Measuring natural selection in viral populations: models with host immunity and high mutation rates

Measuring natural selection in viral populations: models with host immunity and high mutation rates

Carmen H. S. Chan

August 2014

A thesis submitted for the degree of Doctor of Philosophy
School of Biotechnology and Biomolecular Sciences
Faculty of Science

PLEASE TYPE

THE UNIVERSITY OF NEW SOUTH WALES
Thesis/Dissertation Sheet

Surname or Family name: Chan

First name: Carmen Other name/s: Hoi Shan

Abbreviation for degree as given in the University calendar: PhD

School: Biotechnology and Biomolecular Sciences Faculty: Science

Title: Measuring natural selection in viral populations: models with host immunity and high mutation rates

Abstract 350 words maximum: (PLEASE TYPE)

Measuring selective pressures shaping the evolution of viral populations is important for preventing and controlling the spread of disease, as well as for understanding evolutionary processes in general. Traditional methods for detecting and quantifying selection assume that a single segregating allele is under constant selection in a population of constant size. However, viruses frequently violate these assumptions due to (i) their high mutation rates and (ii) their complex epidemiological dynamics. We examine the effect of these factors using computational models describing evolution at protein-coding regions, under various population dynamics. In Chapter 2 we show, assuming population sizes are constant, that linkage-induced interference between segregating mutations distort commonly used statistics such as d_N/d_S and the McDonald-Kreitman (MK) statistic. We propose three alternative statistics to detect the effect of background selection, hitch-hiking and clonal interference. In Chapter 3, we examine selection acting in the context of an epidemiological multi-strain SIRS model, by explicitly modelling the effect of cross-immunity between related strains, competition for susceptible hosts, and decaying host immunity. By studying the probability of antigenic reversion, we show that time-varying antigenic selection mediated by host immunity has a qualitatively different effect from constant selective constraint, which is observable from changing frequencies at antigenic sites over time. In Chapter 4, we apply both of these methods to avian influenza, demonstrating their utility in comparing selection between different lineages. In combination, these methods allow us to distinguish between different forms of selection, which may allow us to discriminate between potential biological mechanisms shaping viral populations.

Declaration relating to disposition of project thesis/dissertation

……………………………………………………………… ……………………………………..……………… ………..……………………..……..…
Signature Witness Date

The University recognises that there may be exceptional circumstances requiring restrictions on copying or conditions on use. Requests for restriction for a period of up to 2 years must be made in writing. Requests for a longer period of restriction may be considered in exceptional circumstances and require the approval of the Dean of Graduate Research.

FOR OFFICE USE ONLY Date of completion of requirements for Award:

THIS SHEET IS TO BE GLUED TO THE INSIDE FRONT COVER OF THE THESIS

Abstract

Measuring selective pressures shaping the evolution of viral populations is important for preventing and controlling the spread of disease, as well as for understanding evolutionary processes in general. Traditional methods for detecting and quantifying selection assume that a single segregating allele is under constant selection in a population of constant size. However, viruses frequently violate these assumptions due to (i) their high mutation rates and (ii) their complex epidemiological dynamics. We examine the effect of these factors using computational models describing evolution at protein-coding regions, under various population dynamics. In Chapter 2 we show, assuming population sizes are constant, that linkage-induced interference between segregating mutations distort commonly used statistics such as d_N/d_S and the McDonald-Kreitman (MK) statistic. We propose three alternative statistics to detect the effect of background selection, hitch-hiking and clonal interference. In Chapter 3, we examine selection acting in the context of an epidemiological multi-strain SIRS model, by explicitly modelling the effect of cross-immunity between related strains, competition for susceptible hosts, and decaying host immunity. By studying the probability of antigenic reversion, we show that time-varying antigenic selection mediated by host immunity has a qualitatively different effect from constant selective constraint, which is observable from changing frequencies at antigenic sites over time. In Chapter 4, we apply both of these methods to avian influenza, demonstrating their utility in comparing selection between different lineages. In combination, these methods allow us to distinguish between different forms of selection, which may allow us to discriminate between potential biological mechanisms shaping viral populations.

Contents

Preface

This thesis consists of three studies on the evolution of viral populations.

In Chapter 2, I construct a computer simulation model to analyse the effect of interfering mutations on the evolution of codon sequences. This work was done in collaboration with my supervisor, Mark Tanaka, and Steven Hamblin. I played a major role in the study conception and design of this chapter, which were developed through many discussions with Mark and Steven. I implemented the model, performed the analysis, and wrote the manuscript, with contributions from Mark and Steven in organising and editing the text. This chapter has been published in *BMC Evolutionary Biology* (Chan et al., 2013).

In Chapter 3, I develop a model to explain variation in patterns of reversion at antigenic sites. This work was done in collaboration with Mark Tanaka and Lloyd Sanders. The ideas in this study were developed jointly by Mark and me. The study was designed primarily by me, in discussion with Mark. I implemented the model, with assistance from Lloyd, and carried out all of the analysis. I also wrote the manuscript with contributions from Lloyd and Mark in editing the text. This chapter has been submitted to a journal for publication.

In Chapter 4, I apply the methods developed in the previous chapters, to study the evolutionary dynamics of avian influenza A. The ideas in this study were developed in discussion with Mark Tanaka. I carried out all of the analyses and wrote the manuscript. The work in this chapter has not yet been submitted for publication.

Chapter 1

Introduction

RNA viral populations provide a powerful model for studying molecular evolution. The high mutation rates, in combination with large population sizes and short generation times, result in rapid evolutionary changes on observable time-scales. In this thesis, we study the influenza A virus, as an example of viral evolution. The segmented genome of this virus consists of eight gene segments encoding eleven proteins; of these, the most extensively studied is the hemagglutinin (HA) gene, which is the primary target of antigenic recognition. Reassortment between segments can occur, but recombination is very rare (Boni et al., 2008), so that for any given segment, genetic diversity is accumulated clonally.

The increasing availability of sequence data provides the opportunity to study the effect of molecular changes on viral fitness in natural populations. However, many of the most widely used methods are based on simple population models which assume that only a single mutation segregates at any time in the population, and do not account for changes in population dynamics. How do these assumptions affect inference of selection in viral populations?

In this chapter, we provide a brief description of the theory and application for inferring selection in viral populations. In Section 1.1, we describe some theoretical models of population dynamics that underlie these estimation methods. The basic Wright-Fisher model is described in Section 1.1.1; we then describe extensions that examine the effect of multiple mutations segregating at the same time in the population (Section 1.1.2) and changing population dynamics (Section 1.1.3). In Section 1.2, we describe methods of inference, with particular focus on applications to human influenza A. This virus is one of the most extensively studied examples of positive selection, and a comparison of the methods which have been applied illustrates how the development of population models can provide further insight into viral evolution. Finally, in Section 1.3, we will summarise some of the unresolved questions regarding inference of viral selection and outline the goals of this thesis.

1.1 Models of evolving populations

1.1.1 The Wright-Fisher model

One of the most basic models in population genetics is the Wright-Fisher model (Wright, 1931). It assumes a population of constant size N, which evolves in discrete generations so that in each generation the population is replaced by its progeny. A mutation with a selection coefficient s changes the reproductive fitness of an individual by a factor of $1 + s$. Mutations that are beneficial ($s > 0$) are more likely to reach fixation, so that it is acquired by all individuals in the population due to positive selection; whereas, deleterious mutations ($s < 0$) are more likely to be removed from the population by negative selection. However, stochasticity also affects the fate of the mutation. Genetic drift occurs as a result of random sampling of a finite number of individuals in each generation, so that mutations (even beneficial ones) at low frequencies may be stochastically lost. The relative influence of genetic drift is determined by the population size N, and we denote the effective strength of selection as $S = Ns$.

In each generation, $\theta = UN$ new mutations are introduced into the population, where U is genomic mutation rate, and some of the existing genetic diversity is lost during the reproductive process. Assuming a population of constant size at equilibrium, the probability that a new mutation with a selective coefficent s survives genetic drift to reach fixation is given by (Kimura, 1962)

$$\begin{aligned} p_{\text{fix}} &= \frac{1 - e^{-2s}}{1 - e^{-2Ns}} \, , \\ &\approx \frac{2s}{1 - e^{-2Ns}} \, . \end{aligned} \tag{1.1}$$

In particular, the probability that a neutral ($s = 0$) mutation reaches fixation is $1/N$, while the fixation probability of a beneficial mutation ($s > 0$) in a large (haploid) population approaches $2s$. The second value approaches the result obtained by Haldane (1927), using a branching process approximation (s small but not too close to 0).

The rate of substitution, which is commonly used to describe long-term evolutionary dynamics, is simply given by the product of the rate at which mutations are introduced into the population θ and the probability they are retained in the population p_{fix}.

1.1.2 Population models with interference

Under the conditions described in Section 1.1.1, the behaviour of a mutation at low frequency is mainly influenced by genetic drift, but tends to follow more deterministic trajectories once the mutation reaches high frequencies. However, in a population where the rate of mutational input θ is large, multiple mutations may occur and perturb the trajectory of the mutation, even when it reaches high frequencies. This is particularly relevant for understanding evolution in viral populations which have both rapid mutation rates and large population sizes.

Historically, this issue has been addressed from a number of different perspectives, resulting in wide variation in terminology, which can differ depending on the number and composition of new mutations and the effect of interest. For example, background selection (Charlesworth et al., 1993) refers to the effect of deleterious mutations on neutral diversity, hitch-hiking (Maynard-Smith et al., 1974) refers to the effect of beneficial mutations on deleterious and neutral mutations, and clonal interference (Gerrish and Lenski, 1998) refers to the effect of multiple beneficial mutations on different sequences. In all cases, the general effect of interference is to reduce the effect of selection (both positive and negative) and to increase the variance of the offspring distribution. We describe these effects in more detail in Chapter 2; the aim of this section is only to provide some intuition regarding the distorting effects of interference on the dynamics of the population.

Gerrish and Lenski (1998) derives an expression for the probability of fixation which combines both the probability of surviving extinction by drift, and the probability that no mutation with a higher selection coefficient emerges. Mutations can therefore be categorized as strongly beneficial or driver mutations that influence the trajectory of other mutations, or passenger mutations which are only weakly influenced by the selection coefficient of the mutation (Schiffels et al., 2011).

Interference does not only perturb the trajectory of a single mutation; it distorts the frequency of all mutations carried on the same lineage. For strong selection, this effect is transitionary. Strongly deleterious mutations are usually removed before they reach high frequencies (Charlesworth et al., 1993), and strongly beneficial mutations sweep rapidly to fixation (Maynard-Smith et al., 1974). However, when mutation rates are high and selection is weak, multiple lineages can persist. In this selection regime, the lineage in which a mutation occurs is more important than its selection coefficient. Mutations that occur in the most fit lineages are more likely to occur at high frequencies in the population, even if they are neutral or weakly deleterious. This is expected to occur in both populations with many beneficial mutations (Desai et al., 2013; Neher and Hallatschek, 2013), and populations with many deleterious mutations (Walczak et al., 2012).

1.1.3 Population models with varying population size

The population model described in Section 1.1.1 assumes a constant coefficient of selection s, and a population of constant size N, which is at equilibrium. Alternatively, it is possible to derive fixation probabilities and passage times for a time-inhomogeneous process using branching processes to incorporate the effect of changing population sizes (Patwa and Wahl, 2008) and selection coefficients (Uecker and Hermisson, 2011) over time.

Fixation probabilities computed using the branching process differ from the results of the Wright-Fisher model. This is due to the fact that the branching process has no density dependence, unless it is explicitly incorporated into the birth or death rate (Uecker and Hermisson, 2011). In the case of positive selection in a large population, the probability of fixation derived from the branching processes (Haldane, 1927) is similar to the probability in the Wright-Fisher model, but changes in population size can affect the fixation probability of selected mutations. One simple case is where the population is growing exponentially at rate r; this increases the probability of fixation of a beneficial mutation to $2(s + r)$ (Otto and Whitlock, 1997). However, where rates of birth and death vary over time, the expression for the probability of fixation is also time-dependent (Uecker and Hermisson, 2011). Where population size changes rapidly, it is common to approximate the effective population size as the harmonic mean of the population size over time (Ewens, 1967), but this approximation performs well only if the fluctuations are much more rapid than the time to fixation (on the order of $1/s$) (Otto and Whitlock, 1997).

The results above suggest that the probability of fixation in an epidemiological model should be time-dependent. However, if we assume that stochastic extinction primarily occurs when a mutation is at low frequency, we can approximate the fixation probability of a new mutation using the initial rates of pathogen transmission, by assuming that the number of susceptible hosts are not depleted. Under these simplifying conditions, the probability of fixation can be expressed in terms of the basic reproductive ratio of the virus (Keeling and Rohani, 2008). This probability is developed and described in more detail in Chapter 3.

1.2 Inference of natural selection in influenza A

Having outlined some of the relevant results which guide our expectations about how selection can influence samples of sequences, we now review how some of these methods have been applied to human influenza A, and show how the use of different methods have provided insights into different aspects of the evolutionary dynamics of this virus.

1.2.1 Inferring selection from substitutions

A common approach for detecting selection involves comparing the rate of non-synonymous substitution d_N to the synonymous rate of substitution d_S. Synonymous nucleotide changes do not affect the encoded amino acid, and are therefore assumed to be neutral. Based on this comparison, $d_N/d_S > 1$ indicates the effect of positive selection favouring a change in the amino acid of a protein, while $d_N/d_S < 1$ is a sign of negative selection due to functional constraints (Hughes and Nei, 1988).

When values of d_N/d_S are computed across a whole gene, positive selection at a small number of sites may be masked by negative selection at a larger of number of sites. Thus, an analysis of human influenza H1N1 (Sugita et al., 1991) found that the evolution of the HA gene is mainly characterised by selective constraint. However, Fitch et al. (1991) showed that the number of non-synonymous changes varies considerably between codon sites of the HA gene of human influenza A (H3N2). Antigenic changes (experimentally determined) were also observed to occur more frequently along of the trunk of the phylogeny, indicating enhanced survival of the strain.

The low sensitivity of gene-wide tests for selection has motivated the development of methods for identifying positive selection at individual codon sites. Fitch et al. (1997) tested for sites where d_N/d_S exceeds the binomial expectation, computed by pooling changes at all codon sites. Maximum likelihood methods (Kosakovsky Pond and Frost, 2005; Nielsen and Yang, 1998) assign sites to different classes and estimate d_N/d_S for each site-class according to their *posterior* probability. Results from d_N/d_S-based studies have varied in terms of the number of sites identified [31 (Fitch et al., 1997), 18 (Bush et al., 1999) and 12 (Yang, 2000)], but a small number of sites in the HA gene show strong signs of positive selection and has been identified in multiple studies.

More refined substitution models have been developed in the maximum-likelihood framework (see Kosakovsky Pond and Frost (2005)). These allow for codon-specific bias in the rates of mutation and base frequencies, but a single parameter $\omega = d_N/d_S$ describes the relative rate of all non-synonymous substitutions. However, the HA gene shows a strong directional effect, with a preference for a limited number of amino acids at any site. This directionality effect can be modelled by including codon-specific bias factors into the substitution model (Kosakovsky Pond et al., 2008; Seoighe et al., 2007), or by computing an additional bias statistic (Kryazhimskiy et al., 2008). These methods detected directional selection in both the HA gene (Kryazhimskiy et al., 2008) and the internal segments (Kosakovsky Pond et al., 2008). Many of the sites identified by Kryazhimskiy et al. (2008) overlap with previously identified antigenic sites (Bush et al., 1999; Fitch et al., 1997; Yang, 2000), which suggests that antigenic sites in influenza evolve under both selective constraint and positive selection.

Phylogentic methods have also been developed to characterise the distribution of fitness effects (DFE) across the whole gene, including both positively and negatively selected sites. Nielsen and Yang (2003) showed that if evolution is assumed to follow a Wright-Fisher process, than we can describe the estimator $\omega = d_N/d_S$, as a ratio of the fixation probabilities [Equation (1.1)]. The estimator ω can then be expressed as function of the effective strength of selection $S = Ns$. Assuming the DFE followed a gamma distribution, Nielsen and Yang (2003) estimated shape and rate parameters to be respectively, $\hat{\alpha} = 0.306$ and $\hat{\beta} = 0.298$. This was quite similar to estimates for the HIV-1 *env* gene ($\hat{\alpha} = 0.373$, $\hat{\beta} = 0.523$) and very different from a similar analysis on mammalian mitochondrial data ($\hat{\alpha} = 3.22$, $\hat{\beta} = 1.62$) (Nielsen and Yang, 2003). In particular, the smaller shape parameter for the viral proteins indicates that only a small proportion of sites are weakly deleterious.

Alternatively, the DFE can be estimated by relating the relative frequency of an amino acid to the probability of fixation [Equation (1.1)]. This result follows from the reversibility of the evolution process (Halpern and Bruno, 1998). Tamuri et al. (2009) show that residue frequency is more informative of selective constraint than d_N/d_S. This parametrisation more clearly reflects the codon-specific nature of selective constraint; negative selection drives the population towards a static equilibrium, whereas positive selection is a shift in equilibrium (Mustonen and Lässig, 2009). This contrasts with d_N/d_S, which assumes that positive selection occurs recurrently so that an unfixed amino acid is always favoured. Comparison of selective constraint between avian and human branches shows that the strength of selective constraint is similar for both host types but different residues are preferred (Tamuri et al., 2012). They found that deleterious changes are distributed bi-modally with peaks at strongly deleterious ($S < -10$) and nearly neutral ($-2 < S < 2$) mutations. At the adaptive equilibrium, there is a third well-defined peak for strongly beneficial ($S > 10$) mutations.

The assumption of a specific model enables a quantitive estimation of fitness. In the absence of a model, however, it is still possible to determine the relative effect of specific factors. Meyer et al. (2013) compared estimates of d_N/d_S at each site between lineages of avian and human HA of subtype H1, H3 and H5. They found that similarity in structural constraints between lineages explain 20–40% of variation in d_N/d_S values.

1.2.2 Inferring selection from sequence polymorphism

Phylogenetic d_N/d_S-based methods were originally developed for analysing divergence between different species. Their application to viral sequences sampled from a single population relies on the simplifying but incorrect assumption that polymorphic mutations are equivalent to substitutions. As demonstrated by Kryazhimskiy and Plotkin (2008), this results in a over-estimation of d_N/d_S, and generates biased estimates of the effective strength of selection S.

One modification is to use the McDonald-Kreitman (MK) test (McDonald and Kreitman, 1991), which compares the ratio of non-synonymous to synonymous substitutions to the ratio of non-synonymous to synonymous polymorphisms. This comparison assumes that deleterious mutations are rapidly removed and surviving polymorphisms are neutral (Smith and Eyre-Walker, 2002). An additional modification, using only polymorphisms segregating at medium frequencies (Bhatt et al., 2011), was proposed to reduce bias due to non-neutral mutations segregating in the population. A study by Messer and Petrov (2013) suggests that the comparative approach of the MK test is relatively robust to the effects of linkage. Bhatt et al. (2011), using a modified form of the MK test, found that positive selection occurs on all segments of human influenza A (H3N2); this contrasts to a d_N/d_S study (Suzuki, 2006), which detected positive selection in only the HA, NA and NP gene segments.

Alternatively, there are sequence analysis methods which do not explicitly differentiate between frequency and fixation. These methods are less established, and the theoretical framework is less developed, but they can be highly descriptive about the evolutionary dynamics of the virus. While phylogenetic methods have difficulty incorporating time-dependent or branch-specific effects (Rodrigue, 2013), these effects can be incorporated quite easily using frequency-based methods. One of the simplest approaches to frequency data is developed in a study by Shih et al. (2007). No model is explicitly stated, but rapid increases in frequency are taken as an indication of strong selection. These events were observed to occur mainly at antigenic sites, and to occur continuously thoughout the period of study. A more model-based interpretation was used by Illingworth and Mustonen (2012); frequency trajectories of each allele were fitted to a model of logistic growth where the selection coefficient is given by the growth rate (Wright, 1931).

Consistent with the observation that almost no recombination occurs in influenza A, a number of frequency-based studies have shown that interference between mutations has a strong effect on the population dynamics of human influenza A. Illingworth and Mustonen (2012) found a large discrepancy between an "effective" selection coefficient and an inherent selection coefficient, indicating strong interference effects at many sites. A related study (Strelkowa and Lässig, 2012) found that many positively selected sites reach high frequencies but are subsequently out-competed, which is an indication of clonal interference.

1.2.3 Inference accounting for population dynamics

The methods described in Section 1.2.1–1.2.2 largely assumed fixed population sizes, but as discussed in Section 1.1.3, changing population sizes can affect evolutionary dynamics. Coalescent theory describes the probability of observing a particular genealogy (Kingman, 1982). The likelihood constructed from this process can be used to estimate the effective population size (Felsenstein, 1992; Kuhner et al., 1995). Furthermore, the coalescent has the property that inter-coalescent intervals are independent (Griffiths and Tavare, 1994), so that it is possible to estimate different population sizes at each coalescent time. (Pybus et al., 2000).

Using extensions that allow the effective population size to vary nonparametrically (Minin et al., 2008; Pybus et al., 2000), Bahl et al. (2011) reconstructs the population dynamics of human influneza A (H3N2). They found that seasonal fluctuations in the effective population size occur at temperate locations, while more stable dynamics are observed in tropical and subtropical regions. Globally, the effective population size is small, but lineages continuously die out and are replaced by reseeding from different locations. No single population acts as the source, but genetic diversity is maintained by constant migration between populations. Interestingly, Southeast Asia, which is one of the major hubs maintaining diversity between seasons (Russell et al., 2008) showed smaller genetic diversity than other regions. As noted by the author, this may represent the effect of selection rather than a smaller census population size (Bahl et al., 2011).

This demonstrates two difficulties to applying coalescent methods to viral sequence data. First, it does not include selection; second, it is unclear how the effective population size should be biologically interpreted. In regards to the first problem, we note that theoretical extensions of the coalescent (Desai et al., 2013; Kaplan et al., 1988; Walczak et al., 2012) have been developed to describe changes in the rate of coalescence between lineages of different fitness classes. However, these models have not been implemented for inference, as they require the fitness of each node in the genealogy to be known.

In response to the second problem, a range of phylodynamic models (Grenfell et al., 2004; Pybus and Rambaut, 2009; Volz et al., 2009), have been developed to link the coalescent to epidemiological models. This involves reparametrization of the rate of coalescence in terms of the rate of transmission (Frost and Volz, 2010; Koelle and Rasmussen, 2012), as well as computational methods to incorporate epidemiological data (Rasmussen et al., 2011). However, these methods cannot incorporate sudden changes in population dynamics due to molecular (antigenic) changes.

In view of these limitations of the coalescent, an alternative approach is to explicitly model both epidemiological and molecular changes. These models can incorporate mechanisms specific to the virus of interest, but the resulting complexity can make it difficult to evaluate the influence of different effects. For example, multiple models have been proposed to explain the mechanism shaping the linear, "cactus-like" phylogeny of human influenza A, where genetic diversity at any time is limited despite continual generation of new strains. Ferguson et al. (2003) showed that observations could be explained by the introducing a short-lived strain transcending immunity to restrict the growth of viral diversity in the population. However, an alternative model, in which antigenic substitutions occur in only punctuated bursts, was also able to generate the a cactus-like phlyogeny without invoking additional immunity (Koelle et al., 2006). A recent model incorporating both strain-transcending immunity and mutation-limited antigenic drift (Zinder et al., 2013) found that phylogenetic patterns were unable to distinguish between these two hypotheses.

1.3 Goals of this thesis

In this chapter, we have outlined the theoretical underpinnings and analytical methods developed to infer selection from viruses. Standard models of molecular evolution used in phylogenetic and coalescent models assume the process of substitution is time homogeneous and instantaneous along branches. Changes in rates of substitutions can be described in terms of different parameters such as the mutation rate, frequency bias and mutation bias, to distinguish between mechanisms affecting evolutionary rates. Modifications in the substitution model also allow parameters to be varied between branches, but these models assume there is no change in population dynamics along a single branch, and that there is no interaction between concurrent lineages. These assumptions affect how we identify populations evolving under selection, how we quantify the amount of selection, and how we generalise from the population under study to similar viral populations that may differ in population dynamics.

In this thesis, we address two interaction effects which complicate the process by which mutations emerge and reach fixation in viral populations. In Chapter 2, we examine the effect of interference between co-segregating mutations using computer simulations. Previous papers have examined different effects of particular forms of interference such as background selection (Charlesworth et al., 1993; Walczak et al., 2012), hitch-hiking (Bustamante et al., 2001) and clonal interference (Desai et al., 2013; Strelkowa and Lässig, 2012). Here, we examine their combined effects and compare how different conditions of linked selection affect sequence statistics, such as d_N/d_S and the MK statistic.

In Chapter 3, we examine the effect of interactions between the molecular substitution process and the epidemiological dynamics by studying antigenic reversion. We first develop an analytical model for a simple three-strain SIRS model (Hethcote, 2000) to provide insight into the relative influence of the various underlying parameters describing the viral and host population. We then extend the analysis using computer simulations and consider whether frequency data can be used to allow inference of parameters of antigenic selection and selective constraint.

We have focused on the application of these methods to human influenza A, which has been the subject of extensive study. The interest in this virus is not only because it imposes a significant health burden, but also because of its distinctive evolutionary dynamics with clear signs of strong positive selection (Fitch et al., 1991). As such, human influenza is a canonical example of positive selection, but this raises the question of whether these selective mechanisms are particular to human influenza A, or occur generally in other viruses. In Chapter 4, we examine the role of interference and epidemiology in the evolution of avian influenza A. As avian influenza does not have same ladder-like shape of human influenza, multiple lineages co-circulate, allowing us to compare the effect of different population dynamics on evolutionary trajectories across a similar fitness landscape.

In Chapter 5, we summarise the conclusions of this thesis, and discuss some general implications for modelling viral evolution and possible directions for future work.

Chapter 2

The effects of linkage on comparative estimators of selection

2.1 Introduction

Understanding the mechanisms by which natural selection shapes the evolution of genes is one of the major aims of molecular evolution. One commonly used approach for the detection of positive selection in protein-coding sequences is based on comparing the frequency of non-synonymous or amino-acid (A) changes to the frequency of synonymous (S) changes (Nielsen, 2005). For simplicity, synonymous nucleotide changes that do not affect the protein are generally assumed to be neutral. In the absence of selection and accounting for the genetic code, we expect both types of changes to be equally probable so that the rate of non-synonymous substitutions per site (K_A) is equal to the rate of synonymous substitutions per site (K_S); a ratio of $K_A/K_S > 1$ indicates positive selection favouring a change in the protein (Hughes and Nei, 1988). However, this test is heavily conservative as proteins are generally under negative selection against amino acid changes that may affect protein function. Positive selection at a small number of sites may be masked by negative selection removing non-synonymous changes in the rest of the protein (Hughes, 2007). In this study, we use the K_A/K_S notation of Li (1993) rather than the more commonly used d_N/d_S to make explicit that we compute the ratio of the number of substitutions, not the ratio of substitution rates.

The McDonald-Kreitman (MK) test (McDonald and Kreitman, 1991) attempts to account for the presence of negatively selected sites by comparing K_A/K_S to f, the proportion of nearly neutral sites in the sequence (Eyre-Walker and Keightley, 2007). If selection is strong, deleterious and beneficial mutations are expected to make little contribution to polymorphism; deleterious mutations are removed by selection and beneficial mutations reach fixation rapidly. Polymorphic sites are expected to consist largely of neutral variation, and the ratio of the number of neutral non-synonymous polymorphic sites (P_A) to the number of synonymous polymorphic sites (P_S) can be used as an estimator of f (Smith and Eyre-Walker, 2002). In the MK test, positive selection is inferred when $K_A/K_S > P_A/P_S$. Following similar reasoning, K_A/K_S measured in a related sample can be used as a measure of selective constraint so that an increase in the K_A/K_S ratio implies positive selection (Czelusniak et al., 1982; Toll-Riera et al., 2011).

With the increasing availability of sequence data, various modifications of K_A/K_S methods have been developed to quantify the prevalence (Smith and Eyre-Walker, 2002), strength (Nielsen and Yang, 2003; Sawyer and Hartl, 1992) and dynamics of positive selection (Lemey et al., 2007; Yang, 1998). These methods rely on the assumption that sites segregate independently; that is, the change in frequency at one site will not affect the change in frequency at another site. In a large population with a high mutation rate, however, multiple mutations co-occur in the population and the change in frequency of one mutation also depends on selection acting at linked sites. Depending on the type of selection, linkage can have different effects; background selection, hitch-hiking and clonal interference can both increase or decrease fixation probability or polymorphism frequency relative to expected levels, which we describe below.

Background selection is the reduction in genetic variability caused by linkage to negatively selected sites (Charlesworth et al., 1993). The effect of background selection on the probability of fixation is qualitatively similar to a reduction in effective population size (Birky and Walsh, 1988; Charlesworth, 1994; Charlesworth et al., 1993; Peck, 1994), which implies a higher than expected value of K_A/K_S under negative selection and a lower than expected value of K_A/K_S under positive selection relative to expectations under independently segregating sites (Birky and Walsh, 1988). Background selection also reduces the number of neutral polymorphic sites (Charlesworth et al., 1995), and can result in a non-monotonic site-frequency spectrum, similar to the effect of continual adaptation (Neher and Hallatschek, 2013; Neher and Shraiman, 2011). Linkage between sites introduces dependencies in the site frequency spectrum, increasing the covariance even if the mean is unchanged (Bustamante et al., 2001). Recent work with the structured coalescent (Walczak et al., 2012) with a model of only negative selection, provides analytical expressions for the number of both neutral and deleterious mutations showing that the effective population size varies, both going back in time, and between individuals in different fitness classes.

When both positive and negative selection operate on a locus, the dynamics of linked neutral and deleterious mutations will also be affected by hitch-hiking (Maynard-Smith et al., 1974). Birky and Walsh (1988) showed that hitch-hiking does not affect the fixation probability at neutral sites but increases the fixation probability at negatively selected sites, which implies that K_A/K_S values are elevated relative to expectation under independently segregating sites. For the MK statistic, the effect of hitch-hiking depends on its effect on polymorphism relative to its effect on divergence. The effect of hitch-hiking on neutral polymorphism has been described by Braverman et al. (1995), but has not been characterised on a selected background. Previous findings (Braverman et al., 1995; Kim and Stephan, 2000; Kim and Wiehe, 2009) were largely based on coalescent simulations which allow only a small number of sites to be under selection and model the trajectory of beneficial mutations deterministically. Forward simulation studies (Birky and Walsh, 1988; Comeron and Kreitman, 2002; Li, 1987; Zeng and Charlesworth, 2010) which begin with a number of positively selected sites and evolve towards mutation-selection equilibrium show that linkage affects a number of frequency-based statistics including Tajima's D and heterozygosity.

Clonal interference (interactions between positively selected mutations) has also been predicted to reduce the fixation probability of beneficial mutations and promote the fixation of deleterious mutations; this was demonstrated in several experimental systems (Miralles et al., 1999; Rozen et al., 2002). More recently, theoretical models assuming continual adaptation with a high supply of beneficial mutations have been used to obtain analytical expressions characterising genetic diversity. These models predict a non-monotonic site frequency spectrum with a large number of both low and high-frequency mutations (Desai et al., 2013; Neher and Hallatschek, 2013; Neher and Shraiman, 2011). This is equivalent to large number of lineages coalescing simultaneously, and is often described as multiple-mergers (Desai et al., 2013; Neher and Hallatschek, 2013; Neher and Shraiman, 2011).

Here, we examine the joint effects of background selection, hitch-hiking and clonal interference on the K_A/K_S and MK statistic. Based on theoretical studies (Desai et al., 2013; Neher and Hallatschek, 2013; Neher and Shraiman, 2011; Walczak et al., 2012), we expect different forms of distortion in the site-frequency spectrum due to these effects. Previous simulation studies (Birky and Walsh, 1988; Comeron and Kreitman, 2002; Li, 1987; Messer and Petrov, 2013) have often considered these effects together, but here we distinguish between them by allowing both the strength of selection and the level of interference to vary. We do this using forward simulations with finite sites, allowing positive selection to occur at different times. Finally, we propose three diagnostic statistics to indicate the degree to which (a) hitch-hiking of deleterious mutations (b) background selection and (c) clonal interference affect a sample of protein-coding sequences.

2.2 Methods

2.2.1 Simulation of sequence evolution under linkage

We simulate the evolution of a population, represented as a sequence of length $L = 500$ codons (nucleotide triplet). Each codon site is associated with a selection coefficient, s_d, which is drawn from the distribution of fitness effects (DFE; see Section 2.2.1.1). The DFE affects both the extent of background selection and hitch-hiking. To model a well-adapted population, each simulation is initialised so that all non-synonymous changes from the ancestral sequence are negatively selected, reducing fitness by a factor of $1 - s_d$. All synonymous changes are neutral. Throughout the simulation, positive selection is introduced at a specified number of sites at fixed times. After the introduction of positive selection, an individual carrying a non-synonymous change from the ancestral sequence at the positively selected site undergoes a change of fitness by a factor of $1 + s_b$. The timing of the introduction of positive selection and the strength of selection (see Section 2.2.1.2) control the extent of clonal interference. The extent of hitch-hiking is determined by the interaction between the DFE and positive selection.

Each simulation is initialised with a haploid population of $N = 10^4$ monomorphic individuals. The mutation process follows a Kimura (1980) two-parameter model, with the transition-transversion ratio fixed at $\kappa = 3$. Ancestral sequences are generated randomly assuming that the base frequency of all 61 non-stop codons are equal, and all 27 one-step mutations at a codon are allowed. For $\kappa > 1$, the mutation probabilities are not equal. Individuals carrying stop-codons have fitness set to zero.

In each generation, the total number of mutations introduced into the population follows a Poisson distribution with mean uNL, where the mutation rate per site per generation is $u = 10^{-6}$ or $u = 10^{-5}$ and occurs uniformly across all sites and all sequences. We assume non-overlapping generations, and individuals reproduce by multinomial sampling with probability proportional to their fitness, as in a Wright-Fisher process.

2.2.1.1 Distribution of deleterious effects

The selection coefficient at each site is drawn from a continuous distribution of fitness effects (DFE), which we model using the gamma distribution following previous studies (Charlesworth and Eyre-Walker, 2008; Nielsen and Yang, 2003; Piganeau and Eyre-Walker, 2003),

$$\rho(x, \beta, \bar{s}) = \frac{(\beta/\bar{s})^{\beta} e^{-(\beta/\bar{s})x} x^{\beta-1}}{\Gamma(\beta)}, \tag{2.1}$$

where β is the shape parameter and $\bar{s}$ is the mean selective coefficient. We consider shape parameters of $\beta = 0.25, 0.5, 1, 2$, which is similar to the range used by Charlesworth and Eyre-Walker (Charlesworth and Eyre-Walker, 2008). Estimated values in the literature range from 0.23 (Eyre-Walker et al., 2006) to 3.22 (Nielsen and Yang, 2003). The mean strength of selection was set at $\bar{s} = 4.4 \times 10^{-1}, 8.5 \times 10^{-3}, 1.5 \times 10^{-3}, 7.0 \times 10^{-4}$, each of which in combination with the respective β value above gives $\omega_0 \approx 0.1$ in the presence of linkage for $u = 10^{-6}$.

The shape parameter β controls the proportion of weakly deleterious mutations, and therefore the extent of hitch-hiking, and in combination with u, the amount of background selection. For small values of β, the distribution of selection coefficients is broadly distributed with a larger proportion of both nearly neutral and strongly deleterious mutations; large values of β give a more strongly peaked DFE centred at nearly neutral to weakly deleterious values. Background selection is primarily mediated by the deleterious mutations that are sufficiently weakly selected that they are able to persist to appreciable frequencies but accumulate to increase the extinction probability of linked neutral and beneficial mutations. This range of selective coefficients is given approximately by $0.5 < U_d/s_d < 5$ (Seger et al., 2010), where U_d is the genomic mutation rate at selected sites. Equating U_d with the genomic mutation rate gives a range of $6.7 \times 10^{-5} < s_d < 6.7 \times 10^{-4}$ for $u = 10^{-6}$, but U_d is generally smaller than U for finite values of β. For $\beta = 0.25$, less than 5% of sites lie within this range so strong negative selection dominates and most deleterious mutations are rapidly removed from the population. For $u = 10^{-5}$, all mutations with $6.7 \times 10^{-4} < s_d < 6.7 \times 10^{-3}$ contribute to background selection, which covers the range around $1/N$, so that much high levels of background can be observed. Similarly, the extent of hitch-hiking is controlled by the proportion of sites with weak deleterious effects relative to the strength of positive selection, with the specific range varying according to the strength and prevalence of positive selection.

2.2.1.2 Positive selection

To examine the effect of linked positive selection, we introduce positive selection at a small number of codon sites in the sequence. Unlike negatively selected sites that individually have small effect but cumulatively can have strong effect due to the large number of negatively selected sites, positive selection is expected to be rare, but a single site can have a strong effect. Thus we model all positively selected sites to have the same fixed selective effect s_b.

At regular time intervals, we randomly choose a site and change the selective coefficient to s_b to generate recurrent sweeps. This models a scenario of continuous positive selection, with beneficial mutations arising at different times. By varying the interval between each introduction of positive selection, we can model full selective sweeps that occur successively (Kim, 2006) or interfering sweeps (Coop and Ralph, 2012). Note that unlike coalescent simulations (Coop and Ralph, 2012; Kim, 2006), we control the rate at which beneficial mutations are introduced rather than the sweep rate. The selective sweep may occur considerably later than the time at which positive selection is introduced because genetic drift, background selection and hitch-hiking can affect the time required for beneficial mutations to reach establishment.

For a low supply rate of beneficial mutations, we expect beneficial mutations to fix primarily in successive sweeps with rare occurrences of clonal interference, whereas clonal interference will occur with high probability when the supply rate of beneficial mutations is high. The expected time for a beneficial mutation to become established in the population is given by $t_{est} = 1/(uNL_b s_b)$ (Desai and Fisher, 2007); after establishment, the beneficial mutation behaves almost deterministically, increasing rapidly in frequency and is expected to fix in $t_{fix} = \log(Ns_b)/s_b$ generations (Desai and Fisher, 2007). For population size $N = 10^4$ and $u = 10^{-6}$, a single beneficial mutation of strength $s_b = 0.01$ is expected to have establishment and fixation times of $t_{est} \approx 2857$ and $t_{fix} \approx 460$ generations. To obtain a high supply rate of beneficial mutations, we introduce positive selection at high frequency, specifically at one site in every $\tau = 1000$ generations, which is faster than the rate of establishment. For a low supply rate of beneficial mutations, we set $\tau = 10000$ generations, so that establishment and fixation of one beneficial mutation is likely to occur before a second positively selected site is introduced. Note that varying the timing of positive selection controls the supply rate of beneficial mutations (generally parameterised as $U_b N = uL_b N$) indirectly. After positive selection is introduced at a site, L_b is increased by one; however, L_b is also decreased when a beneficial mutation reaches fixation.

2.2.2 Simulations under independently segregating sites

To compare sequence statistics obtained under complete linkage with those obtained under the assumption of independently segregating sites, we simulate the number of polymorphic and divergent sites according to the Poisson Random Field (PRF) model (Sawyer and Hartl, 1992). The PRF model assumes a Wright-Fisher population at equilibrium with an infinite number of sites so that all new mutations occur on distinct sites. Under these assumptions, Sawyer and Hartl (1992) showed that number of sites carrying a derived mutation follows a Poisson random field, with expectations that are functions of the mutation and selection parameters. We use the PRF as it is the basis of a number of inference methods (Nielsen and Yang, 2003; Piganeau and Eyre-Walker, 2003; Sawyer and Hartl, 1992; Smith and Eyre-Walker, 2002), and therefore provides a better reference than a finite-site model with independently segregating sites.

In the PRF framework (Sawyer and Hartl, 1992), the number of derived sites can be simulated as independent Poisson variables. We can then use the number of divergent and polymorphic sites to calculate sequence statistics $\hat{\omega}$, $\hat{a}_{MK}$ and $\hat{a}_F$ as described in the main text. In the following section, we give the equations used to calculate the mean number of divergent and polymorphic sites.

