The within- and among-host evolution of chronically-infecting human RNA viruses

A thesis submitted for the degree of D.Phil. at the University of Oxford

Trinity Term 2008

Joe Parker

Linacre College

Abstract

The within- and among-host evolution of chronically-infecting human RNA viruses

Joe Parker, Linacre College, University of Oxford

D.Phil. thesis, Trinity 2008

This thesis examines the evolutionary biology of the RNA viruses, a diverse group of pathogens that cause significant diseases. The focus of this work is the relationship between the processes driving the evolution of virus populations within individual hosts and at the epidemic level. First, Chapter One reviews the basic biology of RNA viruses, the current state of knowledge in relevant topics of evolutionary virology, and the principles that underlie the most commonly used methods in this thesis. In Chapter Two, I develop and test a novel framework to estimate the significance of phylogeny-trait association in viral phylogenies. The method incorporates phylogenetic uncertainty through the use of posterior sets of trees (PST) produced in Bayesian MCMC analyses. In Chapter Three, I conduct a comprehensive analysis of the substitution rate of hepatitis C virus (HCV) in within- and between-host data sets using a relaxed molecular clock. I find that within-host substitution rates are more rapid than previously appreciated, that heterotachy is rife in within-host data sets, and that selection is likely to be a primary driver. In Chapter Four I apply the techniques developed in Chapter Two to successfully detect compartmentalization between peripheral blood and cervical tissues in a large data set of human immunodeficiency virus (HIV) patients. I propose that compartmentalization in the cervix is maintained by selection. I extend the framework developed in Chapter Two in Chapter Five and explore the Type II error of the statistics used. In Chapter Six I review the findings of this thesis and conclude with a general discussion of the relationship between within- and among-host evolution in viruses, and some of the limitations of current techniques.

Acknowledgements

"Without people, you're nothing"

- Joe Strummer

I have not been a model student; and so I have been very lucky that in Dr. Oli Pybus I had a supervisor of incredible energy, skill, vision and above all, patience. Because Oli showed me the gulf between bodging-up a pretty idea and ruthlessly hammering out a great one (sadly none yet myself) I cannot repay him, so these thanks must do instead. I am grateful also to Dr. Andrew Rambaut who first took me on, a scruffy biology graduate with some vague ideas about programming. I might not look any neater four years later, but if my coding is any crisper or my mind any sharper it is by following his example.

I have also been fortunate to count on some cracking colleagues. We have worked and played and their jokes kept me going, insights kept me ticking, and Darwin's cards kept me fed. Abby, Alex, Alexei, Aris, Beth, Colin, Dino, Iain, James, Liz, Patty, Peter, the Philippes (laBoite et laViande), Polly, Rob, Sam and Tulio were the ringleaders but there were hundreds more. In terror of missing anyone out I'd better stop there – but you all know who you are.

The assistance of Mr. T. W. G. Dawson is appreciated; as he well knows, this thesis is one volume in a story that began some time ago. And finally, this thesis would not have been written at all but for the encouragement, support, goading and love of all my friends and family. Thank you very much.

Table of Contents

The within- and among-host evolution of chronically-infecting human RNA viruses

Chapter One

Introduction: The Evolutionary Biology of RNA viruses

1.1 – Introduction to the introduction

Viruses have been studied for only a century, yet the diseases they cause have shaped human populations through history, and continue to do so to this day. Likewise, the evolutionary behaviour of viruses has only recently opened itself to scientific scrutiny, and methods for the investigation of viral evolution remain under-developed in many areas. In this thesis I present work on the evolutionary biology of RNA viruses (including the retroviruses), a diverse group of viruses responsible for significant morbidity and mortality in humans and economically important organisms.

In this chapter I review current topics in viral evolutionary biology, with particular reference to the unique problems that viral biology presents to the theoretical and empirical understanding of virus evolution. For example, it appears that the evolutionary processes dominating the evolution of viral populations within hosts differ substantially from those that predominate at the level of host populations (Grenfell *et al*., 2004). As a result, it is useful to consider the evolution of viruses at two levels; within- and between-hosts. A major goal of this thesis is to explicitly demonstrate the different results arising from viral evolution at these two scales.

In Chapter Two I present a new methodology to detect association between phenotypic traits (*e.g*. morphology, geography, pathology) and phylogeny, and illustrate the importance of this method to viral sequence analysis. Chapter Three provides the first rigorous whole genome study of the rate of hepatitis C virus (HCV) evolution. The analysis presented uses serially-sampled (heterochronous) sequence data to investigate HCV evolution at the within- and between-host levels. In Chapter Four I use a range of techniques, including the trait-association approach developed in Chapter Two, to test the hypothesis that human immunodeficiency virus type 1 (HIV-1) exhibits significant population structure among tissues within infected individuals (known as viral 'compartmentalization'). Chapter Five explores the addition of branch length information to measures of phylogeny-trait

association, and introduces a randomisation method to test the statistical power of such approaches.

In Appendix One I present a software tool developed during my research that quantifies sitewise diversity in sequence alignments using a variety of measures, including Shannon's entropy score. In Appendices Two and Three I present two multi-author studies of viral evolution to which I contributed. In Appendix Four I present BaTS, my implementation of the method developed in Chapter Two.

1.1.1 Why study RNA viruses?

Despite being some of the smallest replicons known to biology, RNA viruses have nevertheless attracted considerable scientific interest since their discovery. Many viral diseases cause substantial, and in some cases increasing, morbidity and mortality in humans. These range from acute infections that generate seasonal epidemics, such as influenza (up to 500,000 deaths annually; WHO 2006) and measles (over 300,000 deaths in 2005; Wolfson *et al.,* 2007) to endemic diseases like Dengue (~50-100m cases and 500,000 deaths worldwide annually; Gubler, 2006) . In addition, many viruses cause persistent or chronic infections, including the hepatitis C virus (up to 180m people infected worldwide; WHO,1999) and HIV, the aetiological agent of acquired immunodeficiency syndrome, or AIDS (>30m infections worldwide; ~2.5 million new infections in 2007; UNAIDS 2007). Some viruses are considered to be 'emerging infections' as they have only recently been identified or begun to spread epidemically among humans (e.g. HIV/AIDS, HCV, West Nile Virus, human T-cell lympotropic virus, SARS-coronavirus and many more).

RNA viruses are also responsible for non-human disease with important economic consequences. For instance, the 2001 UK outbreak of foot-and-mouth disease, caused by the Group IV Aphthovirus foot and mouth disease virus, is thought to have cost the UK economy over £8 billion in direct and indirect costs (NAO, 2002). RNA viruses exert an economic cost on societies across the world through morbidity and mortality of poultry and livestock and through reduced crop yields. Finally, but by no means insignificantly, the rapid rate at which some viruses evolve can provide biologists with an opportunity to study fundamental evolutionary processes on timescale of months or years (*e.g.* Drummond *et al.*, 2002; Shankarappa *et al.*, 1999). Furthermore, their small genome sizes facilitate genetic investigation (the first gene sequenced was an RNA bacteriophage; Jou *et al.*, 1972).

1.1.2 The Biology of RNA viruses

1.1.2.1 RNA virus structure

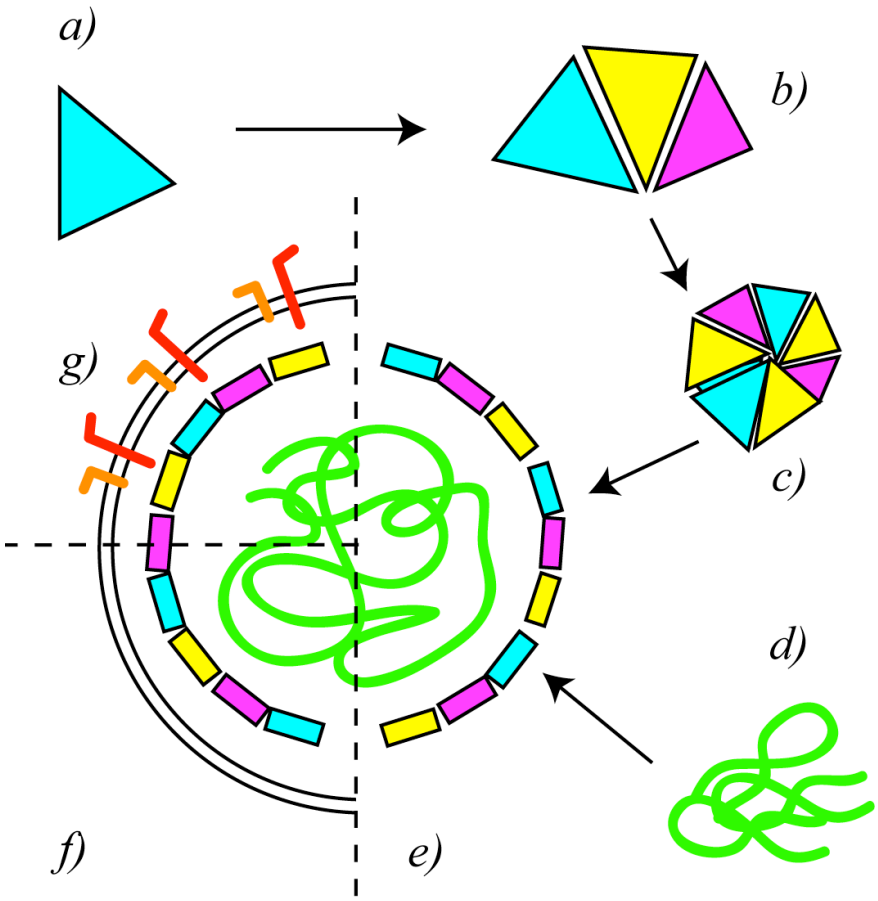


Figure 1.1. Diagram illustrating the typical structure of an RNA virus. The capsid is comprised of *a)* one or more sub-units, that are assembled *b) & c)* to form the virion, which packages and protect one or more RNA molecules *d)*. The genome may be further surrounded by an envelope *f)*, a lipid bilayer derived from host cell membranes. This bilayer is studded with surface proteins *g)* to effect interactions with host cell membranes. Non-enveloped virion capsids may also display additional surface proteins. After Alberts, B. *et al*. (2002) and Fauquet *et al*., (2005).

As a group, the RNA viruses are the smallest replicons known to science. Most DNA viruses are larger, whilst prions or selfish genetic elements have less claim to be independent

replicating entities. RNA virus are usually either helical (rod-shaped) or icosohedral (roughly spherical); icoohedral viruses vary in size from parvovirus (18-26nm in diameter) to the considerably larger henipavirus (up to 600nm diameter). Most common spherical viruses are around 30-100nm in diameter, whilst a typical rod-shaped virus is about 150nm long. Ebolavirus, at up to 1400nm long, is one of the largest rod-shaped viruses. In rod-shaped viruses, the capsid packs around the genome in a spiral, forming a long cylinder, whereas in the spherical viruses the capsid forms a hollow space enclosing the genome (see Figure 1.1). Other peptides or enzymes may also be packaged into the capsid, which may be naked, or surrounded by a lipid envelope. Surface proteins (glycoproteins or other peptides) stud the envelope or capsid and are required to effect target cell entry (Figure 1.1).

1.1.2.2 RNA virus genomes

Viruses are classified by the International Committee on Taxonomy of Viruses (ICTV) in conjunction with the "Baltimore" system which groups viruses primarily on the basis of their mode of mRNA production (Fauquet *et al*., 2005). RNA viruses are arranged into the double-stranded RNA viruses (Group III), the positive-sense single-stranded RNA viruses (Group IV) and the negative-sense single-stranded RNA viruses (Group V). Positive-sense viruses have genomes that are equivalent to host messenger RNA molecules; negative sense viruses are correspondingly complementary to host mRNA. The retroviruses carry +ssRNA genomes that must be retro-transcribed to duplex DNA prior to replication, and are classified separately as Group VI viruses. RNA virus genomes may be contained on a single RNA molecule, or split over several molecules (a 'segmented' genome).

The genomes of RNA viruses tend to be smaller than those of DNA viruses. Total genome length, whether segmented or unsegmented, ranges in size from around 3.5kb (rous sarcoma virus) to >30kb (mouse hepatitis virus; in contrast several DNA viruses' genomes exceed 100 kb in length; Suzan-Monti *et al*., 2005) Correspondingly, they exhibit a number of mechanisms of "genome compression" by which more information is contained on the same length of nucleic acid, such as overlapping genes, multiple reading frames and alternative

splicing (Belshaw *et al*., 2007; Holmes, 2003b). By its biochemical nature, RNA is less stable than DNA and more readily forms complex secondary structures such as hairpin loops.

1.1.2.3 Viral life cycles

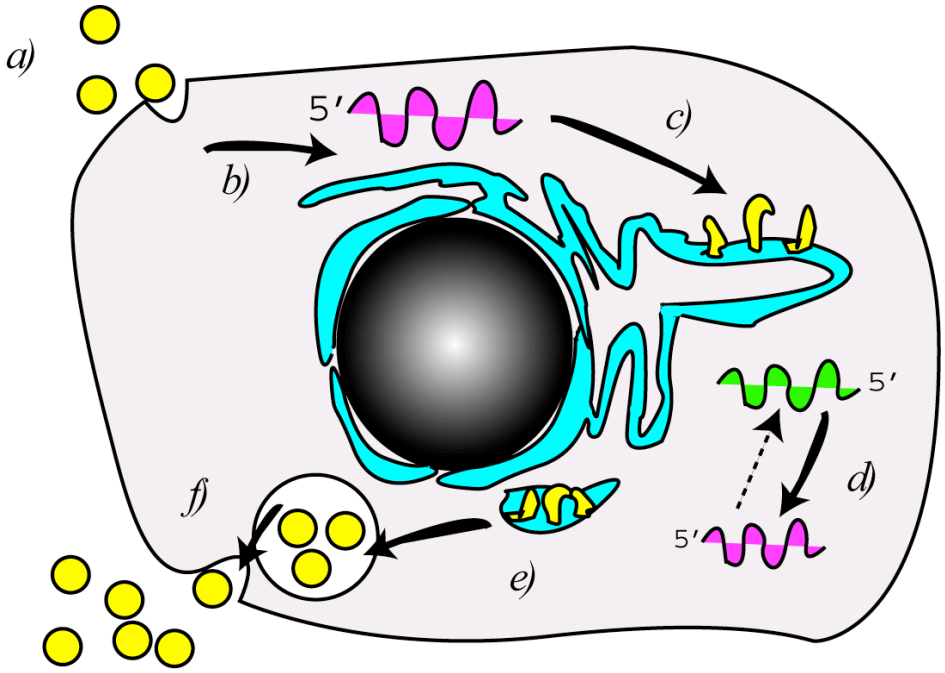


Figure 1.2. Lifecycle of the hepatitis C virus (HCV), a member of the Flaviridae. *a*) membrane fusion and cell entry; *b*) uncoating and release of +ssRNA strand; *c*) viral protein synthesis on the endoplasmic reticulum; *d*) concurrent +ssRNA replication by viral RNA polymerase through –ssRNA intermediate; *e*) capsid assembly and virion coating with lipid bilayer *f*) virion release. After Moradpour *et al*. (2007).

Cell fusion and entry is the first stage of the viral life cycle (*Figure* 1.2*a*). Most viruses exhibit strong tropism for particular cell types (e.g. HIV & $CD4^+$ T-lymphocytes; Berger *et al*, 1999), and this specificity is an important determinant of host range and disease type.

Following capsid uncoating (*Figure 2b*), the virus then replicates within the host. All RNA viruses must generate multiple copies of the genomic RNA from the single infecting virion (*Figure 2d*), both to create more progeny and also simply to increase expression of viral proteins. In Group III (dsRNA) viruses this is accomplished through a helicase and polymerase, while in Group IV (+ssRNA) the genome may be directly translated, or -ssRNA

intermediates may be produced that act as templates for further +ssRNA replication in turn. mRNA can be transcribed directly from the genomic –ssRNA of Group V viruses for peptide synthesis (but they must carry an RNA polymerase to do so since none exists in the host cytosol). The retroviruses must first produce a reverse transcriptase, which then catalyses transcription of RNA to duplex DNA (which may integrate into the host genome as a provirus). This viral DNA is then transcribed using host pathways. Following maturation (*Figure 2e*), assembled virions are released either through budding (*Figure 2f*), or cell lysis.

1.1.3 Transmission of RNA viruses

As obligate cellular parasites, viruses are not thought to persist for significant periods of time outside host cells. However, viruses may be passed from host to host through a variety of transmission methods. From a human perspective, the most important transmission routes are the aerosol route via the respiratory tract (*e.g.* SARS coronovirus; influenza), the fecal-oral route to and from the gastrointestinal tract (poliovirus), sexually-transmitted infections (HIV) and percutaneous transmission through direct contact with broken skin (hepatitis C virus, HCV). Vertical transmission (mother-to-child) is also a significant source of infection (*e.g.*, HIV), as is vector-borne infection (*e.g.* through mosquito vectors in yellow fever virus). In addition to these natural avenues of infection, the last century saw the rise of parenteral and iatrogenic transmission through the extensive use of unsafe injection, injecting drug use (IDU), blood transfusion and blood product usage. Blood-borne viruses such as HIV (Curran *et al.*, 1998) and HCV have spread by these routes (Drucker *et al.,* 2001). In addition, increasing human migration rates and higher human population densities have also aided viral dispersal (Perrin *et al.*, 2003).

1.2 RNA Virus evolution and phylodynamics

Evolutionary biologists have traditionally studied complex organisms, typically eukaryotes, for which the classic evolutionary forces of mutation, recombination, drift & selection operate at the level of populations. For such organisms, the genome is stable over measurable time scales and the boundary between 'individuals' and 'populations' is clear.

In contrast it is possible to observe evolutionary changes in RNA viruses over the same timescale (months or years) as viral population dynamics and epidemiology. Consequently, RNA virus have been defined as measurably evolving populations (MEPs; Drummond *et al.,* 2003). Both genetic drift and natural selection leave their mark on RNA virus genomes; the interplay between these forces and their effects on viral diversity has been termed 'phylodynamics.' (Grenfell *et al.,* 2004) and can be investigated with a range of phylogenetic techniques. As a result viral evolutionary biology is now in a position to make medically and economically relevant predictions, such as the future epidemiology of hepatitis C infection in the United States (Mizokami *et al.,* 2006; Tanaka *et al.,* 2002), or the seasonal generation of novel avian influenza recombinants (Rambaut *et al.*, 2008).

However the very rapidity of virus evolution, coupled with their population dynamics (including multiple severe bottleneck events on transmission; Rambaut *et al.*, 2004), frequent recombination (Worobey & Holmes, 1999) and strong selection by the immune system (Borrow *et al.*, 1997) or drug treatments (Rambaut *et al.*, 2004) considerably complicates viral phylogenetics. Of particular relevance to this thesis is evidence that discrete viral populations can persist in different organs or tissues within the same host ('compartmentalization'; Pillai *et al.,* 2005; Korber *et al.*, 1994). Furthermore, it also appears that at least some of the substitutions and adaptations acquired in one individual may, although transmitted to the next, be lost in the initial phase of infection ('reversion'; Herbeck *et al.*, 2006; Leslie *et al.*, 2004).

1.2.1 RNA Virus Evolutionary Processes

1.2.1.1 Mutation & Substitution

Although they use host machinery for peptide synthesis, transport and assembly, all RNA viruses nonetheless encode their own viral RNA replicase (or reverse transcriptase in the case of the retroviruses) that lacks the proofreading or error-checking activities common to higher organisms. Viruses also replicate rapidly, with short generation times and large population sizes (typically producing 10^5 copies per 10 hours per virion; Moya *et al.*, 2004). Finally RNA is by its nature less stable than DNA, even in duplex or heteroduplex form. As a result, mutation rates for RNA viruses have been shown to be in the range of 0.1 – 1 mutations per genome per replication (Drake *et al.*, 1998).

This basic rate at which new mutations are generated within host cells is acted upon by various forces to give rise to the observed within- and between- host rates of substitution ("evolutionary rates"). Although a large number of mutations are fatal or strongly deleterious and therefore do not accumulate in virus populations (Sanjuán *et al.*, 2004c), nucleotide fixation is sufficiently frequent for the substitution rates of most RNA viruses to be 10^{-4} to 10^{-2} substitutions per site per year (Jenkins *et al.*, 2002), with even more rapid evolution at some sites (Lemey *et al.*, 2005).

1.2.1.2 Selection

Diversity produced through mutation is shaped by natural selection, which is broadly categorized as positive or negative. Those aspects of host biology that can change rapidly – such as antibody specificity or cytotoxic T-lymphocyte activation – can lead to strong positive or 'diversifying' selection. On the other hand, purifying (or 'negative') selection preserves viral phenotypes that interact with slowly-changing components of the host cellular biological machinery, or structurally constrained aspects of the virus. Purifying and

diversifying forms of selection may act concurrently on a viral genome. For instance, the HCV *core* & *NS3* genes are under greater purifying selection, while the *E1/E2 envelope* genes are under diversifying selection (Salemi & Vandamme, 2002; see also Chapter Three).

Many viral genomes are under weak or absent selection (Jenkins *et al*., 2002) but recent developments in methodology allow us to segregate selection analyses into individual sites with much greater power to resolve selection. In fact, it seems that a large number of sites do not evolve neutrally when examined at the within-host level (Borrow *et al*., 1997; Sanjuán *et al*., 2004c). Selection pressures, particularly those imposed by the immune system, are still a major driving force in viral evolution (Rambaut *et al*., 2004; Moya *et al*., 2004; Farci *et al*., 2000; Borrow *et al*., 1997).

The adaptive and innate components of the host immune system represent strong diversifying selective forces that have probably applied to viruses over most of their evolutionary history. As a result, both virus and host have co-evolved a variety of capabilities or defences that come into play on infection (Nousbaum *et al*., 2000). Most importantly, viral peptides presented to the immune system by cells' MHC complexes, or observed by macrophages or dendritic cells on extracellular virions, are in time recognised by the immune system, leading to B and T cell activation and antibody production. This produces a constant pressure for the virus to change antigen characteristics in order to circumvent an effective immune response (Borrow *et al*., 1997; Kuntzen *et al*., 2007), leading to strong positive selection on those antigen sites that are less-constrained (Ross & Rodgrigo, 2002). Diversifying selection has been observed in hypervariable regions (HVR) in the envelope genes of HIV (Hughes *et al*., 1997) and HCV (Simmonds, 2004). The corollary observation – that viral evolution slows when the immune system is ineffective or collapses – has also been made (Kuntzen *et al*., 2007; Williamson, 2003).

As well as the adaptive immune response, selection also mediates viral diversity through innate host defence mechanisms act to counter HIV infection such as the defence proteins

such as TRIM5α, tetherin and the APOBEC family (Neil *et al*., 2008; Huthoff & Towers, 2008). HIV encodes several accessory proteins that appear to be key to overcoming these innate defences (Malim & Emerman, 2008) and both host and viral genes appear to be co-evolving; for example selection has been documented in both tetherin (McNatt *et al*., 2009) and the viral gene it antagonises, *Vpu* (de Oliveira *et al*., 2004).

Viral genetic diversity may also be maintained as a direct result of fluctuating or frequency-dependent selection (Lorenzo *et al*., 2004; Leslie *et al*., 2004; Rambaut *et al*., 2004; Ross & Rodrigo, 2002). Drug treatments contribute another powerful selection pressure (Kitchen *et al*., 2004) and the evolution of drug resistance is a key area of research (Rambaut *et al*., 2004). As in the population genetics of other pathogens, the anthropogenic introduction of steep selective gradients in the form of effective drug treatments or vaccinations is predictably countered by a diversifying response in the viral population (Polyak *et al*., 1998; Rambaut *et al*., 2004). Substitutions that confer resistance (in the case of treatment; White *et al*., 2004) frequently carry an associated fitness cost (Yen *et al*., 2005), and to a degree the duration of efficacy of these interventions may be a function of the rate at which potentially resistant novel strains are produced (Rambaut *et al*., 2004).

Strong purifying selection acts to conserve many sites along the viral genome (Lemey *et al*., 2005; Williamson, 2003), most frequently those concerned with non-structural peptides encoded by the virus and (in the case of some viruses), the 5' and 3' untranslated regions (UTR) that flank the genome.

1.2.1.3 Recombination

Recombination between viruses occurs when two or more viruses co-infect (or super-infect) a single host cell. This may occur in the course of an infection arising from a single transmission event, or following super-infection by a second inoculate after a second or subsequent transmission event(s), in which case the potential to increase viral diversity will be even greater. Template-switching (by which a viral replication complex disengages from

one strand and restarts synthesis on a separate one) is the proposed mechanism for recombination in non-segmented viruses (such as dengue; Worobey *et al*., 1999). Segmented viruses (such as influenza) are much more likely to recombine simply through the reassortment of segments in a coinfected cell (Steinhauer & Skehel, 2002). Available evidence suggests that recombination is frequent between serotypes in dengue (Worobey & Holmes, 1999), while recombination between subtypes is also common in HIV (*c.f.* Charpentier *et al*., 2006; Worobey & Holmes, 1999; Rhodes *et al*., 2003), and well-documented in influenza (Steinhauer & Skehel, 2002). As well as countering Muller's Ratchet (the accumulation of deleterious mutations over time in asexual organisms (Moya *et al*., 2004; Rambaut *et al*., 2004; Muller, 1964)) and continuously preserving fitness, recombination occasionally results in sudden antigenic shift, with clear implications for host immunity (Seo *et al*., 2004; Steinhauer & Skehel, 2002), drug and vaccine design (Rambaut *et al*., 2004); while very high rates of recombination may actually generate new genotypes more rapidly than substitution alone (Worobey & Holmes, 1999). It has been suggested that recombination may allow viruses to escape Muller's Ratchet.

1.2.1.4 Genetic drift & molecular clocks

Many techniques of molecular evolutionary analysis have been developed against the background of the neutral theory of molecular evolution (Kimura, 1968) – which posits that that the majority of observed amino acid substitutions have no substantial fitness effect (subsequently reformulated as the 'nearly' neutral theory to allow for slightly deleterious mutations; Ohta & Gillespie, 1996). Closely related to the neutral theory is the hypothesis of an evolutionary molecular clock – the prediction that when substitution rates are constant, observed sequence divergence is proportional to time (Zukerkandl & Pauling, 1965). The molecular clock hypothesis has allowed the development of powerful techniques for the simultaneous estimation of divergence times, rates of evolution and other population parameters (Drummond *et al*., 2002). Such methods are described in more detail in sections 1.5.2.1 & 1.5.2.2.

However, given the intense nature of selection by the immune system (Borrow *et al*. 1997, Ross & Rodrigo, 2002) and drug treatments (Cohen & Fauci, 1998) on viral populations, there is no *a priori* reason to suppose that viral evolutionary rates remain constant. Empirical studies suggest that at the population level some viruses appear to evolve neutrally or at a constant rate (Jenkins *et al*., 2004; Elena *et al*., 2003; Allain *et al*., 2000) whereas for others a molecular clock can be rejected (Jenkins *et al*., 2002), posing a challenge to evolutionary analysis. In response the 'relaxed' molecular clock model (where the evolutionary rate is not modelled as a single fixed rate, nor allowed to vary freely across the whole tree as in a 'no-clock' model, but as a discretized distribution from which branches' evolutionary rates are drawn: see Chapter 3) has emerged as an alternative to the strict (constant-rate) molecular clock. The relaxed clock incorporates some variation in substitution rates through time or among lineages, whilst allowing phylogenies to be estimated on a temporal timescale (Drummond *et al*., 2006). An important theme of this thesis will be the detection of departure from the strict clock model and the application of the relaxed molecular clock to the analysis of within-host viral data sets.

1.2.1.5 Population structure / Phylogeography

Phylogeographic structure is population structure that is detectable by phylogenetic means; it occurs when the rate of gene flow among subpopulations is lower than the rate of genetic divergence, such that individual lineages from the same subpopulation are more similar compared with those in other areas than we would expect due to chance (Avise, 2000; *e.g.*; Perrin *et al.,* 2003). The rapid evolutionary rate of RNA viruses therefore allows for the fine-scale detection of population structure at different levels; furthermore, both geographic and demographic signals are reflected in viral phylogenies (Grenfell *et al*., 2004).

Structure is most often considered to arise from spatial segregation, and in this context several studies have used phylogeographic methods to determine the rate of transmission between geographic locations (Cochrane *et al*., 2002) or dispersal rates (Carrington *et al*., 2005). However, in viruses, population structure can also arise through host risk factors and treatment regime at the between-host level (Leigh-Brown *et al*., 1997, Kosakovsky Pond *et al*., 2006), and through tissue type and cell tropism at the within-host level (Salemi *et al*., 2005; Potter *et al*., 2004; Wong *et al*., 1997).

Phylogeographic methods provide techniques to estimate the strength of correlation between these variables and viral genetic diversity. An intuitive approach is to simply examine the relationship between the geographical location of virus RNA samples and their position on a phylogenetic tree. However more sophisticated approaches may quantify the degree of association by indices (Cochrane *et al*., 2002) or parsimony mapping of migration events (Carrington *et al*., 2005; Nakano *et al*., 2004) One of the goals of this thesis is to investigate and expand on these methods.

1.2.1.6 Speciation & Extinction

Comparatively little is known about the speciation and extinction of viruses since they leave no physical evidence in the fossil record. A number of authors have speculated on viral origins *e.g*. whether they have been present since, or even before, DNA-based life (Holmes,

2003). However, accurately estimating divergence times for unrelated groups is currently problematic due to saturation of mutations at the nucleotide level (even in DNA viruses; Lemey *et al*., 2002) while many viral groups appear to have emerged in the last million years (Holmes *et al*., 2003). This presents a paradox, since a number of host viral defence mechanisms are expected to be far more ancient (Sawyer *et al*., 2004). Similarly we only have direct evidence for the extinction of one viral species (smallpox, which was eradicated by man from the wild). Viral extinction is normally expected to occur through host adaptation and resultant reduction in viral transmission (Ariën *et al*., 2005; de Groot *et al*., 2002).

However some retroviruses do leave a genetic footprint on their hosts' genomes in the form of endogenous retroviruses (Katzourakis *et al*., 2007). Although many of these elements are so degraded that their relationship to modern exogenous retroviruses is tenuous (Mi *et al*., 2000), enough are sufficiently well conserved that it is likely that they were once active viruses millions of years ago (Xiong & Eickbush, 1990; Switzer *et al*., 2005). Thus the interplay between micro- and macro-evolutionary processes in RNA virus evolution cannot be ignored, at least for the retroviruses.

1.2.2 Within-patient viral evolution: major research themes

Natural selection must operate on individual virions in the context of the viral population within an infected host (Mayr, 1997). The forces affecting virus evolution are therefore those that govern the life-cycle of each virion; immune evasion, cell entry, replication, virion assembly and release. In the next sections of this chapter I will look at some of the important topics arising from RNA virus evolution within hosts.

1.2.2.1 Quasispecies

One of the most contentious areas of evolutionary virology concerns the application of the 'quasispecies' mathematical model to virus evolution within hosts (Gomez *et al.,* 1999) [1]. The model proposes a mutationally-linked population of replicators, where selection acts to maintain the fitness of the quasispecies rather than that of the individual; consequently the fitness of the quasispecies as a whole would not be identical to the arithmetic mean fitness of the individuals that comprise it. The quasispecies model was first proposed by Eigen & Schuster (1979) in the context of replicator dynamics during the orgin of life. Although the idea has attracted considerable attention (particularly with reference to chronic HCV & HIV infection), the predictions it makes can equally well be understood using standard population genetics theory under very high mutation rates, altering the mutation / selection balance (Wilke, 2005; Holmes, 2002; Jenkins *et al*., 2001; Miralles *et al.,* 1999). Nevertheless, it is still useful to conceive of viral populations within hosts as an ensemble of genotypes and phenotypes with a continuous gradation of fitness.

[1] A distinction is drawn here between the specific evolutionary definition considered above, and the term 'quasispecies' as sometimes used in virological literature to refer to the set of viral genotypes or phenotypes represented in a single infection.

1.2.2.2 Compartmentalization

Viruses exhibit strong tropism for certain cell types, even extending to different cell types within the immune system (Berger *et al.,* 1999; Fear *et al.,* 1998; Wong *et al.*,1997), central nervous system (Salemi *et al.,* 2005) and other tissues (Sobeski *et al.,* 2007). Since this observation was first made, various authors have debated whether identifiably distinct viral subpopulations circulate in different host tissues or cell lines; a phenomenon called 'compartmentalization' (Philpott *et al.*, 2005; Cuevas *et al.,* 2003).

Compartmentalization is of practical relevance as well as theoretical interest. Firstly, the consensus or modal genotype observed in a host would depend on the tissue or compartment sampled (Cabot *et al.,* 2001; 1997), with implications for evolutionary analysis (Cabot *et al.,* 1997). Secondly, modelling of the immune response, and treatment and drug development, would need to take into account the existence of multiple viral phenotypes at any one time (Wong *et al.,* 1997), the diversity of which may also be increased by viral archiving.

In the early 1990's it became apparent that HIV infections were not only able to persist for long periods of time within hosts, but also that certain within-host viral lineages were able to exist at very low or undetectable frequencies only to re-appear later. This phenomenon was termed 'viral archiving' (Nunnari *et al.,* 2005) and thought to be partly a consequence of compartmentalization. Small, isolated, tissue-specific compartments of virus within the host (Baccam *et al.,* 2003) could periodically increase in population size and sweep through the main plasma population when selective conditions became favourable. The implications for treatment and disease progression are clear, since this would suggest that the immune system and drug regimes must continuously target all lineages that have ever been observed within a host – even those that appear to have been cleared – to avoid relapse (Noë *et al.,* 2005; Wong *et al.,* 1997; Finzi *et al.*, 1997).

It remains unclear whether the specific principle of viral archiving must be invoked to explain the resurgence of certain lineages or whether, given the rapid rates of substitution and

recombination in RNA viruses, phenotypes similar to previously-observed lineages can be generated *de novo* from the plasma population under appropriate selective contexts (Potter *et al.,* 2006, Sanjuán *et al.,* 2004a).

Finally, the viral diversity transmitted to new hosts would depend on which compartment provided the infectious titre; for instance the population present in a putative cervical compartment, while representing a minority of virions compared with those present in the plasma, might be more likely to be passed on through sexual transmission (Kemal *et al.,* 2003). The equivalent suggestion has also been made for seminal fluid (Gupta *et al.,* 2000).

The existence of *bona fide* compartmentalization (as opposed to stochastic sampling effects among tissues) is supported by some reports (Pillai *et al.,* 2005; Fulcher *et al.,* 2004; Philpott *et al.,* 2002), but others have been more critical (Choudhury *et al.,* 2002; Alves *et al.,* 2002; Hughes *et al.,* 1998; Delwart *et al.,* 1997; van der Hoek *et al.,* 1998). Given the important consequences for evolutionary virology of the compartmentalization hypothesis, a substantial part of this thesis is dedicated to improving, testing and applying techniques that further our knowledge of this phenomenon in RNA viruses.

1.2.2.4 Immune escape mutations

As already mentioned, viruses are subject to intense selection by the immune system (Desrosiers *et al*., 1999) and the evolution of antibody and cytotoxic T-lymphocyte (CTL) viral escape mutants during HIV infection demonstrates many of the aspects of a co-evolutionary classic evolutionary arms race (*c.f.* Günthard *et al*., 1999; Mammano *et al*., 1998). It is suggested that positive selection mediated by the adaptive immune system drives viral evolution to avoid antigen recognition and hence circumvent the immune response. However since a large number of sites are constrained by purifying selection for other reasons, a predictable sequence of changes occur; the appearance of many of these variants and their timing is thought to be strongly correlated with disease progression (Karlsson *et al*., 2007; Klenerman.*et al*., 2002; Borrow *et al*., 1994). Recent studies suggest that some CTL escape mutations may occur in predictable sequences (Iversen *et al*., 2006) with implications for treatment design (Goulder & Watkins, 2004).

1.2.2.5 Wild-type reversion

In HIV infection (Friedrich *et al*., 2004; Leslie *et al*., 2004) the consensus sequence of a viral population that has acquired substitutions in one host has been observed to lose some of those rapidly after transmission to the next host, within which another and different set of substitutions are acquired. This has been termed 'wild-type' reversion (Herbeck *et al*., 2006) and although sometimes wrongly described as 'evolution in reverse' it most likely represents the loss of specific escape mutations (Kuntzen *et al*., 2007; Crawford *et al*., 2007) or other host specific adaptations. One important question remains the degree to which such escape mutations revert on transmission (Leslie *et al*., 2004). An important consequence of this could be a discrepancy between the rates of substitution as measured between and within hosts, suggesting that the range and time of sampling may influence rate estimation (Herbeck *et al*., 2006).

1.2.2.6 Transmission bottlenecks & founder events

An important feature of viral population genetics is their oscillation between large population sizes at peak viraemia and very low population sizes on transmission that constitute severe population bottlenecks (Rambaut *et al*., 2004; Wahl *et al*., 2002; although effective populations sizes at seroconversion may be lower since hardly any time has elapsed since transmission, so diversity is low; Lemey *et al*., 2006). The result of these repeated bottlenecks events is that viral genotypes present even at high frequencies may be lost from time to time (Edwards *et al*., 2006), since such founder events act to stochastically pick certain lineages, decreasing viral diversity (Zhu *et al*., 1993; Delwart, *et al*., 2002). Founder effects may also lead to decreases in fitness (Domingo *et al*., 1996). The strength of this effect will vary, depending on the viral population diversity and effective size prior to the event, and the size of the population at bottleneck (Novella *et al*., 1996; Wahl *et al*., 2002); the effect of bottlenecks due to transmission is probably most pronounced (Rambaut *et al*., 2004). The strength of the founder effect at transmission on viral phylodynamics is probably also closely correlated with the size of the viral titre on transmission and thus transmission method (Dickover *et al*. 2000).

1.3 Practical consequences of virus evolution

1.3.1 Evolution of virulence and drug resistance

The phenotypic traits of virulence and drug resistance are as mutable as any others. In this context, the rapid evolution and high diversity of RNA viruses is a significant cause for concern.

Virulence – the decrease in host fitness due to a infectious disease – is hypothesized to evolve to an evolutionary optimum that is determined by a trade-off between virulence and transmissibility (Moya *et al*., 2004; Lipsitch & Moxon, 1997). There is evidence that virulence can both increase and decrease over time, as hosts evolve in response to the burden of disease and the pathogens correspondingly reply. Increases in virulence can occur rapidly due to pathogen evolution (Brault *et al*., 2007), although host responses typically occur over longer time-scales. For example, the APOBEC viral defence genes (Harris & Liddament, 2004) represent a host evolutionary response to retrovirus infection, and the long-standing SIV infections in non-human primates are typically non-lethal (Santiago *et al*., 2002). Furthermore, there are some hopeful signs that HIV-1 itself may be decreasing in virulence (Ariën *et al*., 2005) and that human populations may already possess some HIV-1 resistance alleles (Tang & Kaslow, 2003) that could over time increase in frequency.

Human attempts to control the prevalence of pests and pathogens through the use of drugs and other biochemical agents represent an extreme selection pressure on those pathogens to circumvent or mollify those agents. As we might expect, both microbial pathogens and pests have been observed to mount effective evolutionary responses to human efforts at their control. However RNA viruses' rapid substitution rates and high rates of recombination, coupled with the difficulty associated with the development of effective therapies in the first place, mean that the evolution of drug resistance is of particular concern. In particular, HIV has been observed to evolve drug resistance *in vivo* (Hecht *et al*., 1998) and the virus' ability to maintain resistance mutations through repeated transmission events (Yerly *et al*., 1999)

and associated fitness penalties outside of the selective environment (*i.e.* in untreated individuals) is subject to intense scrutiny. Futhermore, the high rates of intra- and inter-subtype recombination in HIV have generated fears that resistance to any novel therapies will not only develop rapidly in treated individuals, but also be spread throughout the global HIV epidemic rapidly through recombination (Cohen & Fauci, 1998). Recombination is also thought to allow the virus to resist several drugs (as in HAART or other combination therapies; Rambaut *et al*., 2004).

1.3.2 Disease progression of chronic viral infections

A growing body of evidence suggests that the nature of viral evolutionary dynamics at the within-host level can be a significant determinant of disease outcome for chronic viral infections (*e.g*. Itakura *et al*., 2005; Ross & Rodrigo, 2002; Farci *et al*., 2000; Ray *et al*., 1999; Shankarappa *et al*., 1999; Wolinsky *et al*., 1996). For instance, in HIV infections it has been suggested certain viral escape mutations are associated with rapid progression to AIDS, while others (following a similar titre of inoculate in equally healthy hosts) lead to long-term non-progression of the HIV infection (Iversen *et al*., 2006). Furthermore, the rapidity with which new viral lineages are generated can vary and seems to be important in progression to AIDS (Williamson *et al*., 2004). Similarly, Sheridan *et al*. (2004), investigating disease progression in HCV, found that patients where viral sequences' sites were more likely to be under strong diversifying selection (which usually is mediated by the host immune system) developed acute symptoms less quickly than patients where selection of the viral quasispecies was less robust. Co-infection with more than one virus can also have implications for host disease outcomes; for instance the selective pressures on evolution of HCV seem to be relaxed, and disease progression to be accelerated, when the host is co-infected with HIV (Mao *et al*., 2001).

1.3.3 Classification of RNA viruses

Taxonomy of the RNA viruses is complicated by their small size, poor sampling and frequent host switching, as well as the absence of a coherent species concept analogous to that underpinning classification of cellular forms of life, yet their rapid rate of evolution presents the biggest challenge to effective classification and description. Even related viruses show considerable divergence, either due to neutral or adaptive evolution; while the construction of deep phylogenies of distantly-related virus groups remains unfinished due to genomic rearrangements, drift, and saturation at the nucleotide level.

However some conventions have developed: below the group level of classification under the Baltimore system (see Section 1.1.2.2), RNA viruses are further classified into families, subfamilies and virus species based mainly on morphology, genome composition and host species range.

Below the species level, viruses are typically classified as subtypes, genotypes, lineages or strains on the basis of genetic similarity and phylogenetic analysis. These classification schemes are typically developed in an *ad hoc* manner and their utility has sometimes been questioned (Rambaut *et al.,* 2004), since designated subtypes sometimes differ little in phenotype and may frequently recombine. Furthermore, subtype differences in recently emerging epidemics may only be the footprint of founder events and liable to recede over time (Rambaut *et al*. 2001).

A comprehensive, coherent and robust system for viral classification below the species level has yet to be adopted, partly due to under-sampling of many viral species and partly due to their genomic plasticity. The subtype / genotype scheme used in many viruses nonetheless provides an intuitive shorthand for viral diversity and continues to be widely employed (*c.f* Choisy *et al.,* 2004; Nakano *et al.,* 2004; Pybus *et al.,* 2001)

1.4 Techniques & methodology

The work presented in this thesis relies on the acquisition of viral sequence data and its analysis through phylogenetic and population genetic methodologies. The basic principles that underlie these techniques are reviewed in this section.

1.4.1 Data acquisition

The studies in this thesis have been carried out on simulated data or on viral RNA genomic sequence data obtained from online sequence databases or from collaborators. It is crucial to know the conditions under which a particular DNA sequence has been isolated. Some RNA sequences available represent the genome of an actual virion that existed at some point in time, whereas others may more closely approximate the consensus of an entire virus population.

1.4.1.1 Sample isolation

Samples of host tissue or fluids are collected either opportunistically when an interesting epidemiological situation presents itself (as with the HCV 'Anti-D' dataset in Chapter Three) or according to a pre-determined sampling strategy (as with the HIV 'compartmentalization' dataset presented in Chapter Four). In either case, the samples must be correctly stored prior to sequencing, and the clinical information relating to the patient of origin must be collected as accurately as possible whilst observing relevant ethical guidelines. Furthermore, many of the analyses in this thesis rely on heterochronous sampling – where samples are collected at different points in time – and for this purpose the date of sample isolation should be ascertained as accurately as possible.

1.4.1.2 Data collection – polymerase chain reaction & sequencing techniques

Viral RNA sequences may be obtained from appropriate samples (specimens) by various reverse transcription and polymerase chain reaction (RT-PCR) methods such as direct RT-PCR and 'bulk' sequencing, or direct RT-PCR followed by clonal sequencing. In both instances RNA is first reverse transcribed into DNA and then the resulting DNA is PCR amplified from solutions using primers complementary to specific regions that are expected to be sufficiently conserved. The resulting PCR products contain heterogenetic amplicons roughly representative of the extracted RNA depending on the efficacy of the primers used and the quality of the initial RNA (degraded RNA may not amplify); the PCR amplicons are then sequenced. A variety of sequencing methods are available (*e.g.* sanger sequncing, pyro sequencing, Big Dye Chem, *etc.*) with different biases, however all detect the most-common alleles more strongly. The finished sequence represents a consensus of all viral genotypes present in the sample.

Although this technique is rapid it is of limited use for studies of viral diversity as it does not indicate which polymorphisms occur on which strand - furthermore it does not give an accurate representation of the polymorphic diversity present in the viral population as polymorphic variants with the same individual sequences will be under-represented.

Alternatively, the PCR product may be cloned to obtain sequences of individual viral RNA genomes ('clonal' sequencing). This is done by inserting the product copies into a carrier bacterial plasmid then culturing colonies of clonally identical individual bacteria from which the original insert may be extracted in sufficiently high numbers to detect by direct sequencing. In this case each sequence obtained reflects the genome of an individual viral sequence.

1.4.1.3 Viral sequence databases

While much of the sequence data used in this thesis was simulated or obtained directly from collaborators, extensive use was also made of viral sequences residing in public databases;

both the generalized sequence databases (NCBI, EMBL and DDBJ) and the specialised viral sequences databases (LANL HCV (Kuiken *et al*., 2005) and HIV[2] databases) that contain additional patient clinical and viral phenotype annotations of interest to virologists. In either case appropriate attributions have been made.

Before commencing analysis of empirical sequence data in this thesis, individual sequences were visually inspected for stop codons or obvious signs of contamination or sequencing errors, checked for recombination and where appropriate, subtyped using the automated recombination and subtype detection methods available in the Oxford / BioAfrica Subtype Tools[3] (de Oliveira *et al*., 2005). Multiple sequence alignment was performed with ClustalX (Jeanmougin *et al*., 1998) and subsequently checked and optimised by hand in Se-Al (Rambaut, 2002). Sequences were frequently rejected during this process. Where sequences were downloaded from public databases, every effort was made to verify sequence annotations from the primary literature (or by directly contacting the individual researchers or groups who had published the sequences).

1.4.2 Phylogenetic analysis

The majority of chapters in this thesis rely heavily on methods of phylogenetic and population genetic estimation, particularly those based on maximum likelihood or Bayesian inference frameworks. The basic principles of these approaches are outlined here.

1.4.2.1 Phylogenetic trees

The estimation of a phylogenetic tree is often a precursor to other forms of viral evolutionary analysis as well as a useful source of information about the ancestral relationships among sequences. They are used in all chapters of this thesis. Following sequence alignment, a tree is estimated that represents a set of hypotheses about the sequence of evolutionary events that

[2] http://www.hiv.lanl.gov/

[3] http://www.bioafrica.net/index.html

led to the sampled distribution of sequences (Felsenstein, 2004). As with any statistical inference, it is important to recognise the assumptions, limitations, and statistical error and power of this process.

Phylogenetic estimation can be considered to have two components: a model of sequence change through time (the nucleotide substitution model) and a method of tree reconstruction. Substitution models are probabilistic Markov processes that expresses the rate and conditions under which one nucleotide or codon may be substituted for another. Many different and related models are available. Two of the most commonly-used models are employed in this thesis, the general time reversible (GTR) model of Lanave *et al*. (1984), which specifies a separate rate for every possible nucleotite substitution, and the Haswgawa-Kishino-Yano (HKY) model of Hasegawa *et al*. (1985), which specifies separate rates for transitions and transversions. These models may be further modified by specifiying a distribution of rates to account for rate variation along the genome (commonly a γ-distribution), and also (or separately) specify a certain proportion of sites as invariant over the phylogeny. All such models make two key assumptions, (i) that each nucleotide or codon changes independently of the others and (ii) that changes from state X to state Y occur at the same rate as from Y to X, that is, the model is time reversible. This assumption allows for our lack of knowledge about the relative nucleotide frequencies of ancestral sequences and has the advantage that it allows for the tree to be arbitrarily re-rooted.

The goodness-of-fit of alternative substitution models may be calculated in a likelihood-ratio test; with degrees of freedom equal to the difference in the number of free parameters between competing models (Huelsenbeck & Crandall, 1997); however the LRT may only be used to evaluate the goodness-of-fit of competing substitution models a single topology. The Akaike Information Criterion (AIC) represents an improvement over the likelihood-ratio test since the fit of distinct tree topology models may also be compared (Posada & Buckley, 2004) and is widely implemented in automated model-selection software (Posada & Crandall, 1998) but may wrongly select complex models over simple ones, leading to over-

parametization; newer Bayesian methods of model comparison are less susceptible to this bias (Alfaro & Huelsenbeck, 2006).

The relative plausibility of any specific phylogenetic reconstruction (compared to other possible reconstructions) is assessed by computing the likelihood of the phylogeny - the probability of the specified tree topology and branch lengths, given the data (a set of aligned sequences) and the nucleotide substitution model (Felsenstein, 1981b). Such values are typically expressed in log-likelihood units since the probabilities involved are tiny. Since a very large number of reconstructions are possible for any large set of sequences, it is important that space of all possible trees is searched efficiently when searching for the maximum likelihood tree. The estimated trees presented in this thesis were all constructed either by the neighbour-joining method (when speed is more important than accuracy; Saitou & Nei, 1987) or heuristic maximum likelihood search (when accuracy is more important; Felsenstein, 1981b).

Once a phylogeny has been estimated it is important to estimate its statistical robustness. A common measure of the confidence in the topology of a tree is obtained by the bootstrapping procedure, whereby tree estimated is repeated on a large number of pseudoreplicate sequence alignments, each generated by randomly sampling the original nucleotides sites with replacement (Felsenstein, 1985). The robustness of a node in the original tree is assessed by counting the proportion of bootstrap trees that concur with the grouping at that node.

However a limitation of the single-estimated tree approach is that, at best, bootstrapping can only report the support for clusters in the original topology and cannot not compare alternative tree topologies (Hillis & Bull (1993); Alfaro & Huelsenbeck, 2003).

Once a phylogenetic tree has been reconstructed a variety of 'post-phylogenetic' analyses are possible. The following list provides a short introduction to those most relevant to this thesis:

(1) A distribution of character traits (which may be phenotypic or geographic) may be overlaid on the tips of the phylogeny, and their putative ancestral distribution reconstructed using parsimony or other methods, in order to test hypotheses regarding trait change and evolution (as in phylogeography, *cf*. Carrington *et al*., 2005; Slatkin

(2) Model testing techniques (e.g. LRTs or the AIC) can be employed on single estimated phylogenies to assess different substitution models (Huelsenbeck & Crandall, 1997; Yang, 1997), to test different models of selection (Yang, 1997), or to test different molecular clock models (*e.g*. Jenkins *et al*., 2002).

(3) Inference of effective population size on single phylogenies is possible using the skyline plot (Strimmer & Pybus, 2001) or other methods based on coalescent theory. The coalescent model (Kingman, 1981a, b) provides a powerful framework under which the relationship between population dynamics and phylogenetic history can be explored. Given a neutrally evolving population, the date of the most recent common ancestor of a pair of sequences from a sparsely-sampled subset of the population is expected to be proportional to their divergence, substitution rate and effective population size (Tavaré *et al*., 1997).

(4) The evolutionary rate of dated heterochronous sequences can be estimated by linear regression of root-to-tip genetic distances against sampling time, (Li *et al*., 1988), or, more accurately by maximum likelihood under a phylogenetic model that accounts for heterochronous sampling (Rambaut, 2000).

However, all the analyses listed above are subject to error as they depend on a single phylogeny and therefore ignore the statistical uncertainty inherent in phylogeny estimation (Pie, 2006).

1.4.2.2 Bayesian Markov-chain Monte Carlo (MCMC) techniques

Despite the liabilities inherent in the single-tree approach (Pie, 2006; Holder & Lewis, 2003), the framework described above has been in widespread use for many years due to its computational tractability. However, recent years have seen the development of a new class of phylogenetic methods based on Bayesian inference in an MCMC framework (Huelsenbeck

& Ronquist, 2001; Drummond & Rambaut, 2007). Despite some theoretical concerns (debated by Mossel & Vigoda, 2005; Ronquist *et al.,* 2006; and Mossel & Vigoda, 2006), these methods have been shown to improve on likelihood-based single-tree approaches when applied to empirical data (Ho *et al.,* 2005a) and such methods form the mainstay of this thesis. Although a number of principles involved are identical to those used in likelihood-based single-tree approaches, Bayesian MCMC approaches differ in some key respects which are outlined here.

Bayesian MCMC approaches are based on Bayes' Theorem, which in a simplified form can be written as:

$$\Pr(\Phi \mid D) \propto \Pr(D \mid \Phi).\Pr(\Phi) \quad (1.1)$$

where $Pr(\Phi|D)$ is the posterior probability density of the model or hypothesis Φ. given the data D. This is proportional to the product of $Pr(D|\Phi)$, the probability of the data given Φ , and $\Pr(\Phi)$, the prior probability density of Φ. In an analysis where a coalescent population model is used (coalescent Bayesian MCMC), such as BEAST, the equation in (1.1) is expanded to give:

$$\Pr(\mu,g,\Theta \mid D) \propto \Pr(D \mid \mu,g,\Theta).\Pr(\mu,g,\Theta) \quad (1.2)$$

where the posterior probability is proportional to the product of the likelihood of the molecular clock substitution model (μ), the tree model (g) and coalescent (Θ) model, given the data, D, and the joint prior probability density of (μ, g, Θ). In this context, the coalescent model can be thought of as a prior distribution on phylogenetic topology and branch lengths. The original Kingman coalescent can be extended to any number of models of population size change (e.g. constant-size, exponential growth etc. (Tavaré, 1997). Most usefully, the Bayesian skyline plot defines a very flexible model which provides a 'non-parametric' estimate of population sizes-through-time (Drummond *et al.,* 2005).

In Bayesian evolutionary analysis, an MCMC algorithm is used to draw samples randomly from ($Pr(D|\Phi)$. $Pr(\Phi)$), thereby generating a posterior distribution that is directly

proportional to $Pr(\Phi|D)$. The MCMC algorithm performs a very large number of steps in a MCMC 'chain'; for each step a new phylogeny and other new model parameter values are proposed and accepted if they represent an improvement in probability over the existing values. Values may also be probabilistically accepted or rejected if the new state is worse – this helps to avoid local optima. The MCMC chain is said to have converged to equilibrium when the MCMC algorithm samples the equilibrium posterior distribution (since states are not independent but connected the effective sample size (ESS) is used as a measure of the equivalent number of independent draws from the posterior). At this point ('stationarity'), the tree topologies and model parameters sampled will be drawn from the posterior distribution in proportion to their probability; if a set of taxa appear as a monophyletic clade in 88% of our samples, that is equivalent to saying the probability that they are really monophyletic is 0.88 (Ronquist *et al*. in: eds. Lemey *et al*., 2009; Felsenstein, 2004). This useful property has been exploited before to perform phylogenetic inference. For example, the probability of topologies in a majority-rule consensus tree constructed from the posterior set of trees (PST) produced by the MCMC sampler will be proportional to their frequency in the PST. Similarly, this consequence underpins the novel phylogeny-trait association framework developed in Chapter Two and Chapter Five.

Other informative prior distributions that represent our knowledge about the dataset may be included. For instance, if we have information about the time of transmission of a within-host infection (for instance, in the case of vertical transmission), the tMRCA of the tree might be represented as a uniform or normal distribution centred on the most likely value, between two possible extremes. The likelihood of the substitution and tree models are calculated as the probability of each given the observed sequence data. Nucleotide substitutions are modelled as described above for single-tree methods. There may be a single substitution model, separate substitution models for different codon positions, or for different sections of the alignment. The rate or 'clock' model describes the distribution of rates on the branches of the tree. In the simplest case, a single substitution rate is fitted to every branch on the tree (corresponding to the strict molecular clock hypothesis). These 'strict clock' models were

widely popular in pioneering phylogenetic studies since their simplicity allowed for easy implementation and computational efficiency. However, they have been shown to be biologically unrealistic in a variety of contexts, from vertebrate evolution over geological time-scales to studies of virus evolution (Bromham & Penny, 2003). Alternatively, 'relaxed clock' models allow for individual branches' substitution rates to vary, the substitution rates being drawn from a log-normal or exponential distribution with a mean equal to the overall substitution rate (Drummond *et al*., 2006). Relaxed clock models are intermediate between the strict molecular clock model and the 'no-clock' model of evolution; they allow for more biologically realistic variations in the rate of evolution over time while also incorporating sequences' sampling date information to allow for simultaneous estimation of evolutionary and demographic parameters; Drummond *et al*., 2006; Ho *et al*., 2005. However, they cannot currently accommodate all substitution rate variations that can be envisaged: for instance, a strongly bimodal rate distribution might occur where a phylogeny contains two lineages under differing constraints. 'No-clock' models provide greater flexibility by fitting substitution rates separately and independently to each branch in the phylogeny; however the much larger number of parameters carries with it the attendant risk of over-parametization, and discard heterechoronous time data that can be used to aid inference. Finally the likelihood of the tree model is simply the likelihood of the phylogenetic tree at that state given the sequence data.

The advantages of the Bayesian approach (reviewed in Holder & Lewis, 2003) are that it is computationally simpler to implement alternative hypotheses and models; furthermore relevant *a priori* assumptions can be included as proper Bayesian priors, increasing the power of the analysis. Secondly, the posterior distributions of model parameters provide confidence information on the shape (including skew) of posterior probabilities in the data, not just a single *p*-value. Finally, rather than attempting to identify a single 'best' tree, MCMC sampling algorithms necessarily identify many possible topologies, where each possible topology is sampled in direct proportion to the strength of phylogenetic signal that supports it

(the posterior distribution of trees). Thus the phylogenetic error that plagues single-tree approaches is accommodated, achieving greater accuracy (Ho *et al.,* 2005a).

However the techniques of Bayesian inference are still comparatively new and pose some challenges of their own (Bollback, 2002; Suchard & Weiss, 2001). Firstly, the MCMC process must be run for several million states and a number of independent times in order to be sure that the MCMC chain has converged on the optimal area of the posterior and not become trapped in local optima (Holder & Lewis, 2003). This can be computationally expensive for large data sets. Secondly, it has become clear during the preparation of this thesis that model selection procedures for this type of analysis are under-developed. Recently the work of Suchard & Weiss (2001) in applying the approximate harmonic mean likelihood estimator of posterior goodness-of-fit as a means to compare Bayesian MCMC analyses has begun to be used in preference to likelihood ratio or AIC tests of the posterior likelihood, due to apparent bias in the latter (Alfaro & Huelsenbeck, 2006). However, this measure may itself be subject to bias, and it has yet to be fully evaluated. Furthermore, rigorous model selection procedures within this framework are extremely computationally intensive, as a large number of possible model combinations must each be analyzed several times to ensure convergence. Finally, appropriate prior selection may be problematic and require approximation (*e.g.* the transmission time between two hosts might be modelled as a uniform distribution over probable dates where the exact timing is uncertain.)

1.5 Thesis outline

In the following chapters, I develop the themes explored here. I apply Bayesian MCMC methods to empirical and simulated data, exploring the coherence and tractability of currently-available model selection procedures. I develop methodological approaches that take advantage of the opportunities offered by Bayesian MCMC analyses, principally the posterior set of trees, and apply these methods to the analysis of evolution within and between patients. In Chapter Two I present a new methodological framework for the estimation of the strength of the association between phenotypic trait values and lineages in a phylogeny. In Chapter Three I carry out a detailed analysis of substitution rates in hepatitis C virus, an important viral pathogen. This analysis combines comprehensive Bayesian model testing, whole-genome analysis in discrete windows and relaxed-clock models of evolution. The approach developed in Chapter Two is applied, along with a combination of other novel and existing methods, to examine the evidence for compartmentalization of HIV virus in the cervix. Having demonstrated the practical utility of the phylogeny-trait association framework developed in Chapter Two and applied in Chapter Four, the final research chapter (Chapter Five) further explores the properties of the statistics implemented. My conclusions are presented in Chapter Six; the Appendices relate to other projects in which I have been able to make a significant contribution.

1.6 References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. & Walter, P. (2002) *The molecular biology of the cell (4th Ed.)* Garland Science, New York and London.

Alfaro, M.E. & Huelsenbeck, J.P. (2006). Comparative performance of Bayesian and AIC-based measures of phylogenetic model uncertainty. *Syst. Biol.* **55**(1):89-96.

Altman, J. D. & Feinberg M. B. (2004) HIV escape: there and back again. *Nat. Med.* **10**(3):229-230.

Alves, K., Canzian, M., Delwart, E.L. (2002). HIV type 1 envelope quasispecies in the thymus and lymph nodes of AIDS patients. *AIDS Research and Human Retroviruses,* **18** (2): 161-165.

Ariën, K. K., Troyer, R. M., Gali, Y., Coleblunders, R. L., Arts, E. J. & Vanham, G. (2005) Replicative fitness of historical and recent HIV-1 isolates suggests HIV-1 attenuation over time. *AIDS* **19**:1555-1564.

Armitage, A. E., Katzourakis, A., de Oliveira, T., Welch, J. J., Belshaw, R., Bishop, K. N., Kramer, B., McMichael, A. J., Rambaut, A. & Iversen, A. K. conserved footprints of APOBEC3G on hypermutated HIV-1 and HERV-K(HML2) sequences. *J.Virol.* **82**(17):8743-8761.

Avise, J.C. (2000). *Phylogeography: The History and Formation of Species*. Harvard University Press, Cambridge, MA. 447pp.

Baccam, P., Thompson, R. J., Li, Y., Sparks, W. O., Belshan, M., Dorman, K. S., Wannemuchler, Y., Oaks, J. L., Cornette, J. L. & Carpenter, S. (2003) Subpopulations of equine infectious anemia virus *rev* coexist *in vivo* and differ in phenotype. *J. Virol.* **77**(22):12122-12131.

Belshaw, R., Gardner, A., Rambaut, A. & Pybus, O. G. (2008) Pacing a small cage: mutation and RNA viruses. *TREE* **23**(4):188-193.

Belshaw, R., Pybus, O. G. & Rambaut, A. (2007) The evolution of genome compression and genomic novelty in RNA viruses. *Genome Res.* **17**:1496-1504.

Berger, E. A., Murphy, P. A. & Farber, J. M. (1999) Chemokine receptors as HIV-1 coreceptors: Roles in viral entry, tropism and disease. *Annu. Rev. Immunol.* **17**:657-700.

Bollback, J. P. (2002) Bayesian model adequacy and choice in phylogenetics. *Mol. Biol. Evol.* **19**(7):1171-1180.

Borrow, P., Lewicki, H., Hahn, B. H., Shaw, G. M. & Oldstone, M. B. (1994) Virus-specific $CD8^{=}$ cytotoxic T-lymphocyte activity associated with control of viraemia in primary human immunodeficiency virus Type-1 infection. *J. Virol.* **68**:6103-6110

Borrow, P., Lewicki, H., Wei, X., Horwitz, M. S., Peffer, N., Meyers, H., Nelson, J. A., Gairin, J. E., Hahn, B. H., Oldstone, M. B. & Shaw, G. M. (1997) Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. *Nat. Med.* **3**(2):205-211.

Brault, A. C., Huang, C. Y-H., Langevin, S. A., Kinney, R. M., Bowen, R. A., Ramey, W. N., Panella, N. A., Holmes, E. C., Powers, A. M. & Miller, B. R. (2007) A single positively selected West Nile viral mutation confers increased viriogenesis in American crows. *Nat. Genet.* **39**(9):1162-1166.

Bromham, L. & Penny, D. C. (2003) The modern molecular clock. *Nat. Rev. Genet.* **4**:216-224.

Cabot *et al* (2001) Longditudinal evaluation of the structure of replicating and circulating hepatitis C virus quasispecies in nonprogressive chronic hepatitis C patients. *J Virol,* **75:**12005-12013.

Cabot *et al*. (1996) Structure of replicating Hepatitis C virus (HCV) quasispecies in the liver may not be reflected by analysis of circlulating HCV virions. *J Virol***. 71**:1732-1734

Caragounis, E. C., Gisslén, M., Lindh, M., Nordborg, C., Westergren, S., Hagber, L., Svennerholm, B. (2008) Comparison of HIV-1 *pol* and *env* sequences of blood, CSF, brain and spleen isolates collected ante-mortem and post-mortem. *Acta. Neurol. Scand.* **117**:108-116.

Carrington, C.V.F., Foster, J.E., Pybus, O.G., Bennett, S.N. & Holmes, E.C. (2005). Invasion and maintenance of Dengue Virus Type 2 and Type 4 in the Americas. *J. Virol.* **79**(23): 14680-14687.

Charpentier, C., Nora, T., Tenaillon, O., Clavel, F. & Hance, A. J. (2006) Extensive recombination among human immunodeficiency virus Type 1 quasispecies makes an important contribution to viral diversity in individual patients. *J. Virol*. **80**(5):2472-2482.

Choisy, M., Woelk, C. M., Guégan, J.-L. & Robertson, D. L. (2004) Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. *J. Virol* **78**(4):1962-1970.

Choudhury, B., Pillay, D., Taylor, S. & Cane, P. A. (2002) Analysis of HIV-1 variation in blood and semen during treatment and treatment interruption. *J. Med. Virol*. **68**:467-472.

Cochrane, A., Searle, B., Hardie, A., Robertson, R., Delahooke, T., Cameron, S., Tedder, R.S., Dusheiko, G.M., de Lamballerie, X. & Simmonds, P. (2002). A genetic analysis of Hepatitis C Virus transmission between injection drug users. *J. Infect. Dis*. **186**:1212-21.

Cohen, O. J. & Fauci, A. S. (1998) Transmission of multidrug-resistant human immunodeficiency virus – the wake-up call. *NEJM* **339**:341-343.

Crawford, H., Prado, J. G., Leslie, A., Hué, S., Honeyborne, I., Reddy, S., van der Stok, M., Mncube, Z., Brander, C., Rousseau, C., Mullins, J. I., Kaslow, R., Goepfert, P., Allen, S., Hunter, E., Mulenga, J., Kiepiela, P., Walker, B. D. & Goulder, P. J. R. (20**07**) Compensatory mutation partially restores fitness and delays reversion of escape mutation within the immunodominant HLA-B*5703-restricted *gag* epitope in chronic HIV-1 infection. *J. Virol* **81**(15):8346-8351.

Crotty, S., Cameron, C. E. & Andino, R. (2001) RNA virus error catastrophe: direct molecular test by using ribavirin. *PNAS* **98**(2):6895-6900.

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

Curran, J. W., Jaffe, H. W., Hardy, A. M., Morgan, W. M., Selik, R. M. & Dondero, T. J. (1988) Epidemiology of HIV infection and AIDS in the United States. *Science* **239**(4840):610-616.

de Groot, N. G., Otting, N., Doxiadis, G. G. M., Balla-Jhagjhoorsingh, S. S., Heeny, J. L., van Rood, J. J., Gagneux, P. & Bontrop, R. E. (2002) Evidence for an ancient selective sweep in the MHC class I gene repertoire of chimpanzees. *PNAS* **99**(18):11748-11753.

de Oliveira, T., Salemi, M., Gordon, M., Vandamme, A-M., van Rensburg, E. J., Engelbrecht, S., Coovadia, H. M. & Cassol, S. (2004). Mapping sites of positive selection and amino acid diversification in the HIV genome: an alternative approach to vaccine design? *Genetics* **167**:1047-1058.

de Oliveira, T., Deforche, K., Cassol, S., Salminem, M., Paraskevis, D., Seegregts, C., Snoek, J., van Rensburg, E. J., Wensing,A. M. J., van de Vijver, D. A., Boucher, C. A., Camacho, R. & Vandamme, A.-M. (2005) An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics* **21**(19):3797-3800.

