), 2344–2358.

Van den Bergh, B., Swings, T., Fauvart, M., & Michiels, J. (2018). Experimental Design, Population Dynamics, and Diversity in Microbial Experimental Evolution. *Microbiology and Molecular Biology Reviews*, *82*(3).

van der Avoort HG, van der Ende A, van Arkel GA, Weisbeek PJ (1984) Regions of incompatibility in single-stranded DNA bacteriophages wX174 and G4. *J Virol*, *50*, 533–540.

Van Der Ende, A., Langeveld, S. A., Van Arkel, G. A., & Weisbeek, P. J. (1982). The Interaction of the A and A* Proteins of Bacteriophage ϕX174 with Single-Stranded and Double-Stranded ϕX DNA in vitro. *European Journal of Biochemistry*, *124*(2), 245–252.

Vazquez-Calvo, A., Caridi, F., Sobrino, F., Sandri-Goldin, R.M., & Martin-Acebes, M.A. (2013). An Increase in Acid Resistance of Foot-and-Mouth Disease Virus Capsid Is Mediated by a Tyrosine Replacement of the VP2 Histidine Previously Associated with VP0 Cleavage. *Journal of Virology, 88*, 3039–3042.

Wahl L.M. & Krakauer, D. V. (2000). Models of Experimental Evolution: The Role of Genetic Chance and Selective Necessity. *Genetics*, *156 (3), 1437-1448.*

Wang, E., Ni, H., Xu, R., Barrett, A.D., Watowich, S.J., Gubler, D.J., Weaver, S.C., (2000). Evolutionary relationships of endemic/epidemic and sylvatic dengue viruses. *J. Virol. 74*, 3227-3234.

Weaver, S.C., Costa, F., Garcia-Blanco, M.A., Ko, A.I., Ribeiro, G.S., Saade, G., Shi, P.Y., & Vasilakis, N. (2016). Zika virus: History, emergence, biology, and prospects for control. *Antiviral Research 130*, 69–80.

Weaver, S. C., & Barrett, A. D. T. (2004). Transmission cycles, host range, evolution and emergence of arboviral disease. *Nature Reviews Microbiology*, 2, 789–801.

Webby, R. J., & Webster, R. G. (2001). Emergence of influenza A viruses. In *Philosophical Transactions of the Royal Society B: Biological Sciences, 356*, 1817–1828.

Weinreich, D. M., Lan, Y., Jaffe, J. & Heckendorn, R. B. (2018). The Influence of Higher-Order Epistasis on Biological Fitness Landscape Topography. *Journal of Statistical Physics, 172*(1), 208-225.

Weinreich, D. M., Watson, R. A. & Chao, L. (2005). Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution, 59*(6), 1165-1174.

Weitz, J. S., & Dushoff, J. (2008). Alternative stable states in host - Phage dynamics. *Theoretical Ecology*, *1*(1), 13–19.

World Health Organisation **(**2004**).** Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003. http://www.who.int/csr/sars/country/table2004_04_21/en/index.html. [accessed 21 December, 2019].

World Health Organisation (2019). World Health Organisation website. Availabe at https://www.who.int/ebola/en/ [accessed 13 Febuary, 2019].

World Health Organisation (2019). World Health Organisation website. Availabe at https://www.who.int/emergencies/diseases/zika/zika-epidemiology-update-july-2019.pdf?ua=1 [accessed 13 December, 2019].

Wiens, J. J. (2011). Re-evolution of lost mandibular teeth in frogs after more than 200 million years, and re-evaluating Dollo's Law. *Evolution 65,*1283–1296.

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., & Brown, C.J. (2010). Experimental evolution of viruses: Microviridae as a model system. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *365*(1552), 2495-2501.

Wichman, H. A., Millstein, J., & 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.

Wick, R. R., Judd, L. M., Gorrie, C. L., & Holt, K. E. (2017). Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computational Biology*, *13*(6).

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: Interactive visualization of de novo genome assemblies. *Bioinformatics*, *31*(20), 3350–3352.

Wilcox, A. (2017). Evolution at high imposed mutation rate. (Doctoral dissertation). Nottingham Trent University, Nottingham, United Kingdom.

Witte, A., Isi, B. /, Halfmann, G., Szostak, M., Wanner, G., & Lubitz, W. (1990). *PhiX174 protein E-mediated iysis of Escheriehia coil*. *Biochimie*, *72*, 191–200.

Whitman, W. B., Coleman, D. C., & Wiebe, W. J. (1998) Prokaryotes: The unseen majority. *Proceedings of the Natural Academy Science of the United States of America*, *95*, 6578–6581.

Wolfe, N.D., Dunavan, C.P., & Diamond, J. (2007). Origins of major human infectious diseases. *Nature*, *447*, 279–283.

Wolfson, R., & Eisenberg, S. (2006). Escherichia coli host factor required specifically for the phi X174 stage III reaction: in vitro identification and partial purification. *Proceedings of the National Academy of Sciences of the United States of America*, *79*, 5768–5772.

Wylie, C. S., & Shakhnovich, E. I. (2012). Mutation Induced Extinction in Finite Populations: Lethal Mutagenesis and Lethal Isolation. *PLoS Computational Biology*, *8*(8), e1002609.

Yethon, J. A., Vinogradov, E., Perry, M. B., & Whitfield, C. (2000). Mutation of the lipopolysaccharide core glycosyltransferase encoded by waaG destabilizes the outer membrane of Escherichia coli by interfering with core phosphorylation. *Journal of Bacteriology*, *182*(19), 5620–5623.

Young, L. N., Hockenberry, A. M., & Fane, B. A. (2013). Mutations in the N Terminus of the oX174 DNA Pilot Protein H Confer Defects in both Assembly and Host Cell Attachment. *Journal of Virology*, *88*(3), 1787–1794.

Yu, S.Y., Peng, W., Si, W., Yin, L., Liu, S.G., Liu, H.F., Zhao, H.L., Wang, C.L., Chang, Y.H., & Lin, Y.Z. (2011). Enhancement of bacteriolysis of Shuffled phage PhiX174 gene e. *Virology Journal*, *8*(206).

Zhang, J. (2000). Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *Journal of Molecular Evolution*, *50*(1), 56–68.

Zhang, Z., Boccazzi, P., Choi, H. G., Perozziello, G., Sinskey, A. J., & Jensen, K. F. (2006). Microchemostat - Microbial continuous culture in a polymer-based, instrumented microbioreactor. *Lab on a Chip*, *6*(7), 906–913.
Ziv, N., Brandt, N. J., & Gresham, D. (2013). The use of chemostats in microbial systems biology. *Journal of Visualized Experiments*, (80).

Appendix

Appendix A : Python scripts

A.1 Quality filter reads

```
#!/usr/bin/env bash

usage() {
  NAME=$(basename $0)
  cat <<EOF
Usage:
  ${NAME}
You must define global variables first (via
"0_config_and_run.sh")

EOF
}

# location for log file
LOGFILE=./1_filter.log

# reverse complemented adapter sequence for Nextera (XT)
rv_adapter="CTGTCTCTTATA"

# variables to be used in main loop
reads1=(${FASTQLOC}/*R1*.fastq.gz) # collect each forward read in
array, e.g. "~/FASTQ/A_S1_L001_R1_001.fastq.gz"
reads1=("${reads1[@]##*/}") # [@] refers to array, greedy remove
*/ from left, e.g. "A_S1_L001_R1_001.fastq.gz"
reads2=("${reads1[@]/_R1/_R2}") # substitute R2 for R1, e.g.
"A_S1_L001_R2_001.fastq.gz"

# main loop
pipeline() {

echo [`date +"%Y-%m-%d %H:%M:%S"`] "#> START: " $0 $@

# fastqc analysis of raw reads
mkdir -p ${BASEDIR}/fastqc_before
fastqc -o ${BASEDIR}/fastqc_before ${FASTQLOC}/*.fastq*

# cutadapt loop
for ((i=0; i<=${#reads1[@]}-1; i++)); do # i from zero to one
minus length of array
  fwdrds="${reads1[$i]}" # e.g. "A_S1_L001_R1_001.fastq.gz"
  rvsrds="${reads2[$i]}" # e.g. "A_S1_L001_R2_001.fastq.gz"
  id="${fwdrds%%_*}" # greedy remove _ from right e.g. "A"

  cutadapt --quality-base=33 --quality-cutoff 30,30 \
  -a ${rv_adapter} -A ${rv_adapter} --error-rate=0.2 --overlap=3
\
  --trim-n --pair-filter=any --minimum-length=20 --cores=$NUMCPUS
\
  --output=${BASEDIR}/TRIM/${id}_trimmed_R1.fastq.gz \
  --paired-output=${BASEDIR}/TRIM/${id}_trimmed_R2.fastq.gz \
  ${FASTQLOC}/${fwdrds} ${FASTQLOC}/${rvsrds}
done
```

```
# fastqc analysis after trimming
mkdir -p ${BASEDIR}/fastqc_after
fastqc -o ${BASEDIR}/fastqc_after ${BASEDIR}/TRIM/*.fastq*

# multiqc run for working directory (will create HTML and folder)
multiqc ${BASEDIR}

echo [`date +"%Y-%m-%d %H:%M:%S"`] "#> DONE."
} #pipeline end

pipeline 2>&1 | tee $LOGFILE
```