In the case where there is no positive selection, the expected number of synonymous and non-synonymous divergent sites, as described in Sawyer and Hartl (1992), is given by

$$E(k'_S) = u_S L t \tag{2.2}$$

$$E(k'_A) = u_A L t \int \omega(-s_d, N) \rho(s_d, \beta, \bar{s}) \,, \tag{2.3}$$

where $\omega(.)$ is given by Equation (2.15), $\rho(.)$ is the DFE, t is the divergence time, L is the length of sequence, $u_S = u/(1+c)$ and $u_A = uc/(1+c)$. Using $\rho(.)$ as given in Equation (2.1), this can be simplified to (Charlesworth and Eyre-Walker, 2008)

$$E(k'_A) = u_A L t \beta \left(\frac{\beta}{2N\bar{s}}\right)^{\beta} \zeta\left(\beta+1, \frac{\beta}{2N\bar{s}}+1\right) \tag{2.4}$$

where,

$$\zeta(s,a) = \frac{1}{\Gamma(s)} \int_0^{\infty} \frac{t^{s-1}}{e^{at}(1-e^{-t})} dt \tag{2.5}$$

denotes the Hurwitz zeta function which is provided in the GNU scientific library (Galassi et al., 2003). When $L_b > 0$ sites are positively selected, we generate the number of divergent non-synonymous sites over the deleterious portion of the sequence using Equation (2.4) and the number of divergent beneficial sites is generated from a truncated Poisson distribution with mean $u_A L_b t \omega(s_b, N)$, capped at L_b. This allows comparison with the finite sites model which explicitly does not allow recurrent positive selection at a single site.

The expected number of derived polymorphic sites with selection coefficient s segregating at frequency x in the population is given by (Wright, 1938)

$$\theta\phi(x, Ns) = \frac{\theta}{x(1-x)} \frac{1-e^{-2Ns(1-x)}}{1-e^{-2Ns}}\,, \tag{2.6}$$

where $\theta = 2uNL$ is the mutation input rate. For a sample of size n with a known ancestral sequence, the expected numbers of synonymous and non-synonymous polymorphic sites observed at frequency i, as given in Sawyer and Hartl (1992), are

$$\begin{aligned} E(p_S(i)) &= \theta_S \int_0^1 \binom{n}{i} x^i (1-x)^{n-i} \phi(x, 0) dx \\ &= \frac{\theta_S}{i} & (2.7) \\ E(p_A(i)) &= \theta_A \int_0^1 \binom{n}{i} x^i (1-x)^{n-i} \times \\ & \int_0^\infty \rho(s_d, \beta, \bar{s}) \phi(x, -Ns_d) ds dx & (2.8) \end{aligned}$$

where $\theta_S = 2u_S NL$ and $\theta_A = 2u_A NL$. Applying the gamma DFE used in our model, Equation (2.8) can also be simplified in terms of the Hurwitz zeta function to give

$$\begin{aligned} E(p_A(i)) &= \theta_A \binom{n}{i} \beta \left(\frac{\beta}{2N\bar{s}}\right)^\beta \times \\ & \int_0^1 b(x, i, n-i) \zeta\left(\beta+1, \frac{\beta}{2N\bar{s}} + x\right) dx. & (2.9) \end{aligned}$$

where

$$b(x, a, b) = \int_0^x x^{a-1}(1-x)^{b-1} dx\,, \tag{2.10}$$

denotes the incomplete beta function. To calculate sequence statistics under assumptions of independently segregating sites, we sample the number of segregating synonymous and non-synonymous polymorphisms from Poisson distributions characterised by Equations (2.7) and (2.9). The number of observed divergent sites is given by

$$\begin{aligned} k_S &= k'_S + \frac{1}{n} \sum_{i=1}^{n-1} i p_S(i) & (2.11) \\ k_A &= k'_A + \frac{1}{n} \sum_{i=1}^{n-1} i p_A(i) & (2.12) \end{aligned}$$

where k'_S and k'_A are Poisson random variables described by Equations (2.2) and (2.4).

2.2.3 Selection statistics

In each simulation, we randomly sample $n = 100$ sequences every 2000 generations. Based on each sample and the known ancestral sequence, we then calculate the K_A/K_S and MK statistics as follows. Let $p_A(i)$ denote the number of derived polymorphic codon sites that are non-synonymous (relative to the ancestral codon) and occur i times in the sample of size $n = 100$, and similarly, let $p_S(i)$ denote the number of derived synonymous polymorphic sites that occur i times. Multiple mutations at the same site are counted as distinct polymorphisms. The number of synonymous divergent sites and non-synonymous divergent sites is given respectively by

$$k_S = \frac{1}{n}\sum_{i=1}^{n} i p_S(i) \tag{2.13}$$

$$k_A = \frac{1}{n}\sum_{i=1}^{n} i p_A(i)\,. \tag{2.14}$$

The K_A/K_S statistic is given by (Hughes and Nei, 1988),

$$\hat{\omega} = \frac{k_A}{ck_S}\,. \tag{2.15}$$

The scaling factor $c = 2.4$ accounts for the fact that non-synonymous mutations are more likely than synonymous mutations due to the structure of the genetic code. It is calculated by summing across the substitution matrix, in our case, the Kimura two-parameter model (Kimura, 1980). Standard methods (Yang, 2007) will automatically account for this scaling factor. Using this scaling, $\hat{\omega}$ can be interpreted as a function of the strength of selection s and the population size N, which under the assumptions of a Wright-Fisher population with independently segregating sites is given by (Nielsen and Yang, 2003)

$$\omega(Ns) \approx \frac{2Ns}{1 - e^{-2Ns}}\,. \tag{2.16}$$

This is obtained by taking the ratio between fixation probabilities of a selected and a neutral mutation (Kimura, 1962). In the case where positive selection is not operating, the value of ω summed across the entire sequence is equal to the proportion of effectively neutral sites, denoted f (Eyre-Walker and Keightley, 2007).

We use a modification of the MK test (McDonald and Kreitman, 1991) which provides a quantitative measure of adaptive substitution (Smith and Eyre-Walker, 2002),

$$\hat{a}_{MK} = k_A - k_S \frac{\sum_{i=1}^{n-1} p_A(i)}{\sum_{i=1}^{n-1} p_S(i) + 1} \, . \tag{2.17}$$

The MK statistic does not require a scaling factor c, as it is given in units of the number of non-synonymous substitutions. The offset (+1) term in the denominator means that this estimator is defined in all cases. Smith and Eyre-Walker (2002) found that the offset does not introduce noticeable bias.

The ratio in Equation (2.17) is an estimator of f, under the assumption that all segregating polymorphisms are selectively neutral. This assumption is valid when selection is strong so that selected mutations immediately reach fixation or extinction, but not when weak selection is frequent. This problem is further compounded in the context of linked selection as linkage has the effect of weakening the effective strength of selection so that both deleterious and beneficial mutations can potentially segregate for longer prior to extinction or fixation. Here, we examine two modifications of the MK statistic.

The first is motivated by weakly deleterious mutations that segregate transiently in the population, which are known to inflate the estimate of selective constraint and cause underestimation of the number of adaptive substitutions (Charlesworth and Eyre-Walker, 2008). To correct for this, we exclude low-frequency (< 0.15) derived polymorphisms from the analysis, following Fay et al. (2001), giving

$$\hat{a}_F = k_A - k_S \frac{\sum_{i=[0.15n]}^{n-1} p_A(i)}{\sum_{i=[0.15n]}^{n-1} p_S(i) + 1} \, , \tag{2.18}$$

where the square brackets indicate rounding to the nearest integer. A further modification used by Bhatt et al. (2011) is to exclude high-frequency polymorphisms which are likely to contain beneficial mutations and would, if included, lead to an overestimate of f and therefore underestimation of the number of adaptive substitutions,

$$\hat{a}_B = k_A - k_S \frac{\sum_{i=[0.15n]}^{[0.75n]} p_A(i)}{\sum_{i=[0.15n]}^{[0.75n]} p_S(i) + 1} \, . \tag{2.19}$$

Both $\hat{a}_F$ and $\hat{a}_B$ were developed to account for selected variation segregating in the population on the assumption of independently segregating sites. However, in the context of frequent selection, linkage between sites is also likely to have strong effect, motivating us to consider the performance of these statistics. For comparison with the MK statistics, it is helpful to consider the performance of an estimator that does not use polymorphism information. Based on the $\hat{\omega}$ statistic, we estimate the number of adaptive substitutions using

$$\hat{a}_D = k_A - ck_S\omega_0 \,. \tag{2.20}$$

In fact, $\hat{a}_D$ is not a true estimator as ω_0 is a fixed value (treating f as known) rather than a measurable quantity. Here, ω_0 is obtained using the median value of $\hat{\omega}$ based on simulations with linkage and the same values of β and $\bar{s}$ but no positive selection ($\omega_0 = 0.09, 0.09, 0.11, 0.12$ for $u = 10^{-6}$ and $\omega_0 = 0.10, 0.13, 0.23, 0.33$ for $u = 10^{-5}$). We used simulations rather than the theoretical expectation of f to account for background selection. In practice, ω_0 cannot be estimated from divergence information unless there is a period where it is known positive selection has not occurred. However, we use $\hat{a}_D$, as it provides a comparison showing how $\hat{a}_F$ and $\hat{a}_B$ differ in their estimation of f.

2.3 Results

2.3.1 The effect of background selection

We begin by examining the effect of negative selection and linkage without positive selection in a protein-coding region of 500 codons evolving under a Wright-Fisher process. Negative selection is described by the distribution of fitness effects (DFE) of non-synonymous changes, which are specific to each codon site. The DFE is modelled using a gamma distribution where a large value of the shape parameter β corresponds to a higher proportion of weakly deleterious mutations.

The effect of background selection on the $\hat{\omega} = K_A/K_S$ statistic is shown in in Figure 2.1. The density of estimators with linked selection, computed using Equation (2.15) is shown in solid lines, whereas the corresponding values obtained with independently segregating sites from PRF simulations are shown with dashed lines. Both simulations account for the contribution of segregating polymorphisms. The effect of linkage, therefore, is shown by the difference between simulations with linkage and without linkage.

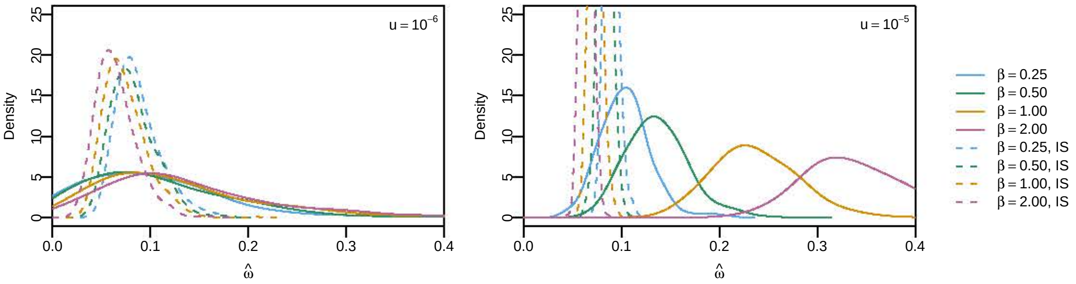

Figure 2.1: Distribution of $\hat{\omega}$. Distribution of $\hat{\omega}$ under only negative selection for DFE shape parameters $\beta = 0.25, 0.5, 1, 2$. Solid curves indicate simulation results under complete linkage and dashed curves indicate results based on independently segregating sites using the PRF. Distributions were calculated from 100 sequences sampled at $6N$ generations with 500 replicates.

As expected, the effect of background selection in reducing $\hat{\omega}$ increases with β and u. Our simulations also show that linkage increases the variance of the estimator due to correlations between linked sites. This is particularly evident for $u = 10^{-5}$ where the distribution of $\hat{\omega}$ visibly broadens with increasing β.

In Figure 2.2, we consider three forms of the MK statistic: (i) the uncorrected estimator $\hat{a}_{MK}$ [Equation (2.17)] and (ii) Fay's corrected estimator $\hat{a}_F$ [Equation 2.18)] which removes low-frequency polymorphisms to reduce the effect of segregating deleterious polymorphisms and (iii) Bhatt's corrected estimator $\hat{a}_B$ [Equation (2.19)] which removes both low and high frequency polymorphisms which are likely to contain deleterious and beneficial mutations. In the absence of positive selection, we expect $\hat{a}_F$ and $\hat{a}_B$ to perform similarly, and this is indeed seen for $u = 10^{-6}$. However, for simulations with a higher mutation rate and correspondingly larger effect of background selection, discrepancies occur between the two statistics due to an increase in the number of high-frequency polymorphisms. Unlike $\hat{\omega}$, the variance of the MK statistics does not seem to be affected by linkage. In fact the performance of the MK statistics (in the absence of positive selection) is slightly improved by background selection which removes weakly deleterious mutations.

2.3.2 The combined effect of background selection, clonal interference and hitch-hiking

In the following section, we examine the combined effect of negative and positive selection. Positive selection is introduced at a fixed number of sites at intervals of τ generations

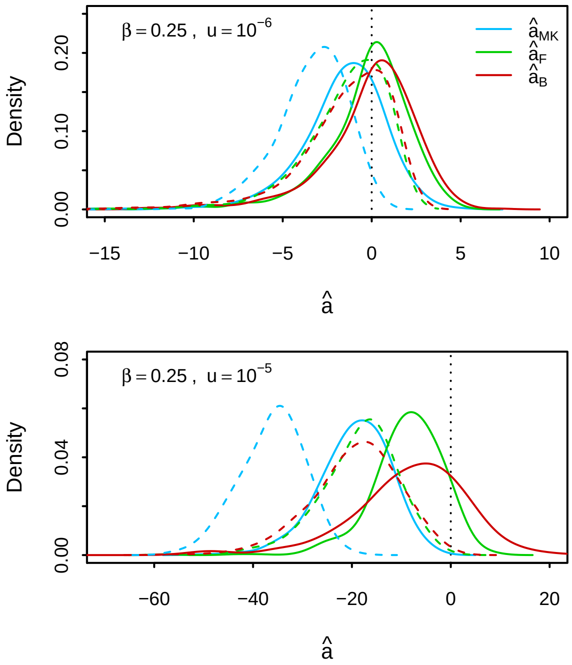

Figure 2.2: Distribution of the MK statistics. Distribution of the MK statistics under only negative selection. Results are shown for simulations with complete linkage (solid lines) and independently segregating sites (dashed lines) for different DFEs and mutation rates. The true number of adaptive substitutions (zero) is indicated by the dotted vertical line.

throughout the simulation, where all positively selected sites have the same selective coefficient s_b. Decreasing τ increases the probability of interfering positive sweeps. A comparison of the effects of different selective conditions on the site frequency spectrum is shown in Figure 2.3. Note that these curves represent averaged levels of polymorphisms, and dynamics can vary rapidly over time (see Figures A.3–A.20).

We show results for low levels of background selection (small u) in the left column and results for high levels of background selection in the right column. The (unscaled) synonymous site frequency spectrum is shown in the top row. When the effect of background selection is small, the synonymous site frequency spectrum is close to the expectation under independently segregating sites (θ/i; black dashed lines). Background selection (bold grey lines) reduces the level of synonymous variation, particularly at medium frequencies, leading to a non-monotonic distribution but the effect is not as severe as clonal interference. Linked positive selection further reduces polymorphism levels; a slow rate of sweeps with strong selection (orange lines) primarily affects high-frequency mutations while a high supply of weak positive selection (green lines) results in smaller levels of reduction at both low and high frequencies. When both the supply rate and the strength

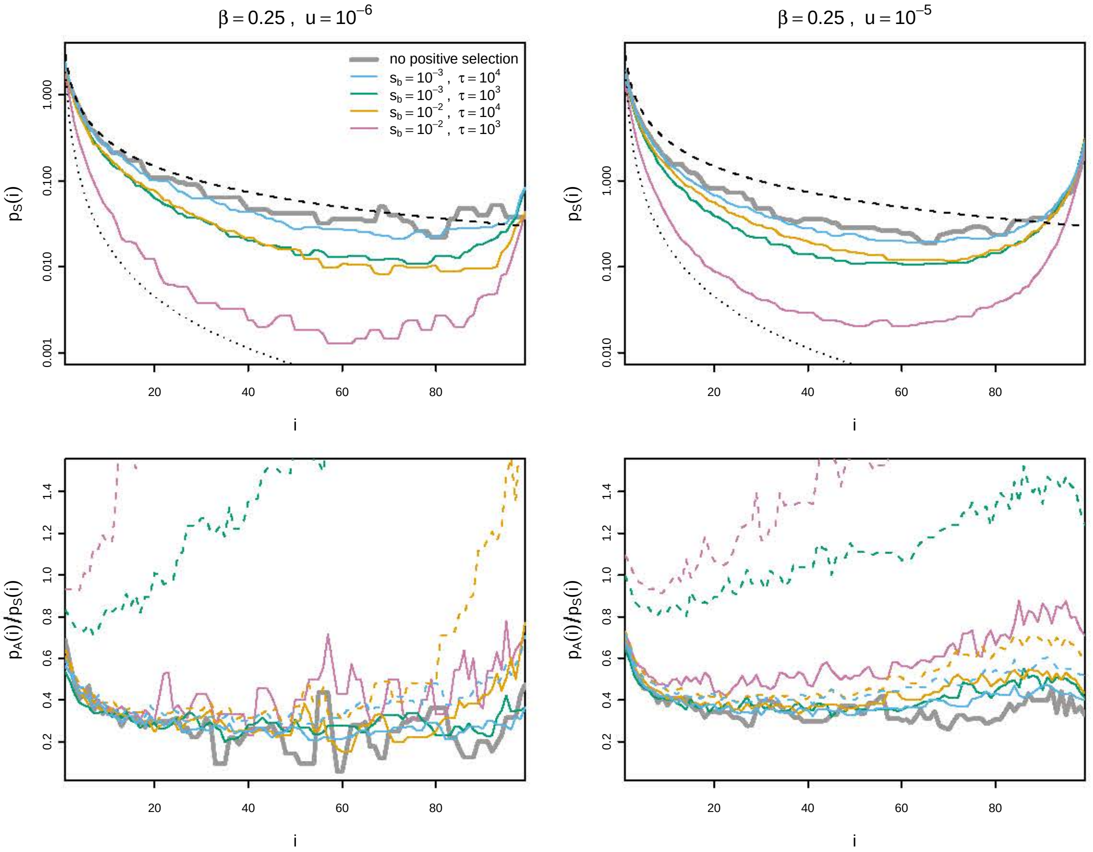

Figure 2.3: The effect of linkage on the site frequency spectrum. The synonymous site frequency spectrum (top row) and the ratio of non-synonymous to synonymous frequency spectrum (bottom) is shown for $\beta = 0.25$ with mutation rates $u = 10^{-6}$ and 10^{-5}. All curves are averaged over 500 replicates, under conditions of only negative selection (grey), and different conditions of positive selection (coloured lines). Black dashed lines show the expected behaviour of the neutral site frequency spectrum under independently segregating sites (θ/i) and under black dotted lines indicate the leading order behaviour expected under constant adaptation (θ/i^2). In the bottom panels, solid lines show the average non-synonymous to synonymous ratio for only negatively selected sites, whereas dashed lines show the ratio across both positively and negatively selected sites.

of positive selection is strong (pink lines), the synonymous site frequency spectrum approaches θ/i^2 (black dotted line) which is the leading behaviour predicted for continual adaptation (Neher and Shraiman, 2011).

To examine how linkage affects selected mutations, we compare the ratio of the averaged frequency spectra for non-synonymous (A) and synonymous (S) sites (Figure 2.3, bottom row). The A/S ratio in the absence of positive selection is indicated by the bold grey line, whereas the A/S ratio for deleterious sites linked to positively selected sites is shown by coloured solid lines. The discrepancy between the grey and coloured lines reflects the effect of hitch-hiking; there is a slight increase in the A/S ratio at high-frequencies due to hitch-hiking. Note that the actual number of deleterious polymorphisms is reduced relative to simulations with no positive selection (Figure A.1) but the number of synonymous polymorphisms is reduced by a relatively greater proportion.

Comparing the A/S ratio with (dashed coloured lines) and without (solid coloured lines) beneficial mutations, it can be seen that beneficial mutations can segregate at all frequencies when the supply rate is high (green and pink lines), but mutations segregating at high frequencies tend to include more beneficial mutations. Comparison of the two panels in the bottom row also shows that higher levels of background selection increases the effect of both hitch-hiking (solid coloured lines) and clonal interference (dashed coloured lines) as distortions in the site-frequency spectrum tend to occur over a wider range of frequencies. Similar results are seen for larger values of β with more pronounced reductions of synonymous polymorphism due to background selection, and changes in the A/S ratio due to hitch-hiking and clonal interference are spread across a broader frequency range (Figure A.1).

The contributions of background selection, hitch-hiking and clonal interference result in qualitatively different behaviour in the site-frequency spectrum and this in turn causes characteristic types of bias in the various forms of the MK statistic. This is summarised in Figure 2.4, where we compare the performance of different forms of the MK statistic in estimating the true number of beneficial mutations in each simulation. Here, we do not consider the uncorrected $\hat{a}_{MK}$ as it was severely biased in all the simulations we examined. An additional MK statistic, $\hat{a}_D$ is considered which uses divergence information from simulations with no positive selection instead of estimating selective constraint from

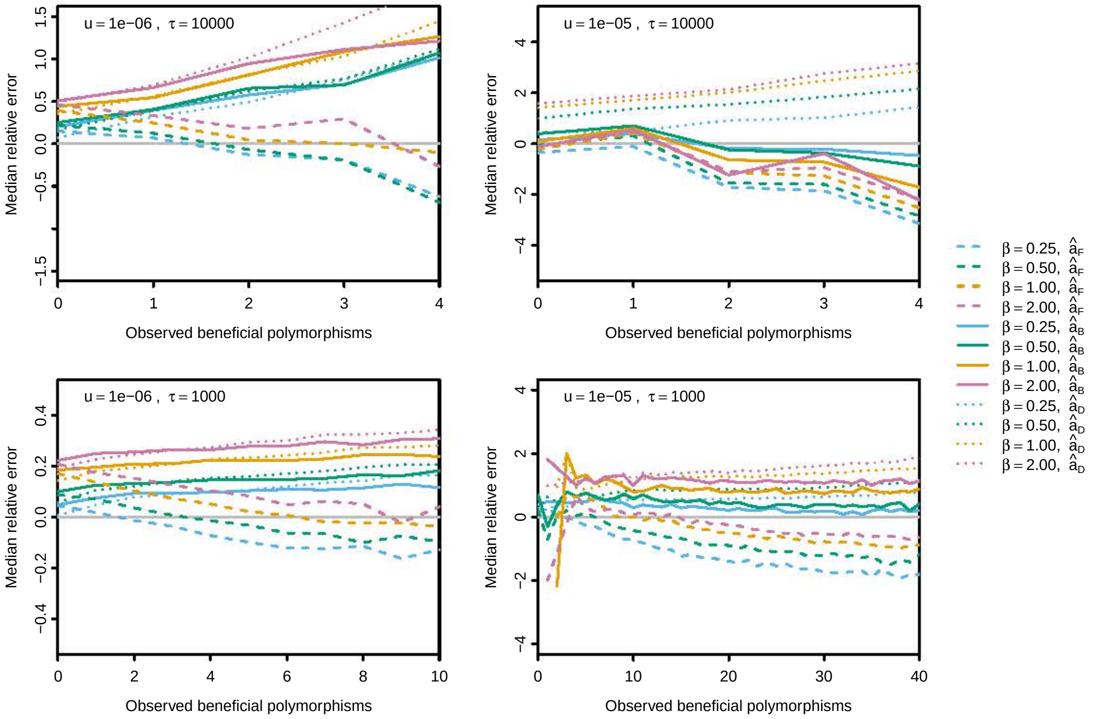

Figure 2.4: Effect of background selection, hitch-hiking and clonal interference on the MK statistics. Effect of background selection, hitch-hiking and clonal interference recurrent sweeps. Lines indicate the median relative error of different forms of the MK statistics for $s_b = 10^{-2}$ across all time points.

polymorphism information. Comparison of $\hat{a}_F$ or $\hat{a}_B$ against $\hat{a}_D$, therefore, shows how much of the bias is due to incorrect estimation of selective constraint.

The different panels in Figure 2.4 correspond to different combinations of positive and negative selection: low levels of background selection (due to strong negative selection) and infrequent positive sweeps (top left), low levels of background selection and frequent positive sweeps (bottom left), high levels of background selection and infrequent positive sweeps (top right) and high levels of background selection with frequent positive sweeps (bottom right). When the effect of background selection is large (top right) both $\hat{a}_F$ and $\hat{a}_B$ tend to underestimate the true number of adaptive substitutions. For low levels of background selection or frequent positive sweeps, the effect of hitch-hiking (controlled by β) and the amount of clonal interference (using the observed number of beneficial mutations as a proxy) has a consistent effect on the MK statistics. For small values of β so that low levels of hitch-hiking occur, $\hat{a}_B$ has smaller bias than $\hat{a}_F$. However, for high levels of hitch-hiking, $\hat{a}_F$ is less biased, particularly when clonal interference is low. Results for different values of s_b were qualitatively similar but with larger relative error for weaker positive selection.

The reason for these biases is intuitively clear from the site frequency spectrum. $\hat{a}_B$ differs from $\hat{a}_F$ only in that it does not use polymorphism data at high frequency. Therefore, $\hat{a}_B$ is more robust against clonal interference (Figure 2.4, bottom row) as beneficial mutations are more likely to segregate at high frequencies. However, when weakly deleterious effects are prevalent (Figure 2.4, solid pink lines), $\hat{a}_B$ is upwardly biased as it does not account for the relaxation of selective constraint due to positive selection. This is confirmed by the similar values obtained for $\hat{a}_B$ and $\hat{a}_D$, suggesting that removal of high and low frequency polymorphisms in the context of linked selection has a similar effect to that expected under independently segregating sites, namely the removal of both positively and negatively selected mutations. Bhatt's correction does not perform well when there are high levels of background selection as distortions in the site frequency spectrum are spread across a wider range of frequencies than without background selection.

2.3.3 Diagnostics for linkage effects

In the previous section, we showed that much of the bias in the comparative estimators can be explained in terms of background selection, hitch-hiking and clonal interference. In order to detect these effects using samples of protein-coding sequences, we construct and examine three diagnostic statistics.

The first diagnostic tests for an excess of low frequency non-synonymous polymorphisms relative to medium frequency polymorphisms. For a sample size of n, we consider a mutation to occur at low frequency if it occurs i times in the sample, where i belongs to the set $\mathcal{I}_L = \{1, 2, \ldots, [0.15n] - 1\}$. Charlesworth and Eyre-Walker (2008) showed that the majority of deleterious polymorphisms occurred in this frequency range even when the sample size is varied. Similarly, we consider a mutation to occur at medium frequencies if the number of times it occurs in the sample belongs to $\mathcal{I}_M = \{[0.15n], [0.15n] + 1, \ldots, [0.75n]\}$. The first diagnostic is given by

$$D_1 = \frac{\sum_{i \in \mathcal{I}_L} p_A(i)}{\sum_{i \in \mathcal{I}_L} p_S(i) + 1} - \frac{\sum_{i \in \mathcal{I}_M} p_A(i)}{\sum_{i \in \mathcal{I}_M} p_S(i) + 1} \,. \tag{2.21}$$

If weak deleterious effects are rare, then we expect that most deleterious mutations are immediately removed from the population. In this case, most polymorphisms would be selectively neutral and we would expect that the ratio of non-synonymous to synonymous polymorphisms, at any frequency range, is simply determined by the mutational bias. The difference of the two ratios in D_1 is therefore expected to equal zero in the absence of weak deleterious effects and large values are indicative of a high frequency of weak deleterious mutations, which results in susceptibility to hitch-hiking.

In Figure 2.5, we show the correlation between D_1 and the amount of hitch-hiking, which we measure as the relative excess of non-synonymous substitutions at non-beneficial sites in simulations with positive selection compared to simulations with no positive selection. A value of 1.0 in the x-axis corresponds to half of all non-synonymous substitutions being due to hitch-hiking. When positive selection is weak so that $\hat{a}_B < 0$ (open circles), D_1 correlates with the β shape parameter so that values of $D_1 > 0$ indicate susceptibility to hitch-hiking. When strong positive selection occurs selective constraint is reduced so that the proportion of mutations that can be considered weakly deleterious may be increased. In this case, we see that D_1 is also increased, even for small values of β. Interpretation of the D_1 statistic, therefore, should depend on both the value of D_1 and the MK statistic. We use $\hat{a}_B$ here as Figure 2.4 indicates that it is less likely to result in underestimation than $\hat{a}_F$.

The second diagnostic tests for an excess of high frequency polymorphisms which is an indication of multiple merger events (Desai et al., 2013; Neher and Hallatschek, 2013) due to interfering selected mutations, which can be either negative (background selection) or positive (clonal interference). We compare the number of high frequency polymorphisms to medium frequency polymorphisms, where a mutation is defined to be at high frequency if the number of times it occurs in the sample belongs to $\mathcal{I}_H = \{[0.75n]+1, \ldots, n-1\}$ and $|x|$ denotes the number of elements in the set x,

$$D_2 = \frac{\sum_{i \in \mathcal{I}_M} i p_A(i)}{|\mathcal{I}_M|} - \frac{\sum_{i \in \mathcal{I}_H} i p_A(i)}{|\mathcal{I}_H|}\,. \tag{2.22}$$

Deleterious mutations are not expected to persist to medium frequencies, so polymorphisms at medium and high frequencies can be assumed to be neutral or beneficial. Under assumptions of neutrality and independently segregating sites, the expected number of polymorphic sites that occur at frequency i is given by $E(p_A(i)) = \theta_A/i$, where

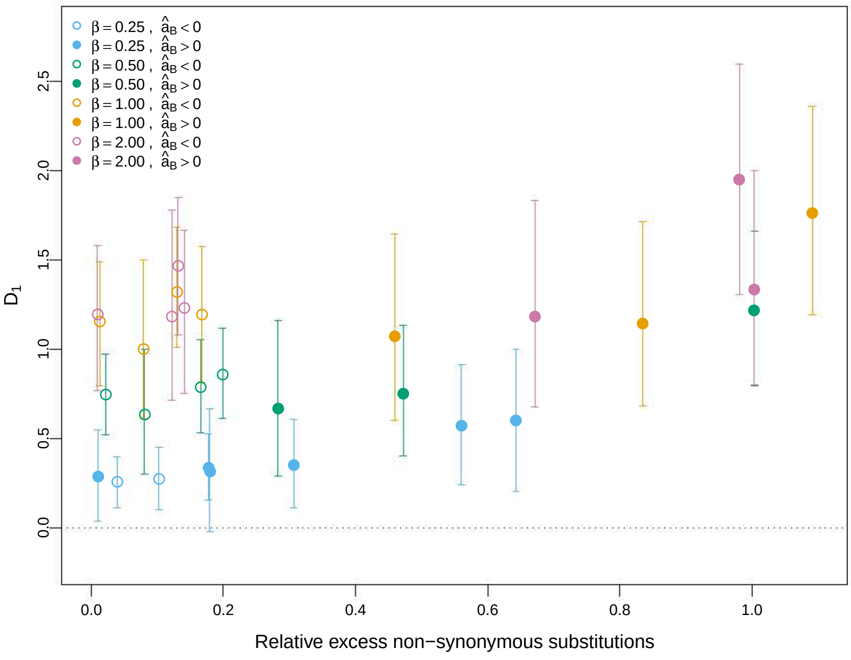

Figure 2.5: Diagnostic for hitch-hiking. Circles and bars indicate median and interquartile ranges from all combinations of recurrent sweeps with $\tau = 10^3, 10^4$, $s_b = 10^{-3}, 10^{-2}$, $u = 10^{-5}, 10^{-6}$ and $\beta = 0.25, 0.5, 1, 2$. Parameter combinations which give a median $\hat{a}_B < 0$ are denoted with an open circle whereas simulation conditions which result in $\hat{a}_B > 0$ are shown with closed circles. The x-axis is the relative excess of non-synonymous substitutions due to linked positive selection, calculated as $(\bar{k}' - \bar{k})/\bar{k}$ averaged across all time points after 40000 generations, where $\bar{k}$ is the mean number of non-synonymous substitution at non-beneficial sites averaged across 500 simulations with no positive selection and the prime indicates the corresponding values in a simulation with both positive and negative selection.

$\theta_A = 2uNLc/(c+1)$ giving an expectation of $D_2 = 0$. Values of $D_2 < 0$ can, therefore, indicate strong linkage effects due to an excess of beneficial or deleterious mutations.

A third statistic can distinguish between the effect of background selection and clonal interference,

$$D_3 = \frac{2\sum_{i\in\mathcal{I}_H} ip_A(i)}{|\mathcal{I}_H|} - \frac{\sum_{i\in\mathcal{I}_M} ip_A(i)}{|\mathcal{I}_M|} - \frac{\sum_{i\in\mathcal{I}_H} ip_S(i)}{|\mathcal{I}_H|} \cdot \frac{\sum_{i=1}^{n-1} p_A(i)}{\sum_{i=1}^{n-1} p_S(i) + 1} . \tag{2.23}$$

This statistic tests for an excess of high-frequency non-synonymous polymorphisms relative to both high frequency synonymous polymorphisms relative to both medium frequency non-synonymous polymorphisms and high-frequency synonymous polymorphisms. As with D_1 and D_2, the expectation under independently segregating neutral sites is $D_3 = 0$ and values of $D_3 > 0$ are indicative of clonal interference. In Figure 2.6, values of D_2 and D_3 are shown for varying levels of background selection and clonal interference. In the left panel, low mutation rates generate only low levels of background selection and values of D_2 and D_3 are strongly correlated, as both are due to clonal interference. In the right panel, a high mutation rate increases levels of both background selection and clonal interference. Simulations with a high supply rate of beneficial mutations (filled red circles) have large values of D_3 and strongly negative D_2 values, whereas simulations with a low supply rate of beneficial mutations and occasional instances of clonal interference tend to small positive values of D_3 with negative values of D_2 (filled blue circles). When only high levels of background selection are acting, both D_3 and D_2 fall below zero (open black circles). The behaviour of these three diagnostics are similar for different sample sizes (Figure A.2) and different population sizes (Figures A.15–A.20).

In Figure 2.7, we show that the bias of $\hat{a}_F$ and $\hat{a}_B$ varies systematically with D_3 (clonal interference) and D_1 (hitch-hiking). Larger values of D_1 and D_3 tend to result in larger values for both statistics; for $\hat{a}_F$ this tends to reduce the magnitude of the bias, but increases bias for $\hat{a}_B$. This suggests that $\hat{a}_F$ performs better for large D_1 but $\hat{a}_B$ performs better for large D_3 and small D_1. The size of the bias for both statistics is larger for higher mutations rates (bottom row, $u = 10^{-5}$) which corresponds to very large D_2 values (Figure 2.6) and larger effects of background selection. In particular, when $D_3 < 0$ and $D_2 \ll 0$, both statistics are expected to heavily underestimate the amount of positive selection that has occurred.

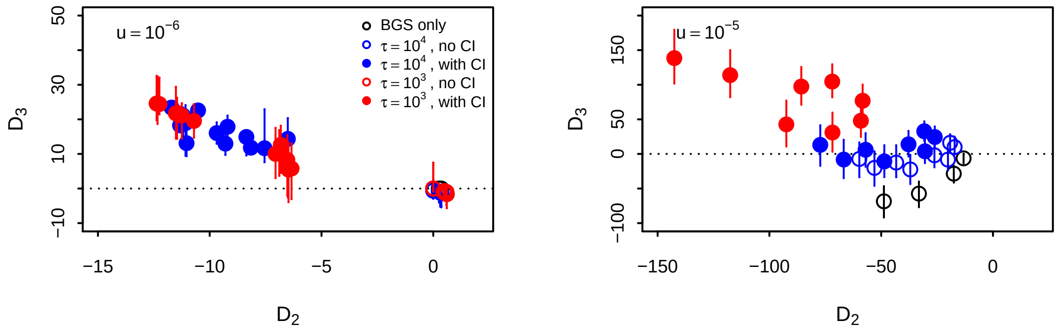

Figure 2.6: Diagnostics for clonal interference and background selection. Median values of D_2 and D_3 are shown for all combinations of $s_b = 10^{-3}, 10^{-2}$ and $\beta = 0.25, 0.5, 1, 2$ for all time points after 40000 generations. Bars represent interquartile ranges for D_3. In the left panel, negative values of D_2 are mostly due to clonal interference but in the right panel, negative values of D_2 are caused by a combination of clonal interference and background selection.

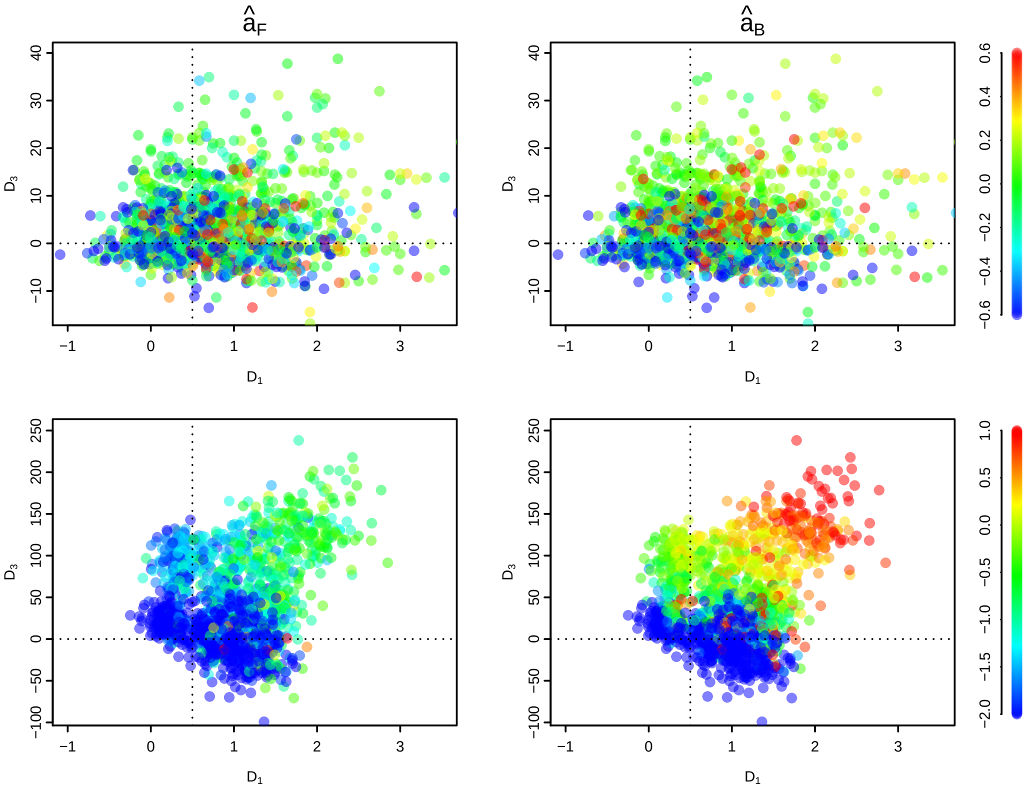

Figure 2.7: Correlation between diagnostics and bias of the MK statistics. Each point in the plot represents a single simulation replicate with the value of bias of $\hat{a}_F$ (left column) and $\hat{a}_B$ (right column) indicated by the colour of the point. Results are shown for $u = 10^{-6}$ (top row) and $u = 10^{-5}$ (bottom row), and each panel consists of 100 replicates from all combinations of recurrent sweeps with $\tau = 10^3, 10^4$, $s_b = 10^{-3}, 10^{-2}$ and $\beta = 0.25, 0.5, 1, 2$. Simulations with bias outside the range of the colour scale were set at the extreme values and points with less than two medium frequency polymorphisms were excluded.

To evaluate whether D_1, D_2 and D_3 differ from zero, we use a non-parametric bootstrap, recalculating statistics after resampling with replacement from the original sequence sample. The scaling factor for mutation bias c, which is omitted from D_1 is automatically accounted for by this method. Confidence intervals for D_1 were constructed from the bootstraps using the 2.5 to 97.5 percentiles. As D_2 is slightly biased confidence intervals for D_2 and D_3 were constructed using the BCA method provided in R (Efron, 1987).

2.3.4 Application of diagnostics to human influenza A (H3N2)

We applied the diagnostics with the bootstrap method to the human influenza A (H3N2) dataset used by Strelkowa and Lässig (2012). The dataset comprises 2030 sequences with a length of 330 codons spanning 1968–2007. The list of accession numbers is provided in the Additional file 1 in Strelkowa and Lässig (2012). Following Strelkowa and Lässig (2012), we used A/Bilthoven/16190/1968 as the ancestral sequence; results using A/Hong Kong/1/1968 were very similar. Diagnostics D_1 and D_2 were computed for samples in each year separately, with sample sizes ranging from 5 to 215. The results are shown in Figure 2.8. There is some variation over time, with wider confidence intervals in the earlier samples due to small sample sizes, but D_1 values are mostly centred around zero, suggesting low levels of hitch-hiking. Hitch-hiking cannot be conclusively ruled out as confidence intervals are quite wide and a number of points reach $D_1 = 1$. However, values of D_1 remain consistently low with a number of time points falling below zero which is more consistent with a low hitch-hiking scenario. In contrast, simulations with prevalent hitch-hiking tend to to have confidence intervals that are consistently above zero and point estimates much higher than 0.5 (Figures A.3–A.9). Values of D_2 are strongly negative, indicating a strong effect due to interfering deleterious or beneficial mutations; the magnitude of D_2 is consistent with a high level of background selection. Multiple time points with $D_3 \gg 0$ also suggests that clonal interference frequently occurs in the evolution of H3N2.

