Statistical analysis of natural selection in RNA virus populations

Samir Bhatt

Brasenose College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hillary 2010

Acknowledgements

I would like to thank my funding body the National Environmental Research Council (NERC) for their financial support.

I owe my deepest gratitude to Dr Oliver Pybus, my supervisor and mentor. Without his help and support I would never have completed this thesis. His role as supervisor went way beyond just providing encouragement and direction with my DPhil, he was a source of inspiration and a true role model to a young scientist.

I have enjoyed my time completing this DPhil, but have only accomplished what I have with the help of my family and friends. I offer to them my deepest gratitude.

Abstract

Statistical analysis of natural selection in RNA virus populations

Samir Bhatt, *Brasenose College, University of Oxford*

A key goal of modern evolutionary biology is the identification of genes or genome regions that have been targeted by natural selection. Methods for detecting natural selection utilise the information sampled in contemporary gene sequences and test for deviation from the null hypothesis of neutrality. One such method is the McDonald Kreitman test (MK test), which detects the the molecular 'footprint' left by natural selection by considering the frequency of observed mutations within the sampled population. In this thesis I investigate the applicability of the MK test to viral populations and develop several new methods based on the original MK test. In chapter 2, I use a combination of simulation and methodological improvements to show that the MK test can have low error when applied to analysis of RNA virus populations. Then, in chapter 3, I develop an extension of the MK test with the purpose of estimating rates of adaptive fixation for all genes of the human influenza A virus subtypes H1N1 and H3N2. My results are consistent with previous studies on selection in influenza virus populations, and provide a new perspective on the evolutionary dynamics of human influenza virus. In chapter 4 I develop a formal statistical framework based, on the MK test, for calculating the number of non neutral sites at any frequency range in the site frequency spectrum. In this framework, I introduce a new method for reconstructing the site frequency spectrum that incorporates sampling error and allows for the inclusion of prior knowledge. Using this new framework I show that the majority of nucleotide sites in hepatitis C virus sequences sampled during chronic infection represent deleterious mutations. Finally, in chapter 5 I use the generalised framework introduced in chapter 4 to develop a statistic for evaluating the deleterious mutation load of a population. I apply this test sequences that represent 96 RNA virus genes and show that my approach has comparable power to equivalent phylogenetic

methods. In this thesis I have developed computationally efficient methods for analysis of genetic data from virus populations. It is my hope that these methods will become useful given the explosion in sequence data that has accompanied recent improvements in sequencing technology

Contents

Chapter 1

Introduction

Charles Darwin's *On the Origin of Species*, published in 1859, laid the foundations for the field of evolutionary biology. In the early part of the twentieth century, a union among the ideas set out in Darwin's seminal work, Mendel's laws of inheritance, and developments in genetic and mathematical theory led to what is termed the modern evolutionary synthesis (Huxley and Baker, 1963). One of the most important questions to emerge from the modern synthesis was how to quantify the 'molecular footprint' left behind in nucleotide and protein sequences by the action of natural selection. Since this question was first posed, a substantial body of research has focused on the application of theory within a rich mathematical framework to analyze information contained in molecular sequence data. However, many current methods for detecting and quantifying the molecular footprint of selection are computationally time consuming. This problem is being compounded by the increasing size of modern genetic data sets as a result of the explosion in sequence data that has accompanied recent improvements in sequencing technology (e.g. Bains and Smith, 1988; Ronaghi et al., 1996; Brenner et al., 2000; Margulies et al., 2005). The recent 2009 H1N1 swine influenza pandemic provides an excellent example of this phenomenon (Smith et al., 2009). Since its detection in April 2009, over 14080 gene segments of this strain have been sequenced and added into GenBank, with more samples being added on a daily basis. Indeed, this rapid growth can be observed for all types of sequence data, with the number of nucleotide sequences in GenBank increasing exponentially and doubling approximately every 18 months (GenBank release notes February 15th 2010).

The motivation behind this thesis is to investigate computationally-efficient methods for detecting selection, and to develop the application of such approaches to RNA virus data. It is hoped that this will in turn lead to an improved capacity to address significant questions relating to the molecular 'footprint' of selection in viral populations.

1.1 General Introduction

1.1.1 Sources of genetic variation and evolutionary forces

The existence of heritable variation is essential for the process of evolution, and the ultimate source of this variation in populations is mutation. It is upon mutations that genetic processes such as *natural selection*, *genetic drift*, and *recombination* act to shape the genetic structure of populations. Viruses, particularly RNA viruses, generally have very high mutation rates and short generation times, which greatly facilitates the creation of genetic diversity (Jenkins, 2002; Belshaw et al., 2008). Although mutations are the ultimate source of genetic variation, they are also the weakest force in shaping genetic variation. In the absence of an evolutionary force that changes the frequency of a mutation, a mutation will increase in a population only by independent mutations occurring repeatedly. This would be an incredibly slow process yielding little evolutionary change. By contrast, evolutionary forces such as *natural selection* or *random genetic drift*, are capable of rapidly changing mutant frequency in a population (e.g. Pybus and Shapiro, 2008).

When a mutation occurs, there are two possible final states: *fixation* (where the frequency of a mutation is 100% in a specific finite population) or *elimination* (when a mutation is removed from the population). The probability of each outcome is dependant on the degree to which the mutation affects the fitness of the organism in the current environment, and the size of the population. Beneficial mutations cause an increase in fitness that can eventually lead to fixation of the mutation (called *positive selection*). Deleterious mutations have the opposite effect, causing a decrease in fitness that can eventually lead to removal from the population (called *negative selection*). Neutral mutations do not affect fitness. In addition to positive and negative selection, populations may display *frequency dependent selection*, which occurs when a rare variant exhibits greater fitness than a common one.

Random genetic drift acts on all mutations. Drift occurs because, in the absence of selection,

each member of a generation is equally likely to be the parent of any member of the subsequent generation. Drift is a random sampling process, which is exacerbated by small population sizes. While drift cannot cause adaption in populations, it can change allele frequency; it has a critical impact on new mutations which begin at low frequencies and so are more susceptible to random effects. The probability at any given time that a neutral mutant will ultimately, by drift, become fixed at its locus is simply its frequency in the population at that time.

1.1.2 Models of selection and drift

Real populations are susceptible to both drift and selection. This combined action can be represented by the Wright -Fisher model, which is a population genetic model that describes the change in frequency of a single mutation in a population over time. The simplest version of the model makes the following assumptions: (1) nonoverlapping generations, (2) constant population size in each generation and (3) random mating. In order to make this model more applicable to RNA viruses with high mutation rates, I will consider here a slightly more complex model that also includes recurrent mutation

Consider a population of N haploid individuals that has a single polymorphic site with two mutations, one ancestral (fitness = 1) and one derived (fitness = $1 + s$), where s, the selection coefficient, is a relative value that compares the fitness of one mutation to another. In this model I define u as the recurrent mutation rate.

Under this model, the frequency of the derived mutation in the current generation is a function of the selection pressure on the mutation and the binomial sampling arising from genetic drift. The probability p_{ij} that there are j derived individuals present at generation $G + 1$ given i derived individuals present at generation G, is given by:

$$p_{ij} = \binom{N}{j} \pi_i^j (1 - \pi_i)^{N-j} \tag{1.1}$$

where π depends on a number of population genetic parameters. In simplest case of no selection or recurrent mutation $\pi_i = \frac{i}{N}$. Under the more general case of drift, selection and mutation this model defines a complex stochastic process, the derivation of which is lengthy and is given in full in section 3.3.2.1.

The behaviour of this stochastic process can be most easily summarised as the stationary distribution of mutation frequencies under selection, mutation and drift.

$$\pi(x|u, N_e, s) = \frac{(1-x)^{2Nu-1}(x)^{2Nu-1}e^{2Nsx}}{\int_0^1 (1-x)^{2Nu-1}(x)^{2Nu-1}e^{2Nsx}dx} \tag{1.2}$$

Where N represents the population size. Equation 1.2 gives the distribution of mutations with population frequencies in the range $(x, x+\delta x)$, which can also be thought of as the amount of time a site spends in the interval between x and $x+\delta x$ when that site is affected by selection, drift and recurrent mutation.

In addition another standard result from population genetic theory is the probability of fixation of a mutation (Kimura, 1983), which is approximated by:

$$\mathbb{P}(x) = \frac{1-e^{-2Ns\,x}}{1-e^{-2Ns}} \tag{1.3}$$

From equations 1.2 and 1.3 it is clear that:

- The evolutionary behaviour of each mutation under selection is dependent on the product, $2Ns$
- Favourable mutations won't necessarily go to fixation when drift is present
- Most mutations, whether beneficial, deleterious or neutral are lost from the population in which they occurred.
- If selection against deleterious mutations is weak or N is small, $2Ns \approx 1$, a deleterious mutation is likely to behave as a neutral mutation

- If N is large, deleterious mutations are much less likely to be fixed than neutral mutations. And the larger N is, the more likely it is that a favourable mutation will be fixed.

While based on simplistic assumptions, these equations provide a framework for understanding the complex dynamics of mutant frequencies, and introduce the usefulness of the site frequency spectrum, which contains the information for the statistical tests used in this thesis.

1.1.3 The site frequency spectrum

The site frequency spectrum is a summary of the frequency of mutations observed in a set of sampled sequences. To begin it is necessary to define the relative terms *derived* and *ancestral*. A derived nucleotide is a nucleotide which is not present in a common ancestor of the population, whereas an ancestral nucleotide was present in that common ancestor.

The proportion of the population in which a mutation is found is its frequency, or *site-frequency*, which changes through time and depends on the action of selection and drift at that site (see equation 1.2 for a model of this process). The *unfolded* site frequency spectrum counts the number of derived polymorphisms in a sequence alignment. Mathematically I define the spectrum as a random vector $X = \{X_1, X_2, ..., X_{m-1}\}$ of nucleotide configurations, where X represents the numbers of singletons, doubletons, tripletons and so on in a sequence alignment and where X_i represents the number of sites that have $n - i$ ancestral and i derived nucleotides among n aligned sequences. To further understand this consider the following example sequence alignment

Sequence 1	A	A	T	A	G	C
Sequence 2	A	A	A	C	A	C
Sequence 3	A	T	A	C	T	C
Sequence 4	A	A	A	C	T	C
Ancestral Sequence	A	A	A	A	T	C

The unfolded site frequency spectrum would be 2 singletons, 1 doubleton and 1 tripleton, hence $X = \{2, 1, 1\}$.

The unfolded spectrum assumes that it is known which base is derived and which is ancestral (Bustamante et al., 2001). In the case where this information is unknown, a *folded* spectrum would be used. The folded site frequency spectrum conflates i with $(n - i)$ thereby effectively counting the frequency of the rarest nucleotide at each site in a sequence alignment. The spectrum can be defined in a similar manner to that above: $X^* = \{X_1^*, X_2^*, .., X_{m/2}^*\}$, where $X^*$is defined as the number of sites where 1 or $(n - 1)$ individuals carry a mutation, 2 or $(n - 2)$ individuals carry a mutation etc. So for the above example site 4 would be reclassified as a singleton not a tripleton, because in the folded spectrum the rarest nucleotide is counted. Thus the folded site frequency would be $X^* = \{3, 1\}$.

Both the folded and unfolded site frequency spectrum only consider polymorphic sites. In chapter 4 I create a different type of site frequency spectrum, which not only includes sites that are polymorphic but also those which are fixed or invariant.

1.1.4 The coalescent

When trying to determine whether sequences from a specific population have undergone selection a null model needs to be specified. The appropriate null model is genetic drift, and the relationship between drift and shared ancestry is provided by the rich statistical framework of the coalescent model (Kingman, 1982b). The coalescent is a retrospective model that describes the characteristics of sampled lineages back in time to a common ancestor. This joining of lineages is referred to as coalescence (Kingman, 1982b). The coalescent has a number of attractive properties; (i) it is mathematically tractable and straightforward to simulate, (ii) it can be used to make inferences under small sample sizes and (iii) its behaviour is largely invariant to the way in which the population reproduces or undergoes mutation.

The simple, well defined framework of the coalescent has made it possible to incorporate more

complex population genetic processes, resulting in models incorporating recombination (Hudson, 1990; Griffiths and Marjoram, 1996, 1997), population subdivision (Nath and Griffiths, 1996) and variable population size (Slatkin and Hudson, 1991; Pybus et al., 2000). One notable exception to the above list are models for natural selection. Coalescent models of selection are very difficult to develop as they violate the key assumptions of the model, namely neutrality. Efforts have been made to introduce selection into coalescent theory (Kaplan et al., 1988; Neuhauser and Krone, 1997) but none are in widespread use.

1.1.4.1 The standard coalescent

Consider a haploid population of N individuals evolving according to the Wright-Fisher model. As time moves to the next generation, (i) the population stays constant in size (ii) each offspring picks a parent at random from the individuals in the previous generation(density is multinomially distributed), and parent and child are linked by a line and (iii) each offspring takes the genetic information of the parent. This is summarised in figure 1.1 on the following page. The model assumes: non overlapping generations, random mating and no selection. From this process it is clear that the probability that a gene comes from a specific parent is $1/N$. In this framework the coalescent process can be derived.

Under the above assumptions, the coalescent process can be derived as follows. Assume that the population varies deterministically through time; we define the number of individuals in generation r as $N(r)$. The size of the population at $r = 0$ generations in the past (the initial size at the present) is $N(0) = N_0$. For a population with constant size, N_0, in a particular generation, label the individuals $\{1, 2, ..., N_0\}$ and let $\{v_1, v_2, ..., v_{N0}\}$ be the respective numbers of offspring that each individual as. It is assumed that $\{v_1, v_2, ..., v_{N0}\}$ are exchangeable random variables with sum $N(r)$ and variance $\sigma^2(r-1)$. In biological terms this interchangeability embodies the properties of a non-structured neutral population - although individuals may differ in their reproductive success, these differences are random and cannot be passed onto

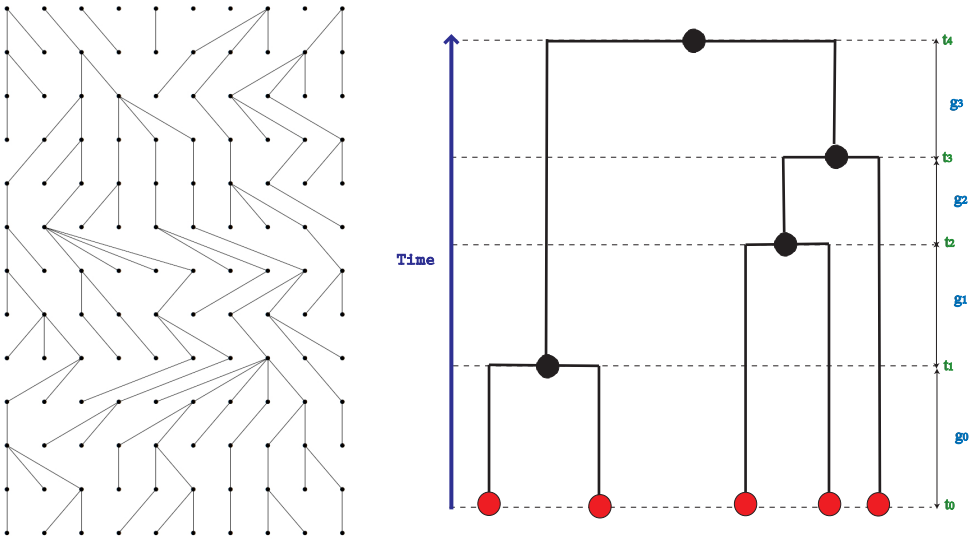


Figure 1.1: *(Left)Example of the wright fisher process for a haploid population. (Right) Realisation of the Coalescent Process: Red dots are Samples from the population. Black dots are coalescence events. t are the times between coalescence event*s

the next generation.

It is clear from this framework that there are two specific functions: (i) *The demographic model* $N(\cdot)$, which describes the change in population size through time and (ii) *The reproductive model* $\sigma^2(\cdot)$, which describes the variance in reproductive success through time. These two processes derive a non-homogeneous continuous time Markov death process (Griffiths and Tavare, 1994b). If a random sample of n individuals is taken at time 0 then the coalescent is a Markov process $\{A_n(t),\, t \geqslant 0\}$ that describes the ancestral relationships among the n individuals. The Markov death process starts from $A_n(0) = n$ and moves down in steps of 1 until the common ancestor is reached. The transition probabilities defined by Kingman (1982a) are

$$\mathbb{P}\left[A_n(r+h) = j \mid A_n(r) = i\right] = \begin{cases} \binom{i}{2}\frac{\sigma^2(r)}{N(r)}h + o(h), & j = i-1 \\ 1 - \binom{i}{2}\frac{\sigma^2(r)}{N(r)}h + o(h), & j = i \\ 0, & \textit{Otherwise} \end{cases} \tag{1.4}$$

If it is assumed that $N \ggg n$ then events of order $o(h)$ can be ignored. When measured in units of generations, the amount of time during which there are j lineages, t_j, is approximately exponentially distributed with mean $\binom{i}{2}$. The waiting times between coalescence events can be derived from the above process. If g_i is the size of the waiting time which contains i lineages (see figure on the previous page) then the distribution of g_i is:

$$\mathbb{P}(g_i|t_i) = \left(\binom{i}{2} \frac{\sigma^2 (g_i + t_i)}{N (g_i + t_i)} \right) \cdot \exp \left[- \int_{x=t_i}^{g_i+t_i} \binom{i}{2} \frac{\sigma^2 (x)}{N (x)} dx \right] \tag{1.5}$$

Equation 1.5 can be thought of as the product of two probabilities. The left hand term of the product is the probability that a coalescence even will occur at time $g_i + t_i$ and the right hand term is the probability no event occurs between time t_i and $g_i + t_i$. Therefore the product is the waiting time until the next coalescence event.

Equation 1.5 can further be simplified. σ^2 is a nuisance parameter that can be removed under the assumption that the reproductive success is constant through time. We can now define a new term, $N_e(x) = {}^{N(x)}/_{\sigma^2}$, which is the effective population size. Variance in reproductive success implies that some individuals get more offspring and some get less, and therefore if σ^2 increases then there is a higher probability that offspring have the same parents. Under a neutral Wright-Fisher model $\sigma^2 = 1$ (due to a symmetric multinomial distribution of offspring numbers) (Donnelly and Tavare, 1995). Therefore Equation 1.5 becomes

$$\mathbb{P}(g_i|t_i) = \left(\binom{i}{2} \frac{1}{N_e (g_i + t_i)} \right) \cdot \exp \left[- \binom{i}{2} \int_{x=t_i}^{g_i+t_i} \frac{1}{N_e (x)} dx \right] \tag{1.6}$$

The demographic model can incorporate changes in population size. Basic demographic models such as constant population size or exponential growth ($N_e(x) = N_e(0)\, e^{-Rx}$) can easily be evaluated in the integral. More complex models of demographic history have also been investigated (e.g. Hudson, 1990; Griffiths and Tavare, 1994a; Pybus et al., 2000).

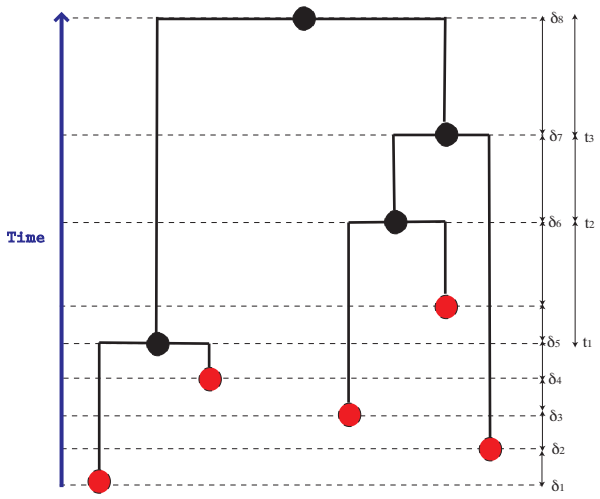


Figure 1.2: Realisation of the s-coalescent process. Red dots are new entry sequences. Black dots are coalescence events. δ are the times of new entry sequences. t are the times of coalescence events

1.1.4.2 The serial coalescent

The tests in this thesis detect signatures of selection by comparing patterns of polymorphism and divergence. In order to simulate neutral sequences for such tests, there is a need to simulate a reference ancestral sequence (see figure 1.2) to determine if a site is derived or ancestral.

The serial coalescent (s-coalescent) is an extension of the n-coalescent model introduced in the previous section that allows incorporation of sequences sampled at different points in time (Rodrigo et al., 1999).

Consider a tree (Figure 1.2), where internal nodes represent coalescent events and external nodes (leaves) represent points in time when which new samples join the genealogy. Define δ_x as the amount of time between nodes x and $x+1$ (see figure 1.2). The probability density of the waiting times between coalescent events is now simply an extension of equation (1.6)

$$\mathbb{P}(g_i|t_i) = \left(\binom{i}{2} \frac{\sigma^2 (g_i + t_i)}{N (g_i + t_i)} \right) \cdot \exp \left\{ - \sum_k \left[- \int_{x=t_i}^{g_i+t_i} \binom{i}{2} \frac{\sigma^2 (x)}{N (x)} dx \right] \delta_k \right\} \tag{1.7}$$

The sum in equation 1.7 is calculated across all δ_k intervals between coalescent events. The serial coalescent does not allow sequences to join the genealogy at exactly the same moment of time that a coalescent event occurs. However, it is possible for several sequences to enter the genealogy simultaneously. It is clear from examining equations 1.5 and 1.7 that the difference between the serial coalescent and the standard coalescent is that with the standard coalescent the number of lineages monotonically decreases as time advances into the past, but with the serial coalescent the number of lineages can increase as new sequences join the genealogy.

1.1.4.3 Simulating the coalescent

A great benefit of the coalescent framework is its ease in computational implementation. Because neutral mutations have no effect on the genealogy of randomly sampled sequences it is possible to separate the genealogical process from the neutral mutation process. When simulating neutral evolution using the coalescent, algorithms for both processes are readily available. For example, Hudson (1990) and Rodrigo and Felsenstein (1999) explain how coalescent trees can be simulated and Yang (1997) describes how to simulate sequences under various codon substitution models.

For the purpose of analysis in this thesis the class *TreeBuild*, available from the *Java Evolutionary Biology Library (JEBL)* was used to simulate serial and standard coalescent trees and the program *Evolver*, a package in *PAML(Yang, 1997)* was used to simulate sequence down phylogenies under various codon substitution models (Yang, 2000).

1.2 Detecting Signatures of Selection

The question of how to quantify the relative contributions of drift and selection to the genetic diversity of a population is of central importance to this thesis. Sequences sampled from populations contain two sources of information about the action of natural selection on

mutations. The first is the ratio of silent to replacement variation per site, ω (e.g. (Nei and Gojobori, 1986)). Here the relative *rates* of silent and replacement fixations are compared. Silent mutations are always considered neutral and therefore a greater rate of fixation for replacement changes (D_n) relative to the rate of silent change (D_s) can only be explained by the action of positive selection ($\omega > 1$). Methods to estimation ω were first introduced in the in the early 1980s and were essentially counting algorithms which counted the number of silent and replacement sites in and between two sequences and then correcting for multiple substitutions at the same site (e.g. Li et al., 1985; Nei and Gojobori, 1986; Ina, 1995; Comeron, 1995). These methods however, were not grounded in a rigorous statistical framework and made simplistic assumptions coupled with *ad hoc* treatment of the data. More recently a rigorous set of methods set in the maximum likelihood framework is used to estimate ω (e.g. Goldman and Yang 1994; Muse 1996; Yang et al. 2000 . This set of methods uses explicit models of codon substitution with varying degrees of parametrisation. The benefit of this approach is that all the model parameters can be estimated in one step within a rigorous statistical framework. However, these methods have been shown to be unreliable for within population serially sampled data (Kryazhimskiy and Plotkin, 2008) and are computationally unfeasible for large data sets (Pond and Frost, 2005).

The second set of methods use statistics to measure the frequency of ancestral and derived nucleotides at variable sites. In essence, these statistics are summaries of the *site frequency spectrum* (e.g. Tajima, 1989 or Fay and Wu, 2000). The methods outlined in this section belong primarily to this second category of methods. Site frequency methods can generally be divided into three main groups (1) polymorphism tests, (2) joint polymorphism and divergence tests and (3) polymorphism and divergence tests that determine numbers of adaptive fixations. This thesis focuses on these methods as a means to quantify the action of natural selection.

Before discussing these tests it is first necessary to introduce θ which is a common term that represents the amount of genetic diversity within a set of sequences (nucleotide diversity). In

neutral Wright-Fisher haploid populations, population genetics theory predicts that $\theta = 2Nu$. The units of θ are substitutions per nucleotide site.

1.2.1 Estimators of θ

Various methods exist for estimating the value of θ from sequence data, the simplest assume an infinite sites model (Kimura, 1969). The infinite sites model assumes that mutation only occurs once at a given site, hence the site frequency observed is the result of a single mutation. This assumption is obviously accurate if the sequence length is long and the mutation rate (u) is low.

1.2.1.1 The Watterson estimator

Under infinite sites, the number of segregating sites observed in a sample is equal to the total number of mutations since the most recent common ancestor, S. If the total length of the branches in a genealogy is, L, then the number of segregating sites has a Poisson distribution with mean $\theta L/2$ (Watterson, 1975). Therefore:

$$\begin{aligned}\mathbb{E}(S) =& \mathbb{E}\left(\mathbb{E}\left(\frac{S}{L}\right)\right)\\ =& \mathbb{E}\left(\frac{\theta L}{2}\right)\\ =& \frac{\theta}{2}\sum_{j=2}^{n} j\,\frac{2}{j(j-1)}\\ =& \theta\sum_{j=1}^{n-1} 1/j\end{aligned}$$

Hence for S segregating sites the estimate of θ is

$$\theta_s = \frac{S}{\sum_{j=1}^{n-1} 1/j} \tag{1.8}$$

This estimator is the Watterson estimate, θ_s.

1.2.1.2 The pairwise differences estimator

θ can also be estimated from the average number of pairwise differences. This is known as Tajima's estimator (Tajima, 1983), or the pairwise differences estimator (θ_k). For an alignment of n sequences the average number of pairwise differences is given by the general formula

$$\theta_k = \frac{1}{n(n-1)} \sum_{i \neq j} \Pi(i,j) \tag{1.9}$$

where $\Pi(i,j)$ is the number of sites that sequences i and j differ, and $\frac{1}{n(n-1)}$ is the total number of pairwise comparisons.

1.2.2 Polymorphism tests

1.2.2.1 Tajima's D (Tajima, 1989)

The Tajima's D test is based on the two different estimates of θ introduced above, (i) the mean number of pairwise differences (θ_k) and (ii) the scaled number of segregating sites (θ_s) (Watterson (1975)). The premise of Tajima's D test is that under neutral evolution these two measures should be equal, hence the difference between them should be zero. As mentioned earlier for a neutrally-evolving haploid population, θ is expected to equal $2N_e\mu$, where N_e is effective population size and μ is the rate of nucleotide substitution. Tajima's D statistic is defined as:

$$D = \frac{\theta_k - \theta_s}{\sqrt{\alpha S + \beta S^2}} \tag{1.10}$$

where $\theta = s\gamma$, S is the number of segregating sites and α, β and γ are constants that depend on the number and length of the sequences. The denominator is a normalizing term equal to the standard error of the numerator. Under neutrality and constant population size, the mean

and variance of the D statistic should be approximately zero and one, respectively. Tajima's D critically depends on the shape of the genealogy that relates the sampled sequences. For a star-like tree (long terminal branches and short internal branches), $\theta_k < \theta_s$, hence D is negative. This may occur during population growth or as a result of a selective sweep, which both generate more low-frequency polymorphisms than expected under neutrality (Simonsen et al., 1995). If the tree has long internal and short terminal branches (which may occur if, for example, there is strong population subdivision) then $\theta_k > \theta_s$ and D is positive, signifying an excess of mid-frequency polymorphisms. Tajima's D does not require an outgroup sequence, that is, the ancestral or derived state of each polymorphism is not relevant.

A third estimate of θ, based on the number of singleton mutations, has been proposed by Fu and Li (1993). This estimate has also been used in conjunction with θ_s and θ_k to create two more statistics, $D^*$and F^* as discussed in Fu and Li (1993).

1.2.3 Joint polymorphism and divergence tests

Under neutral evolution genetic polymorphism in a constant size haploid population is expected to be $\theta = 2N_e\mu$, while the divergence between populations or through time is equal to μt, where t is the divergences time between samples. The methods in this section jointly use information on within species polymorphism and between species divergence and make use of these two measures to test whether empirical data match with neutral expectations.

1.2.3.1 McDonald Kreitman test (McDonald and Kreitman, 1991a)

The McDonald Kreitman (MK) test compares the pattern of polymorphism within a group (population or species) to that between two closely related groups. Under neutrality, the ratio of the number of replacement polymorphisms (r_p) to silent polymorphisms (s_p) within a group should equal the ratio of the number of replacement differences (r_d) to silent differences

(s_d) between groups, such that

$$\frac{r_p}{s_p} = \frac{r_d}{s_d} \tag{1.11}$$

Silent sites are assumed to always be neutral, and therefore if an excess of replacement differences between groups is observed then adaptive fixation and positive selection is inferred (McDonald and Kreitman, 1991a). The MK test is expected to be unaffected by the shape of the underlying genealogy and should therefore be more robust to changes in demography (Nielsen, 2001). To prove this consider a single site and contrast the expected polymorphism and divergence at silent and replacement sites. The ratio of expected divergence and polymorphism between silent (s) and replacement (r) sites under the standard neutral model is:

$$\begin{aligned} \frac{s_d}{r_d} &= \frac{t\mu_s}{t\mu_r} &= \frac{\mu_s}{\mu_r} \\ \frac{s_p}{r_p} &= \frac{2N_e\mu_s}{2N_e\mu_r} &= \frac{\mu_s}{\mu_r} \end{aligned}$$

Hence, the two ratios are expected to be equal. More generally, the term $2N_e$ can be replaced by total length of the within-population coalescent branches, which still cancels out in the ratio.

The MK test requires that sites in a sequence alignment are assigned to one of the four categories defined above. Therefore an additional 'outgroup' sequence (or sequences) is needed to determine which sites are fixed differences and which are invariant (Figure 1.3 on the following page). Typically, this outgroup represents a closely related population or sister species (McDonald and Kreitman, 1991a; (Figure 1.3a). However, for rapidly-evolving viruses sampled at different times, the outgroup can represent the same population sampled at an earlier time point (Williamson, 2003; (Figure 1.3b). The four totals (r_p, s_p, r_d and s_d) are summarized in a contingency table and a non-parametric test of independence, such as the χ^2 test or G-test, can used to test for a statistically significant deviation from neutrality.

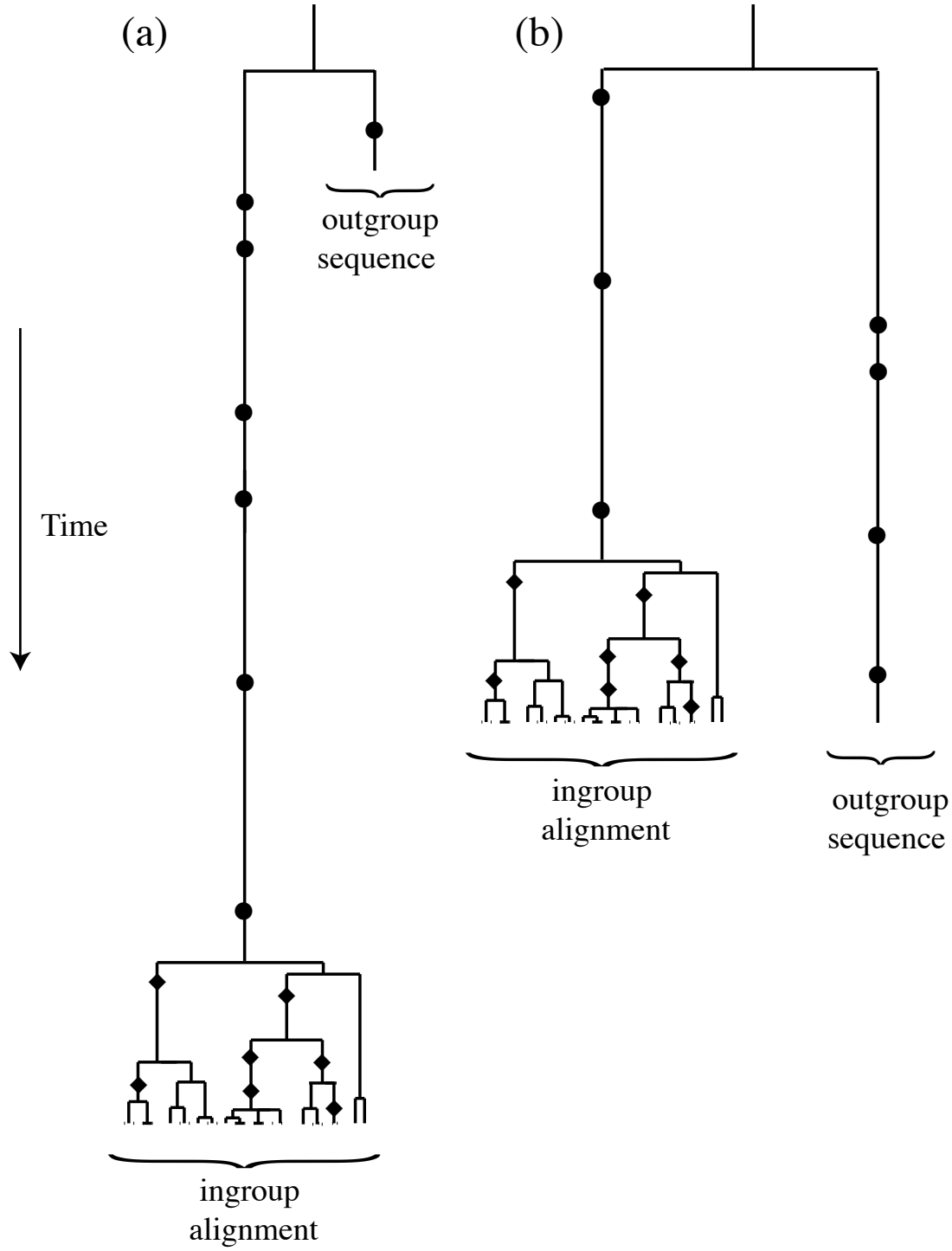


Figure 1.3: *An illustration of the rationale of the McDonald–Kreitman test. Sequences are sampled from the study population (ingroup alignment). In order to identify the direction of evolutionary change, and outgroup sequence is also obtained. (a) Outgroup is sampled from the study population at an earlier time point, sensu Williamson (2003). (b) Outgroup is obtained from a contemporaneous sister population or sister species, sensu McDonald and Kreitman (1991a). The circles represent fixed differences between the ingroup and outgroup. The diamonds represent ingroup polymorphisms.*

It should be noted that the MK test or any test that compares silent changes to replacement changes assumes that all silent changes are neutral with no selective constraints. This assumption is likely to cause bias given that while silent changes have no effect to the primary structure of a protein (due to the redundancy of the genetic code), they have been shown to affect protein secondary structure leading to adaptive changes (Cuevas et al., 2002; Novella et al., 2004) .

1.2.3.2 Poisson random field model (PRF)

Sawyer and Hartl (1992) expanded equation 1.2 for multiple sites. Their model makes the following assumptions: (i) mutations arise through time according to a Poisson distribution (ii) each mutation occurs at a new site and (iii) each mutant follows an independent Wright Fisher process with no linkage (section 1.1.2). Using the model described by Sawyer and Hartl (1992), Bustamante et al. (2001) developed a maximum likelihood method for estimating population genetic parameters N, s and u from observed sequence data.

The basic model of the PRF assumes that all loci are independent and consequently the effect of selection on linked neutral sites is not incorporated into the models. This means that the PRF provides good approximations for analysis of data from multiple locations but less so in analysis of sequence data from few loci. The assumption of independent sites is likely to cause bias, but usefulness of the PRF method comes from its ability to examine the effect of selection on different classes of mutations(Desai and Plotkin, 2008; Nielsen, 2005).

1.2.4 Polymorphism and divergence tests that determine proportions of adaptive sites

1.2.4.1 Smith and Eyre Walker's (2002) method

Smith and Eyre-Walker (2002) developed a simple statistic to estimate the number of adaptive substitutions between two sister species. Their approach was based on earlier work by McDonald and Kreitman (1991). It is based on the following assumptions: (i) all silent mutations are neutral, (ii) all replacement mutations are strongly deleterious, neutral, or strongly advantageous, and (iii) all polymorphism observed in the main alignment is neutral. The rationale for the latter assumption is that effective population sizes and selection coefficients are sufficiently large that advantageous mutations are fixed very rapidly and therefore contribute little to observable polymorphism (Smith and Eyre-Walker, 2002).

Suppose the number of silent and replacement polymorphisms observed in the main alignment are $s_{<1}$ and $r_{<1}$ respectively, and let the number of silent and replacement fixations be s_1 and r_1 respectively. The subscripts refer to the frequency of the derived nucleotide in the main alignment. Under the above assumptions, if a_1 is the number of replacement fixations that are positively selected, then $\frac{(r_1-\alpha)}{r_{<1}} = \frac{s_1}{s_{<1}}$ on average. Thus the number of adaptive fixations is estimated as

$$\alpha = r_1 - s_1 \left(\frac{r_{<1}}{s_{<1}} \right) \tag{1.12}$$

1.2.4.2 Williamson's (2003) method

The assumption that effective population sizes and selection coefficients are large seems reasonable for most chronic viral infections, since the number of individual virions is enormous and selection arising from host immune responses generates significant fitness differences among viral variants. However, Williamson (2003) noted two reasons why polymorphisms may not all be neutral: (i) the co-evolution of the immune system and the virus popula-

tion results in time or frequency-dependent selection and fluctuating selection coefficients, potentially preventing mutations from becoming fixed, and (ii) latently-infected cells in HIV infection release 'ancestral' viral variants that contribute to sampled genetic diversity. Latently-infected cells are a feature particular to retroviruses.

To avoid bias arising from these problems, Williamson (2003) assumed that 'rare' polymorphisms (observed frequency$<$ 50%) were neutral, whereas 'common' polymorphisms (observed frequency$>$ 50%) were treated in the same manner as observed fixations and could therefore be strongly deleterious, neutral or strongly advantageous. Thus two types of nucleotide sites are estimated using this approach: adaptive fixations and adaptive common polymorphisms.

Let $r_{<0.5}$ and $s_{<0.5}$ denote the number of replacement and silent rare polymorphisms in the main alignment, and let $r_{>0.5} > 0.5$ and $s > 0.5$ denote the number of replacement and silent common polymorphisms. Williamson's (2003) estimators for the number of adaptive fixations (α_1) and the number of adaptive common polymorphisms ($\alpha_{<0.5}$) are

$$\alpha_1 = r_1 \left[1 - \left(\frac{s_1}{r_1} \right) \left(\frac{r_{<0.5}}{s_{<0.5}} \right) \right] \tag{1.13}$$

and the rate of polymorphic adaptations is

$$\alpha_{>0.5} = r_{>0.5} \left[1 - \left(\frac{s_{>0,5}}{r_{>0.5}} \right) \left(\frac{r_{<0.5}}{s_{<0.5}} \right) \right] \tag{1.14}$$

These equations differ slightly from those in Williamson (2003), as the original equation explored error bias using a delta method technique, and added an extra error correction term to both above equations. For comparability to Smith and Eyre-Walker (2002) this term has been dropped.

In chapter 3 I apply an extended version of Williamson's (2003) method to human influenza A data. In chapter 4 I develop a generalised version of both Smith and Eyre-Walker's (2002)

and Williamson's (2003) in a rigorous Bayesian framework.

1.3 Thesis Outline

Chapter 2

In chapter 2 I investigate the type I error of the Tajima's D test and the McDonald Kreitman test, and use simulations to determine their applicability to highly diverse viral data. First, I perform extensive simulations to measure type I error under a range of virus like mutational and demographic scenarios. Then, using a data set of 96 RNA virus alignments that represent a diverse range of natural RNA virus populations, I evaluate the applicability of the MK test to real virus data. I also develop a new proportional counting method for the McDonald Kreitman test that improves the test's statistical reliability on typical viral data sets.

Chapter 3

In chapter 3 I provide the first estimates of the rates of adaptive fixation for human influenza A virus subtypes H1N1 and H3N2. I develop a generalisation of the Smith and Eyre-Walker (2002) approach for calculating the number of adaptive sites which is informed by a combination of population genetic theory and empirical viral data. I then apply this method to whole genomes belonging to influenza virus subtypes H1N1 and H3N2 sampled over a 30 year period. The resultant time series provides overall rates of adaptation for all genes in the influenza genome for both subtypes. I also apply my generalised method to a larger data set of HA gene sequences from subtypes H1N1 and H3N2 and calculate rates of adaptation separately for HA subdomains HA1 and HA2. Finally I discuss my results in the context of previous research conducted on influenza virus selection and adaptation.

Chapter 4

In chapter 4 I introduce a new probabilistic approach to estimating the mutational site frequency spectrum from a set of sequences. I develop a new Bayesian counting approach which assigns sites to site classes in a probabilistic manner, thereby incorporating sampling error. Using this method it is possible to estimate separate site-frequency distributions for silent and replacement sites. I then use this approach to further generalise the methods described in Chapter 3, thus allowing the signals of deleterious mutation pressure and positive selection to be estimated within the same data set. Finally I test this method on artificial test data sets and data simulated under neutral evolution and then apply this method to within patient Hepatitis C virus data.

Chapter 5

In chapter 5 I extend the probabilistic counting approach developed in chapter 4 to produce a more powerful statistic specifically for evaluating the magnitude of deleterious mutational load from sequence data. I use a both the the 96 virus data set compiled in Chapter 2 as well as data simulated under neutral evolution to investigate the extent to which deleterious mutations contribute towards RNA virus diversity. I also compare my method to the DNS statistic (Pybus et al., 2007), which is a phylogenetic equivalent of my test.

Chapter 2

Detecting natural selection in RNA virus populations using sequence summary statistics

Published as: Bhatt, S., Katzourakis, A., and Pybus, O. G. (2009). Detecting natural selection in RNA virus populations using sequence summary statistics. *Infection, Genetics and Evolution*, 3:421–430.

AK provided assistance with computational analysis. OGP provided editorial and supervisory assistance.

2.1 Abstract

At present, most analyses that aim to detect the action of natural selection upon viral gene sequences use phylogenetic estimates of the ratio of silent to replacement mutations. Such methods, however, are impractical to compute on large data sets comprising hundreds of complete viral genomes, which are becoming increasingly common due to advances in genome sequencing technology. Here we investigate the statistical performance of computationally efficient tests that are based on sequence summary statistics, and explore their applicability to RNA virus data sets in two ways. Firstly, we perform extensive simulations in order to measure the type I error of two well-known summary statistic methods – Tajima's D and the McDonald-Kreitman test – under a range of virus-like mutational and demographic scenarios. Secondly, we apply these methods to a compilation of ~100 RNA virus alignments that represent natural RNA virus populations. In addition, we develop and introduce a new implementation of the McDonald-Kreitman test and show that it greatly improves the test's statistical reliability on typical viral data sets. Our results suggest that variants of the McDonald-Kreitman test could potentially prove useful in the analysis of very large sets of highly diverse viral genetic data.