A.2 Remove spike-in

```
#!/usr/bin/env bash

usage() {
  NAME=$(basename $0)
  cat <<EOF
Usage:
  ${NAME}
It is important that you run "1_filter_reads.sh" first!

EOF
}

# location for log file
LOGFILE=./2_subspike.log

# variables to be used in main loop
reads1=(${BASEDIR}/TRIM/*_trimmed_R1.fastq.gz) # collect each
forward read in array, e.g. "TRIM/A_trimmed_R1.fastq.gz"
reads1=("${reads1[@]##*/}") # [@] refers to array, greedy remove
*/ from left, e.g. "A_trimmed_R1.fastq.gz"
reads2=("${reads1[@]/_R1/_R2}") # substitute R2 for R1, e.g.
"A_trimmed_R2.fastq.gz"

# main loop
pipeline() {

echo [`date +"%Y-%m-%d %H:%M:%S"`] "#> START: " $0 $@

for ((i=0; i<=${#reads1[@]}-1; i++)); do
  fwdrds="${reads1[$i]}" # e.g. "A_trimmed_R1.fastq.gz"
  rvsrds="${reads2[$i]}" # e.g. "A_trimmed_R2.fastq.gz"
  id="${fwdrds%%_*}" # greedy remove _* from right e.g. "A"

  ## 1. MAP {TRIMMED READS} TO MAIN SPIKE-IN GENOME
  # map reads to the linear spike-in genome (edge effects
expected) and send to TMP/
  bwa mem -t ${NUMCPUS} ${FASTALOC}/pUC18_L09136.fasta
${BASEDIR}/TRIM/${fwdrds} ${BASEDIR}/TRIM/${rvsrds} \
  > ${BASEDIR}/TMP/${id}_tmpspike_1.sam

  # select MAPPED reads (F=flag absent, 4=unmapped), send to SPK/
for later
```

```
  samtools view -O BAM -F 4 -o
${BASEDIR}/SPK/${id}_spikemapped_1.bam
${BASEDIR}/TMP/${id}_tmpspike_1.sam

  # select UNMAPPED reads (f=flag present, 4=unmapped), sort by
read name and keep BAM in TMP/
  samtools view -O SAM -h -f 4
${BASEDIR}/TMP/${id}_tmpspike_1.sam \
  | samtools sort -O BAM -n -o
${BASEDIR}/TMP/${id}_unmapped_1.bam -

  # convert unmapped BAM files to FASTQ (for PE singletons are
discarded by bedtools, which is conservative)
  bedtools bamtofastq -i ${BASEDIR}/TMP/${id}_unmapped_1.bam -fq
${BASEDIR}/TMP/${id}_unmapped_R1.fastq \
  -fq2 ${BASEDIR}/TMP/${id}_unmapped_R2.fastq

  # delete unmapped BAM and temporary SAM (so we only have
unmapped FASTQ in TMP/)
  rm ${BASEDIR}/TMP/${id}_unmapped_1.bam
  rm ${BASEDIR}/TMP/${id}_tmpspike_*.sam

  ## 2. MAP {TRIMMED READS} TO RESECTED GENOME (this is for later
analysis of spike-in genome)
  # map trimmed reads to resected spike-in genome
  bwa mem -t ${NUMCPUS} ${FASTALOC}/pUC18_L09136_resected.fasta
${BASEDIR}/TRIM/${fwdrds} ${BASEDIR}/TRIM/${rvsrds} \
  > ${BASEDIR}/TMP/${id}_tmpspike_2.sam

  # select MAPPED reads (F=flag absent, 4=unmapped), send to SPK/
for later, and delete temporary SAM
  samtools view -O BAM -F 4 -o
${BASEDIR}/SPK/${id}_spikemapped_2.bam
${BASEDIR}/TMP/${id}_tmpspike_2.sam
  rm ${BASEDIR}/TMP/${id}_tmpspike_*.sam

  ## 3. MAP {UNMAPPED READS in TMP/} TO RESECTED GENOME (this is
for maximal accuracy in subtracting spike-in reads)
  # map unmapped reads to resected spike-in genome
  bwa mem -t ${NUMCPUS} ${FASTALOC}/pUC18_L09136_resected.fasta
${BASEDIR}/TMP/${id}_unmapped_R1.fastq \
  ${BASEDIR}/TMP/${id}_unmapped_R2.fastq >
${BASEDIR}/TMP/${id}_tmpspike_3.sam

  # remove input FASTQs
  rm ${BASEDIR}/TMP/${id}_unmapped_*.fastq

  # select UNMAPPED reads (f=flag present, 4=unmapped), use grep
-v to remove MAPPED reads from filter list, sort by read name and
cleanup
  samtools view -O SAM -h -f 4
${BASEDIR}/TMP/${id}_tmpspike_3.sam \
  | samtools sort -O BAM -n -o
${BASEDIR}/TMP/${id}_unmapped_2.bam -
  rm ${BASEDIR}/TMP/${id}_tmpspike_*.sam # delete temporary SAM

  # convert unmapped BAM files to FASTQ, send to UMP/ and cleanup
```

```
  bedtools bamtofastq -i ${BASEDIR}/TMP/${id}_unmapped_2.bam -fq
${BASEDIR}/UMP/${id}_unmapped_R1.fastq \
  -fq2 ${BASEDIR}/UMP/${id}_unmapped_R2.fastq
  rm ${BASEDIR}/TMP/${id}_unmapped_2.bam # remove BAM

  # compress FASTQ files in UMP/
  gzip -f ${BASEDIR}/UMP/${id}_unmapped_*.fastq # zip FASTQs for
space (-f forces deletion of original)
done

echo [`date +"%Y-%m-%d %H:%M:%S"`] "#> DONE."
} #pipeline end

pipeline 2>&1 | tee $LOGFILE
```

A.3 Map to references

```
#!/usr/bin/env bash

usage() {
  NAME=$(basename $0)
  cat <<EOF
Usage:
  ${NAME}
It is important that you run "2_subtract_spike.sh" first!

EOF
}

# location for log file
LOGFILE=./3_map.log

# variables to be used in main loop
reads1=(${BASEDIR}/UMP/*_unmapped_R1.fastq.gz) # collect each
forward read in array, e.g. "UMP/A_unmapped_R1.fastq.gz"
reads1=("${reads1[@]##*/}") # [@] refers to array, greedy remove
*/ from left, e.g. "A_unmapped_R1.fastq.gz"
reads2=("${reads1[@]/_R1/_R2}") # substitute R2 for R1, e.g.
"A_unmapped_R2.fastq.gz"

# main loop
pipeline() {

echo [`date +"%Y-%m-%d %H:%M:%S"`] "#> START: " $0 $@

for ((i=0; i<=${#reads1[@]}-1; i++)); do
  fwdrds="${reads1[$i]}" # e.g. "A_unmapped_R1.fastq.gz"
  rvsrds="${reads2[$i]}" # e.g. "A_unmapped_R2.fastq.gz"
  id="${fwdrds%%_*}" # greedy remove _* from right e.g. "A"

  # choose a reference genome
  echo ${id} processing...
  ref=$(
  awk -F"," -v id=$id '$1 == id { print $3 }'
${FASTALOC}/ref_decoder.csv
  )
  echo ${ref} selected as reference
```

```
  ## MAP TO MAIN GENOME
  # map (unmapped) reads against (concatenated) host and phix
genomes
  # -R = adding read group ID/sample to header
  bwa mem -t ${NUMCPUS} -R '@RG\tID:'"$id"'\tSM:'"$id"
${FASTALOC}/${ref}.fna \
  ${BASEDIR}/UMP/${fwdrds} ${BASEDIR}/UMP/${rvsrds} >
${BASEDIR}/TMP/${id}_refmapped_1.sam

  # SAM>BAM, filter for mapped reads and MAPQ>=20 (1 in 100),
pipe to sort by ref position (=default, don't use -n option);
cleanup
  samtools view -bS -F 4 -q 20
${BASEDIR}/TMP/${id}_refmapped_1.sam \
  | samtools sort -@ 3 -o ${BASEDIR}/MAP/${id}_refmapped_1.bam -
  rm ${BASEDIR}/TMP/${id}_refmapped_*.sam # delete SAM

  # let's index
  samtools index ${BASEDIR}/MAP/${id}_refmapped_1.bam

  # filtering to phix, removing cross-contig read pairs with awk,
removing non-ref @SQ with sed, then index
  samtools view -O SAM -h ${BASEDIR}/MAP/${id}_refmapped_1.bam
AF176034.1 | \
  awk '$7 == "=" || $1 ~ /^@/' | sed '/^@SQ/{/AF176034.1/!d;}' |
\
  samtools view -bS -o ${BASEDIR}/MAP/${id}_phixmapped_1.bam -
  samtools index ${BASEDIR}/MAP/${id}_phixmapped_1.bam

  ## MAP TO RESECTED GENOME
  # map (unmapped) reads against (concatenated) host and RESECTED
phix genomes
  bwa mem -t ${NUMCPUS} -R '@RG\tID:'"$id"'\tSM:'"$id"
${FASTALOC}/${ref}_resected.fna \
  ${BASEDIR}/UMP/${fwdrds} ${BASEDIR}/UMP/${rvsrds} >
${BASEDIR}/TMP/${id}_refmapped_2.sam

  # SAM>BAM with filter and sort - but push to BAM to TMP
(because we won't keep this one)
  samtools view -bS -F 4 -q 20
${BASEDIR}/TMP/${id}_refmapped_2.sam \
  | samtools sort -@ 3 -o ${BASEDIR}/TMP/${id}_refmapped_2.bam -
  rm ${BASEDIR}/TMP/${id}_refmapped_*.sam # delete SAM

  # let's index
  samtools index ${BASEDIR}/TMP/${id}_refmapped_2.bam

  # fetch BAM from TMP and select only those read pairs mapping
to resected phix (as above), index again and cleanup
  samtools view -O SAM -h ${BASEDIR}/TMP/${id}_refmapped_2.bam
RESTART_2694_RESECTED_AF176034.1 | \
  awk '$7 == "=" || $1 ~ /^@/' | sed
'/^@SQ/{/RESTART_2694_RESECTED_AF176034.1/!d;}' | \
  samtools view -bS -o ${BASEDIR}/MAP/${id}_phixmapped_2.bam -
  samtools index ${BASEDIR}/MAP/${id}_phixmapped_2.bam
  rm ${BASEDIR}/TMP/${id}_refmapped_*.* # remove BAM
done
```

```
echo [`date +"%Y-%m-%d %H:%M:%S"`] "#> DONE."
} #pipeline end

pipeline 2>&1 | tee $LOGFILE

#Acknowledgements
#SAM flags: https://broadinstitute.github.io/picard/explain-
flags.html
#considered f2 (proper pairs) but orientation is variable
#for awk: modified (invert match) from
https://www.biostars.org/p/118301/#118308
#for sed: modified (remove in-place) from
https://stackoverflow.com/a/27734472
```