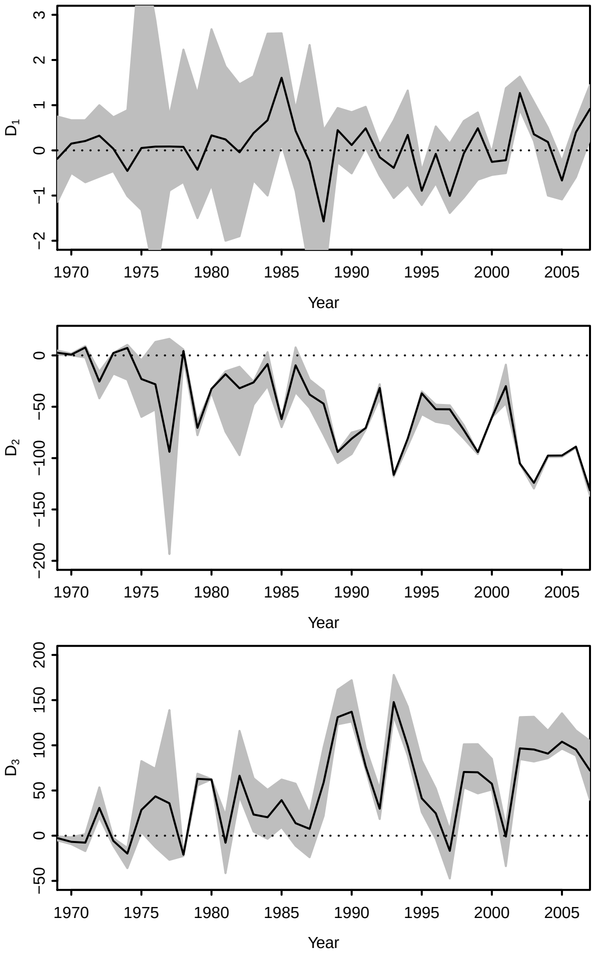

Figure 2.8: Application of diagnostics to human influenza A. Diagnostics D_1, D_2 and D_3 applied to the human influenza A (H3N2) HA1 region. Shaded regions shows (uncorrected) 95% confidence intervals based on 10000 bootstrap replicates, calculated for each time point separately.

2.4 Discussion

It has long been known that linkage influences polymorphism frequencies and fixation probabilities which can adversely affect methods that assume independent segregation of sites. The MK statistic, which compares the number of polymorphic sites rather than using only frequency information, is generally considered to be more robust to linkage effects than frequency-based statistics (Bustamante et al., 2001; Comeron and Kreitman, 2002). In this study, we show that the MK statistic can be affected, depending on the mode of linked selection and its characteristic effect on the site-frequency spectrum.

When background selection has a large effect, distortions in the site-frequency spectrum result in downward bias in the MK statistics. However, when the effect of negative selection is low compared to positive selection, the performance of the MK statistics as a quantitative estimator of the number of adaptive substitutions is determined by the extent of hitch-hiking and clonal interference. When negative selection is strong so that levels of hitch-hiking are low, $\hat{a}_B$ tends to perform better. Specifically, it is more robust against distortions of the site frequency spectrum at high frequencies caused by background selection or clonal interference. However, $\hat{a}_F$ is more robust to hitch-hiking which occurs when weak negative selection is pervasive.

Our results are consistent with that of a recent study by Messer and Petrov (2013) showing that $\hat{a}_F$ tends to underestimate the number of adaptive substitutions. We find this primarily occurs when background selection has a large effect and positive selection is infrequent. However, when positive selection is strong, hitch-hiking can also lead to overestimation as suggested in some empirical studies (Fay, 2011). When interactions between deleterious polymorphisms dominate the dynamics of the populations, the asymptotic correction proposed by Messer and Petrov (2013) can be used to correct for underestimation due to low and medium frequency deleterious polymorphisms. This method corrects for deleterious mutations as the relative abundance of deleterious mutations is reduced at higher frequencies but cannot be applied for beneficial mutations which have increased relative abundance at higher frequencies.

Here, we show that, when background selection is relatively weak, choosing the appropriate form of the MK statistic can reduce estimation bias. Messer and Petrov (2013) results apply for organisms with large genomes and many weakly deleterious mutations, but when genomes are small and selective effects are broadly distributed, as is the case viral populations (Sanjuán et al., 2004; Seger et al., 2010), the considerations raised in this study apply.

Understanding the effects of linked selection also affects our interpretation of these estimators. The number of adaptive substitutions cannot be directly related to either the strength of selection or the supply of beneficial mutations, but it is a combination of both of these factors. For example, Strelkowa and Lässig (2012) and Koelle et al. (2006) raised alternative hypotheses concerning whether periodic positive sweeps in human influenza were due to a limiting supply of beneficial mutations, or by a high supply rate with competition between beneficial mutations limiting the fixation rate.

The selective regime is important, then, for both the application and interpretation of these estimators. We present three statistics for evaluating the effects of linked selection. D_1 signals the presence of weak deleterious mutations that can potentially cause hitch-hiking and is based on identifying an excess of non-synonymous low-frequency polymorphisms. More sophisticated methods to characterise the DFE have been used (Nielsen and Yang, 2003; Piganeau and Eyre-Walker, 2003), but these methods rely on a number of assumptions and have given conflicting results. Consequently, it is useful to have a simple diagnostic that flags when hitch-hiking might be an issue. We have not attempted to use standard site-frequency based indicators of hitch-hiking (e.g., Fay and Wu (2000)) which test for an excess of low and high frequency polymorphisms. As demonstrated by Kim (2006), the excess of high-frequency polymorphisms may not occur in recurrent sweeps. In addition, this effect can be complicated by clonal interference. If no comparative information is available, the excess of low frequency polymorphisms cannot be distinguished from a model of population growth (Hahn et al., 2002).

Our second diagnostic, D_2 detects an excess of high-frequency non-synonymous polymorphisms signalling strong linkage effects on the population, either due to a large number of deleterious mutations causing background selection, or a large number of beneficial mutations causing clonal interference. In both cases, MK statistics are biased and estimators assuming independently segregating sites must be interpreted with care. We can distinguish between the effects of background selection and clonal interference by using a third statistic, D_3. The diagnostic for clonal interference presented here follow a similar reasoning to the method used by Strelkowa and Lässig (2012) in testing for an excess of high-frequency non-synonymous polymorphisms. Our method has more general applicability as it accounts for the effect of deleterious mutations, and can be used even when it is not known which region of the sequence is positively selected.

We have applied these diagnostics to the dataset used by Strelkowa and Lässig (2012). Our results provide further support for their conclusion that clonal interference occurs in human influenza A. The authors also raised the question of whether strong and frequent positive selection would promote the fixation of deleterious mutations, increasing the brittleness of the protein. The values obtained for D_1, however, suggest that strong negative selection is predominant, so that hitch-hiking in the HA1 region is rare, which is in agreement with Shih et al. (2007), who showed that few non-synonymous substitutions occurred outside antigenic epitopes. It is also consistent with a phylogenetic study by Nielsen and Yang (2003) that estimated the DFE shape parameter β in that region as 0.373, indicative of low sensitivity to hitch-hiking in our model. The combination of clonal interference and robustness against hitch-hiking suggests that the modification used by Bhatt et al. (2011) is appropriate for application to the HA1 region.

In this study, we have not considered the effect of population size changes, which are known to affect site-frequency based statistics. However, we expect D_1 and D_3 to be relatively robust, as they are based on comparisons between the non-synonymous and synonymous site frequency spectra (Messer and Petrov, 2013). Population expansions, which are expected to have the strongest effect on low-frequency mutations (Maruyama and Fuerst, 1985), should have minimal effect on D_2 and D_3. Population bottlenecks, however, can remove medium frequency polymorphisms, potentially elevating the magnitude of both D_2 and D_3 to give false positives for clonal interference. We have also not examined the effect of selection against synonymous mutations. We expect, however, that D_1 and D_3 should not be strongly affected as long as selection against synonymous mutations is weaker than against non-synonymous mutations. D_2 does not use information from the synonymous site frequency spectra and should not be affected by selection against synonymous mutations, but negative values of D_2 may also result from background selection at synonymous sites. These effects could be considered in more detail in future simulation studies.

2.5 Conclusions

We have shown that linked selection is responsible for biases in the MK statistics, causing underestimation when there are high levels of interference between selected mutations, and overestimation for strong non-interfering sweeps. The statistics presented in this study can be applied to samples of protein-coding sequences to evaluate the influence of linked selection, for the parameter range studied here. Values of D_1 that are significantly greater than zero signal susceptibility to hitch-hiking while values of D_2 significantly smaller than zero indicate the occurrence of multiple mergers. Multiple mergers due to clonal interference can be distinguished from background selection when $D_3 > 0$.

Based on our simulations, when $D_2 < 0$, $D_3 > 0$ and $D_1 \approx 0$, we recommend using a statistic such as $\hat{a}_B$, which excludes both low- and high-frequency polymorphisms. On the the hand, when high values of D_1 (signalling hitch-hiking) are obtained, it is more appropriate to use $\hat{a}_F$ which uses medium and high-frequency polymorphisms, accounting for change in selective constraint due to hitch-hiking. In cases where $D_2 \ll 0$ and $D_3 <= 0$, both $\hat{a}_F$ and $\hat{a}_B$ are expected to perform poorly.

Chapter 3

Modelling the role of immunity in reversion of viral antigenic sites

3.1 Introduction

Viral evolution is shaped by both epidemiological effects on population dynamics, and molecular effects of mutations in the viral genome (Grenfell et al., 2004). The combination of these effects generates distinctive dynamics at antigenic sites of viral proteins, which are the targets of host immune recognition. Selection for strains carrying antigenic changes that evade immune recognition result in elevated rates of non-synonymous substitution. It is unclear, however, why different dynamics of forward or reverse substitution are observed. Antigenic reversion has been reported frequently in viruses such as HIV (Delport et al., 2008; Fryer et al., 2010; Leslie et al., 2004), respiratory syncytial virus (RSV) (Botosso et al., 2009) and hepatitis C (Bailey et al., 2012; Irausquin and Hughes, 2008), and less frequently in other viruses such as influenza (Wagner, 2014; Wikramaratna et al., 2013), parvovirus (Parrish et al., 1991), hepatitis A (Lemon et al., 1990) and polio (Ketterlinus et al., 1993). Various explanations for occurrence of reversion have been proposed, such as changing immunity (Botosso et al., 2009), a limited antigenic repertoire (Botosso et al., 2009; Wikramaratna et al., 2013), or constraints on function (Bailey et al., 2012; Lemon et al., 1990; Wagner, 2014), but it is not understood how the relative influence of these effects can generate differences in observed rates of reversion.

The difficulty in evaluating the contribution of selective mechanisms is due to the lack of methods that model both epidemiological and molecular dynamics. Phylodynamic approaches (Pybus and Rambaut, 2009) incorporating epidemiological models into a coalescent framework have provided insight into the origins and spread of novel pathogens. However, they assume that molecular changes do not affect epidemiological dynamics, and are uninformative about selection. In contrast, codon-based approaches (Nielsen and Yang, 2003; Tamuri et al., 2012) aim to identify sites that contribute to the adaptation of a virus, but they assume that the population size is constant and that the selection coefficient is constant at each site. Various modifications of the substitution model allow for different selective effects based on directionality or target residue (Delport et al., 2008; Kosakovsky Pond et al., 2008), but retain the assumption that substitution occurs as a time-homogeneous process which is not affected by population dynamics. To understand how the probability of reversion at antigenic sites is affected by both selective constraint against molecular changes and selection to evade immune recognition, there is a need to incorporate the time-dependence imposed by epidemiological dynamics into the substitution process.

Models of pathogen dynamics have shown that reversion probabilities are affected by fitness costs (Fryer et al., 2010; Kent et al., 2005; Petravic et al., 2008; Silva, 2012), at both the within-host and between-host level, and the availability of susceptible hosts (Fryer et al., 2010), at the between-host level. However, these models were developed in the context of HIV escape mutations. HIV infects host chronically, with host susceptibility determined by human leukocyte antigen (HLA) type, which does not vary over time. Due to these infection dynamics the prevalence of each strain changes relatively slowly, and is expected to eventually stabilise (Fryer et al., 2010). In contrast, for acute infections such as human influenza and RSV where transmission occurs frequently and host immunity can last for much longer than the duration of the infection, the structure of host immunity can vary rapidly over time. Due to differences in the dynamics of selection, we expect antigenic selection to have qualitatively different effects on sequence changes at antigenic sites compared to constant selective pressure (Grenfell et al., 2004).

Here, we examine the probability of antigenic reversion in an epidemiological model, which describes the complex ecology of multiple viral strains with cross-immunity competing for susceptible hosts. This model allows us to quantify the relative advantage of an antigenically novel mutation, compared to a reversion which may be antigenically less advantageous, but improves transmission. Using both a simple three-strain model and simulations with multiple codon sites, we examine the effect of the duration of host immunity, selective costs, population size, and the basic reproductive ratio. We show that these effects lead to distinctive dynamics in the frequencies of derived amino acids, which is informative about the duration of host immunity and strength of selective constraint. Time-structured sequence data from influenza and RSV are compared to simulated sequences, and we discuss what these results imply about the relative effects of host immunity and functional constraint.

3.2 Methods

3.2.1 Simple analytical model for antigenic reversion

The simplest model containing reversion is a system where the population has mutated away from the ancestral state, and potentially can mutate either back to the ancestral state (reversion) or to a novel state (forward substitution). In an epidemiological context, we describe the substitution process in terms of competition between three strains of virus which are related through mutation. The viral population is initially of strain 0 (ancestral state), which is then replaced by strain 1, and can subsequently be replaced by either strain 0 (reversion) or strain 2 (forward substitution).

The dynamics of the viral population can be described at the population level using a three-strain SIRS model (Hethcote, 2000). This model accounts for viral competition for available hosts by tracking the number of hosts which are susceptible S_i, infected I_i, and recovered with immunity R_i to strains $i = 0, 1, 2$. Assuming a large host population of constant size N with homogeneous mixing, the epidemiological dynamics can be described by

$$\frac{dI_i}{dt} = \beta_i \frac{S_i}{N} I_i - \delta I_i, \tag{3.1}$$

$$\frac{dR_i}{dt} = \delta I_i - \gamma R_i, \tag{3.2}$$

with transmission rate β_i, recovery rate δ, and immunity that decays at rate γ. Interactions between strains are described by the implicitly defined term S_i, which is the number of hosts susceptible to strain i. Assuming that each host can only be infected by a single strain at a time, and prior infection with strain j reduces susceptibility to strain i by a factor σ_{ij}, the relationship between susceptible and immune hosts is given by

$$S_i = N - \sum_j I_j - \sum_j \sigma_{ij} R_j \, , \tag{3.3}$$

with the constraint that $S_i > 0$ for any strain i. All uninfected hosts ($N - \sum_j I_j$) can be categorised as either susceptible (S_i) or immune ($\sum_j \sigma_{ij} R_j$) to strain i. The similarity between this model and the status-based model with polarised immunity developed by Gog and Grenfell (2002) becomes evident when we differentiate Equation (3.3) to give

$$\begin{aligned} \frac{dS_i}{dt} &= -\sum_j \left(\frac{dI_j}{dt} + \sigma_{ij} \frac{dR_j}{dt} \right) , \\ &= -\sum_j \beta_j \frac{S_j}{N} I_j + \sum_j \delta(1-\sigma_{ij}) I_j + \sum_j \gamma \sigma_{ij} R_j \, . \end{aligned} \tag{3.4}$$

The main difference is that we retain the history of infections accumulated across the population through the additional set of variables, R_i. This allows us to obtain analytical expressions for the number of hosts susceptible to all strains as functions of the same set of variables, as shown in Equation (3.3). In contrast to the Gog and Grenfell (2002) model assuming polarised immunity, we assume a model of partial additive immunity. That is, we do not need to keep track of which hosts have previously been infected with both strain i and strain j, because these hosts simply contribute $\sigma_{ij} + \sigma_{ii}$ immunity against strain i. This approximation overestimates the level of immunity (underestimates S_i) if many hosts are infected with multiple strains, since the level of immunity in any one host should not be greater than 1.

Our model of partial additive immunity generates similar dynamics to the Gog and Grenfell (2002) model. From Equation (3.4), it can seen that hosts infected with strain j are removed from the susceptible class S_i, and then a proportion $1 - \sigma_{ij}$ of all infected hosts are returned to the susceptible class on recovery, so that the overall contribution of immunity is $\sigma_{ij}\beta_j S_j I_j / N$, which is similar to the $\sigma_{ij}\beta_j S_i I_j / N$ term in the Gog and Grenfell (2002) model. The difference in the S_i and S_j term arises because in the Gog and Grenfell (2002) model, immunity arises from exposure, but in our model, immunity is only generated when infection occurs.

The strict exclusion of co-infection involves a second approximation, where an infection by any strain j will always be removed from S_i but not from R_i [first term in Equation (3.4)]. This occurs because while it is possible to distinguish between S_i and R_i at the time of infection from strain j, it is not possible, at the time of recovery from strain j, to determine whether the host was previously susceptible or immune to strain i. Our approximation leads to an underestimation of S_i. We expect this to have a small effect as the bias lasts only for the duration of the infection. In addition, strains which are closest to the current circulating strain j will not be heavily affected ($\sigma_{ij} \approx 1$); the most heavily affected strains are those distant from strain j which are likely to be no longer circulating.

Using this model, we examine the effect of cross-immunity σ_{ij}, immunity duration γ and selective costs incurred by antigenic escape s. The rate of immune decay γ includes the loss of immunity by the death and migration of immune hosts as well as the loss of immunity in individual hosts. The selective cost is parametrized through a reduction in the strain-specific transmission rate so that $\beta_0 = \beta$, $\beta_1 = \beta(1-s)$ and $\beta_2 = \beta(1-s)^2$. To understand the effect of these parameters, we first characterise the number of susceptible hosts to each strain at equilibrium, and use this to determine probabilities of fixation, assuming a single strain appears at a time.

We assume the population is initially infected with only strain 0, which is maintained at equilibrium until strain 1 emerges at time t_1. Strain 1, then replaces strain 0 and equilibrates until time t_2, when a third strain (either strain 0 or strain 2) emerges and can potentially replace strain 1. These equilibrium assumptions allow us to characterise host immunity accumulated due to infection by strain 0 at t_1 (denoted R_0^*), and host immunity accumulated due to infection by strain 1 at t_2 (denoted R_1^*), which then allows us to evaluate the probability of strain 0 or 2 emerging at time t_2.

The equilibrium is obtained by setting the derivative of S_i and I_i to zero. When the viral population consists of only one strain, the endemic equilibrium, which is asymptotically, locally stable when the basic reproductive ratio $\beta_i/\delta > 1$ (Hethcote, 2000), is given by

$$\begin{aligned} S_i^* &= \frac{\delta}{\beta_i} N, && (3.5) \\ I_i^* &= \frac{\gamma N}{\delta\sigma_{ii} + \gamma}\left(1 - \frac{\delta}{\beta_i}\right). && (3.6) \end{aligned}$$

We assume that at time t_1, when strain 1 emerges, the population remains close to equilibrium. As strain 1 has only just emerged and strain 2 has not yet occurred, the cross-immunity terms in Equation (3.3) can be ignored so that it contains only terms of subscript $i = 0$. Substitution of Equations (3.5) and (3.6) into Equation (3.3) gives

$$R_0^* = \frac{\delta N}{\delta\sigma_{00} + \gamma}\left(1 - \frac{\delta}{\beta_0}\right). \qquad (3.7)$$

Now, consider a later time t_2, when a third strain (either 0 or 2) emerges and can potentially replace strain 1. Again, we assume that strain 1 remains close to equilibrium and that the third strain has had negligible effect on immunity. In addition, we assume that immunity due to infection by strain 0 has decayed exponentially since time t_1, so that Equation (3.3) can be approximated as

$$S_1^* = N - I_1^* - \sigma_{11} R_1^* - \sigma_{10} R_0^* e^{-\gamma(t_2 - t_1)}. \qquad (3.8)$$

Substituting Equations (3.5) and (3.6) into (3.8) then gives

$$R_1^* = \frac{\delta N}{\delta\sigma_{11} + \gamma}\left(1 - \frac{\delta}{\beta_1}\right) - \frac{\delta\sigma_{10} N}{\sigma_{11}(\delta\sigma_{00} + \gamma)}\left(1 - \frac{\delta}{\beta_0}\right) e^{-\gamma(t_2 - t_1)}. \qquad (3.9)$$

Having obtained an expression for R_0^* and R_1^*, we can now compute the proportion of hosts that are susceptible to each strain, $p_i(\tau) = S_i(\tau)/N$, where $\tau = t_2 - t_1$ is the time since the emergence of strain 1. Thus,

$$\begin{aligned} p_0(\tau) &= 1 - \frac{I_1^*}{N} - \frac{\sigma_{01}}{N} R_1^* - \frac{\sigma_{00}}{N} R_0^* e^{-\gamma\tau}, && (3.10) \\ p_2(\tau) &= 1 - \frac{I_1^*}{N} - \frac{\sigma_{21}}{N} R_1^* - \frac{\sigma_{20}}{N} R_0^* e^{-\gamma\tau}, && (3.11) \end{aligned}$$

which can be written in the form

$$p_i(\tau) \;=\; A + B_i e^{-\gamma\tau}, \text{ for } i = 0, 2. \tag{3.12}$$

Assuming that cross-immunity is additive with respect to the number of antigenic differences ($\sigma_{ii} = \sigma$, $\sigma_{01} = \sigma_{10} = \sigma_{21} = \sigma/2$ and $\sigma_{20} = 0$), the coefficients simplify to

$$\begin{aligned} A &= 1 - \frac{\delta\sigma + 2\gamma}{2(\delta\sigma + \gamma)} \left(1 - \frac{\delta}{\beta_1}\right), & (3.13) \\ B_0 &= -\frac{3\delta\sigma}{4(\delta\sigma + \gamma)} \left(1 - \frac{\delta}{\beta_0}\right), & (3.14) \\ B_2 &= \frac{\delta\sigma}{4(\delta\sigma + \gamma)} \left(1 - \frac{\delta}{\beta_0}\right). & (3.15) \end{aligned}$$

Note that we expect that prior immunity reduces infection against an unmutated strain at appreciable levels ($\sigma \gg 0.1$) and that immunity lasts for much longer than the infection duration ($\gamma \ll \delta$). Within the parameter range of interest, the fractional terms containing δ, σ and γ in Equations (3.13–3.15) approach constants, so that A is approximately a function of only β_1/δ and B_0 and B_2 are approximately functions of only β_0/δ.

We calculate the probability of a strain generated by reversion or forward mutation at time t_2 giving rise to a new epidemic by approximating the emergence of a new strain as a linear birth-death process. Ignoring initial changes in host susceptibility, the probability that a new strain reaches fixation (Keeling and Rohani, 2008) is given by

$$f_i = \begin{cases} 1 - \frac{1}{r_{e,i}}, & \text{if } r_{e,i} > 1 \\ 0, & \text{otherwise} \end{cases} \tag{3.16}$$

where $r_{e,i} = \beta_i p_i/\delta$ denotes the effective reproductive ratio of the new strain i at the time of emergence. Using Equations (3.12–3.15), at time τ after strain 1 has reached equilibrium, we compute the probability of fixation for strain 0 (reversion) and strain 2 (forward substitution) to be

$$f_i(\tau) = \begin{cases} 1 - \frac{\delta}{\beta_i(A + B_i e^{-\gamma\tau})}, & \text{if } \tau > t_c \\ 0, & \text{otherwise} \end{cases} \tag{3.17}$$

Table 3.1: Table of parameters used in the multi-site simulation model.

Parameter	Description
β	Transmission rate per time-step
δ	Recovery rate per time-step
γ	Decay rate of host immunity per time-step
σ	Strength of immune protection
μ	Mutation rate per site per time-step
s	Cost of immune escape
L_a	Number of antigenic sites
N	Host population size

where the threshold t_c (if considering strain 0) is given by

$$t_c = -\frac{1}{\gamma}\log\left(\frac{\delta}{\beta_0 B_0} - \frac{A}{B_0}\right). \tag{3.18}$$

The probability of reversion given fixation is therefore

$$\rho(\tau) = \frac{f_0(\tau)}{f_0(\tau) + f_2(\tau)}. \tag{3.19}$$

Asymptotically, if all prior immunity against strain 0 has decayed, then the exponential term in the denominator of Equation (3.17) approaches zero, thus giving

$$\begin{aligned} \rho_\infty &= \frac{\beta_0 A - \delta}{2\beta_0 A - \delta(1 + \frac{\beta_0}{\beta_2})} && (3.20)\\ &= \frac{\frac{1}{2}\left[\frac{\beta}{\delta} - (1-s)^{-1}\right] - 1}{\frac{\beta}{\delta} - (1-s)^{-1} - (1-s)^{-2} - 1}. && (3.21) \end{aligned}$$

In summary, Equation (3.19) describes the combined effect of immunity γ and functional constraint s on the probability of reversion at some time τ after immunity has begun to wane from equilibrium levels, whereas the long-term asymptote ρ_∞ shows the effect of functional constraint in the absence of immunity.

3.2.2 Multi-site simulation model

To verify our theoretical model, and to examine the impact of increasing the antigenic space, we develop a stochastic computer simulation model where each infection is associated with a sequence of antigenic sites. Population dynamics are similar to the analytical model (see Table 3.1 for a complete list of parameters), but in the multi-site simulation, we explicitly model the mutation process. In the analytical model, we assumed the emergence of three strains at specified times, and calculated the probability that these strains would reach fixation. In contrast, for the simulation model, we allow mutations to occur stochastically at any antigenic site throughout the simulation; thus, new strains may emerge before the old strain reaches equilibrium and even favourable mutations may be lost due to stochasticity.

We implement two models using different representations of the antigenic space. The first model uses a bit-string representation so that each of the L_a antigenic sites can take values of $\mathbf{v} = \{0, 1\}$, and a change at any site away from the ancestral state (0) will reduce transmissibility. The bit-string model with two sites has a antigenic space similar to the analytical model. In the second model, we use a more realistic codon representation. Sites can mutate to any one of the 64 possible codons, but viral fitness is only affected by non-synonymous changes (i.e., $\mathbf{v}$ consists of the 20 amino acids). Specifically, any amino-acid change will affect cross-immunity, but only changes from the ancestral amino acid to a derived state will reduce transmissibility. The ancestral codon sequence is determined at the beginning of each simulation by randomly sampling L_a non-terminating codons with uniform probability.

Throughout the simulation, we track the number of infected hosts I, the genotype of each infection, and the immune status of the host population. The last variable is stored in the immunity matrix consisting of $2 \times L_a$ elements for the bit-string model, or $20 \times L_a$ for the codon model, where each element $r_{v,j}$ stores the number of people with immunity to a value of v at site j. That is, $r_{v,j}$ stores the site-specific immunity accumulated across the whole population, and we compute the immunity against any viral genotype by summing across these values (described below).

The multi-site model is implemented as a discrete time simulation (Keeling and Rohani, 2008), with a time-step of one day. The system is initialised with a naive population ($r_{v,j} = 0$ for all v and j) and an infected host which carries the ancestral strain. At each time-step, the population changes according to SIRS dynamics, with the following events occurring:

1. Mutation: The number of mutations that occur in the viral population in each time-

step is drawn from a Poisson distribution with mean $\mu I L_a$, where μ is the mutation rate per site per time-step, and occur uniformly across all sites and all individuals. For the codon model, the probability of any codon occurring at the mutated site is specified by the Kimura (1980) two-parameter model with a transition-transversion rate of $\kappa = 3$.

2. Transmission: The number of potential new infections which occur in each time-step is a Poisson random variable $X \sim \text{Pois}(\Lambda)$, where $\Lambda = \sum_{i=1}^{I} \beta(1-s)^{k_i}$ is the force of infection. The scaling factors $(1-s)^{k_i}$ account for the reduction in transmission of genotype i due to the cost of k_i changes away from the ancestral strain. The genotypes of the X potential infections are determined by multinomial sampling according to $(1-s)^{k_i}$, to account for variation in transmissibility within the viral population. We can then calculate the probability of each potential infection i encountering a susceptible host, given by

$$p_i = \frac{N - I - \frac{\sigma}{L_a} \sum_{j=1}^{L_a} r_{v_{ij},j}}{N}, \tag{3.22}$$

where $r_{v_{ij},j}$ is the level of recognition against a particular antigenic site as described above. Equation (3.22) corresponds to Equations (3.10–3.11) in the analytical model. The success of the potential infection is determined using a Bernoulli random variable $U \sim \text{Bernoulli}(p_i)$. If $U = 1$, a new infection is generated with a genotype identical to the parent.

3. Recovery: The number of infected hosts which recover in each time-step is Poisson with mean δI. Each recovered host i is drawn from the infected population with uniform probability and increases immunity to allele v_{ij} at site $j = 1, \cdots, L_a$. That is, for each recovery, we update L_a elements of the immunity matrix

$$r_{v_{ij},j} \quad := \quad r_{v_{ij},j} + 1\,. \tag{3.23}$$

4. Decay of host immunity (across the whole population) is simulated by reducing $r_{v,j}$ for all antigenic states $v \in \mathbf{v}$ at each site $j = 1, \cdots, L_a$ by a binomial random

variable,

$$r_{v,j} \quad := \quad r_{v,j} - V,\ \text{where } V \sim \text{Binom}(r_{v,j}, \gamma)\,. \tag{3.24}$$

3.3 Results

Using the analytical and simulation models, we examine how the epidemiology of the virus affects the probability of reversion at antigenic sites. We first describe the dynamics of the simple three-strain model (Section 3.3.1), before examining the time dependence of this system (Section 3.3.2) and the effect of the epidemiological parameters (Section 3.3.3). The combined effect of these interacting factors on the observed amino acid frequencies is described in Section 3.3.4, and we compare this to sequence data for human influenza A (H3N2) and RSV-A in Section 3.3.5.

3.3.1 Dynamics of changing susceptibility

To provide some intuition about the process, we show an example of forward substitution and reversion in the three-strain model using the two-site bit-string model (Figure 3.1). In both simulations, the ancestral strain 0 is introduced into an initially naive population, and is then replaced by strain 1, which is subsequently replaced by a third strain, either strain 2 [forward substitution; panel (a)] or strain 0 [reversion; panel (b)]. For each simulation, we also show the corresponding proportion of susceptible hosts p_i [panels (c) and (d)], for each strain $i = 0, 1, 2$. The emergence of the ancestral strain 0 in the initially naive population sharply reduces the proportion of susceptible hosts to strain 0, p_0; p_1 is also slightly reduced due to cross-immunity between strains 0 and 1, while p_2 is unaffected. When strain 1 emerges and dominates the population, both p_0 and p_2 are temporarily reduced but p_0 slowly increases above its previous equilibrium.

In the first simulation [panels (a) and (c)], strain 1 is rapidly replaced with strain 2, so that at the time of emergence t_2, susceptibility to strain 0 remains quite low [black line in panel 1(c)]. In this case, forward substitution is favoured because there is a larger pool of susceptible hosts for strain 2. In contrast, in panels (b) and (d), the interval between t_1 and t_2 (vertical grey lines) is longer than the first simulation, providing time for p_0 to reach similar levels to p_2 so that reversion can occur.

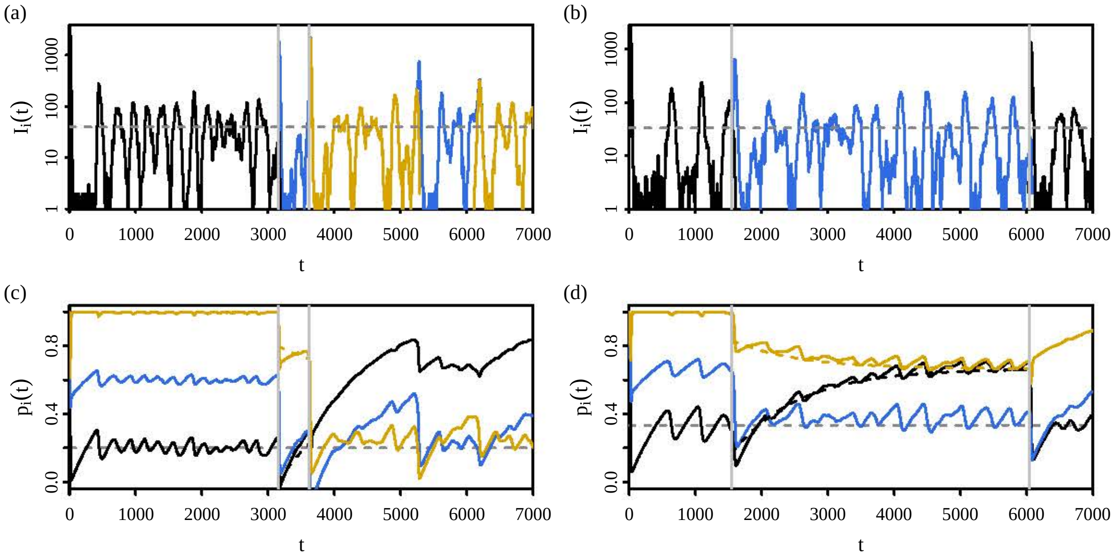

Figure 3.1: An example of forward substitution [panels (a) and (c)] and reversion [panels (b) and (d)] in the two site bit-string epidemiological model. Solid lines show the trajectory from a single simulation of the number of hosts [panels (a) and (b)] infected by, and the proportion of hosts susceptible [panels (c) and (d)] to strain 0 (black), strain 1 (blue) and strain 2 (orange). Simulations are initialised with a small number of hosts infected with strain 0 which tend towards the equilibrium [horizontal grey dashed lines; Equations (3.5–3.6)]. At time t_1, strain 1 emerges and dominates the population until time t_2 when a third strain (either strain 0 or 2) emerges. Times t_1 and t_2 are indicated by vertical grey lines. Between t_2 and t_1, the expected proportion of susceptible hosts [Equations (3.12–3.15)] is shown by dashed lines. Simulations were run with parameters (a) $\beta = 1.0$ day^{-1} and (b) $\beta = 0.6$ day^{-1} and in both panels, $N = 10^4$, $\gamma = 10^{-3}$ day^{-1} and $s = 0.1$.

3.3.2 Time-dependence of the probability of reversion

In Figure 3.2, we show the probability of reversion as a function of $\tau = t_2 - t_1$, the interval between the time of strain emergence (indicated by vertical grey lines in Figure 3.1). The theoretical probability of reversion [Equation (3.19)] is compared to the proportion of reversion events in simulations with a two-site bit-string model. To correspond to the analytical model, only substitution events following transitions between strain 0 to strain 1 are counted. Note that in the analytical model, τ is the interval between the times of emergence; however, in the simulation, it is difficult to determine which of the emerging mutations will reach fixation. As a proxy for τ, the counts from the simulation are binned according to the time between antigenic substitutions.

These results confirm that the reversion probability varies with τ. The probability of reversion is low if substitution occurs rapidly, and gradually increases with τ until it flattens at the asymptote ρ_∞, given by Equation (3.21). This asymptotic value represents the probability of reversion in the absence of cross-immunity. The decay rate of host immunity γ affects the speed at which the asymptotic value is reached, but not the value of the asymptote.

The greatest discrepancy between theoretical and simulated results occurs near the transition t_c [Equation (3.18)]. At $\tau = t_c$, the theoretical model predicts a sharp transition away from $\rho(\tau) = 0$; in the stochastic simulations, the transition is more gradual. The reason for this discrepancy is that the theoretical model assumes that each strain reaches equilibrium before it is replaced. However, in large viral populations, the mutational input rate can be large enough that strain 1 replaces strain 0 before I_0 can reach equilibrium. In these cases, R_0^* will be upwardly biased, so that $\rho(\tau)$ underestimates the probability of reversion.

Based on the form of $\rho(\tau)$, we expect the time-dependent probability to be independent of the viral mutation rate and population size. Consistent with this, we observe that simulation results for different population sizes lie on the same curve, with points from small populations (circles) corresponding to large values of τ and points from larger populations (triangles) corresponding to smaller values of τ.

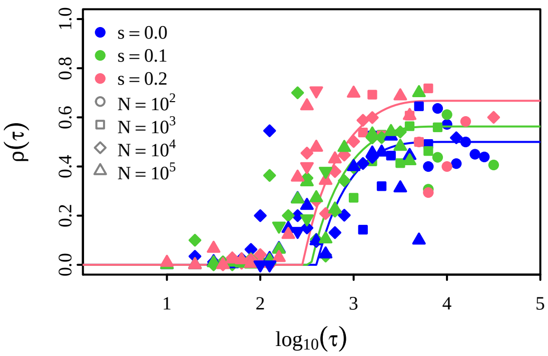

Figure 3.2: The probability of reversion, as a function of the time between strain emergence τ. The blue line shows the reversion probability [Equation (3.19)] of an unconstrained antigenic site as immunity decays, whereas the green ($s = 0.1$) and red ($s = 0.2$) lines show the combined effect of selective cost and immunity. Points show the proportion of reversion events observed from simulations of the two-site bit-string model. The proportion was computed from the binned number of substitution events that occurred immediately after a transition from strain 0 to strain 1, using the time between antigenic substitutions as a proxy for τ. Simulations were run for 5×10^5 time-steps, with a time-step of one day, with 200 replicates for each parameter combination of s and N. All other parameters were set to immune decay: $\gamma = 10^{-3}$ day^{-1}, mutation rate: $\mu = 10^{-5}$ site^{-1} day^{-1}, recovery rate: $\delta = 0.2$ day^{-1} and transmission rate: $\beta = 0.6$ day^{-1}.

3.3.3 The effect of epidemiological parameters

To examine the effects of viral transmission (β, δ, s) and host immunity (γ, σ), we now consider ρ for a fixed τ in the analytical model [Equation (3.19)]. For simplicity of notation, we omit the argument τ in this section. Equations (3.13–3.15) indicate that the strength of immune protection σ affects ρ only through the coefficients A, B_0, B_2, and is expected to have only a weak effect. In Figure 3.3, we confirm that the level of immune protection σ has only a weak effect on ρ unless the typical duration of the infection $1/\delta$ [Figure 3.3(a)] is as long as the immune duration $1/\gamma$ [Figure 3.3(b)], or σ is negligibly small. Throughout the rest of the paper, we set $\sigma = 1.0$.

Figure 3.4(a) shows how the reversion probability varies as a function of the basic reproductive ratio β/δ, for various values of selective cost s, for a fixed level of host immunity ($\gamma\tau$). For sites under no selective constraint (black line), the probability of reversion increases slightly with β/δ, but very different effects are observed for a non-zero selective cost. The effect of a selective cost is strongest for small transmission rates, as slight decreases in infection rates can have a more detrimental impact on the mutant subpopulation.

The interaction between the selective cost s, and the immunity decay rate γ, is shown

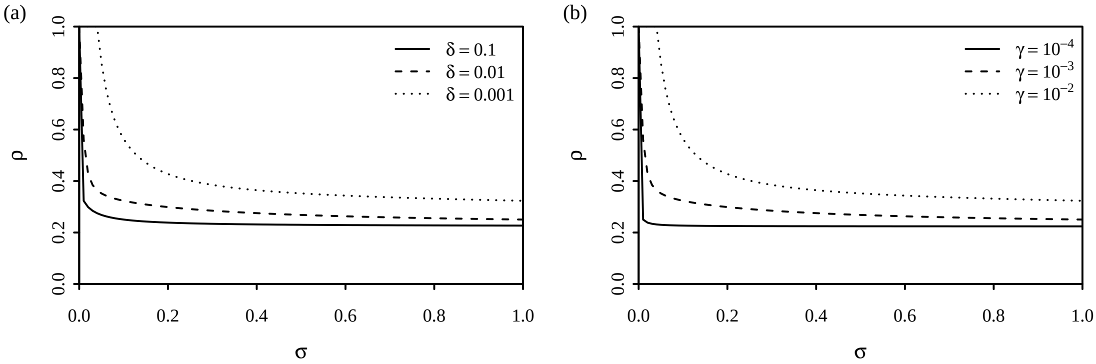

Figure 3.3: A comparison of the sensitivity of the reversion probability [Equation (3.19)] to the strength of immunity σ for different (a) rates of recovery δ and (b) rates of immunity decay γ. Unless otherwise specified, parameters were set to $\gamma = 10^{-3}$ day^{-1}, $\delta = 0.1$ day^{-1}, $\beta/\delta = 5$, $\gamma\tau = 0.5$ and $\mu = 10^{-5}$ site^{-1} day^{-1}.