2.2 Introduction

One of the main goals of viral evolutionary genetics is to understand to what extent natural selection – as opposed to mutation and random genetic drift – determines the genetic variability and evolution of viruses. Various methods of gene sequence analysis have been developed to detect and measure natural selection, the most popular of which can be categorised as either D_n/D_s-based methods (e.g. Nei and Gojobori, 1986) or methods based on site-frequency summary statistics (e.g. Tajima, 1989; McDonald and Kreitman, 1991a). The former calculate the ratio of replacement to silent genetic changes, which is typically denoted D_n/D_s or ω. A ratio greater than one indicates the action of positive selection, while a ratio of less than one can indicate purifying selection. In contrast, summary statistic methods depend on the frequency at which polymorphisms are found in a sample of sequences. These statistics may be computed from within-species polymorphisms (Tajima, 1989) or from both polymorphisms and among-species fixations (McDonald and Kreitman, 1991b).

Currently, most studies of viral genetic data use phylogenetic D_n/D_s methods as a means to detect selection (e.g. Pond and Frost, 2005; Yang, 2007), which are based on statistical models of codon evolution (Goldman and Yang, 1994; Yang, 2000). Examples of this approach are too numerous to list here, but one of the most influential was done by Nielsen and Yang (1998b) for an investigation of positive selection in the HIV-1 env gene. Phylogenetic D_n/D_s methods do not require users to make specific assumptions about the sampled population and can therefore provide robust evidence for the directionality of selection. In addition, simulations show D_n/D_s methods to have good statistical power under models of both positive and negative selection (Zhai et al., 2009), although in practice such methods are likely more powerful in detecting recurrent or reciprocal selection than single, historical selective sweeps (Pybus and Shapiro, 2008). However, the interpretation of D_n/D_s can be potentially misleading when recombination has been operating (Wilson and McVean, 2006) and, crucially, phylogenetic D_n/D_s methods can be very time consuming or impractical to compute on large data sets

(Pond and Frost, 2005). Recent developments in sequencing technology (Margulies et al., 2005) will make commonplace the publication of data sets containing hundreds or thousands of complete viral genomes, and therefore it is sensible to investigate the potential utility of alternative methods.

Site-frequency summary statistics such as Tajima's D (Tajima, 1989) have occasionally been used to analyze viral data sets. For example, Shriner et al. (2004) and Edwards et al. (2006a)applied versions of Tajima's D to HIV-1 and Tsompana et al. (2005) employed the test on the Tomato spotted wilt virus. In addition, tests that consider patterns of both polymorphism and divergence, notably the McDonald Kreitman (MK) test, have been applied to the Bovine immunodeficiency virus (Cooper et al., 1999), beak and feather disease virus (Ritchie et al., 2003) and North American Powassan virus (Ebel et al., 2001). Most pertinent to virus evolution, Williamson (2003) demonstrated that the MK test can be applied to "serially-sampled" sequences that are obtained from the same population at different time points, thereby estimating the rate of viral adaptation through time. Summary statistic methods are computationally very efficient, can potentially be applied to very large whole genome data sets, and perhaps are more robust to the effects of recombination than phylogenetic D_n/D_s methods. However, summary statistic methods typically assume that multiple mutations do not occur at the same nucleotide site, which may explain why they are rarely employed on rapidly-evolving viral data sets, but commonly applied to species with relatively low evolutionary rates, such as Drosophila (McDonald and Kreitman (1991a); Smith and Eyre-Walker (2002); Andolfatto (2005)).

In this chapter we investigate the utility and performance of two common summary statistic methods, Tajima's D statistic (Tajima, 1989, see section 1.2.2.1) and the MK test (McDonald and Kreitman, 1991a), section 1.2.3.1), when applied to RNA virus sequences. We perform extensive simulations of virus-like alignments in order to measure the type I error of these tests (i.e. the chance of falsely rejecting the hypothesis of neutral evolution). Second, we apply

the two tests to a collection of almost 100 RNA virus alignments that represent natural viral populations. Third, we develop and implement a new algorithm for computing the MK test which improves the performance of the test on data sets containing much genetic variation.

2.3 Methods

2.3.1 Investigating the performance of Tajima's D

To explore the reliability and type I error rate of Tajima's D statistic, we simulated alignments of neutrally-evolving sequences under various scenarios. Simulation was a two-step process. First, for each scenario, 500 neutral coalescent trees with 50 taxa were simulated. Second, one alignment of sequences, 6000*nt* in length, was simulated along each tree (see section 1.1.4.3).

Neutral coalescent trees were simulated using standard approaches (e.g. Hudson, 1990) which were implemented in the Java Evolutionary Biology Library (JEBL; available from *http://sourceforge.net/projects/jebl*). Coalescent trees were simulated under two scenarios, constant population size and exponential growth. The latter scenario was chosen because many viral populations of interest undergo a sustained increases in population size, either during an epidemic or, at a smaller scale, immediately following transmission to a new host.

For the constant population size scenario, trees were simulated under 28 logarithmically spaced values of θ, ranging from 0.00001 to 70. For the exponential growth scenario, trees were simulated under the same θ values plus a scaled growth rate $\rho = 200$. [Note that $\rho = {}^{r}/_{\mu}$, where r is the exponential growth rate of the population, hence $\theta\rho = N_e r$. If $\theta\rho \ggg 1$ then very star-like trees are generated; see Pybus et al., 1999]. These parameter ranges were chosen to include the range of values typical for RNA virus data sets (θ_s was calculated for all of the 100 viruses data set, representing the range of θ values present in viral populations).

A codon-based Markov substitution model (Goldman and Yang, 1994) was used to simulate neutrally evolving sequences along the coalescent trees, as implemented in PAML (Yang,

2007). The sequences were generated under $D_n/D_s = 1$ and with equal rates of transitions and transversions. One sequence alignment was generated for every simulated tree, meaning that for each value of θ, 500 alignments of 50 sequences were generated.

Tajima's D statistic was calculated for each simulated data set (section 1.2.2.1). Although Tajima (1989) used the beta distribution to calculate critical values the for test, Simonsen et al. (1995) argue that this approach leads to conservative values and a reduction in statistical power. Therefore we used parametric bootstrapping to obtain a null distribution and 95% critical values for D, as follows: (i) was calculated from the target data set, (ii) given this value, 1000 constant population size coalescent trees were simulated using the methods above, (iii) for each tree generated in step (ii) a sequence alignment was generated under the infinite sites assumption, following the method described in Simonsen et al. (1995), (iv) Tajima's D was calculated for each alignment generated in step (iii), resulting in a null distribution of the statistic, (v) the null hypothesis was rejected if the D value of the target data set fell outside the 95% critical values obtained in step (vi). The type we error of the test was then calculated as proportion of the 500 target data sets that rejected the null hypothesis.

2.3.2 Investigating the performance of the MK test

As explained in section 1.2.3.1, the MK test needs to discriminate between polymorphisms and fixed differences and therefore requires an outgroup sequence, taken from either a closely related species (Figure 1.3b) or an earlier time point (Figure 1.3a). We chose to simulate the latter situation, which can be easily represented using the serial-sample coalescent model (Rodrigo and Felsenstein, 1999) and also corresponds to situation investigated by Williamson (2003). Crucially, the results obtained are applicable to both situations, because the MK test depends on the genetic distance between the outgroup and ingroup, not on their relative positions in time (see Figure 1.3).

As before, simulations were undertaken on both constant population size and exponential

growth scenarios. For the former, two parameters were required to simulate the serial-sample coalescent trees,θ and $\tau = t\mu$, where t is the time elapsed between the earlier time point and the ingroup. A range of 13 logarithmically spaced θ values were chosen, ranging from 0.00001 to 1. For each θ value, 500 trees were simulated under 12 different τ values, ranging from 0.1 to 5. Each tree comprised 50 ingroup sequences plus one outgroup sequence sampled τ time units into the past. These serial-sample coalescent trees were simulated using JEBL (see above). For the exponential growth scenario, phylogenies were simulated under the range of θ and τ values described immediately above, with the addition of a scaled exponential growth rate of $\rho = 200$.

As before, a codon-based Markov substitution model (Goldman and Yang, 1994) was used to simulate neutrally-evolving sequences along the coalescent trees. One sequence alignment ($6000nt$ long) was generated for every simulated serial-sample tree, meaning that for each value of θ or τ, 500 alignments of 51 sequences were generated. As before, sequences were generated under $^{dn}/_{ds} = 1$ and with equal rates of transitions and transversions.

For each simulated alignment, the total number of sites in each category ($r_p, s_p r_d$ and s_d) were computed. We developed a new approach to this computation, explained below, and analysis of simulated alignments was performed using both the standard method and our new approach. A χ^2 test of independence was applied to the site totals for each of the 500 replicate alignments. Hence, for each specific combination of θ and τ, the type I error equals the proportion of the 500 χ^2 tests that were significant at the $p = 0.05$ level.

2.3.3 New proportional counting algorithm for the McDonald Kreitman test

The MK test requires that the number of sites belonging to different categories (r_p, s_p, r_d and s_d) are computed accurately. When sequence diversity (θ) is low this is straightforward, as the majority of sites will be either fixed or 1-state polymorphic (Figure 2.1 on the next page), that is, each mutation occurs at a different site. Furthermore, variable sites are unlikely

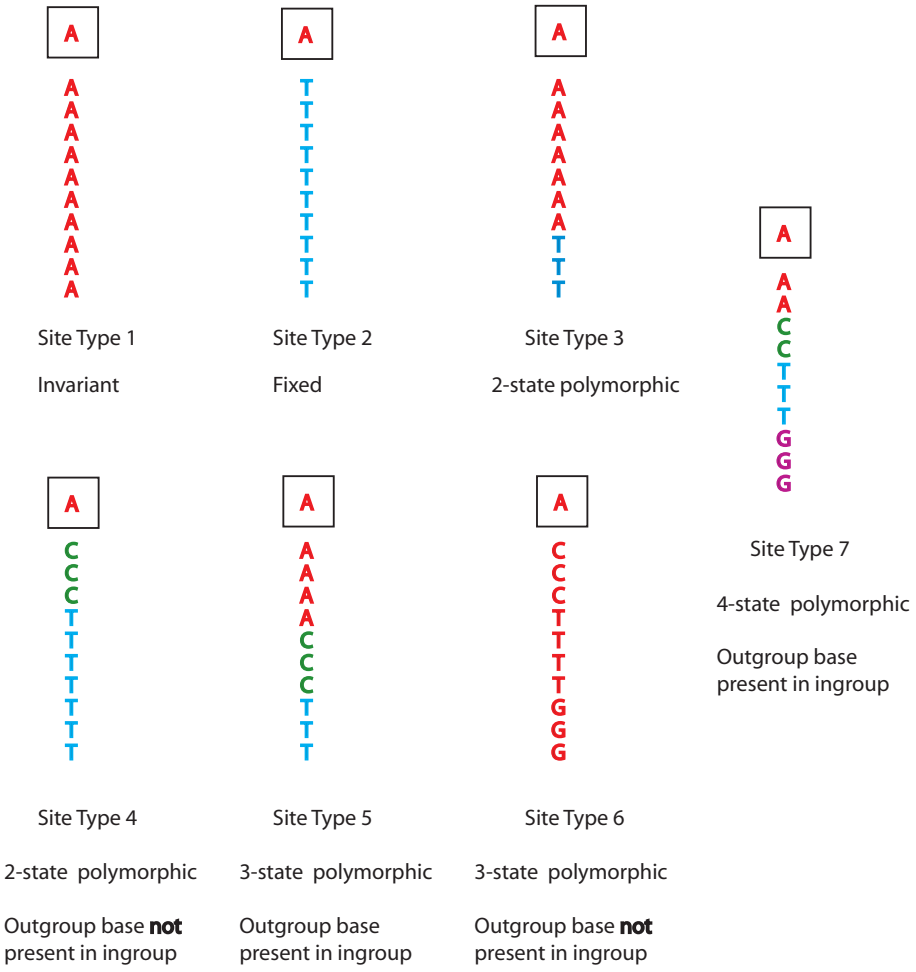


Figure 2.1: *An illustration of the seven "site types" defined in the main text. In each case, the box contains the nucleotide observed in the outgroup sequence. Below this is shown the site pattern observed in the ingroup alignment.*

to fall within the same codon hence each mutation can be easily categorized as silent or replacement. Therefore a simple count will suffice when the test is applied to animal genomes (e.g. McDonald and Kreitman, 1991a; Smith and Eyre-Walker, 2002).

However, a more sophisticated approach is needed for viral alignments, which can be highly diverse. As θ increases, there is a greater chance of observing sites which are 2, 3 or 4-state polymorphic (Figure 2.1), hence multiple mutations may occur at the same site or within the same codon. The categorization of sites therefore becomes more ambiguous – not accounting for this ambiguity could potentially introduce biases into the MK test. To avoid such biases

we have developed a "proportional" counting approach, described below, that incorporates the ambiguity in site categorization. A different, but related, approach was employed by Egea et al. (2008).

For a given ingroup alignment plus outgroup sequence, we define seven 'site types' that describe all the possible nucleotide patterns that could occur, illustrated in Figure 2.1. Rather than unambiguously assigning sites as fixed or polymorphic, we give each site i a "fixation score" F_i and a "polymorphism score" $P_i = (1 - F_i)$. If the site is definitely fixed then $F_i = 1$ and $P_i = 0$. Uncertainty in the status of a site is representing by assigning values between zero and one, as follows.

- SITE TYPE 1: All ingroup bases identical to the outgroup (invariant sites). $F_i = 0$ and $P_i = 0$.
- SITE TYPE 2: All ingroup bases identical but different from the outgroup (fixed sites). $F_i = 1$ and $P_i = 0$.
- SITE TYPE 3: Ingroup contains two bases, one of which is identical to the outgroup. $F_i = 0$ and $P_i = 1$.
- SITE TYPE 4: Ingroup contains two bases, neither of which is identical to the outgroup. McDonald and Kreitman (1991a) would classify this site as polymorphic (i.e. $F_i = 0, P_i = 1$). However, as no ancestral base is observed, the most plausible explanation is that an earlier fixation event has been followed another mutation at the same site. Classifying such sites as polymorphic would underestimate the number of fixations. Therefore $F_i = 0.5$ and $P_i = 0.5$.
- SITE TYPE 5: Ingroup contains three bases, one of which is identical to the outgroup. Observing an outgroup base increases the likelihood that neither of the two polymorphic bases has yet fixed. Therefore $F_i = 0$ and $P_i = 1$.

- SITE TYPE 6: Ingroup contains three bases, none is identical to the outgroup. As with site type 4, no ancestral bases are observed hence the most likely scenario is an earlier fixation followed by further mutations are the same site. Therefore $F_i = 1/3$ and $P_i = 2/3$.

- SITE TYPE 7: Ingroup contains all four bases. No reliable conclusion can be drawn, so we conservatively assign the site as $F_i = 0$ and $P_i = 1$. If such sites are common then the MK test should not be applied.

We also developed a proportional approach to evaluating whether a variable site is silent or replacement. Rather than unambiguously assigning sites as fixed or polymorphic, we give each site i a "silent score" S_i and a "replacement score" $R_i = (1 - S_i)$. For each site, the silent score is simply the proportion of ingroup bases that, if hypothetically inserted into the outgroup sequence, would not change the amino acid coded by the corresponding codon.

For a given alignment of k sites, the number of site in different categories (r_p, s_p, r_d and s_d) are straightforwardly computed from the proportional site scores as follows:

$$\begin{aligned} r_p &= \sum_{i=1}^{k} P_i R_i \\ s_p &= \sum_{i=1}^{k} P_i S_i \\ r_d &= \sum_{i=1}^{k} F_i R_i \\ s_d &= \sum_{i=1}^{k} F_i S_i \end{aligned}$$

2.3.4 Comparative analysis of RNA virus data sets

To investigate the performance of Tajima's D and the MK test on viral sequences, we utilized a previously published and curated collection of alignments from 100 different RNA virus species (see Shapiro et al., 2006 and Pybus et al., 2007for details). The alignments represent

partial or complete structural gene sequences and should well represent the behavior and diversity of RNA virus data sets. For each alignment, gene diversity (θ) was calculated using the Watterson estimator (θ_s) and Tajima's D statistic was calculated as described above.

In order to perform the MK test, it was first necessary to identify an outgroup sequence for each data set. we chose to use sister-species for outgroups, as serial-sampled outgroups were less common. Sister species outgroups were identified as follows: (i) representative sequences from each species were used as queries in a nBLAST search against the non-redundant database, resulting in a set of candidate outgroups; (ii) distance-based phylogenies and the viral taxonomic literature were used to choose the most closely-related candidate outgroup; (iii) the chosen outgroup was profile-aligned to the curated alignment using ClustalW2 (Larkin et al., 2007) and subsequently inspected and edited by hand, paying particular attention to codon structure. Using this approach, 96 RNA virus data sets were given a reliable sister-species outgroup and were subjected to the MK test, as described above. In addition, the mean pairwise genetic distance between the outgroup and ingroup sequences was calculated for each data set, using the Jukes-Cantor method (Jukes and Cantor, 1969).

2.4 Results

2.4.1 Investigating the performance of Tajima's D

Figure 2.2 on the following page shows the performance of Tajima's D test on neutral sequences simulated under different θ values and sampled from a constant-sized population. Figure 2.2a shows the type I error of the test, Figure 2.2b shows the average D value and Figure 2.2c shows the mean values of $\theta_s$and θ_k for each simulated value of θ. The statistical performance of the test depends greatly on θ. For explanatory convenience, we divide the range of θ into three regions.

- REGION ONE ($\theta < 10^{-4}$): Alignments generated under these low θ values have very

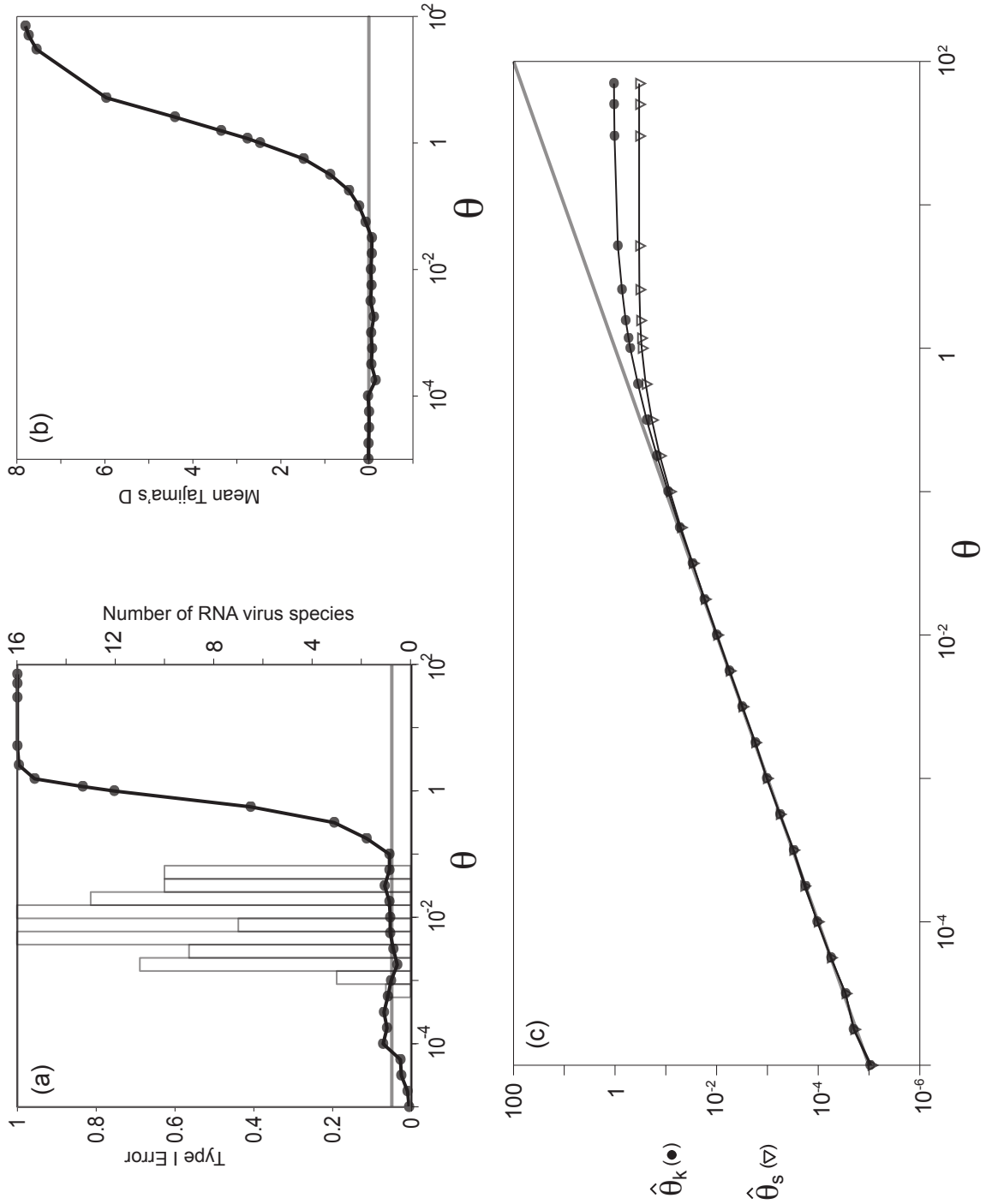


Figure 2.2: *The behaviour of Tajima's D under a constant population size coalescent model. Simulations were conducted under a range of θ values (horizontal axes). (a) The type I error of Tajima's D test for different values of θ. Each point represents the mean error of 500 simulations, and the grey line marks the expected 5% error rate. The superimposed histogram represents the distribution of θ values from 96 empirical RNA virus data sets (see text and Table 2.1). (b) The mean values of Tajima's D statistic for each values of θ. Each point represents the average of 500 simulations. The grey line marks the expected value of D, zero. (c) Mean values θ_s and θ_k for each simulated value of θ. Each point represents the average of 500 simulations. The gray line represents the expected relationship $\theta_k = \theta_s = \theta$.*

few polymorphic sites (0 to 4 per alignment). Although the error rate of the test appears low in this region (Figure 2.2a), alignments with such small amounts of variation are not suitable for analysis. Furthermore, simulations in this region are conditionally distributed, not random, because alignments with zero polymorphisms are discarded. Therefore the simulations from this region are ignored.

- REGION TWO ($10^{-4} < \theta < 0.1$): Tajima's D test performs very well in this region, with type I error rates close to 5% (Figure 2.2a) and a mean D value close to zero (Figure 2.2b). A small amount of measurement error is noticeable, as only 500 simulations were performed for each point.

- REGION THREE ($\theta > 0.1$): The error rate in this region rises rapidly as θ increases. If $\theta > 5$, the error rate is 100% (Figure 2.2a) and reach maximal values because all sites are polymorphic (Figure 2.2c). In this region multiple changes at the same site are observed, violating the 'infinite sites' assumption of the test and generating error. Both θ_s and θ_k under-estimate true θ. However the under-estimation is greater for θ_s, hence mean D> 0 (Figure 2.2b).

Figure 2.3 on the next page shows the performance of the Tajima's D test on neutral sequences sampled from exponentially-growing populations. These results differ from those simulated under constant population size (Figure 2.2 on the preceding page) in several ways. Firstly, at high θ values, we no longer observe multiple mutations at the same site. This is because, on average, the underlying phylogeny becomes shorter as $\theta\rho$ increases and therefore fewer polymorphisms are seen in the sample (Slatkin and Hudson, 1991). Secondly, as θ rises, average D becomes increasingly negative (Figure 2.3b) because exponential growth causes the phylogeny to become more star-like (Slatkin and Hudson, 1991). Therefore, as $\theta\rho$ increases, the level of diversity stabilizes (Figure 2.3c) and polymorphisms are more commonly seen at low frequencies, which results in comparatively lower values for θ_k than for θ_s (Figure 2.3c). Previous studies have shown that the transition from structured 'constant-size' phylogenies to

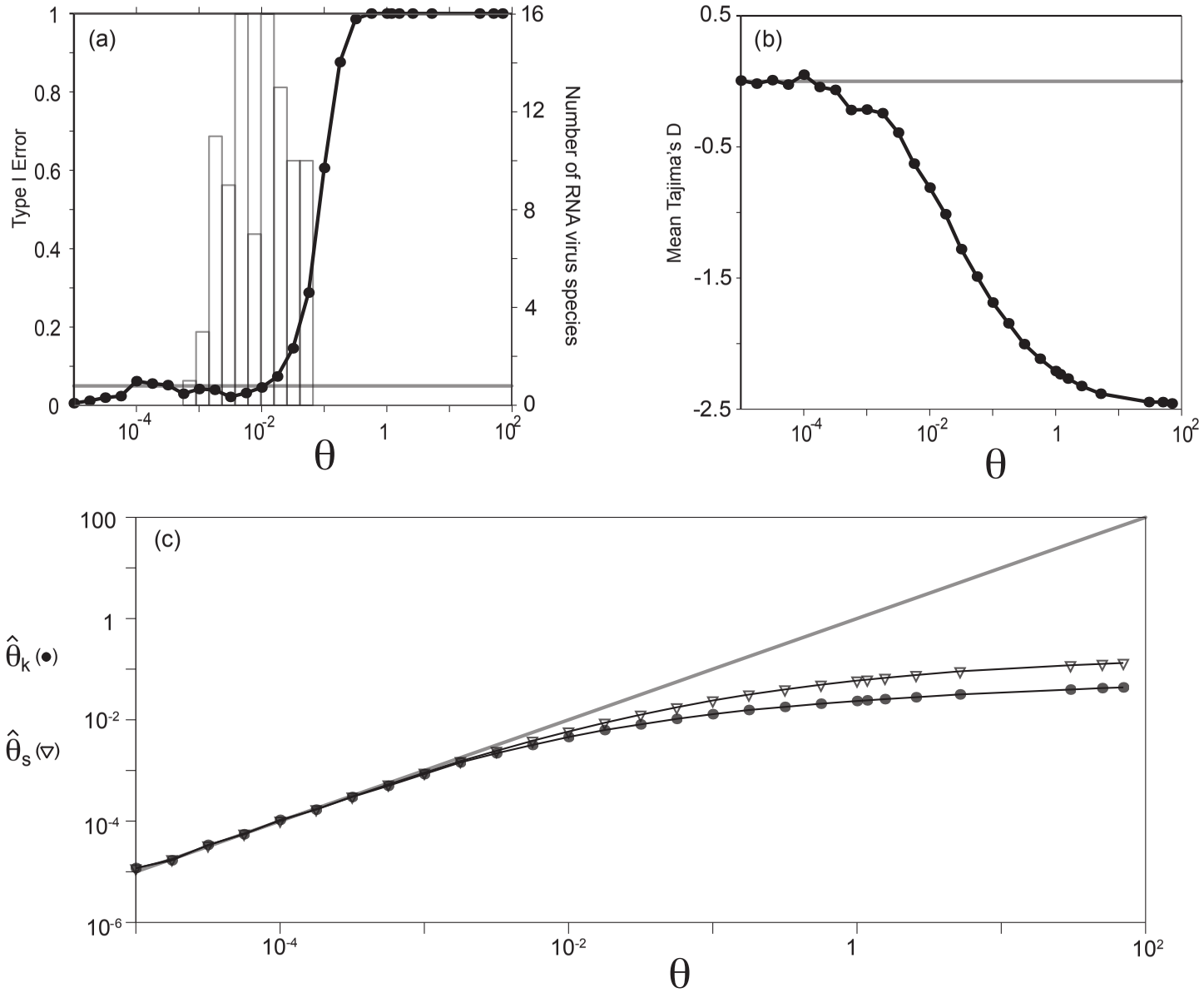


Figure 2.3: *The behaviour of Tajima's D under an exponential growth coalescent model. Simulations were identical to those shown in Figure 2.2 on page 44, except that a scaled population growth rate of ϱ=200 was used. See figure 1.3 on page 27 legend for further details*

star-like 'exponential growth' phylogenies occurs around $\theta\rho = 1$ (Slatkin and Hudson, 1991; Pybus et al., 1999). In our simulations we used $\rho = 200$, hence in Figure 2.3 this transition occurs around $\theta = 0.005$. Above this value, the error rate rises rapidly (Figure 2.3 a) and mean D values become significantly negative (Figure 2.3b), as expected by theory.

To test whether real RNA virus data sets are suitable for analysis using Tajima's D, we calculated θ and Tajima' D statistic for each of 96 RNA virus data sets (Table 2.1 on page 51). We find that 17 out of 96 (17.7%) of empirical data sets rejected the null hypothesis of neutral evolution. In addition, we superimposed the frequency distribution of these empirical θ values onto the type I error plots (Figure 2.3a and Figure 2.2a). If the sequences are assumed to come from a constant-sized population then all of the empirical data sets have θ values that lie within the working range of Tajima's D test (Figure 2.2a). If the sequences are assumed to come from an exponentially-growing population then the suitability of the test drops dramatically (Figure 2.3a). Hence the primary problem arising when Tajima's D test is applied to RNA viruses is not the invalidation of the infinite sites assumption caused by high mutation rates, but rather the sensitivity of the test to changing population size or population structure (Simonsen et al., 1995).

2.4.2 Investigating the performance of the MK test

Figure 2.4 on the following page shows the performance of the MK test on simulated neutral sequences, performed using both the standard counting method (McDonald and Kreitman, 1991a; Smith and Eyre-Walker, 2002) and our new "proportional" counting approach (see section 2.3.3). For both counting methods, the test was applied to neutral sequences from both constant-sized (Figures 2.4a and 2.4b) and exponentially-growing (Figures 2.4c and 2.4d) populations. In order to directly compare the simulation results with our empirical RNA virus data sets, we plot θ against the mean pairwise genetic distance between the ingroup and outgroup sequences (rather than against τ, which is unknown for our real data). The

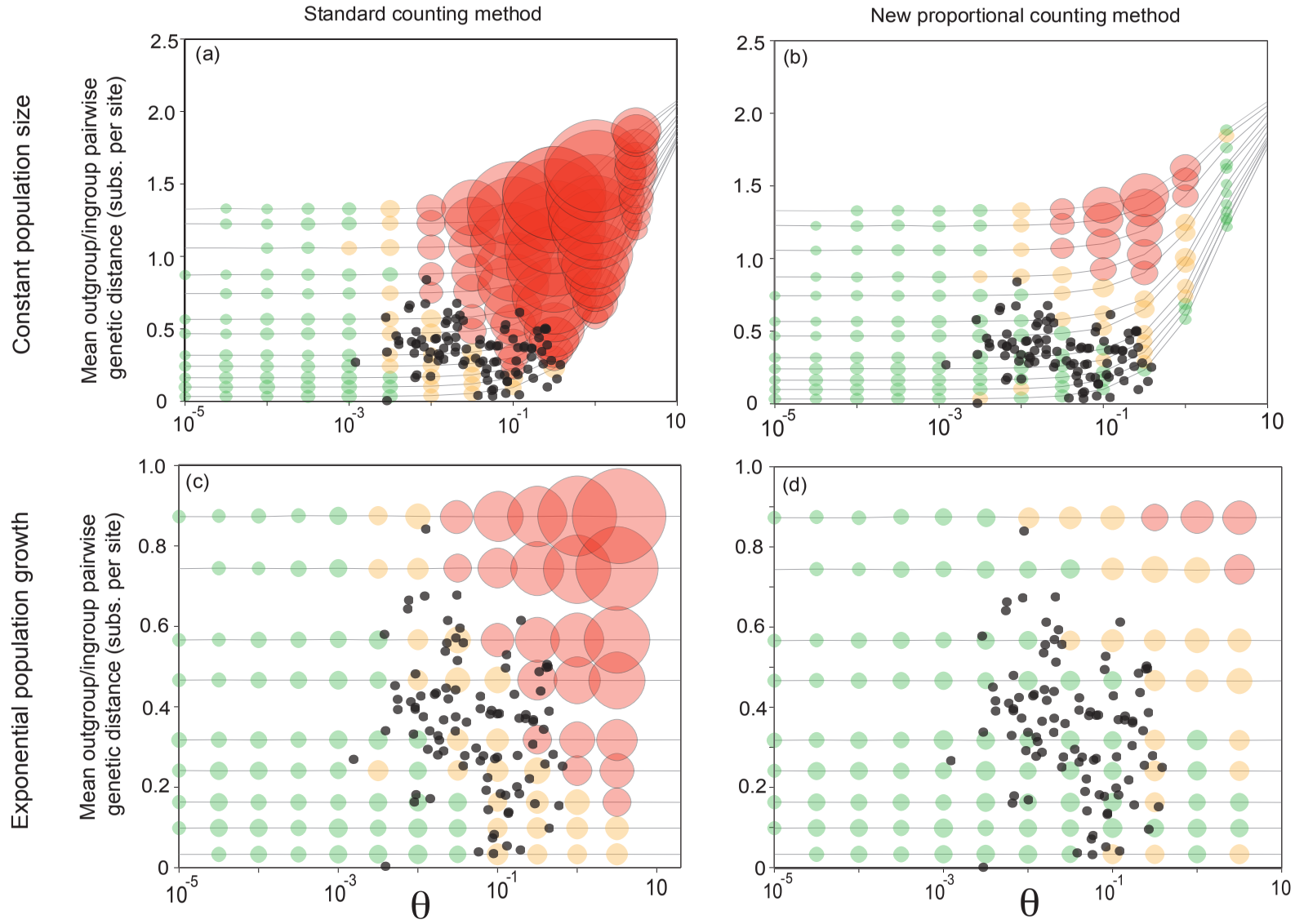


Figure 2.4: *The behaviour on the MK test on neutral sequences. Each plot shows the type I error of the MK test under different parameter values. For a given value of θ and τ, the diameter of each circle is proportional to the type I error of the test, averaged across 500 simulations. Green circles represent type I error rate <5%, orange circles between 5% and 10%, and red circles >10%. Thin grey lines connect simulations performed under the same value of τ but different values of θ. The superimposed black points represent the distribution in parameter space of the 96 empirical RNA virus data sets (see text & Table 2.1). (a) Standard counting method applied to sequences simulated under constant population size. (b) New proportional counting method applied to sequences simulated under constant population size. (c) Standard counting method applied to sequences simulated under exponential population growth. (d) New proportional counting method applied to sequences simulated under exponential population growth.*

performance of the MK test under each set of parameters (θ and τ) is represented as a circle whose diameter is proportional to type I error. Error rates between 5% and 10% are coloured orange and error rates greater than 10% are coloured red.

We begin by considering the results obtained for constant-sized populations (Figures 2.4a and 2.4b). For explanatory convenience, we again divide parameter space into different regions:

- REGION 1 ($\theta < 10^{-2}$): Alignments generated under these θ values exhibit few differences between the ingroup and outgroup, and of this variation, almost all is described by site types 2 and 3 (see figure 2.1 on page 40). As discussed in Methods, the interpretation of these is unambiguous, resulting in low type I error rates (below or around 5%) for both implementation methods (Figures 2.4a and 2.4b). Many non-viral data sets that have been analysed using the MK test have θ values that lie in this region (e.g. McDonald and Kreitman, 1991a; Smith and Eyre-Walker, 2002; Andolfatto, 2005).

- REGION 2 ($10^{-2} < \theta < 3$): In this range of θ values, the number of polymorphisms and fixed differences increases as $\theta\tau$ increases, as does the frequency of potentially ambiguous sites (site types 4, 5 and 6). Therefore type I error rises towards the upper-right hand corner of this region. When $\theta\tau$ is comparatively low, the standard counting method produces satisfactory error rates, but as $\theta\tau$ increases this method becomes unusable (Figure 2.4a). In contrast, our new proportional counting method performs very well on all but the very highest values of $\theta\tau$ (Figure 2.4b).

- REGION 3 (θ >3): This region is characterized by exceptionally high θ values, such that almost every ingroup site is polymorphic and contains multiple mutations (site types 6 and 7). Under these conditions it is impossible to accurately estimate the number of fixed differences (r_d, s_d). MK test is unlikely to generate meaningful and robust results in this region, even if the type I error sometimes appears low. It should be noted that, at high values of θ, our estimate of the mean pairwise genetic distance is overestimated

due to violation of the infinite sites assumption, causing the plot lines to deviate from linearity and increase.

The performance of the MK test on neutral sequences from exponentially-growing populations is shown in figure 2.4c and 2.4d. Under exponential growth the underlying phylogeny becomes star-like and its total length is reduced. As a result of the latter change, fewer mutations accrue and potentially ambiguous sites (site types 4, 5, 6 and 7; see figure 2.1 on page 40) are rarer, hence type I errors are lower than those obtained for constant populations (section2.4). Under exponential growth, only 40% of the sites were variable at the highest values of $\theta\tau$, whereas under constant population size, 95-100% of sites were variable when $\theta\tau$ was very high. As before, our proportional counting method exhibits lower type I error than the corresponding values obtained using the standard method.

We also calculated and the mean outgroup-ingroup pairwise genetic distance for 96 empirical RNA virus data sets (Table 2.1). These empirical values are superimposed as black dots on each plot in Fig. 5. It is clear that most RNA viral populations have comparatively high θ values ($0.001<\theta<0.5$ substitutions per site). Using the standard counting method, many data sets correspond to regions of parameter space with high type I error (Figures 2.4a and 2.4c, red and orange circles). However, our proportional counting method reduces type I error in this region so that almost all empirical data sets correspond to parameter regions with low error (Figures 2.4b and 2.4d, green circles). Table 2.1 also shows that the MK test rejected the null hypothesis of neutrality for 70 out of 96 (78.9 %) empirical RNA virus data sets.

Table 2.1: *Summary statistics and McDonald–Kreitman test p-values for RNA virus data sets.*

RNA virus (gene)	Sequence diversity $\frac{\theta_s}{2}$	Tajima's D statistic	Mean pairwise ingroup to outgroup Genetic Distance	*pValue* of MK test
Australian bat lyssavirus (G)	0.0088	-1.6204	0.3398	**0.0004**
Acute bee paralysis virus (C)	0.0142	0.4417	0.3273	<**0.0001**
Akabane virus (NP)	0.0127	-0.1987	0.3808	<**0.0001**
Avian influenza A, serotype H5N1 (NP)	0.0264	-0.6054	0.2120	0.2382
Avian influenza A, serotype H7N1 (HA)	0.0541	-0.1222	0.3353	0.1121
Avian pneumovirus (N)	0.0070	0.8327	0.2779	**0.0008**
Barley yellow mosaic virus (CP)	0.0116	-1.0804	0.3923	**0.0189**
Bean yellow mosaic virus (CP)	0.0490	1.2207	0.3666	**0.0335**
Bluetongue virus (VP7)	0.0462	2.7994	0.5042	<**0.0001**
Bovine rotavirus (VP7)	0.0706	1.8978	0.3615	**0.0001**
Crimean-Congo haemorrhagic fever virus (NP)	0.0515	-0.2500	0.4981	<**0.0001**
Canine distemper virus (H)	0.0363	-1.1216	0.3627	**0.0079**
Chikungunga virus (E1)	0.0355	-0.3933	0.2935	0.7942
Classical swine fever virus (E2)	0.0485	0.7598	0.4454	<**0.0001**
Clover yellow vein virus (CP)	0.0431	0.1443	0.3433	<**0.0001**
Coxsackievirus B4 (VP1)	0.0398	2.0165	0.4862	<**0.0001**
Curcurbit yellow stunting disease virus (CP)	0.0184	0.6448	0.4175	<**0.0001**
Dengue virus, serotype 1 (E)	0.0336	-0.4862	0.3672	<**0.0001**
Dengue virus, serotype 1 (CM)	0.0289	0.0696	0.3928	<**0.0001**
Dengue virus, serotype 2 (E)	0.0408	-0.6938	0.4410	<**0.0001**
Dengue virus, serotype 3 (E)	0.0282	-0.2544	0.3700	<**0.0001**
Dengue virus, serotype 4 (E)	0.0254	-0.5659	0.4298	<**0.0001**
Dobrava virus (N)	0.0356	0.4805	0.2731	<**0.0001**
Eastern equine encephalitis virus (C)	0.0461	-1.5826	0.2968	**0.0044**
Eastern equine encephalitis virus (E1)	0.0508	-1.1961	0.5139	<**0.0001**
Enterovirus 71 (VP1)	0.0377	1.7037	0.5572	<**0.0001**
Equine influenza, serotype H3N8 (HA)	0.0245	-1.1273	0.2701	<**0.0001**
Feline immunodeficiency virus (Gag)	0.0412	1.2206	0.4259	<**0.0001**
Human influenza A virus, serotype H3N2 (HA)	0.0307	-0.7710	0.2645	<**0.0001**
Human influenza A virus, serotype H3N2 (NP)	0.0198	-0.6619	0.2149	**0.0088**
Garlic latent virus (CP)	0.0588	2.2698	0.6762	<**0.0001**
Hepatitis C virus 1b (C)	0.0361	-1.3132	0.1630	0.6081
Hepatitis C virus 1b (E1E2)	0.0551	-0.1017	0.4250	0.3703
HIV type 1, subtype B (Env-Gap)	0.0552	-1.1956	0.2203	0.4129
HIV type 1, subtype B (Env)	0.0536	-1.0352	0.1875	**0.0277**
HIV type 1, subtype B (Gag)	0.0377	-1.6711	0.2527	0.1574
Human polio virus type 2 (VP)	0.0463	-0.3438	0.3390	<**0.0001**
Human respiratory syncytial virus A (G)	0.0436	-1.0333	0.4381	**0.2953**
Human respiratory syncytial virus A (N)	0.0416	2.0114	0.2383	**0.0322**

RNA virus (gene)	Sequence diversity $\frac{\theta_s}{2}$	Tajima's D statistic	Mean pairwise ingroup to outgroup Genetic Distance	*pValue* of MK test
Human respiratory syncytial virus B (G)	0.0245	-0.6424	0.4301	**0.1642**
Hantaan virus (G1)	0.0603	0.4124	0.3797	<**0.0001**
Hantaan virus (N)	0.0460	1.5131	0.3060	<**0.0001**
Viral hemorrhagic septicaemia virus (GP)	0.0148	-0.9545	0.6637	**0.0211**
Viral hemorrhagic septicaemia virus (N)	0.0254	-1.2692	0.6738	<**0.0001**
Highlands J virus (E1)	0.0060	-1.8739	0.2687	**0.5030**
Human astrovirus (C)	0.0617	2.1755	0.2958	0.2097
Human parainfluenza virus type 1 (HN)	0.0185	-0.7862	0.4116	**0.0363**
Human parainfluenza virus type 3 (HN)	0.0171	-0.2453	0.6420	<**0.0001**
Infectious pancreatic necrosis virus (VP2)	0.0435	2.0159	0.3011	**0.6452**
Japanese encephalitis virus (CP)	0.0240	-0.2458	0.3979	<**0.0001**
Japanese encephalitis virus (E)	0.0419	-0.2765	0.3743	<**0.0001**
Junin virus (NP)	0.0291	-1.0405	0.2784	**0.2412**
Leek yellow stripe virus (CP)	0.0614	1.4954	0.5699	<**0.0001**
Lettuce mosaic virus (CP)	0.0392	-0.9932	0.4807	**0.0652**
Maize dwarf mosaic virus (CP)	0.0401	0.2082	0.2797	<**0.0001**
Measles virus (HA)	0.0188	-0.2067	0.4955	<**0.0001**
Measles virus (N)	0.0307	-1.3780	0.3815	**0.0081**
Mumps virus (NP)	0.0142	-0.0931	0.5786	<**0.0001**
Onion yellow dwarf virus (CP)	0.0601	2.1301	0.4228	<**0.0001**
Oropouche virus (NP)	0.0156	0.1407	0.3822	**0.0022**
Pea seed-borne mosaic virus (CP)	0.0336	-0.2746	0.5284	<**0.0001**
Peanut stripe virus (CP)	0.0189	-0.3502	0.1226	**0.5789**
Polio virus, serotype 1 (VP1)	0.0481	1.7750	0.3881	<**0.0001**
Porcine rotavirus (VP7)	0.0742	2.4927	0.4624	<**0.0001**
Potato virus A (CP)	0.0288	-1.1946	0.4298	0.0065
Potato virus S (CP)	0.0453	-1.0140	0.4966	<**0.0001**
Potato virus X (CP)	0.0259	-1.1872	0.8398	<**0.0001**
Prunus necrotic ringspot virus (CP)	0.0244	-1.1962	0.6136	**0.4305**
Puumala virus (G2)	0.0511	1.4784	0.4278	<**0.0001**
Puumala virus (N)	0.0418	2.2214	0.3981	<**0.0001**
Rabies virus (G)	0.0554	0.3381	0.3610	**0.0033**
Rabies virus (N)	0.0505	0.7230	0.2571	0.7655
Rice black streaked dwarf virus (CP)	0.0205	-0.2376	0.1347	0.8355
Ross River Virus (E2)	0.0119	-0.2745	0.3393	**0.0120**
Rotavirus A (VP7)	0.0374	-1.7368	0.1587	**0.0005**
Rotavirus C (VP7)	0.0187	-1.3955	0.1797	**0.1248**
St. Louis encephalitis virus (E)	0.0328	-0.3246	0.4187	<**0.0001**
Sendai virus (NP)	0.0214	1.5404	0.3313	**0.5678**

RNA virus (gene)	Sequence diversity $\frac{\theta_s}{2}$	Tajima's D statistic	Mean pairwise ingroup to outgroup Genetic Distance	$pValue$ of MK test
Simian foamy virus (Env)	0.0107	-1.5075	0.4515	0.3615
Soybean mosaic virus (CP)	0.0290	0.0102	0.1811	<**0.0001**
Sugarcane mosaic virus (CP)	0.0333	0.4798	0.2232	0.2193
Sweet potato feathery mottle virus (CP)	0.0496	0.0511	0.3707	**0.0183**
Swine influenza virus, serotype H3N2 (HA)	0.0439	0.5441	0.3351	**0.0001**
Tick-borne encephalitis virus (E)	0.0603	0.8008	0.4280	<**0.0001**
Tomato spotted wilt virus (N)	0.0151	-0.8911	0.2742	**0.4981**
Tula virus (NP)	0.0479	1.4770	0.3505	0.0000
Turnip mosaic virus (CP)	0.0287	-1.6289	0.3746	**0.0873**
Venezuelan equine encephalitis virus (C)	0.0647	1.7385	0.6133	<**0.0001**
Venezuelan equine encephalitis virus (E)	0.0464	0.4981	0.5943	<**0.0001**
Western equine encephalitis virus (E1)	0.0248	-1.6399	0.2774	**0.6705**
West Nile virus (E)	0.0117	-0.9405	0.3944	**0.0026**
Wheat streak mosaic virus (CP)	0.0292	-2.1307	0.3162	**0.0003**
Wheat yellow mosaic virus (CP)	0.0129	-1.5503	0.3900	0.4392
Yellow fever virus (E)	0.0487	2.0571	0.5578	<**0.0001**
Yam mosaic virus (CP)	0.0476	0.4151	0.5368	<**0.0001**
Zucchini yellow mosaic virus (CP)	0.0436	-0.0075	0.4334	<**0.0001**

2.5 Discussion

It is widely acknowledged that Tajima's D is sensitive to changes in population size or the existence of population structure (e.g. Simonsen et al., 1995; Nielsen, 2005). A further concern with using Tajima's D test on viral populations is that their high evolutionary rates would invalidate the test's key assumption that each mutation occurs at a different site (the 'infinite sites' assumption). In our study we simulated sequences under an exhaustive range of θ values to assess the type I error of Tajima's D and also analysed a compilation of 96 alignments in order to determine the empirical range of θ values for RNA viruses (Table 2.1). We conclude that for constant size populations, violation of the infinite sites assumption is not the primary problem – all our RNA viral data sets have θ values that lie in the working range of the Tajima's D test. In contrast, We find that exponential population growth causes a reduction

in the working range of Tajima's D test, which is expected given that critical values for the test are obtained under the assumption of constant population size. It is usually impossible to know *a priori* if a sampled population meets this and other assumptions of Tajima's D test. For example, it is likely that many viral populations of interest will have experienced a complex form of population growth (during an epidemic or directly following transmission to a new host) or been subject to population structure. One possible solution to this problem was developed by Edwards et al. (2006b), who, for each data set, simulated a null distribution of D upon phylogenies whose shapes are highly supported by the data. This approach will reduce the error rate of Tajima's D test, but at the cost of reduced computational efficiency.