A.4 Calling SNPs

```
#!/usr/bin/env bash

usage() {
  NAME=$(basename $0)
  cat <<EOF
Usage:
  ${NAME}
It is important that you run "3_map_reference.sh" first!

EOF
}

# location for log file
LOGFILE=./4_call.log

# main loop
pipeline() {

echo [`date +"%Y-%m-%d %H:%M:%S"`] "#> START: " $0 $@

# create key variables
# RegEx: zero or more any character followed by C = .*C
# RegEx cont(+): then one or more digits = [0-9]+
# RegEx cont(*): then straight onto _ or A, B or S first = [A-
B]*_
cregex1='.*C[0-9]+[A-B,S]*_phixmapped_1.bam'
cregex2='.*C[0-9]+[A-B,S]*_phixmapped_2.bam'
# RegEx: zero or more any character followed by S = .*S
# RegEx cont(+): then one or more digits before _ = [0-9]+_
sregex1='.*S[0-9]+_phixmapped_1.bam'
sregex2='.*S[0-9]+_phixmapped_2.bam'

# create BAM list files
ls -1 ${BASEDIR}/MAP/A_phixmapped_1.bam > ${BASEDIR}/C_list1.txt
find ${BASEDIR}/MAP/ -regextype egrep -regex ${cregex1} >>
${BASEDIR}/C_list1.txt
ls -1 ${BASEDIR}/MAP/A_phixmapped_2.bam > ${BASEDIR}/C_list2.txt
find ${BASEDIR}/MAP/ -regextype egrep -regex ${cregex2} >>
${BASEDIR}/C_list2.txt
find ${BASEDIR}/MAP/ -regextype egrep -regex ${sregex1} >
${BASEDIR}/S_list1.txt
```

```
find ${BASEDIR}/MAP/ -regextype egrep -regex ${sregex2} >
${BASEDIR}/S_list2.txt

# all files
cat C_list1.txt S_list1.txt > ALL_list1.txt
cat C_list2.txt S_list2.txt > ALL_list2.txt

# Run FreeBayes (C/S and 1/2)
for host in C S; do
  for ((i=1; i<=2; i++)); do
    # set reference
    reffile=phix_AF176034.fasta
    if [[ ${i} -eq 2 ]]; then
      reffile=phix_AF176034_resected.fasta
    fi
    # FreeBayes
    freebayes --fasta-reference ${FASTALOC}/${reffile} \
    --pooled-continuous --min-alternate-fraction 0.01 \
    --min-alternate-count 1 --min-mapping-quality 20 \
    --min-base-quality 30 --bam-list
${BASEDIR}/${host}_list${i}.txt \
    --vcf ${BASEDIR}/VCF/${host}_1STCALL_phix_${i}.vcf
  done
done

# Using GNU parallel bgzip and index the VCF outputs
parallel "bgzip {}" ::: ${BASEDIR}/VCF/*.vcf
parallel "bcftools index {}" ::: ${BASEDIR}/VCF/*.vcf.gz

#¬†Obtain union of sites using bcftools merge (and a text summary
with isec)
# Output called ignore to remind user not to interpret
for ((i=1; i<=2; i++)); do
  bcftools isec --nfiles -11 --collapse all \
  ${BASEDIR}/VCF/C_1STCALL_phix_${i}.vcf.gz \
  ${BASEDIR}/VCF/S_1STCALL_phix_${i}.vcf.gz \
  --output ${BASEDIR}/VCF/unionsites_${i}.txt

  bcftools merge --merge both \
  ${BASEDIR}/VCF/C_1STCALL_phix_${i}.vcf.gz \
  ${BASEDIR}/VCF/S_1STCALL_phix_${i}.vcf.gz \
  --output-type v --output ${BASEDIR}/VCF/ignore_${i}.vcf
done

# Run FreeBayes again with ignore file (so it inspects union of
sites)
for ((i=1; i<=2; i++)); do
  # set reference
  reffile=phix_AF176034.fasta
  if [[ ${i} -eq 2 ]]; then
    reffile=phix_AF176034_resected.fasta
  fi
  # FreeBayes
  freebayes --fasta-reference ${FASTALOC}/${reffile} \
  --pooled-continuous --min-alternate-fraction 0.01 \
  --min-alternate-count 1 --min-mapping-quality 20 \
  --min-base-quality 30 --variant-input
${BASEDIR}/VCF/ignore_${i}.vcf \
```

```
  --bam-list ${BASEDIR}/ALL_list${i}.txt \
  --vcf ${BASEDIR}/VCF/ALL_UNIONCALL_phix_${i}.vcf
done

# Clean up: remove ignore outputs
rm ${BASEDIR}/VCF/ignore_*.vcf

echo [`date +"%Y-%m-%d %H:%M:%S"`] "#> DONE."
} #pipeline end

pipeline 2>&1 | tee $LOGFILE
```

A.5 Annotate

```
#!/usr/bin/env bash

usage() {
  NAME=$(basename $0)
  cat <<EOF
Usage:
  ${NAME}
It is important that you run "3_map_reference.sh" first!

EOF
}

# location for log file
LOGFILE=./5_annotate.log

# main loop
pipeline() {

echo [`date +"%Y-%m-%d %H:%M:%S"`] "#> START: " $0 $@

# Let's operate within the VCF directory
cd ${BASEDIR}/VCF/

# Run filter on 1st call samples
echo "   * Filtering 1st call samples."
for host in C S; do
  for ((i=1; i<=2; i++)); do
    echo Host ${host}. File ${i}.
    gunzip ${host}_1STCALL_phix_${i}.vcf.gz
    vcffilter -f "SRP > 20" -f "SAP > 20" -f "EPP > 20" -f "QUAL
> 30" -f "DP > 30" \
    ${host}_1STCALL_phix_${i}.vcf >
Filtered_${host}_1STCALL_phix_${i}.vcf
    echo
  done
done

# Recount the resected VCFs
echo "   * Correcting resected VCFs."
for host in C S; do
  ${BASEDIR}/python_scripts/2_recount_resected.py
${host}_1STCALL_phix_2.vcf
  ${BASEDIR}/python_scripts/2_recount_resected.py
Filtered_${host}_1STCALL_phix_2.vcf
```

```
done
${BASEDIR}/python_scripts/2_recount_resected.py
ALL_UNIONCALL_phix_2.vcf

# Run annotation pipeline (echo statement "*" is passing default
ARF character to script)
echo "   * Generating annotation files from VCFs."
for ind in "1" corrected; do
  echo Host C, 1st pass. File ${ind}.
  echo "*" | python ${BASEDIR}/python_scripts/3_vcf_parser.py \
  ${BASEDIR}/FASTA/phix_AF176034.fasta
${BASEDIR}/FASTA/phix_coord.txt \
  Filtered_C_1STCALL_phix_${ind}.vcf
Filtered_C_1STSUMMARY_${ind}.tsv
  echo
  echo Host S, 1st pass. File ${ind}.
  echo "*" | python ${BASEDIR}/python_scripts/3_vcf_parser.py \
  ${BASEDIR}/FASTA/phix_AF176034.fasta
${BASEDIR}/FASTA/phix_coord.txt \
  Filtered_S_1STCALL_phix_${ind}.vcf
Filtered_S_1STSUMMARY_${ind}.tsv
  echo
  echo All, union calls. File ${ind}.
  echo "*" | python ${BASEDIR}/python_scripts/3_vcf_parser.py \
  ${BASEDIR}/FASTA/phix_AF176034.fasta
${BASEDIR}/FASTA/phix_coord.txt \
  ALL_UNIONCALL_phix_${ind}.vcf ALL_UNIONSUMMARY_${ind}.tsv
  echo
done
# N.B. the annotation script will handle multi-sample VCFs and
multi-allelic
# sites, but multi-nucleotide sites will be rejected. Please
examine log file.

# housekeeping
mkdir TABLES
mv *.tsv TABLES
mkdir FILTER_VCFS
mv Filtered_* FILTER_VCFS/
mkdir FIRST_VCFS
mv *1STCALL* FIRST_VCFS/
mkdir UNION_VCFS
mv *UNIONCALL* UNION_VCFS/

echo [`date +"%Y-%m-%d %H:%M:%S"`] "#> DONE."
} #pipeline end

pipeline 2>&1 | tee $LOGFILE
```

A.6 Resected genome recount

```
#!/usr/bin/env python

__author__ = "Ben Dickins"
__version__ = "1.0"

import sys
```

```
from pysam import VariantFile

if __name__ == "__main__":

    # genome name
    vcf_name = sys.argv[1].split('.vcf')[0]
    vcf_in = VariantFile(vcf_name+'.vcf')  # auto-detect input
format
    vcf_out = VariantFile(vcf_name+'_reindexed.vcf', 'w',
header=vcf_in.header)

    # resect coord
    print(list((vcf_in.header.contigs)))
    coord = 2694 # 1-based
    glen = 5386

    # main loop
    for rec in vcf_in.fetch():
        rec.pos += (coord-1) #¬†so 1 becomes 2694
        if rec.pos > glen:
            rec.pos -= glen # so 5387 becomes 1
        vcf_out.write(rec)
        #print(rec, file=vcf_out)
    vcf_in.close()
    vcf_out.close()
```