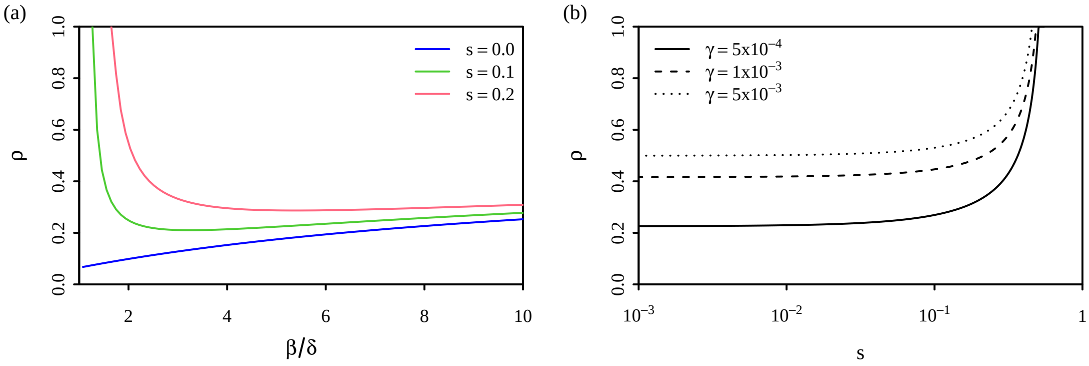

Figure 3.4: The effect of the cost of immune escape s and decay rate of immunity γ on the probability of reversion [Equation (3.19)]. In (a), we hold the level of immunity ($\gamma\tau = 0.5$) constant to show how varying basic reproductive ratio β/δ changes the effect of s. In (b), we show the effect of varying s for different values of γ with a fixed time between strain emergence $\tau = 3 \times 365$ days and $\beta/\delta = 5$. Other parameters were set to $\delta = 0.2$ day^{-1} and $\mu = 10^{-5}$ site^{-1} day^{-1}.

for a fixed τ [Figure 3.4(b)]. We showed in Figure 3.2 that for large $\gamma\tau$, $\rho(\tau)$ plateaus at ρ_∞, which is independent of γ; however, when the rate of strain replacement is comparable to the decay rate of host immunity, there are strong dependencies. The effect of varying γ, in the absence of selective constraint ($s \approx 0$), can be seen in the difference between ρ where the curves plateau. Further increases in selective cost leads to a rapid increase in the probability of reversion, with more rapid increases for longer lasting immunity (solid line).

3.3.4 Fluctuating frequencies at antigenic sites

In Sections 3.3.1–3.3.3, we observed that τ had a strong effect on whether reversions occur or not. In fact, where τ is known, no further information on mutation rate μ or population size N is required. However, in practice this quantity is difficult to measure. It is possible to account for variation in τ by integrating over the distribution of τ, but this can remove important information; under certain parameter ranges, the stochasticity of τ is sufficient to cause noticeable variation in reversion probabilities.

To observe the effect of fluctuations in ρ, we measure the frequency of the ancestral allele π_0 at each antigenic site. The frequency of an allele is informative about its fixation probability (Kimura, 1962), and the rate of change in frequency is proportional to the strength of selection s (Kent et al., 2005; Zhao et al., 2013). Under directional selection, we expect any allele to eventually reach fixation or extinction. Thus fluctuations between $\pi_0 = 0$ to $\pi_0 = 1$ indicates changes in selection. We measure the frequencies of each antigenic site separately, as immunity against each site may vary depending on the history of previous circulating strains.

In Figure 3.5, we show frequency trajectories π_0, under conditions of both antigenic selection and selective constraint, so that antigenic changes away from the ancestral sequence imposes a cost. To account for inaccuracies due to sampling, π_0 was computed from sequences sampled at discrete intervals, and the earliest sequence sampled after the burn-in period was used as the ancestral sequence. In all panels, we observe fluctuations in frequency levels as reversion probabilities vary due to the stochasticity of the time between antigenic substitutions, although there is no change in μ, N, or s during a simulation. The pattern of fluctuations in π_0 differs depending on the host population size N (varying along columns) or the decay rate of host immunity γ (varying along rows). Faster changes in π_0 are observed for larger N and fixation of the ancestral allele becomes less likely. Tracking frequency over time also provides information on γ that would not be available in the time-averaged approach. Comparison between columns in Figure 3.5 indicates that increasing γ tends to reduce both the frequency and amplitude of π_0. This effect is particularly evident for larger population sizes [panels (c)–(f)], where the rate of substitution is not limited by the rate of mutational input.

The effect of removing the selective cost ($s = 0$) is shown in Figure 3.6. Although

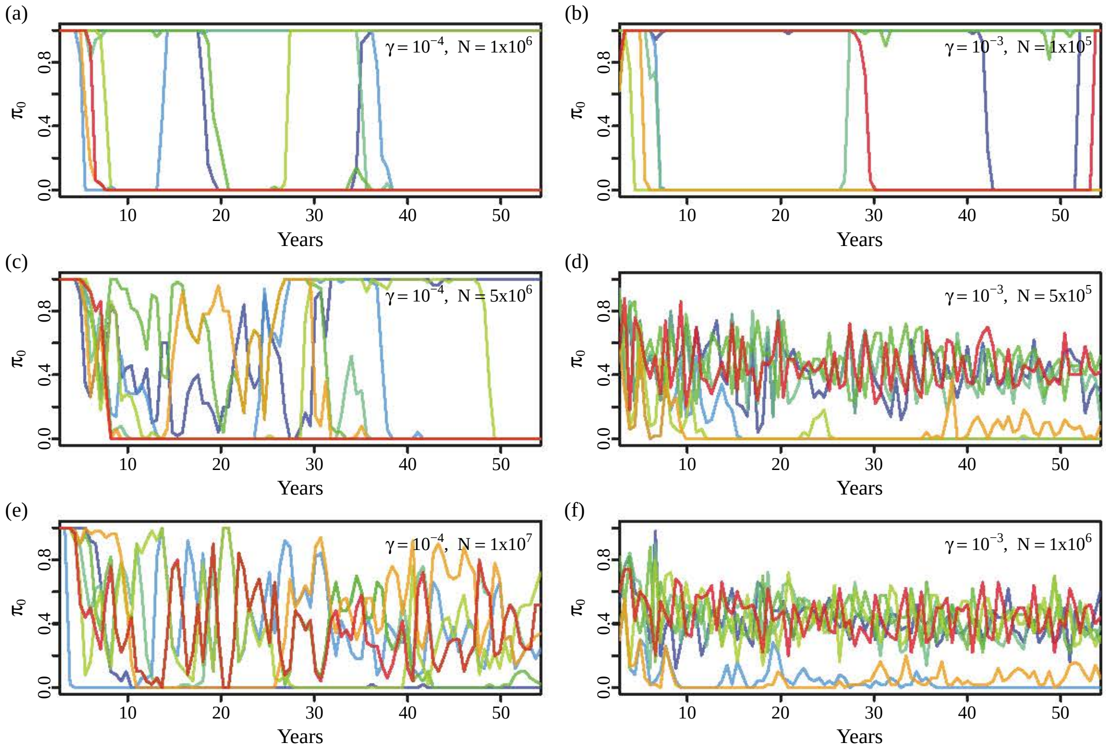

Figure 3.5: The effect of population size and duration of immunity on the frequency of the ancestral allele π_0 at antigenic sites under selective constraint ($s = 0.2$). Each line represents the changing frequencies, at a single antigenic site, of the ancestral allele estimated to be the earliest sampled amino acid residue after the burn-in period (1000 days). For high rates of mutational input [panels (d) and (f)], the earliest sequence may not be the true ancestral sequence (set to be the most transmissible), which in some cases results in low observed values of π_0. Each panel represents the dynamics of a single simulation, with π_0 computed from samples of 20 sequences taken every 200 days. All simulations were run with a time-step of one day and parameters $\beta = 1.0$ day^{-1}, $\delta = 0.2$ day^{-1}, $\mu = 10^{-5}$ site^{-1} day^{-1}, $L_a = 7$, to match parameters used for human influenza A (H3N2) (Koelle et al., 2009).

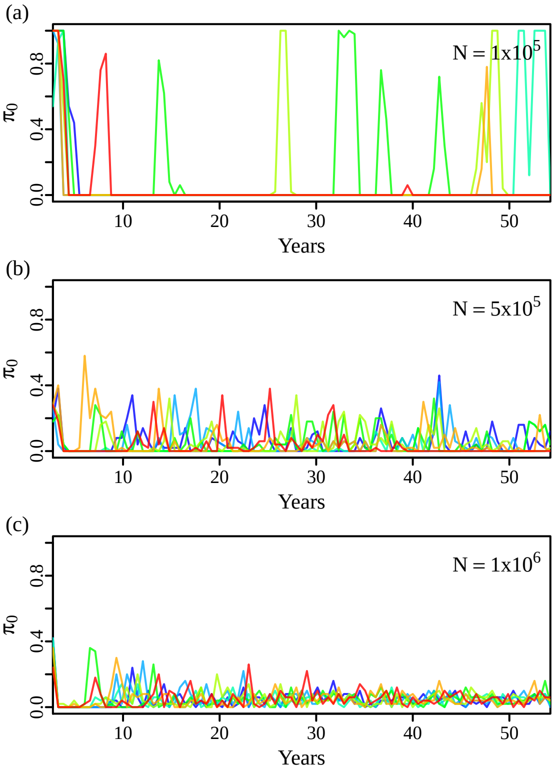

Figure 3.6: The frequency of the the ancestral allele π_0 at antigenic sites under no selective constraint ($s = 0$). Lines in each panel show the changing frequency of the ancestral (earliest sampled) allele at each antigenic site in a single simulation. π_0 was computed from samples of 20 sequences taken every 200 days, discarding all sequence data from the burn-in period of 1000 days. All simulations were run with $\gamma = 1 \times 10^{-3}$ day^{-1}, $\beta = 1.0$ day^{-1}, $\delta = 0.2$ day^{-1}, $\mu = 10^{-5}$ site^{-1} day^{-1} , and $L_a = 7$.

fluctuations can still occur, the ancestral allele at the antigenic site rarely returns to fixation ($\pi_0 = 1$) and, if so, does not remain fixed for long. This effect occurs even for small population sizes [panel (a)] which favour reversion. Continual antigenic selection drives further substitutions to other derived amino acid residues, that have not induced prior immunity. That is, multiple instances of increasing π_0 as an indication of high selective costs s is robust to misspecification of the ancestral allele. However, consistently low values of π_0 may simply be due to using an misspecified ancestral allele (an alternative interpretation is that π_0 correctly identifies that an unfavoured amino acid is unconstrained).

3.3.5 Application to influenza and RSV

In Figure 3.7, we show π_0 changing over time for the human influenza A virus subtype H3N2 and the respiratory syncytial virus (RSV) subtype A at antigenic and non-antigenic sites. The H3N2 data set consists of all HA sequences for human H3N2 from the influenza virus database (Bao et al., 2008) where the year of sampling is known. The accession numbers surface G protein sequences of RSV-A sequences that we used were listed in Botosso et al. (2009). In total, we analysed 5831 H3N2 sequence spanning 45 years and 538 RSV sequences spanning 19 years. The temporal distribution of both datasets are shown in Figure B.1.

We computed π_0 for antigenic sites which have been identified by experimental methods, as sequence-based methods are also designed to identify sites with variation in amino acid composition. For H3N2, we used the seven sites (145, 155, 156, 158, 159, 189, 193) listed in a recent study (Koel et al., 2013) which used antigenic cartography which integrates information over multiple pairs of antigen and antisera in order to evaluate overall antigenic change (Smith et al., 2004). For RSV-A, experimental studies with monoclonal antibodies have identified a large number of sites which react to different monoclonal antibodies (García et al., 1994; Martínez et al., 1997). More recent studies have used phylogenetic analysis of natural isolates to identify potential antigenic sites (Pretorius et al., 2013; Zlateva et al., 2004). Note that there is an ascertainment bias in using sites identified on the basis of frequent amino acid changes. Here, we have restricted the analysis to eight sites (225, 226, 233, 237, 244, 274, 280, 290) which were identified as reducing antigenic recognition in multiple studies (García et al., 1994; Martínez et al., 1997; Pretorius et al., 2013; Zlateva et al., 2004), with at least one being experimental (García et al., 1994; Martínez et al., 1997). Including a larger number of sites does not affect the results, but will obscure features of distinct trajectories.

For both viruses, we obtain oscillating patterns of π_0 that are consistent with our expectations for antigenic sites evolving under both immune selection and functional constraint. Non-antigenic sites [Figure 3.7(c) and (d)] generally do not exhibit these fluctuations, but some non-antigenic sites in RSV-A may experience frequency fluctuations due to linkage to antigenic sites [Figure 3.7(d)]. For further comparison, the dynamics of additional number of non-antigenic sites are shown in Figures B.2–B.3. Patterns of frequency change in H3N2 and RSV-A differ considerably from each other. H3N2 frequencies have sharper and slower oscillations, which are suggestive of both a smaller population size and longer lasting immunity. At least four antigenic sites in H3N2 revert and fix at the ancestral state which indicates very strong selective constraint. RSV-A shows more rapid oscillations, suggesting faster decaying immunity and moderate selective constraint. The relatively short time that the ancestral allele is at high frequencies suggests that selective constraint has a smaller influence than for H3N2.

3.4 Discussion

We have shown that for acute, recurrent infections, the probability of reversion at antigenic sites depends on the interaction between the cost of immune escape and the duration of host immunity. Similar to models for HIV (Kent et al., 2005), we find that a higher cost of immune escape increases the probability of antigenic reversion. The impact of the cost of immune escape on the reversion probability is greater when the basic reproductive ratio is low, as small reductions in transmissibility have a more detrimental effect. This is in agreement with a previous study on the effect of selective constraint on antigenic drift (Kucharski and Gog, 2011). In addition to these two parameters, we find that longer lasting immunity can also reduce the probability of reversion, but the precise extent of this reduction depends on the time between antigenic substitutions.

The time between antigenic substitutions, which is inversely proportional to the viral population size and mutation rate, is closely related to the rate of mutational input θ, a

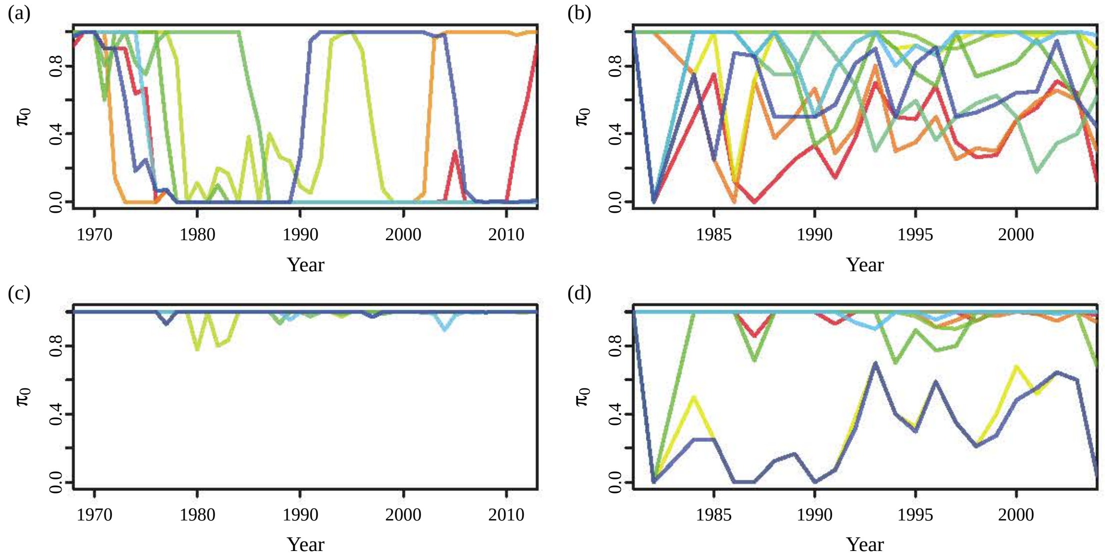

Figure 3.7: Trajectory of the frequency of the ancestral allele π_0 computed at antigenic sites of (a) human H3N2 and (b) human RSV-A show fluctuations which are distinct from randomly chosen non-antigenic sites [panels (c) and (d)]. Each coloured line shows the frequency of a single site. Frequencies were computed at (a, c) seven sites in the HA segment of H3N2 computed with A/Aichi/2/1968 as the ancestral strain, and (b, d) eight sites in the C-terminal hyper-variable region of the surface G protein of RSV-A using strain AF065406 (sampled in 1981) as the ancestral strain. Sequences were pooled according to the year of isolation, with years in which fewer than five sequences were sampled were excluded.

parameter commonly used in population genetics to describe the time-scale of selection and drift. In the epidemiological model, it affects the balance between selective constraint and antigenic selection by determining the extent to which prior immunity has decayed. When the interval between antigenic substitutions is small, immunity against the ancestral strain remains high at the time of substitution so that antigenic selection reduces the reversion probability. For larger intervals between antigenic substitutions, prior immunity will have decayed to a greater extent and the basic reproductive ratio and cost of immune escape become stronger determinants of the reversion probability. In the context of phylodynamic models, θ is also the parameter which is used to link the coalescent to epidemiological models (Koelle and Rasmussen, 2012).

Previous studies have described varying levels of reversion in a range of viruses and speculated on the influence of host immunity (Botosso et al., 2009; Delport et al., 2008; Wagner, 2014), but it has been unclear how the level of reversion should be quantified and how these results should be interpreted. In contrast to previous studies (Botosso et al., 2009; Palmer et al., 2013) based on phylogenetic methods, we propose using temporal patterns of frequency change to quantify reversion. Where sequence data from multiple time-points is available, a frequency-based approach can more easily show the time-dependent effect of antigenic selection. Simulation results predict that varying parameters controlling population size, transmission rate, immunity decay and selective constraint have qualitative effects on the frequency of the ancestral allele π_0 which are consistent with the analytical model, providing a means for interpretation. As our approach uses site-frequency data rather than a phylogeny, it is amenable to the application of large time-structured data sets, but is also more sensitive to effects such as biased sampling and spatial structure.

In this paper, we compared patterns of π_0 for two viruses that induce acute respiratory infection which recurrently infect human populations and induce long-term immunity: influenza A (H3N2) and RSV-A. For both viruses, we observed fluctuations in frequency at antigenic sites suggesting the presence of both immune memory and selective constraint. Without the continuously changing balance between these two effects, we would expect an allele for a particular site to reach fixation and remain in that state (Zhao et al., 2013). While RSV has been reported to experience high levels of reversions (Botosso et al., 2009), previous phylogentic studies have not identified reversion in H3N2. However, a recent study (Wagner, 2014) showed that changes at antigenic sites in H3N2 occur as cycles in a genotype network; that is, mutations to multiple states occur before reversion to the ancestral allele, so that the reversion is not identifiable along a the phylogeny.

Our model suggests that the higher rates of reversion in RSV-A compared to H3N2 is due mainly to more rapidly decaying immunity rather than stronger selective constraint. Fluctuations in frequency are more rapid and complete fixation of the ancestral amino acid does not occur for most antigenic sites of RSV-A. In contrast, for H3N2, we observe multiple occasions where a fixation of the ancestral amino acid occurs, and long periods where $\pi_0 = 1$ is maintained, suggesting strong selective constraints. This is consistent with the location of the sites within the receptor binding region of the HA gene, so that any antigenic change is also likely to affect viral transmissibility (Koel et al., 2013). Comparison between the frequency of the oscillations also suggests that H3N2 induces more long-lasting immunity than RSV-A. RSV-A exhibits more rapid fluctuation while several of the antigenic sites in H3N2 were fixed for long periods (> 10 years) at a derived amino acid, supporting the hypothesis that immune pressure against reversion is maintained for long periods. Frequency patterns of H3N2 frequency patterns are consistent with multi-site codon simulations (Figure 3.5) with host immunity decay rate γ on the order of 10^{-4}, whereas a value of $\gamma \approx 10^{-3}$ is more compatible with frequency patterns for RSV-A. These values are in agreement with reinfection experiments which estimate immunity for H3N2 lasting 8 years ($\gamma = 3 \times 10^{-4}$ day^{-1}) (Couch and Kasel, 1983) compared to 1.8 years ($\gamma = 1.5 \times 10^{-3}$ day^{-1}) for RSV-A (Hall et al., 1991).

Our study shows that the frequency of the ancestral allele, π_0, which can be easily calculated for time-stamped viral sequences, is informative about the immune dynamics and cost of escape. In particular, sharp fluctuations in frequency is indicative of immune selection occurring at a comparable time-scale to substitutions at antigenic sites. However, a small number of linked sites may also display similar patterns as they co-segregate with antigenic sites. That is, frequency patterns should not be used as a method to identify antigenic sites; but where the antigenic sites are known, frequency patterns provide information about the epidemiology of the virus as a whole.

The approach outlined here provides a qualitative description rather than estimates of the epidemiological parameters. Analytical expressions, relating the probability of reversion to the parameters underlying the viral dynamics for the three-strain model, rely on the assumption that each strain reaches equilibrium before it is replaced. This assumption tends to be violated when population sizes and mutation rates become large, so that we generally underestimate the probability of reversion. To address the restrictions of the equilibrium assumptions and the assumption of only three strains, we used computer simulations describing sequence dynamics in a multi-site model. Formal inference using a complex computational model is a challenge for future research. Despite the simplicity, our approach is useful in providing a scheme to consider both epidemiological and molecular effects simultaneously. As such, it is complementary to both coalescent approaches (Frost and Volz, 2010, 2013; Stadler and Bonhoeffer, 2013) which assume epidemiological dynamics are largely unaffected by molecular changes, and to codon-based methods (Nielsen and Yang, 2003; Tamuri et al., 2012) which assume that substitution occurs instantaneously as a time-homogeneous process along branches of the phylogeny.

Our model highlights the importance of understanding the interaction between epidemiological and molecular effects. The results imply that different evolutionary trajectories are expected in viral populations with the same distribution of fitness effects but differing population size and contact rates. In particular, we expect that viral populations in larger cities with denser populations undergo less reversion and are more likely to generate antigenically novel variants.

Chapter 4

A comparison of evolutionary dynamics between different lineages of avian influenza

4.1 Introduction

Influenza A infects a wide range of hosts, including humans, pigs, horses and birds, but their natural host is wild aquatic birds (Webster et al., 1992). All 16 hemagglutinin (HA) and 11 neuraminidase (NA) subtypes have been isolated in avian hosts. The evolutionary dynamics of avian influenza differ considerably from human influenza; multiple strains co-circulate and frequent reassortment occurs between segments (Macken et al., 2006). Avian influenza, therefore, forms a reservoir gene pool with potential to generate outbreaks in humans and other mammals (Webster et al., 1992).

Infection in avian hosts is generally asymptomatic and subject to weaker selective pressures (Chen and Holmes, 2006; Webster et al., 1992), but over recent years, there has been an increase in outbreaks of avian influenza (Alexander, 2007). Of particular concern has been the persistence of high-pathogenic H5N1 and low-pathogenic H9N2, and the continual generation of novel variants in these lineages (Alexander, 2007). Due their epidemiological impact, both subtypes have been the focus of particular attention. A number of subclades have been identified within both H5N1 (Neumann et al., 2010; World Health Organization, 2009, 2012) and H9N2 (e.g. Fusaro et al., 2011), and substitutions affecting phenotype have been identified across a number of gene segments (e.g. Fan et al., 2009; Hulse-Post et al., 2007; Neumann et al., 2010; Shi et al., 2009). However, it is difficult to understand the evolutionary significance of these substitutions. We know that while there is structural similarity between influenza genes of different host types (Meyer et al., 2013), there are also considerable differences in the distribution of site-specific selective effects (Meyer et al., 2013; Tamuri et al., 2012, 2009). To quantify natural selection, d_N/d_S has been estimated by pooling across the whole avian phylogeny (Chen and Holmes, 2006), and other studies have observed differences in d_N/d_S measured within different lineages of avian influenza (e.g Bahl et al., 2009). However, from these disparate studies, it remains unclear how much selection varies between lineages of avian influenza as a whole. Should the distinctive epidemiological behaviour of H5N1 and H9N2 be attributed to singular molecular characteristics? Or should we expect similar epidemiological behaviour in other influenza subtypes given particular conditions?

In this study, we attempt to provide a broad comparison of the dynamics of selection acting on lineages of avian influenza by examining not only sequence changes that become fixed in the population (substitution), but also transient sequence changes that are eventually removed (polymorphisms). As demonstrated in Chapter 2, selection is expected to distort patterns of polymorphism by interfering with linked sites. These indirect effects can be observed even if standard branch-site specific codon models are unable to identify specific sites under positive selection. Hence, the statistics developed in Chapter 2 (Chan et al., 2013) can be used to identify lineages evolving under different forms of selection. We analyse the coding region of the six internal gene segments; this consists of the polymerase genes PB2, PB1 and PA (segments 1–3, as ordered by size), the nucleoprotein (NP; segment 5), the matrix protein (M1; segment 7) and the non-structural protein (NS1; segment 8). The internal genes have evolved under stronger selective constraint (Chen and Holmes, 2006) and share a more recent common ancestor (Worobey et al., 2014) than the surface glycoproteins HA and NA (segments 4 and 6). Frequent reassortment should also reduce the effect of linked selection from other gene segments, so that selection at each segment can be considered independently.

4.2 Methods

4.2.1 Influenza sequence data

All full-length avian influenza sequences with a known year of isolation, excluding lab isolates were taken from the Influenza Virus database (Bao et al., 2008) (date accessed: 1 December 2013). This consisted of approximately 8000 sequences for each gene segment, with similar numbers of American and Eurasian sequences. However, the composition of the sequences was strongly biased in terms of subtype, with much larger numbers of H5, H6 and H9 sequences compared to the other subtypes. See Tables C.1–C.2 for more details. Sequences from each internal segment were aligned using Muscle (Edgar, 2004) with default parameters and GTR substitution model phylogenies were constructed using RAxML (Stamatakis, 2006) with a GTR substitution model. Sequence analysis was performed using R with the ape (Paradis et al., 2004) and seqinr (Charif and Lobry, 2007) packages.

4.2.2 Identifying lineages in the internal segments

Previous studies have tended to separate lineages based on bootstrap confidence or branch-length (e.g., Obenauer et al., 2006); however, these methods tend to split up lineages after a period of reduced sampling, or if particularly rapid divergence has occurred in a population. The second criterion tends to separate lineages along branches associated with selective effects. In contrast, the analysis here aims to identify populations or "lineages" evolving separately (i.e. reduced gene flow) due to both selective or non-selective (e.g. spatial separation) factors. This allows us to examine whether we observe similar or different selective conditions between lineages.

To define these lineages, we separate strains based on amino acid polymorphisms that are retained for long periods. Theoretical results from population genetics (Wright, 1931) predict that polymorphic sites are not expected to be stably maintained; the frequency of a polymorphism fluctuates until the segregating site reaches either fixation or extinction. Sequences for each of the internal segments were first divided into two groups based on whether they were isolated from the American or Eurasian hemisphere. Amino acid sites that were polymorphic at frequencies between 0.25–0.75 for over five years were then identified in sequences from each hemisphere. Further examination of these sites show find that they are generally polymorphic for more than five years, and much of this time is spent at stable frequencies (see Figures C.1–C.6).

To ensure that we only identify lineages with common ancestry, polymorphic sites were mapped back onto the maximum likelihood phylogenies, and sites at which mutation to the same amino acid residue occur along multiple branches were excluded. After removing redundant sites that have the same mutation pattern, the sequences of each gene segment were grouped according to whether they were isolated from the American or Eurasian hemisphere and the identity of the amino acid at a small number of sites: PB2 (64, 340), PB1 (59, 149), PA (272, 348, 400, 545), NP (34), M1 (27, 166), NS1 (6, 55, 60, 83, 87). These groups were then trimmed by identifying the node most distant from the root (A/chicken/Brescia/1902) with a high proportion of descendants belonging to a single lineage. At branch points where two lineages split off, the point where the lineages split is relatively robust to the cutoff (80%), but where a single lineage emerges from the end of an old one (e.g., lineage 2 of M1 and lineage 6 of NS1), a higher cut-off is required (99%). This step both removes outliers which were included into the lineage due to homoplasy and re-incorporates singletons that may have been excluded due to random mutations.

Note that incorrectly splitting a population into two lineages based on a polymorphism slowly reaching fixation does not overly affect our analysis. This is because, from the phlyogenetic information, we can assume that the lineages genuinely do reflect different ancestry; as such, similar site-frequencies distributions, and therefore summary statistics, should be obtained within each lineage (Wright, 1938). The overall result of splitting within a lineage would be that we overestimate the number of populations with similar selection dynamics.

4.2.3 Estimating d_N/d_S between different lineages

We estimated d_N/d_S for each lineage of all internal genes using HyPhy (Kosakovsky Pond et al., 2005). For computational efficiency, and to reduce the distorting effect of deleterious polymorphisms (Kryazhimskiy and Plotkin, 2008), we constructed sparse trees for each gene using a subsample of sequences with a maximum of ten sequences from each year. Trees were built using RAxML (Stamatakis, 2006) with a GTR substitution model. Branches in each subtree are labelled according to the lineage in which they occur, and d_N/d_S values for all lineages of a gene were estimated simultaneously by maximum likelihood. For each gene segment, two phylogenetic codon models were fitted: (i) a global model where d_N/d_S along all branches were constrained to the same value, and (ii) a local model where d_N/d_S in branches in the same lineage were constrained to be the same but could differ between lineages. The significance of the separation between lineages was evaluated using a Chi-squared test with $n-1$ degrees of freedom, where n is the number of lineages (Yang, 1998). For the local model, 95% confidence intervals for d_N/d_S values were estimated assuming asymptotic normality (Kosakovsky Pond et al., 2005).

4.2.4 Analysis of interference effects

To test for the presence of interfering mutations, we computed D_1, D_2 and D_3, as described in Equations (2.21)–(2.23) in Chapter 2 (Chan et al., 2013). For each lineage, statistics were computed with sequences pooled according to their year of isolation and compared against the ancestral sequence. The earliest sampled sequence in each lineage was taken to be the ancestral sequence. Bootstrap intervals (Efron, 1987) are constructed by resampling sequences with replacement for 500 replicates and recomputing the statistics.

4.2.5 Analysis of antigenic selection

The alignment for each HA subtype was constructed separately using Muscle (Edgar, 2004) with a GTR substitution model and default parameters. The separate alignments were then profile-aligned using Muscle with a gap penalty of 1000 to determine H3 numbering. For subtypes of class 1 (H1, H5, H6), we computed π_0 at known antigenic sites of H5 (Koel et al., 2013), while a different set of sites was used for H9 (Fusaro et al., 2011). For each antigenic site we computed the frequency of the most favoured amino acid residue π_0, taking the most frequently observed between 1990 and 2002 for that subtype to be the favoured residue.

4.3 Results

4.3.1 Distinct lineages within American and Eurasian influenza populations

A small number of lineages were identified for each gene segment, ranging from three to seven, which is shown in Figure 4.1. For all segments apart from PA, we observe a greater number of lineages in the Eurasian hemisphere. The more distinct clustering patterns in Eurasia can be mainly attributed to rapidly diversifying lineages of subtypes H5N1 and H9N2. Almost all sequences in lineage 4 of PB2, lineage 4 of PB1, lineage 3 of NP, lineage 4 of M1 and lineage 2 of NS1 belong to the H5 subtype. Similarly, the H9 lineage forms a separate lineage in the M1 (lineage 3) and NS1 (lineage 3) gene segment (see Tables C.1–C.2 for more details). A summary of the lineages is provided in Table 4.1, and we describe the results in further detail in the following section.

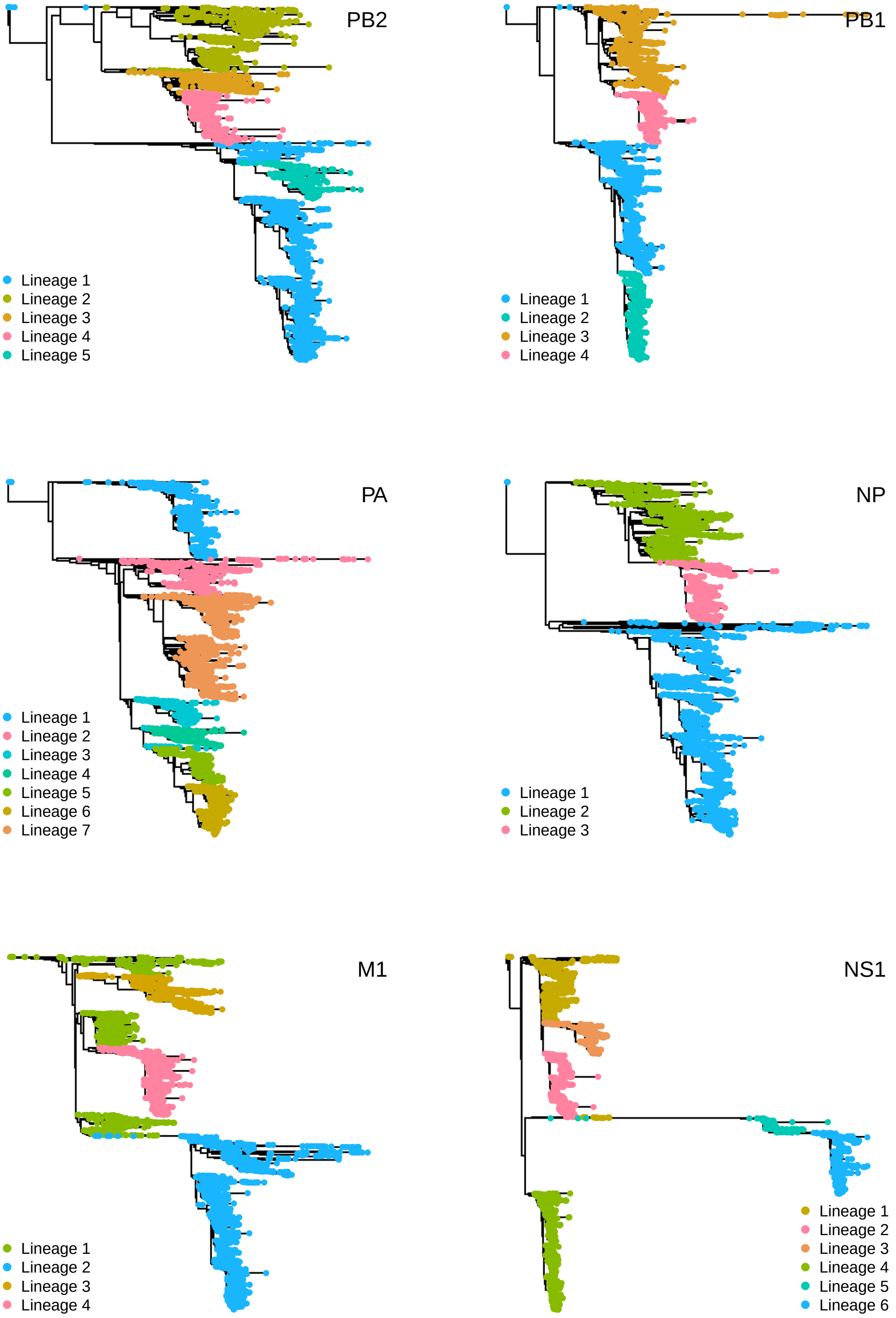

Figure 4.1: Lineages of the internal segments of avian influenza A. Phylogenies are shown rooted at the sequence A/chicken/Brescia/1902.

Table 4.1: Summary of selection statistics for lineages of the internal gene segments.

	Lineage	Region	Years [a]	D_1 [b]	D_2 [c]	D_3 [d]	d_N/d_S	95% CI
PB2	1	A	39	25	18	0	0.035	(0.032, 0.038)
	2	E	29	15	15	0	0.032	(0.029, 0.036)
	3	E	20	10	1	0	0.050	(0.044, 0.057)
	4 (H5)	E	14	9	9	0	0.042	(0.035, 0.049)
	5	A	26	11	5	1	0.049	(0.042, 0.056)
PB1	1	A	41	30	16	0	0.034	(0.032, 0.038)
	2	A	15	9	6	0	0.028	(0.022, 0.035)
	3	E	33	18	14	0	0.032	(0.029, 0.035)
	4 (H5)	E	14	10	11	1	0.041	(0.034, 0.049)
PA	1	A	26	10	13	4	0.032	(0.027, 0.037)
	2	E	33	15	8	1	0.046	(0.042, 0.050)
	3	A	29	6	6	1	0.047	(0.040, 0.055)
	4	A	22	3	1	0	0.066	(0.055, 0.078)
	5	A	23	8	4	1	0.047	(0.038, 0.057)
	6	A	15	8	3	0	0.036	(0.027, 0.046)
	7	E	19	15	7	0	0.054	(0.048, 0.054)
NP	1	A	39	24	24	1	0.030	(0.028, 0.033)
	2	E	34	16	13	1	0.043	(0.038, 0.049)
	3 (H5)	E	13	11	4	0	0.040	(0.032, 0.049)
M1	1	E	42	11	13	3	0.050	(0.042, 0.059)
	2	A	39	19	1	0	0.023	(0.019, 0.028)
	3 (H9)	E	19	7	6	5	0.057	(0.042, 0.076)
	4 (H5)	E	15	6	8	9	0.064	(0.046, 0.088)
NS1	1	E	35	9	16	0	0.188	(0.169, 0.208)
	2 (H5)	E	14	6	6	0	0.249	(0.201, 0.304)
	3 (H9)	E	17	6	0	0	0.296	(0.244, 0.356)
	4	A	38	22	13	1	0.150	(0.132, 0.170)
	5 (Allele B)	E	16	5	4	1	0.160	(0.113, 0.218)
	6 (Allele B)	A	32	13	10	1	0.099	(0.085, 0.114)

[a] Total number of years in which lineage was sampled.
[b] Number of years in which D_1 is significantly greater than zero.
[c] Number of years in which D_2 is significantly smaller than zero.
[d] Number of years in which D_3 is significantly greater than zero.

4.3.2 Comparison of selection between lineages

We compare selection acting on each lineage by examining d_N/d_S, which is a measure of the relative evolutionary rate over time, and by computing the interference statistics introduced in Chapter 2 (Chan et al., 2013). These statistics identify the presence of selected mutations in the population and characterise the distortion of the site-frequency spectrum. Briefly, values of $D_1 > 0$ indicate the presence of deleterious mutations that can potentially hitch-hike, $D_2 < 0$ indicates the presence of interfering mutations (both beneficial or deleterious), and the composition of segregating mutations is indicated by D_3. Specifically, the presence of multiple beneficial mutations (i.e., clonal interference) is indicated by the combination of $D_2 < 0$ and $D_3 > 0$.

Values of d_N/d_S for each lineage are reported in Table 4.1. For all six internal gene segments, the phylogenetic model with separate lineages fit the data better than a model where a global d_N/d_S value was used. The model fit was significantly improved ($p \ll \times 10^{-4}$) for all segments with the exception of PB1, where the improvement was only at the 0.05 level.

As mentioned in Section 4.3.1, a number of lineages predominantly consist of H5 and H9 sequences. These lineages are also associated with higher d_N/d_S values in PB1, M1, and NS1. A higher d_N/d_S value can indicate positive selection, or relaxed selective constraint. Lineage 4 of PB1 shows signs of excess beneficial mutations ($D_3 > 0$; Figure 4.2), which suggests that positive selection may contribute to the higher d_N/d_S value, but some portion of this may also be due to deleterious mutations. We observe indications of clonal interference only briefly after the emergence of lineage 4, but values of $D_2 < 0$ (Figure 4.3) are maintained subsequently over multiple years.

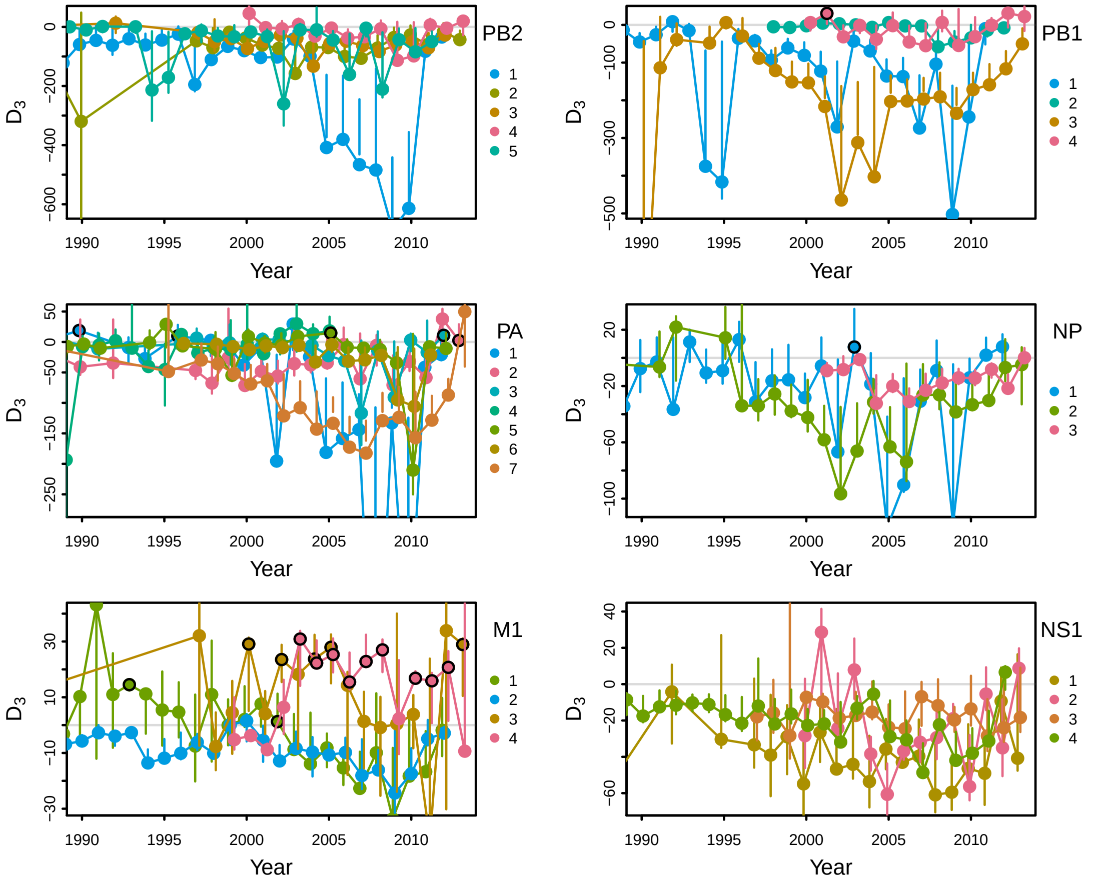

Figure 4.2: Clonal interference ($D_3 > 0$), in the internal segments of avian influenza A. Each lineage is shown with colours corresponding to Figure 4.1. Lineages from NS1 allele B were omitted for clarity. Vertical bars represent 95% intervals computed by bootstrapping with resampled sequences. Years in which D_3 is significantly greater than zero are indicated by black outlines.

In contrast, for lineage 4 of M1, there are signs of clonal interference ($D_3 > 0$) in multiple years, indicating that the higher d_N/d_S value does reflect positive selection. For lineage 2 of NS1, we do not observe $D_3 > 0$, but we observe multiple years where $D_2 > 0$. This indicates an excess of mid-frequency mutations, which suggests that balancing selection is operating on NS1 of the H5 lineage. We observe a similar pattern for lineages associated with H9 (orange), where lineage 3 of M1 shows signs of clonal interference and lineage 3 of NS1 shows signs of balancing selection.

Lineage 4 of PB2 and lineage 3 of NP, which are associated with H5, do not show particularly high values of d_N/d_S, but they also exhibit distinct population dynamics which differ from other other lineages of the same gene segment. PB2 shows more signs of interfering mutations ($D_2 < 0$), while in contrast, lineage 3 of NP seems to have a reduced level of variation in fitness.

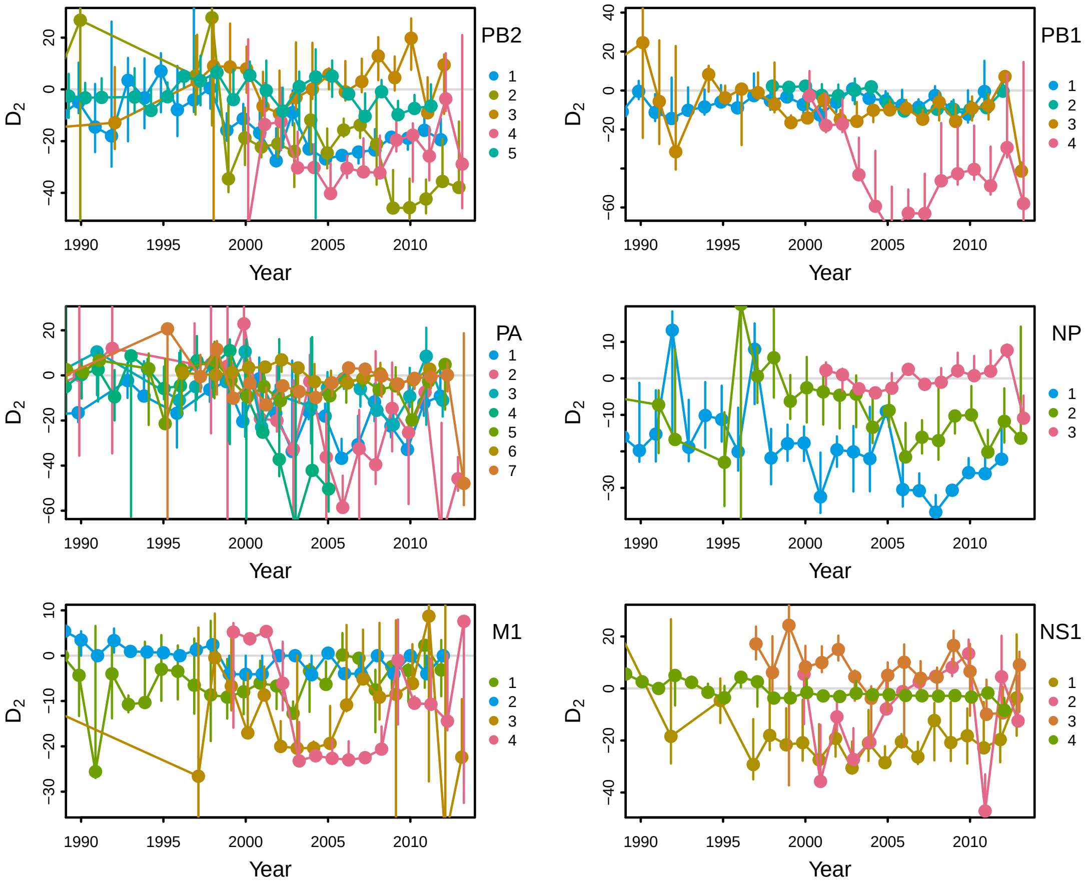

Figure 4.3: The presence of interfering mutations ($D_2 < 0$), in the internal segments of avian influenza A. Each lineage is shown with colours corresponding to Figure 4.1. Lineages from NS1 allele B were omitted for clarity. Vertical bars represent 95% intervals computed by bootstrapping with resampled sequences.

We focused on lineages associated with H5 and H9 as these showed the most distinctive patterns of selection, but other differences between lineages can also be seen. We observe occasional signs of clonal interference in multiple lineages of PA and NP (Table 4.1). For all lineages, D_1 is greater than zero a considerable proportion of the time (Table 4.1), indicating the presence of deleterious mutations in the population. However, deleterious mutations are maintained for varying durations in the population for different gene segments, as seen by the variation in patterns of D_2 and D_3 over time.

4.3.3 Comparison of site-specific patterns between lineages

In Figure 4.4, we consider differences between lineages in more detail by comparing amino acid frequencies at each site. Halpern and Bruno (1998) showed that at equilibrium, the frequency of an allele is proportional to its probability of fixation. Thus, the frequency

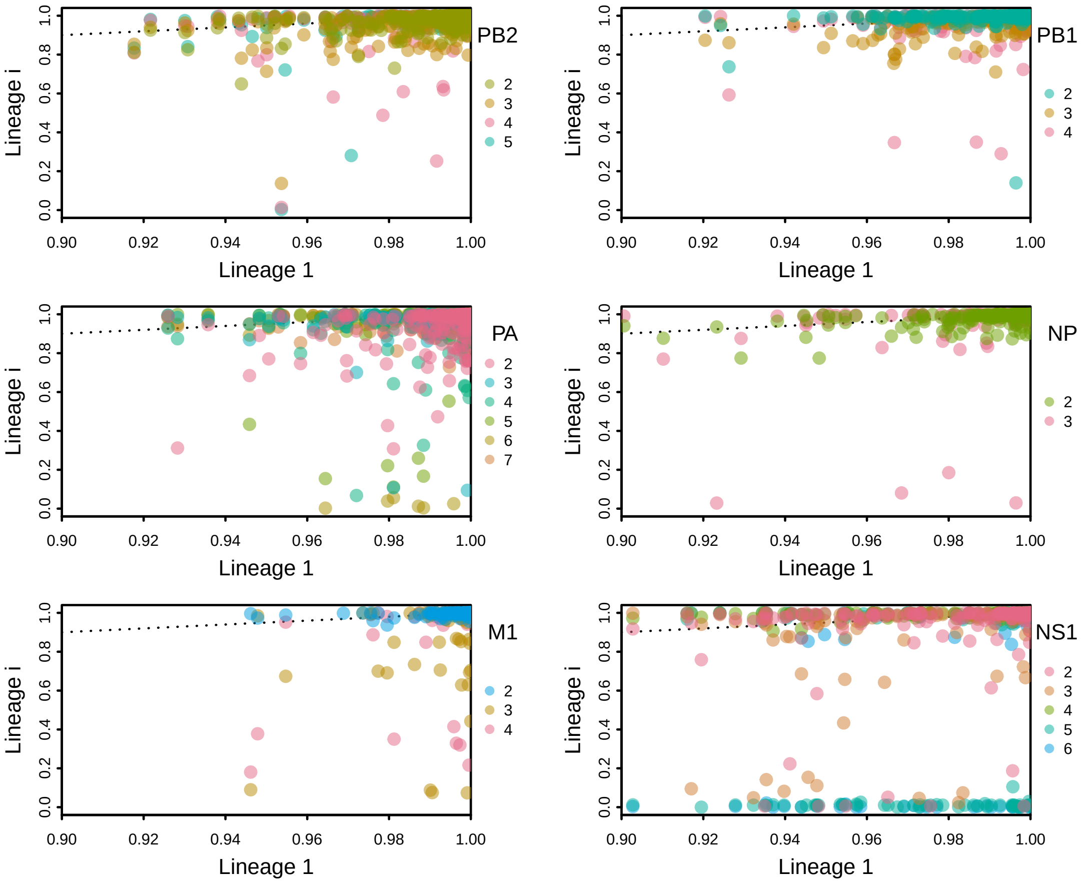

Figure 4.4: Comparison of site-specific constraint between lineage 1 and all other lineages i. Constraint is measured as the frequency of the favoured amino acid, averaged over all years from 1990 onwards. Each point corresponds to a single amino acid site, with the lineage indicated by the colour of the point, following the same colouring as Figure 4.1. The dotted black line represents $x = y$.

of a given allele is expected to increase with the degree of selective constraint for that allele. For each site, we determine the residue which occurs most frequently in lineage 1, which is closest to the root of the phylogeny. Taking this to be the amino acid that was favoured ancestrally, we measure the degree of constraint for the favoured amino acid by averaging its frequency across all years from 1990 onwards.

Assuming that constraint on protein structure induces similar levels of constraint between lineages, we would expect that the site-specific patterns in one lineage correspond to similar patterns in a different lineage. Lineages under similar selective conditions should therefore have sites falling mostly around the $x = y$ line in Figure 4.4. Relaxed selective constraint in a lineage should have a relatively stronger effect on sites which are less constrained, so that they fall off more sharply than $x = y$.

We indeed observe the expected pattern for relaxed selective constraint in lineage 3 of PB2, lineages 2 and 4 of PA, lineage 3 of M1 and lineage 3 of NS1. However, this is not the pattern observed for lineage 3 and lineage 6 of PA and lineage 3 of NP, where sites which are strongly constrained in lineage 1 are fixed at a derived state in a different lineage. These lineages also have values of $D_2 \approx 0$ (Figure 4.3), which is suggestive of reduced genetic variation due to single sweeps of positive selection.

4.3.4 Antigenic dynamics of H5 and H9

In Section 2.2, we saw that the emergence of the Eurasian H5N1 and H9N2 lineage was associated with specific lineages of the internal segments. Here, we examine, how the antigenic dynamics of the H5 and H9 Eurasian subtypes compare to other subtypes. The analysis in this section is based on the frequency of the ancestral amino acid at antigenic sites, denoted π_0, which was described in Chapter 3. Strong selective constraint for the ancestral amino acid is expected to π_0 to values close to 1, while fluctuations in π_0 reflect the effect of immunity and its duration in the host population.

In Figure 4.5, we show π_0 values for subtypes H1, H5 and H6 in both the American and Eurasian lineages. All 3 subtypes belong to the H1 clade (Nobusawa et al., 1991), and are expected to be very similar in structure. For H1 and H6, we observed that each antigenic site tends to behave similarly in both the American and Eurasian lineage. Although the duration of immunity in the Eurasian lineage of H6 seems slightly longer than in the American lineage, we still observe fluctuating dynamics indicating the presence of both antigenic selection and selective constraint. However, for H5, a sharp difference in dynamics is observed at sites 137 and 193; these sites do not revert to their favoured state despite indications of strong constraint in the American lineage of H5.

For H9, a similar comparison between related subtypes is not possible as subtypes of the same clade, H8 and H12 have rarely been sampled, and only a small number of H9 sequences have been isolated from the American hemisphere. However, for H9 sequences isolated from Eurasia, we observe temporal patterns of π_0 at antigenic sites resembling those of H5 in Eurasia. These patterns suggest that the Eurasian lineages of both H5 and H9 undergo relaxation of selective constraint leading to a reduced cost of immune escape and promoting antigenic drift.

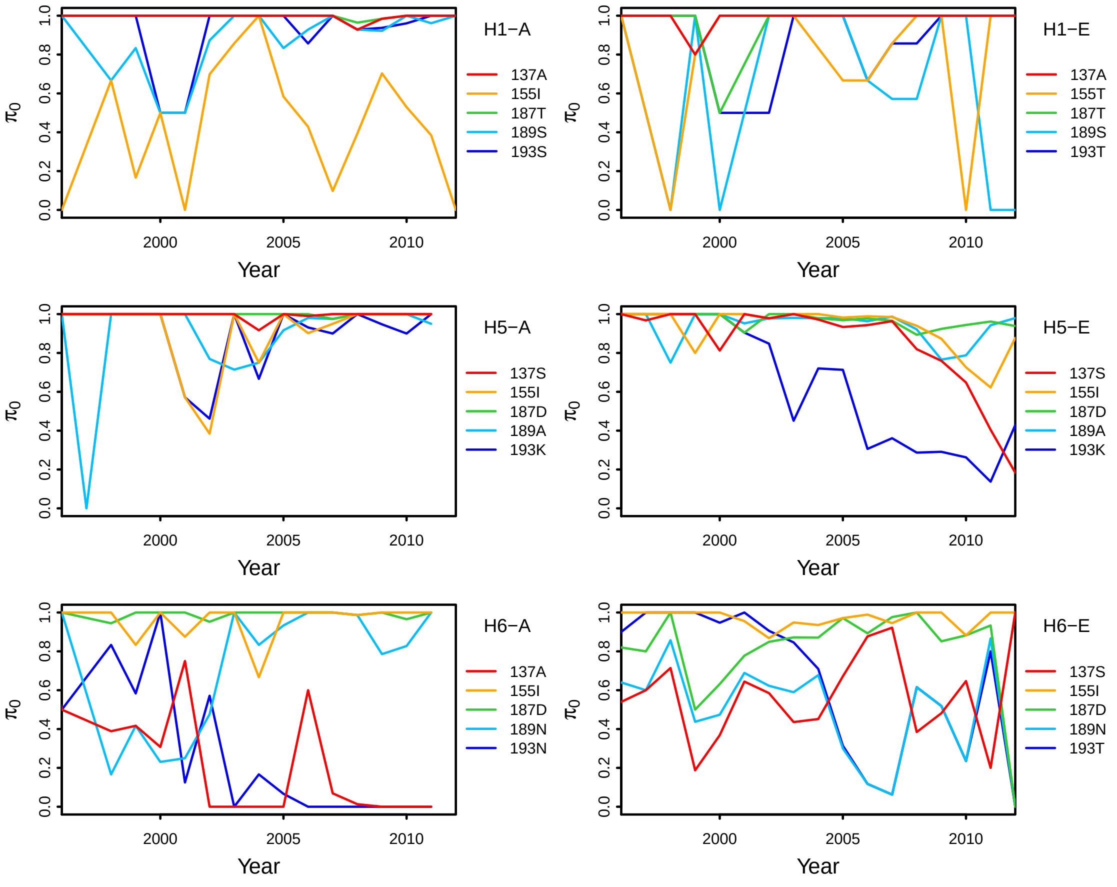

Figure 4.5: Trajectory of the favoured allele π_0 at antigenic sites in HA subtypes of clade 1. The legend to the right of each panel denotes the subtype and hemisphere of the HA sequences, and the favoured amino acid at each site. In all panels, we analysed the same sites, which are known to be antigenic sites of H5 (Koel et al., 2013)

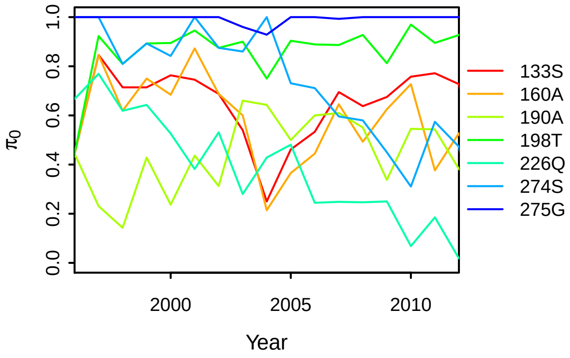

Figure 4.6: Trajectory of the favoured amino acid of H9 in Eurasia at antigenic sites (Fusaro et al., 2011).

4.4 Discussion

In this study, we have identified distinct lineages in the internal segments of avian influenza and have shown that the evolutionary dynamics of these lineages differ. Differences in values of d_N/d_S values are typically interpreted as reflecting a change in either the strength of selection s, or the effective population size N_e (Nielsen and Yang, 2003). Further analysis using interference statistics D_1, D_2 and D_3, indicated that much of the variation in d_N/d_S can be attributed to the indirect effects of selection at linked sites. As shown by a comparison of site-specific frequencies between lineages, interference has an effect similar to reducing population size by weakening selective constraint across all sites in the lineage. However, relaxed selective constraint due to interference differs from a reduced population size in that it reflects high rates of mutational input. The presence of interfering mutations ($D_2 < 0$), particularly beneficial mutations ($D_3 > 0$), therefore indicates increased potential for the emergence of new strains. Consistent with this, we find that two lineages showing signs of clonal interference in the M1 gene correspond to outbreaks of subtype H5N1 and H9N2.

In contrast to human influenza, which is mainly characterised by clonal interference in HA (Chan et al., 2013; Strelkowa and Lässig, 2012), we find that patterns of linked selection vary between genes and between lineages in avian influenza. Multiple lineages of PA and NP show intermittent signs of clonal interference, and comparison between sites suggests that single-sweeps of positive selection may occur in other lineages of PA and NP. There is also considerable variation in the behaviour of deleterious mutations. All lineages show some signs of deleterious mutation ($D_1 > 0$), but the extent to which they persist in the population to generate background selection ($D_2 < 0$) varies. Most genes, particularly PB1 and PB2, have values of D_2 which fluctuate between negative values and values close to zero, suggesting that deleterious mutations are slowly but eventually removed.

Much of the analysis in this study focuses on the dynamics of evolution in the lineages associated with H5 and H9 outbreaks. The larger number of H5 and H9 sequences may be an artefact of the more extensive monitoring of these subtypes due to their potential for generating epidemics. Lineages which are sampled at only low frequencies, or persist for only short durations are unlikely to be distinguishable using our method of identifying lineages. With more extensive sampling of avian influenza, it may be possible to identify more lineages and characterise their dynamics to compare selection in avian influenza more comprehensively.

Nevertheless, this study was able to identify some distinctive characteristics common to Eurasian lineages of H5 and H9, which have become endemic in avian populations (Alexander, 2007). Consistent with experimental studies, our results suggest that the emergence and spread of these lineages have been influenced by adaptive changes in the internal genes, including M1 (Fan et al., 2009), NS1 (Li et al., 2006) and PA and PB1 (Hulse-Post et al., 2007). In most of these genes, only a small number of changes is likely to have occurred, but the trajectory of the interference statistics suggests that continuous adaptive changes in the M1 gene occurred for both the H5 and H9 lineage; this is consistent with the close association of the matrix protein with the surface glycoproteins (Scholtissek et al., 2002). Both lineages also show signs of fluctuating selection at NS1. The precise function of NS1 is unclear, but it is known to have an important role in evading host immunity (reviewed in Dundon and Capua, 2009). Previous studies have noted that these two subtypes mainly reassort within only their own subtype (Lu et al., 2014; Neumann et al., 2010), suggesting that segments from different lineages are incompatible. One reason for this incompatibility, our results suggest, is that deleterious mutations may have reached fixation by hitch-hiking to adaptive substitutions.

The dynamics of antigenic selection in H5 and H9 are distinctive according to the temporal patterns of frequency π_0 at antigenic sites. Both subtypes show a decrease in reversion, which indicates reduced selective constraint. This contrasts with the observed patterns of fluctuating frequencies at antigenic sites of other subtypes. In fact, for H5 strains in the American hemisphere, we observe fluctuations at the same sites which contribute to antigenic drift in Eurasian H5 strains. However, the trajectory of π_0 in American strains of H5 show that these sites are under strong selective constraint.

Theoretical results from Chapter 3 suggest that any increase in the basic reproductive ratio of the virus will increase the ability of the antigenic sites under selective constraint to escape immune recognition. This suggests that adaptations in the internal genes of Eurasian H5 and H9 may differ in their biological effect, but any change that increases viral transmissibility may be able to explain the distinctive evolutionary dynamics of these two lineages. It may also explain why attempts to control outbreaks through vaccination have been largely unsuccessful in both H5N1 (Cattoli et al., 2011) and H9N2 (Park et al., 2011); in contrast, vaccination of poultry populations against subtype H7N3 and H5N2 did not induce antigenic drift (Marangon et al., 2008).

Chapter 5

Conclusions and future directions

In this thesis, we have developed models to describe some complexities of viral evolution caused by different forms of selection acting at similar timescales. Standard models of molecular evolution assume the fixation process of any mutation is independent of other mutations in the population and is largely determined by genetic drift and the fitness effect of the mutation. However, in viral populations, the trajectory of a mutation in a viral population can be affected by additional factors, such as interfering mutations and epidemiological dynamics. In Chapter 2, we presented statistics to identify and distinguish between the effects of background selection, hitch-hiking and clonal interference. In Chapter 3, we showed that temporal patterns of frequency change at antigenic sites can be used to distinguish between the effects of epidemiological dynamics and selective constraint. Specific points about these models and statistics have already been addressed in the previous chapters, but here we discuss some general implications for modelling of viral evolution, and directions for future work.

Our results show that demographic processes cannot be clearly separated from selection as population dynamics are influenced by the indirect effects of selection at linked sites, and selection in viral populations are often frequency-dependent due to viral-host interactions. One of the most influential parameters is the rate of mutational input θ. In the Wright-Fisher model, the rate of mutational input controls the effect of drift and rate of neutral evolution. In viral populations, the effects can be more complex, affecting the level of interference between different strains (Chapter 2), and how the viral population responds to epidemiological changes (Chapter 3).

In theoretical models (Desai and Fisher, 2007; Rouzine et al., 2003), the rate of mutational input is presented as a simple parameter, but it is actually composite of many different effects. Biologically, it would be of interest to be able to distinguish these effects. For example, in Chapter 4, we found that lineages of avian influenza showed variation in both the level of interference, and in the dynamics of antigenic selection. Interestingly, we found that two lineages involved in the generation of multiple outbreaks, H5N1 and H9N2, are associated with higher rate of mutational input. However, it remains unclear whether the change in evolutionary dynamics is due to demographic effects, such as changes in poultry breeding practices, or to a change in the fitness landscape so that there is greater number of adaptive sites. This underscores the necessity of epidemiolgical models for understanding of viral evolution.

The ability to evaluate the relative influence of a potential selective mechanism relies crucially on the use of informative sequence statistics (Gray et al., 2011; Zinder et al., 2013). For example, previous ecological models of human influenza A have been unable to discriminate between the influence of the rate of mutational input (Koelle et al., 2006) and the role of host immunity (Ferguson et al., 2003) based on comparisons of the shape of the phylogeny, rates of substitution, and the level genetic diversity. Here we use frequency-based statistics, which can make use of densely sampled sequences to provide information on a finer time-scale about the dynamics of transitory mutations (Chapter 2), and to distinguish between the dynamics of different forms of selection (Chapter 3).

In ecological models, it is often not appreciated that selective constraint is the most common form of natural selection (Ohta, 1973), which has implications for the choice of summary statistics. In the classic two-allele (ancestral and derived) model, there is no difference between negative and positive selection because both forms of selection involve favouring one allele over the other. However, where sites can mutate to multiple alleles, the dynamics of positive selection and selective constraint are very different. As seen in Chapter 3, this difference is crucial to how we distinguish between the effects of host immunity and selective constraint. Recent models of molecular evolution (Halpern and Bruno, 1998; Kosakovsky Pond et al., 2008; Kryazhimskiy et al., 2008; Seoighe et al., 2007) have attempted to incorporate the directionality of negative selection. However ecological models of viral evolution have generally focused on positive selection. The distinctive directionality of selective constraint is often overlooked when an infinite-sites (Koelle et al., 2009) or two-allele representation (Tria et al., 2005) is used, or when sites are assumed by default to be evolving neutrally (Zinder et al., 2013).

An additional aspect for future work is better statistical inference. The primary concern of this thesis was to develop models which were able to qualitatively describe the dynamics of viral evolution. But to obtain more detailed understanding of why different viruses show different evolutionary dynamics, it is important to develop better means of quantification which can account for temporal and spatial bias in sampling. The use of temporal patterns of frequency to infer selection has only been recently developed and lacks a full statistical framework. The method presented by Illingworth et al. (2012) accounts for variance due to sampling, but assumes the trajectory of the mutation is deterministic. Analytical models describing the variance structure of the site-frequency spectrum (Sawyer and Hartl, 1992) is based on the assumption of independence between sites, which, as described in Chapters 1 and 2, is inappropriate for viral populations (see also Bustamante et al., 2001; Neher and Shraiman, 2011). While an analytical expression for variance may be difficult to obtain, there are statistical methods that can circumvent this problem. One promising approach is the approximate Bayesian computation method presented by Rasmussen et al. (2011), in which a stochastic SIR model was fitted to simulated genealogies and time series data to infer key epidemiological parameters.

The work in this thesis provides a basis for future work to elucidate the molecular and epidemiological mechanisms that shape viral populations.

Bibliography

Alexander, D. J. (2007). An overview of the epidemiology of avian influenza. *Vaccine 25*(30), 5637–5644.

Bahl, J., M. I. Nelson, K. H. Chan, R. Chen, D. Vijaykrishna, R. A. Halpin, T. B. Stockwell, X. Lin, D. E. Wentworth, E. Ghedin, Y. Guan, J. S. M. Peiris, S. Riley, A. Rambaut, E. C. Holmes, and G. J. D. Smith (2011). Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans. *Proc. Natl Acad. Sci. U.S.A. 108*(48), 19359 –19364.

Bahl, J., D. Vijaykrishna, E. C. Holmes, G. J. Smith, and Y. Guan (2009). Gene flow and competitive exclusion of avian influenza A virus in natural reservoir hosts. *Virology 390*(2), 289–297.

Bailey, J. R., S. Laskey, L. N. Wasilewski, S. Munshaw, L. J. Fanning, E. Kenny-Walsh, and S. C. Ray (2012). Constraints on viral evolution during chronic hepatitis C virus infection arising from a common-source exposure. *J. Virol. 86*(23), 12582–12590.

Bao, Y., P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman (2008). The influenza virus resource at the National Center for Biotechnology Information. *J. Virol. 82*(2), 596–601.

Bhatt, S., E. C. Holmes, and O. G. Pybus (2011). The genomic rate of molecular adaptation of the human influenza A virus. *Mol. Biol. Evol. 28*(9), 2443 –2451.

Birky, C. W. and J. B. Walsh (1988). Effects of linkage on rates of molecular evolution. *Proc. Natl. Acad. Sci. U.S.A. 85*(17), 6414–6418.

Boni, M. F., Y. Zhou, J. K. Taubenberger, and E. C. Holmes (2008). Homologous recombination is very rare or absent in human influenza A virus. *J. Virol. 82*(10), 4807–4811.

Botosso, V. F., P. M. d. A. Zanotto, M. Ueda, E. Arruda, A. E. Gilio, S. E. Vieira, K. E. Stewien, T. C. T. Peret, L. F. Jamal, M. I. d. M. C. Pardini, J. R. R. Pinho, E. Massad, O. A. Sant'Anna, E. C. Holmes, E. L. Durigon, and and the VGDN Consortium (2009). Positive selection results in frequent reversible amino acid replacements in the G protein gene of human respiratory syncytial virus. *PLoS Pathog. 5*(1), e1000254.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan (1995). The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics 140*(2), 783–796.

Bush, R. M., W. M. Fitch, C. A. Bender, and N. J. Cox (1999). Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol. Biol. Evol. 16*(11), 1457 –1465.

Bustamante, C. D., J. Wakeley, S. Sawyer, and D. L. Hartl (2001). Directional selection and the site-frequency spectrum. *Genetics 159*(4), 1779–1788.

Cattoli, G., A. Fusaro, I. Monne, F. Coven, T. Joannis, H. S. A. El-Hamid, A. A. Hussein, C. Cornelius, N. M. Amarin, M. Mancin, E. C. Holmes, and I. Capua (2011). Evidence for differing evolutionary dynamics of A/H5N1 viruses among countries applying or not applying avian influenza vaccination in poultry. *Vaccine 29*(50), 9368–9375.

Chan, C. H., S. Hamblin, and M. M. Tanaka (2013). The effects of linkage on comparative estimators of selection. *BMC Evol. Biol. 13*(1), 244.

Charif, D. and J. R. Lobry (2007). SeqinR 1.0-2: A contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In D. U. Bastolla, P. D. M. Porto, D. H. E. Roman, and D. M. Vendruscolo (Eds.), *Structural Approaches to Sequence Evolution*, Biological and Medical Physics, Biomedical Engineering, pp. 207–232. Springer Berlin Heidelberg.

Charlesworth, B. (1994). The effect of background selection against deleterious mutations on weakly selected, linked variants. *Gent. Res. 63*(03), 213–227.

Charlesworth, B., M. T. Morgan, and D. Charlesworth (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics 134*(4), 1289–1303.

Charlesworth, D., B. Charlesworth, and M. T. Morgan (1995). The pattern of neutral molecular variation under the background selection model. *Genetics 141*(4), 1619.

Charlesworth, J. and A. Eyre-Walker (2008). The McDonald-Kreitman test and slightly deleterious mutations. *Mol. Biol. Evol. 25*(6), 1007–1015.

Chen, R. and E. C. Holmes (2006). Avian influenza virus exhibits rapid evolutionary dynamics. *Mol. Biol. Evol. 23*(12), 2336 –2341.

Comeron, J. M. and M. Kreitman (2002). Population, evolutionary and genomic consequences of interference selection. *Genetics 161*(1), 389–410.

Coop, G. and P. Ralph (2012). Patterns of neutral diversity under general models of selective sweeps. *Genetics 192*(1), 205–224.

Couch, R. B. and J. A. Kasel (1983). Immunity to influenza in man. *Annu. Rev. Microbiol. 37*(1), 529–549.

Czelusniak, J., M. Goodman, D. Hewett-Emmett, M. L. Weiss, P. J. Venta, and R. E. Tashian (1982). Phylogenetic origins and adaptive evolution of avian and mammalian haemoglobin genes. *Nature 298*(5871), 297–300.

Delport, W., K. Scheffler, and C. Seoighe (2008). Frequent toggling between alternative amino acids is driven by selection in HIV-1. *PLoS Pathog. 4*(12), e1000242.

Desai, M. M. and D. S. Fisher (2007). Beneficial mutation–selection balance and the effect of linkage on positive selection. *Genetics 176*(3), 1759 –1798.

Desai, M. M., A. M. Walczak, and D. S. Fisher (2013). Genetic diversity and the structure of genealogies in rapidly adapting populations. *Genetics 193*(2), 565–585.

Dundon, W. G. and I. Capua (2009). A closer look at the NS1 of influenza virus. *Viruses 1*(3), 1057–1072.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res. 32*(5), 1792–1797.

Efron, B. (1987). Better bootstrap confidence intervals. *J. Amer. Statist. Assoc. 82*(397), 171–185.

Ewens, W. J. (1967). The probability of survival of a new mutant in a fluctuating environment. *Heredity 22*, 438–443.

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

Eyre-Walker, A., M. Woolfit, and T. Phelps (2006). The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics 173*(2), 891–900.

Fan, S., G. Deng, J. Song, G. Tian, Y. Suo, Y. Jiang, Y. Guan, Z. Bu, Y. Kawaoka, and H. Chen (2009). Two amino acid residues in the matrix protein M1 contribute to the virulence difference of H5N1 avian influenza viruses in mice. *Virology 384*(1), 28–32.

Fay, J. C. (2011). Weighing the evidence for adaptation at the molecular level. *Trends Genet. 27*(9), 343–349.

Fay, J. C. and C.-I. Wu (2000). Hitchhiking under positive Darwinian selection. *Genetics 155*(3), 1405–1413.

Fay, J. C., G. J. Wyckoff, and C.-I. Wu (2001). Positive and negative selection on the human genome. *Genetics 158*(3), 1227–1234.

Felsenstein, J. (1992). Estimating effective population size from samples of sequences: Inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Gent. Res. 59*(02), 139–147.

Ferguson, N. M., A. P. Galvani, and R. M. Bush (2003). Ecological and immunological determinants of influenza evolution. *Nature 422*(6930), 428–433.

Fitch, W. M., R. M. Bush, C. A. Bender, and N. J. Cox (1997). Long term trends in the evolution of H3 HA1 human influenza type A. *Proc. Natl. Acad. Sci. U.S.A. 94*(15), 7712 –7718.

Fitch, W. M., J. M. Leiter, X. Q. Li, and P. Palese (1991). Positive Darwinian evolution in human influenza A viruses. *Proc. Natl. Acad. Sci. U.S.A. 88*(10), 4270 –4274.

Frost, S. D. W. and E. M. Volz (2010). Viral phylodynamics and the search for an 'effective number of infections'. *Philos. T. Roy. Soc. B 365*(1548), 1879 –1890.

Frost, S. D. W. and E. M. Volz (2013). Modelling tree shape and structure in viral phylodynamics. *Phil. Trans. R. Soc. B 368*(1614), 20120208.

Fryer, H. R., J. Frater, A. Duda, M. G. Roberts, R. E. Phillips, A. R. McLean, and The SPARTAC Trial Investigators (2010). Modelling the evolution and spread of HIV immune escape mutants. *PLoS Pathog. 6*(11), e1001196.

Fusaro, A., I. Monne, A. Salviato, V. Valastro, A. Schivo, N. M. Amarin, C. Gonzalez, M. M. Ismail, A.-R. Al-Ankari, M. H. Al-Blowi, O. A. Khan, A. S. M. Ali, A. Hedayati, J. G. Garcia, G. M. Ziay, A. Shoushtari, K. N. A. Qahtani, I. Capua, E. C. Holmes, and G. Cattoli (2011). Phylogeography and evolutionary history of reassortant H9N2 viruses with potential human health implications. *J. Virol. 85*(16), 8413–8421.

Galassi, M., J. Davies, J. Theiler, B. Gough, G. Jungman, M. Booth, and F. Rossi (2003). *GNU Scientific Library Reference Manual.*

García, O., M. Martín, J. Dopazo, J. Arbiza, S. Frabasile, J. Russi, M. Hortal, P. Perez-Breña, I. Martínez, and B. García-Barreno (1994). Evolutionary pattern of human respiratory syncytial virus (subgroup A): cocirculating lineages and correlation of genetic and antigenic changes in the G glycoprotein. *J. Virol. 68*(9), 5448–5459.

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

Gog, J. R. and B. T. Grenfell (2002). Dynamics and selection of many-strain pathogens. *Proc. Natl. Acad. Sci. U.S.A. 99*(26), 17209 –17214.

Gray, R. R., O. G. Pybus, and M. Salemi (2011). Measuring the temporal structure in serially sampled phylogenies. *Method. Ecol. Evol. 2*(5), 437–445.

Grenfell, B. T., O. G. Pybus, J. R. Gog, J. L. N. Wood, J. M. Daly, J. A. Mumford, and E. C. Holmes (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science 303*(5656), 327–332.

Griffiths, R. C. and S. Tavare (1994). Sampling theory for neutral alleles in a varying environment. *Philos. T. Roy. Soc. B 344*(1310), 403–410.

Hahn, M. W., M. D. Rausher, and C. W. Cunningham (2002). Distinguishing between selection and population expansion in an experimental lineage of bacteriophage t7. *Genetics 161*(1), 11–20.

Haldane, J. B. S. (1927). A mathematical theory of natural and artificial selection, part V: Selection and mutation. *Math. Proc. Cambridge 23*(07), 838–844.

Hall, C. B., E. E. Walsh, C. E. Long, and K. C. Schnabel (1991). Immunity to and frequency of reinfection with respiratory syncytial virus. *J. Infect. Dis. 163*(4), 693–698.

Halpern, A. L. and W. J. Bruno (1998). Evolutionary distances for protein-coding sequences: Modeling site-specific residue frequencies. *Mol. Biol. Evol. 15*(7), 910–917.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Review 42*(4), 599–653.

Hughes, A. L. (2007). Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity 99*(4), 364–373.

Hughes, A. L. and M. Nei (1988). Pattern of nucleotide substitution at major histocompatibility complex class i loci reveals overdominant selection. *Nature 335*(6186), 167–170.

Hulse-Post, D. J., J. Franks, K. Boyd, R. Salomon, E. Hoffmann, H. L. Yen, R. J. Webby, D. Walker, T. D. Nguyen, and R. G. Webster (2007). Molecular changes in the polymerase genes (PA and PB1) associated with high pathogenicity of H5N1 influenza virus in mallard ducks. *J. Virol. 81*(16), 8515–8524.

Illingworth, C. J. R. and V. Mustonen (2012). Components of selection in the evolution of the influenza virus: Linkage effects beat inherent selection. *PLoS Pathog. 8*(12), e1003091.

Illingworth, C. J. R., L. Parts, S. Schiffels, G. Liti, and V. Mustonen (2012). Quantifying selection acting on a complex trait using allele frequency time series data. *Mol. Biol. Evol. 29*(4), 1187–1197.

Irausquin, S. J. and A. L. Hughes (2008). Distinctive pattern of sequence polymorphism in the NS3 protein of hepatitis C virus type 1b reflects conflicting evolutionary pressures. *J. Gen. Virol. 89*(8), 1921–1929.

Kaplan, N. L., T. Darden, and R. R. Hudson (1988). The coalescent process in models with selection. *Genetics 120*(3), 819–829.

Keeling, M. J. and P. Rohani (2008). *Modeling infectious diseases in humans and animals.* Princeton University Press.

Kent, S. J., C. S. Fernandez, C. Jane Dale, and M. P. Davenport (2005). Reversion of immune escape HIV variants upon transmission: insights into effective viral immunity. *Trends Microbiol. 13*(6), 243–246.

Ketterlinus, R., K. Wiegers, and R. Dernick (1993). Revertants of poliovirus escape mutants: new insights into antigenic structures. *Virology 192*(2), 525–533.

Kim, Y. (2006). Allele frequency distribution under recurrent selective sweeps. *Genetics 172*(3), 1967–1978.

Kim, Y. and W. Stephan (2000). Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics 155*(3), 1415–1427.

Kim, Y. and T. Wiehe (2009). Simulation of DNA sequence evolution under models of recent directional selection. *Brief. Bioinform. 10*(1), 84–96.

Kimura, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics 47*(6), 713–719.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol. 16*(2), 111–120.

Kingman, J. F. C. (1982). On the genealogy of large populations. *J. Appl. Probab. 19*, 27.

Koel, B. F., D. F. Burke, T. M. Bestebroer, S. v. d. Vliet, G. C. M. Zondag, G. Vervaet, E. Skepner, N. S. Lewis, M. I. J. Spronken, C. A. Russell, M. Y. Eropkin, A. C. Hurt, I. G. Barr, J. C. d. Jong, G. F. Rimmelzwaan, A. D. M. E. Osterhaus, R. A. M. Fouchier, and D. J. Smith (2013). Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science 342*(6161), 976–979.

Koelle, K., S. Cobey, B. Grenfell, and M. Pascual (2006). Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science 314*(5807), 1898–1903.

Koelle, K., M. Kamradt, and M. Pascual (2009). Understanding the dynamics of rapidly evolving pathogens through modeling the tempo of antigenic change: Influenza as a case study. *Epidemics 1*(2), 129–137.