We examined the error rate of the MK test using both the standard site counting method (McDonald and Kreitman, 1991a) and our new proportional counting approach (see Methods). Using the former method, we observed raised type I errors when θ was larger than 0.01 (figure 2.4a). In this region of parameter space most variable sites are 2, 3 or 4-state polymorphic (site types 4-7; figure 2.1 on page 40) and the standard method classifies such sites as polymorphic. McDonald and Kreitman (1991b), in response to Whittam and Nei (1991) and Graur (1991), justified their implementation by arguing that any algorithm for typing substitutions as fixed or polymorphic will affect the numbers of replacement and silent substitutions equally and therefore not affect their test, which depends on ratios. This argument appears correct when the infinite sites assumption is valid. However, if $\theta\tau$ is large then nucleotide saturation occurs and the alignment contains more sites of types 4 and 6 (figure 2.1). Because these sites are, on average, more likely to be replacement sites than silent, the standard counting method (which always classes such sites as polymorphic) will tend to overestimate the number of replacement polymorphisms relative to silent polymorphisms. Simulation results show that our new proportional counting method is more robust, allowing the MK test to be applied even when $\theta\tau$ is very high (figure 2.4b) and reducing the average type I error (over all parameter space) from 20.4% to 5.3%.

The MK test has a lower error rate on sequences sampled from a growing population than on corresponding sequences sampled from a constant population (figure 2.4). This is because population growth results in shorter trees that accrue fewer mutations, leading to proportionally fewer 2, 3 and 4-state polymorphic sites (figure 2.1). Lastly, we note that the empirical RNA virus data sets are placed in a region of parameter space associated with high type I error if the standard counting method is used (figure 2.4a). However, the MK test has good statistical properties in this region if the new proportional counting method is used, indicating that this approach can be reliably applied to RNA virus genomes (figure 2.4b).

Our analyses focussed on the type I error of Tajima's D and the MK test and did not directly measure the probability of failing to reject the null hypothesis when it is false (type II error, or statistical power). Measuring the statistical power of neutrality tests is a formidable task, owing to the computational difficulties of simulating sequences under selection. Under selection it is not possible to separately simulation the mutational and genealogical process (which is possible under the neutral coalescent), making simulation of sequences only possible by computationally intensive forward simulations. Furthermore, the multiplicity of possible scenarios under which selection could occur (see Zhai et al. (2009)) make it difficult to obtain results of general applicability. However, we applied Tajima's D and the MK test to a compilation of 96 RNA virus data sets (Table 2.1) and found that the MK test rejected the null hypothesis of neutrality more than four times as often as Tajima's D (72.9%. and 17.7% of data sets, respectively). Since our simulation results demonstrate that the MK test (when used with the proportional counting method) has correct type I error, we conclude that the MK test has significantly greater statistical power than the Tajima's D test. We observed a positive correlation between the p-value of the MK test and the mean ingroup/outgroup pairwise genetic distance. The failure of the MK test to reject neutrality for ~30% of the RNA virus data sets may be a consequence of the presence of low frequency, slightly deleterious mutations that have yet to be purged by purifying selection. Such mutations appear to be common in RNA virus populations (Pybus et al., 2007; Hughes and Hughes, 2007). Charlesworth and

Eyre-Walker (2008) show that deleterious mutations can reduce the power of the MK test to detect adaptive evolution and propose the removal of low frequency polymorphisms as a solution to this problem.

In summary, our results indicate that the MK test with proportional site counting is suitable for analysis on RNA virus data sets. This test is quick to compute and makes a minimal number of assumptions, making it potentially more useful for the analysis of very large scale genomic datasets than D_n/D_s methods. However, unlike D_n/D_s methods, the MK test cannot be used to pinpoint specific sites under selection, although it can estimate the rate of adaptive substitution of a gene (Smith and Eyre-Walker, 2002). The MK test can also be applied to intra-species data sets that have been sampled serially through time (Williamson, 2003). When applied to intra-species data, the MK test will likely have low type I error (as $\theta\tau$ will be small) but could be statistically weak if the ingroup/outgroup genetic distance is low.

Chapter 3

Estimating the genomic rate of Influenza A adaptation

3.1 Introduction

3.2 Introduction to influenza

In chapter 3 I develop a generalised method based on the Smith and Eyre-Walker (2002) approach, for estimating the rate of adaptive fixations in a viral population. I apply this method to human influenza A virus data from subtype H1N1 and H3N2. Therefore in this section I cover the basics of influenza biology and highlight the sources of influenza variability and the mechanisms that shape this variability.

3.2.1 General introduction to the influenza virus

Human influenza is a classic example of a viral disease for which continued evolution and adaptation is of paramount importance for annual epidemics and occasional pandemics. Influenza virus is transmitted though the air, from the respiratory tract of an infected person or by direct contact with respiratory droplets. Once infected the incubation period ranges from one to five days, after which symptoms begin to develop. Typical influenza disease is characterised by an abrupt onset of fever, aching muscles, sore throat and dry non-productive cough (Webster et al., 1992). Most infected people recover within two weeks without requiring medical treatment, however, in the very young, the elderly, and those with other serious medical conditions, infection can lead to severe complications of the underlying condition, pneumonia and death. It should also be noted that, while rates of infection are highest among children, and death rates are highest among the elderly, influenza viruses causes disease to persons of all ages (Webster et al., 1992).

The high virulence and emergence of new influenza viruses is attributed to two dominant processes that cause surface antigens to change: antigenic drift and antigenic shift. Antigenic drift is a mechanism that involves that involves the accumulation of mutations within the antibody-binding sites of the viral envelope genes, so that the resulting viruses cannot be

inhibited by antibodies against previous strains, making it easier for the mutated strain to spread throughout a partially immune population. Antigenic shift is a sudden shift in the antigenicity resulting from the reassortment of the genomes of coinfecting viral strains.

Influenza spreads in seasonal epidemics: according to the United State Center for Disease Control a serious seasonal epidemic can cost an estimated $12 billion. More seriously, epidemics can spread across large regions (continental or worldwide) and become devastating pandemics. There have been several major pandemics throughout human history but undoubtedly the most severe was the 'Spanish Flu' Pandemic (1918-1920) (Crosby, 1976; Beveridge, 1991). An estimated one third of the worlds population (500 million people) were infected and had clinically apparent illnesses, and case fatality rates were $> 2.5\%$ compared to $< 0.1\%$ in other influenza pandemics (Frost, 1920). The total death toll is estimated between 50 - 100 million persons (Reid, 1999; Johnson and Mueller, 2002). Since the 'Spanish Flu' (branded one of the 'the greatest medical holocausts in history' (Potter, 2001)), three other pandemics have occurred (1957 'Asian Flu', 1968 'Hong Kong Flu' and 2009 'Swine Flu'). Previous influenza pandemics have occurred unpredictably and show great variation in mortality.

The severity of seasonal influenza epidemics and the continual threat of serious influenza pandemics, makes research into the evolutionary and epidemiological dynamics of influenza of critical importance.

3.2.2 Influenza biology

Influenza is an RNA virus belonging to the family Orthomyxoviridae, a family comprised of five genera: influenza A,B,C, Thogotovirus and Isavirus. Of these, influenza A is of greatest interest to researchers as infections have been reported in a variety of animal species including humans, pigs, horses, sea mammals, mustelids (known more commonly as the weasel family) and birds, and is also the cause of all human influenza pandemics. Influenza viruses A, B and C are very similar in structure with a roughly spherical viral particle of diameter between

80-120nm (Webster et al., 1992). Two large surface gylcoprotiens are present on the viral particle: *Haemagglutinin* (HA) and *Neuraminidase* (NA). Within the bilipid layer envelope each influenza A viral particle contains 8 different anti-sense RNA segments (figure 3.1 on the following page), which encodes 10 proteins: *HA,NA,NP,M1,M2,NS1,NS2,PA,PB1* and *PB2* (Palese, 1977; Lamb, 1989).

3.2.2.1 Influenza virus proteins

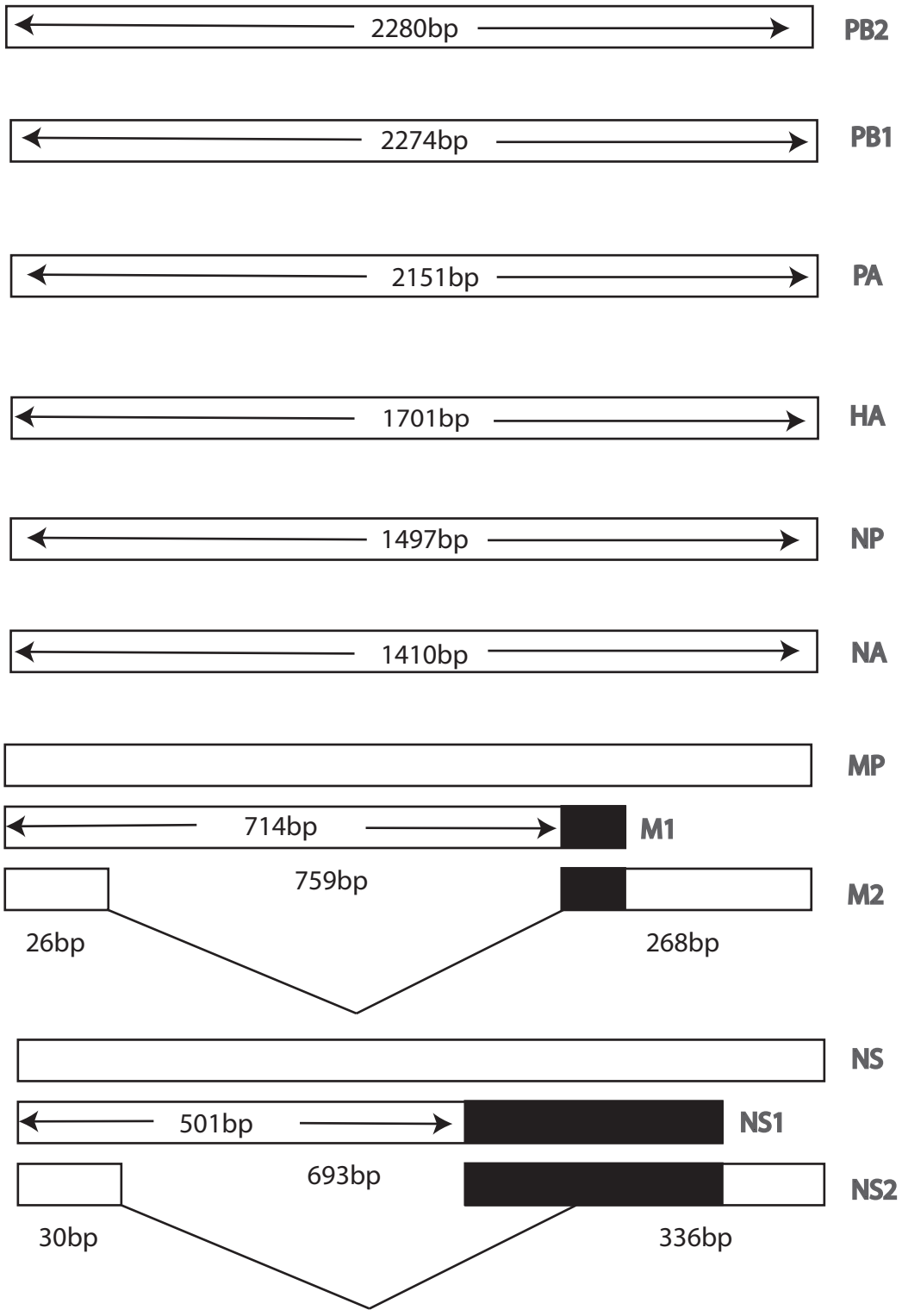
The virus RNA polymerase consists of a hetrotrimer formed by the PB1, PB2 and PA proteins, which are responsible for viral RNA synthesis (Detjen et al., 1987).

- *PB2 polymerase :* Encoded by RNA segment 1, it functions during the transcription initiation of viral mRNA and recognises and binds to the cap of the primer RNA(Perales and Ortin, 1997).
- *PB1 polymerase :* Encoded by RNA segment 2 and functions in the RNA polymerase complex as the protein that catalyzes the addition of nucleotides to growing viral mRNA chains (Webster et al., 1992).
- *PA polymerase :* Encoded by RNA segment 3, Its role in viral RNA synthesis is unknown, but there is evidence for a possible role as a protein kinase (Webster et al., 1992; Hara et al., 2006).

The two integral membrane proteins are HA and NA, and are responsible for host cell binding and subsequent fusion of viral and host membranes in the endosome after the virus has been taken up by endocytosis. The viral RNA transcribed into mRNA's are attached to ribonucleoprotiens (RNPs).

- *HA (Hemagglutinin) protein :* Encoded by RNA segment 4. HA is an integral membrane protein, which is responsible for mediating the binding of virons to host cell antibody

Figure 3.1: *Genome Map of influenza A. Gene lengths are in nucleotide base pairs from start to stop codon. For genes PB2,PB1,PA,HA,NP and NA, protein sequences are derived from colinear transcription. Genes MP and NS yield two proteins, one colinear and one spliced.*



receptors and for fusion between the viron envelope and the host cell (Wiley and Skehel, 1987). HA is the major surface antigen of the virus against which neutralising antibodies are produced and,therefore, undergoes considerable antigenic variation which leads to recurrent epidemics and pandemics (see section 3.2.3.2).

- *NP (Nucleoprotein) protein :* Encoded by RNA segment 5. NP functions by encapsidating the virus genome to form a ribonucleoprotein (RNP) particle for the purposes of transcription, replication and packaging (Portela and Digard, 2002). The protein is the second most abundant protein in the influenza viral viron, and is a target of host T-cell immune responses (Webster et al., 1992).

- *NA (Neuraminidase) protein :* Encoded by RNA segment 6. NA is an integral membrane protein, a second major surface antigen of the viron (after HA). Its function is to free viral particles from host cell receptors. In doing so NA mediates the release of virons from the cells in which they arose, to facilitate viral spread. In addition like HA, NA experiences considerable immune pressure from host antibodies and therefore undergoes antigenic variation (Varghese et al., 1983).

- *M1 and M2 protein :* Encoded by RNA segment 7. Colinear transcription of segment 7 yields mRNA for the M1 matrix protein (a structural protein that links the viral envelope with the virus core). The M1 protein forms a shell surrounding the viron nucleocapsid, underneath the viron envelope. It is the most abundant protein in the influenza virus virion. M2 is an integral membrane protein derived from splicing, which serves as a signal for transport to the cell surface (Webster et al., 1992).

- *Nonstructural NS1 and NS2 :* Encoded by RNA segment 8. NS1 mRNA is colinear with viral RNA and NS2 mRNA is derived from splicing. NS1 and NS2 are multifunctional proteins that are speculated to play roles in virus replication and in evading both innate and adaptive host immune responses (Fernandez-Sesma et al., 2006; Lin et al., 2007). It has also been proposed that NS1 performs several regulatory functions during the

viral replication cycle, including the regulation of synthesis, transport, splicing, and translation of mRNAs (Garcia-Sastre et al., 1998).

3.2.2.2 Influenza virus infection and replication cycle

The first step of infection involves the influenza virus virion binding to receptors on the surface of the host cell. Specifically, HA binds to the salic acid sugars found on glycoprotiens or glycolipid receptors of the host epithelial cells (Skehel and Wiley, 2000) forming a trimer which is responsible not only for the binding of virions to the cell surface but also for entry of the virus genome into the cell via fusion. After binding to the cell receptors, virions enter the cell within an endosomal vesicle. Acidification of the endocytic vesicle occurs (M2 facilitates acidification of the vesicle by a small proton channel in the viral membrane (Cros and Palese, 2003) causing two conformational changes. In HA, acidification results in the protein moving a short fusion peptide out of the hydrophobic region of HA and towards the endosomal membrane, where is inserts itself and causes fusion between viral and endosomal membranes. The second irreversible conformational change occurs in M1, the effect of which causes the release of viral Ribonucleoproteins (RNPs : a complex of proteins NP,PB1,PB2,PA) into the cytoplasm of the cell (Sugrue and Hay, 1991).

Mediated by nuclear transport signals of NP, the viral RNPs then migrate into the host cell nucleus (O'Neill et al., 1995) and begin primary transcription of mRNA, to produce necessary proteins for replication. These primary transcripts are exported out of the nucleus into the cytoplasm for translation of viral proteins used in early stages of the infection; NP and NS1. The increased concentration of free NP triggers a shift from mRNA synthesis to vRNA (viral RNAs) synthesis by the infecting viral genome. It is these vRNAs which are encapsidated by NP causing the formation of genome containing RNPs. The RNPs then function as templates for the secondary transcription of viral mRNAs which produce the principle translation products: M1,M2,HA and NA. The two membrane proteins, HA and

NA, are posttranslationally processed and transported to the cell surface (Nayak et al., 2004).

Morphogenesis of influenza virus is a complex process. First all viral sub-components are directed to the plasma membrane (NS2 comes into play here as the protein contains a nuclear transport signal). The subcomponents are: The vRNP viral core, M1 forming the bridge between the envelope and vRNP and the envelope (HA,NA and M2). The interaction and concentration of these subcomponents initiates bud formation (Nayak et al., 2004). After bud completion the influenza virus virions are released from salic acid by NA and the host cell dies. Cell to cell spread and transmission then continues.

3.2.2.3 Immune responses to Influenza virus infection

Both Mucosal and systemic immunity contribute to resistance to influenza infection. Antibodies secreted locally in the upper respiratory tract are a major factor in resistance to viral infection (Clements et al., 1986; Clements and Murphy, 1986). The immune response induced by infection protects against reinfection with the same virus or an antigenically similar viral strain (cross protection). Infection may lead to long-lasting immunity to the infecting virus, as demonstrated by the reappearance of the influenza A H1N1 subtype in 1977, where only subjects under the age of 20 years became infected (Cox et al., 2004) . However, the combination of frequent reassortments coupled with high mutation rates (see section 3.2.3.2) allows the virus to undergo unpredictable changes which greatly reduces the effective period of cross protection provided by the hosts immunity.

The immune response to influenza virus infection consists of two components: the humoral and the cell mediated response. B cell lymphocytes play a large role in the humoral immune response (first line of defence) and is therefore a major contributing factor to immunity from influenza infection. The costimulation of helper T cells and B cells leads to the production of antibodies against different influenza antigens, of which HA-specific antibody's are important for the neutralisation of the virus, and NA-specific antibodies are important in lessening the

release of the virus from infected cells. The cell mediated response plays a role in recovery from influenza infection but does not seem to contribute to preventing infection (McMichael et al., 1983) . The cell mediated response is mediated by cytotoxic T lymphocytes (CTLs) along with specific antibodies and the complement system (Jurgensen et al., 1973). CTLs express receptors that can recognise antigenic peptides bound to class I MHC molecules, and as a result the CTL response is mainly directed at internal proteins such as NP ((McMichael et al., 1983, 1986)).

3.2.3 Epidemiology and influenza variability

3.2.3.1 Seasonality of influenza

Influenza incidence exhibits strong seasonal fluctuations in temperate regions throughout the world, concentrating the medical and economic burdens into a few months every year. The underlying cause of seasonal oscillations in influenza incidence is unclear, but these fluctuations are possibly due to some mechanism that causes seasonality in the effective transmission rate of the virus. This rate could be affected by mixing patterns such as travel, school terms, more time spent indoors or due to other factors such as humidity/temperature conditions (increases viral production) (Dushoff et al., 2004).

In temperate regions, there are clear seasonal variations in the occurrence of influenza, the disease generally exists at low levels throughout the year with a marked seasonal peak in cold winter months. In contrast, seasonality is less defined in tropical regions, where there is high background influenza activity throughout the year, on top of epidemics occurring at intermediate months between the influenza seasons in temperate countries (Northern and Southern hemispheres) (Viboud et al., 2006). Generally local epidemics begin suddenly, peak in 2 to 3 weeks, and can last for up to 10 weeks (Glezen and Couch, 1978).

3.2.3.2 Variability of influenza

Influenza variability and diversity occurs by the actions of the evolutionary mechanisms, mutation and reassortment (section 1.1.1). Mutation, the ubiquitous and main source of RNA virus diversity, occurs due to the infidelity of RNA polymerases during RNA synthesis, however, viruses such as influenza with segmented genomes can also receive variability through reassortment when more than one virus has infected the cell. During co-infection, the segments of all virus particles are copied in the nucleus (section 3.2.2.2) and then assembled at the plasma membrane, where the 8 RNA segments may originate from different infecting viruses thereby introducing variability.

The combination of mutation and reassortment has lead to circulation of a large number of subtypes (defined according to membrane proteins HA and NA) and strains (variation in subtypes). There are 15 known subtypes of *HA* and 9 subtypes of *NA* in the wild, but only *HA* subtypes 1,2 and 3 and *NA* subtypes 1 and 2 are responsible for stable human and swine infections (Lewis, 2006). Pathogenic and non-pathogenic avian influenza viruses are ubiquitous in aquatic birds (particularly in wild migratory avian populations) where all subtypes (H1 to H14 and N1 to N9) are circulated and maintained.

The diversity in influenza is modulated by three evolutionary forces (section 1.1.1), natural selection, random genetic drift, and migration. Natural selection primarily acts on the major antigenic proteins *HA* and *NA*, which can incur non-conservative amino acid substitutions and still maintain function (Bush, 1999); a property that is not shared by other influenza proteins, such as those involved in replication and packaging. Continual change in antigenic structure is brought about primarily by positive selection and to a lesser extent negative selection (Suzuki, 2006). From a phylogenetic viewpoint there is clear evidence for the effect of positive selection on membrane protein HA, where the continual selective turn over produces a distinctive phylogenetic tree depicting the pathway of positively selected mutations (competitive exclusion) (Nelson and Holmes, 2007). However, natural selection is

not the only driving force shaping influenza diversity, Nelson et al. (2006) found that adaptive evolution is infrequent within individual influenza seasons allowing stochastic processes such as random genetic drift to play an important role in influenza virus evolution. The action of selection and random genetic drift shaping the continuous process of genetic change (introduced earlier as *antigenic drift*), and allows the virus to evade host immunity, infect more hosts and proliferate, thereby reducing the susceptibility of the virus to neutralising antibodies induced by previous influenza infections or immunisations consequently causing seasonal influenza epidemics. It should be noted however that while there is correspondence between antigenic and genetic evolution, antigenic evolution is more punctuated, and genetic changes can sometimes cause a disproportionately large antigenic effect (Smith et al., 2004). The current view (Grenfell model (Grenfell et al., 2004)) is that influenza epidemics arise through the incremental accumulation of viral mutations, culminating in a novel antigenic type that is able to escape host immunity. It is thought this process leads to punctuated antigenic evolution (Smith et al., 2004).

Migration of livestock, human travel and wild bird migrations facilitate influenza genome reassortment, which can occur from exchange between different human strains, or even between a human strains and avian and swine strains. This reassortment brought about through migration (*antigenic shift)* is responsible for major changes in the influenza virus, which can lead to the emergence of novel, potentially pandemic strains where the complete lack of immunity allows the virus to spread more rapidly than seasonal strains. While subtypes originate from avian populations and there is evidence of avian-human transmission (Subbarao and Katz, 2000), all influenza pandemics of the 20th and 21st century have been generated by a series of multiple reassortment events in swine or humans (Smith et al., 2009).

3.2.4 Detecting natural selection in influenza virus populations

Influenza viruses are amongst of the most rapidly adapting organisms known, and exhibit high levels of genetic diversity that are shaped by a complex interplay of natural selection, genome reassortment and transmission. Each gene within the influenza virus genome evolves under different selective pressures and evolutionary constraints. Most notably, strong positive selection acts on the HA (Hemagglutinin) and NA (Neuraminidase) genes, which code for the viral envelope proteins that contain many antigenic sites targeted by humoral immune responses. Of these, the HA protein is known to contain the highest concentration of antibody epitopes, and therefore experiences the most intense positive selection pressure (Fitch et al., 1991), resulting in *antigenic drift*. However, internal proteins such as NP and NS1, which are not subject to strong antibody-mediated selective pressure, are still are thought to undergo considerable selection from cytotoxic T-lymphocyte (CTL) specific responses (Townsend et al., 1984; Ludwig et al., 1991; Berkhoff et al., 2005; Fernandez-Sesma et al., 2006; Suzuki, 2006; Lin et al., 2007).

The detection of influenza adaptation is of particular interest from both a practical and an evolutionary perspective. The high rates of adaptation make influenza a good model organism for evolutionary study, and when coupled with the virus' impact on global public health, understanding the history of influenza's adaptive dynamics becomes of critical importance. However, despite the recent availability of abundant complete genome sequence data for human influenza, very little research has been conducted to determine how the rate of adaptive change varies across time, among different gene segments and among subtypes.

Most research into influenza adaptation has concentrated on the use of phylogenetic D_n/D_s methods to identify codons that are targeted by immune responses, with a primary focus on the selective pressure acting on HA (particularly the immunogenic HA1 domain) with much fewer studies conducted on the whole genome (see below). HA gene phylogenies display a characteristic structure comprising a single main 'trunk' lineage through time

(see figure 3.5 on page 87) that represents the pathway of fixed mutations, with terminal side branches representing isolates that do not leave descendants in subsequent epidemics, perhaps because they are not sufficiently antigenically favoured to evade host immunity (Fitch et al., 1991; Ina and Gojobori, 1994). The surviving trunk lineage experiences greater positive selective pressure at antigenic sites than at non-antigenic sites, with replacement substitutions concentrated at a few hyper-variable codons (Fitch et al., 1997; Bush, 1999). Bush et al. (1999) proposed that within each epidemic year, strains with the most amino acid replacements at hyper-variable codons would out-compete other lineages as they would likely be the most antigenically distinct from the current strain. That is, the lineage with the greatest additional amino acid replacements at known positively selected codons would likely be the progenitor of future epidemics, and therefore could represent a candidate strain for vaccine selection (Bush et al., 1999). However, this phylogenetic approach used by Bush et al. (1999), is potentially plagued by low statistical support and, as a result, alternative approaches for detecting positive selection have also been developed (Plotkin et al., 2002). Other studies have used phylogenetic D_n/D_s methods to investigate the tempo and mode of positive selection in HA. For example Wolf et al. (2006) asked whether interpandemic evolution is characterised by a period of neutral evolution punctuated by short bursts of rapid fitness increase and Shih et al. (2007) investigated whether antigenic change is an ongoing process or is sporadic, and whether multiple mutations at antigenic sites cumulatively enhance antigenic drift.

A few studies have attempted whole genome analysis. For example Suzuki (2006) applied D_n/D_s methods to 100 complete H3N2 genomes, and found that $D_n/D_s < 1$ for the majority of codons in all proteins, and that $D_n/D_s > 1$ for the PB2, HA and NS1 genes. Suzuki (2006) also concluded that positive selection operates on both B-cell epitopes and T-cell epitopes and suggested that negatively selected T-cell epitopes and sites under strong functional constraints are useful candidate targets for vaccines and anti-viral drugs. Rambaut et al. (2008) applied a phylogenetic method to both H1N1 and H3N2 genomes sampled over a 12 year period. They compared the ratio of substitution rates at 1^{st} and 2^{nd} codon positions to the rate at 3^{rd} codon

positions (which will be strongly correlated D_n/D_s), and found that this ratio was correlated with the overall evolutionary rate for both subtypes.

Despite the wealth of research on detecting adaptation in influenza gene sequences, almost all work has been carried out using D_n/D_s methods. These methods have been proven to be effective in detecting selection from genetic sequences from divergent species (Yang and Bielawski, 2000). However, for closely related within-population data sampled serially through time (like that available for influenza) the relationship between selection and D_n/D_s is not necessarily a simple monotonic function, making it difficult to accurately interpret selective forces from D_n/D_s values (Rocha et al., 2006; Kryazhimskiy and Plotkin, 2008). This point is shown by Nielsen and Yang (2003) who demonstrate that any inferences made using D_n/D_s methods will be strongly dependent on the specific details of the model of selection that is assumed, making it difficult to estimate the distribution of selection coefficients from D_n/D_s values. In addition, D_n/D_s methods are inherently conservative in that they require recurrent selection (repeated changes at the same codon) to demonstrate positive selection, but any mutations that occur only once on a single lineage will not be detected (a single selective sweep). Therefore it would seem prudent to investigate the use of alternative methods for detecting selection, such as those developed by Smith and Eyre-Walker (2002) and Williamson (2003) which do not seek to identify particular sites under selection, but rather quantify an overall rate of amino acid fixation. The evident paucity of whole genome studies of influenza virus may be partly due to the inability of D_n/D_s methods to cope with huge data sets - in this respect site frequency methods are more suitable as they can tractably handle data sets of many thousand genomes.

Here, I develop a generalisation of Williamson's (2003) method for evaluating rates of adaptive evolution and I apply this to data sets of several thousand complete human influenza A genomes from subtypes H1N1 and H3N2. Specifically, I aim to answer the following questions:

(i) To what degree does the adaptive rate vary among genes? As discussed above, research has shown that the HA gene experiences considerable positive selection from antibody mediated selective pressures, but relatively little research has been carried out on the selective pressures on NA. I aim to compare the rates of adaptation for HA and NA. In addition, I investigate whether genes coding for internal proteins experience adaptive fixations. Finally, because of the use of complete genome data, rates among genes can be directly compared to assess their relative contributions to the overall adaptation rate. Selective fixations may occur in all genes due to non-immune selection, and immune-mediated selection in the envelope proteins HA/NA may result in compensatory changes elsewhere in the genome.

(ii) Do adaptive rates differ between subtypes? Because data for both subtypes over a long time scale (31 years) are available, it is possible to compare rates of adaptation for influenza genes between human subtypes H1N1 and H3N2. Rambaut et al. (2008) showed that both these two subtypes exhibit very different evolutionary dynamics, and speculated that antigenic evolution proceeds at a reduced pace in H1N1 than in H3N2. I seek to test this hypothesis in this chapter.

(iii) For both subtypes and for all genes, I test whether the rate of adaptive fixation is constant or changes through time.

3.3 Methods

3.3.1 Data collection

I obtained all available whole human influenza genomes sampled between 1977-2008 from subtypes H1N1 and H3N2. The transmission of H1N1 in humans stopped in 1957, after which the dominant subtype was H2N2, but the 1957 human H1N1 strain was reintroduced into circulation in 1977 (caused by strain *Influenza A/USSR/90/77 (H1N1)*). H3N2, as described in section 3.2.1, first appeared in 1968. Therefore H1N1 and H3N2 have co-circulated in humans

continuously since 1977 which thus defines the period of observation for this study.

3.3.1.1 Whole genome data

Whole genome sequences were obtained from the National Institute of Allergy and Infectious Diseases (NSAID) Influenza Genome Sequencing Project (website: http:// www.ncbi. nlm. nih. gov/ genomes / FLU / FLU.html). Excluding the recent 2009 swine flu pandemic sequences, I selected all available complete human influenza A genomes belonging to subtypes H1N1 and H3N2 from all geographical regions. In addition, for each genome, I ensured that all 8 segments came from the same host species, and that no laboratory generated strains were included. In total 759, complete H1N1 genomes and 1603 complete H3N2 genomes were selected.

The 8 segments per genome code for 10 different proteins (PB2, PB1, PA, HA, NP, NA, M1, NS1, NS2). Therefore correspondingly, for both subtypes, I split the whole genome data set into 10 smaller data sets (one for each protein) and profile aligned these using Muscle (Edgar, 2004). These sequences were then trimmed to remove non-coding regions. Finally neighbour joining phylogenetic trees were constructed using QuickTree (Howe et al., 2002) for each data set, and were manually inspected to identify and remove any genomes that contained non human sequences or were incorrectly labeled as H1N1 or H3N2.

The methods developed later in this chapter are based on determining whether sites are silent or replacement, and as a result genome regions that contain overlapping reading frames are not suitable for analysis (since a silent change in one frame could be replacement in the other). Segments 1-6 each encode one protein (PB2, PB1, PA, HA, NP, NA) on a single non-overlapping reading frame. However, segments 7 (MP) and 8 (NS) code for two proteins: one colinear (M1,NS1) and one spliced (M2,NS2) (figure 3.2 on the following page), and both have overlapping regions on different reading frames. For M1 and M2 there is a small overlap (44nt) (figure 3.2 black area) and therefore overlapping sections were trimmed from both

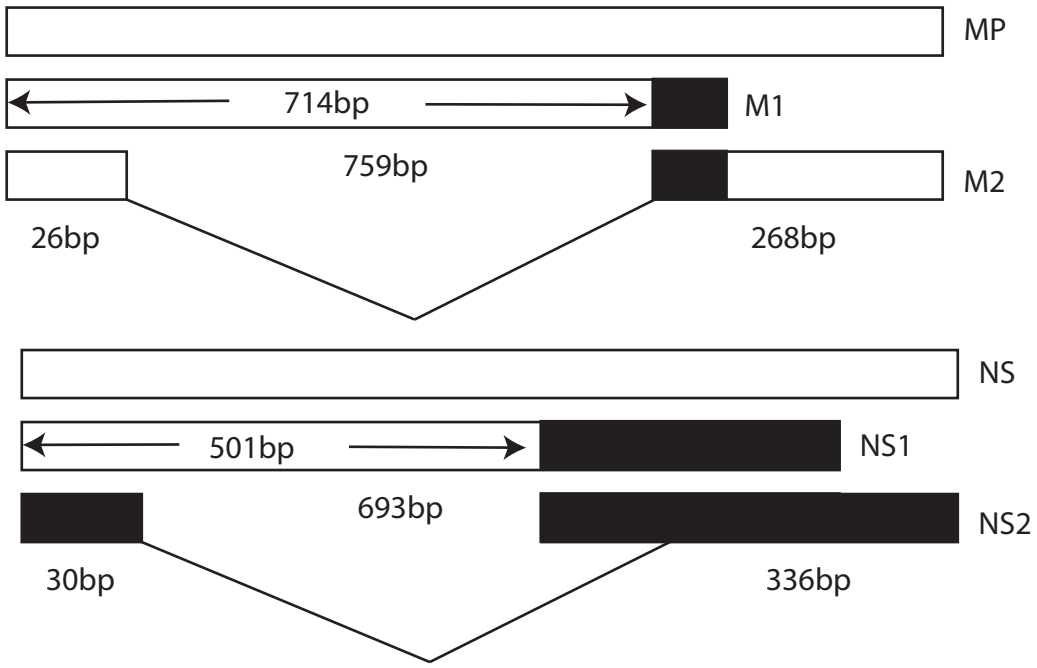


Figure 3.2: *The proteins encoded by influenza segment 7 (MP) and segment 8 (NS). M1 and NS1 are colinear and M2 and NS2 are derived by splicing. The black area shows sequence regions omitted from my analysis.*

proteins. For NS1 and NS2 there is a large overlap (191nt) which occurs over most of NS2, I therefore trimmed the overlap region from NS1 but chose not to use NS2 for analysis, as the resulting sequence after trimming would be too short for meaningful results.

The data sets from the NSAID database span different periods of time: the H1N1 data set ranges from 1918 to 2008, whereas as the H3N2 data set ranges from 1968 to 2008. For the analysis, for both subtypes, I only use data from years 1977-2008 (see section 3.3.1.1). Years with small sample sizes are merged with adjacent years to reduce statistical variance. If sequences were sampled from a single year then it was assumed that the sequences all date from the middle of the year (e.g. 2005.5), however, if sequences from adjacent years were merged then the dates were defined as the average date of all the merged sequences. The final data sets comprise 723 H1N1 genomes and 1520 H3N2 genomes (table 3.1)

Table 3.1: *Adjusted dates and numbers of whole genomes per year for H1N1 and H3N2*

H1N1		H3N2	
1977.50	**5**	**1977.50**	**7**
1978.50	**12**	**1978.67**	**6**
1980.25	**4**	**1980.63**	**8**
1982.25	**4**		
1983.52	**48**	**1983.21**	**7**
		1985.20	**10**
		1986.64	**7**
1987.93	**7**	**1988.50**	**6**
1990.50	**3**	**1990.30**	**5**
		1991.90	**5**
		1993.50	**56**
		1994.50	**28**
1995.50	**25**	**1995.50**	**28**
1996.54	**28**	**1996.50**	**46**
		1997.50	**27**
		1998.50	**49**
1999.50	**6**	**1999.50**	**118**
2000.50	**78**	**2000.50**	**146**
2001.50	**113**	**2001.50**	**49**
2002.50	**6**	**2002.50**	**161**
2003.50	**20**	**2003.50**	**196**
		2004.50	**206**
2005.45	**19**	**2005.50**	**153**
2006.50	**12**	**2006.50**	**9**
2007.50	**306**	**2007.50**	**120**
2008.50	**27**	**2008.50**	**67**
Total	**723**	**Total**	**1520**

3.3.1.2 Heamaglutinin gene data sets

In addition to the whole genome study I also performed an additional analysis of HA genes belonging to subtypes H1N1 and H3N2, for which many more sequences are available. Sequences were obtained from the NSAID database as above. As before, for both subtypes, profile alignment was performed using Muscle, and then alignments were trimmed such that the sequences contained the coding region only. Neighbour joining phylogenetic trees were constructed to identify and remove any HA sequences that were incorrectly labelled. As above, adjacent years with small sample sizes were merged with adjacent years. In total,

the final HA alignments contained 1013 H1N1 sequences and 2169 H3N2 sequences sampled from 1977 to 2008 (table 3.2)

Table 3.2: *Dates and Numbers of Samples per Date for H1N1 and H3N2, HA gene only*

H1N1		H3N2	
1978.18	**19**	**1978.10**	**10**
1979.93	**7**	**1980.38**	**8**
1982.25	**4**	**1982.30**	**5**
1983.55	**43**	**1983.50**	**5**
		1985.23	**11**
1986.67	**9**	**1986.63**	**8**
		1988.50	**11**
1990.13	**8**	**1990.10**	**10**
		1992.00	**6**
		1993.50	**47**
		1994.50	**55**
1995.50	**27**	**1995.50**	**40**
1996.50	**28**	**1996.50**	**44**
		1997.50	**54**
1998.33	**6**	**1998.50**	**79**
1999.50	**8**	**1999.50**	**148**
2000.50	**72**	**2000.50**	**163**
2001.50	**106**	**2001.50**	**70**
2002.50	**7**	**2002.50**	**195**
2003.50	**24**	**2003.50**	**252**
		2004.50	**226**
2005.38	**34**	**2005.50**	**176**
2006.50	**52**	**2006.50**	**79**
2007.50	**409**	**2007.50**	**212**
2008.50	**150**	**2008.50**	**155**
Total	**1013**	**Total**	**2069**

3.3.2 Estimating rates of natural selection and adaptation

Smith and Eyre-Walker (2002) developed a simple statistic to estimate the number of adaptive substitutions between two sister species. Their approach was based on earlier work by McDonald and Kreitman (1991a) and was based on the following assumptions: (i) all silent mutations are neutral, (ii) all replacement mutations are strongly deleterious, neutral, or strongly advantageous, and (iii) all polymorphism observed in the main alignment is neutral.