A.7 Annotated tabular files

```
#! /usr/bin/env python
__author__ = "Ben Dickins"
__status__ = "Prototype"
__version__ = "0.1"

# pretty print function
def prettyprint(x, y, handle):
    if len(x) == 0:
        print("INT", end="\t", file=handle)
    elif len(x) == 1:
        print(x[0], end="\t", file=handle)
    elif len(x) == 2:
        if y == 0:
            print( "{}({})".format(x[0],x[1]), end="\t",
file=handle )
        else:
            print( "{}/{}".format(x[0],x[1]), end="\t",
file=handle )
    else:
        print( "{}/{}".format(x[0],x[1]), end="", file=handle )
        for item in x[2:]:
            handle.write('(' + item + ')')
        handle.write('\t')

# DNA/protein position reporter - N.B. all inputs and outputs are
1-based!
def withincheck(position, gene_start, gene_end, length):
    if gene_start > gene_end: # adjustments for origin breakers
        if position <= gene_end:
            position += length
```

```
        gene_end += length
    if gene_start <= gene_end: # now the main logic
        if position >= gene_start and position <= gene_end:
            dna_pos = position - gene_start + 1
            prot_pos = -(-dna_pos//3)
            return(dna_pos, prot_pos)
        else:
            return(False)
    else:
        raise UserWarning("Gene start/end conflict (even
correcting for genome length).")
# note the // operator used for prot_pos rounds down when numbers
are negative
# hijacked it here with double negative for rounding up

# main loops (also accessible if called as vcf_parser.main)
def main():
    # handle command line input from user
    import argparse
    parser = argparse.ArgumentParser()
    parser.add_argument("genome_fa", type=str, help="FASTA genome
file")
    parser.add_argument("feat_table", type=str, help="feature
table file")
    parser.add_argument("vcf_file", type=str, help="FreeBayes VCF
output")
    parser.add_argument("output", type=str, help="Output file")
    args = parser.parse_args()

    # read single fasta sequence file
    with open(args.genome_fa,'rU') as file:
        genome = ''
        for line in file:
            if not line:
                pass
            elif not line.startswith('>'):
                genome += line.rstrip()
        length = len(genome)
        print("Measured genome length is", length, "bases.")

    # read coordinates from feature table
    genecoords, subsequences = {}, {}
    with open(args.feat_table,'rU') as file:
        for line in file:
            if not line:
                pass
            elif "Product" not in line:
                line = line.split('\t')
                gen = line[0].strip()
                beg, ter = int(line[1].strip()),
int(line[2].strip())
                genecoords[gen] = (beg, ter)
                # coordinates so far are 1-based, so seq slices
must accommodate this
                if beg <= ter:
                    subseq = genome[ beg-1:ter ] # remember
python slicing is exclusive of end
                elif beg > ter:
```

```
                    subseq = genome[ beg-1:length] + genome[
0:ter ]
                subsequences[gen] = subseq

    # genetic code #11, but not annotating alternative start
codons
    # reference:
http://www.bioinformatics.org/JaMBW/2/3/TranslationTables.html#SG
11
    codons = {
        'TTT': 'F', 'TCT': 'S', 'TAT': 'Y', 'TGT': 'C',
        'TTC': 'F', 'TCC': 'S', 'TAC': 'Y', 'TGC': 'C',
        'TTA': 'L', 'TCA': 'S', 'TAA': '*', 'TGA': '*',
        'TTG': 'L', 'TCG': 'S', 'TAG': '*', 'TGG': 'W',
        'CTT': 'L', 'CCT': 'P', 'CAT': 'H', 'CGT': 'R',
        'CTC': 'L', 'CCC': 'P', 'CAC': 'H', 'CGC': 'R',
        'CTA': 'L', 'CCA': 'P', 'CAA': 'Q', 'CGA': 'R',
        'CTG': 'L', 'CCG': 'P', 'CAG': 'Q', 'CGG': 'R',
        'ATT': 'I', 'ACT': 'T', 'AAT': 'N', 'AGT': 'S',
        'ATC': 'I', 'ACC': 'T', 'AAC': 'N', 'AGC': 'S',
        'ATA': 'I', 'ACA': 'T', 'AAA': 'K', 'AGA': 'R',
        'ATG': 'M', 'ACG': 'T', 'AAG': 'K', 'AGG': 'R',
        'GTT': 'V', 'GCT': 'A', 'GAT': 'D', 'GGT': 'G',
        'GTC': 'V', 'GCC': 'A', 'GAC': 'D', 'GGC': 'G',
        'GTA': 'V', 'GCA': 'A', 'GAA': 'E', 'GGA': 'G',
        'GTG': 'V', 'GCG': 'A', 'GAG': 'E', 'GGG': 'G'}

    # types of amino acids - N.B. this is a minimal
classification
    # could switch to this EBI classification:
https://en.wikipedia.org/wiki/Conservative_mutation
    groups = {
        'G':'NP', 'A':'NP', 'V':'NP', 'L':'NP', 'I':'NP',
        'F':'NP', 'M':'NP', 'P':'NP', 'W':'NP', 'S':'PO',
        'T':'PO', 'Y':'PO', 'C':'PO', 'N':'PO', 'Q':'PO',
        'D':'AC', 'E':'AC', 'H':'BA', 'K':'BA', 'R':'BA',
        '*':'ST'}

    # ask the user to identify a character that describes
overlapping genes in the same ORF
    orf_char = input("Enter character found in non-ARF
overlapping genes (default: *): ") or "*"

    # read vcf file checking each (alternative base at each)
position against all genes
    with open(args.vcf_file, 'rU') as file:
        outfile = open(args.output,'w')
        print("Site\tProtein(s)\tAmino
acid(s)\tRadicality\tCoverage\tMAC\tMAF", file=outfile)
        for line in file:
            if not line:
                pass
            elif not line.startswith("#"):
                line = line.split("\t")
                refpos = int(line[1]) # very important to note
this is 1-based!!
                refnuc, altnucs = line[3], line[4].split(',')
                field_names = line[8].split(':')
```

```
                dp_idx, ad_idx = field_names.index('DP'),
field_names.index('AD')
                field_num = line[9].split(':')
                dp_num, ad_num = field_num[dp_idx],
field_num[ad_idx].split(',')[1:] # 0th val is ref allele

                # we may have >1 alternative allele so we must
loop through these + their # of reads
                assert len(altnucs) == len(ad_num) # these should
be the same length
                for base, num in zip(altnucs, ad_num):
                    prot_list, change_aa, change_type = [], [],
[]
                    orf_check = 0 # will be passed to prettyprint
as y (positional) argument
                    change_pos = refnuc + str(refpos) + base
                    for gen, coord in genecoords.items():
                        beg, ter = coord # tuple unpacking
                        within_gene = withincheck(refpos, beg,
ter, length)
                        if within_gene:
                            prot_list.append(gen)
                            if orf_char in gen:
                                orf_check = 1
                            dna_pos, prot_pos = within_gene #
tuple unpacking
                            dna_pos -= 1 # this is now 0-based!!
                            codon_position = dna_pos % 3 # codon:
0, 1, 2

                            subseq = subsequences[gen]

                            if codon_position == 0:
                                refcodon =
subseq[dna_pos:dna_pos+3]
                                newcodon = base + refcodon[1:3]

                            elif codon_position == 1:
                                refcodon = subseq[dna_pos-
1:dna_pos+2]
                                newcodon = refcodon[0] + base +
refcodon[2]

                            else:
                                refcodon = subseq[dna_pos-
2:dna_pos+1]
                                newcodon = refcodon[0:2] + base

                            oldaa = codons[refcodon]
                            newaa = codons[newcodon]
                            change_aa.append(oldaa +
str(prot_pos) + newaa)
                            if oldaa == newaa:
                                change_type.append("SYN")
                            elif groups[oldaa] == groups[newaa]:
                                change_type.append('CON')
                            else:
                                change_type.append('RAD')
```

```
                    print(change_pos, end="\t", file=outfile)
                    prettyprint(prot_list, orf_check, outfile)
                    prettyprint(change_aa, orf_check, outfile)
                    prettyprint(change_type, orf_check, outfile)
                    print(dp_num, num, float(num)/float(dp_num),
sep="\t", end="\n", file=outfile)
        outfile.close()

if __name__ == "__main__":
    main()
```

Appendix B – Summary of alleles, allele frequencies and samples

Site	Protein(s)	Amino acid(s	Radicality	SCSC2_cvrg	SCSC2_altdp	SCSC2_AF	S2_cvrg	S2_altdp	S2_AF	2NDS2_cvrg	2NDS2_altdp	2NDS2_AF	CSCS10_cvrg
C199T	K(C)	T50I(L23L)	RAD(SYN)	1008	7	0.00694444	855	0	0	12927	0	0	18758
A323G	C	D64G	RAD	910	0	0	773	0	0	15518	0	0	16574
C530T	D	D47D	SYN	1122	0	0	894	0	0	16229	1	6.16E-05	20839
T572C	D(E)	G61G(V2A)	SYN(CON)	1041	0	0	806	0	0	16061	2	0.00012453	19005
A590G	D(E)	G67G(D8G)	SYN(RAD)	985	0	0	790	0	0	14264	1	7.01E-05	18287
C648G	D(E)	H87D(F27L)	RAD(CON)	1017	264	0.25958702	837	832	0.99402628	15856	15813	0.99728809	18780
G944T	J	V33L	CON	877	0	0	851	0	0	18476	1	5.41E-05	18148
C1137T	F	A46V	CON	904	0	0	859	0	0	15671	0	0	19722
C1148A	F	L50I	CON	933	0	0	899	0	0	15205	0	0	20320
T1307C	F	Y103H	RAD	957	8	0.00835946	902	0	0	18002	9	0.00049994	20435
A1317G	F	H106R	CON	889	0	0	851	0	0	17142	0	0	18666
C1460A	F	Q154K	RAD	1187	475	0.40016849	1045	49	0.04688995	16362	675	0.04125413	23895
A1548G	F	Q183R	RAD	1113	0	0	974	398	0.40862423	16097	6439	0.40001242	22658
A1637G	F	M213V	CON	1151	0	0	1045	0	0	15668	5	0.00031912	22933
G1639T	F	M213I	CON	1145	1	0.00087336	1038	0	0	15833	4	0.00025264	22739
C1727T	F	L243F	CON	1159	0	0	1053	0	0	16796	1	5.95E-05	23451
T1956G	F	V319G	CON	1087	267	0.24563017	1004	1003	0.99900398	17224	17183	0.9976196	22121
T2083A	F	P361P	SYN	1004	0	0	910	0	0	15886	25	0.00157371	20206
C2085T	F	A362V	CON	1026	238	0.23196881	938	241	0.25692964	16510	3536	0.21417323	20856
C2093A	F	L365I	CON	989	5	0.00505561	900	0	0	17628	0	0	20023
G2179T	F	Q393H	RAD	847	1	0.00118064	827	1	0.00120919	19286	0	0	18366
G2275A	F	M425I	CON	912	660	0.72368421	839	1	0.0011919	15861	31	0.00195448	18500
A2276T	F	T426S	CON	897	0	0	820	0	0	16082	4	0.00024873	18167
T2321C	INT	INT	INT	956	3	0.00313808	829	10	0.01206273	15040	65	0.00432181	18729
C2971T	H	A14V	CON	921	243	0.26384365	858	857	0.9988345	16849	16784	0.9961422	18935
G3071A	H	M47I	CON	848	0	0	762	32	0.04199475	18523	387	0.02089294	16236
G3111A	H	V61I	CON	928	0	0	819	0	0	17927	3	0.00016735	17967
C3120T	H	P64S	RAD	912	25	0.02741228	840	0	0	17456	3	0.00017186	18538
G3129T	H	A67S	RAD	876	225	0.25684932	800	800	1	16850	16800	0.99703264	17310
G3132A	H	A68T	RAD	858	0	0	808	3	0.00371287	16748	44	0.00262718	17551
G3339A	H	D137N	RAD	1065	1065	1	891	891	1	14897	14886	0.9992616	19902
A5360C	A/A*(B)	I460I/I288I(K	SYN/SYN(RA	802	214	0.26683292	679	675	0.99410898	14217	14164	0.99627207	15630