Koelle, K. and D. A. Rasmussen (2012). Rates of coalescence for common epidemiological models at equilibrium. *J. R. Soc. Interface 9*(70), 997–1007.

Kosakovsky Pond, S. L. and S. D. W. Frost (2005). Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol. 22*(5), 1208 –1222.

Kosakovsky Pond, S. L., S. D. W. Frost, and S. V. Muse (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics 21*(5), 676–679.

Kosakovsky Pond, S. L., A. F. Poon, A. J. Leigh Brown, and S. D. Frost (2008). A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. *Mol. Biol. Evol. 25*(9), 1809 –1824.

Kryazhimskiy, S., G. A. Bazykin, J. Plotkin, and J. Dushoff (2008). Directionality in the evolution of influenza A haemagglutinin. *P. Roy. Soc. B: Biol. Sci. 275*(1650), 2455 –2464.

Kryazhimskiy, S. and J. B. Plotkin (2008). The population genetics of dN/dS. *PLoS Genet. 4*(12), e1000304.

Kucharski, A. and J. R. Gog (2011). Influenza emergence in the face of evolutionary constraints. *P. Roy. Soc. B: Biol. Sci. 279*(1729), 645–652.

Kuhner, M. K., J. Yamato, and J. Felsenstein (1995). Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics 140*(4), 1421–1430.

Lemey, P., S. L. Kosakovsky Pond, A. J. Drummond, O. G. Pybus, B. Shapiro, H. Barroso, N. Taveira, and A. Rambaut (2007). Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. *PLoS Comput. Biol. 3*(2), e29.

Lemon, S. M., L. N. Binn, R. Marchwicki, P. C. Murphy, L.-H. Ping, R. w. Jansen, L. V. S. Asher, J. T. Stapleton, D. G. Taylor, and J. W. LeDuc (1990). In vivo replication and reversion to wild type of a neutralization-resistant antigenic variant of hepatitis A virus. *J Infect Dis. 161*(1), 7–13.

Leslie, A. J., K. J. Pfafferott, P. Chetty, R. Draenert, M. M. Addo, M. Feeney, Y. Tang, E. C. Holmes, T. Allen, J. G. Prado, M. Altfeld, C. Brander, C. Dixon, D. Ramduth, P. Jeena, S. A. Thomas, A. S. John, T. A. Roach, B. Kupfer, G. Luzzi, A. Edwards, G. Taylor, H. Lyall, G. Tudor-Williams, V. Novelli, J. Martinez-Picado, P. Kiepiela, B. D. Walker, and P. J. R. Goulder (2004). HIV evolution: CTL escape mutation and reversion after transmission. *Nat. Med. 10*(3), 282–289.

Li, W.-H. (1987). Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J. Mol. Evol. 24*(4), 337–345.

Li, W. H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol. 36*(1), 96–99.

Li, Z., Y. Jiang, P. Jiao, A. Wang, F. Zhao, G. Tian, X. Wang, K. Yu, Z. Bu, and H. Chen (2006). The NS1 gene contributes to the virulence of H5N1 avian influenza viruses. *J. Virol. 80*(22), 11115–11123.

Lu, L., S. J. Lycett, and A. J. L. Brown (2014). Reassortment patterns of avian influenza virus internal segments among different subtypes. *BMC Evol. Biol. 14*(1), 16.

Macken, C. A., R. J. Webby, and W. J. Bruno (2006). Genotype turnover by reassortment of replication complex genes from avian influenza a virus. *J. Gen. Virol. 87*(10), 2803–2815.

Marangon, S., M. Cecchinato, and I. Capua (2008). Use of vaccination in avian influenza control and eradication. *Zoonoses and Public Health 55*(1), 65–72.

Martínez, I., J. Dopazo, and J. A. Melero (1997). Antigenic structure of the human respiratory syncytial virus G glycoprotein and relevance of hypermutation events for the generation of antigenic variants. *J. Gen. Virol. 78*(10), 2419–2429.

Maruyama, T. and P. A. Fuerst (1985). Population bottlenecks and nonequilibrium models in population genetics. II. Number of alleles in a small population that was formed by a recent bottleneck. *Genetics 111*(3), 675–689.

Maynard-Smith, J., J. Haigh, et al. (1974). The hitch-hiking effect of a favourable gene. *Genet. Res. 23*(1), 23–35.

McDonald, J. H. and M. Kreitman (1991). Adaptive protein evolution at the *Adh* locus in *Drosophila. Nature 351*(6328), 652–654.

Messer, P. W. and D. A. Petrov (2013). Frequent adaptation and the McDonald-Kreitman test. *Proc. Natl. Acad. Sci. U.S.A. 110*(21), 8615–8620.

Meyer, A. G., E. T. Dawson, and C. O. Wilke (2013). Cross-species comparison of site-specific evolutionary-rate variation in influenza haemagglutinin. *Phil. Trans. R. Soc. B 368*(1614).

Minin, V. N., E. W. Bloomquist, and M. A. Suchard (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol. 25*(7), 1459–1471.

Miralles, R., P. J. Gerrish, A. Moya, and S. F. Elena (1999). Clonal interference and the evolution of RNA viruses. *Science 285*(5434), 1745–1747.

Mustonen, V. and M. Lässig (2009). From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends Genet. 25*(3), 111–119.

Neher, R. A. and O. Hallatschek (2013). Genealogies of rapidly adapting populations. *Proc. Natl. Acad. Sci. U.S.A. 110*(2), 437–442.

Neher, R. A. and B. I. Shraiman (2011). Genetic draft and quasi-neutrality in large facultatively sexual populations. *Genetics 188*(4), 975–996.

Neumann, G., M. A. Green, and C. A. Macken (2010). Evolution of highly pathogenic avian H5N1 influenza viruses and the emergence of dominant variants. *J. Gen. Virol. 91*(8), 1984–1995.

Nielsen, R. (2005). Molecular signatures of natural selection. *Annu. Rev. Genet. 39*, 197–218.

Nielsen, R. and Z. Yang (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics 148*(3), 929 –936.

Nielsen, R. and Z. Yang (2003). Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol. Biol. Evol. 20*(8), 1231–1239.

Nobusawa, E., T. Aoyama, H. Kato, Y. Suzuki, Y. Tateno, and K. Nakajima (1991). Comparison of complete amino acid sequences and receptor-binding properties among 13 serotypes of hemagglutinins of influenza A viruses. *Virology 182*(2), 475–485.

Obenauer, J. C., J. Denson, P. K. Mehta, X. Su, S. Mukatira, D. B. Finkelstein, X. Xu, J. Wang, J. Ma, Y. Fan, K. M. Rakestraw, R. G. Webster, E. Hoffmann, S. Krauss, J. Zheng, Z. Zhang, and C. W. Naeve (2006). Large-scale sequence analysis of avian influenza isolates. *Science 311*(5767), 1576–1580.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature 246*(5428), 96–98.

Otto, S. P. and M. C. Whitlock (1997). The probability of fixation in populations of changing size. *Genetics 146*(2), 723–733.

Palmer, D., J. Frater, R. Phillips, A. R. McLean, and G. McVean (2013). Integrating genealogical and dynamical modelling to infer escape and reversion rates in HIV epitopes. *Proc. R. Soc. B 280*(1762), 20130696.

Paradis, E., J. Claude, and K. Strimmer (2004). APE: Analyses of phylogenetics and evolution in r language. *Bioinformatics 20*(2), 289–290.

Park, K. J., H.-i. Kwon, M.-S. Song, P. N. Q. Pascua, Y. H. Baek, J. H. Lee, H.-L. Jang, J.-Y. Lim, I.-P. Mo, H.-J. Moon, C.-J. Kim, and Y. K. Choi (2011). Rapid evolution of low-pathogenic H9N2 avian influenza viruses following poultry vaccination programmes. *J. Gen. Virol. 92*(1), 36–50.

Parrish, C. R., C. F. Aquadro, M. L. Strassheim, J. F. Evermann, J. Y. Sgro, and H. O. Mohammed (1991). Rapid antigenic-type replacement and DNA sequence evolution of canine parvovirus. *J. Virol. 65*(12), 6544–6552.

Patwa, Z. and L. M. Wahl (2008). The fixation probability of beneficial mutations. *J. R. Soc. Interface 5*(28), 1279–1289.

Peck, J. R. (1994). A ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex. *Genetics 137*(2), 597–606.

Petravic, J., L. Loh, S. J. Kent, and M. P. Davenport (2008). CD4+ target cell availability determines the dynamics of immune escape and reversion in vivo. *J. Virol. 82*(8), 4091–4101.

Piganeau, G. and A. Eyre-Walker (2003). Estimating the distribution of fitness effects from DNA sequence data: Implications for the molecular clock. *Proc. Natl. Acad. Sci. U.S.A. 100*(18), 10335 –10340.

Pretorius, M. A., S. v. Niekerk, S. Tempia, J. Moyes, C. Cohen, S. A. Madhi, and M. Venter (2013). Replacement and positive evolution of subtype A and B respiratory syncytial virus G-protein genotypes from 1997–2012 in south africa. *J Infect Dis. 208*(suppl 3), S227–S237.

Pybus, O. G. and A. Rambaut (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet. 10*(8), 540–550.

Pybus, O. G., A. Rambaut, and P. H. Harvey (2000). An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics 155*(3), 1429–1437.

Rasmussen, D. A., O. Ratmann, and K. Koelle (2011). Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput. Biol. 7*(8), e1002136.

Rodrigue, N. (2013). On the statistical interpretation of site-specific variables in phylogeny-based substitution models. *Genetics 193*(2), 557–564.

Rouzine, I. M., J. Wakeley, and J. M. Coffin (2003). The solitary wave of asexual evolution. *Proc. Natl. Acad. Sci. U.S.A. 100*(2), 587–592.

Rozen, D. E., J. G. de Visser, and P. J. Gerrish (2002). Fitness effects of fixed beneficial mutations in microbial populations. *Curr. Biol. 12*(12), 1040–1045.

Russell, C. A., T. C. Jones, I. G. Barr, N. J. Cox, R. J. Garten, V. Gregory, I. D. Gust, A. W. Hampson, A. J. Hay, A. C. Hurt, J. C. de Jong, A. Kelso, A. I. Klimov, T. Kageyama, N. Komadina, A. S. Lapedes, Y. P. Lin, A. Mosterin, M. Obuchi, T. Odagiri, A. D. M. E. Osterhaus, G. F. Rimmelzwaan, M. W. Shaw, E. Skepner, K. Stohr, M. Tashiro, R. A. M. Fouchier, and D. J. Smith (2008). The global circulation of seasonal influenza A (H3N2) viruses. *Science 320*(5874), 340–346.

Sanjuán, R., A. Moya, and S. F. Elena (2004). The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc. Natl. Acad. Sci. U.S.A. 101*(22), 8396 –8401.

Sawyer, S. A. and D. L. Hartl (1992). Population genetics of polymorphism and divergence. *Genetics 132*(4), 1161–1176.

Schiffels, S., G. J. Szöllősi, V. Mustonen, and M. Lässig (2011). Emergent neutrality in adaptive asexual evolution. *Genetics 189*(4), 1361–1375.

Scholtissek, C., J. Stech, S. Krauss, and R. G. Webster (2002). Cooperation between the hemagglutinin of avian viruses and the matrix protein of human influenza A viruses. *J. Virol. 76*(4), 1781–1786.

Seger, J., W. A. Smith, J. J. Perry, J. Hunn, Z. A. Kaliszewska, L. L. Sala, L. Pozzi, V. J. Rowntree, and F. R. Adler (2010). Gene genealogies strongly distorted by weakly interfering mutations in constant environments. *Genetics 184*(2), 529–545.

Seoighe, C., F. Ketwaroo, V. Pillay, K. Scheffler, N. Wood, R. Duffet, M. Zvelebil, N. Martinson, J. McIntyre, L. Morris, and W. Hide (2007). A model of directional selection applied to the evolution of drug resistance in HIV-1. *Mol. Biol. Evol. 24*(4), 1025–1031.

Shi, W.-f., A.-s. Dun, Z. Zhang, Y.-z. Zhang, G.-f. Yu, D.-m. Zhuang, and C.-d. Zhu (2009). Selection pressure on haemagglutinin genes of h9n2 influenza viruses from different hosts. *Virol. Sin. 24*(1), 65–70.

Shih, A. C.-C., T.-C. Hsiao, M.-S. Ho, and W.-H. Li (2007). Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *Proc. Natl. Acad. Sci. U.S.A. 104*(15), 6283 –6288.

Silva, J. d. (2012). Antibody selection and amino acid reversions. *Evolution 66*(10), 3079–3087.

Smith, D. J., A. S. Lapedes, J. C. de Jong, T. M. Bestebroer, G. F. Rimmelzwaan, A. D. M. E. Osterhaus, and R. A. M. Fouchier (2004). Mapping the antigenic and genetic evolution of influenza virus. *Science 305*(5682), 371–376.

Smith, N. G. C. and A. Eyre-Walker (2002). Adaptive protein evolution in drosophila. *Nature 415*(6875), 1022–1024.

Stadler, T. and S. Bonhoeffer (2013). Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Phil. Trans. R. Soc. B 368*(1614).

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics 22*(21), 2688–2690.

Strelkowa, N. and M. Lässig (2012). Clonal interference in the evolution of influenza. *Genetics 192*(2), 671–682.

Sugita, S., Y. Yoshioka, S. Itamura, Y. Kanegae, K. Oguchi, T. Gojobori, K. Nerome, and A. Oya (1991). Molecular evolution of hemagglutinin genes of H1N1 swine and human influenza A viruses. *J. Mol. Evol. 32*(1), 16–23.

Suzuki, Y. (2006). Natural selection on the influenza virus genome. *Mol. Biol. Evol. 23*(10), 1902–1911.

Tamuri, A. U., M. Dos Reis, and R. A. Goldstein (2012). Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics 190*(3), 1101–1115.

Tamuri, A. U., M. dos Reis, A. J. Hay, and R. A. Goldstein (2009). Identifying changes in selective constraints: Host shifts in influenza. *PLoS Comput. Biol. 5*(11), e1000564.

Toll-Riera, M., S. Laurie, and M. M. Albà (2011). Lineage-specific variation in intensity of natural selection in mammals. *Mol. Biol. Evol. 28*(1), 383–398.

Tria, F., M. Lässig, L. Peliti, and S. Franz (2005). A minimal stochastic model for influenza evolution. *J. Stat. Mech. 2005*(07), P07008–P07008.

Uecker, H. and J. Hermisson (2011). On the fixation process of a beneficial mutation in a variable environment. *Genetics 188*(4), 915 –930.

Volz, E. M., S. L. K. Pond, M. J. Ward, A. J. L. Brown, and S. D. W. Frost (2009). Phylodynamics of infectious disease epidemics. *Genetics 183*(4), 1421–1430.

Wagner, A. (2014). A genotype network reveals homoplastic cycles of convergent evolution in influenza a (H3N2) haemagglutinin. *Proc. R. Soc. B 281*(1786), 20132763.

Walczak, A. M., L. E. Nicolaisen, J. B. Plotkin, and M. M. Desai (2012). The structure of genealogies in the presence of purifying selection: A fitness-class coalescent. *Genetics 190*(2), 753–779.

Webster, R. G., W. J. Bean, O. T. Gorman, T. M. Chambers, and Y. Kawaoka (1992). Evolution and ecology of influenza A viruses. *Microbiol. Mol. Biol. Rev. 56*(1), 152–179.

Wikramaratna, P. S., M. Sandeman, M. Recker, and S. Gupta (2013). The antigenic evolution of influenza: drift or thrift? *Phil. Trans. R. Soc. B 368*(1614), 20120200.

World Health Organization (2009). Continuing progress towards a unified nomenclature for the highly pathogenic H5N1 avian influenza viruses: divergence of clade 2-2 viruses. *Influenza Other Respir. Viruses 3*(2), 59–62.

World Health Organization (2012). Continued evolution of highly pathogenic avian influenza A (H5N1): updated nomenclature: H5N1 clade nomenclature update. *Influenza Other Respir. Viruses 6*(1), 1–5.

Worobey, M., G.-Z. Han, and A. Rambaut (2014). A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature 508*, 254–257.

Wright, S. (1931). Evolution in mendelian populations. *Genetics 16*(2), 97–159.

Wright, S. (1938). The distribution of gene frequencies under irreversible mutation. *Proc. Natl. Acad. Sci. U.S.A. 24*(7), 253.

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol. 15*(5), 568–573.

Yang, Z. (2000). Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *J. Mol. Evol. 51*(5), 423–432.

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol. 24*(8), 1586–1591.

Zeng, K. and B. Charlesworth (2010). The effects of demography and linkage on the estimation of selection and mutation parameters. *Genetics 186*(4), 1411–1424.

Zhao, L., M. Lascoux, A. D. J. Overall, and D. Waxman (2013). The characteristic trajectory of a fixing allele: A consequence of fictitious selection that arises from conditioning. *Genetics 195*(3), 993–1006.

Zinder, D., T. Bedford, S. Gupta, and M. Pascual (2013). The roles of competition and mutation in shaping antigenic and genetic diversity in influenza. *PLoS Pathog. 9*(1), e1003104.

Zlateva, K. T., P. Lemey, A.-M. Vandamme, and M. V. Ranst (2004). Molecular evolution and circulation patterns of human respiratory syncytial virus subgroup A: Positively selected sites in the attachment G glycoprotein. *J. Virol. 78*(9), 4675–4683.

Appendix A

Additional figures for linkage model

In this appendix, we provide additional information about the model described in Chapter 2. They were included as an additional file in Chan et al. (2013). The figures show the typical behaviour of transitory mutations (Figure A.1) and the behaviour of the statistics D_1, D_2, D_3 in populations evolving under different conditions (Figures A.2–A.20).

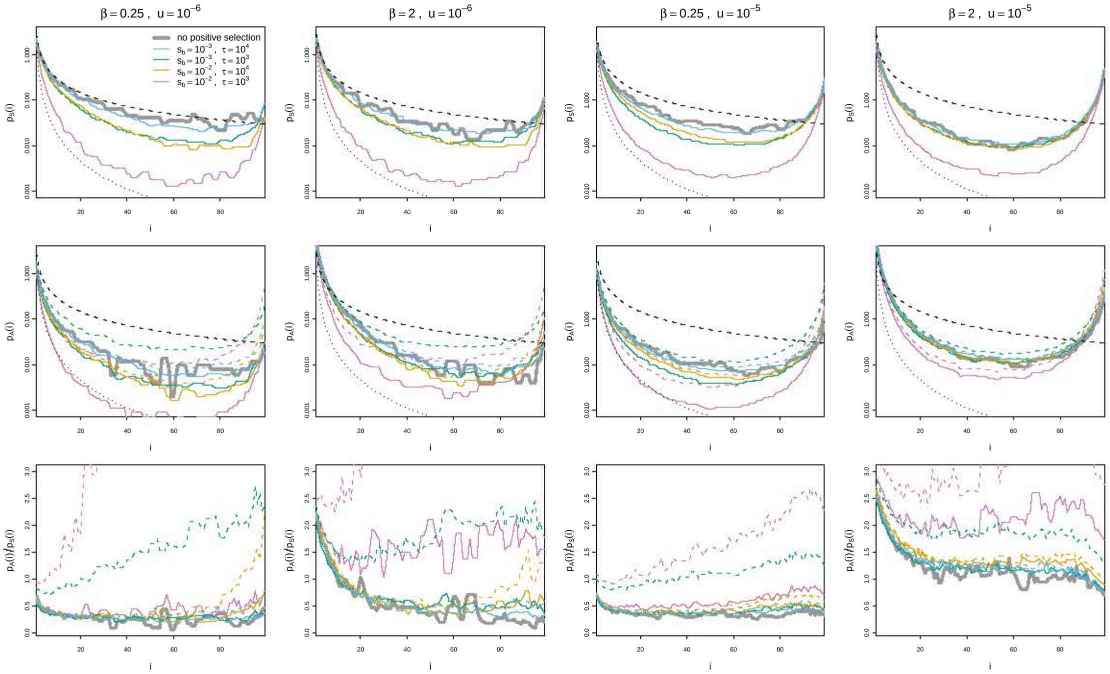

Figure A.1: The effect of linkage of the site frequency spectrum. The synonymous site frequency spectrum (top row), non-synonymous site-frequency spectrum (middle row), and the ratio of non-synonymous to synonymous frequency spectrum (bottom) is shown for $\beta = 0.25$ with mutation rates $u = 10^{-6}$ and 10^{-5}. All curves are averaged over 500 replicates, under conditions of only negative selection (grey), and different conditions of positive selection (coloured lines). Black dashed lines show the expected behaviour of the neutral site frequency spectrum under independently segregating sites (θ/i) and under black dotted lines indicate the leading order behaviour expected under constant adaptation (θ/i^2). In the bottom two rows, solid lines show the average non-synonymous to synonymous ratio for only negatively selected sites, whereas dashed lines show the ratio across both positively and negatively selected sites.

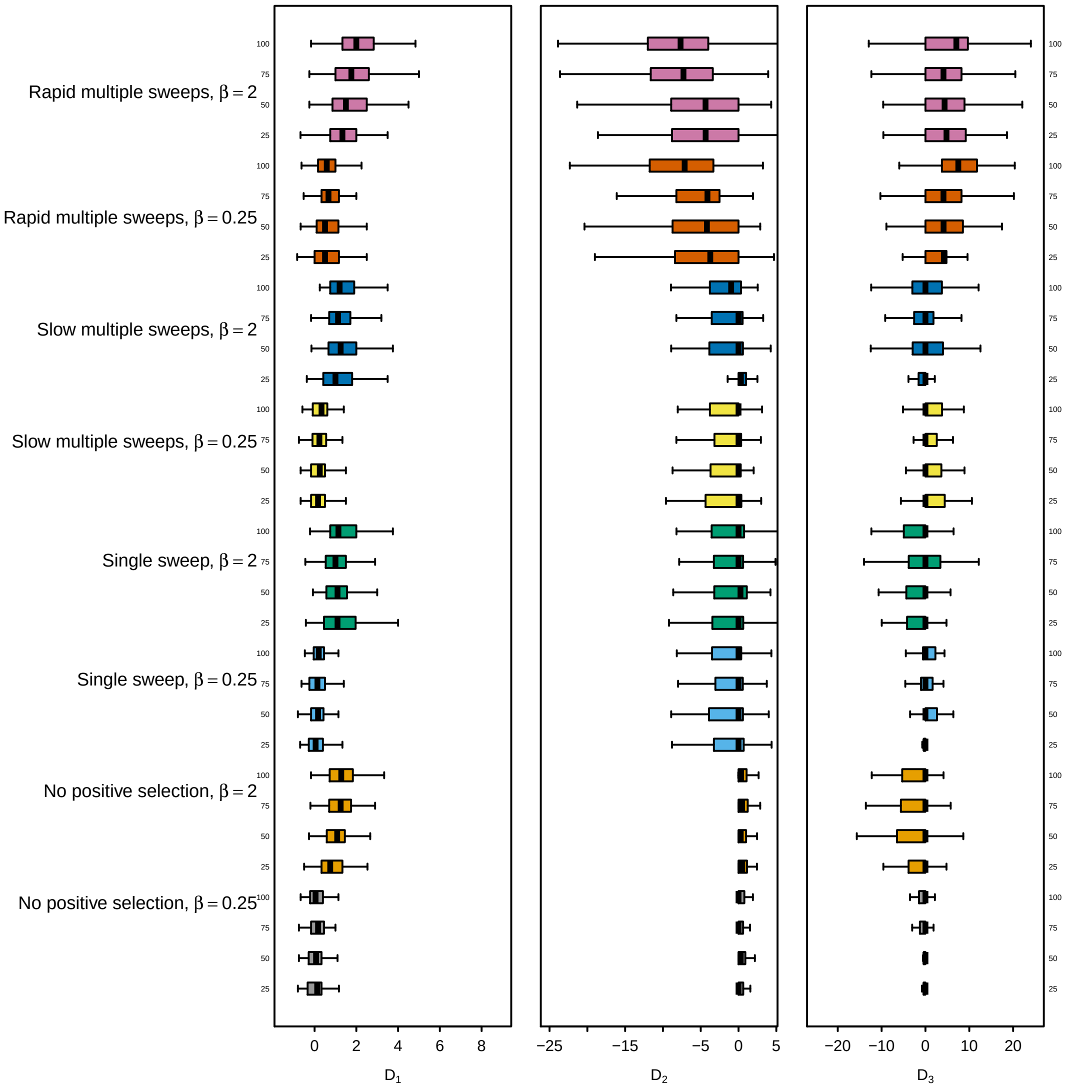

Figure A.2: The effect of sample size on D_1, D_2, D_3. Boxplots summarise D_1, D_2 and D_3 values for different sample sizes from independent simulations at $t = 6N$ and $u = 10^{-6}$. For each parameter combination (indicated by the colour), we show results for sample sizes of 25, 50, 75 and 100 (y-axis)

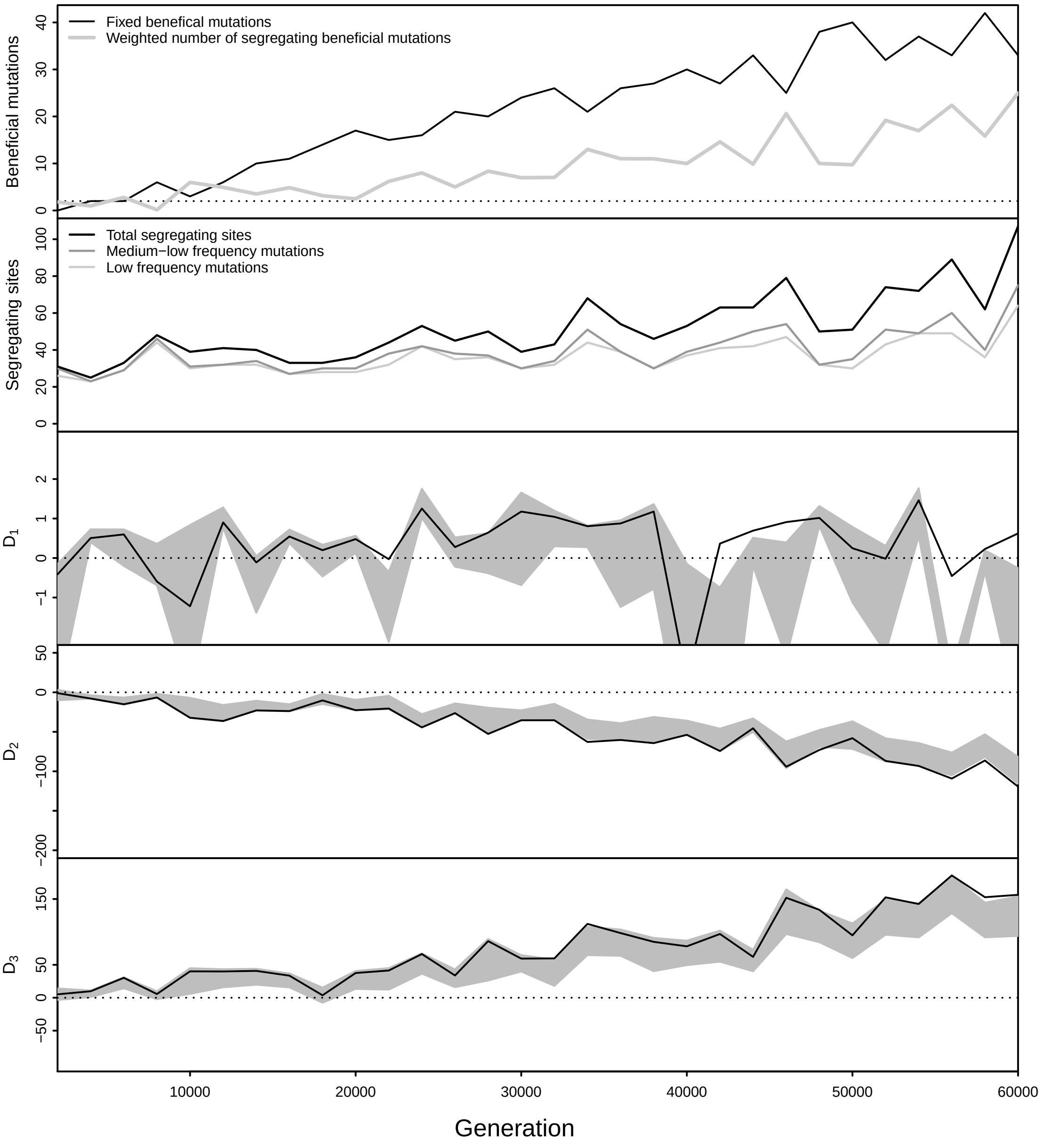

Figure A.3: Sequence statistics of a population evolving with interfering recurrent sweeps background selection: $u = 10^{-5}$, $\tau = 1000$, $s_b = 10^{-2}$, $\beta = 0.25$, $\bar{s} = 4.4 \times 10^{-1}$ and $N = 10000$. Bootstraps for D_1, D_2 and D_3 were constructed using the percentile method with 1000 replicates (grey shaded area).

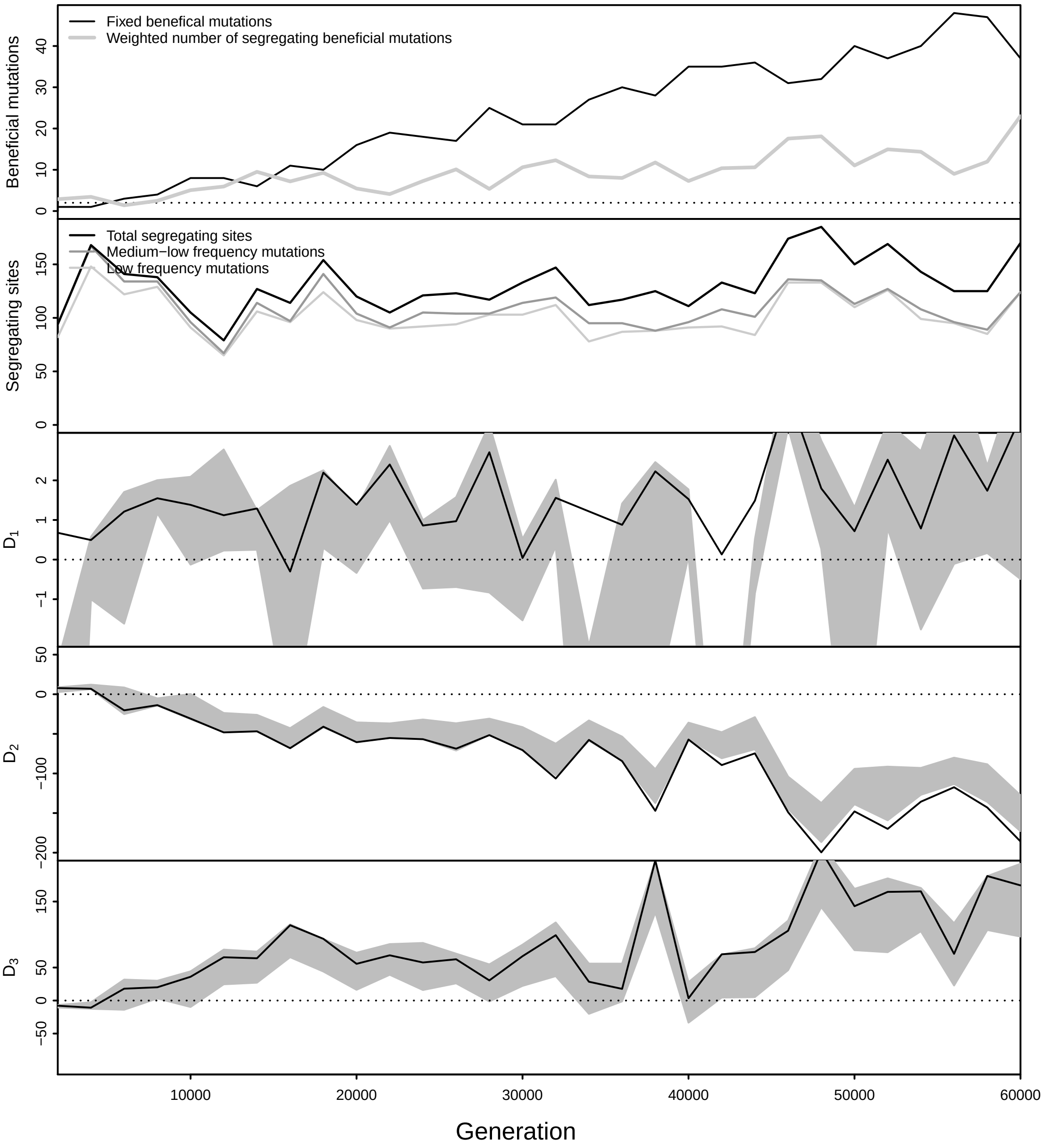

Figure A.4: Sequence statistics of a population evolving with interfering recurrent sweeps and background selection and hitch-hiking: $u = 10^{-5}$, $\tau = 1000$, $s_b = 10^{-2}$, $\beta = 2$, $\bar{s} = 7 \times 10^{-4}$ and $N = 10000$.

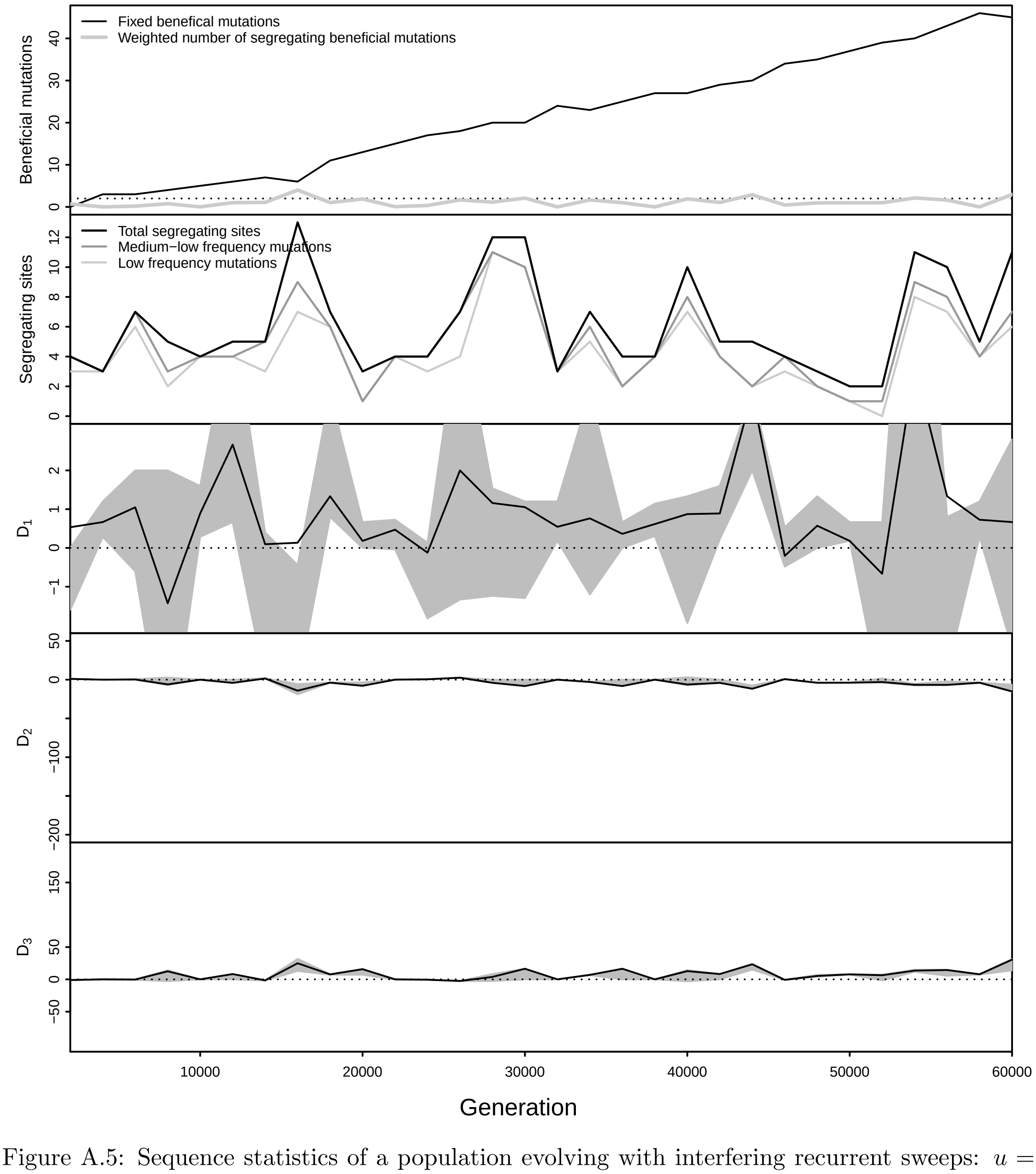

Figure A.5: Sequence statistics of a population evolving with interfering recurrent sweeps: $u = 10^{-6}$, $\tau = 1000$, $s_b = 10^{-2}$, $\beta = 0.25$, $\bar{s} = 4.4 \times 10^{-1}$ and $N = 10000$.

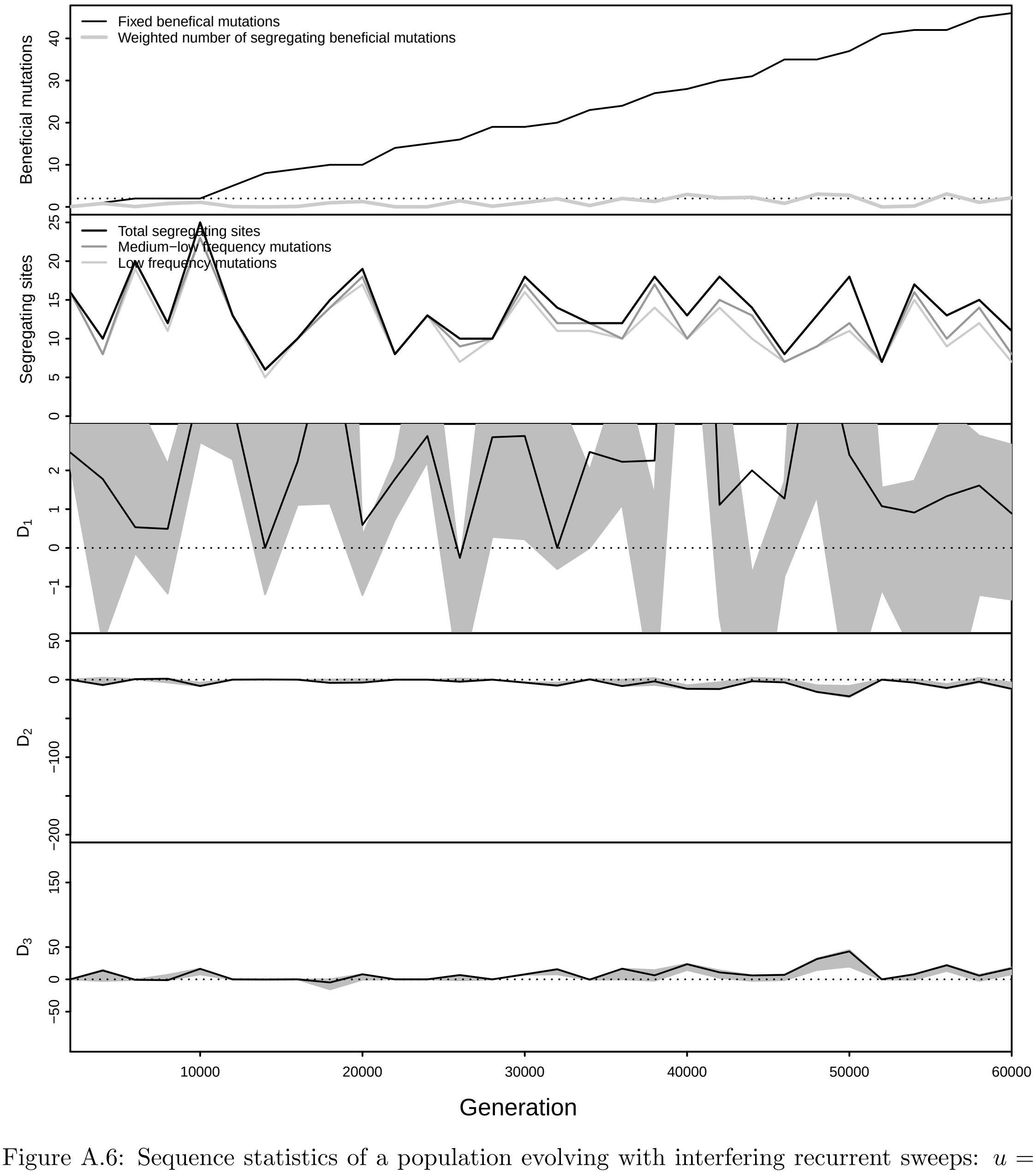

Figure A.6: Sequence statistics of a population evolving with interfering recurrent sweeps: $u = 10^{-6}$, $\tau = 1000$, $s_b = 10^{-2}$, $\beta = 2$, $\bar{s} = 7 \times 10^{-4}$ and $N = 10000$.

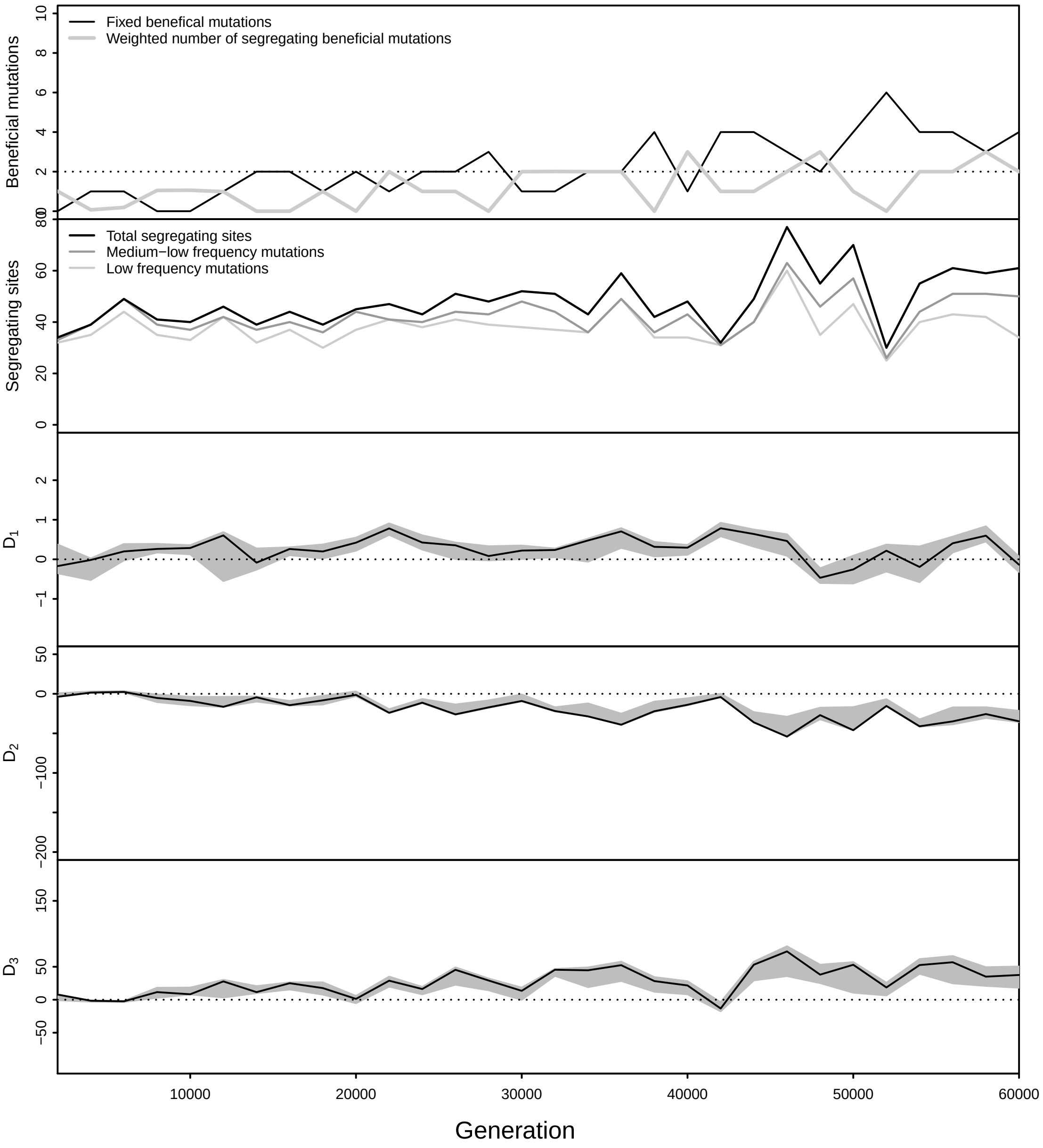

Figure A.7: Sequence statistics of a population evolving with slow recurrent sweeps and background selection, $u = 10^{-5}$, $\tau = 10000$, $s_b = 10^{-2}$, $N = 10000$, $\beta = 0.25$, $\bar{s} = 4.4 \times 10^{-1}$.

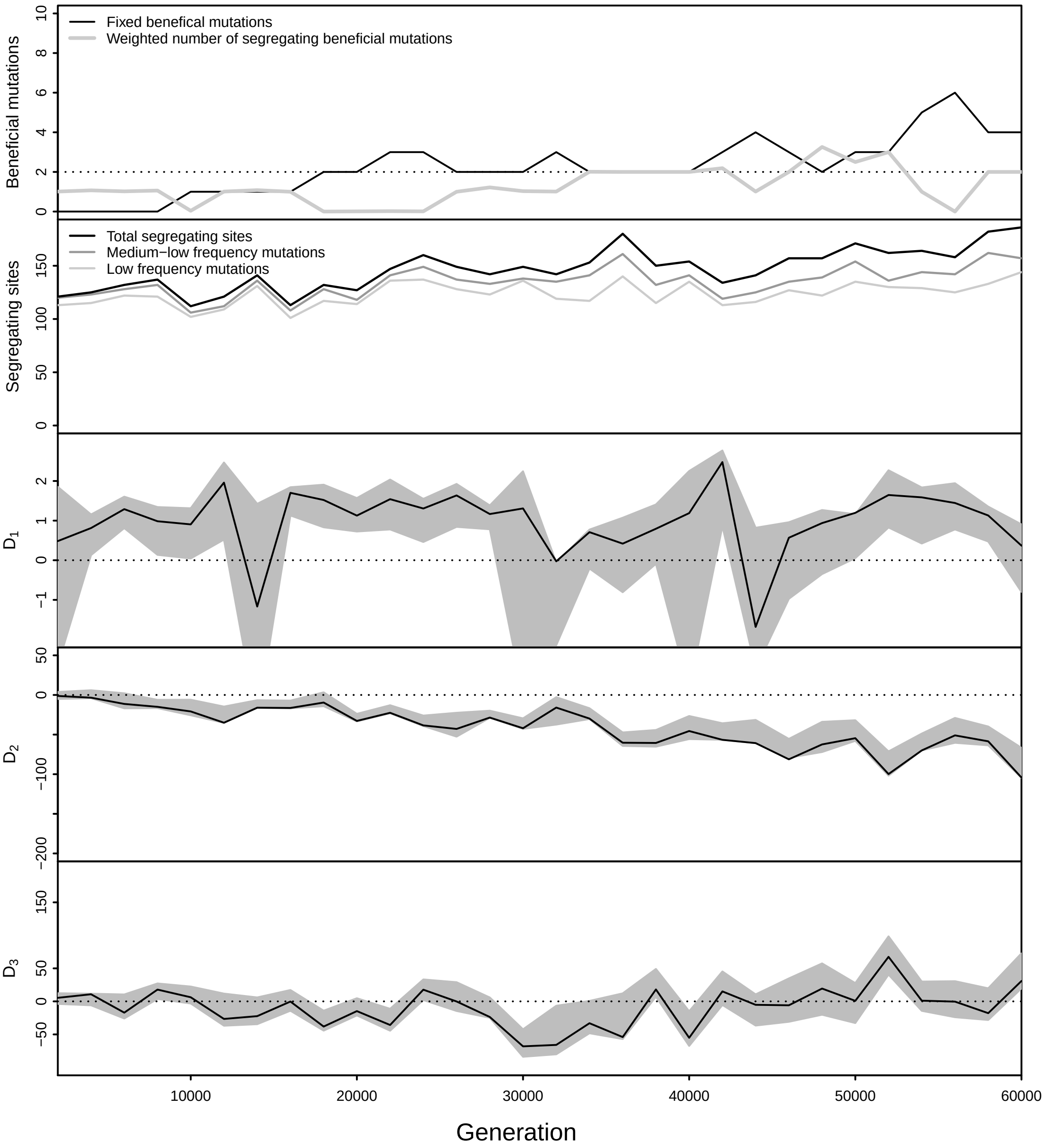

Figure A.8: Sequence statistics of a population evolving with slow recurrent sweeps, background selection and hitch-hiking, $u = 10^{-5}$, $\tau = 10000$, $s_b = 10^{-2}$, $N = 10000$, $\beta = 2$, $\bar{s} = 7 \times 10^{-4}$

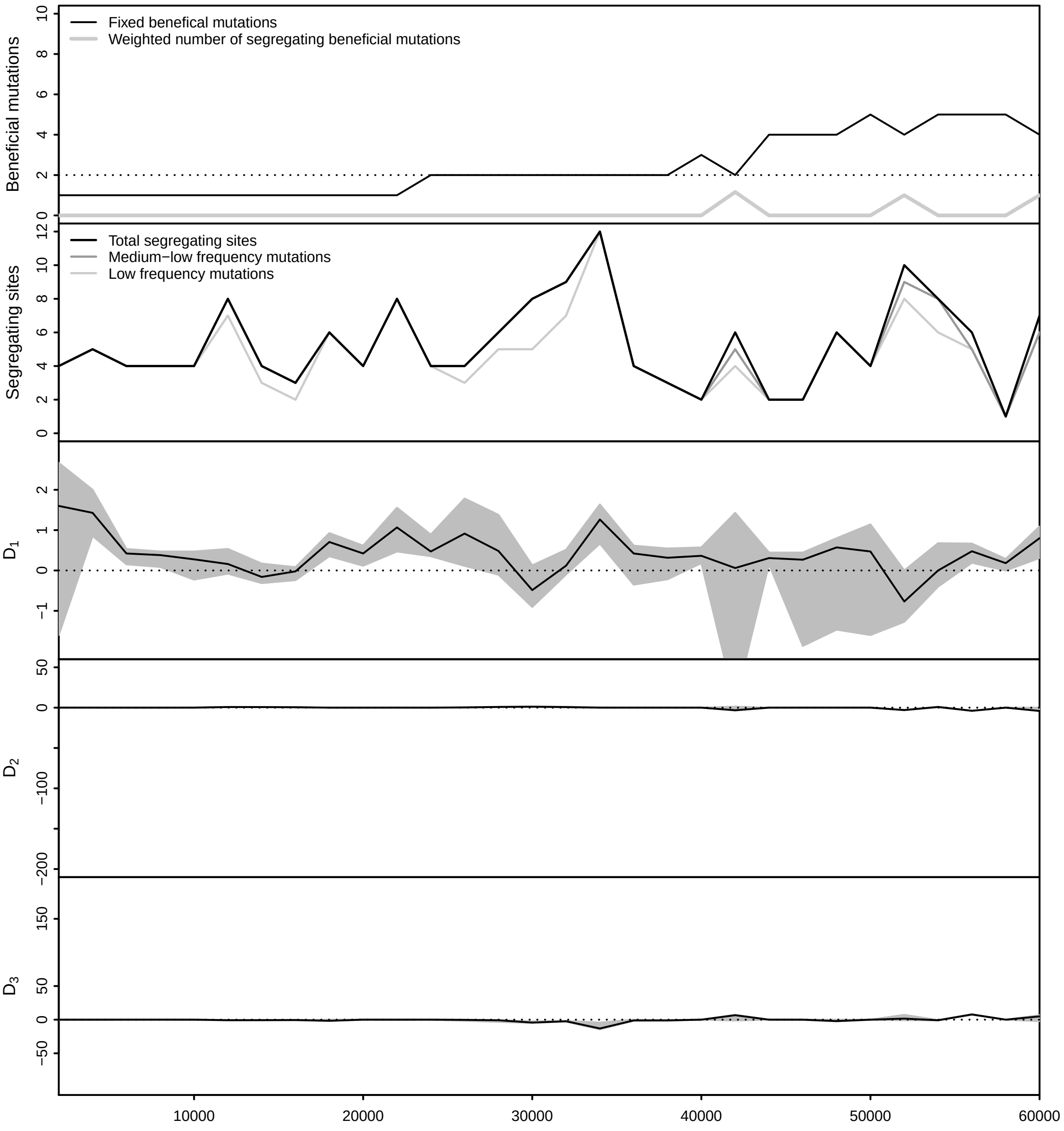

Figure A.9: Sequence statistics of a population evolving with slow recurrent sweeps, $u = 10^{-6}$, $\tau = 10000$, $s_b = 10^{-2}$, $N = 10000$, $\beta = 0.25$, $\bar{s} = 4.4 \times 10^{-1}$

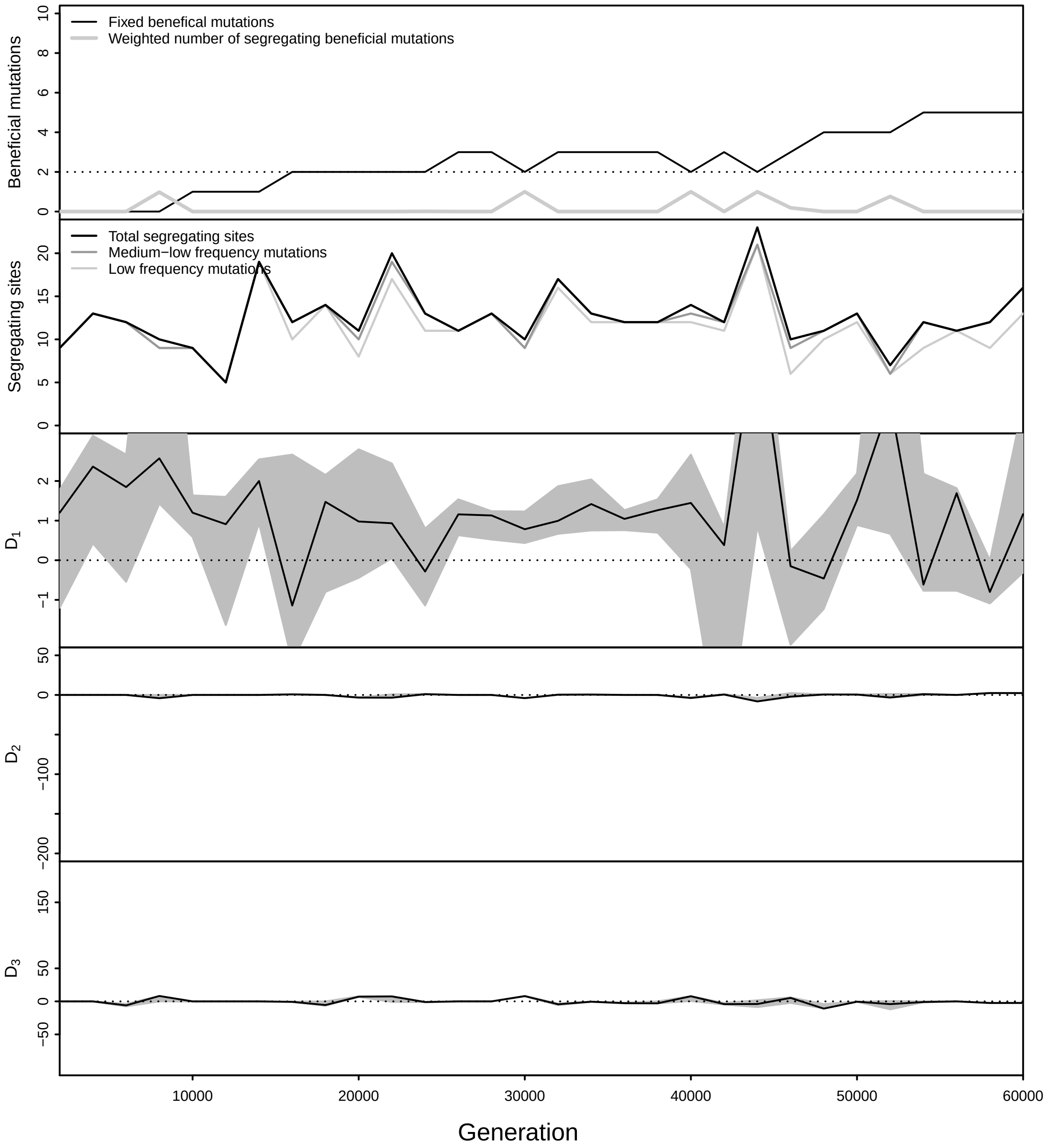

Figure A.10: Sequence statistics of a population evolving with slow recurrent sweeps, $u = 10^{-6}$, $\tau = 10000$, $s_b = 10^{-2}$, $N = 10000$, $\beta = 2$, $\bar{s} = 7 \times 10^{-4}$

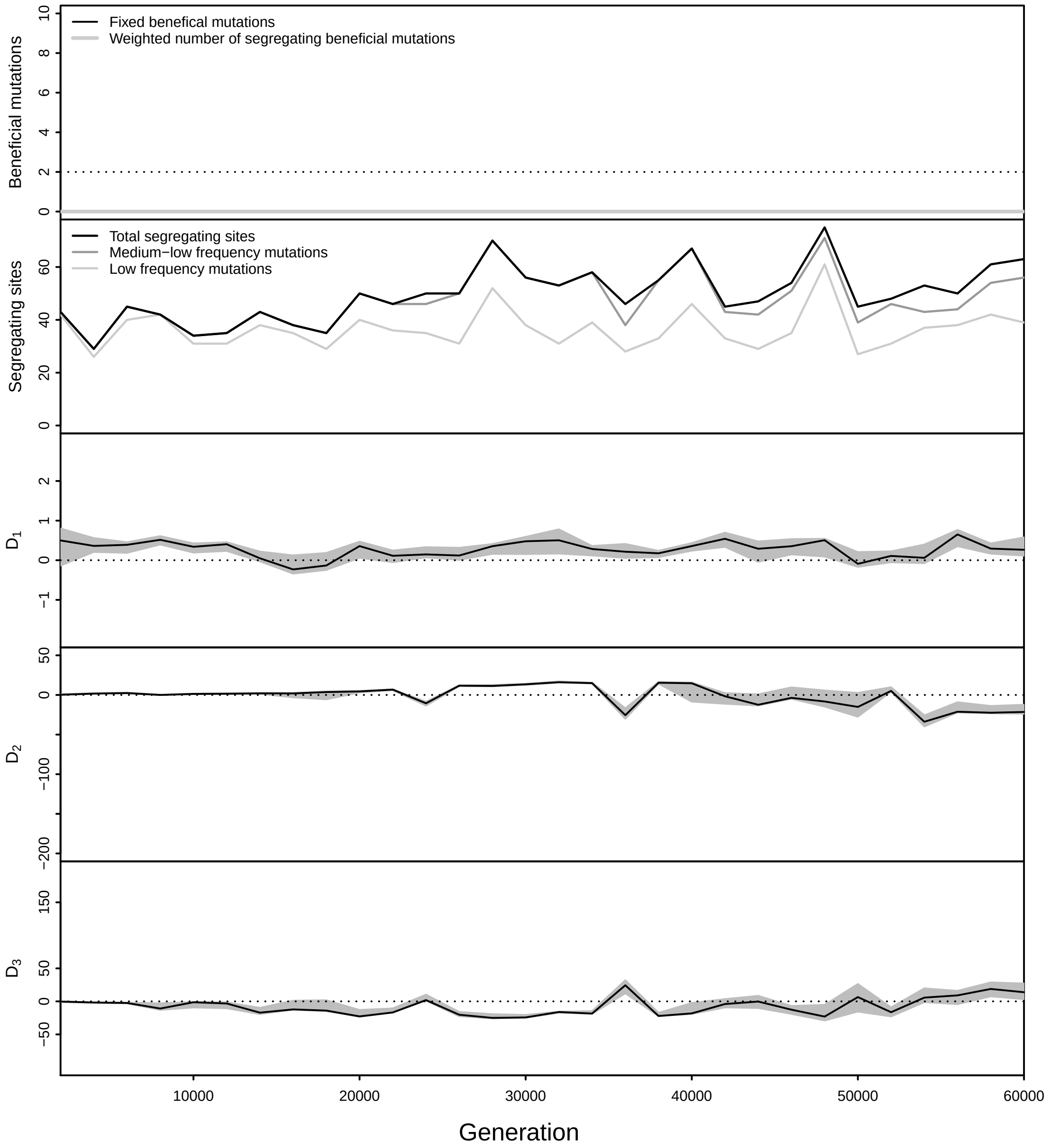

Figure A.11: Sequence statistics of a population evolving with no positive selection and background selection: $N = 10000$, $\beta = 0.25$, $\bar{s} = 4.4 \times 10^{-1}$, $u = 10^{-5}$.

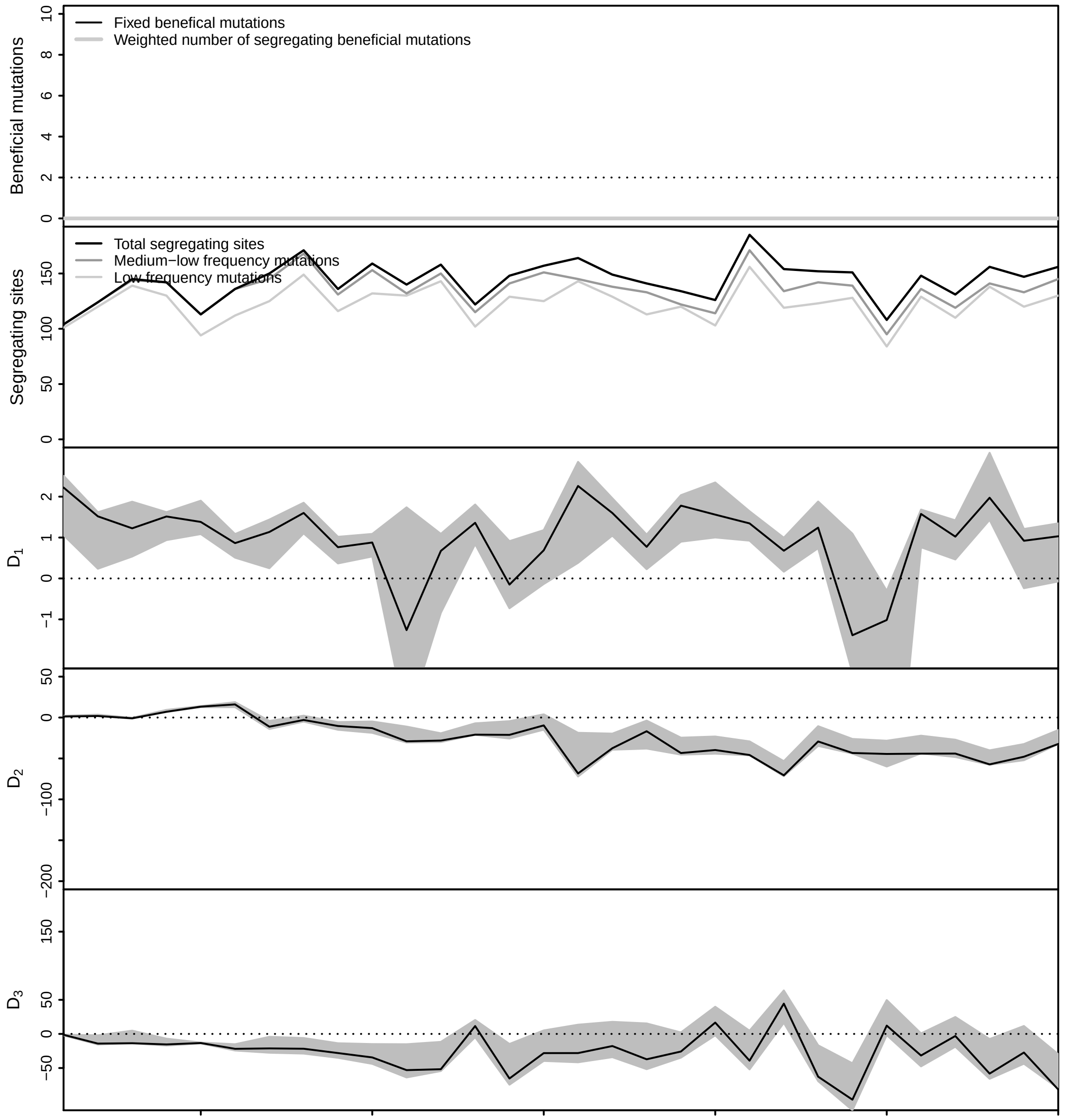

Figure A.12: Sequence statistics of a population evolving with no positive selection and high levels of background selection: $N = 10000$, $\beta = 2$, $\bar{s} = 7 \times 10^{-4}$, $u = 10^{-5}$.

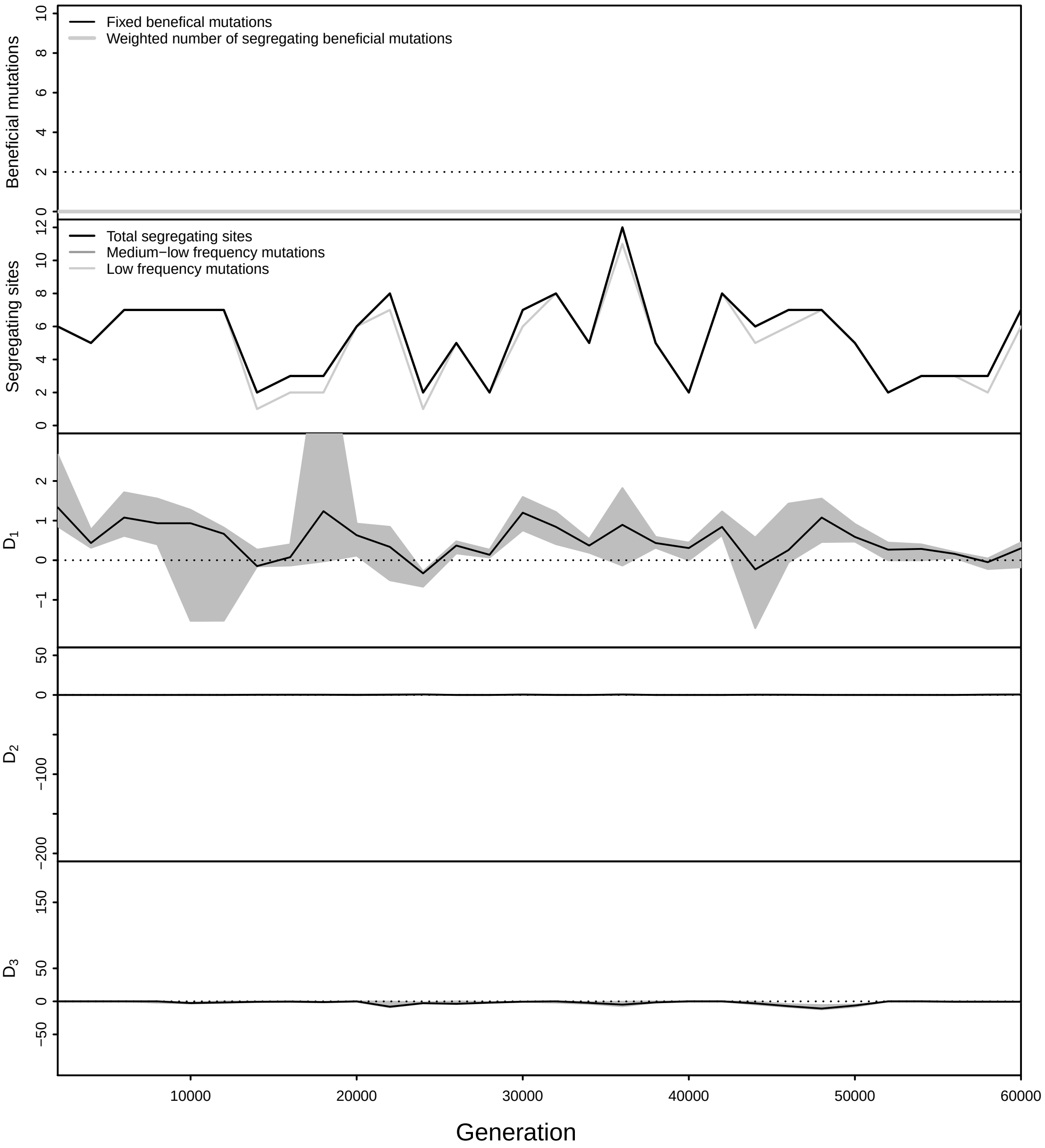

Figure A.13: Sequence statistics of a population evolving with no positive selection, $N = 10000$, $\beta = 0.25$, $\bar{s} = 4.4 \times 10^{-1}$, $u = 10^{-6}$.

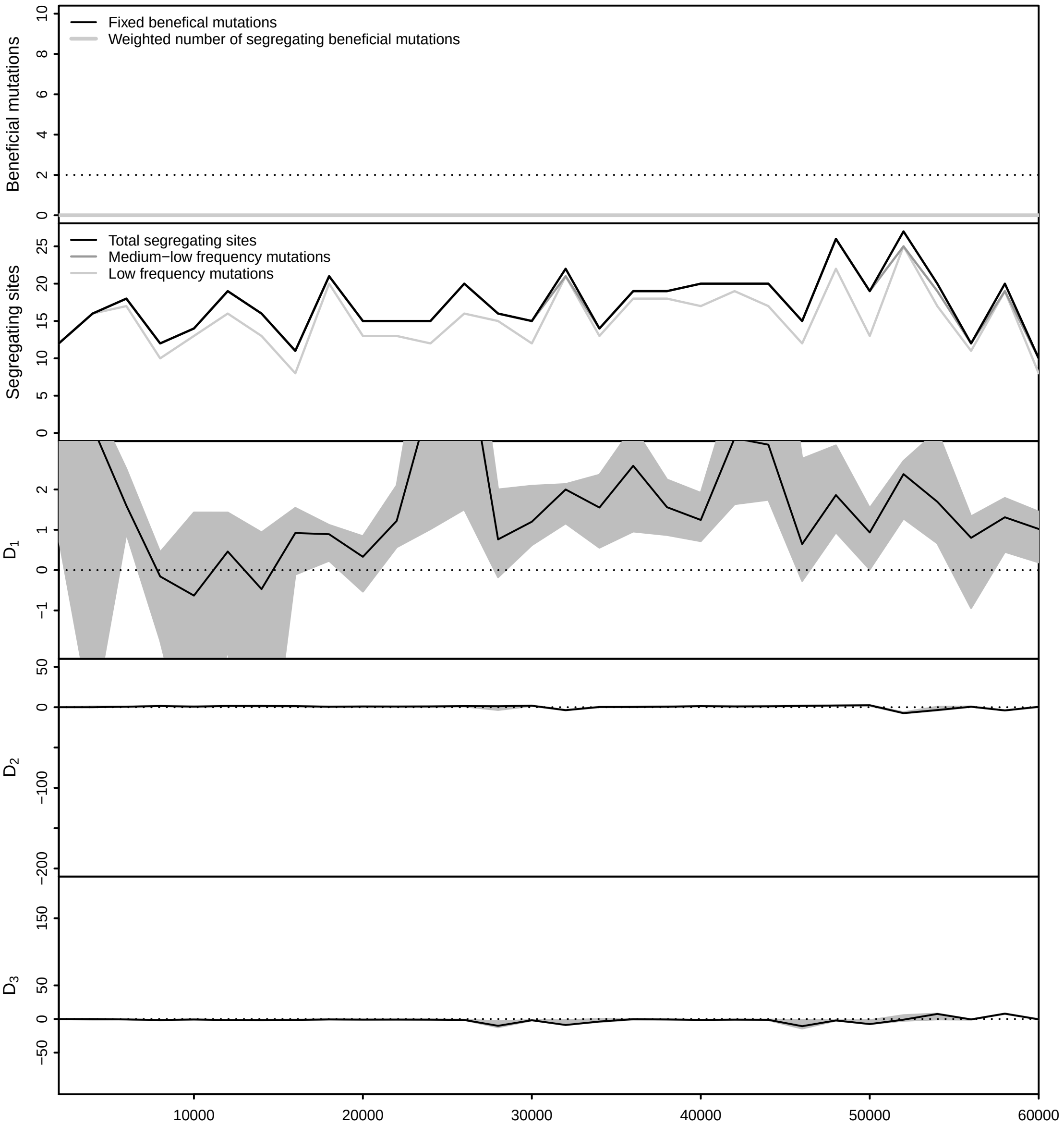

Figure A.14: Sequence statistics of a population evolving with no positive selection, $u = 10^{-6}$, $N = 10000$, $\beta = 2$, $\bar{s} = 7 \times 10^{-4}$, $u = 10^{-6}$.

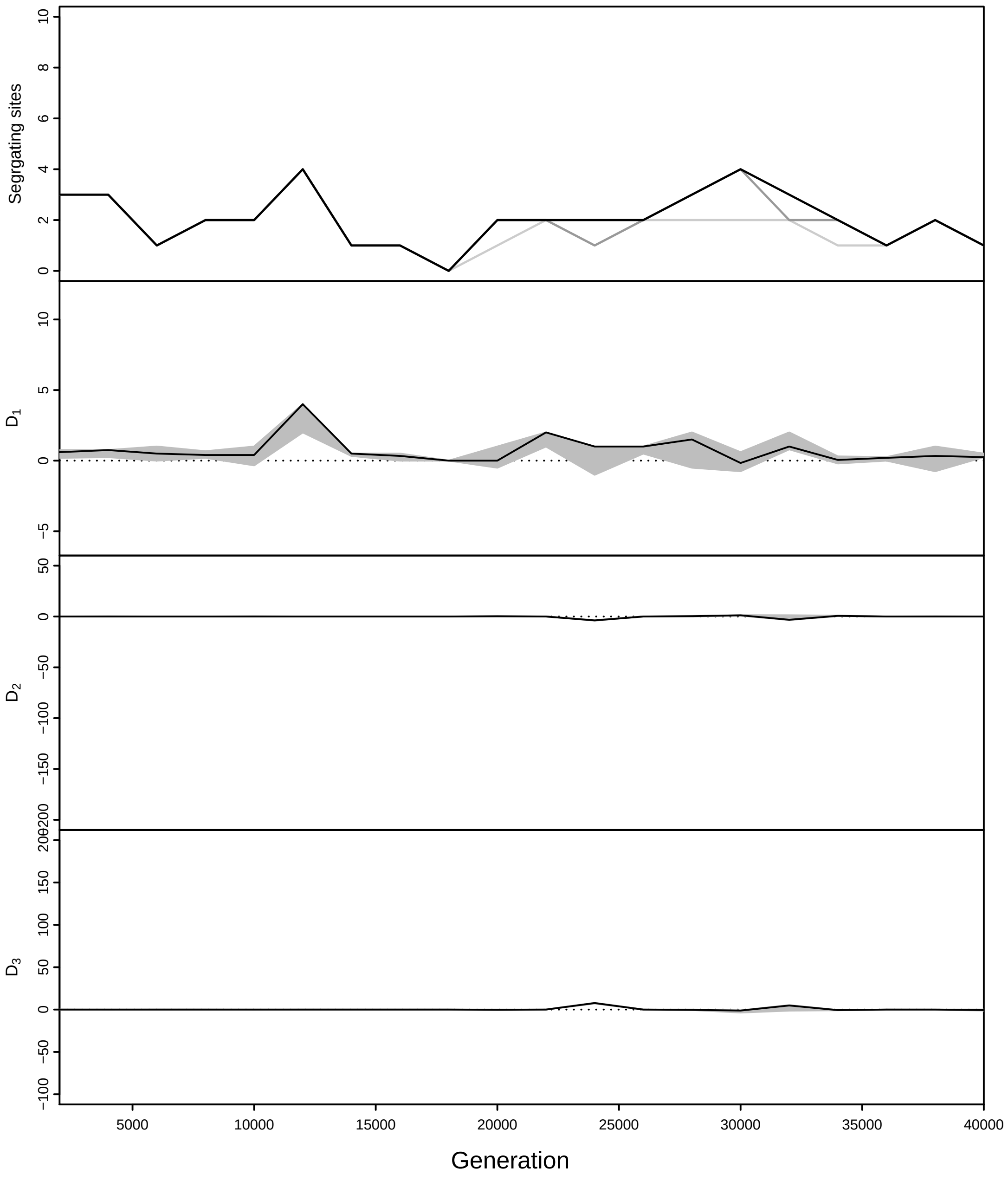

Figure A.15: Sequence statistics of a population evolving with no positive selection, $N = 2500$, $\beta = 0.25$, $\bar{s} = 4.4 \times 10^{-1}$, $u = 10^{-6}$.

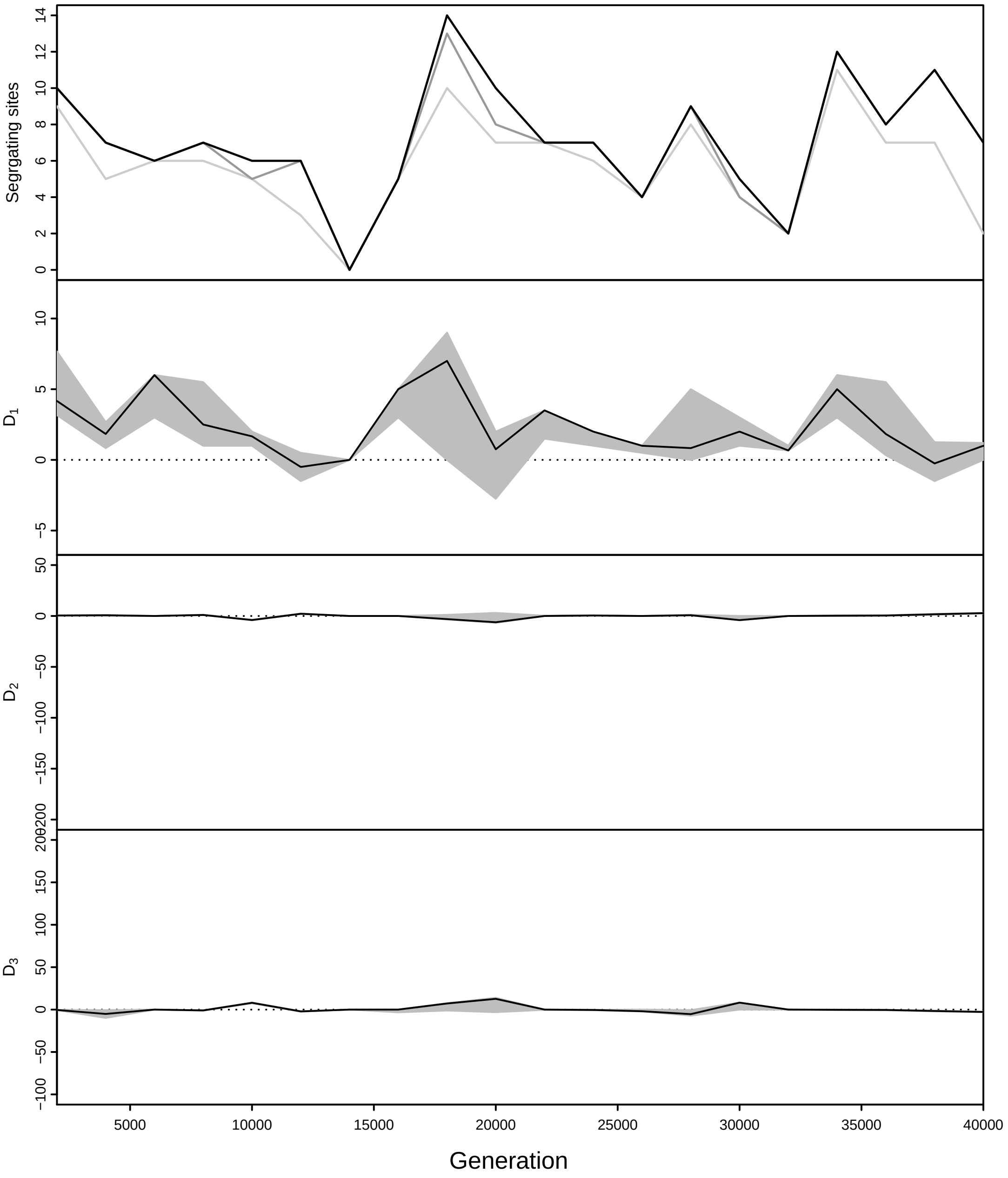

Figure A.16: Sequence statistics of a population evolving with no positive selection, $N = 2500$, $\beta = 2$, $\bar{s} = 7 \times 10^{-4}$, $u = 10^{-6}$.