Suppose the number of silent and replacement polymorphisms observed in the main alignment are $s_{<1}$ and $r_{<1}$ respectively, and the number of silent and replacement fixations are s_1 and r_1 respectively (subscripts refer to the frequency of the derived nucleotide in the main alignment). Under the above assumptions, if α is the number of replacement fixations that are positively selected, then $\frac{(r_1-\alpha)}{r_{<1}} = \frac{s_1}{s_{<1}}$ on average. Thus the number of adaptive fixations is estimated as

$$\alpha = r_1 - s_1 \left(\frac{r_{<1}}{s_{<1}} \right) \tag{3.1}$$

I develop a generalised version of the Smith and Eyre-Walker (2002) estimator similar to the approach taken by Williamson (2003). Williamson (2003) employed the same assumptions and fundamental principles as Smith and Eyre-Walker (2002), except relaxed he relaxed the assumption that all polymorphism in the main alignment must be neutral. Instead, polymorphic sites are divided into two observed frequency classes: 'rare' polymorphisms (observed frequency$<$ 0.5) that are assumed to be neutral, and 'common' polymorphisms (observed frequency$>$ 0.5) that are treated in the same manner as observed fixations, and can therefore be neutral or advantageous. The justification for this change is that positive selection may contribute to high frequency polymorphisms as well as fixations. For example, viral populations subjected to fluctuating selection may maintain selected polymorphisms at high frequencies thus preventing their fixation.

I however, would argue that Williamsons approach is insufficient: in viruses where high mutation rates generate considerable mutation pressure, one would expect to observe large numbers of low frequency variants, many of which will be deleterious or slightly deleterious, thus the Williamson (2003) assumption that all 'rare' (observed frequency <0.5) polymorphic sites are neutral is unlikely to be correct in such cases. Furthermore the presence of slightly deleterious mutations will cause an underestimation in the levels of adaptive evolution (Fay et al., 2001; Charlesworth and Eyre-Walker, 2008). Therefore it would seem that the site frequency range where neutral polymorphisms are most likely to be found is at intermediate

frequencies. However, rather than assuming this *a priori*, in the following section I use a Wright-Fisher diffusion model to test this idea using empirical data on the selection coefficients of mutations in viral populations. I then use these results to formulate a generalised approach for determining the number and overall rate of adaptations. The steps involved are as follows

(i) I outline the derivation of the Wright-Fisher stationary distribution which gives the probability that a derived mutation will be observed at a specific site frequency.

(ii) I use empirical data to construct a probability mass function describing the distribution of selection coefficients for viral populations.

(iii) Using the distributions in (i) and (ii), I determine the relative frequencies of advantageous, deleterious and neutral mutations at different site frequencies.

(iv) Using the results from (iii) I introduce a method for estimating the rate of adaptive fixation through time.

3.3.2.1 The stationary distribution of mutational site frequencies

A model of the joint action of random genetic drift and selection is known as the Wright-Fisher model (e.g. Wright, 1945). The Wright-Fisher model describes the change in frequency of a single mutation in a population over time. The simplest version of the model makes the following assumptions: (1) nonoverlapping generations, (2) constant population size in each generation, and (3) random mating. This model was introduced in section 1.1.2, but here I show its complete derivation. My derivation follows the standard approach for determining the stationary distribution from a diffusion equation (e.g. Gardiner, 1985).

Consider a population with an effective population size of N_e haploid individuals that has a single polymorphic site with two alleles, one ancestral (wild type, labelled a) and one derived (mutant, labelled A). Under this model, the frequency of the derived allele in the current generation is a function of the selection pressure on this allele and the binomial sampling

effect, with transition probabilities proportional to the frequency of this allele in the previous generation. This process is defined by a Markov chain. If there are j copies of the derived mutation in generation $X+1$ given i copies in generation X, then the Markov chain transition probabilities are given by:

$$P_{ij} = \mathbb{P}(X(t+1) = j | X(t) = i) \tag{3.2}$$

$$P_{ij} = \binom{N_e}{j} (\phi_i)^j (1-\phi_i)^{2N_e - j} \tag{3.3}$$

where ϕ is a function of the relative fitness of the derived mutation. Under the assumption that recurrent mutation doesn't occur:

$$\phi_i = \frac{x(1+s)}{x(1+s)+(1-x)} \tag{3.4}$$

where $1+s$ is the fitness of the derived mutation relative to the ancestral mutation, and x is the derived mutation frequency in generation X (which is i/N_e for a new mutation). It is easy to see that, given no selection, $\phi = x = i/N_e$. For the Markov process defined in equation 3.2, the mean and variance are:

$$\begin{aligned} E[X(t+1) - X(t) | X(t) &= i] = N_e\phi - i \\ E[(X(t+1) - E[X(t+1)])^2 | X(t) = i] &= N_e \frac{i}{N_e}\left(1 - \frac{i}{Ne}\right) \end{aligned} \tag{3.5}$$

To understand how these moments work in a biological context, consider the case of no selection, in which case the mean becomes zero ($N_e(i/N_e) - i = 0$), which shows that in the absence of selection there is no directional tendency in site frequency. Because the variance does not include a selection or mutation term, it can be thought of as the stochastic component representing random genetic drift, which is only dependant on effective population size.

Next, I introduce recurrent mutation. I define v as the mutation rate for $\{a \rightarrow A\}$ and u

as the mutation rate for $\{A \rightarrow a\}$. Under this scenario a diffusion process for the relative frequency of an allele at time t: $\phi(x,t)$ can be obtained. This diffusion process is described by the Kolmogorov forward equation.

$$\partial\phi\left(x,t|x_0,t_0\right) = \frac{-\partial\left(\mu_x \cdot \phi\left(x,t|x_0,t_0\right)\right)}{\partial x} + \frac{1}{2}\frac{\partial^2\left(\sigma_x^2 \cdot \phi\left(x,t|x_0,t_0\right)\right)}{\partial x^2} \tag{3.6}$$

Equation 3.6 describes the allele frequencies at each point in time ($t > 0$) given an initial frequency (x_0) at time t_0. The term μ_x represents the direct evolutionary change due to selection and σ_x is the random, undirected evolutionary change due to drift.

While for the Wright-Fisher model there is no general solution for equation 3.6, a stationary distribution $\pi(x)$ can be found. The stationary distribution does not change over time, ${}^{\partial\pi(x)}/_{\partial(t)} = 0$, and if the system is stationary at t_0, then the density at future times will also be stationary and satisfy $\pi(x) = \int(\pi(x_0)\phi(x,t|x_0,t_0)$ when integrated over all possible values of x_0. Given these conditions, it is possible to multiply equation 3.6 through by $\pi(x_0)$ and integrate over all possible states of x_0.

$$\begin{aligned}\int \pi(x_0)\partial\phi\left(x,t|x_0,t_0\right)dx_o &= \int \pi(x_0)\frac{-\partial\left(\mu_x \cdot \phi\left(x,t|x_0,t_0\right)\right)}{\partial x}dx_0 \\ &\quad + \frac{1}{2}\int \pi(x_0)\frac{\partial^2\left(\sigma_x^2 \cdot \phi\left(x,t|x_0,t_0\right)\right)}{\partial x^2}dx_0\end{aligned}$$

Because integration occurs over x_0 (which is not time dependant) it is possible to interchange the order of integration and differentiation as follows:

$$\begin{aligned}\partial\int \pi(x_0)\phi\left(x,t|x_0,t_0\right)dx_o &= \frac{-\partial\int \mu_x \cdot \pi(x_0)\phi\left(x,t|x_0,t_0\right)dx_0}{\partial x} \\ &\quad + \frac{1}{2}\frac{\partial^2\int \sigma_x^2 \cdot \pi(x_0)\phi\left(x,t|x_0,t_0\right)dx_0}{\partial x^2}\end{aligned}$$

From the conditions described above, specifically $\pi(x) = \int(\pi(x_0)\phi(x,t|x_0,t_0)$ and ${}^{\partial\pi(x)}/_{\partial(t)} =$

0, I obtain

$$0 = -\frac{d\left(\mu_x \pi(x)\right)}{dx} + \frac{1}{2}\frac{d^2\left(\sigma_x^2 \pi(x)\right)}{dx^2} \tag{3.7}$$

This is now an ordinary linear differential equation (ODE). Using the general solution for this type of ODE the stationary distribution can be found as:

$$\pi(x) = \frac{e^{\frac{2\mu_x}{\sigma_x^2}}\left(c_1 \int e^{-\frac{2\mu_x}{\sigma_x^2}} dx + c2\right)}{\sigma_x^2} \tag{3.8}$$

For the Wright-Fisher model described above, if I substitute equation 3.4 into equation 3.5, then genetic parameters μ_x and σ_x^2 are found as:

$$\begin{aligned} \mu_x &= -N_e v x + N_e u(1-x) + N_e s x(1-x) \\ \sigma_x^2 &= x(1-x) \end{aligned}$$

And if these terms are substituted into 3.7 and the differential equation is solved, then the stationary distribution becomes

$$\pi(x|u, v, N_e, s) = \frac{(1-x)^{2N_e v - 1}(x)^{2N_e u - 1} e^{2N_e s x}}{\int_0^1 (1-x)^{2N_e v - 1}(x)^{2N_e u - 1} e^{2N_e s x} dx} \tag{3.9}$$

This equation describes the probability of a derived mutation assuming site frequency between $x, x + \partial x$ in the presence of selection, drift and mutation. It should be noted that equation 3.9 is derived from a Markov process where there are no absorbing states (due to recurrent mutation); it therefore does not allow for the permanent fixation or loss of mutations, and as a result equation 3.9 is not always defined for site frequency values of 0 or 1. An example of this can be seen in the case where $(Ne\, v - 1) < 0$. Therefore in the analysis that follows, I choose to use 0.001 and 0.999 as our lowest and highest site frequency values. This is reasonable, given that frequencies lower or higher are unlikely to be seen in real data.

In this chapter I solve Equation 3.9 by numerical integration using the trapezium rule. To aid

computational, it is evaluated over evenly spaced intervals $\{0.001 < m < n < 0.999\}$

$$\pi(m < x < n \,|u, v, N_e, s) = \int_m^n \pi(x|u, v, N_e, s) = \frac{\int_m^n (1-x)^{2N_e v-1}(x)^{2N_e u-1} e^{2N_e sx} dx}{\int_{0.001}^{0.999} (1-x)^{2N_e v-1}(x)^{2N_e u-1} e^{2N_e sx} dx} \quad (3.10)$$

This equation gives the probability mass function that a derived mutation assumes a site frequency between m and n in the presence of selection, drift and mutation.

3.3.2.2 The distribution of selection coefficients

Equation 3.10 is conditional on 4 variables - the mutation rates (u, v), effective population size (N_e) and the selection coefficient (s). In this section I explore the behaviour of the stationary distribution under parameter values representative of RNA virus populations. Literature on the distribution of mutational fitness effects is very sparse, however, Sanjuan et al. (2004) performed site-directed mutagenesis to create 91 single mutant clones of vesicular stomatitis virus (VSV), each derived from a common ancestral cDNA, and then performed competition experiments to measure the relative fitness of each mutant. Here, I use the results of Sanjuan et al. (2004) to develop a discrete probability distribution describing the probability mass function of selection coefficients.

By fitting various statistical models to their results, Sanjuan et al. (2004) concluded that advantageous (a) and deleterious (d) mutations follow an exponential distribution with parameters $\lambda_a = 0.044$ and $\lambda_d = 0.14$ respectively. I also assume that neutral mutations follow a uniform distribution in the interval $-0.01 < s < 0.01$. Sanjuan et al. (2004) also noted that the probabilities of random *non lethal* mutations being advantageous, deleterious or neutral are $\theta_a = 0.069$, $\theta_d = 0.483$ or $\theta_n = 0.448$ respectively.

From these distributions I derive a discrete probability density for the distribution of selection coefficients as follows. I begin by assuming that the distribution of selection coefficients takes the range $s = \{-0.70, ..., 2.42\}$, where the upper and lower bounds were chosen as

the maximum values observed when sampling 10^6 random variables from the distributions $\lambda_a e^{-\lambda_a x}$ and $\lambda_d e^{-\lambda_d x}$, respectively. Imposing bounds in this manner means that a very small probability mass is ignored, but these extreme tail values are unlikely to affect the resulting distribution of selection coefficients. By combining the above, the probability mass function of observing a mutation with selection coefficient within an interval $s = \{x, y\}$ can be defined as:

$$\mathbb{P}(s)_{\{x,y\}} = \begin{cases} \mathbb{P}(s_d)_{\{x,y\}} = \frac{\left(-e^{-y\lambda_d} + e^{x\lambda_d}\right)\theta_d}{\int_{0.01}^{0.70} \lambda_d e^{-\lambda_d s} ds} & if \ -0.70 < \{x, y\} < -0.01 \\ \mathbb{P}(s_n)_{\{x,y\}} = \frac{0.02\,\theta_n}{x-y} & if \ -0.01 < \{x, y\} < 0.01 \\ \mathbb{P}(s_a)_{\{x,y\}} = \frac{\left(-e^{-y\lambda_a} + e^{x\lambda_a}\right)\theta_a}{\int_{0.01}^{2.42} \lambda_a e^{-\lambda_a s} ds} & if \ 0.01 < \{x, y\} < 2.42 \end{cases} \tag{3.11}$$

Because this is a probability distribution, the sum of probabilities $\mathbb{P}(s)_{\{-0.70,\,2.42\}} = 1$. Figure 3.3 on the next page graphically illustrates this discrete probability distribution.

3.3.2.3 Picking classes from the site frequency spectrum

Given the two distributions described by equations 3.9 and 3.11, I can define a probability distribution that describes the probability of observing a mutation at a frequency between $x, x + \partial x$. I assume equal forward and backward mutation rates ($u = v$) and assign the mutation rate as 2×10^{-5} mutations per site per replication cycle (Sanjuan et al., 2010). Therefore the probability that a mutation lies in interval $\{m, n\}$ is given by

$$\pi(m < x < n | v, N_e) = \int_{x=m}^{n} \int_{s=-0.70}^{2.42} \mathbb{P}(s)\pi(x|v, Ne, s)\, ds\, dx \tag{3.12}$$

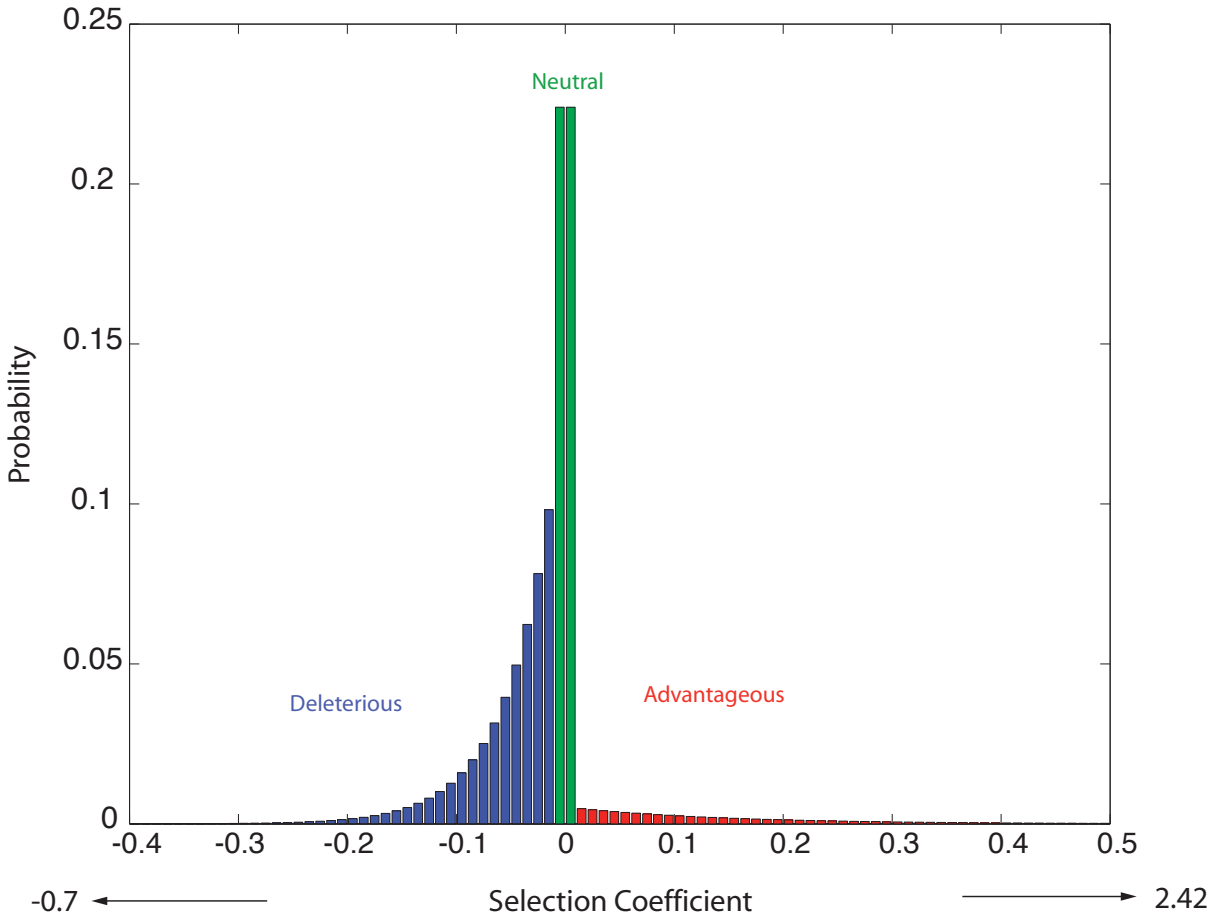


Figure 3.3: *Discrete probability distribution described by equation 3.11. Blue represents deleterious mutations, green represents neutral mutations and red represents advantageous mutations*

The probability that a mutation observed at a frequency between $\{m, n\}$ AND is deleterious, neutral or advantageous is given by π_d, π_n and π_n, as follows:

$$\begin{aligned}
\pi(m < x < n|v, N_e)_d &= \int_{x=m}^{n} \int_{s=-0.70}^{-0.01} \mathbb{P}(s_d)\pi(x|v, Ne, s_d)\, ds\, dx \\
\pi(m < x < n|v, N_e)_n &= \int_{x=m}^{n} \int_{s=-0.01}^{0.01} \mathbb{P}(s_n)\pi(x|v, Ne, s_n)\, ds\, dx \\
\pi(m < x < n|v, N_e)_a &= \int_{x=m}^{n} \int_{s=0.01}^{2.42} \mathbb{P}(s_a)\pi(x|v, Ne, s_a)\, ds\, dx
\end{aligned} \tag{3.13}$$

Numerical integration of equation 3.10 was performed in MATLAB using the Trapezium rule with trapezium widths of 0.01 for s and 0.001 for x. At each point in the discrete site frequency spectrum I divide the probability of a mutation being advantageous, deleterious or neutral by the total probability at that frequency ($\pi_d + \pi_n + \pi_a$), thereby evaluating the relative proportion of each type of mutation at each site frequency. I then plot these values for six values of N_e (figure 3.4 on the following page). The upper bound of $N_e < 1000$ is not chosen to represent the maximum effective population observed in influenza virus populations, but rather, due to computational precision limits, it is only possible to evaluate equation 3.10 for values of $N_e < 1000$. This problem can possibly be resolved by estimating equation 3.10 in logarithmic space to keep within precision limits.

It is clear from figure 3.4 that Williamson's (2003) assumption that all mutations at frequencies below 0.5 are neutral is unjustified. Neutral mutations are observed appreciably throughout the site frequency spectrum, but as the effective population size becomes larger and the effect of genetic drift becomes smaller, neutral mutations are most likely to be seen at intermediate frequency ranges ($0.25 - 0.75$). Deleterious mutations are also likely to be seen, but only at low site frequencies ($\lessapprox 0.25$) whereas advantageous mutations are always rare and are only seen at high site frequencies ($\gtrapprox 0.75$). The continuous year round transmission of influenza infections in the tropics (Viboud et al., 2006) and the frequent occurrence of selective sweeps

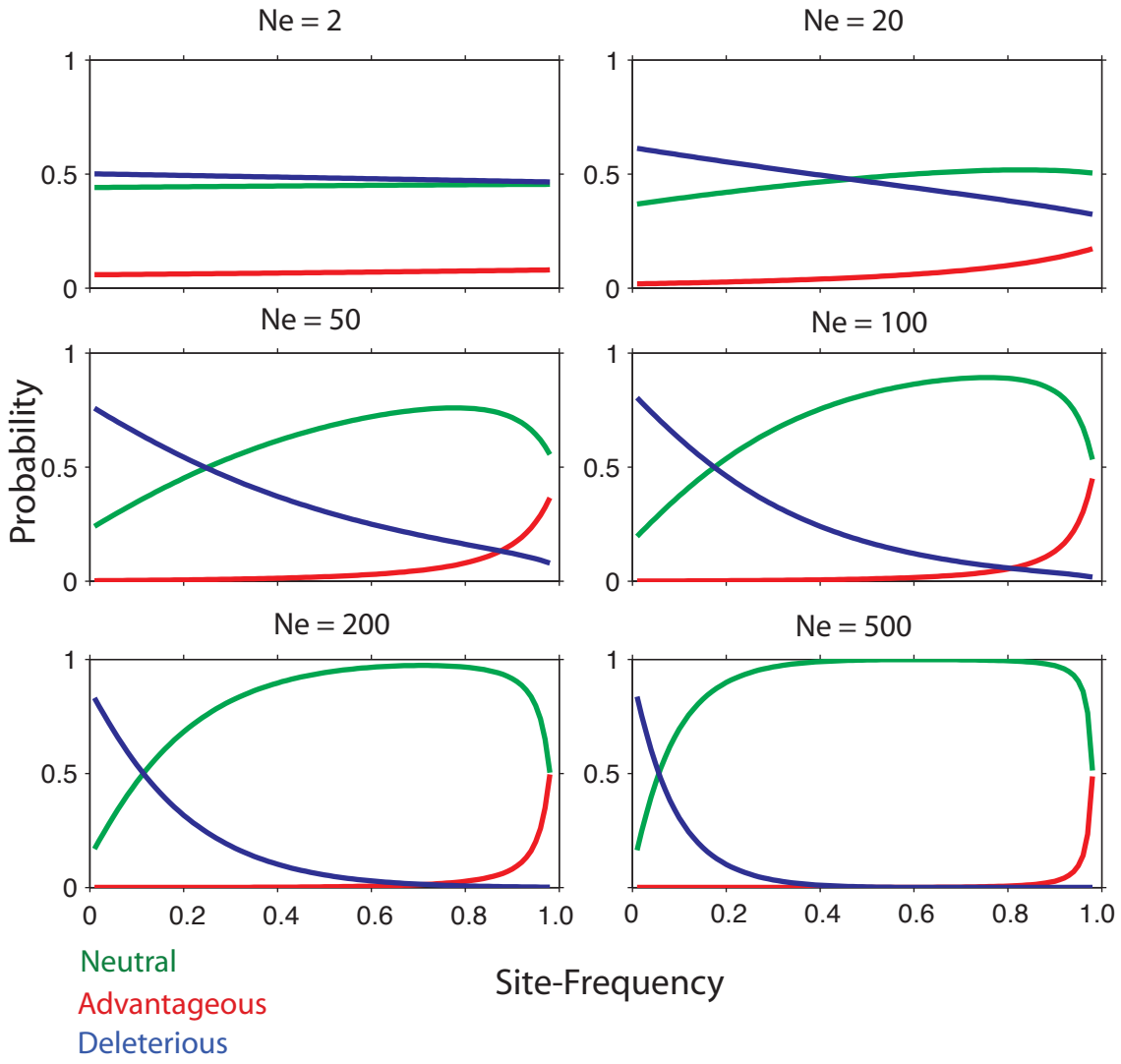


Figure 3.4: *For each site frequency, the lines show the probability that mutations at that frequency are advantageous (red), deleterious (blue) or neutral (green) as determined using equation 3.13*

suggest that effective population sizes for influenza are large. Therefore it is most likely that mutations in influenza populations will follow the distributions in the bottom row of figure 3.4, for which N_e is large.

Because any method based on equation 3.1 requires the definition of a class of sites that is selectively neutral, the results in figure 3.4 can be used to pick a statistically suitable class. Given this data it would seem more suitable to define three polymorphic site frequency classes: the low range $(0 - 0.25)$, middle range $(0.25 - 0.75)$ and the high range $(0.75 - 1)$. Eyre-Walker (2006) and Williamson (2003) used classes based on *a priori* assumptions, whereas here I have used theory and empirical data to show a preponderance of neutral mutations in the middle range. Deleterious mutations are found in very low numbers at mid frequencies due to purifying selection, and advantageous mutations will sweep rapidly through this range fast. I therefore choose to use the mid-frequency range as the neutral range. It should be noted that this range may contain few polymorphisms compared to the low and high frequency ranges, but this should not be a problem if sufficient data is available.

3.3.2.4 Calculating the number of adaptive sites per time point:

To estimate the number of adaptive substitutions between two time points for influenza, I use equations similar in form to equation 3.1 except that the site frequency spectrum is split over 4 classes (i) low-frequency sites ($0 < l < 0.25$), (ii) middle-frequency sites ($0.25 < m < 0.75$), (iii) high-frequency sites ($0.75 < h < 1$) and (iv) fixed sites $f = 1$. Following the results above I assume that middle frequencies sites are neutral. The estimators for the number of non neutral sites in each frequency class are therefore

$$\begin{aligned} \alpha_l &= r_l - s_l \frac{r_m}{s_m} \\ \alpha_h &= r_h - s_h \frac{r_m}{s_m} \\ \alpha_f &= r_f - s_f \frac{r_m}{s_m} \end{aligned}$$

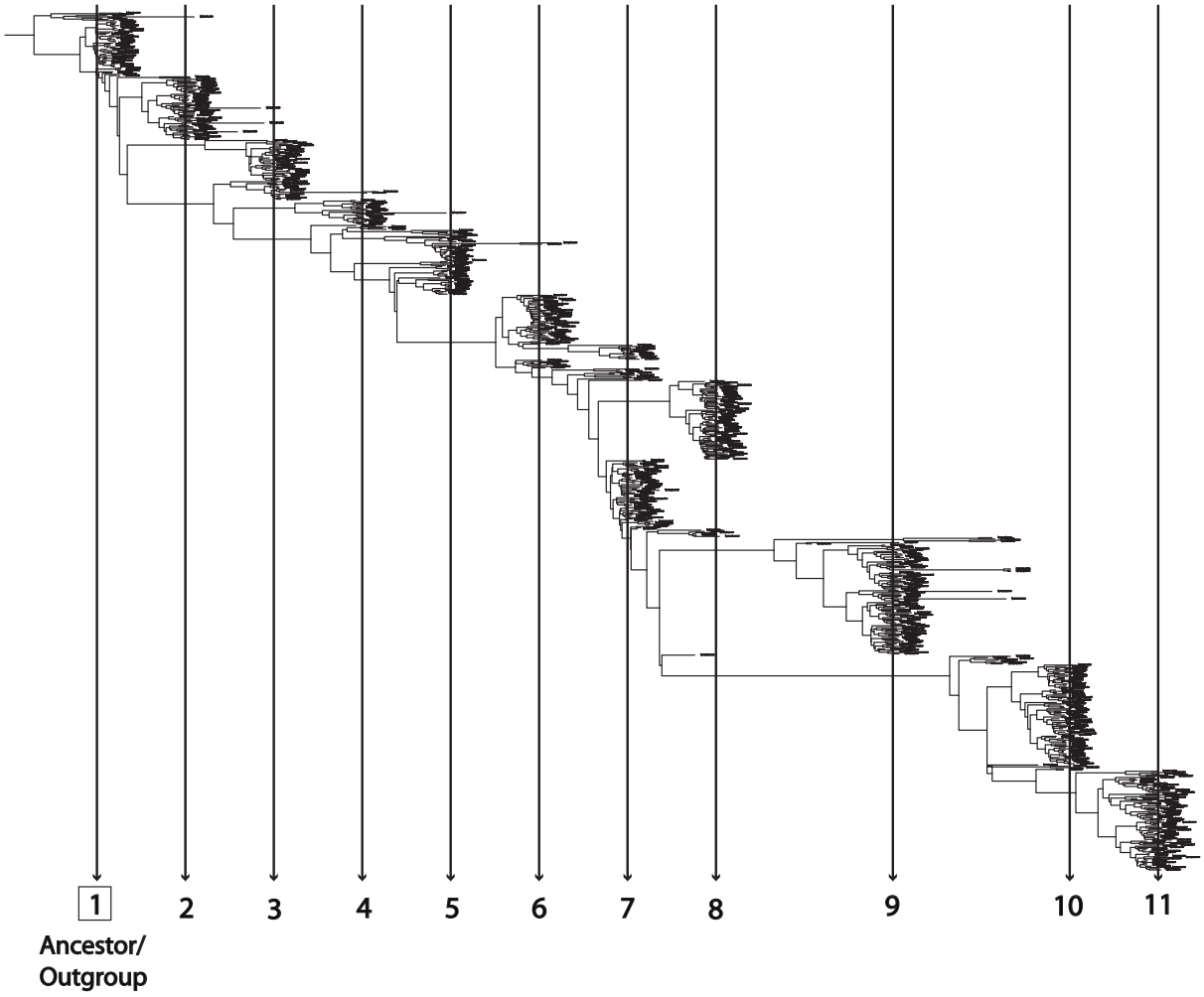


Figure 3.5: *Example HA phylogenetic tree from Rambaut et al. (2008). The consensus of the first set of sequences (timepoint 1) is taken as the ancestral sequence. The numbers of adaptive sites at each subsequent time point is then calculated using the method described in section 3.3.2.4.*

Where α is the number of non neutral sites in each class, s and r are the number of silent and replacement sites respectively, and the subscripts denote the site frequency range over which they are evaluated. These estimators are computed in the same way as those described in section 2.3.3. When calculating the total number of adaptive mutations I ignore the low frequency range as non neutral sites in this range are more likely the result of mutation pressure than positive selection (Charlesworth and Eyre-Walker, 2008). The equation for the total number of adaptive sites is therefore:

$$\alpha = \left(\alpha_h + \alpha_f\right) = \left(r_h + r_f\right) - \left(s_h + s_f\right) \frac{r_m}{s_m} \tag{3.14}$$

When using equation 3.14, the consensus sequence of the first time point (figure 3.5 on the previous page) is taken to be the ancestral sequence. Then equation 3.14 is used to calculate the number of adaptive substitutions at each time point in tables 3.2 and 3.1.

3.3.2.5 Estimating the rate of adaptation

Evaluating equation 3.14 for all time points gives a vector (Ω) of the number of adaptive changes. If time points are labelled $t = \{2, ...T\}$ then $\Omega \in \{\alpha_2, ..., \alpha_T\}$. Each point in Ω thus represents the number of adaptive changes incurred since the ancestral sequence. To summarise Ω and to estimate an overall rate of adaptation over the time series, I used an linear ordinary least squares (OLS) regression. The regression model takes the form

$$\alpha_i = \beta_1 t_i + \beta_2 + \epsilon_i, \quad i = 2, ...T \tag{3.15}$$

The gradient of this linear regression (β_1) thus represents the rate of viral adaptation.

3.3.2.6 Distribution percentiles

The analysis described above may suffer from serial autocorrelation, which is common for such time series analysis. This autocorrrelation should not bias the OLS regression coefficients, but affect standard errors, thereby making parametric tests of assessing confidence, such as the T-test, unsuitable. I therefore choose to use a bootstrap approach, which assesses the estimation variance non-parametrically (Hardle et al., 2003). Distribution percentiles are found for both Ω and β_1 using the following approach:

1. Codons are sampled with replacement from the consensus ancestral alignment at $t = 1$ to create new ancestral sequence alignment, of the same length and sample size.
2. For all time points $t = \{2, ...T\}$, a new sequence alignment of the same size are created by sampling codons from the original alignments according to the same codon order as defined in (1)
3. From the new bootstrapped time series created in (1) and (2) I calculate and record a bootstrapped vector $\Omega^* \in \{\alpha_2^*, ..., \alpha_T^*\}$,
4. I Perform a linear regression on Ω^* and record the gradient, β_1^*
5. I Repeat the steps (1) to (4) 1000 times to build bootstrap distributions of Ω^* and β_1^*.
6. Obtain 95% percentiles for the two distributions created in step (5)

3.3.2.7 The neutral ratio

In equation 3.14, the ratio r_m/s_m is the ratio of replacement to silent sites at mid-frequencies, which are assumed to be neutral, I call r_m/s_m the 'neutral ratio'. Although r_m/s_m can be calculated separately for each time-point and gene, this will result in unacceptably high variances for some calculations due to the small number of available sites. Previous studies have reduced variance by combining site counts across genes (e.g. Smith and Eyre-Walker, 2002), but this

Table 3.3: *Ratio of r_m/s_m from summed values of r_m and s_m across all years with samples sizes greater than 10.*

		PB2	PB1	PA	HA	NP	NA	M1	M2	NS1
H1N1	r_m	12.00	9.50	13.50	38.75	6.50	28.50	0.00	7.00	13.50
	s_m	89.50	87.00	91.00	104.25	48.00	74.50	14.00	4.50	16.50
	$\frac{r_m}{s_m}$	0.134	0.109	0.148	0.372	0.135	0.383	0.000	1.556	0.818
H3N2	r_m	17.00	13.00	38.00	70.50	31.50	69.00	13.00	1.00	10.00
	s_m	141.28	239.67	147.94	91.50	103.44	104.00	38.50	5.00	17.00
	$\frac{r_m}{s_m}$	0.120	0.054	0.257	0.770	0.305	0.663	0.338	0.200	0.588

may introduce bias if r_m/s_m varies significantly among genes (see Welch, 2006). However, the temporal structure of the data provides a solution: for any given gene, r_m/s_m is not expected to vary through time provided that long-term effective population sizes remain sufficiently large (a condition almost certainly met in this case). The data supports this conclusion: for both subtypes, r_m/s_m varies significantly among genes but not among time-points (2-way ANOVA; $p<0.01$). Therefore, for each subtype and gene, we calculate M by combining site counts among time-points.

To determine the neutral ratio for a given gene, I sum the values r_m and s_m across all years that have samples sizes greater than 10 (Table 3.1 and 3.2) i.e. if there are n years with sample sizes greater than 10 the neutral ratio is:

$$\frac{\bar{r_m}}{\bar{s_m}} = \frac{\sum_{i=1}^{n} r_{m,i}}{\sum_{i=1}^{n} s_{m,i}} \tag{3.16}$$

My justification for taking the ratio of the summed counts is that (i) estimates of r_m/s_m evaluated for each individual year may incur considerable sample variance, especially for years with small samples sizes, and (ii) population genetic theory suggests that the neutral ratio for a protein coding gene should be the same through time. Table 3.3 shows the neutral ratios calculated for each genes.

By examining table 3.3 it is clear that the ratio $\bar{r_m}/\bar{s_m}$ changes among genes. For the HA and NA genes exceptionally frequent selective sweeps or fluctuating selection means that

many mid-frequency polymorphisms may be selected, not neutral. In addition the genes M1, M2 and NS1 are short (< 750bp) and not very diverse, hence the values of r_m and s_m are possibly biased The polymerase genes PB1,PB2 and PA are perhaps the most reliable gene for estimating the neutral ratio. They are suitably long (> 2000bp) and are strongly conserved and therefore mid-frequency polymorphisms in these genes are unlikely to represent repeated selective sweeps or fluctuating selection. I therefore calculate the neutral ratio as the average of $\frac{\bar{r_m}}{\bar{s_m}}$ for the genes PB1, PB2 and PA. The estimates for the neutral ratio are 0.131 for H1N1 and 0.144 for H3N2.

It should be noted that due to functional constraints on each gene (Welch, 2006) this approach is likely to introduce bias into the estimates of $\frac{\bar{r_m}}{\bar{s_m}}$. For future work it is preferable to use the calculated value of $\frac{\bar{r_m}}{\bar{s_m}}$ for each gene. The problem of exceptionally frequent selective sweeps or fluctuating selection in HA and NA can be overcome by using non antigenic regions to calculate the neutral ratio $\frac{\bar{r_m}}{\bar{s_m}}$. For HA the subdomain HA2 can be used to calculate $\frac{\bar{r_m}}{\bar{s_m}}$, and for NA, because no such subdomain exists, antigenic and non antigenic regions can be isolated from structural analysis. For M1,M2 and NS1 the bias incurred due to the short length of these sequences can be avoided by further joining adjacent years to increase sample sizes.

3.4 Results

3.4.1 Whole genome results

Figure 3.6 on page 93 shows the results for subtype H1N1. Each point is the number of adaptive substitutions calculated using equation 3.14, and the blue line through these points represents the OLS best fit linear regression. The 95% percentile intervals for each time point are shown as error bars, and the bootstrap replicates for the linear regression are shown as the red lines. Each subplot in figure 3.6 is a different gene. The same plots are shown for subtype H3N2 in figure 3.7 on page 94. A comparison of the different rates among genes for

both subtypes is shown in figure 3.8 on page 95 where the error bars represent, the bootstrap distribution.

It is clear from figures 3.6, 3.7 and 3.8 that the rates of adaptation vary greatly between genes and subtypes. As expected the highest rates of adaptation are observed on membrane proteins HA and NA which exhibit rates of 1.12 and 1.16 adaptive fixations per year respectively in H1N1 and 1.68 and 1.32 adaptive fixations a year in H3N2. However, while the rates for HA and NA for H1N1 are almost the same (3% difference), for H3N2 the HA rate is considerably (21% difference) higher than that of NA. In addition to this the rates for HA and NA are both higher for H3N2 than for H1N1 (figure 3.8).

Aside from the membrane proteins, non zero rates of adaptation are observed in NP and NS1, with rates slightly more elevated for H1N1 than for H3N2 (figures 3.6 and 3.7). As expected, for both subtypes, the more conserved polymerase genes (PB1,PB2 and PA), show very low rates of adaptation. For PB2 and PB1, H1N1 exhibits higher rates than H3N2, however, for PA, H3N2 shows a greater rate of adaptation than H1N1 (figure 3.8). The non structural protein M1 experiences very little adaptation for both subtypes, however M2 shows a difference between subtypes, with a greater rate observed for H1N1 than for H3N2 (7% difference).

3.4.2 HA results

As highlighted in section 3.2.4, previous studies have found a significantly higher rate of amino acid replacements in antigenic positions of the HA1 subdomain compared to the HA2 domain. My analysis of the two subdomains of HA (figure 3.9 on page 96) also demonstrates this: the rates of adaptation for the HA1 subdomain are considerably higher than those for HA2 for both subtypes. As with the whole genome analysis above, a large rate difference in the HA1 subdomain rate is observed between subtypes H1N1and H3N2.