Site	Protein(s)	Amino acid(s	Radicality	CSCS10_altd	CSCS10_AF	C1_cvrg	C1_altdp	C1_AF	SCSC3_cvrg	SCSC3_altdp	SCSC3_AF	SC10_cvrg	SC10_altdp	SC10_AF	SCSC10A_cvr	SCSC10A_alt	SCSC10A_AF
C199T	K(C)	T50I(L23L)	RAD(SYN)	0	0	2055	15	0.00729927	16483	60	0.00364011	5757	28	0.00486364	23841	100	0.00419445
A323G	C	D64G	RAD	0	0	1828	0	0	14245	0	0	5335	0	0	20864	2	9.59E-05
C530T	D	D47D	SYN	0	0	2185	1	0.00045767	18254	0	0	6660	0	0	26693	1	3.75E-05
T572C	D(E)	G61G(V2A)	SYN(CON)	0	0	2086	0	0	16723	1	5.98E-05	6157	91	0.01477993	24735	0	0
A590G	D(E)	G67G(D8G)	SYN(RAD)	68	0.00371849	1967	0	0	15874	43	0.00270883	5855	5	0.00085397	23713	536	0.02260364
C648G	D(E)	H87D(F27L)	RAD(CON)	18280	0.97337593	2116	30	0.01417769	16697	10922	0.65412948	6119	256	0.0418369	24739	3400	0.13743482
G944T	J	V33L	CON	2	0.0001102	1941	4	0.00206079	15765	0	0	5745	0	0	22911	1	4.36E-05
C1137T	F	A46V	CON	1	5.07E-05	2021	0	0	16996	0	0	6030	0	0	24798	1755	0.07077184
C1148A	F	L50I	CON	1	4.92E-05	2081	0	0	17539	3	0.00017105	6177	0	0	25532	0	0
T1307C	F	Y103H	RAD	5	0.00024468	2042	10	0.00489716	17843	66	0.00369893	6418	153	0.0238392	24714	180	0.00728332
A1317G	F	H106R	CON	1	5.36E-05	1877	0	0	16229	0	0	5849	71	0.01213883	22480	2	8.90E-05
C1460A	F	Q154K	RAD	1011	0.04231011	2356	1392	0.59083192	21167	3301	0.1559503	7374	4085	0.55397342	29480	12722	0.43154681
A1548G	F	Q183R	RAD	8436	0.37231883	2255	0	0	20142	1	4.96E-05	6759	0	0	28272	2	7.07E-05
A1637G	F	M213V	CON	1	4.36E-05	2454	32	0.01303993	20689	3	0.000145	7370	1	0.00013569	29984	4	0.0001334
G1639T	F	M213I	CON	0	0	2426	22	0.00906843	20583	3	0.00014575	7290	105	0.01440329	29768	0	0
C1727T	F	L243F	CON	0	0	2407	0	0	21034	4	0.00019017	7469	0	0	30538	2278	0.07459559
T1956G	F	V319G	CON	21576	0.97536278	2497	25	0.01001201	19920	12889	0.64703815	7365	272	0.03693143	29617	4238	0.14309349
T2083A	F	P361P	SYN	0	0	2321	0	0	18558	1	5.39E-05	6811	0	0	26862	1	3.72E-05
C2085T	F	A362V	CON	5606	0.26879555	2389	0	0	18987	12004	0.63222205	6980	2	0.00028653	27565	2007	0.07280972
C2093A	F	L365I	CON	0	0	2335	0	0	18117	277	0.01528951	6784	2	0.00029481	26398	1480	0.05606485
G2179T	F	Q393H	RAD	1	5.44E-05	2122	0	0	16396	239	0.01457673	6210	0	0	23741	1334	0.05618971
G2275A	F	M425I	CON	442	0.02389189	2262	2070	0.91511936	16765	5889	0.35126752	6372	5999	0.94146265	24313	20933	0.86097972
A2276T	F	T426S	CON	0	0	2203	58	0.02632773	16383	0	0	6273	0	0	23812	0	0
T2321C	INT	INT	INT	127	0.00678093	2198	18	0.00818926	16791	139	0.00827824	6289	42	0.00667833	24034	171	0.00711492
C2971T	H	A14V	CON	18479	0.97591761	1947	20	0.01027221	16736	11305	0.67548996	5811	225	0.03871967	23412	3414	0.14582266
G3071A	H	M47I	CON	405	0.02494457	1796	0	0	14678	0	0	5333	0	0	21401	1	4.67E-05
G3111A	H	V61I	CON	3	0.00016697	2019	0	0	16240	2	0.00012315	5859	0	0	23490	1523	0.0648361
C3120T	H	P64S	RAD	2	0.00010789	2041	7	0.00342969	16580	73	0.0044029	5955	33	0.00554156	24136	413	0.01711137
G3129T	H	A67S	RAD	16847	0.97325246	1998	0	0	15724	10416	0.66242686	5700	3	0.00052632	22937	2871	0.12516894
G3132A	H	A68T	RAD	42	0.00239303	1965	0	0	15894	12	0.000755	5664	76	0.01341808	22816	68	0.00298036
G3339A	H	D137N	RAD	19900	0.99989951	2434	2225	0.91413311	18219	18217	0.99989022	6759	6655	0.98461311	25778	25767	0.99957328
A5360C	A/A*(B)	I460I/I288I(K	SYN/SYN(RA	15219	0.97370441	1849	26	0.01406165	13365	8819	0.65985784	4722	185	0.03917831	19844	2767	0.13943761

Site	Protein(s)	Amino acid(s	Radicality	SC1_cvrg	SC1_altdp	SC1_AF	C10_cvrg	C10_altdp	C10_AF	SCSC1_cvrg	SCSC1_altdp	SCSC1_AF	SCS1_cvrg	SCS1_altdp	SCS1_AF	S8_cvrg	S8_altdp
C199T	K(C)	T50I(L23L)	RAD(SYN)	6417	15	0.00233754	6320	25	0.0039557	8878	27	0.00304123	13695	0	0	20179	0
A323G	C	D64G	RAD	6025	1	0.00016598	5769	89	0.01542728	7771	1	0.00012868	12071	1	8.28E-05	17787	0
C530T	D	D47D	SYN	7716	0	0	7401	3	0.00040535	9615	1	0.000104	15019	1	6.66E-05	22080	0
T572C	D(E)	G61G(V2A)	SYN(CON)	7051	0	0	6795	0	0	8936	2	0.00022381	13974	2	0.00014312	20426	1
A590G	D(E)	G67G(D8G)	SYN(RAD)	6639	0	0	6383	6	0.00094	8525	7	0.00082111	13493	1	7.41E-05	19606	0
C648G	D(E)	H87D(F27L)	RAD(CON)	7126	2966	0.41622228	6649	53	0.00797112	9159	104	0.01135495	13921	13893	0.99798865	20256	20096
G944T	J	V33L	CON	6364	0	0	6355	0	0	8597	2	0.00023264	13641	1	7.33E-05	19563	0
C1137T	F	A46V	CON	6538	0	0	6664	1	0.00015006	9412	3	0.00031874	14832	0	0	21199	0
C1148A	F	L50I	CON	6741	3	0.00044504	6903	4701	0.68100826	9656	0	0	15467	2	0.00012931	21904	0
T1307C	F	Y103H	RAD	7030	42	0.0059744	6860	13	0.00189504	9281	91	0.00980498	15854	5	0.00031538	21639	15
A1317G	F	H106R	CON	6429	1	0.00015555	6268	0	0	8357	0	0	14601	5	0.00034244	19882	3
C1460A	F	Q154K	RAD	8528	3820	0.44793621	8066	5872	0.72799405	10753	7505	0.69794476	18528	632	0.03411054	25561	860
A1548G	F	Q183R	RAD	8116	0	0	7341	0	0	10243	1	9.76E-05	17588	6647	0.37792813	24356	9373
A1637G	F	M213V	CON	8394	0	0	7984	0	0	11079	0	0	18056	0	0	25345	3
G1639T	F	M213I	CON	8321	3	0.00036053	7934	43	0.00541971	10958	0	0	17867	1	5.60E-05	25158	14
C1727T	F	L243F	CON	8256	0	0	7917	12	0.00151573	11186	0	0	18080	1	5.53E-05	25701	0
T1956G	F	V319G	CON	7956	3313	0.41641528	8272	46	0.00556093	11051	113	0.01022532	16870	16841	0.99828097	23833	23679
T2083A	F	P361P	SYN	7533	0	0	7615	2	0.00026264	10025	1	9.98E-05	15495	1	6.45E-05	22074	2
C2085T	F	A362V	CON	7759	3	0.00038665	7771	9	0.00115815	10269	1	9.74E-05	15923	3904	0.24517993	22677	6269
C2093A	F	L365I	CON	7481	0	0	7552	0	0	9853	2	0.00020298	15398	1	6.49E-05	21610	2
G2179T	F	Q393H	RAD	6809	0	0	6749	2	0.00029634	9035	0	0	13969	0	0	19889	1
G2275A	F	M425I	CON	7137	4168	0.58399888	7064	1932	0.27349943	9522	9400	0.98718757	14347	20	0.00139402	20407	124
A2276T	F	T426S	CON	6992	4	0.00057208	6965	4995	0.71715721	9328	0	0	14025	0	0	19981	7
T2321C	INT	INT	INT	6922	61	0.00881248	6922	48	0.00693441	9250	68	0.00735135	14689	208	0.01416026	20670	140
C2971T	H	A14V	CON	6767	2964	0.43800798	6128	38	0.00620104	8415	106	0.01259655	14226	14208	0.99873471	19611	19481
G3071A	H	M47I	CON	6002	0	0	5686	2	0.00035174	7818	0	0	12398	313	0.02524601	17498	373
G3111A	H	V61I	CON	6647	0	0	6466	0	0	8707	2	0.0002297	13277	2	0.00015064	19585	3
C3120T	H	P64S	RAD	6721	0	0	6611	5	0.00075632	8846	23	0.00260005	13563	3	0.00022119	20048	2
G3129T	H	A67S	RAD	6377	0	0	6338	0	0	8395	20	0.00238237	12745	11518	0.90372695	18858	18717
G3132A	H	A68T	RAD	6314	1	0.00015838	6271	0	0	8338	17	0.00203886	12885	34	0.00263873	19178	52
G3339A	H	D137N	RAD	7293	7289	0.99945153	7381	2105	0.28519171	9864	9862	0.99979724	14583	14581	0.99986285	22083	22048
A5360C	A/A*(B)	I460I/I288I(K	SYN/SYN(RA	5405	2238	0.41406105	5318	50	0.00940203	7740	121	0.01563307	11522	11508	0.99878493	16925	16787