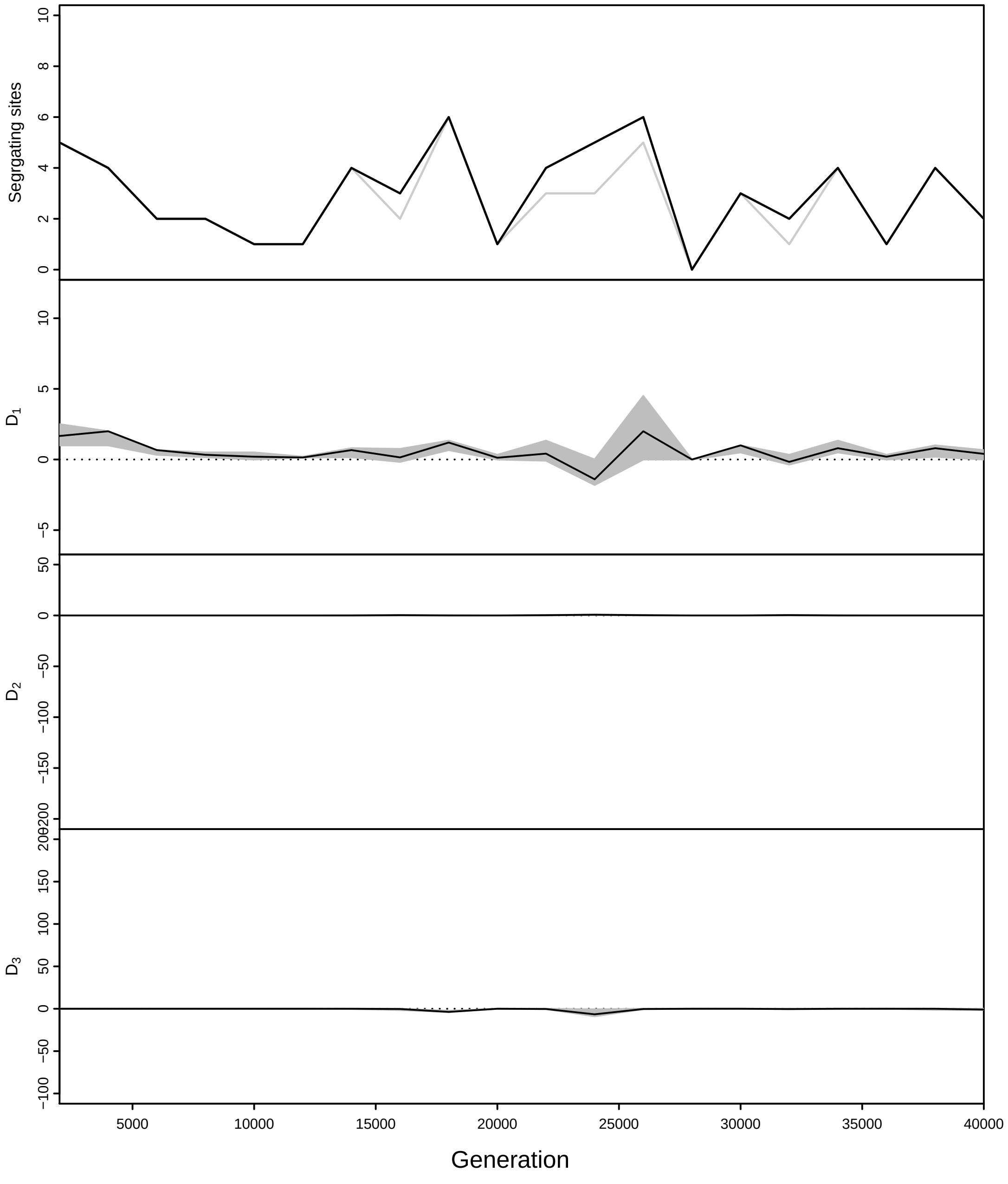

Figure A.17: Sequence statistics of a population evolving with no positive selection, $N = 5000$, $\beta = 0.25$, $\bar{s} = 4.4 \times 10^{-1}$, $u = 10^{-6}$.

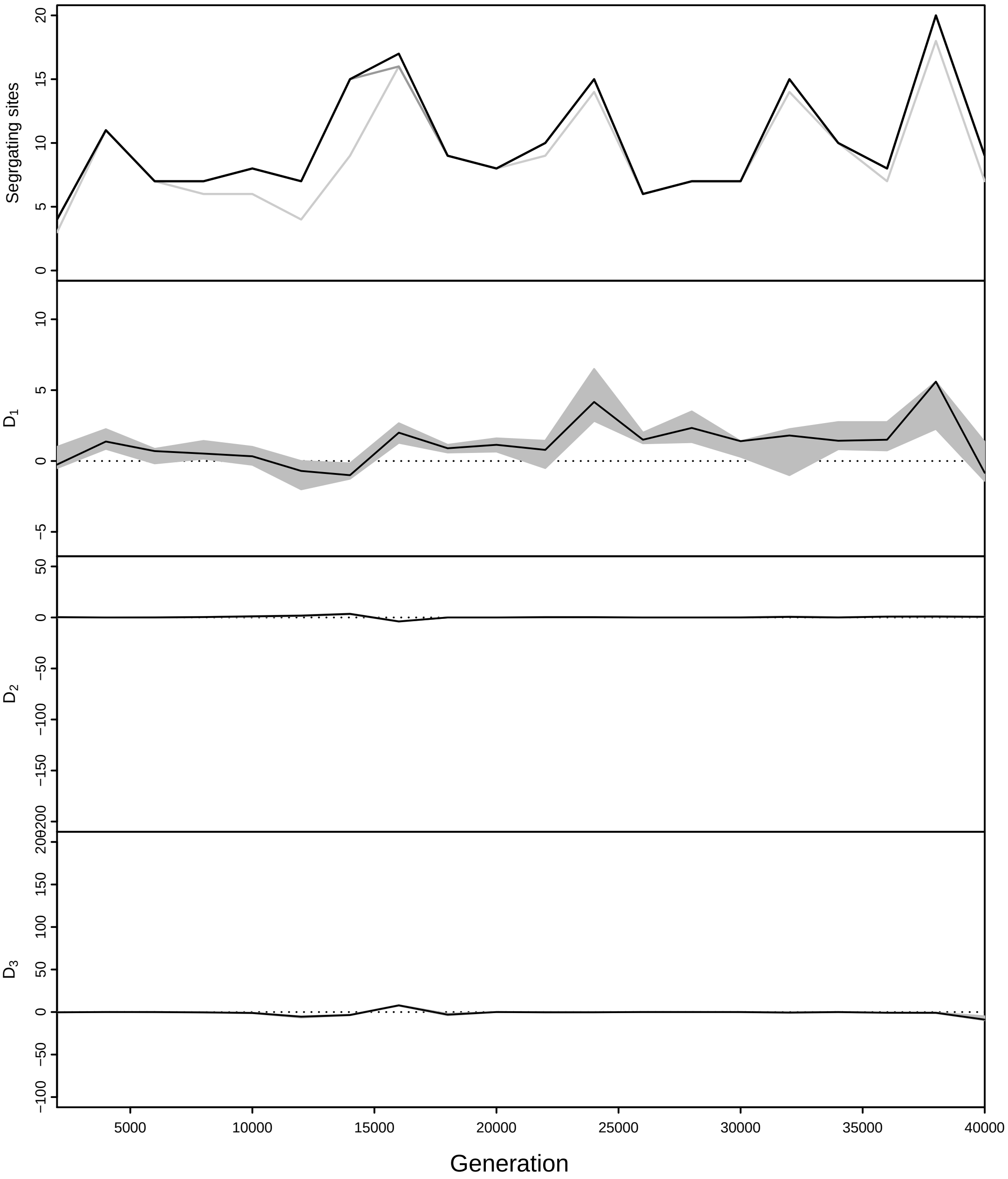

Figure A.18: Sequence statistics of a population evolving with no positive selection, $N = 5000$, $\beta = 2$, $\bar{s} = 7 \times 10^{-4}$, $u = 10^{-6}$.

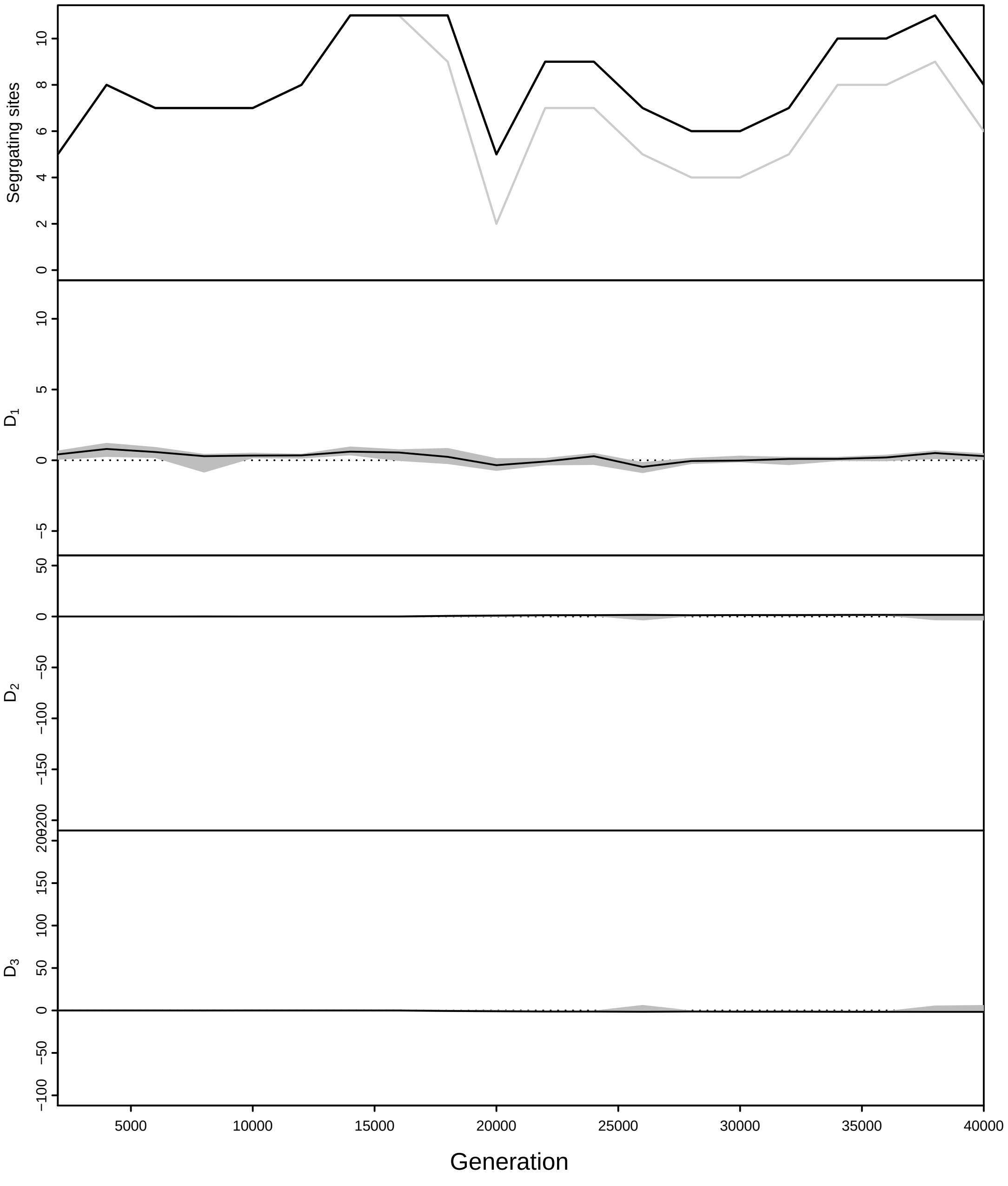

Figure A.19: Sequence statistics of a population evolving with no positive selection, $N = 20000$, $\beta = 0.25$, $\bar{s} = 4.4 \times 10^{-1}$, $u = 10^{-6}$.

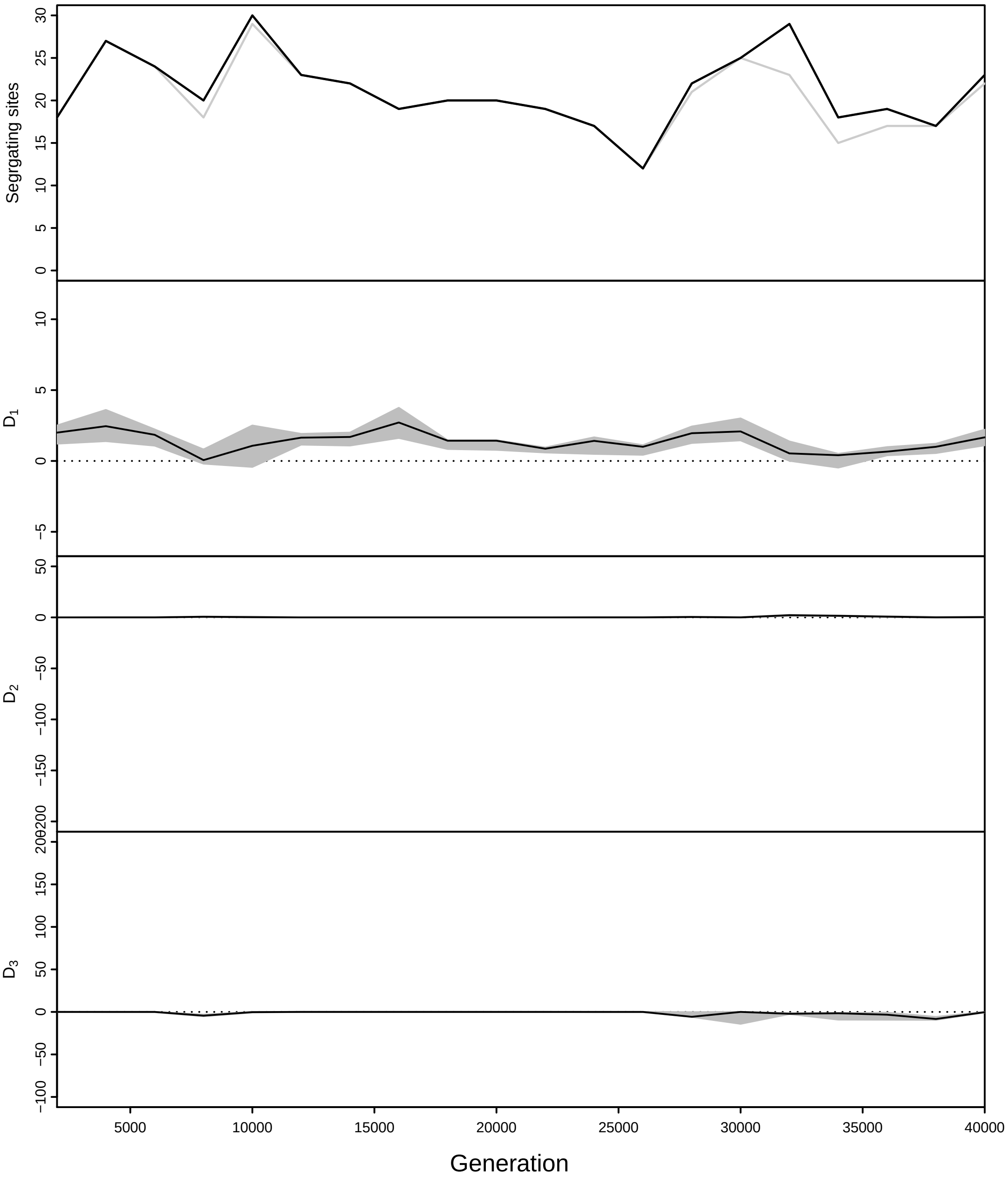

Figure A.20: Sequence statistics of a population evolving with no positive selection, $N = 20000$, $\beta = 2$, $\bar{s} = 7 \times 10^{-4}$, $u = 10^{-6}$.

Appendix B

Additional figures for the analysis of antigenic reversions

In this appendix, we provide additional information about the influenza A (H3N2) and RSV-A data sets used to examine antigenic reversions (Chapter 3). In Figure B.1 we show the number of sequences sampled in each year. To provide a sense of the typical dynamics of polymorphic sites, we show the frequency trajectory of the first observed amino acid at sites which are expected to be non-antigenic for H3N2 (Figure B.2) and RSV-A (Figure B.3).

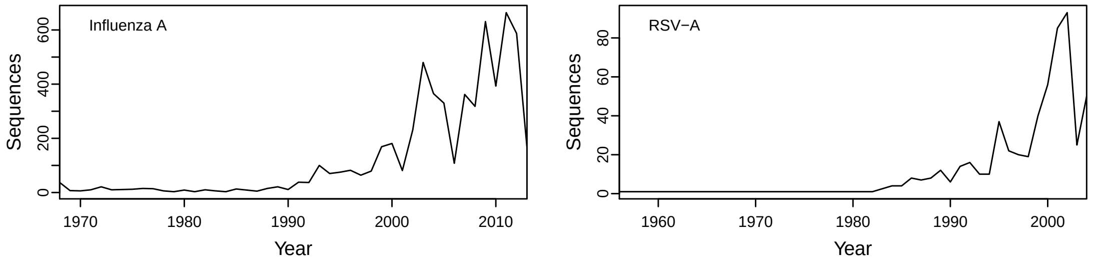

Figure B.1: The number of sequences sampled in each year for human influenza A (H3N2) and human RSV-A.

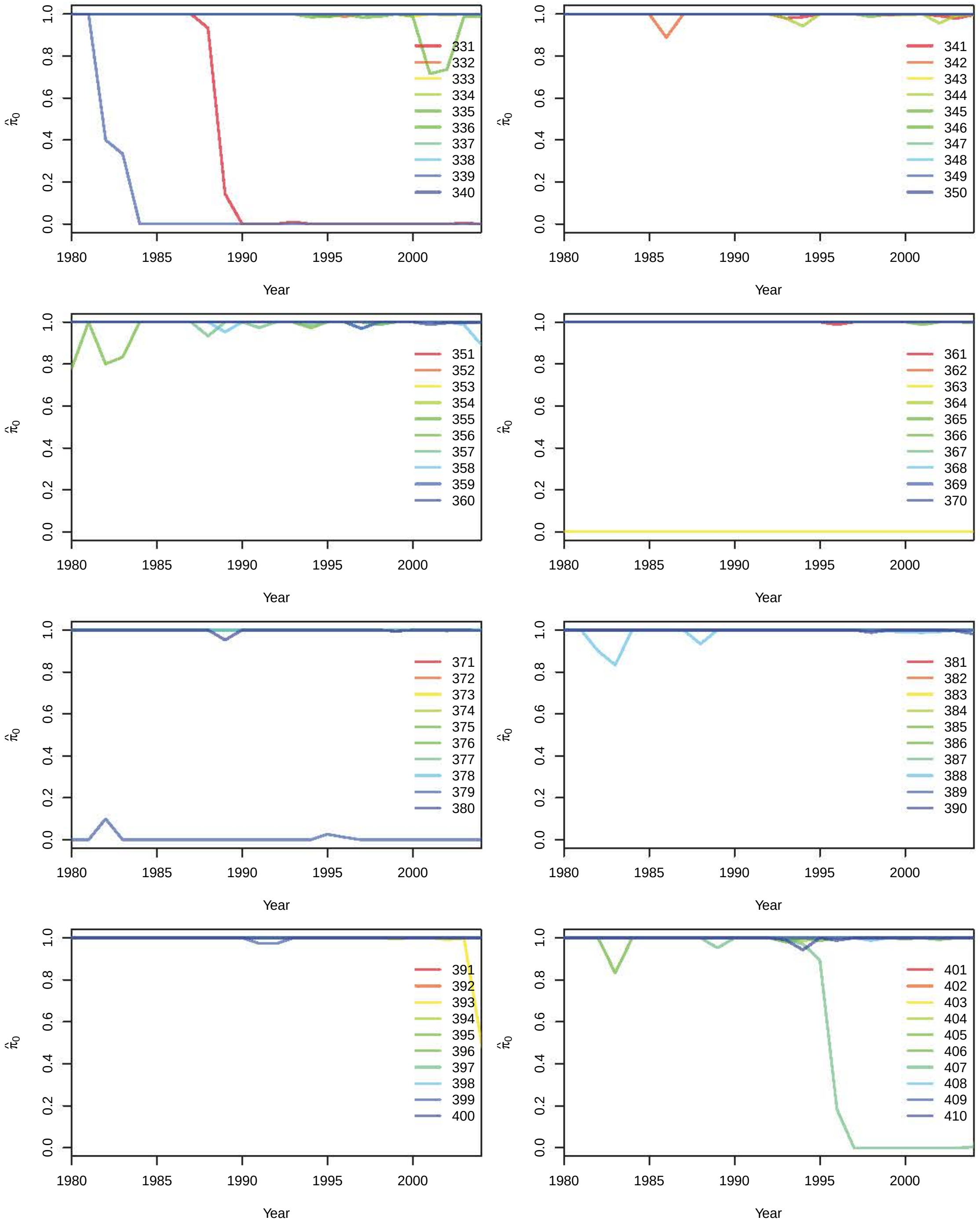

Figure B.2: Frequency trajectory of an additional 80 sites in the HA2 region of the H3N2, which does not contain antigenic regions. Each line indicates a single site, with the position given in the legend. Sites have been arbitrarily separated into different panels for visibility.

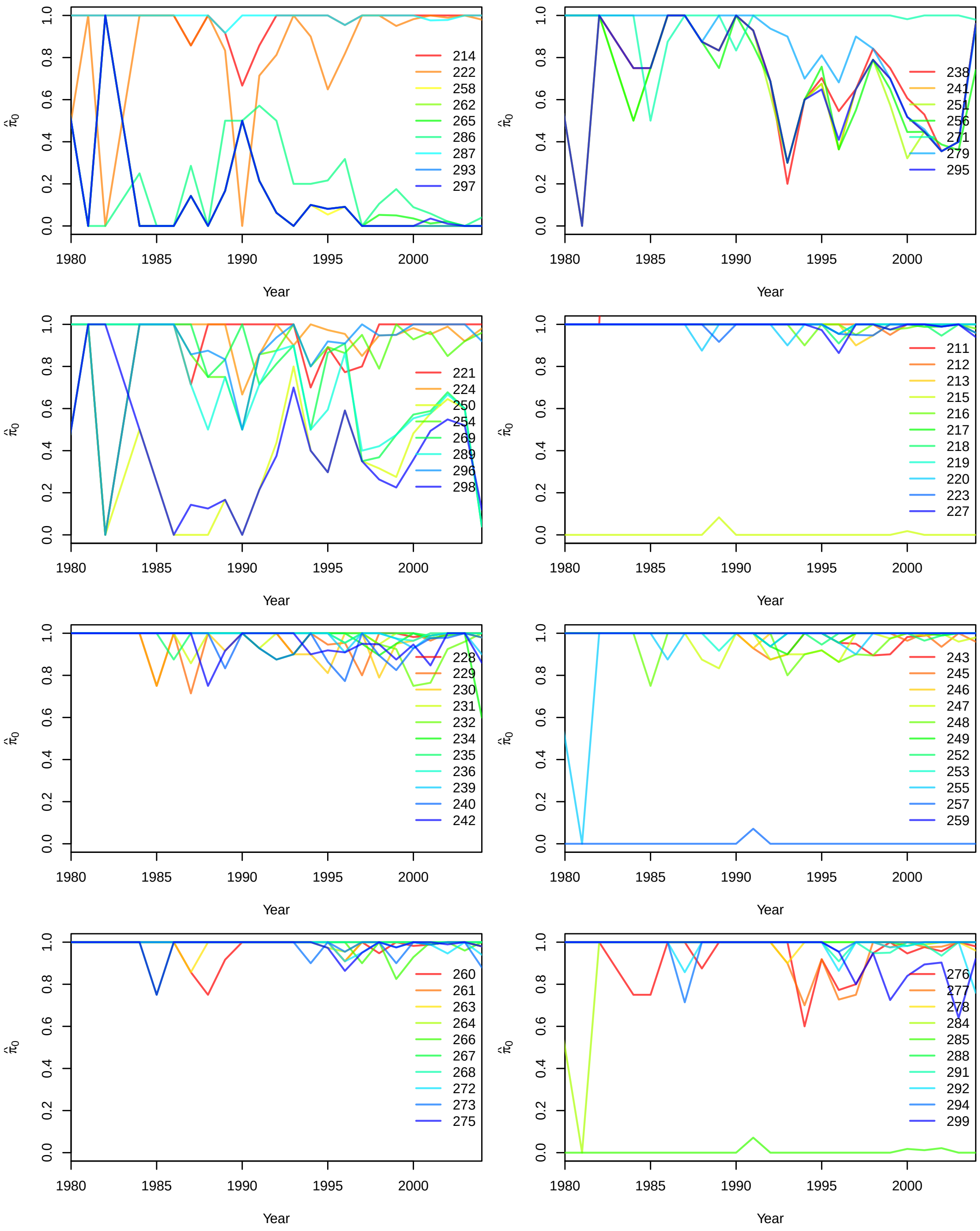

Figure B.3: Frequency trajectory of the other "non-antigenic" sites in the hypervariable region of the surface G protein in RSV-A. We note that the sequences contain only the hypervariable region (89 codon sites in total), and some of the sites which we did not include in the analysis may, in fact, be antigenic. In the first three panels, we group sites which show similar frequency patterns, possibly due to linkage; the remaining sites are ordered by position and arbitrarily separated into different panels for visibility.

Appendix C

Avian influenza sequences

In this appendix, we show the composition of the avian influenza sequences used in Chapter 4. All sequences were extracted from the Influenza Virus Resource (Bao et al., 2008). We also show changes in frequency at amino acid sites over time (Figures C.1–C.6) at internal proteins PB2, PB1, PA, NP, M1 and NS1.

Table C.1: Number of PB2, PB1 and PA sequences in each lineage by subtype.

Segment	Lineage	Region	Total	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16
PB2	1	A	3748	294	177	773	625	271	361	404	89	55	383	197	105	2	7	0	5
PB2	2	E	1449	28	94	54	40	162	436	241	8	259	21	33	12	35	1	9	16
PB2	3	E	466	14	9	32	30	72	64	41	1	180	12	10	1	0	0	0	0
PB2	4	E	1133	11	0	12	20	994	34	22	0	13	9	18	0	0	0	0	0
PB2	5	A	700	60	69	138	89	85	72	38	6	27	32	77	7	0	0	0	0
PB1	1	A	2919	235	162	491	382	247	302	316	58	73	331	191	80	30	1	6	14
PB1	2	A	1905	121	99	488	410	129	158	144	44	25	139	97	38	4	6	0	3
PB1	3	E	2063	55	97	101	82	237	531	328	12	469	45	59	11	24	3	1	8
PB1	4	E	1155	0	3	3	0	1125	12	0	0	8	1	3	0	0	0	0	0
PA	1	A	1726	117	63	432	289	126	145	207	43	22	137	90	47	8	0	0	0
PA	2	E	890	23	28	27	38	73	317	106	4	243	8	12	4	1	1	5	0
PA	3	A	744	42	43	184	129	19	74	42	6	13	141	28	16	0	5	0	2
PA	4	A	401	9	34	27	7	69	65	136	4	2	1	30	1	9	0	0	7
PA	5	A	845	94	49	151	141	82	77	32	19	30	72	85	10	1	2	0	0
PA	6	A	1068	97	40	190	210	82	86	114	36	8	112	47	44	2	0	0	0
PA	7	E	2570	45	111	95	62	1361	241	222	7	265	34	60	10	37	4	1	15

Table C.2: Number of NP, M1 and NS1 sequences in each lineage by subtype.

Segment	Lineage	Region	Total	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12	H13	H14	H15	H16
NP	1	A	4542	356	261	846	713	379	437	461	91	87	435	285	106	50	7	7	21
NP	2	E	1860	57	107	172	97	214	542	260	11	256	56	69	11	1	6	1	0
NP	3	E	1382	0	0	1	0	1074	3	61	0	242	0	0	1	0	0	0	0
M1	1	E	2120	96	133	234	174	210	535	297	13	151	63	109	20	53	4	7	21
M1	2	A	4740	371	247	952	747	387	492	494	102	105	445	263	103	15	7	0	10
M1	3	E	1118	3	3	10	9	47	117	66	1	857	5	0	0	0	0	0	0
M1	4	E	1961	10	1	20	9	1838	52	17	0	3	2	8	0	0	0	0	1
NS1	1	E	1965	70	109	180	136	173	483	265	6	262	67	91	16	67	7	8	25
NS1	2	E	1885	0	0	0	0	1878	1	0	0	6	0	0	0	0	0	0	0
NS1	3	E	913	1	0	1	0	29	160	63	0	657	0	2	0	0	0	0	0
NS1	4	A	3150	245	201	497	476	232	337	295	93	87	404	201	80	2	0	0	0
NS1	5	E	424	16	15	55	40	98	52	104	4	11	11	13	3	1	0	1	0
NS1	6	A	1664	130	61	452	270	196	162	212	10	20	42	77	27	0	5	0	0

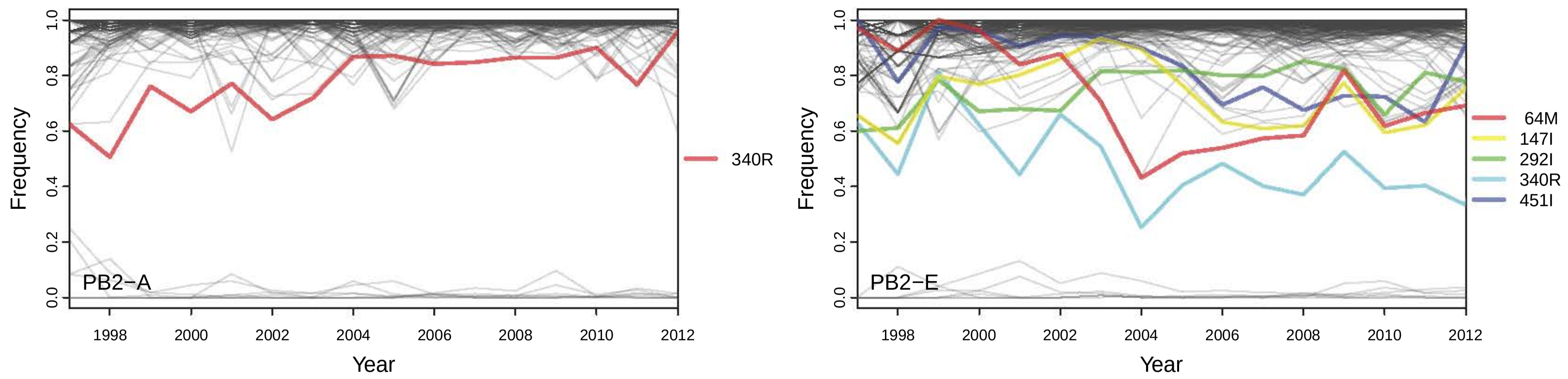

Figure C.1: Frequency trajectory of amino acid polymorphisms in PB2 in the American (A) or Eurasian (E) hemisphere. Polymorphic sites maintained for over five years at mid-frequency (0.25–0.75) are shown in coloured lines, whereas all other sites are shown as grey lines.

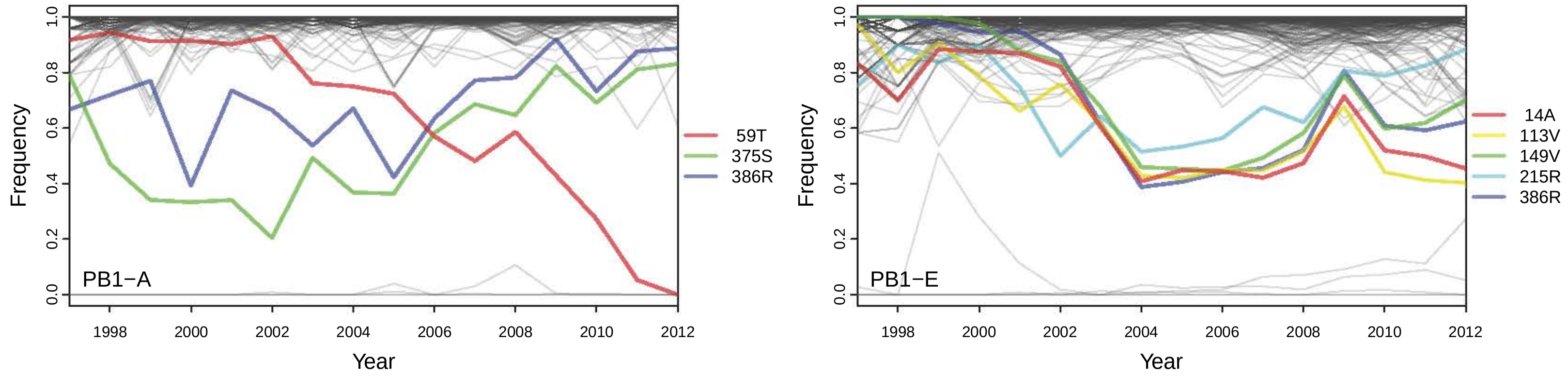

Figure C.2: Frequency trajectory of amino acid polymorphisms in PB1.

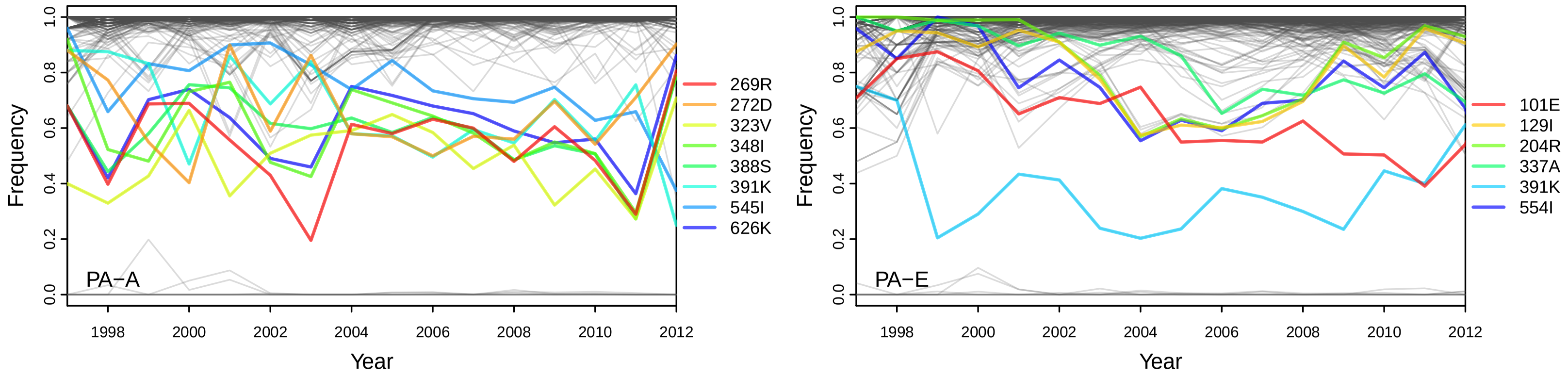

Figure C.3: Frequency trajectory of amino acid polymorphisms in PA.

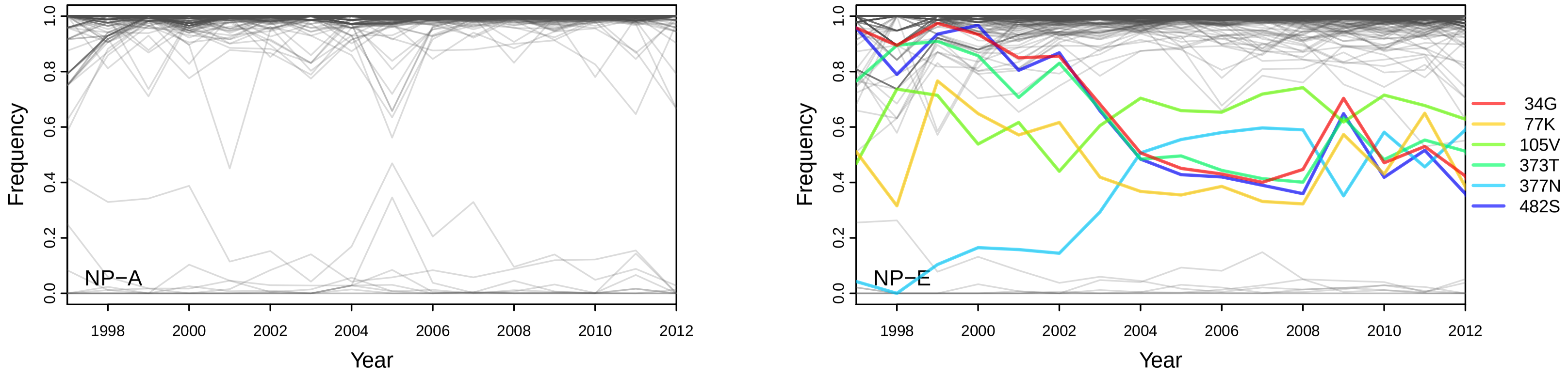

Figure C.4: Frequency trajectory of amino acid polymorphisms in NP. No polymorphic sites maintained for over five years were observed in the American hemisphere.

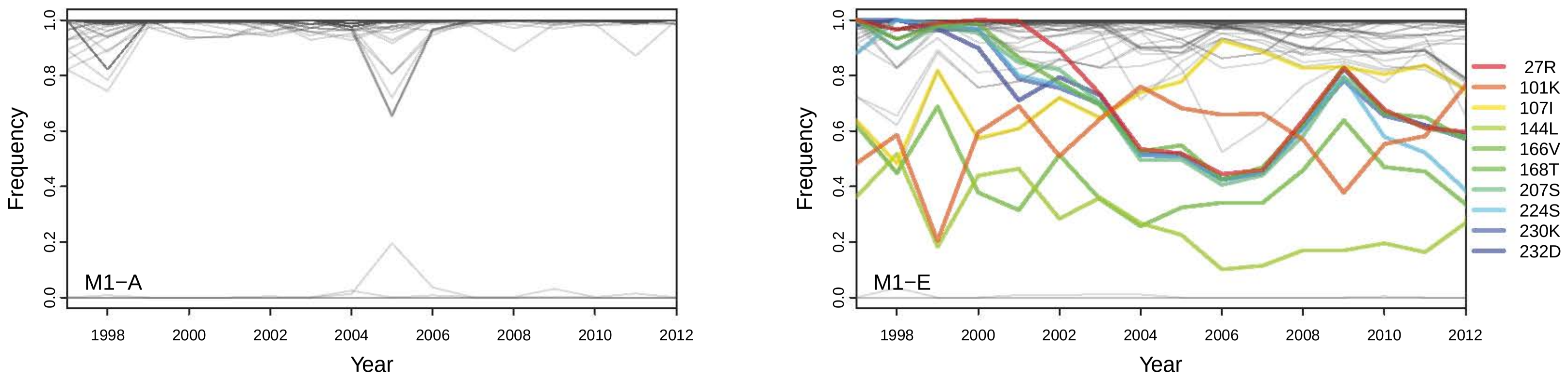

Figure C.5: Frequency trajectory of M1. No polymorphic sites maintained for over five years were observed in the American hemisphere.

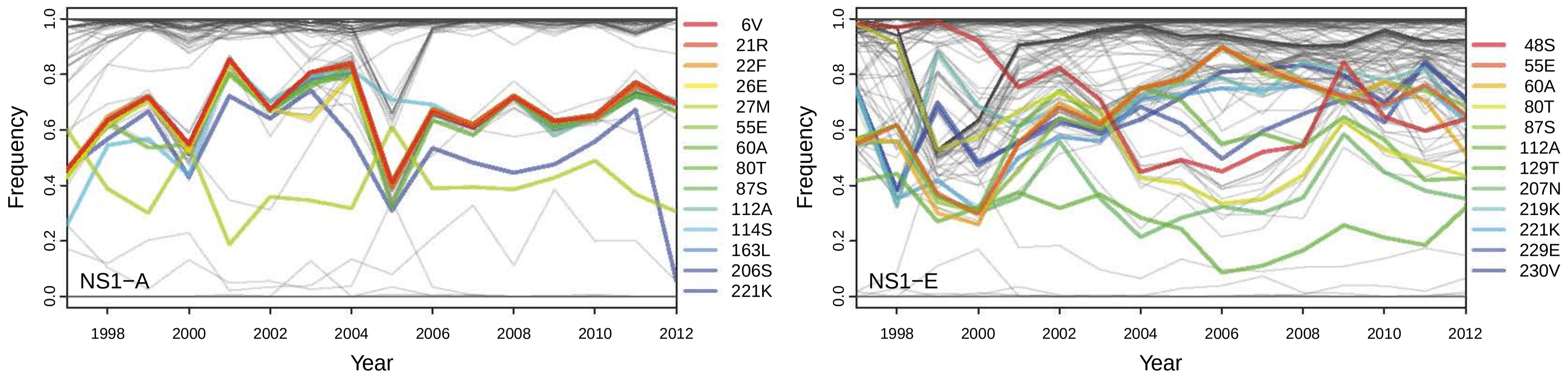

Figure C.6: Frequency trajectory of amino acid polymorphisms in NS1. For visibility, we removed a number of polymorphic sites which shared the same trajectory; the large number of polymorphic sites is due to the deep divergence between alleles A and B of NS1.