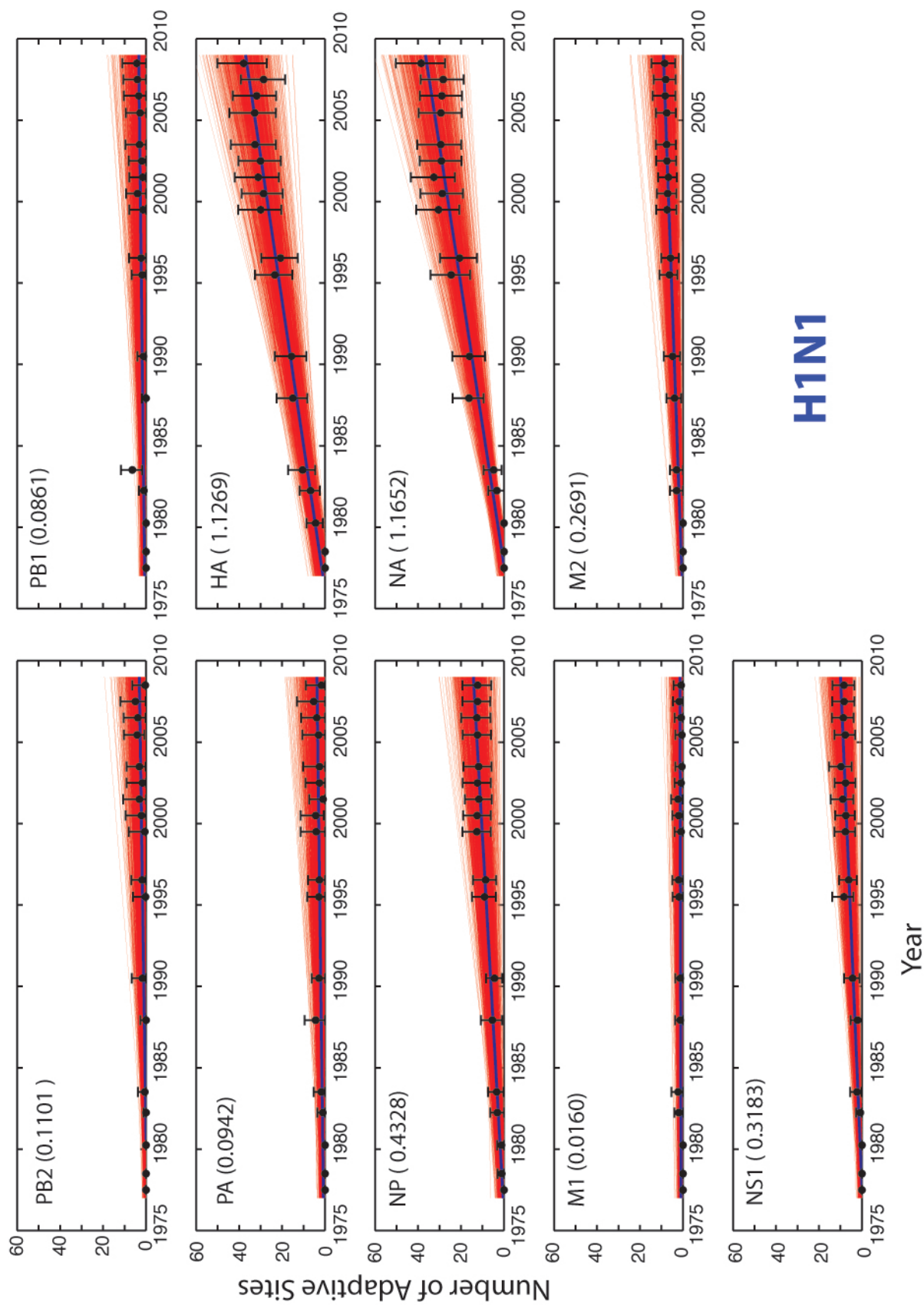


Figure 3.6: *Number of adaptive changes in H1N1 from 1977 to 2008. Black points are the number of adaptive changes calculated using equation 3.14. At each point the error bars are the 95% bootstrap percentiles. The blue line is the best fit OLS regression and the red lines are the bootstrap regressions. Numbers in brackets are the rates of adaptive fixation (β_1).*

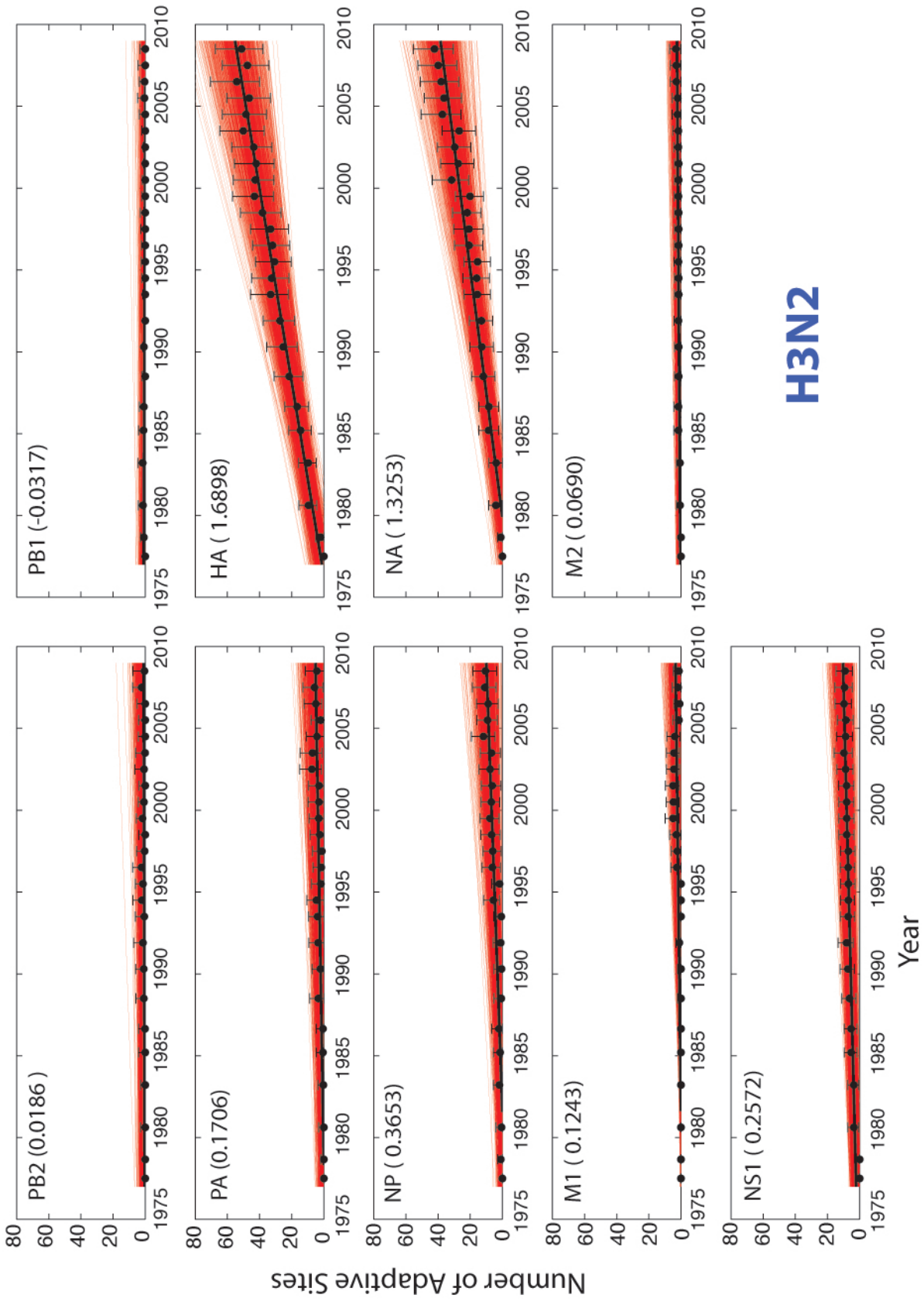


Figure 3.7: *Number of adaptive changes in H3N2 from 1977 to 2008. Black points are the number of adaptive changes calculated using equation 3.14. At each point the error bars are the 95% bootstrap confidence intervals. The blue line is the best fit OLS regression and the red lines are the bootstrap regressions. Numbers in brackets are the regression gradient coefficients.*

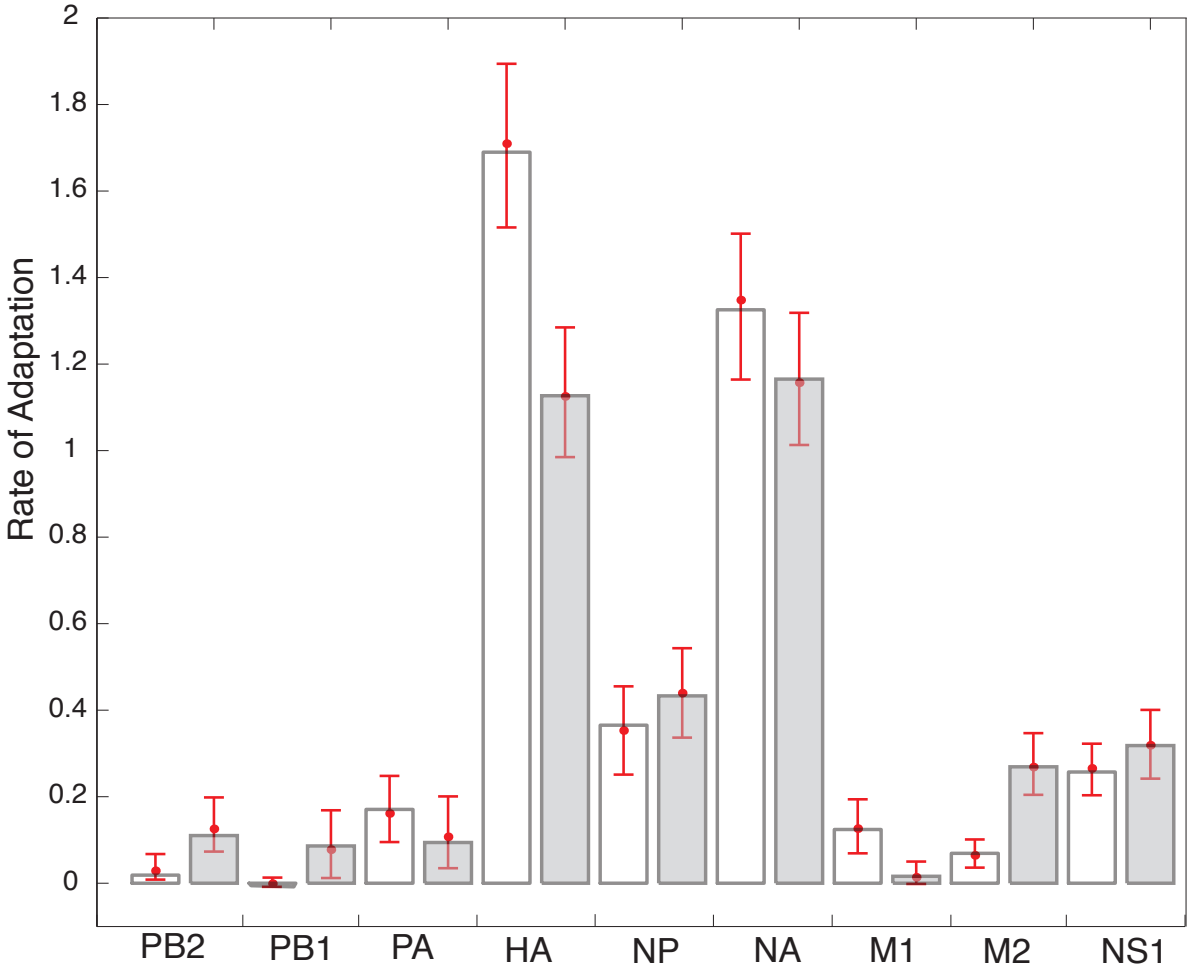


Figure 3.8: *Rates of adaptive fixations from a best fit OLS regression across all data points, and bootstrap percentiles. The error bars show the median (red point) and the bootstrap distribution (red error bars). H1N1 is coloured grey and H3N2 coloured white (Section 3.3.2.6)*

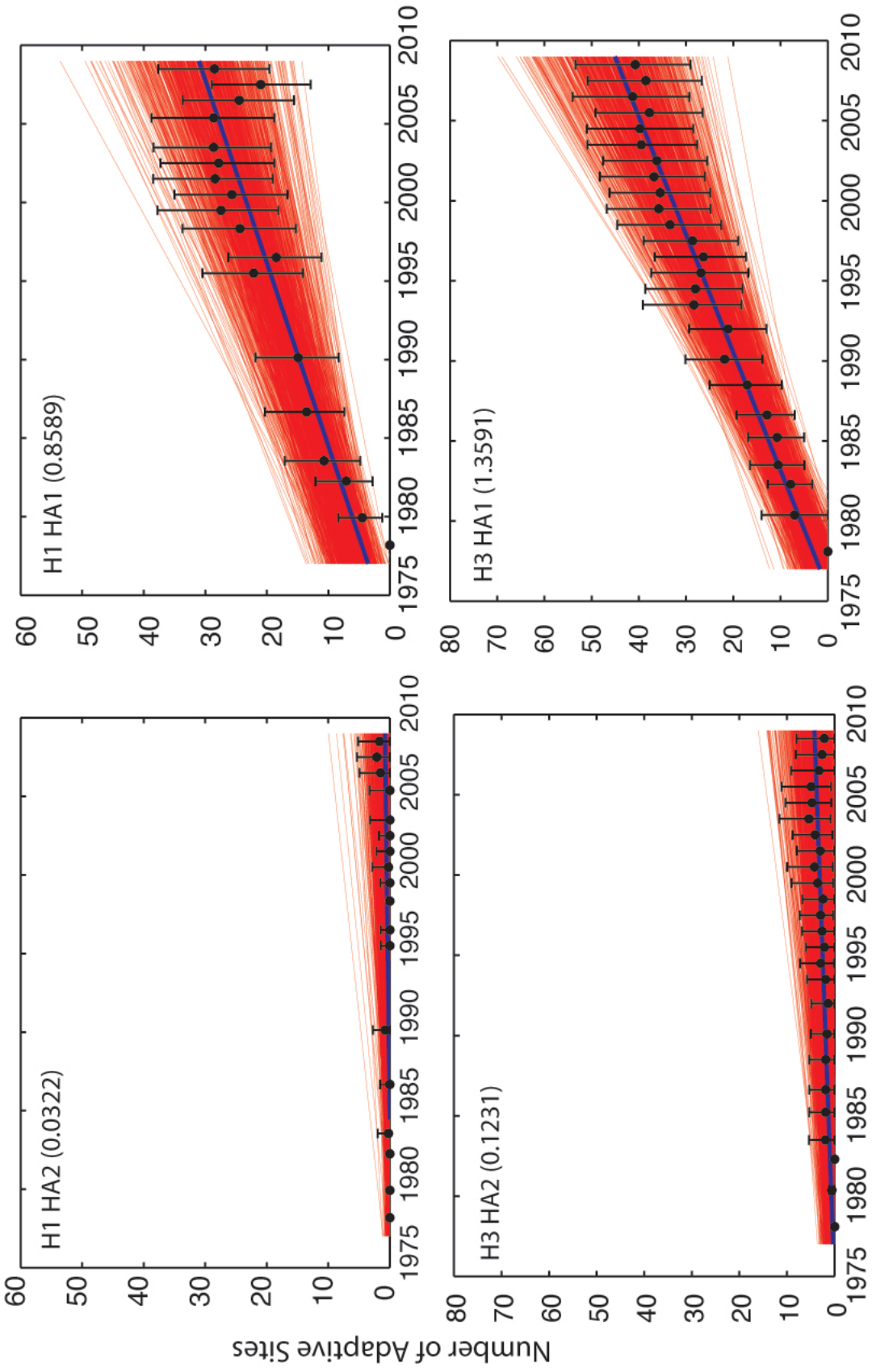


Figure 3.9: *Number of adaptive changes for H1N1 and H3N2 HA1 and HA2 from 1977 to 2008. Black points are the number of adaptive changes calculated using equation 3.14. At each point the error bars are the 95% bootstrap percentiles. The blue line is the best fit OLS regression and the red lines are the bootstrap regressions.Numbers in brackets are the rates of adaptive fixation*

3.5 Discussion

The methods described in this chapter provide a computationally tractable approach to evaluating the overall rate of adaptation of viral populations from exceptionally large sequence data sets. In Chapter 1 I showed that a low type I error is observed when employing the McDonald Kreitman test to sequence data that is comparable in diversity to influenza A data. The methods developed and described in this chapter are essentially a simple extension of the McDonald Kreitman test, and therefore are also likely to exhibit good type I error rates.

It has been well established that antigenic variation occurs predominantly in the HA1 subdomain of HA, while the HA2 subdomain remains relatively conserved (Skehel and Wiley, 2000). This expected pattern is repeated by my method, as shown in figure 3.9 on the preceding page, where a clear difference in rates of adaptation is observed, with a much higher rate observed in the HA1 subdomain for both subtypes. This suggests that the method I have introduced has reasonable statistical power for calculating an overall rate of adaptation. A further indication of the reliability of this approach is the low rates of adaptation (for both subtypes) observed in the highly conserved polymerase genes (figure 3.8 on page 95). However, it should be noted that two factors may reduce the statistical power of the method. Firstly, adaptive sites in the mid frequency ranges would cause underestimates of the true levels of adaptation occurring. I attempted to account for this by calculating the 'neutral ratio' using only polymerase genes, which are unlikely to contain adaptive sites at mid frequencies. Secondly, my approach is likely to underestimate adaptation caused by multiple selective sweeps at the same site, but this factor seems to be insignificant given that I observed a constant rate of adaptive fixation through time, even for the HA gene.

These observations indicate that the results of the whole genome analysis (figure 3.8) are largely reliable, especially if they are interpreted as a conservative, minimum estimate of the rate of molecular adaptation. The key result from this study is the variation in HA and NA adaptive dynamics between subtypes. As discussed in section 3.2.4 the main focus of

research into the adaptation of influenza has been on the HA gene. For both H3N2 and H1N2, the highest rates of adaptation are observed in HA, but the rates for NA are also very high. Rambaut et al. (2008) found that HA and NA have similar evolutionary rates; the high rate of adaptation in NA for both subtypes may be due to antibody mediated immune responses, or due to changes associated with functional compatibility and coevolution with HA (e.g. for efficient virus replication (Mitnaul et al., 2000)). In either case, these results confirm the importance of the use of NA inhibition assays in vaccine selection, and suggest that more research is needed into investigating positive selection and the evolutionary histories of NA.

There is a marked difference in rates of adaptation for both HA and NA between subtypes, with the rates of adaptation for both genes being higher in H3N2 than in H1N1 (0.56 higher for HA and 0.16 higher for NA). Since the reintroduction of H1N1 into human populations in 1977, the H1N1 subtype has exhibited a lower virulence and mortality rate compared to H3N2, particularly if the strains are circulating in close temporal association (Wright et al., 1980; Kaji et al., 2003). In addition Rambaut et al. (2008) showed that the genetic diversity H1N1 undergoes less severe seasonal genetic bottlenecks than H3N2. This observation could suggest weaker immune selection on H1N1 compared to H3N2, resulting in weaker or less frequent selective sweeps. The difference in rates of adaptation shown in figure 3.8 supports this hypothesis. Furthermore the difference between HA and NA is greater for H3N2 than for H1N1, which may suggest that immune selection on HA is the driving force causing the higher rate of adaptation in H3N2.

Aside from the envelope proteins, genes NP and NS1 show high rates of adaptation, with higher rates being seen in NP for both subtypes (0.36 in H3N2 and 0.41 in H1N1). The cause of this adaptation is not obvious, but some hypotheses can be suggested. One possible cause is immune pressure mediated by CTLs, which these genes are known to undergo (McMichael et al., 1983; Townsend et al., 1984, 1986; Ludwig et al., 1991; Voeten et al., 2000; Gog et al., 2003; Berkhoff et al., 2005; Fernandez-Sesma et al., 2006; Suzuki, 2006; Lin et al., 2007). There has

been evidence that responses against NP may play a role in partial cross protection between subtypes (Tumpey et al., 2005; Epstein et al., 2005; Ahmed et al., 2007) which is a possible explanation for the high rates of adaptive evolution I observed. These results have important implications for vaccine design, as gene-based vaccination with NP may help contribute to protective immunity against diverse influenza viruses. The adaptation observed in NS1 could be explained by viral evasion of host innate immune response (Baigent and McCauley, 2003; Li et al., 2006). High rates of adaptation are also seen in M2 but curiously only in H1N1. M2 has been implicated many times in adamantane resistance (Weinstock and Zuccotti, 2006; Bright et al., 2006; Deyde et al., 2007), but mainly in H3N2 (Simonsen et al., 2007)

Recent analysis by Nelson et al. (2006) indicated that positive selection on the HA1 domain occurs in a punctuated manner. However my results in figures 3.9, 3.6 and 3.7 suggest that adaptive fixation in HA, NA and internal proteins NS1 and NP occurs at a constant rate. This discrepancy is most likely due to the difference in sample sizes and methods used. The time period of sampling used in this chapter is considerably longer than that used in Nelson et al. (2006), and therefore it is unlikely that any short time punctuated events are seen.

It is clear that genes NP,NS1 and M2 contribute to influenza adaptation and that research into the evolutionary dynamics of these genes has important practical applications in influenza drug research. Furthermore, recent research by Rambaut et al. (2008) has shown that genetic reassortment plays a key role in influenza evolution, particularly if changes among segments increase fitness (Rimmelzwaan et al., 2004). In this light future research on the evolutionary dynamics of influenza should be considered on a genomic level, and not just on the evolutionary dynamics of HA and NA.

The methods developed in this chapter are developed with next generation sequencing technology in mind (Brenner et al., 2000; Ronaghi et al., 1996; Bains and Smith, 1988). The methods in this chapter become more informative and accurate with greater amounts of data, and as such data becomes available, improvements in my methodology can be readily applied

without any reductions in computational speed.

Chapter 4

A general framework for investigating nucleotide site frequencies in viral populations

4.1 Introduction

RNA viruses have the potential to adapt exceptionally rapidly as a result of their high mutation rates, large population sizes and short generation times. In viruses such as HIV and HCV, which replicate continuously during chronic infection, viral adaptation arises within individual infections and can lead to drug resistance and escape from host immune responses. Therefore the development of successful drug and vaccine strategies against such viruses requires an understanding of the evolutionary dynamics of adaptation. Significant progress in measuring positive selection and adaptation in viral genes, and codons, has been made, much of which uses methods based on the D_n/D_s ratio (e.g. Nielsen and Yang, 1998a or Suzuki et al., 2001). As discussed in chapter 2, in studies of chronic viral infection during which viral gene sequences are typically obtained from an infected individual serially through time, the interpretation of D_n/D_s becomes difficult and other methods are worth investigating. In addition, the use of phylogenetic D_n/D_s methods becomes practically challenging, or even unfeasible, when applied to very large data sets.

The alternative methods presented throughout this thesis are based on nucleotide site frequencies, such as the MK test. In chapter 1 I showed that the MK test gives acceptably low type I error when applied to RNA virus like data, and that the type I error can be improved by a 'proportional' method of site counting. The ability to make small methodological improvements without changing the robustness and computational tractability of the MK test makes it a very appealing starting point. Such modifications of the MK test, such as those reported by Smith and Eyre-Walker (2002) and Williamson(2003), allow an overall rate of viral adaptation to be estimated from large data sets of thousands of sequences. In Chapter 3, I used a combination of population genetics theory and empirical data to extend Williamson's (2003) method. The large influenza data sets investigated in Chapter 3 allowed for accurate estimation of the 'neutral ratio', and I obtained acceptably low variances when estimating numbers of adaptive sites and rates of adaptation.

The approach taken in Chapter 3, however, suffers from two problems: (i) the counting method used is likely to be influenced by sampling error when looking at data with small sample sizes. Sometimes comparatively few samples are available for a particular time point, a problem that was resolved in Chapter 3 by grouping samples from different time points. A more appropriate solution would be to incorporate sampling error into the method. Sampling error will likely be significant for samples of less than 10 sequences, as there will be considerable uncertainty in determining the true frequency of a derived mutation, or in classifying a site as being invariant (see case 1 figure 2.1 on page 40) or fixed. (see case 2 figure 2.1 on page 40) (where every base in the main alignment is the same but different from the outgroup) (ii) In Chapter 3 I used theory and empirical data to define the 'most neutral' site-frequency range as the range 0.25 - 0.75. I then evaluated the ratio of silent to replacement polymorphisms within this range from the highly conserved polymerase genes. This approach may not be valid for all data sets and instead a more suitable approach would be to evaluate the 'most neutral' range directly from the data.

To solve these problems I introduce a new probabilistic approach to estimating the mutational site frequency spectrum from a set of sequences. I develop a new Bayesian counting approach which assigns sites to site-frequencies and to site classes (i.e. silent vs. replacement) in a probabilistic manner, thereby incorporating sampling error. The method can estimate separate site-frequency distributions for silent and replacement sites. Using this approach I generalise the methods described in Chapter 3, thus allowing the signals of deleterious mutation pressure and positive selection to be observed within the same data set. I then test this new method on simulated neutrally-evolving data and apply it to within-patient Hepatitis C Virus (HCV) sequences.

4.2 Methods

Williamson (2003) developed a simple statistic to estimate the number of adaptive substi-

tutions from serially-sampled sequences (Chapter 3.3.2). A fundamental disadvantage of this method is that is does not account for sampling error when counting the number of polymorphic and fixed sites in the data. This is particularly important for studies of within-patient virus evolution, as sample sizes at each time point are typically small — about 2 to 20 sequences. At a given site i in the main alignment, only empirical frequencies of bases A, T, G and C are observed. However, these frequencies have large variances when sample sizes are small and therefore may not represent the true frequencies that exist at site i, that is, even though a site may appear 'fixed' in a sample of $5-10$ sequences, the true frequency in the population may be considerably less than 1.0. Similarly, if a site is invariant and no derived nucleotides are observed in the main alignment, there is a non-zero probability that it may actually be present in the population. To correct for this sampling error, I introduce a probabilistic counting method that incorporates multinomial sampling error. This probabilistic counting gives rise to a new type of site-frequency spectrum: the conventional unfolded site-frequency spectrum is composed only from sites that are polymorphic in the main alignment (Site types 3-7 figure 2.1 on page 40). But if probabilistic counting is used, then all sites in the main alignment, including those that are fixed or invariant, contribute towards the site-frequency spectrum, thereby producing a more accurate estimate of the true unfolded spectrum.

When calculating the number of adaptive sites, individual site frequencies are required, but there is also a need to evaluate whether a site is silent or replacement. In section 2.3.3 a proportional counting method was introduced which, rather than unambiguously assigning sites as fixed or replacement, gave proportional score. This proportional score is also prone to be biased by sampling error, and therefore I also introduce a binomial counting method for calculation of the site ratio of silent bases to replacement bases which similar to that used for evaluating the site frequency. Using these counting methods the data sets can be split into site frequency ranges, and used to develop a formula calculating the number of adaptive substitutions at any site frequency range.

4.2.1 Estimation of site frequency

Consider a main alignment consisting of N viral gene sequences, k nucleotides in length. For a given site i in this alignment, there exists an empirical observed frequency of each nucleotide base A, T, G and C. For example, consider a fixed site i in the main alignment ($N = 10$), such that the ancestral base is A and every base at site i in the main alignment is G. In this case the observed empirical frequencies of $\{A, T, G, C\}$ would be $\{0, 0, 10, 0\}$. These values are expectations based on the data but are not necessarily the true base frequencies. To estimate the true frequencies, I use a Bayesian framework. I define the set of true base frequencies as $\bar{\theta}_i$ (also referred to as the model parameters) and the set of observed frequencies as $\bar{\lambda}_i$ (the data) or more formally:

$$\begin{aligned} \bar{\theta}_i &\in \{\theta_A, \theta_C, \theta_G, \theta_T\} \\ \bar{\lambda}_i &\in \{\lambda_A, \lambda_C, \lambda_G, \lambda_T\} \end{aligned}$$

hence the probability of the true base frequencies ($\bar{\theta}_i$) given the data ($\bar{\lambda}_i$) is defined by Bayes rule:

$$\mathbb{P}(\bar{\theta}|\bar{\lambda}) = \frac{\mathbb{P}(\bar{\theta})\mathbb{P}(\bar{\lambda}|\bar{\theta})}{\mathbb{P}(\bar{\lambda})} \tag{4.1}$$

where $\mathbb{P}(\bar{\theta}|\bar{\lambda})$ is called the posterior probability, and represents the joint probabilities of the model parameters given the data. $\mathbb{P}(\bar{\theta})$ is the joint prior probability distribution of the model parameters. $\mathbb{P}(\bar{\lambda}|\bar{\theta})$ is the probability of the data given the model, also known as the likelihood. $\mathbb{P}(\bar{\lambda})$ is called the marginal distribution or evidence, and is the probability of the data.

To model the sampling error of the empirical site frequencies I use the multinomial distribution, for which the natural choice of prior distribution is the *Dirichlet* distribution. Due to conjugacy (Raiffa and Schlaifer, 1972), if these two distributions are chosen then the posterior distribution will have the same probability distribution as the prior and therefore also follow

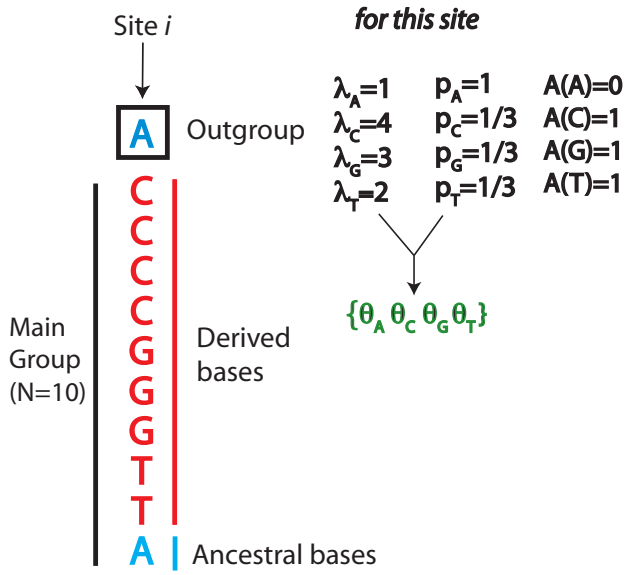


Figure 4.1: *Figure explaining notation used in this chapter. θ are the model parameters, λ is the data, p are the hyperparameters of the prior distribution , and A is the indicator function returning 0 is a base in the main alignment is ancestral and 1 if derived.*

a *Dirichlet* distribution. This relationship causes the marginal distribution $\mathbb{P}(\bar{\lambda}) = 1$, allowing for simpler evaluation of the posterior probabilities.

For a given site i, the likelihood $\mathbb{P}(\bar{\lambda}|\bar{\theta})$ is defined by the multinomial distribution

$$\mathbb{P}\left(\bar{\lambda}_i|\bar{\theta}_i\right) \propto \prod_j \theta_j^{\lambda_j} \tag{4.2}$$

Where $j \in \{A, C, G, T\}$ and $\lambda_A, +\lambda_C, +\lambda_G + \lambda_T = N$ (see figure 4.1 for an illustration of the notation used in this chapter). The Dirichlet prior distribution is given by

$$\mathbb{P}\left(\bar{\theta}_i|\bar{p}_i\right) \propto \prod_j \theta_j^{p_j - 1} \tag{4.3}$$

Where $\theta_j > 0$ for all bases j and $\theta_A + \theta_C + \theta_G + \theta_T = 1$. The set $\bar{p}_i \in \{p_A, p_C, p_G, p_T\}$ contains the hyperparameters of the prior probability distribution. They are referred to as hyperparameters to distinguish them from parameters of the underlying model ($\bar{\theta}$) or the data ($\bar{\lambda}$) (figure. 4.1).

Because of conjugacy mentioned above, the expected joint probability distribution of $\bar{\theta}_i$ also follows a Dirichlet distribution, which when written in full is given by:

$$\begin{aligned}
\mathbb{P}\left(\bar{\theta}_i|\bar{p}_i,\bar{\lambda}_i\right) &\propto \prod_j \theta_j^{\lambda_j} \prod_j \theta_j^{p_j-1} && (4.4)\\
&\propto \prod_j \theta_j^{\lambda_j+p_j-1} && (4.5)\\
&= \left[\frac{\prod_j \Gamma(p_j+\lambda_j)}{\Gamma\left(N+\sum_j p_j\right)}\right] \prod \theta_j^{p_j+\lambda_j-1} && (4.6)
\end{aligned}$$

where Γ is the gamma function and the term in square brackets in equation 4.4 is the multinomial beta function, which is a normalising constant. The contribution of the prior hyperparameters (p) to the posterior distribution is very intuitive, they simply represent additive parameters that represent prior belief in the frequencies of each base.

Using equation 4.4 it is possible to calculate the posterior probability that the *derived* bases at site i have a frequency D_i in the population, by summing all the posterior probabilities of non-ancestral bases in the probability vector $\mathbb{P}\left(\bar{\theta}_i|\bar{p}_i,\bar{\lambda}_i\right)$. I define $A_i(j)$ as an indicator function, which returns 0 if base j is present in the ancestral alignment and 1 if base j is derived. Therefore

$$\mathbb{P}(D_i|\bar{\lambda_i},\bar{p_i}) = \sum_j A_i(j) \cdot \mathbb{P}\left(\bar{\theta}_i|\bar{p}_i,\bar{\lambda}_i\right) \tag{4.7}$$

At this point I have not yet set values for hyper parameters $\bar{p}_i$. Theoretically, in the most general case $\bar{p}_i$ could be set individually for each site, but in what follows I will use a set of fixed values across all sites (figure 4.1). In the case of no strong prior belief it is usually preferable to use a uniform distribution as the prior. For equation 4.7, there exists a certain combination of hyperparameters (P^*) such that the shape of the prior distribution for each model parameter $\{\theta_A, \theta_C, \theta_G\theta_T\}$ becomes uniformly distributed. These hyperparameters are

$$p^* = \begin{cases} 1 & if\ A_i(j) = 0 \qquad i.e\ if\ base\ j\ is\ ancestral \\ \frac{1}{3} & if\ A_i(j) = 1 \qquad i.e\ if\ base\ j\ is\ derived \end{cases} \tag{4.8}$$

Using these hyperparameters, the Dirichlet prior distribution scales to a joint uniform distribution. For the rest of this chapter, the values p^* are used for every site in the main alignment.

Using equation 4.7, the probability that the population frequency of derived bases at site i lies within an interval $[u, v]$ is calculated as:

$$\mathbb{P}(u < D < v | p^*, \bar{\lambda_i}) = \int_{D_i=u}^{v} \mathbb{P}(D_i | \bar{\lambda_i}, p^*) \tag{4.9}$$

The integral can numerically evaluated in several ways - I choose to use Monte Carlo integration (Hammersley, 1960). To calculate the total number of sites with a derived base frequency in interval $[u, v]$, equation 4.9 is summed across all k sites in the main alignment

$$\hat{f}_{u,v} = \sum_{i=1}^{k} \mathbb{P}(u < D < v | p^*, \bar{\lambda_i}) \tag{4.10}$$

The values of u and v are the site-frequency range that contains $\hat{f}_{u,v}$ sites. Since by definition $\hat{f}_{0,1} = k$, the interval $[0, 1]$ can be split into any number of non-overlapping site frequency ranges. The number and width of the site-frequency ranges can be chosen to suit the analysis at hand. A plot of $\hat{f}_{u,v}$ against site-frequency is similar to a plot of the unfolded mutational site-frequency spectrum of the sequences. The difference between a $\hat{f}_{u,v}$ plot and an unfolded site-frequency spectrum is that the former contains all k sites in the main alignment, including those that are fixed and invariant with respect to the outgroup alignment, whereas the 'traditional' site frequency spectrum shows the frequencies of polymorphic sites only. The distinction arises from the inclusion of sampling error in my approach

	Sample Size: 4	Sample Size: 8	Sample Size: 24	Sample Size: 48	Sample Size: 192
Site frequency Mean	0.500	0.500	0.500	0.500	0.500
Site frequency Variance	$3.57e^{-3}$	$2.26e^{-3}$	$9.26e^{-4}$	$4.91e^{-4}$	$1.28e^{-4}$

Table 4.1: *Mean and standard deviation for artificial data set (i) where all derived bases segregate at observed frequency* 0.5*. The mean site frequency stays constant but the variance in site frequency reduces according to sample size.*

4.2.1.1 Application of probabilistic counting to artificial data sets

To illustrate how this probabilistic counting works, I consider two test nucleotide alignments: (i) an alignment that has 300 sites, all of which have a derived base at an observed frequency 0.5 (i.e site type 3, in figure 2.1 on page 40), and (ii) an alignment of 320 sites, half of which are fixed and half of which are invariant with respect to an outgroup sequence (i.e site types 1 and 2 in figure 2.1). Equation 4.10 is used to calculate a site frequency spectrum using a the uniform prior (equation 4.8). From figure 4.2 on the next page it is clear that for both data sets (i) and (ii), as the number of samples increases, the variance shrinks and uncertainty in the distribution of the site frequencies reduces (figure 4.2, table 4.1).

These artificial data sets illustrate how this new method can reconstruct the site frequency spectrum of mutations. Each site, including those which are fixed or invariant, is not assigned single frequency values, but rather follows a probability distribution whose variance represents the sampling uncertainty. This approach therefore not only provides a more complete site frequency spectrum, but also accounts for sampling error when sample sizes are small. My results in figure 4.2 indicate that the computational implementation of my method is correct.

4.2.1.2 Use of the prior distribution

Consider another artificial nucleotide alignment which contains an outgroup sequence of which has 300 sites, all of which are nucleotide A and a main alignment of 12 sequences

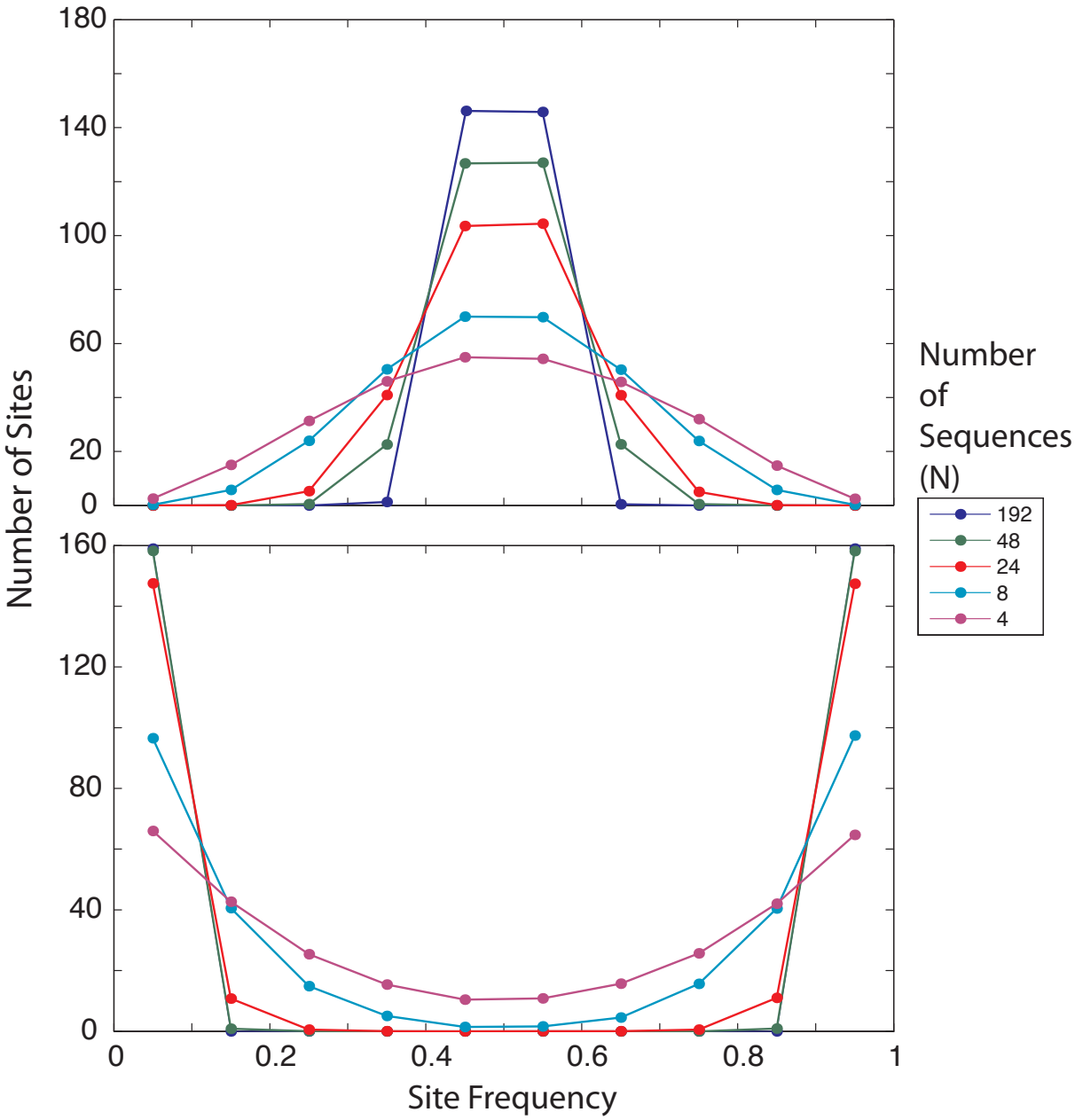


Figure 4.2: *Effect of site frequency counting on artificial nucleotide alignments with the sample size shown in the legend. (a) Alignment of 300 bases of which all are derived and at observed frequency* 50%. *(b) Alignment of 320 bases half of which are fixed and half invariant. The site frequency is evaluated over 10 bins spaced in intervals of size* 0.1, *i.e* $[0, 0.1]$, $[0.1, 0.2]$, ..., $[0.9, 1.0]$.

with 300 sites, all of which are fixed with respect to the outgroup (every base is nucleotide G). Suppose there is a prior belief that, in the main alignment, every site should in fact be invariant with respect to the outgroup (i.e every base is A).

This prior belief can be represented by using a Dirichlet ($\mathbb{D}$) prior (figure 4.3b on the next page) parameterised as

$$p = \mathbb{D}(A,\, C\,, G,\, T) = \mathbb{D}(1, 0.01, 0.01, 0.01) \tag{4.11}$$

Using Equation 4.10 to reconstruct a site frequency spectrum with this prior results in a posterior distribution which is a compromise between the data (figure 4.3a) and the prior distribution (figure 4.3b). However, as shown in equation 4.4, there is an additive relationship between the data and the prior (the posterior is given by a Dirichlet distribution parametrised as $\mathbb{D}(p_A + \lambda_A, p_C + \lambda_C, p_G + \lambda_G, p_T + \lambda_T)$). Therefore I introduce a term, called the confidence parameter (also a hyperparameter), which is a multiplicative constant that scales the hyperparameters of the prior distribution. For example, for the above prior, a confidence parameter of 5 would give a set of scaled hyperparameters $p = (5, 0.05, 0.05, 0.05)$, which would cause the prior distribution to contribute more towards the posterior distribution than the data.

Using equation 4.7 the posterior distribution is evaluated for confidence parameters 1, 5, 20, 50, 100 and 200. The posterior distributions shown in Figure 4.3c changes depending on the confidence parameter. At low confidence parameters, the data contributes more to the posterior distribution than the prior distribution, but as the confidence parameter increases, the prior shifts the posterior to lower site frequencies, reflecting the prior belief that all sites are in fact invariant.

The choice of the confidence parameter and prior depends on the data being analysed. Fixed hyperparameters can be set for the entire data set (such as the uniform prior mentioned earlier) or sites can each be given individual hyperparameters based on knowledge of the data. Given

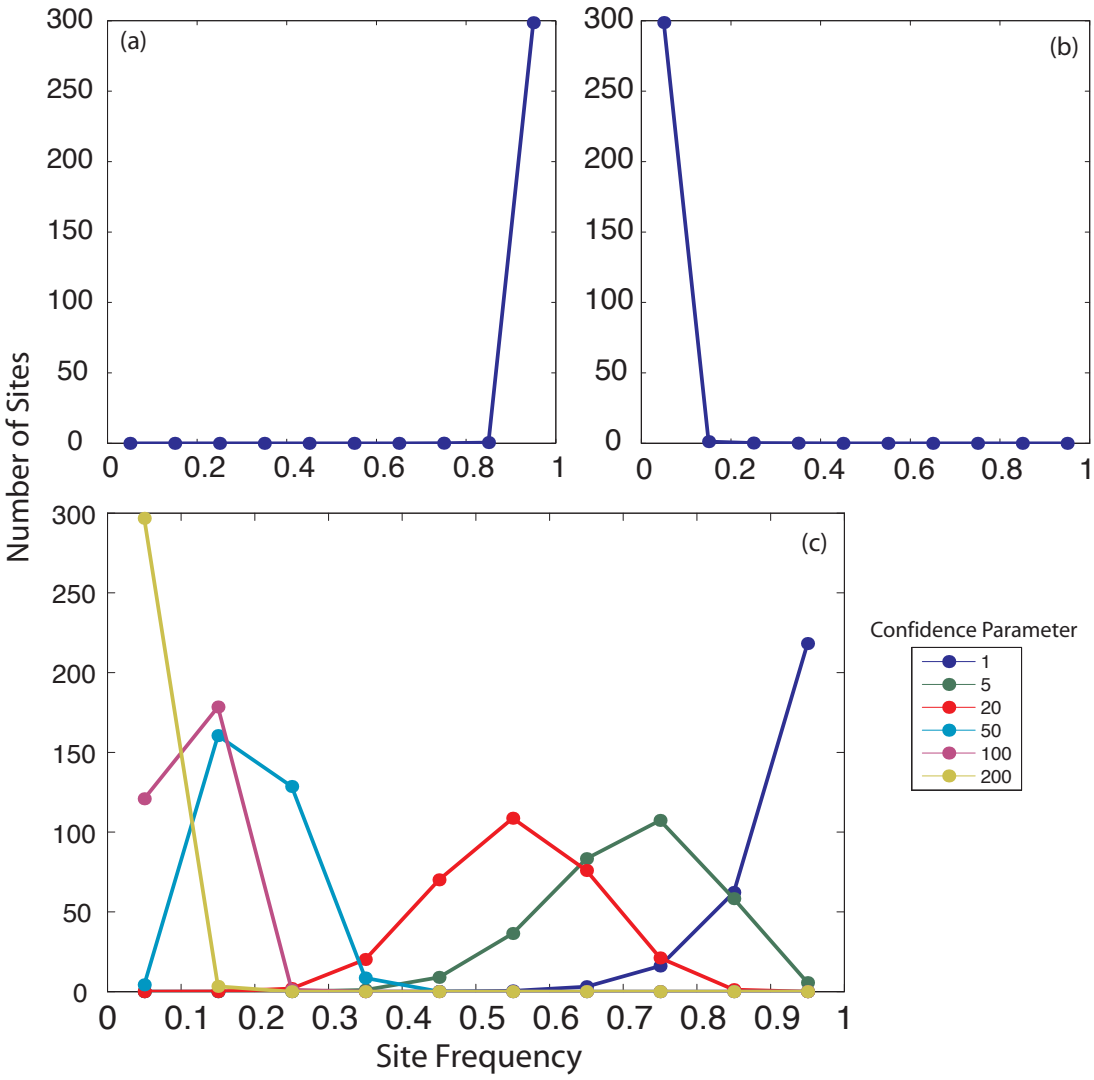


Figure 4.3: *Effect of the prior distribution on the site frequency (a) shows the site frequency spectrum of the data which comprises an alignment of 12 samples and 300 sites, all of which are fixed with respect to the outgroup. (b) shows the site frequency generated by the prior distribution (c) Shows the posterior using different 'confidence parameters' on the prior. The site frequency is evaluated using equation 4.7over* 10 *bins, i.e* $[0, 0.1], [0.1, 0.2], ..., [0.9, 1.0]$.

no strong prior knowledge it is safest to assume a uniform distribution for the prior. For determining the confidence parameter, a vast amount of literature exists on optimal selection of the confidence parameter, dating back to the early 1970s. However, the choice remains a very difficult question and is in general unsolved (Golub et al., 1979). Prior information could be obtained directly from empirical data combined with easily parameterisable distributions such as the gamma distribution or the log normal distribution. For the analysis in the rest of this chapter I choose to use a uniform prior. This uniform prior takes the values defined in equation 4.8 and therefore does not require a scaling confidence parameter.