Delwart, E., Magierowska, M., Royz, M., Foley, B., Peddada, L., Smith, R., Heldebrant, C., Conrad, A. & Busch, M. (2002) Homogeneous quasispecies in 16 out of 17 individuals during very early HIV-1 primary infection. *AIDS* **16**(2):189-195.

Desrosiers, R. C. (1999) Strategies used by human immunodeficiency virus that allow persistent viral replication. *Nat. Med.* **5**:723-725

Dickover, R. E., Garratty, E. M., Plaeger, S. & Bryson, Y. J. (2000) Perinatal transmission of major, minor and multiple maternal human immunodeficiency irus Type 1 variants *in utero* and intrapartum. *J. Virol.* **75**(5):2194-2203.

Domingo, E., Escarmís, C., Sevilla, N., Moya, A., Elena, S. F., Quer, J., Novella, I. S., & Holland, J. J. (1996) Basic concepts in RNA virus evolution. *FASEB J.* **10**:859-864.

Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. (1998) Rates of spontaneous mutation. *Genetics* **148**:1667-1686.

Drucker, E., Alcabes, P. G. & Marx, P. A. (2001) The injection century: massive unsterile injections and the emergence of human pathogens. *Lancet* **358**:1989-1992.

Drummond, A. J. & Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**:214-226.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G. & Solomon, W. (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**:1307-1320.

Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R. & Rodrigo, A. G. (2003) measurably evolving populations. *TREE* **18**(9): 481-488.

Drummond, A. J., Rambaut, A. , Shapiro, B. & Pybus, O. G. (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22**(5):1185-1192.

Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. (2004b) Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22**(5):1185-1192.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J. & Rambaut, A. (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biology* **4**:e88.

Edwards, C. T. T., Holmes, E. C., Wilson, D. J., Viscidi, R. P., Abrams, E. J., Phillips, R. E. & Drummond, A. J. (2006) Population genetic estimation of the loss of genetic diversity during horizontal transmission of HIV-1. *BMC Evol. Biol.* **6**:28-38.

Eigen, M. & Schuster, P. (1979) *The hypercycle: a principle of natural self-organization*. Springer, Berlin.

Elena, S. F., Codoñer, F. M., & Sanjuán, R. (2003) Intraclonal variation in RNA viruses: generation, maintenance & consequences. *Biol. J. Linn. Soc.* **79**:17-26.
Farci, P., Shimoda, A., Coiana, A., Diaz, G., Peddis, G., Melpolder, J.C., Strazzera, A., Chien, D.Y., Munoz, S.J., Balestrieri, A., Purcell, R.H., Alter, H.J. (2000) The outcome of acute hepatitis C predicted by the evolution of the viral quasispecies. *Science* **288**:339-344.

Fauquet, C. M., Mayo, M. A., Maniloff, J., Desselberger, U. & Ball, L. A. (ed) (2005) *Virus taxonomy: Eighth report of the International Committee on Taxonomy of Viruses*. Elsevier Academic, London.

Fear, W. R. , Kesson, A. M., Naif, H., Lynch, G. W. & Cunningham, A. L. (1998) Differential tropism and chemokine receptor expression of human immunodefficiency virus Type 1 in neonatal monocytes, monocyte-derived macrophages, and placental macrophages. *J. Virol.* **72**(2):1334-1344.

Felsenstein, J. (1981b) Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17**:368-376.

Felsenstein, J. (1985) Confidence intervalson phylogenies: an approach using the bootstrap. *Evolution* **39**:783-791.

Felsenstein, J. (2004) *Inferring phylogenies*. Sinauer, Massachusetts, USA.

Finzi, D., Hermankova, M., Pierson, T., Carruth, L. M., Buck, C., Chaisso, R. E., Quinn, T. C., Chadwick, K., Margolick, J., Brookmeyer, R., Gallant, J., Markowitz, M., Ho, D. D., Richman, D. D. & Siciliano, R. F. (1997) Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278**:1295-1300.

Friedrich, T. C., Dodds, E. J., Yant, L. J., Vojnov, L., Rudersorf, R., Cullen, C., Eans, D. T., Desrosiers, R. C., Mothé, B. R., Sidney, J., Sette, A., Kunstman, K., Wolinky, S., Piatak, M., Lifson, J., Hughes, A. L., Wilson, N., O'Connor, D. H. & Watkins, D. I. (2004) Reversion of CTL escape-variant immunodeficienct viruses *in vivo*. *Nat. Med*. **10**(3):275-281.

Friedrich, T., Dodds, E., Yant, L., ojnov, L., Rudersdorf, R., Cullen, C., Evans, D., Desrosiers, R., Mothe, B., Sidney, J., Sette, A., Kunstmann, K., Wolinsky, S., Piatak, M., Lifson, J., Wilson, Froissart, R., Roze, D., Gailbert, L, Blanc, S. & Michalakis, Y. (2005). Recombination every day: Abundant recombination in a virus during a single multicellular host infection. *PLoS Biol*., **3**(3):389-395.

Froissart, R., Roze, D., Uzest, M., Gailbert, L., Blanc, S. & Michalakis, Y. (2005) Recombination every day: abundant recombination in a virus during a single multi-cellular host infection. *PLoS Biol*. **3**(3):389-395.

Fulcher, J. A., Hwangbo, Y., Zioni, R., Nickle, D., Lin, X., Heath, L., Mullins, J. I., Corey, L. & Zhu, T. (2004) Compartmentalization of human immunodeficiency virus Type 1 between blood monocytes and $CD4^{=}$ T cells during infection. *J. Virol*. **78**(15):7883-7893.

Futuyama, D. J. (1998) *Evolutionary Biology, 3rd ed*. Sinauer, NY.

Gaydos, J. C., Top, F. H., Hodder, R. A. & Russell, P. K. (2006) Swine influenza A outbreak, Fort Dix, New Jersey, 1976. *Emerg. Infect. Dis*. **12**(1):23-28.

Gibbs, E. P. J. (2005) emerging zoonotic epidemics in the interconnected global community. *Vet. Rec*. **157**:673-679.

Gitlin, L. & Andino, R. (2003) Nucleic-acid based immune system: the antiviral potential of mammalian RNA silencing. *J. Virol*. **77**(13):7159-7165.

Gómez, J., Martell, M., Quer, B., Cabot, J. & Esteban, J. I. (1999) Hepatitis C viral quasispecies. *J. Vir. Hepatitis*. **6**:3-16.

Goulder, P. J. R. & Watkins, D. I. (2004) HIV & SIV CTL escape: implications for vaccine design. *Nat. Rev. Immun*. **4**(8):630-640.

Goulder, P. J., Phillips, R. E., Colbert, R. A., McAdam, S., Ogg, G., Nowak, M. A., Giagrande, P., Luzzi, G., Morgan, B., Edwards, A., McMichael, A. J. & Rowland-Jones, S. (1997) Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS. *Nat. Med*. **3**(2):212-217.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A. & Holmes, E. C. (2004) Uniting the epidemiological and evolutionar dynamics of pathogens. *Science* **303**:327-332.

Gubler, D. J. (2006) Dengue / dengue haemorrhagic fever: history and current status. *Novartis Found. Symp*. **277**:3-16.

Günthard, H. F., Leigh-Brown, A. J., D'Aquila, R. T., Johnson, V. A., Kuritzkes, D. R., Richman, D. D. & Wong, J. K. (1999) Higher selection pressure from antiretroviral drugs *in vivo* results in increased evolutionary distance in HIV-1 *pol*. *Virology* **259**:154-165.

Gupta, P., Leroux, C., Patterson, B. K., Kingsley, L., Rinaldo, C., Ding, M., Chen, Y., Kulka, K., Buchanan, W., McKeon, B. & Montelaro, R. (2000) Human immunodeficiency virus Type 1 shedding pattern in semen correlates with the compartmentalization of viral quasi species between blood and semen. *J. Infect. Dis*. **182**:79-87.

Hahn, B. H., Shaw, G. M., de Cock, K. M. & Sharp, P. M. (2000) AIDS as a zoonosis: scientific and public health implications. *Science* **287**:607-614.

Halkett, F., Simon, J-C. & Balloux, F. (2005) Tackling the population genetics of clonal and partially clonal organisms. *TREE*, **20**(4):194-201

Harris, R. S. & Liddament, M. T. (2004) Retroviral restriction by APOBEC proteins. *Nat. Rev. Immun*. **4**:868-877.

Harris, R. S., Bishop K., Sheehy, A., Craig, H., et al (2002) RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. Mol Cell 10:1247-1253.

Harvey, P. H., & Pagel., M. D. (1991) *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford.

Hasegawa, M., Kishino, H. & Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**:132-147.

Hecht FM, Grant RM, Petropoulos CJ, *et al*. (1998) Sexual transmission of an HIV-1 variant resistant to multiple reverse-transcriptase and protease inhibitors. *N. Engl. J. Med.* **339**:307-311.

Herbeck, J. T., Nickle, D. C., Learn, G. H., Gottlieb, G. S., Curlin, M. E., Heath, L. & Mullins, J. I. (2006) Human immunodeficiency virus Type 1 *env* evolves towards ancestral states upon transmission to a new host. *J. Virol.* **80**(4):1637-1644.

Hillis, D. M. & Bull, J. J. (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* **42**(2):182-192.

Ho, S. Y. W., Philips, M. J., Cooper, A. & Drummond, A. J. (2005) Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol. Biol. Evol.* **22**(7):1561-1568.

Ho, S. Y. W., Phillips, M. J., Drummond, A. J. & Cooper, A. (2005) Accuracy of rate estimation using relaxed-clock methods with a critical focus on the early metazoan radiation. *Mol. Biol. Evol.* **22**(5):1355-1363.

Holder, M. & Lewis, P. O. (2003) Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.* **4**:275-284.

Holmes, E. C. & Moya, A. (2002) Is the quasispecies concept relevant to RNA viruses? *J. Virol.* **76**(1):460-462.

Holmes, E. C. (2003a) Molecular clocks and the puzzle of RNA virus origins. *J. Virol* **77**(7):3893-3897.

Holmes, E. C. (2003b) Error thresholds and the constraints to RNA virus evolution. *Trends Microbiol.* **11**(12):543-546.

Holmes, E. C. (2004) The phylogeography of human viruses. *Mol. Ecol.* **13**:745-756.

Huelsenbeck, J. P. & Crandall, K. A. (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu. Rev. Ecol. Syst.* **28**:437-66.

Huelsenbeck, J.P., & Ronquist., F. (2001). MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* **17**:754-755.

Hughes, E. S., Bell, J. E. & Simmonds, P. (1997) Investigation of population diversity of human immunodeficiency virus type 1 *in vivo* by nucleotide sequencing and length polymorphism analysis of the V1/V2 hypervariable region of *env*. *J. Gen. Virol.* **78**:2871-2882.

Huthoff, H. & Towers, G.J. (2008). Restriction of retroviral replication by APOBEC3G/F and TRIM5α. *Trends Microbiol.* **16**(12):612-619.

Itakuara, J., Nagayama, K., Enomoto, N., Hamano, K., Sakamoto, N., Fanning, L.J., Kenny-Walsh, E., Shanahan, F. & Watanabe, M. (2005) Viral load change and sequential evolution of entire hepatitis C viral genome in Irish recipients of single source-contaminated anti-D immunoglobulin. *Journal of Viral Hepatitis*, **12**:594-603.

Iversen, A. K. N., Stewart-jones, G., Learn, G. H., Cristie, N., Sylvester-Hviid, C., Armitage, A. E., Kaul, R., Beattie, T., Lee, J. K., Li, Y., Chotiyarnwong, P., Dong, T., Xu, X., Luscher, M., A., MacDonald, K., Ullum, H., Klarlund-Pedersen, B., Skinhøj, P., Fugger, L., Buus, S., Millins, J. I., Jones, E. Y., van der Merwe, P. A. & McMichael, A. J. (2006) Conflicting selective forces affect T cell receptor contacts in an immunodominant human immunodeficiency virus epitope. *Nat. Immu.* **7**(2):179-189.

Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G. & Gibson, T. J. (1998) Multiple sequence alignment with Clustal X. *Trends Biochem Sci*, **23**, 403-405.

Jenkins, G. M., Worobey, M., Woel, C. H. & Holmes, E. C. (2001) Evidence for the non-quasispecies evolution of RNA viruses. *Mol. Biol. Evol.* **18**:987-994.

Jenkins, G.M., Rambaut, A., Pybus, O.G. & Holmes, E.C. (2002) Rates of molecular evolution in RNA viruses: A quantitative phylogenetic analysis. J. *Mol. Evol.* **54**:156-165.

Joint United Nations Programme on HIV/AIDS (UNAIDS) & World Health Organisation (WHO) (2007) Aids Epidemic Update: December 2007. *UNAIDS,* Geneva.

Jou, W. M., Hageman, G., Ysebaert, M. & Fiers, W. (1972) Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* **237**:82-88.

Karlsson, A. C., Iversen, A. I., Chapman, J. M., de Oliveira, T., Spotts, G., McMichael, A. J., Davenport, M. P., Hecht, F. M. & Nixon, D. F. (2007) Sequential broadening of CTL responses in early HIV-1 infection is associated with viral escape. *PLoS ONE* **2**(2):e225.

Katzourakis A, Tristem M, Pybus OG & Gifford RJ (2007) Discovery and analysis of the first endogenous lentivirus. *Proc Natl Acad Sci USA* **104**:6261-6265.

Kemal, K. S., Foley, B., Burger, H., Anastos, K., Minkoff, H., Kitchen, C., Philpott, S. M., Gao, W., Robison, E., Holman, S., Dehner, C., Beck, S., Meyer III, W. M., Landay, A., Kovacs, A., Bremer, J. & Weiser, B. (2003) HIV-1 in genital tract and plasma of women: compartmentalization of viral sequences, coreceptor usage and glycosylation. *PNAS* **100**(22):12972-12977.

Killbourne, E.D. (2006) Influenza pandemics of the 20th century. *Emerg. Infect. Dis.* **12**(1):9-14.

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

Kingman, J. F. C. (1982a) The coalescent. *Stoch. Proc. App.* **13**:235-248.

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

Kitchen, C. M. R., Philpott, S. R. , Burger, H., Weiser, B., Anastos, K., & Suchard, M. A. (2004) Evolution of human immunodeficiency virus Type 1 coreceptor usage during antiretroviral therapy: a Bayesian approach., *J. Virol.* **78**(20):11296-11302.

Klenerman, P., Wu, Y. & Phillips, R. (2002) HIV: Current opinion in escapology. *Curr. Op. Microbiol.* **5**:408-413.

Koenig, S., Conley, A. J., Brewah, Y. A. , Jones, G. M., Leath, S., Boots, L. J., Davey, V., Pantaleo, G., Demarest, J. F., Carter, C. *et al.* (1995) Transfer of HIV-1 specific cytotoxic T lymphocytes to an AIDS patient leads to selection for mutant HIV variants and subsequent disease progression. *Nat. Med.* **1**(4):330-336.

Korber, B., Kunstman, K.J., Patterson, B.K., Furtado, M., McEvilly, M.M., Levy, R. & Wolinsky, S.M. (1994). Genetic differences between blood- and brain-derived viral sequences from human immunodeficiency virus Type 1-infected patients: evidence of conserved elements in the V3 region of the envelope protein of brain-derived sequences. *J. Virol*. **68**(11):7464-7481.

Kosakovsky Pond, S. L., Frost, S. W., Grossman, Z., Gravenor, M. B., Richman, D. D. & Leigh Brown, A. J. (2006) Adaptation to different populations by HIV-1 revealed by codon-based analyses. *PLoS Comp. Biol*. **2**(6):530-536.

Krakauer, D. C. & Plotkin, J. B. (2002) Redundancy, antiredundancy, and the robustness of genomes. *PNAS* **99**(3):1405-1409.

Kuiken C, Yusim K, Boykin L, Richardson R. (2005) The Los Alamos HCV Sequence Database. *Bioinformatics*, **21**(3):379-84.

Kuntzen, T., Timm, J., Berical, A., Lewis-Zimenez, L. L., Jones, A., Nolan, B., Schulze zur Wiesch, J., Li, B., Schneidewind, A., Kim, A. Y., Chung, R. T., Lauer, G. M. & Allen, T. M. (2007) Viral sequence evolution in acute hepatitis C virus infection. *J. Virol*. **81**(21):11658-11668.

Lanave, C., Preparata, G., Saccone, C. & Serio, G. (1984). A new method for calculating evolutionary substitution rates. *J. Mol. Evol*. **20**:86-93.

Lauer, G. M., Nguyen, T. N., Day, C. L., Robbins, G. K., Flynn, T., McGowan, K., Rosenberg, E. S., Lucas, M., Klenerman, P., Chung, R. T. & Walker, B. D. (2002). Human immunodeficiency virus type-1 hepatitis C virus coinfection: intraindividual comparison of cellular immune responses against tow persistent viruses. *J. Virol*. **76**(6):2817-2826.

Leigh Brown, A.J., Lobidel, D., Wade, C.M., Rebus, S., Philips, A.N., Brettle, R.P., France, A.J., Leen, C.S., McMenamin, J., McMillan, A., Maw, R.D., Mulcahy, F., Robertson, J.R., Sankar, K.N., Scott, G., Wyld, R. & Peutherer, J.F. (1997). The molecular epidemiology of human immunodeficiency virus Type 1 in six cities in Britain and Ireland. *Virology* **235**:166-177.

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

Lemey, P., Rambaut, A. & Pybus, O. G. (2006) HIV evolutionary dynamics among and within hosts. *AIDS Rev*. **8**:125-140.

Lemey, P., Pybus, O. G., Rambaut, A., Drummond, A. J., Robertson, D. L., Roques, P., Worobey, M. & Vandamme, A.-M.. (2004) The molecular population genetics of HIV-1 Group O. *Genetics* **167**:1059-1068.

Lemey, P., Salemi, M., Bassit, L. & Vandamme, A.-M. (2002) Phylogenetic classification of TT virus groups based on the N22 region is unreliable. *Virus Res*. **85**:47-59.

Lemey, P., Van Dooren, S., & Vandamme, A.-M. (2005) Evolutionary dynamics of human retroviruses investigated through full-genome scanning. *Mol. Biol. Evol*. **22**(4):942-951.

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

Li, W.-H., Tanimura, M. & Sharp, P. M. (1988) Rates and dates of divergence between AIDS virus nucleotide sequences. *Mol. Biol. Evol*. **5**(4):313-331.

Lipsitch, M. & Moxon, E. R. (1997) Virulence & transmissibility of pathogens: what is the relationship? *Trends Microbiol*. **5:**31-37.

Lorenzo, E., Colon, M. C., Almodovar, S., Malodonado, I. M., Gonzalez, S., Costa, S. E., Hill, M. D., Mendoza, R., Sepulveda, G., Yanagihara, R., Nerurkar, V., Kumar, R., Yamamura, Y., Scott, W. A. & Kumar, A. (2004) Influence of $CD4^+$ T cell counts on viral evolution in HIV-infected individuals undergoing suppressive HAART. *Virology* **330**:116-126.

McNatt, M.W., Zang, T., Hatziioannou, T., Bartlett, M, Fofana, I.B., Johnson, W.E., Neil, S.J.D. & Bienasz, P.D. (2009). Species-specific activity of HIV-1 Vpu and positive selection of transmembrane domain variants. *PLoS Pathogens* **5**(2):e1000300.

Malim, M.H. & Emerman, M. (2008) HIV-1 accessory proteins – ensuring survival in a hostile environment. *Cell Host & Microbe* **3**(6):388-398.

Mammano, F., Petit, C. & Clavel, F. (1998) Resistance-associated loss of viral fitness in human immunodeficiency virus Type 1: A phenotypic analysis of protease and *gag* coevolution in protease inhibitor-treated patients. *J. Virol* **72**:7632-7637.

Mao, Q., Ray, S., Laeyendecker, O., Ticehurst, J. R., Strathdee, S. A., Vlahov, D. & Thomas, D. L. (2001) Human immunodeficiency virus seroconversion and evolution of the hepatitis C virus quasispecies. *J. Virol*. **75**(7):3259-3267.

Mayr, E. (1997). The objects of selection. *PNAS* **94**:2091-2094

McMichael, A. J. & Hanke, T. (2003) HIV vaccines 1983-2003. *Nat. Med*. **9**(7):874-880.

Mi, S., Lee, X., Li, X, Veldman, G. M., Finnerty, H., Racie, L., LaVallie, E., Tang, X-Y., Edouard, P., Howes, S., Keith, J.C. Jnr & McCoy, J. M. (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403**:785-789.

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

Mizokami, M., Tanaka, y. & Miyakawa, Y. (2006) Spread times of hepatitis C virus estimated by the molecular clock differ among Japan, the United States and Egypt in reflection of their distinct socioeconomic backgrounds. *Intervirology* **49**:28-36.

Moradpour, D., Penin, F. & Rice, C. M. (2007) Replication of hepatitis C virus. *Nat. Rev. Microbiol*. **5**:453-463.

Mossel, E. & Vigoda, E. (2005) Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science* **309**:2207-2209.

Mossel, E. & Vigoda, E. (2006) Response to Comment on "Phylogenetic MCMC algorithms are misleading on mixtures of trees" *Science* **312**:367b.

Moya, A., Holmes, E. C. & González-Candelas, F. (2004) The population genetics and evolutionary epidemiology of RNA viruses. *Nat. Rev. Microbiol*. **2**:279-287.

Muller, H. J. (1964). The relation of recombination to mutational advance. *Mutat. Res*. **106**:2-9.

Nakano, T., Lu, L., Liu, P. & Pybus, O.G. (2004). Viral gene sequences reveal the variable history of hepatitis C virus infection among countries. *Journal of Infectious Disease* **190**:1098-1108.

Navas *et al*. (1997) Genetic diversity and tissue compartmentalization of the Hepatitis C virus genome in blood mononuclear cells, liver, and serum from chronic hepatitis C patients. *J Virol* **72**(2):1640-1646.

Neil, J. D., Zang, T. & Bieniasz, P. D. (2008). Tetherin inhibits virus release and is antagonized by HIV-1 Vpu. *Nature* **451**:425-430.

Noë, A. Plum, J. & Verhofstede, C. (2005) The latent reservoir in patients undergoing HAART: an archive of pre-HAART drug resistance. *J. Antimicrob. Chemoth.* **55**:410-410.

Nousbaum, J.-B., Polyak, S. J., Ray, S. C., Sullivan, D. G., Larson, A. M., Carithers, R. L. Jr & Gretch, D. R. (2000) Prospective characterization of full-length hepatitis C virus NS5a quasispecies during induction and combination antiviral therapy. *J. Virol.* **74**(19):9028-9038.

Novella, I. S., Elena, S. F., Moya, A., Domingo, E. & Holland, J. J. (1996) Size of genetic bottlenecks leading to virus fitness loss is determined by mean initial population fitness. *J. Virol.* **69**(5):2869-2872.

Nowak, M. (2006) Evolutionary dynamics: Exploring the equations of life. Belknap Press, Cambridge, MA, USA.

Nunnari, G., Sullivan, J., Xu, Y. et al. (2005). HIV type 1 cervicovaginal reservoirs in the era of HAART. *AIDS Research and Human Retroviruses,* **21** (8): 714-718.

O'Neill, R. J. W., O'Neill, M. J. & Graves, J. A. M. (1998) Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. *Nature* **393**:68-72.

Ohagen, A., Devitt, A., Kunstman, K. J., Gorry, P. R., Rose, P. P., Korber, B., Taylor, J., Levy, R., Murphy, R. L., Wolinsky, S. M. & Gabuzda, D. (2003) Genetic and functional analysis of full-length human immunodeficiency virus Type-1 *env* genes derived from brain and blood of patients with AIDS. *J. Virol.* **77**(22):12336-12345.

Ohta, T. & Gillespie, J. H. (1996) Development of neutral and nearly neutral theories. *Theoret. Pop. Biol.* **49**:128-142.

Pavesi, A. (2001) Origin and evolution of GBV-C/Hepatitis G virus and relationships with ancient human migrations. *J. Mol. Evol.* **53**:104-113.

Perrin, L., Kaiser, L. & Yerly, S. (2003). Travel and the spread of HIV-1 genetic variants. *Lancet Infec. Dis.* **3**:22-26.

Philpott, S., Burger, H., Tsoukas, C., Foley, B., Anastos, K., Kitchen, C. & Weiser, B. (2005) Human immunodeficiency virus Type 1 genomic RNA sequences in the female genital tract and blood: compartmentalization and intrapatient recombination. *J. Virol.* **79**(1):353-363.
Pie, M. R. (2006) The influence of phylogenetic uncertainty on the detection of posititve Darwinian selection. *Mol. Biol. Evol.* **23**(12):2274-2278.

Pillai, S. K., Good, B., Kosakovsky Pond, S., Wong, J. K., Strain, M. C., Richman, D. D. & Smith, D. M. (2005) Semen-specific genetic characteristics of human immunodeficiency virus Type 1 *env*. *J. Virol.* **79**(3):1734-742.

Polyak et al (1998). Evolution of Hepatitis C virus quasispecies in hypervariable region 1 and the puatative interferon sensitivity-determining region during interferon therapy and natural infection, J. Virol 72:4288-4296.

Posada, D. & Buckley, T. R. (2004) Model selection and model averaging in phylogenetics: advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. *Syst. Biol.* **53**(5):793-808.

Posada, D. & Crandall, K. A. (1998) MODELTEST: testing the models of DNA substitution. *Bioinformatics* **14**(9):817-818.

Potter, S. J. , Lemey, P., Achaz, G., Chew, C. B., Vandamme, A.-M., Dwyer, D. E. & Saksena, N. K. (2004) HIV-1 compartmentalization in diverse leukocyte populations during antiretroviral therapy. *J. Leukocyte. Biol.* **76**:562-570.

Potter, S. J., Lemey, P., Dyer, W. B., Sullivan, J. S., Chew, C. B., Vandamme, A.-M., Dwyer, D. E. & Saksena, N. K. (2006) Genetic analyses reveal structured HIV-1 populations in serially sampled T lymphocytes of patients receiving HAART. *Virology* **348**(1):35-46.

Pybus, O. G., Rambaut, A., Belshaw, R., Freckleton, R. P., Drummond, A. J. & Holmes, E. C. (2007) Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. *Mol. Biol. Evol.* **24**(3):845-852.

Pybus, O.G., Charleston, M.A., Gupta, S., Rambaut, A., Holmes, E.C. & Harvery, P.H. (2001) The epidemic behaviour of the hepatitis C virus. *Science* **292**:2323-2325.

Rambaut, A., Robertson, D., Pybus, O. G., Peeters, M. & Holmes, E. C. (2001) Phylogeny and the origin of HIV-1. *Nature* **410**:1047-1048.

Rambaut, A (2002) Se-Al v2.0a11. Available from: http://tree.bio.ed.ac.uk/software/seal/
Rambaut, A. (2000) Estimating the rate of molecular evolution: incorporating non-contemperaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16**(4):395-399.

Rambaut, A., Posada, D., Crandall, K.A. & Holmes, E.C. (2004) The causes and consequences of HIV evolution. *Nat. Rev. Genet.* **5**:52-61.

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

Ray, S.C., Wang, Y.-M., Laeyendecker, P., Ticehurst, J. R., Villano, S. A. & Thomas, D. L. (1999) Acute hepatitis C virus structural gene sequences as predictors of persistent viremia: hypervariable region 1 as a decoy. *J. Virol.* **73**(4):2938-2946.

Rhodes, T., Wargo, H. & Hu, W-S. (1999). High rates of human immunodeficiency virus Type 1 recombination: Near-random segregation of markers one kilobase apart in one round of viral replication. *J. Virol.* **77**(20):11193-11200.

Ronquist, F., van der Mark, P. & Huelsenbeck, J. P. (2009) *Bayesian phylogenetic analysis using MRBAYES: theory*. In Lemey, P., Salemi, M. & Vandamme, A.-M. (eds.) *The Phylogenetic Handbook (2nd ed.)* Cambridge Univ. Press, Cambridge.

Ronquist, F., Larget, B., Huelsenbeck, J. P., Kadane, J. B., Simon, D. & van der Mark, P. (2006) Comment on "Phylogenetic MCMC algorithms are misleading on mixtures of trees" *Science* **312**:376a.

Ross HA & Rodrigo AG. (2002) Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. *J Virol.* **76**(22):11715-20.

Ross, H. A. & Rodrigo, A. G. (2002) Immune-mediated positive selection drives human immunodeficiency virus Type 1 molecular variation and predicts disease duration. *J. Virol.* **76**(2):11715-11720.

Saitou, N. & Nei, M. (1987) The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**:406-425.
Salemi, M. & Vandamme, A.-M. (2002) Hepatitis C virus evolutionary patterns studies through analysis of full-genome sequences. *J. Mol. Evol.* **54**:62-70.

Salemi, M., Lamers, S. L., Yu, S., de Oliveira, T., Fitch, W. M. & McGrath, M. S. (2005) Phylodynamic analysis of human immunodeficiency virus Type 1 in distinct brain compartments provides a model for the neuropathogenesis of AIDS. *J. Virol.* **79**(17):11343-11352.

Sanjuán, R., Codoñer, F. M., Moya, A. & Elena, S. F. (2004a) Natural selection and the organ-specific differentiation of HIV-1 V3 hypervariable region. *Evolution* **58**(6):1185-1194.

Sanjuán, R., Moya, A. & Elena, S. F. (2004b) The contribution of epistasis to the architecture of fitness in an RNA virus. *PNAS* **101**:15376-15379.

Sanjuán, R., Moya, A. & Elena, S. F. (2004c) The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *PNAS* **101**(22):8396-8401.

Santiago, M. L., Rodenburg, C.M., Kamenya, S., Bibollet-Ruche, F., Gao, F., Bailes, E., Meleth, S., Soong, S.,-J., Kilby, J. M., Moldoveanu, N., Fahey, B., Muller, M. N., Ayouba, A., Nerrienet, E., McClure, H. M., Heeney, J. L., Pusey, A. E., Collins, D. A., Boesch, C., Wrangham, R. W., Goodall, J., Sharp, P. M., Shaw, G. M. & Hahn, B. M. (2002) SIVcpz in wild chimpanzees. *Science* **295**(5554):465.

Sawyer, S., Emerman, M, & Malik, H. (2004) Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS Biol.* **2**(9):0001-0008.

Seo, T.-K., Thorne, J. L., Hasegawa, M. & Kishino, H. (2002) A viral sampling design for testing the molecular clock and estimating evolutionary dates and divergence times. *Bioinformatics* **18**(1):115-123.

Shackleton & Holmes (2004) The evolution of large DNA viruses: combining genomic evolution of viruses and their hosts. Trend in Micro. 12(10):458-465

Shankarappa, R., Margolick, J. B., Gange, S. J., Rodrigo, A. G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C. R., Learn, G. H., He, X., Huang, X.-L. & Mullins, J. I. (1999). Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus Type 1 infection. *J. Virol.* **73**(12):10489-10502.

Sheridan, I., Pybus, O. G., Holmes, E. C. & Klenerman, P. (2004) High-resolution phylogenetic analysis of hepatitis C virus and its relationship to disease progression. *J. Virol.* **78**(7):3447-3454.

Simmonds, P. (2004) Genetic diversity and evolution of hepatitis C virus – 15 years on. *J. Gen. Virol.* **85**:3173-3188.

Slatkin, M. & Maddison, W. P. (1989) A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics* **123**:603-613.

Sobesky, R., Feray, C., Rimlinger, F., Derian, N., Dos Santos, A., Roque-Alonso, A.-M., Samuel, D., Bréchot, C. & Thiers, V. (2007) Distinct hepatitis C virus core and F protein quasispecies in tumoral and nontumoral hepatocytes isolated via microdissection. *Hepatology* **46**:1704-1712.

Steinhauer, D. A. & Skehel, J. J. (2002) Genetics of influenza viruses. *Annu. Rev. Genet.* **36**:305-32.

Strimmer, K. & Pybus, O. G. (2001) Exploring the demographic history of DNA sequences using the generalized skyline plot. *Mol. Biol. Evol.* **18**(12):2298-2305.

Suchard, M.A., Weiss, R.E. & Sinsheimer, J.S. (2001) Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.* **18**:1001:1013.

Suzan-Monti, M., La Scola, B. & Raoult, D. (2005) Genomic and evolutionary aspects of *Mimivirus*. *Virus Res.* **117**(7):145-155.

Switzer, W.M., Salemi, M., Shanmugam, V., Gao, F., Cong, M., Kuiken, C., Bhullar, V., Beer, B. E., Vallet, D., Gautier-Hion, A., Tooze, Z., Villinger, F., Holmes, E. C. & Heneine, W. (2005) Ancient co-speciation of simian foamy viruses and primates. *Nature* **434**:376-380.

Tanaka, Y, Hanada, K., Mizokami, M., Yeo, A.E.T., Shih, J.W.-K., Gojobori, T. & Alter, H.J. (2002). A comparison of the molecular clock of hepatitis C virus in the United States and Japan predicts that hepatocellular carcinoma incidence in the United States will increase over the next two decades. *PNAS* **99**(12):15584-15589.

Tang, J. & Kaslow, R. A. (2003) The impact of host genetics on HIV infection and disease progression in the era of highly active antiretroviral therapy. *AIDS* **17** (suppl 4):S51-S60.
Taubenberger, J.K., Reid, A.H., Lourens, R.M., Wang, R., Jin, G. & Fanning, T.G. (2005) Characterization of the 1918 influenza virus polymerase genes. *Nature* **437**(6):889-893.

Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. (1997) Inferring coalescence times from DNA sequence data. *Genetics* **145**:505-518.

The National Audit Office (NAO) (2002) *The 2001 outbreak of foot and mouth disease*. National Audit Office, HC Session 2001-2002:21 June 2002.

Tristem, M., Purvis, A. & Quicke, D. L. Q. (1998) Complex evolutionary history of lentiviral vpr genes. *Virology* **240**, 232-237.

van der Hoek, L., Goudsmit, J., Maas, J. & Sol, C. J. A. (1998) Human immunodeficiency virus type 1 in faeces and serum: evidence against independently evolving populations. *J. Gen. Virol.* **79**:2455-2459.

van Nimwegen, E., Crutchfield, J. P. & Huynen, M. (1999) Neutral evolution of mutational robustness. *PNAS* **96**:9716-9720.

Vignussi, M., Stone, J. K., Arnold, J. J., Cameron, C. E. & Andino, R. (2006) Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* **439**:344-348.
Wahl, L. M., Gerrish, P. J. & Saika-Voivod, I. (2002) Evaluating the impact of experimental bottlenecks in experimental evolution. *Genetics* **162**:961-971.

Wain, L. V., Bailes, E., Bibollet-Ruche, F., Decker, J. M., Keele, B. F. , van Heuverswyn, F., Li, Y., Takehisa, J., Ngole, E. M., Shaw, G. M., Peeters, M., Hahn, B. H. & Sharp, P. M. (2007) Adaptation of HIV-1 to its human host. *Mol. Biol. Evol.* **24**(8):1853-1860.

White, N. C., Israel-Biet, D., Coker, R. J., Mitchell, D. N., Weber, J. N. & Clarke, J. C. (2004) Different resistance mutations can be detected simultaneously in the blood and the lung of HIV-1 infected individuals on antiretroviral therapy. *J. Med. Virol.* **72**:352-357.

Wilke, C. O. (2005) Quasispecies theory in the context of population genetics. *BMC Evol. Biol,* **5**:44-52.

Williamson *et al* (2004) A statistical characterization of consistent patterns of human immunodeficiency virus evolution within infected patients. *MBE* **22**:456-468.

Williamson, S. (2003) Adaptation in the *env* gene of HIV-1 and evolutionary theories of disease progression. *Mol. Biol. Evol.* **20**(8):1318-1325.

Wolfson, L. J., Strebel, P. M., Gacic-Dobo, M., Hoekstra, E. J. McFarland, J. W. & Hersh, B. (2007) Has the 2005 measles mortality reduction goal been achieved? A natural history modelling study. *The Lancet* **369**:191-200.

Wong, J. K., Ignacio, C. C., Torriani, F., Havlir, D., Fitch, N. J. S. & Richman, D. D. (1997) *In vivo* compartmentalization of human immunodeficiency virus: evidence from the examination of *pol* sequences from autopsy tissues. *J. Virol.* **71**(3):2059-2071.

Worobey, M. & Holmes, E. C. (1999) Evolutionary aspects of recombination in RNA viruses. *J. Gen. Virol.* **80**:2535-2543.

Worobey, M., Rambaut, A. & Holmes, E. C. (1999) Widespread intra-serotype recombination in natural populations of dengue virus. *PNAS* **96**:7352-7357.
Wyatt, R. & Sodroski, J. (1998) The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens. *Science* **280**:1884-1888.

Xiong, Y. & Eickbush, T. H. (1990) Origin and evolution of retroelements based upon their reverse transcruiptase sequences. *The EMBO Journal* **9**(10):3353-3362.

Yang, Z. (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. CABIOS 13:555-556 (http://abacus.gene.ucl.ac.uk/software/paml.htm)

Yen, H.-L., Monto, A. S., Webster, R. G. & Govorka, E. A. (2005) Virulence may determine the necessary duration and dosage of Oseltamivir tratment for highly pathogenic A/Vietnam/1203/04 influenza virus in mice. *J. Infect. Dis.* **192**:665-672.

Yerly, S., Kaiser, L., Race, E., Bru, J.-P., Clavel, F. & Perrin, L. (1999). Transmission of antiretroviral-drug-resistant HIV-1 variants. *Lancet* **354**:729-733.

Yu, X., Yu, Y., Liu, B., Kong W., et al (2003). Induction of APOBEC3G ubiquitination and degredation by HIV-1 Vif-Cul5-SCF complex. Science 302:1056-1060.

Zhang, J. & Temin, H. M. (1993) Retrovirus recombination depends on the length of sequence identity and is not error prone. *J. Virol.* **68**(4):2409-2414.

Zhang, L., Rowe, L., He, T., Chung, C., Yu, J., Yu, W., Talal, A., Markowitz, M. & Ho, D. D. (2002) Compartmentalization of surface envelope glycoprotein of human immunodeficiency virus Type 1 during acute and chronic infection. *J. Virol.* **76**(18):9465-9473.

Zheng, Y.-H., Irwin, D., Kurosu, T., Tokunaga, K., Sata, T. & Peterlin, B. M. (2004) Human APOBEC3F is another host factor that blocks human immunodeficiency virus Type 1 replication. *J. Virol.* **78**(11):6073-6076.

Zhu, T., Mo, H., Wang, N., Nam, D.S., Cao, Y., Koup, R. A. & Ho, D. D. (1993) Genotypic and phenotypic characterzation of HIV-1 patients with primary infection. *Science* **261**(5125):1179-1181.

Zuckerkandl, E., & Pauling, L. (1965) Evolutionary divergence and convergence of proteins. In Kasha M. & Pullman, B. (eds.), *Horizons in Biochemistry*, pp189-225. Academic Press, NY.

Chapter Two

Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty

This chapter has been published as:

Parker, J., Rambaut, A.R. & Pybus, O.G. (2008) Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *MEEGID* **8**(3):239-246.

A.R. provided editorial assistance.

O.G.P. provided supervisory support and editorial assistance.

2.1 Abstract

Many recent studies have sought to quantify the degree to which viral phenotypic characters (such as epidemiological risk group, geographic location, cell tropism, drug resistance state etc.) are correlated with shared ancestry, as represented by a viral phylogenetic tree. Here we present a new Bayesian Markov-Chain Monte Carlo approach to the investigation of such phylogeny-trait correlations. This method accounts for uncertainty arising from phylogenetic error and provides a statistical significance test of the null hypothesis that traits are associated randomly with phylogeny tips. We perform extensive simulations to explore and compare the behaviour of three statistics of phylogeny-trait correlation. Finally, we re-analyse two existing published data sets as case studies. Our framework aims to provide an improvement over existing methods for this problem.

2.2 Introduction

In recent years explosions in the availability of molecular sequence data and of statistical methods for evolutionary analysis have given new insights in the field of molecular epidemiology. For example, the processes of natural selection, recombination, mutation and migration have all been studied to great effect at different levels of biological organization. However, despite recent increases in computing power, analytical approaches for some classes of problem are still in need of further improvement and rigorous statistical validation.

One such under-developed area concerns the association of phenotypic characters (e.g. geographic locations, physical characteristics, behavioural traits, etc.) with the shared ancestry of a sample of organisms from which gene sequences have been obtained. The individuals sampled may represent different cells, virions, organisms, populations, species, or even higher phyla. We may wish to know whether a particular phenotype has arisen independently in different organism, or whether it is the result of common ancestry from a single ancestral individual. Another common application of phylogeny-trait correlation is the investigation of spatial population structure; that is, do sequences cluster on a phylogeny according to their geographic location (e.g. Avise, 2000; Starkman *et al*, 2003; Holmes 2004)? In all such cases, analyses are complicated by the lack of statistical independence – the phenotypic traits associated with each phylogenetic tip may not be independent as a result of the shared ancestry among sampled individuals (Harvey & Pagel, 1991). It is therefore inappropriate to use standard general linear models to statistically test the null hypothesis that the phenotypes are uncorrelated with the genetic distances among sampled individuals.

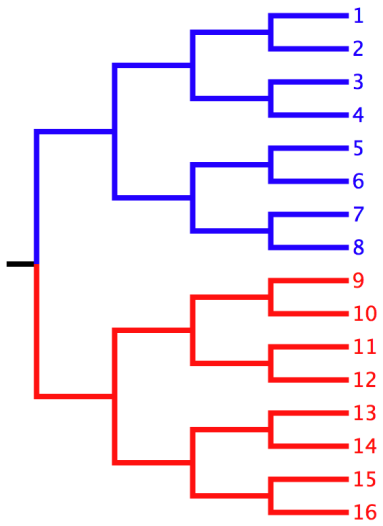
A number of previous studies in the field of molecular epidemiology have investigated the association between a virus phylogeny and viral traits. These have included investigations of population structure, resulting from geographic location (e.g. Cochrane *et al*, 2002; Nakano *et al*, 2004; Carrington *et al*, 2005), epidemiological risk group (e.g. Holmes *et al*, 1995; Leigh Brown *et al*, 1997) or compartmentalization, either among different host tissues (e.g. Salemi *et al*, 2005; Pillai *et al*, 2006) or among different host cell types (e.g. Fulcher *et al*, 2004). The detection of within-host compartmentalization has been an issue of particular interest for the Human Immunodeficieny Virus, HIV (McGrath *et al*, 2001; Kemal *et al*, 2003; Sanjuan *et al*, 2004). In addition, phylogeny-trait associations have been used to investigate antibody and T-cell escape during chronic Hepatitis C Virus (Sheridan et al. 2004; Komatsu et al. 2006) and HIV (Bhattacharya et al. 2007) infection.

A closely related and long-standing technique is the estimation of gene flow among subpopulations using explicit population genetic models (e.g. Wright, 1952; Beerli & Felsenstein, 2001). Such approaches have recently been implemented in a Bayesian framework and applied to virus genetic data (Beerli & Felsenstein, 2001; Wilson & Rannala, 2003; Ewing *et al*, 2004). However, given the complexity of such methods it may be desirable to first demonstrate that the sequences are indeed phylogenetically structured according to the trait of interest. In this paper we investigate and develop statistical tests for this preliminary null hypothesis.

What exactly do we mean by phylogeny-trait correlation? Given a discrete character for each tip of a phylogenetic tree, we are asking if more closely related taxa are more likely to share the same trait values than we would expect by chance alone, i.e. if the characters were randomly assigned to the phylogeny tips. As illustrated in Figure 2.1, the tip characters may be tightly correlated with phylogeny (Fig. 2.1a) or

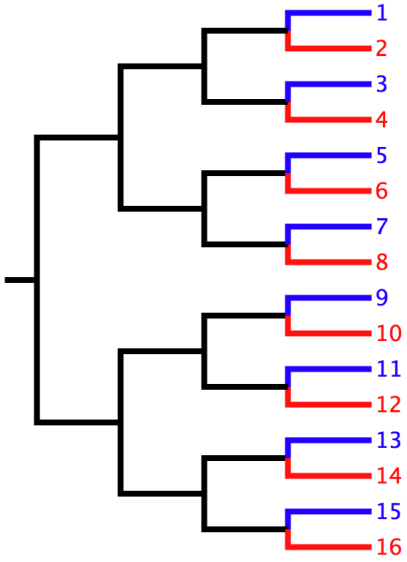
they may be fully interspersed (Fig. 2.1b). The biological significance of either situation will depend on the nature of the phenotypic trait under investigation. If, for example, the trait represents geographic location then a tight correlation reflects low lineage dispersal or migration, and the opposite represents panmixis or high rates of gene flow. Alternatively, if the trait is thought to be under strong selection – pathogen drug resistance, for example – then interspersed traits may indicate that drug resistance has independently evolved several times or that this phenotype is not under strong evolutionary constraints. However, in many situations the phylogenetic distribution of the traits may be intermediate and their correlation with phylogeny less clear (Fig. 2.1c). Therefore the strength of the association needs to be quantified and statistically tested against the distribution of characters expected by chance.

Figure 2.1a



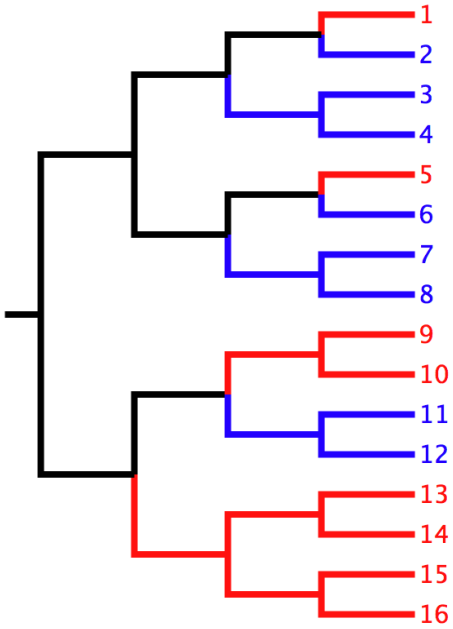
AI	<0.0001
PS	1
MC (blue)	8
MC (red)	8

Figure 2.1b



AI	4.51
PS	15
MC (blue)	1
MC (red)	1

Figure 2.1c



AI	0.762
PS	4
MC (blue)	2
MC (red)	4

Figure 2.1

(a) Strong association: There is clear association between the characters at the tips (represented by taxon colours) and the phylogeny. Taxa 1-8 have inherited the 'blue' character state, whilst taxa 9-16 have inherited 'red.' *(b) Maximally interspersed:* This tree clearly shows no clear association between phenotype and phylogeny. The 'blue' or 'red' characters have been acquired or lost multiple times in the ancestry of these taxa. *(c) Intermediate situation:* In some areas of the tree, such as taxa 13-16, it does appear that sister taxa share characters of interest. However, in other areas, such as sister taxa 1-2 and 5-6, the trait values look more interspersed. An analytical method is needed to decide if the association between tips and characters is significant.

A variety of metrics have been proposed to quantify phylogeny-trait correlation. An early approach by Hudson *et al* (1992) used a range of sequence-summary statistics calculated directly from a sequence alignment. Unfortunately such techniques suffer from a lack of independence due to shared ancestry, as explained above, and as a result more recent techniques have tended to employ some form of phylogeny, either estimated from a molecular sequence alignment or from morphological information.

The most common phylogenetic method is to use a parsimony approach (such as the Fitch 1971b algorithm) to reconstruct the character states at ancestral nodes. The number of state changes in the phylogeny is then calculated (the parsimony score statistic; PS). However, although PS quantifies phylogeny-trait correlation, it provides no information on whether the value obtained is statistically significant or not. Slatkin and Madison (1989) addressed this problem by randomizing tip-character associations a large number of times and calculating the PS statistic from each randomization, thereby providing a null distribution of the PS statistic with critical values at the p=0.05 confidence level, against which the observed PS value can be compared. A second metric is the association index (AI) statistic, which explicitly takes into account the shape of the phylogeny by measuring the imbalance of internal phylogeny nodes (Wang *et al* 2001; see Methods). As with the PS statistic, a randomization approach can be used to generate a null distribution for the AI statistic (Wang *et al*, 2001).

The methods outlined above involve calculating metrics from a single phylogeny that is assumed to be correct. In reality, this single tree is estimated from gene sequences with phylogenetic error, and there may be a large set of different trees that do not differ significantly from the ML tree (Jermiin *et al*, 1997). Single-tree

approaches do not incorporate this error and thus underestimate the true statistical variance. Wang *et al* (2001) attempted to address this issue by calculating the AI statistic across a set of bootstrap replicate trees; although this approach provides for confidence estimation through bootstrapping, results are still conditional on a single topology, and hence susceptible to (unaccounted-for) phylogenetic error.

In this paper we concentrate on the methods most commonly applied to viral phylogenies, namely the PS and AI statistics. However, we note that several related statistics have been developed in the field of community ecology. Faith (1991) defined the 'phylogenetic diversity' (PD) of a set of taxa as the sum of the shortest paths between all taxa in the set. Webb (2000) and Webb *et al* (2002) developed two related methods, the net relatedness index (NRI) and nearest taxa index (NTI), which combine nodal distances (the number of nodes between two taxa that share a trait value) with branch lengths. Lastly, Lozupone & Knight (2005) introduced the UniFrac statistic, which measures the proportion of phylogeny branch lengths that can be unambiguously associated with a particular trait value. A common feature of the PD, NTI/NRI and UniFrac statistics is that they all depend on both the tree topology and the tree branch lengths. The PS and AI statistics, in contrast, depend only on the former. Both types of statistic measure the degree to which taxa with the same trait values cluster together, but the PD, NTI/NRI and UniFrac statistics also measure the genetic similarity among clustering taxa.

In this study we accommodate phylogenetic error in the calculation of phylogeny-trait correlations using Bayesian Markov chain Monte Carlo (MCMC) methods. Such methods have become increasingly popular and practical over recent years (Holder & Lewis, 2003). Programs that implement MCMC sampling can be used to obtain a posterior distribution of phylogenies, from which the posterior

distributions of phylogeny shape statistics can be calculated. Our method correctly incorporates statistical error arising from phylogenetic uncertainty and provides error intervals for hypothesis testing, as well as returning the posterior distribution of the statistics, which provides greater detail than the traditional single 'p-value'. Additionally, we perform extensive simulations to test the relative statistical power of different statistics for the first time. We also investigate a new phylogeny-trait statistic, the 'MC size' statistic (described below). Finally, we investigate the performance of our new method by re-analysing the data published in Carrington *et al* (2005) and Salemi *et al* (2005); these data were previously investigated using other phylogeny-trait correlation methods.

2.3 Methods

We start by defining how phylogeny-trait statistics are calculated from a single phylogeny. Figure 2.1 provides the values of several different statistics for three example trees. We then explain how phylogenetic uncertainty can be incorporated into this calculation.

The parsimony score (PS) statistic can be calculated using the Fitch (1971b) parsimony algorithm. If the gain/loss of the trait under investigation does occur parsimoniously, then the observed PS value should be inversely related to the strength of tip-character association. The PS statistic for a given trait takes the range $1 \leq PS \leq n$, where n is the number of tips in the phylogeny. Low PS scores represent strong phylogeny-trait association. Note that for a single tree, PS (unlike AI) takes integer values and hence is a discrete metric.

The Association Index (AI) statistic introduced by Wang *et al*. (2001) is the sum:

$$AI = \sum_{i=1}^{k} \frac{(1 - f_i)}{2^{m_i - 1}} \qquad (2.1)$$

The AI is a sum across all the internal nodes in the phylogeny; k is the number of internal nodes. For each internal node i, f_i is defined as the frequency of the most common trait value among the tips subtended by that node; m_i is the number of tips subtended by node i. Thus low AI values represent strong phylogeny-trait association.

We also define a new statistic that was used in Salemi *et al.* (2005) but which has not been previously investigated. Intuitively, stronger phylogeny-trait associations should produce larger monophyletic clades whose tips all share the same

trait. This property is quantified by the monophyletic clade ('MC') size statistic for a particular trait value *x*, defined as:

$$MC(x) = \max_{i=1}^{k} (m_i I_i) \qquad (2.2)$$

where m_i is the number of tips subtended by node *i and* I_i is an indicator function that equals 1 if all tips subtended by node *i* have trait value x, and equals zero otherwise. *k* is the number of internal nodes in the phylogeny, including the root. MC is a discrete integer metric for a single tree and is bounded by $1 \leq MC \leq n_x$, where n_x is the number of tips that have trait value *x*. MC will be positively correlated with the strength of the phylogeny-trait association.

2.3.1 Incorporating phylogenetic error

The above methods all require a fully resolved phylogeny to be specified *a priori*. In practice, the tree is estimated from sequences and has an associated statistical error. To account for phylogenetic uncertainty, we developed a Bayesian MCMC approach. Programs such as BEAST (Drummond *et al*, 2001; Drummond & Rambaut, 2003) or MrBayes (Huelsenbeck & Ronquist, 2001) calculate a posterior sample of trees (PST) that approximates the true posterior distribution of phylogenies given the sequences, with more likely phylogenies being sampled more frequently, and less likely ones less so. By calculating and averaging phylogeny-trait statistics across all trees in the posterior sample, we integrate over (marginalize) the phylogeny and thus incorporate phylogenetic uncertainty.

We combine phylogenetic error and significance testing in the following way. First, the value of the statistic concerned is calculated for every tree in the posterior sample, forming the posterior distribution of the statistic, and the median of this

posterior distribution is denoted μ. Next, from the observed set of taxon-character associations C, we generate n randomized sets of taxon-character associations $\{C_1, C_2, C_3 \ldots C_n\}$. Each randomized set C_i is simply the observed set of associations C resampled without replacement. The set $\{C_1, C_2, C_3 \ldots C_n\}$ therefore constitutes a null distribution of taxon-character associations; in our analyses we used $n = 100$ (this was determined by simulation to be a sufficiently large number of replicates to capture the null distribution; see the Appendix to this Chapter – Appendix Six). Then, for each C_i the median posterior estimate of the statistic is calculated from the PST using the same method as for the observed data, and denoted μ_i,. The distribution of the μ_i obtained from the n randomized sets therefore corresponds to an estimate of the null distribution of the statistic. The significance p is then obtained from this null distribution by simply calculating the proportion of μ_i values that are more extreme than the observed value μ (low values being extreme for AI and PS; high values being extreme for MC).

We have developed a computer program BaTS (Bayesian Tip-association Significance testing; available on request) that takes the PST output from an existing program such as BEAST or MrBayes and performs these randomizations.

2.3.2 Simulations

We performed a set of simulations to test the type 1 statistical error of our Bayesian MCMC approach (i.e. the probability of rejecting a null hypothesis when the null hypothesis is true). Theoretically, if data are repeatedly simulated under the null hypothesis, then the distribution of the resulting p-values should follow a unit uniform distribution. If this is so, then the type 1 error of the test will correct for all levels of statistical significance (e.g. p=0.05, p=0.01, etc.). We therefore simulated a large

number of data sets with random tip-trait associations and computed the *p*-values for each test statistic on each dataset. The *p*-values were then collated to create an cumulative density function, which was compared against the expected unit uniform distribution using the two-sample Kolmogorov-Smirnoff test.

Data sets were simulated in two steps; (i) a large set of simulated alignments were generated and (ii) character traits were randomly associated with the simulated sequences. For the first step, 3436 random trees of 32 taxa were generated under a pure-birth process using the package Phyl-O-Gen (available from http://evolve.zoo.ox.ac.uk). This set therefore includes phylogenies with a wide variety of tree shapes, branch lengths and node imbalances. Next, the Seq-Gen (Rambaut & Grassly, 1997) program was used to create sequence alignments, by simulating down each tree using a substitution model typical of empirical HIV-1 *env* gene data sets (transition/transversion ratio=2.4; base frequencies A=0.426, C=0.152, G=0.182, T=0.24; evolutionary rates in substitutions/site/year of 0.0152 for codon position 1, 0.0142 for codon position 2 and 0.0215 for codon position 3). The evolutionary rates used were estimated from HIV-1 subtype A, B, C, D, and G sequences, collected in the 1990s from Scandinavia (A. Iversen, personal communication). These rates are entirely typical of HIV evolution and are comparable to previously published estimates (e.g. Lemey *et al* 2004, 2005). This step produced 3436 alignments of 32 sequences, each 300nt in length. A posterior sample of trees (PST) was then obtained for each alignment using BEAST (Drummond *et al*, 2001; Drummond & Rambaut, 2003).

In the second step, the simulated sequences and PSTs were used to investigate a hypothetical binary character trait, denoted ‘black’ / ‘white’. To investigate the null hypothesis, taxa were associated with character traits by randomly sampling without

replacement from 16 ‘black’ and 16 ‘white’ states. The assignments of traits to taxa were then applied to the full set of 3436 PSTs, with the assignments being resampled and randomized for each PST. The 3436 PSTs, each with a specified taxon-character association matrix, were analysed in our package BaTS. For each dataset, we calculated the AI, PS and MC statistics for the binary character trait and calculated the *p*-value of the null hypothesis test. The *p*-values for each statistic were then collated to generate an cumulative density function (CDF) of *p*-values for each statistic.

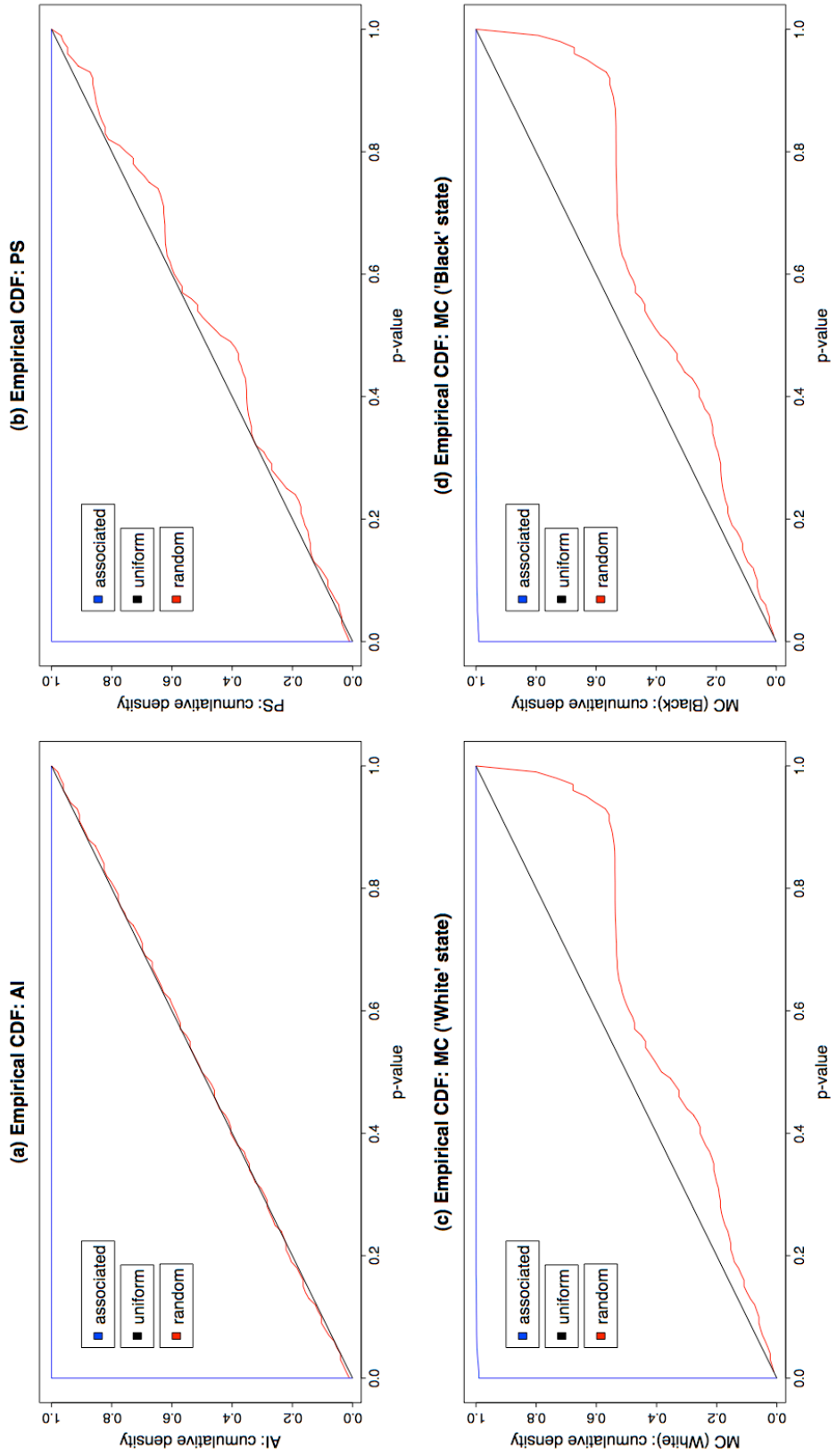


Figure 2.2

The expected unit uniform cumulative density function is a black line. The cumulative density functions for each statistic are shown (a) AI statistic, (b) PS statistic (c) MC(white) statistic (d) MS(black) statistic. 3436 simulated data sets were obtained under two models: the null 'random' model of taxon-character association (red crosses) or the 'completely associated' model (blue circles).

The CDF plots for data sets generated under the null hypothesis are shown in Figure 2.2. While the AI plot is a smooth curve that very closely matches the expected unit uniform distribution, the PS and MC plots deviate from the unit uniform distribution and feature several inflections. Using a two-sample Kolmogorov-Smirnov test (see Lilliefors 1969), we found that the CDFs of the AI and PS statistics did not depart from the theoretically expected unit uniform distribution; the MC statistic, however, did. The type 1 error rates of the AI and PS statistics, as implemented in a Bayesian MCMC framework, are therefore largely correct.

Finally, to explore the power of each statistic, we repeated the analysis above, but this time specifying a taxon-character matrix corresponding to a very strong phylogeny-trait correlation: the first 16 taxa were assigned the 'white' trait and the last 16, 'black'. Because the null hypothesis should be rejected in every case, the proportion of rejections at the p=0.05 level provides an estimate of the statistical power of our approach on 300nt sequences. The results were as follows: the null hypothesis was rejected for all 3436 simulated data sets when the AI and PS statistics were used. When the MC statistics were used, the null hypothesis was accepted only for 0.35% and 0.5% of the simulated data sets. These results are also contained in Figure 2.2, which shows the CDF of p-values equals 1.0 for almost all values of p. Therefore all the statistics show high statistical power when a strong tip-trait correlation does exist. The AI and PS statistics show slightly greater statistical power than the MC statistic.

The differences in type 1 error, above, likely result from the fact that AI is a more continuous metric than PS, which in turn can take more possible values than the MC statistic. Here, our simulated phylogenies all had 32 tips, so the range of values that discrete statistics can take on a single tree is limited, hence the MC and PS

statistics suffer from a lack of resolution. Furthermore, possible values of the MC statistic are further constrained to the sizes of the monophyletic clades in the tree whose tips all share a trait value. For instance, a perfectly symmetrical tree of 32 tips, 16 of which are ‘white’ and 16 of which are ‘black’, can only take MC(white) and MC(black) values of 1, 2, 4, 8 or 16. Such constraints may explain the significant departure from uniformity by the MC statistic observed here and the lesser degree of departure shown by the PS statistic. Hence researchers conducting multiple tests with the MC statistic should compare their observed CDF against an appropriate simulated null distribution, rather than the expected uniform distribution.

The statistics investigated here can be considered to span a continuum. At one end, the MC statistic is intuitive, can be scored by hand on a single tree, but has low resolution, reduced power and incorrect type 1 error rates. At the other end, the AI statistic is less intuitive and harder to calculate, but is a better-behaved statistic.

2.3.3 Empirical data sets

To evaluate the performance of our method on an empirical set of sequences associated with two trait values, we used two data sets presented by Carrington *et al* (2005) that represent the spread of Dengue virus Type 2 (DENV-2) and Dengue virus Type 4 (DENV-4) in the Americas. Each viral sequence is labeled as 'island' if it was sampled from a Caribbean island or 'mainland' if it was sampled from a Central or South American continental nation. In the original paper the authors calculated the PS statistic on single phylogenies that were estimated using maximum likelihood (ML), and concluded there was sufficient correlation between these geographical characters and the phylogeny to suggest that geography had been a key factor in the spread of Dengue virus in the Americas. We used these published ML trees to also calculate the AI and MC statistics for these data. Next, we reanalysed the DENV-4 and DENV-2 data sets using all three statistics implemented in a Bayesian MCMC framework (see Table 2.1). Overall, the agreement between the values obtained using the MCMC method and the values calculated from single ML trees was good. Our *p*-values validate Carrington *et al's* (2005) conclusions, with the exception of the MC(island) statistic for the DENV-2 dataset, which we found not to be significantly larger than that expected by chance.

Table 2.1: *DENV-2 and DENV-4 dataset results*

Statistic	single ML tree estimate	BaTS estimate (95% HPD CIs)	P-value (BaTS null hypothesis test)
DENV-2 data set			
AI	1.33	1.48 (1.07, 1.93)	<0.005
PS	12[a]	11.77 (11,12)	<0.005
MC (island)	4	5.62 (4,8)	0.185
MC (mainland)	15	16.04 (15,18)	0.01
DENV-4 data set			
AI	0.397	0.796 (0.336, 1.27)	<0.005
PS	7[a]	8.51 (7,10)	<0.005
MC (island)	14	16.08 (14,21)	0.01
MC (mainland)	4	4.66 (4,7)	0.03

[a] As reported in Carrington et al. (2005)
HPD CIs = highest posterior density confidence intervals (credible sets)

To evaluate the performance of our method on an empirical dataset with more than two character states, we re-analysed the data set published in Salemi *et al* (2005) This study examined HIV-1 sequences isolated from several tissue compartments of the central nervous system immediately after death. Among other aims, they sought to test the hypothesis of compartmentalization among the seven tissues sampled using the Slatkin-Maddison test (Slatkin & Maddison, 1989). This test was only performed for a subset of the data, so here we have calculated the PS and AI statistics from the ML tree presented in the original paper. Salemi et al. (2005) rejected the null hypothesis of no structure using the Slatkin-Maddison test and also reported the MC size statistic for each tissue sampled. Our results (Table 2.2) agree very closely with the original analysis; not only in the significance of the PS and MC statistics, but the actual MC sizes also matched closely. However, the AI value obtained from their ML tree fell outside our 95% CIs. It is likely that the large number of polytomies in the ML tree are responsible for this. Alternatively, due to phylogenetic bias, the ML tree may be a rather poor representative of the PST as a whole.

Table 2.2: *HIV dataset results*

Statistic	single ML tree estimate	*P-value (Finkelstein test)*	BaTS estimate (95% HPD CIs)	P-value (BaTS null hypothesis test)
AI	0.25[a]	-	1.78 (1.12, 2.51)	0.0056
PS	19[a]	-	19.34 (17,22)	0.0056
MC(frontal lobe)	15	7 x 10e-7	11.71 (6,16)	0.01
MC(occipital lobe)	19	3 x 10e-7	18.99 (18,19)	0.01
MC(meninges)	12	9 x 10e-7	12.32 (12,13)	0.01
MC(lymph nodes)	10	8 x 10e-7	8.76 (5,10)	0.0056
MC(temporal lobe)	11	6 x 10e-7	10.98 (10,11)	0.01
MC(seminal vesicles)	2	0.82	3.19 (2,5)	0.01
MC(spinal cord)	5	5 x 10e-4	5.01 (5,6)	0.01

[a] *Scored from the published ML tree.*

HPD CIs = highest posterior density confidence intervals (credible sets)

2.4 Discussion

Here we have developed a method for investigating phylogeny-tip correlation that accounts for phylogenetic uncertainty. Integrating over the set of all posterior trees is a qualitative and quantitative improvement over single-tree methods. It should produce better estimates of tree statistics and more accurate significance values. While it is reassuring that the published analyses' values fall within our 95% intervals, we also noticed that they did not always fall centrally (Tables 1 and 2). This indicates that statistics estimated from single trees may not accurately reflect the location of the bulk of the posterior probability. Maximum likelihood point estimates are known to be biased in many cases (Edwards, 1972), hence the PST may provide a more 'unbiased' estimate of a tree statistic than the value obtained from a single ML tree. In addition, Bayesian MCMC methods are thought to provide a better estimate of phylogenetic accuracy than the bootstrap or jacknife methods commonly used to assess ML trees (Alfaro *et al* 2003; Huelsenbeck & Rannala 2004).