Site	Protein(s)	Amino acid(s	Radicality	S8_AF	CSC1_cvrg	CSC1_altdp	CSC1_AF	CSCS1_cvrg	CSCS1_altdp	CSCS1_AF	2NDSCSC2A_	2NDSCSC2A_	2NDSCSC2A_	2NDSCSC2B_	2NDSCSC2B_	2NDSCSC2B_	S3_cvrg
C199T	K(C)	T50I(L23L)	RAD(SYN)	0	5885	36	0.00611725	23322	2	8.58E-05	44821	611	0.013632	49276	662	0.01343453	15192
A323G	C	D64G	RAD	0	5165	0	0	20478	0	0	51443	3	5.83E-05	58910	0	0	13670
C530T	D	D47D	SYN	0	6535	0	0	26103	0	0	57530	5	8.69E-05	66059	7	0.00010597	15656
T572C	D(E)	G61G(V2A)	SYN(CON)	4.90E-05	6075	1	0.00016461	23969	2	8.34E-05	59387	11	0.00018523	66927	5	7.47E-05	14498
A590G	D(E)	G67G(D8G)	SYN(RAD)	0	5812	5	0.00086029	22996	25	0.00108715	54149	190	0.00350884	60602	340	0.00561038	13750
C648G	D(E)	H87D(F27L)	RAD(CON)	0.99210111	6185	135	0.021827	23116	23019	0.99580377	61912	1968	0.03178705	69550	2179	0.03132998	14226
G944T	J	V33L	CON	0	5573	0	0	21957	0	0	73698	0	0	82261	0	0	15234
C1137T	F	A46V	CON	0	5998	1	0.00016672	24402	1	4.10E-05	60996	3	4.92E-05	69989	3	4.29E-05	16314
C1148A	F	L50I	CON	0	6157	1	0.00016242	25043	4	0.00015973	60994	2	3.28E-05	69257	2	2.89E-05	17043
T1307C	F	Y103H	RAD	0.00069319	6078	50	0.00822639	24008	25	0.00104132	68638	68	0.0009907	75899	114	0.001502	18262
A1317G	F	H106R	CON	0.00015089	5520	0	0	22153	1	4.51E-05	64015	4	6.25E-05	71208	5	7.02E-05	16864
C1460A	F	Q154K	RAD	0.03364501	7377	3968	0.53788803	28380	1011	0.03562368	54695	2	3.66E-05	62675	3	4.79E-05	20980
A1548G	F	Q183R	RAD	0.38483331	6962	1	0.00014364	27157	10306	0.379497	57817	3	5.19E-05	64739	4	6.18E-05	19882
A1637G	F	M213V	CON	0.00011837	7332	0	0	28056	2	7.13E-05	56326	12	0.00021305	64131	8	0.00012474	20728
G1639T	F	M213I	CON	0.00055648	7300	0	0	27940	3	0.00010737	56841	1	1.76E-05	64645	4	6.19E-05	20497
C1727T	F	L243F	CON	0	7470	0	0	28234	0	0	59918	1	1.67E-05	67446	3	4.45E-05	20778
T1956G	F	V319G	CON	0.99353837	6987	132	0.01889223	27059	26940	0.9956022	54834	2037	0.03714848	62921	2267	0.03602931	17924
T2083A	F	P361P	SYN	9.06E-05	6382	0	0	24972	0	0	51608	5	9.69E-05	57513	2	3.48E-05	17328
C2085T	F	A362V	CON	0.2764475	6538	0	0	25683	7167	0.27905619	53975	269	0.00498379	60043	332	0.00552937	17780
C2093A	F	L365I	CON	9.25E-05	6303	0	0	24863	3	0.00012066	51783	2	3.86E-05	57352	7	0.00012205	17018
G2179T	F	Q393H	RAD	5.03E-05	5845	1	0.00017109	22851	1	4.38E-05	55953	0	0	61187	4	6.54E-05	15766
G2275A	F	M425I	CON	0.00607635	6290	6165	0.98012719	23188	112	0.00483008	49897	48014	0.96226226	55336	53315	0.96347766	15864
A2276T	F	T426S	CON	0.00035033	6170	0	0	22778	0	0	49798	4	8.03E-05	55311	2	3.62E-05	15538
T2321C	INT	INT	INT	0.0067731	6291	58	0.00921952	23406	363	0.01550884	46882	12	0.00025596	53202	33	0.00062028	15969
C2971T	H	A14V	CON	0.99337107	5633	131	0.02325581	22343	22242	0.99547957	58033	2132	0.03673772	68743	2565	0.03731289	16377
G3071A	H	M47I	CON	0.02131672	5199	0	0	19788	500	0.02526784	63677	7	0.00010993	72821	6	8.24E-05	14681
G3111A	H	V61I	CON	0.00015318	5834	0	0	22201	3	0.00013513	57738	9	0.00015588	66429	17	0.00025591	16023
C3120T	H	P64S	RAD	9.98E-05	5951	2	0.00033608	22863	6	0.00026243	57438	144	0.00250705	65224	207	0.00317368	16217
G3129T	H	A67S	RAD	0.99252307	5679	0	0	21466	21381	0.99604025	53866	1574	0.02922066	62079	1769	0.02849595	15113
G3132A	H	A68T	RAD	0.00271144	5580	0	0	21791	51	0.00234042	53285	12	0.0002252	61439	21	0.0003418	15358
G3339A	H	D137N	RAD	0.99841507	6758	6758	1	25123	25116	0.99972137	46978	46969	0.99980842	55713	55706	0.99987436	17216
A5360C	A/A*(B)	I460I/I288I(K	SYN/SYN(RA	0.99184638	5029	91	0.01809505	19946	19843	0.99483606	43468	1415	0.03255268	50277	1649	0.0327983	12177

Site	Protein(s)	Amino acid(s	Radicality	S3_altdp	S3_AF	2NDS1_cvrg	2NDS1_altdp	2NDS1_AF	CSC10_cvrg	CSC10_altdp	CSC10_AF	SCSC10B_cvr	SCSC10B_alt	SCSC10B_AF	S1_cvrg	S1_altdp	S1_AF
C199T	K(C)	T50I(L23L)	RAD(SYN)	3	0.00019747	6207	0	0	15993	84	0.0052523	15993	82	0.00512724	21569	6	0.00027818
A323G	C	D64G	RAD	2	0.00014631	7540	0	0	14111	2	0.00014173	14790	0	0	20191	0	0
C530T	D	D47D	SYN	0	0	7725	0	0	18219	2	0.00010978	17371	1	5.76E-05	23038	2	8.68E-05
T572C	D(E)	G61G(V2A)	SYN(CON)	0	0	7565	1	0.00013219	16993	1	5.88E-05	16125	1	6.20E-05	21347	2	9.37E-05
A590G	D(E)	G67G(D8G)	SYN(RAD)	3	0.00021818	6855	1	0.00014588	16184	20	0.00123579	15459	311	0.02011773	20337	1	4.92E-05
C648G	D(E)	H87D(F27L)	RAD(CON)	14209	0.998805	7653	7534	0.98445054	16812	335	0.01992624	16619	2444	0.14706059	21895	161	0.00735328
G944T	J	V33L	CON	1	6.56E-05	8847	11	0.00124336	15799	0	0	16828	0	0	23104	6	0.0002597
C1137T	F	A46V	CON	4	0.00024519	7034	0	0	16726	3	0.00017936	17958	1295	0.07211271	24929	2	8.02E-05
C1148A	F	L50I	CON	6	0.00035205	6864	0	0	17272	2	0.00011579	18689	0	0	25976	5	0.00019249
T1307C	F	Y103H	RAD	16	0.00087614	8351	8	0.00095797	16628	143	0.00859995	18513	116	0.00626587	27859	16	0.00057432
A1317G	F	H106R	CON	0	0	7835	0	0	15077	3	0.00019898	16790	1	5.96E-05	25365	4	0.0001577
C1460A	F	Q154K	RAD	807	0.0384652	8227	436	0.05299623	19175	11187	0.58341591	21231	9240	0.43521266	30585	1	3.27E-05
A1548G	F	Q183R	RAD	7488	0.37662207	7501	3106	0.41407812	18015	1	5.55E-05	19981	5	0.00025024	28923	3	0.00010372
A1637G	F	M213V	CON	2	9.65E-05	7459	1	0.00013407	19140	1	5.22E-05	21514	1	4.65E-05	30568	5	0.00016357
G1639T	F	M213I	CON	2	9.76E-05	7508	8	0.00106553	19004	0	0	21329	1	4.69E-05	30286	14	0.00046226
C1727T	F	L243F	CON	1	4.81E-05	7743	0	0	19723	0	0	21972	1850	0.08419807	30909	1	3.24E-05
T1956G	F	V319G	CON	17890	0.9981031	8614	8510	0.98792663	19904	374	0.01879019	20079	2877	0.14328403	28270	161	0.00569508
T2083A	F	P361P	SYN	0	0	7771	80	0.01029469	18321	1	5.46E-05	19010	3	0.00015781	26024	0	0
C2085T	F	A362V	CON	4781	0.26889764	8043	1172	0.14571677	18771	0	0	19486	1527	0.07836395	26631	0	0
C2093A	F	L365I	CON	0	0	9133	0	0	18303	0	0	18625	1024	0.05497987	25624	6	0.00023416
G2179T	F	Q393H	RAD	0	0	9695	0	0	16434	1	6.08E-05	17262	954	0.0552659	23148	0	0
G2275A	F	M425I	CON	34	0.00214322	7821	26	0.00332438	16775	16448	0.98050671	17785	15226	0.8561147	23907	23729	0.99255448
A2276T	F	T426S	CON	0	0	7909	23	0.00290808	16461	0	0	17414	0	0	23419	19	0.00081131
T2321C	INT	INT	INT	150	0.0093932	7449	103	0.01382736	16763	141	0.00841138	17674	137	0.0077515	23369	156	0.00667551
C2971T	H	A14V	CON	16346	0.9981071	7910	7813	0.98773704	15191	321	0.02113093	16770	2509	0.1496124	23115	155	0.0067056
G3071A	H	M47I	CON	334	0.02275049	9846	214	0.02173471	14263	1	7.01E-05	14876	1	6.72E-05	21201	3	0.0001415
G3111A	H	V61I	CON	7	0.00043687	9717	0	0	15755	0	0	16143	1108	0.06863656	23126	1	4.32E-05
C3120T	H	P64S	RAD	6	0.00036998	9496	0	0	15969	7	0.00043835	16177	324	0.02002844	23252	22	0.00094616
G3129T	H	A67S	RAD	15076	0.99755178	9172	9026	0.98408199	15256	84	0.00550603	15335	1945	0.12683404	22005	122	0.00554419
G3132A	H	A68T	RAD	41	0.00266962	9144	19	0.00207787	15097	2	0.00013248	15277	76	0.0049748	21752	1	4.60E-05
G3339A	H	D137N	RAD	17213	0.99982574	7362	7270	0.9875034	17984	17977	0.99961077	18046	18043	0.99983376	24700	24644	0.99773279
A5360C	A/A*(B)	I460I/I288I(K	SYN/SYN(RA	12151	0.99786483	6843	6771	0.9894783	13938	280	0.02008897	13451	1978	0.14705226	17613	104	0.00590473

Site	Protein(s)	Amino acid(s	Radicality	SCSC8_cvrg	SCSC8_altdp	SCSC8_AF	SCS10_cvrg	SCS10_altdp	SCS10_AF	S10_cvrg	S10_altdp	S10_AF	CS10_cvrg	CS10_altdp	CS10_AF	2NDSCSC3_c	2NDSCSC3_a
C199T	K(C)	T50I(L23L)	RAD(SYN)	12802	47	0.0036713	11628	1	8.60E-05	27709	2	7.22E-05	33509	0	0	-2.147E+09	-2.147E+09
A323G	C	D64G	RAD	11121	2	0.00017984	10265	0	0	25104	4	0.00015934	30009	2	6.66E-05	1	0
C530T	D	D47D	SYN	14279	0	0	12449	0	0	30127	0	0	36456	1	2.74E-05	1	1
T572C	D(E)	G61G(V2A)	SYN(CON)	13243	1	7.55E-05	11451	0	0	27746	2	7.21E-05	33679	4	0.00011877	1	0
A590G	D(E)	G67G(D8G)	SYN(RAD)	12666	2766	0.21837991	10925	0	0	26683	3	0.00011243	32093	2	6.23E-05	-2.147E+09	-2.147E+09
C648G	D(E)	H87D(F27L)	RAD(CON)	13236	4562	0.34466606	11365	11325	0.99648042	27866	27196	0.97595636	33124	32835	0.99127521	1	0
G944T	J	V33L	CON	12405	0	0	11072	0	0	27785	1	3.60E-05	32810	1	3.05E-05	-2.147E+09	-2.147E+09
C1137T	F	A46V	CON	13251	620	0.04678892	11768	1	8.50E-05	29746	2	6.72E-05	36476	2	5.48E-05	-2.147E+09	-2.147E+09
C1148A	F	L50I	CON	13598	0	0	12218	0	0	30827	0	0	37745	1	2.65E-05	-2.147E+09	-2.147E+09
T1307C	F	Y103H	RAD	11705	85	0.00726185	12650	8	0.00063241	30592	16	0.00052301	38877	87	0.00223783	-2.147E+09	-2.147E+09
A1317G	F	H106R	CON	10650	1	9.39E-05	11694	0	0	27807	1	3.60E-05	35674	3	8.41E-05	-2.147E+09	-2.147E+09
C1460A	F	Q154K	RAD	14811	5683	0.3837013	14880	684	0.04596774	34805	1010	0.02901882	44461	630	0.01416972	-2.147E+09	-2.147E+09
A1548G	F	Q183R	RAD	14584	2	0.00013714	14361	5310	0.36975141	32820	12014	0.36605728	42060	15550	0.36970994	-2.147E+09	-2.147E+09
A1637G	F	M213V	CON	15507	1	6.45E-05	14991	1	6.67E-05	34931	1	2.86E-05	44117	4	9.07E-05	-2.147E+09	-2.147E+09
G1639T	F	M213I	CON	15361	3	0.0001953	14903	2	0.0001342	34543	560	0.01621168	43825	0	0	-2.147E+09	-2.147E+09
C1727T	F	L243F	CON	15873	769	0.04844705	14798	0	0	35735	4	0.00011194	44683	1	2.24E-05	-2.147E+09	-2.147E+09
T1956G	F	V319G	CON	15331	5226	0.34087796	13617	13578	0.99713593	33261	32445	0.97546676	41335	41003	0.99196807	-2.147E+09	-2.147E+09
T2083A	F	P361P	SYN	14109	0	0	12975	0	0	30512	0	0	37965	0	0	1	0
C2085T	F	A362V	CON	14405	740	0.05137105	13375	3740	0.27962617	31188	8553	0.27424009	38951	7633	0.19596416	1	0
C2093A	F	L365I	CON	13847	3838	0.27717195	12910	0	0	29916	2	6.69E-05	37312	0	0	1	0
G2179T	F	Q393H	RAD	12490	3548	0.28406725	11927	0	0	27275	0	0	33487	1	2.99E-05	1	0
G2275A	F	M425I	CON	12775	8534	0.66802348	12321	48	0.00389579	27586	189	0.0068513	33979	271	0.00797551	-2.147E+09	-2.147E+09
A2276T	F	T426S	CON	12499	0	0	12104	0	0	27013	1	3.70E-05	33329	7	0.00021003	-2.147E+09	-2.147E+09
T2321C	INT	INT	INT	12903	101	0.00782764	12463	124	0.00994945	27946	236	0.00844486	33773	279	0.00826104	-2.147E+09	-2.147E+09
C2971T	H	A14V	CON	12004	4280	0.35654782	11664	11622	0.99639918	28622	27971	0.97725526	34969	34711	0.99262204	-2.147E+09	-2.147E+09
G3071A	H	M47I	CON	10942	0	0	10461	278	0.0265749	25004	607	0.02427612	30965	777	0.02509285	1	0
G3111A	H	V61I	CON	12212	0	0	11599	2	0.00017243	27920	13	0.00046562	33822	1	2.96E-05	1	0
C3120T	H	P64S	RAD	12487	53	0.00424441	11957	2	0.00016727	28317	0	0	34713	3	8.64E-05	1	0
G3129T	H	A67S	RAD	11803	4043	0.34254003	11233	11188	0.99599395	26387	25754	0.97601091	32416	32126	0.9910538	1	0
G3132A	H	A68T	RAD	11797	20	0.00169535	11417	20	0.00175177	26795	58	0.00216458	32854	54	0.00164364	1	0
G3339A	H	D137N	RAD	14302	14300	0.99986016	13552	13550	0.99985242	29889	29409	0.98394058	36232	36212	0.999448	1	1
A5360C	A/A*(B)	I460I/I288I(K	SYN/SYN(RA	10895	3686	0.33832033	9987	9944	0.9956944	23209	22701	0.97811194	27539	27288	0.99088565	-2.147E+09	-2.147E+09

Site	Protein(s)	Amino acid(s	Radicality	2NDSCSC3_A	2NDSCSC2S_	2NDSCSC2S_	2NDSCSC2S_	A_cvrg	A_altdp	A_AF	CS1_cvrg	CS1_altdp	CS1_AF
C199T	K(C)	T50I(L23L)	RAD(SYN)	1	4294	0	0	7288	35	0.00480241	14671	0	0
A323G	C	D64G	RAD	0	5107	0	0	6442	0	0	13245	0	0
C530T	D	D47D	SYN	1	5508	0	0	8082	0	0	16475	0	0
T572C	D(E)	G61G(V2A)	SYN(CON)	0	5375	0	0	7415	0	0	15261	0	0
A590G	D(E)	G67G(D8G)	SYN(RAD)	1	4859	5	0.00102902	7070	3	0.00042433	14557	2	0.00013739
C648G	D(E)	H87D(F27L)	RAD(CON)	0	5400	5223	0.96722222	7550	87	0.01152318	15164	15121	0.99716434
G944T	J	V33L	CON	1	6042	0	0	6873	110	0.01600466	14771	2	0.0001354
C1137T	F	A46V	CON	1	5241	0	0	6906	0	0	15952	0	0
C1148A	F	L50I	CON	1	5051	0	0	7159	0	0	16599	1	6.02E-05
T1307C	F	Y103H	RAD	1	5748	1	0.00017397	7073	96	0.01357274	16926	11	0.00064989
A1317G	F	H106R	CON	1	5439	1	0.00018386	6422	1	0.00015571	15493	1	6.45E-05
C1460A	F	Q154K	RAD	1	5093	168	0.03298645	8031	4613	0.5743992	19335	731	0.03780709
A1548G	F	Q183R	RAD	1	5127	2127	0.41486249	7523	2	0.00026585	18127	7479	0.41258896
A1637G	F	M213V	CON	1	5070	1	0.00019724	8101	2	0.00024688	19084	3	0.0001572
G1639T	F	M213I	CON	1	5111	0	0	8071	0	0	18979	3	0.00015807
C1727T	F	L243F	CON	1	5322	0	0	8444	0	0	19246	0	0
T1956G	F	V319G	CON	1	5751	5591	0.97217875	8983	111	0.01235667	18101	18047	0.99701674
T2083A	F	P361P	SYN	0	5268	16	0.00303721	8400	1	0.00011905	17058	1	5.86E-05
C2085T	F	A362V	CON	0	5520	967	0.17518116	8616	0	0	17534	4320	0.24637846
C2093A	F	L365I	CON	0	5561	0	0	8412	0	0	16835	4	0.0002376
G2179T	F	Q393H	RAD	0	5781	0	0	7484	0	0	15371	0	0
G2275A	F	M425I	CON	1	4959	121	0.02440008	7770	7208	0.92767053	15754	41	0.00260251
A2276T	F	T426S	CON	1	4996	0	0	7622	1	0.0001312	15358	0	0
T2321C	INT	INT	INT	1	4769	10	0.00209688	7829	163	0.02082003	15750	155	0.00984127
C2971T	H	A14V	CON	1	5718	5552	0.97096887	6469	94	0.01453084	15504	15452	0.99664603
G3071A	H	M47I	CON	0	5978	110	0.0184008	6249	0	0	13821	353	0.02554084
G3111A	H	V61I	CON	0	5843	2	0.00034229	6902	0	0	15086	2	0.00013257
C3120T	H	P64S	RAD	0	5653	1	0.0001769	7026	0	0	15486	2	0.00012915
G3129T	H	A67S	RAD	0	5463	5316	0.97309171	6749	1	0.00014817	14485	14450	0.99758371
G3132A	H	A68T	RAD	0	5424	23	0.00424041	6668	1	0.00014997	14681	48	0.00326953
G3339A	H	D137N	RAD	1	5043	5040	0.99940512	8161	7834	0.95993138	17063	17062	0.99994139
A5360C	A/A*(B)	I460I/I288I(K	SYN/SYN(RA	1	4744	4620	0.97386172	6222	83	0.01333976	12678	12639	0.99692381

B.2 Multi-nucletide events – 1301-1304

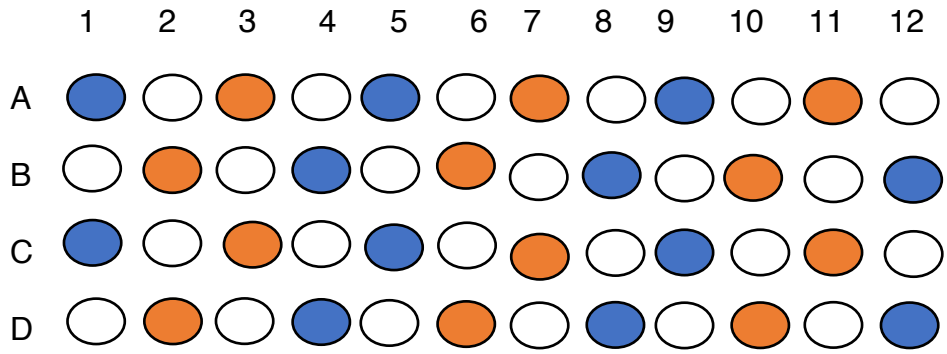
Sample	Coverage	WT ACTG	Double Mutant GCTC	1301 Mutant GCTG	NULL G	1304 Mutant ACTC
2NDS1	8205	96	0	2	0	8103
2NDS2	17516	45	2	0	0	17455
2NDSCSC2A	65911	63837	0	8	0	2025
2NDSCSC2B	73453	70987	0	18	0	2410
2NDSCSC2S	5628	186	0	0	0	5438
2NDSCSC3						
A	6667	6606	0	54	0	0
C1	1929	1914	0	14	0	1
C10	6456	6425	0	28	0	0
CS1	15871	41	0	2	0	15816
CS10	36542	349	2	0	0	36159
CSC1	5712	5710	0	0	0	0
CSC10	15749	15682	1	3	0	60
CSCS1	22546	118	2	0	0	22417
CSCS10	19211	412	0	0	0	18787
S1	26188	25926	0	109	0	140
S10	28746	190	1	483	0	28051
S2	850	7	0	0	0	843
S3	17073	31	1	0	0	17029
S8	20384	120	4	0	0	20248
SC1	6599	6595	0	0	0	0
SC10	5999	5924	0	72	0	0
SCS1	14922	1430	0	0	0	13483
SCS10	11920	32	0	0	0	11876
SCSC1	8801	8784	1	0	0	13
SCSC2	890	690	0	0	0	200
SCSC3	16802	5814	1	1	0	10970
SCSC8	10803	8520	1573	0	0	671
SCSC10A	23102	20734	473	4	0	1728
SCSC10B	17302	15441	360	2	0	1391

WT ACTG	Double Mutan GCTC	1301 Mutant GCTG	NULL G	1304 Mutant ACTC
0.01170018	0	0.00024375	0	0.98756856
0.00256908	0.00011418	0	0	0.99651747
0.96853333	0	0.00012138	0	0.03072325
0.96642751	0	0.00024505	0	0.0328101
0.03304904	0	0	0	0.96624023
#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
0.99085046	0	0.0080996	0	0
0.99222395	0	0.00725765	0	0.0005184
0.99519827	0	0.00433705	0	0
0.00258333	0	0.00012602	0	0.99653456
0.00955065	5.4732E-05	0	0	0.98951891
0.99964986	0	0	0	0
0.99574576	6.3496E-05	0.00019049	0	0.00380977
0.00523374	8.8708E-05	0	0	0.99427836
0.02144605	0	0	0	0.97792931
0.98999542	0	0.00416221	0	0.00534596
0.00660962	3.4787E-05	0.01680234	0	0.97582272
0.00823529	0	0	0	0.99176471
0.00181573	5.8572E-05	0	0	0.99742283
0.00588697	0.00019623	0	0	0.9933281
0.99939385	0	0	0	0
0.98749792	0	0.012002	0	0
0.09583166	0	0	0	0.90356521
0.00268456	0	0	0	0.99630872
0.9980684	0.00011362	0	0	0.0014771
0.7752809	0	0	0	0.2247191
0.34603023	5.9517E-05	5.9517E-05	0	0.65289846
0.78866981	0.1456077	0	0	0.06211238
0.89749805	0.02047442	0.00017315	0	0.07479872
0.89244018	0.02080684	0.00011559	0	0.08039533

B.3 Multi-nucletide events – 1346-1347

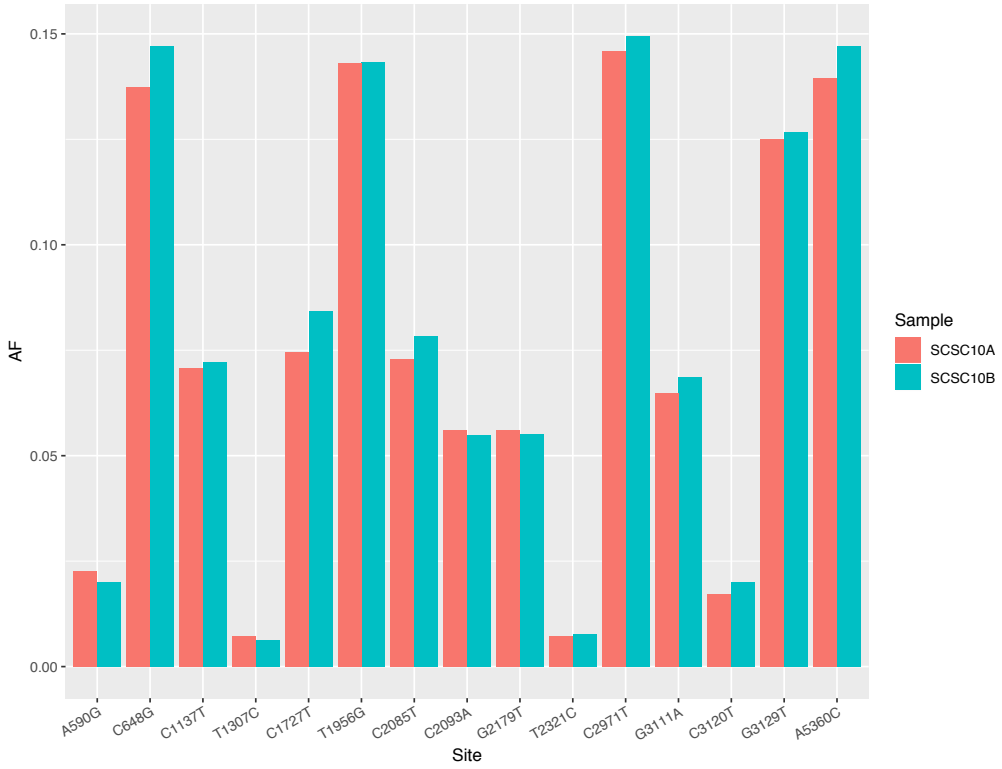
Sample	Coverage	WT GA	Double Mutant AG	1346 mutant A	WT GA	Double Muta AG	1346 mutant A
2NDS1	7672	7669	0	0	0.99960897	0	0
2NDS2	16397	16394	0	0	0.99981704	0	0
2NDSCSC2A	60283	60265	0	0	0.99970141	0	0
2NDSCSC2B	67613	67598	0	0	0.99977815	0	0
2NDSCSC2S	5252	5251	0	0	0.9998096	0	0
2NDSCSC3					#DIV/0!	#DIV/0!	#DIV/0!
A	6703	6700	0	0	0.99955244	0	0
C1	1923	1923	0	0	1	0	0
C10	6502	6502	0	0	1	0	0
CS1	15908	15906	0	0	0.99987428	0	0
CS10	37003	36994	0	0	0.99975678	0	0
CSC1	5728	5728	0	0	1	0	0
CSC10	15528	15528	0	0	1	0	0
CSCS1	23105	23101	0	0	0.99982688	0	0
CSCS10	19323	19321	0	0	0.9998965	0	0
S1	26249	26238	0	0	0.99958094	0	0
S10	28757	28750	0	0	0.99975658	0	0
S2	873	872	0	0	0.99885452	0	0
S3	17248	17240	0	0	0.99953618	0	0
S8	20509	20501	0	0	0.99960993	0	0
SC1	6694	6693	0	0	0.99985061	0	0
SC10	6083	6083	0	0	1	0	0
SCS1	14969	14967	0	0	0.99986639	0	0
SCS10	12112	12109	0	0	0.99975231	0	0
SCSC1	8713	8710	1	0	0.99965569	0.00011477	0
SCSC2	926	925	0	0	0.99892009	0	0
SCSC3	16927	16922	1	0	0.99970461	5.9077E-05	0
SCSC8	11121	9264	1804	0	0.83301861	0.16221563	0
SCSC10A	23446	23140	302	0	0.98694873	0.01288066	0
SCSC10B	17345	17125	214	0	0.98731623	0.01233785	0

Appendix C - 96 well plate layout used during the DNA library preparation



A	C1	C10	SC1	SC10	CSC1
CSC10	SCSC1	SCSC2	SCSC3	SCSC8	SCSC10A
SCSC10B	S1	S2	S3	S8	S10
CS1	CS10	SCS1	SCS10	CSCS1	CSCS10

Blue coloured wells represent DNA + pUC18 spike-in, orange wells represent DNA only, while non-colured wells are blanks left in-between DNA



Allele frequency at high-frequency sites (sites with frequencies less than 15%) compared between within-run replicates from the Nextera XT run (repeated from the DNA preparation stage). The identity of substitutions is given on the x axis and allele frequencies are shown on the y axis (note that the y axis begins at a frequency of 0%).