4.2.2 Estimation of the replacement/silent ratio

In section 2.3.3 a proportional counting method was introduced to evaluate the number of sites that are silent or replacement. As with determining site frequencies, these silent-to-replacement proportions are also affected by sampling error. To account for this I introduce a binomial probabilistic counting method to estimate, at a given site, the proportion of derived bases that are silent or replacement. This method is effectively the same as that used in section 4.2.1 except performed on a two state (silent or replacement), not a four state (A, T, G, C), model. For a binomial likelihood the natural choice of prior is a beta distribution. Both the binomial and the beta distribution are special cases of the multinomial and Dirichlet distributions, and as a result are also conjugate with respect to each other (Raiffa and Schlaifer, 1972).

To begin, consider a main alignment consisting of N viral gene sequences, k nucleotides in length. For a given site i in this alignment there exists an empirical observed frequency for the number of derived substitutions which are silent (s) or replacement (r) when compared to an outgroup. Values s and r are therefore the parameters of the data, where their sum is the number of derived bases i.e $s_i + r_i = D_i$. However, if site i is invariant (site type 1, section 2.3.3, figure 2.1 on page 40), it is not possible to decisively classify a site as silent or

Table 4.2: *The genetic code. N stands for any nucleotide (U,C,A,G), Y stands for any pyrimidine (U,C), R stands for any purine (A,G) and H stands for not G (U,C,A)*

		2nd POSITION							
		U		C		A		G	
1st POSITION	U	UUY UUR	PHE LEU	UCN	SER	UAY UAR	TYR STOP	UGY UGA UGG	CYS STOP TRP
	C	CUN	LEU	CCN	PRO	CAY CAR	HIS GLN	CGN	ARG
	A	AUH AUG	ILE MET	CAN	THR	AAY AAR	ASN LYS	AGY AGR	SER ARG
	G	GUN	VAL	GCN	ALA	GAY GAR	ASP GLU	GGN	GLY

replacement. I will first consider this case.

4.2.2.1 Invariant sites

For invariant sites, depending on the codon position, I assign a fixed probability of whether the site is silent or replacement. This probability is determined from observing the codon degeneracy inherent in the genetic code (tables 4.2 and 4.3 on the following page). Probabilities are assigned depending on whether a change at a site is silent (no amino acid change), replacement (causes an amino acid change) or ambiguous (depending on the nucleotide, the change is either silent or replacement).

To evaluate probabilities of silent or replacement changes at each codon position, I assume random mutation, equal codon frequencies and equal transitions to transversions. Using table

Table 4.3: *Codon degeneracy table assuming equal codon frequencies and an equal transition transversion ratio. At a given position codons which are always silent regardless of what nucleotide change occurs are assigned a* 1*, replacement changes are assigned a* 0 *and ambiguous changes are assigned a* 0.5

Amino Acid	Codon	pos 1	pos 2	pos 3	Amino Acid	Codon	pos 1	pos 2	pos 3
Lysine	**AAA**	0	0	0.5	**Glutamate**	**GAA**	0	0	0.5
Asparagine	**AAC**	0	0	0.5	**Apartic acid**	**GAC**	0	0	0.5
Lysine	**AAG**	0	0	0.5	**Glutamate**	**GAG**	0	0	0.5
Asparagine	**AAU**	0	0	0.5	**Apartic acid**	**GAU**	0	0	0.5
Threonine	**ACA**	0	0	1	**Alanine**	**GCA**	0	0	1
Threonine	**ACC**	0	0	1	**Alanine**	**GCC**	0	0	1
Threonine	**ACG**	0	0	1	**Alanine**	**GCG**	0	0	1
Threonine	**ACU**	0	0	1	**Alanine**	**GCU**	0	0	1
Arginine	**AGA**	0.5	0	0.5	**Glycine**	**GGA**	0	0	1
Serine	**AGC**	0	0	0.5	**Glycine**	**GGC**	0	0	1
Arginine	**AGG**	0.5	0	0.5	**Glycine**	**GGG**	0	0	1
Serine	**AGU**	0	0	0.5	**Glycine**	**GGU**	0	0	1
Isoleucine	**AUA**	0	0	0.5	**Valine**	**GUA**	0	0	1
Isoleucine	**AUC**	0	0	0.5	**Valine**	**GUC**	0	0	1
Methionine	**AUG**	0	0	0	**Valine**	**GUG**	0	0	1
Isoleucine	**AUU**	0	0	0.5	**Valine**	**GUU**	0	0	1
Glutamine	**CAA**	0	0	0.5	**Stop**	**UAA**	0	0	0
Histidine	**CAC**	0	0	0.5	**Tyrosine**	**UAC**	0	0	0.5
Glutamate	**CAG**	0	0	0.5	**Stop**	**UAG**	0	0	0
Histidine	**CAU**	0	0	0.5	**Tyrosine**	**UAU**	0	0	0.5
Proline	**CCA**	0	0	1	**Serine**	**UCA**	0	0	1
Proline	**CCC**	0	0	1	**Serine**	**UCC**	0	0	1
Proline	**CCG**	0	0	1	**Serine**	**UCG**	0	0	1
Proline	**CCU**	0	0	1	**Serine**	**UCU**	0	0	1
Arginine	**CGA**	0.5	0	1	**Stop**	**UGA**	0	0	0
Arginine	**CGC**	0	0	1	**Cysteine**	**UGC**	0	0	0.5
Arginine	**CGG**	0.5	0	1	**Tryptophan**	**UGG**	0	0	0
Arginine	**CGU**	0	0	1	**Cysteine**	**UGU**	0	0	0.5
Leucine	**CUA**	0.5	0	1	**Leucine**	**UUA**	0.5	0	0.5
Leucine	**CUC**	0	0	1	**Phenylalanine**	**UUC**	0	0	0.5
Leucine	**CUG**	0.5	0	1	**Leucine**	**UUG**	0.5	0	0.5
Leucine	**CUU**	0	0	1	**Phenylalanine**	**UUU**	0	0	0.5

4.3the probability of an invariant site being silent is given by:

$$\mathbb{P}(s_i) = \begin{cases} 0.071 & position\ 1 \\ 0.0 & position\ 2 \\ 0.653 & position\ 3 \end{cases} \tag{4.12}$$

Consequently the probability of an invariant site i being replacement is given by $\mathbb{P}(s_i) = 1 - \mathbb{P}(r_i)$. It is also possible to calculate the probabilities for unequal codon frequencies by scaling the values in table 4.3, or account for different transition transversion ratios by using table 4.2. Alternatively these probabilities can also be set *a priori* or evaluated directly from empirical data.

4.2.2.2 Variant sites

For sites that are variable in the main alignment (site types 2-7, section 2.3.3, figure 2.1 on page 40), the posterior probabilities that they are silent or replacement can be calculated using a Bayesian approach similar to that used in section 4.2.1, except using a simpler two state case. To begin, consider a site i with D_i derived bases, of which s_i are silent and r_i are replacement. I define the true probability of the site being silent as π_i, and replacement as $1 - \pi_i$.

Using the same approach as in section 4.2.1, I use the binomial distribution to represent the likelihood that the true proportion of derived bases at site i being silent is equal to π_i

$$\mathbb{P}(s_i, D_i|\pi_i) \propto \pi^{s_i}(1-\pi_i)^{D-s_i} \tag{4.13}$$

Using a conjugate beta distributed prior, the prior probability is thus given by:

$$\mathbb{P}(\pi_i) \propto \pi_i^{p_s-1}(1-\pi_i)^{p_r-1} \tag{4.14}$$

Hyperparameters p_s and p_r represent the prior knowledge in the *frequencies* of silent bases and replacement bases respectively. Using these two distributions the resulting posterior density of π is also beta distributed and given by:

$$\begin{aligned}
\mathbb{P}(\pi_i|s_i, D_i) &\propto \pi^{s_i}(1-\pi)^{D_i-s_i}\pi^{p_s-1}(1-\pi)^{p_r-1} \\
&\propto \pi^{s_i+p_s-1}(1-\pi)^{D_i-s_i+p_r-1} \\
&= \left[\frac{\Gamma(D_i-2)}{\Gamma(S_i+p_s-1)\Gamma(D_i-S_i+p_r-1)}\right]\pi^{s_i+p_s-1}(1-\pi)^{D_i-s_i+p_r-1} \quad (4.15)
\end{aligned}$$

where Γ is the gamma function, and the term in the square brackets is a normalising constant. I choose to use the mean of the posterior distribution as the probability that site i is silent:

$$\mathbb{P}(s_i) = \frac{s_i + p_s - 1}{D_i + p_s + p_r - 2} \quad (4.16)$$

Correspondingly the probability that site i is replacement is $1 - \mathbb{P}(s_i)$.

If a unit uniform prior of $p_s = p_r = 1$ is required then equation 4.15 reduces to

$$\mathbb{P}(\pi_i|s_i, D_i)_{unif} = \frac{(D_i+1)!}{(D_i-s_i)!s_i!}\pi^{s_i}(1-\pi)^{D_i-s_i} \quad (4.17)$$

However as discussed above, the probability of a site being silent or replacement varies according to codon position. For example, in the standard genetic code, no change at codon position 2 is silent, and therefore using a uniform prior could introduce significant bias. In section 4.2.1, where the site frequency spectrum is estimated, it is very difficult to assign a prior belief to each site because each has its own empirical base frequencies. However, when calculating whether a site is silent or replacement the genetic code can provide a strong prior belief in what the distribution should look like. Prior probabilities for each codon position can be found from empirical data or by using codon degeneracy as in equation 4.8. For analysis in this chapter I use the values defined in equation 4.12 as prior probabilities. As introduced

in section 4.2.1.2, hyperparameters that determine the prior distribution can be scaled by the use of a confidence parameter. I assign a confidence parameter *a priori* of 2. The choice of this confidence parameter is chosen such that the prior mean only affects the posterior mean when the number of silent or replacement changes is low at a given site (site frequencies $[0.0 - 0.2]$.

4.2.3 The split site frequency spectrum

Using equation 4.10 and 4.15 the number of sites in each frequency range can be calculated separately for silent and replacement sites. If $\rho_{u,v}$ and $\sigma_{u,v}$ define the estimated number of replacement and silent sites with a frequency between $[u, v]$, then

$$\begin{aligned} \sigma_{u,v} &= \sum_{i=1}^{k} \mathbb{P}(s_i)\, \mathbb{P}(u < D < v | p^*, \bar{\lambda}_i) \\ \rho_{u,v} &= \sum_{i=1}^{k} (1 - \mathbb{P}(s_i))\ \mathbb{P}(u < D < v | p^*, \bar{\lambda}_i) \end{aligned} \tag{4.18}$$

If the sampled sequences contain $S = \sum_{i=1}^{k} s_i$ silent sites and $R = \sum_{i=1}^{k} r_i$ replacement sites then $\sigma_{0,1} = S$, $\rho_{0,1} = R$ and $S + R = k$.

4.2.4 Estimating the number of non neutral sites

Here I derive a new estimator of the number of non neutral sites in a set of sequences which is a generalisation of the methods described in chapter 3 section 3.3.2.4. First consider a site frequency range, denoted X with lower and upper bounds equal to x and x^* respectively. This range X contains both silent (σ_{x,x^*}) and replacement (ρ_{x,x^*}) sites. As defined above S and R are the total number of silent and replacement sites in the sequence alignment. In what follows all silent sites are assumed neutral. If q_X is the *proportion* of neutral sites that fall within range X, then the expected number of silent sites that fall within this range is $\sigma_{x,x^*} = Sq_X$ (see figure 4.4 on the following page).

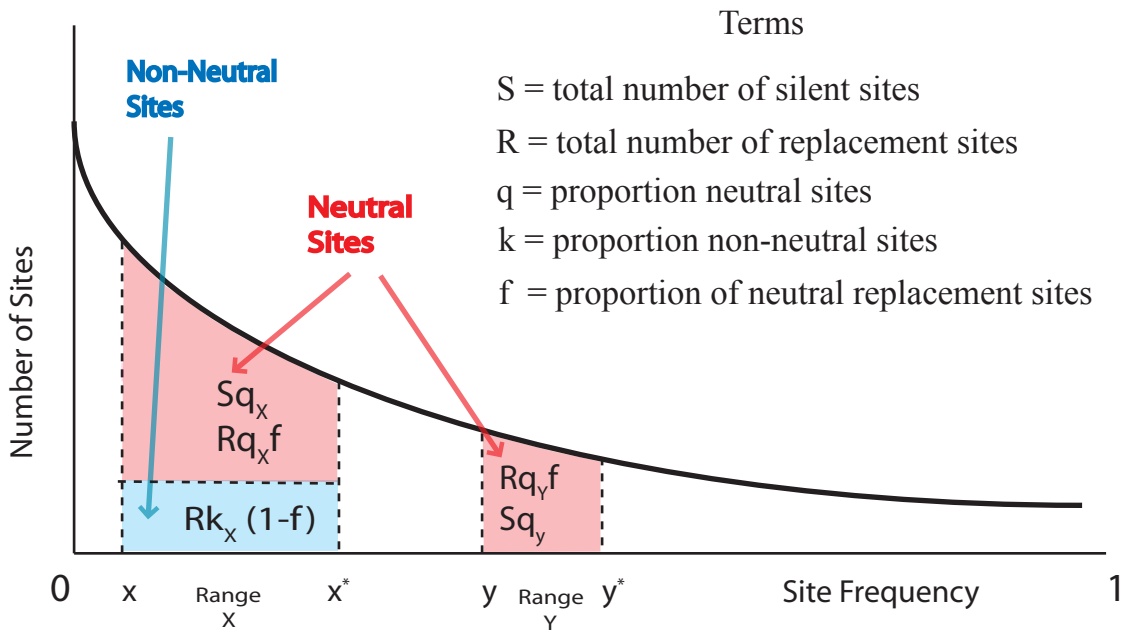


Figure 4.4: *Figure showing terms used in the derivation of equation4.21 . The area under the solid black curve represents the number of sites. The regions in dotted lines represent two hypothetical regions, one with both non neutral and neutral sites (X) and one with only neutral sites(Y).*

Next consider replacement sites in range X, some of these will be neutral and some will be non neutral. The proportion of these that are neutral is denoted f. The expected number of neutral replacement sites that fall within range X is therefore $Rq_X f$. The remaining $R(1-f)$ replacement sites are non neutral, and the proportion of these that fall in range X is denoted k_X (see figure 4.4). From this the total number of replacement sites in range X is just the sum of these two terms

$$\rho_{x,x^*} = Rq_X f + Rk_X(1-f) \tag{4.19}$$

where $Rk_X(1-f)$ is the expected number of non neutral sites in range X.

Now suppose there exists a site frequency range Y whose replacement sites are exclusively neutral. In other words it is unlikely that any $R(1-f)$ segregate at this frequency. The number of replacement sites and silent sites in range Y (with upper and lower bounds y and

y^*) is

$$\begin{aligned} \rho_{y,y^*} &= Rq_Y f \\ \sigma_{y,y^*} &= Sq_Y \end{aligned}$$

The ratio $\frac{\rho_Y}{\sigma_Y} = z$ or the 'neutral ratio' and Y the 'most neutral' range or simply the neutral range (introduced in chapter 3). To calculate the number of adaptive sites in site frequency range X (α_X), I follow the standard Smith and Eyre-Walker (2002) derivation.

$$\frac{\rho_X - \alpha_X}{\sigma_X} = \frac{\rho_Y}{\sigma_Y} \tag{4.20}$$

And from rearranging this equation α_X is given by:

$$a_X = \rho_X \left(1 - \frac{\sigma_X}{\rho_X}\frac{\rho_Y}{\sigma_Y}\right) \tag{4.21}$$

This formula has the same form as that used in chapter 3 section 3.3.2.4, except in that earlier work z was set to a mid frequency range of $0.25 - 0.75$. Using the above approach, z can represent the ratio of replacement to silent sites in any site-frequency range. Using equation 4.18 the 'most-neutral' site-frequency range can be evaluated from data which can subsequently used to estimate z. And from z the numbers of non neutral sites can be calculated for any frequency range X across the whole site frequency spectrum. In this framework non neutral sites at low frequencies are likely to represent deleterious or slightly deleterious mutations, and low frequency advantages mutations, whereas high frequency non neutral sites are likely to represent advantageous mutations. Using this approach provides a generalisation of the approach taken in chapter 3.

4.3 Results

4.3.1 Neutral simulations

4.3.1.1 Site frequency spectrum of simulated neutral data

I initially applied the probabilistic counting method method developed above (4.2.1) to alignments simulated under neutral evolution and under a demographic model of constant population size (a subset of the data set used in section 2.4.2). A range of 7 logarithmically spaced θ values were chosen ($\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1, 10\}$), and for each θ value a single neutral coalescent tree was simulated. Each tree comprised 50 taxa sampled at a single time point in the present and 1 taxa sampled τ in the past. Three different τ values were chosen ($\{0.3, 1.85, 4.5\}$). These values were chosen to represent varying genetic diversity under three different scenarios: a close outgroup, a distant outgroup and an intermediate outgroup. In each case a uniform prior was used and the site frequency spectrum evaluated over 10 equally spaced ranges (see section 4.2.1).

The effect of the probabilistic counting method on neutral data is shown in Figure 4.5 on the next page. First consider the case of a close ancestral sequence outgroup (figure 4.5a). Under low/mid θ ($10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}$) values, there are few differences between the ingroup and outgroup. As a result, while there are some fixed sites ($\sim 5\%$), the majority of sites are either invariant or contain polymorphisms at low frequencies. As θ becomes larger $(0.1, 1)$, sites become more variable and more fixed sites and polymorphic sites are seen (figure 4.5a). For extreme θ values (10) every site in the alignment is polymorphic, and the vast majority of these containing multiple mutations. In this scenario most sites have polymorphic frequencies in middle ranges, and very few sites are invariant or fixed.

F or low θ values, as τ increases and the outgroup becomes more distant, fewer invariant or low frequency polymorphic sites are observed and more fixations occur due to the effect of random genetic drift (figures 4.5b and 4.5c). For higher θ values the act of increasing the

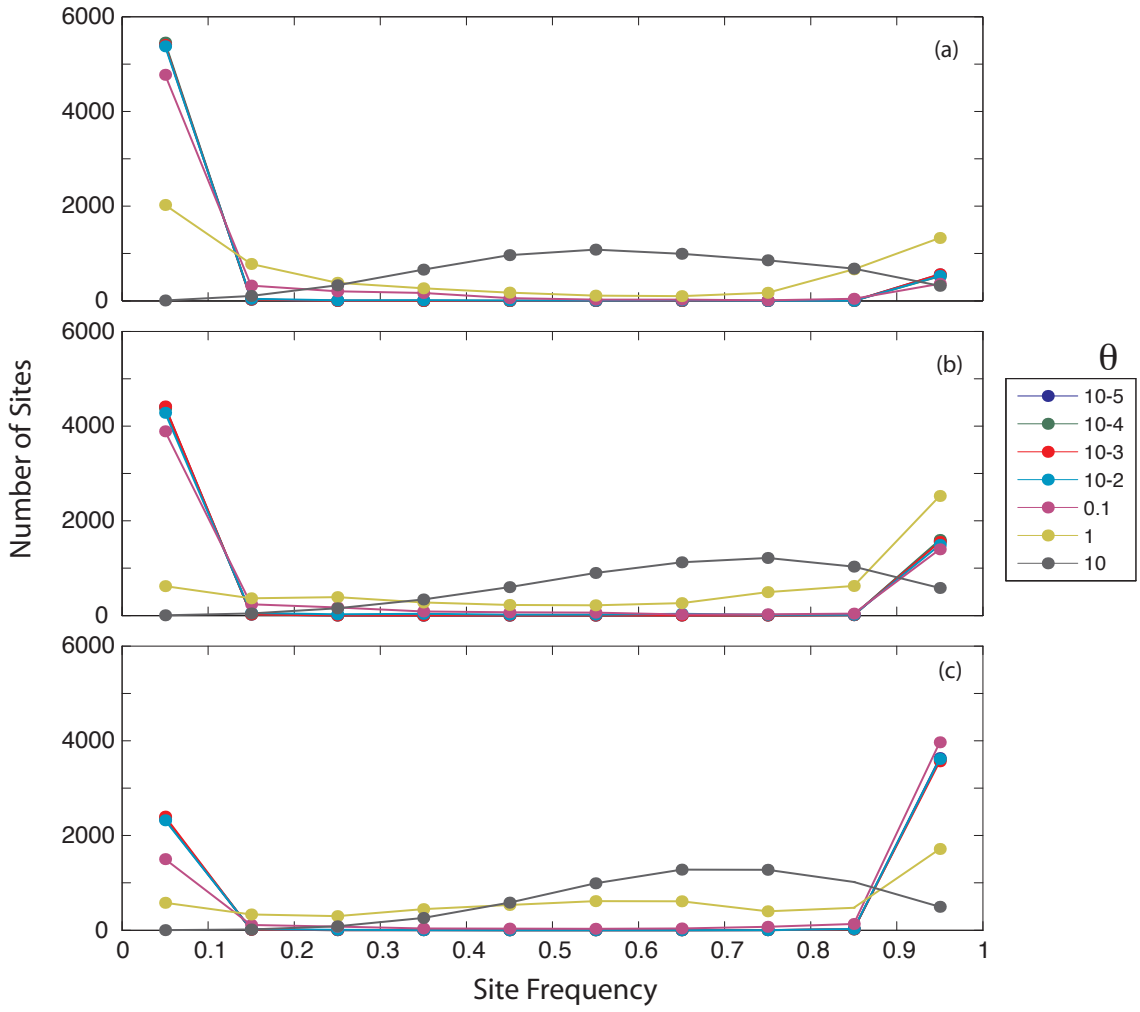


Figure 4.5: *Reconstructed site frequency spectrum for neutral data simulated under a model of constant population size. Figure (a) is simulated with a close outgroup.* $\tau = 0.3$*, (b) is simulated with an intermediate outgroup* $\tau = 1.85$ *and (c) is simulated with a distant outgroup* $\tau = 4.5$*.*

distance to the outgroup is to cause polymorphisms to segregate at higher frequencies. This is because many sites have experienced fixation but are seen as polymorphic by the very high level of variation in the main group alignment. This is distinct from the site frequency spectrum in literature Bustamante et al. (2001); Nielsen (2005) (which are constructed on an infinite sites model) because multiple mutations can occur at fixed sites, giving rise to high frequency polymorphisms.

4.3.1.2 The effect of probabilistic counting on neutrally simulated data

Nucleotide alignments were simulated under neutral evolution and under a demographic model of constant population size. 5 coalescent trees were generated with $\theta = 0.1$ and $\tau = 1.85$. Each tree comprised 10 taxa sampled at a single timepoint in the present and 1 taxa sampled τ in the past. For each tree a nucleotide alignment was simulated using EVOLVER (Yang, 1997). Then for each nucleotide alignment two site frequency spectrum's were constructed. The first site frequency spectrum was reconstructed using the new probabilistic approach. The second site frequency spectrum was reconstructed using just counting the observed empirical base frequencies from the data i.e if $M_j$represents the number of sites that have derived polymorphisms segregating at frequency j then the site frequency spectrum over 10 equally spaced frequency ranges is defined by vector $\{M_{[0,0.1]}, M_{[0.1,0.2]}, ..., M_{[0.9,1]}\}$.

Figure 4.6 on the following page shows the effect of probabilistic counting on neutral data. It is clear that the probabilistic approach has 'smoothed' the site frequency spectrum. The effect of sampling error, which is absent in figure 4.6a, is seen as uneven fluctuations in figure 4.6b. The 'smoothed' effect in figure 4.6a represents the correction of sampling error as the number of samples per nucleotide alignment is low (10). As the number of taxa increases the two spectrum's will converge to the same distribution.

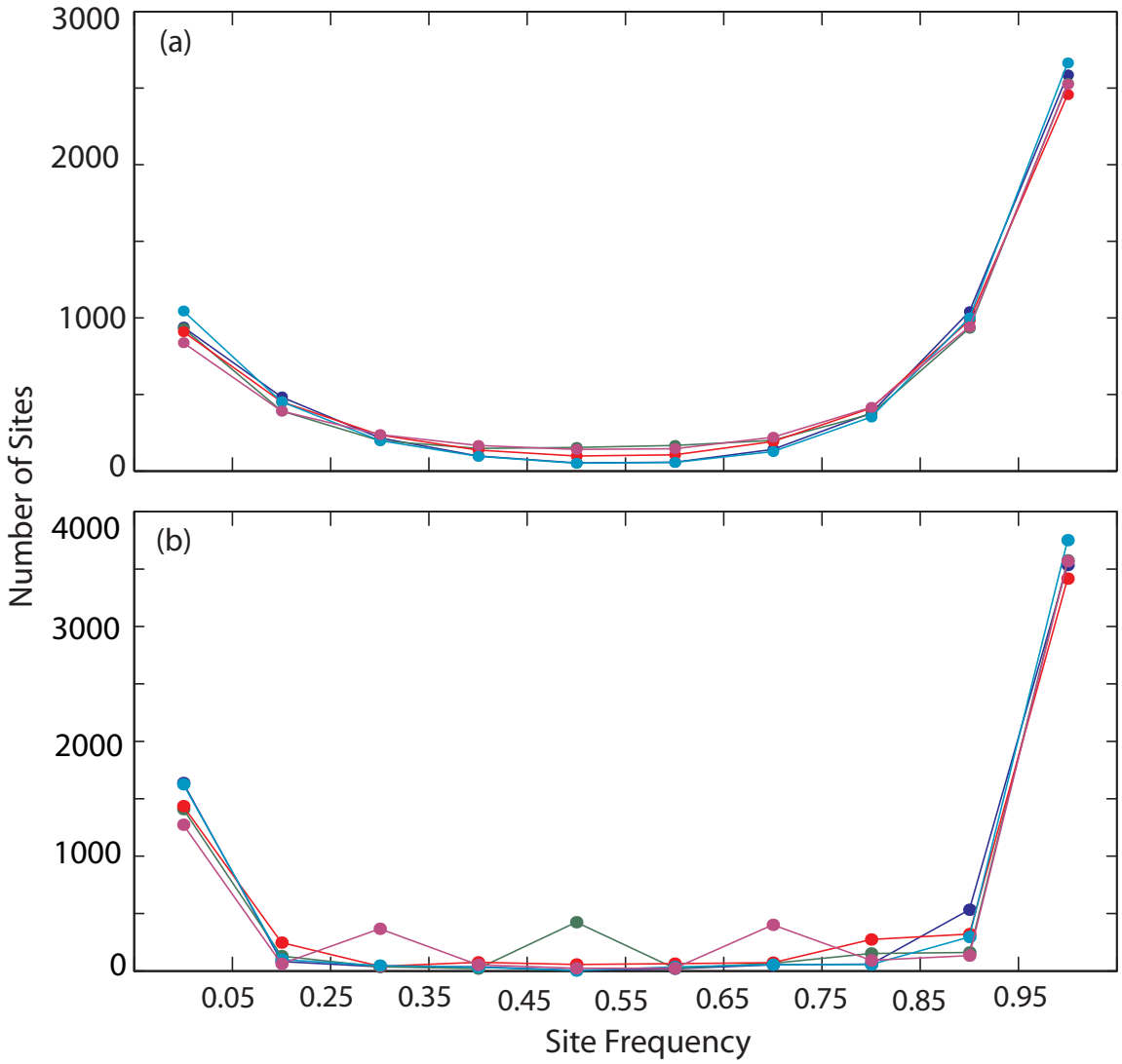


Figure 4.6: *Reconstructed site frequency spectrum (a) verses traditional site frequency spectrum (b) for neutral data simulated under a model of constant population size.*

4.3.1.3 The Replacement/Silent ratio of neutrally simulated data

To test the methods developed above (section 4.2.2) for estimating the number of silent and replacement sites in an alignment, I generated simulated data sets of neutrally evolving sequences under a demographic model of constant population size. 500 neutral coalescent trees were simulated with $\theta = 0.1$. Each tree comprised 50 taxa sampled at the present and 1 taxon sampled $\tau = 2.5$ time units into the past. For each tree, an alignment of sequences, each 6000 codons long, was generated using EVOLVER (Yang, 1997). Simulations were performed using equal codon frequencies. For each of these 500 alignments, the site frequency spectrum was calculated over 10 equally spaced ranges. The number of silent and replacement sites in each range was then calculated using equation 4.18. Finally for each frequency range the median, 5% and 95% percentiles were are calculated across all alignments.

As expected the replacement/silent (ρ/σ) ratio for neutral sequences does not change over the site frequency spectrum (Figure 4.7 on the next page). This is because both silent and replacement changes are neutral, hence their frequency spectrum's should be identical in shape. The mean value of ≈ 3 is due to the nature of the genetic code, where replacement sites are more common than silent ones. The size of the percentiles, which represent the variance, is proportional to the number of sites in each bin, and as a result the distribution intervals for fixed and invariant sites are very tight, reflecting the large numbers of these sites ($\approx 45\%$ are invariant and $\approx 52\%$ are fixed). For middle frequency bins $(0.4, 0.5)$, which only account for $\approx 0.1\%$ of the sites, there is a greater variance resulting in large confidence intervals.

4.3.2 Analysis of within-patient HCV Data

The exceptionally rapid rate of evolution of RNA viruses means that viral evolution within a single host can be studied over the duration of an infection. Within host evolution is the source of viral genetic diversity and therefore crucial in understanding viral evolutionary dynamics (Pybus and Rambaut, 2009). Here I apply the methods introduced in section 4.2 to

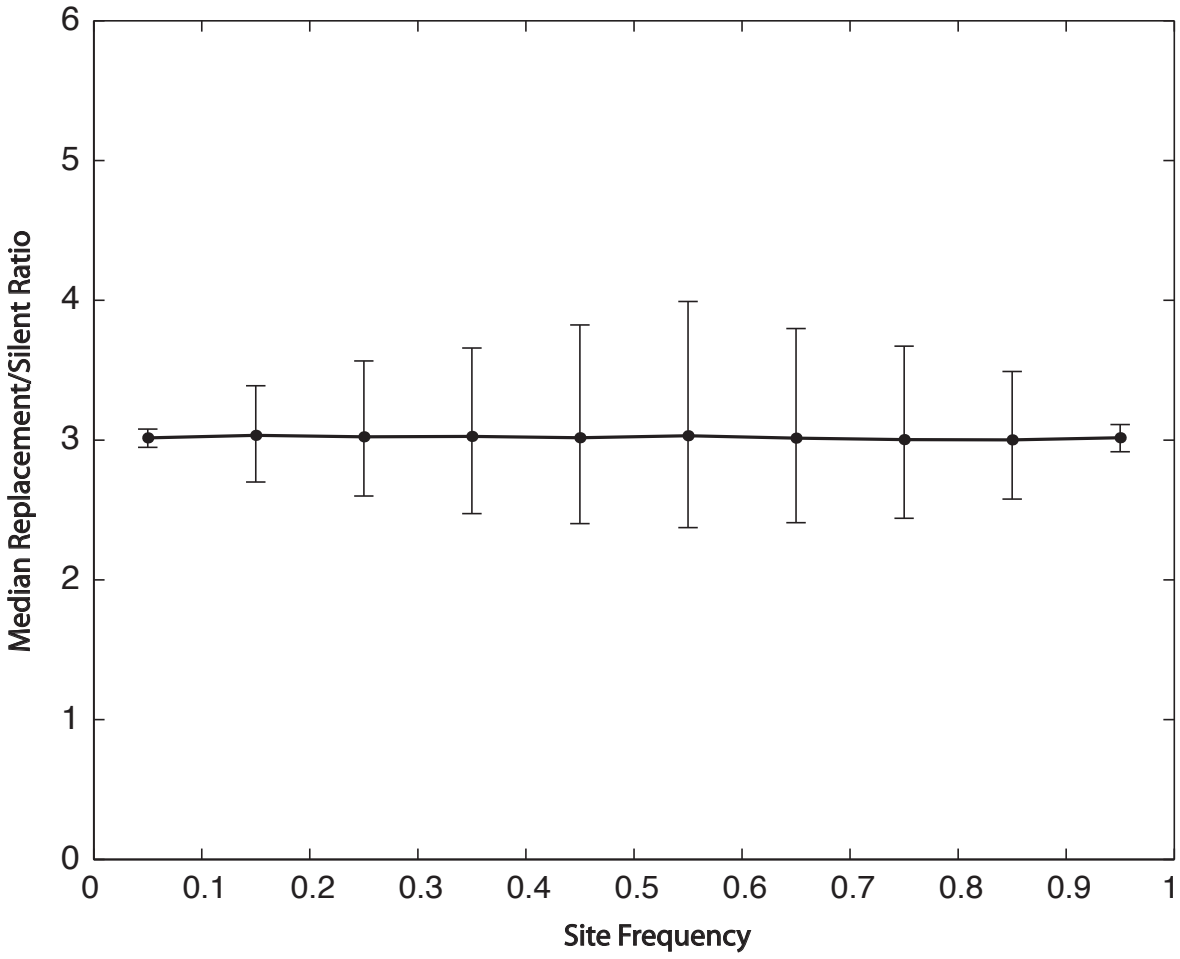


Figure 4.7: *Median replacement to silent ratio,for 500 neutral alignments with $\theta = 0.1; \tau = 2.5$ under a model of constant population size. Error bars represent the 5% and 95% percentiles. The replacement/silent ratio is evaluated using the methods described in section 4.2.2*

study the within-patient evolution of the Hepatitis C Virus (HCV). HCV is a positive-sense RNA virus with a 9.6kb genome that infects over 170 million people worldwide and is a leading global cause of liver disease and liver cancer (Memon and Memon, 2002). In most cases, HCV transmission results in long-term chronic infection that can last for decades. HCV is thought to persist in the body during chronic infection by selecting for escape mutants that evade the host's immune response. Although HCV has been shown to have a very high mutation rate (Jenkins, 2002), recombination occurs rarely (Cristina and Colina, 2006). The rate of disease progression during chronic infection varies among individuals and is not fully understood, but it has been suggested that rapid substitution of the amino acid sequence in the hypervariable region of the E1/E2 region (which is the target of antibody responses) may be one of the mechanisms of persistent HCV infection(Yamaguchi et al., 1994; Flint and McKeating, 2000; Sheridan et al., 2004a). Using sequences viral sequences from individuals infected with HCV, I apply methods developed in this chapter to investigate the mechanisms of natural selection acting on HCV virus, and test whether these mechanisms vary among patients.

4.3.2.1 HCV data

The viral sequences investigated here come from a group of individuals who all became infected in 1977 after being treated with a blood product (anti-D immunoglobulin) that was contaminated with HCV. The sequences were originally published in McAllister et al. (1998) and Duffy et al. (2002). All sequences represent the E1/E2 genome region. E1 and E2 are membrane-bound glycoproteins that form an integral part of the HCV virion envelope. In addition the E1/E2 region sequenced includes the hypervariable region (HVR1) that is thought to be the primary target of the host's antibody response and is therefore expected to be under strong immune selection.

In total, 21 main group alignments were constructed, with each alignment representing a

sample of viruses sampled from one patient at one point in time (table 4.4). Each patient was infected by one of two batches of the original contaminated blood product (McAllister et al., 1998; Duffy et al., 2002). These alignments were found to contain very little variation, which suggests that that each individual in the group was infected with the same "founding strain". The outgroup sequence consists of the consensus of sequences obtained from the contaminated batches.

4.3.2.2 Analysis

Table 4.4 provides a summary of the data used in the analysis. Sample sizes ranged from 6 to 17 sequences, each 357 nucleotides in length. To perform the analysis, 10 equally-spaced site-frequency ranges were chosen, with ranges [0-0.1], [0.1-0.2], [0.2-0.3],..., [0.9-0.1]. The number of sites in each site frequency range for each data set were calculated using equation 4.10 and the number of silent and replacement sites calculated using the methods described in section 4.2.2. All codons were used except those with gaps in any sequences (6% of sites, see table 4.4 for details of the data sets used).

The ratio of replacement to silent sites, ρ/σ, in each site-frequency range is shown in figure 4.8 on page 130. It is clear that at low site-frequencies $(0 - 0.2)$ there is little variation among data sets (ρ/σ ranges from 3.57 to 2.47) due to the contribution of the large number of invariant sites to the ρ/σ ratio. These low-frequency sites, representing invariant sites or mutations that have arisen only recently, reflect the underlying ratio of replacement to silent sites in the sequence as a whole, which is 3.43 on average (see table 4.4). As the site frequency increases from low to mid frequencies $(0.4 - 0.6)$, the average ρ/σ decreases (red line figure 4.8) for most data sets. The average ρ/σ is 1.72 in the $0.4 - 0.5$ range, and is 1.76 in the $0.5 - 0.6$ range. This decrease most likely demonstrates the removal of deleterious replacement mutations by negative selection, as such mutations are unlikely to be observed at mid frequencies. The average ρ/σ then increases as the site frequency rises from 0.6 to 1.0. It is unlikely

Table 4.4: *Details of the data sets used*

Data set (original patient code)	Year sample taken	Number of sequences (N)	Number of Sites used (k)	Number Sites Ignored	Total number of replacement sites,R	Total number of silent sites,S	Ratio R/S
R1	1994	7	357	0	271.5	77.6	3.50
R10	1994	7	282	75	214.3	61.1	3.51
R12	1994	7	352	5	265.3	76.3	3.48
R15	1994	7	354	3	266.1	78.2	3.40
R2	1994	9	329	28	249.8	73.0	3.42
R29217	1998	10	357	0	270.2	77.6	3.48
R35016	1998	10	351	6	270.0	74.9	3.61
R35028	1998	12	249	108	187.8	55.6	3.38
R35041	1998	9	354	3	270.9	75.9	3.57
R35044	1998	11	357	0	266.7	82.2	3.24
R35047	1998	11	357	0	272.5	76.8	3.55
R35060	1998	12	357	0	269.7	80.0	3.37
R35076	1998	13	339	18	252.0	76.9	3.28
R5	1994	10	321	36	239.2	73.0	3.28
R16	1994	7	354	3	268.6	77.1	3.48
R344	1994	16	297	60	221.7	66.8	3.32
R68	1994	8	339	18	254.7	74.3	3.43
R69	1994	8	327	30	245.9	72.7	3.38
R78	1994	17	339	18	254.5	76.4	3.33
R818	1994	11	342	15	262.5	73.3	3.58
R988	1994	8	339	18	257.8	72.9	3.54

that mutations segregating at these frequencies are deleterious, and therefore these high frequency replacement sites likely represent positively selected sites and fixations which are then followed by secondary mutations, giving rise to high frequency polymorphisms. However, as the site frequency increases so does variance in ρ/σ among data sets, possibly due to different levels of adaptation among patients.