Two situations common in studies of pathogen evolution may be particularly vulnerable to the errors that a single-tree approach can introduce. Firstly, data sets may have weak phylogenetic signal and large phylogenetic error (i.e. viral genetic diversity is low due to very strong negative selection, population bottlenecks, or low mutation rates). In such cases single tree estimates will have low bootstrap support values and a number of alternative branching orders may be equally plausible. By integrating over the PST, all these possible topologies are taken into account and, importantly, weighted by their posterior probability. Secondly, rapidly-growing or epidemic viral populations are common in viral epidemiology and typically give rise to star-like sample phylogenies. A single tree estimate of such trees may contain numerous polytomies, reflecting the lack of phylogenetic information about branching

order near the root. Again, by integrating over the PST we are able to take this uncertainty into account.

Our method is not without limitations. Firstly, the requirement for a PST means that the researcher must first carry out a Bayesian MCMC analysis of the data, which can be time consuming. Secondly, the low resolution of the MC statistic means that some care should be taken in interpreting multiple trials, as previously discussed. Finally, at present, we have only implemented statistics based on tree topology, and we have yet to investigate statistics that use both tree topology and branch length information (e.g. the PD, NTI/NRI and UniFrac measures). However, by placing the PS, AI and MC statistics in a Bayesian inference framework, our method does incorporate statistical variance arising from phylogeny estimation from sequences, which includes variation in both topology and branch lengths. This information is totally discarded by single-tree approaches. It would be useful to consider the PD, NTI/NRI and UniFrac statistics in a similar manner, and we plan to implement and evaluate these metrics in the near future.

Although the two empirical studies presented here focused on spatial characters, the methods could as easily be applied to any other phenotypic traits. Examples might include different risk groups or routes of infection in transmission networks, the different HLA genotypes of infected individuals, or the different disease symptoms of viral infections. The method is readily expandable to larger multi-state data sets, and could also be extended to consider continuously variable traits or ordered discrete states.

2.5 References

Alfaro, M.E., Zoller, S. & Lutzoni, F. (2003) Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov Chain Monte Carlo samplingand bootstrapping in assessing phylogenetic confidence. *Mol. Biol. Evol.* **20**(2):255-266.

Avise, J.C. (2000). *Phylogeography: The History and Formation of Species*. Harvard University Press, Cambridge, MA. 447pp.

Beerli, P. & Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *PNAS* **98**(8):4563-4568.

Bhattacharya T, Daniels M, Heckerman D, Foley B, Frahm N, Kadie C, Carlson J, Yusim K, McMahon B, Gaschen B, Mallal S, Mullins JI, Nickle DC, Herbeck J, Rousseau C, Learn GH, Miura T, Brander C, Walker B, Korber B. (2007) Founder effects in the assessment of HIV polymorphisms and HLA allele associations. Science 315:1583-6.

Carrington, C.V.F., Foster, J.E., Pybus, O.G., Bennett, S.N. & Holmes, E.C. (2005). Invasion and maintenance of Dengue Virus Type 2 and Type 4 in the Americas. *J. Virol.* **79**(23): 14680-14687.

Cochrane, A., Searle, B., Hardie, A., Robertson, R., Delahooke, T., Cameron, S., Tedder, R.S., Dusheiko, G.M., de Lamballerie, X. & Simmonds, P. (2002). A genetic analysis of Hepatitis C Virus transmission between injection drug users. *J. Infect. Dis.* **186**:1212-21.

Drummond, A.J. & Rambaut, A. (2003) BEAST v1.0, Available at http://evolve.zoo.ox.ac.uk/beast/

Drummond, A.J., Nicholls, G.K., Rodrigo, A.G. & Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**: 1307-1320.

Edwards, A.W.F. (1972). *Likelihood*. Cambridge University Press, Cambridge.

Ewing, G., Nicholls, G., Rodrigo, A. (2004) Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations. *Genetics*, **168**(4): 2407-2420.

Faith, D.P. (1992) Conservation evaluation and phylogenetic diversity. *Biol. Cons.* **61**:1-10.

Finkelstein, M., Fitch, W.M., Lanciani, C.A. & Miyamoto, M.M. (1998) Estimating the probabilities of identical events within biological sequences. *Mol. Biol. Evol.* **15**(4):470-472.

Fitch, W.M. (1971b). Toward defining the course of evolution: Minimal change for a specific tree topology. *Syst. Zool.* **20**: 406-416.

Fulcher, J.A., Hwangbo, Y., Zioni, R., Nickle, D., Lin, X., Heath, L., Mullins, J.I., Corey, L. & Zhu, T. (2004). Compartmentalization of Human Immunodeficiency Virus Type 1 between blood monocytes and CD4(+) T cells during infection. *Journal of Virology*, **78**(15):7883-7893.

Harvey, P. & Pagel, M. (1991). *The comparative method in Evolutionary Biology*. Oxford University Press, Oxford. 239pp.

Holder, M. & Lewis, P.O. (2003) Phylogeny estimation: Traditional and Bayesian approaches. *Nature Reviews Genetics* **4**:275-284.

Holmes, E.C., Zhang, L.Q., Robertson, P., Cleland, A., Harvey, E., Simmonds, P. and Leigh Brown, A.J. (1995) The molecular epidemiology of HIV-1 in Edinburgh, Scotland. *J. Infect. Dis.* **171**: 45-53.

Holmes, E.C. (2004). The phylogeography of human viruses. *Molecular Ecology* **13**:745-756.

Hudson, R.R., Boos, D.D. & Kaplan, N.L. (1992). A statistical test for detecting geographic subdivision. *Molecular Biology and Evolution* **9**(1):138-151.

Huelsenbeck, J.P., & Ronquist., F. (2001). MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* **17**:754-755.

Huelsenbeck, J.P., & Rannala, B. (2004). Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst. Biol.* **53**(6):904-913.

Jermiin, L.S., Olsen G., Mengersen K.L., Easteal S. (1997) Majority-rule consensus of phylogenetic trees obtained by maximum-likelihood analysis. *Molecular Biology and Evolution* **14**:1296-1302.

Kemal, K.S., Foley, B., Burger, H., Anastos, K., Minkoff, K., Kitchen, C., Philpott, S.M., Gao, W., Robison, E., Holman, S., Dehner, C., Beck, S., Meyer, W.A., Landay, A., Kovacs, A., Bremer, J. & Weiser, B. (2003). HIV-1 in genital tract and plasma of women: compartmentalization of viral sequences, coreceptor usage, and glycosylation. *Proceedings of the National Academy of Sciences of the United States of America,* **100**(22):12972-12977.

Komatsu H, Lauer G, Pybus OG, Ouchi K, Wong D, Ward S, Walker B & Klenerman P. (2006). Do antiviral CD8+ T cells select hepatitis C virus escape mutants? Analysis in diverse epitopes targeted by human intrahepatic CD8+ T lymphocytes. Journal of Viral Hepatitis 13:121-30.

Leigh Brown, A.J., Lobidel, D., Wade, C.M., Rebus, S., Philips, A.N., Brettle, R.P., France, A.J., Leen, C.S., McMenamin, J., McMillan, A., Maw, R.D., Mulcahy, F., Robertson, J.R., Sankar, K.N., Scott, G., Wyld, R. & Peutherer, J.F. (1997). The molecular epidemiology of human immunodeficiency virus Type 1 in six cities in Britain and Ireland. *Virology* **235**:166-177.

Lemey, P., O. G. Pybus, A. Rambaut, A. J. Drummond, D. L. Robertson, P. Roques, M. Worobey, & A. M. Vandamme. (2004). The molecular population genetics of HIV-1 group O. *Genetics* **167**:1059–1068.

Lemey P., van Dooren, S. & Vandamme, A-M. (2005) Evolutionary dynamics of human retroviruses investigated through full-genome scanning. *Molecular Biology and Evolution* **22**(4):942-951.

Lilliefors, H.W. (1967). On the Kolmogorov-Smirnov test for Normality with mean and Variance unknown. *Journal of the American Statistical Association* **62**(318): 399-402.

Lozupone, C. & Knight, R. (2005) UniFrac: A new method for comparing microbial communities. *App. & Environ. Microbiol.* **71**(12):8228-8235.

McGrath, K.M., Hoffman, N.G., Resch, W., Nelson, J.A.E. & Swanstrom, R. (2001). Using HIV-1 sequence variability to explore virus biology. *Virus Biology* **76**:137-160.

Nakano, T., Lu, L., Liu, P. & Pybus, O.G. (2004). Viral gene sequences reveal the variable history of hepatitis C virus infection among countries. *Journal of Infectious Disease* **190**:1098-1108.

Pillai, S.K., Kosakovsky Pond, S.L., Lui, Y., Good, B.M., Strain , M.C., Ellis, R.J., Letendre, S., Smith, D., Gunthard, H.F., Grant, I., Marcotte, T.D., McCutchan, J.A., Richmann, D. & Wong, K. (2006). Genetic attributes of cerebrospinal fluid-derived HIV-1 *env*. *Brain* **129**: 1872-1883.

Rambaut, A. & Grassly, N.C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics* **13**(3):235-238.

Rambaut, A. (2001). Phyl-O-Gen. Available at http://evolve.zoo.ox.ac.uk

Salemi, M., Lamers, S.L., Yu, S., de Oliveira, T., Fitch, W.M. & McGrath, M.S. (2005). Phylodynamic analysis of Human Immunodeficiency Virus Type 1 in distinct brain compartments provides a model for the neuropathogenesis of AIDS. *J. Virol* **79**(17): 11343-11352.

Sanjuan, R., Codoner, F.M., Moya, A. & Elena, S.F. (2004) Natural selection and the organ-specific differentiation of HIV-1v3 hypervariable region. *Evolution* **58**(6): 1185-1194.

Sheridan I, Pybus OG, Holmes EC, Klenerman P. 2004. High resolution phylogenetic analysis of hepatitis C virus adaptation and its relationship to disease progression. Journal of Virology 78:3447-54.

Slatkin, M., & Maddison, W.P. (1989). A cladistic measure of gene flow measured from the phylogenies of alleles. *Genetics* **123**(3):603-613.

Starkman, S.E., MacDonald, D.M., Lewis, J.C.M., Holmes, E.C. & Simmonds, P. (2003). Geographic and species association of hepatitis B virus genotypes in non-human primates. *Virology* **314**:381-393.

Wang, T.H., Donaldson, Y.K., Brettle, R.P., Bell, J.E. & Simmonds, P. (2001). Identification of shared populations of Human immunodeficiency Virus Type 1 infecting microglia and tissue macrophages outside the central nervous system. *J. Virol.* **75** (23): 11686-11699.

Webb, C.O. (2000) Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *Am. Nat.* **156**(2):145-155

Webb, C.O., Ackerly, D.D, McPeek, M.A. & Donoghue, M.J. (2002) Phylogenies and community ecology. Annu. *Rev. Ecol. Syst.* **33**:475-505

Wilson, G.A. & Rannala, B. (2003). Bayesian inference of recent migration rates using multilocus genotypes. *Genetics* **163**:1177-1191.

Wright, S. (1952). The theoretical variance within and among subdivision of a population that is in a steady state. *Genetics* **37**:312-321.

Chapter Three

Detailed analysis of within- and between-host rates of evolution in Hepatitis C virus using a relaxed-clock model.

Abstract

Most previous studies of Hepatitis C virus (HCV) have assumed that the rate of molecular evolution is constant through time (a 'strict clock' model of evolution.) We have analysed four HCV datasets under a state-of-the-art 'relaxed clock' model that allows for rate variation through time and between lineages. We find evidence to reject the strict clock in many cases, but with differences between the rate of evolution estimated for different HCV genes in within-host datasets when compared to between-host datasets. We conclude that the evolution of HCV during the course of infection is more rapid and complex than hitherto appreciated and that some key aspects of HCV evolution are poorly explained by our current understanding of its underlying biology.

3.1 Introduction

Determining the substitution rate of a lineage through time is a common goal of evolutionary biologists. Not only does is it a key parameter for understanding the processes of molecular evolution, but it is also of wider significance to a number of other disciplines. Most commonly, estimated substitution rates are used to estimate the likely divergence time of separate lineages (Bromham *et al.,* 1998; Hasegawa *et al,* 1985 Brown, 1980). In the context of pathogen evolution and epidemiology, evolutionary rates are used to estimate transmission rates (e.g. Pybus *et al.*, 2007; Pybus *et al.*, 2001 Mizokami, 2006) and the spatial spread of pathogens (ie. phylogeography; Verbeeck *et al.,* 2006; Carrington *et al.,* 2005; Holmes 2004), thereby helping to inform public health and clinical decisions. For example, the increasingly abundant nature of pathogen sequence data has enabled researchers to pinpoint the origin of the human immunodeficiency virus type-1 (HIV-1) epidemic (Worobey *et al.*, 2008; Korber *et al.*, 2000), to estimate the evolutionary relationships among circulating human influenza strains (Rambaut *et al.,* 2008; Taubenberger *et al.,* 2005) and to provide insights into HCV epidemiology (Pybus *et al*., 2004).

The hepatitis C virus (HCV) has been a virus of significant interest for nearly two decades since its discovery (Choo, 1989). Increased HCV transmission contributes to a growing number of cases of hepatocellular carcinoma (HCC) in many countries (Simmonds, 2004), and the virus infects around 170 million people worldwide (WHO (1999)) and is a significant cause of mortality. An estimated 4 million persons are

infected the United States alone, of whom 1-5% are expected to die as a result (CDC (2008))

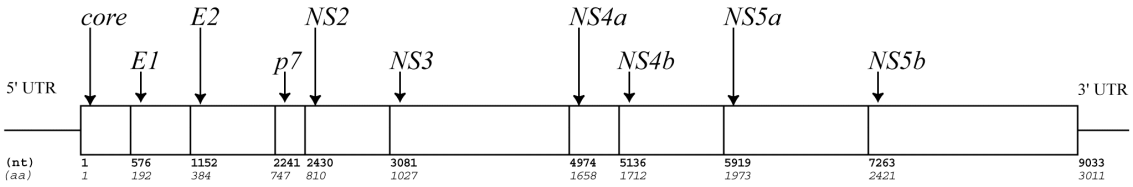


Figure 3.1: Schematic representation of HCV genome with characterized genes above and approximate nucleotide & amino acid positions (relative to the *core* gene N-terminus in the H77 numbering sequence) below. The first hypervariable region (HVR-1) lies at the N-terminus of the *E2* gene, at H77 positions 1152-1230.

HCV is an unsegmented positive-sense enveloped RNA virus, approximately 50nm in diameter, which belongs to the family Flaviviridae (genus: Hepacivirus). The genome (see Fig. 3.1) comprises three structural genes (*core*, *E1* & *E2*) that encode the core, nucleocapsid & envelope proteins, and seven non-structural genes (*p7*, *NS2*, *NS3*, *NS4a*, *NS4b*, *NS5a* & *NS5b*) that are thought to encode proteins responsible for RNA polymerase activity (*NS5b*), transmembrane proteins (p7), protease (*NS2/3*) and RNA helicase (*NS3*) activities, amongst other functions (see Moradpour *et al* (2007) and references therein). This central coding region (which totals around 9kb) is flanked by two untranslated regions (UTR). Despite the great relevance of HCV evolution to public health and the wealth of sequence data now publicly available, HCV evolutionary rates remain poorly-characterised across the genome as a whole, compared to other significant viruses such as human immunodeficiency virus (HIV), dengue or influenza.

Previous studies of HCV evolutionary rates have developed along with available methodologies (Supplementary Tables S1 & S2); early studies approximated the rate of molecular evolution in HCV lineages by simply counting the number of observed site changes in a sequence relative to some baseline. For instance, an early study by Ogata *et al* (1991) used this method in a single patient to estimate an average evolutionary rate for the whole genome of ~1.92 x 10^{-3} nucleotide substitutions / site / year (presciently, they found the majority of changes to have occurred in what we now know to be hyper-variable region 1 (HVR-1) of the E2 gene). Following a similar method, though in a chimpanzee host, Abe *et al* (1992) reported a rate of 0.9 x 10^{-3} nucleotide substitutions / site / year. Another study using a chimpanzee model similarly estimated the genomic evolutionary rate by a two-time-point comparison as ~1.44 x 10^{-3} nucleotide substitutions / site / year (Okamoto *et al* 1992). A later study by Booth *et al* (1998) of the the HVR-1 region only used similar methodology on a larger number of patients to arrive at higher rates of ~6.1 x 10^{-2} substitutions / site / year in control patients (n = 5) and ~5.14 x 10^{-3} substitutions / site / year in immunocompromised patients (n = 4) with common variable immunodeficiency. Allain *et* al (2000) investigated evolution in the E2 gene in six transmission networks, and obtained an evolutionary rate of 1.55 x 10^{-3} substitutions / site / year. Smith *et al* (1997) applied this technique more comprehensively in an investigation of the well-known "Irish cohort" of pregnant women, who were all infected after being treated with the same batches of HCV-contaminated anti-D immunoglobulin (see Power *et al* 1994 & 1995). Smith et al.'s (1997) analysis of the Irish anti-D dataset produced E1 and NS5B gene rate estimates of 7.4 x 10^{-4} and 4.1 x 10^{-4} nucleotide substitutions / site / year, respectively. Mizokami *et al* (2006) used a linear regression of

difference in sampling time against pairwise genetic distance to estimate the long-term within-host rate of HCV evolution in a chronically-infected human host (the 'Hutchinson' strain) over an interval of two decades, leading to an average rate estimate for the *core, E1* and *NS5* regions of 5.8 x 10^{-4} nucleotide substitutions / site / year. Pybus *et al* (2001) also analysed the Anti-D dataset, under a maximum likelihood phylogenetic model that incorporates time of sampling (Rambaut 2000) to estimate rates for the E1 gene (7.9 x 10^{-4} substitutions / site / year) and NS5b gene (5.0 x 10^{-4} substitutions / site / year.)

The mean substitution rate can differ greatly between genes within a genome, due to heterogenous selection pressures and genomic constraints (Graur & Li, 2000). It is therefore important to appreciate the relative evolutionary rate among genes. For HCV, great rate variation seems to occur in the two envelope genes, particularly in a series of hypervariable regions near the N-terminus of *E2* (Weiner *et al*, 1991; Kato *et al*, 1992; Troesch *et al* 2006). The precise function of the first and most-studied hypervariable region (HVR-1) is unknown, but HVR-1 substitutions are associated with increased immune escape and improved cell entry (Bartosch *et al.*, 2005). To this end, several studies have sought to investigate the variation in evolutionary rate along the HCV genome. Ina *et al* (1994) made an early attempt to compare evolutionary rates in different areas of the genome in a single study, finding wide variation in substitution rates from 4-7 x 10^{-3} substitutions / site / year over two time-points in a single patient. Itakura *et al* (2005) revisited the Anti-D data set, measuring the absolute number of substitutions that occurred over a known time-period as above but introducing a sliding-window approach. They estimated the rate of substitutions in *E1* & *E2* to be elevated nearly tenfold over

other regions of the genome. Furthermore, they estimated average genomic nucleotide substitution rates from donor to patient (~2.75 x 10^{-3} substitutions / site / year) and reported an elevated rate from donor to patients' later sequences (~9 x 10^{-3} substitutions / site / year) Since recombination in HCV is very rare, information about the relative rate of evolution in different genome regions can be very simply investigated by comparing local genetic diversity (e.g. Simmonds, 2004; Salemi & Vandamme, 2002) Both these studies revealed that across HCV genotypes the relative evolutionary rate is lowest in the UTRs, highest in the envelope gene regions, and intermediate in the non-structural genes, with the NS5a gene (which incorporates a V3 loop domain and protein kinase R activity) and the NS5b gene (which encodes the viral RNA polymerase) appearing to evolve slightly over and under the non-structural gene average, respectively. A separate gene-specific study of within-patient evolution from 25 blood donor / recipient pairs (Cantaloube *et al.*, 2003) estimated rates (by calculating the mean distance to an ancestral sequence) separately for donors (who had been chronically infected for some time) and recipients (who had only recently been infected). They found donor and recipient rates respectively to be 0.6 x 10^{-3} and 1.3 x 10^{-3} substitutions / site / year in the E1 gene region and slower at 0.4 x 10^{-3} and 0.8 x 10^{-3} substitutions / site / year in the NS5b gene.

As well as suffering from a relative paucity of data, previous studies of the evolutionary rate of HCV were based on analysis methods with significant drawbacks when applied to viral sequence data. All phylogenetic analyses using molecular sequence data have at their core a model of evolution at the nucleotide level: a set of assumptions about how, through mutation, one nucleotide may be substituted for another (see Chapter One,

section 1.4.2.1) Previously, it has been common to assume that the rate of nucleotide substitution is constant among lineages and through time. This is referred to as the 'strict' molecular clock hypothesis and is an attractive model since it promises that if the evolutionary rate is known then sequence divergence will linearly scale with time (Graur & Li, 2000). However, although substitution rates in HCV are known to vary across genes (Cantaloube *et al.*, 2003; Abe *et al.*, 1992) the previous HCV evolutionary rates analyses outlined above have all assumed a strict clock, either implicitly, in the case of distance-over-time based methods (e.g. Itakura *et al.,* 2005; Cantaloube *et al.,* 2003; Farci *et al.,* 2000; Smith *et al.,* 1997; Abe *et al.,* 1992; Ogata *et al.,* 1991) or explicitly, in the case of root-to-tip regression approaches (e.g. Mizokami *et al.,* 2006) or maximum-likelihood treatments (e.g. Pybus *et al.,* 2001).

However there are good reasons to believe that evolutionary rates can and do vary among lineages in viral populations. Firstly, the effective population sizes of pathogenic viruses can undergo rapid changes over short timespans (Kuntzen *et al.*, 2007; Shankarappa *et al*, 1999) – profoundly affecting the rate of fixation of slightly deleterious or slightly advantageous polymorphisms (Bromham & Penny, 2006; Bromham *et al.*, 2000) Also, viruses can undergo rapid molecular substitutions that may be strongly punctuated; for example, viral populations infecting new hosts may carry adaptations to the previous host, leading to a flurry of substitutions as the viral population reverts (Li *et al.,* 2007). Alternatively, antigenic pressures due to humoral or cellular immune pressure or treatment may cause the strength of selection to vary through time. Such fluctuations in the direction and strength of selection may lead to changes in the substitution rate

(reviewed most recently by Duffy *et al.,* 2008). Variation in generation times among lineages is also thought to contribute significantly to variations in substitution rate (Holmes, 2003). Crucially, empirical studies indicate that the molecular clock hypothesis may be rejected frequently for very many RNA viruses (Jenkins *et al.*, 2002) and some studies have specifically rejected the strict-clock hypothesis for within patient HCV evolution (Salemi & Vandamme, 2002; Pybus *et al.,* 2001).

A further weakness of previous studies (e.g. Salemi & Vandamme, 2002) is that they have evaluated the goodness-of-fit of the strict clock model on a single phylogenetic tree which is assumed to be correct – even though phylogeny estimation is subject to statistical error. In the past this limitation has been circumvented through bootstrapping procedures that provide information only on the confidence in the single best tree (Salemi & Vandamme, 2002). Fortunately, recently introduced methods based on Bayesian Markov-chain Monte Carlo (MCMC) evolutionary analysis (Drummond *et al.*, (2006) allow the variation in the substitution rate among lineages to be taken into account and estimated. Furthermore, Bayesian MCMC estimation methods take phylogeny uncertainty into account, by sampling a large number of phylogenies that, taken together, form the posterior distribution of trees given the sequence data (Drummond *et al* (2006)).

This study revisits existing sequence data to examine the substitution rate of HCV within and between patients. For the first time, a Bayesian MCMC technique with a relaxed-clock model is employed, using the most comprehensive model selection procedures available. A wide range of genomic regions (covering the whole genome in the within-

patient data set) are separately analyzed, but jointly contribute to phylogeny estimation, increasing precision. The goodness-of-fit of equivalent strict clock analyses is compared to test for heterotachy – variation in the rate of substitution among lineages.

3.2 Methods and Materials

3.2.1 Structure of this study.

In order to compare within- and between- host evolutionary processes I required two datasets; one that minimised the relative amount of within-host evolution observed and another that maximised it. Furthermore, because I sought separate independent estimates of the evolutionary rate in each dataset, as well as an estimate of the degree of heterotachy, I aimed to obtain data sets that comprise near continuous full-length sequences that were sampled at significantly different times. Fortunately, the between-host dataset assembled in Tanaka *et al* (2002) and well-known within-host 'Irish cohort' dataset of contaminated Anti-D immunoglobulin recipients (Power *et al* (1994, 1995) provided an excellent starting point. Although both these two datasets have been previously described in the literature, their structure is key to the present study and are summarized below for convenience.

The between-host datasets.

The 'Tanaka' dataset (Fig. 3.2) comprises two sequence alignments; one spans the *core, E1* genes & includes the first half of *E2*, whilst the other is a partial section of the *NS5b* gene. These sequences were published previously in Tanaka *et al* (2002) and their dates of sampling range from 1976 to 2000. These sequences can be considered a population-level representation of the HCV genotype 1a epidemic, that is, the data set represents among-host evolution. Because some of the sequences in this dataset are obtained from the same hosts, we discarded all but the first sample from any such patients. This gave an

alignment of nine sequences spanning sampling years 1976-1992 for the *NS5b* gene region. In the case of the *core-E1-E2* dataset, this left only six sequences, with a limited range of sampling dates. To maximise sampling time depth I therefore supplemented the core-E1-E2 dataset with additional sequences chosen from the Los Alamos HCV Sequence Database (http://hcv.lanl.gov) according to the following criteria: (i) sequences had to belong to the 1a genotype, (ii) were not recombinant, (iii) were sampled from human hosts in the US. Having identified a number of sequences that matched these criteria, I aligned the newly-incorporated sequences by hand in Se-Al (A. Rambaut; available from http://tree.bio.ed.ac.uk/software/seal/) and evaluated the phylogenetic structure of the whole dataset with a distance-based neighbour-joining tree (estimated under the HKY+γ+I model using PAUP* 4.0; Swofford, 1997). To reduce the computational load for the final analysis, I then down-sampled the additional sequences until 10 remained, which were genetically diverse (based on their positions in the estimated phylogeny, midpoint-rooted) and covered a wide range of sampling times. This 'enhanced Tanaka' *core-E1-E2* dataset therefore comprised 19 sequences (of 339nt), sampled from 1977-2005.

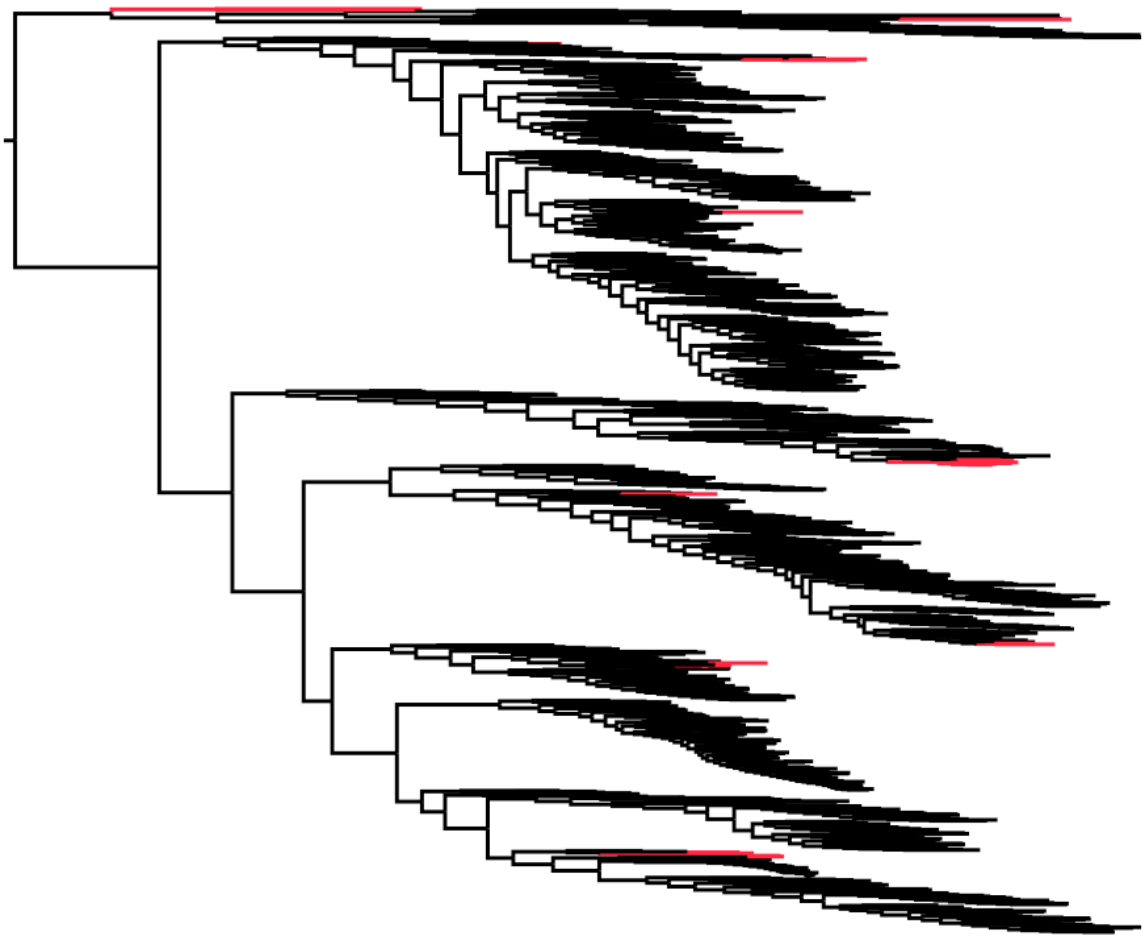


Figure 3.2: Diagrammatic representation of the context of the 'Tanaka' dataset of HCV sequence data, as reported in Tanaka *et al* (2002). From the U.S. national HCV epidemic (background sequences; black) a smaller number of patients were pseudo-randomly sampled and a single sequence isolated from each patient (foreground sequences; red.) Samples were collected over the period 1977 – 2000. This data set is expected to be reasonably indicative of the HCV genotype 1a population across the US.

To analyse the NS5b gene I combined the Tanaka *et al* (2002) dataset with that of Cochrane *et al* (2002), having obtained sequence sampling date information for the latter from the authors (A. Cochrane, *pers. comm.*). These two studies sampled the same region of NS5b and are complementary in date range. I also re-analysed the Europe-wide within-patient dataset published in van Asten *et al* (2004.)

The within-host dataset

To investigate the rate of evolution within infected individuals, I turned to the Anti-D dataset (Fig. 3.3), first reported by Power *et al* (1995). In this ethnically-homogeneous cohort, a number of pregnant women were all infected within 2 years of each other (1977-8) via the same batch of blood product generated from a single HCV-infected blood donation. Following identification of HCV virus in plasma and sample collection (1994-6) all were managed identically and a second sample taken between 1998-2000. The data set consists of full-length genomic HCV sequences from 15 patients, sampled at two time-points, plus an additional sequence sampled from the original HCV-infected blood donation from 1977 (n = 31; date range 1977-2000.) Since very low diversity was observed in the contaminated batch of blood product (McAllister *et al* (1998)), all recipients were infected with essentially the same viral sequence, and the resulting represents 15 independent within-host evolutionary histories. Furthermore, because the data set comprises full genome sequences I was able to estimate substitution rates over the whole HCV genome.

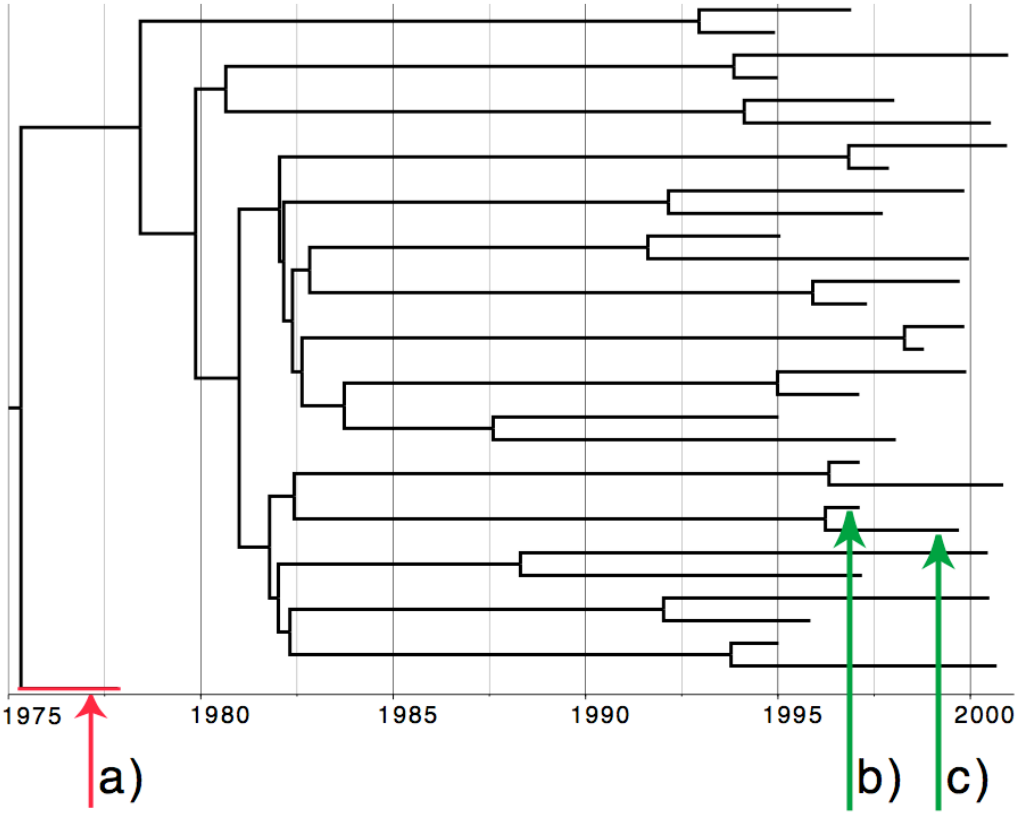


Figure 3.3: Diagram illustrative of the Anti-D dataset structure. The 1977 donor sequence is represented in red ('a'.) Following blood donation a single batch of contaminated blood product was prepared and administered to a number of recipients. A subset of 15 patients were sampled at early ('b') and late ('c') timepoints. This phylogeny is the highest-likelihood reconstruction from the posterior set of trees generated by a BEAST MCMC analysis of the Anti-D dataset under the most-favoured substitution, demographic and clock models (GTR+γ+I with the UCLN relaxed clock model, under the constant population-size demographic model; see 3.2.2 and 3.3.1 for details of model selection procedure and results).

3.2.2 BEAST partition model.

In order to estimate separate molecular clock and other evolutionary parameters whilst also accommodating phylogenetic uncertainty, I implemented a customised sequence partition model in BEAST version 1.4.6 (Drummond & Rambaut (2007)). Each alignment of sufficient length was partitioned a priori (the Tanaka *NS5b* dataset was not

long enough to be partitioned); I used non-overlapping partitions of 300bp, beginning from the start of the *core* gene. Separate evolutionary rate and clock model parameters were then estimated for each partition – but all partitions shares the same underlying nucleotide substitution model and phylogenetic tree model, thus minimising the variance in estimating these nuisance parameters.

Model selection procedure for BEAST analyses.

Although the theoretical framework for statistical selection of evolutionary models is well developed in a maximum likelihood framework, using methods such as the likelihood ratio test (LRT) and Akaike information criterion (AIC) and programs such as ModelTest (Posada & Crandall, 1998, 2001), Bayesian phylogeny estimation is a comparatively new technique, and methodologies for Bayesian evolutionary model selection are still undergoing some refinement. For instance, although Bayesian analyses performed by BEAST could be evaluated by applying LRT or AIC methods to the resulting posterior probability densities, such methods are incorrect in an MCMC setting and would be subject to bias (Posada & Crandall, (2001); Suchard *et al,* (2001); Alfaro & Huelsenbeck, (2006)). Fortunately, two models may be compared in a Bayesian setting by calculating the ratio of their marginal posterior probabilities given the data, otherwise known as the Bayes Factor (BF). A approximate method of approximating marginal posterior densities has been developed (Suchard *et al,* 2001) and is extensively used here to estimate BFs and perform model selection. An outline of the major steps in the analysis is provided below:

1. I began by investigating neighbour-joining trees for each data set. As a result of this preliminary analysis I discarded two sequences from the *core-E1-E2* Tanaka dataset which exhibited unusually long branches.
2. I then chose the best BEAST model combination by analysing whole data sets (i.e.with no partitions) under 16 different substitution model combinations (GTR or HKY85 models, with or without gamma-distributed rate variation, with and without invariant sites.) I estimated the log marginal posterior density of each model using Tracer v1.4 (http://tree.bio.ed.ac.uk/software/tracer) and subsequently calculated the Bayes' Factor for each model against the most-favoured model.
3. I retained those models with a BF < 10 decibans (dB) (Bayes' factors > 10 are generally considered to represent strong support for the favoured model (Kass & Raftery (1995); Suchard *et al* (2001)).
4. Having chosen a substitution model, I then repeated the model selection process to select the most appropriate coalescent models (from among the constant-size, exponential growth and Bayesian skyline plot models). For the Tanaka dataset (for which computational constraints were weaker due to shorter alignment length) it was possible to test all 16x3=48 model combinations directly. Where necessary I manually optimised MCMC operators to obtain good chain mixing and discarded models that failed to adequately converge after 10^8 states. Input (XML files) and output (log and tree files) are available on request.

5. Lastly I divided up the Anti-D and Tanaka *core-E1-E2* datasets into 300-nt partitions, as described above, for the final analysis. For the Anti-D data set, I also constrained all the recipient sequences to be a single monophyletic clade and placed a strong prior on the root height of this clade (normal distribition with mean=May 1976 and variance ± 1 year)

In all analyses, sites were further partitioned into 1st & 2nd codon versus a separate partition for 3rd codon positions only.

Maximum-likelihood tests of the strict clock hypothesis.

In the relaxed-clock model, among-lineage variation in the rate of evolution is modelled by assigning branches in the phylogeny with separate substitution rates drawn from either the uncorrelated exponential (UCED) or uncorrelated log-normal distribution (UCLN). The UCLN is used in the present study. The standard deviation of the relaxed clock UCLN posterior distribution branch rate parameter is sampled as well as the mean rate, and gives an indication of the degree of 'clocklikeness' in the dataset. If a standard deviation of, or near zero does not appear, we can infer that the branch-specific rates in the posterior set of phylogenies were all different. This strongly suggests that it is unlikely the sequences have evolved due to a strict molecular clock. However I also explicitly tested the molecular clock hypothesis for the Anti-D dataset in baseml (Yang (1997)), using the ML trees from our BEAST runs as input trees and calculating the likelihood of these trees under the estimated rate and substitution models under a strict (single rate on all branches) or a no-clock (individual, uncorrelated rates on all branches) molecular clock hypothesis. The ratio of these hypotheses' likelihoods is evaluated as a chi-square distribution with degrees of freedom equal to ($2n-1$) where n = nodes in the phylogeny (Yang, Z (1997)). Although the BEAST uncorrelated log-normal relaxed-clock (UCLN) coefficient of variation represents a less extreme model of clock variation than the 'no-clock' model in PAUP* (it allows for some branches to be, in fact, correlated (Drummond *et al* (2006))) and so the two methods represent slightly different tests of the strict clock hypothesis, I nonetheless expected that their results would largely agree.

Sitewise selection.

I investigated the ratio of nonsynonymous:synonymous amino acid replacements (dN/dS) in the Anti-D dataset, using the single-likelihood ancestor counting (SLAC) method implemented in HYPHY (Pond *et al.*, 2005.) In this method the most likely ancestor state for each ancestral codon is estimated by maximum-likelihood using the HKY model. I used the majority-rule (>50%) maximum clade credibility consensus tree calculated from the PST of the most-favoured BEAST PST. dN / dS was directly calculated for each site from the estimated number of synonymous and nonsynonymous changes, and the mean dN/dS calculated for consecutive 100aa (300nt) windows.

Additional materials. All sequences are available from online nucleotide sequence databases and accession numbers are listed in Supplementary Table S3 for convenience along with references and subtype, H77 position location and sampling dates.

3.3 *Results*

3.3.1 Between-host evolution: 'Enhanced Tanaka' data set.

The best-fitting model combination was the GTR + γ +I substitution model, with an uncorrelated lognormal relaxed-clock model and a constant-population size demographic model. This combination was not significantly favoured over the other two model combinations that differed only in demographic model selected (exponential growth and Bayesian skyline plot), however, all three were better-favoured than the three next most-favoured models, which shared similar molecular clock and substitution models (GTR + γ with UCLN relaxed-clock), by at least 9db. Similarly these in turn were followed in preference by the three coalescent models with a HKY + γ +I substitution model and UCLN relaxed-clock model (Table 3.1). All model combinations produced very similar posterior distributions of nucleotide substitution rate (Fig. 3.4).

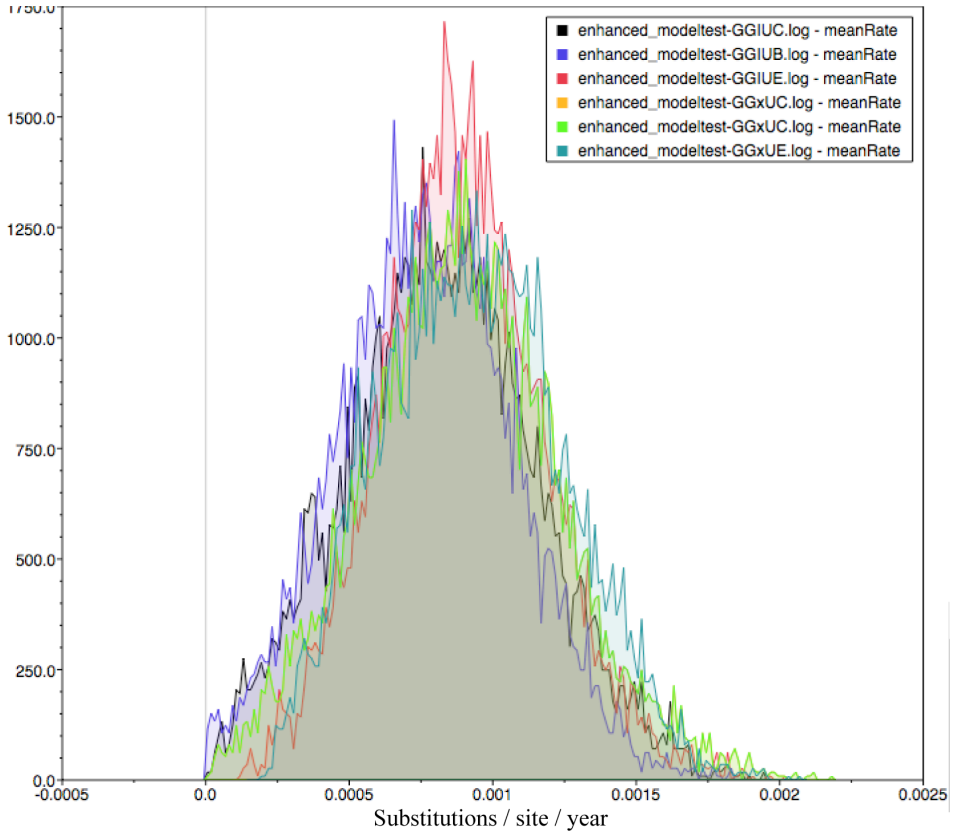


Figure 3.4: Horizontal axis: estimated evolutionary rates in substitutions.site^{-1}.year^{-1} in the *core-E1-E2* 'enhanced' Tanaka dataset. Vertical axis: log-posterior density of estimate.

Substitution model	Log marginal posterior density*	Substitution rate†	'Un-Clocklikeness'‡
GTR + γ + I, UCLN, Constant	-6742.315 *(0.199)*	8.18 *(1.19, 14.1)*	0.533 *(0.283. 0.848)*
GTR + γ + I, UCLN, Skyline plot	-6742.329 *(0.211)*	7.66 *(1.49, 13.5)*	0.369 *(0.251. 0.497)*
GTR + γ + I, UCLN, Exponential	-6742.633 *(0.214)*	8.94 *(3.46, 14.6)*	0.402 *(0.261. 0.556)*
GTR + γ, UCLN, Constant	-6753.456 *(0.185)*	8.92 *(1.85, 15.4)*	0.518 *(0.291. 0.815)*
GTR + γ , UCLN, Skyline plot	-6753.976 *(0.196)*	7.92 *(1.61, 13.8)*	0.365 *(0.243. 0.490)*
GTR + γ , UCLN, Exponential	-6754.325 *(0.227)*	9.31 *(3.39, 15.2)*	0.401 *(0.264. 0.556)*

Table 3.1: Evolutionary parameters of averaged over the *core-E1-E2* region of the Tanaka dataset, constant population size and UCLN clock models. *Standard deviations in parentheses. The difference in two models' log harmonic mean likelihoods is their Bayes' Factor and equivalent to a likelihood ratio test (LRT) in a maximum likelihood setting. A difference of greater than 8 is considered strong support for the preferred model. †in 10^{-4} substitutions.site^{-1}.yr^{-1}. Lower and upper 95% HPD intervals respectively in parentheses. ‡As measured by the coefficient of variation hyperparameter of the UCLN relaxed clock; a coefficient of variation of zero corresponds to zero branch rate heterogeneity and agrees with a strict clock model; whilst if the 95% HPD interval excludes zero a strict clock model can be rejected. Lower and upper 95% HPD intervals respectively in parentheses.

3.3.2 Between-host evolution: Cochrane / Tanaka *NS5b* data set.

The two best-fitting substitution models selected from those tested were HKY + γ and GTR + γ. The choice of demographic and clock model made little difference to the marginal posterior density (not shown) - mean posterior substitution rates reported by these model combinations agreed with the most favoured model combination, which gave a mean posterior estimate of substitution rate of 9.81×10^{-4} substitutions/site/year (95% HPDs: 6.9×10^{-4}, 12.9×10^{-4}). For all relaxed-clock models, the lower 95% credible interval for the "coefficient of variation" hyperparameter was 0.26 or higher, suggesting I could reject the strict molecular clock hypothesis for this data set.

3.3.3 Between-host evolution: Van Asten NS5b data set.

Under the most favoured model combination for this dataset (GTR + γ; relaxed-clock; Bayesian skyline plot) I estimated substitution rate to be 6.68×10^{-4} (95% HPDs: 2.17, 13.9×10^{-4}) substitutions / site / year. Other common model combinations gave similar estimates (Fig. 3.5). Again, for all relaxed-clock models, the lower 95% credible interval for the "coefficient of variation" hyperparameter was >0.3, suggesting the strict molecular clock hypothesis could be rejected for this data set.

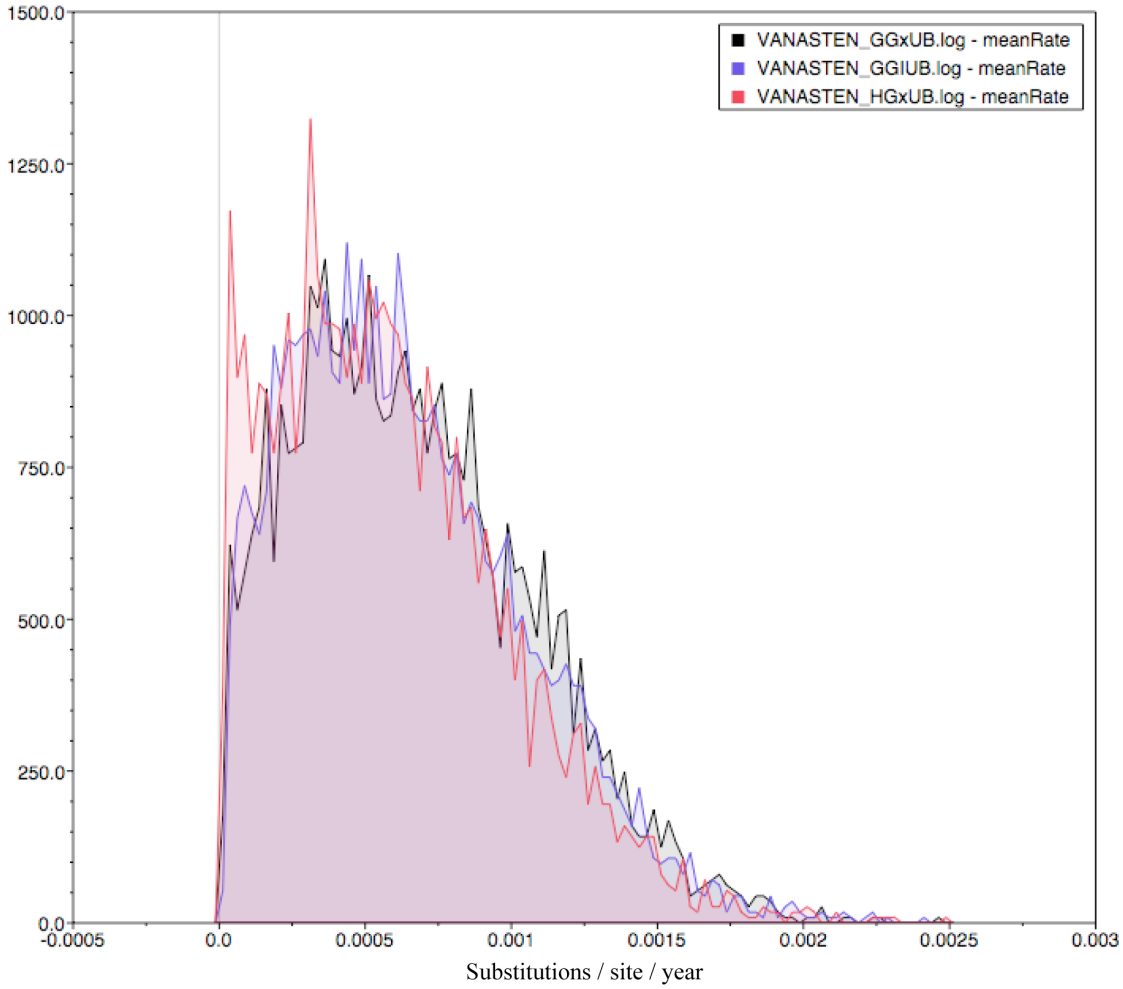


Figure 3.5: van Asten dataset, NS5b gene. Horizontal axis: posterior density of mean evolutionary rates (in substitutions.site^{-1}.year^{-1}). Vertical axis: log-posterior density of estimate.

3.3.4 **Within-host evolution (Anti-D dataset):**

Model selection procedures carried out on the entire coding region identified the GTR + γ with invariant sites and UCLN relaxed clock model, exponential growth demographic model as the most favoured model combination, by a substantial Bayes' factor of >15 db compared to the next-most-favoured model combination (Table 3.2.) Accordingly, this model was used as the basis for the subsequent partitioned analysis of the Anti-D data set, except that a strict clock model was used in place of the UCLN relaxed clock model in one comparison to determine whether or not choice of clock model affected mean posterior rate estimates. It did not: rate estimates obtained by UCLN and strict clock models agreed for all partitions (Fig. 3.6); note a particularly disproportionate rate magnitude compared with other partitions in partitions 4 & 5 (E1/E2 gene regions; H77 positions 1243-1843; including HVR-1 (H77 positions 1560-1638 (Abe *et al*, 1992).)) Codon rate-ratios (CRR) were plotted as (relative 1^{st} & 2^{nd} codon position rate : relative 3^{rd} codon position rate); these showed an uneven pattern of low ratios (higher 3^{rd} codon position relative rates) for most of the genome, except, again, for partitions 4 & 5. These displayed CRRs > 1, indicating that more substitutions occurred in the (1^{st} & 2^{nd}) codon positions than the 3^{rd} codon positions of these partitions (Fig. 3.6) A linear regression (Fig. 3.7) of CRRs and absolute mean substitution rate suggested ($R^2 = 0.82$) a positive correlation. Sitewise (dN/dS) values are shown in Fig. 3.6.)

The posterior distributions of partitions' 'un-clocklikeness', the tendency to reject the strict clock model as measured by the UCLN clock model coefficient of variation hyperparameter, fell into three classes: those that clearly failed to provide support for the strict clock (Fig. 3.8.); showed only weak support; or did not appear to reject the strict

clock model. A partition-wise plot of coefficient of variation showed little pattern in clocklikeness along the genome. Regression of lower 95% HPD estimates of coefficient of variation obtained by BEAST against odds-ratio of the 'no-clock' versus strict clock models in PAML gave a modest positive correlations ($R^2 = 0.61$; Fig. 3.9.)

Model[a]	Harmonic mean $\log_{10}$ Pr(tree model|data)[b]	Standard deviation[c]	Mean substitution rate[d]
GTR + γ + I (relaxed)	-30987.0323	*0.292*	1.16E-03
HKY + γ + I (relaxed)	-31031.9848	*0.2005*	1.15E-03
GTR + γ + I (strict)	-31045.775	*0.6073*	1.11E-03
HKY + γ + I (strict)	-31094.8224	*0.1582*	1.15E-03
GTR + I (relaxed)	-31129.1347	*0.2661*	1.10E-03
HKY + I (relaxed)	-31168.3883	*0.1855*	1.06E-03
GTR + I (strict)	-31188.3242	*0.5922*	1.08E-03
GTR + γ (relaxed)	-31227.624	*0.5866*	1.21E-03
HKY + I (strict)	-31251.3551	*0.3668*	1.08E-03
HKY + γ (relaxed)	-31299.9857	*0.179*	1.21E-03
GTR + γ (strict)	-31312.9716	*0.1552*	1.18E-03
HKY + γ (strict)	-31363.1455	*0.5643*	1.18E-03
GTR (relaxed)	-32149.4374	*0.4004*	1.07E-03
HKY (relaxed)	-32200.9575	*0.2942*	1.03E-03
GTR (strict)	-32210.6578	*0.1478*	1.08E-03
HKY (strict)	-32261.1739	*0.43*	1.04E-03

Table 3.2: Model comparison of substitution and clock models' performance on the Anti-D data set in BEAST v1.4.6. The whole genome was analysed in one analysis with 1st &2nd codon positions and the 3rd codon position sites evolving under separate substitution models. An exponential demographic growth model was specified as the coalescent prior. [a]Where 'HKY' and 'GTR' are the Hasegawa-Kishino-Yano 85 and General Time Reversible substitution models, and 'relaxed' and 'strict' the BEAST uncorrelated log-normally distributed (UCLN) relaxed and single-rate (strict) molecular clock models respectively. [b]The difference in harmonic mean log-likelihoods of tree likelihoods in BEAST is the Bayes' Factor and analogous to a likelihood ratio test (LRT; M. Suchard, pers. Comm..) A difference of more than 8 log-likelihoods (decibans) is considered strong support for the preferred model. [c]Of harmonic mean log-likelihood. [d]In substitutions.site^{-1}.year^{-1}.

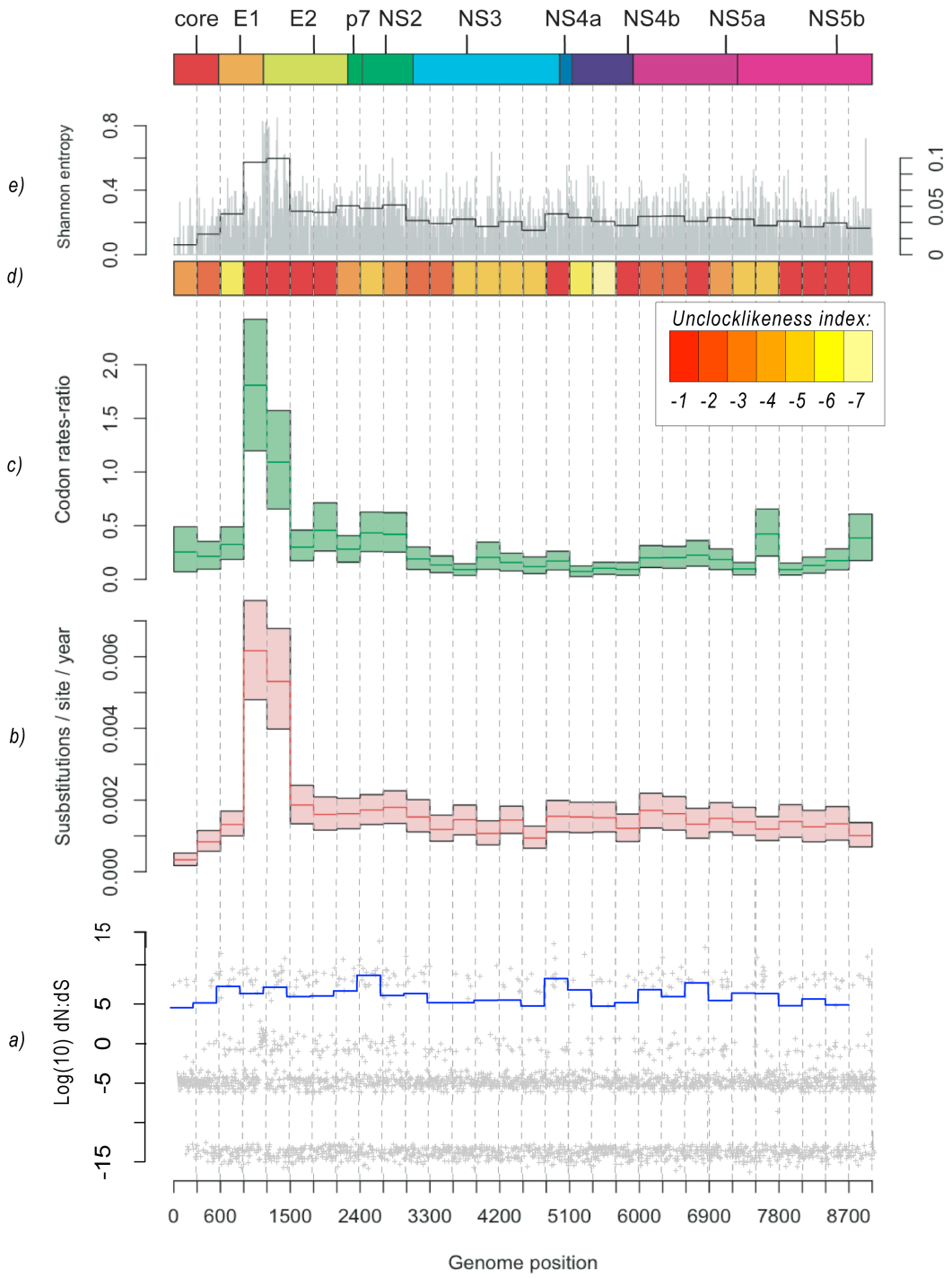


Figure 3.6 (legend on following page.)

Figure 3.6: Aspects of the evolutionary characteristics of within-host evolution in HCV (Anti-D dataset), plotted with respect to genome position (in nucleotides (nt); genome overview: top schematic). **Bottom plots (*a-e)*: *a*) Ratio of nonsynonymous : synonymous amino acid substitutions by individual codon (grey) and 100 amino acid windows (bue). *b)* (red)** Mean nucleotide substitution rates by position in 300-nt partitions. Upper and lower 95% confidence intervals represented by dotted lines. The mean posterior strict clock rate of evolution is always within the 95% HPD interval of the relaxed clock estimate of evolutionary rate.***c*)** (green): codon rate ratios (1^{st} + 2^{nd} codon positions' relative rate : 3^{rd} codon position relative rate) in within-patient Anti-D dataset. Compared with the ratio in the rest of the genome, a strikingly large number of substitutions occur in partitions IV & V in the first and second codon positions relative to the third codon position. *d*) clocklikeness of the Anti-D dataset as measured by the coefficient of variation of the uncorrelated lognormal relaxed clock in BEAST. Under this model, the strict clock is supported if coefficient of variation = 0, and cannot be rejected if the 95% CI of the posterior distribution includes 0. Lower 95% HPD of relaxed clock coefficient of variation orders of magnitude, from 10^{-6} (pale yellow; highly clocklike) to 10^{-1} (red; highly unclocklike). ***e*) plot:** mean alignment diversity, as measured by the Shannon entropy (Shannon (1948); Korber *et al.* (1994); see also Appendix One) measured sitewise (grey; primary y-axis) or averaged across each 300 bp partition (black; secondary y-axis.)

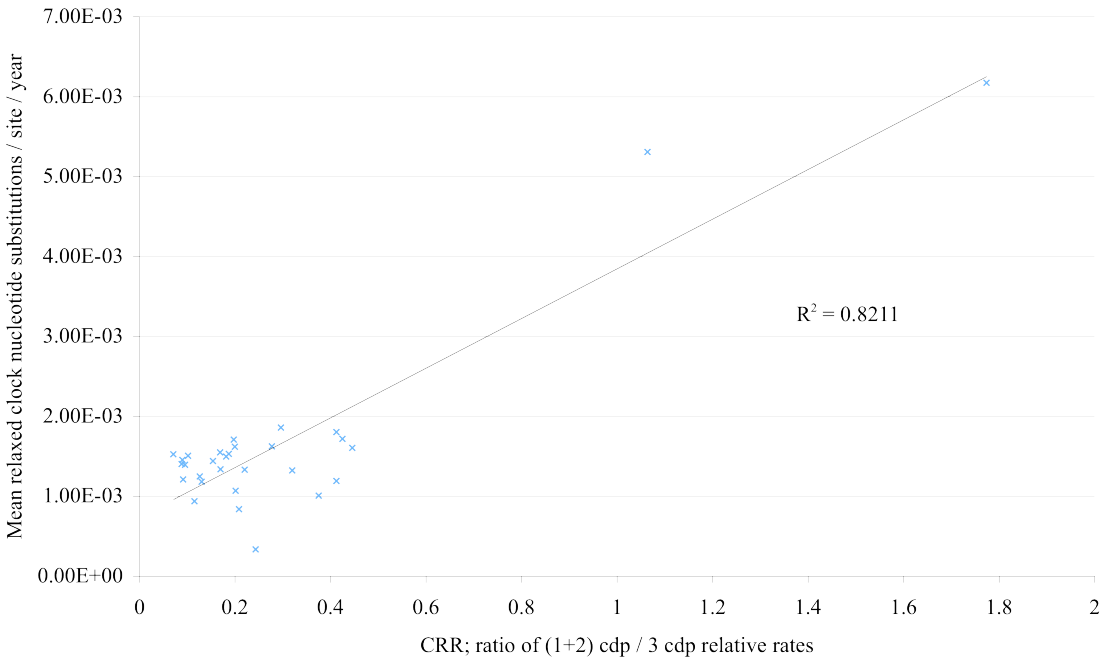
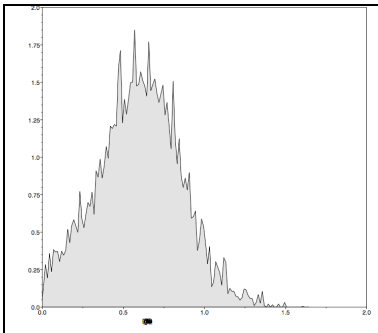
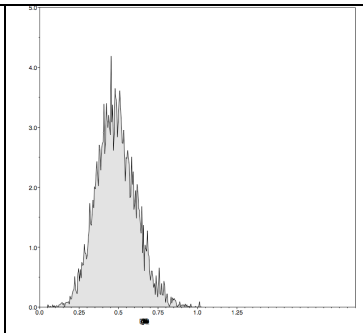


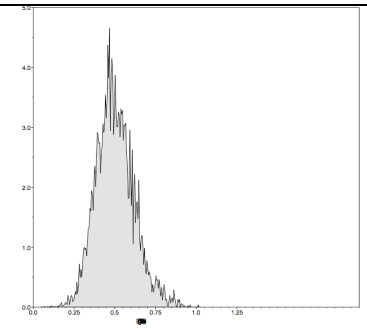
Figure 3.7: Scatterplot of codon rate ratios' correlation with absolute substitution rate in final selected partitioned Anti-D model. There is a strong linear correlation (R^2 – 82.1%) that suggests the increased numbers of substitutions likely to be found in rapidly-evolving partitions are also more likely to occur at first or second codon positions than in slower-evolving partitions



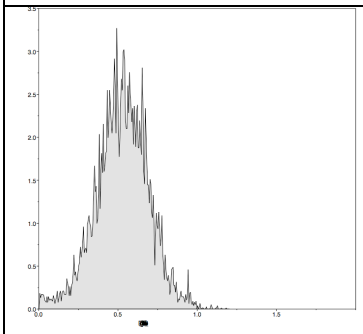
Partition 2



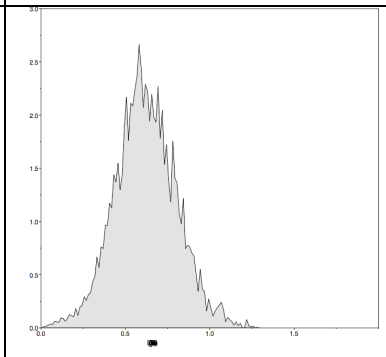
Partition 4



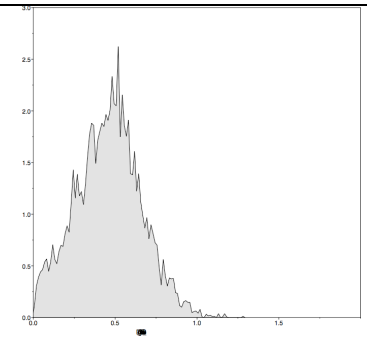
Partition 5



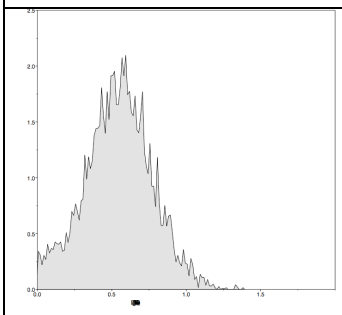
Partition 6



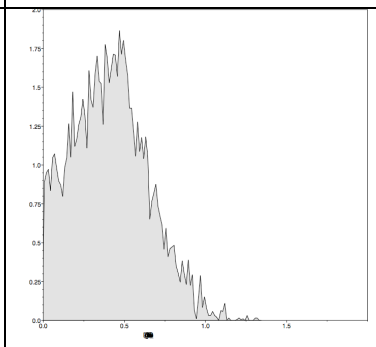
Partition 7



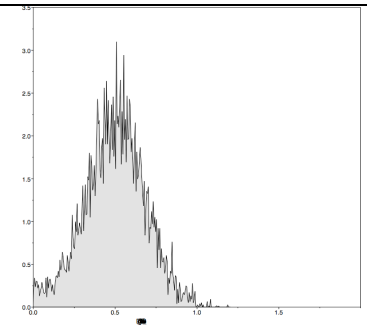
Partition 11



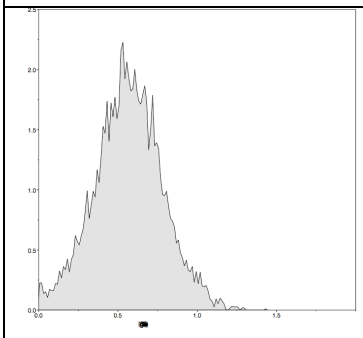
Partition 12



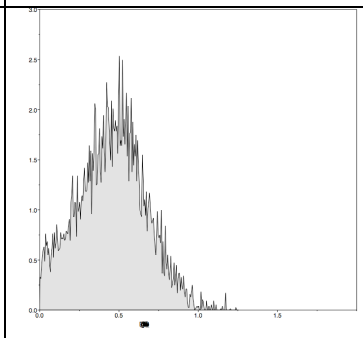
Partition 16



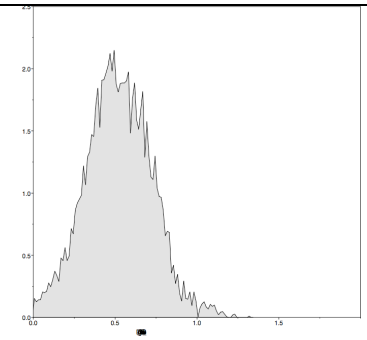
Partition 17



Partition 20



Partition 21



Partition 22

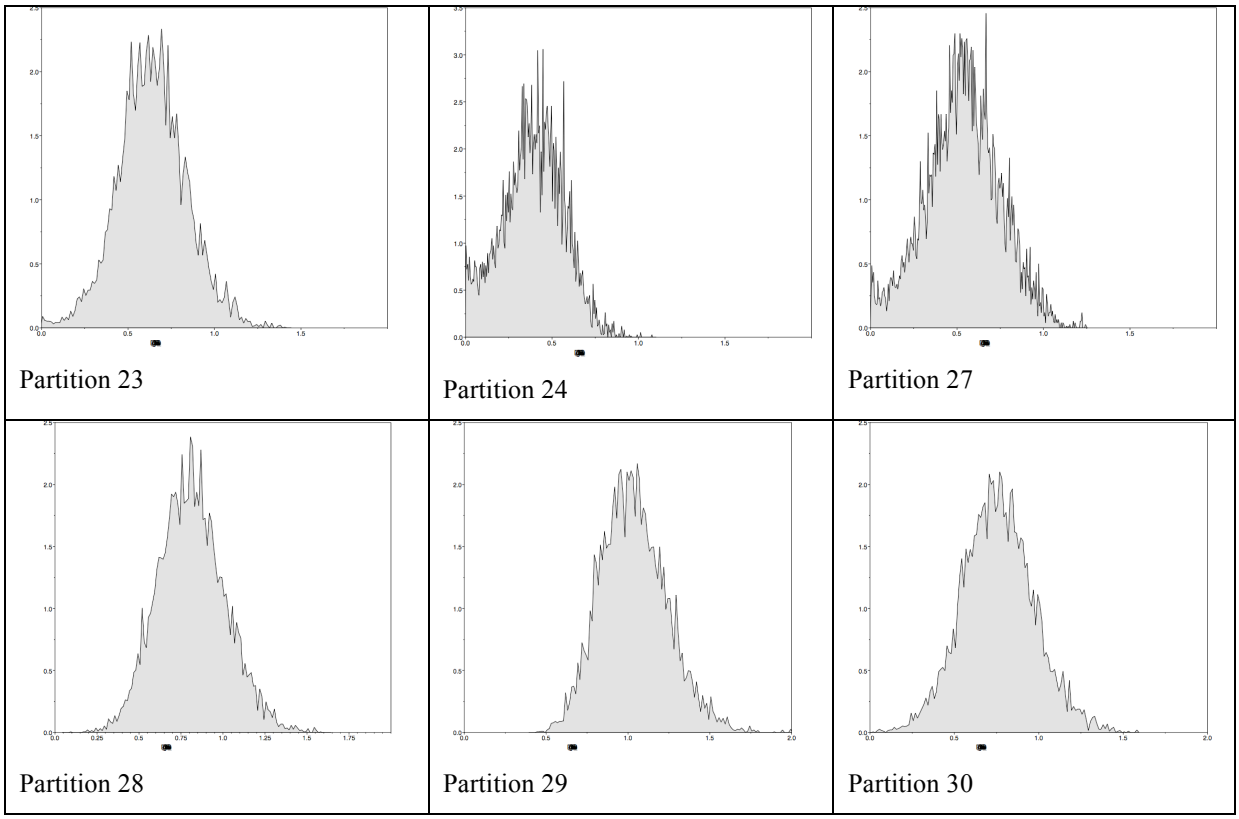


Figure 3.8: Posterior distributions of partitions that clearly reject the relaxed clock hypothesis. In these Partitions mean posterior coefficient of variation is large (0.5 or greater) and lower 95% HPD does not abut zero. The *E1*, *E2* & NS5b genes are all included (Partitions 2,4,5,6,7,11,12,16,17,20,21,22,23,24,27,28,29,30; a full table giving partition locations against the standard H77 reference strain can be found in Supplementary Sable S4)

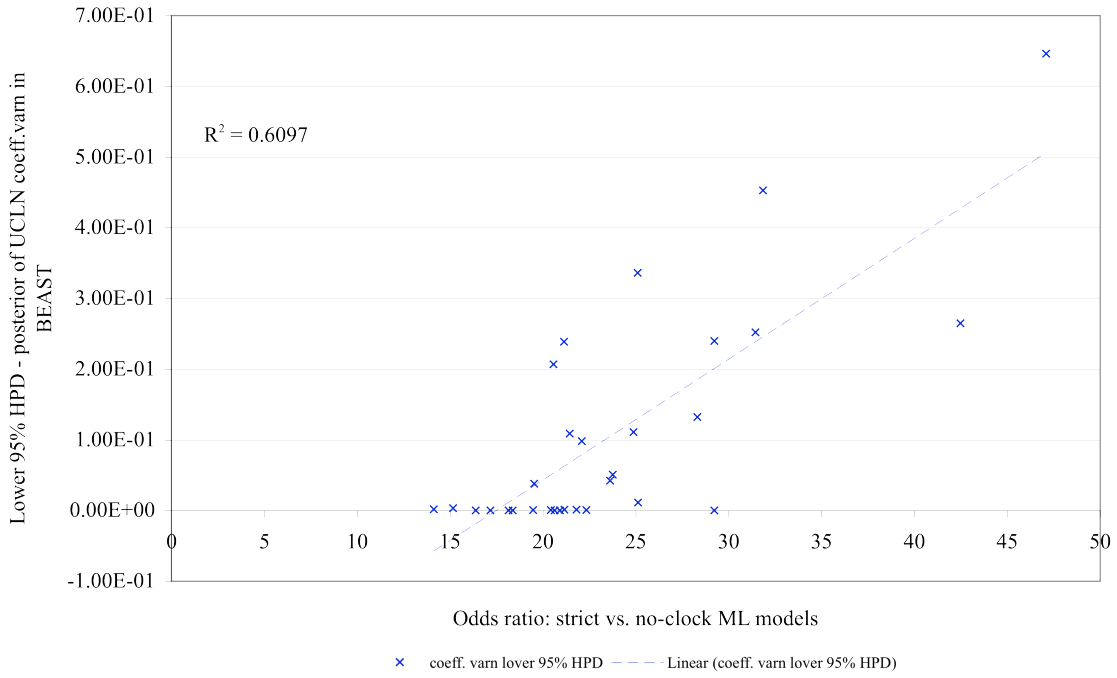


Figure 3.9: By taking the maximum likelihood tree from each Partition of the BEAST output and using this tree as a specified topology in baseml (Yang, 1997) we were able to conduct a likelihood test of the strict vs. no-clock hypotheses. Comparing this to the 'clocklikeness' data from the UCLN coefficient of variation hyperparameter shows a degree of correlation of support for clock hypothesis by BEAST and baseml LRT respectively.

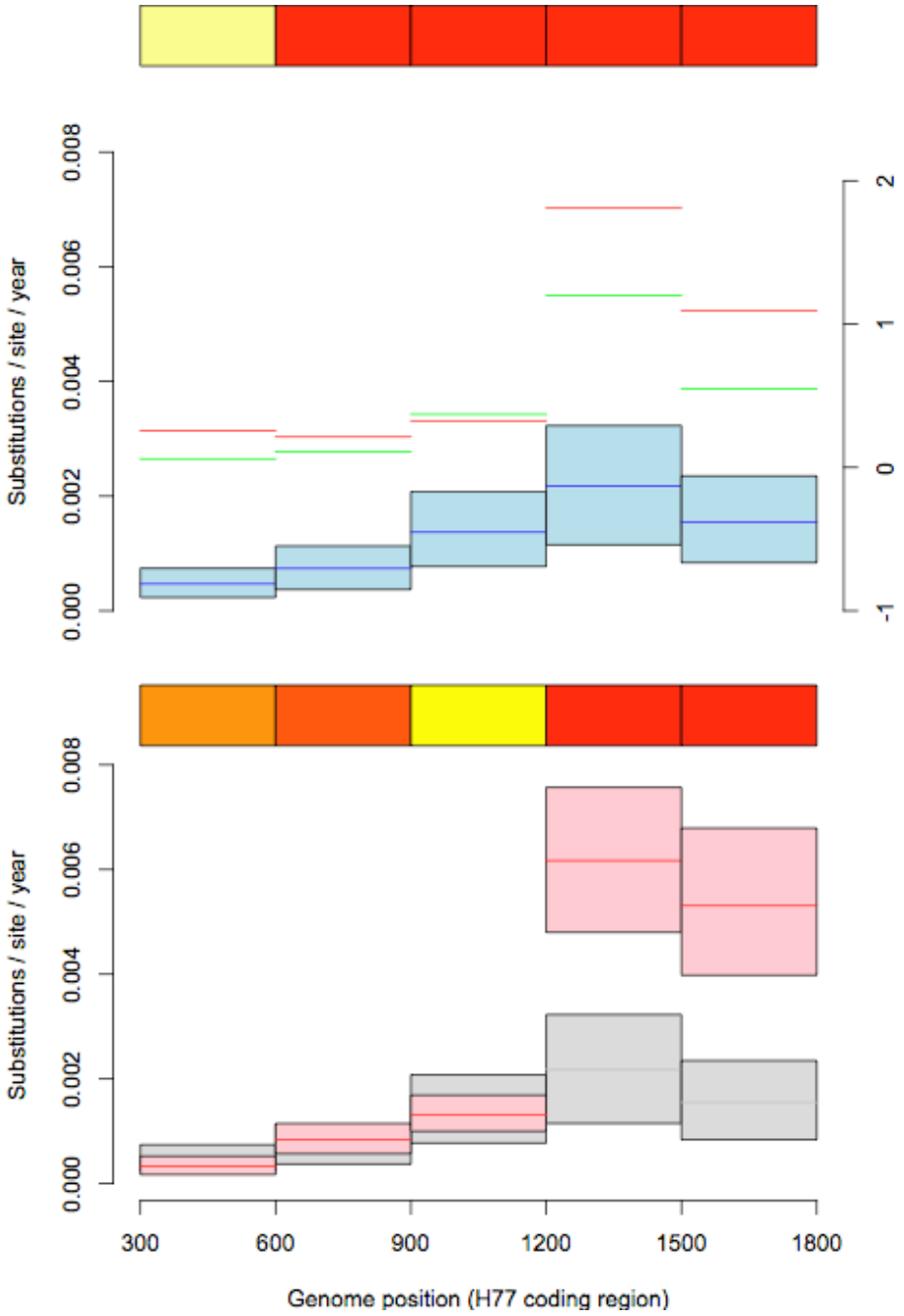
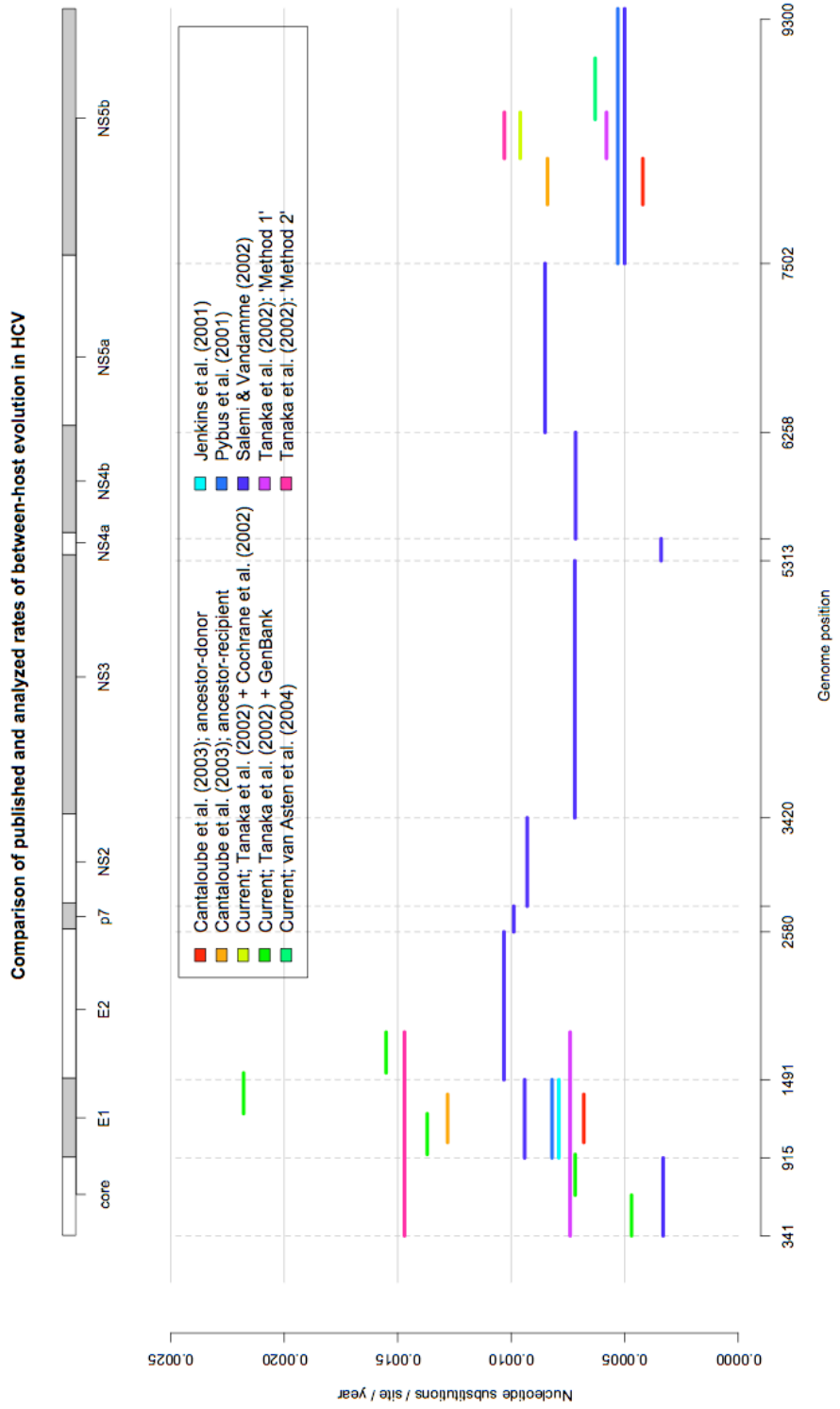
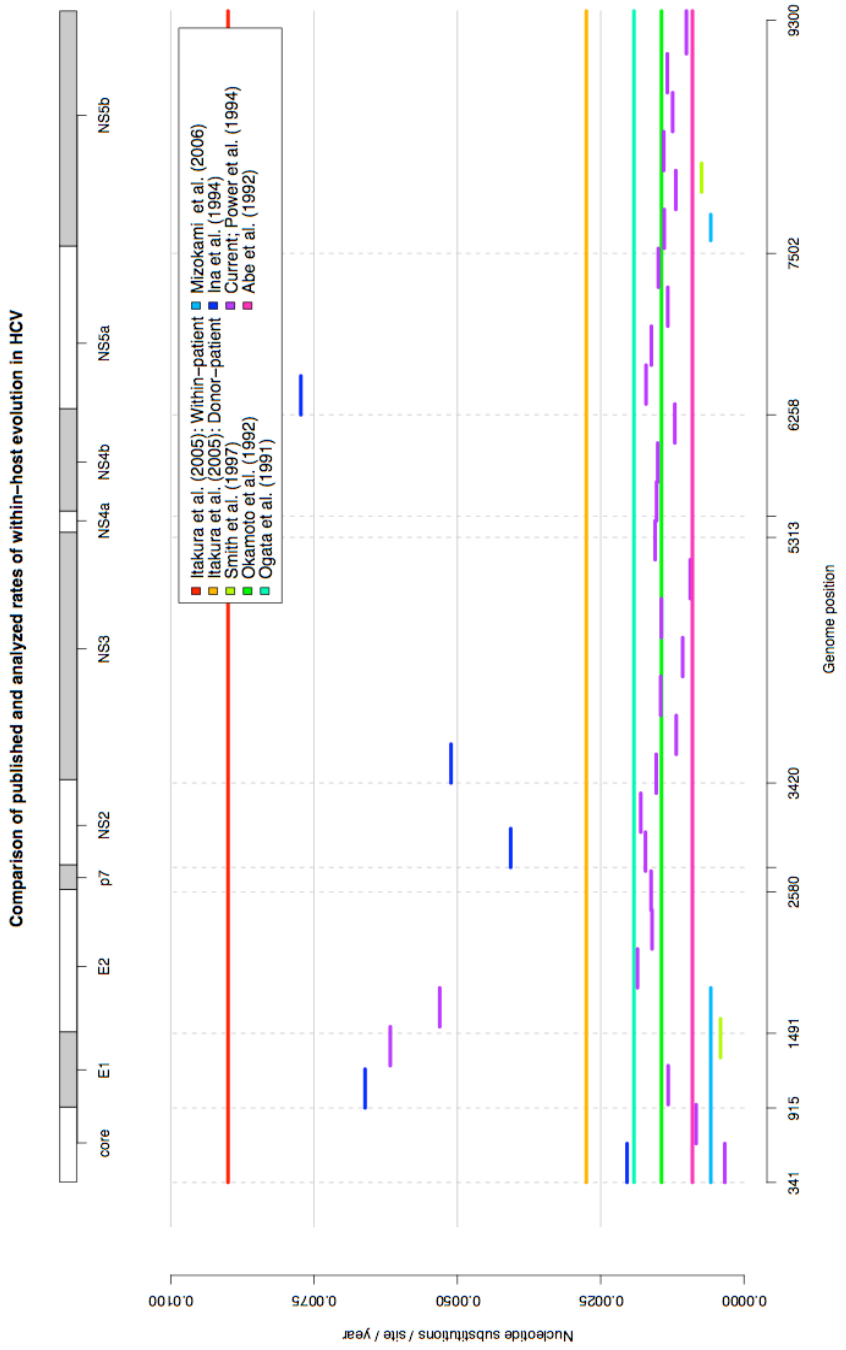


Figure 3.10: Comparison of between- and within- host evolutionary rates (upper; blue, and lower; red, plots respectively) in the *core-E1-E2* structural gene region of HCV. Shaded pink and light blue boxes represent 95% HPD confidence intervals. In the upper plot codon rate-ratios of (1^{st} & 2^{nd}) : 3^{rd} codon position relative rates are plotted on the secondary *y*-axis for between- (green) and within-host (red) data (confidence intervals not shown). In the lower plot within-host rates are superimposed over between-host rates (grey) for comparison. Temperature plots (from top) represent 'unclocklikeness', departure from the strict clock model, for between- and within-host data respectively (scaled as in previous figure.



: A summary of between-host HCV evolutionary rates in this and other studies. Horizontal axis shows genome position. Evolutionary rates shown in substitutions.site^{-1}.year^{-1}. Where more than one model has been used to estimate evolutionary rates over the same genome in primary literature, comparative value given is the arithmetic mean of those estimates. See text for details. N.B. Relative-rates analysis of Salemi & Vandamme (2002) included; scaled to μ(NS5b) = 0.0005 nucleotide substitutions / site / year

Figure 3.11



A summary of within-host HCV evolutionary rates in this and other studies. Horizontal axis shows genome position. Evolutionary rates shown in substitutions.site^{-1}.year^{-1}. Where more than one model has been used to estimate evolutionary rates over the same genome in primary literature, comparative value given is the arithmetic mean of those estimates. See text for details.

Figure 3.12

3.4 Discussion

Substitution rates in HCV are elevated at the within-host level.

Gratifyingly, the overall substitution rates estimated for the between-host datasets were clearly and consistently in line with previous expectations based on the literature (Fig. 3.11). All model combinations repeatedly showed a substitution rate for the *NS5b* gene of 5-10 x 10^{-4} substitutions.site^{-1}.year^{-1}, commensurate with previous studies' estimates. The more rapid rate of evolution in the 'Tanaka / Cochrane' dataset may be explained by the presence of some multiply-sampled patients in the Tanaka (2002) study, equivalent to a degree of within-host evolution. Similarly, the *core-E1-E2* section , although evolving more rapidly, can be considered to conform to expectations; the increase in rate (to around 9x10^{-4} substitutions.site^{-1}.year^{-1}) presumably reflects the inclusion of the *E2* hypervariable regions in this dataset. Other studies (notably Salemi & Vandamme (2002)) have noted a similar pattern. Significantly, substitution rates in the (within-host) Anti-D dataset were observed that were considerably faster than accepted rates for between-host evolution of HCV (Fig. 3.12). The genome-wide average estimate for this data set – around 1.3x10^{-3} substitutions.site^{-1}.year^{-1} – is two or three times higher than estimates of host population-level evolutionary rates, and the *E2* gene HVR appears to be evolving nearly an order of magnitude faster than published average genomic rates for inter-host HCV evolution.

The higher evolutionary rate observed in the Anti-D data set likely reflects the rapid nature of evolution on infection in pathogenic RNA viruses within their hosts. Just as the substitution rate is lower than the genomic per-replication mutation rate, since many

highly mutated copies are not viable virions and incapable of cell entry (Domingo *et al* (1996)), similarly over multiple transmission events a purifying selection effect works to revert many host-specific (and transmission-fitness-decreasing) mutations acquired (Li *et al*, 2007; Leslie *et al*, 2007) in response to selection pressures imposed by the host immune system, target cell epitopes, or drug treatments (Sheridan *et al* (2004)). This results in a lower apparent rate, seen in the between-host data set.

Evolutionary rates vary along the genome.

The differences between different regions' evolutionary rates in the within-patient (Anti-D) dataset are striking. The most salient point is the marked and sudden rate increase in *E2* (and especially the HVR) but other trends are also important. The slowest rate was observed in the *core* gene, which evolved within patients at less than half the genomic average. Secondly, the *E1* gene, although evolving more rapidly than the non-structural genes and *core* gene, nonetheless does so at a more pedestrian pace than *E2*. Lastly, and perhaps most surprisingly of all given relative rate variability in the structural genes, these results suggest that much less variation in relative rates exists between the non-structural genes. (c.f. Salemi & Vandamme (2002); Simmonds (2004); Itakura *et al* (2005))

Codon rates-ratios, not dN:dS, drive high mean substitution rates in the E1/E2 region.

Third codon substitution rates are typically expected to be greater than those for the first or second codon positions, since a larger range of possible substitutions in this position

are silent, producing synonymous amino acids on translation. This tends to be the case in the Anti-D dataset, although the pattern varies with gene. The non-structural (NS) genes appear to be more highly conserved overall, with relative codon rate ratios all below 0.2, indicating that 3rd codon rates in these positions are roughly a fifth of 1st and 2nd codon positions. The remaining genes appear less well conserved with higher ratios. However, it is surprising to notice that the *E2* gene shows markedly higher relative substitution rates in the important 1st & 2nd codon positions. We further analysed these by means of a simple linear regression of codon rate ratios (relative 1st + 2nd codon positions' rates : 3rd codon position rate) against absolute rate, for each Partition in the Anti-D dataset. Surprisingly, we found a strong (R^2 = 0.82) correlation. It therefore appears that it is an elevated rate of 1st & 2nd codon position nucleotide substitutions driving the very high overall absolute substitution rates in this area (N-terminus of the *E2* gene, including HVR-1 & HVR-2), rather than an increase in rates at the 3rd codon position. Since fewer possible nucleotide changes in the 1st & 2nd codon positions are degenerate, we would expect that this area is under strong positive selection. However the selection analysis showed that dN/dS was only modestly elevated in E1/E2 and NS5b regions as a whole, compared with the rest of the genome, and codon-rates ratios and dN/dS were not strongly correlated.

Heterotachy in HCV is extensive and only weakly explained by known virus biology.

Our results suggest poor goodness-of-fit for the strict molecular clock as a model for HCV evolution within hosts. In common with other reports (Salemi & Vandamme, 2002), this study was unable to confirm the molecular clock hypothesis robustly in any of the

datasets testes. 'Clocklikeness' in the Anti-D genomic data set did not correlate well with currently-availably biological information. Recalling that un-clocklike behaviour in this dataset represents variation in within-host evolutionary rates between hosts, two clear areas of rate heterogeneity may be explained by existing understanding: the E2 gene data supports strong rate heterogeneity, whilst three consecutive Partitions towards the C-terminus of NS5a also suggest variation in within-host substitution rates among patients. Although these two regions evolve at different mean rates over the whole data set and encode a very different set of biological activities, they can both be explained in terms of host environment heterogeneity. The E2 gene is known to be highly immunogenic and in our data (as in other datasets: Farci *et al* (2000); Salemi & Vandamme (2002); Simmonds (2004); Itakura *et al* (2005)) exhibited very strong evidence for positive selection and high site-specific diversity. These attributes reflect the close interaction of this gene with host biology, particularly the immune system. HCV ætiology frequently displays considerable variation among hosts in the form of differing disease progression outcomes and in strength of immune response; in fact four of the 15 patients in the Anti-D study had stable or decreasing viral loads, while the remainder displayed viral loads that increased over the sampling period. Data observed by Booth *et al* (1998) showing slower evolution in *E2* HVR1 in immunocompromised patients would seem to further support this idea.

The similar, though less marked, heterogeneity seen in the C-terminus of *NS5a* probably also reflects variations in host biology; this region of *NS5*a includes regions that determine interferon sensitivity (ISDR), protein kinase R (PKR) interaction activity, and

a variable V3 loop. Interpretation of the C-terminus of the *NS5b* gene results is less straightforward; this region combines very similar substitution rates to regions in the rest of the genome coding for the other non-structural genes with extremely strong evidence for non-clocklike behaviour. This suggests that there is very considerable variation in the nature of this gene's interactions between individual hosts; unlike the *E2* or *NS5a* regions, specific biological explanations in the literature are scarce – in fact, the *NS5b* gene is expected to evolve relatively stably, since it been used to infer deep viral phylogenies including retroviruses and even non-viral retroelements due to considerable homology to the reverse transcriptases of retroviruses (Xiong & Eickbush, 1990). Because four consecutive partitions displayed this property in this analysis it is unlikely that a handful of variant sites were responsible alone. However, it is possible that the ancestral (donor) sequence at this locus was very well adapted with respect to some hosts, but poorly adapted with respect to others. In this scenario, purifying selection would strongly constrain evolution in some hosts (where the viral quasispecies, being already optimally-adapted, was tightly clustered round a narrow adaptive peak); in other hosts (due to variations in host biology) the virus would be poorly-adapted; here the positive selection would be the dominant driver of evolution as the quasispecies expanded and moved towards optimal adaptation. It is conceivable that a combination of both scenarios lead to a balancing-out of both (non-neutral) modes of selection over the whole dataset, and hence might pass undetected by the codon rates-ratio indicator of selection, since this is a summary measure over the whole phylogeny. Further work is needed to investigate this possibility; related phylogenetic methods to detect selection over a whole phylogeny or alignment-based frequency methods would both be subject to the same potential

problems as the codon rates-ratio technique used here however. Therefore a lineage-specific phylogenetic approach is needed. Alternatively it is possible that secondary structural constraints operate on this gene at the RNA level that are not codon-position-specific.

The Tanaka data set of between-host samples supported for the strict clock hypothesis as measured between hosts across the US HCV genotype 1a epidemic more strongly. Lower HPDs of the UCLN relaxed-clock coefficient of variation hyperparameter clearly abutted zero, indicating clock-like evolution (although the mean and median coefficient of variation posterior estimates were larger suggesting more work may be needed to confirm this.)

3.6 Conclusion

This study is the first to explicitly compare within- and between-host datasets with large sample sizes and ranges of sampling times in this context, and represents a good estimate of these key evolutionary parameters. The biological picture they paint is a complicated one.

It seems that within-host evolution, represented by the Anti-D data set, is far more heterogeneous than between-host evolution. Heterotachy appears to be rife at this level of evolution. It may be that this results from phylodynamic behaviours that reflect the sometimes stochastic nature of the underlying population dynamics of an acute viral infection in a single host that transitions to a chronic infection, with cyclical periods of increased viræmia and bouts of strong selection mediated by the host immune system (e.g. Drummond *et al* (2003); Edwards *et al* (2006); Lemey *et al*, (2006)). In some circumstances, it may be safe to assume the strict molecular clock: although our results suggest only weak rate agreement or correlation between lineages over time during the course of host infections for most parts of the HCV genome, we also found that (with the notable exception of E2 HVR) estimates of the rate of evolution in individual Partitions arrived at by a strict- or relaxed-clock model agreed.

Drawing on these conclusions, I suggest clinicians, practitioners and researchers concerned about the rate of evolution of this virus ought to be mindful that the rate of evolution within a patient is more rapid than the reported rate of evolution between

patients, with much of the pace of the evolutionary arms race driven by the envelope genes. Furthermore, the rate of evolution itself may well vary widely over time, both in response to selection pressures from the immune system and treatments as well as population size.

In future a lineage-specific selection analysis following on this work may reveal interesting patterns of selection between hosts. Finally, this study raises some interesting biological questions; in particular, while known aspects of HCV-host interactions in the NS5a gene may explain observed rate heterogeneity, virus-host interactions as currently understood are not sufficient to account for the very strong rate heterogeneity observed in the NS5b gene.

3.7 References

Abe, K., Inchauspe, G. & Fujisawa, K. (1992) Genomic characterization and mutation rate of hepatitis C virus isolated from a patient who contracted hepatitis during an epidemic of non-A, non-B hepatitis in Japan. *J. Gen. Virol.* **73**:2725-2729.

Alfaro, M.E. & Huelsenbeck, J.P. (2006). Comparative performance of Bayesian and AIC-based measures of phylogenetic model uncertainty. *Syst. Biol.* **55**(1):89-96.

Allain, J.-P., Dong, Y., Vandamme, A-M., Moulton, V. & Salemi, M. (2000). Evolutionary rate and genetic drift of hepatitis C virus are not correlated with host immune reponse: studies of infected donor-recipient clusters. *J. Virol.* **74**(6):2541-2549.

Bartosch, B., Verney, G., Dreux, M., Donot, P., Morice, Y., Penin, F., Pawlotsky, J.-M., Lavillette, D. & Cosset, F.-L. (2005) An interplay between hypervariable region 1 of the hepatitis C virus E2 glycoprotein, the scavenger receptor BI, and high-density lipoprotein promotes both enhancement of infection and protection against neutralizing antibodies. *J. Virol.* **79**(13):8217-8229.

Booth, C.L., Kumar, U., Webster, D., Monjardino, J. & Thomas, H.C. (1998) Comparison of the rate of sequence variation in the hypervariable region of E2/NS1 region of hepatitis C virus in normal and hypogammaglobulinemic patients. *Hepatology* **26**:223-227.

Bromham, L., Rambaut, L., Fortey, R., Cooper, A. & Penny, D. (1998). Testing the Cambrian explosion hypothesis by using a molecular dating technique. *PNAS* **95**:12386-12389.

Bromham, L., Rambaut, A., Hendy, M.D. & Penny, D. (2000) The power of relative tests depends on the data. *J. Mol. Evol.* **50**:296-301.

Bromham, L. & Penny, D. (2003) The modern molecular clock. *Nature Rev. Genet.* **4**(3):216-224.

Brown, W. M. (1980) Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease assays. *PNAS* **77**:3605-3609.

Cantaloube, J.-F., Biagini, P., Attoui, H., Gallian, P., de Micco, P. & de Lamballerie, X. (2003) Evolution of hepatitis C virus in blood donors and their respective recipients. *J. Gen. Virol.* **84**:441-446.

Centres for Disease Control (2008) Surveillance for acute viral hepatitis. *CDC Morbidity and mortality weekly report,* **57:**SS-2.

Carrington CV, Foster JE, Pybus OG, Bennett SN & Holmes EC (2005) Invasion and maintenance of dengue virus type 2 and type 4 in the Americas *J Virol* **79**:14680-14687.

Choo, Q.L., Kuo. G., Weiner. A.J., Overby, L.R., Bradley. D.W. & Houghton, M. (1989) Isolation of a cDNA clone derived drom a blood-borne non-A, non-B viral hepatitis. *Science* **244**(4902): 359-362.

Choo, Q.-L., Richman, K.H., Han, J.H., Berger, K., Lee, C., Dong, C., Gallegos, C., Coit, D., Medina-Selby, A., Barr, P.J., Weiner, A.J., Bradley, D.W., Kuo, G. & Houghton, M. (1991) Genetic organization and diversity of the hepatitis C virus. *PNAS* **88**:2451-2455.

Cochrane, A., Searle, B., Hardie, A., Robertson, R., Delahooke, T., Cameron, S., Tedder, R. S., Dusheiko, G. M., De Lamballerie, X. & Simmonds, P. (2002) A genetic analysis of hepatitis C virus transmission between injection drug users. *J Infect Dis* **186**, 1212-21

Duffy, S., Shackleton, L.A. & Holmes, E.C. (2008) Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* **9**:267-276.

Domingo, E., Escarmís, C., Sevilla, N., Moya, A., Elena, S.F., Quer, I.S. & Holland, J.J. (1996). Basic concepts in RNA virus evolution. *FASEB Journal* **10**:859-864.

Drummond, A.J., Nicholls, G.K., Rodrigo, A.G. & Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**: 1307-1320.

Drummond, A.J., Pybus, O.G., Rambaut, A., Forsberg, R. & Rodrigo, A.G. (2003). Measurably evolving populations. *TREE* **18**(9):481-488.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J. & Rambaut, A. (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biology* **4**, e88

Drummond, A.J., Rambaut, A. (2007) "BEAST: Bayesian evolutionary analysis by sampling trees." *BMC Evolutionary Biology* **7**:214

Edwards, C.T.T., Holmes, E.C., Pybus, O.G., Wilson, D.J., Viscidi, R.P., Abrams, E.J., Philips, R.E. & Drummond, A.J. (2006) Evolution of the human immunodeficiency virus envelope gene is driven by purifying selection. *Genetics* **174**(3):1441-1453.

Farci, P., Shimoda, A., Coiana, A., Diaz, G., Peddis, G., Melpolder, J.C., Strazzera, A., Chien, D.Y., Munoz, S.J., Balestrieri, A., Purcell, R.H., Alter, H.J. (2000) The outcome of acute hepatitis C predicted by the evolution of the viral quasispecies. *Science* **288**:339-344.

Graur, D. and Li, W-H (2000). *Fundamentals of Molecular Evolution*, (2nd edition). Sinauer Associates.

Hasegawa, M., Kishino, M., Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**(2):160-174.

Holmes, E. C. (2003) Molecular clocks and the puzzle of virus origins. *J. Virol.* **77**(7):3893-3897.

Holmes, E.C. (2004). The phylogeography of human viruses. *Mol. Ecol.* **13**:745-756.

Ina, Y., Mizokami, M., Ohba, K. & Gojobori, T. (1994). Reduction of synonymous substitutions in the core protein gene of hepatitis C virus. *J. Mol. Evol.* **38**:50-56.

Itakuara, J., Nagayama, K., Enomoto, N., Hamano, K., Sakamoto, N., Fanning, L.J., Kenny-Walsh, E., Shanahan, F. & Watanabe, M. (2005) Viral load change and sequential evolution of entire hepatitis C viral genome in Irish recipients of single source-contaminated anti-D immunoglobulin. *Journal of Viral Hepatitis*, **12**:594-603.

Izopet, J., Rostaing, L., Sandres, K., Cisterne, J.-M., Pasquier, C., Rumeau, J.-L., Duffaut, M., Durand, D. & Puel, J. (2000) Longitudinal analysis of hepatitis C virus replication and liver fibrosis progression in renal transplant patients. *J. Infect. Dis.* **181**:852-858.

Jenkins, G.M., Rambaut, A.R., Pybus, O.G. & Holmes, E.C. (2002) Rates of molecular evolution in RNA viruses: A quantative phylogenetic analysis. *J. Mol. Evol.* **54**:156-165.

Kass, R.E & Raftery, A.E. (1995) Bayes factors. *J. Am. Stat. Ass.* **90**(430):773-795.

Kato, N., Ootsuyama, Y., Tanaka, T., Nakagawa, M., Nakazawa, T., Muraiso, K., Ohkoshi, S., Hijikata M. & Shimotohno, K. (1992) Marked sequence diversity in the putative envelope proteins of hepatitis C viruses. *Virus Res.* **22**:107–123.

Kenny-Walsh, E. (1999.) Clinical outcomes after hepatitis C infection from contaminated anti-D immune globulin. *NEJM* **340**:1228-1233.

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

Korber, B., Kunstman, K.J., Patterson, B.K., Furtado, M., McEvilly, M.M., Levy, R. & Wolinsky, S.M. (1994). Genetic differences between blood- and brain-derived viral sequences from human immunodeficiency virus Type 1-infected patients: evidence of conserved elements in the V3 region of the envelope protein of brain-derived sequences. *J. Virol.* **68**(11):7464-7481.

Korber, B., Muldoon, M., Thelier, J., Gupta, R., Lapedes, A., Hahn, B.A., Wolinsky, S., Bhattacharya, T. (2000) Timing the origin of the HIV-1 pandemic strains. *Science* **288**:1789-1796.

Kuntzen, T., Timm, J., Berical, A., L. Lewis-Ximenez, L.L., Jones, A., Nolan, B., Schulze zur Weisch, J., Li, B., Schneidewind, A., Kim, A., Chung, R.T., Lauer, G.M. & Allen, T.M. (2007) Viral sequence evolution in acute HCV infection. *J. Virol.* **81**:11658-11668.

Lemey, P., Rambaut, A. & Pybus, O.G. (2006) HIV evolutionary dynamics among and within hosts. *AIDS Rev.* **8**:125-140.

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

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

Li, B., Gladden, A.D., Altfeld, M., Kaldor, J.M., Cooper, D.A., Kelleher, A.D. & Allen, T.M. (2007). Rapid reversion of sequence polymorphisms dominates early human immunodeficiency virus type 1 evolution. *J. Virol.* **81**(1):193-201.

Li, W-H. (1997) *Molecular Evolution.* Sinauer, New York.

McAllister, J., Casino, C., Davidson, F., Power, J., Lawlor, E., Yap, P.L., Simmonds, P. & Smith, D.S. (1998) Long-term evolution of the hypervariable region of hepatitis C virus in a common-source infected cohort. *J. Virol.* **72**:4893-4905.

Mizokami, M., Tanaka, y. & Miyakawa, Y. (2006) Spread times of hepatitis C virus estimated by the molecular clock differ among Japan, the United States and Egypt in reflection of their distinct socioeconomic backgrounds. *Intervirology* **49**:28-36.

Moradpour, D., Penin, F. & Rice, C.M. (2007). Replication of hepatitis C virus. *Nat. Rev. Microbiol.* **5**:453-463.

Nakano, T., Lu, L., Liu, P. & Pybus, O.G. (2004) Viral gene sequences reveal the variable history of hepatitis C infection among countries. *J. Infect. Dis.* **190**:1098-1108.

Ogata, N., Alter, J.H., Miller, R.H. & Purcell, R.H. (1991) Nucleotide sequence and mutation rate of the H strain of hepatitis C virus. *PNAS* **88**:3392-3396.

Okamoto, H., Kojima, M., Okada, S.-I., Yoshizawa, H., Iizuka H., Tanaka, Y., Muchmore, E.E., Peterson, D.A., Ito, Y. & Mishiro, S. (1992) Genetic drift of hepatitis C virus during an 8.2-year infection in a chimpanzee: variability and stability. *Virology* **190**:894-899.

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

Posada, D. & Crandall, K.A. (1998). Modeltest: testing the model of DNA substitution. *Bioinformatics* **14** (9): 817-818.

Posada, D. & Crandall, K.A. (2001). Selecting the best-fit model of nucleotide substitution. *Syst. Biol.* **4**(1):580-601.

Power JP, Davidson F, O'Riordan J, Simmonds P, Yap PL, Lawlor E. (1995) Hepatitis C infection from anti-D immunoglobulin. *Lancet* **346**(8971):372-373.

Pybus, O.G., Charleston, M.A., Gupta, S., Rambaut, A., Holmes, E.C. & Harvery, P.H. (2001) The epidemic behaviour of the hepatitis C virus. *Science* **292**:2323-2325.

Pybus, O.G., Cochrane, A., Holmes, E.C. & Simmonds, P. (2004) The hepatitis C epidemic among injecting drug users. *Infect. Genet. Evol.* **5**(2):131-139.

Pybus, O.G., Markov, P.V., Wu, A. & Tatem, A.J. (2007) Investigating the endemic transmission of the hepatitis C virus. *Int. J. Parasitol.* **37**:839-849.

Rambaut, A. (2000). Estimating the rate of molecular evolution: incorporating non-contemperaneous sequences into maximum-likelihood phylogenies. *Bioinformatics* **16**:395-399.

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

Salemi, M. & Vandamme, A.-M. (2002) Hepatitis C virus evolutionary patterns studies through analysis of full-genome sequences. *J. Mol. Evol.* **54**:62-70.

Shankarappa, R., Margolick, J.B., Gnage, S.J., Rodrigo, A.G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C.R., Learn, G.H., He, X., Huang, X.L. & Mullins, J.I. (1999)

Constant viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* **73**:10489-10582

Shannon, C.E. (1948). A mathematical theory of communication *The Bell System Tech. J.* **27**:379-423

Sheridan, I., Pybus, O.G., Holmes, E.C. & Klenerman, P. (2004) High-resolution analysis of hepatitis C virus and its relationship to disease outcome. **78**(7):3447-3454.

Simmonds, P. (2004) Genetic diversity and evolution of hepatitis C virus – 15 years on. *J. Gen. Virol.* **85**:3173-3188.

Suchard, M.A., Weiss, R.E. & Sinsheimer, J.S. (2001) Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.* **18**:1001:1013.

Swofford, D.L. (1997), *PAUP**, Sinauer, Sunderland, MA.

Sy, T. & Jamal, M.M. (2006). Epidemiology of hepatitis C virus (HCV) infection. *Int. J. Med. Sci.* **3**(2):41-46.

Tanaka, Y, Hanada, K., Mizokami, M., Yeo, A.E.T., Shih, J.W.-K., Gojobori, T. & Alter, H.J. (2002). A comparison of the molecular clock of hepatitis C virus in the United States and Japan predicts that hepatocellular carcinoma incidence in the United States will increase over the next two decades. *PNAS* **99**(12):15584-15589.

Taubenberger, J.K., Reid, A.H., Lourens, R.M., Wang, R., Jin, G. & Fanning, T.G. (2005) Characterization of the 1918 influenza virus polymerase genes. *Nature* **437**(6):889-893.

Tavaré, S. (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.* **17**:57–86.

Troesch, M., Meunier, I., Lapierre, P., Lapointe, N., Alvarez, F., Boucher, M. & Soudeyns, H. (2006). Study of a novel hypervariable region in hepatitis C virus (HCV) E2 envelope glycoprotein. *Virology* **352**(2):357-367.

van Asten, L., Verhaest, I., Lamzira, S., Hernandez-Aguado, I., Zangerle, R., Boufassa, F., Rezza, G., Broers, B., Robertson, J.R., Brettle, R.P., McMenamin, J., Prins, M., Cochrane, A., Simmonds,, P., Coutinho, R.A. & Bruisten, S. (2004) European and Italian seroconverter studies. Spread of hepatitis C virus among European injection drug users infected with HIV: a phylogenetic analysis. *J Infect Dis* **189**: 292-302

Verbeeck, J., Maes, P., Lemey, P., Pybus, O.G., Wollants, E., song, E., Nevens, F., Fevery, J., Delport, W., Van der Merwe, S. & Van Ranst, M. (2006). Investigating the origin and spread of hepatitis C virus genotype 5A. *J. Virol.* **80**(9):4220-4226.

Weiner, A.J., Brauer, M.J., Rosenblatt, J., Richman, K.H., Tung, J., Crawford, K., Bonino, F., Saracco, G., Choo, Q.L., Houghton, M., *et al* (1991) Variable and hypervariable domains are found in the regions of HCV corresponding to the flavivirus envelope and NS1 proteins and the pestivirus envelope glycoproteins. *Virology* **180**(2):842-848.

Worobey, M., Gemmel, M., Teuwen, D. E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J.-J., Kabongo, J.-M. M., Kalengayi, R. M., Van Marck, E., Gilbert, M. T. P. & Wolinsky, S. M. (2008). Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* **455**:661-664.

Xiong, Y. & Eickbush, T.H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9**(10):3353-3362.

Yang, Z. (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. CABIOS 13:555-556 (http://abacus.gene.ucl.ac.uk/software/paml.htm)

.

Supplementary Information

Gene	Reference	Rate	Dataset	Comments
Core	(this study)	0.6	Tanaka *et al.* (2002) + GenBank seqs	(mean of BEAST partition models)
	Tanaka *et al.* (2002)	0.74	ibid.	Method 1; Root-tip regresssion; Core, E1, E2, & NS5b (combined)
		1.47	ibid.	Method 2; Root-tip regresssion; Core, E1, E2, & NS5b (combined)
E1	(this study)	1.77	Tanaka *et al.* (2002) + GenBank seqs	(mean of BEAST partition models)
	Cantaloube *et al.* (2003)	0.68	ibid. Ancestor -> donor	25 patients; donor-recipient pairs
		1.28	ibid. Ancestor -> recipient	25 patients; donor-recipient pairs; distance to recipient.
	Jenkins *et al.* (2001)	0.79	ibid.	ML tree / TipDate. Dataset origin unknown.
	Pybus *et al.* (2001)	0.79	Power *et al.* (1994;1995)	Tip-Date
	Tanaka *et al.* (2002)	0.74	ibid.	Method 1; Root-tip regresssion; Core, E1, E2, & NS5b (combined)
		1.47	ibid.	Method 2; Root-tip regresssion; Core, E1, E2, & NS5b (combined)
E2	(this study)	1.87	Tanaka *et al.* (2002) + GenBank seqs	(mean of BEAST partition models)
	Tanaka *et al.* (2002)	0.74	ibid.	Method 1; Root-tip regresssion; Core, E1, E2, & NS5b (combined)
		1.47	ibid.	Method 2; Root-tip regresssion; Core, E1, E2, & NS5b (combined)
	Allain *et al.* (2000)	1.55	ibid. (HVR – 1 only)	Range 0.34 -> 4.51 (6 donor-recipient networks)
NS5b	(this study)	0.63	van Asten *et al* (2004)	range 0.3 -> 13.9
		0.96	Tanaka *et al* (2002); Cochrane *et al.* (2002)	range 0.69 -> 12.5
	Cantaloube *et al.* (2003)	0.42	ibid. Ancestor -> donor	25 patients; donor-recipient pairs
		0.84	ibid. Ancestor -> recipient	25 patients; donor-recipient pairs; distance to recipient.
	Pybus *et al.* (2001)	0.5	Power *et al.* (1994;1995)	Tip-Date
	Tanaka *et al.* (2002)	0.58	ibid.	Method 1; Root-tip regresssion
		1.03	ibid.	Method 2; Root-tip regresssion
		0.74	ibid.	Method 1; Root-tip regresssion; Core, E1, E2, & NS5b (combined)
		1.47	ibid.	Method 2; Root-tip regresssion; Core, E1, E2, & NS5b (combined)

Table S1: comparison of published and analyzed between-host nucleotide substitution rates.

Gene	Reference	Rate	Dataset	Comments
Core	(this study)	0.59	Power *et al. (*1994, 1995); Smith *et al.* (1997); Itakura *et al.* (2005)	(mean of BEAST partition models)
	Ina *et al.* (1994)	2.04	ibid.	2 timepoints, H77 -> H90. Used other DDBJ seqs to build tree. Distance calculated as difference in distances from seq -> outgroup.
	Mizokami *et al.* (2006)	0.58	Choo *et al.* (1989; 'Hutchinson strain')	Root-tip regression; 5 samples (single patient) range 1977-1996. Core, E1 & NS5b (combined)
E1	(this study)	3.75	Power *et al. (*1994, 1995); Smith *et al.* (1997); Itakura *et al.* (2005)	(mean of BEAST partition models)
	Mizokami *et al.* (2006)	0.58	Choo *et al.* (1989; 'Hutchinson strain')	Root-tip regression; 5 samples (single patient) range 1977-1996. Core, E1 & NS5b (combined)
	Smith *et al.* (1997)	0.74	Power *et al.* (1994;1995)	Average over patients (*n*=23).
E2	(this study)	3.31	Power *et al. (*1994, 1995); Smith *et al.* (1997); Itakura *et al.* (2005)	(mean of BEAST partition models)
	Ina *et al.* (1994)	6.61	ibid.	2 timepoints, H77 -> H90. Used other DDBJ seqs to build tree. Distance calculated as difference in distances from seq -> outgroup. Labelled 'E' in paper.
	Booth *et al.* (1998)	60.6	ibid.; HVR-1 only; control patients	Five patients, 2-3 timepoints. Average over patients.
		5.38	ibid; HVR-1 only; CVID patients	Four patients, 2-3 timepoints. Average over patients.
P7	(this study)	1.67	Power *et al. (*1994, 1995); Smith *et al.* (1997); Itakura *et al.* (2005)	(mean of BEAST partition models)
NS2	(this study)	1.68	Power *et al. (*1994, 1995); Smith *et al.* (1997); Itakura *et al.* (2005)	(mean of BEAST partition models)
	Ina *et al.* (1994)	4.07	ibid.	2 timepoints, H77 -> H90. Used other DDBJ seqs to build tree. Distance calculated as difference in distances from seq -> outgroup. Labelled 'NS1' in paper.
NS3	(this study)	1.31	Power *et al. (*1994, 1995); Smith *et al.* (1997); Itakura *et al.* (2005)	(mean of BEAST partition models)
	Ina *et al.* (1994)	5.11	ibid.	2 timepoints, H77 -> H90. Used other DDBJ seqs to build tree. Distance calculated as difference in distances from seq -> outgroup.
NS4a	(this study)	1.54	Power *et al. (*1994, 1995); Smith *et al.* (1997); Itakura *et al.* (2005)	(mean of BEAST partition models)
NS4b	(this study)	1.41	Power *et al. (*1994, 1995); Smith *et al.* (1997); Itakura *et al.* (2005)	(mean of BEAST partition models)
NS5	Ina *et al.* (1994)	7.73	ibid.	2 timepoints, H77 -> H90. Used other DDBJ seqs to build tree. Distance calculated as difference in distances from seq -> outgroup. Labelled 'NS5' in paper.
NS5a	(this study)	1.46	Power *et al. (*1994, 1995); Smith *et al.* (1997); Itakura *et al.* (2005)	(mean of BEAST partition models)

NS5b	(this study)	1.26	Power *et al. (*1994, 1995); Smith *et al.* (1997); Itakura *et al.* (2005)	(mean of BEAST partition models)
	Mizokami *et al.* (2006)	0.58	Choo *et al.* (1989; 'Hutchinson strain')	Root-tip regression; 5 samples (single patient) range 1977-1996. Core, E1 & NS5b (combined)
	Smith *et al.* (1997)	0.41	Power *et al.* (1994;1995)	Average over patients (*n*=26).
Whole genome	(this study)	1.66	Power *et al. (*1994, 1995); Smith *et al.* (1997); Itakura *et al.* (2005)	(mean of BEAST partition models)
	Abe *et al.* (1992)	0.9	ibid.	*P.Trogodoytes* model, two dated sequences
	Itakura *et al* (2005)	2.759	Power *et al.* (1994;1995)	30nt Partitions; approx. genomic mean; donor-recipient 30nt Partitions; approx. genomic mean; within-recipient
	Ogata *et al*. (1991)	1.92	ibid.	Single patient, two dated sequences
	Okamoto *et al.* (1992)	1.44	ibid.	*P.Trogodoytes* model, two dated sequences

Table S2: comparison of published and analyzed within-host nucleotide substitution rates.

Accession number[a]	Subtype[b]	Length (nt)	H77 5' position[c]	H77 3' position	Sampling date	Dataset
AF313916	1b	9030	343	9373	1977.25	Anti-D
AB154177.1	1b	9030	343	9373	1996.72	Anti-D
AB154178.1	1b	9030	343	9373	1999.13	Anti-D
AB154179.1	1b	9030	343	9373	1996.52	Anti-D
AB154180.1	1b	9030	343	9373	1999.3	Anti-D
AB154181.1	1b	9030	343	9373	1996.53	Anti-D
AB154182.1	1b	9030	343	9373	1999.11	Anti-D
AB154183.1	1b	9030	343	9373	1997.29	Anti-D
AB154184.1	1b	9030	343	9373	2000.36	Anti-D
AB154185.1	1b	9030	343	9373	1996.53	Anti-D
AB154186.1	1b	9030	343	9373	2000.26	Anti-D
AB154187.1	1b	9030	343	9373	1995.25	Anti-D
AB154188.1	1b	9030	343	9373	1999.9	Anti-D
AB154189.1	1b	9030	343	9373	1994.42	Anti-D
AB154190.1	1b	9030	343	9373	2000.09	Anti-D
AB154191.1	1b	9030	343	9373	1997.13	Anti-D
AB154192.1	1b	9030	343	9373	1999.25	Anti-D
AB154193.1	1b	9030	343	9373	1996.59	Anti-D
AB154194.1	1b	9030	343	9373	1999.86	Anti-D
AB154195.1	1b	9030	343	9373	1994.4	Anti-D
AB154196.1	1b	9030	343	9373	2000.39	Anti-D
AB154197.1	1b	9030	343	9373	1997.43	Anti-D
AB154198.1	1b	9030	343	9373	1999.94	Anti-D
AB154199.1	1b	9030	343	9373	1994.42	Anti-D
AB154200.1	1b	9030	343	9373	1997.47	Anti-D
AB154201.1	1b	9030	343	9373	1994.33	Anti-D
AB154202.1	1b	9030	343	9373	1996.31	Anti-D
AB154203.1	1b	9030	343	9373	1998.2	Anti-D
AB154204.1	1b	9030	343	9373	1999.25	Anti-D
AB154205.1	1b	9030	343	9373	1994.47	Anti-D
AB154206.1	1b	9030	343	9373	1999.37	Anti-D
AB079076.1	1a	1463	376	1839	1991	Tanaka
AB079077.1	1a	1463	376	1839	1999	Tanaka
AB079078.1	1a	1463	376	1839	1978	Tanaka
AB079079.1	1a	1463	376	1839	1986	Tanaka
AB079080.1	1a	1463	376	1839	1999	Tanaka
AB079081.1	1a	1463	376	1839	1977	Tanaka
AB079082.1	1a	1463	376	1839	1990	Tanaka
AB079083.1	1a	1463	376	1839	1995	Tanaka
AB079084.1	1a	1463	376	1839	1999	Tanaka
AB079085.1	1a	1463	376	1839	1983	Tanaka
AB079086.1	1a	1463	376	1839	1989	Tanaka
AB079087.1	1a	1463	376	1839	1998	Tanaka
AB079088.1	1a	1463	376	1839	1977	Tanaka
AB079089.1	1a	1463	376	1839	1995	Tanaka
AB079090.1	1a	1463	376	1839	1981	Tanaka
AB079091.1	1a	1463	376	1839	1989	Tanaka
AB079092.1	1a	1463	376	1839	1995	Tanaka
AB079693.1	1a	339	8279	8617	1991	Tanaka
AB079694.1	1a	339	8279	8617	1999	Tanaka
AB079695.1	1a	339	8279	8617	1992	Tanaka
AB079696.1	1a	339	8279	8617	1999	Tanaka
AB079697.1	1a	339	8279	8617	1978	Tanaka

AB079698.1	1a	339	8279	8617	1983	Tanaka
AB079699.1	1a	339	8279	8617	1989	Tanaka
AB079700.1	1a	339	8279	8617	1998	Tanaka
AB079701.1	1a	339	8279	8617	1978	Tanaka
AB079702.1	1a	339	8279	8617	1982	Tanaka
AB079703.1	1a	339	8279	8617	1986	Tanaka
AB079704.1	1a	339	8279	8617	1986	Tanaka
AB079705.1	1a	339	8279	8617	1999	Tanaka
AB079706.1	1a	339	8279	8617	1981	Tanaka
AB079707.1	1a	338	8279	8617	1988	Tanaka
AB079708.1	1a	338	8279	8617	1991	Tanaka
AB079709.1	1a	338	8279	8617	1995	Tanaka
AB079710.1	1a	338	8279	8617	1996	Tanaka
AB079711.1	1a	339	8279	8617	1977	Tanaka
AB079712.1	1a	339	8279	8617	1990	Tanaka
AB079713.1	1a	339	8279	8617	1995	Tanaka
AB079714.1	1a	339	8279	8617	1999	Tanaka
AB079715.1	1a	339	8279	8617	1976	Tanaka
AB079716.1	1a	339	8279	8617	1995	Tanaka
AB079717.1	1a	339	8279	8617	1977	Tanaka
AB079718.1	1a	339	8279	8617	1981	Tanaka
AB079719.1	1a	339	8279	8617	1985	Tanaka
AB079720.1	1a	339	8279	8617	1978	Tanaka
AB079721.1	1a	339	8279	8617	1982	Tanaka
AB079722.1	1a	339	8279	8617	1987	Tanaka
AB079723.1	1a	339	8279	8617	1994	Tanaka
EU155216.2	1a	1464	375	1387	2001	Broad
EU155214.2	1a	1464	375	1387	2005	Broad
EU155215.2	1a	1464	375	1387	2004	Broad
EU155245.2	1a	1464	375	1387	2003	Broad
EU155312.2	1a	1464	375	1387	2000	Broad
EU155338.2	1a	1464	375	1387	1996	Broad
EU155213.2	1a	1464	375	1387	2004	Broad
EU155298.2	1a	1464	375	1387	2002	Broad
EU155341.2	1a	1464	375	1387	1992	Broad
EU155339.2	1a	1464	375	1387	1990	Broad
EU155340.2	1a	1464	375	1387	1991	Broad
EU155342.2	1a	1464	375	1387	1989	Broad
AY131387	1a	459	8359	8817	1990	van Asten
AY131392	1a	459	8279	8817	1992	van Asten
AY131397	1a	459	8279	8817	1989	van Asten
AY131400	1a	459	8279	8817	1990	van Asten
AY131401	1a	459	8279	8817	1989	van Asten
AY131403	1a	459	8279	8817	1989	van Asten
AY131404	1a	459	8279	8817	1989	van Asten
AY131406	1a	459	8279	8817	1990	van Asten
AY131407	1a	459	8279	8817	1991	van Asten
AY131410	1a	459	8279	8817	1994	van Asten
AY131411	1a	459	8279	8817	1996	van Asten
AY131412	1a	459	8279	8817	1997	van Asten
AY131416	1a	459	8279	8817	1988	van Asten
AY131417	1a	459	8279	8817	1989	van Asten
AY131418	1a	459	8279	8817	1991	van Asten
AY131419	1a	459	8279	8817	1991	van Asten
AY131421	1a	459	8279	8817	1997	van Asten

AY131425	1a	459	8279	8817	1997	van Asten
AY131430	1a	459	8279	8817	1990	van Asten
AY131432	1a	459	8279	8817	1996	van Asten
AY131433	1a	459	8279	8817	1993	van Asten
AY131435	1a	459	8279	8817	2001	van Asten
AY131345	1a	459	8279	8817	1994	van Asten
AY131346	1a	459	8279	8817	1994	van Asten
AY131350	1a	459	8279	8817	1997	van Asten
AY131354	1a	459	8279	8817	1995	van Asten
AY131361	1a	459	8279	8817	1994	van Asten
AY131362	1a	459	8279	8817	1993	van Asten
AY131366	1a	459	8279	8817	1997	van Asten
AY131367	1a	459	8279	8817	1997	van Asten
AY131368	1a	459	8279	8817	1998	van Asten
AY131370	1a	459	8279	8817	1997	van Asten
AY131371	1a	459	8279	8817	1997	van Asten
AY131374	1a	459	8279	8817	1997	van Asten
AY131375	1a	459	8279	8817	1996	van Asten
AY131376	1a	459	8279	8817	1997	van Asten
AY131377	1a	459	8279	8817	1997	van Asten
AY131378	1a	459	8279	8817	1997	van Asten
AY131380	1a	459	8279	8817	1997	van Asten
AY131381	1a	459	8279	8817	1998	van Asten
AY131382	1a	459	8279	8817	1997	van Asten
AY131383	1a	459	8279	8817	1998	van Asten
AY131386	1a	459	8279	8817	1997	van Asten
AF516391.1	1a	339	8279	8617	1999	Cochrane
AF516388.1	1a	339	8279	8617	1999	Cochrane
AF516385.1	1a	339	8279	8617	1999	Cochrane
AF516392.1	1a	339	8279	8617	1999	Cochrane
AF516387.1	1a	339	8279	8617	1999	Cochrane
AF516389.1	1a	339	8279	8617	1999	Cochrane
AF516390.1	1a	339	8279	8617	1999	Cochrane
AF516386.1	1a	339	8279	8617	1999	Cochrane
AY100024.1	1a	339	8279	8617	2000	Cochrane
AY100025.1	1a	339	8279	8617	2000	Cochrane
AY100026.1	1a	339	8279	8617	1999	Cochrane
AY100027.1	1a	339	8279	8617	1999	Cochrane
AY100028.1	1a	339	8279	8617	1999	Cochrane
AY100029.1	1a	339	8279	8617	1999	Cochrane
AY100030.1	1a	339	8279	8617	1999	Cochrane
AY100031.1	1a	339	8279	8617	2000	Cochrane
AF516367.1	1a	339	8279	8617	2000	Cochrane
AF516368.1	1a	339	8279	8617	1998	Cochrane
AF516369.1	1a	339	8279	8617	1999	Cochrane
AF516370.1	1a	339	8279	8617	1999	Cochrane
AF516371.1	1a	339	8279	8617	2000	Cochrane
AY100041.1	1a	339	8279	8617	2001	Cochrane
AY100085.1	1a	339	8279	8617	2000	Cochrane
AY100086.1	1a	339	8279	8617	1998	Cochrane
AY100087.1	1a	339	8279	8617	1998	Cochrane
AY100088.1	1a	339	8279	8617	1999	Cochrane
AY100091.1	1a	339	8279	8617	2001	Cochrane
AY100092.1	1a	339	8279	8617	2001	Cochrane
AY100138.1	1a	339	8279	8617	1999	Cochrane

AY100139.1	1a	339	8279	8617	2001	Cochrane
AY100140.1	1a	339	8279	8617	1998	Cochrane
AY100141.1	1a	339	8279	8617	2001	Cochrane
AY100142.1	1a	339	8279	8617	1998	Cochrane
AY100143.1	1a	339	8279	8617	1999	Cochrane
AY100144.1	1a	339	8279	8617	1999	Cochrane
AY100145.1	1a	339	8279	8617	1999	Cochrane
AY100146.1	1a	339	8279	8617	2001	Cochrane
AY100147.1	1a	339	8279	8617	2001	Cochrane
AY100148.1	1a	339	8279	8617	2001	Cochrane
AY100149.1	1a	339	8279	8617	2001	Cochrane
AY100150.1	1a	339	8279	8617	2001	Cochrane
AY100151.1	1a	339	8279	8617	2001	Cochrane
AF516393.1	1a	339	8279	8617	1999	Cochrane
AF516395.1	1a	339	8279	8617	1999	Cochrane
AF516394.1	1a	339	8279	8617	1999	Cochrane
AY100076.1	1a	339	8279	8617	1999	Cochrane
AY100077.1	1a	339	8279	8617	1998	Cochrane
AY100078.1	1a	339	8279	8617	2000	Cochrane
AY100079.1	1a	339	8279	8617	2000	Cochrane
AY100042.1	1a	339	8279	8617	1999	Cochrane
AY100051.1	1a	339	8279	8617	1999	Cochrane
AY100052.1	1a	339	8279	8617	1999	Cochrane
AY100053.1	1a	339	8279	8617	1999	Cochrane
AY100054.1	1a	339	8279	8617	1999	Cochrane
AY100055.1	1a	339	8279	8617	1999	Cochrane
AY100043.1	1a	339	8279	8617	1999	Cochrane
AY100044.1	1a	339	8279	8617	1999	Cochrane
AY100045.1	1a	339	8279	8617	1999	Cochrane
AY100046.1	1a	339	8279	8617	1999	Cochrane
AY100047.1	1a	339	8279	8617	1999	Cochrane
AY100048.1	1a	339	8279	8617	1999	Cochrane
AY100049.1	1a	339	8279	8617	1999	Cochrane
AY100050.1	1a	339	8279	8617	1999	Cochrane
AY100094.1	1a	339	8279	8617	2000	Cochrane
AY100095.1	1a	339	8279	8617	2000	Cochrane
AY100096.1	1a	339	8279	8617	2000	Cochrane
AY100097.1	1a	339	8279	8617	1999	Cochrane
AY100098.1	1a	339	8279	8617	2000	Cochrane
AY100099.1	1a	339	8279	8617	2001	Cochrane
AY100100.1	1a	339	8279	8617	2000	Cochrane

Table S3: List of sequences used in this study. [a]Genbank accession number. [b]As evaluated by REGA HCV subtype tool. [c]As determined by the Los Alamos HCV Sequence Database Sequence Locator Tool after manual alignment to other sequences *in this study, [http://hcv.lanl.gov/content/sequence/LOCATE/locate.html]. [d]Dataset references: 'Anti-D', Smith *et al* (1997); 'Tanaka', Tanaka *et al* (2002); 'Broad', unpublished / Los Alamos HCV Sequence Database, http://hcv.lanl.gov; 'van Asten', van Asten *et al*, (2004); 'Cochrane', Cochrane *et al*, (2002).

	In Anti-D alignment		In H77 reference strain	
Partition	5' start	3' end	5' start	3' end
1	1	300	342	641
2	301	600	642	941
3	601	900	942	1241
4	901	1200	1242	1541
5	1201	1500	1542	1841
6	1501	1800	1842	2141
7	1801	2100	2142	2441
8	2101	2400	2442	2741
9	2401	2700	2742	3041
10	2701	3000	3042	3341
11	3001	3300	3342	3641
12	3301	3600	3642	3941
13	3601	3900	3942	4241
14	3901	4200	4242	4541
15	4201	4500	4542	4841
16	4501	4800	4842	5141
17	4801	5100	5142	5441
18	5101	5400	5442	5741
19	5401	5700	5742	6041
20	5701	6000	6042	6341
21	6001	6300	6342	6641
22	6301	6600	6642	6941
23	6601	6900	6942	7241
24	6901	7200	7242	7541
25	7201	7500	7542	7841
26	7501	7800	7842	8141
27	7801	8100	8142	8441
28	8101	8400	8442	8741
29	8401	8700	8742	9041
30	8701	9000	9042	9341

Supplementary table S4: List of H77 reference sequence co-ordinates for the Anti-D partitions used in BEAST analysis.

Chapter Four

Testing for genotypic compartmentalisation by tissue type in HIV-1

4.1 - Abstract

I studied 41 female HIV-1-positive patients (Subtypes A1, A2, B, C, D, E, G, CRF03_AB, CRF08_BC, CRF10_CD, CRF11_cpx, CRF13_cpx & CRF14_BG), serially sampled over 1-3 years, infected through a number of routes and sampled from the peripheral blood mononuclear cells ('PBMC') or cervical ('Cervix') populations and from the *env* and *gag* genes. A minority of datasets, mainly of the *env* gene, gave strong individual support to a tissue compartment model which in *env* was associated with long-term non-progression. I conclude that cervical compartmentalisation is not the norm for HIV-1 populations but is driven by immune-mediated selection where it occurs. I suggest that compartmentalization in the cervix is a sympatric process while CNS compartment formation represents an allopatric process.

4.2 Introduction

Human immunodeficiency (HIV) viruses primarily infect $CD4^+$ T-lymphocytes (Wyatt & Sodrowski, 1998). In addition to CD4+, the virus also binds the CCR5 and / or CXCR4 co-receptors in order to effect cell entry (Berger *et al.*, 1999). The co-receptor preference of a within-host HIV population may change during infection, with most viruses initially showing preference towards CCR5 tropism (dendritic cells, macrophages & $CD4^+$ T cells) on infection (Shankarappa *et al.*, 1999). Subsequently viral populations tend towards evolution of CXCR4 tropism (activated T cells) around the time of progression to AIDS (Shankarappa *et al.*, 1999). CCR5 tropism of HIV sampled from the cervix (Kemal *et al.*, 2003) may therefore be explained by the abundance of macrophages present in the cervix mucosa (Pudney *et al.*, 2005), and hence the preferential transmission of CCR5- over CXCR4-tropic HIV.

Iversen *et al.* (2005) reported an intriguing clinical case study concerning a woman co-infected with HIV subtypes A and C'. In this patient, subtype C' virus was preferentially detected in cervical cells over Subtype A, even though superinfection with Subtype C' occurred after primary infection and both subtypes were present in peripheral blood. This striking anecdotal observation suggests that cervical compartmentalization of HIV populations may occur – that is, a distinct viral subpopulation may exist in host tissues most closely associated with onward viral transmission. If a separate cervical HIV sub-population is associated with significant genetic and phenotypic differences then the implications for HIV transmission and evolution level would be profound.

The within-host evolution of HIV is characterized by heterogeneous selection pressures that are driven by various factors, including the host immune system (Borrow *et al.*, 1997), drug

treatment (Kitchen *et al.*, 2004) and cell tropism. HIV within-host populations exhibit high rates of mutation and undergo substantial fluctuations in population size, leading to the rapid accumulation of genetic divergence and swift changes in viral diversity (Rambaut *et al.*, 2004). As a result, phylogenetic methods should be able to detect genetic differences between viral sequences sampled from different organs or tissues. Here, compartmentalization is defined as the existence of genetically distinguishable HIV subpopulations in different host tissues or cell lines. A consequence of this definition is that mean diversity among compartments should exceed that within compartments. Phylogenetically, viral sequences isolated from the same compartment should appear as monophyletic clades on a tree.

From the outset, the central nervous system (CNS) has been proposed as an ideal candidate for compartmentalization, owing to the relative impermeability to infection of the blood-brain barrier. Early results (Korber *et al.*, 1994) were promising – in fact strong *a priori* evidence linking CNS HIV infection with neurodegenerative disorders motivated such research – and subsequent studies have generally confirmed the existence of compartmentalization in the CNS (Wong *et al.*, 1997; Ohagen *et al.*, 2003; Salemi *et al.*, 2005), though with some exceptions (Caragounis, *et al.*, 2008). This finding is of distinct importance to the neuropathogenesis of HIV infection, since HIV-associated cognitive impairment is thought to arise from primary HIV infection of microglia and macrophages in the brain. Furthermore, compartmentalization more generally is of concern since it may help generate and sustain 'archive' viral subpopulations that retain ancestral sequence variation that is distinct from the viral consensus (Bello *et al.* 2004). For example, drug resistant strains could potentially be 'stored' between treatment episodes (Noë *et al.*, 2005).

The existence of separate compartmentalized viral subpopulations has also been suggested to occur in a wide range of organ, tissue and cell types, including PBMCs, cell-free plasma, bone marrow, lymph nodes, spleen, brain & cerebrospinal fluid, lung, kidney, liver, rectum, gastrointestinal tract, testes, semen and cervicovaginal tract (Sanjuán *et al.*, 2004; Alves *et al.*, 2002; Zhang *et al.* 2002; Van t'Wout *et al.*, 1998; Wong *et al.*, 1997; although see also van der Hoek *et al.*, 1998). Potter *et al.* (2004) have also reported compartmentalization between leukocytes. Simian models that have been experimentally infected with HIV also show genetic heterogeneity among a wide set of organs, tissues and cells (Magierowska *et al.*, 2004).

More recently, attention has focused on compartmentalization in seminal fluid, the primary route of transmission through heterosexual contact. Furthermore, HIV is present in cervicovaginal lavage (Adal *et al*, 2005; Kovacs *et al.*, 2001; Iversen *et al*, 1998), even during effective highly active antiretroviral therapy (HAART; Nunnari *et al.*, 2005). Genetic heterogeneity between the viral population in genital and non-genital compartments could therefore have important implications for worldwide HIV epidemiology, since different selective forces could operate in each compartment. To date, the evidence has suggested compartmentalization in semen (Pillai *et al.*, 2005; Luizzi *et al*., 2004), though there appears to be a complex relationship between HIV diversity and titre in seminal fluid (Gupta *et al.*, 2000).

Although HIV infection of cervical tissues has been extensively studied (*cf.* Sullivan *et al.,* 2005; Tirado *et al.*, 2004), previous studies are split as to whether compartmentalization actually occurs. Genetically distinct HIV populations have been observed to occur in specific tissues *post-mortem* (Wong *et al.*, 1997), while *in vitro* studies of separately cultured HIV

isolates show organ- and cell-line specific tropisms (Fear *et al.*, 1998). However, while numerous studies appear to show genetic compartmentalization (Iversen *et al*., 2005; Philpott *et al.*, 2005), others disagree (Poss *et al.*, 1998) and some studies show both trends across different patients in the same study (Kemal *et al.*, 2003).

Furthermore, the vast majority of studies undertaken to date have suffered from low sample sizes and statistically weak methods of analysis (Philpott *et al.*, 2005; Poss *et al.*, 1998). For example, methods such as AMOVA (Sullivan *et al.*, 2005) treat genetic polymorphisms as statistically independent observations and therefore ignore phylogenetic shared ancestry, (Harvey & Pagel, 1991). To avoid this, later studies have included phylogenetic analyses in which each taxon is labelled with its compartment of origin, but typically only qualitative and comparative conclusions have been drawn (Potter *et al.*, 2004; Daniels *et al.*, 2004; van der Hoek *et al.*, 1998). A better approach is to use ancestral state reconstruction: changes in phenotypic traits on a tree are statistically independent, even though the states at the tips are not. For example, in the Slatkin–Maddison test (Slatkin & Maddison, 1989), a null distribution of trait-phylogeny associations is generated through randomization so that a significance value for the observed number of migrations may be calculated. Observed migrations (changes in the 'state', or location, of the geographical character) are obtained through state reconstruction by Fitch parsimony (Fitch 1971b; implemented in Simmonds *et al.*, 2005; Fulcher *et al*., 2004). An alternative statistic, called the association index (AI), was introduced by Wang *et al.* (2001); significance can be calculated for observed AI values in the same way as the Slatkin–Maddison test.

However, previous phylogenetic analyses of compartmentalisation have utilised a single-tree approach – a maximum-likelihood (ML) or maximum–parsimony (MP) algorithm is used to

estimate a "best" phylogenetic tree that is assumed to be correct and used in subsequent ancestral state reconstructions. This approach is attractive since it is computationally quick to undertake (Holder & Lewis, 2003). In practice, tree estimation is subject to phylogenetic error, hence single-tree methods could result in erroneous assessments of phylogeny-trait association. Fortunately, Bayesian Markov-chain Monte Carlo (MCMC) methods have increased in popularity and attainability in recent years, largely thanks to improvements in computing resources (Holder & Lewis, 2003). In these methods a posterior distribution of trees is obtained, which represents the statistical uncertainty inherent in phylogeny estimation, such that more probable topologies are sampled more often, and unlikely ones less so. This approach is particularly well suited to the compartmentalization problem, for which the tree topology must be accurately estimated. In Chapter Two I developed these approaches, and that methodology is applied here.

In this study I investigate evidence for compartmentalization in a large, serially sampled dataset of *gag* and *env* gene sequence data, sampled from a cohort of 41 HIV infected women. HIV was obtained from two compartments: blood plasma and cervical tissues. Information on method of transmission, duration of infection, $CD4^+$ count, human leukocyte antigen (HLA) type, viral load and viral subtype were also collected where available. Because of the very high rate of recombination in HIV (Rhodes *et al*, 2003; Charpentier *et al*, 2006) the two genes were considered separately at all times. I use Bayesian MCMC methods to test the hypothesis that the cervix harbours genetically distinct viral strains.

4.3 Data collection

gag & *env* sequences were collected from 41 infected women over one, two or three timepoints. Viruses were sampled from either peripheral blood mononuclear cells (PBMC) or from cervical tissue swabs. The patients were designated using two-letter codes (*e.g.* 'AA', 'BE' etc). Full details of the sampling procedure and clinical information can be found in the Appendix to this chapter.

Patient alignments for each gene were produced using CLUSTAL and edited by hand using Se-Al. Sequences were subtyped using the REGA HIV Subtyping Tool (de Oliveira *et al*, 2005) and sequences that could not be unambiguously assigned to a subtype or recognised circulating recombinant form (CRF) excluded. Finally, datasets with fewer than two sequences sampled from the cervix were removed such that the final datasets as tested comprised 34 (*env*) and 31 (*gag*) patients. I calculated the Shannon entropy as a measure of the diversity in each alignment using a custom-made tool (see Appendix 2 of this thesis).

4.4 Methods

4.4.1 Phylogenetic analysis

Individual Bayesian MCMC phylogenetic analyses were first performed in order to obtain posterior sets of trees (PST) for each patient. The PSTs are an estimate of the posterior distribution of phylogenies supported by each patient alignment, and were used as the basis of subsequent phylogeny-trait association analyses.

Bayesian MCMC analysis using BEAST v1.4 (Drummond & Rambaut, 2007; 2003) was carried out with constant-size and exponential growth coalescent models, under the Hasegawa-Kishino-Yano 85 substitution model with gamma-distributed rate heterogeneity (HKY+Γ). I allowed the codon positions' relative rates to vary and the UCLN relaxed clock and strick clock models were tested.

Only three patients (AB, AF & AV) had sufficient temporal structure to attempt simultaneous estimation of the tree, substitution rate & coalescent parameters. These were run for 10,000,000 states with the strict molecular clock enforced (the UCLN relaxed clock models were found to lead to poor MCMC chain mixing). Of these, only AB reliably returned rates in general agreement with the literature (e.g. Lemey *et al* 2005) for both genes, due to the smaller numbers of sequences (and hence weaker phylogenetic and temporal signals) in the AF & AV data sets. However, this study is primarily concerned with tree topology, not substitution rates or coalescent inference. As a result, all patient datasets were analyzed using the same substitution and coalescent models as used for patient AB. The posterior distribution of substitution rates obtained for patient AB was used to provide a timescale the other data sets. Specifically, the posterior distribution of rates for each codon position was fitted to a normal distribution[1] and then the means and standard deviations of these distributions were used to define a normally distributed prior on the codon position rates for the other data sets. A strict molecular clock model was used throughout. All MCMC runs

[1] In substitutions per site per year:
env gene data set:
mean substitution rate at first codon position (μ_1) = 0.0152 (standard deviation (σ_1) = 0.004); mean substitution rate at second codon position (μ_2) = 0.0142 (standard deviation (σ_2) = 0.004); mean substitution rate at third codon position (μ_3) = 0.0215 (standard deviation (σ_3) = 0.005):
gag gene data set:
μ_1 = 0.0051 (σ_1 = 0.0015); μ_2 = 0.0035(σ_2 = 0.001); μ_3 = 0.0151 (σ_3 = 0.003).

were inspected using Tracer version 1.4 (Rambaut & Drummond, 2007) to check convergence and an appropriate proportion of burn-in states discarded from further analysis. A selection of data sets were re-run several times to check that the MCMC had not converged on local optima. I also used the 'taxon set' option in BEAST to specify separate taxon sets for the PBMC and cervix sequences. This does not alter the behaviour of the coalescent to tree models, but allowed me to estimate separate most recent common ancestor dates (tMRCA) for the ancestor of each taxon set ('compartment sub-tree'), as well as the number of migrations between compartments. This statistic is calculated the same way as the parsimony score statistic implemented in Chapter Two.

4.4.2 Single tree analyses

For comparison with the Baysian MCMC analyses, phylogeny-trait analyses were also performed on single trees, as this methodology has commonly been employed in similar studies (Potter *et al.*, 2004; Daniels *et al.*, 2004; van der Hoek *et al.*, 1998). The maximum-likelihood tree for each data set was obtained using a computer script to identify the tree with the highest likelihood (tree likelihoods are written to the BEAST output) in each PST. In these single trees each taxon was labelled with its tissue of origin (PBMC or 'Cervix') and visually inspected for monophyly of either compartment.

4.4.3 Compartment sub-tree tMRCA categorization

A compartmentalized subpopulation of sequences ought to share a younger tMRCA than the tMRCA of the whole viral population, even if interspersed by sequences from other compartments. This distinction suggests a test for the directionality of lineage migration

events among compartments. I exploited the compartment sub-tree tMRCA information from the BEAST output, which logs these values separately for each tree in the posterior set of trees (PST). For every tree in the PST, I obtained three ages from the BEAST data: dates of MRCA of cervix sequences (T_c), date of MRCA of PBMC sequences (T_p), date of MRCA of all sequences (T_a). Each data set was then categorized using the following conditions:

Category	MRCA values	Possible interpretation
A	$T_a > T_c$ and $T_a > T_p$	Total compartmentalization
B	$T_a > T_c$ and $T_a = T_p$	Cervical compartmentalization
C	$T_a = T_c$ and $T_a > T_p$	PBMC compartmentalization
D	$T_a = T_c = T_p$	No compartmentalization

4.4.4 Phylogeny-trait correlations

To estimate the strength of phylogeny-trait association I used the methodology developed in Chapter Two, as implemented in the BaTS package. Three tree topology statistics were used: the association index (AI), the parsimony score (PS) and the maximum exclusive (monophyletic) clade size (MC). Briefly, given a PST obtained for each data set using BEAST and given the set of sequence-trait mappings, BaTS calculates the values of the association statistics for each credible topology, thereby estimating the posterior distribution of each association statistic. The significance of each statistic is calculated by comparison with a null expected posterior distribution, obtained by randomizing the sequence-trait mappings a large number of times (1000 replicates in this study) for each tree in the PST.

4.4.5 Multiple test correction

Phylogeny-trait association tests were conducted on 31 (*env*) and 21 (*gag*) data sets. Since this experimental design involves multiple tests, it is subject to increased Type I error, or a higher false discovery rate (FDR). To control for false positives, I employed an FDR correction procedure as described in Benjamini & Hochberg (1995). The set of *p*-values $P = \{p_0, p_1, \ldots p_n\}$ is ordered by significance. Starting with the most significant trial, each p_i-*th* result was accepted if it met the criterion:

$$p_i \leq \frac{m_0}{m} q^* \leq q^* \qquad \text{(Eqn 4.1)}$$

where m_0 / m = the fraction of null hypotheses (including the m_i of significance p_i) that would be rejected and q^* is the FDR (the unknown realisation of the random proportion of positives that are false). For this study I sought to control the FDR to 0.05 and have calculated the number of those positives that could be accepted by this procedure using equation 4.1.

4.4.6 Multiple-patient analyses

As an alternative to individual tests on each data set (which necessitates multiple test correction), I performed a single analysis of phylogeny-correlation on two master alignments. This approach should also yield more power to detect compartmentalization (deviation from the null hypothesis of random trait-phylogeny association) over the whole dataset. The first master alignment contains *env* sequences from all patients (n=661 sequences in total) and the second contains *gag* sequences from all patients (n = 478 sequences). I generated PSTs in

BEAST v1.4.6 under constant population-size coalescent models. As with individual patients, this analysis is dependent on accurate sampling of posterior tree topologies more than accurate substitution and clock model estimation; a model and parametisation typical of HIV studies in the literature was chosen. I used the Hasegawa-Kishino-Yano 85 substitution model with gamma-distributed rate heterogeneity and invariant-sties (HKY+Γ+I). I allowed the 1^{st}+2^{nd} and 3^{rd} codon positions' relative rates to vary and employed the uncorrelated lognormal relaxed-clock model with mean rates fixed at $4x10^{-3}$ substitutions / site / year (*env*) and $2x10^{-3}$ substitutions / site / year (*gag*.) Following inspection of the output logfiles an appropriate burn-in proportion was discarded and the output tree files uniformly downsampled to 68 trees (*env*) and 951 trees (*gag*) due to computational constraints. The resulting PSTs were tested for phylogeny-trait correlation using the AI, PS & MC statistics implemented in BaTS (as described above, using 1000 replicates to form the null distributions). I also inspected the maximum-likelihood trees of the two master alignments to verify that sequences from individual patients were monophyletic.

4.5 Results

Samples were collected and sequenced from a total of 41 patients. Of these, sufficient samples were collected both from PBMC and cervical compartments to analyse 34 *env* data sets and 31 *gag* data sets.

Results are given in Tables 4.1 & 4.3 (*env*) and Tables 4.2 & 4.4 (*gag*). Table 4.1 gives the individual patients' *env* data sets' observed trait-phylogeny association scores under the AI, PS, MC_p & MC_c statistics as well as their significance as calculated by the BaTS program using the PST from the BEAST analyses. The 'cervix' and 'PBMC' taxon set migration counts calculated in BEAST are identical to the observed 'PS' score. Table 4.3 gives the proportion of trees in each patient's *env* gene posterior set that fell under each compartment sub-tree tMRCA category. Table 4.3 also records the total sitewise Shannon entropy (a measure of diversity) for each patient alignment and which patients' ML *env* trees showed total compartmentalization.

Similarly, Table 4.2 gives the individual patients' *gag* data sets' observed trait-phylogeny association scores under the AI, PS, MC_p & MC_c statistics as well as their significance as calculated by the BaTS program. Table 4.4 gives the proportion of trees in each patient's *gag* gene posterior set that fell under each compartment sub-tree tMRCA category. Table 4.4 also records the total sitewise Shannon entropy for each patient alignment and which patients' ML *gag* trees showed total compartmentalization.

Phylogeny-trait association by BaTS (individual data sets): The total number of patients that showed significant phylogeny-trait association by each statistic in the BaTS analysis is

given at the bottom of Table 4.1 (*env)* & Table 4.2 (*gag*). Four patients' results showed significant association across all statistics in the *env* data set (AD, AU, BM & BO); only Patient AH showed significant association by all statistics in the *gag* data set.

Following FDR correction Patients AD, AG, BM & BO still showed significant compartmentalization by the AI statistic (*env*; Table 4.1), as did Patients AH, AN & AW (*gag*; Table 4.2). None of the MC analyses were significant in either the *env* or *gag* datasets; however I demonstrated by simulation in Chapter Two that the MC statistic is not expected to be statistically powerful in small data sets. Most of the patient data sets in this study may suffer from this problem since on average only 16 sequences were sampled.

In the *env* dataset, in addition to Patients AD, AU, BM & BO already mentioned, Patients AR & AG also showed significant association as measured by both the AI & PS statistics and at least one of the MC measures. The ML trees for these patients are shown in Figure 4.1.

In the *gag* data set, only Patient AH showed significant association between compartment and phylogeny by all statistics (though this data set comprised only 8 samples). Data sets taken from Patients AW (n = 12) & AZ (n = 13) also showed significant association by the AI or PS statistic and at least one MC statistic. These three data sets' ML trees are shown in Figure 4.2.

Patient[1] (*n*)		Time[2]	Phylogeny – trait association analyses[3]							
			AI	*P* –value (AI)	PS	*P* –value (PS)	MC_P	*P* –value (MC_P)	MC_c	*P* –value (MC_c)
AA	*28*	NA	1.20	0.160	12.29	0.210	2.67	0.180	3.25	0.430
AB	*47*	4	1.61	0.050	18.15	0.090	2.73	0.130	5.40	0.310
AD	*19*	10	0.01	**p<0.005**	2.75	**p<0.005**	7.00	**0.040**	4.00	**0.040**
AF	*24*	6	1.28	0.490	10.78	0.340	2.24	0.890	2.00	0.540
AG	*18*	NA	0.01	**p<0.005**	1.53	**p<0.005**	3.87	0.200	4.97	**0.010**
AH	*11*	4	0.54	0.460	4.73	0.460	3.99	0.060	1.02	1.000
AI	*23*	12	0.75	0.050	10.02	0.280	3.41	0.590	2.01	0.500
AJ	*14*	12	0.36	0.060	4.07	**0.030**	2.19	0.730	3.00	0.200
AK	*27*	NA	0.90	0.130	9.28	0.200	2.02	0.290	5.17	0.360
AL	*34*	9	1.36	0.140	16.00	0.390	2.61	0.530	2.90	0.140
AM	*34*	11	1.31	0.230	12.90	0.290	2.03	0.500	4.00	0.520
AO	*11*	4	0.41	0.360	4.44	0.080	1.68	0.220	2.10	0.710
AP	*10*	13	0.35	0.240	3.74	0.140	2.71	0.140	3.01	0.120
AQ	*11*	NA	0.35	0.140	2.52	**p<0.005**	2.40	0.350	4.93	**0.010**
AR	*12*	3	0.19	**0.020**	2.97	**0.010**	4.66	**0.040**	2.02	0.300
AS	*24*	6	0.85	0.220	8.96	0.090	2.55	0.290	6.00	**0.020**
AT	*9*	13	0.17	0.230	4.38	0.560	2.30	0.730	1.04	1.000
AU	*28*	NA	0.38	**0.010**	6.38	**p<0.005**	5.55	**0.030**	7.56	**0.010**
AV	*31*	7	1.26	0.130	13.63	0.310	2.12	0.970	3.01	0.310
AW	*10*	5	0.32	0.270	2.96	0.900	3.32	0.770	1.00	1.000
AX	*11*	12	0.51	0.550	5.06	0.240	1.32	1.000	2.00	0.330
AY	*17*	8	0.51	**0.030**	5.36	**0.020**	2.10	0.100	6.39	0.080
AZ	*13*	8	0.00	0.070	1.00	0.070	12.00	0.080	1.00	1.000
BA	*10*	NA	0.13	**0.010**	2.83	0.130	3.18	0.250	2.00	0.100
BB	*10*	6	0.26	**0.040**	2.86	0.090	3.01	0.040	4.00	**0.010**
BE	*9*	7	0.33	0.160	3.60	0.410	1.01	1.000	4.09	0.120
BF	*7*	5	0.11	0.070	1.93	0.160	2.04	0.790	1.75	**0.010**
BH	*19*	NA	0.84	0.510	6.46	0.300	5.71	0.270	1.15	1.000
BI	*6*	5	0.30	0.660	1.48	**p<0.005**	1.00	1.000	1.95	0.670
BJ	*12*	12	0.15	**0.030**	3.95	0.070	2.00	0.250	4.60	0.070
BK	*9*	5	0.41	0.270	3.96	0.300	2.28	0.660	2.01	0.320
BL	*13*	9	0.51	0.110	5.58	0.170	1.83	0.530	2.34	0.270
BM	*20*	NA	0.36	**p<0.005**	3.80	**p<0.005**	8.00	**0.010**	3.25	**0.040**
BN	*10*	7	0.06	**0.010**	2.80	0.050	2.12	0.210	3.03	0.320
BO	*20*	10	0.03	**p<0.005**	1.84	**p<0.005**	3.74	**0.010**	13.66	**0.010**
Total positives:				*11*		*10*		*5*		*9*
Total positives remaining[4] following FDR correction:				**4**		**8**		**0**		**0**
Significance in combined data set:				0.00		0.00		0.09		

Table 4.1: Combined results for the *env* data set. [1]Patient codes and number of samples. [2]Time: duration of infection (in years) where known. [3]As calculated by the BaTS program – see 'Methods'. AI, association index; PS, parsimony score (migrations); MC_{PBMC}, largest monophyletic PBMC clade; MC_{Cervix}, largest monophyletic cervix clade. [4]The total number of positives detected by each statistic in the BaTS analysis was corrected for false positives using FDR correction as described in the text.

			Phylogeny – trait association analyses[3]							
Patient[1] (*n*)		**Time[2]**	**AI**	***P* –value (AI)**	**PS**	***P* –value (PS)**	MC_P	***P* –value (MC_P)**	MC_c	***P* –value (MC_c)**
AA	28	NA	0.96	0.100	8.91	**0.020**	2.12	0.630	4.99	0.190
AB	43	4	1.94	0.110	18.55	0.180	3.54	0.230	3.58	0.180
AD	9	10	0.00	0.100	1.00	0.100	8.00	0.110	1.00	1.000
AF	20	6	0.98	0.710	7.26	0.490	3.06	0.750	1.11	1.000
AH	8	4	0.01	**p<0.005**	1.26	**p<0.005**	5.00	**0.010**	2.52	**0.010**
AI	28	12	1.33	0.290	11.64	0.440	1.97	0.400	4.43	0.240
AJ	10	12	0.24	0.830	1.49	**p<0.005**	5.67	**0.010**	1.00	1.000
AK	14	NA	0.83	0.700	5.00	0.210	3.08	0.670	1.68	0.060
AL	25	9	1.68	0.830	12.88	0.770	1.98	0.690	2.03	1.000
AM	23	11	1.55	0.820	11.48	0.530	1.91	0.740	2.41	1.000
AN	5	5	0.06	**p<0.005**	1.19	**p<0.005**	3.62	**0.010**	1.00	1.000
AP	10	13	0.64	0.430	4.17	0.120	3.00	0.110	1.01	1.000
AQ	5	NA	0.23	0.580	2.36	0.120	2.13	0.290	1.00	1.000
AR	7	3	0.24	0.190	2.63	0.280	2.00	0.420	2.18	0.110
AS	19	6	1.43	0.930	9.22	0.700	1.64	0.340	2.05	0.990
AT	9	13	0.43	0.230	4.00	0.260	1.87	0.210	2.13	0.510
AU	11	NA	0.31	0.640	1.49	**p<0.005**	7.32	0.390	1.00	1.000
AV	15	7	0.36	0.190	4.12	0.200	6.66	0.050	1.12	1.000
AW	12	5	0.11	**p<0.005**	2.77	0.080	4.81	**0.040**	2.02	0.070
AY	14	8	0.79	0.630	6.08	0.350	1.03	1.000	2.71	0.340
AZ	13	8	0.28	0.070	3.01	**0.020**	4.40	**0.010**	2.97	**0.010**
BA	10	NA	0.20	0.820	1.51	1.000	4.12	0.510	1.00	1.000
BD	5	1	0.27	0.290	1.46	**p<0.005**	2.74	**0.010**	1.00	1.000
BF	8	5	0.43	0.320	3.83	0.720	1.55	1.000	2.09	0.310
BH	16	NA	0.83	0.560	6.86	0.810	1.19	1.000	3.39	0.170
BJ	8	12	0.17	0.120	1.67	0.080	3.59	0.160	1.93	0.060
BK	8	5	0.60	1.000	3.07	0.620	1.45	1.000	1.00	1.000
BL	10	9	0.74	0.530	4.73	0.640	1.62	0.270	2.00	0.840
BM	8	NA	0.29	0.330	2.83	0.420	3.00	0.690	1.32	1.000
BO	8	10	0.54	1.000	3.00	0.690	1.00	1.000	1.79	0.810
Total positives:				*3*		*7*		*6*		*2*
Total positives remaining[4] following FDR correction:				**3**		**5**		**0**		**0**
Significance in combined data set:				**0.00**		**0.00**		**0.04**		0.84

Table 4.2: Combined results for the *gag* data set. Significance values < 0.0005 shown rounded to zero. [1]Patient codes and number of samples. [2]Time: duration of infection (in years) where known. [3]As calculated by the BaTS program – see 'Methods'. AI, association index; PS, parsimony score (migrations); MC_{PBMC}, largest monophyletic PBMC clade; MC_{Cervix}, largest monophyletic cervix clade. [4]The total number of positives detected by each statistic in the BaTS analysis was corrected for false positives using FDR correction as described in the text.

Distribution of p values among patients: I used a binning procedure (n = 100, uniformly distributed between 0 and 1) to calculate cumulative distribution functions of the p-values of each statistic among patients. This was done separately for the complete *env* and *gag* data sets. If the null hypothesis (no population structure) is correct, the p values should follow a unit uniform distribution. The expected distribution and observed cumulative density functions are shown in Figure 4.1 (for *env*) and Figure 4.2 (for *gag*). For the *env* datasets the cumulative density functions of all statistics were significantly different from the uniform distribution (Kolmogorov-Smirnov test, n=100; $p < 0.0001$), with an excess of low p-values, particularly p<0.25. In the *gag* dataset only the cumulative density functions of the PS ($p < 0.001$) and MC_{PBMC} ($p < 0.05$) statistics were significantly greater than the uniform distribution.

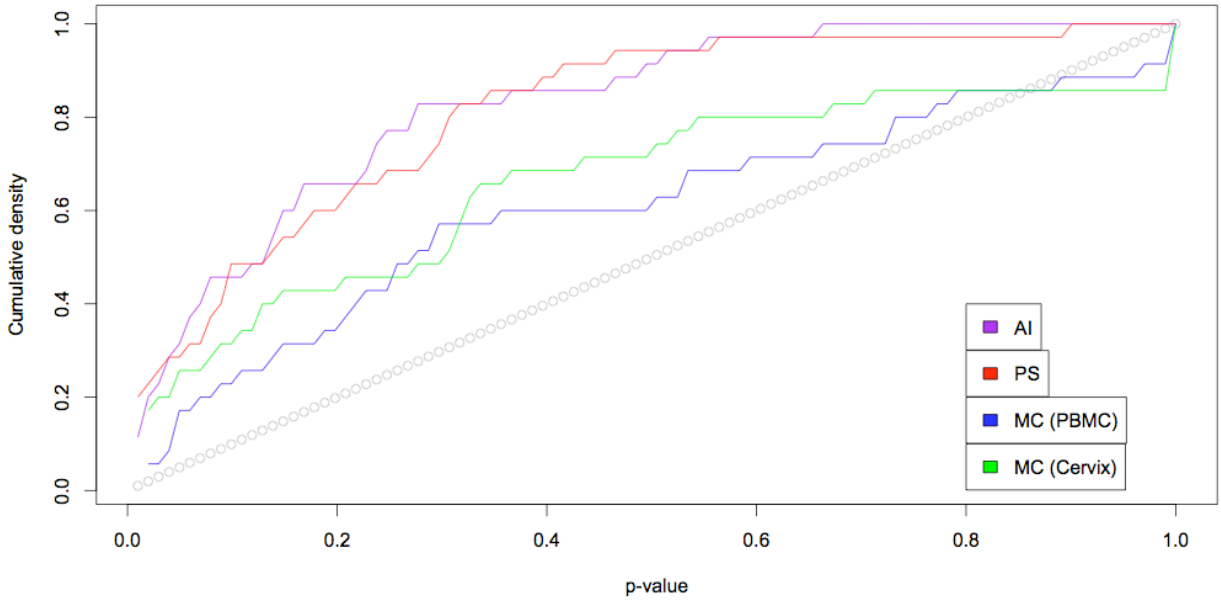


Figure 4.1: Cumulative density functions of BaTS statistics' *p*-values observed in *env* gene data sets. The expected null (uniform) distribution of *p*-values is shown (grey circles). Solid lines represent the calculated cumulative density functions for: AI (purple); PS (red); MC_{PBMC} (blue); MC_{Cervic} (green).

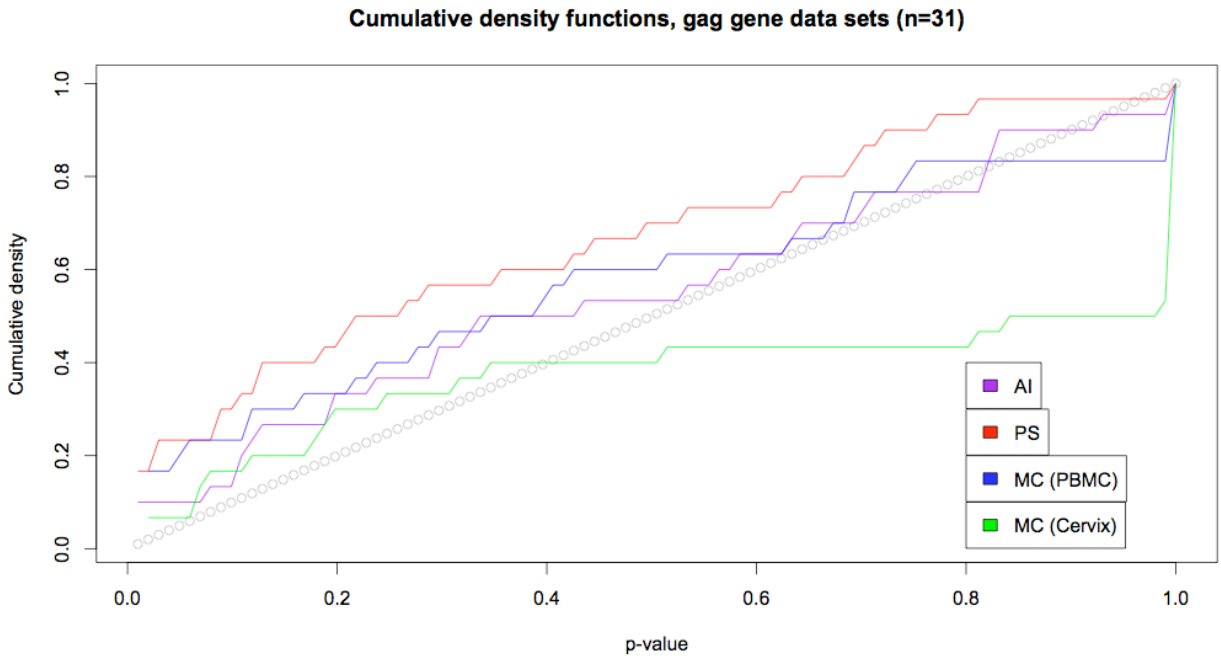


Figure 4.2: Cumulative density functions of BaTS statistics' *p*-values observed in *gag* gene data sets. The expected null (uniform) distribution of *p*-values is shown (grey circles). Solid lines represent the calculated cumulative density functions for: AI (purple); PS (red); MC_{PBMC} (blue); MC_{Cervic} (green).

Phylogeny-trait association by BaTS (master data sets): Both *env* and *gag* master data sets showed significant ($p < 0.005$) compartmentalization by AI and PS statistics between PBMC and cervical samples when analysed in BaTS. Monophyletic PBMC clades tended to occur significantly often in the *gag* data set ($p < 0.04$). The results for *env* and *gag* are given at the bottom of Tables 4.1 & 4.2 respectively.

Compartment sub-tree tMRCA categorization on individual patient data sets: The proportion of trees in PST that fell into each sub-tree tMRCA category are given in Table 4.3 (*env*) and Table 4.4 (*gag*). Of the six *env* and three *gag* data sets identified above as showing evidence for compartmentalization, half (*env* – AG, AR, BM; *gag* – AW) exhibited Category 'B' tMRCA relationships that were considered to be indicative of cervical compartmentalization against a background PBMC diversity, while two suggested compartmentalization with no suggestion of directionality (Category 'A' sub-tree tMRCA relationship; *env* – AD; *gag* – AZ). Patient BO (*env*) reversed this trend, showing a Category 'C' sub-tree tMRCA relationship, indicative of PBMC compartmentalization against a background population of cervically-dervived samples. This patient had been infected for a considerable period of time (at least 10 years.) Although Patient AH showed strong compartmentalization in the BaTS analysis, this was not reflected in the compartment sub-tree tMRCA categorization**.** From the topology of the ML tree (Figure 4.1a) it seems possible that this is because the tree is approximately symmetrical.

The Shannon alignment entropy scores for the sequences sampled from each patient are also shown in Table 4.3 (*env*) & Table 4.4 (*gag*). No clear relationship between either duration of infection or dominant compartment sub-tree tMRCA categorization is apparent.

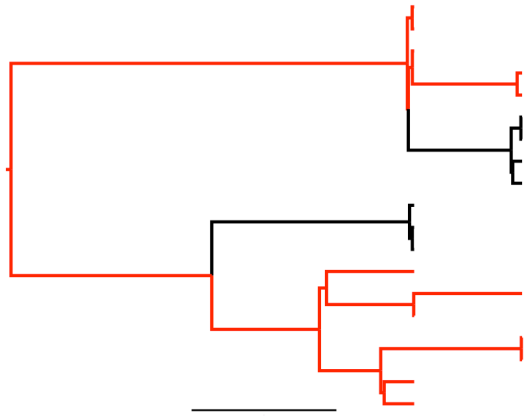
Patient[1] (*n*)		Time[2]	ML tree[3]	Shannon entropy	Compartment sub-tree categorization[3]			
					A	B	C	D
AA	*28*	NA		15.73	0.00	0.00	1.00	0.00
AB	*47*	4		30.81	0.46	0.00	0.54	0.00
AD	*19*	10		37.14	1.00	0.00	0.00	0.00
AF	*24*	6		24.49	0.00	1.00	0.00	0.00
AG	*18*	NA	*	12.81	0.00	1.00	0.00	0.00
AH	*11*	4		7.05	0.78	0.14	0.07	0.00
AI	*23*	12		11.42	0.29	0.65	0.06	0.00
AJ	*14*	12		8.58	0.95	0.03	0.02	0.00
AK	*27*	NA		19.52	0.97	0.00	0.03	0.00
AL	*34*	9		11.11	0.00	0.00	1.00	0.00
AM	*34*	11		22.42	0.13	0.03	0.84	0.00
AO	*11*	4		30.15	0.00	0.00	1.00	0.00
AP	*10*	13		20.18	0.00	0.99	0.01	0.00
AQ	*11*	NA		11.32	0.03	0.04	0.92	0.00
AR	*12*	3		11.12	0.07	0.91	0.02	0.00
AS	*24*	6		22.06	0.00	0.00	1.00	0.00
AT	*9*	13		15.34	0.01	0.98	0.02	0.00
AU	*28*	NA		48.45	1.00	0.00	0.00	0.00
AV	*31*	7		51.78	0.00	0.00	1.00	0.00
AW	*10*	5		6.25	0.58	0.38	0.03	0.00
AX	*11*	12		36.24	0.02	0.00	0.98	0.00
AY	*17*	8		23.15	0.09	0.65	0.26	0.00
AZ	*13*	8		22.62	0.00	0.00	0.00	1.00
BA	*10*	NA		13.37	0.60	0.25	0.14	0.00
BB	*10*	6		6.01	0.08	0.57	0.35	0.00
BE	*9*	7		6.18	0.32	0.47	0.20	0.00
BF	*7*	5	*	9.27	0.00	0.99	0.00	0.00
BH	*19*	NA		17.26	0.50	0.00	0.50	0.00
BI	*6*	5		12.68	0.00	1.00	0.00	0.00
BJ	*12*	12		38.53	0.20	0.76	0.04	0.00
BK	*9*	5		10.41	0.57	0.20	0.23	0.00
BL	*13*	9		12.23	0.21	0.73	0.07	0.00
BM	*20*	NA		46.71	0.00	1.00	0.00	0.00
BN	*10*	7		6.58	0.42	0.38	0.20	0.00
BO	*20*	10	*	5.83	0.05	0.02	0.87	0.06
Mean observed frequency in posterior set of trees (PST):					*0.27*	*0.38*	*0.33*	*0.03*

Table 4.3: Combined results for the *env* data set. [1]Patient codes and number of samples. [2]Time: duration of infection (in years) where known. [3]ML trees where either compartment appeared monophyletic are marked with an asterisk. [4]Proportion of trees in the posterior set corresponding to that sub-tree root height categorization as defined in 'Methods.'

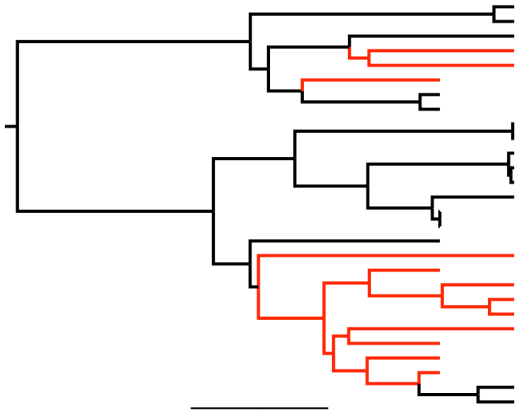
Patient[1] (*n*)		Time[2]	ML tree[3]	Shannon entropy	Compartment sub-tree categorization[4]			
					A	**B**	**C**	**D**
AA	28	NA		12.30	0.32	0.00	0.67	0.00
AB	43	4		10.85	0.93	0.05	0.01	0.00
AD	9	10		18.78	0.00	0.00	0.00	1.00
AF	20	6		9.35	0.73	0.26	0.01	0.00
AH	8	4	*	24.86	0.00	0.13	0.12	0.75
AI	28	12		17.09	0.69	0.20	0.10	0.00
AJ	10	12		7.90	0.00	0.98	0.00	0.02
AK	14	NA		6.35	0.72	0.28	0.00	0.00
AL	25	9		11.10	0.95	0.02	0.03	0.00
AM	23	11		6.07	0.95	0.02	0.03	0.00
AN	5	5		3.50	0.00	0.83	0.00	0.17
AP	10	13		11.18	0.44	0.50	0.06	0.00
AQ	5	NA		6.75	0.08	0.26	0.63	0.03
AR	7	3		4.98	0.15	0.63	0.20	0.02
AS	19	6		9.10	0.94	0.02	0.04	0.00
AT	9	13		10.27	0.85	0.04	0.11	0.00
AU	11	NA		9.20	0.00	1.00	0.00	0.00
AV	15	7		9.47	0.37	0.52	0.11	0.00
AW	12	5		3.86	0.12	0.67	0.20	0.00
AY	14	8		8.12	0.71	0.26	0.03	0.00
AZ	13	8		6.46	0.57	0.40	0.03	0.00
BA	10	NA		4.80	0.00	0.99	0.00	0.01
BD	5	1		1.33	0.00	0.98	0.00	0.02
BF	8	5		2.41	0.00	0.98	0.00	0.02
BH	16	NA		16.00	0.00	0.95	0.00	0.05
BJ	8	12	*	17.10	0.00	1.00	0.00	0.00
BK	8	5		4.30	0.86	0.13	0.01	0.00
BL	10	9		8.31	0.64	0.12	0.24	0.00
BM	8	NA	*	21.61	0.00	1.00	0.00	0.00
BO	8	10		63.34	0.42	0.02	0.55	0.00
Mean observed frequency in posterior set of trees (PST):					*0.381*	*0.442*	*0.106*	*0.070*

Table 4.4: Combined results for the *gag* data set. Significance values < 0.0005 shown rounded to zero. [1]Patient codes and number of samples. [2]Time: duration of infection (in years) where known. [3]ML trees where either compartment appeared monophyletic are marked with an asterisk. [4]Proportion of trees in the posterior set corresponding to that sub-tree root height categorization as defined in 'Methods.'

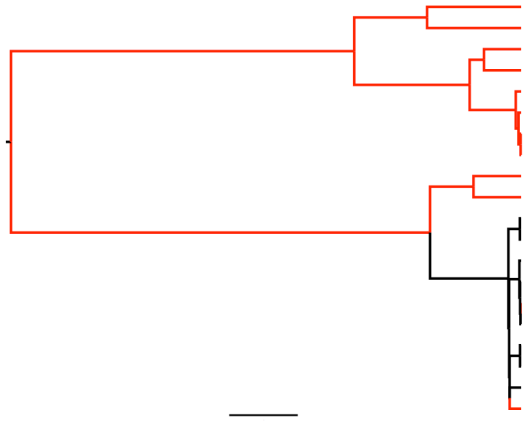
.



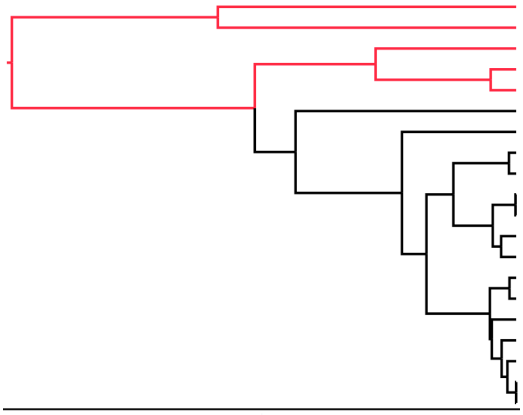
a) AD ($AI = 0.01$, $p < 0.005$)



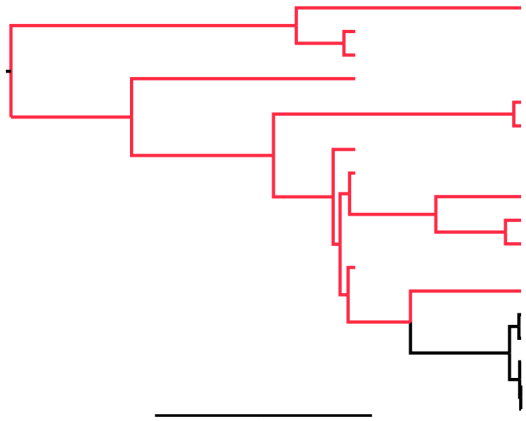
b) AU ($AI = 0.38$, $p < 0.01$)



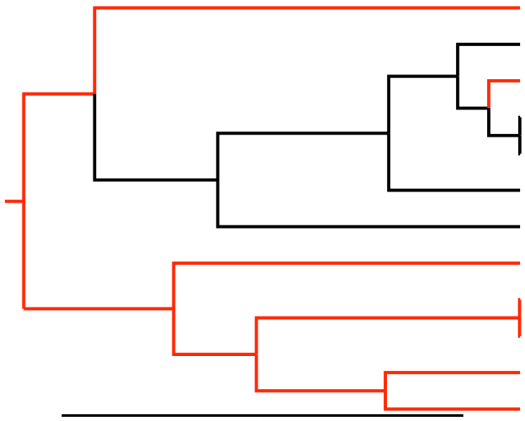
c) BM ($AI = 0.36$, $p < 0.005$)



d) BO ($AI = 0.03$, $p < 0.005$)



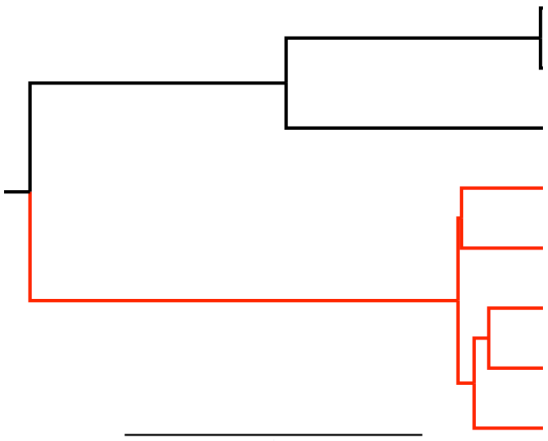
e) AG ($AI = 0.01$, $p < 0.005$)



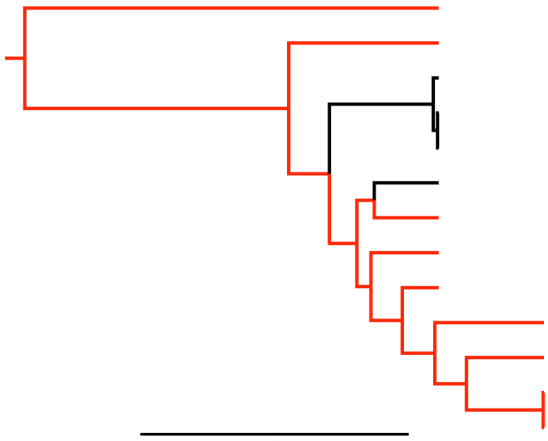
f) AR ($AI = 0.19$, $p < 0.02$)

Figure 4.3: Midpoint-rooted ML trees from significantly compartmentalized patients, *env* gene data set, *a*) Patient AD; *b*) Patient AU; *c*) Patient BM; *d*) Patient BO; *e*) Patient AG; *f*) Patient AR. The scale bars represent one substitution per site per year. Sequences isolated from PBMC (red) and cervical (black) samples. Association Index (AI) posterior observed values and significance given.

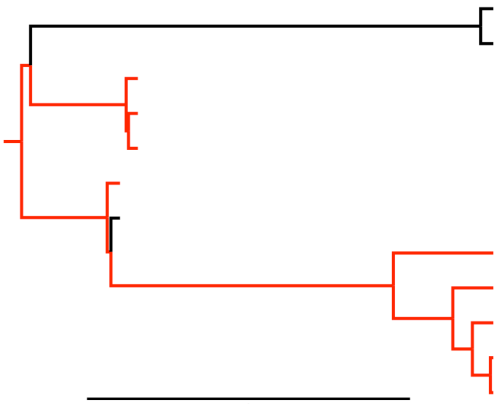
Maximum-likelihood trees: Only 3 *env* and 3 *gag* ML trees showed complete compartmentalization (reciprocal monophyly of cervix and PBMC sequences). Furthermore, of these, only two trees (AG and BO, both from the *env* data set) comprised more than ten taxa. Maximum likelihood trees estimated from the master data sets (not shown) confirmed that samples from each patient were monophyletic. ML trees from patients identified as significantly associated by the BaTS method above are shown in: Figure 4.3 (*env*); *a*) Patient AD, *b*) Patient AU, *c*) Patient BM, *d*) Patient BO, *e*) Patient AG, *f*) Patient AR – Figure 4.4 (*gag*); *a*) Patient AH, *b*) Patient AZ, *c*) Patient AW.



a) AH ($AI = 0.01$, $p < 0.005$)



b) AZ ($AI = 0.28$, $p < 0.07$)



c) AW ($AI = 0.11$, $p < 0.005$)

Figure 4.4: midpoint-rooted ML trees from significantly compartmentalized patients, *gag* data set, *a*) Patient AH; *b*) Patient AZ; *c*) Patient AW. The scale bars represent one substitution per site per year, except *c*) scale bar is 0.5 substitutions per site per year. Sequences isolated from PBMC (red) and cervical (black) samples. Association Index (AI) posterior observed values and significance given. AI was not significant to $\alpha < 0.05$ in Patient AH although all other statistics were.

Correlation between clinical and phylogenetic data: Many studies have shown a relationship between duration of infection and $CD4^+$ cell counts in HIV-1 disease progression (*e.g.* Shankarappa *et al*., 1999). Although complete longditudinal data was not available, a negative correlation was observed between duration of infection and $CD4^+$ count ($p < 0.05$; Figure 4.5). Since migration might be expected to be a time-dependent process we might expect to find a relationship between observed number of migrations and duration of infection (Figure 4.5). However number of migrations (observed value of the PS statistic in the BaTS analysis) was only weakly correlated ($p > 0.05$) in both *env* and *gag* data sets.

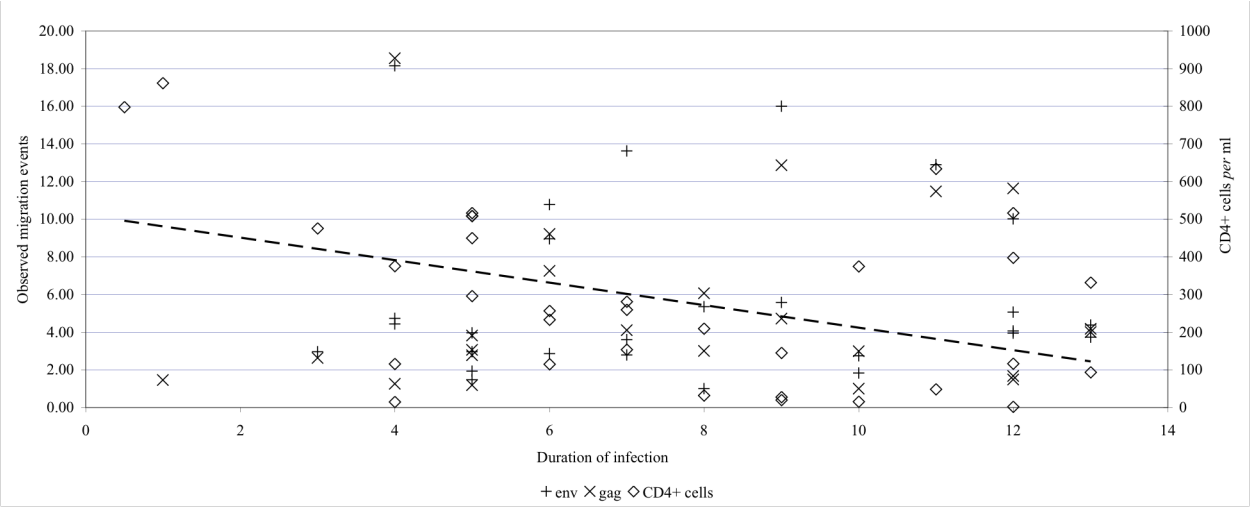


Figure 4.5: Correlation of infection duration, CD4+ counts and number of observed migrations between compartments. Infection duration and CD4+ cell counts were correlated ($p < 0.05$). Infection duration and number of migrations in either env or gag data sets were very weakly correlated ($p > 0.05$).

Following similar reasoning, we might expect that the strength of compartmentalization, as measured by the p-value of the AI statistic, might show some time-dependency (Figure 4.6). However they were only weakly correlated ($p > 0.05$) in both *env* and *gag* data sets.

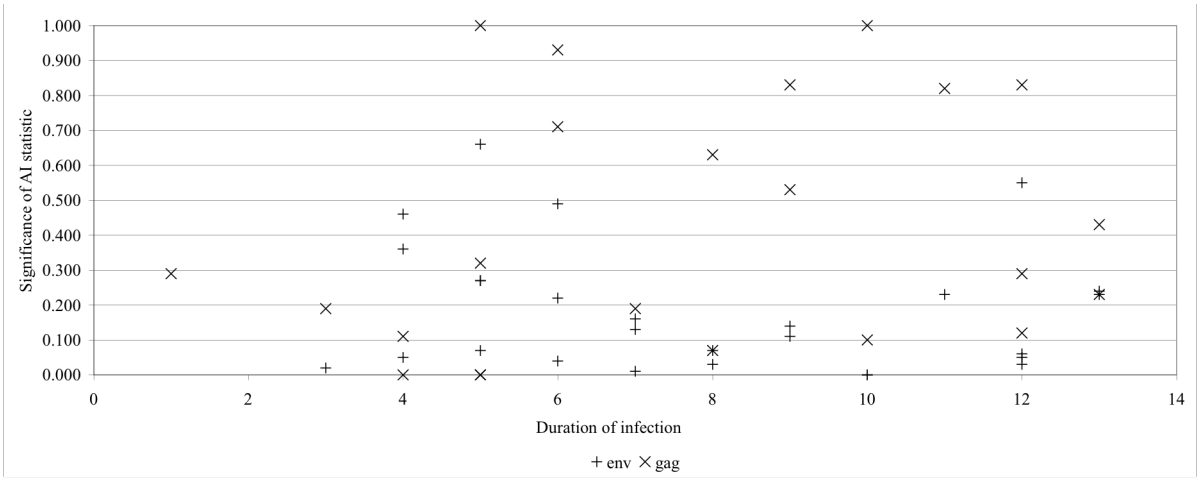


Figure 4.6: Correlation of disease progression (years) with significance of the observed AI statistic. Neither *env* nor *gag* data sets showed a significant correlation with duration of infection ($p > 0.05$).

Diversity in HIV infection has previously been shown to increase over time (Shankarappa *et. al.*, 1999). To investigate this possible relationship I measured Alignment diversity using the Shannon entropy score (see Appendix One), normalized as mean entropy per taxon in order to correct for alignments of different sizes (Figure 4.7). There was only a weak relationship between duration and diversity/entropy score ($p > 0.05$) in both *env* and *gag* data sets.

Finally, I explored the relationship between diversity and compartmentalization (Figure 4.8) to see if stronger evidence for compartmentalisation (measured as the p-value of the AI statistic test) was associatiated with greater diversity, which may reflect greater phylogenetic signal in a data set. There was only a weak correlation between *env* or *gag* data sets' diversity and AI statistic significance ($p > 0.05$).

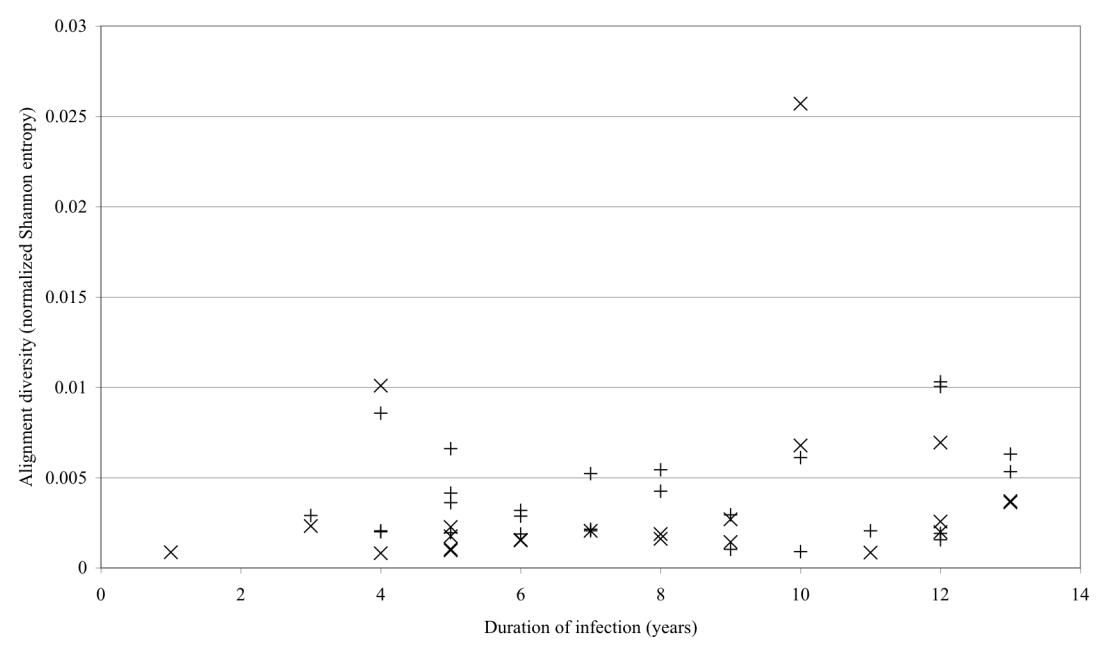


Figure 4.7: Correlation of alignment diversity with duration of infection, from *env* (plus signs) and *gag* (crosses) patients' data sets. Diversity given as the sum of Shannon information entropies at every site in the alignment, normalized by number of taxa (all alignments were the same length.)

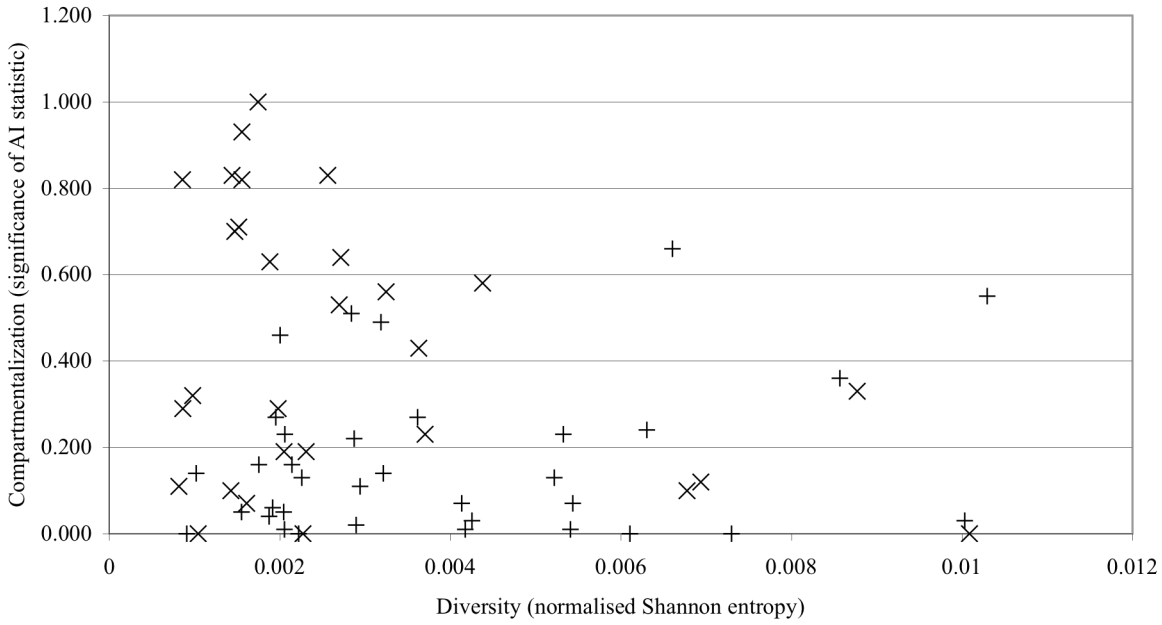


Figure 4.8: Correlation between *env* (plus signs) and *gag* (crosses) patients' degree of compartmentalization (significance of AI statistic) and alignment diversity (normalized Shannon entropy).

Of all the significantly compartmentalized *env* data sets for which progression information was available were long-term non-progressors (LTNP). Of the three compartmentalized *gag* data sets, one was an LTNP, one a standard progressor and one had no information available.

4.6 Discussion

This study employed a variety of statistical methods to detect compartmentalization in the *env* and *gag* genes of HIV sampled from PBMC and cervical tissues. I found evidence for the existence of a genotypically distinct HIV viral population in the cervix among individual *env* gene data sets moderately frequently, though fewer individual data sets showed sufficient individual support for genetic structure by compartment in the *gag* gene. This finding was repeated in the two separate meta-analyses of combined patients' data.

However this study has demonstrated the value of the methodological approaches selected; that some compartmentalization does occur, more regularly observed in the *env* gene than *gag*; and tentatively inferred that the direction of migrations tends to be from the blood to the cervix; furthermore these results detected compartmentalization in the *env* gene mainly in patients displaying with long-term non-progression of HIV infection to AIDS. Unfortunately, this information on disease progression was only available for a minority of patients; the relationship between compartmentalization in cervical tissue and disease progression indicators merits further study. Strength of compartmentalization is only weakly correlated with clinical indicators ($CD4^+$ count & duration of infection) and alignment diversity.

These results represent a validation of the approach introduced in Chapter Two. A single-tree approach such as the Slatkin-Maddison test would have failed to detect significance in this dataset following FDR correction; thus the Bayesian trait-phylogeny approach (implemented in BaTS) increased the power to detect significance at the same time as including phylogenetic error. The large number of patients sampled by this study also allowed two effective meta-analyses to be carried out (comparison of *p*-values & null distributions on

individual results, and BaTS analysis of master alignments) which gave concordant results. The compartment sub-tree categorization also informed the exploration of migration direction. Furthermore, this approach facilitated inference of the frequency of each category of migration direction in the posterior set of trees; this approach incorporates phylogenetic error in a way that more common methodologies based on parsimony reconstruction in single-tree data sets do not.

Where compartmentalization was identified by the BaTS approach, the results of the compartment sub-tree categorization approach suggest that the main direction of migration is from the blood to the cervix. This is not unanticipated: the cervical virus population is many times smaller than that of the blood, intermittently decreasing to the limit of detection (around 1 copy / ml; Nunnari *et al.*, 2005.) Cervical population sizes seem to be at least partly driven by those in the blood (Kovacs *et al.,* 2001) – in fact this simple observation suggests that compartmentalization in the cervix is rare because the established population size is so low. There may even be a number of cervical compartments; Pudney *et al.* (2005) suggest that environmental heterogeneity in the cervix contributes to the presence of a wide variety of cells of the immune system.

Compartmentalization was far more frequently and robustly detected in the *env* data set. This is more significant since more *env* data sets, which generally contained more samples and greater diversity (giving a stronger phylogenetic signal), were analyzed than *gag* ones. Given the variation in evolutionary rates and strength of selection between genes in HIV (Lemey *et al.*, 2005) the discrepancy between genes this result suggests is not surprising. There appears to be only a very weak correlation between alignment diversity (as measured by Shannon entropy) and strength of compartmentalization, however, so the observed

compartmentalization in *env* must be accounted for by another means. One explanation may be the strength of immune selection in *env* evolution (Ross & Rodrigo, 2002). Since the immune system macrophage populations in blood and cervix differ (Pudney *et al*, 2005), it is not unreasonable to expect compartmentalization in response and many studies have focussed on the *env* gene for this reason (Lorenzo *et al*, 2004; Philpott *et al*, 2005). In support of this hypothesis, *env* compartmentalization was seen in long-term non-progressors only, a disease profile thought to be associated with an effective immune response (Carotenuto *et al.*, 1998; Delwart *et al.*, 1998). Similarly Kemal *et al.* (2003) found a correlation (though not repeated here) between $CD4^+$ cell counts and *env* compartmentalization, while Sullivan *et al*. (2005) also reported a close correlation between $CD4^+$ cell population dynamics and cervical HIV infection.

These results stand in contrast to compartmentalization of the CNS and brain, where genetically distinct subpopulations have been well-characterized (Salemi *et al.* (2005); Wang *et al.* (2001); Korber *et al.*, (1994)). One clearly evident difference between cervical populations and those of the brain and CNS – aside from their ease of detection – is the absence of a physical barrier to infection in the cervix, which is highly vascularised. In contrast Korber *et al.* (1994) suggest that the blood-brain barrier is an effective barrier to viral migration.

I propose that separate evolutionary processes are responsible for the generation and maintenance of HIV compartmentalization in the cervix and the brain / CNS. Two principal models for compartment formation and maintenance have been advanced, roughly analogous to the sympatric and allopatric models of speciation. In the first, compartmentalization is driven by selection for different viral phenotypes (possibly due to altered cell tropism or

immune evasion requirements.) In this case, repeated selection and compartment-specific viral genotypes ought to be detected in separate hosts. In the second compartmentalization occurs due to lowered migration between compartments (due to physical barriers) with a resultant founder effect or genetic bottleneck leading to stochastic compartment subpopulation formation. This model predicts that repeated correlation of viral genotype with compartment within a single host would be repeated in other hosts, but with different polymorphisms occurring.

It follows that compartment formation in the brain and CNS constitutes an allopatric process, stochastically established by rare founder events when the blood-brain barrier is penetrated by chance. On the other hand, compartments in the cervix may be formed by a sympatric process, driven by immune-mediated selection (and possibly drug resistance (Devereux *et al.*, 2002)). If so, I predict that viral diversity and strength of selection would be stronger in cervical compartments than in brain / CNS compartments. Furthermore it is possible that (as with CTL escape mutations in the blood; Goulder & Watkins, 2004; Leslie *et al.*, 2004) repeated sets of substitutions occur in the cervix in different hosts, though not in the brain or CNS. Just as CTL escape mutations in the plasma have important implications for preventative and theraputic therapy design (Klenerman *et al.*, (2002)), this would suggest that evolution in the cervix – a major route of transmission – might follow repeated, even predictable, pathways.

In conclusion, I have not found evidence to confidently conclude that compartmentalization frequently occurs in cervical HIV populations. Although many individual data sets did not show strong evidence for compartmentalization, the distribution of *P*-values obtained, and the number of significant results remaining following multiple hypothesis test correction suggest

that genotypic subpopulation structure differs between the cervix and plasma.

Although conclusive evidence of compartmentalization was not detected in every dataset, I have developed and applied a variety of novel techniques to compartmentalization detection that are applicable to similar problems in other areas of virology. Furthermore, the methods used differ from existing ones not only in improved power to detect the specific hypothesis of compartmentalization, but also allow us to examine some of the details of cervical compartment formation.

From this basis I have been able to formulate a comprehensive model of compartment formation as it applies to HIV populations of the cervix and brain / CNS, with testable predictions. Work in the near future should seek to test these hypotheses, mainly with an analysis on the strength of selection in these two compartments. Bearing in mind the heterogeneity of compartmentalization among hosts seen in this study, a large number of subjects should be intensively sampled in order to adequately accept or reject this hypothesis. In the longer-term it seems clear that more work should be focused on molecular evolution of HIV in the cervix generally, particularly with respect to the existence or not of cervix-specific escape mutations as well as the impact of cervical compartmentalization on between-host phylodynamics.

4.7 – References

Adal, M., Ayele, W., Wolday, D. *et al*. (2005). Evidence of genetic variability of Human Immunodeficiency Virus Type 1 in plasma and cervicovaginal lavage in ethiopian women seeking care for sexually transmitted infections. *AIDS Research and Human Retroviruses*, **21** (7): 649-653.

Alves, K., Canzian, M., Delwart, E.L. (2002). HIV type 1 envelope quasispecies in the thymus and lymph nodes of AIDS patients. *AIDS Research and Human Retroviruses,* **18** (2): 161-165.

Bello, G., Casado, C., Garcia, S. et al. (2004). Co-existence of recent and ancestral nucleotide sequences in viral quasispecies of Human Immunodeficiency Virus Type 1 patients. *Journal of General Virology*, **85**: 399-407 part 2.

Benjamini, Y., and Hochberg Y. (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *Journal of the Royal Statistical Society. Series B (Methodological)* **57** (1), 289–300.

Berger, E. A., Murphy, P. A. & Farber, J. M. (1999) Chemokine receptors as HIV-1 coreceptors: Roles in viral entry, tropism and disease. *Annu. Rev. Immunol.* **17**:657-700.

Borrow, P., Lewicki, H., Wei, X., Horwitz, M. S., Peffer, N., Meyers, H., Nelson, J. A., Gairin, J. E., Hahn, B. H., Oldstone, M. B. & Shaw, G. M. (1997) Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. *Nat. Med.* **3**(2):205-211.

Caragounis, E. C., Gisslén, M., Lindh, M., Nordborg, C., Westergren, S., Hagber, L., Svennerholm, B. (2008) Comparison of HIV-1 *pol* and *env* sequences of blood, CSF, brain and spleen isolates collected ante-mortem and post-mortem. *Acta. Neurol. Scand.* **117**:108-116.

Carotenuto P, Looij D, Keldermans L, de Wolf F, Goudsmit J. (1998). Neutralizing antibodies are positively associated with CD4+ T-cell counts and T-cell function in long-term AIDS-free infection. *AIDS.* **12**(13):1591-600.

Charpentier C, Nora T, Tenaillon O, Clavel F, Hance AJ. (2006) Extensive recombination among human immunodeficiency virus type 1 quasispecies makes an important contribution to viral diversity in individual patients. *J. Virol.* **80**(5):2472-2482.

Daniels, R. S., Wilson, P., Patel, D., Longhurst, H. & Patterson, S. (2004) Analysis of full-length HIV Type-1 *env* genes indicates differences between the virus infecting T cells and dendritic cells in peripheral blood of infected patients. *AIDS Res. Hum. Retrovir.* **20**(4):409-413.

Delwart, E.L., Mullins, J.I., Gupta, P. *et al*. (1998).Human Immunodeficiency Virus Type 1 populations in blood and semen. *Journal of Virology*, **72** (1): 617-623.

De Oliveira, T., Deforche, K., Cassol, S., Salminen, M., Paraskevis, D., Seebregts, C., Snoek, J., van Rensburg, E.J., Wensing, A.M.J., van de Vijver, D.A., Boucher, C.A., Camacho, R. & Vandamme, A-M. (2005). An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics,* **21**(19):3797-3800.

Devereux, H.L., Burke, A., Lee, C.A. *et al*. (2002) *in vivo* HIV-1 compartmentalisation: drug resistance-associated mutation distribution. *Journal of Medical Virology*, **66** (1): 8-12.

Drummond AJ. (2006.) Population genetic estimation of the loss of genetic diversity during horizontal transmission of HIV-1. *BMC Evol Biol.* **23**;6:28.

Drummond, A.J. & Rambaut, A. (2003) BEAST v1.4, Available from http://beast.bio.ed.ac.uk/beast/.

Drummond, A. J. & Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**:214.

Fear, W. R. , Kesson, A. M., Naif, H., Lynch, G. W. & Cunningham, A. L. (1998) Differential tropism and chemokine receptor expression of human immunodefficiency virus Type 1 in neonatal monocytes, monocyte-derived macrophages, and placental macrophages. *J. Virol*. **72**(2):1334-1344.

Fitch, W.M. (1971b). Toward defining the course of evolution: Minimal change for a specific tree topology. *Syst. Zool.* **20**: 406-416.

Fulcher, J.A., Hwangbo, Y., Zioni, R. *et al*. (2004). Compartmentalization of Human Immunodeficiency Virus Type 1 between blood monocytes and CD4(+) T cells during infection. *Journal of Virology*, 78 (15): 7883-7893.

Goulder, P. J. R. & Watkins, D. I. (2004) HIV & SIV CTL escape: implications for vaccine design. *Nat. Rev. Immun.* **4**(8):630-640

Gupta, P., Leroux, C., Patterson, B. K., Kingsley, L., Rinaldo, C., Ding, M., Chen, Y., Kulka, K., Buchanan, W., McKeon, B. & Montelaro, R. (2000) Human immunodeficiency virus Type 1 shedding pattern in semen correlates with the compartmentalization of viral quasi species between blood and semen. *J. Infect. Dis.* **182**:79-87.

Harvey, P. H., & Pagel., M. D. (1991) *The Comparative Method in Evolutionary Biology.* Oxford University Press, Oxford.

Holder, M. & Lewis, P. O. (2003) Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.* **4**:275-284.

Iversen AK, Larsen AR, Jensen T, Fugger L, Balslev U, Wahl S, Gerstoft J, Mullins JI, Skinhoj P. (1998). Distinct determinants of human immunodeficiency virus type 1 RNA and DNA loads in vaginal and cervical secretions. *J Infect Dis.* **177**(5):1214-20.

Iversen, A.K.N., Learn, G.H., Skinhoj, P. *et al*. (2005) Preferential detection of HIV Subtype C ' over Subtype A in cervical cells from a dually infected woman. *AIDS,* **19** (9): 990-993.

Kemal, K.S., Foley, B., Burger, H. *et al*. (2003). HIV-1 in genital tract and plasma of women: compartmentalization of viral sequences, coreceptor usage, and glycosylation. *Proceedings of the National Academy of Sciences of the United States of America*, **100** (22): 12972-12977.

Kitchen, C.M.R., Philpott, S., Burger, H. *et al*. (2004) Evolution of Human Immunodeficiency Virus Type 1 coreceptor usage during antiretroviral therapy: a Bayesian approach. *Journal of Virology*, **78** (20): 11296-11302.

Klenerman, P., Wu, Y. & Phillips, R. (2002) HIV: Current opinion in escapology. *Curr. Op. Microbiol.* **5**:408-413.

Korber, B. T. M., Kunstman, K. J., Patterson, B. K., Furtado, M., McEvilly, M. M., Levy, R. & Wolinsky, S. M. (1994) Genetic differences between blood- and brain-derived viral sequences from human immunodeficiency virus Type 1-infected patients: Evidence of conserved elements in the V3 region of the envelope protein of brain-derived sequences. *J. Virol.* I**68**(11):7467-7481.

Kovacs, A., Wasserman, S.S., Burns, D. *et al*. (2001). Determinants of HIV-1 shedding in the genital tract of women. *Lancet,* **358** (9293): 1593-1601.

Lemey, P., Van Dooren, S. & Vandamme, A.M. (2005). Evolutionary dynamics of human retroviruses investigated through full-genome scanning. *Mol Biol Evol.* **22**(4):942-51.

Leslie AJ, Pfafferott KJ, Chetty P, Draenert R, Addo MM, Feeney M, Tang Y, Holmes EC, Allen T, Prado JG, Altfeld M, Brander C, Dixon C, Ramduth D, Jeena P, Thomas SA, St John A, Roach TA, Kupfer B, Luzzi G, Edwards A, Taylor G, Lyall H, Tudor-Williams G, Novelli V, Martinez-Picado J, Kiepiela P, Walker BD, Goulder PJ. (2004) HIV evolution: CTL escape mutation and reversion after transmission. *Nat Med.* **10**(3):282-9.

Liuzzi, G., Chirianni, A., Zaccarelli, M. *et al*. (2004) Differences between semen and plasma of nucleoside reverse transcriptase resistance mutations in HIV-infected patients, using a rapid assay. *in vivo*, **18** (4): 509-512.

Lorenzo, E., Colon, M.C., Almodovar, S. *et al*. (2004) Influence of CD4(+) T cell counts on viral evolution in HIV-infected individuals undergoing suppressive HAART. *Virology,* **330** (1): 116-126.

Magierowska, M., Bernardin, F., Garg, S. et al. (2004). Highly uneven distribution of tenofovir-selected simian immunodeficiency virus in different anatomical sites of rhesus macaques. *Journal of Virology*, **78** (5): 2434-2444.

Noë, A., Plum, J. & Verhofstede, C. (2005) The latent HIV-1 reservoir in patients undergoing HAART: an archive of pre-HAART drug resistance. *Journal of Antimicrobial Chemotherapy* **55**:410–412.

Nunnari, G., Sullivan, J., Xu, Y. et al. (2005). HIV type 1 cervicovaginal reservoirs in the era of HAART. *AIDS Research and Human Retroviruses,* **21** (8): 714-718.

Ohagen, A., Devitt, A., Kunstman, K. J., Gorry, P. R., Rose, P. P., Korber, B., Taylor, J., Levy, R., Murphy, R. L., Wolinsky, S. M. & Gabuzda, D. (2003) Genetic and functional analysis of full-length human immunodeficiency virus Type-1 *env* genes derived from brain and blood of patients with AIDS. *J. Virol.* **77**(22):12336-12345.

Philpott, S., Burger, H., Soukas, C. *et al*. (2005) Human Immunodeficiency Virus Type 1 genomic RNA sequences in the female genital tract and blood: compartmentalization and intrapatient recombination. *Journal of Virology*, **79** (1): 353-363.

Pillai, S.K., Good, B., Pond, S.K. *et al.* (2005) Semen-specific genetic characteristics of Human Immunodeficiency Virus Type 1 env. *Journal of Virology*, **79** (3): 1734-1742.

Poss, M., Rodrigo, A.G. Gosink, J.J. *et al.* (1998). Evolution of envelope sequences from the genital tract and peripheral blood of women infected with Clade A Human Immunodeficiency Virus Type 1. *Journal of Virology*, **72** (10): 8240-8251.

Potter, S. J. , Lemey, P., Achaz, G., Chew, C. B., Vandamme, A.-M., Dwyer, D. E. & Saksena, N. K. (2004) HIV-1 compartmentalization in diverse leukocyte populations during antiretroviral therapy. *J. Leukocyte. Biol.* **76**:562-570.

Pudney, J., Quayle, A.J., & Anderson, D.J. (2005). Immunological Microenvironments in the Human Vagina and Cervix: Mediators of Cellular Immunity Are Concentrated in the Cervical Transformation Zone. *Biology of Reproduction*. **73**:1253-1263.

Rambaut, A. & Drummond, A.J. (2007) Tracer v1.4, Available from http://beast.bio.ed.ac.uk/Tracer

Rambaut, A., Posada, D., Crandall, K.A. & Holmes, E.C. (2004) The causes and consequences of HIV evolution. *Nat. Rev. Genet*. **5**:52-61.

Rhodes, T., Wargo, H. & Hu, W.-S. (2003) High rates of human immunodeficiency virus Type 1 recombination: near random segregation of markers one kilobase apart after one round of viral replication. *J. Virol.* **77**(20):11193-11200.

Ross HA, Rodrigo AG. (2002) Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. *J Virol*. **76**(22):11715-20.

Salemi, M. & Vandamme, A.-M. (2002) Hepatitis C virus evolutionary patterns studies through analysis of full-genome sequences. *J. Mol. Evol.* **54**:62-70.

Sanjuán, R., Codoner, F.M., Moya, A. *et al*. (2004) Natural selection and the organ-specific differentiation of HIV-1v3 hypervariable region. *Evolution,* **58** (6): 1185-1194.

Shaffer, J.P. (1995). Multiple hypothesis testing. *Ann Rev. Psychol*. **46**:561-84

Shaheen, F., Sison, A.V., McIntosh, L. *et al*. (1999). Analysis of HIV-1 in the cervicovaginal secretions and blood of pregnant and nonpregnant women. *Journal of Human Virology*, **2** (3): 154-166.

Shankarappa, R., Margolick, J. B., Gange, S. J., Rodrigo, A. G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C. R., Learn, G. H., He, X., Huang, X.-L. & Mullins, J. I. (1999). Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus Type 1 infection. *J. Virol.* **73**(12):10489-10502.

Simmonds, P. (2004) Genetic diversity and evolution of hepatitis C virus – 15 years on. *J. Gen. Virol.* **85**:3173-3188.

Slatkin, M. & Maddison, W. P. (1989) A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics* **123**:603-613.

Sullivan, S.T., Mandava, U., Evans-Strickfaden, T. *et al*. (2005). Diversity, divergence, and evolution of cell-free Human Immunodeficiency Virus Type 1 in vaginal secretions and blood of chronically infected women: associations with immune status. *Journal of Virology*, **79** (15): 9799-9809.

Tang J, Tang S, Lobashevsky E, Myracle AD, Fideli U, Aldrovandi G, Allen S, Musonda R, Kaslow RA; Zambia-UAB HIV Research Project. (2002) Favorable and unfavorable HLA class I alleles and haplotypes in Zambians predominantly infected with clade C human immunodeficiency virus type 1. *J Virol.* **76**(16):8276-84.

Tirado, G. Jove, G., Kumar, R. *et al*. (2004). Differential virus evolution in blood and genital tract of HIV-infected females: evidence for the involvement of drug and non-drug resistance-associated mutations. *Virology*, **324** (2): 577-586.

van der Hoek, L., Goudsmit, J., Maas, J. *et al.* (1998). Human Immunodeficiency Virus Type 1 in faeces and serum: evidence against independently evolving subpopulations. *Journal of General Virology*, **79**: 2455-2459 part 10.

van't Wout, A.B., Ran, I.J., Kuiken, C.L. *et al.* (1998). Analysis of the temporal relationship between Human Immunodeficiency Virus Type 1 quasispecies in sequential blood samples and various organs obtained at autopsy. *Journal of Virology*, **72** (1): 488-496.

Wang, T.H., Donaldson, Y.K., Brettle, R.P., Bell, J.E. & Simmonds, P. (2001). Identification of shared populations of Human immunodeficiency Virus Type 1 infecting microglia and tissue macrophages outside the central nervous system. *J. Virol.* **75** (23): 11686-11699.

Wong, J.K., Ignacio, C.C., Torriani, F. *et al.* (1997). *in vivo* compartmentalization of Human Immunodeficiency Virus: evidence from the examination of pol sequences from autopsy tissues. *Journal of Virology*, **71** (3): 2059-2071.

Wyatt, R. & Sodroski, J. (1998) The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens. *Science* **280**:1884-1888.

Zhang, I.Q., Rowe, L., He, T. *et al.* (2002). Compartmentalization of surface envelope glycoprotein of Human Immunodeficiency Virus Type 1 during acute and chronic infection. *Journal of Virology*, **76** (18): 9465-9473.

Chapter Five

Error rate and statistical power of distance-based measures of phylogeny-trait association.

5.1 Abstract

Building on work presented in Chapter Two, I study here a number of more complex measures of phylogeny-trait association, which take into account the branch lengths of a phylogenetic tree in addition to the topological relationship between taxa. Extensive simulation is performed to measure the Type II error rate (statistical power) of these statistics including those introduced in Chapter Two, as well as the relationship between power and tree shape. The technique is applied to an empirical hepatitis C virus data set presented by Sobesky *et al*. (2007); their original conclusion that compartmentalization exists between viruses sampled from tumorous and non-tumorous cirrhotic nodules and the plasma is upheld. The association index (AI), migration (PS), phylodynamic diversity (PD) and unique fraction (UF) statistics offer the best combination of Type I error and statistical power to investigate phylogeny-trait association in RNA virus data, while the maximum monophyletic clade size (MC) and nearest taxon (NT) statistics suffer from reduced power in some regions of tree space.

5.2 Introduction

In Chapter Two, I reviewed many areas of viral evolutionary biology where more accurate estimation of the degree of association between the phylogenetic structure of a data set and the distribution of trait values of some character of interest at the tips of that phylogeny is desireable. These included viral phylogeography (Holmes, 2004; Starkman, 2003); population structure (Carrington *et al*., 2005; Nakano *et al.*, 2004); epidemiology (Leigh Brown *et al*., 1997) and compartmentalization (Pillai *et al.*, 2006; Salemi *et al*., 2005; Fulcher *et al.*, 2004) as well as T-cell escape (Bhattacharya *et al.*, 2007; Komatsu *et al*., 2006; Sheridan *et al*, 2004).

However, it was also noted in Chapter Two that previously adopted methodologies such as AMOVA (Sullivan *et al.*, 2005), single tree estimation (Potter *et al.*, 2004) or the Slatkin-Maddison test (Skatkin & Maddison, 1989), were deficient in some respects; significantly they failed to correctly incorporate phylogenetic error due to reliance on single-tree approaches to phylogeny-trait correlation. As a result, these methods were unable to assign significance to observed phylogeny-trait correlations. To address these concerns, Chapter Two presented a novel implementation ('BaTS') of three measures of phylogeny-trait association – the Association Index ('AI'; Wang *et al,* 2001); parsimony score ('PS'; following Fitch, 1971b); and introduced the new maximum monophyletic clade size statistic ('MC'). BaTS calculates these statistics in a Bayesian MCMC framework that takes into account phylogenetic uncertainty by 'averaging' over the posterior distribution of trees. The Type I error rate of these statistics was also measured through simulation and found to be correct.

In Chapter Four a large human immunodeficiency virus Type-1 (HIV-1) data set was analyzed using BaTS to determine the evidence for genetic compartmentalization of viral sequences (*env* & *gag* genes) in cervical tissue of 41 HIV-positive women. The differences in the relative performance of the statistics on an empirical data set were clear, with the AI and PS statistics appearing to be more statistically powerful than the MC statistic. Overall, there was sufficient evidence from the BaTS phylogeny-trait analysis to support the compartmentalization hypothesis; the single-tree approaches employed in Chapter Four were statistically weaker.

The conclusions of Chapter Two form the starting point for this study. An incorrect Type I error rate (false rejection of the null hypothesis) is generally taken to be a more serious flaw in any statistical approach than a Type II error rate (failure to correctly reject the null hypothesis where a significant result exists) since a definitive rejection of the null hypothesis leads us to modify our model. However, in studies of viral evolution large amounts of sequence data are often generated at considerable financial and scientific expense in order to investigate a particular hypothesis (*e.g.*, viral compartmentalization). In this light it seems clear that high statistical power (low Type II error) is also desirable in a statistical test. Accordingly, this study uses extensive simulations to quantify the Type II error rate of phylogeny-trait association statistics, as implemented in a Bayesian framework.

The AI, PS and MC statistics investigated in Chapter Two depend only on tree topology; they take into account only the branching order of taxa, not the absolute evolutionary distance between them. However, RNA viruses are capable of very rapid evolution (Jenkins *et al.*,

2002; Drake *et al*., 1998) and their phylogenies exhibit a wide range of tree shapes, from highly 'comb'-like (internal nodes distributed towards the terminal taxa) in dengue virus, to stat-like phylogenies with very long external branches (as in HIV population-level phylogenies) and highly unbalanced trees (*e.g.* influenza virus A population phylogenies; Grenfell *et al.*, 2004). It is therefore reasonable to consider the relevance of branch length information to the estimation of phylogeny-trait correlation.

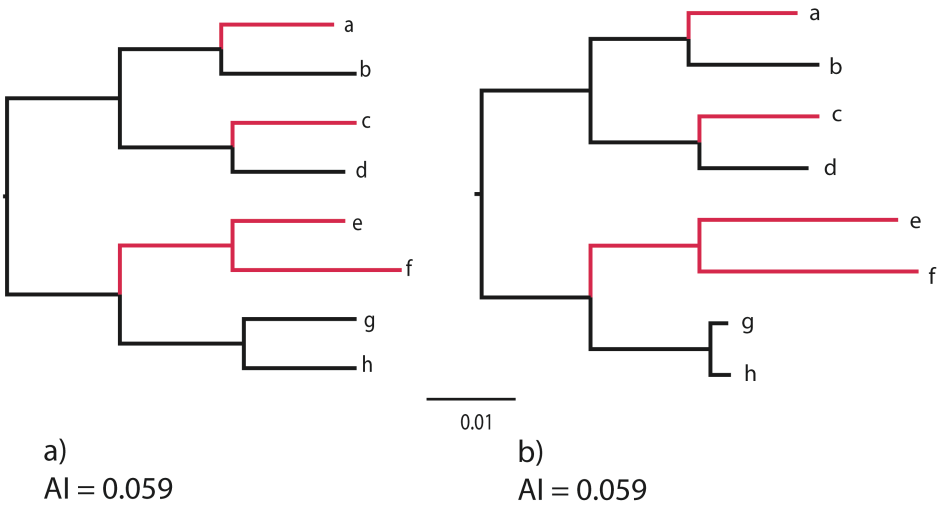


Figure 5.1: Trees *a*) and *b*) have identical topologies. The association between the 'red' and 'black' traits and phylogeny, as measured by the AI statistic, is necessarily the same for both.

Figure 5.1 gives an example of two trees that differ in tree branch lengths but share a topology, and have the same distribution of a hypothetical 'red / blue' trait at their terminal taxa. The AI statistic introduced by Wang *et al.* (2001) here measures the strength of association between the red or black traits' distribution and the phylogeny (higher values reflect a stronger association). Both the trees in Figure 5.1 would be calculated to have an AI of 0.059; this suggests that the red / blue trait is equally correlated with phylogeny, and of equal biological significance, in both data sets. However, the 'red' trait's association with phylogeny has been maintained through a considerable period of evolution and time in the

clade containing taxa 'e' and 'f' in Figure 5.1*b*, while the same correlation has so far been maintained over a much shorter period of evolution in Figure 5.1*a*. We might reasonably conclude that the association pattern seen in Figure 5.1*b* is more significant than that seen in Figure 5.1*a* – yet because the AI statistic ignores branch length information, we have no grounds to do so.

This chapter investigates four new statistics that include branch length information as well as taking into account the topological relationships among taxa. They are the phylogenetic diversity ('PD') measure of Faith (1992); the Net Relatedness ('NR') and Nearest Taxa ('NT') indices of Webb (2000; 2002); and the Unique Fraction ('UniFrac' or 'UF') statistic of Lozupone & Knight (2005).

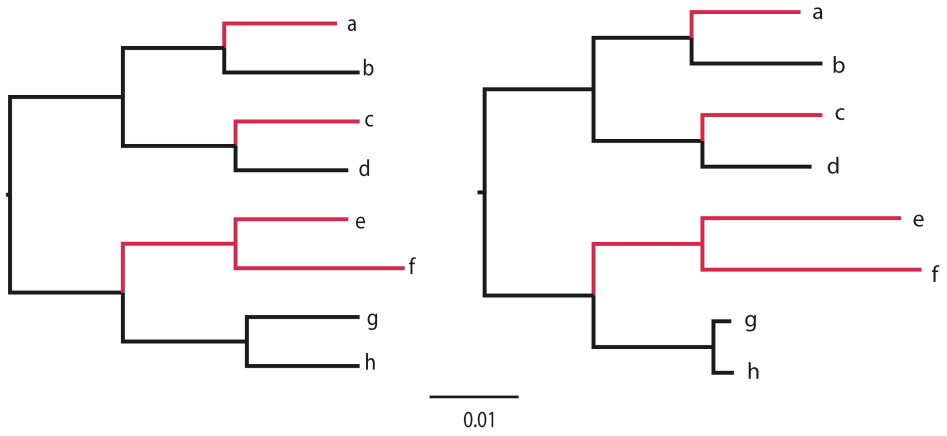


Figure 5.2: The trees presented in Figure 5.2; this time phylogeny-trait association is measured by four statistics (UniFrac, Nearest Taxon ('NT'), Net Relatedness ('NR') & Phylogenetic Diversity ('PD')). The value of the statistic is proportional to the strength of association; higher values are more strongly associated. Tree *b*) has stronger phylogeny-trait association than tree *a*).

By including branch length information these statistics may be able to discriminate between the two trees presented in Figure 5.1; Figure 5.2 shows the same phylogenies, but this time values for the new statistics are given. This time tree *b*) shows a stronger phylogeny-trait association than tree *a*) – the UniFrac, NT, NR and PD values are all higher.

This chapter seeks to investigate, through extensive simulation, the Type I and Type II error rates of all the statistics introduced in this chapter and those introduced in Chapter Two. The influence of tree shape on the Type I error rate is also investigated: since this technique is implemented in a Bayesian framework, the observed and null distributions of the association statistics are calculated from the posterior set of trees (PST). This is sampled from the true posterior distribution of topologies (topologies are sampled in proportion to their posterior probability) so power should be maintained equally well in topologies that are traditionally problematic for evolutionary parameter estimation (*e.g.* star-like trees). To illustrate the use of these statistics, I apply them to an empirical data set of within-patient HCV sequences, sampled from a number of different tissues by Sobesky *et al.* (2007). I re-visit their central hypothesis of genetic compartmentalization between tumoral and non-tumoral HCV-infected hepatocytes.

5.3 Methods

In this Chapter I add a number of new statistics to the BaTS package, first introduced in the course of Chapter Two. The new statistics differ from those implemented in Chapter Two; they incorporate branch length information as well as tree topology. Therefore it is more important in this Chapter to ensure the model of substitution is correctly selected and estimated to obtain accurate estimates of genetic distance, in addition to efficient sampling of the posterior distribution of tree topologies.

5.3.1 The Statistics

In the foregoing descriptions, *s* is defined as a subset of taxa on phylogenetic tree that only and exclusively possess a given discrete phenotypic trait value. They are not assumed to be monophyletic.

Phylogenetic Diversity ('PD'): The PD statistic was first proposed by Faith (1992) and is a simple intuitive measure of the amount of 'diversity', or genetic distance, captured by a subset *s* of taxa in a phylogeny. The PD of *s* here equals the sum of branch lengths (including terminal branches) in the subtree connecting all taxa in *s* but excluding any branches (internal or external) leading only to taxa that are not in *s* (the 'minimum spanning path', or MSP). To give an estimate of the strength of phylogeny-trait association in a data set, the PD_s of *s* is divided by the sum of all branch lengths in the phylogeny. This measure is summed for all subsets in of taxa present to give an estimate of the strength of association; in a completely-associated case the MSP of each subset will be shorter (and PD_s smaller) than in an interspersed case.

Nearest Taxon (NT): The NT score of *s* is defined as the sum, over all taxa in *s*, of branch lengths between each taxon and the nearest taxon that is also in *s*. This definition is modified from that proposed by Webb (2000) in two ways: Firstly, I use branch lengths rather than nodal distances. Secondly, and importantly, I do not divide the sum of NT distances by the maximum possible sum of nearest taxa distances in a tree to create an index. Instead, I simply measure the sum of NT distance for all taxa subsets in a tree. It is not necessary in the context of this study to create an index as Webb (2000) originally did, since BaTS generates a correct null distribution for the statistic through randomization of taxa trait allocations. Furthermore, calculating the maximum possible value exactly is computationally expensive in the current BaTS implementation, especially for large data sets.

Net Relatedness (NR): The net relatedness is defined as the sum of all pairwise distances between all members of *s*. As with the NT statistic, Webb (2000) introduced the statistic using nodal distances for calculation, and divided the NT by a maximum possible value of this statistic for any equally-sized subset of taxa to create an index. Again, the statistic is implemented here using estimated branch lengths in place of nodal distances and not as an index, instead calculating the significance of the observed NR value by generating an appropriate null distribution by simulation.

Unique Fraction ('UniFrac', or 'UF'): This simple measure, introduced by Lozupone & Knight (2005) is the proportion of internal branches on a phylogeny that connect nodes whose trait values are unambiguously resolved following trait value reconstruction by

parsimony (Fitch, 1971b). The sum of UF values for *s* is expressed as a ratio of the sum of internal branch lengths of the tree.

5.3.2 Incorporating phylogenetic uncertainty

Phylogenetic uncertainty (statistical error in phylogenetic estimation arising from sequence data) is taken into account using the approach developed in Chapter Two. The expanded computer package, Befi-BaTS (Bayesian Tip-association Significance) is available on request or from http://www.lonelyjoeparker.com

5.3.3 Simulation

In Chapter Two, I estimated the Type I statistical error (*i.e.* the probability of falsely rejecting the null hypothesis) through simulation. If the statistic is correct then the distribution of *p*-values of a set of randomly drawn phylogeny-trait associations should follow a unit uniform distribution. Here, I repeat that approach to investigate the Type I statistical error of the newly-introduced PD, NT, NR & UF statistics.

In addition, I conduct a new series of simulations to test the Type II error rate of all phylogeny-trait association statistics. The Type II error rate is defined as the frequency at which a method fails to reject the null hypothesis when it is false. This is also known as the 'power' of a statistical method; a statistic may have a correct Type I error rate, but its applicability to analysis will be limited if it is weak or overly conservative (of diminished power) since it may ignore too many significant results.

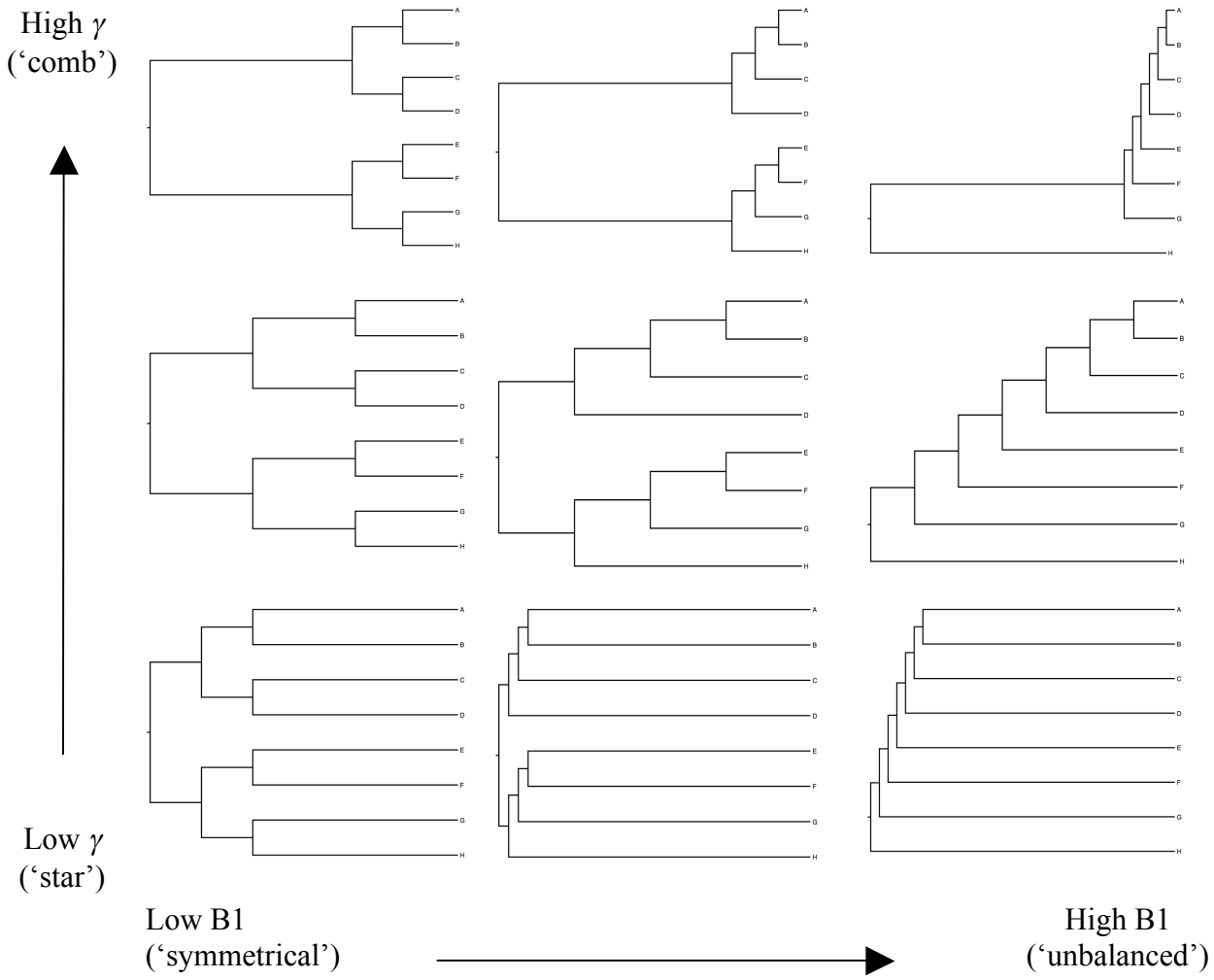


Figure 5.3: Diagram of the spread of tree shapes represented by the nine master topologies used in simulation, ordered by their node spread (γ statistic, vertical axis) and tree imbalance, (B1, horizonal axis).

The set of test phylogenies simulated in Chapter Two were used to explore the power of these statistics. Firstly, a set of test alignments were generated and analyzed in BEAST to obtain a set of PSTs with which to test Befi-BaTS:

1. 1000 phylogenies were generated under a pure-birth process using Phylo-O-Gen (available from http://evolve.zoo.ox.ac.uk). The tree imbalance (Colless, 1982) and

node spread (γ, Pybus & Harvey, 2000) statistics were calculated for each tree in the set. Nine 'master' topologies were selected that reflected all possible combinations of tree imbalance and node spread for tree imbalance values of (0, 0.125, 0.5) and γ values of (-2, 0, 2). Figure 5.3 shows a diagram of the range of tree shapes thus selected.

2. A large set (n = 1000) of alignments were simulated from each of the nine master tree topologies by Seq-Gen (Rambaut & Grassly, 1997). Substitution model parameters derived from typical human immunodeficiency virus Type 1 (HIV-1) data were used[1]. Each alignment contained 32 taxa and was 300 nucleotides long.
3. The PST for each alignment was then estimated using BEAST v1.4 (Drummond & Rambaut, 2007). An HKY85 + Γ substitution model with codon-position-specific substitution rates and the strict molecular clock enforced (rate fixed to μ= 0.017) under a constant population-size demographic model.
4. The set of simulations was down-sampled (to *n* = 897) to reduce computation. The first 10% of each PST was removed as burn-in. The PSTs produced were used for the shuffling procedure below.
5. Statistics that measure tree spread tree imbalance and node spread (two measures that together, describe most aspects of tree topology) were calculated for these source trees using code from the TreeStat program (Drummond & Rambaut, 2007. Available

[1] The substitution model parameters were derived from analysis of the *env* gene data set sampled from Patient AB in BEAST analysis (Chapter Four). Transition : transversion ratio = 2.54; Nucleotide frequencies, A=0.426, C=0.152, G=0.182; specific substitution rates for first, second and third codon positions respectively, μ_1 = 0.0152, μ_1 = 0.0142, μ_1 = 0.0215 (in substitutions per site per year).

from: http://tree.bio.ed.ac.uk); I developed a modified command-line interface to facilitate batch processing (author's work, available on request). The statistics calculated were: B1 (Kirkpatrick & Slatkin, 1993); Tree-imbalance (Colless, 1982); Cherry count (Steel & Mackenzie, 2001); γ and δ (Pybus & Harvey, 2000) and Fu & Li's D (Fu & Li, 1993).

In the second stage, the 897 PSTs generated in Step 4 above were used to investigate the power of the phylogeny-trait association statistics. In order to measure the Type II error rate it was necessary to generate data sets with different levels of phylogeny-trait association as follows:

1. Each taxon in each PST of the set of PSTs was initially labelled with a hypothetical binary character trait (*e.g.*, 'black' / 'white') using the known master topology (the underlying 'true' tree) in step 1 above to ensure maximal phylogeny-trait association. These phylogeny-trait labellings are referred to as 'completely associated'.
2. A new set of phylogeny-trait associations were generated by selecting two taxa at random and exchanging their trait values. This is referred to as a 'shuffle'. Note that the posterior set of trees remains unchanged; only the taxon-trait labelling is modified.
3. Re-arrangements were carried out to give multiple data-sets, each comprising 897 PSTs with the same trees but varying numbers of shuffles. As the number of shuffles increases, the tip-trait associations become more random, from the completely associated set (0 shuffles) to a set with random taxon trait labels (10,000 shuffles). Data sets of 1, 2, 3…33, 60, 70, 80, 90, 100, 500, 1000, 5000 & 10000 shuffles were produced.

4. Each shuffled data set was analysed with Befi-BaTS (using 100 replicates to calculate the null distribution) to determine: *a*) the frequency of positives in each statistic (statistics whose observed values $p \leq 0.05$) and *b*) the mean significance (*p*-value) of each statistic. In addition, the cumulative density function (CDF) of each statistic for every shuffled set was determined by ordering and binning the *p*-values obtained. These CDFs were compared to a unit uniform distribution using the Kolmogorov-Smirnov test (Lilliefors, 1969; Massey, 1951) to investigate the transition between the completely associated, interspersed, and random cases of phylogeny-trait association.

5.3.4 Empirical Data

To illustrate the application of this technique to viral sequence data, I analysed an empirical hepatitis C virus (HCV) data set reported by Sobesky *et al.* (2007). The authors sought to determine whether significant genetic compartmentalization existed between HCV virus populations sampled from peripheral blood and from cirrhotic nodules (two normal and one cancerous) of a post-transplant human liver. Individual hepatocytes were sampled by microdissection whilst serum samples were taken *in vivo*. Data was collected from seven patients and alignments spanned 573 nucleotides of the *core* gene.

To investigate the hypothesis of compartmentalization using the new methods introduced here, a PST was calculated from the data (aligned using Se-Al; http://evolve.zoo.ox.ac.uk) using BEAST 1.4 (Drummond & Rambaut, 2007) for two patients from the data set: P1 (n = 70 sequences) and P7 (n = 68 sequences). Substitution, clock and demographic models were selected based on the most likely models identified for similar data (the *core* gene window of

the ‘Anti-D’ within-patient data set in Chapter Three): a constant population-size model of demographic growth and an HKY85 + Γ model of nucleotide substitution with the strict molecular clock enforced at 0.005 substitutions / site / year. Six MCMC analyses were independently performed for 10,000,000 states each to check convergence. Taxa were labelled with their tissue of origin, and analyzed in Befi-BaTS with 100 replicates used to calculate the null distribution.

5.4 Results

5.4.1 Type I Error rate

The number of significant results ($p \leq 0.05$) obtained using each statistic when taxon trait labels were shuffled 10,000 times is given in Table 5.1. This simulates random taxon trait allocation (the null hypothesis), so equals the Type 1 error rate of these statistics. The CDFs of all statistics were not significantly different from a unit uniform distribution in the 10,000 shuffles data set.

Statistic	Type I rate
AI	0.051
PS	0.046
UF	0.028
PD	0.041
NR	0.062
NT	0.041
MC	0.029

Table 5.1: Type I error rate of statistics implemented in the Befi-BaTS package. Error rate given is the proportion of significant results ($p \leq 0.05$) observed in a data set of 897 randomly assigned tip trait values (binary character, 10,000 shuffles).

5.4.2 Type II Error rate

Figures 5.4 – 5.10 give the results for the AI, PS, PD, UF, NR, NT & MC statistics respectively. In each figure, the top plot shows the cumulative density function (CDF) of the statistic for increasingly shuffled (more weak phylogeny-trait association) simulations, the centre plot shows the proportion of rejections of H_0 with increasing shuffles and the bottom

plot shows the mean p-value of the test with increasing shuffles. A red dashed line is drawn at $p = 0.05$.

CDF curves for most statistics show a smooth transition from maximal association (no shuffles) to random tip-trait associations (approximately those simulations with more than 100 shuffles). The randomly associated simulations have CDFs that are unit uniformly distributed (diagonal grey line). However, the MC statistic CDFs quickly fall below the diagonal line, even at low numbers of shuffles, indicating that the MC statistic is a weak measure. In contrast the NR statistic CDF never reaches the diagonal line, suggesting the Type I error of this statistic may not be correct at some levels of α.

The Kolmogorov-Smirnov test (Lilliefors, 1969; Massey, 1951) was used to calculate the significance of difference between *p*-values CDF of each simulation and a unit uniform distribution (the expected distribution of *p*-values under the null hypothesis). The value of the Kolmogorov-Smirnov statistic, D^+, and significance, are given in Figure 5.11. Across the range of shuffles used, the NR statistic showed the weakest departure from uniformity, while the NT and PS statistics showed greatest departure from uniformity.

The number of significant tests and the mean significance of each test that are given in Figures 5.4 – 5.10 for each statistic are presented together for visual comparison in Figure 5.12 and Figure 5.13. Figure 5.12 shows that the proportion of significant tests ($p \leq 0.05$) obtained using the MC and NT statistics declines more rapidly with the number of shuffles than other statistics, indicative of weak statistical power. The PS and NR statistics, on the other hand, continue to strongly reject H_0 even in large numbers of shuffles. Equally, in

Figure 5.13 the mean *p*-values of the tests (probability of accepting the null hypothesis) rapidly increases with increasing shuffles for the MC and NT statistics. In contrast, the PS and particularly, NR, statistics show a lower mean significance.

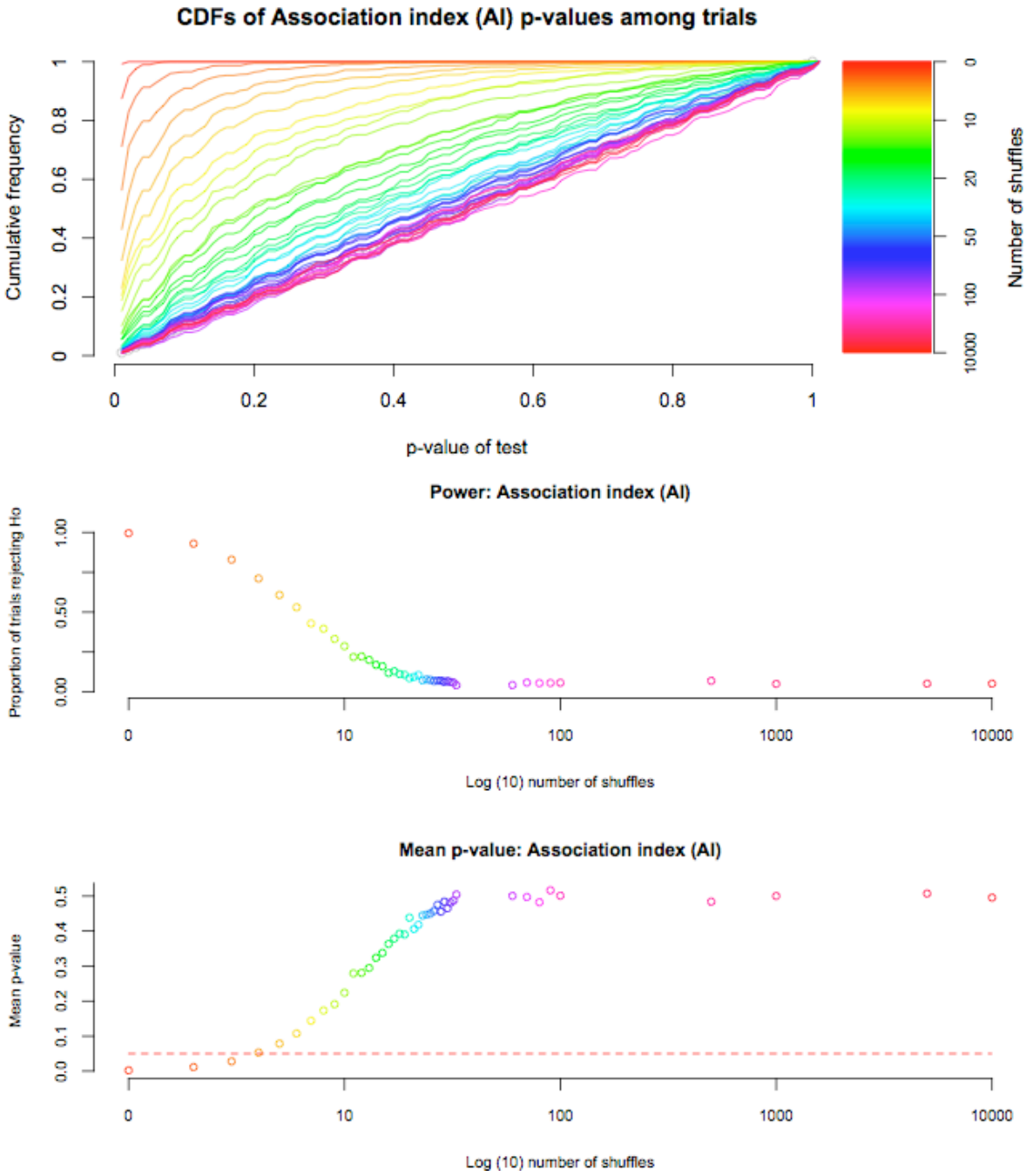


Figure 5.4: CDFs and performance of AI statistic on simulated data. Top panel: CDFs of each simulation, from no shuffles, or completely associated (red) to 10,000 shuffles (violet). The unity (unit uniform distribution) is shown in grey. Centre panel: proportion of simulations rejecting H_0 (out of 897 possible) with increasing trait re-arrangements ($\log_{10}$). Lower panel: mean significance of observed AI statistic.

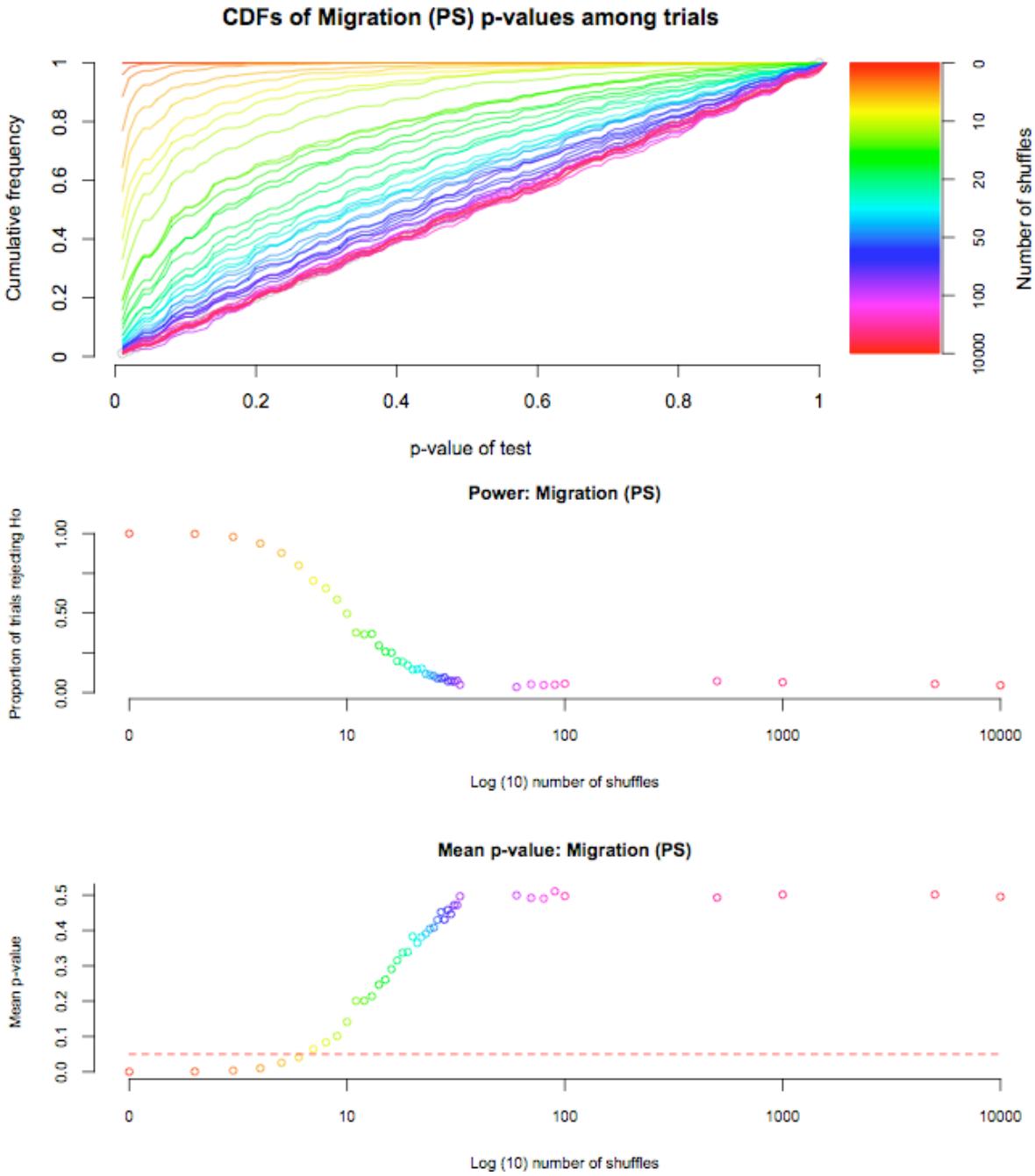


Figure 5.5: CDFs and performance of parsimony statistic (PS) on simulated data. Top panel: CDFs of each simulation, from no shuffles, or completely associated (red) to 10,000 shuffles (violet). The unity (unit uniform distribution) is shown in grey. Centre panel: proportion of simulations rejecting H_0 (out of 897 possible) with increasing trait re-arrangements ($\log_{10}$). Lower panel: mean significance of observed parsimony statistic.

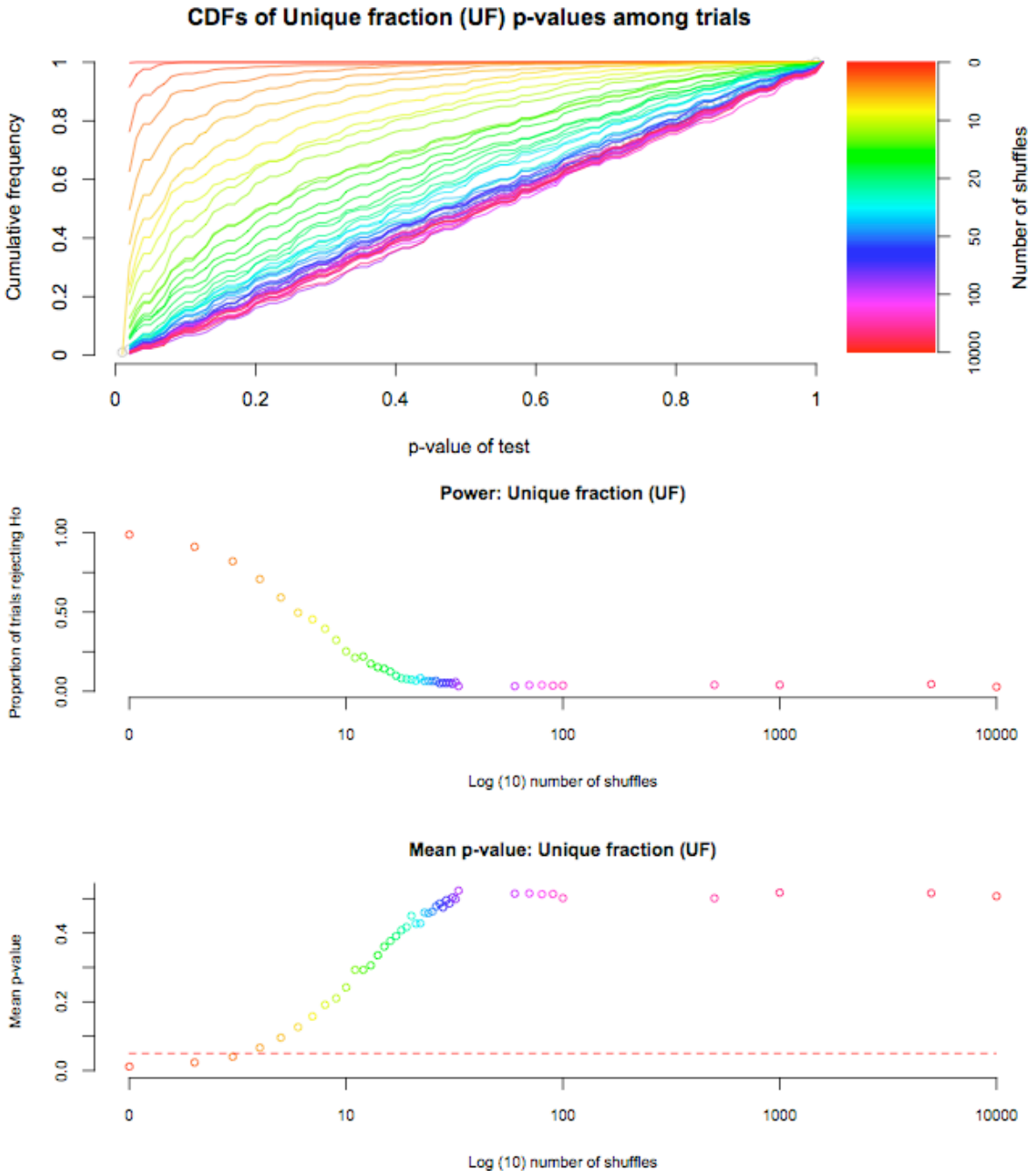


Figure 5.6: CDFs and performance of unique fraction (UniFrac) statistic on simulated data. Top panel: CDFs of each simulation, from no shuffles, or completely associated (red) to 10,000 shuffles (violet). The unity (unit uniform distribution) is shown in grey. Centre panel: proportion of simulations rejecting H_0 (out of 897 possible) with increasing trait re-arrangements ($\log_{10}$). Lower panel: mean significance of observed UniFrac statistic..

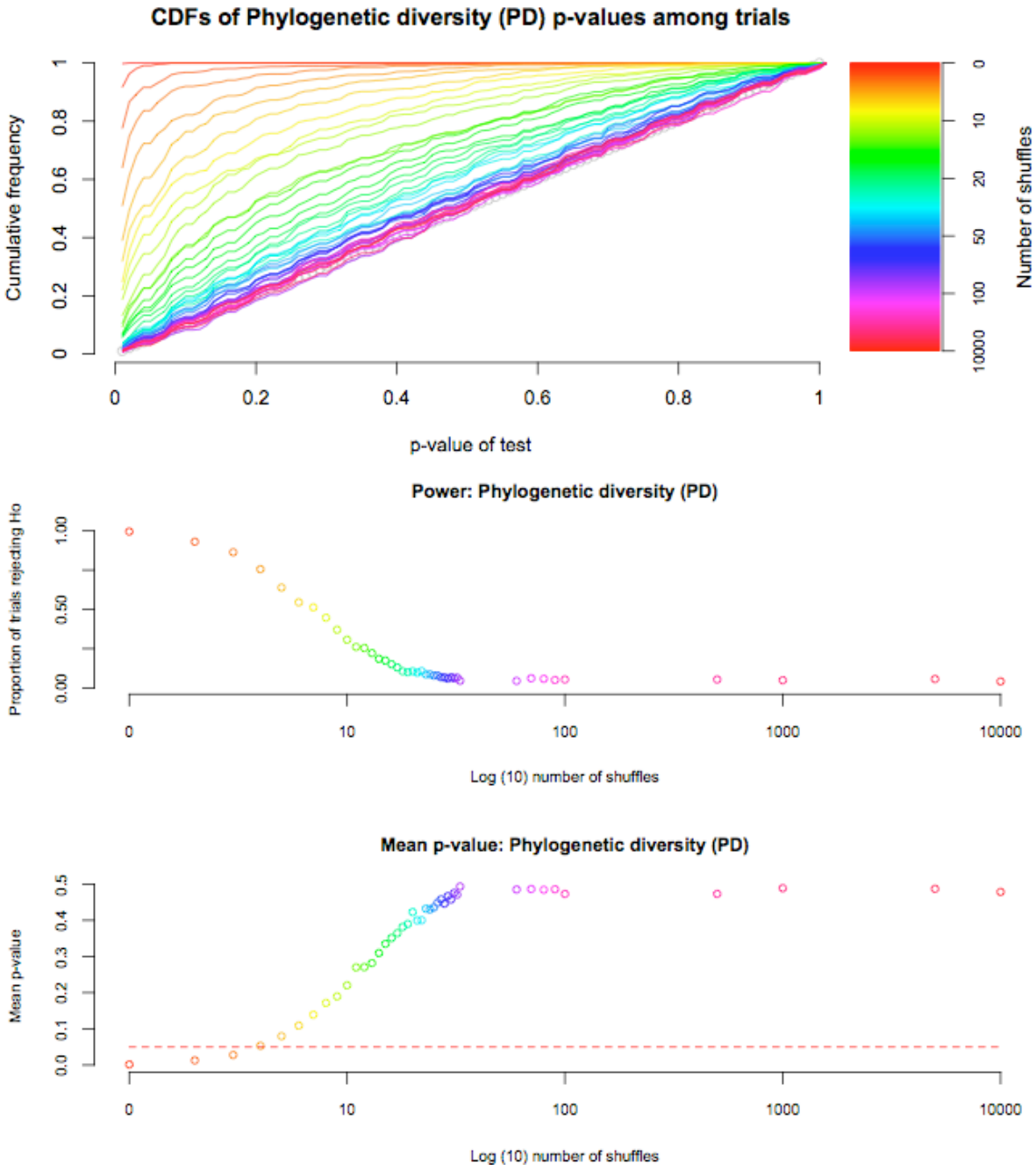


Figure 5.7: CDFs and performance of phylogenetic diversity (PD) statistic on simulated data. Top panel: CDFs of each simulation, from no shuffles, or completely associated (red) to 10,000 shuffles (violet). The unity (unit uniform distribution) is shown in grey. Centre panel: proportion of simulations rejecting H_0 (out of 897 possible) with increasing trait re-arrangements ($\log_{10}$). Lower panel: mean significance of observed PD statistic.

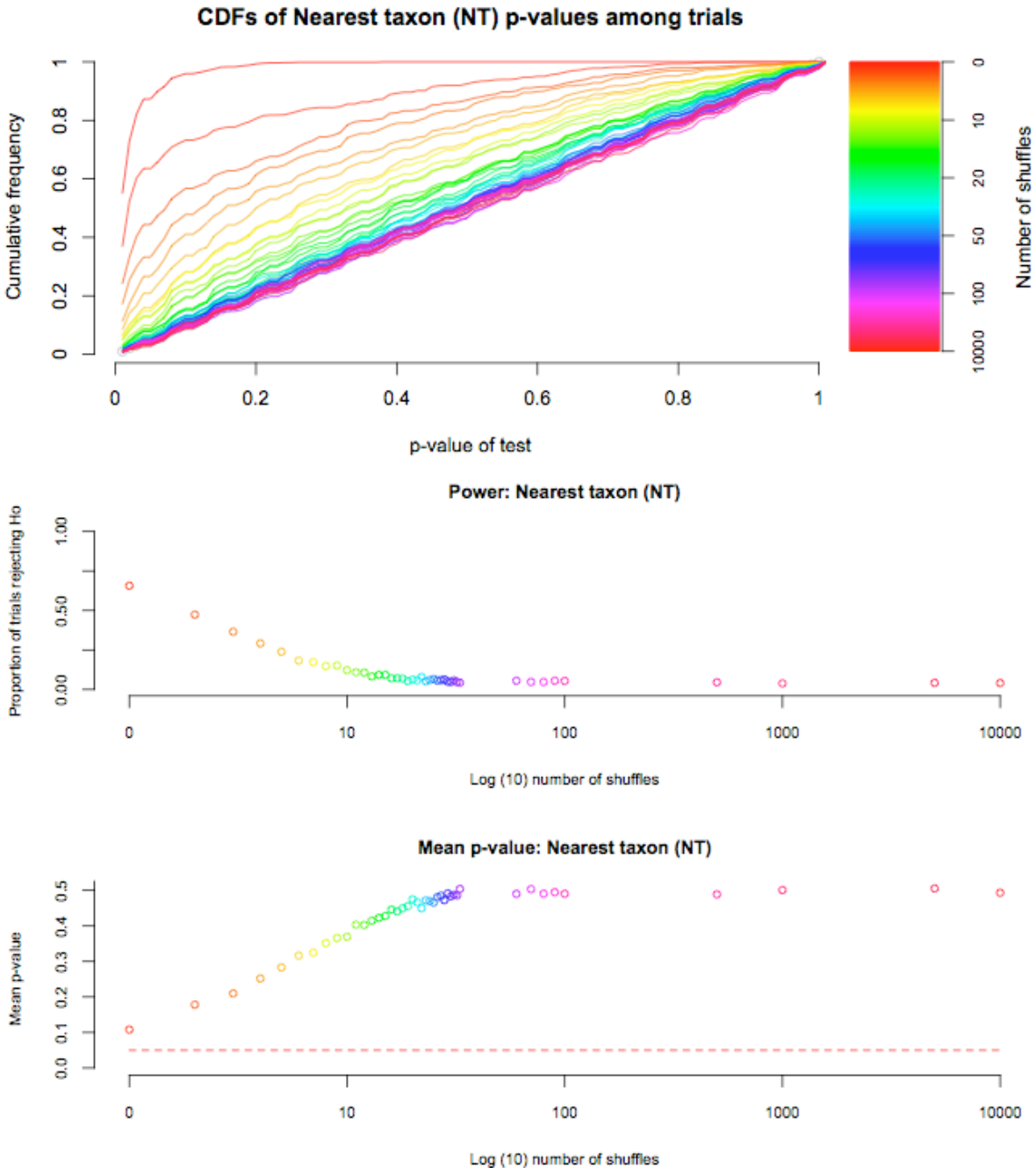


Figure 5.8: CDFs and performance of nearest taxon (NT) statistic on simulated data. Top panel: CDFs of each simulation, from no shuffles, or completely associated (red) to 10,000 shuffles (violet). The unity (unit uniform distribution) is shown in grey. Centre panel: proportion of simulations rejecting H_0 (out of 897 possible) with increasing trait re-arrangements ($\log_{10}$). Lower panel: mean significance of observed NT statistic.

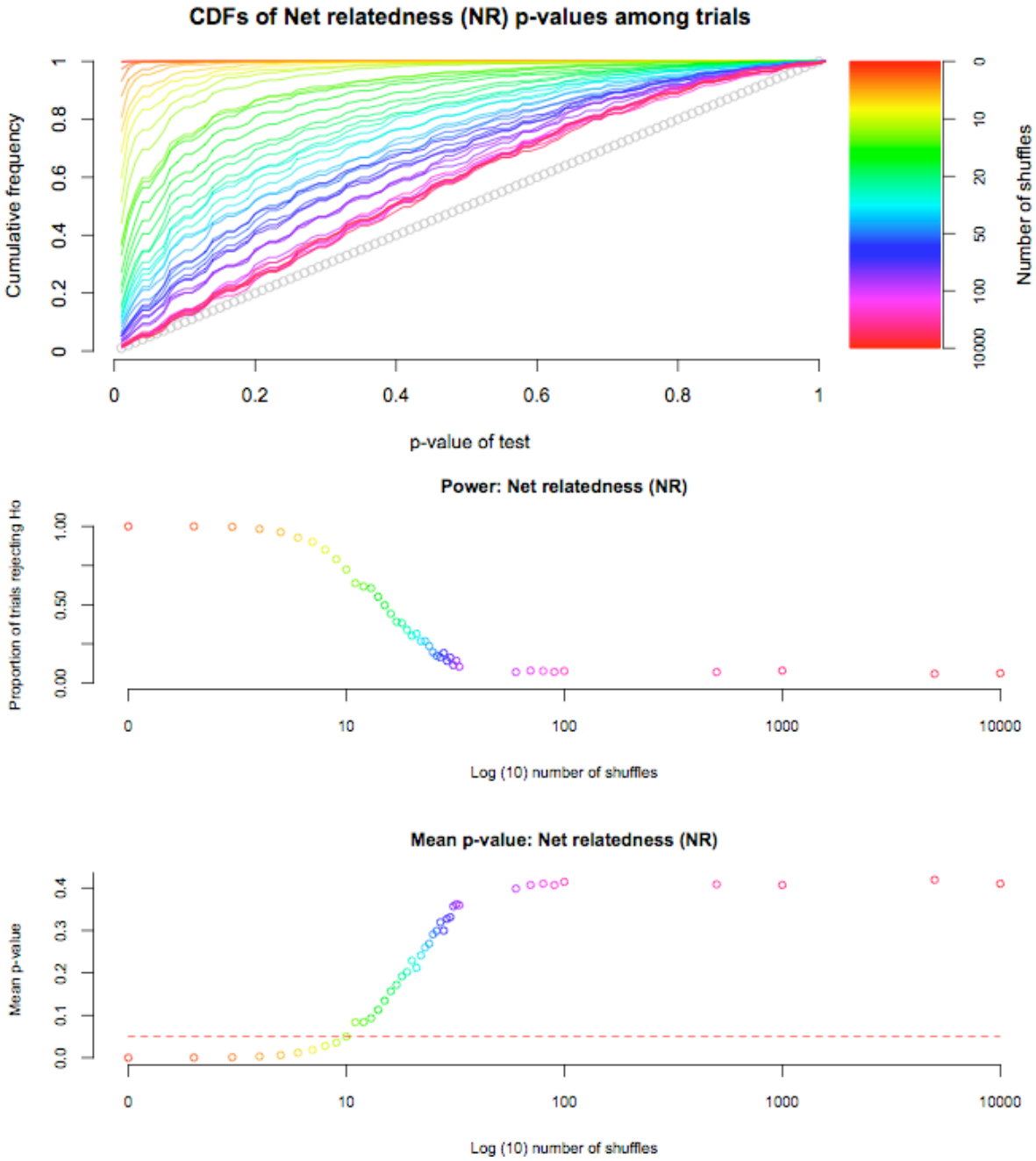


Figure 5.9: CDFs and performance of net relatedness (NR) statistic on simulated data. Top panel: CDFs of each simulation, from no shuffles, or completely associated (red) to 10,000 shuffles (violet). The unity (unit uniform distribution) is shown in grey. Centre panel: proportion of simulations rejecting H_0 (out of 897 possible) with increasing trait re-arrangements ($\log_{10}$). Lower panel: mean significance of observed NR statistic.

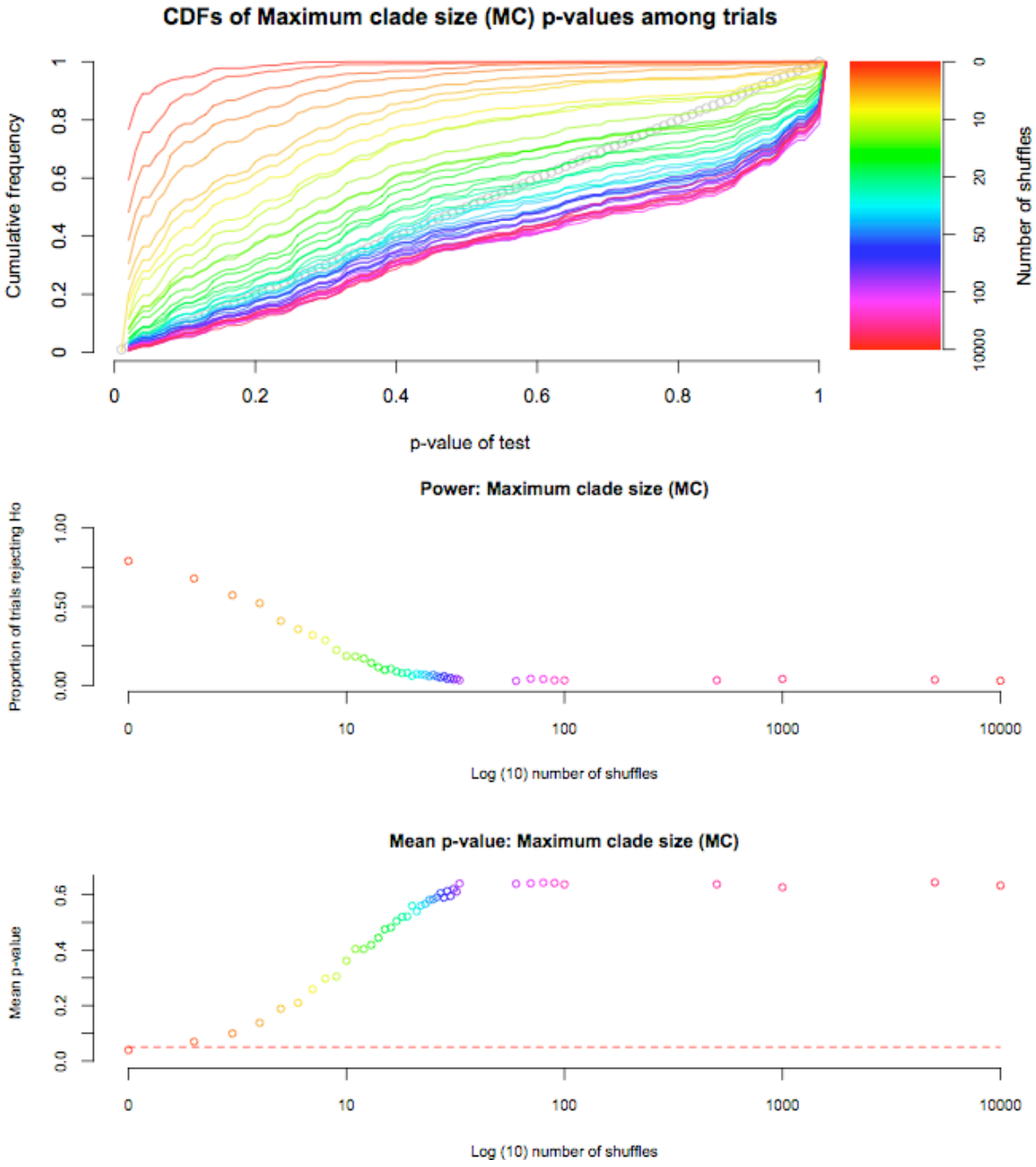


Figure 5.10: CDFs and performance of MC statistic on simulated data. Top panel: CDFs of each simulation, from no shuffles, or completely associated (red) to 10,000 shuffles (violet). The unity (unit uniform distribution) is shown in grey. Centre panel: proportion of simulations rejecting H_0 (out of 897 possible) with increasing trait re-arrangements ($\log_{10}$). Lower panel: mean significance of observed MC statistic.

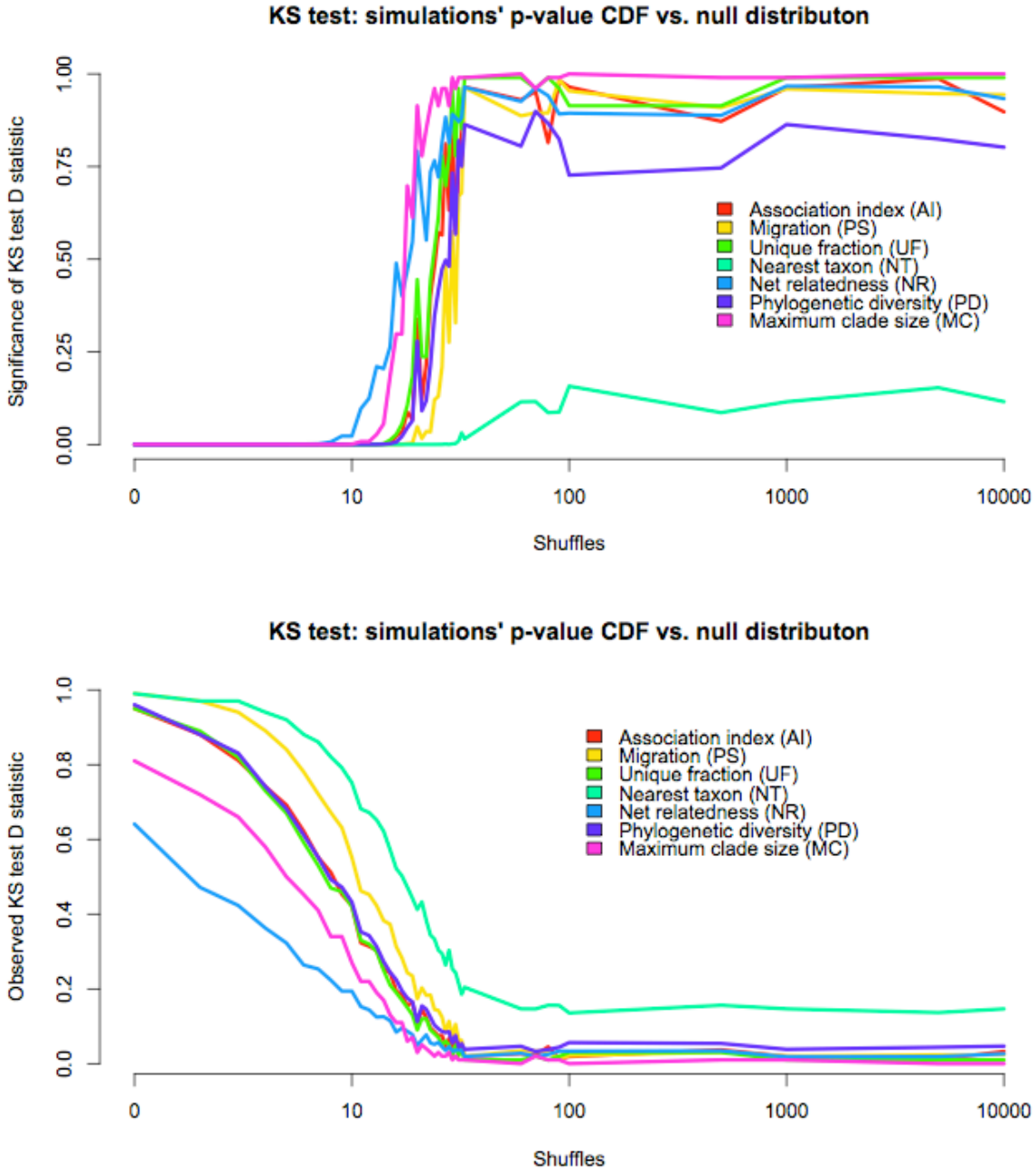


Figure 5.11: The CDF for each statistic was compared to a unit uniform distribution under increasing numbers of taxon rearrangements using a Kolmogorov-Smirnoff test. Shown are the value of the difference statistic (lower plot) and *p*-value (upper plot) in each separate simulation replicate ($\log_{10}$(taxon rearrangements)).

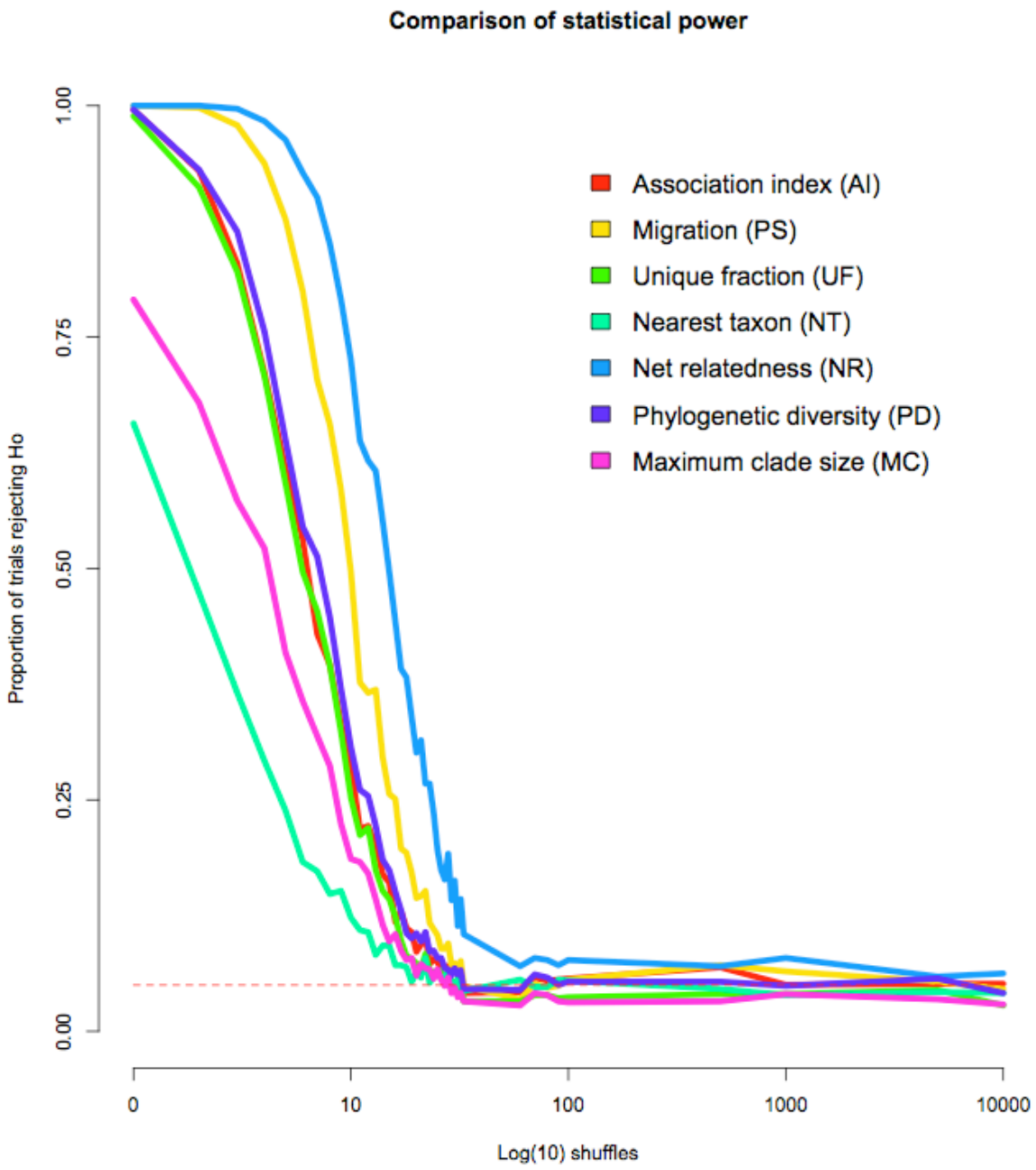


Figure 5.12: Proportion of rejections of H_0 ($p \leq 0.05$) with increasing numbers of random taxon trait-value rearrangements (log scale) in different statistics. The dashed red line is at 0.05 (5%), the proportion of trials expected to reject H_0 under the null hypothesis at $\alpha = 0.05$ if the Type I error rate is correct.

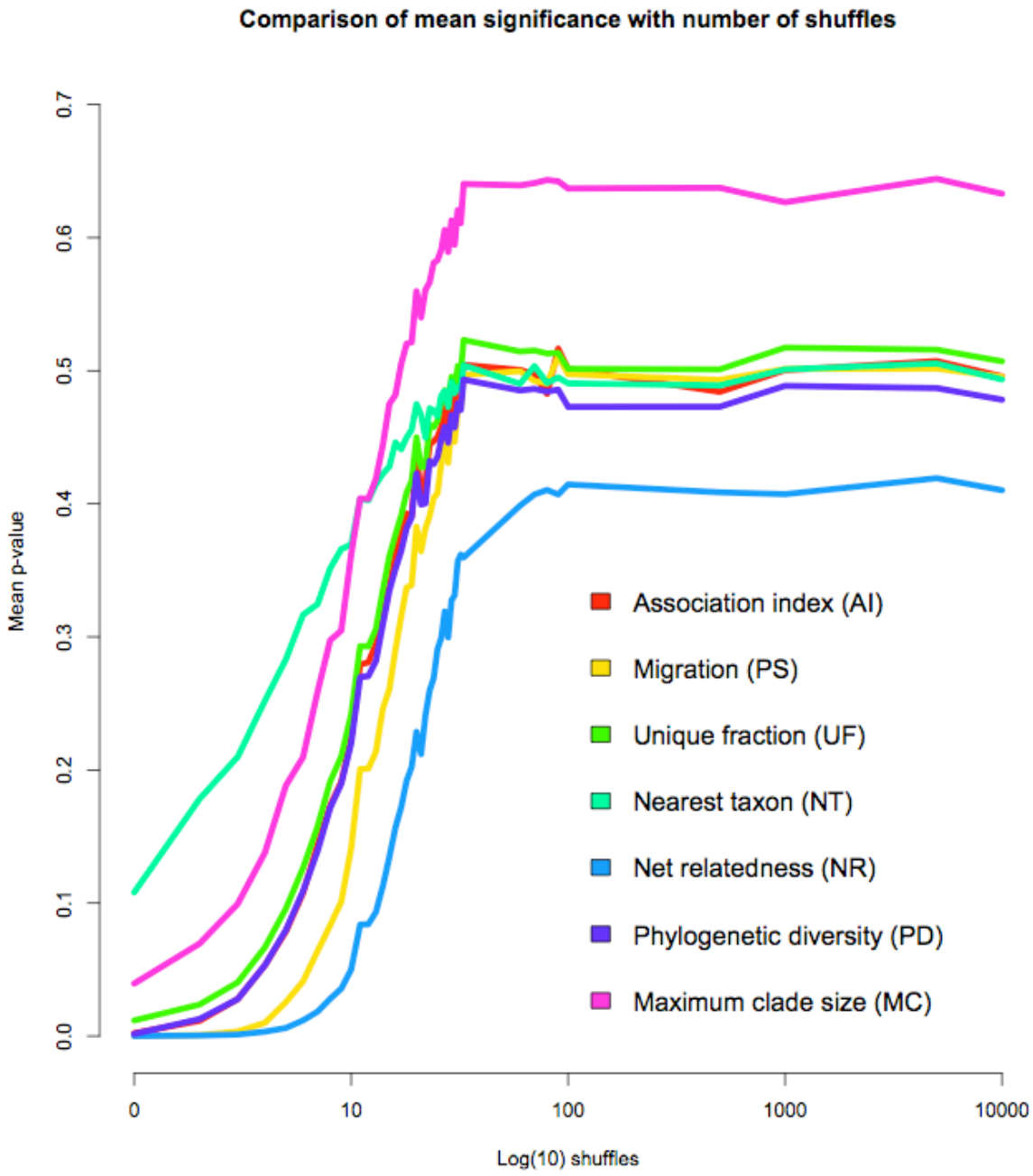


Figure 5.13: Mean significance of observed trait-association values by different statistics with increasing numbers of random taxon trait-value rearrangements (log scale).

5.4.3 Sensitivity of phylogeny-trait association measures to tree shape

The distribution of common tree shape statistics on the set of PSTs used in each simulated data set to test the phylogeny-trait association statistics (n = 897) is shown in Figure 5.14. The nine topologies used to simulate the initial sequence alignments can be discerned as discrete clusters.

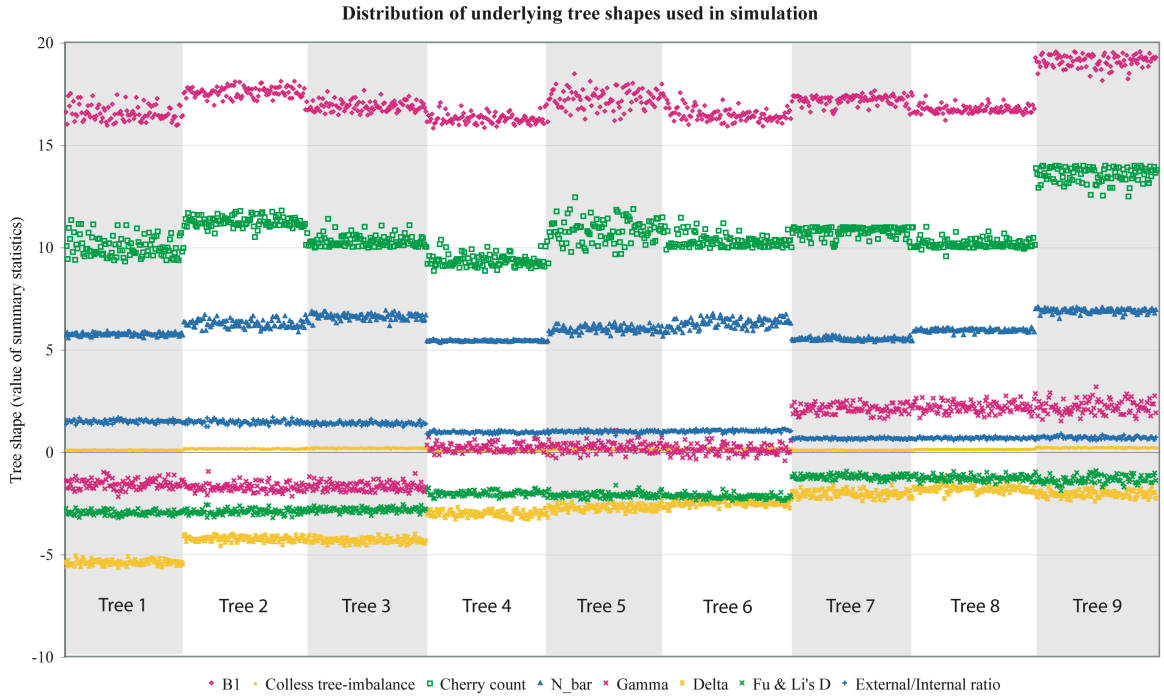


Figure 5.14: Distribution of tree shape statistics of 897 simulated data sets used in this study. Each alignment was simulated from one of nine master topologies picked to give a range of tree topologies typical of human immunodeficiency virus (HIV) evolution. Simulated alignments were analysed in BEAST version 1.4.6 (see Section 5.3.3 for details). Mean tree shape statistics given were calculated from the posterior set of trees (PST) in each analysis using code from the FigTree version 1.1 package (retrieved from http://beast-mcmc.googlecode.com; my implementation is available on request).

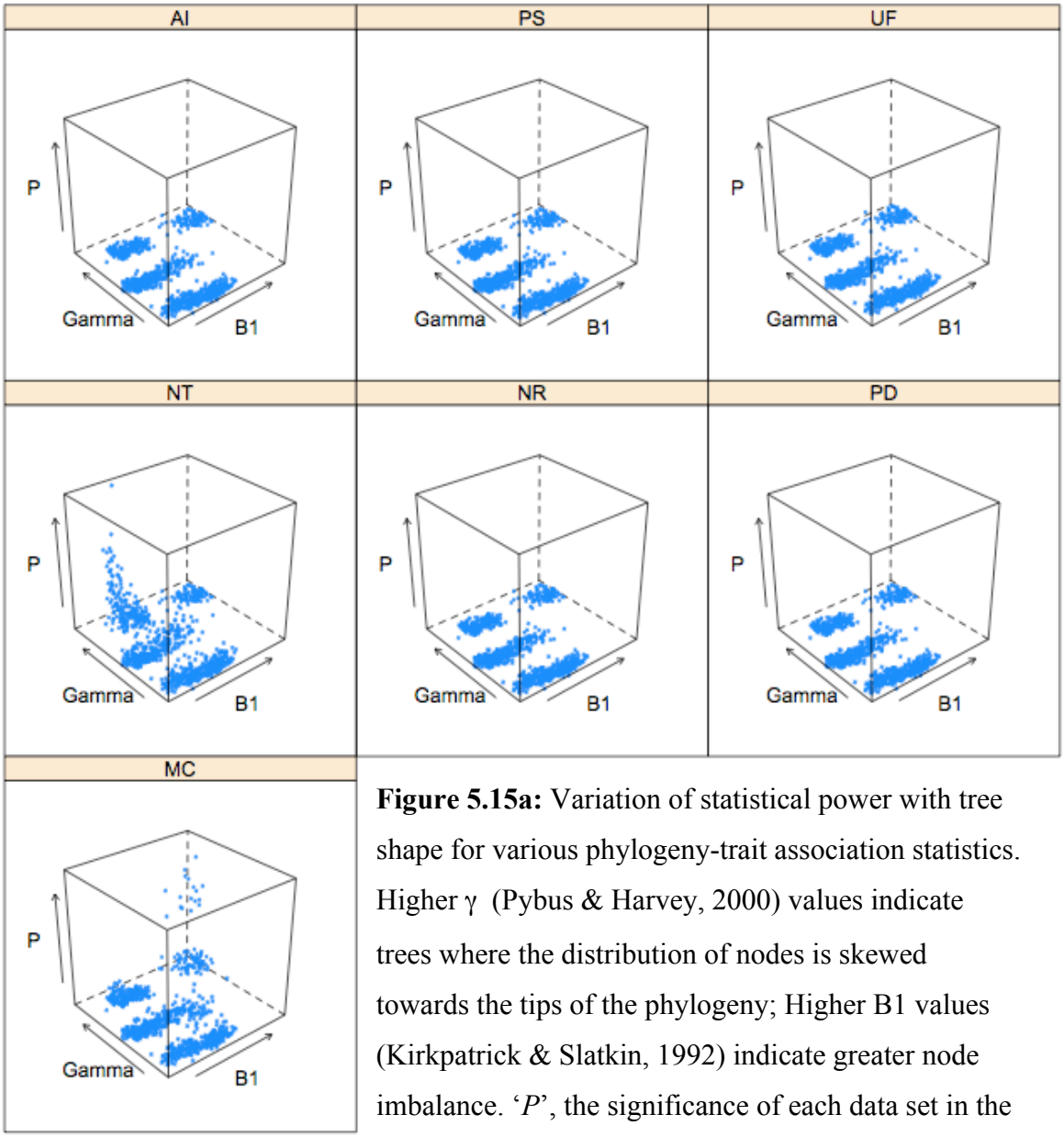


Figure 5.15a: Variation of statistical power with tree shape for various phylogeny-trait association statistics. Higher γ (Pybus & Harvey, 2000) values indicate trees where the distribution of nodes is skewed towards the tips of the phylogeny; Higher B1 values (Kirkpatrick & Slatkin, 1992) indicate greater node imbalance. '*P*', the significance of each data set in the totally associated model.

Figure 5.15a shows the distribution of *p*-values for each phylogeny-trait statistic when applied to data sets with maximal phylogeny-trait association (*i.e.*, no trait shuffles between tips). The majority of statistics show no distinct pattern of failures to reject the null hypothesis ($p > 0.05$) with tree shape, but the MC and NT statistics appear to do so at conditions of high γ values ('comb-like' topologies, with a distribution of nodes pushed towards the tips of the tree) and either high B1 values (strong node imbalance; NT statistic) or low B1 values (balanced trees; MC statistic.) These figures are reproduced in more detail in Figure 5.15b; it can be seen that a large proportion of simulations in these two cases accept

H_0. In fact, under this completely associated simulation, the NT statistic rejected H_0 in 10% of trials while the MC statistic rejected H_0 in 8.5% of trials. It is possible that the discrete nature of these statistics gives rise to this behaviour; none of the other statistics rejected the null hypothesis in any trials under this simulation.

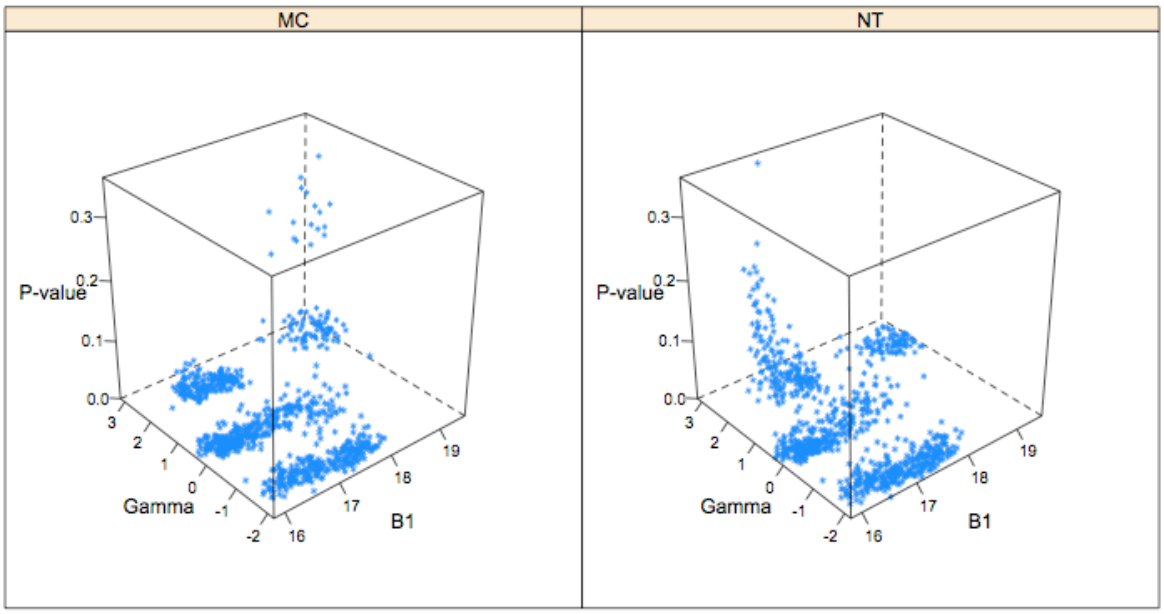


Figure 5.15b: A more detailed look at dependence of power on tree shape in MC and NT statistics. The MC statistic, left, shows weaker power in trees with strong node imbalance (high B1 statistic) and a distribution of nodes that is skewed towards the tips of the tree (high γ). The NT statistic, right, is also weaker in topologies with high γ, but in trees with evenly-balanced nodes.

5.4.4 Compartmentalization in the liver during chronic HCV infection

Sobesky *et al.* (2007) studied compartmentalization between HCV viruses sampled from the peripheral blood and two types of cirrhotic nodules (tumorous and non-tumorous) in seven patients with chronic hepatitis C infection and hepatocellular carcinoma (HCC). 573nt sequences were obtained from the *core* gene by clonal PCR; Patients P1 (n=70) and P7 (n=68) from the original data set were re-analyzed in this study to examine the evidence for compartmentalization with Befi-BaTS (see Section 5.3.4). The Befi-BaTS analysis identified significant compartmentalization by all methods (Table 5.2), except in the MC measurements in Patient 1, where only clades of sequences sampled from tumorous nodules were found to be significantly larger than expected due to chance. I also measured the γ and B1 tree shape statistics in these patients with TreeStat (Table 5.2).

	Patient 1 γ = -2.34, B1 = 35.5			Patient 7 γ = 3.20, B1 = 35.4		
Statistic[1]	Mean posterior estimate	95 % HPD[2] (*lower, upper*)	*P*[3]	Mean posterior estimate	95 % HPD[2] (*lower, upper*)	*P*[3]
AI	2.83	*2.07, 3.58*	**0.000**	0.03	*0.00, 0.09*	**<0.005**
PS	29.72	*25, 34*	**0.000**	6.03	*4, 8*	**<0.005**
UniFrac	0.45	*0.38, 0.52*	**0.010**	0.85	*0.77, 0.92*	**0.010**
NT	442	*373.16, 516.11*	**0.000**	60.18	*45.29, 76.86*	**<0.005**
NR	17330	*14185, 20894*	**0.090**	2324	*1758, 2984*	**<0.005**
PD	1400	*1193, 1631*	**0.000**	290	*226.12, 361.47*	**<0.005**
MC_{N1}	1.57	*1, 2*	0.080	9.96	*10, 10*	**0.010**
MC_{N2}	2.09	*2, 3*	0.190	5.93	*6, 6*	**0.010**
MC_{serum}	4.36	*3, 6*	0.270	31.33	*31, 33*	**0.010**
MC_{tumour}	4.09	*2, 7*	**0.010**	10.85	*6, 15*	**0.010**

Table 5.2: Compartmentalization during hepatitis C virus (HCV) infection; data from Sobesky *et al.*, 2007. [1]Statistics: AI, association index; PS, parsimony score; UF, unique fraction; NT, nearest taxon; NR, net relatedness; PD, phylogenetic diversity; MC statistics, maximum monophyletic clade sizes of: N1, first non-tumorous cirrhotic nodule; N2, second non-tumorous cirrhotic nodule; serum, serum sample; tumour, tumorous cirrhotic nodule. [2]Estimated upper and lower 95% highest posterior densities of each statistic. [3]Significance of observed mean posterior estimate of the statistic.

5.5 Discussion

Empirical data: In their original report, Sobesky *et al.* (2007) visually compared single neighbour-joining (NJ) trees and calculated within- and between-compartment genetic distances. By the visual comparison method, they detected clear compartmentalization in Patient P7 but only limited clustering in Patient P1. They also used Mantell's test (Mantell, 1967) to detect the significance of correlation between pairwise distances and compartment location; again there was significant evidence for compartmentalization in P7 but only for some compartments in P1. The Befi-BaTS analysis conducted here showed significant compartmentalization ($p < 0.05$, all statistics) in P7 and also in P1 ($p < 0.05$, all statistics except MC). Therefore Befi-BaTS not only incorporates phylogenetic error correctly, but also has more power to reject the null hypothesis in empirical data sets.

Performance of phylogeny-trait association statistics: This study shows the importance of rigorous validation in phylogenetic statistics development. The Type I error rates of the MC and NT statistics were correct; however on further inspection, they were shown to be statistically weak; furthermore, their Type II error rate seems to be linked in some way to tree shape – further work is needed to explore this relationship and until that time their behaviour on other topologies may be considered too unpredictable. The NR statistic, though powerful and not sensitive to tree shape, displayed a slightly elevated Type I error rate. It may be that, with further refinement, this will become a valuable statistic but for now its incorrect Type I error means it should be employed with caution. Of the remaining statistics, the AI, PD & UF statistics have very similar Type II error rates, though differing Type I error rates (AI having a slightly high Type I error rate, at 0.051) while the PS statistic is slightly more powerful, but does not include branch length information as PD and UF do.

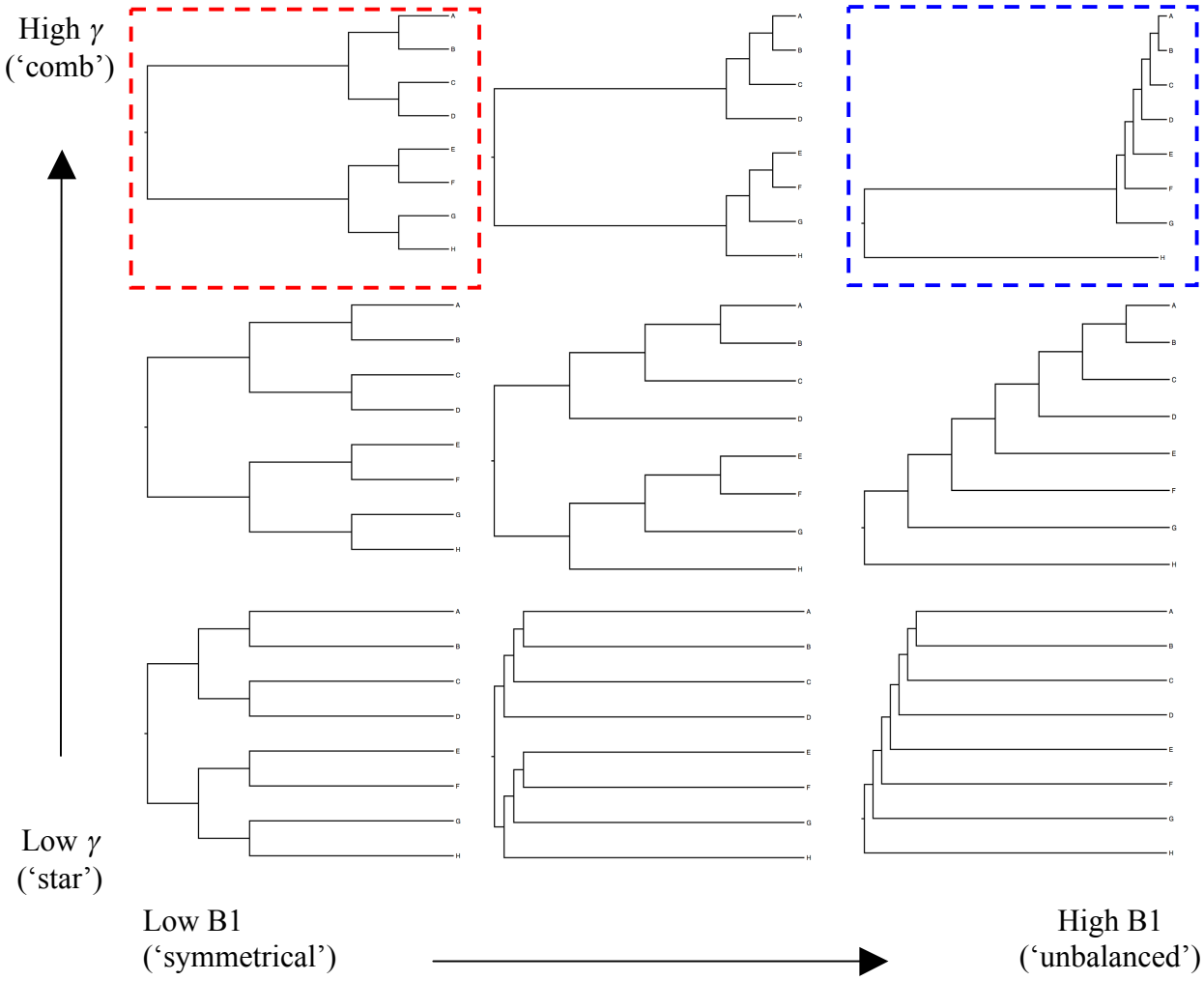


Figure 5.16: Representation of typical tree shapes for certain combinations of γ and B1. The NT statistic exhibited weak power in symmetrical, comb-like trees (red dashed box). The MC statistic exhibited weak power in unbalanced, comb-like trees (blue dashed box).

The statistics' sensitivity to tree shape was also investigated; the MC and NT statistics both appear to suffer from reduced power under certain conditions, illustrated in Figure 5.16. The MC statistic was weak when trees were comb-like (internal nodes distributed toward the tips of the tree) in balanced trees (such as in the top-right hand corner (blue box) of Figure 5.16).

The NT statistic was weak in unbalanced comb-like trees (such as in the top-left corner (red box) of Figure 5.16). What both cases have in common is that in very comb-like trees, internodal distances among the immediate ancestors are often minimal, reflecting low sequence divergence. As a result, reconstructing phylogenetic relationships in these cases may be problematic: single ML trees often represent these relationships as soft polytomies. In a posterior set of trees this will manifest itself as a wider variation in tree branching orders. However, both the MC and NT statistics are most sensitive to changes in branching order near the tips of a phylogeny: the MC statistic because the largest clade monophyletic for a given trait value in a phylogeny rarely extends deeply to the root, as can be verified by comparing the observed MC size with number of tips in total; the NT statistic by implication since it calculates the nearest taxon of the same trait value over all taxa – which will frequently traverse the tree no deeper than the first or second ancestor node.

Where large variance exists this may result in lower observed mean MC clade sizes than in less comb-like trees. Furthermore the observed MC clade sizes may be further lowered since in unbalanced phylogenies monophyletic clades arise under a narrower range of possible trait associations than in balanced phylogenies. To illustrate this point, consider two trees where one, *C* (which might be similar to the tree in the top-left corner of Figure 5.16), is completely symmetrical, and the other, *U*, is unbalanced (similar to the tree in the top-right corner of Figure 5.16). Now suppose we begin with no character traits assigned to any of the tips, and assign a hypothetical ‘white’ trait to four of the tips in such a way as to maximise phylogeny-trait association. However, the first ‘white’ trait must be assigned at random.

It can be seen that the position of the first trait value on *C* is irrelevant; a monophyletic clade of 'white' traits can still be created. However, any monophyletic clade in *U* must include the two uppermost taxa. In other words, for any tree of more than three taxa, more phylogeny trait associations leading to monophyletic clades of size two or larger are possible in balanced trees than in unbalanced trees. The MC statistic therefore suffers from reduced power in unbalanced comb-like trees because observed mean MC clade sizes tend to be smaller, increasing the potential overlap between observed and null distributions.

The NT statistic is expected to correlate with strength of trait-phylogeny association because phylogenetically related taxa should be separated by minimal evolutionary distance. This can usefully be considered here as the sum of the two external branch lengths in question (which will not depend on their phylogenetic proximity) and the internal branch distance separating them, which will depend on their evolutionary relationship. In comb-like trees, the nearest-neighbour distance between two taxa of the same trait value (as calculated in the observed NT size) will be largely determined by their external branch lengths, since, as in the MC statistic, they will rarely be separated by more than a few internal nodes. However, the expected NT distances will vary, depending on the degree of tree imbalance. In symmetrical comb-like trees, the nearest-neighbour distances of any randomly-chosen pair of taxa will vary little; in other words, observed and expected NT values will be similar, since the distribution of possible NT distances is relatively smooth. I therefore suggest that the power of the NT statistic could be improved by considering only internal branch lengths. These results underscore the importance of exploring the effect of likely parameter values on statistical power.

Furthermore, on reflection the distance-based statistics (UF, NT, NR and PD) may generally suffer from another drawback. The null distribution for all these statistics is calculated by random allocation of trait values on the tips of the phylogeny (see Chapter Two, specifically Section 2.3.1). Effectively, this method only randomizes the association of trait values with branching order, not branch length. The null hypothesis is that there is no evolutionary association between taxa with identical trait values; that two taxa are as likely to have the same trait value if they are selected at random or if they share phylogenetic ancestry.

Where shared phylogenetic ancestry is represented by common topology (as in the AI, PS and MC statistics introduced in Chapter Two) it is necessary and sufficient to generate the null distribution through randomizing branch orders since power to reject the null hypothesis arises from lower-than-expected numbers of internal nodes separating associated traits. However, in the case of statistics that incorporate branch length information (as in the UF, PD, NT & NR statistics introduced in this chapter) it may not be sufficient to simply randomize branching order as in Chapter Two to calculate a null distribution. A more appropriate null distribution would randomize both branch order and branch lengths in the tree – Freckleton & Pybus (2006) followed a similar approach to test trait association. Alternatively, a new phylogeny could be generated *de novo*. Pybus and Harvey (2000) used birth-death models to usefully simulate phylogenetic trees; alternatively the coalescent (Kingman 1982a, b) might provide a suitable null model. Clearly further work is needed to establish how the null distribution for distance-based phylogeny-trait association statistics may be most efficiently calculated.

I have developed this technique in order to take advantage of Bayesian MCMC processes that more adequately estimate the true topology of a phylogeny, as they incorporate phylogenetic error in the estimation process through the posterior set of trees. In Chapter Two it was not important to accurately estimate the substitution model and molecular clock model, since the measures of phylogeny-trait association (AI, PS, MC) were purely topological.
However with respect to phylogeny-trait association statistics incorporating branch length information (PD, NT & NR, UF) branch lengths must be more accurately estimated. This presents a challenge since model selection procedures in Bayesian MCMC methods are laborious and in the process of development. That is, although Bayesian MCMC methods explore the parameter space of a given substitution model well, the actual choice of model used may be subject to misspecification (Suchard *et al.*, 2001). Since these measures depend on accurate branch length estimation, misspecification of the substitution model may lead to serious consequences for the accuracy of these statistics.

Accordingly, I suggest that the best available model selection procedures should be followed when these statistics are used to quantify phylogeny-trait association. Furthermore, work needs to be done to quantify the sensitivity of these statistics to substitution model misspecification. More generally, this conclusion (and the result seen in Chapter Three, where a large number of substitution models tested disagreed in predicted substitution rates) strongly suggests that substantial further work is needed to put model selection in Bayesian MCMC phylogenetic analyses on a more rigourously-tested footing, with commonly-accepted standards of model selection.

In conclusion, this study suggests that a combination of PD, UF AI and PS statistics should be used in studies of phylogeny-trait association. These combine correct Type I error rates, reasonable power that is evenly spread across the range of tree shapes tested, and utilize both branching order (topology) and length (in the case of UF and PD) information.

5.6 References

Bhattacharya T, Daniels M, Heckerman D, Foley B, Frahm N, Kadie C, Carlson J, Yusim K, McMahon B, Gaschen B, Mallal S, Mullins JI, Nickle DC, Herbeck J, Rousseau C, Learn GH, Miura T, Brander C, Walker B, Korber B. (2007) Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science* **315**:1583-6.

Carrington, C.V.F., Foster, J.E., Pybus, O.G., Bennett, S.N. & Holmes, E.C. (2005). Invasion and maintenance of Dengue Virus Type 2 and Type 4 in the Americas. *J. Virol.* **79**(23): 14680-14687.

Colless, D.H. (1982) Phylogenetics: the theory and practice of phylogenetic systematics. Part II, pp. 100–104.

Drake, J. W., Charlesworth. B., Charlesworth, D. & Crow, J. F. (1998) Rates of spontaneous mutation. *Genetics* **148**:1667-1686.

Drummond, A. J. & Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**:214-226.

Faith, D.P. (1992) Conservation evaluation and phylogenetic diversity. *Biol. Cons.* **61**:1-10.

Fitch, W.M. (1971b). Toward defining the course of evolution: Minimal change for a specific tree topology. *Syst. Zool.* **20**: 406-416.

Freckleton, R. P. & Harvey, P. H. (2006) Detecting non-brownian trait evolution in adaptive rasiations. *PLoS Biol.* **4**(11):3373.

Fu, Y. X. & Li, W. H. (1993) Statistical tests of neutrality of mutations. *Genetics* **48**:91-103.

Fulcher, J.A., Hwangbo, Y., Zioni, R., Nickle, D., Lin, X., Heath, L., Mullins, J.I., Corey, L. & Zhu, T. (2004). Compartmentalization of Human Immunodeficiency Virus Type 1 between blood monocytes and CD4(+) T cells during infection. *Journal of Virolog*y, **78**(15):7883-7893.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A. & Holmes, E. C. (2004) Uniting the epidemiological and evolutionar dynamics of pathogens. *Science* **303**:327-332.

Holmes, E.C. (2004). The phylogeography of human viruses. *Molecular Ecology* **13**:745-756.

Jenkins, G.M., Rambaut, A., Pybus, O.G. & Holmes, E.C. (2002) Rates of molecular evolution in RNA viruses: A quantitative phylogenetic analysis. J. *Mol. Evol.* **54**:156-165.

Kingman, J. F. C. (1982a) The coalescent. *Stoch. Proc. App.* **13**:235-248.

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

Kirkpatrick and Slatkin (1993). M. Kirkpatrick and M. Slatkin, Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution* **47** :1171–1181.

Komatsu H, Lauer G, Pybus OG, Ouchi K, Wong D, Ward S, Walker B & Klenerman P. (2006). Do antiviral CD8+ T cells select hepatitis C virus escape mutants? Analysis in diverse epitopes targeted by human intrahepatic CD8+ T lymphocytes. *Journal of Viral Hepatitis* **13**:121-30.

Leigh Brown, A.J., Lobidel, D., Wade, C.M., Rebus, S., Philips, A.N., Brettle, R.P., France, A.J., Leen, C.S., McMenamin, J., McMillan, A., Maw, R.D., Mulcahy, F., Robertson, J.R., Sankar, K.N., Scott, G., Wyld, R. & Peutherer, J.F. (1997). The molecular epidemiology of human immunodeficiency virus Type 1 in six cities in Britain and Ireland. *Virology* **235**:166-177.

Lilliefors, H. W. (1969) On the Kolmogorov-Smirnov test for Normality with mean and variance unknown. *J. Am. Stat. Ass.* **62**(318):399-402.

Lozupone, C. & Knight, R. (2005) UniFrac: A new method for comparing microbial communities. *App. & Environ. Microbiol.* **71**(12):8228-8235.

Massey, F. J. (1951) The Kolmogorov-Smirnov test for goodness of fit. *J. Am. Stat. Ass.* **46**(253):68-78.

McKenzie, A., & Steel, M (2000) Distributions of cherries for two models of trees. *Math. Biosci.* **164**: 81–92.

Nakano, T., Lu, L., Liu, P. & Pybus, O.G. (2004). Viral gene sequences reveal the variable history of hepatitis C virus infection among countries. *Journal of Infectious Disease* **190**:1098-1108.

Pillai, S.K., Kosakovsky Pond, S.L., Lui, Y., Good, B.M., Strain , M.C., Ellis, R.J., Letendre, S., Smith, D., Gunthard, H.F., Grant, I., Marcotte, T.D., McCutchan, J.A., Richmann, D. & Wong, K. (2006). Genetic attributes of cerebrospinal fluid-derived HIV-1 *env*. *Brain* **129**: 1872-1883.

Potter, S. J. , Lemey, P., Achaz, G., Chew, C. B., Vandamme, A.-M., Dwyer, D. E. & Saksena, N. K. (2004) HIV-1 compartmentalization in diverse leukocyte populations during antiretroviral therapy. *J. Leukocyte. Biol.* **76**:562-570.

Pybus OG & Harvey PH (2000) Testing macro-evolutionary models using incomplete molecular phylogenies. *Proc Roy Soc B* **267**:2267-2272.

Rambaut, A. & Grassly, N.C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics* **13**(3):235-238.

Rambaut, A. (2001). Phyl-O-Gen. Available at http://evolve.zoo.ox.ac.uk

Salemi, M., Lamers, S.L., Yu, S., de Oliveira, T., Fitch, W.M. & McGrath, M.S. (2005). Phylodynamic analysis of Human Immunodeficiency Virus Type 1 in distinct brain compartments provides a model for the neuropathogenesis of AIDS. *J. Virol* **79**(17): 11343-11352.

Sheridan I, Pybus OG, Holmes EC, Klenerman P. (2004). High resolution phylogenetic analysis of hepatitis C virus adaptation and its relationship to disease progression. *Journal of Virology* **78**:3447-54.

Slatkin, M., & Maddison, W.P. (1989). A cladistic measure of gene flow measured from the phylogenies of alleles. *Genetics* **123**(3):603-613.

Sobesky, R., Feray, C., Rimlinger, F., Derian, N., Dos Santos, A., Roque-Alonso, A.-M., Samuel, D., Bréchot, C. & Thiers, V. (2007) Distinct hepatitis C virus core and F protein quasispecies in tumoral and nontumoral hepatocytes isolated via microdissection. *Hepatology* **46**:1704-1712.

Starkman, S.E., MacDonald, D.M., Lewis, J.C.M., Holmes, E.C. & Simmonds, P. (2003). Geographic and species association of hepatitis B virus genotypes in non-human primates. *Virology* **314**:381-393.

Sullivan, S.T., Mandava, U., Evans-Strickfaden, T. *et al*. (2005). Diversity, divergence, and evolution of cell-free Human Immunodeficiency Virus Type 1 in vaginal secretions and blood of chronically infected women: associations with immune status. *Journal of Virology*, **79** (15): 9799-9809.

Suchard, M.A., Weiss, R.E. & Sinsheimer, J.S. (2001) Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.* **18**:1001:1013.

Wang, T.H., Donaldson, Y.K., Brettle, R.P., Bell, J.E. & Simmonds, P. (2001). Identification of shared populations of Human immunodeficiency Virus Type 1 infecting microglia and tissue macrophages outside the central nervous system. *J. Virol.* **75** (23): 11686-11699.

Webb, C.O. (2000) Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *Am. Nat.* **156**(2):145-155

Webb, C.O., Ackerly, D.D, McPeek, M.A. & Donoghue, M.J. (2002) Phylogenies and community ecology. Annu*. Rev. Ecol. Syst.* **33**:475-505

Chapter Six

Concluding Remarks

6.1 Concluding remarks

This thesis set out to investigate the within-host evolutionary rates and population structure of chronic RNA virus infections, reviewed in Chapter One: in this thesis I have adapted existing methodologies for the analysis of RNA virus evolution and developed new ones. Furthermore, by applying these approaches to sequence data I have been able to investigate within-host rate variation in HCV and cervical compartmentalization in HIV in greater detail than has previously been possible. Taken together, my results suggest a number of further research directions: consequences for models of virus evolution, recommendations for future experiments and reveal a number of opportunities to easily improve existing methodologies, summarized below.

6.1.1 Incorporating phylogenetic uncertainty into measures of phylogeny-trait association

In Chapter Two I explored current approaches to the study of viral compartmentalization, a process with important consequences for viral evolution and epidemiology. I developed a novel framework utilizing the posterior sets of trees (PST) generated by recently popularized Bayesian MCMC methods, such as MRBAYES (Huelsenbeck & Ronquist, 2001) and BEAST (Drummond & Rambaut, 2007). This framework not only correctly incorporates phylogenetic uncertainty into estimates of observed phylogeny-trait association, but also generates correct null distributions, against which the statistical significance of observed associations may be established. This framework implements two existing statistics (AI & PS); I also defined a third (MC). Through simulation I established that the Type I (false positive) error rates at the 5 per cent significance level were correct for these statistics.

Applying the methodology to two published data sets previously analyzed using competing approaches, I was able to demonstrate the value of the PST-based approach, and demonstrated the shortcomings of existing approaches that rely on single-tree phylogenetic reconstruction.

6.1.2 Estimating the rate of HCV evolution within and between hosts.

In Chapter Three, I reported the first study of hepatitis C virus (HCV) to use a coherent relaxed-clock approach on serially-sampled data, with the aim of estimating nucleotide substitution rates from within- and between-host alignments. This analysis determined separate substitution rates, codon rates-ratios and relaxed molecular clock parameters in discrete windows along the genome, but used information from all windows jointly to infer the underlying phylogeny, increasing statistical power. Substitution rates estimated by this method from between-host data agreed with existing literature. However, I found that HCV substitution rates at the within-host level – which are rarely estimated directly but instead interpolated from between-host data sets – vary considerably along the genome, are subject to heterotachy in many genome regions, and are more rapid than previously appreciated; in fact on a par with human immunodeficiency virus (HIV) infection.

Furthermore, the most rapidly-evolving genome regions displayed an excess of substitutions at first- and second-codon positions, indicative of strongly diversifying selection. While positive selection in these regions had been previously reported (Kuntzen *et al.*, 2007; Ray *et al.*, 1999), in Chapter Three I simultaneously compared within- and between-host rates of evolution in HCV, making it clear that the elevation of within-host rates in the *E2* gene

region is more marked than previously appreciated. This chapter also highlighted the importance of correct model selection procedures in Bayesian phylogenetic analyses, and I found that at present, substitution, demographic and clock models in this area are not efficiently and robustly tested and rejected: investigating a wide range of models is highly computationally expensive if the currently-available model selection techniques are followed. The alternative is to either specify models *a priori* or select models for Bayesian MCMC analysis in the maximum likelihood (ML) framework; it is not clear that this approach would not risk model mis-specification.

6.1.3 Detecting cervical compartmentalization in HIV infection

In Chapter Four I was able to apply the methods developed in Chapter Two to a large, newly generated empirical data set of human immunodeficiency virus (HIV) sequences. This study focused on the hypothesis of compartmentalization of the viral population in cervical tissues (significant since the cervix surface is a major source of heterosexual transmission throughout the world: UNAIDS, 2007). Surprisingly, although pervasive and marked HIV compartmentalization has been reported before in the brain and central nervous system (CNS) for this virus (Salemi *et al.*, 2005; Korber *et al.*, 1994), I found evidence of compartmentalization in only a minority of individual patients; existing methods based on single ML phylogenetic trees only found evidence for compartmentalization in three individual data sets (of 65 compared). However, the statistics developed in Chapter Two had greater statistical power to reject the null hypothesis and I was able to definitively identify cases of cervical compartmentalization. This study was the first to do so in a data set of this scale, using a phylogenetic analysis of this complexity.

In addition, it was found that evidence for compartmentalization was stronger for *env* than for *gag*, and that the direction of migration between compartments was predominantly from the blood to the cervix. I argued that compartmentalization in the cervix is maintained despite an obvious physical barrier, and in conclusion I suggested a model for the establishment and maintenance of genetic compartmentalization by tissue type in RNA viruses through either allopatric mechanisms (physical separation) in the CNS or sympatric mechanisms (heterogeneous selection pressures) in the cervix. In support of this model, I noted that evidence for compartmentalization was stronger among *env* data sets (where selection is known to act strongly (Ross & Rodrigo, 2002)) than *gag* data sets. A specific prediction arising from this work is that selection should be more strongly detected in cervical datasets where compartmentalization exists than in compartments of infection in the brain or CNS.

6.1.4 Distance-based measures of phylogeny-trait association and evaluating their statistical power.

Having demonstrated the tractability of the compartmentalization problem on empirical data in Chapter Four, in Chapter Five I sought to expand and extend the framework developed in Chapter Two. I implemented several new statistics that incorporated branch length information as well as topological information. In order to more comprehensively characterize the behaviour of phylogeny-trait statistics, I also measured their Type II error rate (*i.e.* statistical power) and investigated bias arising from different tree topologies through extensive simulation. I was able to verify the largely correct Type I error rate of these statistics, and discovered that they differ greatly in statistical power over some regions of parameter space. Some statistics (*i.e.* AI, NR, PD, UF) performed equally well on the full

range of tree shapes that I simulated. However others (*i.e.* MC, NT) suffered from reduced power in trees where nodes were spread towards the terminal taxa; these statistics typically incorporate topological information from internal nodes near the tips of phylogenetic trees. This study highlights the importance of validation of phylogenetic statistics using appropriate parameters. However, further work is needed to expand the range of parameter space used in simulation. In addition, on reflection, although randomization without replacement represents an appropriate and intuitive null model for trait distributions, the precise properties of the null models used for distance-based phylogeny-trait association statistics may need to be reviewed; a random distribution of branch lengths may be biologically unrealistic and ignores present phylogenetic information.

6.2 Discussion: evolution on separate levels?

Chapter Three showed different rates of HCV nucleotide substitution and evidence for heterotachy among and within hosts. Chapter Four found evidence to suggest that within a host separate populations are maintained by both spatially-isolated processes (similar to allopatric speciation through genetic drift) and selection mediated processes (similar to clinal variation, or sympatric speciation). These two results suggest that processes of evolution within-hosts are not well-described by between-host data sets.

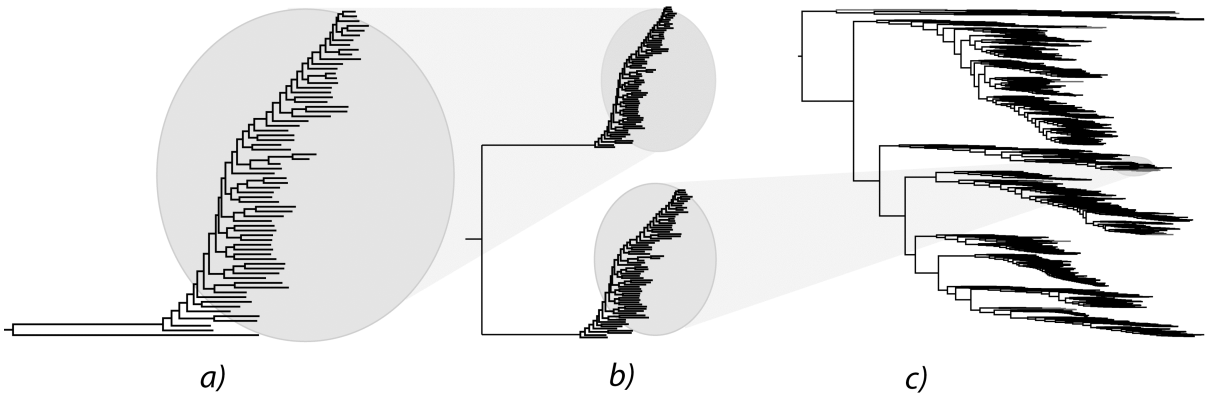


Figure 6.1: Diagrammatic comparison of phylogenies at different spatial scales. *a)* Serially-sampled within-host phylogeny. *b)* Phylogeny of two within-host populations infected from a common source. *c)* A typical between-host phylogeny; individual hosts are rarely sampled more than once and each process within-host evolution is represented as a single branch.

Figure 6.1 illustrates the relationship between within- and between-host data in a series of schematic phylogenies. Figure 6.1a shows a within-host data set. Here selection plays an important role in shaping diversity; consequently lineages are frequently replaced, giving rise to a 'ladder-like' tree (Grenfell *et al*., 2004). Figure 6.1b the phylogeny of two separate viral within-host populations sharing a common source; as long as the transmission bottleneck is low (as is frequently the case; Wahl *et al.*, 2002) they will share single branches to their common ancestor. Finally, Figure 6.1c shows the phylogeny of a number of hosts. Taken together, the internal portion of their population phylogenies collapse and only the connecting branches are seen. These phylogenies are characteristically seen in populations where selection is less evident (Grenfell e*t al.*, 2004). However, note that as we have seen in Chapter Three, our ability to resolve phylogenetic detail is related to the sampling intensity; in particular that to investigate within-host phylodynamics it is necessary to sample within-host data sets at least as thoroughly as between-host data sets.

In this section I consider explanations for, and consequences of, this result for the biological interpretation and methodological practice of viral evolutionary biology.

6.2.1 Evolution within and among hosts.

Most of the viral sequence data sets currently used in phylogenetic studies are comprised of individual sequences collected from separate hosts. Data obtained by direct PCR sequencing represents a majority-rule consensus on the total genotypic diversity in a single individual; even clonal sequences may not give an accurate picture of the viral diversity within an individual, given the rapid nature of virus evolution.

Such data sets represent among-host evolutionary processes and have been used to study rates of substitution (*e.g.* Jenkins *et al.*, 2002), recombination (*e.g.* Worobey & Holmes, 1999), the evolution of drug resistance (*e.g.* Boden *et al.*, 1999), detection of epistasis (Shapiro *et al.*, 2006), determination of the mutational load on viruses (Pybus *et al.*, 2007), and geographic dispersal (*e.g.* Holmes, 2004), often with considerable precision. To apply the above results to within-host processes interpolation is required. The evidence presented here suggests that this may not always be valid. Other studies have also reported complex evolutionary interactions within hosts; separate sub-populations can co-exist (Salemi *et al*., 2005); viral archiving suggests that certain viral lineages can fluctuate in abundance within a single host over time (Nunnari *et al.*, 2005). A proportion of substitutions are lost on transmission from one host to another by wild-type reversion (Herbeck *et al.*, 2006; Friedrich *et al.*, 2004).

As discussed in Chapter One, the range of population genetic forces that affect viral diversity principally act at the within-host level. Substitution and drift occur only among cells; initial effective population sizes are determined by the size of the infectious titre (Rambaut *et al.*, 2004; Domingo *et al*., 1996); recombination can only occur when a single host is dually-infected or super-infected; and since resistance mutations often incur a fitness penalty (Back *et al.*, 1996), it is expected that they will only be acquired or maintained in hosts undergoing treatment, although evidence to the contrary is emerging (Hué *et al*., 2009). The processes of viral evolution at the within-host level therefore profoundly affect the evolution of viruses between hosts; the selective environment that a viral lineage passes through as it expands into successive hosts through repeated cycles of transmission, replication and attack by the man-made interventions or by the immune system is highly heterogeneous over time and space. A critical question is the nature of the transmission effect on viral evolution; how it acts to shape viral diversity among hosts.

Nonetheless, the practice of measuring between-host rates of evolution, and of analyzing evolutionary systems at the between-host level, persists. There are three reasons for this: firstly, there is vastly more sequence data available for between-host data sets (where sampling effort is allocated towards a lower number of sequences from a greater number of individuals; often consensus sequences obtained by direct sequencing) than within-host ones (where effort is allocated to generate a large number of clonal sequences from a small number of individual hosts), for numerous pragmatic reasons. It may be difficult to obtain consent to repeatedly sample the same individual; furthermore direct "population sequencing" by PCR is far easier, cheaper and quicker than clonal sequencing, but perceived to be of equal value for many clinical applications such as drug resistance screening. In fact,

the availability of intensively-sampled within-patient data sets is so poor that the few that exist are known by colloquial names, and have been analyzed many times over. For instance, the HIV data set first reported by Shankarappa *et al*. (1999) has been re-analyzed in numerous studies (*e.g.* Lemey *et al.*, 2007; Watabe *et al.*, 2006; Herbeck *et al.*, 2006; Williamson, 2003). Secondly, the area of viral phylogenetics is comparatively new, and most evolutionary biologists are trained in the context of eukaryote evolution for which the genetic diversity generated *de novo* within an individual is negligible (cancer notwithstanding). Lastly, viral evolution is mainly centred on the 'between-host' model of evolution because by treating within-host evolution as a 'black box', predictions of genuine utility can be made (Rambaut *et al.*, 2004). For instance, although the acquisition of drug resistance mutations happens within individuals, it certainly cannot occur unless those individuals undergo therapy. Similarly, the primary driver of between-clade recombination must be co- or super-infection frequency, since it is the fundamental prerequisite for recombination.

However, while drug resistance, the immune response and recombination are powerful forces that shape evolution of RNA viruses, they are by no means the only ones. I have shown in this thesis that processes of evolution within hosts give rise to variation in sub-population structure, selection and substitution rate that are wide-ranging compared to the between-host processes, but measurable. It is clear that there is a pressing need to expand the state of knowledge of within-host evolution. In particular, the most basic measurement in molecular evolutionary biology, the rate of molecular evolution, is poorly-characterized for within-host data sets. For instance, the study of substitution rates in RNA viruses by Jenkins *et al.* (2002) remains a gold-standard for this problem, even though available data and techniques have moved on. The results presented in Chapter Three are of importance here, since they indicate

that within-host evolution in HCV is considerably faster than the current among-host estimates would suggest. This result echoes the findings from HIV analysis of Lemey *et al*. (2006).

6.2.1 Coupling within- and between-host rates of evolution

Despite the concerns raised above, the rate of substitution continues to be measured at the between-host level. Given that this is so, it is worth investigating the relationship between the within-host substitution rate, μ_W, and the between-host substitution rate, μ_B. In doing so, we are able to incorporate the contribution of the other processes of evolution. In general, I suggest that the relationship between μ_B and μ_W be modelled as:

$$\mu_B = \mu_W + (A - \Omega) \quad \textbf{(6.1)}$$

where A describes those processes that increase divergence observed between hosts, and Ω describes those that limit divergence of the virus as it moves between hosts. Factors in A will include recombination, viral archiving, since 'snapshots' of earlier diversity may be periodically recalled to the circulating population and diversifying selection, whether due to drug resistance or immune escape, where those pressures are constant between hosts. On the other hand, diversifying selection pressures whose directionality varies between hosts (selection for different CTL escape mutation pathways are one example) will lead to reversion of mutations which, along with purifying selection and repeated population bottlenecks, will contribute to Ω and decrease the apparent substitution rate. Conceptualized in this way, it is clear that key parameters of interest, such as the reversion rate to wild-type on transmission, might be estimated.

In Figure 6.2, I further explore the relationship between within- and between-host diversity.

Figure 6.2a shows the relationship between divergence and time within a single host, where the substitution rate is constant. Because the relationship is linear, we may sample the line at any two points in time (*e.g.* points 'I' and 'II') and by comparing the observed sequence divergence with elapsed time, accurately estimate the substitution rate. In Figure 6.2b the substitution rate is not constant over time, but rather accelerates. In this case our estimate of the substitution rate is sensitive to the location of the sampling points. For instance, two samples taken at (I) and (II) would suggest a similar rate to that estimated in Figure 6.2a; while a sample taken at (III) would suggest a less rapid substitution rate.

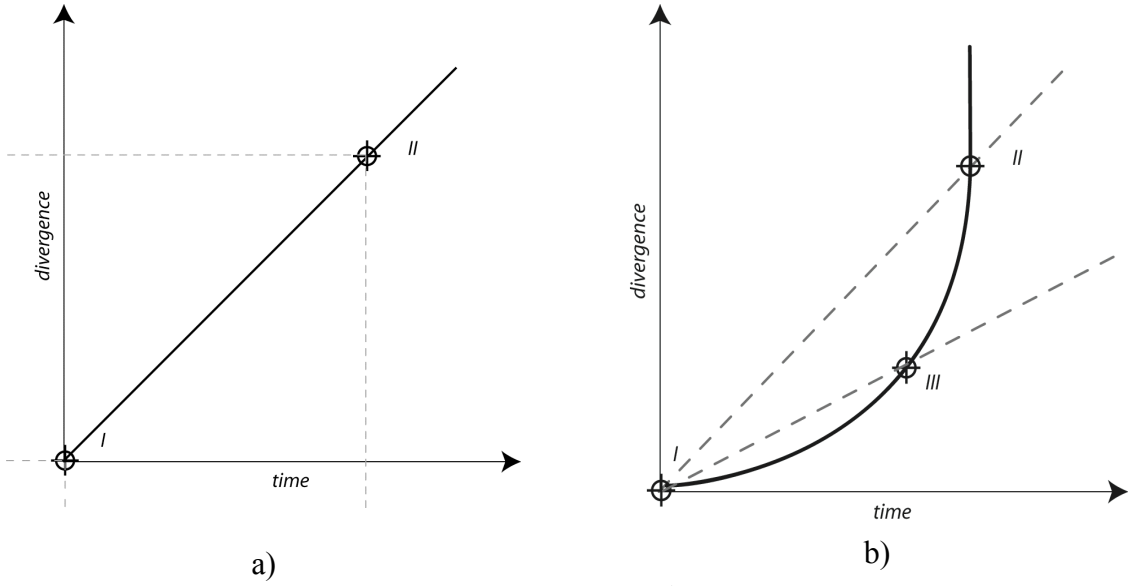


Figure 6.2: The estimated substitution rate may depend on sampling time when the underlying within-host rate is non-linear, as may occur in chronic infections where population sizes fluctuate.

Thus far we have considered sampling the within-host diversity in the context of phylogenetic investigation. However, the series of transmission events that constitute a transmission chain also constitute a series of samples of the within-host evolutionary process. In fact, when we talk of the 'between-host' or 'population' rate of evolution it is typically this process that is meant. I suggest that the between-host rate of evolution, for the reasons outlined above, is sensitive to within-host heterogeneity in the absolute rate of evolution (mediated by strength and direction of selection, population sizes and virulence), to the linearity or otherwise of the within-host rate of evolution (*e.g*, overall nonsynonymous substitution rates might increase if selection relaxes, as in immune collapse, or decrease if purifying selection strengthens, as in drug treatment) and wild-type reversions.

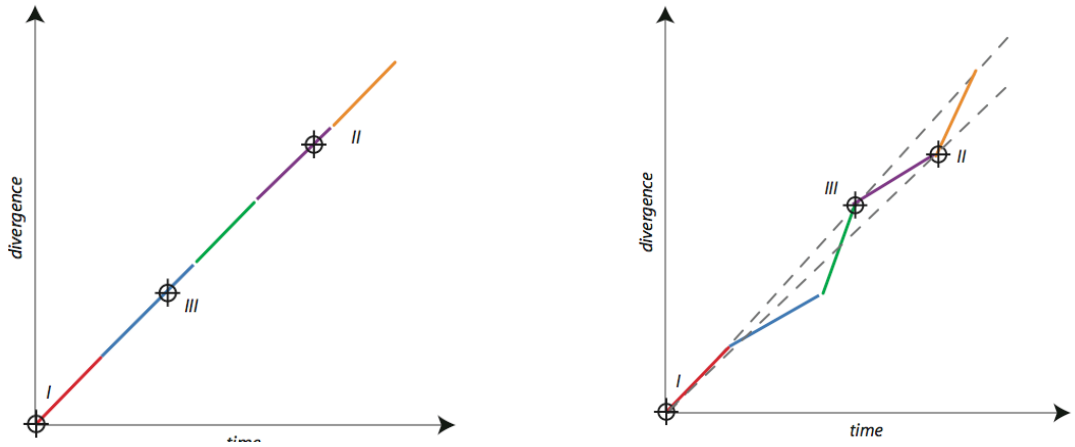
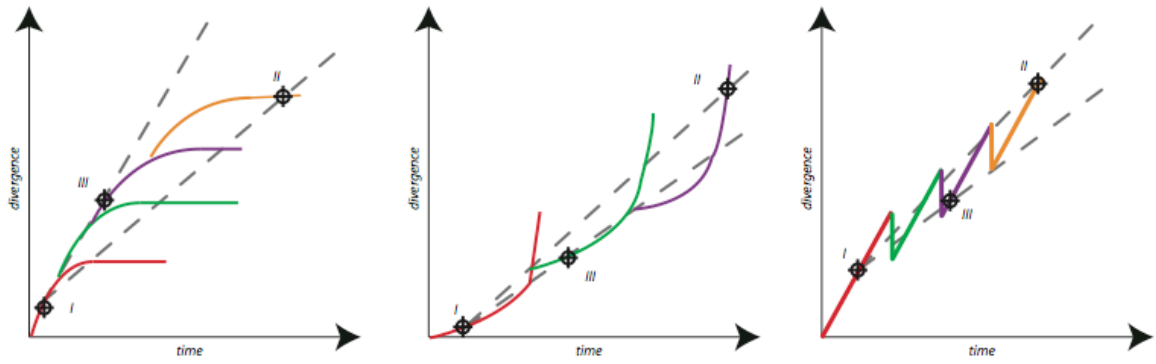


Figure 6.3: The relationship between sampling time and estimated within-host substitution rate in between-host data sets. Coloured lines indicate evolution within different individuals in the transmission chain.

These consequences are explored in Figure 6.3. Each plot is similar to that introduced in Figure 6.2, however successive new within-host populations 'sample' from the time / divergence curve in the previous host, while the sample points represent hypothetical measurements of diversity to estimate the substitution rate. In Figure 6.3a, the substitution

rate is identical between hosts and constant within them. As in Figure 6.2a, we can easily recover the underlying rate regardless of sample location. In Figure 6.3b, the rate within any given host is constant, but the mean within-host rate is heterogeneous. In this case, the substitution rate estimated will depend on sample location; however as long as we sample sufficiently intensively over time we can approximate the within-host rate.



a) *b)* *c)*

Figure 6.4: The between-host substitution rate depends on the point at which each individual infection is sampled or transmitted.

Figure 6.4 demonstrates three situations where this result does not hold true. In Figure 6.4a, the within-host rate of evolution decelerates over time. If the curve is sampled (whether by 'sample' we mean infection or measurement) early in the course of each infection, estimates will tend to be higher; if sampling occurs later in the within-host process, estimates will tend to be lower. The opposite is true of Figure 6.4b, where an accelerating within-host process occurs. Lastly, Figure 6.4c represents the effect of wild-type reversion. For simplicity, the within-host rate of evolution is assumed to be constant over time and homogeneous over hosts. However, it can be seen that regardless of sampling location, the observed between-host rate of evolution is lower than the within-host rate. In fact, by sampling sufficiently soon

after infection, a zero or negative rate of evolution may be inferred. This does not represent 'evolution in reverse' but rather is the result of inappropriate sampling design.

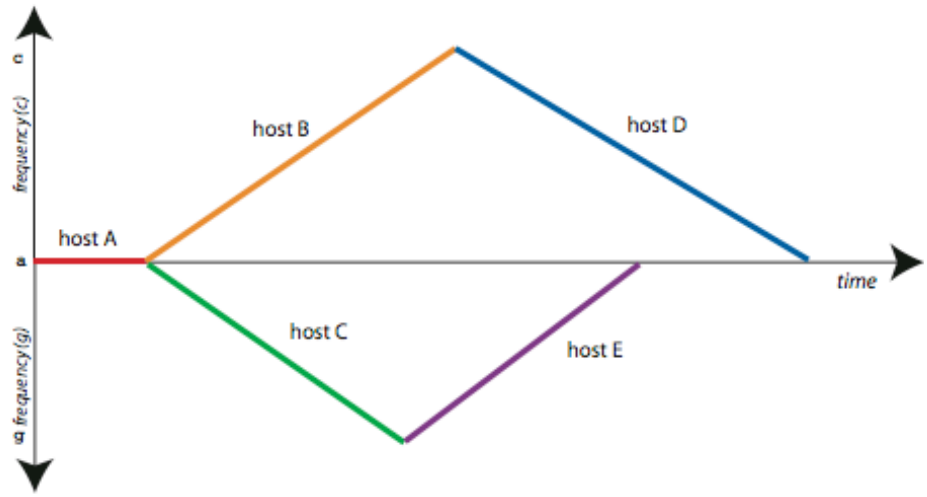


Figure 6.5: Impact of constricted escape pathways.

As noted in Chapter One, it is thought that, due to constraints on the sequence space available to viruses, certain chains of mutations occur sequentially under particular selective contexts, notably evasion of the cytotoxic T lymphocyte response (CTL escape; Karlsson *et al*., 2007; Iversen *et al.*, 2006; Klenerman *et al.*, 2002). In Figure 6.5, I consider the implications of escape pathways for divergence at one single site. In the initial host A, the allele is fixed. On transmission to either of the intermediate hosts B or C, another allele is selected for; as in the examples given above, the distance we measure (or the frequency of the new allele in the population) will be a function of time and the substitution rate. Finally, each of the intermediate hosts infect a new host (D & E), where the original allele is selected for. Here we see a reversion to the allele that was fixed in A. The key result of this system is that even over a considerable number of transmission cycles, where the rate of within-host evolution is rapid and where the allele is strongly selected, it is still possible that no net divergence is

observed between hosts. My second observation is that the strength of phylogenetic bias this represents depends on the rate of evolution at the loci or sites other than the one in question; but that the concept of 'escape mutations' necessarily implies that sequence space is limited to the virus (*i.e.* other sites are conserved). Lastly, the most important prerequisite for this system is heterogeneity between hosts – and host heterogeneity would produce similar effects, albeit less marked, in other strongly selected sites. Considering the rate-heterogeneity I observed in the within-host HCV data set in Chapter Three, it seems reasonable to suggest that host heterogeneity may play an important part in viral evolution.

A further consideration is that a key determinant of diversity will be the strength of these effects in those tissues associated with routes of transmission. For example, I showed in Chapter Four that compartmentalization can occur in cervical tissues in HIV infections and speculated that this compartmentalization was maintained by selection for macrophage-tropic variants. In this case, the strength of selection for CCR5 co-receptor usage, and the effective population size of the cervical compartment, would be relevant factors to be considered (Kemal *et al*, 2003).

6.3 Implications for methodological development

The evolution of serially-sampled within-host viral populations is a relatively new area of study and the analytical tools available are in need of improvement. For instance, although viral sequence data can be analyzed in a matter of days using advanced Bayesian MCMC methods in their simplest form, adequate model selection for Bayesian MCMC methods using Bayes Factors is currently a laborious procedure that can take many days to complete

rigorously. This dichotomy perhaps reflects the historical preoccupation of taxonomists with phylogenetic inference aimed primarily at topology estimation; although the posterior distributions of tree topologies and substitution model parameters are estimated by popular Bayesian MCMC programs, substitution models themselves are usually either selected *a priori* based on previous analyses, or on model test procedures carried out on the same data in a maximum likelihood (ML) context. Choosing the most appropriate model for Bayesian MCMC analysis based on an ML analysis (*e.g.* in MODELTEST; Posada & Crandall, 1998) might be better than an *a priori* approach, but that is by no means clear and potential sources of bias in this procedure are unknown. One useful avenue of study would be to use simulation to compare optimal models for ML and Bayesian MCMC analyses of the same data set. As a minimum, the existing model test procedure for Bayesian MCMC models should be put into a standard framework. This should include the systematic comparison of popular substitution and site-heterogeneity models using Bayes Factors; ideally integrated into existing analysis packages.

Furthermore, the collection and clonal sequencing of viral samples can take a considerable amount of time, even when established organisation, sampling and sequencing procedures exist. As a result, many published experimental designs have yet to catch up with available analysis methodologies. For instance, although serial sampling greatly expands the inferential possibilities of a given data set (*e.g.* the simultaneous estimation of substitution and demographic parameters; Drummond *et al.*, 2002), these data sets are still comparatively rare. Ideally, serially-sampled data sets would include isolates sampled immediately after seroconversion; Seo *et al.* (2002) have suggested a sampling protocol appropriate to viral phylogenies. Taken together, these measures would increase our power to detect variations in

the within-host rate of evolution since it is at this stage that viral populations undergo their most rapid expansion.

6.4 Implications for applied virology

The two findings in this thesis of the greatest clinical relevance are the variation in HCV evolutionary rate within hosts, and the compartmentalization of the cervical HIV population. It seems likely that both are driven by selection, and as such may provide important avenues of future research, aimed either at drug treatment, vaccination, or theraputic interventions. The verification of selection, and measurement of its direction and strength by appropriate[1] phylogenetic and comparative methods therefore presents a pressing challenge. With sufficient serially-sampled within-patient data it may finally be possible to propose specific sites or epitopes that might serve as targets for treatment or vaccine design, a long-held goal of this research (McMichael & Hanke, 2003). The discovery that evolutionary rates vary and are generally lower between hosts is of more general relevance, since estimates of the rate of viral adaptation to intra-host immune pressure or drug treatment may need to be revised; similarly the conclusive evidence produced for compartmentalization in the cervix suggests that more attention may need to be paid to the specific set of phylodynamic and physiological processes affecting this compartment. More generally, it may be that the focus of applied viral phylogenetics should move towards more intensive study of the evolutionary dynamics of transmission routes.

[1] Adequate model selection is important in selection analyses, since model misspecification can lead impact power and precision of synonymous and non-synonymous rate estimates.

6.5 In conclusion

I have been able to present a number of key findings of direct relevance to the field that pose a number of urgent challenges for research in the future. This work has encompassed both computational and biological challenges, which have complemented each other. Despite the ever-increasing complexity of theoretical models of RNA virus evolution and the analytical programs that investigate them, a 'dual-tropic' or multidisciplinary approach to investigator training will continue to yield significant insights into this complex, critical and fascinating area of biology.

6.6 References

Back, N. K. T., Nijhuis, M., Keulen, W., Boucher, C. A. B., Essink, B. B. O., van Kuilenberg, A. B. P., van Gennip, A. H. & Berkhout, B. (1996) Reduced replication of 3TC-resistant HIV-1 variants in primary cells due to a processivity defect of the reverse transcriptase enzyme. *EMBO J.* **15**(15):4040-4049.

Boden, D., Hurley, A., Zhang, L., Cao, Y., Guo, Y., Jones, E., Tsay, J., Ip, J., Farthing, C., Limili, K., Parkin, N. & Markowitz, M. (1999) HIV-1 drug resistance in newly infected individuals. *NEJM* **282**(12):1135-1141.

Cochrane, A., Searle, B., Hardie, A., Robertson, R., Delahooke, T., Cameron, S., Tedder, R. S., Dusheiko, G. M., De Lamballerie, X. & Simmonds, P. (2002) A genetic analysis of hepatitis C virus transmission between injection drug users. *J Infect Dis* **186**, 1212-21

Domingo, E., Escarmís, C., Sevilla, N., Moya, A., Elena, S. F., Quer, J., Novella, I. S., & Holland, J. J. (1996) Basic concepts in RNA virus evolution. *FASEB J.* **10**:859-864.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G. & Solomon, W. (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**:1307-1320.

Drummond, A. J. & Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**:214.

Duffy, S., Shackleton, L.A. & Holmes, E.C. (2008) Rates of evolutionary change in viruses: patterns and determinants. *Nat. Rev. Genet.* **9**:267-276.

Friedrich, T. C., Dodds, E. J., Yant, L. J., Vojnov, L., Rudersorf, R., Cullen, C., Evans, D. T., Desrosiers, R. C., Mothé, B. R., Sidney, J., Sette, A., Kunstman, K., Wolinky, S., Piatak, M., Lifson, J., Hughes, A. L., Wilson, N., O'Connor, D. H. & Watkins, D. I. (2004) Reversion of CTL escape-variant immunodeficienct viruses *in vivo*. *Nat. Med.* **10**(3):275-281.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A. & Holmes, E. C. (2004) Uniting the epidemiological and evolutionar dynamics of pathogens. *Science* **303**:327-332.

Herbeck, J. T., Nickle, D. C., Learn, G. H., Gottlieb, G. S., Curlin, M. E., Heath, L. & Mullins, J. I. (2006) Human immunodeficiency virus Type 1 *env* evolves towards ancestral states upon transmission to a new host. *J. Virol.* **80**(4):1637-1644.

Huelsenbeck, J. & Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**(8):754-755.

Hué, S., Gifford, R. J., Dunn, D., Fernhill, E. & Pillay, D. on Behalf of the UK Collaborative Group on HIV Drug Resistance (2009). Demonstration of sustained drug-resistant human immunodeficiency virus Type-1 lineages circulating among treatment-naïve individuals. *J. Virol* **83**(6):2645-2654.

Holmes, E. C. (2004) The phylogeography of human viruses. *Mol. Ecol.* **13**:745-756.

Iversen, A. K. N., Stewart-jones, G., Learn, G. H., Cristie, N., Sylvester-Hviid, C., Armitage, A. E., Kaul, R., Beattie, T., Lee, J. K., Li, Y., Chotiyarnwong, P., Dong, T., Xu, X., Luscher, M., A., MacDonald, K., Ullum, H., Klarlund-Pedersen, B., Skinhøj, P., Fugger, L., Buus, S., Millins, J. I., Jones, E. Y., van der Merwe, P. A. & McMichael, A. J. (2006) Conflicting selective forces affect T cell receptor contacts in an immunodominant human immunodeficiency virus epitope. *Nat. Immu.* **7**(2):179-189.

Jenkins, G.M., Rambaut, A., Pybus, O.G. & Holmes, E.C. (2002) Rates of molecular evolution in RNA viruses: A quantitative phylogenetic analysis. J. *Mol. Evol.* **54**:156-165.

Joint United Nations Programme on HIV/AIDS (UNAIDS) and World Health Organization (WHO) (2007) *AIDS epidemic update: December 2007*. UNAIDS, Geneva.

Karlsson, A. C., Iversen, A. I., Chapman, J. M., de Oliveira, T., Spotts, G., McMichael, A. J., Davenport, M. P., Hecht, F. M. & Nixon, D. F. (2007) Sequential broadening of CTL responses in early HIV-1 infection is associated with viral escape. *PLoS ONE* **2**(2):e225.

Kemal, K.S., Foley, B., Burger, H. *et al*. (2003). HIV-1 in genital tract and plasma of women: compartmentalization of viral sequences, coreceptor usage, and glycosylation. *Proceedings of the National Academy of Sciences of the United States of America*, **100** (22): 12972-12977.

Klenerman, P., Wu, Y. & Phillips, R. (2002) HIV: Current opinion in escapology. *Curr. Op. Microbiol.* **5**:408-413.

Korber, B. T. M., Kunstman, K. J., Patterson, B. K., Furtado, M., McEvilly, M. M., Levy, R. & Wolinsky, S. M. (1994) Genetic differences between blood- and brain-derived viral sequences from human immunodeficiency virus Type 1-infected patients: Evidence of conserved elements in the V3 region of the envelope protein of brain-derived sequences. *J. Virol.* **68**(11):7467-7481.

Kuntzen, T., Timm, J., Berical, A., L. Lewis-Ximenez, L.L., Jones, A., Nolan, B., Schulze zur Weisch, J., Li, B., Schneidewind, A., Kim, A., Chung, R.T., Lauer, G.M. & Allen, T.M. (2007) Viral sequence evolution in acute HCV infection. *J. Virol.* **81**:11658-11668.

Lemey, P., Rambaut, A. & Pybus, O. G. (2006) HIV evolutionary dynamics among and within hosts. *AIDS Rev.* **8**:125-140.

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

McMichael, A. J. & Hanke, T. (2003) HIV vaccines 1983-2003. *Nat. Med.* **9**(7):874-880.

Mideo, N., Alizon, S. & Day, T. (2008) Linking within- and between-host dynamics in the evolutionary epidemiology of infectious diseases. *Trends Ecol. Evol.* **23**(9):511-517.

Moya, A., Holmes, E. C. & González-Candelas, F. (2004) The population genetics and evolutionary epidemiology of RNA viruses. *Nat. Rev. Microbiol.* **2**:279-287.

Nunnari, G., Sullivan, J., Xu, Y. et al. (2005). HIV type 1 cervicovaginal reservoirs in the era of HAART. *AIDS Research and Human Retroviruses,* **21** (8): 714-718.

Posada, D. & Crandall, K. A. (1998) MODELTEST: testing the models of DNA substitution. *Bioinformatics* **14**(9):817-818.

Power, J.P., Lawlor, E., Davidson, F., Holmes, E.C., Yap, P.L. & Simmonds, P. (1995) Molecular epidemiology of an outbreak of infection with hepatitis C virus in recipients of anti-D immunoglobulin. *The Lancet* **345**(8959):1211-1213.

Pybus, O. G., Rambaut, A., Belshaw, R., Freckleton, R. P., Drummond, A. J. & Holmes, E. C. (2007) Phylogenetic estimation of deleterious mutation load and its contributon to RNA virus evolution. *Mol. Biol. Evol.* **24**:845-852.

Ray, S.C., Wang, Y.-M., Laeyendecker, P., Ticehurst, J. R., Villano, S. A. & Thomas, D. L. (1999) Acute hepatitis C virus structural gene sequences as predictors of persistent viremia: hypervariable region 1 as a decoy. *J. Virol.* **73**(4):2938-2946.

Rambaut, A., Posada, D., Crandall, K.A. & Holmes, E.C. (2004) The causes and consequences of HIV evolution. *Nat. Rev. Genet*. **5**:52-61.

Ross, H. A. & Rodrigo, A. G. (2002) Immune-mediated positive selection drives human immunodeficiency virus Type 1 molecular variation and predicts disease duration. *J. Virol.* **76**(2):11715-11720.

Salemi, M., Lamers, S. L., Yu, S., de Oliveira, T., Fitch, W. M. & McGrath, M. S. (2005) Phylodynamic analysis of human immunodeficiency virus Type 1 in distinct brain compartments provides a model for the neuropathogenesis of AIDS. *J. Virol.* **79**(17):11343-11352.

Seo, T.-K., Thorne, J. L., Hasegawa, M. & Kishino, H. (2002) A viral sampling design for testing the molecular clock and estimating evolutionary dates and divergence times. *Bioinformatics* **18**(1):115-123.

Shankarappa, R., Margolick, J.B., Gnage, S.J., Rodrigo, A.G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C.R., Learn, G.H., He, X., Huang, X.L. & Mullins, J.I. (1999) Constant viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* **73**:10489-10582.

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

Wahl, L. M., Gerrish, P. J. & Saika-Voivod, I. (2002) Evaluating the impact of experimental bottlenecks in experimental evolution. *Genetics* **162**:961-971.

Watabe, T., Kishino, H., Okuhara, Y. & Kitazoe, Y. (2006) Fold recognition of the human immunodeficiency virus Type 1 V3 loop and flexibility of its crown structure during the course of adaptation to a host. *Genetics* **172**(3):1385-1396.

Williamson, S. (2003) Adaptation in the *env* gene of HIV-1 and evolutionary theories of disease progress. *Mol. Biol. Evol.* **20**(8):1318-1325.

Worobey, M. & Holmes, E. C. (1999) Evolutionary aspects of recombination in RNA viruses. *J. Gen. Virol.* **80**:2535-2543.

Appendix One

Shannon Heterogeneity in Alignments Tool

The tool provides a convenient graphical user interface for the calculation of genetic diversity by the Shannon information entropy, allelic heterozygosity, consensus-frequency and Hamming distance methods. This tool is publicly available and was used in Chapter Three and Appendix Two.

Shannon Heterogeneity In Alignments tool

Version 1.1 Manual

Joe Parker
Kitson Consulting
87 Clevedon Road
Bristol
BS8 3UL
United Kingdom

e: joe@kitson-consulting.co.uk
w: phylo.lonelyjoeparker.com
+44 (0)1865 281 987 (tel)
+44 (0)1865 271 249 (fax)

Contents

Introduction

Welcome to the documentation for this version of the Shannon tool – the first public release of this software. I hope you find the package useful but it is still in the early stages of development. As such please let me know if you find any bugs or have suggestions for improvements. This package has undergone only limited testing so you use it at your own risk – see the disclaimer below.

License & Disclaimer

License

This software is supplied under the GNU Lesser General Public License, Version 3. This is an open-source software liscence, and others are authorised and encouraged to examine and modify code if they see fit, as long as the contribution of previous workers is recognised. For more details see
http://www.gnu.org/licenses/lgpl.html

Disclaimer

No guarantee of the **functionality** of this software, or of the **accuracy of results** obtained using it is made, expressed or implied. The programmers, authors and editors of this documentation and the institutions they represent **will not be held responsible** for any errors of analysis, damage to software or hardware, or other losses incurred as a result of using this programme.

What is it?

This software aims to provide a quick and easy method by which the amount of amino-acid or nucleotide heterogeneity within an alignment may be quantified and compared between alignments.

What can it do?

A number of heterogeneity measures are implemented in this package. Some are scored sitewise along an alignment, giving a position-by-position score of alignment heterogeneity. Others give a summary (Hamming Distance) of heterogeneity in the alignment as a whole that can be compared for similar alignments.

System requirements

Java

Shannon is written and compiled for Java 1.5.0, ("J2SE 5.0"). You will need a computer capable of running this version of Java or higher. For most platforms it is sufficient to download the required version of Java directly from java.sun.com Mac OS X users should note however that on versions 10.4.5 and lower the procedure for upgrading to Java 1.5.0 (from 1.4.2, the default on these systems) differs. They should consult http://www.apple.com/downloads/macosx/apple/macosx_updates/index_abc.html for further instructions.
If unsure, typing '`java -version`' from a Terminal session will tell you which version of Java is currently used by the operating system. Max OS X 10.5.x ('Leopard') users are lucky – Java 1.5 is installed on these systems as standard!

Hardware

We have not identified any specific minimum hardware requirements; these in any case scale with the number of taxa and the length of the sequences. Generally speaking, at least 256 Mb of system memory (physical RAM, not virtual memory or swap file cache) should be available for each separate instance of the program that is running. Note that in some rare cases this may not be sufficient and users will need to increase the amount of memory available to the Java Virtual Machine (JVM) using the `–Xms` command; for more information type '`man java`' from the command-line or see http://edocs.bea.com/wls/docs70/perform/JVMTuning.html

Important: a 'progress bar' for long operations is not currently implemented. This means that sometimes the package may appear to 'hang' for periods while, for instance, calculating pairwise Hamming distances in large alignments. Please be patient.

Installation

Given the correct JVM (1.5.0 or higher) is available, installation of is simple. This package contains two files: a Macintosh application folder for Mac OS X users, and a java ".jar" file for all other operating systems.

Installation – Mac OS X

Simply drag the 'Shannon Heterogeneity In Alignments v1.1' application into your Applications folder, or wherever you keep phylogenetic software on your machine.

Installation – Other platforms

Drag or copy and paste the .jar file to a location on your hard drive. If you normally need to run Java applications from the command-line, make sure you make a note of the file path.

Usage: input file requirements

Alignment assumptions

Sequences must be aligned and of the same length, though gaps are allowed. Gaps are treated as informative. IUPAC ambiguity codes are included in the Hamming distance calculation but disregarded in the sitewise heterogeneity measures' calculation.

Format

Input files must be in fasta format. Users are discouraged from using non-standard characters, e.g. those other than alphanumerics (a-z, A-Z, 0-9) and some punctuation characters, such as [',', '.', '-', '_']. This is because the behaviour of some of these characters can be hard to predict.

Usage: running an analysis

Mac OS X

Double-click on the application.

Other platforms

Depending on your system settings, you may be able to double-click on the .jar file to launch Java and load the application (you may be prompted to verify that you wish to launch Java.)
If this doesn't work you will need to launch the application from your command-line (also variously referred to as 'command-prompt', 'MS-DOS', 'terminal', or 'console' depending on the platform.) Open a command-line window and navigate to the folder where the application is installed. Then type:

```
java –jar Shannon_v1.1_sealed.jar
```

The application should then launch. Note that since this method opens the Shannon package as a 'child' of the command-line window you have opened, closing the command-line window will on some platforms also exit Shannon, and any analyses you have generated will be lost (input alignments should be unaffected though.)

Usage: interpreting analyses & general operations

The heterogeneity measures

Four measures are implemented:

- Shannon information entropy[1] (sitewise by position)
- 1-(consensus frequency)[2] (sitewise by position)
- 1-(heterozygosity index H)[3] (sitewise by position)
- Hamming Distance[4] (summed pairwise across alignment)

The Shannon, consensus and H index scores are also summed across the alignment, but not pairwise. For all these measures, higher scores reflect more heterogeneity; the Shannon entropy naturally scales with the number of sequences in the alignment so as a crude normalization to compare alignments we also calculate the 'normalised' Shannon entropy as the mean entropy per sequence. We have not yet ascertained whether there is a more theoretically sound way to compare entropies in alignments of differing sizes.

Loading an alignment

To load a new alignment, choose 'Add alignment' from the File menu. Shannon will calculate sitewise and summary heterogeneity scores for this alignment and after a brief delay update the output.
On some operating systems the graphs may only refresh when the mouse is moved over them, or need the alignment to be loaded a second time.

---

[1] This is the Shannon Informtion entropy, *H(x)* as introduced by Shannon (1948.) It is calculated per position as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_m p(x_i)$$

Where $p(x_i)$ = observed probability (frequency) of amino acid or nucleotide *x* for all *i* possible amino acids or nucleotides. Note firstly that absent amino acids or nucleotides will be included in the equation (though with an entropy of zero) and secondly that the base of the logarithm taken is *m*, the number of possible 'information states.' For nucleotide data this is taken to be 5 (a,c,g,t plus gaps) and for universal amino acid data this is taken to be 23 (20 amino acids plus start, stop and gaps.)
Shannon, C. E. (1948) "A Mathematical Theory of Communication", *Bell Syst. Tech. J.* (**27**):379–423, 623-656

[2] Simply the frequency of the most common nucleotide or amino acid at that position; an alignment with {"a", "a", "c", "g", "t"} at a particular position for instance has a consensus frequency of 0.4 ("a" is the most common nucleotide observed at 2/5 nucleotides.)

[3] The Heterozygosity index, long used in genetics to quantify the allelic diversity, *h*, in a population. This is given by:

$$h = 1 - \sum_{i=1}^{m} x_i^{\,2}$$

Where x_i^2 is the frequency of each observed amino acid or nucleotide, squared.

[4] Hamming distance is the mean of all pairwise distances in an alignmnent, where the distance of any pair of sequences is defined as the number of positions at which they have differing amino acids or nucleotides.

Output panes

The main application window is divided into two panes – the Graphical View and the Data View. The Graphical View provides a quick visual overview of the alignment heterogeneity as it relates to other previously-loaded alignments analyzed in the same session, while the Data View returns the numerical heterogeneity information for the most recently-loaded alignment.

Data View

The Data View output contains heterogeneity information from the most recently-loaded alignment. You should be able to select, copy and paste this successfully to a variety of other applications, such as Microsoft Excel or Minitab.

Graphical View

This shows the value of the Shannon, consensus, and h index measures at each position in the alignment for all traces alignments loaded since the Shannon tool was started (or since the last 'clear all' command), as well as a graph (‘Compare measures’) summarising the relative values of each index in the last alignment that has been successfully loaded. These graphs copy straight to clipboard in some applicaions; others will need a third-party screen capture program (such as 'File>Grab>Selection...' in Preview on Mac OS.)

Clearing the output

Selecting ‘Clear all’ from the File menu removes all traces from all graph plots (note that on some operating systems axis labels may remain, and the graph axes may not re-size.)

Help

As well as this documentation a summary live help page is available through the ‘Help’ menu.

Quitting

To exit the application, choose ‘quit’ from the File menu.

Usage: caveats and warnings

Computational constraints:

Users should be aware that although this package has received limited testing on our development machines, this is the first public release of the package. As such, the real-world performance of the core packages (which in any case are highly dependent on hardware architecture, and software architecture to a degree) is unknown at this point; in fact, we would appreciate any feedback concerning performance.

As a guide, most alignments should load in a few minutes on a 1.25 GHz Apple PowerPC machine under Java 1.5.0 / Mac OS 10.4.10.

As usual, physical factors increasing compute time performance will include:

- Slower system CPUs
- Slower system bus speeds
- Slower system RAM access

In particular, because Shannon uses a lot of
memory at present heavy reliance on virtual memory coupled with a slow hard drive is likely to adversely affect performance)

Problem parameters that will slow the analysis include:

- Number of taxa - increasing numbers of taxa will increase compute time and memory usage. This particularly affects calculation of the Hamming distance since this is a pairwise measure – the number of possible pairwise comparisons rises much faster than the number of sequences.
- Sequence length – long sequences will take more time.
- Amino acid alignments will take longer to compute than nucleotide ones, since more potential residue probabilities (23 versus 5) must be evaluated (this is unlikely to be a major constraint)

Biological constraints

We have developed this method to analyse heterogeneity information so that researchers can quickly and simply evaluate the relative variability in an alignment as an exploratory analysis, perhaps to identify highly variable regions of a gene, or to compare the variability of a locus in separate patients' viral populations where the same or similar numbers of viral samples have been isolated.

This package is not a substitute for a proper phylogenetic analysis of selection or distance – in particular users should bear in mind that these methods all treat individual sites and sequences as independent pieces of information in the statistical sense. This assumption is violated both along sequences (by epistasis or secondary or tertiary structural constraints may operate) and between sequences (lack of independence due to shared phylogenetic ancestry – see Harvey & Pagel, 1991.)

FAQ

Note: this is just a preliminary list of FAQs; for any persistent problems, or if you have any other comments, please contact the author.

Q: Why can't I open the package?
A: Check you have the appropriate version of Java (1.5.0 or higher) installed. If you are having persistent problems contact the author.

Q: Why can't I load my alignments?
A: Try running the sample alignment included in this package. If the sample files don't run either you may well have the wrong version of java installed; check and install the correct version. If the example alignments load but your data does not, it is likely you have parsed the .fasta files incorrectly. If you are editing your alignments by hand, check they load into a sequence editor program such as BioEdit, Se-Al or Geneious and then try re-exporting the alignment as a .fasta file from there. If problems persist, contact the author.

Q: Why do my alignments take a long time to load / the program crashes when I load my sequences?
A: Your alignment is probably too big. You can try increasing the memory allocated to Shannon (see below) but if this does not solve the problem you may need to analyse a subset of the alignment. Hopefully these issues will be addressed in a future release.

Q: Why do I get an error message saying I've run out of memory?
A: Typically this will manifest itself with a message such as "`Exception in thread 'main': java.lang.OutOfMemoryError: java heap space`" but may also take the form of hangs or crashes.
This error arises then the JVM doesn't have enough system memory ('RAM') available to hold all the data it needs to. Unfortunately increasing your computer's virtual memory allocation will not solve this problem. You can try increasing the default amount of RAM allocated to the JVM with the '`-Xms`' command (see the 'System Requirements > Hardware' section of this documentation for details.)

Q: Why are there three separate sitewise statistics? Which is best?
A: Each statistic is sensitive to slightly different types of diversity. For instance, the consensus frequency measure doesn't pick up on the diversity of non-consensus amino acids or nucleotides – an alignment with sequences displaying bases with {"a", "a", "a", "a", "c", "g", "t"} at a particular position has the same consensus frequency as an alignment displaying bases {"a", "a", "a", "a", "t", "t", "t"}, but different Shannon and heterozygosity index scores.
As to which is 'best' work is ongoing to evaluate the utility of each statistic. For now we suggest users compare all statistics and inspect their alignments where they significantly diverge.

Contact

The author of this documentation and software is Joe Parker. You can contact him at:

Joe Parker
Viral Evolution Group
Department of Zoology, University of Oxford
South Parks Road
OX1 3PS
United Kingdom

Or by email: joe.parker@zoo.ox.ac.uk

Appendix Two:

Detailed characterisation of the of the HCV genotype-3a envelope 2 protein in hepatitis C virus reveals two novel hypervariable regions under selection pressure early in acute infection

NOT AVAILABLE ONLINE DUE TO COPYRIGHT RESTRICTIONS:

Published as:

Humphreys I, Fleming V, Fabris P, Parker J, Schulenberg B, Brown A, Demetriou C, Gaudieri S, Pfafferott K, Lucas M, Collier J, Huang KH, Pybus OG, Klenerman P, Barnes E. (2009) Detailed characterisation of the of the HCV genotype-3a envelope 2 protein in hepatitis C virus reveals two novel hypervariable regions under selection pressure early in acute infection. *J Virol* **83**(22):11456-66.

I wrote the SHiAT tool used to quantify genetic diversity, advised on the selection analysis and provided editorial assistance.

Appendix Three

Estimating the date of origin of an HIV-1 circulating recombinant form

NOT AVAILABLE ONLINE DUE TO COPYRIGHT RESTRICTIONS:

Published as:

Tee, K. K., Pybus, O. G. Parker, J. Ng, K. P. Kamarulzaman, A. & Takebe, Y. (2009) Estimating the date of origin of an HIV-1 circulating recombinant form. *Virology*. **387**(1):229-34.

I contributed the phylogenetic analysis.

Appendix Four

BaTS – Bayesian Tip-association Significance testing

The software package implements a number of phylogeny-trait association statistics in a Bayesian framework as introduced in Chapter Two.

An executable jarfile is available from evolve.zoo.ox.ac.uk and lonelyjoeparker.com

Full listing available on request.

BaTS – Bayesian Tip-association Significance testing

Joe Parker, Viral Evolution Group, Department of Zoology, University of Oxford.

Version 1.0 Documentation

Introduction

Welcome to the documentation for this version of BaTS, version 1.0. This is the first publically-available version to be released. I hope you find BaTS accessible and of use to you in your research.

BaTS was essentially concieved in response to a specific problem encountered in my own studies and although a number of changes have been incorporated in this (beta) release that allow a wider range of problems to be addressed, it remains a fairly inflexible tool, both in terms of technical requirements and logical problems that can be solved. Of course, over time, other researchers may yet find BaTS useful in situations as-yet unthought of by us.

I would gratefully like to acknowledge Oli Pybus and Andrew Rambaut, my supervisors, whose direction and input brought BaTS to this point, and Aris Katzourakis, Rob Belshaw, Philippe Lemey, Simon Ho and Beth Shapiro for adivce and help.

(For the specific Bayesian approach to quantifying phylogeny-trait associations, as well as an exploration of the three statistics and discussion of their merits, users are encouraged to read the Parker et al. (*2008) which should also be used as the preferred reference when citing* BaTS*.)*

Licence & Disclaimer

Licence

BaTS is supplied under the GNU Lesser General Public Licence, Version 1.3. This is an open-source software licence, and others are authorised and encouraged to examine and modify code if they see fit, *as long as* the contribution of previous workers is recognised. For more details see
http://en.wikipedia.org/wiki/GNU_Lesser_General_Public_License

Disclaimer

No guarantee of the **functionality** of this software, or of the **accuracy of results** obtained using it is made, expressed or implied. The programmers, authors and editors of this documentation and the institutions they represent **will not be held responsible** for any errors of analysis, damage to software or hardware, or other losses incurred as a result of using this programme.

What is BaTS?

This software aims to provide a method by which the degree to which phenotypic traits seen in a phylogeny are associated with ancestry are correlated. In other words, where a set of character states are seen on the tip of a phylogenetic tree, is any given taxon more likely to share a character state with a sister taxon than we would expect due to chance?

This problem has been posed in a variety of contexts over the last three decades, particularly molecular epidemiology and phylogeography. A number of approaches have been developed over the years, of which the method of Slatkin & Maddison (1989) is perhaps the best known.

BaTS uses two established statistics (the Association Index, AI, and Fitch parsimony score, PS) as well as a third statistic (maximum exclusive single-state clade size, MC) introduced by us in the BaTS citation, where the merits of each of these are discussed. What sets BaTS aside from previous methods, however, is that we incorporate uncertainty arising from phylogenetic error into the analysis through a Bayesian framework. While other many other methods obtain a null distribution for significance testing through tip character randomization, they rely on a single tree upon which phylogeny-trait association is measured for any observed or expected set of tip characters.

To improve on this basic approach we use posterior sets of trees (PSTs), obtained through earlier Bayesian MCMC analysis of the data, that integrate over all likely phylogenies and incorporate the phylogenetic uncertainty arising from the data. Although a Bayesian MCMC analysis is therefore a precondition to using BaTS, we do not feel that this is likely to deter potential users since these analyses are increasingly common.

What can it do?

BaTS can estimate phylogeny-trait associations for *n* different states of a single character at a time using the AI, PS and MC statistics, and provide 95% confidence intervals and significance estimation for these. BaTS is also capable of performing batch analyses of large numbers of datasets.

System requirements

Java

`BaTS` is written and compiled for Java 1.5.0, ("J2SE 5.0"). You will need a computer capable of running this version of Java or higher. For most platforms it is sufficient to download the required version of Java directly from java.sun.com Mac OS X users should note however that on versions 10.4.5 and lower the procedure for upgrading to Java 1.5.0 (from 1.4.2, the default on these systems) differs. They should consult http://www.apple.com/downloads/macosx/apple/macosx_updates/index_abc.html for further instructions. If unsure, typing '`java -version`' from a Terminal session will tell you which version of Java is currently used by the operating system.

Hardware

We have not identified any specific minimum hardware requirements; these in any case scale with the number of taxa in the tree, number of possible states observed and number of null sets used to form the null distribution. Generally speaking, at least 256 Mb of system memory (physical RAM, not virtual memory or swap file cache) should be available for each separate instance of `BaTS` running. Note that in some cases this may not be sufficient and users will need to increase the amount of memory available to the Java Virtual Machine (JVM) using the `–Xms` command; for more information type '`man java`' from the command-line or see http://edocs.bea.com/wls/docs70/perform/JVMTuning.html

Installing BaTS

Given the correct JVM (1.5.0 or higher) is available, installation of `BaTS` is simple. The complete `BaTS` package is hosted at evolve.zoo.ox.ac.uk/software for download as an archived jarfile. Simply download the jarfile to some memorable location on your hard drive.

Using BaTS: input file requirements

Preconditions

Because BaTS uses the PST from a Bayesian MCMC analysis to integrate over all credible trees to account for error in the phylogenetic signal, users must first use an appropriate package to produce a PST. This is a single treefile containing many trees from the posterior set, weighted by the MCMC sampler so that the most likely topologies are sampled more often. MrBayes (Huelsenbeck & Ronquist, 2001; mrbayes.csit.fsu.edu) and BEAST (Drummond & Rambaut, 2003; evolve.zoo.ox.ac.uk/software) are ideal packages available for this task, and produce these treefiles automatically.

Burnin period

Before they begin to efficiently sample from the posterior likelihood distribution of interest, MCMC samplers typically require an initial period of optimisation during which they arrive in the vicinity of highest-likelihood and tune weighting parameters, etc. During this initial 'burnin' period the posterior likelihood fluctuates wildly; once the lieklihood becomes more stable the MCMC chain can be said to be sampling efficiently. It is common to discard the initial burnin, but currently no burnin is automatically discarded from treefiles in BaTS. Users **must** therefore decide on an appropriate burnin period using external data analysis software such as Tracer (evolve.zoo.ox.ac.uk/software) and remove these trees from the start of their treefile accordingly.

Format

Input files for BaTS follow the popular NEXUS file format, with a small modification: Instead of the normal '`taxa`' and '`translate`' blocks, a single '`states`' block is present. The formatting for this is shown below. It is currently necessary to format these by hand; we are working on a graphical interface to parse these input files, please check for updates at evolve.zoo.ox.ac.uk/software

Typical NEXUS format:

```
#NEXUS

Begin taxa;
    Dimensions ntax=8;
    Taxlabels
        HIV_env_JP2000a
        HIV_env_JP2000b
        HIV_env_JP2001
        HIV_env_JP2002
        HIV_env_JP2003
        HIV_env_JP2003b
        HIV_env_JP2005
        HIV_env_JP2005b
        ;
End;

Begin trees;
    Translate
        1 HIV_env_JP2000a
        2 HIV_env_JP2000b
        3 HIV_env_JP2001
        4 HIV_env_JP2002
        5 HIV_env_JP2003
        6 HIV_env_JP2003b
        6 HIV_env_JP2005
        8 HIV_env_JP2005b
        ;
tree STATE_0 = [&R] ((((20:35.3479176569581,((
tree STATE_11000 = [&R] ((((24:19.959266963075
tree STATE_22000 = [&R] (((((12:80.83442567419
tree STATE_33000 = [&R] (((((9:13.267831008023
tree STATE_44000 = [&R] ((((3:55.6268027053323
```

BaTS format for the same data:

```
#NEXUS

begin states;
1 island
2 island
3 main
4 island
5 main
6 main
7 main
8 main
End;

begin trees;
tree STATE_1011000 = [&R] ((((((8:1.442671720049141
tree STATE_1022000 = [&R] ((((((((((4:2.19177960366
tree STATE_1033000 = [&R] ((((((((((8:1.77960366104
tree STATE_1044000 = [&R] ((((((((((8:1.85759597599
```

The formatting is relatively simple. The key difference in the BaTS-formatted example is that the numbered taxon labels found in the trees no longer correspond, through the '`translate`' block, to individual taxon names. Instead, through the '`begin states`' block, each taxon is assigned a character state. Here for instance two states are present, 'island' and 'main'. The state coding used, number of states, and their biological nature are all irrelevant to BaTS; furthermore, all states are weighted to be equally different to each other in state reconstruction.

Note that the '`begin states`' and '`begin trees`' statements are case-sensitive and that no whitespace characters (spaces or tabs) appear at the start of any line within the '`begin states`' block.

Using BaTS: running an analysis

Versions

Two versions of `BaTS` are available: the `GeneralizedSingleBaTS` estimates significance values for a single dataset and also provides 95% CIs as well as *p*-values; while the `GeneralizedMassBaTS` batch-analyses datafiles a directory at a time and only provides a summary set of *p*-values. This version is useful for analysing a very large number of datasets, for example those derived by simulation.
It is also planned to include a `DetailedSingleBaTS` in an imminent future release; this will report the entire posterior observed distribution and composite posterior expected distribution. Users requiring this functionality before then should contact the author directly.

Usage: GeneralizedSingleBaTS

To use `GeneralizedSingleBaTS` from the command-line, type:

```
java –jar GeneralizedSingleBaTS_v_1_0.jar <treefile_name> <reps> <states>
```

where
`<treefile_name>` is the name and full location of the treefile to be analysed,
`<reps>` is the number (an integer > 1, typically 100 at least) of state randomizations to perform to yield a null distribution, and
`<states>` is the number of different states seen.

Usage: GeneralizedMassBaTS

To use `GeneralizedMassBaTS` from the command-line, type:

```
java –jar GeneralizedMassBaTS_v_1_0.jar <dataset_dir> <reps> <max_states>
```

where
`<dataset_dir>` is the name and full location of the directory of treefiles to be analysed (`BaTS`) will attempt to analyse all files with the extension `.trees` in that directory,
`<reps>` is the number (an integer > 1, typically 100 at least) of state randomizations to perform to yeild a null distribution, and
`<max_states>` is the maximum number of different states seen in any one treefile.

Using BaTS: interpreting analyses

The test statistics

BaTS currently includes implementation of three test statistics used to quantify the strength of phylogeny-trait association. These are the parsimony score ('PS') statistic of Slatkin & Maddison (1989), association index ('AI') of Wang *et al* (2001) and a new measure introduced by Parker *et al* (2008), the maximum monophyletic clade ('MC') size. On a single tree these measures all use the tree node structure and a Fitch (1967b) parsimony agorithm to reconstruct internal node trait values {…} For a full discussion of the statistics, see Parker *et al* (2007, submitted.)

Release note: An imminent future release will also include the Phylogenetic Diversity ('PD') statistic of Hudson *et al* (1992), the Nearest Taxa and Net Relatedness indices ('NTI' & 'NRI') of Webb *et al* (2000; 2002) and the Unique Fraction ('UniFrac') index of Lozupone *et al* (2005.) These indices also include branch length information as well as tree topology, hence weighting related clades by the strength of their relatedness.

The null hypothesis

The null hypothesis under test is one of random phylogeny-trait association; that is, that

> *"No single tip bearing a given character trait is any more likely to share that trait with adjoining taxa than we would expect due to chance"*

As implemented in BaTS, each statistic is scored on the PST and a null distribution generated against which the true posterior statistic distribution is compared. The p-value reported is the proportion of trees from the null distribution equal to, or more extreme than, the median posterior estimate of the statistic from the PST. By convention therefore, we reject the null hypothesis at the desired level of significance α where $p \leq \alpha$, e.g. $p \leq 0.05$ for a significance level of 0.05. We leave it to the user to decide what level of significance is appropriate. Computationally, the p-value is stored, manipulated and printed to output as a Java `float`, a 32-bit floating-point number of variable precision. Because of the way Java handles and rounds floating-point numbers, it is possible that, for instance, a number reported as:

```
0.0500
```

might actually represent the number:

```
0.0500953…
```

in the analysis. For this reason we advise that users accept decimal numbers to at least 3 decimal places; preferably more.

Output

The `BaTS` output varies depending on whether a Single or Batch analysis has been carried out.

Single `BaTS` analyses output a table of information with rows corresponding to posterior estimates of observed and expected values for the PS, AI and MC metrics respectively. Regardless of the number of character states (trait values) only one row of information is presented for the PS and AI metric, but there will be as many MC metric rows as trait values. The MC metrics appear in the order in which they occur in the input .trees file. Columns in the table correspond to the mean, median, and upper and lower 95% HPD intervals of the observed values of the metric, followed by those of the expected (null set), then a p-value. This is the proportion of trees in the observed set equal to, or more extreme than, the median value of the expected (null) set.

Batch `BaTS` analyses are not conducted any differently to single analyses. However they are intended to support large-scale analysis, such as simulation or automated sequence analysis, where the investigator is more interested in the behaviour of a set of metadata than individual datasets. To this end, in order reduce computation time where possible and simplify output data, only p-values are recorded. The output takes the form of a table with one row per dataset where columns give the p-values for PS, AI and MC*{0..n}* statistics respectively. If you intend to carry out a large scale `Batch BaTS` analysis it is recommended that you first carry out a number of trial `Single BaTS` analyses of a random selection of the input datasets to check parsing of the input files is correct and that posterior observed and expected values are approximately in line with those predicted by other methods or *a priori* expectation.

Note that the expected (null) distributions generated are a function of the PST and terminal taxon trait values in combination; even if the same taxa are used the null distributions `BaTS` generates are not valid for PSTs generated under a different tree model, nor for different tip labels. If you reanalyse your data under a different model, or change your tip labelling scheme, you *must* re-analyse the data.

Using BaTS: caveats and warnings

Computational constraints:

Users should be aware that although `BaTS` has been rigorously tested on our development machines, this is the first public release of the package.
As such, the real-world performance of the core packages (which in any case are highly dependent on hardware architecture, and software architecture to a degree) is unknown at this point; in fact, we would appreciate any feedback concerning performance.

As a guide, a `GeneralizedSingleBaTS` of 32 taxa with a binary (2-value) trait on a PST of 10,000 trees with expected distributions assembled from 100 replicates per tree typically takes approximately 5 minutes on a 1.25 GHz Apple PowerPC machine under Java 1.5.0 / Mac OS 10.4.10.

As usual, physical factors increasing compute time performance will include:

- Slower system CPUs
- Slower system bus speeds
- Slower system RAM access (in particular, because `BaTS` uses a lot of memory at present heavy reliance on virtual memory coupled with a slow hard drive is likely to adversely affect performance

Problem parameters that will slow the analysis include:

- Number of replicates to construct expected posterior distribution. We have not specified a default value, since we prefer that users take responsibility for this key parameter; however we have not noticed significantly better power or Type I performance when the number of replicates increases above 100 and typically use this value. Increasing the number of replicate sets dramatically increases both compute time and memory use. We're not sure by how much (depends on platform) but it is expected to be a linear increase at best (e.g. 1000 replicates will take at least 10 times as long, and use 10 times as much memory as 100 replicates)
- Number of taxa. Increasing numbers of taxa will increase compute time and memory usage.
- Increasing sizes of PST. Longer tree sets linearly increase compute time, though memory usage increases by a lesser amount since most of the memory allocated to each tree is re-used for subsequent ones. We have often found it useful to downsample PSTs obtained from long MCMC chains (e.g. to 1000 trees from 90,000 states following a 10,000 state discard as burnin from an original 100,000 state chain); this should be done at regular (not random) intervals.

- Large numbers of different trait values may slow performance (e.g. an analysis with 20 rather than 2 trait values may suffer) because more tree traversals are required. We have not collected substantial data on this effect though, and would particularly welcome feedback.

Biological constraints

We have developed this method to analyse multi-state data on a single character. While there is no computational reason why large numbers of states cannot be analysed, there seems little point in, say, analysing the phylogeny-trait association of a character with 20 discrete states on a 25-taxon tree; any association seen may be as due to sampling error as genuine data signal. Users must use their judgement as to whether `BaTS` is an appropriate way to analyse their data, but are encouraged to contact the author for help.

The MCMC requirements given at the start of this documentation must be obeyed, since the analysis depends on an accurately estimated PST. In particular, users must have confidence in the MCMC, which should have reached stationarity and have no limits, priors or transformations on tree topology or branch lengths, except where there is a good *a priori* reason to apply them.

Lastly, the method uses unweighted parsimony reconstruction. This assumes that transitions between states are all equally likely, totally reversible, and independent of branch length. If your data does not obey these criteria then unfortunately a `BaTS` analysis should not be performed (though future versions of `BaTS` using the PD, NTI/NRI and UniFrac indices will include branch length information.)

FAQ

Note: this is just a preliminary list of FAQs; for any persistent problems, or if you have any other comments, please contact the author.

Why don't my analyses run?
Try running the sample datasets included in this package. If the sample files don't run either you may well have the wrong version of java installed; check and install the correct version. If they run but your data does not, it is likely you have parsed the input files incorrectly. Refer to the 'Input File Requirements' section of this documentation. If problems persist, contact the author.

Why do I get an error message saying I've run out of memory?
Typically this will manifest itself with a message such as
"`Exception in thread 'main': java.lang.OutOfMemoryError: java heap space`"
This error arises then the JVM doesn't have enough system memory ('RAM') available to hold all the data it needs to. Unfortunately invreasing your computer's virtual memory allocation will not solve this problem. You can try increasing the default amount of RAM allocated to the JVM with the '`-Xms`' command (see the 'System Requirements > Hardware' section of this documentation for details.)

Why do I get an error message saying `(ArrayList casting error)`*?*
This is because you are using an earlier version of Java (below 1.5.0) that does not support the way this software handles `java.util.ArrayList`s. Check you have the correct version of Java installed.

Can I run BaTS on a bootstrapped set of trees?
No. This package and the discussion of statistics evaluated with it (Parker *et al*, 2007) is designed to be used in a Bayesian MCMC context. You can evaluate any set of trees you like as long as they are parsed properly, be they a bootstrapped set or some deliberately chosen set: but if you do, all assumptions about the power and behaviour of the statistics are invalid.

Which package should I use to produce a PST first?
Both `MrBayes` and `BEAST` produce acceptable output PSTs. We are unaware of any others.

Contact

The author of this documentation and software is Joe Parker. You can contact him at:

Joe Parker
Kitson Consulting
87 Clevedon Road
Bristol
BS8 3UL
United Kingdom

Or by email: joe@kitson-consulting.co.uk

References

Drummond, A.J. & Rambaut, A. (2003) BEAST v1.0, Available at http://evolve.zoo.ox.ac.uk/beast/

Drummond, A.J., Nicholls, G.K., Rodrigo, A.G. & Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**: 1307-1320.

Faith, D.P. (1992) Conservation evaluation and phylogenetic diversity. *Biol. Cons.* **61**:1-10.

Fitch, W.M. (1971b). Toward defining the course of evolution: Minimal change for a specific tree topology. *Syst. Zool.* **20**: 406-416.

Huelsenbeck, J.P., & Ronquist., F. (2001). MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* **17**:754-755.

Lozupone, C. & Knight, R. (2005) UniFrac: A new method for comparing microbial communities. *App. & Environ. Microbiol.* **71**(12):8228-8235.

Parker, J., Rambaut, A.R. & Pybus, O.G. (2008) Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *MEEGID* **8**(3):239-246

Slatkin, M., & Maddison, W.P. (1989). A cladistic measure of gene flow measured from the phylogenies of alleles. *Genetics* **123**(3):603-613.

Wang, T.H., Donaldson, Y.K., Brettle, R.P., Bell, J.E. & Simmonds, P. (2001). Identification of shared populations of Human immunodeficiency Virus Type 1 infecting microglia and tissue macrophages outside the central nervous system. *J. Virol.* **75** (23): 11686-11699.

Webb, C.O. (2000) Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *Am. Nat.* **156**(2):145-155

Webb, C.O., Ackerly, D.D, McPeek, M.A. & Donoghue, M.J. (2002) Phylogenies and community ecology. Annu. *Rev. Ecol. Syst.* **33**:475-505

Appendix to Chapter Four

Details of clinical information, sample collection & sequencing.

Attribution

The data analyzed in Chapter Four was prepared by A. Iversen (Weatherall Institute of Molecular Medicine, Oxford University, John Radcliffe Hospital, OX3 9DS, UK), who collected and sequenced samples and performed the HLA typing analysis and viral load quantification.

Collection

The samples used in this study were collected by A. Iversen from infected women. Clinical information, including duration of infection and reported method of transmission can be found in the table below.

Viral load quantification

HIV-1 RNA load in plasma was measured using the Amplicor HIV Monitor kit, which amplify a part of the *gag* gene (Roche Diagnostic Systems, NJ, USA). The sensitivity of the test on non-B HIV-1 subtypes has been evaluated on a limited number of isolates (2-5 isolates/subtype) by the manufacturer. The sensitivity of detection relative to HIV-1 subtype B was: subtype A, 0-15%; subtype C, 50-97%; subtype D, 90-130%; subtype G, 75-115%.

Sequencing

DNA was extracted from PBMC and cervical cell pellets using the QIAamp®Blood kit (Qiagen Incorporated, CA, USA). Aliquots from all DNA samples were amplified using b-globin specific primers in order to determine if the samples contained strong inhibitors of PCR amplification (Iversen, 1998.) HLA typing was done using the AllSet^{TM+} low resolution A, B and Cw kit (Dynal Biotech A.S.A., Norway). Quantification of HIV-1 proviral DNA from patients was carried out by limiting dilution using nested PCR to amplify parts of the p17/p24 *gag* and C2-V5 *env* genes (Iversen, 1998; Holmes, 1995.) Approximately 10000 PCR reactions were performed in order to obtain the 1140 sequences. Direct DNA sequencing was done on PCR amplicons obtained from single copy PBMC and cervix derived HIV templates as recommended by the manufacturer (DyeDeoxy^{TM}Terminators, Applied Biosystems Inc., CA, USA).

Disease progression status

Many studies have tried to correlate the magnitude and/or breadth of the CTL response with control of viremia in humans and for the majority of HLA restriction elements no such simple relationship exist. In chronic HIV-1 infection both broad and narrow high frequency CTL responses have been seen in patients whose CD4 count are rapidly declining, or who are dying from AIDS, as well as in asymptomatic patients with stable CD4 counts and low levels of viremia (Goulder, 1997; Feeney, 2004; Hay, 1999; Draenert, 2004; Migueles, 2000.) Immunogenetic studies have clearly shown that certain HLA alleles, e.g. HLA-B27, -B57 and –B58, are associated with long-term non-progression (LTNP) while HLA-B35 is associated with rapid disease progression (Carrington, 1999; Kaslow, 1996; Goulder, 1996; O'Brien,

2001; Migueles, 2000; Tang, 2002; Gao, 2001). This classification was used to predict disease progression status for patients along with available clinical data.

Clinical information

Patient	Sample dates	Subtype[1]	CD4+ count	Duration[2]	DNA load[3]	RNA load[4]	DP[5]	Trans-mission mode[6]
AA	06/06/1995 12/03/1996	A (gag), D (env)	49	NA	41666 111111	8328 15064	-	HSC, IVDU
AB	02/06/1995 27/11/1995 20/05/1996	B	376	4	41666 31250 30722	77611 50634 101135	STD	HSC
AC	01/06/1995	C	20	NA	45454	737234	STD	HSC
AD	01/06/1995 28/02/1996	B	16	10	45454 500000	10294 18564	-	HSC
AE	06/06/1995 13/03/1996	B	634	11	70.6 79	301 198	-	HSC, IVDU
AF	08/06/1995 22/06/1995 20/06/1996	CRF 13	115	6	24390 15928 36496	*7185 *2025 *3180	-	HSC
AG	08/06/1995 13/03/1996	A	267	NA	5747 8065	*409 *288	-	HSC
AH	14/06/1995	C	15	4	5514	166996	LTNP	HSC (Rape)
AI	15/06/1995 14/12/1995	B	2	12	921658 11363	147621 139653	RP	HSC
AJ	16/06/1995 23/01/1996	B	398	12	4353 2724	74088 19071	STD	BT
AK	16/06/1995 20/11/1995	CRF 10	180	NA	11319 5128	*22278 *12906	-	HSC
AL	20/06/1995 14/12/1995	B	20	9	430108 2000000	655335 296153	-	HSC
AM	22/06/1995 28/11/1995	B	49	11	41666 80645	363174 282759	RP	HSC
AN	22/06/1995	B	509	5	1792	8687	-	HSC

AO	23/06/1995	CRF 10	116	4	187003	317235	LTNP	HSC
AP	30/06/1995	B	94	13	37037	733713	LTNP	HSC
AQ	30/06/1995	B	45	NA	37037	130952	-	HSC, IVDU
AR	30/06/1995	B	476	3	44	3787	LTNP	HSC
AS	30/06/1995 05/01/1996	B	257	6	7791 5780	32992 65369	STD	HSC
AT	30/06/1995	B	332	13	18181	83182	STD	HSC
AU	03/07/1995 16/01/1996	D	79	NA	29868 16949	13004 16956	LTNP	HSC, IVDU, BT
AV	04/07/1995 29/11/1995 20/06/1996	C	154	7	1529 2666 4608	2347 1146 11328	RP	HSC
AW	05/07/1995 01/02/1996	B	450	5	144 104	657 431	STD	HSC
AX	13/07/1995	B	117	12	3145	16485	-	HSC, IVDU
AY	04/09/1995	A	32	8	250000	*87337	STD	HSC
AZ	10/10/1995 31/05/1996	B	210	8	1479 1300	8958 5475	-	BT, IVDU
BA	30/11/1995 31/05/1996	G	500	NA	1700 860	521 0	LTNP	HSC
BB	21/11/1995	B	233	6	3106	3699	STD	HSC, IVDU
BC	30/11/1995 30/05/1996	B	28	9	11947 12658	16370 31889	-	HSC
BD	19/12/1995	A	862	1	213	*489	-	HSC
BE	10/01/1996	B	281	7	2646	7198	STD	HSC
BF	29/01/1996	B	296	5	1706	8084	STD	HSC, IVDU
BG	01/02/1996	B	798	0.5	1182	33389	-	HSC
BH	08/02/1996	B	80	NA	38462	103293	-	HSC
BI	13/02/1996	B	509	5	4854	2770	-	HSC
BJ	21/03/1996	B	517	12	17543	60169	-	HSC
BK	22/04/1996	B	517	5	489	4065	LTNP	HSC, IVDU
BL	30/04/1996	B	145	9	6897	10299	-	IVDU
BM	06/05/1996	C (gag),	396	NA	2805	*36099	LTNP	HSC

		A (env)						
BN	15/05/1996	B	260	7	2933	523	LTNP	HSC
BO	06/06/1996	C	375	10	12658	11419	-	HSC (rape)

[1]As determined by the BioAfrica subtyping tool; RF: circulating recombinant form.

[2]Approximate; NA: not available.

[3]Copies per 10^6 CD4+ cells.

[4]Copies per ml plasma. Starred entries (*) from subtypes not well amplified by this method.

[5]Where available. STD: standard; LTNP: long term non-progression; RP: rapid progression.

[6]HSC: heterosexual contact; IVDU: intravenous drug use; BT: blood transfusion.

References

Carrington M, Nelson GW, Martin MP, Kissner T, Vlahov D, Goedert JJ, Kaslow R, Buchbinder S, Hoots K, O'Brien SJ. (1999) HLA and HIV-heterozygote advantage and B*35-Cw*04 disadvantage. *Science* **283**(5408):1748-1752.

Draenert R, Le Gall S, Pfafferott KJ, Leslie AJ, Chetty P, Brander C, Holmes EC, Chang SC, Feeney ME, Addo MM, Ruiz L, Ramduth D, Jeena P, Altfeld M, Thomas S, Tang Y, Verrill CL, Dixon C, Prado JG, Kiepiela P, Martinez-Picado J, Walker BD, Goulder PJ. (2004) Immune selection for altered antigen processing leads to cytotoxic T lymphocyte escape in chronic HIV-1 infection. *J Exp Med.* **199**(7):905-915.

Feeney ME, Tang Y, Roosevelt KA, Leslie AJ, McIntosh K, Karthas N, Walker BD, Goulder PJ. (2004) Immune escape precedes breakthrough human immunodeficiency virus type 1 viremia and broadening of the cytotoxic T-lymphocyte response in an HLA-B27-positive long-term-nonprogressing child. *J Virol.* **78**(16):8927-8930.

Gao X, Nelson GW, Karacki P, Martin MP, Phair J, Kaslow R, Goedert JJ, Buchbinder S, Hoots K, Vlahov D, O'Brien SJ, Carrington M. (2001) Effect of a single amino acid change in MHC class I molecules on the rate of progression to AIDS. *N Engl J Med.* **344**(22):1668-1675.

Goulder PJ, Bunce M, Krausa P, McIntyre K, Crowley S, Morgan B, Edwards A, Giangrande P, Phillips RE, McMichael AJ. (1996) Novel, cross-restricted, conserved, and immunodominant cytotoxic T lymphocyte epitopes in slow progressors in HIV type 1 infection. *AIDS Res Hum Retroviruses.* **12**(18):1691-1698.

Goulder PJ, Bunce M, Luzzi G, Phillips RE, McMichael AJ. (1997) Potential underestimation of HLA-C-restricted cytotoxic T-lymphocyte responses. *AIDS* **11**(15):1884-1886.

Hay CM, Ruhl DJ, Basgoz NO, Wilson CC, Billingsley JM, DePasquale MP, D'Aquila RT, Wolinsky SM, Crawford JM, Montefiori DC, Walker BD. (1999) Lack of viral escape and defective in vivo activation of human immunodeficiency virus type 1-specific cytotoxic T lymphocytes in rapidly progressive infection. *J Virol.* **73**(7):5509-5519.

Holmes, E.C., *et al* (1995). The molecular epidemiology of human immunodeficiency virus type 1 in Edinburgh, *J Infect Dis*. **171**(1):45-53 Iversen AK, Larsen AR, Jensen T, Fugger L, Balslev U, Wahl S, Gerstoft J, Mullins JI, Skinhoj P. (1998). Distinct determinants of human immunodeficiency virus type 1 RNA and DNA loads in vaginal and cervical secretions. *J Infect Dis.* **177**(5):1214-1220.

Kaslow RA, Carrington M, Apple R, Park L, Munoz A, Saah AJ, Goedert JJ, Winkler C, O'Brien SJ, Rinaldo C, Detels R, Blattner W, Phair J, Erlich H, Mann DL. (1996) Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection. *Nat Med.* **2**(4):405-411.

Migueles SA, Sabbaghian MS, Shupert WL, Bettinotti MP, Marincola FM, Martino L, Hallahan CW, Selig SM, Schwartz D, Sullivan J, Connors M. (2000) HLA B*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors. *Proc Natl Acad Sci U S A.* **97**(6):2709-2714.

O'Brien SJ, Gao X, Carrington M. (2001) HLA and AIDS: a cautionary tale. *Trends Mol Med.* **7**(9):379-381.

Tang J, Tang S, Lobashevsky E, Myracle AD, Fideli U, Aldrovandi G, Allen S, Musonda R, Kaslow RA; Zambia-UAB HIV Research Project. (2002) Favorable and unfavorable HLA class I alleles and haplotypes in Zambians predominantly infected with clade C human immunodeficiency virus type 1. *J Virol.* **76**(16):8276-8284.

Appendix Six

Appendix to Chapter Two:

Variance of null distributions generated in BaTS

A6.1 Variance of null distributions

The BaTS method developed in Chapter Two (and employed again in Chapters Four & Five) compares the distribution of observed phylogeny-trait association values in the posterior set of trees against a proper null distribution of phylogeny-trait associations generated by randomly resampling without replacement from the observed distribution of character states. Since the replicates generated in this way are effectively sampled from the true null distribution, we must be certain that we have generated a proper null distribution that reflects the true null distribution. In order to be certain of doing so, it is necessary to generate a large number of replicates; however it is inefficient to generate more replicates than required. This trade-off will be approximately optimized when the variance in the mean expected value of the statistic estimated from separate null distributions is low. Here I use simulation to investigate the variance of the mean of null distributions of the AI statistic, and show that 100 replicates are a sufficiently high number to recover the null distribution for most purposes.

6.1.1 Method

To examine the variance in the expected mean AI statistic, I used 3 of the simulated posterior sets of trees (PST) selected at random from each of the 9 master tree topology sets in Chapter 2, giving 27 test PSTs (each simulated PST consists of 909 trees of 32 tips each, randomly coloured with either ‘black’ or ‘white ‘ character traits). For each PST I constructed null distributions in BaTS with varying sample intensity (number of replicates to construct the null distribution: 1, 2, 3, 4, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400 & 500 replicates were used). 30 null distributions were constructed at each sampling intensity for each PST, and the variance of the mean AI statistic values recorded. The mean variance in

the expected mean AI statistic over all 27 test PSTs was then calculated for each sampling intensity. Variance of AI, PS, MC, UniFrac, NT, NR & PD statistics at 10, 20, 25, 30, 50, 75, 100, 250, 500, 750 and 1000 replicates for an empirical data-set was also estimated and did not differ significantly from the curve shown for AI in Fig. A6.1 (available on request).

6.1.2 Results

Figure A6.1 shows the variance in mean expected AI values under each sampling intensity for each of the 27 test data sets. It can be seen that as larger numbers of replicates are used to construct the null distribution, the variance in the mean AI values observed between different constructed null distributions decreases asymptotically.

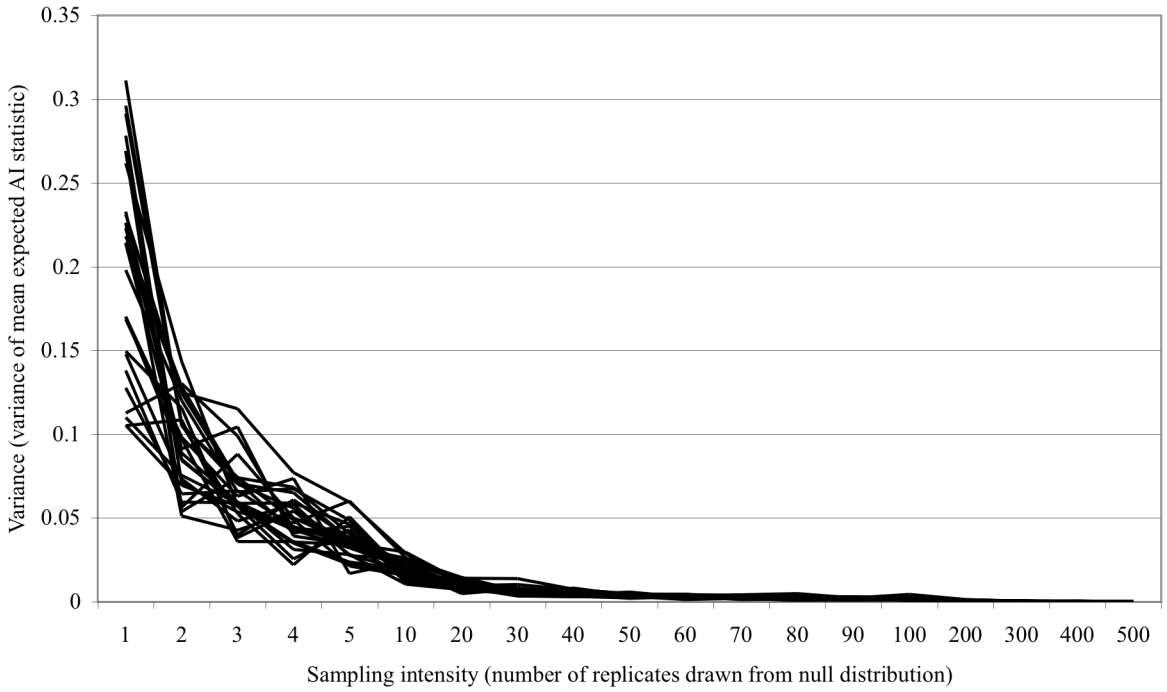


Figure A6.1: Asymptotic decline of variance in generated null distributions with increased sampling intensity. At each sampling intensity, 30 null distributions were generated by random draws from the true null distribution. This was repeated for each of 27 test posterior sets of trees.

This information is also summarized in Table A6.1, which gives details of the mean variance in mean AI scores at each sampling intensity.

Sampling intensity	Mean variance in mean expected AI value
1	0.204
2	0.093
3	0.065
4	0.049
5	0.037
10	0.019
20	0.010
30	0.007
40	0.005
50	0.004
60	0.003
70	0.003
80	0.002
90	0.002
100	0.002
200	0.001
300	<0.001
400	<0.001
500	<0.001

Table A6.1
Mean variance in expected AI values at increasing sampling densities.

6.1.3 Conclusion

I conclude that, for most purposes, 100 replicates are a sufficiently large number to correctly sample from the true null distribution when constructing the expected null distribution in BaTS while minimising the computational load, though for significantly larger trees than those tested here (n = 32 tips) it may be preferable to use up to 500 replicates.