In equation 4.21 a neutral range needs to be specified in order to calculate the numbers of adaptive sites in other ranges. From figure 4.8 it is clear that, on average, middle frequencies between 0.4 and 0.6 contain the least number of replacement substitutions and therefore is most suitable for the neutral range. Therefore I choose to use range 0.4-0.6 as the neutral

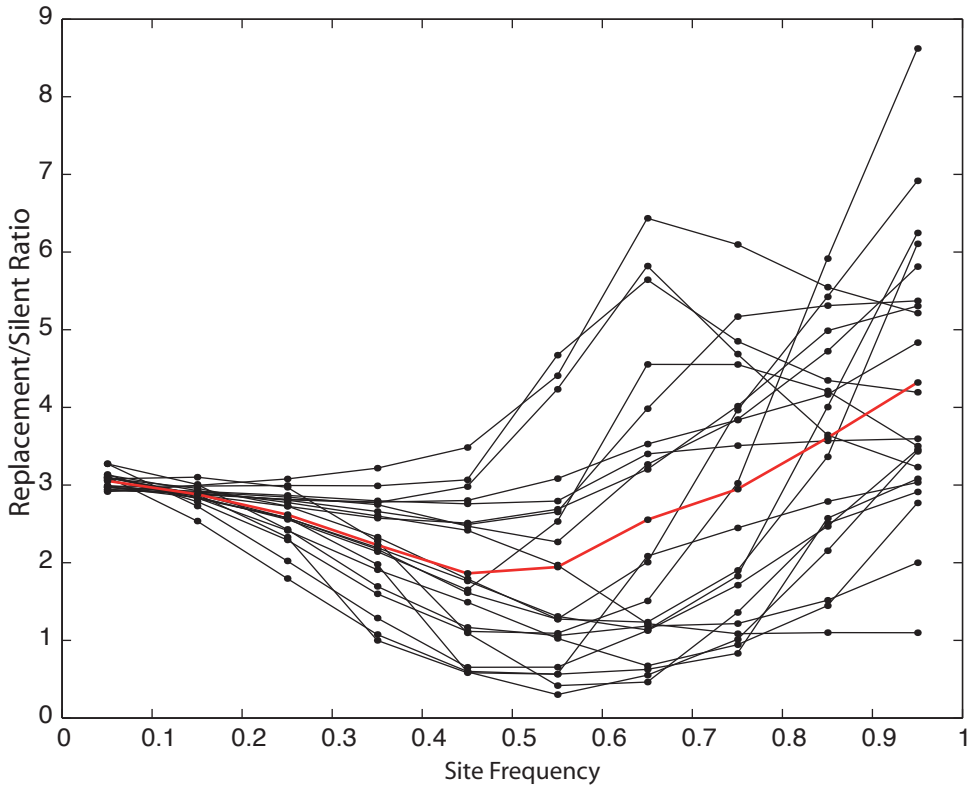


Figure 4.8: *The ratio of replacement to silent sites (ρ/σ) in each site-frequency range. Values from the same data set are joined by a thin grey lines. The mean ratio across data sets is shown as a thick red line.*

Table 4.5: *Estimated numbers of non neutral sites obtained using for an average estimate of z across all data sets ($\hat{z}$)*

Original Patient Code	Number of Low Frequency non neutral sites	Number of high Frequency non neutral sites	z
R1	109.046	10.249	2.842
R10	83.568	10.950	2.302
R12	97.185	14.516	2.443
R15	103.969	10.412	1.285
R2	109.788	4.490	0.653
R29217	102.507	20.433	0.950
R35016	110.404	17.295	4.058
R35028	79.513	2.083	1.042
R35041	109.196	14.148	2.305
R35044	112.650	1.484	0.964
R35047	104.989	18.784	1.972
R35060	111.385	9.019	0.488
R35076	92.448	11.385	0.352
R5	101.998	0.000	1.875
R16	106.424	10.255	3.519
R344	93.308	2.012	0.326
R68	96.379	12.030	1.068
R69	96.640	8.535	1.308
R78	113.666	1.203	0.475
R818	110.027	11.712	3.869
R988	100.812	13.929	2.765
		$\hat{z}$	1.755

range to estimate z.

$$z = \frac{1}{2}\left[\frac{\rho_{0.4,0.5}}{\sigma_{0.4,0.5}} + \frac{\rho_{0.5,0.6}}{\sigma_{0.5,0.6}}\right] \tag{4.22}$$

In order to obtain a more a accurate estimate, I also average z across data sets ($\hat{z}$) (table 4.5). Using this average neutral ratio ($\hat{z} = 1.755$) I estimate the number of non neutral sites in all frequency ranges using equation 4.21. Table 4.5 shows the total number of non neutral sites at 'low site-frequencies' ($[0 - 0.4]$) and the total number of non neutral sites at 'high site-frequencies' ($[0.6 - 1.0]$).

It is clear from table 4.5 that the number of non neutral sites at low-frequencies is an order of magnitude greater (mean=102) than the number of high frequency non neutral sites

(mean=10), with most low frequency non neutral sites falling within the $0 - 0.1$ site frequency range. It is likely that most of the low-frequency non neutral sites are transient deleterious or slightly deleterious mutations, although a small fraction of such sites may be advantageous mutations that have only just started to spread in the population. The variation in the number of low-frequency non neutral sites possibly reflects the different viral effective population sizes in each individual, or sampling variance among data sets.

Conversely the high-frequency non neutral sites almost certainly represent fixed or nearly-fixed advantageous mutations, therefore the variation among data sets could represent different rates of viral adaptation in different individuals. Furthermore since all members of the group were infected with effectively the same viral genome, the variation in viral adaptive fixations must reflect differences in viral-immune system dynamics among patients. Unfortunately there is not enough information on immune response or disease progression for these individuals to explore this relationship further.

These results suggest that much of the amino acid variation observed in chronic HCV viral infections represents transient deleterious polymorphism and not adaptive substitution.

4.4 Discussion

The various variants of the McDonald Kreitman test provide an alternative to computationally costly phylogenetic D_n/D_s methods. These variants make fundamental assumptions and have been applied in several different ways, as shown by the approaches taken by Smith and Eyre-Walker (2002), Williamson (2003) and the approach I took in chapter 3. However, to date, there exists no general framework for applying MK-like tests. Furthermore, the MK test has so far been used almost exclusively as a test for adaptation by positive selection, thereby ignoring the contribution of other forms of selection, such as purifying selection. Charlesworth and Eyre-Walker (2008) have shown that the presence of deleterious and slightly deleterious mutations will cause the MK test to underestimate the amount of adaptation,

limiting the effectiveness of the test. In this chapter I developed a generalised framework for reconstructing the full site frequency spectrum, and apply the fundamental logic of the MK test to develop a statistic for calculating the number of non neutral sites at any site frequency range. My approach not only accounts for sampling error but provides simultaneous measures of both deleterious mutation pressure and positive selection acting on a population. I applied this method to within patient HCV data, which showed that most of the non neutral evolution observed in chronic HCV viral infections represents transient deleterious polymorphism and not adaptive substitution. In addition I showed that variation in adaptation among patients varies greatly, reflecting differences in viral-immune system dynamics.

My generalised framework represents a new approach to the estimation and interpretation of the site frequency spectrum. It does not make the artificial distinction between variable sites and invariant sites, the number of which will depend on the sample size. In natural within-host viral populations, the actual number of virus particles ranges from tens of millions to billions. When combined with high mutation rates, such populations are very unlikely to contain any invariant sites. Rather, statistically speaking, all sites are variable, and the number of variant sites is not a direct reflection of the evolutionary process but of the sample size. In my probabilistic method no sites are decisively classed as invariant, but rather every site has a non-zero probability of being variable, a probability that is a function of the sample size. This perspective is more realistic when looking at viral populations.

The generalised framework I have developed is very flexible and allows for the inclusion of prior knowledge (a consequence of the Bayesian approach taken) which should provide better estimates of both the reconstructed site frequency spectrum and non neutral evolution. In this chapter I have not researched the effect of different priors or an informative criteria for choosing hyperparamers, but this is an interesting area for future research.

A key benefit of my generalised framework, is that it does not require a parametric model of selection in order to detect non neutral sites, and therefore is more robust that explicitly

model based approaches such as the Poisson random field model (Sawyer and Hartl, 1992). However, the method developed here is likely statistically weaker than the Poisson random field under the true model. Improvement in counting algorithms (e.g. Nei and Gojobori, 1986) or prior distributions can also readily be made without sacrifice to computational efficiency.

My method can also be used (as in chapter 3) to evaluate rates of adaptation from large data sets of serially sampled data. Such data sets are not only sometimes difficult to interpret using phylogenetic D_n/D_s methods (Kryazhimskiy and Plotkin, 2008) but also computationally demanding. The ability to investigate the action of selection very large data sets may prove very beneficial as such data becomes commonplace due to advances in next generation sequencing technology (Bains and Smith, 1988; Ronaghi et al., 1996; Brenner et al., 2000). However, my method is still based on summary statistics, and while MK type tests have good power for detecting recent or ongoing selective sweeps at different sites, phylogenetic D_n/D_s methods generally have more power to detect selection, particularly if selection at the same site is recurrent, even if the overall level of divergence is low (Zhai et al., 2009). In addition my method is likely to not be as robust as phylogenetic D_n/D_s methods to assumptions regarding mutation rates and the mutational process more generally (Andolfatto, 2001; Wall et al., 2002). In particular violations of the infinite sites model can lead to type I errors, and make results difficult to interpret.

Chapter 5

Evaluating deleterious mutation load using probabilistic site frequencies

5.1 Introduction

RNA viruses exploit all known mechanisms of genetic variation to ensure their survival. In particular, RNA virus populations are an example of an evolutionary system dominated by mutation (Domingo and Holland, 1997). However, while there is clear evidence for the high rates of adaptation required to evade host immune pressure (Domingo, 2000), this does not necessitate that the majority of mutations are advantageous. In fact, the vast majority of mutations in viral populations are either lethal, neutral, or deleterious (Sanjuan et al., 2004). Methods for estimating adaptation in nucleotide sequences, such as phylogenetic D_n/D_s methods or the methods described in Chapters 3 and 4, are unlikely to be affected by the presence of lethal and neutral mutations. This is because lethal mutations are instantly removed from populations and so are never seen, and because neutral mutations will not change the ratio of replacement to silent changes. Deleterious mutations, however, do affect estimates of adaptation and positive selection. While strongly deleterious mutations are removed from populations by purifying (negative) selection, the fate of slightly deleterious mutations is not only dependent on selection but also on the action of random genetic drift.

For MK-like tests, slightly deleterious mutations can contribute to artifactual evidence of adaptive evolution, especially in cases where the current effective population size is larger than the long-term effective population size e.g. in the case of a strong genetic bottleneck. In such cases, slightly deleterious mutations can become fixed or segregate at higher frequencies thereby elevating replacement mutation counts, resulting in underestimates of the true number of adaptive substitutions (McDonald and Kreitman, 1991a; Charlesworth and Eyre-Walker, 2008). For phylogenetic D_n/D_s methods, the presence of deleterious mutations will also lead to underestimation of the levels of positive selection, when D_n/D_s is calculated over the whole phylogeny.

Several methods have been developed to evaluate deleterious mutation pressure from sequence data. Hasegawa et al. (1998) estimated D_n/D_s ratios within and among species branches

in molecular phylogenies. Their method was based on the observation that the average age of replacement mutations increases with their selective advantage, and therefore deleterious mutations are more likely to fall on the external branches, whereas advantageous mutations are more likely to fall deeper in the genealogy. Using this method Hasegawa et al. (1998) showed a preponderance of slightly deleterious mutations throughout the mitochondrial genome of hominids. This method was subsequently applied by Sharp et al. (2001) to HIV and SIVcpz and Holmes (2003) to dengue virus. More recently Pybus et al. (2007) used this approach to create a 'deviation from neutrality statistic' (DNS) and applied the statistic to a data set of 143 RNA viruses showing that a substantial proportion of variation in RNA viral populations comprises transient deleterious mutations. Non phylogenetic methods of evaluating deleterious mutation pressure have been developed by Hughes (2005) who estimated Tajima's D (Tajima, 1989) separately for silent (D_{syn}) and replacement polymorphism (D_{non}). This method was applied to 149 population data sets from 84 species of bacteria. Hughes (2005) found a significant preponderance of negative D_{non} values, indicating the wide spread occurrence of rare replacement polymorphisms caused by deleterious mutations. The Hughes (2005) approach has more recently applied by Hughes and Hughes (2007) to 222 independent viral data sets, and highlighted a prevalence of purifying selection in RNA viruses.

In Chapter 4, I developed methods to reconstruct the site frequency spectrum in a statistical and probabilistic framework, and used the resultant distribution to evaluate the ratio of replacement to silent mutations at different site frequencies. The application of this method to HCV data indicated that the method can be used to extract information about segregating deleterious mutation from an alignment from the reduction in the replacement to silent ratio as site frequencies increase. In this chapter I extend this approach and develop a more powerful statistic specifically for evaluating the magnitude of deleterious mutational load from on sequence data. I then apply this method to the 96 virus data set created in Chapter 2 as well as to neutrally simulated data.

5.2 Methods

5.2.1 Detecting deleterious mutational pressure using probabilistic counting

The site frequency spectrum can be thought of as a continuous frequency distribution, where any point on the frequency curve represents the number of sites in a nucleotide alignment which are segregating at that frequency. In chapter 4, I introduced a probabilistic method to reconstruct the site frequency spectrum, which, for a series of site-frequency intervals, shows the number of sites expected to exist at those frequencies. Then, by assigning each site in the main alignment a probability of being silent or replacement, I developed equations that describe the site frequency distributions of silent and replacement sites. Specifically, at any frequency interval $\{u, v\}$ the number of silent sites (σ) and the number of replacement sites (ρ) are evaluated as:

$$\begin{aligned} \sigma_{u,v} &= \sum_{i=1}^{k} \mathbb{P}(s_i)\, \mathbb{P}(u < D < v | p^*, \bar{\lambda}_i) \\ \rho_{u,v} &= \sum_{i=1}^{k} (1 - \mathbb{P}(s_i)) \cdot \mathbb{P}(u < D < v | p^*, \bar{\lambda}_i) \end{aligned} \tag{5.1}$$

Under neutrality, the ratio of replacement sites to silent sites should be the same for any interval $\{u, v\}$, i.e

$$\frac{\rho_{0,0.1}}{\sigma_{0,0.1}} = \frac{\rho_{0.1,0.2}}{\sigma_{0.1,0.2}} = ... = \frac{\rho_{u,v}}{\sigma_{u,v}} \tag{5.2}$$

Under selection the relation in equation 5.2 can change. As site frequencies increase from low to mid frequencies, a successive reduction of ρ/σ suggests a removal of deleterious replacement mutations by negative selection. Conversely a successive increase in ρ/σ from low to mid frequencies suggests fluctuating or frequency dependent selection which causes an excess of replacement changes at mid frequencies. At higher frequencies changes in ρ/σ can be attributed to the presence of advantageous mutations.

Following from these trends, I use the relation in equation 5.2 as the basis for a test of the

presence and strength of deleterious mutation pressure. I choose to use site frequency range $\{0, 0.7\}$. Owing to a combination of their rarity and fitness, advantageous mutations are most likely to be seen at high frequencies (Sanjuan et al., 2004). This observation is confirmed by population genetic models used in chapter 3 section 3.3.2.3, which show that, regardless of the effective population size, the vast majority of advantageous mutations are only seen at high site frequencies. Therefore it is unlikely that a departure from equality in equation 5.2, in frequency range $\{0, 0.7\}$, is more likely a consequence of either negative, frequency dependent or fluctuating selection than a result of advantageous mutations caused by positive selection..

To quantify the strength of removal of deleterious mutations I use a first order auto regressive model.

$$\left(\frac{\rho}{\sigma}\right)_i = \beta \cdot \left(\frac{\rho}{\sigma}\right)_{i-1} + \epsilon \tag{5.3}$$

Where i is the current site frequency range and $i-1$ is the previous frequency range. ϵ is a noise term and β is the autoregresison coefficient. I evaluate coefficient β by maximising the Gaussian likelihood function

$$L\left(\beta, \sigma | \chi\right) = \prod_{i=1}^{N} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \left(\frac{\left(\left(\frac{\rho}{\sigma}\right)_i - \beta \cdot \left(\frac{\rho}{\sigma}\right)_{i-1} \right)^2}{2\sigma^2} \right) \right] \tag{5.4}$$

Under this model, values of $\beta < 1$, can be interpreted as a successive decrease in ρ/σ caused by the removal of replacement deleterious mutations from low frequencies for mid frequencies by negative selection. Values of $\beta > 1$ this suggests the effect of fluctuating or frequency dependent selection holding replacement mutations at mid frequencies.

However, when evaluating ρ/σ from data, even with probabilistic counting, some site frequencies contain very little site information and are therefore prone to statistical error. To avoid this problem, I instead compare the cumulative sum of ρ to the cumulative sum of σ. Under neutrality, the behaviour of the ratio of these cumulative sums is the same as that in equation

5.2.

$$\frac{\rho_{0,0.1}}{\sigma_{0,0.1}} = \frac{\rho_{0.0,0.2}}{\sigma_{0.0,0.2}} = ... = \frac{\rho_{0,v}}{\sigma_{0,v}} \tag{5.5}$$

Equation 5.5 is essentially the same as equation 5.2, except defines how the ratio of the *sum* of the number of replacement sites to the *sum* of the number of silent sites changes across the site frequency spectrum. Under neutrality this ratio will be constant, but if a population is undergoing purifying selection, the ρ/σ ratio will reduce with each successive addition to the sum. This approach is likely to be more robust than using equation 5.2.

5.2.2 A statistical test for the presence of deleterious mutations

I apply the above method to the data set of 96 viruses each with an appropriate outgroup sequence compiled and used in Chapter 2 section 2.3.4. To test for a significant presence of deleterious mutations in the real viral data sets, I also apply the above method on two neutrally simulated data sets, created using the PAML package EVOLVER (Yang, 1997). These neutral data sets constitute the null hypothesis of the test.

To create the simulated data sets, maximum likelihood phylogenies for each real virus alignment were estimated using GARLI (Zwickl, 2006). These phylogenies were estimated under the HKY substitution model, using empirical base frequencies and ignoring among site rate variation (to avoid over-parametrisation). Then for each of these data sets, CODEML (Yang, 1997) was used to estimate the maximum likelihood ω ($^{D_n}/_{D_s}$) ratio and κ (transition transversion rate) across the whole tree using the F3X4 codon substitution model.

Simulated data sets were created to match the real empirical virus data as closely as possible. Therefore for both simulated data sets, sequence lengths and sample sizes were set to those of the corresponding real virus alignment. For the first simulated data set, the maximum likelihood phylogeny and κ were set in the empirical values, but ω was set to 1, thereby representing strict neutrality. For the second simulated data set the maximum likelihood

phylogeny, ω and κ were all set to the empirical values, thereby representing general neutrality. If it is assumed that silent sites are neutral then, under general neutrality, when $\omega < 1$ (which is the case for all empirical data sets), ω represents the ratio of lethal to neutral mutations (lethal mutations are, by definition, not observed). For both the general and strict neutrality models, I simulate 300 replicates in EVOLVER using equilibrium codon frequencies described in the F3X4 model. These replicates represent the null distribution.

To test whether a population has an excess of deleterious mutations, I calculate the β statistic for the real data and for the simulated data sets under the general and strict neutrality model. P values are then calculated as the proportion of real β values which are more extreme than the simulated values.

5.2.3 Comparison to the DNS statistic

I compare the calculated P-values to the DNS statistic (Pybus et al., 2007), which is phylogenetic measure of the strength of purifying selection. The DNS statistic is calculated by comparing ω for internal (ω_i) and external tree branches (ω_e) to ω for the whole tree. If some polymorphic sites are deleterious or slightly deleterious then there will be an excess of replacement changes at external branches and $\omega_e > \omega$, conversely if there is strong recurrent positive selection then $\omega_i > \omega$. Using this observation Pybus et al. (2007) defined DNS statistic as:

$$\begin{aligned} &sin(45) \cdot [log(\omega_i) - log(\omega)] \quad for \quad Internal\ Branches \\ &sin(45) \cdot [log(\omega_e) - log(\omega)] \quad for \quad External\ Branches \end{aligned} \tag{5.6}$$

The DNS statistic for external branches is a phylogenetic equivalent of equation 5.4. By calculating these statistics on the same data sets I intend to evaluate the differences between my site frequency based approach and the more traditional phylogenetic approach.

5.3 Results

The methods described above were applied to 96 alignments representing viral sequence diversity from different RNA virus species. Using equation 5.4, test statistic β was calculated for the real virus data, and the data sets simulated under strict and general neutrality. Figure 5.1 on the next page shows statistic β for all data sets. For the empirical virus data, β values had the greatest range and took values from 0.998 to 0.774, for simulated data under general neutrality β ranged from 1.004 to 0.923 and under strict neutrality beta ranged from 1.006 to 0.999. Under both general and strict neutrality ρ/σ is not expected to change across the site frequency spectrum (equation 5.5) and therefore β should equal 1. As expected, the β values from simulated data under strict neutrality have small variances and have median values very close to 1. In addition under strict neutrality the 5% and 95% percentiles always contain the expected value of 1. However, for general neutrality, while the majority of data sets have β values close to 1, some data sets show large deviations from the expected value of 1 (general neutrality data sets are displayed in figure 5.1 with green crosses).

The deviation from the expected value of 1 in the general neutrality data sets is a systematic error caused by a violation of the infinite sites model. Some data sets have phylogenetic shapes that include long terminal branches, which are likely to experience repeat mutations. When combined with low ω values (< 0.2), repeat mutations can introduce systematic error (ω values are positively correlated with the median β values from general neutrality data sets). Low ω values indicate that replacement changes are very rare in comparison to silent changes, and when combined with long phylogeny terminal branches, there is likely to be a a systematic underestimation of the number of low frequency silent sites compared to replacement sites, thereby reducing β values. This systematic error will also be present in estimation of β in the empirical data, but given that the general neutrality simulated alignments are generated down the same phylogeny as the real data and under the same ω ratio, differences between the empirical data and the general neutrality simulations must be due to a violation of neutrality,

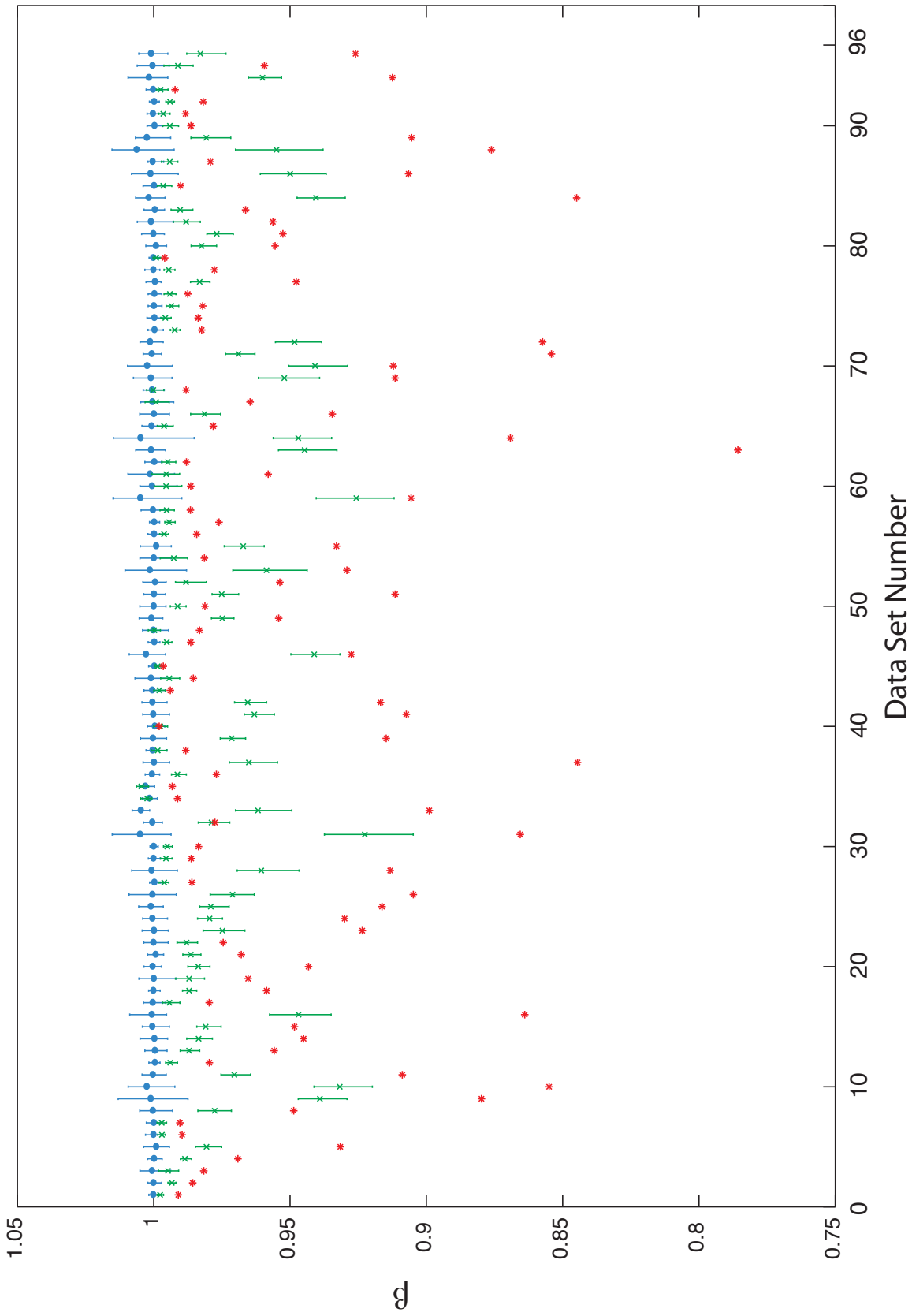


Figure 5.1: *Each number on the x axis is a separate virus data set (refer to table 5.1 for the corresponding virus names). The y axis represents β values evaluated using equation 5.4, from empirical data (red), neutral data sets simulated under the strict neutrality model (blue), and neutral data sets simulated under the general neutrality model (green). For the simulated data the error bars represent the complete distribution of β values for 300 replicates and the points represent the median value of this distribution.*

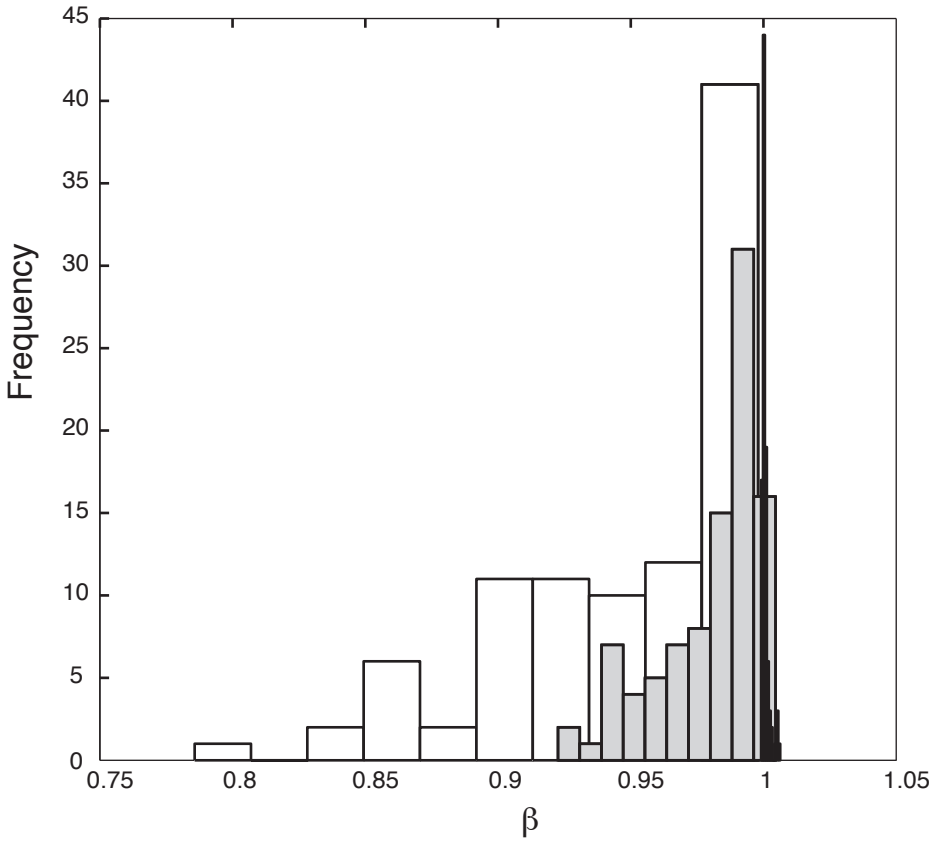


Figure 5.2: *Histogram of β values for the empirical data sets (white) and median β values for alignments simulated under the general neutrality model (grey) and the strict neutrality model (black)*

rather than to systematic bias caused by repeat mutations.

Figure 5.2 shows a histogram of β values for the empirical and simulated data. There is a clear difference between the mean value of β for the empirical virus data sets (0.948) and the mean value of β for the simulated data sets (0.980 for general neutrality and 1 for strict neutrality). These differences indicate that real viral populations have an excess of low frequency replacement changes caused by deleterious or slightly deleterious mutations. The β values for the empirical data sets are also positively correlated (P-value < 0.001) with ω estimated for external branches of the phylogenetic trees. This correlation demonstrates that detection of the magnitude of the deleterious mutation load using my site frequency method is comparable with that estimated phylogenetically using ω on external branches.

Table 5.1 shows the P-values and the DNS statistic for all 96 virus data sets. 78% of data sets

have DNS<0 for external branches, and correspondingly 95% of data sets have P-values <0.05 for my method. It is expected that a substantial proportion of amino acid variation observed in RNA virus populations comprises transient deleterious mutations that are later purged by purifying selection. Therefore these results suggest that my novel approach possible has more power in detecting purifying selection than the phylogenetic $^{D_n}/_{D_s}$ methods (given that we expect all viral data sets to experience some degree of deleterious mutational load). However, 3% of data sets have DNS<0 but P-values >0.05 (data sets 6,40 and 46), which suggests that my method likely experiences type I error, possible introduced by violations of the infinite sites assumption.

Table 5.1: *Table of ω_e, DNS for external branches (Equation 5.6) and P-values of my statistic (bold values show data sets with P-value > 0.05) .*

Name	Data Set Number	ω_e	DNS External Branches	P-values
Australian bat lyssavirus (G)	**1**	0.027	-0.160	<0.003
Acute bee paralysis virus (C)	**2**	0.022	-0.104	<0.003
Akabane virus (NP)	**3**	0.035	-0.143	<0.003
Avian influenza A, serotype H5N1 (NP)	**4**	0.054	-0.084	<0.003
Avian influenza A, serotype H7N1 (HA)	**5**	0.051	-0.196	<0.003
Avian pneumovirus (N)	**6**	0.051	-0.442	**0.12**
Barley yellow mosaic virus (CP)	**7**	0.108	-0.094	<0.003
Bean yellow mosaic virus (CP)	**8**	0.067	-0.261	<0.003
Bluetongue virus (VP7)	**9**	0.003	-0.346	<0.003
Bovine rotavirus (VP7)	**10**	0.024	-0.310	<0.003
Crimean-Congo haemorrhagic fever virus (NP)	**11**	0.029	-0.082	<0.003
Canine distemper virus (H)	**12**	0.215	-0.115	<0.003
Chikungunga virus (E1)	**13**	0.026	-0.397	<0.003
Classical swine fever virus (E2)	**14**	0.130	-0.120	<0.003
Clover yellow vein virus (CP)	**15**	0.028	-0.267	<0.003
Coxsackievirus B4 (VP1)	**16**	0.006	-0.363	<0.003
Curcurbit yellow stunting disease virus (CP)	**17**	0.064	-0.057	<0.003
Dengue virus, serotype 1 (E)	**18**	0.050	-0.164	<0.003
Dengue virus, serotype 1 (CM)	**19**	0.063	-0.150	<0.003
Dengue virus, serotype 2 (E)	**20**	0.035	-0.350	<0.003
Dengue virus, serotype 3 (E)	**21**	0.068	-0.041	<0.003
Dengue virus, serotype 4 (E)	**22**	0.043	-0.217	<0.003
Dobrava virus (N)	**23**	0.024	0.021	<0.003
Eastern equine encephalitis virus (C)	**24**	0.007	-0.384	<0.003
Eastern equine encephalitis virus (E1)	**25**	0.080	0.106	<0.003
Enterovirus 71 (VP1)	**26**	0.015	-0.349	<0.003
Equine influenza, serotype H3N8 (HA)	**27**	0.304	-0.051	<0.003
Feline immunodeficiency virus (Gag)	**28**	0.014	-0.148	<0.003
Human influenza A virus, serotype H3N2 (HA)	**29**	0.429	0.052	<0.003
Human influenza A virus, serotype H3N2 (NP)	**30**	0.131	-0.097	<0.003
Garlic latent virus (CP)	**31**	0.012	-0.081	<0.003
Hepatitis C virus 1b (C)	**32**	0.157	0.022	**0.1033**
Hepatitis C virus 1b (E1E2)	**33**	0.267	0.087	<0.003
HIV type 1, subtype B (Env-Gap)	**34**	0.879	0.082	<0.003
HIV type 1, subtype B (Env)	**35**	1.079	0.109	<0.003
HIV type 1, subtype B (Gag)	**36**	0.491	0.175	<0.003
Human polio virus type 2 (VP)	**37**	0.035	-0.067	<0.003
Human respiratory syncytial virus A (G)	**38**	0.453	-0.044	<0.003
Human respiratory syncytial virus A (N)	**39**	0.012	-0.257	<0.003

Name	Data Set Number	ω_e	DNS External Branches	P-values
Human respiratory syncytial virus B (G)	**40**	0.377	-0.017	**0.62**
Hantaan virus (G1)	**41**	0.028	-0.069	<0.003
Hantaan virus (N)	**42**	0.010	-0.232	<0.003
Viral hemorrhagic septicaemia virus (GP)	**43**	0.309	0.077	0.01
Viral hemorrhagic septicaemia virus (N)	**44**	0.242	0.238	0.09
Highlands J virus (E1)	**45**	0.051	-0.315	0.02
Human astrovirus (C)	**46**	0.042	-0.021	**0.0567**
Human parainfluenza virus type 1 (HN)	**47**	0.131	-0.113	<0.003
Human parainfluenza virus type 3 (HN)	**48**	0.069	-0.280	<0.003
Infectious pancreatic necrosis virus (VP2)	**49**	0.068	-0.095	<0.003
Japanese encephalitis virus (CP)	**50**	0.039	-0.226	<0.003
Japanese encephalitis virus (E)	**51**	0.018	-0.460	<0.003
Junin virus (NP)	**52**	0.060	0.062	<0.003
Leek yellow stripe virus (CP)	**53**	0.127	0.031	<0.003
Lettuce mosaic virus (CP)	**54**	0.186	-0.033	<0.003
Maize dwarf mosaic virus (CP)	**55**	0.048	-0.025	<0.003
Measles virus (HA)	**56**	0.152	-0.109	<0.003
Measles virus (N)	**57**	0.140	-0.089	<0.003
Mumps virus (NP)	**58**	0.021	-0.392	<0.003
Onion yellow dwarf virus (CP)	**59**	0.056	0.058	0.0433
Oropouche virus (NP)	**60**	0.031	-0.528	<0.003
Pea seed-borne mosaic virus (CP)	**61**	0.039	-0.248	<0.003
Peanut stripe virus (CP)	**62**	0.101	-0.077	<0.003
Polio virus, serotype 1 (VP1)	**63**	0.016	-0.308	<0.003
Porcine rotavirus (VP7)	**64**	0.115	0.061	<0.003
Potato virus A (CP)	**65**	0.171	0.057	<0.003
Potato virus S (CP)	**66**	0.035	-0.167	<0.003
Potato virus X (CP)	**67**	0.067	-0.120	<0.003
Prunus necrotic ringspot virus (CP)	**68**	0.252	-0.145	<0.003
Puumala virus (G2)	**69**	0.018	-0.095	<0.003
Puumala virus (N)	**70**	0.015	-0.162	<0.003
Rabies virus (G)	**71**	0.088	-0.074	<0.003
Rabies virus (N)	**72**	0.031	-0.182	<0.003
Rice black streaked dwarf virus (CP)	**73**	0.024	-0.359	<0.003
Ross River Virus (E2)	**74**	0.057	-0.151	<0.003
Rotavirus A (VP7)	**75**	0.098	-0.211	<0.003
Rotavirus C (VP7)	**76**	0.051	-0.136	<0.003
St. Louis encephalitis virus (E)	**77**	0.028	-0.162	<0.003
Sendai virus (NP)	**78**	0.609	0.680	<0.003
Simian foamy virus (Env)	**79**	0.290	0.122	0.0067

Name	Data Set Number	ω_e	DNS External Branches	P-values
Soybean mosaic virus (CP)	**80**	0.041	-0.018	<0.003
Sugarcane mosaic virus (CP)	**81**	0.023	-0.282	<0.003
Sweet potato feathery mottle virus (CP)	**82**	0.074	-0.168	<0.003
Swine influenza virus, serotype H3N2 (HA)	**83**	0.345	0.011	<0.003
Tick-borne encephalitis virus (E)	**84**	0.030	-0.071	<0.003
Tomato spotted wilt virus (N)	**85**	0.121	-0.066	0.0033
Tula virus (NP)	**86**	0.011	-0.144	<0.003
Turnip mosaic virus (CP)	**87**	0.082	-0.145	<0.003
Venezuelan equine encephalitis virus (C)	**88**	0.040	0.031	<0.003
Venezuelan equine encephalitis virus (E)	**89**	0.074	0.001	<0.003
Western equine encephalitis virus (E1)	**90**	0.082	-0.186	<0.003
West Nile virus (E)	**91**	0.027	-0.440	<0.003
Wheat streak mosaic virus (CP)	**92**	0.148	0.110	<0.003
Wheat yellow mosaic virus (CP)	**93**	0.097	-0.347	0.03
Yellow fever virus (E)	**94**	0.010	-0.359	<0.003
Yam mosaic virus (CP)	**95**	0.166	-0.027	<0.003
Zucchini yellow mosaic virus (CP)	**96**	0.065	-0.108	<0.003

5.4 Discussion

As a result of the error-prone nature of their viral RNA polymerases, RNA viruses are known to have very high mutation rates per site per replication. Given that most of these mutations in coding regions are either deleterious or neutral (Sanjuan et al., 2004), it is important to account for the effect of purifying selection when looking at how natural selection shapes RNA virus genetic diversity. My results in this chapter confirm those obtained using different approaches (Hughes and Hughes, 2007; Pybus et al., 2007): RNA virus populations experience a significant excess of replacement substitutions at low site frequencies caused by deleterious or slightly deleterious changes. While the magnitude of this excess varies greatly depending on the species and the gene (figure 5.1), these results reveal that deleterious mutation pressure is major component of the genetic variation observed in RNA virus populations. Therefore it is likely that the common approaches of evaluating the dynamics of selective pressure acting on a population, such as a global ω ratio or P-Values from the McDonald Kreitman test will give incomplete estimates of the true selection dynamics in a population (Charlesworth and

Eyre-Walker, 2008).

As discussed in section 5.1, the methods introduced by Hasegawa et al. (1998) and Pybus et al. (2007) provide estimates of both adaptation and deleterious mutation pressure in a population, but are subject to the limitations of phylogenetic D_n/D_s methods. As an alternative, I have used the probabilistic counting method introduced in chapter 4 as the basis of a test for the presence of a deleterious mutation load in a population. An extensive study conducted by Zhai et al. (2009) found that the MK tests has substantial power to detect purifying selection, and suggest that an important role of the MK test in population genetics might be to test for negative selection. I have shown that my method appears to have more power than phylogenetic D_n/D_s methods in detecting deleterious mutation pressure at least on the empirical virus data sets investigated here. However, my method is likely to have a higher type I error profile and overestimate the magnitude of this pressure. Type I errors originate from two possible causes (i) advantageous mutations that have only started to spread in the population will also contribute to an excess of replacement changes at low frequencies, thereby overestimating the number changes caused by deleterious mutations, and (ii) as discussed above, my method's use of simple counting techniques is influenced by violations of the infinite sites assumption, a problem that is exacerbated by the extreme phylogenetic tree shapes often seen in virus populations. This violation of the infinite sites assumption will overestimate the amount of replacement changes in populations experiencing strong purifying selection (which is the case for most RNA virus populations). Despite this drawbacks, the method I have proposed here is an example of how the generalised framework I developed in chapter 4 can be applied to evaluate deleterious mutation pressure. The flexibility of the general framework I have developed allows for more complex counting method, such as those proposed by Nei and Gojobori (1986) and Egea et al. (2008). In addition, substitution models (e.g. Jukes and Cantor, 1969) can be introduced to correct the number of silent sites or replacement sites for multiple hits (Tamura et al., 2007; Rozas and Rozas, 1999). These additions, which can be easily added without sacrifice to computational tractability, may help to increase robustness and reduce

type I error.

Another non phylogenetic method is that introduced by Hughes (2005), who estimated Tajima's D (Tajima, 1989) separately for silent (D_{syn}) and replacement polymorphism (D_{non}). Using this method negative D_{non} values indicate the occurrence of rare replacement polymorphisms caused by deleterious mutations. However, while Hughes's (2005) approach will have the same computational benefits over D_n/D_s methods that my method offers, his approach will not only also suffer from violations of the infinite sites assumption, but also be confounded by the influence of population history, a problem inherent to the Tajima's D statistic (Chapter 2). My method uses a ratio based approach where ρ/σ is a sufficient statistic that is ρ and σ are equally affected by demography and genetic drift (Nielsen, 2001), and as a result will have a significant advantage over Hughes's (2005) approach.

Chapter 6

Conclusions and final thoughts

Quantifying the relative contributions of natural selection, genetic drift and mutation in shaping viral genetic variation is a complex and multifaceted problem. Much of the evolutionary literature focuses on positive selection and its association with adaptation and with the evolution of new forms or functions. However, negative or purifying selection may also be of great interest to help detect regions or residues of functional importance. Aside from the theoretical interest in understanding viral evolutionary dynamics, sites and regions under selection are of practical importance as their identification can assist in the development of vaccines and in anti-viral drug design (Sheridan et al., 2004b; Gaschen et al., 2002).

Various tests of neutral evolution have been developed and applied to empirical data, and although many are able to reject strict neutrality, they rarely provide decisive evidence of selection under all circumstances. The majority of previous research concerning viruses has been conducted using phylogenetic D_n/D_s methods. These methods have been shown to have good statistical power to detect positive and negative selection, both across whole genes and at specific codons under selection (Zhai et al., 2009). However, phylogenetic D_n/D_s methods can be computationally expensive when sample sizes are large. Additionally, in studies of viral infection where sequences are obtained from an infected individual serially through time, the correct interpretation of D_n/D_s becomes difficult to discern (Kryazhimskiy and Plotkin, 2008; Sheridan et al., 2004b). The alternatives presented in this thesis make use of the mutational site-frequency spectrum as a means for detecting the molecular 'footprint' of natural selection.

6.1 Power, robustness and error

In chapter 2, I investigated the type I error of two site-frequency based methods for detecting selection: the MK test (McDonald and Kreitman, 1991a) and Taijma's D (Tajima, 1989). I showed that these methods have reasonably low type I error when applied to virus populations with constant population sizes. However, this demographic scenario is rarely seen

in viral populations; instead, it is likely that most viral populations will have experienced a complex history of population size change (during an epidemic, or directly following transmission to a new host) and will have likely been subject to some form of population structure. Under a more realistic scenario of exponentially-growing population size, Tajima's D incurs unacceptably high levels of type I error. This error stems from the inability of the test to distinguish between the act of selection and demographic processes, such as population growth, subdivision or bottlenecks. The MK test, however, is largely unaffected by the demographic history of a population, because the test does not look at the frequency of mutations in a population, but rather the ratio of the frequency of silent mutations to replacement mutations (Eyre-Walker, 2002). Changes in population demography will affect the ratio equally, but the effect of selection will not, thereby making the test robust to changes in population demography.

In chapter 2 I noted that type I errors still occur when using the MK test on highly variable populations, or when genetically distant outgroups are used. I showed that this error is a systematic error caused by the underestimation of the numbers of silent and replacement sites, arising from the 'infinite sites' assumption. To correct this problem I also introduced a new proportional site-counting method that provided better estimates of the numbers of silent sites and replacement sites, thereby reducing type I error. Alternative counting methods such as those developed by Nei and Gojobori (1986) use substitution models to correct for multiple mutations at the same site (e.g. Jukes and Cantor, 1969). However, further work is needed to investigate whether these approaches are better than the proportional site counting method I implemented here.

I applied both the MK test and Tajima's D to empirical RNA virus data representing a broad range of RNA virus populations. I demonstrated that on these empirical data the MK test has more power to reject the null hypothesis of neutral evolution than Tajima's D. However, there is still a need to fully investigate the power and robustness of these tests when they

are applied to non neutrally evolving sequences. Although neutral evolution can be easily simulated using the coalescent, the simulation of non neutral evolution requires much more complex individual-based simulation approaches. In the following discussion of statistical power and robustness I do not consider Tajima's D test any further, as it is not directly relevant to the remainder of the thesis.

Zhai et al. (2009) investigated the differences in statistical power between the MK test and phylogenetic D_n/D_s methods under different strengths of recurrent negative and positive selection. Zhai et al. (2009) found that when attempting to detect recurrent positive selection, the MK test has significantly less power than phylogenetic D_n/D_s methods, but has substantial power to detect recurrent negative selection. However, Zhai et al. (2009) failed to conduct simulations of populations undergoing infrequent single selective sweeps at different sites. In such cases, the MK test is likely to have considerably more power for detecting selection than phylogenetic D_n/D_s methods. Phylogenetic D_n/D_s methods are unable to detect single historical selective sweeps as they rely on recurrent replacement changes at a single locus to provide evidence of selection, while the MK test quantifies the number of adaptive fixations and maybe more sensitive to selective sweeps that occur in codons whose D_n/D_s ratio is considerably less than one.

With regards to robustness, phylogenetic D_n/D_s methods are likely to be sensitive to assumptions regarding evolutionary independence or free recombination among sites. These assumptions have been made in previous studies using D_n/D_s (Nielsen and Yang, 2003), and are expected to be reasonable when for mutation rates are small and selection pressures are weak (Kryazhimskiy and Plotkin, 2008). However, for higher mutation rates and populations undergoing strong selection (conditions common in RNA viruses) the effects of linkage of linkage on D_n/D_s methods are difficult to quantify and free recombination may be an unrealistic assumption in such settings. In comparison, the MK test allows for robust inference in presence of recombination (Sawyer and Hartl, 1992) or non-equilibrium demography (Andol-

fatto, 2008; Nielsen, 2001) as the test is independent of the shape of the underlying population genealogy.

There is currently no single selection-detection method that can provide a complete picture of both the strength and directionality of selection from sequence data, without making high restrictive assumptions about the population under study. Critically, current methods tell very little about why adaptation has taken place or the nature of the fitness benefit to the organism. Therefore when choosing methods for analysing selective forces, it is important to consider the statistical issues of power, error and robustness in combination with functional experiments that can provide a comprehensive biological explanation of evolution and adaptation.

6.2 Rates of adaptation

In chapter 3 I extended and generalised the methods developed by Smith and Eyre-Walker (2002) and Williamson (2003), which are based on the standard MK test. These methods have an advantage over phylogenetic D_n/D_s methods in that they can be used to evaluate rates of adaptation through time, rather than just the presence or absence of selection in a gene or at a codon. In addition, MK-based methods are computationally efficient and can be applied to very large serially-sampled data sets. Previously, applications of MK-based tests to serially-sampled data (i.e. Williamson (2003)) were based on evolutionary assumptions made *a priori*, specifically, the assumption that all polymorphisms with observed population frequencies below 50% are selectively neutral. In contrast, in chapter 3 I combined population genetic theory and empirical viral data from vesicular stomatitis virus (VSV) in order to objectively identify the most neutral class of sites. This approach makes the assumption that the fitness distribution of VSV is similar to that of influenza virus, or indeed any over viral species. However, the Sanjuan et al. (2004) study of experimental VSV evolution considers mutations that arise in only a single cell type, where as influenza viruses in natural populations replicate in a variety of cell types and are transmitted among biologically-variable hosts. It therefore

seems likely the distribution of mutational fitness will differ between experimental and natural populations, with possibly a larger proportion of deleterious mutations in the latter. This difference in mutational fitness distributions may affect the assumptions I used in the generalised method for estimating rates of adaptive fixation. Indeed the choice of fitness distribution is likely to be an important factor populations with large effective population sizes as these populations are largely unaffected by drift and dominated by selection. Future use of this approach should consider more carefully the robustness of conclusions to changes in the underlying distribution of mutation fitness.

In chapter 3 I applied a generalised form of the methods developed by Smith and Eyre-Walker (2002) and Williamson (2003) to human influenza A genomes belonging to subtypes H1N1 and H3N2. In my analysis I estimated the number of adaptive fixations at different points in time for each gene in the influenza genome, and then evaluated rates of adaptation through time using common statistical techniques (linear regression and parametric bootstrapping). My results concurred with those already conducted on adaptation in the influenza HA gene, thus suggesting that my new methodology is reliable. My results also highlighted the important contribution of envelope protein NA and internal proteins NP and NS1 to influenza virus evolution. Finally my approach provided the first direct test of the hypothesis by Rambaut et al. (2008) that immune selection on H1N1 is weaker compared to H3N2. My results supported this hypothesis and suggested that the HA gene might be the driving force causing higher rates of adaptation in H3N2.

The only notable previous study of selection across the whole influenza genome was conducted by Suzuki (2006) using phylogenetic D_n/D_s methods on 259 whole influenza genomes from subtype H3N2 only. That study detected adaptation mainly in the envelope proteins HA and NA. In contrast my study was conducted using a much larger sample size of 723 H1N1 genomes and 1520 H3N3 genomes, and found evidence of adaptation in the internal proteins NS1 and NP. The inability of the phylogenetic D_n/D_s approach to detect noteworthy

selection in the internal proteins could stem from a variety of reasons. First, as discussed above, my method is more likely to detect recent single or ongoing selective sweeps than recurrent selection at a single locus, providing a different perspective on natural selection in the influenza genome. Secondly, most phylogenetic D_n/D_s methods, for example those implemented in the CODEML software package (Yang, 1997) assume that the topology of the phylogeny is known without error. This introduces the potential for bias if the assumed phylogeny is incorrect. The correct statistical approach would be to include phylogenetic error when evaluating codon substitution models (e.g. Ren et al., 2005; Suchard and Rambaut, 2009). However, this approach will be exceptionally time-consuming for the sample size used in my study. Finally, although my method is probably less statistically powerful than phylogenetic D_n/D_s approaches (Zhai et al., 2009), I was able to investigate a data set in my study that was six times larger than that used by Suzuki (2006). My method therefore represents an alternative compromise between computational efficiency and statistical power.

6.3 Generalised framework

In this section I will discuss the generalised framework for estimating site-frequencies and site-frequency ratios that was introduced in chapter 4. The methods that I developed and tested in chapters 2 and 3 are essentially heuristic techniques. That is, they are algorithmic and rule-based in nature and are not rooted in a rigorous statistical framework. Throughout the history of population genetics and phylogenetics there have been several occasions when methods have been initially introduced as simple heuristics, which are then subsequently developed and placed into a more formal statistical framework. For example, phylogenetic tree construction was initially performed using heuristic methods such as UPGMA (Sneath and Sokal, 1973) and neighbour-joining (Saitou and Nei, 1987), but later progressed to tree estimation using rigorous statistical techniques such as maximum likelihood and Bayesian inference (e.g. Huelsenbeck and Ronquist, 2001; Swofford, 2002; Zwickl, 2006). Methods

for detecting selection using D_n/D_s methods are another example: such methods started out as simple heuristic counting approaches (e.g. Nei and Gojobori, 1986) which were later developed into a more complex likelihood framework using codon substitution models (e.g. Yang et al., 2000).

For methods that investigate nucleotide site-frequencies, one formal statistical framework already exists: the Poisson Random Field (PRF) model (Sawyer and Hartl, 1992). However the PRF model typically makes a number of strong assumptions about the population under study, for example: (i) an infinite sites model of evolution (ii) the population is constant in size, (iii) there is free recombination among sites and (iv) selective pressure is constant across sites. When populations do not conform to these assumptions estimates from the PRF can be misleading (Bustamante et al., 2001). Efforts have been taken to relax some of these assumptions within the PRF framework, these include allowing for population subdivision (Wakeley, 2003), changing population size (Williamson et al., 2005), linkage between sites (Zhu and Bustamante, 2005) and finite sites models (Desai and Plotkin, 2008), but no general framework incorporating all these modifications exist. In addition, each extension of the basic PRF framework requires the new process (e.g. population size change, population structure) to be modelled explicitly in a parametric form, which could lead to error if the model does not adequately reflect the population under investigation.

It is with these limitations in mind that in chapter 4 I introduced an alternative approach to reconstructing the site frequency spectrum within a Bayesian framework. Using this framework I developed a generalised form of the Smith and Eyre-Walker (2002) and Williamson (2003) methods, which is capable of calculating the number of non neutral sites at any range in the site frequency spectrum. This method is different from the PRF framework in that it does not require any parametric models of selection (or other population genetic processes) to detect non neutral evolution, making it potentially more robust but, in most cases, less statistically powerful. It is my hope that the increasing availability of large genomic data sets in the future

will compensate for this loss of power. The robustness of my method is inherited from the MK test, and arises from the use of ratios of silent-to-replacement sites, rather than the absolute number of these sites. The Bayesian approach taken reconstructs the site frequency spectrum in a probabilistic manner, and therefore incorporates sampling error. This is particularly relevant in alignments with <10 samples, where there is considerably uncertainty in inferring the true site frequency spectrum, or the distribution of silent and replacement sites. The application of my new approach to within patient HCV data to calculate the number of non neutral substitutions at high and low site frequency ranges demonstrates that, but using my approach, it is possible to infer both adaptation and deleterious mutation load on the same data set. My results showed that much of the amino acid variation observed in chronic HCV viral infections represents transient deleterious polymorphism and not adaptive substitution.

A key feature of my framework is its flexibility: improvements in site-counting methods and inclusion of knowledge in prior distributions can be incorporated. As an example of this malleability I developed a new method in chapter 5 to quantify the deleterious mutation load in a population. This method uses the estimated site frequency spectrum to test the null hypothesis that, under neutrality, the ratio of replacement-to-silent sites should be the same for any frequency range in the site-frequency spectrum. I tested this method on empirical data sets of 96 RNA virus genes and rejected the null hypothesis for 92 data sets. I also directly compared the performance of my deleterious mutation statistic method with an analogous phylogenetic D_n/D_s approach, which looked at the difference in D_n/D_s ratios among internal and external phylogeny branches (Pybus et al., 2007). In the analysis of 96 RNA virus analysis, my method appeared to have at least as much power to detect deleterious mutation load as the equivalent phylogenetic approach. Chapter 5 therefore illustrates how new methods can be readily derived from the general framework that I introduced in Chapter 4.

6.4 Future directions

To conclude, I will highlight some of the potential extensions and applications of the framework developed in this thesis. One current advantage of phylogenetic D_n/D_s methods (e.g. those implemented in the CODEML (Yang, 1997) and HyPHY (Pond et al., 2005) computer packages) is that they are able to identify specific codons that have undergone selection. At present my framework cannot replicate this functionality. However, as part of chapter 2, I split an alignment of influenza HA genes into two subdomains, HA1 and HA2 (figure 3.9 on page 96), and by doing so showed that the vast majority of adaptive fixation occurred in the HA1 region. In this case, the HA gene was partitioned using pre-existing knowledge about the functional differences between the HA1 and HA2 regions. However, in instances where the sites within the gene that are under selection are not known, it should be theoretically possible to investigate all possible site partitions, and thereby identify the partition that contains the most adaptive fixations. A combinatorial or heuristic optimisation algorithm could be developed with an appropriate optimality criterion to search among all possible partitions. Such a 'brute force' computational approach is practical in this case because the evaluation of each partition is very quick. An important extension of my framework may therefore be the development of methods which, given a set of gene or genome sequence alignments, are capable of detecting gene regions that are under significantly different selection forces.

Second-generation (or 'nextgen') sequencing techniques based on pyrosequencing methods (Margulies et al., 2005) are already established and are being rapidly adopted by researchers worldwide. This new technology has already been applied to the sequencing of virus genomes, e.g. Wang et al. (2007) for HIV-1, Bright et al. (2006) for influenza virus and Adelson et al. (2005) for herpes simplex virus types 1 and 2. New sequencing methods have the ability to produce whole genome data at unprecedented rates. All likelihood-based phylogenetic methods of evolutionary analysis will become increasingly difficult to use once sample sizes increase into the thousands. Searching for the maximum-likelihood tree has not yet been shown to be NP-

complete (non-deterministic polynomial time) (Felsenstein, 2003), but nevertheless practical searches among tree topologies with sample sizes of the order of thousands is extremely difficult. Advances in computational analysis, such as the innovative use of algorithms utilising graphics processing units (GPUs) that contain large numbers of processing cores, can enable more efficient computation of codon substitution models within a likelihood or Bayesian MCMC framework (Suchard and Rambaut, 2009). New methods utilising many-core algorithms can also benefit from the assiduous increase in computational power still growing exponentially every 18 months (Moore, 1998). However, even these computational advances are unlikely to keep up with the increases of sequence data in online databases such as Genbank (Benson et al., 2008). I therefore hope that the methods I have developed and used in this thesis will prove useful in the context of these advances in sequencing technology. In addition to being computationally efficient, I hope that the methods I have introduced will be used in conjunction with phylogenetic D_n/D_s methods to provide a richer view of the selective forces acting on viral populations.

Chapter 7

References

Adelson, M. E., Feola, M., Trama, J., Tilton, R. C., and Mordechai, E. (2005). Simultaneous detection of herpes simplex virus types 1 and 2 by real-time PCR and pyrosequencing. *Journal of Clinical Virology*, 33:25–34.

Ahmed, R., Oldstone, M. B. A., and Palese, P. (2007). Protective immunity and susceptibility to infectious diseases: lessons from the 1918 influenza pandemic. *Nat Immunol*, 8:1188–1193.

Andolfatto, P. (2001). Adaptive hitchhiking effects on genome variability. *Current Opinion in Genetics & Development*, 11:635–641.

Andolfatto, P. (2005). Adaptive evolution of non-coding dna in drosophila. *Nature*, 437:1149–1152.

Andolfatto, P. (2008). Controlling type-I error of the McDonald-Kreitman test in genomewide scans for selection on noncoding DNA. *Genetics*, 180:1767.

Baigent, S. J. and McCauley, J. W. (2003). Influenza type a in humans, mammals and birds: determinants of virus virulence, host-range and interspecies transmission. *Bioessays*, 25:657–671.

Bains, W. and Smith, G. C. (1988). A novel method for nucleic acid sequence determination. *Journal of Theoretical Biology*, 135:303.

Belshaw, R., Gardner, A., Rambaut, A., and Pybus, O. G. (2008). Pacing a small cage: mutation and RNA viruses. *Trends in Ecology & Evolution*, 23:188–193.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2008). GenBank. *Nucleic Acids Research*, 36:D25–30.

Berkhoff, E. G. M., de Wit, E., Geelhoed-Mieras, M. M., Boon, A. C. M., Symons, J., Fouchier, R. A. M., Osterhaus, A. D. M. E., and Rimmelzwaan, G. F. (2005). Functional constraints of influenza a virus epitopes limit escape from cytotoxic T lymphocytes. *Journal of Virology*, 79:11239–11246.

Beveridge, W. I. (1991). The chronicle of influenza epidemics. *Hist Philos Life Sci*, 13:223–34.

Brenner, S., Williams, S. R., Vermaas, E. H., Storck, T., Moon, K., McCollum, C., Mao, J. I., Luo, S., Kirchner, J. J., Eletr, S., et al. (2000). In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proceedings of the National Academy of Sciences*, 97:1665.

Bright, R. A., Shay, D. K., Shu, B., Cox, N. J., and Klimov, A. I. (2006). Adamantane resistance among influenza a viruses isolated early during the 2005-2006 influenza season in the united states. *Journal of the American Medical Association*, 295.

Bush, R. M. (1999). Positive selection on the h3 hemagglutinin gene of human influenza virus a. *Molecular Biology and Evolution*, 16:1457–1465.

Bush, R. M., Bender, C. A., Subbarao, K., Cox, N. J., and Fitch, W. M. (1999). Predicting the evolution of human influenza a. *Science*, 286:1921.

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

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

Clements, M. L., Betts, R. F., Tierney, E. L., and Murphy, B. R. (1986). Serum and nasal wash antibodies associated with resistance to experimental challenge with influenza a wild-type virus. *Journal of clinical microbiology*, 24:157.

Clements, M. L. and Murphy, B. R. (1986). Development and persistence of local and systemic antibody responses in adults given live attenuated or inactivated influenza a virus vaccine. *Journal of clinical microbiology*, 23:66.

Comeron, J. M. (1995). A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *Journal of molecular evolution*, 41:1152Ð1159.

Cooper, C. R., Hanson, L. A., Diehl, W. J., Pharr, G. T., and Coats, K. S. (1999). Natural selection of thePolGene of bovine immunodeficiency virus. *Virology*, 255:294–301.

Cox, R. J., Brokstad, K. A., and Ogra, P. (2004). Influenza virus: immunity and vaccination strategies. comparison of the immune response to inactivated and live, attenuated influenza vaccines. *Scandinavian journal of immunology*, 59:1Ð15.

Cristina, J. and Colina, R. (2006). Evidence of structural genomic region recombination in hepatitis c virus. *Virology Journal*, 3:53.

Cros, J. F. and Palese, P. (2003). Trafficking of viral genomic RNA into and out of the nucleus: influenza, thogoto and borna disease viruses. *Virus Research*, 95:3–12.

Crosby, A. W. (1976). *Epidemic and peace, 1918*. Greenwood Press.

Cuevas, J. M., Elena, S. F., and Moya, A. (2002). Molecular basis of adaptive convergence in experimental populations of RNA viruses. *Genetics*, 162:533–542.

Desai, M. M. and Plotkin, J. B. (2008). The polymorphism frequency spectrum of finitely many sites under selection. *Genetics*, 180:2175.

Detjen, B. M., StAngelo, C., Katze, M. G., and Krug, R. M. (1987). The three influenza virus polymerase (p) proteins not associated with viral nucleocapsids in the infected cell are in the form of a complex. *Journal of Virology*, 61:16–22.

Deyde, V. M., Xu, X., Bright, R. A., Shaw, M., Smith, C. B., Zhang, Y., Shu, Y., Gubareva, L. V., Cox, N. J., and Klimov, A. I. (2007). Surveillance of resistance to adamantanes among influenza a (H3N2) and a (H1N1) viruses isolated worldwide. *The Journal of infectious diseases*, 196:249–257.

Domingo, E. (2000). Viruses at the edge of adaptation. *Virology*, 270:251–253.

Domingo, E. and Holland, J. J. (1997). RNA virus mutations and fitness for survival. *Annual Reviews in Microbiology*, 51:151–178.

Donnelly, P. and Tavare, S. (1995). Coalescents and genealogical structure under neutrality. *Annual Reviews in Genetics*, 29:401–421.

Duffy, M., Salemi, M., Sheehy, N., Vandamme, A. M., Hegarty, J., Curry, M., Nolan, N., Kelleher, D., McKiernan, S., and Hall, W. W. (2002). Comparative rates of nucleotide sequence variation in the hypervariable region of E1/E2 and the NS5b region of hepatitis c virus in patients with a spectrum of liver disease resulting from a common source of infection. *Virology*, 301:354–364.

Dushoff, J., Plotkin, J. B., Levin, S. A., and Earn, D. J. D. (2004). Dynamical resonance can account for seasonality of influenza epidemics. *Proceedings of the National Academy of Sciences of the United States of America*, 101:16915–16916.

Ebel, G. D., Spielman, A., and Telford, S. R. (2001). Phylogeny of north american powassan virus. *J Gen Virol*, 82:1657–1665.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32:1792.

Edwards, C. T. T., Holmes, E. C., Pybus, O. G., Wilson, D. J., Viscidi, R. P., Abrams, E. J., Phillips, R. E., and Drummond, A. J. (2006a). Evolution of the human immunodeficiency virus envelope gene is dominated by purifying selection. *Genetics*, 174:1441–1453.

Edwards, C. T. T., Holmes, E. C., Pybus, O. G., Wilson, D. J., Viscidi, R. P., Abrams, E. J., Phillips, R. E., and Drummond, A. J. (2006b). Evolution of the human immunodeficiency virus envelope gene is dominated by purifying selection. *Genetics*, 174:1441–1453.

Egea, R., Casillas, S., and Barbadilla, A. (2008). Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucl. Acids Res.*, 36:157–162.

Epstein, S. L., pui Kong, W., Misplon, J. A., Lo, C., Tumpey, T. M., Xu, L., and Nabel, G. J. (2005).

Protection against multiple influenza a subtypes by vaccination with highly conserved nucleoprotein. *Vaccine*, 23:5404–5410.

Eyre-Walker, A. (2002). Changing effective population size and the McDonald-Kreitman test. *Genetics*, 162:2017.

Eyre-Walker, A. (2006). The genomic rate of adaptive evolution. *Trends in Ecology and Evolution*, 21:569–575.

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

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

Felsenstein, J. (2003). *Inferring phylogenies*. Sinauer Associates.

Fernandez-Sesma, A., Marukian, S., Ebersole, B. J., Kaminski, D., Park, M. S., Yuen, T., Sealfon, S. C., Garcia-Sastre, A., and Moran, T. M. (2006). Influenza virus evades innate and adaptive immunity via the NS1 protein. *The Journal of Virology*, 80:6295.

Fitch, W. M., Bush, R. M., Bender, C. A., and Cox, N. J. (1997). Long term trends in the evolution of h (3) HA1 human influenza type a. *National Acad Sciences*, 94:7712–7718.

Fitch, W. M., Leiter, J. M. E., Li, X., and Palese, P. (1991). Positive darwinian evolution in human influenza a viruses. *Proceedings of the National Academy of Sciences*, 88:4270–4274.

Flint, M. and McKeating, J. A. (2000). The role of the hepatitis c virus glycoproteins in infection. *Reviews in medical virology*, 10:101–117.

Frost, W. H. (1920). Statistics of influenza morbidity: With special reference to certain factors in case incidence and case fatality. *Public Health Reports (1896-1970)*, pages 584–597.

Fu, Y. X. and Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics*, 133:693–709.

Garcia-Sastre, A., Egorov, A., Matassov, D., Brandt, S., Levy, D. E., Durbin, J. E., Palese, P., and Muster, T. (1998). Influenza a virus lacking the NS1 gene replicates in Interferon-Deficient systems. *Virology*, 252:324–330.

Gardiner, C. W. (1985). *Handbook of stochastic methods*. Springer Berlin.

Gaschen, B., Taylor, J., Yusim, K., Foley, B., Gao, F., Lang, D., Novitsky, V., Haynes, B., Hahn, B. H., Bhattacharya, T., et al. (2002). Diversity considerations in HIV-1 vaccine selection. *Science*, 296:2354.

Glezen, W. P. and Couch, R. B. (1978). Interpandemic influenza in the houston area, 1974-76. *N Engl J Med*, 298:587–592.

Gog, J. R., Rimmelzwaan, G. F., Osterhaus, A. D., and Grenfell, B. T. (2003). Population dynamics of rapid fixation in cytotoxic T lymphocyte escape mutants of influenza a. *Proceedings of the National Academy of Sciences*, 100:11143.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*, 11:725–736.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223.

Graur, D. (1991). Neutral mutation hypothesis test. *Nature*, 354:114–115.

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

Griffiths, R. C. and Marjoram, P. (1996). Ancestral inference from samples of dna sequences with recombination. *J Comput Biol*, 3:479–502.

Griffiths, R. C. and Marjoram, P. (1997). An ancestral recombination graph. *Institute for Mathematics and Its Applications,*, 87:257.

Griffiths, R. C. and Tavare, S. (1994a). Ancestral inference in population genetics. *Stat. Sci*, 9:307–319.

Griffiths, R. C. and Tavare, S. (1994b). Sampling theory for neutral alleles in a varying environment. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 344:403–10.

Hammersley, J. M. (1960). Monte carlo methods for solving multivariable problems. *Ann. New York Acad. Sci*, 86:1960.

Hara, K., Schmidt, F. I., Crow, M., and Brownlee, G. G. (2006). Amino acid residues in the n-terminal region of the PA subunit of influenza a virus RNA polymerase play a critical role in protein stability, endonuclease activity, cap binding, and virion RNA promoter binding. *The Journal of Virology*, 80:7789.

Hardle, W., Horowitz, J., and Kreiss, J. (2003). Bootstrap methods for time series. *International Statistical Review*, 71:435–459.

Hasegawa, M., Cao, Y., and Yang, Z. (1998). Preponderance of slightly deleterious polymorphism in mitochondrial dna: Nonsynonymous/synonymous rate ratio is much higher within species than between species. *Molecular Biology and Evolution*, 15:1499.

Holmes, E. C. (2003). Patterns of intra-and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. *The Journal of Virology*, 77:11296.

Howe, K., Bateman, A., and Durbin, R. (2002). QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics*, 18:1546.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, 7:1–44.

Huelsenbeck, J. P. and Ronquist, F. (2001). MrBayes: a program for the bayesian inference of phylogeny. *Bioinformatics*, 17:754–755.

Hughes, A. L. (2005). Evidence for abundant slightly deleterious polymorphisms in bacterial populations. *Genetics*, 169:533.

Hughes, A. L. and Hughes, M. A. K. (2007). More effective purifying selection on RNA viruses than in dna viruses. *Gene*, 404:117–25.

Huxley, J. and Baker, J. R. (1963). *Evolution: the modern synthesis*. Allen and Unwin London.

Ina, Y. (1995). New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *Journal of Molecular Evolution*, 40:190Ð226.

Ina, Y. and Gojobori, T. (1994). Statistical analysis of nucleotide sequences of the hemagglutinin gene of human influenza a viruses. *Proceedings of the National Academy of Sciences*, 91:8388–8392.

Jenkins, G. M. (2002). Rates of molecular evolution in RNA viruses: A quantitative phylogenetic analysis. *Journal of Molecular Evolution*, 54:156–165.

Johnson, N. and Mueller, J. (2002). Updating the accounts: Global mortality of the 1918-1920" spanish" influenza pandemic. *Bulletin of the history of medicine*, 76:105–120.

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. *Mammalian Protein Metabolism*, 3:21–132.

Jurgensen, P. F., Olsen, G. N., III, J. E. J., Swenson, E. W., Ayoub, E. M., Henney, C. S., and Waldman, R. H. (1973). Immune response of the human respiratory tract. II. cell-mediated immunity in the lower respiratory tract to tuberculin and mumps and influenza viruses. *The Journal of Infectious Diseases*, 128:730–735.

Kaji, M., Watanabe, A., and Aizawa, H. (2003). Differences in clinical features between influenza a H1N1, a H3N2, and b in adult patients. *Respirology*, 8:231–233.

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

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61:893–903.

Kimura, M. (1983). *Neutral Theory of Molecular Evolution*. Cambridge University Press.

Kingman, J. F. C. (1982a). The coalescent. *Stochastic Process. Appl*, 13:235–248.

Kingman, J. F. C. (1982b). On the genealogy of large populations. *J. Appl. Prob*, 19:27–43.

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

Lamb, R. A. (1989). Genes and proteins of the influenza viruses. *The Influenza Viruses*, pages 1–87.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., and Lopez, R. (2007). ClustalW2 and ClustalX version 2. *Bioinformatics*, 23:2947–2948.

Lewis, D. B. (2006). Avian flu to human influenza. *Annual Review of Medicine*, 57:139–54.

Li, S., Min, J. Y., Krug, R. M., and Sen, G. C. (2006). Binding of the influenza a virus NS1 protein to PKR mediates the inhibition of its activation by either PACT or double-stranded RNA. *Virology*, 349:13–21.

Li, W. H., Wu, C. I., and Luo, C. C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution*, 2:150.

Lin, D., Lan, J., and Zhang, Z. (2007). Structure and function of the NS1 protein of influenza a virus. *Acta Biochimica et Biophysica Sinica*, 39:155.

Ludwig, S., Schultz, U., Mandler, J., Fitch, W. M., and Scholtissek, C. (1991). Phylogenetic relationship of the nonstructural (NS) genes of influenza a viruses. *Virology*, 183:566–577.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., and Chen, Z. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380.

McAllister, J., Casino, C., Davidson, F., Power, J., Lawlor, E., Yap, P. L., Simmonds, P., and Smith, D. B. (1998). Long-term evolution of the hypervariable region of hepatitis c virus in a common-source-infected cohort. *The Journal of Virology*, 72:4893.

McDonald, J. H. and Kreitman, M. (1991a). Adaptive protein evolution at the adh locus in drosophila. *Nature*, 351:652–654.

McDonald, J. H. and Kreitman, M. (1991b). Neutral mutation hypothesis test. *Nature*, 354:116.

McMichael, A. J., Gotch, F. M., Noble, G. R., and Beare, P. A. (1983). Cytotoxic T-cell immunity to influenza. *New England Journal of Medicine*, 309:13.

McMichael, A. J., Michie, C. A., Gotch, F. M., Smith, G. L., and Moss, B. (1986). Recognition of influenza a virus nucleoprotein by human cytotoxic T lymphocytes. *Journal of general virology*, 67:719.

Memon, M. I. and Memon, M. A. (2002). Hepatitis c: an epidemiological review. *Journal of viral hepatitis*, 9:84–100.

Mitnaul, L. J., Matrosovich, M. N., Castrucci, M. R., Tuzikov, A. B., Bovin, N. V., Kobasa, D., and Kawaoka, Y. (2000). Balanced hemagglutinin and neuraminidase activities are critical for efficient replication of influenza a virus. *Journal of Virology*, 74:6015–6020.

Moore, G. E. (1998). Cramming more components onto integrated circuits. *Proceedings of the IEEE*, 86:82–85.

Muse, S. V. (1996). Estimating synonymous and nonsynonymous substitution rates. *Molecular biology and evolution*, 13:105.

Nath, H. B. and Griffiths, R. C. (1996). Estimation in an island model using simulation. *Theoretical Population Biology*, 50:227–253.

Nayak, D. P., Hui, E. K., and Barman, S. (2004). Assembly and budding of influenza virus. *Virus Research*, 106:147–65.

Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular biology and evolution*, 3:418–26.

Nelson, M. I. and Holmes, E. C. (2007). The evolution of epidemic influenza. *Nature Reviews Genetics*, 8:196–205.

Nelson, M. I., Simonsen, L., Viboud, C., Miller, M. A., Taylor, J., George, K. S., Griesemer, S. B., Ghedin, E., Sengamalay, N. A., Spiro, D. J., Volkov, I., Grenfell, B. T., Lipman, D. J., Taubenberger, J. K., and Holmes, E. C. (2006). Stochastic processes are key determinants of Short-Term evolution in influenza a virus. *PLoS Pathogens*, 2.

Neuhauser, C. and Krone, S. M. (1997). The genealogy of samples in models with selection. *Genetics*, 145:519–534.

Nielsen, R. (2001). Statistical tests of selective neutrality in the age of genomics. *Heredity*, 86:641–647.

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

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

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

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

Novella, I. S., Zarate, S., Metzgar, D., and Ebendick-Corpus, B. E. (2004). Positive selection of synonymous mutations in vesicular stomatitis virus. *Journal of molecular biology*, 342:1415Ð1421.

O'Neill, R. E., Jaskunas, R., Blobel, G., Palese, P., and Moroianu, J. (1995). Nuclear import of influenza virus RNA can be mediated by viral nucleoprotein and transport factors required for protein import. *Journal of Biological Chemistry*, 270:22701.

Palese, P. (1977). The genes of influenza virus. *Cell*, 10:1–10.

Perales, B. and Ortin, J. (1997). The influenza a virus PB2 polymerase subunit is required for the replication of viral RNA. *Journal of Virology*, 71:1381–1385.

Plotkin, J. B., Dushoff, J., and Levin, S. A. (2002). Hemagglutinin sequence clusters and the antigenic evolution of influenza a virus. *Proceedings of the National Academy of Sciences of the United States of America*, 99:6263–6268.

Pond, S. L. and Frost, S. D. (2005). DATAMONKEY: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics*, 21:2531.

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

Portela, A. and Digard, P. (2002). The influenza virus nucleoprotein: a multifunctional RNA-binding protein pivotal to virus replication. *The Journal of General Virology*, 83:723–34.

Potter, C. (2001). A history of influenza. *Journal of Applied Microbiology*, 91:572–579.

Pybus, O. G., Holmes, E. C., and Harvey, P. H. (1999). The mid-depth method and HIV-1: a practical approach for testing hypotheses of viral epidemic history. *Molecular biology and evolution*, 16:953.

Pybus, O. G. and Rambaut, A. (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews Genetics*, 10:540–550.

Pybus, O. G., Rambaut, A., Belshaw, R., Freckleton, R. P., Drummond, A. J., and Holmes, E. C. (2007). Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. *Molecular biology and evolution*, 24:845.

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

Pybus, O. G. and Shapiro, B. (2008). *Natural Selection and Adaptation of Molecular Sequences*, chapter 6. in Phylogenetics Handbook. Cambridge University Pess, first edition.

Raiffa, H. and Schlaifer, R. (1972). Applied statistical decision theory. *Boston; MIT Press, 1972, 382 p. Ilus, tablas*.

Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K., and Holmes, E. C. (2008). The genomic and epidemiological dynamics of human influenza a virus. *Nature*, 453:615–619.

Reid, A. H. (1999). Origin and evolution of the 1918 spanish influenza virus hemagglutinin gene.

Ren, F., Tanaka, H., and Yang, Z. (2005). An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. *Systematic biology*, 54:808–818.

Rimmelzwaan, G. F., Berkhoff, E. G. M., Nieuwkoop, N. J., Fouchier, R. A. M., and Osterhaus, A. (2004). Functional compensation of a detrimental amino acid substitution in a cytotoxic-T-lymphocyte epitope of influenza a viruses by comutations. *Journal of virology*, 78:8946.

Ritchie, P. A., Anderson, I. L., and Lambert, D. M. (2003). Evidence for specificity of psittacine beak and feather disease viruses among avian hosts. *Virology*, 306:109–115.

Rocha, E. P., Smith, J. M., Hurst, L. D., Holden, M. T., Cooper, J. E., Smith, N. H., and Feil, E. J. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of theoretical biology*, 239:226–235.

Rodrigo, A. G. and Felsenstein, J. (1999). *Coalescent approaches to HIV population genetics*, chapter 8. in Coalescent Approaches to HIV Population Genetics. Johns Hopkins University Press, first edition.

Rodrigo, A. G., Shpaer, E. G., Delwart, E. L., Iversen, A. K. N., Gallo, M. V., Brojatsch, J., Hirsch, M. S., Walker, B. D., and Mullins, J. I. (1999). Coalescent estimates of HIV-1 generation time in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 96.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M., and Nyren, P. (1996). Realtime dna sequencing using detection of pyrophosphate release. *Analytical biochemistry*, 242:84–89.

Rozas, J. and Rozas, R. (1999). DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics*, 15:174.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4:406.

Sanjuan, R., Moya, A., and Elena, S. F. (2004). The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proceedings of the National Academy of Sciences of the United States of America*, 101:8396–8401.

Sanjuan, R., Nebot, M. R., Chirico, N., Mansky, L. M., and Belshaw, R. (2010). Viral mutation rates. *Journal of virology*, 84:9733.

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

Shapiro, B., Rambaut, A., Pybus, O. G., and Holmes, E. C. (2006). A phylogenetic method for detecting positive epistasis in gene sequences and its application to RNA virus evolution. *Molecular Biology and Evolution*, 23:1724–1730.

Sharp, P. M., Bailes, E., Chaudhuri, R. R., Rodenburg, C. M., Santiago, M. O., and Hahn, B. H.

(2001). The origins of acquired immune deficiency syndrome viruses: where and when? *Philosophical Transactions: Biological Sciences*, 356:867–876.

Sheridan, I., Pybus, O. G., Holmes, E. C., and Klenerman, P. (2004a). High-resolution phylogenetic analysis of hepatitis c virus adaptation and its relationship to disease progression. *Journal of Virology*, 78:3447.

Sheridan, I., Pybus, O. G., Holmes, E. C., and Klenerman, P. (2004b). High-resolution phylogenetic analysis of hepatitis c virus adaptation and its relationship to disease progression. *The Journal of Virology*, 78:3447.

Shih, A. C., Hsiao, T. C., Ho, M. S., and Li, W. H. (2007). Simultaneous amino acid substitutions at antigenic sites drive influenza a hemagglutinin evolution. *Proceedings of the National Academy of Sciences*, 104:6283.

Shriner, D., Rodrigo, A. G., Nickle, D. C., and Mullins, J. I. (2004). Pervasive genomic recombination of HIV-1 in vivo. *Genetics*, 167:1573.

Simonsen, K. L., Churchill, G. A., and Aquadro, C. F. (1995). Properties of statistical tests of neutrality for dna polymorphism data. *Genetics*, 141:413–429.

Simonsen, L., Viboud, C., Grenfell, B. T., Dushoff, J., Jennings, L., Smit, M., Macken, C., Hata, M., Gog, J., Miller, M. A., et al. (2007). The genesis and spread of reassortment human influenza A/H3N2 viruses conferring adamantane resistance. *Molecular biology and evolution*, 24:1811.

Skehel, J. J. and Wiley, D. C. (2000). Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. *Annual Review of Biochemistry*, 69:531–69.

Slatkin, M. and Hudson, R. R. (1991). Pairwise comparisons of mitochondrial dna sequences in stable and exponentially growing populations. *Genetics*, 129:555–562.

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

Smith, G. J., Vijaykrishna, D., Bahl, J., Lycett, S. J., Worobey, M., Pybus, O. G., Ma, S. K., Cheung, C. L., Raghwani, J., Bhatt, S., et al. (2009). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza a epidemic. *Nature*, 459:1122–1125.

Smith, N. G. and Eyre-Walker, A. (2002). Adaptive protein evolution in drosophila. *Nature*, 415:1022–4.

Sneath, P. H. and Sokal, R. R. (1973). *Numerical taxonomy: the principles and practice of numerical classification*. WH Freeman San Francisco.

Subbarao, K. and Katz, J. (2000). Avian influenza viruses infecting humans. *Cellular and Molecular Life Sciences*, 57:1770–1784.

Suchard, M. A. and Rambaut, A. (2009). Many-core algorithms for statistical phylogenetics. *Bioinformatics*, 25:1370.

Sugrue, R. J. and Hay, A. J. (1991). Structural characteristics of the m2 protein of influenza a viruses: evidence that it forms a tetrameric channel. *Virology*, 180:617–24.

Suzuki, Y. (2006). Natural selection on the influenza virus genome. *Molecular biology and evolution*, 23:1902.

Suzuki, Y., Gojobori, T., and Nei, M. (2001). ADAPTSITE: detecting natural selection at single amino acid sites. *Bioinformatics*, 17:660.

Swofford, D. L. (2002). PAUP*: phylogenetic analysis using parsimony (* and other methods), version 4.0 b10. *Sunderland, MA: Sinauer Associates*.

Tajima, F. (1983). Evolutionary relationship of dna sequences in finite populations. *Genetics*, 105:437–460.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123:585–595.

Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular biology and evolution*, 24:1596.

Townsend, A. R., Rothbard, J., Gotch, F. M., Bahadur, G., Wraith, D., and McMichael, A. J. (1986). The epitopes of influenza nucleoprotein recognized by cytotoxic T lymphocytes can be defined with short synthetic peptides. *Cell*, 44:959–968.

Townsend, A. R. M., McMichael, A. J., Carter, N. P., Huddleston, J. A., and Brownlee, G. G. (1984). Cytotoxic T cell recognition of the influenza nucleoprotein and hemagglutinin expressed in transfected mouse l cells. *Cell*, 39:13–25.

Tsompana, M., Abad, J., Purugganan, M., and Moyer, J. W. (2005). The molecular population genetics of the tomato spotted wilt virus (TSWV) genome. *Molecular Ecology*, 14:53–66.

Tumpey, T. M., Garcia-Sastre, A., Taubenberger, J. K., Palese, P., Swayne, D. E., Pantin-Jackwood, M. J., Schultz-Cherry, S., Solorzano, A., Rooijen, N. V., Katz, J. M., and Basler, C. F. (2005). Pathogenicity of influenza viruses with genes from the 1918 pandemic virus: Functional roles of alveolar macrophages and neutrophils in limiting virus replication and mortality in mice. *Journal of Virology*, 79:14933–14944.

Varghese, J. N., Laver, W. G., and Colman, P. M. (1983). Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9 a resolution. *Nature*, 303:35–40.

Viboud, C., Alonso, W. J., and Simonsen, L. (2006). Influenza in tropical regions. *PLoS Medicine*, 3:e89.

Voeten, J. T. M., Bestebroer, T. M., Nieuwkoop, N. J., Fouchier, R. A. M., Osterhaus, A., and Rimmelzwaan, G. F. (2000). Antigenic drift in the influenza a virus (H3N2) nucleoprotein and escape from recognition by cytotoxic T lymphocytes. *Journal of virology*, 74:6800.

Wakeley, J. (2003). Polymorphism and divergence for island-model species. *Genetics*, 163:411.

Wall, J. D., Andolfatto, P., and Przeworski, M. (2002). Testing models of selection and demography in drosophila simulans. *Genetics*, 162:203.

Wang, C., Mitsuya, Y., Gharizadeh, B., Ronaghi, M., and Shafer, R. W. (2007). Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome research*, 17:1195.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*, 7:256–76.

Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M., and Kawaoka, Y. (1992). Evolution and ecology of influenza a viruses. *Microbiology and Molecular Biology Reviews*, 56:152–179.

Weinstock, D. M. and Zuccotti, G. (2006). Adamantane resistance in influenza a. *Jama*, 295:934.

Welch, J. J. (2006). Estimating the genomewide rate of adaptive protein evolution in drosophila. *Genetics*, 173:821–837.

Whittam, T. S. and Nei, M. (1991). Neutral mutation hypothesis test. *Nature*, 354:115–116.

Wiley, D. C. and Skehel, J. J. (1987). The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *Annual Reviews in Biochemistry*, 56:365–394.

Williamson, S. (2003). Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. *Molecular Biology and Evolution*, 20:1318–1325.

Williamson, S. H., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R., and Bustamante, C. D. (2005). Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proceedings of the National Academy of Sciences*, 102:7882.

Wilson, D. J. and McVean, G. (2006). Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics*, 172:1411.

Wolf, Y. I., Viboud, C., Holmes, E. C., Koonin, E. V., and Lipman, D. J. (2006). Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza a virus. *Biology Direct*, 1:34.

Wright, P. F., Thompson, J., and Karzon, D. T. (1980). Differing virulence of H1N1 and H3N2 influenza strains. *American journal of epidemiology*, 112:814.

Wright, S. (1945). The differential equation of the distribution of gene frequencies. *PNAS*, 31:382.

Yamaguchi, K., Tanaka, E., Higashi, K., Kiyosawa, K., Matsumoto, A., Furuta, S., Hasegawa, A., Tanaka, S., and Kohara, M. (1994). Adaptation of hepatitis c virus for persistent infection in patients with acute hepatitis. *Gastroenterology*, 106:1344.

Yang, Z. (1997). Paml: a program package for phylogenetic analysis by maximum likelihood. 13:555–556.

Yang, Z. (2000). Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus a. *Journal of molecular evolution*, 51:423–432.

Yang, Z. (2007). Paml 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24:1586.

Yang, Z. and Bielawski, J. P. (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, 15:496–503.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A. M. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155:431–449.

Zhai, W., Nielsen, R., and Slatkin, M. (2009). An investigation of the statistical power of neutrality tests based on comparative and population genetic data. *Molecular Biology and Evolution*, 26:273–283.

Zhu, L. and Bustamante, C. D. (2005). A composite-likelihood approach for detecting directional selection from DNA sequence data. *Genetics*, 170:1411.

Zwickl, D. (2006). Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion.