Identifying the Molecular Signatures of Adaptive Evolution

Thesis submitted for the degree of Doctor of Philosophy
UCL

Christopher Monit

Division of Infection and Immunity
UCL

Declaration

I, Christopher Monit, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Using both novel and established molecular evolutionary modelling techniques, we have investigated the evolution of primate lentiviruses and interactions with their hosts. Firstly, we studied SAMHD1, a restriction factor of HIV-1 which is neutralised by lentiviral proteins. SAMHD1 has previously been shown to be under positive selection in primates, ostensibly due to pressure to escape recognition by lentiviral antagonists. We show positive selection is not unique to primates but has occurred throughout chordate evolution. In mammals, we unexpectedly find SAMHD1 sites under positive selection are clustered in the domain controlling enzymatic activation. We hypothesise that positive selection is driven by undiscovered animal viruses and/or precise regulation of SAMHD1 activity. Secondly, we analysed the capsid proteins of pandemic HIV-1 and its chimpanzee progenitor, SIVcpz. We looked for sites evolving under different selective constraints with the aim of discovering host specific adaptation. We identify sites in the domain bound by host cofactors, which govern crucial events in virus replication and prevent immune sensing, suggesting host specific responses to cofactor interaction. Thirdly, we apply this same approach to pandemic HIV-1 and SIVcpz accessory proteins, which mitigate host immunity. Surprisingly, we identify sites in regions of *nef* and *vpr* involved in putatively conserved interactions with host proteins, suggesting unexpected host specific adaptation. In *vpu*, we identify sites involved in antagonism of the restriction factor tetherin — a function acquired by pandemic HIV-1 on adaptation to humans — together with sites which we hypothesise are similarly involved. Finally, lentiviruses and other organisms possess overlapping coding sequences, for which existing codon selection models are unsuitable. We propose a novel approach which models nucleotide substitution. In synthetic data tests, one of four candidate models was accurate and we developed a mixture model for identifying positive selection at codon sites, which we also tested with synthetic data.

Acknowledgements

I must give my heartfelt thanks to my supervisor, Richard Goldstein. I have greatly enjoyed my time working with him and I am proud to have been a member of his group.

I am grateful also for the training and encouragement from my secondary supervisor, Stéphane Hué, and for the advice from my other thesis committee members Willie Taylor, Deenan Pillay and Greg Towers. The past and present members of Richard's group and others in the Division have made these years greatly enjoyable and I am thankful for their friendship. I am especially grateful to Martin Godány for guiding me in my early days of computational research. My thanks also to my examiners, Michael Malim (KCL) and Ziheng Yang (UCL).

Collaborators

- **SAMHD1:** Chris Ruis assisted with SAMHD1 sequence collection and alignment; Richard Goldstein wrote software for assessing statistical support of residue clustering; Ariberto Fassati assisted with data analysis and Richard Goldstein supervised the study; this work was based on a pilot study with contributions from Grant Thiltgen and N. Avrion Mitchison. Collaborations with the Jonathan Stoye and Ian Taylor laboratories are ongoing.
- **Selection analysis of lentiviral capsid**: Richard Goldstein and members of the Greg Towers laboratory. I am especially grateful for insightful advice from Katsiaryna Bichel.
- **Selection analysis of lentiviral accessory genes**: Supervised by Richard Goldstein, with technical assistance from Asif Tamuri. Based in part on a pilot study also involving Chris Ruis and Stéphane Hué.
- **Overlapping coding sequences**: Supervised by Richard Goldstein.

This work was funded by the Medical Research Council.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Molecular Evolution

Molecular evolution is the study of organisms changing across generations, by statistical inference from molecular sequence data. By analysing how protein and genetic sequences differ between related organisms we can come to understand the process of evolution at the molecular level, or reconstruct the history of past evolutionary events. The capabilities of biological evolutionary research have expanded dramatically with the ongoing increase in genetic sequences available from diverse organisms, together with continuously expanding computational resources for storing and analysing them. Aside from an interest in evolution itself, studying the molecular adaptation which gave rise to the organisms we see today can guide the research of our primary interest as biologists, namely how viruses, butterflies or oak trees are able to function.

1.1.1 Modelling Molecular Evolution

A typical analysis begins with molecular sequences from different organisms being aligned to one another, compensating for sequence insertions or deletions, such that related (homologous) sites in each sequence are correctly compared. Simple probabilistic models are devised to represent how nucleotides, codons or amino acids states undergo substitution by the process of mutation and then fixation by natural selection. A probabilistic approach is superior to reconstructing past events by identifying most parsimonious explanations because it accounts for unobserved multiple substitutions and reversions. These models must balance our understanding of bi-

ology with mathematical tractability. Transitions between states are conventionally modelled as a continuous-time Markov process, expressed as a matrix $\boldsymbol{Q}$ of instantaneous substitution rates between states, populated by parameters representing our assumptions about the substitution process. For example, the HKY85 nucleotide model (Hasegawa et al., 1985):

$$\boldsymbol{Q} = \begin{pmatrix} * & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & * \end{pmatrix} \tag{1.1}$$

where π_i is the equilibrium probability of nucleotide i, a mathematical contrivance corresponding to the proportion of time that i would occupy the site over an infinite duration. κ represents the bias for substitutions of nucleotides of the same chemical structure (*transitions*, i.e. purine to purine or pyrimidine to pyrimidine) over those changing the chemical structure (*transversions*, e.g. purine to pyrimidine). This parameter therefore represents a known biological phenomenon. Different models can be created by adding or removing parameters from $\boldsymbol{Q}$. The rows of the matrix must sum to 0, and so the diagonal elements are defined

$$q_{ii} = -\sum_{\{j|j\neq i\}} q_{ij}. \tag{1.2}$$

The models are reversible, meaning the flux of probability density from state i to j must be identical in the reverse; formally, $\pi_i q_{ij} = \pi_j q_{ji}$. Ultimately we wish to compute the probability of substitution from i to j in a given interval of species divergence t, equivalent to a branch length on a phylogenetic tree topology. For convenience $\boldsymbol{Q}$ is multiplied by a scalar ν such that one unit of t represents one substitution per site. Ordinarily a single $\boldsymbol{Q}$ applies to all sites, in which case

$$\nu = \frac{1}{-\sum_i \pi_i q_{ii}}. \tag{1.3}$$

These Markov models assume that evolution at each site has the statistical property of being 'independent and identically distributed' (i.i.d.), for the purposes of reducing the number of parameters involved (Yang, 2006). However, we may wish to have multiple matrices, each applicable to a subset of sites; e.g. one could have a

separate matrix for each of the three codon positions. In which case,

$$\nu = \frac{\sum_k N_k}{-\sum_k N_k \sum_i \pi_i q_{ii,k}} \tag{1.4}$$

where N_k is the number of sites in subset k, to which $\boldsymbol{Q}_k$ is applicable. (Note that $\sum_k N_k$ is the total number of sites in the alignment.) In this case sites within each set k are i.i.d., but not between sets.

From the scaled rate matrix, we compute the matrix of substitution probabilities for the interval t:

$$\boldsymbol{P}(t) = e^{\boldsymbol{Q}t} = \sum_{n=0}^{\infty} \boldsymbol{Q}^n \frac{t^n}{n!}. \tag{1.5}$$

1.1.2 Model Likelihood and Hypothesis Testing

The fit of a model and its parameter values can be assessed by the principle of likelihood. The likelihood of the model in light of the data is the probability of observing the data assuming the model were true. Given a model $\boldsymbol{Q}$ comprising parameter values $\{\theta\}$, a tree topology with branch lengths $\{t\}$ and a sequence alignment, $\boldsymbol{P}(t)$ is computed for each branch and the likelihood can be computed with Felsenstein's pruning algorithm (Felsenstein, 1981). This is a dynamic programming approach which efficiently sums over substitution probabilities for all possible evolutionary histories. Finding values for θ which maximise the likelihood gives the parameter maximum likelihood estimates (MLEs; $\hat{\theta}$). For all but the simplest models the maximisation requires numerical optimisation algorithms. Tree topologies themselves are also estimated by maximum likelihood (ML), using software such as RAxML (Stamatakis, 2006).

In an ML framework, biological hypotheses can be tested by constructing two models which differ by one or more parameters, corresponding to null and alternative hypotheses. For example, one may test whether there is a bias for transitions over transversions by constructing the HKY85 model (eq. 1.1) as an alternative model and also a null model which has κ fixed to 1. The models are said to be 'nested', as the null is a special case of the alternative. Both models are fitted to the data and the null can be rejected if the alternative has a significantly higher likelihood. A p value is computed from the test statistic $D = 2(\ln L_1 - \ln L_0)$, where L_1 and L_0 are the likelihood under the alternative and null models, respectively. The null

distribution of D is approximated by a χ^2 with degrees of freedom equal to the difference in free parameters (1 in the HKY85 example above). This is called the likelihood ratio test (LRT).

1.2 Modelling Selection

1.2.1 Types of Selection

Often we study molecular sequences in order to identify the action of natural selection. In population genetics theory, natural selection is divided into three types: negative (purifying) selection, in which genotypic change negatively affects the organism's fitness; neutral selection, where there is little or no effect on fitness; or positive selection, where change is favoured by natural selection. Positive selection can be further divided into diversifying selection and directional selection. Diversifying selection describes change from the present state being favoured, with little preference for the target state. For example, in the presence of antibodies, an amino acid change on a virus' exterior which prevents antibody binding may be favoured, regardless of what the new amino acid is. (The result is increased diversity, because each descendent lineage will likely make different changes to differentiate itself from the ancestral type.) Conversely, directional selection describes specific changes being favoured, perhaps because the organism's environment begins imposing new selective constraints, for example if a virus begins infecting a new host species.

1.2.2 Codon Models and the d_N/d_S Ratio

To identify selection in molecular data, one must distinguish substitutions which have occurred by chance, rather than as adaptive change. A powerful family of models do so by exploiting the degeneracy of the genetic code. Codon substitutions which do not change the amino acid (called 'synonymous') are assumed to be neutral (have no effect on fitness), while those that do ('nonsynonymous') are assumed to be subject to selection. Comparing the number of substitutions of each type estimated to have occurred between sequences gives a measure of selection. The ratio of these 'genetic distances' d_N and d_S being < 1 indicates negative selection, while ≈ 1 indicates neutrality and > 1 indicates positive selection.

By defining a parameter $\omega = d_N/d_S$, Goldman and Yang (1994) devised a codon

substitution model which can account for selection, where the rate of substitution from codon I to codon J is

$$q_{IJ} = \begin{cases} 0, & \text{if } I \text{ and } J \text{ differ by 2 or 3 nucleotides,} \\ \pi_J, & \text{if } I \text{ and } J \text{ differ by synonymous transversion,} \\ \kappa\pi_J, & \text{if } I \text{ and } J \text{ differ by synonymous transition,} \\ \omega\pi_J, & \text{if } I \text{ and } J \text{ differ by nonsynonymous transversion,} \\ \omega\kappa\pi_J, & \text{if } I \text{ and } J \text{ differ by nonsynonymous transition} \end{cases} \tag{1.6}$$

where π_I is the equilibrium probability of codon I and κ is as defined above. Given a codon alignment, the likelihood of the model is computed as described above.

To assume a single ω is suited to all sites is unrealistic, however, due to varied structural constraints in most proteins. Indeed, the majority of sites will be under negative selection in any coding sequence, meaning positive selection is unlikely to be identified. Nielsen and Yang (1998) therefore extended this approach to a mixture model allowing for three classes of sites, represented as three rate matrices featuring parameters which are constrained to represent the types of selection described above: $\omega_0 < 1$, $\omega_1 = 1$ (i.e. fixed) and $\omega_2 > 1$. Each site class i has a probability p_i of being applicable. The likelihood for the data X observed at site h is then sum over components of the mixture:

$$L(X_h) = \sum_i p_i P(X_h|\omega_i) \tag{1.7}$$

where $P(X_h|\omega_i)$ is the likelihood of the model given data X_h computed with the rate matrix populated with ω_i. Each site remains i.i.d., but using a mixture model accounts for the ambiguity associated with the selection pressure at each codon site, without resorting to a large number of site-specific parameters. A test for positive selection can be constructed by fitting to the data both a null model which has only the negative and neutral selection site classes and an alternative model which features the positive selection site class in addition. These correspond to models M1a and M2a respectively (Wong et al., 2004). The LRT is used to assess significance, with two degrees of freedom. More complex mixture models allow flexibility by having a range of ω values represented as a beta distribution (Yang et al., 2000), corresponding to models M7 (null) and M8 (alternative); again the LRT provides a test for positive selection. A further extension allows tests on specific lineages in a

phylogeny, such that episodic positive selection can be identified (Yang and Nielsen, 2002). Each of these models are implemented in codeml, in the PAML software package (Yang, 2007).

1.2.3 Site-Wise Mutation Selection Model

The codon substitution models inheriting from Goldman and Yang (1994) are well suited to identifying diversifying positive selection, since the ω parameter makes no distinction between nonsynonymous changes. Arguably these models may be therefore suited to only a few biological situations (Hughes, 2007).

Tamuri et al. (2009) developed a modeling framework explicitly for studying directional positive selection. Their interest lay in identifying adaptation of a virus to new selective constraints following transmission from an animal reservoir to humans (a zoonosis). They developed an amino acid substitution model, in which amino acid equilibrium frequencies were specific to each protein site (i.e. site-wise). In their null model, a single set of equilibrium frequencies is applied to all lineages in the phylogeny, representing the hypothesis that the evolutionary process is consistent. Their alternative model permits a separate set of frequencies for a clade of viruses infecting a different host, representing the hypothesis that the selective constraints imposed by the host cause the evolutionary process to change. Statistical support is assessed with the LRT. This approach was applied to influenza virus, which persists as an avian reservoir but routinely establishes epidemics in humans.

Subsequently these authors (Tamuri et al., 2012) developed a site-wise codon substitution model (swMutSel) which incorporates both processes of mutation at the genetic level and selection acting on the protein, with reference to population genetics theory (Halpern and Bruno, 1998; Yang and Nielsen, 2008; Kimura, 1962). The substitution rate between codons I and J at site K is

$$q_{IJ,K} = q_{ij}^{\mathrm{HKY}} \times \frac{S_{IJ,K}}{1 - e^{-S_{IJ,K}}} \tag{1.8}$$

where q_{ij}^{HKY} is as given in equation 1.1 and represents the rate of mutation between the nucleotides i and j which differ between the codons (multiple nucleotide changes in a codon have a rate of 0); this is shared across sites. $S_{IJ,K}$ is the selection coefficient (difference in codon Malthusian fitnesses f) scaled by the effective population size N: $S_{IJ,K} = 2N(f_{J,K} - f_{I,K}) = F_{J,K} - F_{I,K}$. Codon fitnesses are

assumed to be defined solely by the amino acids they encode and are expressed relative to a single arbitrarily chosen residue, meaning there are 19 scaled fitness parameters. Mutation parameters and branch lengths are estimated separately by a related mutation-selection model which is less computationally demanding (Yang and Nielsen, 2008).

swMutSel has been extended by Asif Tamuri (European Bioinformatics Institute, UK) to test for site-wise divergent selective constraints similar to the amino acid model (Tamuri et al., 2009) in the context of a codon mutation-selection model (Tamuri et al., 2012). The alternative model permits the sublineage of interest independent amino acid fitness parameters. Significance is assessed with the LRT, using $N - 1$ degrees of freedom, where N is the number of residues present at the site. The performance of swMutSel in identifying directional selection has been assessed in a simulation study and found to be accurate and conservative, while also having greater sensitivity than d_N/d_S codon models for this type of positive selection (Thiltgen et al., 2016).

1.3 Primate Lentivirus Biology

We are interested in the evolution of primate lentiviruses and the interactions with their hosts. Lentiviruses are retroviruses (Baltimore class VI), meaning their genomes are positive sense single stranded RNA and replicated through a DNA intermediate which is integrated into the host cell genome. Among these are the human immunodeficiency viruses (HIV), which are causative agents of acquired immunodeficiency syndrome (AIDS).

1.3.1 AIDS and its Impact

AIDS was first identified in the 1980s when previously healthy young men in urban centres presented with infections usually found in immunocompromised patients (Gottlieb et al., 1981) and a retrovirus was identified as the infectious agent responsible (Barre-Sinoussi et al., 1983). HIVs are primarily transmitted sexually and infection is established in mucosal tissue before progressing to the primary target cells, CD4+ T cells which form an integral part of the adaptive immune system. As the infection progresses the population of these cells is destroyed, leaving the patient

vulnerable to infection by opportunistic pathogens (reviewed by Deeks et al., 2015). Antiretroviral drugs have been developed which inhibit the viral enzymes (reverse transcriptase, protease and integrase), suppress viral replication and delay progression to AIDS. However, the virus is not eliminated and will rebound if treatment is withdrawn. Drug resistance emerges routinely, meaning drugs must be used in combination (Clavel and Hance, 2004). No successful therapeutic or prophylactic vaccine has been developed (Esparza, 2013), for reasons discussed below.

AIDS is a pandemic (worldwide epidemic) and its impact on global health has been profound. 37 million people are estimated to be infected with HIV (UNAIDS, 2016), equivalent to the population size of Canada. Since the emergence of HIV 35 million people are estimated to have died from AIDS-related disease. Globally, deaths are now declining, having peaked in 2005, partly due to increased availability of antiretroviral therapy (ART; UNAIDS, 2013). However, still only 37% of infected people receive ART, which is particularly scarce in developing countries. AIDS is damaging to communities and societies because it disables young people who would be otherwise productive. There is a substantial economic impact in the Western world (Trapero-Bertran and Oliva-Moreno, 2014), but AIDS has significantly hindered the progress of developing nations, particularly in sub-Saharan Africa (Dixon, 2002). Without exaggeration, confronting the AIDS pandemic can be described as one of the greatest collective challenges in human history.

1.3.2 Virus Particle, Genome and Life Cycle

The lentivirus virion comprises a viral core surrounded by a lipid bilayer, through which protrude protein spikes (fig. 1.1A). The core is made from the viral capsid protein assembled into hexamers and pentamers, which in turn assemble into a cone structure (fig. 1.1B), housing the two copies of the RNA genome and the viral enzymes. The $\sim$ 10 kilobase genome encodes 9 genes which produce 15 protein products: *gag* encodes structural proteins as a polyprotein, including capsid; *pol* encodes the viral enzymes, also as a polyprotein; *env* encodes the spike protein; *tat* and *rev* are regulatory proteins involved in viral gene expression; *nef*, *vpu*, *vpr* and *vif* are accessory proteins involved in mitigating cellular immunity. The HIV-1 genome contains 7 instances of overlapping coding sequences, where a single locus contains more than one translated reading frame. At one locus the maximum 3

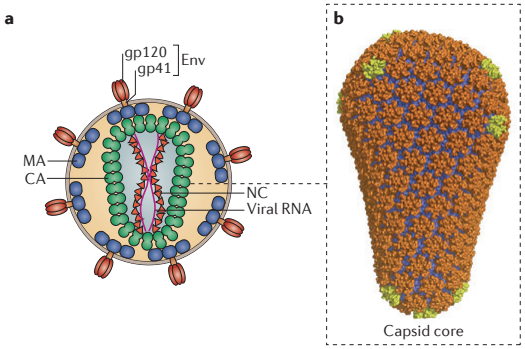

Figure 1.1: Structure of the lentivirus virion. (A) Cartoon of the virion. The viral core comprises the capsid protein (CA) and houses the viral RNA genome and stabilising nucleocapsid protein (NC). The matrix protein (MA) is positioned between the core and the lipid envelope, through which the envelope protein (Env, made from glycoproteins gp120 and gp40) protrudes. (B) Protein model of the core, showing capsid proteins arranged in hexamers (orange) and pentamers (yellow). From Campbell and Hope (2015).

reading frames are occupied, by *tat*, *rev* and *env*.

The life cycle begins with the virion binding via its envelope protein the CD4 receptor on a macrophage or CD4+ T cell (fig. 1.2). Interaction with a coreceptor (typically CCR5) mediates fusion of the viral lipid envelope with the target cell plasma membrane. The core traffics to the nucleus and the reverse transcribed genetic material (cDNA) is imported through the nuclear pore. The details and timing of core uncoating and reverse transcription remain unclear, though it is likely reverse transcription occurs within an intact or partially intact core (Campbell and Hope, 2015). The viral cDNA is integrated into the host genome by the viral integrase, and viral transcription and translation proceed as for cellular genes. Polyproteins of Gag, Gag-Pol and Env aggregate at the plasma membrane, together with viral genomic RNAs. The nascent virion 'buds' from the cell, taking with it a portion of the plasma membrane, which forms the viral envelope. The particle then undergoes a maturation event where the vial protease cleaves the polyproteins, introducing significant conformational changes which gives rise to an infectious virion.

1.3.3 Lentivirus Diversity and Evolution

HIV type 1 (HIV-1) is divided into four groups, each originating from independent cross-species transmissions. Group M accounts for the vast majority of infections worldwide, and is itself divided into 9-11 monophyletic subtypes. Group O infects a few tens of thousands, largely confined to West Africa, while group N has been

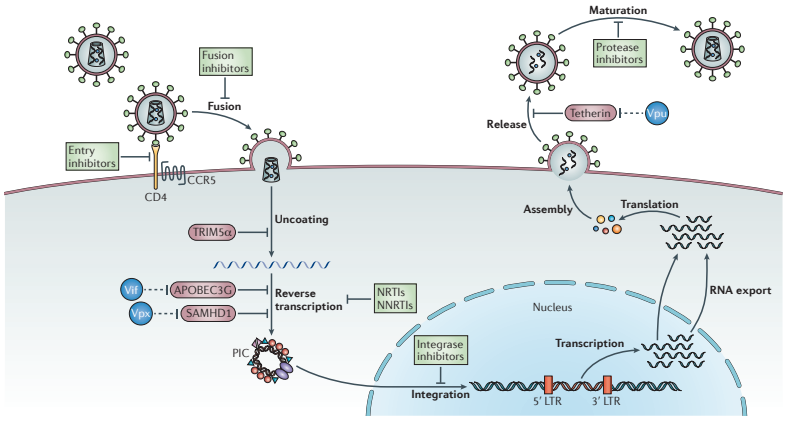

Figure 1.2: Cartoon illustrating the lentivirus life cycle, narrated in the text. Hypothetical and actual drug targets are indicated, as well as host restriction factors targeting various stages of the life cycle. While the figure suggests uncoating occurs prior to reverse transcription, recent evidence suggests this is not the case (reviewed by Campbell and Hope, 2015). From Barré-Sinoussi et al. (2013).

isolated from a handful of individuals in the same region and group P has been found in only two individuals. Phylogenetic evidence shows M and N each arose from cross-species transmission around the beginning of the 20th century (Korber et al., 2000) from chimpanzees of the West African subspecies *Pan troglodytes troglodytes*, where the virus reservoir is the simian immunodeficiency virus SIVcpz *Ptt* (Gao et al., 1999; Keele et al., 2006). Phylogenetically distinct SIVcpz isolates have also been found in the SIVcpz *Pan troglodytes schweinfurthii*. Groups O and P have arisen from transmission of SIVgor (infecting gorillas) to humans, which itself originates from SIVcpz *Ptt* (D'arc et al., 2015; Plantier et al., 2009; Van Heuverswyn et al., 2006). SIVs have now been found infecting more than 40 non-human primate hosts, mostly African monkey species, where cross-species transmission appears common (Sharp and Hahn, 2011). HIV type 2 (HIV-2) has arisen from multiple transmissions of SIVs infecting sooty mangabeys (Hirsch et al., 1989).

Beyond primates, lentiviruses have been identified infecting some domesticated species (horses, cows, sheep, goats and cats), several species wild cats and hyenas (VandeWoude and Apetrei, 2006) and wild ibexes (Erhouma et al., 2008). The discovery of endogenous lentiviruses in several mammalian genomes (e.g. RELIK; Katzourakis et al., 2007) suggests an ancient origin and therefore the distribution of lentiviruses across wild mammal species may be presently underestimated (Gifford, 2012).

Lentiviruses are among the fastest evolving organisms known, with a mutation

rate of 4×10^{-3} per base per replication (Cuevas et al., 2015). Mutations are introduced by: (1) base incorporation errors by the viral reverse transcriptase; (2) errors by the host RNA polymerase III (which produces viral genomic RNAs from integrated proviruses); and (3) editing by the host restriction factor APOBEC3G (which causes G→A hypermutation via cytosine deamination; Mangeat et al., 2004). Further variation is generated by recombination, caused by the reverse transcriptase switching between genomic RNA templates. Positive diversifying selection has been observed across the genome, driven at least in part by escape from the cytotoxic adaptive immune response, where cells present viral antigens ('epitopes') on their surface to signal that they are infected (Phillips et al., 1991; Yang et al., 2003). The antibody response also drives continual diversifying selection in the exposed envelope protein (Nielsen and Yang, 1998; Joos et al., 2007) and for this reason developing a protective vaccine is extremely challenging (Baum, 2010).

1.3.4 HIV Selection Analyses

Molecular evolution techniques have been applied to study selection in HIV-1 previously, using codon models like those of Nielsen and Yang (1998) which employ d_N/d_S ratios. These studies have mostly sought signatures of diversifying selection as the virus tries to escape the adaptive immune response, such as continual substitutions at epitopes in viral proteins recognised by cytotoxic lymphocytes (Yang et al., 2003; Price et al., 1997). The same approaches have been used to identify antibody binding sites in HIV-1 proteins, which can inform vaccine design and even show diversifying selection in the viral population infecting a single patient over time (Nielsen and Yang, 1998; de Oliveira et al., 2004). d_N/d_S methods have also been used to identify selection for resistance to antiretrovirals (Chen et al., 2004), and modeified d_N/d_S codon substitution models have been developed specifically to identify the episodic positive selection associated with viral drug resistance (Murrell et al., 2012).

1.3.5 Interaction with Host Restriction Factors

Recent research has found a suite of innate immunity proteins which restrict the growth of viruses in cells, forming an intracellular immune response. Often these are expressed in response to interferon (IFN), an immune signalling factor, which is

itself expressed in response to the cell 'sensing' the presence of a pathogen, through recognition of pathogen associated molecular patterns (PAMPS; Lee, 2013). Restriction factors target different stages of the lentivirus life cycle (fig. 1.2; Kirchhoff, 2010; Malim and Emerman, 2008). In turn, primate lentiviruses have evolved accessory proteins whose primary function appears to be mitigating the cellular immune response, often by direct binding to a restriction factor and targeting it for degradation by cellular mechanisms (Sheehy et al., 2003; Neil et al., 2008; Van Damme et al., 2008; Hrecka et al., 2011; Laguette et al., 2011).

Recognition of a restriction factor by an accessory protein will presumably negatively affect the fitness of the host. Studies using molecular evolution techniques have found that all known restriction factors targeting the lentiviruses show evidence of having undergone positive diversifying selection. Significantly, sites under positive selection often correlate with the region of the protein bound by the accessory protein, and in some cases the positive selection analysis has guided the mutagenesis experiments which identified the region of interaction (Sawyer et al., 2004, 2005; Gupta et al., 2009; McNatt et al., 2009; Laguette et al., 2012; Lim et al., 2012). Often striking species-specificity of the interaction is observed. Diversifying selection in these regions is consistent with the restriction factor continually changing to evade recognition, matched by change in the accessory protein to restore it. This is called the Red Queen hypothesis (Van Valen, 1973), after the character in Lewis Carroll's *Through the Looking Glass and What Alice Found There* who tells Alice to run as the world rushes past them so they can stay in the same position. Similarly, the restriction factor continually changes in this regions to maintain the same fitness (reviewed by Duggal and Emerman, 2012). The difference in evolutionary timescales means a single virus lineage may not be responsible for the effect on restriction factor evolution, but instead successive infections throughout evolutionary history.

1.3.6 The Present Work

In this work, we wish to apply techniques from the study of molecular evolution to investigate the evolution of lentiviruses and interactions with their hosts.

Previous studies have identified positive diversifying selection in restriction factors in primates which can restrict lentivirus replication. It has been implicitly

assumed that primate lentiviruses and their well characterised accessory proteins have been responsible (Laguette et al., 2012; Lim et al., 2012). We have investigated that assumption by testing for positive selection in the restriction factor and metabolic enzyme SAMHD1 from more diverse species, including groups of mammals and other vertebrates. We find this adaptive evolution has occurred in almost every group analysed, indicating primate lentiviruses are not the only possible cause. In mammals, we find a clustering of sites under positive selection around the region of SAMHD1 involved in regulating this enzyme's activity which leads us, unexpectedly, to the hypothesis that nonsynonymous changes have occurred to modulate its catalytic activity.

The lentivirus capsid plays several important roles in the virus life cycle (Campbell and Hope, 2015) and has been implicated in interactions with cellular immunity (Rasaiyaah et al., 2013). The HIV-1 M transmitted from SIVcpz within the past century (Gao et al., 1999; Korber et al., 2000) and has become pandemic. We are interested in the adaptations in capsid following transmission to humans and have used swMutSel, a relatively new approach, to identify sites in capsid which are experiencing different selective constraints in the two virus groups. We identify sites in regions bound by host cofactors and potentially relevant to the dynamic structure of the protein, suggesting host-specific behaviour.

To date, diversifying selection in host restriction factors has been well characterised, consistent with the Red Queen hypothesis. Few studies concentrate on the evolution of the viral accessory proteins which antagonise them, however. We have investigated the accessory proteins of HIV-1 M and SIVcpz for indications of host-specific selective constraints, again with swMutSel. We have identified sites known to be in regions binding to host factors, suggesting species-specificity in these interactions and adaptation following transmission to humans.

Finally, to comprehensively study selection in lentivirus genomes we must include the overlapping coding sequences. No existing approaches are adequate for this situation, however, as swMutSel and the d_N/d_S approaches described above are codon substitution models and incapable of accounting for selection on multiple frames. Studying adaptive evolution in overlapping coding sequences therefore requires a novel approach. We have developed several models for this purpose and have assessed their performance in simulation studies.

Chapter 2

Positive Selection in Chordate SAMHD1

2.1 Summary

SAMHD1 is a deoxynucleoside triphosphorylase, which becomes catalytically active in a tetrameric form. Phosphorylation of threonine (T) 592 destabilises tetramers and reduces catalytic activity. It is also considered a restriction factor of HIV-1, which it prevents infecting some cell types by reducing dNTP levels below those necessary for reverse transcription. Some primate lentiviruses possess accessory proteins which relieve SAMHD1 restriction by targeting it for proteosomal degradation. Positive selection has been observed in primate SAMHD1, and this has been attributed to selection pressure imposed by primate lentiviruses. We have performed a comprehensive positive selection analysis of chordate SAMHD1 and find positive selection across all mammals and every clade tested within mammals. We observe a significant clustering of sites under positive selection around T592, suggesting adaptation in the propensity for tetramerisation. We also observe positive selection in still more divergent animals, including birds and fish. Our results indicate positive selection in SAMHD1 is a broader phenomenon than previously realised — extending to all chordates — and may be explained by 'fine-tuning' of tetramerisation regulation or antagonism by undiscovered animal viruses.

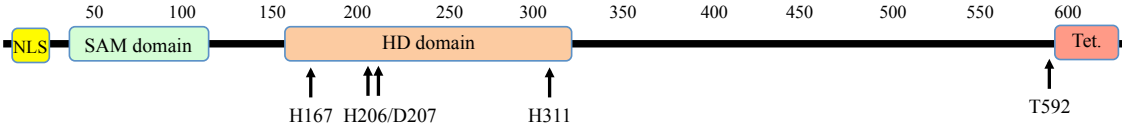

Figure 2.1: Schematic illustration of the SAMHD1 protein and its functional domains. Numbers indicate amino acid positions of human SAMHD1 protein. The SAM domain, HD domain, nuclear localisation signal ('NLS') and the C-terminal region essential for tetramerisation ('Tet.') are indicated. Four residues required for catalysis and the phospho-acceptor threonine residue at site 592 are also shown.

2.2 Introduction

2.2.1 Identification of SAMHD1

The *SAMHD1* gene was first identified in a cDNA library screen in dendritic cells, where it was described as an orthologue of a mouse interferon (IFN)-γ induced gene, *MG11* (Li et al., 2000). Attention from the immunology community came when disruptive mutations in *SAMHD1* were linked with Aicardi-Goutières syndrome (AGS), a Mendelian genetic disease whose symptoms mimic congenital virus infection, including increased IFN-α production (Rice et al., 2009). Other genes linked to AGS were known to have roles in nucleic acid degradation, which had suggested the disease was caused by accumulation of nucleic acids leading to an autoimmune response (Crow et al., 2006). SAMHD1 was therefore suggested to be similarly involved in nucleic acid metabolism (Rice et al., 2009). It posses an HD domain spanning residues 115-562 (see fig. 2.1), making it a member of the HD domain superfamily of metal-dependent phosphohydrolases. SAM (sterile α motif) domains are found across eukaryotes and mediate protein-protein interactions in complex-formation or binding of RNA (Qiao and Bowie, 2005).

2.2.2 Tetrameric SAMHD1 as a deoxynucleoside triphosphorylase

The suspected nucleic acid degrading activity was later confirmed on finding that SAMHD1 hydrolyses deoxynucleoside triphosphates (dNTPs) [1], yielding a deoxynu-

[1] A note on terminology. A nucleoside refers to a nitrogenous base (e.g. adenine, cytosine, guanine or thymine) covalently bonded to the 5-carbon sugar ribose. If the alternative sugar deoxyribose (possessing one less hydroxyl group) is used, the molecule is a deoxynucleoside. The (deoxy)nucleoside can bond to one or more covalently linked phosphate groups, yielding a (deoxy)nucleoside mono-, di- or triphosphate. The abbreviation (d)NTP refers to (deoxy)nucleoside triphosphate, containing any base. dNTPs are the substrate for synthesis of the polymer DNA

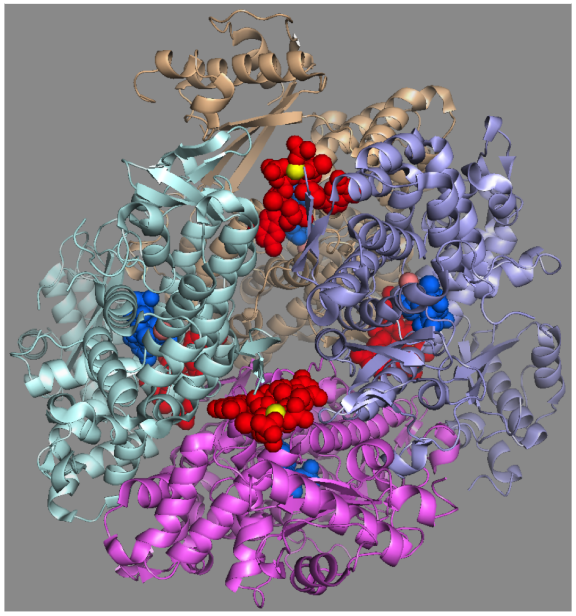

Figure 2.2: Tetrameric SAMHD1 (residues 114-599) bound to dGTPs. Monomers are shown as ribbons and coloured pale cyan, wheat, purple and pink respectively. dGTPs at the allosetic sites (two for each monomer) are shown as spheres in red, with their connecting magnesium ions shown as yellow spheres. dGTPs in active sites are shown as blue spheres, away from the tetramer interfaces, with their associated zinc ions shown as brown spheres. PDB ID: 4BZC; Ji et al. (2013).

cleoside and inorganic triphosphate (Goldstone et al., 2011).

Catalytically active SAMHD1 is a tetramer, which assembles in the presence of dNTPs and guanosine-containing molecules (either dGTP or GTP; Yan et al., 2013; Ji et al., 2014) and is therefore an enzyme regulated by the concentration of its substrate. Crystal structures of the catalytic core show tetramers comprising two juxtaposed allosteric sites and one catalytic site for each monomer (fig. 2.2; Ji et al., 2013). Tetramerisation induces dramatic conformational changes at the subunit interfaces, allowing tighter packing between monomers, in turn inducing alterations at the catalytic sites. While monomeric or dimeric forms are capable of some dNTPase activity, they have much lower enzymatic efficiency (Arnold et al., 2015). The C terminal end of SAMHD1 folds into a 'major lobe' and a 'minor lobe', and the latter is involved in protein-protein interactions between monomers which stabilise the tetramer (Goldstone et al., 2011).

(dexoyribonucleic acid), while NTPs such as GTP (whose base is guanine) are used to synthesise RNA (ribonucleic acid) and also act as signalling molecules or energy carriers. Triphosphorylation involves hydrolysis of a (deoxy)nucleoside triphosphate to produce a (dexoxy)nucleoside and inorganic triphosphate.

2.2.3 Non-substrate regulation of SAMHD1 activity

SAMHD1 appears to be almost ubiquitously expressed[2], its mRNA having been identified in 25 diverse human tissues and protein in several of these (Schmidt et al., 2015). Several studies have investigated SAMHD1 induction by interferons and found different responses in different cell types: while IFN-induced mRNA upregulation has been reported in microarray analysis of peripheral blood mononuclear cells (PBMCs; Crow and Wohlgemuth, 2003), and increased protein expression observed in the HEK 293T, HeLa cell lines and monocytes (Berger et al., 2011), no change in protein levels are observed in dendritic cells, unactivated CD4+ T cells or monocyte-derived macrophages (St Gelais et al., 2012; Schmidt et al., 2015).

SAMHD1 phosphorylation is more significant. Threonine-592 has been identified as a phospho-acceptor site, reported to be phosphorylated by CDK1 or CDK2 (Cribier et al., 2013; Pauls et al., 2014; Yan et al., 2015). CDKs and their associated cyclins mediate control of cell division and their phosphorylation of SAMHD1 appears to regulate its activity throughout the cell cycle. Mutating residues such that SAMHD1 phosphorylation is inhibited was found to reduce dATP levels in cycling cells and was associated with lower incorporation of labelled bases in newly synthesised nuclear DNA (Yan et al., 2015). This suggests SAMHD1 phosphorylation downregulates its dNTPase activity, which is timed to prevent dNTPase activity during S phase, when nuclear DNA is replicated. While Cribier et al. (2013) reported that type 1 IFN reduced T592 phosphorylation in PBMCs and CD4+ T cells (suggesting SAMHD1 is activated in response to virus infection), phosphorylation was not observed in response to any of a panel of cytokines, including interferons, by Schmidt et al. (2015).

Phosphorylated SAMHD1 (or a phosphomimetic T592E/D mutant) retains its dNTPase activity when in the catalytically active tetrameric form, with similar k_{cat} and K_M values (Tang et al., 2015; Arnold et al., 2015). The stability of the phosphorylated tetramer, however, is severely compromised, with the majority of species in the weakly active monomeric or dimeric forms. The structural environment of T592 is predominately negatively charged and therefore the introduction of the negatively charged phosphate group is suggested to provoke significant structural rearrangements. Indeed, the whole C terminal region (residue 581 onwards) is disordered in

[2]A notable exception are activated CD4+ T cells, with great relevance for HIV-1 infection (Ruffin et al., 2015); discussed in §2.2.4.

phospho-SAMHD1 or $SAMHD1_{T592E}$ and this appears to disrupt interactions which stabilise the tetramer, leaving the dNTPs at allosteric sites more exposed to solvent (Tang et al., 2015) and possibly therefore more likely to disassociate at lower dNTP concentrations (Arnold et al., 2015).

Residues in the SAMDH1 C terminus determine its propensity for phosphorylation and affect the stability of the unphosphorylated tetramer. Yan et al. (2015) found deleting the C terminal 31 residues (595-626) abolishes both phosphorylation of T592 in cells and interaction with cyclinA2, which associates with the CDKs to mediate the modification; mutation of specific residues in the C terminus also abolish or weaken cyclinA2 interaction. Deleting the C terminal 43 residues (583-626) abolishes tetramer formation and dNTase entirely (Yan et al., 2013; Arnold et al., 2015).

2.2.4 SAMHD1 and HIV-1 restriction

Recent interest in SAMHD1 biology has been motivated by its identification as a restriction factor of human immunodeficiency virus type 1 (HIV-1; Hrecka et al., 2011; Laguette et al., 2011). While HIV-1 replicates poorly in monocytes, dendritic cells, nondividing macrophages and quiescent CD4+ T cells, HIV-2 and SIVsm (infecting sooty mangabes) suffer no such block, and this is dependent on their accessory protein Vpx (Goujon et al., 2007, 2008; Baldauf et al., 2012). HIV-1 does not possess a Vpx gene and supplying Vpx during HIV-1 infection allows it and similarly restricted retroviruses to replicate successfully in these cells (Goujon et al., 2006). Vpx interacts with the complex responsible for targeting proteins for proteosomal degradation with ubiquitin, comprising the DDB1 and DCAF1 subunits of the Cullin-4 based E3 ubiquitin ligase machinery (Srivastava et al., 2008; Bergamaschi et al., 2009). Vpx is therefore an antagonist, acting as an adaptor which binds at separate interfaces both DCAF1 and SAMHD1, targeting it for destruction (fig. 2.3; Schwefel et al., 2014).

Hrecka et al. (2011) and Laguette et al. (2011) had shown SAMHD1 hinders the accumulation of viral DNA and the latter authors found mutating the putative catalytic H/D residues abolished restriction, suggesting either nuclease or dNTPase activity was responsible. Lahouassa et al. (2012) went on to show that SAMHD1-mediated restriction could be relieved in primary monocyte-derived macrophages by

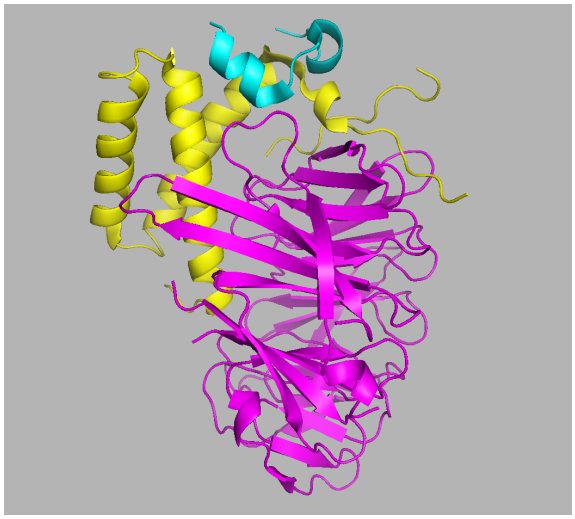

Figure 2.3: Crystal structure of C terminal SAMHD1 (residues 606-624, cyan), SIVsm Vpx (yellow) and human DCAF1 (residues 1073-1392, magenta). PBD ID: 4CC9, Schwefel et al. (2014).

supplying exogenous dNTP precursors, concluding that SAMHD1 reduced the dNTP concentration to below the level necessary for viral cDNA synthesis. Moreover, mutations in the HIV-1 reverse transcriptase making it less processive or conferring lower affinity for dNTPs leave the virus more susceptible to SAMHD1 restriction, further suggesting SAMHD1's control of dNTP concentration is directly responsible (Lahouassa et al., 2012; Arnold et al., 2015).

The dNTPase model of restriction was questioned, however, when phosphorylation of SAMHD1 appeared to negatively regulate HIV-1 restriction without affecting dNTPase activity (White et al., 2013). Ryoo et al. (2014) reported RNase activity by SAMHD1 was responsible for restriction, while its dNTPase function was dispensable, and proposed that SAMHD1 prevents reverse transcription by degrading the viral RNA template. This model naturally requires the enzyme to possess RNase activity, but this is disputed (Goldstone et al., 2011; Powell et al., 2011; Goncalves et al., 2012) and reports of it may be attributable to incomplete SAMHD1 purification (Seamon et al., 2015; Arnold et al., 2015). The restriction model of Arnold et al. (2015) proposes that stable SAMHD1 tetramers inhibit reverse transcription by keeping dNTP concentration low and that SAMHD1 phosphorylation causes tetramers to disassemble; though limited dNTPase activity is maintained by phosphorylated SAMHD1, perhaps explaining the observations of White et al., it is not sufficient for restriction.

It is perhaps surprising that HIV-1 does not antagonise SAMHD1 like so many

of its relatives. This may be explained by a lower sensitivity to SAMHD1 restriction, as the reverse transcriptase of HIV-1 is more efficient than that of HIV-2, allowing replication in environments with lower dNTP concentration (Lenzi et al., 2015). SAMHD1 expression in myeloid cells may reduce rather than ablate HIV-1 infection and indeed these cells may still play a significant role in HIV-1 infection, as they have greater longevity compared with CD4+ T cells and contact other cells in antigen presentation. They may therefore allow efficient onward dissemination despite being inefficiently infected themselves (Aggarwal et al., 2013). Restricted replication in myeloid cells may even be advantageous, by preventing their potent innate immune sensing mechanisms being activated and thus avoiding an immune response (Laguette et al., 2011; Ballana and Esté, 2015).

2.2.5 SAMHD1 and restriction of other exogenous or endogenous viruses

Restriction by SAMHD1 need not be specific to HIV, as all retroviruses and DNA viruses require dNTPs. Reverse transcription by both the alpha retrovirus Rous sarcoma virus and the beta retrovirus Mason-Pfizer monkey virus is also inhibited in the presence of human SAMHD1 (Gramberg et al., 2013); interestingly, human T-lymphotropic virus (HTLV)-1 was reported to be insensitive, perhaps because it has a means of mitigating SAMHD1-mediated restriction, though there was no evidence of its being degraded. Choi et al. (2015) found the lentiviruses feline immunodeficiency virus (FIV) and equine infectious anemia virus (EIAV), as well as the gamma retrovirus Friend murine leukemia virus (FMLV), were all susceptible to restriction by human SAMHD1 in a myeloid cell line (U937). Restriction was reported to be dependent on SAMHD1's putative RNase activity, though restriction of RNA viruses (other than retroviruses) was not observed.

The DNA viruses vaccinia virus and herpes-simplex virus (HSV)-1 replicate more efficiently in the absence of SAMHD1 expression in primary MDMs or DCs, respectively (Hollenbaugh et al., 2013; Kim et al., 2013), though its expression does not entirely restrict infection. Neither virus degrades SAMHD1 but each encode their own nucleoside biosynthesis enzymes, perhaps mitigating the effect of SAMHD1-mediated dNTP depletion.

SAMHD1 has also been shown to restrict proliferation of the LINE-1 retrotrans-

poson, by a mechanism associated with a block to reverse transcription (Zhao et al., 2013b). The effect was also seen in the presence of SAMHD1 from other mammals, suggesting repression of retrotransposition is a conserved function. However, mutation at the SAMHD1 catalytic site (D311A) was reported to have no effect on LINE-1 restriction and the authors therefore suggest retrotransposition is inhibited by a different means to HIV-1, i.e. other than by dNTP depletion. Hu et al. (2015), by contrast, report that dNTPase activity is required for LINE-1 repression, and therefore more work is needed to clarify the underlying mechanism.

2.2.6 Coevolution of primate lentiviruses and primate SAMHD1

Several nonhuman primate lentiviruses encode Vpx orthologues, including SIVsm (infecting sooty mangabey), SIVmac (macaque, derived from SIVsm), SIVrcm (red-capped mangabey) and SIVmnd2 (mandrill). Studies employing a mixture of phylogenetic modelling and experimental work have uncovered striking adaptation of viral antagonists to their host species' SAMHD1, and positive (diversifying) selection acting on SAMHD1 in turn. Using a site-specific model (site model 8 implemented in codeml; Yang, 2007), Laguette et al. (2012) identified 17 sites in primate SAMHD1 as being under positive selection, five of which in the C terminus, which the authors then confirmed is required for degradation by SIVmac Vpx. A more complex picture emerged, however, as species-specific antagonism was revealed: for example, human SAMHD1 is successfully degraded by Vpx of HIV-2 and SIVsm/mac, but not by SIVrcm or SIVmnd2 Vpx.

Using the same model, Lim et al. (2012) identified six sites as under positive selection in Old World monkey SAMHD1, clustering at the N-terminus. They too had observed species-specificity of degradation, with, for example, SIVmnd2 Vpx unable to degrade African green monkey (AGM) SAMHD1 and SIVagm.Gri (grivet African green monkey) unable to degrade mandrill SAMHD1. Residues at sites 46 and 69 — predicted to be under positive selection — correlated with degradation susceptibility: while mandrill SAMHD1 appeared to have 'ancestral' residues (compared with other species), AGM SAMHD1 had changed at these sites. Mutating AGM SAMHD1 to the ancestral residues conferred susceptibility to SIVmnd2 Vpx, while mutating mandrill SAMHD1 to the AGM residues made it resistant to SIVmnd2 Vpx, suggesting the changes in AGM SAMHD1 were driven by pressure from Vpx

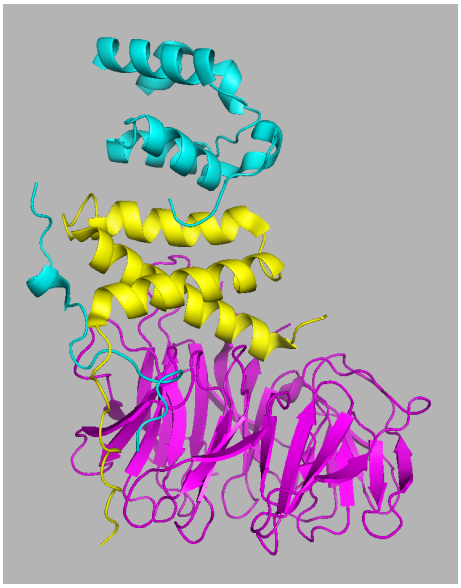

Figure 2.4: Crystal structure of the N-terminus of mandrill SAMHD1 (showing residues 1-22, 34-88 and 93-109) bound to SIVmnd2 Vpx and human DCAF1 (showing residues 1075-1392). Colours as in fig. 2.3. PDB ID: 5AJA (Schwefel et al., 2015).

proteins.

SAMHD1 antagonism is not exclusive to Vpx-endowed viruses, as several SIV Vpr proteins are capable of SAMHD1 degradation by the same mechanism (Lim et al., 2012). Vpx is believed to have arisen by duplication of Vpr (Tristem et al., 1990), and indeed Lim et al. (2012) demonstrate by rooting of the Vpr/Vpx phylogenetic tree that SAMHD1 antagonism is a monophyletic phenotype, indicating that it arose once, prior to the creation of Vpx.

Striking data from Spragg and Emerman (2013) show that SAMHD1 antagonism may be specific not only to the host species, but to SAMHD1 variants within populations. SAMHD1 was found to be polymorphic within all four species of African green monkey (which diverged from a common ancestor less than 3 million years ago) and each species is infected by a distinct SIV. No one Vpr from these viruses could degrade all SAMHD1 variants, but each could degrade the most abundant variant in their own host species. That this variation in the host SAMHD1 has apparently arisen and been maintained in these populations over relatively short time suggests antagonism significantly affects fitness and that the 'arms race' between restriction factor and antagonist is ongoing.

Interestingly, while SAMHD1 antagonism has been conserved in several primate lentiviruses, the site of interaction has been more flexible. HIV-2 and SIVmac Vpx require the C-terminus of their target SAMDHD1 for successful binding and degra-

dation, while SIVmnd2/SIVrcm Vpx and SIVsyk/SIVmus Vpr bind the N-terminus (Fregoso et al., 2013). Furthermore, the determinants for host-specificity can be mapped to these domains. Figure 2.4 shows the N-terminus of mandrill SAMHD1 in complex with SIVmnd2 Vpx and DCAF1, in contrast to the C terminal binding by SIVsm Vpx in figure 2.3. The divergent interaction positions could account for the mix of sites found to be under positive selection at both the N and C termini. Fregoso et al. (2013) propose a model whereby Vpr/Vpx proteins have 'toggled' between binding the N and C termini in response to substitutions in SAMHD1 which confer protection from antagonism. Because these ends are adjacent in tetrameric SAMHD1, each Vpr/Vpx may have strong binding to one terminus and weak binding to the other, such that sequence change at the preferred binding site does not destroy SAMHD1 binding entirely and adaptation of the antagonist restores strong interaction.

2.2.7 Function of SAMHD1

The body of literature describing SAMHD1 various activities leads to the question, what is SAMHD1's primary function? What is it for?

SAMHD1 may be best described simply as a metabolic regulator, rather than an innate immunity protein. Franzolin et al. (2013) take this view, having observed that silencing SAMHD1 expression slows cell proliferation and radically skews the relative proportions of the four dNTPs, which can increase mutation rates in DNA replication. These observations suggest a role with little connection to virus infection or autoimmunity. Whether SAMHD1 should be called a restriction factor which regulates dNTP levels, or a dNTP regulator which restricts viruses, will be decided by examining its effects in wider ranges of cell types and environments.

2.2.8 The present work

SAMHD1 has been shown to be under positive (diversifying) selection in primate species (Laguette et al., 2012; Lim et al., 2012) and previously this has been specifically attributed to an evolutionary 'arms race' with primate lentivirus antagonists. We are interested to know whether this signature of positive selection is unique to *primate* SAMHD1 or common to other animal groups, and therefore part of a broader phenomenon.

We have performed a comprehensive selection analysis of SAMHD1 from mammals, clades within mammals, birds, reptiles, amphibians and marine animals (fish) and find positive selection in the majority of the groups studied. In our analyses of mammals, the majority of sites identified as under positive selection cluster at the C-terminus, in close proximity to the phosphorylation site T592. We speculate that (1) unidentified viral antagonists have driven diversifying selection, and/or (2) the signature has resulted from continual calibration of SAMHD1's dNTPase activity.

2.3 Results

2.3.1 Positive selection across mammals

To investigate the selective pressures acting on SAMHD1 throughout its evolution in mammals, we compiled a dataset of publicly available SAMHD1 coding sequences using NCBI BLAST [3] (Ostell and McEntyre, 2007), listed in appendix A. These searches also yielded HD domain sequences from highly divergent organisms (including plants), indicating that all available animal SAMHD1 sequences were within the scope of the search. These were aligned as codon sequences and a phylogenetic tree was estimated by maximum likelihood. Our initial analyses focussed on mammalian SAMHD1 and other sequences were not included (fig. 2.5; see methods). The mammal SAMHD1 tree was broadly concordant with the previously reported mammalian phylogeny (Meredith et al., 2011) and the majority of nodes had non-parametric bootstrap support values above 70%.

Positive selection identified

We employed the site-specific selection models implemented in codeml of the PAML package (Nielsen and Yang, 1998; Yang, 2007). The M1a (null) model permits classes of sites under negative (purifying) and neutral selection, while the M2a (alternative) model also allows a class of sites evolving under positive selection for which $\omega > 1$ (i.e. $dN/dS > 1$). M2a can be used to compute the posterior probability of each site belonging to the positive selection class, according to the Bayes empirical Bayes calculation (Yang et al., 2005). To guard against identification of local optima by the optimisation routine, analyses were performed multiple times with different initial parameter values (see methods). We also attempted analyses of mammalian SAMHD1 with the more complex M7 (null) and M8 (alternative) models which permit distributions of ω values, but M8 optimisations consistently failed to converge; we therefore discuss results from models M1a and M2a only.

Using the likelihood ratio test, the alternative M2a model had a significantly superior fit to the data, supporting the alternative hypothesis of positive selection in mammalian SAMHD1 ($p << 0.01$; appendix B). 36 of 626 sites had posterior probability > 0.95 of belonging to the positive selection class (table 2.1). These

[3] Sequence collection was done by Christopher Ruis. Alignment was done by Christopher Monit and Christopher Ruis together.

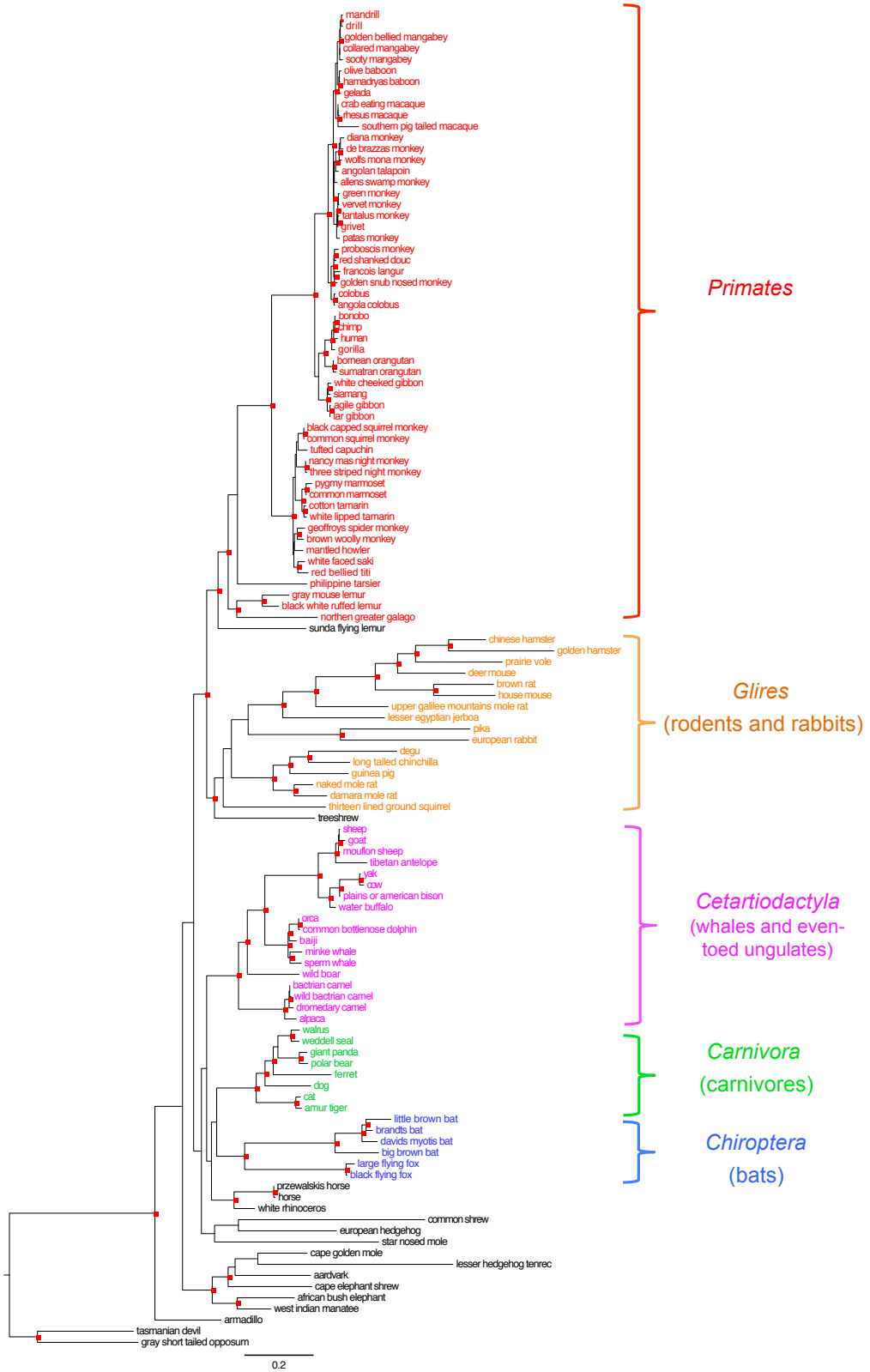

Figure 2.5: Maximum likelihood phylogeny for mammalian SAMHD1, estimated with RAxML. Red squares mark nodes supported by $\geq 70\%$ non-parametric bootstrapping. Clades of five groups used in some analyses are highlighted and remaining mammal species are shown in black. Branch lengths expressed as expected nucleotide substitutions per codon, estimated by codeml model M0 (Yang, 2007) with this topology. When studying specific mammals groups, irrelevant taxa were either pruned to yield subtrees comprising only the taxa of interest (site-specific analyses) or the branches of interest were set as foreground (branch-site-specific analyses).

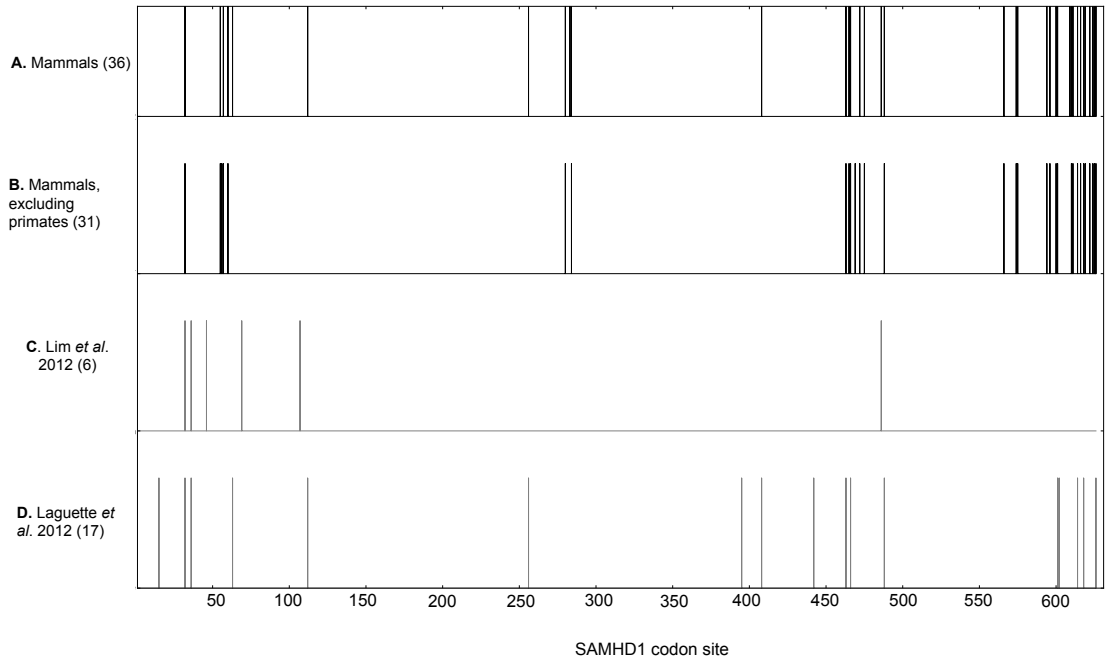

Figure 2.6: Codon sites under positive selection in mammalian SAMHD1. Panels represent the linear sequence and vertical lines mark sites where the probability (Bayes empirical Bayes calculation) is greater than 0.95 that the site is under positive selection; in parentheses are the total numbers of sites thus identified. Site numbering is based on the human sequence. The height of the lines is of no significance. **(A)** Results from analysis of all mammals with codeml's model M2a. **(B)** Similar analysis of mammals with primates excluded. **(C, D)** Published results of Lim et al. (2012) and Laguette et al. (2012), both from analyses of primate SAMHD1.

were distributed amongst four clusters across the SAMHD1 sequence (fig. 2.6a): at the N-terminus (between sites 32 and 112), in the HD domain (between 256 and 284), central domain (between 408 and 488) and at the C-terminus (between sites 566 and 626).

Sites in these regions had previously been identified as being under positive selection in primate SAMHD1 (Lim et al., 2012; Laguette et al., 2012; table 2.1). Therefore to test whether the positive selection signal identified in our mammalian dataset could be attributed to primates, we repeated the site-specific analysis with primate clade excluded (pruned from the same topology) as a control. Again, model M8 failed to converge and we discuss the results from the M1a/M2a analysis only. The alternative model was statistically justified ($p << 0.01$) and 32 sites were identified as under positive selection in non-primate mammals (fig. 2.6b; table 2.1). Of these, 30 had been identified in the analysis of the whole mammals dataset. These data indicate that non-primate mammalian SAMHD1 is under positive selection.

Site	Res.	Ma.	No Pr.	Cn.	Cp.	Gl.	Pr.	Wu.	Re. Ma.	Lag.	Lim
3	R				+						
15	C									+	
32	W	+	+							+	+
36	L									+	+
46	G										+
52	S	+						+			
55	R	+	+		+						
56	R	+	+			+		+			
57	G	+	+						+		
60	E	+	+		+	+					
62	P							+			
63	V	+	+							+	
67	N							+			
68	I				+						
69	R										+
79	P				+						
106	R							+			
107	L										+
112	V	+								+	
114	T				+						
115	M				+						
256	Q	+			+		+			+	
266	C				+						
280	V	+	+						+		
282	D								+		
283	S	+			+		+				
284	L	+	+			+			+		
346	E				+						
395	D									+	
405	K							+			
408	R	+					+	+		+	
442	R						+			+	
463	T	+	+		+			+		+	
464	G							+			
465	Q	+	+								
466	I	+	+					+		+	
469	K		+		+						
472	D	+	+		+						
473	Y				+						
475	S	+	+	+				+	+		
482	S			+							
486	K	+									+
488	L	+	+		+	+			+	+	
563	Y				+						
566	R	+	+					+			
574	A	+	+		+				+		
575	D	+	+						+		
585	D							+			
586	V							+			
588	A							+			
594	Q	+	+					+			
596	K	+	+		+			+	+		
600	D	+	+					+			
601	S	+	+	+						+	
602	T									+	
604	V							+			
605	Q					+					
609	R	+									
610	L	+	+			+					
611	R	+	+								
614	S	+	+		+			+	+	+	
616	S	+	+		+			+	+		
618	V	+	+	+				+		+	
619	Q	+	+		+						
622	K	+	+		+			+			
623	D				+			+			
624	D	+	+		+						
625	P	+	+			+					
626	M	+	+		+	+	+	+		+	

Table 2.1: Sites identified in mammalian SAMHD1, using codeml site models M2a or M8. '+' indicates Bayes empirical Bayes probability > 0.95. Res., human SAMHD1 residue; Ma., all mammals (M2a); No Pr., all mammals without primates (M2a); Cn., *Carnivora* (M8); Cp., *Chiroptera* (M8); Gl., *Glires* (M8); Wu., *Cetiartiodactyla* (M8); Re. Ma., remaining mammals (M8); Lag., sites from Laguette et al. (2012); Lim, sites from Lim et al. (2012).

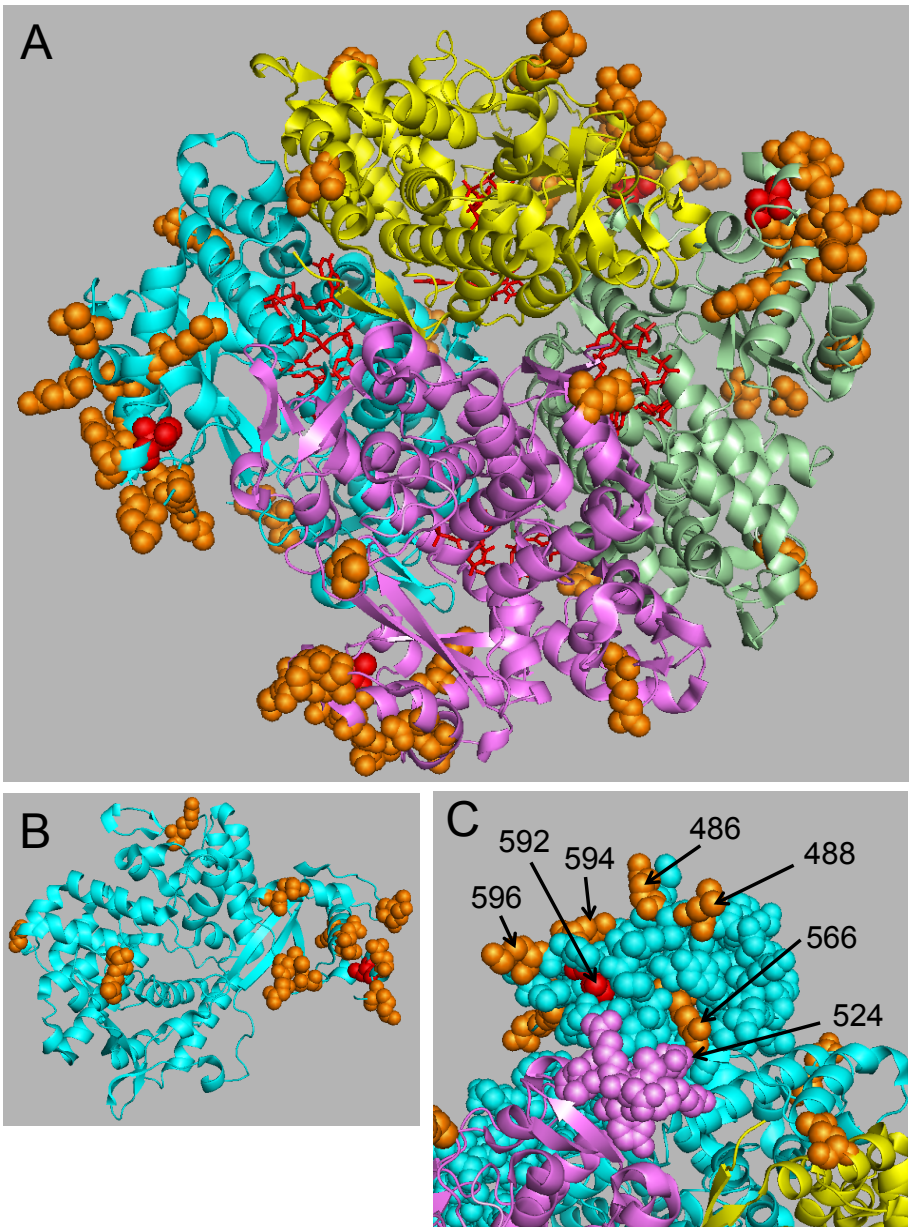

Figure 2.7: Sites identified as under positive selection in mammalian SAMHD1 shown on a published crystal structure (PDB 4TNP, Ji et al. (2014); comprising sites 114-276 and 282-599), shown in orange and threonine 592 in red, with both shown as spheres. Monomers in this structure include 15 or 16 of 36 sites identified in the analysis. **(A)** SAMHD1 tetramer, with dNTPs (bound at allosteric and catalytic sites) shown as red sticks **(B)** SAMHD1 monomer. **(C)** Enlarged view of contacting SAMHD1 molecules in tetramer; residues 480-598 (cyan) and residues 521-530 (magenta) are shown as spheres, with sites identified as under positive selection and T592 coloured orange and red, respectively; residues of special interest are labelled. Note that the labelled residue 525 is present in the neighbouring monomer (magenta).

Sites mapped onto tetramer crystal structure

To assess the biological significance of the sites identified with the mammalian SAMHD1 dataset, we mapped them onto published crystal structures. Structure 4TNP (Ji et al., 2014) comprises sites 113-276 and 282-598 of human tetrameric SAMHD1, including 16 sites identified in this analysis (fig. 2.7). The clusters of residues between sites 408-488 and 566-626, while apparently separate in the linear sequence, are brought into close proximity in the structure. Considering tetrameric SAMHD1 (fig. 2.7a), none of the identified sites were located at the interface between monomers, nor at the dNTP-binding catalytic or allosteric sites. Instead these sites are positioned on the exterior of the tetramer and many are concentrated in the C-terminal minor lobe (fig. 2.7c).

Significant clustering around T592

Noticing the preponderance of sites under positive selection around the phosphoacceptor site T592 (red in fig. 2.7), we asked whether this clustering could be expected by chance. Taking the residue coordinates for the first peptide chain in structure 4TNP (i.e. a SAMHD1 monomer, containing 15 of the identified sites), we calculated the Euclidean distance of each identified residue from T592 and then computed the harmonic mean distance (whereby a small change at close distance influences the mean more than a large change further away). Statistical significance was assessed by bootstrapping: 15 sites were randomly drawn from the set of residues present in the chain, and their harmonic mean distance calculated, 10^5 times. This yields a distribution of unbiased harmonic mean distances. The p value for clustering around T592 is then given by the proportion of the bootstrap samples whose harmonic mean distance from T592 is lower than that of the positively selected sites. The computed value was $p \approx 4 \times 10^{-4}$, indicating that sites under positive selection are disproportionately clustered around T592 [4].

Sites identified are predicted to be structurally important

Collating our positive selection data, phylogeny and the published tetramer structure, we found four sites under positive selection to be of particular biological interest based on positions in the minor lobe (fig. 2.7c) and the different chemical properties

[4] This calculation was performed using (unpublished) software written by Richard A. Goldstein.

of the residues observed. Tree topologies annotated with residues observed at these sites are presented in appendix C.

At site 566 (arginine in humans), which is mostly buried in the minor lobe, many nonconservative residue substitutions have occurred (fig. C.1) and we see residues of greatly varied size, including glycine, leucine and histidine. Site 574 (alanine in humans) is almost entirely buried in the minor lobe, with only 2 $Å^2$ exposed to solvent in structure 4TNP (Ji et al., 2014); yet residues with much larger side chains including leucine, isoleucine and even the aromatic phenylalanine (present in aadvark SAMHD1) are observed (fig. C.2). The hydrophilic serine also appears to have arisen multiple times independently.

Sites 594 and 596 (glutamate and lysine in humans, respectively) are both close to T592 (fig. 2.7c); again, chemically diverse residues are observed at these positions (fig. C.3 and C.4). Interestingly, residues at site 596 mark the divergence of New World monkeys (aspartate) from Old World monkeys and apes (lysine) by a charge difference.

Sites mapped onto Vpx-DCAF1-SAMHD1 crystal structures

Structure 4CC9 (Schwefel et al., 2014), comprises SIVsm Vpx and human DCAF1 in complex with C-terminal human SAMHD1 (residues 606-624), including nine sites identified as under positive selection (fig. 2.8a). Of these, five have side chains in close contact with Vpx residues (609, 610, 611, 618 and 622). All but one of these (609) were also identified when primate sequences were excluded from the analysis. Structure 5AJA (Schwefel et al., 2015) includes human DCAF1, SIVmnd2 Vpx and N-terminal (residues 1-22, 34-88 and 93-109) mandrill SAMHD1. This includes four sites identified as under positive selection across mammals (55, 57, 60 and 63), of which 55 is in particularly close proximity to Vpx (fig.2.8b). All but one of these (63) were also identified with primate sequences excluded. Seven sites found to be under positive selection are not present in any of the three structures presented (32, 112, 280, 600, 601, 625 and 626), only one of which (626) is not also identified when primate sequences are excluded.

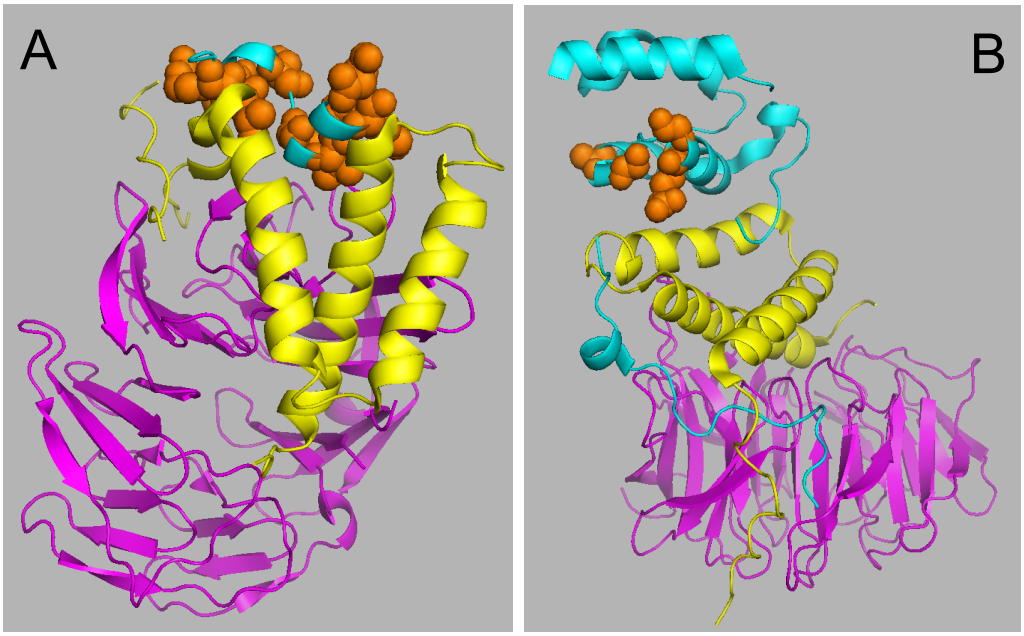

Figure 2.8: Sites identified as under positive selection in mammalian SAMHD1 (orange spheres), shown on published crystal structures of SAMHD1 (cyan) in complex with Vpx (yellow) and DCAF1 (magenta). **(A)** Human SAMHD1 (residues 606-624), Vpx from SIVsm and human DCAF1 (PDB 4CC9, Schwefel et al., 2014); includes 9 of 36 sites identified. **(B)** Mandrill SAMHD1 (residues 1-22, 34-88 and 93-109), SIVmnd-2 Vpx and human DCAF1 (PDB 5AJA, Schwefel et al., 2015); includes 4 of 36 sites identified.

Repeating with alternative tree topologies

Maximum likelihood (ML) selection analyses condition on the tree topology being exactly correct, but since topologies are themselves estimated by phylogenetic methods their accuracy cannot be guaranteed. We therefore repeated the M1a/M2a analyses with three alternative topologies, derived from non-parametric bootstrapping in the initial tree estimation (see methods), to assess the dependence of the results on the topology used. (The need to repeat each analysis with varying start values, to lower the risk of identifying local optima, precluded the use of more topologies.) The M2a model had a statistically significant better fit to the data for each of the three topologies (appendix B). Comparing the sites identified as under positive selection (having BEB estimated probability > 0.95), there was almost complete agreement when using the alternative topologies and the ML topology (table 2.2).

2.3.2 Positive selection within specific mammal groups

We next wanted to determine whether the positive selection signal could be attributed to a specific group within mammals. The mammalian SAMHD1 dataset

Topology	T_i total	ML only ($\mathrm{ML} \setminus T_i$)	Intersection ($\mathrm{ML} \cap T_i$)	T_i only ($T_i \setminus \mathrm{ML}$)
T_1	38	0 sites	36 sites	2 sites (*56, 464*)
T_2	36	3 (*112, 594, 609*)	33 sites	3 sites (*36, 56, 602*)
T_3	37	0 sites	36 sites	1 site (*36*)

Table 2.2: Comparison of sites found to be under positive selection between maximum likelihood (ML) tree topology and three alternative topologies, derived from non-parametric bootstrap sampling of the same mammalian SAMHD1 sequence alignment (see methods). Columns show sites found in ML tree analysis only (not found with alternative topology), found with both (intersection) or with alternative topology only (not found with ML topology), respectively. The three groups also represented with set notation. Where there are discrepancies, the relevant sites are given in parentheses (human sequence numbering).

was divided into 5 clades: *Primates* (55 taxa), *Carnivora* (dogs, bears etc.; 8 taxa), *Cetartiodactyla* (whales and even-toed ungulates; 18 taxa), *Glires* (rodents, rabbits and hares; 16 taxa) and *Chiroptera* (bats; 6 taxa). A sixth group comprised the remaining species not belonging to a well represented monophyletic set (17 taxa), hereafter called 'remaining mammals'. These subgroups were then subjected to the same positive selection analyses as was done for the whole dataset, using the same sequence alignments and subtree topologies. Unlike the whole dataset, model M8 successfully converged when fitted to each subgroup dataset.

Positive selection identified

The alternative models (M2a and M8) had a significantly better fit to the data than their respective null models (M1a and M7) for each of the six subgroups (appendix B) and sites with > 0.95 probability of belonging to the positive selection class were identified (fig. 2.9). For the *Carnivora* and *Primates* subgroups, sites identified with M2a were the same as with M8 and for *Chiroptera*, *Glires* and *Cetartiodactyla*, the M2a sites were a subset of the M8 sites identified; for the remaining mammals, only one site (625) was found with M2a but not M8. The M8 results are therefore prioritised in the following discussion.

The identified sites for each subgroup (table 2.1) fell into the same four clusters as sites from the whole dataset analysis, though not every cluster was represented in each subgroup's results (fig. 2.9). The *Chiroptera* (bats) had the greatest preponderance of sites, with 26 identified by M8 and a distribution across the sequence resembling that of the whole dataset. The *Carnivora* had the fewest sites (4), all of

Mammal subgroup	p value
Carnivora	0.15807
Chiroptera	0.05771
Glires	0.20767
Primates	0.87726
Cetartiodactyla	0.00129
Remaining mammals	0.03504

Table 2.3: Probabilities of the sites identified as being under positive selection clustering around threonine 592 if their distribution throughout the monomeric SAMHD1 crystal structure 4TNP (Ji et al., 2014) were random. p values were computed for the sets of sites identified in each analysis of a mammalian subgroup's SAMHD1 which also are present in 4TNP. Taxonomic groups: *Carnivora* (carnivores), *Chiroptera* (bats), *Glires* (rodents, rabbits and hares), *Primates* and *Cetartiodactyla* (whales and even-toed ungulates).

which fell at the C-terminus. The 15 sites identified in the *Cetartiodactyla* (whales and ungulates) were similarly clustered at the C-terminus, while the eight sites found in the *Glires* (rodents and rabbits) were clustered at the N and C termini, with one additional site in each of the central cluster positions. Nine sites were identified in the remaining mammals, falling in each of the four clusters except at the N-terminus. Surprisingly, only five sites were identified in the *Primates*, with none identified at the N-terminus.

Sites mapped onto crystal structures and tests for clustering around T592

We mapped the sites identified in each subgroup onto the three SAMHD1 crystal structures referred to in the previous section (fig. 2.10). In almost all subgroups, sites identified in the N- and C-terminal structures 5AJA and 4CC9 are in direct contact with Vpx proteins, with the exception of the remaining mammals, where the two identified sites are oriented away from Vpx. Considering structure 4TNP — as with the whole mammals dataset — several of the subgroup results appeared to show clustering of identified sites in the C-terminal minor lobe, in close proximity to T592. We therefore computed the probability of observing this distribution by chance, as described above. Only the *Cetartiodactyla* and remaining mammals had $p < 0.05$, though the value for the *Chiroptera* was close to this threshold (table 2.3).

Branch-site tests of positive selection within mammal groups

We were interested to compare the results found in the subgroups using models M2a/M8 with those from branch-site tests of positive selection, by using the whole

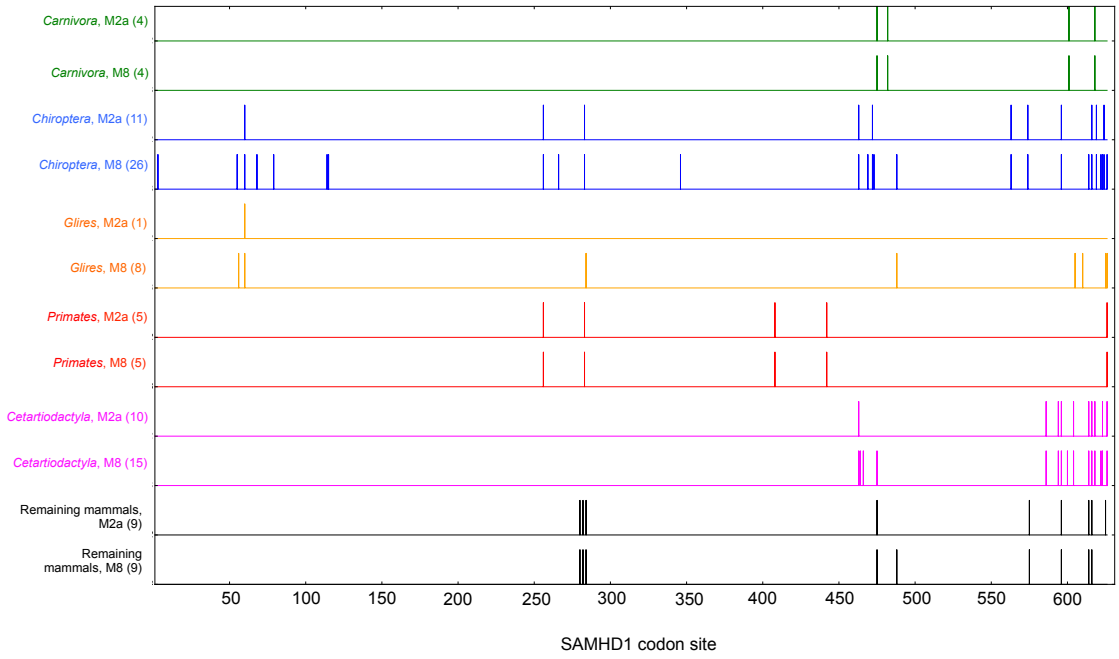

Figure 2.9: SAMHD1 sites identified as under positive selection in subgroups of mammals, presented as in figure 2.6.

Subgroup	Branch-site total	M8 only (M8 \ BS)	Intersection (M8 ∩ BS)	Branch-site only (BS \ M8)
Carnivora	3	2	2	1 (*41*)
Chiroptera	8	19	7	1 (*546*)
Glires	2	6	2	0
Primates	4	3	2	2 (*107, 539*)
Cetartiodactyla	8	10	5	3 (*443, 585, 588*)
Remaining mammals	6	6	3	3 (*235, 590, 597*)

Table 2.4: Numbers of sites identified as being under positive selection using the branch-site model, with the given group's clade/branches as foreground, compared with the M8 model. Sites identified with the branch-site model only are listed in parentheses (human sequence numbering).

mammal phylogeny with clades/branches of interest permitted $dN/dS > 1$ (set as 'foreground'; Yang and Nielsen, 2002; Zhang et al., 2005). For all subgroups, the alternative model permitting positive selection only along foreground branches was found to significantly better fit the data than the null model where $\omega = 1$ on these branches (appendix B). For each of the subgroups, fewer sites were identified as having BEB probability > 0.95 by the branch-site method than with M8 (table 2.4). Indeed most sites thus identified were also identified by M8, with only 1-3 sites being found by the branch-site model alone (listed in table 2.4).

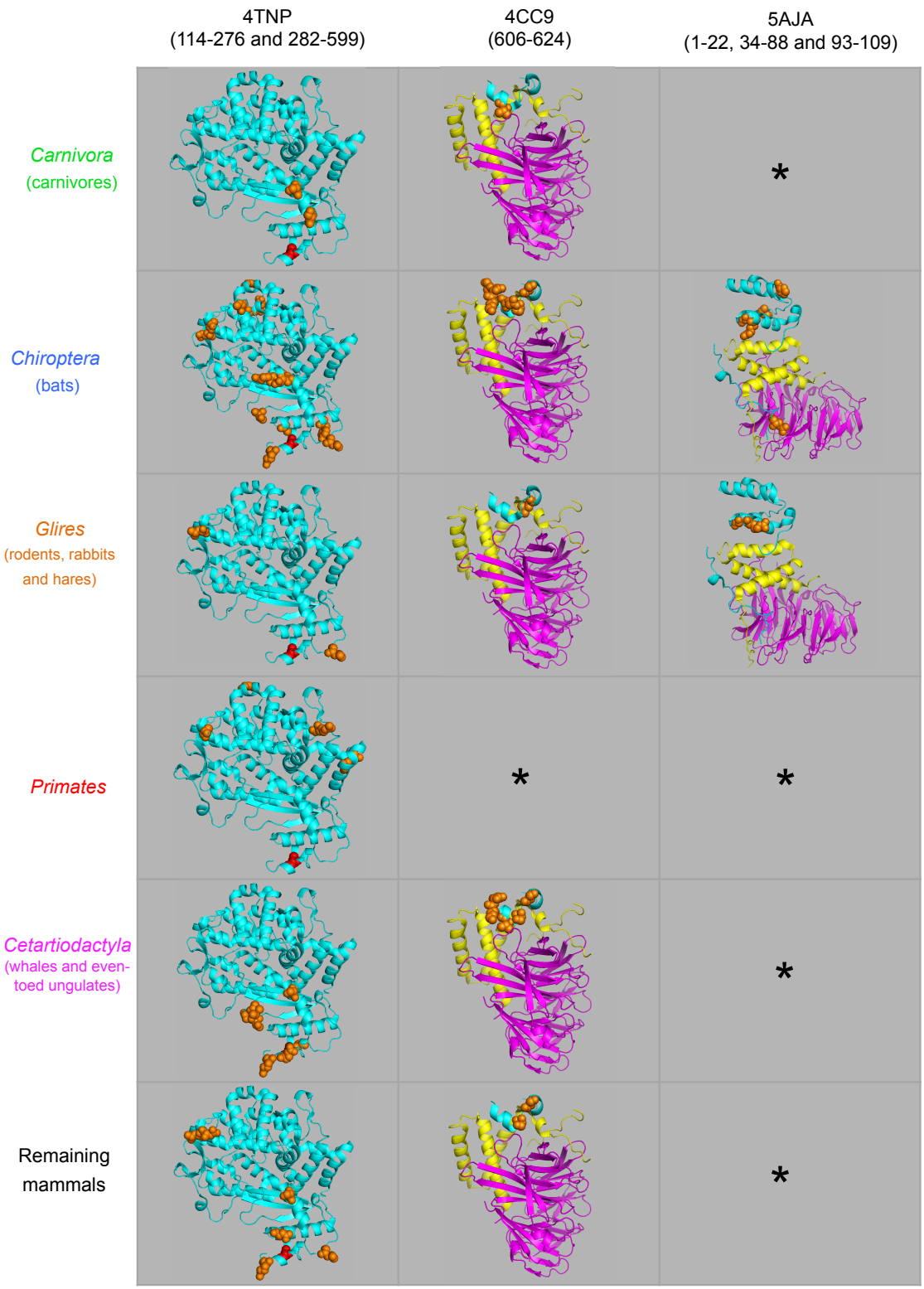

Figure 2.10: Sites identified as under positive selection in mammalian SAMHD1 (orange spheres), shown on published crystal structures of SAMHD1, as in figures 2.7 and 2.8. Threonine 592 is again shown in red. The PDB identifier and the ranges of sites included in the structures are given at the top of each column of figures. Asterisks mark analyses which did not identify sites present in the given structure.

2.3.3 SAMHD1 under positive selection in diverse chordate groups

Having found positive selection in SAMHD1 extends beyond primates to other mammals, we next asked whether positive selection could be observed among still wider chordate groups. We had obtained SAMHD1 coding sequences for 51 bird, 7 reptile, 2 amphibian and 19 marine animal species (appendix A). Amphibians were the least well represented, both sequences coming from the same genus (*Xenopus*). The marine animal group (which, though not a formal taxonomic rank, we will call 'fish') comprised 16 ray-finned fishes (class *Actinopteri*) and one species from each of the divergent classes *Chondrichthyes*, *Sarcopterygii* and *Leptocardii*. The last of these — the lancelet, or amphioxus (*Branchiostoma floridae*) — is considered the most basal chordate group (Delsuc et al., 2006). Our dataset therefore contains a wide diversity of chordate SAMHD1, representing more than 550 million years of evolutionary divergence. A 'master' tree topology comprising the mammal species (discussed above) and these non-mammals species was estimated by maximum likelihood (fig. 2.11, see methods).

Positive selection identified across chordates

We began with site-specific analyses of the master SAMHD1 dataset. Model M2a had a statistically significant better fit to the data than model M1a (appendix B), while, as with the mammals dataset, model M8 failed to converge. 11 sites were estimated to have probability > 0.95 of belonging to the positive selection site class (fig. 2.12, top panel). All of these had previously been identified in the M2a analysis of mammals alone (i.e. these 11 are a subset of the 36 described above).

To test for positive selection signal outside of the mammal clade, we repeated the site-specific analysis with mammals excluded. The M2a model of positive selection was again statistically supported (and again M8 did not converge), but no sites had > 0.95 probability of being under positive selection (the highest probability was 0.924, for site 601). The 11 sites identified with all sequences included could therefore be attributed to signal in mammals alone. However, that the model of positive selection was supported when mammals are excluded indicated positive selection was present outside of mammals, and suggested that the failure to identify specific sites may be because the exact sites responsible for the signal differ between

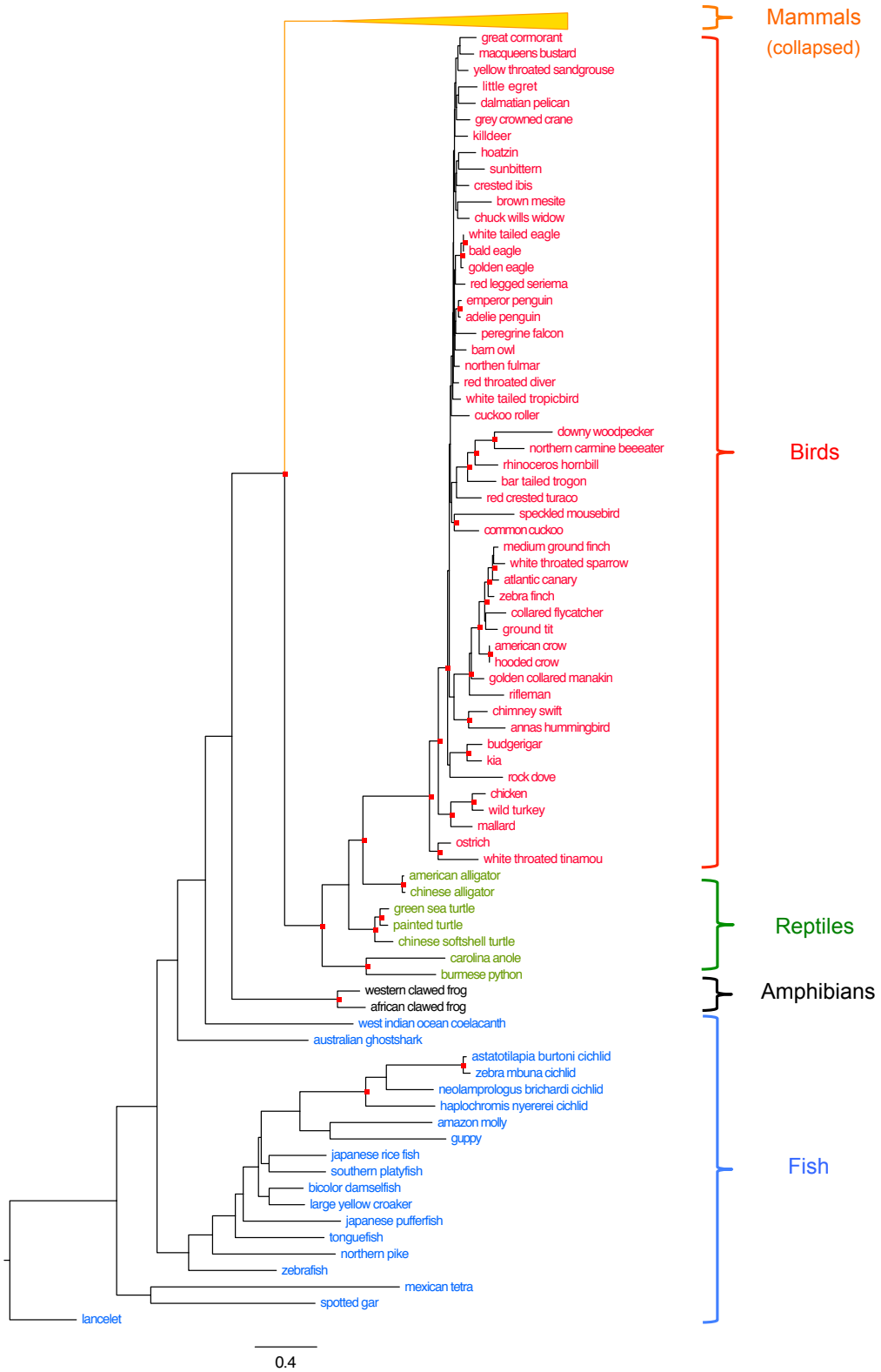

Figure 2.11: Maximum likelihood phylogeny for chordate SAMHD1; see legend fig. 2.5. The mammal clade is collapsed for clarity (this subtree is identical to fig. 2.5). The tree is rooted on the lancelet *Branchiostoma floridae* branch.

Subgroup	M2a supported	M8 supported
Birds	+	−
Reptiles	+	+
Amphibians	−	−
Fish	+	No convergence

Table 2.5: Summary of statistical support for site-specific models of positive selection, assessed by the likelihood ratio test with two degrees of freedom for each (i.e. the difference in the number of parameters between null and alternative models). + indicates $p < 0.05$ and − otherwise, from likelihood ratio test. See appendix B for log-likelihood and p values.

groups.

Evidence of positive selection within chordate groups from site models

Using site-specific models on sublineages, we investigated positive selection within the chordate groups birds, reptiles, amphibians, fish and (using branch-specific models only, having used sites-specific already) mammals. Results from the site-specific models are summarised in table 2.5; support for positive selection was observed in birds, reptiles and fish. Only one site in one chordate group had probability of > 0.95 for being under positive selection with either site model (site 266 in birds, with model M2a). Together, these data suggested positive selection has occurred in these three groups but the models used were not sufficiently powerful to detect the sites responsible. It is unsurprising that no signal was observed in amphibians since only two, highly related, species are represented.

More sites identified with branch-site models

We then used branch-site specific models to analyse the 'master' dataset, with groups of interest set to foreground. There was statistical support for positive selection in all groups tested (the mammal, bird, reptile, amphibian and fish lineages, appendix B). Multiple sites were identified as having probability > 0.95 of being under positive selection, though no sites were identified in amphibians (fig. 2.12 and table 2.6). Unsurprisingly, there was agreement with the previous analysis of mammal sequences: all 36 sites identified previously in the site-specific analysis of mammals alone were included in the 39 sites identified in this branch-site analysis with mammals as foreground (i.e. the former is a subset of the latter). The sets of sites found with mammals, birds, reptiles and fish as foreground were completely non-

overlapping, accounting for the failure to identify sites using site-specific models and the whole 'master' dataset. In fish the sites identified were evenly spread across the sequence and in birds the sites were evenly spread through the HD domain (fig. 2.12). Sites identified in reptiles were clustered at the N-terminus. These distributions were in contrast with the densely clustered arrangement of sites found in analyses of mammals.

Mapping onto crystal structures

We then mapped the sites identified onto the same crystal structures discussed above. There was no clear clustering in Vpx-interacting regions of sites identified with non-mammal groups as foreground (fig. 2.13, centre and right). As with the analysis of mammalian SAMHD1 previously, sites identified within the HD domain (fig. 2.13, left) were positioned away from tetramer interfaces and mostly exposed to solvent. Some sites found in birds and fish were positioned somewhat near to T592 and for completeness we computed p values for clustering around this residue as before, but a significant value was found only for mammals (table 2.7).

Repeating with alternative tree topologies

The 'master' tree topology used did not have high bootstrap support for all nodes and was particularly poor in the fish lineages (fig. 2.11). We therefore wanted to determine whether the positive results obtained were robust to variation in topology. We repeated the site-specific analysis of the whole master dataset and the branch-site specific tests with each subgroup as foreground using three alternative (bootstrap) topologies. All tests with alternative topologies were consistent with results from the maximum likelihood topology (see appendix B). Mostly there was concordance between the sites identified with the maximum likelihood and alternative topologies (table 2.8), with the exception of just one topology in the analysis with mammals as foreground where several more sites were found with the maximum likelihood than the alternative topology. Together these data indicate that the identification of sites under positive selection in these analyses is not dependent on the topology used.

Site	Residue	Mammals	Avian	Reptiles	'Fish'
11	K			+	
12	R				+
32	W	+			
41	D			+	
55	R	+			
57	G	+			
60	E	+			
63	V	+			
92	S				+
93	S			+	
95	G	+			
103	Y	+			
112	V	+			
225	A		+		
226	R				+
247	S				+
256	Q	+			
262	E			+	
280	V	+			
283	S	+			
284	L	+			
338	A		+		
357	G				+
368	S	+			
379	G				+
397	I		+		
408	R	+			
425	N				+
436	P				+
463	T	+			
465	Q	+			
466	I	+			
471	E		+		
472	D	+			
475	S	+			
486	K	+			
488	L	+			
501	D				+
504	N				+
522	C				+
561	S		+		
566	R	+			
568	Y				+
574	A	+			
575	D	+			
594	Q	+			
596	K	+			
600	D	+			
601	S	+			
609	R	+			
610	L	+			
611	R	+			
614	S	+			
616	S	+			
618	V	+			
619	Q	+			
622	K	+			
624	D	+			
625	P	+			
626	M	+			

Table 2.6: Sites identified as under positive selection in chordate SAMHD1 using branch-site models with the specified groups as foreground branches (Yang and Nielsen, 2002; Zhang et al., 2005). Alternative model of positive selection supported for all groups. '+' indicates Bayes empirical Bayes probability > 0.95.

Foreground chordate subgroup	p value
Mammals	0.0005
Birds	0.48279
Reptiles	0.95519
Fish	0.45508

Table 2.7: Probabilities of the sites identified as being under positive selection clustering around threonine 592 if their distribution throughout the monomeric SAMHD1 crystal structure 4TNP (Ji et al., 2014) were random. See table 2.3.

Analysis	**Topology**	T_i **Total**	**ML only**	**Intersection**	T_i **only**
Site model (M2)	T_1	7	4	7	0
	T_2	11	1	10	1
	T_3	10	1	10	0
Mammals f.g.	T_1	17	22	17	0
	T_2	40	1	38	2
	T_3	38	1	38	0
Birds f.g.	T_1	6	0	5	1
	T_2	7	0	5	2
	T_3	7	0	5	2
Reptiles f.g.	T_1	3	1	3	0
	T_2	3	1	3	0
	T_3	3	1	3	0
Fish f.g.	T_1	11	2	10	1
	T_2	12	2	10	2
	T_3	9	3	9	0

Table 2.8: Comparison of SAMHD1 sites identified as under positive selection with the maximum likelihood (ML) tree topology and each of three alternative tree topologies, from site specific and branch-site specific analyses. For branch-site specific analyses, the group designated foreground (f.g.) is indicated. Presented as in table 2.2.

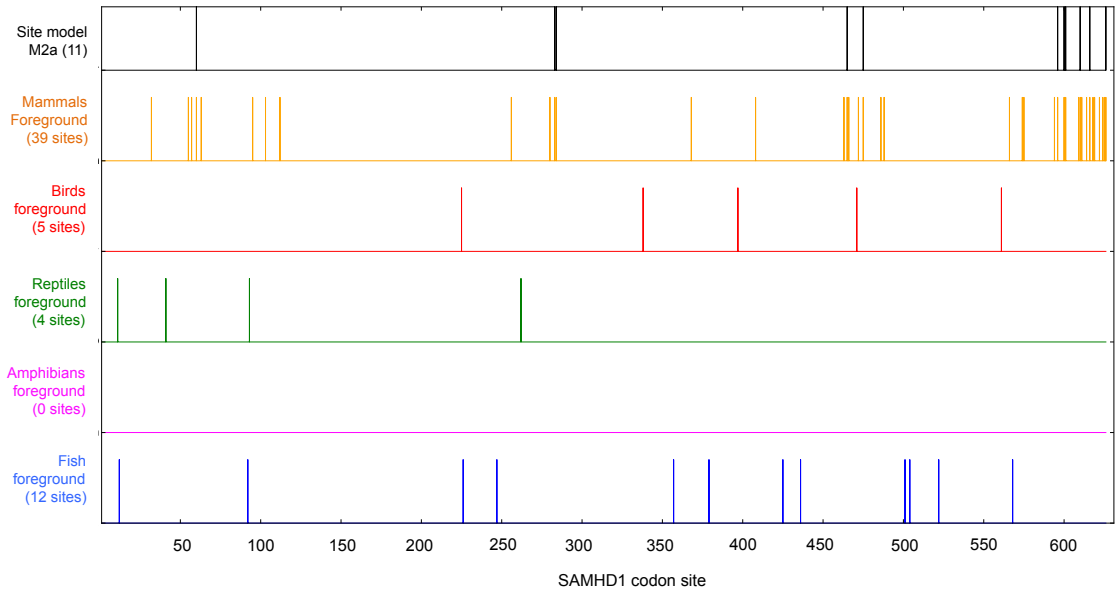

Figure 2.12: Codon sites under positive selection in chordate SAMHD1, displayed as in figures 2.6 and 2.9. All analyses used the whole chordate dataset; panels labelled as 'foreground' refer to results from branch-site specific tests with the given lineages set as foreground branches. No sites were found in the analysis of amphibians, though the alternative model was statistically supported.

2.4 Discussion

Positive selection in SAMHD1 a broader phenomenon than previously appreciated

Previous studies had found SAMHD1 to be under positive selection in primate species and this was attributed to escape from antagonistic proteins carried by primate lentiviruses (Laguette et al., 2012; Lim et al., 2012). We have found, firstly, that this signature extends throughout the mammalian lineage and we identify 36 sites under positive selection (fig. 2.6). With primates excluded, we still identify 31 sites under positive selection, compared with 17 sites observed by Laguette et al. in primates alone; this is not explained by differences in the models used, since codeml's M2a (used here to study the mammals and mammals excluding primates datasets) is less powerful, or more conservative than M8, used by Laguette et al. (Wong et al., 2004). These results show positive selection in SAMHD1 is not unique to primates.

We also identified positive selection within every subgroup of mammals present in our dataset, showing that no one group is responsible for the overall signal (fig. 2.9). Surprisingly few sites were identified in our analysis of primate SAMHD1: only

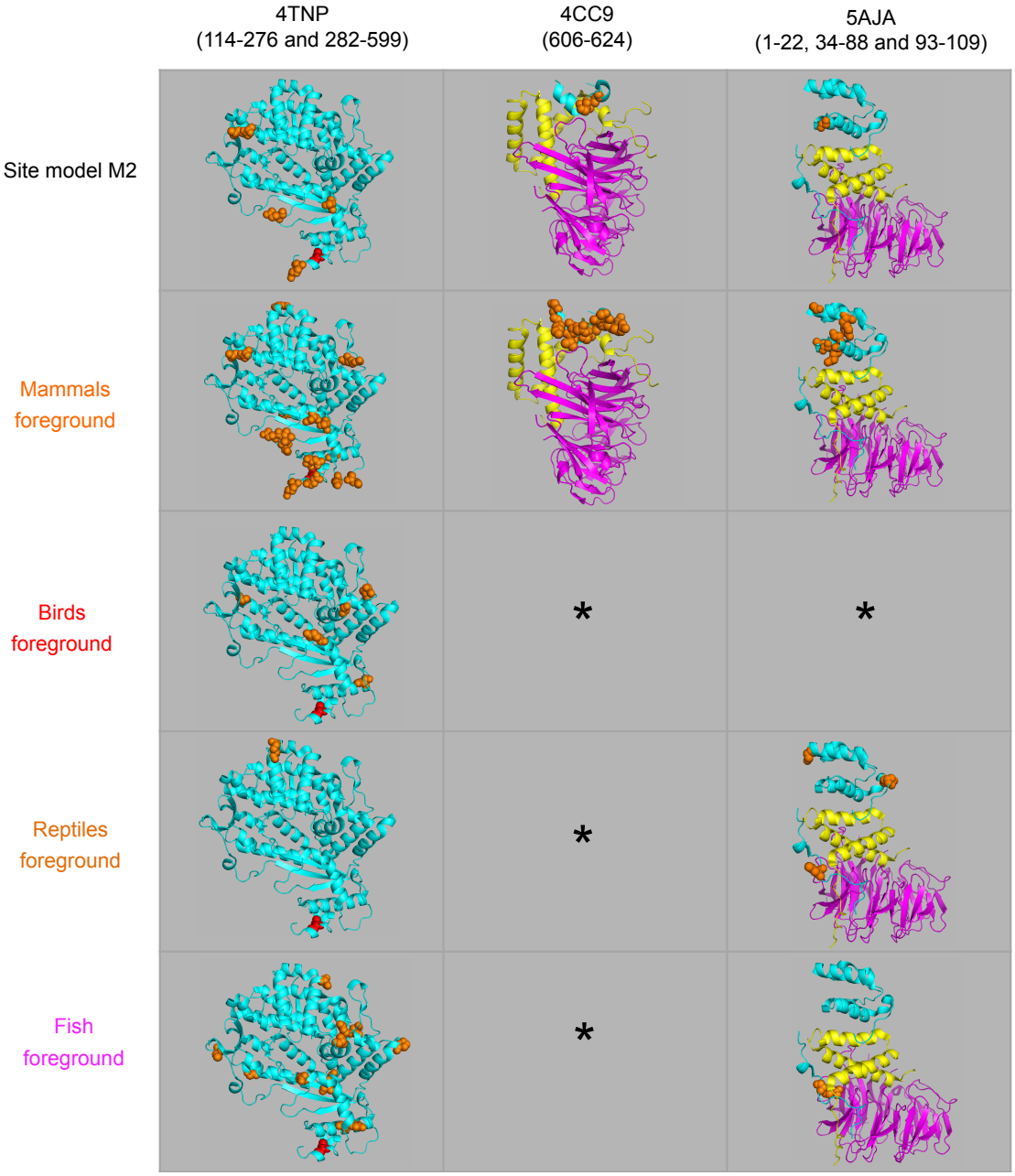

Figure 2.13: SAMHD1 sites identified as under positive selection in branch-site tests with animal groups set as foreground, presented as in figure 2.10.

5 with both M2a and M8, in contrast to 6 found by Lim et al. and the 17 by Laguette et al. The discrepancy may be explained by differences in sequence alignment, our approach having been very conservative (see methods). An abundance of sites was found in bats, despite their being the least well represented mammal subgroup (6 taxa) and it is tempting to speculate that this is connected with bats' unusually high burden of endemic virus infection (Luis et al., 2013), to which they may have adapted to control through innate immunity (Baker et al., 2013).

Apart from mammals, we have found positive selection in SAMHD1 from still more divergent groups of species, representing all chordates for which reliable sequences are presently available (fig. 2.12). Though some of our datasets include considerable sequence divergence, simulation studies show branch-site methods are robust to saturation of synonymous codon substitutions, which is associated with lack of power rather than increased false positive rate (Yang and Dos Reis, 2011; Gharib and Robinson-Rechavi, 2013; Thiltgen, G., dos Reis, M. and Goldstein, R. A., in press). Moreover, we consistently observe positive selection with the site models when analysing the much less divergent subgroups of mammals or other chordates in isolation.

Together, these data show that positive selection in SAMHD1 has been a feature of its evolution in all advanced animals. This is in stark contrast to the implicit assumption that positive selection has occurred in response to primate lentiviruses specifically (Laguette et al., 2012; Lim et al., 2012; Spragg and Emerman, 2013; Fregoso et al., 2013). A much larger pattern could similarly apply to other restriction factors currently assumed to be under positive selection due to primate lentiviruses, such as tetherin (Gupta et al., 2009; McNatt et al., 2009) or APOBEC proteins (Compton et al., 2012).

Sites of structural interest

We identified four sites under positive selection in mammals as a whole which were of special interest because of their position in the minor lobe and the chemical properties of the residues seen in different species. With site 566 being mostly buried, the range of residue sizes observed is surprising, and could have dramatic effects on the shape and stability of the minor lobe. The site is also in close contact with residue T525 in the neighbouring monomer and therefore repeated substitution

at 566 could further influence monomer-monomer interactions.

Site 574 is still more deeply buried in the minor lobe and substitution at the position might be expected to be conservative, since, like 566, the introduction of bulkier side chains might disrupt the packing of residues; yet we observe almost the widest possible range of side chain sizes, from alanine to phenylalanine. The hydrophilic serine also appears to have arisen multiple times independently, despite this site being hidden from solvent in the human structure.

Sites 594 and 596, near the phospho-acceptor T592, are occupied by residues with diverse chemical properties in different species, which may lead to significantly different reactions when the negatively charged phosphate group is introduced. The charge difference observed at 596 in Old World monkeys and apes (lysine, positive) and New World monkeys (aspartate, negative) may mark a substantial phenotypic difference within the primate lineage.

Mutagenesis Experiments

We hypothesise that mutations at the four sites discussed above would be particularly likely to influence the activities of SAMHD1. With reference to residues observed elsewhere in the mammalian phylogeny at these sites, we have designed a set of candidate mutants which we would expect to exhibit altered, but not destroyed, SAMHD1 activity. We have collaborated with the Jonathan Stoye and Ian Taylor laboratories (Crick Institute, UK), who have constructed the mutant proteins in a human SAMHD1 background sequence.

Preliminary results indicate the mutant proteins remain enzymatically active and exhibit variable levels of HIV-1 restriction. Strikingly, for most mutants tested the fraction of SAMHD1 molecules engaged in tetramers increases, suggesting residues at these sites indeed influence tetramer stability. This work is ongoing.

Drivers of positive selection

The question then arises: what has driven this preference for change over conservation? We propose two, mutually compatible models for evolution in chordate SAMHD1: (1) evasion of antagonistic proteins from diverse animal viruses and (2) continual ‘fine-tuning’ of enzymatic activity to maintain balanced dNTP concentration.

Available data from experimental and sequence analysis work strongly suggest positive selection in primate SAMHD1 is driven — at least partly — by the need to escape antagonism by SIV Vpr/Vpx proteins (Laguette et al., 2012; Lim et al., 2012; Fregoso et al., 2013; Spragg and Emerman, 2013). Similar signatures in non-primate animal lineages may therefore be due to equivalent 'arms races' with viruses infecting those species. Several sites identified within subgroups of mammals are positioned in regions known to be targeted by SIV Vpr/Vpx proteins, at both N- and C-termini (fig. 2.10), and these same domains could be targeted by other viral proteins. However, there is significant variation in the regions targeted by Vpr/Vpx (Fregoso et al., 2013) and therefore hypothetical antagonists carried by other viruses could plausibly bind different regions of the protein. This would be consistent with the different distributions of sites identified in each mammal subgroup studied (fig. 2.6 and 2.12).

The nature of the animal viruses potentially responsible for the selective pressure is uncertain, since animal pathogens are less well studied, but SAMHD1 antagonism by Vpr/Vpx-like factors may be shared by other lentiviruses. Indeed, all known lentiviruses infecting humans (HIVs), other primates (SIVs), horses (equine infectious anemia virus), sheep (ovine maedi-visna virus), goats (caprine arthritis and encephalitis virus) and cats (feline immunodeficiency virus) all encode accessory proteins, some of which bear similarities to Vpr (Villet et al., 2003). Still more lentiviruses infecting other animals may remain undiscovered. Moreover they need not be the only candidates, since entirely different viruses could have adapted to antagonise the same restriction factor; just as several lentiviruses (HIV-1 M and O, HIV-2), filoviruses (Ebola virus and Marburg virus) and a herpes virus (human herpes virus 8) have all independently evolved mechanisms for antagonising the restriction factor tetherin (Neil et al., 2008; Kluge et al., 2014; Hauser et al., 2010; Bartee et al., 2006; Kaletsky et al., 2009). Several diverse non-lentiviruses are reported to be sensitive to SAMHD1 restriction, suggesting the evolution of antagonists could offer fitness advantages to other viruses, and the insensitive deltaretrovirus HTLV-1 has already been suggested to encode its own SAMHD1 antagonising factor (Gramberg et al., 2013). Since it has been shown to repress LINE-1 (Zhao et al., 2013b), SAMHD1 may also play a role in limiting spread of endogenous retroelements in general and escape antagonism from them.

However, the primary function of SAMHD1 may be as a metabolic regulator

(Franzolin et al., 2013), and we cannot exclude refinement of the enzyme's dNTPase activity being responsible for the positive selection signature. In mammals, we have observed a statistically significant clustering of sites under positive selection around the (absolutely conserved) phosphoacceptor T592, in the minor lobe of the C terminus (fig. 2.7), which is involved in regulating the tetramerisation, and therefore dNTPase activity, of SAMHD1. Residues in the minor lobe (including 618 and 619, which we find to be under positive selection) are known to affect SAMHD1 interaction with cyclinA2, which together with associated CDKs mediates T592 phosphorylation (Yan et al., 2015). Upon phosphorylation, the minor lobe becomes disordered and prevents formation of stable tetramers (Arnold et al., 2015). Our identification of positive selection in this region suggests the control of tetramerisation and enzymatic activity has adapted throughout mammalian evolution.

The distributions of residues at the example sites 566, 574, 594 and 596 show that dramatic changes in sequence context have occurred throughout mammalian SAMHD1 evolution, and some residues observed appear inconsistent with what we know of the structural context in the human tetramer. It is therefore plausible that the structure of the minor lobe itself has changed significantly, with profound consequences for both the impact of phosphorylation and tetramer stability in different species. This could have been driven by a need to alter the sensitivity for activating dNTPase activity in accordance with the different dNTP cellular concentrations required by each species.

Finally, we note that these two models need not be exclusive. Positive selection may be driven by both escaping viral antagonism and calibrating enzymatic activation simultaneously. Indeed, the substitutions needed to evade pathogen recognition may have necessitated further change to compensate for disruptive effects on the SAMHD1 structure.

2.5 Methods

2.5.1 Sequence gathering and alignment

Chordate SAMHD1 DNA sequences were collected using NCBI BLAST (blastn algorithm) with human SAMHD1 coding sequence (accession NM_015474.3) as the query. Where more than one sequence was available from a single species (usu-

ally transcript variants), sequences most closely matching the human sequence were selected. The sequence for Tasmanian devil (*Sarcophilus harrisii*) was found to be divided into two sequence records (accessions XM_003758997.2 and XM_012553363.1); these were concatenated to give a full length sequence. The list of species and accession numbers for sequences used are listed in appendix A. Sequences which were less than 70% of the length of the human SAMHD1 sequence were excluded. Sequences were collected by Christopher Ruis, following a strategy devised jointly by Christopher Monit and Christopher Ruis.

Dividing sequences by animal group (mammal, avian, reptile, amphibian and fish), these nucleotide sequences were aligned as translated protein using MUSCLE 3.8.31 (Edgar, 2004), as implemented in the alignment editor SEAVIEW 4.4.0 (Gouy et al., 2010). Misalignment greatly increases the chance of identifying false positives in selection analyses, as the apparent rate of nonsynonymous substitution increases. The resulting codon alignments were therefore further edited manually, with a highly conservative approach: sections within sequences which could not be aligned with high confidence were masked, such that they would be treated as missing data (equivalent to alignment gaps) by phylogenetics tools. To create a master alignment comprising all sequences, the separate animal group alignments were again translated to amino acid sequences and aligned to each other as blocks, using the 'merge' function of the MAFFT 7 online alignment tool (Katoh and Standley, 2013) and the resulting alignment pattern was mapped onto the original codon sequences. Alignment columns containing no data (either gaps or masked codons) in $\geq 20\%$ of sequences were removed. Sequences were aligned by Christopher Monit and Christopher Ruis together.

2.5.2 Phylogeny estimation

A phylogenetic tree was estimated by maximum likelihood using multithreaded RAxML HPC-PTHREADS-SSE3, version 7.7.2 (Stamatakis, 2006; Ott et al., 2007) with the general time reversible (GTR) substitution model (Tavare, 1986) and gamma-distributed rate heterogeneity. Confidence in the tree topology was assessed by estimating trees from 1000 non-parametric bootstrap samples, also using RAxML 7.7.2. The platypus (*Ornithorhynchus anatinus*) sequence was found to incorrectly cluster well outside of the mammalian clade and was therefore excluded from sub-

sequent analysis.

RAxML saves each of the non-parametric bootstrap trees it estimates. For repeating selection analyses with alternative tree topologies, we used three of these bootstrap trees (finite computing resources precluded using more). To speed up bootstrap tree estimation, RAxML only computes a new initial estimate tree (by maximum parsimony) for every 10th bootstrap dataset, while interim trees use the previously estimated bootstrap tree as the starting estimate. To ensure the alternative topologies used were independent and identically distributed, we used bootstrap trees which had had maximum parsimony trees as initial estimates (the 1st, 100th and 200th tree produced).

When investigating selection on subgroups from this large ('master') phylogeny using site-specific models, taxa not of interest were pruned from the topology and the sequences omitted from the analysis. Custom software (written by Christopher Monit) for formatting phylogenetic trees made use of the Phylo package in the Biopython library (Talevich et al., 2012). Tree figures were produced using FigTree 1.3.1 (Rambaut, 2006).

2.5.3 Selection analysis

Phylogenetic analyses were performed using the codeml program, of the PAML package (version 4.7a; Yang, 2007). We used the site-specific tests of positive selection M1a/M2a (Nielsen and Yang, 1998) and M7/M8 (Yang et al., 2000). In some analyses we also used the branch-site specific test where clades of interest were set as foreground (meaning positive selection is permitted) in the alternative model (Yang and Nielsen, 2002; Zhang et al., 2005).

To reduce the risk of the log-likelihood optimisation reaching a local optimum, all program runs were performed five times with different initial parameters for the transition/transversion ratio (κ) and dN/dS ratio (ω): 0.1, 1 and 10. The exception was the null model for branch-site specific analyses: since the omega ratio is fixed to 1.0 along foreground branches the user cannot specify an initial ω value other than 1.0. Tree branch lengths were first optimised with codeml's model 0 (which allows a single ω value) with the corresponding initial parameter values, and these branch lengths were used as starting values in subsequent analyses with more complex models. Codon stationary frequencies were estimated using the F1x4 model by

default, but F3x4 was used if optimisation errors were encountered with F1x4 (i.e. if the null model had a superior fit to the alternative, which should not be possible if both models are fitted correctly).

Statistical justification of the alternative model was assessed using the likelihood ratio test, where the null distribution of the test statistic D is assumed to follow a χ^2 distribution, with degrees of freedom equal to the difference in the number of free parameters in the null and alternative models: this was 2 for both site-specific tests and 1 for branch-site specific tests.

Sites were identified as being under positive selection if the computed Bayes empirical Bayes probability for the site belonging to the positive selection class was > 0.95. These sites were mapped onto crystal structures, visualised using PyMOL 1.3 (Schrödinger, 2010).

2.5.4 Clustering around T592 of mammalian SAMHD1 sites under positive selection

The distance of a given residue from T592 was taken as the distance between their respective α carbon atoms. Positions of atoms in protein databank (PDB) files are expressed as co-ordinates (in Ångstroms) and the distance d between two residues is therefore given as the Euclidean distance between α carbons a and b (one of which is T592) in dimensions x, y and z:

$$d = \sqrt{(a_x - b_x)^2 + (a_y - b_y)^2 + (a_z - b_z)^2}. \tag{2.1}$$

Computing this distance for each of n sites identified as under positive selection, we computed the harmonic mean (reciprocal of the mean of reciprocals) of distances H:

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{d_i}} \tag{2.2}$$

We then assessed whether this harmonic mean distance was lower than expected by chance by randomly sampling n sites from the structure, computing d for each site and then computing H for the sample. We did this 10^5 times to generate a distribution of H values for unbiased collections of sites. The p value was thus the proportion of samples for which H was less than the H found with the sites identified as under positive selection. These computations were performed using (unpublished)

software written by Richard A. Goldstein and co-ordinates from PDB file 4TNP (Ji et al., 2014).

Chapter 3

Divergent Selective Constraints in HIV-1 M/SIVcpz Capsid

3.1 Summary

The lentiviral capsid (CA) protein comprises the viral core, which houses the RNA genome and viral enzymes. Recent research has shown the capsid determines many aspects of the early replication cycle, including prevention of innate immune sensing and influence over the genome integration site. Many of these activities are dependent on interaction with host cofactors, such as cyclophilin A (CypA). To both characterise evolutionary differences and inform experimental studies, we sought to identify the CA sites evolving under different selective constraints in the pandemic HIV-1 group M and SIVcpz, from which HIV-1 M arose following transmission from chimpanzees. Using a site-wise mutation selection model, we have identified 23 such sites. These are found in the external CypA binding domain, in positions deep within the CA structure and in regions which stabilise CA-CA interactions in the core. Our results suggest host-specific behaviours, with significant implications for understanding the establishment of a pandemic human pathogen. Unexpectedly, we also find evolutionary differences between SIVcpz isolates and between subtypes of HIV-1 M, suggesting diverse CA behaviours in viruses infecting the same host species. In collaboration with the Greg Towers laboratory, these results will form the basis of comparative experimental investigations into lentiviral CA function.

3.2 Introduction

Near the end of the human immunodeficiency virus type 1 (HIV-1) life cycle, Gag polyproteins assemble to form immature viral particles, and then are cleaved by the viral protease. The resulting 231 residue capsid (CA) proteins, comprising a $\sim$ 150 residue N terminal domain (NTD) and $\sim$ 80 residue C terminal domain (CTD), assemble into mostly hexamers, with 12 pentamers included in specific positions to form the closed viral core (Ganser et al., 1999). In a subsequent infection, the core serves to enclose the genomic RNA template and reverse transcriptase enzyme and probably shields the nascent reverse transcribed DNA from innate immunity sensors, such as cyclic GMP-AMP synthase (cGAS; Gao et al., 2013). Recent observations indicate that CA has a considerably more complex role than as an inert shell, however, as it has been shown to orchestrate both the passage of an incoming viral core through the cytoplasm and delivery of the reverse transcribed genome into the nucleus for integration (reviewed by Campbell and Hope, 2015).

CA mediates these effects through interaction with several host cofactors. CPSF6 (cleavage and polyadenylation specificity factor subunit 6) is ordinarily involved in mRNA processing and shuttles between the cytoplasm and the nucleus (Ruepp et al., 2009). It binds to CA at a conserved interface on the CA hexamer (Price et al., 2012) and may be responsible for directing the core to the nuclear pore via its nuclear localisation signal domain (Lee et al., 2010). TNPO3 (transportin 3) is usually involved in nuclear import of cellular proteins (Kataoka et al., 1999) but has been reported to bind CA at the same interface and may also be involved in nuclear import of the core (Zhou et al., 2011). Cyclophilin A (CypA) is a cytoplasmic peptidyl isomerase which binds to CA via a loop on the exterior of the core (Gamble et al., 1996). It catalyses the cis-trans isomerisation of the G89-P90 peptide bond in CA, which induces conformational changes in the NTD away from the binding loop (Bosco et al., 2002), with possible implications for uncoating. The same binding loop is used for interaction with the nuclear pore protein NUP358 (Schaller et al., 2011), which possesses its own cyclophilin domain and projecting filaments which recruit proteins for passage into the nucleus. CA also associates with NUP153 (Matreyek et al., 2013), a nuclear pore protein whose filaments project into the nucleoplasm.

Failing to make interactions between CA and cyclophilins (CypA or NUP358-Cyp) or CPSF6 has been shown to disturb the core's nuclear import pathway and

is able to determine whether the virus integrates in more or less transcriptionally active regions of the host genome (Schaller et al., 2011). These same cofactors are also required to prevent innate immune sensing of viral DNA in macrophages, as has been demonstrated with CA mutants P90A and N74D which are incapable of CypA or CPSF6 interaction, respectively. This leads to detection of viral DNA by cGAS and type 1 interferon induction, dramatically impeding virus replication (Rasaiyaah et al., 2013).

Recent work has shown CA hexamers form pores into the viral core, through which dNTPs are electrostatically drawn, for use in reverse transcription (Jacques et al., 2016). Their influx is regulated by a hairpin loop comprising short β-sheets on the surface of the hexamer which can move to block the pore, also involving altered arrangements of α-helices within the NTD. These conformational changes are pH dependent, but *in vivo* may be induced by cofactor recruitment and are possibly used to prevent untimely reverse transcription which could trigger DNA sensing. Together, the recent research into the activities of CA show it is deeply involved in the early phases of infection, making it an attractive drug target (Domenech and Neira, 2013).

The pandemic HIV-1 group M entered the human population from a single cross species transmission from the chimpanzee subspecies *Pan troglodytes troglodytes* (*Ptt*; Gao et al., 1999), where the virus reservoir is known as simian immunodeficiency virus from chimpanzee (SIVcpz). Another chimpanzee subspecies, *Pan troglodytes schweinfurthii* (*Pts*) harbours a phylogenetically distinct virus. Separate cross species transmissions of SIVs to humans have occurred several times, but all have spread much less successfully in the human population (reviewed by Sharp and Hahn, 2011). Understanding the molecular basis for high or low degrees of lentiviral adaptation to the human host is of pressing interest.

Unpublished observations from the Greg Towers laboratory suggest HIV-1 group O (which transmitted to humans from chimpanzees via gorillas as an intermediary host; D'arc et al., 2015) may be incapable of regulating access through the CA hexamer pore and preliminary data indicate SIVcpz *Pts* may be similarly impaired, together suggesting CA functions differ between virus groups and hosts. SIVcpz is dependent on the same cofactors as HIV-1 M (Laura Hilditch and Greg Towers, manuscript in preparation) but the nature of these interactions and their consequences remain obscure.

We wished to characterise the regions of CA involved in the adaptation of HIV-1 M to the human host, with the intention of informing comparative studies of lentivirus life cycles. We report findings from a phylogenetic selection analysis of HIV-1 M and SIVcpz CA using the site wise mutation selection model (swMutSel), which is suited to identifying amino acid sites evolving under different evolutionary constraints, following a cross species transmission event (Tamuri et al., 2009, 2012). swMutSel allows us to test the hypothesis that the evolutionary process is different in the virus groups specified, against the null hypothesis that the selective constraints are the same. We have also analysed HIV-1 M and SIVcpz CA with the selection models M7/M8 (Yang et al., 2000) which use d_N/d_S ratios to identify positive (diversifying) selection, without distinguishing lineages. This approach can be useful for identifying sites targeted by cytotoxic lymphocytes (part of the adaptive immune system), which have been observed in all regions of the CA sequence (Llano et al., 2013) and would expected to promote diversifying selection as in other HIV-1 genes (Price et al., 1997). In addition, we have estimated which CA amino acid substitutions occurred following the cross species transmission which established HIV-1 M in the human population.

Several of the sites identified with swMutSel are located in regions involved in CA-CA interactions, the cyclophilin binding loop and around the β-hairpin, suggesting divergent activities of CA in the two virus groups. Unexpectedly, we observe striking differences in CA residues between SIVcpz *Ptt* and *Pts*, as well as HIV-1 M subtypes at sites of functional importance. Together, results presented here will inform experimental studies in an ongoing collaboration with the Greg Towers laboratory.

3.3 Results

3.3.1 Sites Identified as Under Different Selective Constraints

We obtained a dataset comprising 1332 HIV-1 and 21 SIVcpz CA sequences from the Los Alamos HIV sequence database (see Methods). We aligned these manually as codons and estimated a phylogeny by maximum likelihood (ML; fig. 3.1), in which the HIV-1 subtypes mostly formed clear monophyletic groups. As expected, SIVcpz isolates were divided into 13 originating from the *Pan troglodytes troglodytes* (*Ptt*)

chimpanzee subspecies, from which the HIV-1 lineage descends, and a separate clade comprising the 8 sequences derived from *Pan troglodytes schweinfurthii* (*Pts*; Sharp and Hahn, 2011).

Using our codon alignment and ML tree we applied the site-wise mutation selection model (swMutSel; Tamuri et al., 2012), testing the hypothesis that the substitution process differs for sites evolving in the human (HIV-1) or chimpanzee (SIVcpz) hosts. We did not analyse sites which are wholly conserved (where a single residue is seen in all sequences) as we know *a priori* the null hypothesis cannot be rejected in these cases; these were sites 145, 151, 157, 158, 176 and 192. Using the likelihood ratio test (LRT) we identified 23 sites for which the null hypothesis of homogeneous selective constraints was rejected with $p < 0.05$, correcting for multiple hypothesis testing (see methods; table 3.1 and table D.1). We had no swMutSel data for site 120 as log-likelihood optimisation failed to converge.

The 23 sites were distributed across the CA sequence, but there was greater concentration in the N terminal domain (NTD) than the C terminal domain (CTD) and particularly dense clustering in the cyclophilin binding loop (fig. 3.2, red). 11 of the 23 were positioned in linker regions between secondary structure elements.

For comparison, we fitted models M7 and M8 implemented in codeml (Yang, 2007) and found the M8 model of positive selection was statistically justified with the LRT ($p < 0.05$). 7 sites had estimated Bayes empirical Bayes probabilities of 1.0 for belonging to a positive selection site class (table 3.1; fig. 3.2, yellow; table D.1) and no others had BEB probability > 0.5. Four of these had also been identified with swMutSel. All but one of the sites identified with M8 were positioned in the NTD, two of which are in or near the cyclophilin binding loop.

3.3.2 Protein Structure Context

We mapped the identified sites onto a published CA crystal structure, which had been assembled into both hexamer and multiple hexamer units by molecular dynamics simulation (Zhao et al., 2013a; fig. 3.3). Sites 5 and 13 are in or near the β-hairpin and sites 86, 87, 91, 94 and 98 are in or near the cyclophilin binding loop. Pairs of sites 68/141 and 41/131 are each positioned within helices and have side chains in close proximity, suggesting chemical interaction.

To examine whether the identified sites were involved in CA-CA interactions,

CA Site	HXB2 Res.	swMutSel	M8 (BEB >0.95)
5	N	+	
13	Q	+	
14	A	+	+
15	I	+	
31	A	+	
41	S	+	
47	A	+	
50	Q		+
58	T	+	
68	M	+	
83	V	+	+
86	V	+	
87	H	+	
91	I	+	+
94	G	+	
98	E	+	
110	T	+	
116	G	+	+
120	N	n/a	+
131	K	+	
141	I	+	
171	T	+	
177	A	+	
194	A	+	
204	A	+	
225	G		+

Table 3.1: Sites identified in HIV-1 M and SIVcpz capsid (CA) with swMutSel (likelihood ratio test with degrees of freedom $N - 1$ where N is the number of residues at the site) and/or codeml model M8 (Bayes empirical Bayes probability for belonging to positive selection site class > 0.95). Amino acid residues are shown from the HIV-1 reference sequence HXB2 (group M, subtype B; accession K03455). There was no data for site 120 from swMutSel as the model failed to converge while fitting to data.

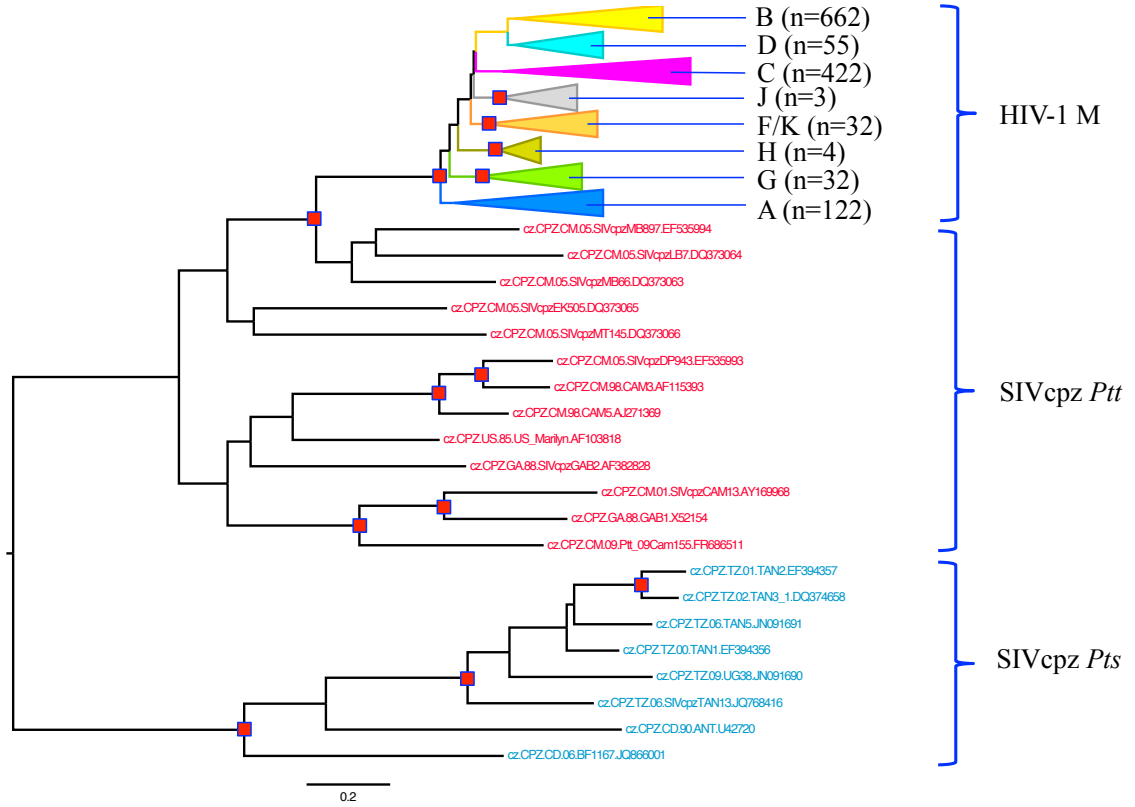

Figure 3.1: Maximum likelihood phylogeny for 1332 HIV-1 and 21 SIVcpz capsid sequences, estimated by RAxML. Clades for HIV-1 subtypes (denoted by letter name) are collapsed for clarity, and the numbers of taxa contained in each are indicated. The SIVcpz sequences are divided by the subspecies of their respective hosts, *Pan troglodytes troglodytes* (*Ptt*; red) and *Pan troglodytes schweinfurthii*; blue (*Pts*), and the tree is rooted with the latter as the outgroup (Sharp and Hahn, 2011). Nodes with $> 70\%$ support (from 1000 non-parametric bootstrap datasets) are indicated by red squares. Branch lengths are shown as nucleotide substitutions per site.

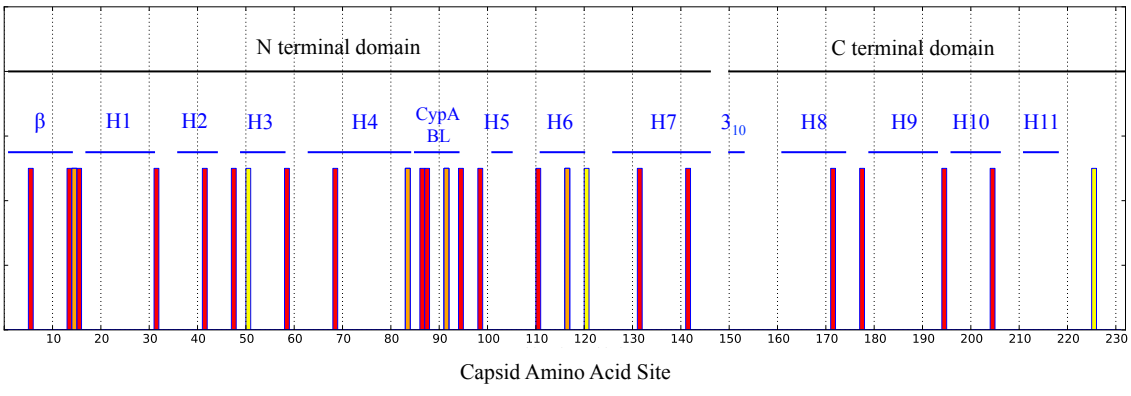

Figure 3.2: Sites identified in HIV-1 M and SIVcpz capsid (CA) with swMutSel (red), codeml's M8 (yellow) or both (orange). The height of the bars is of no significance. Secondary structure elements (as defined by Gres et al., 2015) are indicated: β, β-hairpin loop; H, α-helices; CypA BL, cyclophilin A (and other cyclophilins) binding loop; 3_{10}, short helix defining boundary between domains. The N terminal domain (sites 1-145) and C terminal domain (150-231) are indicated.

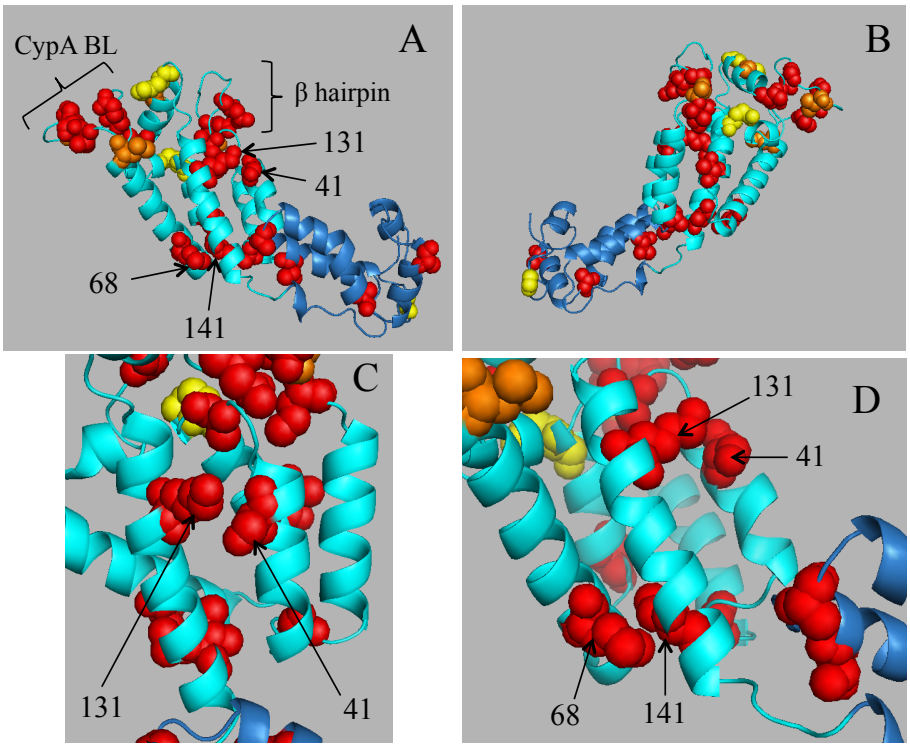

Figure 3.3: Sites mapped onto capsid monomer structure (PDB ID 3J3Q; Zhao et al., 2013a). N terminal domain and connecting linker (sites 1-149) are shown in cyan and the C terminal domain is shown in blue. Sites identified with swMutSel (red), codeml M8 (yellow) or both (orange) have their side chains shown as spheres. The cyclophilin A binding loop (CypA BL) and β-hairpin are indicated. Pairs of sites whose side chains are in close contact are indicated. Panel B is a 180° rotation of A; panel C shows an enlarged and 90° clockwise rotation of A; panel D is an enlarged view of A.

we mapped them onto the same assembled CA structure multimerised to form hexamers. Considering two interacting monomers (fig. 3.4), we saw sites 68, 58 and 15 (identified with swMutSel) are in close contact with 212, 35 and 42 respectively (not identified with either method) in the neighbouring monomer. Sites 5 and 14 were both identified with swMutSel and similarly are in close proximity between monomers. In the hexamer as a whole (fig. 3.5) we saw that the greatest density of sites is on the external side of the hexamer (fig. 3.5A, D) with no side chains of identified residues protruding into the lumen of the hexamer chamber (fig. 3.5B). All sites identified with M8 were positioned on the exterior of the hexamer, mostly on the front as viewed from outside of the viral core, or on periphery of the CTD (fig. 3.5C, D). Strikingly, identified sites 5 and 13 in and near the β-hairpin project into the pore at the centre of the hexamer. Mapping the identified sites onto a CA pentamer model (PDB ID: 3P05; Pornillos et al., 2011; data not shown) showed the same pattern of interacting side chains as for the CA hexamer.

We went on to consider involvement in hexamer-hexamer interactions (fig. 3.6). Site 204 in the CTD identified with swMutSel is in close proximity with site 231, an interaction which stabilises the hexamer lattice by taking part in hydrophobic interactions between CTDs in neighbouring hexamers (Gres et al., 2015; Zhao et al., 2013a). The sites identified in the cyclophilin binding loop, such as site 86 at the apex of the loop, project into the space bounded by interacting hexamers.

3.3.3 Alternative Tree Topologies

We went on to investigate the sensitivity of our results to the tree topology used, as the ML tree did not have very high branch support at the deeper internal nodes (fig. 3.1) and low support at shallow nodes within HIV-1 M subtype clades (not shown). We repeated the swMutSel and codeml M8 analyses with 5 alternative tree topologies, originating from bootstrap replicate datasets produced while estimating support for the ML tree. The sets of sites identified with the ML tree compared with the alternative topologies were mostly concordant; the greatest discrepancy was a difference of 7 out of 23 sites found using swMutSel with the ML topology but not found with an alternative (T_1; table 3.2).

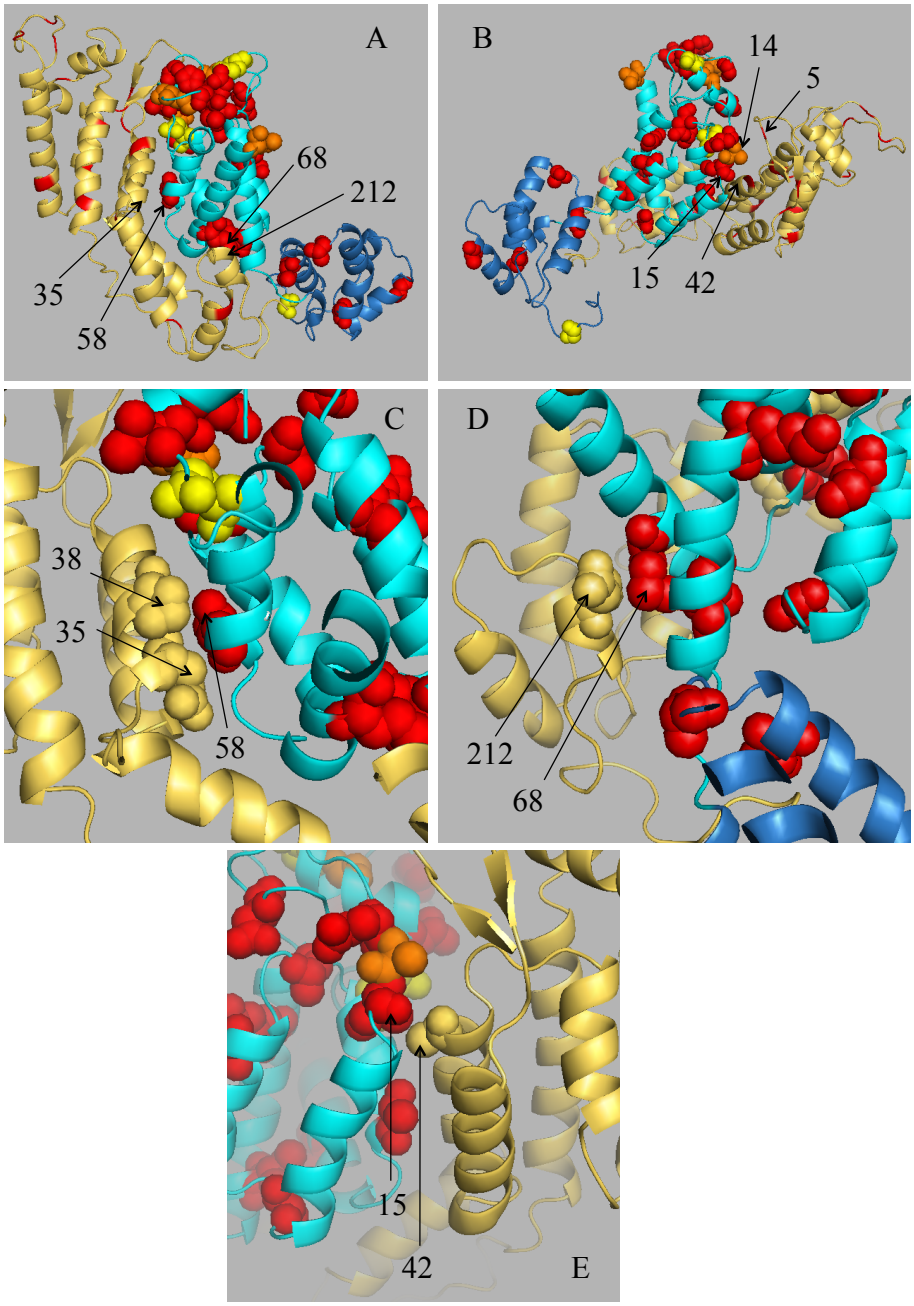

Figure 3.4: Identified sites mapped onto a pair of interacting capsid monomer structures, which form part of a CA hexamer (PDB ID 3J3Q; Zhao et al., 2013a). (A, right) N terminal domain and connecting linker (sites 1-149) are shown in cyan and the C terminal domain is shown in blue. Sites identified with swMutSel (red), codeml M8 (yellow) or both (orange) have their side chains shown as spheres. The interacting monomer (left) is shown in dark yellow, with identified sites similarly coloured but with side chains hidden for clarity. Pairs of sites whose side chains are in close contact between monomers are indicated. Panel B is a 180° rotation of A; panels C-E are enlarged images showing sites interacting between monomers, with residues of interest shown as spheres in the second monomer (dark yellow).

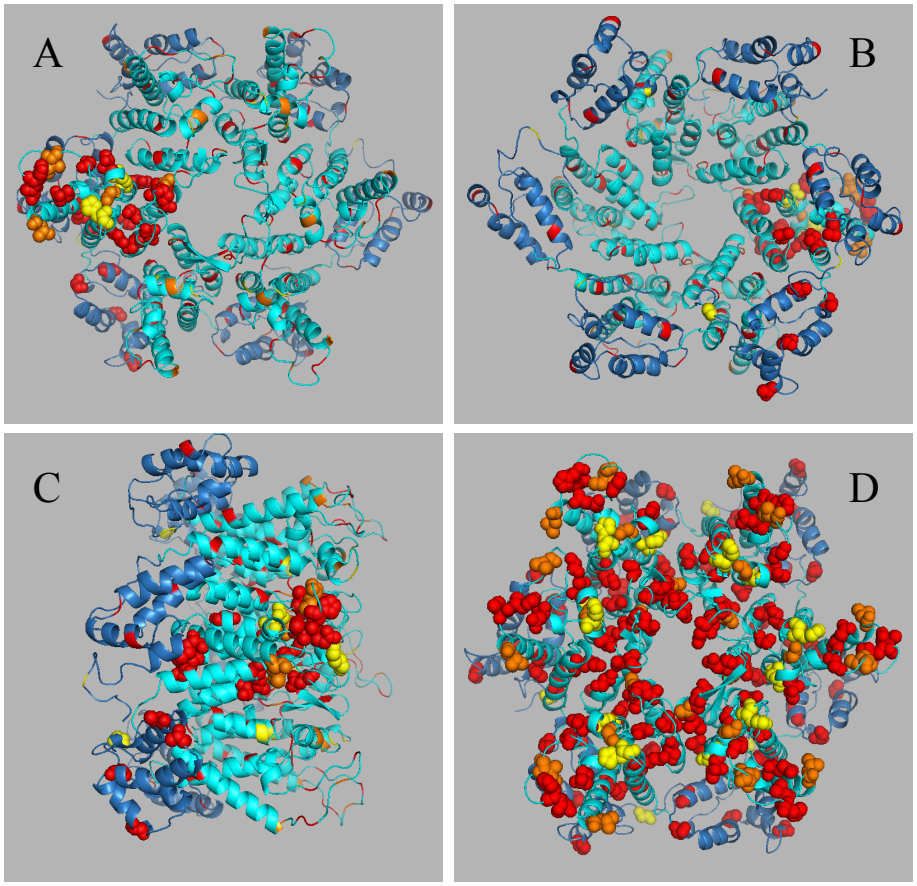

Figure 3.5: Identified sites mapped onto capsid hexamer structure (PDB ID 3J3Q; Zhao et al., 2013a). For each monomer the N terminal domain and connecting linker (sites 1-149) are shown in cyan and the C terminal domain is shown in blue. Sites identified with swMutSel (red), codeml M8 (yellow) or both (orange) are coloured on all monomers and have their side chains shown as spheres on a single monomer (panels A, B and C) or all monomers (D). (A) shows the front of the hexamer, viewed from outside the viral core; (B) shows the rear of the hexamer, viewed from inside the core; (C) shows a view from the side; and (D) shows the same view as (A).

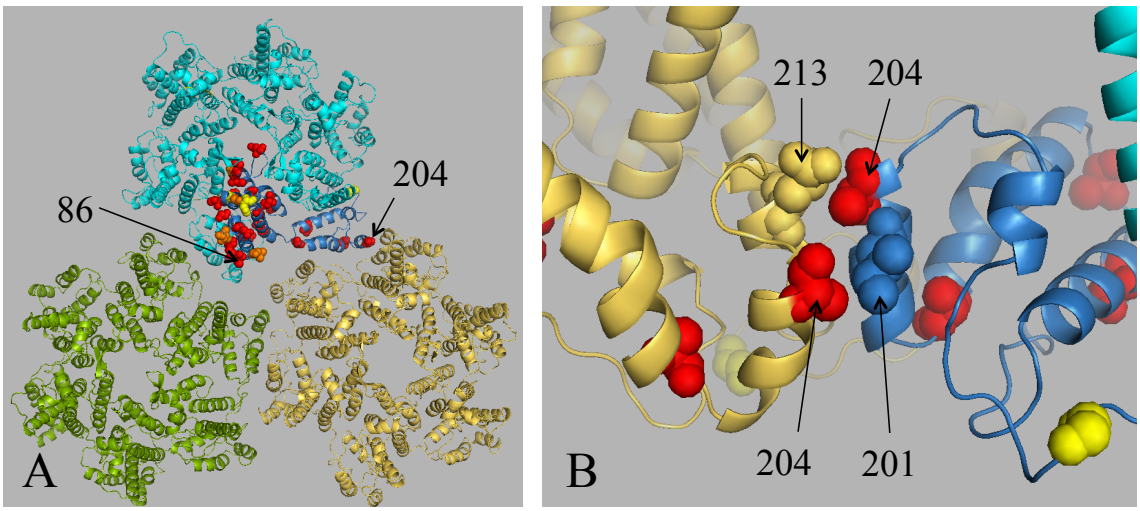

Figure 3.6: Identified sites mapped onto a capsid structure in complex with other monomers forming interacting hexamers (PDB ID 3J3Q, Zhao et al., 2013a; shown in cyan, green and dark yellow, respectively). Both N and C terminal domains of the monomer of interest are coloured blue, with side chains of sites identified with swMutSel (red), codeml M8 (yellow) or both (orange) shown as spheres. (A) Three interacting hexamers, with site 86 in the cyclophilin binding loop and site 204 in the C terminal domain indicated. (B) Enlarged view of interacting C terminal domains in neighbouring hexamers, showing pairs of interacting sites, of which site 204 was identified with swMutSel.

Model	Topology	T_i Total	ML $\setminus$ T_i	ML $\cap$ T_i	T_i $\setminus$ ML
swMutSel	T_1	21	7 *(13, 14, 47, 86, 116, 171, 194)*	16	5 *(6, 27, 148, 208, 210)*
”	T_2	23	3 *(13, 116, 171)*	20	3 *(10, 123, 208)*
”	T_3	24	5 *(14, 86, 110, 116, 171)*	18	6 *(10, 123, 148, 154, 208, 210)*
”	T_4	25	3 *(14, 83, 116)*	20	5 *(64, 92, 120, 208, 210)*
”	T_5	26	3 *(13, 14, 116)*	20	6 *(64, 123, 148, 154, 208, 210)*
codeml M8	T_1	6	1 *(50)*	6	0
”	T_2	11	0	7	4 *(6, 15, 96, 180)*
”	T_3	6	1 *(50)*	6	0
”	T_4	10	0	7	3 *(15, 96, 154)*
”	T_5	6	1 *(50)*	6	0

Table 3.2: Sites identified with alternative tree topologies, $T_{1\text{-}5}$, using swMutSel (top; multiple hypothesis test adjusted $p < 0.05$ from LRT with the degrees of freedom $N - 1$, where N is the number of residues observed at the site) or codeml's M8 (bottom; BEB probability > 0.95). Columns: [T_i Total], total number of sites found with topology T_i; [ML $\setminus$ T_i], number of sites found with ML tree but not T_i; [ML $\cap$ T_i], number of sites in intersection; [T_i $\setminus$ ML], number of sites found with T_i but not ML topology. Where there are discrepancies in the sites found with an alternative topology compared with the ML topology, the capsid site indices are given in italics.

3.3.4 Observed Residue Distributions

We examined the distributions of amino acids observed at the alignment sites we had identified with swMutSel and the ML topology. At many of the sites the frequencies of residues differed between the HIV-1 M and SIVcpz groups but with some overlap in the residues observed; e.g. at site 68 (fig. 3.7), where methionine was observed in virtually all HIV-1 M sequences (with a small number of paraphyletic isoleucine residues) while a mixture of valine, isoleucine and methionine are observed in SIVcpz sequences. At other sites there was a more pronounced difference between SIVcpz *Pts* and SIVcpz *Pts*, such as site 141 (fig. 3.8), where leucine was observed in all but one SIVcpz *Pts*, while valine or isoleucine was seen in SIVcpz *Pts* and isoleucine was seen in almost every HIV-1 M sequence. At a minority of identified sites the distributions of amino acids were very similar between HIV-1 M and SIVcpz, such as site 204 (fig. 3.9), where alanine was observed in nearly every sequence, except for a minority of paraphyletic glycine residues in HIV-1 M; strikingly, not a single glycine was observed in HIV-1 M subtype B despite this group being better represented in the dataset. Site 204 is involved in inter-hexamer interactions (fig. 3.6; Gres et al., 2015).

3.3.5 Substitutions Involved in Host Shift

To further investigate the evolutionary events in the zoonosis which established HIV-1 M, we sought to estimate which specific amino acid changes had occurred around the time of the cross species transmission. Using the WAG amino acid substitution model (Whelan and Goldman, 2001) and our capsid alignment translated to protein sequences, we computed the probabilities of each pair of amino acids states existing at the node representing the most recent common ancestor (MRCA) of both a chimpanzee virus and the human viruses together and the node representing the MRCA of only human viruses, conditional on model parameters and previously estimated branch lengths. To provide more evolutionary context, we used a larger ML tree comprising the HIV-1 M and SIVcpz sequences described above, and in addition sequences from HIV-1 groups N, O and P, together with sequences from viruses infecting gorillas (SIVgor), all of which ultimately originate from independent SIVcpz cross species transmissions (Sharp and Hahn, 2011).

Eight sites had a non-identical amino acid pair with transition probability > 0.5

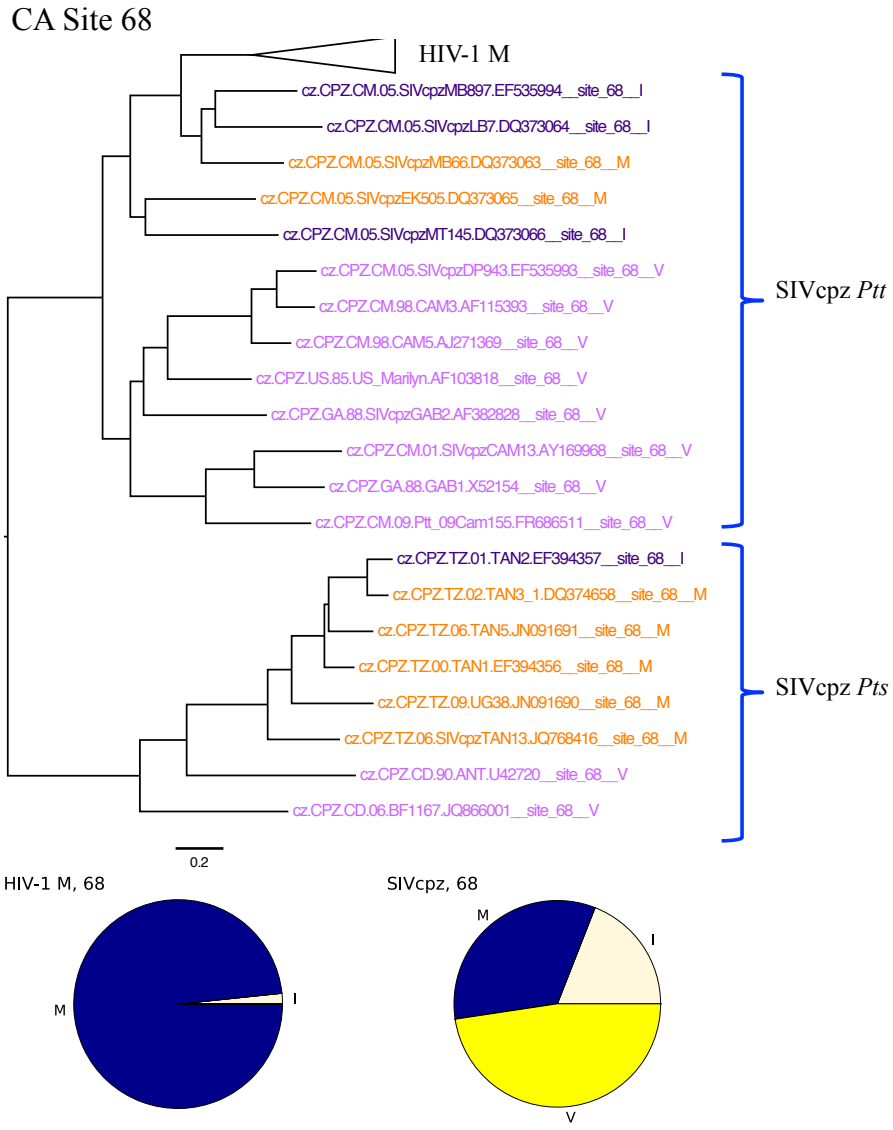

Figure 3.7: Residues observed at capsid site 68, on annotated maximum likelihood phylogeny (top, with HIV-1 M clade collapsed for clarity) and in pie charts (below) showing residue frequencies in 1332 HIV-1 M sequences and 21 SIVcpz sequences. Amino acids are specified by single letter code at the end of the taxon names or annotating the pie charts. The sets of SIVcpz sequences isolated from each chimpanzee subspecies (*Pan troglodytes troglodytes, Ptt* and *Pan troglodytes schweinfurthii, Pts*) are indicated. Branch support is indicated in figure 3.1. (NB the amino acid colour schemes are not consistent between the tree and the pie charts.)

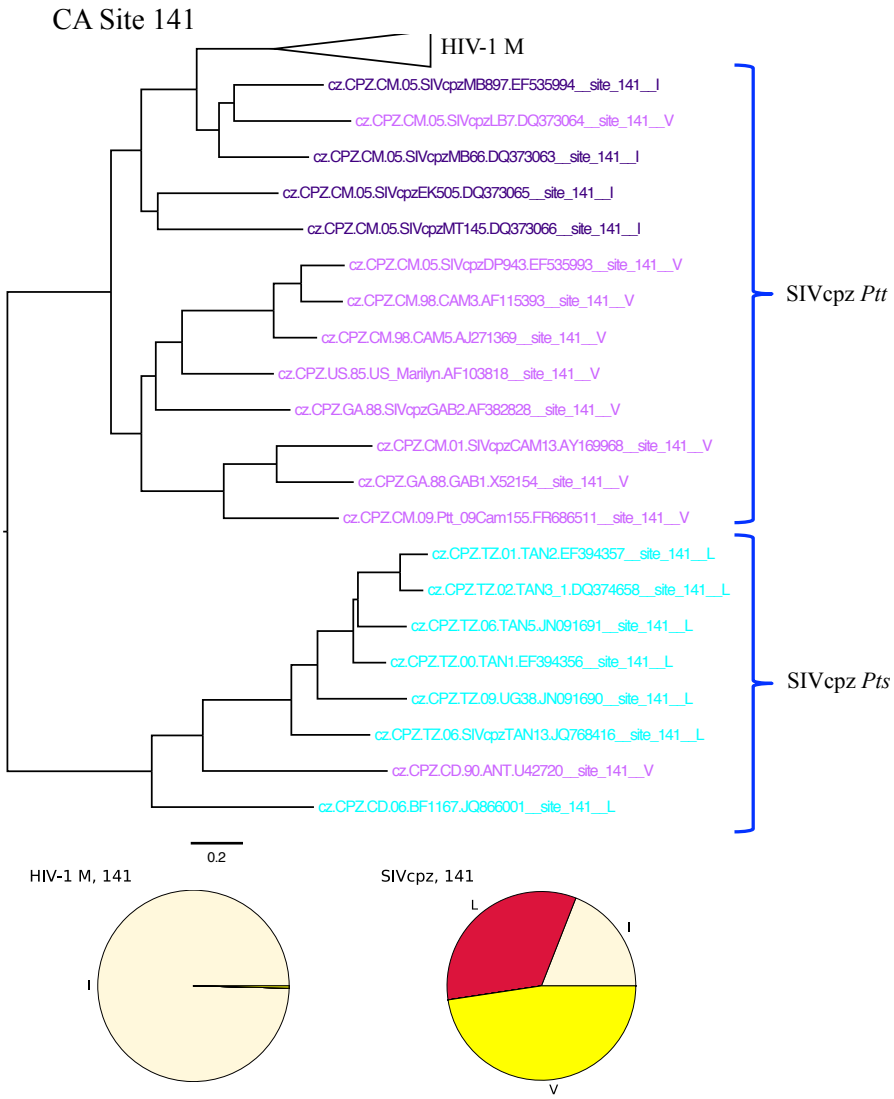

Figure 3.8: Residues observed at capsid site 141, on annotated maximum likelihood phylogeny (top, with HIV-1 M clade collapsed for clarity) and in pie charts (below) showing residue frequencies in 1332 HIV-1 M sequences and 21 SIVcpz sequences. Amino acids are specified by single letter code at the end of the taxon names or annotating the pie charts. The sets of SIVcpz sequences isolated from each chimpanzee subspecies (*Pan troglodytes troglodytes, Ptt* and *Pan troglodytes schweinfurthii, Pts*) are indicated. Branch support is indicated in figure 3.1. (NB the amino acid colour schemes are not consistent between the tree and the pie charts.)

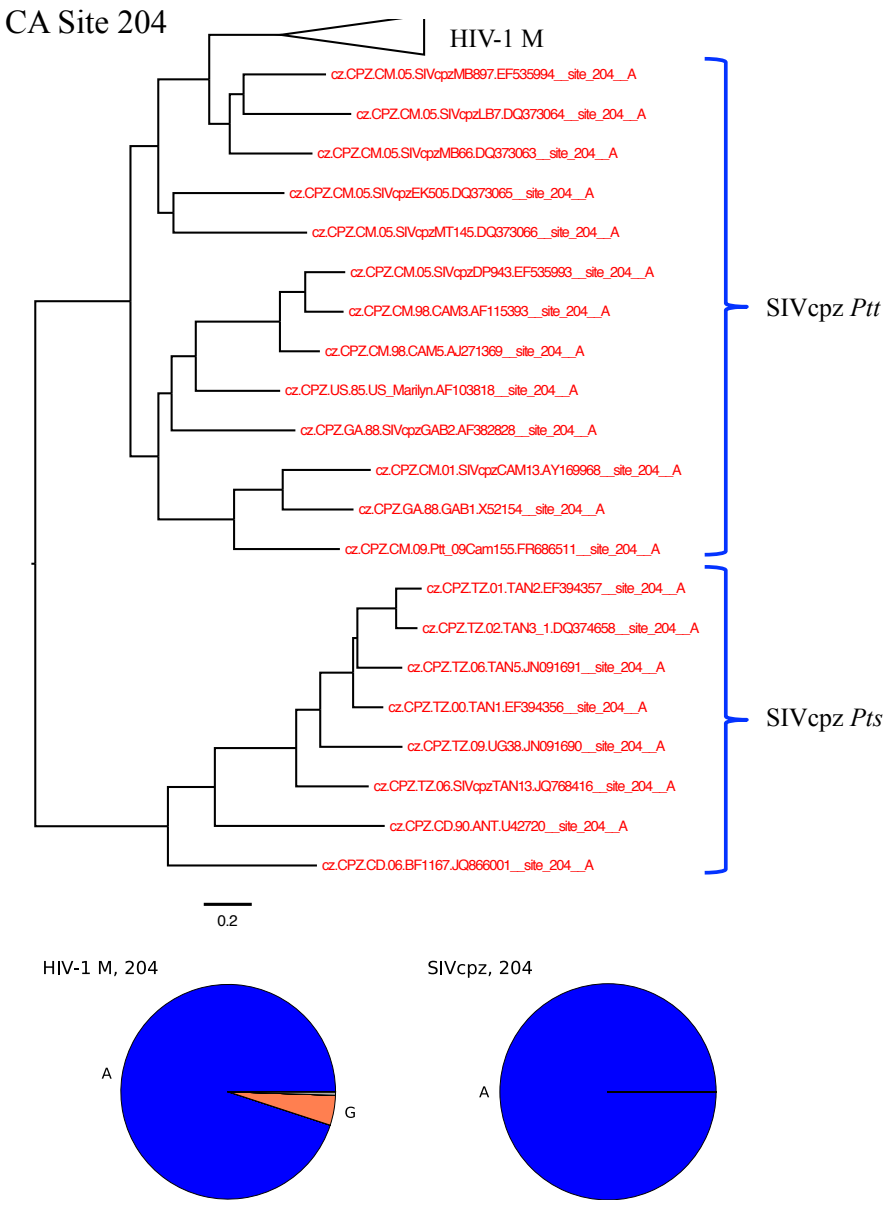

Figure 3.9: Residues observed at capsid site 204, on annotated maximum likelihood phylogeny (top, with HIV-1 M clade collapsed for clarity) and in pie charts (below) showing residue frequencies in 1332 HIV-1 M sequences and 21 SIVcpz sequences. Amino acids are specified by single letter code at the end of the taxon names or annotating the pie charts. The sets of SIVcpz sequences isolated from each chimpanzee subspecies (*Pan troglodytes troglodytes, Ptt* and *Pan troglodytes schweinfurthii, Pts*) are indicated. Branch support is indicated in figure 3.1. (NB the amino acid colour schemes are not consistent between the tree and the pie charts.)

CA Site	Transition	Probability
6	A → L	0.999
41	M → T	0.813
72	V → T	0.999
115	V → I	0.789
120	S → H	0.822
128	D → E	0.855
200	S → T	0.536
203	R → K	0.994

Table 3.3: Conditional probabilities for capsid amino acid transitions along the HIV-1 M ancestral branch, computed with the WAG substitution model, having first estimated branch lengths with the same model implemented in aaml (Whelan and Goldman, 2001; Yang, 2007). Shown are transitions with probability > 0.5 at sites where at least one transition nonidentical transition met this threshold. Transitions are presented as if *from* the most recent common ancestor (MRCA) of a SIVcpz and HIV-1 M *to* the MRCA of HIV-1 M only; amino acid states are shown by conventional letter codes.

(table 3.3). Several of these were chemically conservative changes (e.g. negatively charged aspartate to glutamate transition at site 128), with the exception of sites 41 (hydrophobic methionine to polar threonine), 72 (hydrophobic valine to polar threonine) and 120 (polar serine to positive histidine). From the set, site 41 was identified with swMutSel and 120 with M8 (table 3.1).

We mapped these 8 sites onto the capsid hexamer structure (fig. 3.10). Most sites are positioned on the external side, on or around the β-hairpin loop (such as site 6), but none on the cyclophilin binding loop. Two sites (200 and 203) project out from the CTD and are involved in inter-hexamer interactions (Zhao et al., 2013a; see fig. 3.6.)

3.4 Discussion

We have sought to identify positions within the viral CA protein which have evolved under different selective constraints in SIVcpz compared with HIV-1 M, with the aim of informing experimental investigation. Using swMutSel, we have identified 23 out of 231 CA sites as having statistical support for undergoing different evolutionary processes in each virus group. Importantly, we sought to distinguish sites experiencing distinct constraints from sites undergoing diversifying selection: we also identify 7 sites with codeml's model M8 (which identifies positive selection using a distribution of d_N/d_S ratios without distinguishing lineages). The sites found

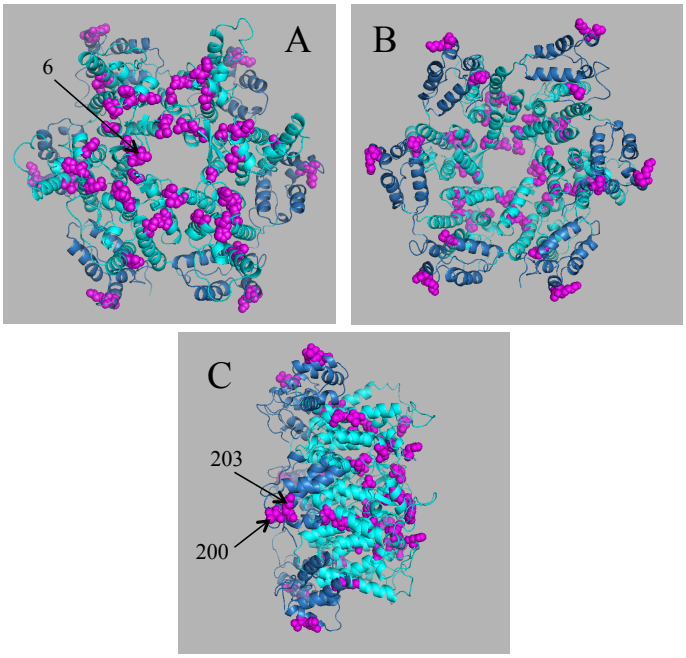

Figure 3.10: Sites with transitional probabilities > 0.5 for non-identical residues mapped onto capsid hexamer structure (PDB ID 3J3Q; Zhao et al., 2013a), shown as magenta spheres. For each monomer the N terminal domain and connecting linker (sites 1-149) are shown in cyan and the C terminal domain is shown in blue. (A) shows the front of the hexamer, viewed from outside the viral core; (B) shows the rear of the hexamer, viewed from inside the core; (C) shows a view from the side. Sites 6, 200 and 203 are indicated.

with M8 are positioned on the exterior of CA hexamers (fig. 3.5), consistent with diversifying selection being confined to less structured regions. This may be driven by the need to escape from cellular immunity; indeed epitopes recognised by cytotoxic lymphocytes (CTL) have been identified in all regions of the HIV-1 M capsid (Llano et al., 2013).

Four of the seven sites found with M8 were also identified with swMutSel (table 3.1) and the question arises what type of selection is active at these sites, since M8 is likely to be suited to identifying diversifying selection, while swMutSel was developed to identify directional selection. Data from a simulation study indicate codeml's site models such as M8 are poor identifiers of directional selection, though the lineage-specific model has higher sensitivity (Thiltgen et al., 2016). It therefore seems most likely that these are sites undergoing diversifying selection in one virus group or the other (e.g. due to different CTL epitope positions), accounting for the identification by swMutSel as experiencing different selective constraints.

Several of the sites identified with swMutSel are positioned within the loop on the exterior of the hexamer to which cyclophilins bind, namely cyclophilin A (CypA)

and NUP358 (fig. 3.3; Thali et al., 1994; Gamble et al., 1996). Catalysis of the cis-trans isomerization of the Gly89-Pro90 peptide bond of CA by CypA, which induces conformational rearrangements away from the binding site itself, may influence core disassembly (Bosco et al., 2002). Like HIV-1 M, SIVcpz is dependent on cyclophilin cofactors (Laura Hilditch and Greg Towers, manuscript in preparation) and finding these sites to be under different selective constraints in the two groups suggests they each mediate different interactions with the host cyclophilins, or that the outcome of cyclophilin interaction is different. As interaction with CypA is required for HIV-1 M to avoid innate immune sensing in macrophages (Rasaiyaah et al., 2013) and interaction with NUP358-Cyp determines the region of the host genome where the virus integrates (Schaller et al., 2011), the significance of a host-specific CypA or NUP358-Cyp interaction could be profound. Experimental studies comparing HIV-1 M and SIVcpz CA behaviour should prioritise investigation of the identified sites, for example in mutation experiments.

We also identify sites in (i.e. site 5) and at the base of (i.e. site 13) the β-hairpin loop, the position of which modulates dNTP flux through a pore at the centre of the CA hexamer in HIV-1 M, probably to enable reverse transcription within the viral core (Jacques et al., 2016). Unpublished observations by the Greg Towers laboratory suggest the β-hairpin in HIV-1 group O CA is inflexible by comparison with group M CA, which suggests it may be unable to regulate the flux of dNTPs into the viral core. This may make it more vulnerable to innate immunity DNA sensors in macrophages, accounting for poorer growth kinetics they observe. Moreover, mutation of site 12 in HIV-1 M has been observed to dramatically impair flexibility of the β-hairpin, indicating that chemical properties of the residues surrounding the loop are important for regulating the activity of the hexamer pore. That these sites are evolving under different constraints in HIV-1 M and SIVcpz suggests that their β-hairpin loops may have different levels of flexibility. Indeed, it remains uncertain whether the β-hairpin in SIVcpz CA is flexible at all and, should this be a property unique to HIV-1 M, substitutions at the identified sites may have been responsible.

None of the identified sites are in the region bound by CPSF6 and TNPO3. This is unsurprising as this hydrophobic pocket has been noted to be conserved across diverse primate lentiviruses (Price et al., 2012). Like the cyclophilin-containing proteins, SIVcpz interacts with these cofactors (Laura Hilditch and Greg Towers, manuscript in preparation), but we find no evidence suggesting divergent interac-

tions by SIVcpz and HIV-1 M CA.

The majority of identified sites fall in the linking regions between helices (fig. 3.2), consistent with the helices themselves being conserved to maintain the overall shape of the CA hexamer, while sites in the linking regions could affect helix orientations in a dynamic structure. This may be because interactions with cofactors are common to both HIV-1 M and SIVcpz, but different structural implications of cofactor interaction may mediate host-specific CA activities (Greg Towers, personal communication). Movement of helices is also involved in the opening and closing of the hexamer pore (Jacques et al., 2016). Several of the identified sites are buried within the hexamer, potentially because because allosteric regulation of pore opening, or other putatively co-factor dependent behaviours, require different structural rearrangements in the two viruses. Some identified sites are within helices, however, and have side chains in close proximity to helices in neighbouring monomers (fig. 3.4). We also find sites in the CTD which are involved in hydrophobic interactions between hexamers (fig. 3.6; Gres et al., 2015), suggesting potentially divergent regulation of viral uncoating.

We also identified sites with high probability of having undergone amino acid substitution in the interval in which cross species transmission had occurred (table 3.3, fig. 3.10). Interestingly, site 6 is at the top of the β-hairpin and projects into the lumen of the hexamer pore (fig. 3.10A); the predicted substitution of the small alanaine to the bulkier leucine could have significantly reduced the rate at which dNTPs are trafficked into the viral core. Furthermore, substitutions at sites 200 and 203 in the CTD (fig. 3.10C) could influence the stability of inter-hexamer interactions (Gres et al., 2015), though these predicted substitutions were chemically conservative (polar S $\rightarrow$ T and positively charged R $\rightarrow$ K). Highly probably changes at other positions, such as sites 41 (M $\rightarrow$ T) and 72 (V $\rightarrow$ T) are both buried deep in the hexameric structure and chemically nonconservative, with potentially significant effects on hexamer structure. Furthermore, site 41 was identified with swMutSel, suggesting this site underwent a substitution following zoonotic transmission and continued experiencing different constraints in the new host. We note, however, this analysis is arguably less informative than our swMutSel results, as it does not account for ongoing evolutionary process in either group, and may identify substitutions which were present in the HIV-1 ancestor by chance ('founder effects') rather than due to novel selective constraints.

Unexpectedly, we found a difference in the distributions of residues in SIVcpz from *Pan troglodytes troglodytes* (*Ptt*) or *Pan troglodytes schweinfurthii* (*Pts*) at some of the identified sites, such as site 141 (fig. 3.8). In this work we have chosen to consider SIVcpz as a single group, because few sequences are available and we want to maximise statistical power, as swMutSel is a parameter-rich model. However, it may be appropriate to consider them as separate viruses (just as HIV-1 is divided into groups M, N, O and P), either by investigating divergent constraints in HIV-1 M and the more closely related SIVcpz *Ptt* only, or even comparing SIVcpz *Ptt* and SIVcpz *Pts* directly. The considerable sequence divergence between SIVcpz isolates (fig. 3.1) suggests in fact there may be sufficient statistical power from even a small dataset, and may be investigated in future work.

Again unexpectedly, we found sites where mostly the same residue is seen in both SIVcpz and HIV-1 M, but we observe a remarkable difference between HIV-1 M subtypes; e.g. at site 204 (fig. 3.9) alanine is predominant in both groups with few independently occurring glycine residues in all HIV-1 M subtypes except subtype B, where exclusively alanine is observed, despite this subtype being the best represented in our dataset ($n = 662$). This site is important for hexamer-hexamer hydrophobic interactions (fig. 3.6; Gres et al., 2015) and its identification with swMutSel suggests HIV-1 M as a whole is more tolerant of the flexible glycine at this position than SIVcpz. But in addition it appears there is selective pressure for greater conservation within subtype B; an observation which suggests future work should address divergent selective constraints between HIV-1 M subtypes and may yield indications of different strategies, such as alternative cofactor recruitment or timing of uncoating, specific to certain subtypes. Biological differences between subtypes are likely to be under-appreciated at present, since most experimental studies use laboratory virus strains derived from subtype B only.

In future work, it will be interesting to investigate whether these sites are undergoing interdependent substitution processes; that is, if they are co-varying. Unpublished software by Richard Goldstein is able to assess statistical support for co-evolution of a subset of protein sites and could be applied to the sites identified with swMutSel. Co-varying sites are likely to be engaged in the same function, or indicate sites where compensatory changes must be made.

In addition, it would be informative to examine a large set of CA structures solved under different conditions (such as pH, which affects pore opening; Jacques et al.,

2016), to quantify the extent to which the positions and side chain interactions of the identified sites are altered during secondary structure rearrangement. This may reveal connections involving the identified sites not apparent in a single structure.

Most significantly, the identification of this set of sites warrants investigation by mutagenesis experiments comparing the early events in SIVcpz and HIV-1 M replication. This will form the basis of an ongoing collaboration with the Greg Towers laboratory.

In conclusion, we have found evidence of sites in CA experiencing divergent selective constraints in SIVcpz and HIV-1 M, several of which are in positions known to be of significant biological importance. We have also found unexpected differences between SIVcpz *Ptt* and *Pts* and HIV-1 M subtypes. The identification of these sites contributes to our understanding of the evolutionary events involved in establishing a pandemic human pathogen, and will be used to guide experimental investigations into virus-host interactions.

3.5 Methods

3.5.1 Sequence Data and Phylogeny Estimation

A dataset of full length genome HIV-1 (groups M, N, O and P), SIVcpz and SIVgor sequences was obtained from the Los Alamos HIV Sequence Database, the 2012 curated dataset (`http://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html`). Known HIV-1 M inter-subtype recombinants and SIVcpz sequences known to be derived from the same molecular clone were excluded, leaving 1332 and 21 HIV-1 M and SIVcpz sequences, respectively. Capsid sequences (reference sequence HXB2 co-ordinates 1186-1879) were excised and aligned manually (using the alignment editor SEAVIEW version 4.4.0 Gouy et al., 2010) as protein sequences and mapped back to produce a codon alignment.

From the codon alignment, a phylogeny of all HIV-1 groups, SIVcpz and SIVgor was estimated by maximum likelihood (ML) using RAxML 7.7.2 HPC-HYBRID-AVX (Stamatakis, 2006; Ott et al., 2007), with GTR model of nucleotide substitution and Gamma-distributed rates. Clade support was assessed by 1000 non-parametric bootstrap replicates. For analyses involving HIV-1 M and SIVcpz only the same tree topology was used, with irrelevant sequences pruned.

RAxML saves each of the non-parametric bootstrap trees it estimates. For repeating selection analyses with alternative tree topologies, we used five of these bootstrap trees (limited computing resources precluded using more). RAxML only computes a new initial estimate tree (by maximum parsimony) for every 10th bootstrap dataset, while interim trees use the previously estimated bootstrap tree as the starting estimate. To ensure the alternative topologies used were independent and identically distributed, we used bootstrap trees which had had independent maximum parsimony trees as initial estimates.

Custom software written for rooting and pruning phylogenetic trees made use of the Phylo package in the Biopython library (Talevich et al., 2012). Tree figures were produced using FigTree 1.3.1 (Rambaut, 2006).

3.5.2 Selection Analysis

The site wise Mutation-Selection model (swMutSel) was written by Asif Tamuri and is open source (`https://github.com/tamuri/swmutsel`; Tamuri et al., 2009, 2012). All amino acid fitness parameters were estimated by ML, with mutation parameters (transition/transversion rate ratio and nucleotide equilibrium frequencies) and branch lengths estimated separately by FMutSel0 (Yang and Nielsen, 2008), implemented in codeml of the PAML package (version version 4.7a; Yang, 2007). The swMutSel alternative model allows a separate rate matrix for each group of taxa studied; all branches in the HIV-1 M clade were assigned to the HIV-1 M specific matrix and all remaining branches to the SIVcpz matrix, with the connecting branch divided in half between the two. For each site analysed the likelihood ratio test (LRT) was used to assess support for the alternative model, with $N-1$ degrees of freedom, where N is the number of amino acids observed at the site. Resulting p values were corrected for multiple hypothesis testing at each site analysed separately by adjusting p value threshold such that the false discovery rate is reduced to 5% (Benjamini and Hochberg, 1995).

Codeml (PAML) models M7 (null) and M8 (alternative) were also applied, with support for M8 assessed with the LRT using 2 degrees of freedom. Default starting parameter values were used.

Sites identified were mapped onto a capsid crystal structure (PDB ID 3J3Q; Zhao et al., 2013a) and images produced with PyMol 1.3 (Schrödinger, 2010).

3.5.3 Computing Ancestral Transition Probabilities

Probabilities for specific amino acid transitions along the HIV-1 M ancestral branch at each site h in dataset X were computed with a modified implementation of the pruning algorithm (Felsenstein, 1981), where we compute the conditional probability of amino acid I being the state existing at node A and amino acid J being the state existing at node B, given the substitution model parameters θ and the data. We compute this for each of 20^2 pairs of states and normalise by the total probability of the data conditional on the model parameters only (i.e. as conventionally computed with the pruning algorithm):

$$P(I_A, J_B|X_h, \theta) = \frac{P(I_A, J_B|\theta)P(X_h|I_A, J_B, \theta)}{P(X_h|\theta)} = \frac{P(I_A)P(I_B|I_A, \theta)P(X_h|I_A, J_B, \theta)}{\sum_{I'_A}\sum_{J'_B} P(I'_A)P(I'_B|I'_A, \theta)P(X_h|I'_A, J'_B, \theta)} \tag{3.1}$$

Note that $P(I) = \pi_I$ is the equilibrium frequency of I and $P(I_B|I_A, \theta) = P(I_A \rightarrow I_B|\theta)$ is the transition probability along the branch connecting A and B. We used the WAG matrix substitution rates and equilibrium frequencies (Whelan and Goldman, 2001) and branch lengths estimated with this same matrix, as implemented in aaml in the PAML package (Yang, 2007).

Our implementation of this computation was written in Java and made use of the Phylogenetic Analysis Library for phylogenetics specific data structures (Drummond and Strimmer, 2001) and Apache Commons Math for linear algebra (Apache-Commons, 2015).

Chapter 4

Divergent Selective Constraints in HIV-1 M/SIVcpz Accessory Proteins

4.1 Summary

Primate lentiviruses posses accessory proteins which serve to manipulate the innate immune response, often by neutralising host innate immunity factors which restrict virus replication. Interactions with host proteins are often species specific, and therefore when a virus spreads to a new host adaptation may be required. These restriction factors show signs of diversifying selection, as amino acid changes occur to prevent recognition by the accessory protein. Much less studied is the evolution of the accessory proteins themselves. We aimed to investigate the species specific adaptation of pandemic HIV-1 group M, compared with its progenitor SIVcpz, which infects chimpanzees. Using a site-wise mutation selection model, we have analysed the accessory genes *nef*, *vpu* and *vpr* for evidence of evolving under different selective constraints in HIV-1 M compared with SIVcpz. Surprisingly, in *nef* and *vpr* we identify sites involved in putatively conserved interactions with host proteins, suggesting unexpected host specific adaptation. In *vpu*, we identify sites involved in the antagonism of the restriction factor tetherin — a function acquired by HIV-1 M during its adaptation to humans — together with sites which we hypothesise are similarly involved. This work demonstrates adaptation of a pandemic pathogen to the human host and should inform experimental investigation of virus-host interactions.

Protein	Functions	Reference
Nef	CD4 downregulation	Garcia and Miller (1991)
	MHC-1 downregulation	Schwartz et al. (1996)
	Signalling disruption	Saksela et al. (1995)
	Tetherin antagonism (SIVcpz)	Sauter et al. (2009)
	SERINC3/5 antagonism	(Usami et al., 2015)
Vpu	CD4 downregulation	Willey et al. (1992)
	Tetherin antagonism (HIV-1 M)	Neil et al. (2008); Van Damme et al. (2008)
Vpr	Cell cycle arrest	Re et al. (1995)
	Apoptosis	Stewart et al. (1997)
	Immune signalling mitigation	Laguette et al. (2014)

Table 4.1: Summary of known functions for the HIV-1 M/SIVcpz accessory proteins studied in this work.

4.2 Introduction

All retroviruses possess three essential genes, known as *gag*, *pol* and *env*, whose products form structural or enzymatic components of viral particles. Lentiviruses are genetically distinguished from more simple retroviruses by also possessing genes involved in viral gene expression (*tat* and *rev*) and accessory, or auxiliary genes, so called because they can be dispensable for infection in laboratory cells lines. HIV-1 possesses four accessory proteins: *vif*, *nef*, *vpu* and *vpr* (table 4.1). Research in recent years has shown that these accessory proteins serve to mitigate the host innate immune response to enable viral replication. A common theme is antagonism of 'restriction factors', host proteins which appear to have evolved specifically to restrict virus replication, and are often induced by interferon (Malim and Emerman, 2008). The first of these to be recognised was APOBEC3G, a cytosine deaminase which can introduce hypermutation of the retroviral genome. By targeting it for destruction through the E3 ubiquitin ligase pathway, the HIV-1 Vif protein degrades APOBEC3G (Sheehy et al., 2002, 2003; Lecossier et al., 2003; Yu et al., 2003). Diversifying selection has been identified in many restriction factors, apparently driven by pressure to escape recognition by antagonistic accessory proteins (Sawyer et al., 2004, 2005; Gupta et al., 2009; McNatt et al., 2009; Laguette et al., 2012; Lim et al., 2012). The evolution of the accessory proteins themselves, however, is much less studied.

The HIV-1 M Nef protein has been shown to influence progression to AIDS in both animal models and patients (Kestler et al., 1991; Kirchhoff et al., 1995) and

has been implicated in several functions. Nef downregulates cell surface expression of CD4, the primary virus receptor in T cells and macrophages (Garcia and Miller, 1991), by interactions with both CD4 and a clathrin adaptor protein complex (AP-2), which causes CD4 to be absorbed into endosomal compartments (Aiken et al., 1994; Lindwasser et al., 2008; Chaudhuri et al., 2009). The purpose of the downregulation is suggested to be for preventing superinfection, allowing efficient release of new virions by preventing interaction with the receptor and/or disrupting T cell activation. This is a conserved behaviour, as SIVcpz can similarly downregulate human CD4 (Sauter et al., 2009). Nef also downregulates the major histocompatability complex 1 (MHC-1; Schwartz et al., 1996), again through a clathrin adaptor protein (AP-1; Greenberg et al., 1998), which prevents antigen presentation and therefore prevents the cell being destroyed by the cytotoxic immune response (Collins et al., 1998). This too is apparently conserved, as MHC-1 downregulation by HIV-1 O and SIVgor has been reported (Kluge et al., 2014), which, like HIV-1 M, both originate from SIVcpz. In addition, Nef interferes with cell signalling by interactions with Src family tyrosine kinases, specifically their common SH3 domains (Saksela et al., 1995), which induces gene expression changes in T cells resembling immune activation. Recent observations have implicated HIV-1 M Nef with the downregulation of the putative restriction factor proteins SERINC3 and SERINC5 (Usami et al., 2015). This permits more efficient virus replication, though the exact mechanism of restriction by SERINC proteins or their antagonism by Nef remain unclear (Trautz et al., 2016).

Vpu comprises a single helix transmembrane domain and a cytoplasmic domain made from two short helices (Maldarelli et al., 1993). Though by a different mechanism to Nef, Vpu also induces CD4 downregulation (Willey et al., 1992) and this appears to be a conserved Vpu behaviour in all HIV-1 groups and SIVs infecting both apes and some monkeys (Sauter et al., 2009). The most important function of HIV-1 M Vpu, however, appears to be antagonism of the host restriction factor tetherin (BST2), which in the absence of Vpu 'tethers' budding virions to the producer cell surface (Neil et al., 2008; Van Damme et al., 2008). Vpu and tetherin interact through their transmembrane domains and E3 ubiquitin ligase is recruited to target tetherin for proteosomal degradation. Surprisingly, human tetherin antagonism by Vpu was apparently acquired following the cross species transmission from chimpanzees which established HIV-1 M, as SIVcpz Vpu antagonises neither human

nor chimpanzee tetherin. Instead SIVcpz uses its Nef protein (Sauter et al., 2009). SIVcpz Nef is ineffective against human tetherin due to a small deletion relative to its chimpanzee and gorilla tetherin orthologues.

Vpr remains the least understood accessory protein, having not been associated with any specific function with a clear benefit to the virus life cycle. Putative apoptosis induction has been described (Stewart et al., 1997), together with incorporation into virions, suggesting a role in the early stages of infection (Cohen et al., 1990). The most prominent activity attributed to Vpr is arrest of the cell cycle at the G2/M checkpoint in cycling cells (Re et al., 1995), the purpose of which is still unclear. This is a conserved behaviour, having been observed originally in HIV-1 M Vpr and then Vprs from HIV-2 and SIVs infecting sooty mangabeys and African green monkeys (Planelles et al., 1996; Stivahtis et al., 1997).

Via residues 60-68 in its core helix bundle, Vpr interacts with the E3 ubiquitin ligase machinery by binding DCAF1 (Zhao et al., 1994), which acts to recognise substrates to be targeted for proteosomal degradation, and this activity is required for cell cycle arrest (Belzile et al., 2007; DeHart et al., 2007; Hrecka et al., 2007). Similarly, the related Vpx protein found in the HIV-2/SIVsm lentivirus lineage interacts with DCAF1, with the purpose of degrading the restriction factor SAMHD1 (Hrecka et al., 2011; Laguette et al., 2011); however, no target for degradation by Vpr has been reported.

Recent work (Laguette et al., 2014) has uncovered some detail of the mechanism underlying Vpr-mediated cell cycle arrest, identifying Vpr's interaction with factors comprising a complex surrounding the SLX4 scaffold protein (collectively, SLX4com), which ordinarily coordinates DNA repair events during S phase. Vpr activates the complex prematurely, disrupting its role in DNA repair and therefore preventing progress through the G2/M checkpoint. Significantly, Vpr interaction with SLX4com was associated with preventing interferon induction, which suggests a role for Vpr in avoiding innate immune sensing. The arrest of the cell cycle and interaction with both DCAF1 and SLX4 have since been replicated with a panel of Vprs from SIVs infecting several Old World monkeys encompassing wide lentiviral divergence, suggesting the activities were present in the ancestral primate lentivirus Vpr (Berger et al., 2015). Moreover, species specificity was observed, with some SIVs incapable of cell cycle arrest and SLX4 interaction in human cell lines, while able to do so in African green monkey cells.

As species specificity is a common feature of accessory proteins' interactions with host factors, it suggests adaptation is required following each transmission to a new host, as has occurred multiple times throughout primate lentivirus evolution (Sauter et al., 2009; Laguette et al., 2012; Lim et al., 2012; Kluge et al., 2014; Sharp and Hahn, 2011). Antagonism by these viral proteins involves direct interactions, and the target restriction factors have shown phylogenetic signatures of diversifying selection at binding interfaces, as non-specific amino acid change is favoured to prevent recognition by the antagonist (Duggal and Emerman, 2012). Models such as those implemented in codeml (Yang, 2007) which identify elevated rates of indiscriminate nonsynonymous codon substitutions have proved highly successful in identifying this type of selection (Sawyer et al., 2004; McNatt et al., 2009; Laguette et al., 2012; Lim et al., 2012). Conversely, we would expect adaptation by the accessory protein to involve specific changes to fulfill the new selective constraints associated with the change of host, as only a narrower set of amino acids in the binding sites will be able to mediate an interaction.

We have sought to characterise the host-specific adaptation of the HIV-1 M accessory genes *nef* and *vpu* following transmission to humans from chimpanzees, as these genes have contrasting interactions with the restriction factor tetherin depending on the host and virus. In addition we have analysed HIV-1 M and SIVcpz *vpr* as this is of special interest given recent investigations into Vpr function (Laguette et al., 2014). Using a site-wise mutation selection model (swMutSel; Tamuri et al., 2012), we have tested the hypothesis that the selective constraints in these proteins differ in HIV-1 M compared with SIVcpz. From this we have identified sites undergoing divergent constraints in each accessory gene. Surprisingly, in *nef* and *vpr*, many of these sites are in regions of the proteins involved in conserved functions, suggesting previously unrecognised species specificity of the interaction. In *vpu*, we identify sites known to be involved the interaction between HIV-1 M Vpu and human tetherin, suggesting these sites underwent directional selection as Vpu acquired a new function. We hypothesise other identified sites in Vpu may be similarly involved, which will be of interest to experimentalists investigating Vpu-tetherin interactions.

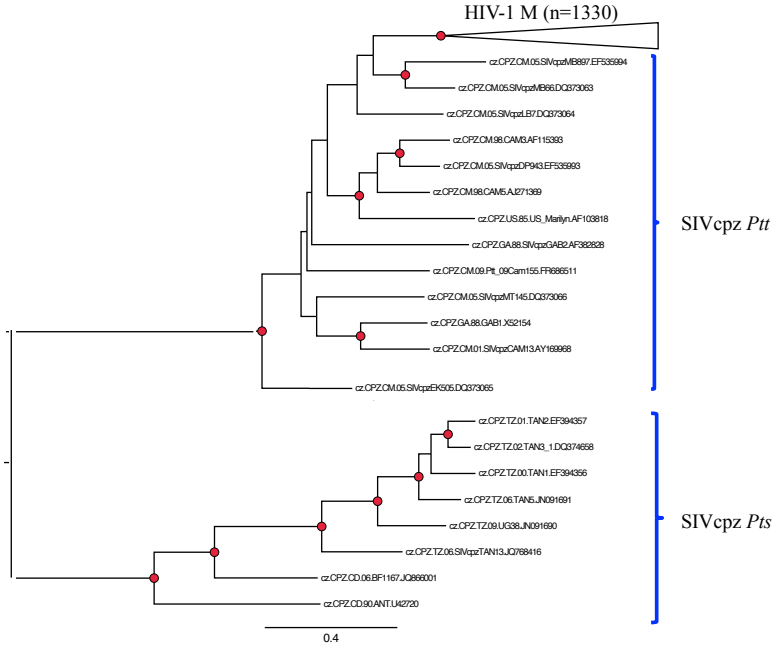

Figure 4.1: Maximum likelihood phylogeny for 1330 HIV-1 and 21 SIVcpz Nef sequences, estimated by RAxML. The HIV-1 clade is collapsed for clarity. The SIVcpz sequences are divided by the subspecies of their respective hosts, *Pan troglodytes troglodytes* (*Ptt*) and *Pan troglodytes schweinfurthii*; (*Pts*), and the tree is rooted with the latter as the outgroup (Sharp and Hahn, 2011). Nodes with $> 70\%$ support (from 1000 non-parametric bootstrap datasets) are indicated by red circles. Branch lengths are shown as nucleotide substitutions per site.

4.3 Results

4.3.1 Nef

34 Sites Identified as Experiencing Divergent Selective Constraints

We obtained a dataset comprising 1330 HIV-1 M and 21 SIVcpz Nef codon sequences from the Los Alamos HIV sequence database, which we aligned manually. We estimated a phylogeny by maximum likelihood (ML; fig. 4.1), for which non-parametric bootstrapping indicated high support for deep internal nodes, but lower support within the HIV-1 clade (not shown). With this codon alignment and ML topology, we applied the site-wise mutation selection model (swMutSel; Tamuri et al., 2009, 2012), testing the hypothesis that the substitution process differs for sites evolving in the human or chimpanzee hosts.

Using the likelihood ratio test (LRT) we identified 34 sites for which the null hypothesis of homogeneous selective constraints was rejected with $p < 0.05$, correcting for multiple hypothesis testing (table 4.2 and table E.1). Separately, we fitted models M7 and M8 implemented in codeml (Yang, 2007) for comparison and found the

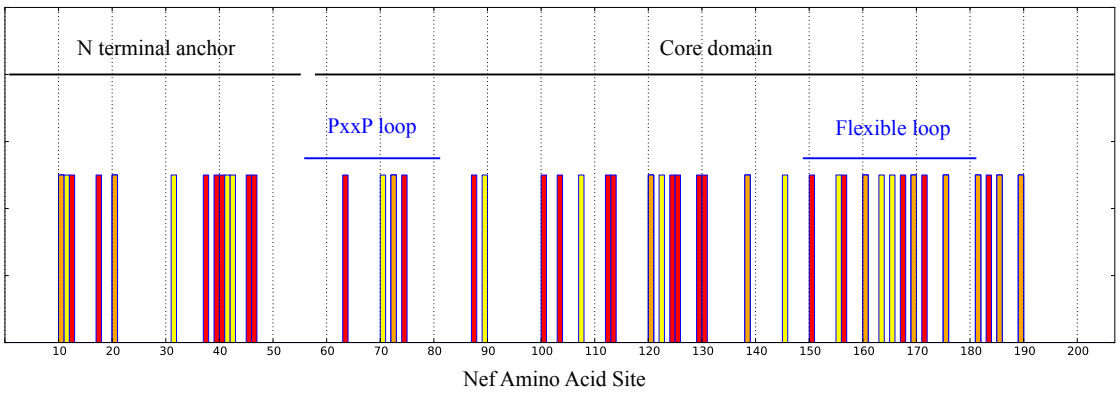

Figure 4.2: Sites identified in HIV-1 M and SIVcpz Nef with swMutSel (red), codeml's M8 (yellow) or both (orange). The height of the bars is of no significance. Secondary structure elements (as defined by Geyer et al., 2001) are indicated.

M8 model of positive selection was statistically justified with the LRT ($p < 0.05$). 23 sites had estimated Bayes empirical Bayes (BEB) probabilities > 0.95 for belonging to a positive selection site class (table 4.2) and no others had BEB probability > 0.5. 11 of these had also been identified with swMutSel. The sites found with either method were distributed across the Nef primary sequence (fig. 4.2).

Interaction with Host Factors

We mapped the identified sites onto Nef protein structures. The amphipathic nature of the protein has prevented a full-length structure being determined, but a composite structure produced by modelling interactions of Nef domains has been described (Geyer and Peterlin, 2001). The identified sites are distributed across the structure, without obvious clustering (fig. 4.3). The CD4 interaction surface of Nef comprises a hydrophobic patch on the core domain (fig. 4.3, cyan), which is highly conserved in HIV-1 M (Geyer and Peterlin, 2001). Consistent with this activity being conserved in both viruses, none of these sites are identified in our analyses, though sites 54, 100 and 113 identified with swMutSel only are in close proximity (fig. 4.3).

Since CD4 downregulation is dependent on interaction with subunits of AP-2, we considered the same sites in the context of AP-2 binding. The interaction is mediated by sites 160-165, with contributions from sites 174-175, which are mostly conserved within HIV-1 M (Lindwasser et al., 2008; fig. 4.4, cyan). Surprisingly, site 163 was identified by swMutSel as undergoing divergent selective constraints. This interacts with $\sigma 2$ residues N97 and V98 which form part of a loop connecting two

Nef site	HXB2 Res.	swMutsel ($p < 0.05$)	M8 BEB > 0.95
14	P	+	+
15	T		+
16	V	+	
21	R	+	
24	E	+	+
39	K		+
45	S	+	
47	N	+	
48	T	+	
49	A		+
50	A		+
53	A	+	
54	A	+	
76	L	+	
83	A		+
85	V	+	+
87	L	+	
100	L	+	
102	H		+
113	W	+	
116	H	+	
120	Y		+
125	Q	+	
126	N	+	
133	V	+	+
135	Y		+
137	L	+	
138	T	+	
142	C	+	
143	Y	+	
151	D	+	+
158	K		+
163	S	+	
168	V		+
169	S	+	
173	M	+	+
176	P		+
178	R		+
180	V	+	
182	E	+	+
184	R	+	
188	R	+	+
194	V	+	+
196	R	+	
198	L	+	+
202	Y	+	+

Table 4.2: Sites identified in HIV-1 M and SIVcpz Nef with swMutSel (likelihood ratio test with degrees of freedom $N - 1$, where N is the number of residues at the site; corrected for multiple hypothesis testing) and/or codeml model M8 (Bayes empirical Bayes probability for belonging to positive selection site class > 0.95). Nef amino acid residues are shown the HIV-1 reference sequence HXB2 (accession K03455).

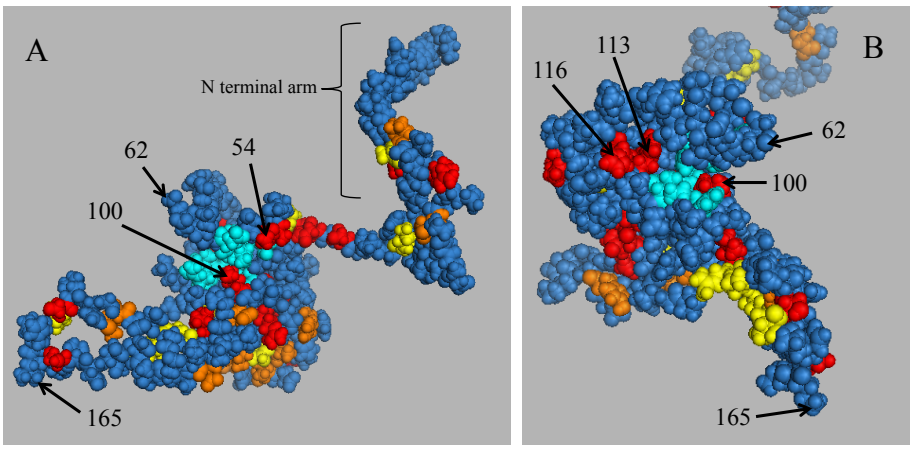

Figure 4.3: Sites identified with swMutSel and M8 mapped onto HIV-1 M Nef composite structure, described by Geyer and Peterlin (2001) and kindly provided by Matthias Geyer (personal communication). All atoms are shown as spheres. Cyan, sites implicated in CD4 binding (reviewed by Foster et al., 2011); red, sites identified with swMutSel; yellow, sites identified with M8; orange, sites identified with both. Sites 54, 100, 113 and 116 identified with swMutSel and close to the CD4 interacting region are indicated. Sites 62 and 165 are indicated for reference between panels. (A) The N terminal domain is indicated, shown in an 'open' conformation, away from the core domain in the centre of the structure. (B) Rotated view of (A), approximately 45° to the right, with part of the N terminal domain obscured.

helices (fig. 4.4B). In addition, Nef site 169, also identified with swMutSel interacts with σ2 residue A63 (fig. 4.4C) and Nef site 173, identified with both swMutSel and M8, interacts with residue Q301 of subunit α2 (fig. 4.4D).

We next considered interactions relevant to MHC-1 downregulation, which we would expect to be conserved in SIVcpz as well as HIV-1 M, as this has been observed in HIV-1 O and SIVgor (Kluge et al., 2014). We mapped the identified sites onto a crystal structure comprising Nef, the cytoplasmic domain of MHC-1 and subunit μ1 of AP-1 (Ren et al., 2014; fig. 4.5). Unexpectedly, the identified sites 76, 138 and 139 were found to be positioned along the MHC-1 binding interface (fig. 4.5A), interacting with MHC-1 sites A323/A324, A329 and Q330, respectively. Moreover, Nef sites 126 and 137 found with swMutSel, together with site 202 found with both swMutSel and M8, form a binding pocket in which resides a loop within subunit μ1 (fig. 4.5B).

Nef Alternative Tree Topologies

To test the sensitivity of our results to the tree topology used, we repeated the analyses with swMutSel and M8 using 5 additional tree topologies, originating from estimates from non-parametric bootstrap datasets. Using swMutSel, most sites

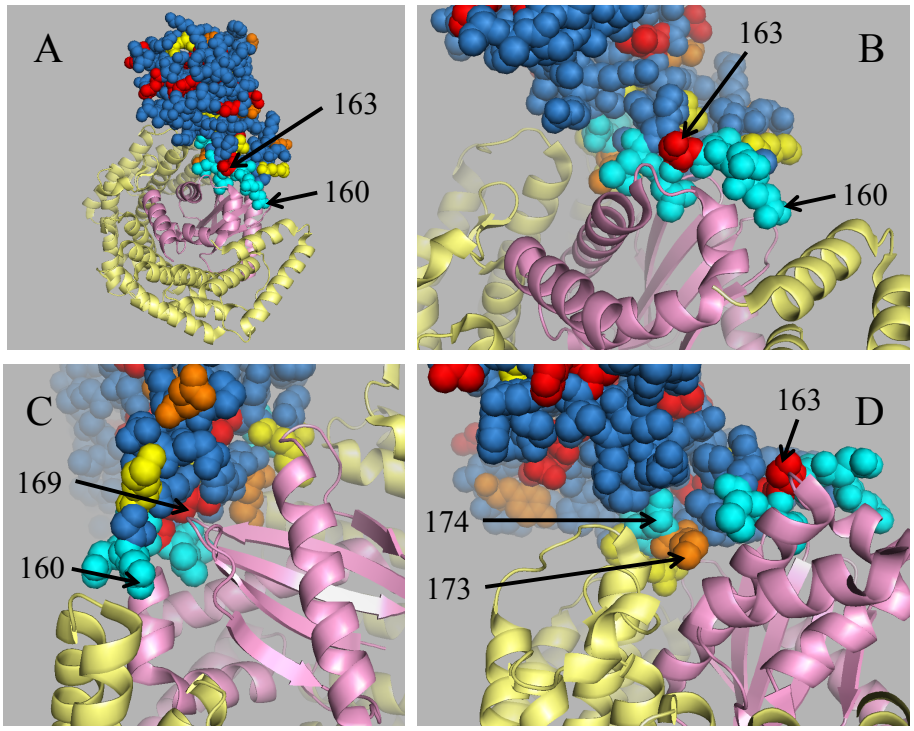

Figure 4.4: Sites identified with swMutSel and M8 mapped onto HIV-1 M Nef (approximately residues 56-203, blue spheres) in complex with subunits σ2 (pink cartoon) and α2 (pale yellow cartoon) of adaptor protein 2 (AP-2). Nef residues 160-165 and 174-175 required for AP-2 interaction are shown in cyan. Nef sites identified by swMutSel are coloured in red, M8 in yellow and sites identified by both are in orange. (A) A view of the whole complex; (B) Enlarged in view of (A); (C) Approximately 90° clockwise rotation of (B); (D) Approximately 90° anticlockwise rotation of (B). Crystal structure PDB ID 4NEE; Ren et al. (2014).

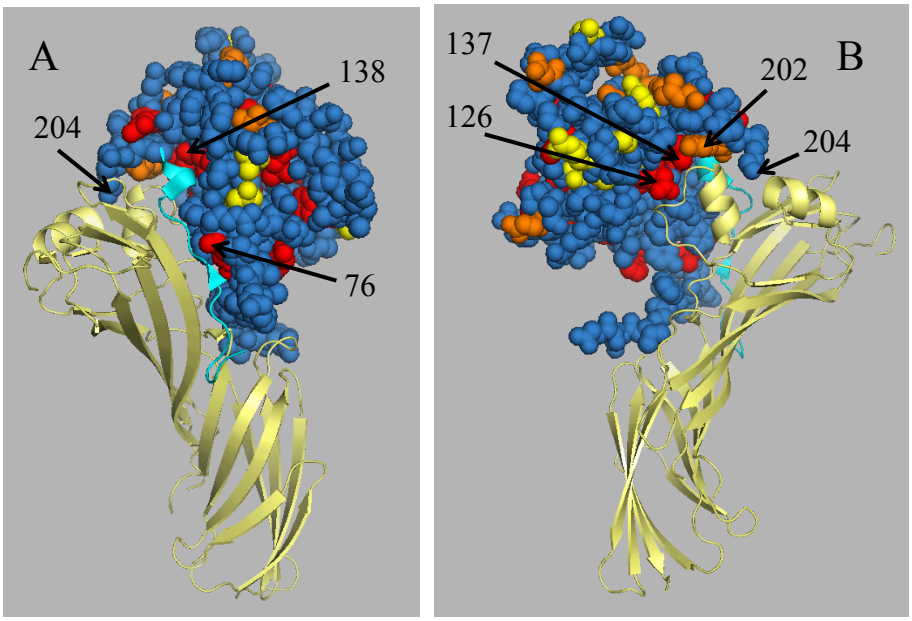

Figure 4.5: Sites identified with swMutSel (red), M8 (yellow) or both (orange) in Nef (spheres) bound to the cytoplasmic tail of MHC-1 (cyan cartoon, residues 314 to 332) and subunit μ1 of adaptor protein 1 (AP-1, pale yellow cartoon, residues 160 to 423). Nef sites identified with either method in interaction with either partner are indicated. (B) is a 180° rotation of (A); site 204 is indicated for reference between panels. PDB ID 4EN2; Ren et al. (2014).

found with an alternative topology had also been identified with the ML topology (table 4.3, top). Only 2 sites were found with the ML topology and not at least one alternative.

The alternative model M8 was found to be statistically supported over the null M7 model for each alternative topology ($p < 0.05$). There was greater range in the numbers of sites identified across topologies (table 4.3, bottom) and several sites were found with the ML topology but not found with one or more alternative; none of these were implicated in host factor binding.

4.3.2 Vpu

13 Sites Identified as Experiencing Divergent Selective Constraints

Following a similar protocol as for Nef, we analysed HIV-1 M and SIVcpz Vpu. This dataset comprised 1333 HIV-1 M and 21 SIVcpz *vpu* codon sequences from the Los Alamos HIV sequence database. Due to alignment uncertainty at the gene's 5' end and coding sequence overlap with the *env* gene at the 3' end, we confined our analysis to codon sites 5-53. From our original estimations of phylogeny by ML, groups expected to be monophyletic were found to be mixed throughout the

Model	Topology	T_i Total	ML $\setminus$ T_i	ML $\cap$ T_i	$T_i \setminus$ ML
swMutSel	T_1	40	4 *(48, 76, 138, 151)*	30	10 *(33, 49, 50, 51, 104, 105, 120, 135, 157, 177)*
”	T_2	37	5 *(48, 76, 138, 194, 198)*	29	8 *(32, 49, 89, 105, 120, 168, 176, 177)*
”	T_3	42	3 *(48, 138, 198)*	31	11 *(3, 15, 32, 39, 49, 55, 104, 108, 120, 135, 177)*
”	T_4	38	5 *(48, 76, 138, 151, 198)*	29	9 *(18, 33, 49, 51, 89, 102, 157, 170, 177)*
”	T_5	45	2 *(48, 138)*	32	13 *(3, 18, 33, 49, 98, 105, 107, 108, 120, 157, 168, 170, 177)*
codeml M8	T_1	11	12 *(24, 39, 120, 133, 135, 151, 158, 168, 173, 176, 188, 202)*	11	0
”	T_2	24	0	23	1 *(3)*
”	T_3	14	9 *(39, 120, 133, 135, 168, 173, 178, 188, 202)*	14	0
”	T_4	15	8 *(39, 135, 151, 173, 176, 178, 188, 202)*	15	0
”	T_5	24	1 *(173)*	22	2 *(3, 170)*

Table 4.3: Nef sites identified with alternative tree topologies, $T_{1\text{-}5}$, using swMutSel (multiple hypothesis adjusted $p < 0.05$, LRT with the degrees of freedom $N - 1$, where N is the number of residues observed at the site; top) or codeml's M8 (BEB probability > 0.95; bottom). Columns: [T_i Total], total number of sites found with topology T_i; [ML $\setminus$ T_i], number of sites found with ML tree but not T_i; [ML $\cap$ T_i], intersection; [$T_i \setminus$ ML], number of sites found with T_i but not ML topology. Where there are discrepancies in the sites found with an alternative topology compared with the ML topology, the Nef site indices are given in italics.

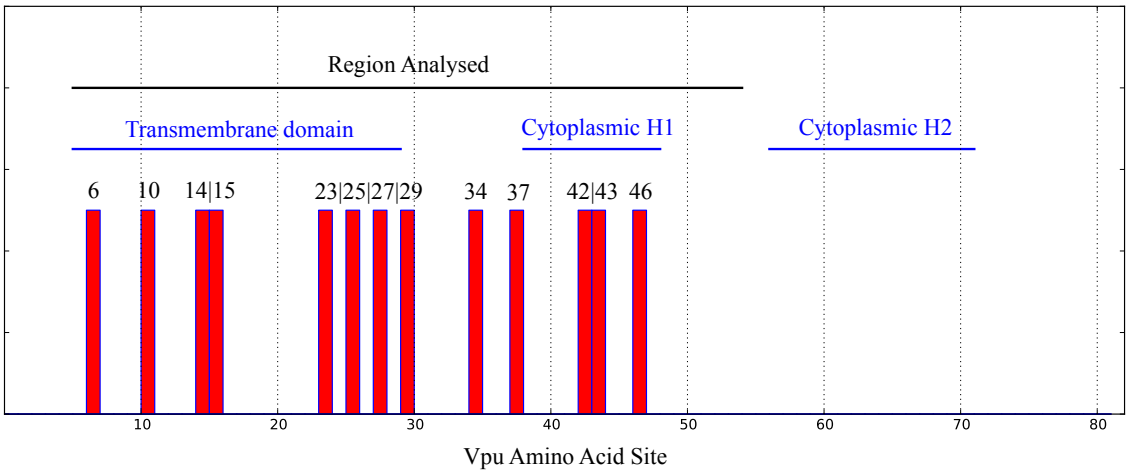

Figure 4.6: Sites identified in HIV-1 M and SIVcpz Vpu with swMutSel. The height of the bars is of no significance, and site indices are indicated (delimited by '|' if close together, for clarity). Secondary structure elements are indicated. Some sites were excluded from the analysis due to alignment uncertainty or overlapping coding sequences; the region analysed (sites 5-53) is indicated.

ML topology and with poor non-parametric bootstrap support (data not shown), probably due to so few sites being present in the alignment. As the topology itself is not of primary interest, we estimated the phylogeny by ML while imposing a constraint topology which required HIV-1 M subtypes and SIVcpz *Pts* each to be monophyletic.

Using swMutSel to compare evolutionary processes in HIV-1 M with SIVcpz Vpu, we identified 13 sites as having support for divergent selective constraints ($p < 0.05$ with LRT; table E.2). 7 of these were positioned in the transmembrane domain, which is involved in tetherin interaction by HIV-1 M Vpu. A further 3 in the first helix of the cytoplasmic domain were identified (fig. 4.6), which has similarly been implicated in tetherin antagonism. Interestingly, sites 14, 25 and 27, identified as experiencing different selective constraints, have previously been implicated in tetherin binding by HIV-1 M Vpu (Vigan and Neil, 2010; McNatt et al., 2013).

No data were available from codeml models M7 and M8 as M8 failed to converge after $>$ 2 weeks' running time. Models M1a and M2a were also used but convergence with M1a was not achieved after the same duration.

Protein Structure Context

We mapped these positions onto available Vpu nuclear magnetic resonance (NMR) structures. Sets of sites 10/14/15 and 23/25 each comprise a small cluster in the

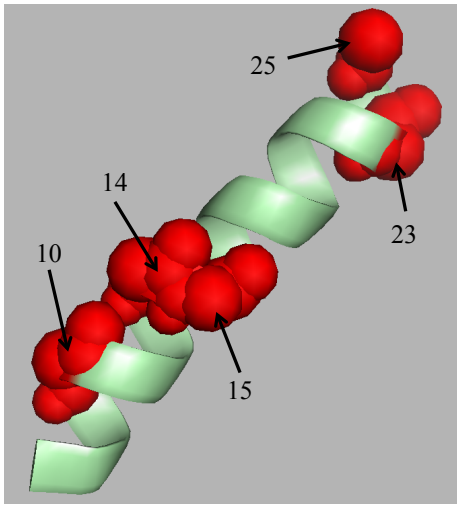

Figure 4.7: NMR structure of the Vpu transmembrane domain (including residues 7-25). Five sites identified with swMutSel have side chains shown as red spheres. PDB ID 2GOF; Park et al. (2003).

transmembrane domain (fig. 4.7). Side chains of residues at sites 42 and 46 in the first helix of the cytoplasmic domain, identified with swMutSel, interact with residues at sites 61, 65 and 68 in the second helix, which were not analysed due to gene overlap (fig. 4.8).

Vpu Alternative Tree Topologies

To investigate the effect of the topology used, we repeated the analyses with 5 alternative tree topologies, again applying the constraint that HIV-1 M subtypes and SIVcpz *Pts* each be monophyletic. The sites found with the ML topology were consistently identified when using most alternative topologies, the exceptions being sites 6 and 46 (in the transmembrane and cytoplasmic domains, respectively) which were not found with more than one alternative (table 4.4). As with the ML topology, model M8 failed to converge when using each alternative topology.

4.3.3 Vpr

The same procedure was followed for analysing HIV-1 M and SIVcpz Vpr as Vpu, involving 1333 and 21 HIV-1 M and SIVcpz sequences, respectively, from the Los Alamos HIV sequence database. Again it was necessary to use introduce a constraint tree to force subtypes and SIVcpz *Pts* each to be monophyletic. We excluded the 5' and 3' ends of the gene as they overlap with *vif* and the first exon of *tat* in the genome respectively (*tat* was also found to begin 6 nucleotides earlier in most

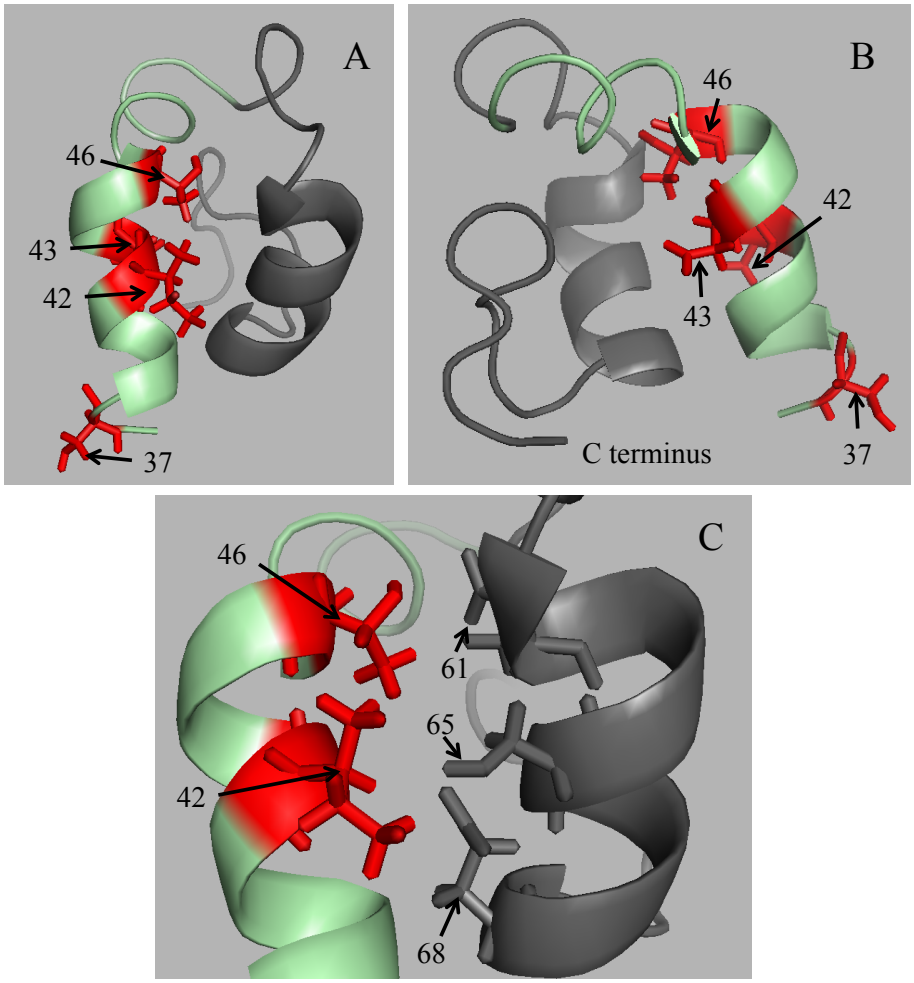

Figure 4.8: NMR structure of the Vpu cytoplasmic domain (including residues 36-81). Four sites in this region identified with swMutSel (red) and interacting partners in the second helix have side chains shown as sticks. Regions which could not be analysed due to overlapping coding sequences are coloured grey. (B) is a 180° rotation of (A); (C) an enlarged view of (A), with side chains in cytoplasmic domain helix 2 (not analysed) interacting with identified sites shown as sticks. PDB ID 2K7Y; (Wittlich et al., 2009).

Topology	T_i Total	ML $\setminus$ T_i	ML $\cap$ T_i	$T_i \setminus$ ML
T_1	26	0	13	13 *(7, 9, 11, 12, 16, 17, 22, 24, 26, 28, 31, 36, 41)*
T_2	17	5 *(6, 25, 42, 43, 46)*	8	9 *(7, 11, 12, 17, 24, 26, 31, 33, 40)*
T_3	24	0	13	11 *(7, 9, 12, 17, 20, 24, 26, 28, 31, 36, 40)*
T_4	24	1 *(46)*	12	12 *(7, 9, 11, 12, 16, 22, 26, 28, 36, 40, 41, 49)*
T_5	22	2 *(6, 46)*	11	11 *(7, 9, 16, 22, 24, 28, 30, 31, 36, 40, 41)*

Table 4.4: Numbers of Vpu sites identified with alternative tree topologies, $T_{1\text{-}5}$, using swMutSel (multiple hypothesis test adjusted $p < 0.05$ from LRT with the degrees of freedom $N - 1$, where N is the number of residues observed at the site). Columns: [T_i Total], total number of sites found with topology T_i; [ML$\setminus T_i$], number of sites found with ML tree but not T_i; [ML $\cap$ T_i], number of sites in intersection; [$T_i \setminus$ ML], number of sites found with T_i but not ML topology. Where there are discrepancies in the sites found with an alternative topology compared with the ML topology, the Vpu site indices are given in italics.

SIVcpz *Pts* isolates, and these sites were excluded in these sequence). We therefore analysed codon sites 22-84, of the total 96.

Using swMutSel, two sites were found to have statistical support for divergent constraints ($p < 0.05$ with LRT; table E.3). Surprisingly, these were sites 61 and 63 and positioned on helix 3 (fig. 4.9), which forms part of the interaction interface with DCAF1, which would be expected to be a conserved interaction (Berger et al., 2015). Site 61 is isoleucine in most HIV-1 M sequences, and engages in hydrophobic interactions with sites 23 and 26, which are each leucine, stabilising the three-helix bundle comprising the Vpr core (fig. 4.9B). As with Vpu, codeml models M1a and M8 failed to converge and therefore no comparison was available.

We repeated the analyses with 5 alternative topologies. More sites were identified with each alternative than the ML topology, and sites 61 and 63 were identified consistently (table 4.5). Again M8 failed to converge with each alternative topology.

4.4 Discussion

In this study, we have investigated evolutionary constraints acting on the lentiviral accessory genes *nef*, *vpu* and *vpr*. Numerous studies of primate lentiviruses have demonstrated species specific interactions involving viral accessory proteins and host restriction factors (e.g. Sauter et al., 2009; Lim et al., 2012; Laguette et al., 2012).

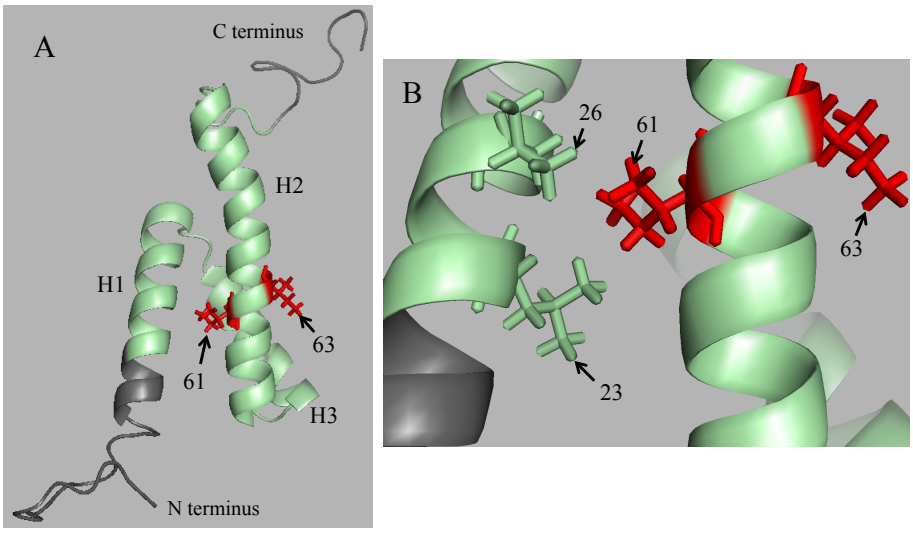

Figure 4.9: NMR structure of the Vpr, showing the two sites identified with swMutSel as red sticks. Regions which could not be analysed due to overlapping coding sequences are coloured grey. (A) Whole structure with helix designations ('H') indicated; (B) is an enlarged view of (A), with additional residue side chains shown as sticks. PDB ID 1M8L; (Morellet et al., 2003).

Topology	T_i Total	ML $\setminus T_i$	ML $\cap T_i$	$T_i \setminus$ ML
T_1	7	0	2	5 *(37, 41, 45, 60, 66)*
T_2	6	0	2	4 *(37, 58, 60, 66)*
T_3	5	0	2	3 *(37, 60, 84)*
T_4	4	0	2	2 *(37, 60)*
T_5	6	0	2	4 *(58, 60, 66, 77)*

Table 4.5: Vpr sites identified with alternative tree topologies, $T_{1\text{-}5}$, using swMutSel (multiple hypothesis test adjusted $p < 0.05$ from LRT with the degrees of freedom $N-1$, where N is the number of residues observed at the site). Columns: [T_i Total], total number of sites found with topology T_i; [ML $\setminus T_i$], number of sites found with ML tree but not T_i; [ML $\cap T_i$], number of sites in intersection; [$T_i \setminus$ ML], number of sites found with T_i but not ML topology. Where there are discrepancies in the sites found with an alternative topology compared with the ML topology, the Vpr site indices are given in italics.

Restriction factors are often found to be experiencing diversifying selection to evade recognition by accessory proteins. To investigate evolution of the accessory proteins themselves, we have asked whether we can identify divergent evolutionary processes in these genes in the pandemic virus HIV-1 M and its animal reservoir SIVcpz, to characterise the sites involved in the host-specific adaptation.

The novelty of this approach is to use swMutSel, a selection analysis tool specifically designed for studying divergent selective constraints. Positive selection in lentiviral accessory proteins has been investigated previously, by Soares et al. (2008) using models employing d_N/d_S parameters (including codeml models), though none of the sites identified in their study were identified in our analyses with swMutSel or M8 for any gene. These authors used much smaller sequence datasets (maximum of 43 sequences) with different proportions of HIV-1 M subtypes, perhaps explaining the discrepancies with our M8 results. We would not necessarily expect their data to concur with our observations with swMutSel since the respective modelling frameworks are significantly different.

Nef

In *nef*, we identified 34 sites with statistical support for divergent selective constraints (table 4.2; fig. 4.2). As expected, sites likely to be directly involved in CD4 binding are not identified, consistent with this function being conserved in the two viruses (fig. 4.3; Sauter et al., 2009). Unexpectedly, however, we identify *nef* sites 163, 169 and 173 which are involved in the AP-2 complex interaction which mediates CD4 downregulation (fig. 4.4). The mechanism by which SIVcpz Nef downregulates CD4 has not been investigated experimentally and it could be expected to closely match that of HIV-1 M Nef. However, these data suggest the putative interaction between SIVcpz Nef and chimpanzee AP-2 subunits is not dependent on the same side chain interactions as for HIV-1 M Nef. In this case, the sites we have identified may have undergone species specific directional selection following transmission to the human host, in order to maintain CD4 downregulation via interaction with AP-2.

Whether SIVcpz Nef downregulates MHC-1 surface expression has not been determined, though this activity has been demonstrated for HIV-1 groups M, O and SIVgor (infecting gorillas; Schwartz et al., 1996; Kluge et al., 2014), suggesting their

common precursor SIVcpz possesses the same phenotype. It is therefore surprising that with swMutSel we identify sites 76 and 138 which form part of the interface with MHC-1 itself (fig. 4.5A). As with CD4 downregulation, other identified sites interact with the recruited complex which mediates MHC-1 downregulation, namely subunit μ1 of AP-1 (fig. 4.5B). Together these data suggest the exact interaction between MHC-1 and AP-1 is not conserved in HIV-1 M and SIVcpz Nef proteins, and adaptation at the identified sites may have accompanied the spread to humans. However, we note in both MHC-1 and μ1, the interaction interfaces involve many more Nef sites than we identify, and it is not certain to what extent the identified sites contribute to the binding affinity.

While HIV-1 M Vpu is responsible for antagonising human tetherin, SIVcpz uses its Nef protein to counter the chimpanzee orthologue (Sauter et al., 2009). HIV-1 M Nef has no such activity and therefore the sites evolving under different constraints may be involved in tetherin antagonism by SIVcpz. The molecular details of the interaction have not been investigated and therefore the sites identified here may serve to inform experimental inquiry into this function.

We also identified 23 sites with M8, suggesting pervasive diversifying selection in *nef*. HIV-1 M Nef is known to be highly immunogenic, containing several epitopes used in antigen presentation for cellular immunity (Llano et al., 2013). Diversifying selection has previously been identified in HIV-1 M and also SIVmac (infecting rhesus macaques; Price et al., 1997; Evans et al., 1999; Cavalieri et al., 2009).

Vpu

Our analysis of *vpu* was motivated by experimental observations showing that while SIVcpz Vpu does not antagonise chimpanzee tetherin, HIV-1 M Vpu has apparently acquired this function on adapting to the human host (Sauter et al., 2009). In the Vpu transmembrane domain (TMD), which mediates interactions with human tetherin (Iwabu et al., 2009), we identified 7 sites undergoing divergent selective constraints with swMutSel (fig. 4.6; fig. 4.7). In mutagenesis experiments, Vigan and Neil (2010) identified TMD sites 14, 18 and 22 as required for tetherin engagement. They noted that both sites 14 and 18 are conserved in HIV-1 M Vpu, which antagonises tetherin, but not in SIVcpz or HIV-1 O Vpu, which do not, suggesting a change of selective constraints at these sites which allowed for tetherin antagonism following

establishment of HIV-1 M in humans. Consistent with this, we have identified site 14 with swMutSel. McNatt et al. (2013) went on to implicate experimentally TMD sites 4, 7, 20, 21, 25, 26 and 27 in the tetherin interaction. Of these, we have found sites 25 and 27 with swMutSel, again consistent with a gain of function inducing a change in selective constraints. Moreover, the greatest density of identified sites is in same region, between sites 20-29 (fig. 4.6).

We also identified 5 sites in and around the first helix of the Vpu cytoplasmic domain (fig. 4.8) which like the TMD has been implicated in tetherin antagonism by HIV-1 M Vpu (McNatt et al., 2013), though the exact sites have not been mapped.

Together, our analysis of HIV-1 M and SIVcpz Vpu has identified several sites already known to be involved in tetherin antagonism. We hypothesise that the remaining identified sites are similarly under different selective constraints due to the gain of anti-tetherin function by HIV-1 M Vpu. This list will be of interest to experimentalists using mutagenesis techniques to probe the mechanism of Vpu-tetherin interaction.

However, while we have emphasised the acquisition of tetherin antagonism as a putative driver for the divergent selective constraints in Vpu, we note that HIV-1 M and SIVcpz Vpu also downregulates CD4 (Willey et al., 1992; Sauter et al., 2009) and has other activities which are poorly defined (Strebel, 2014). We therefore cannot rule out that the identified sites are involved in divergent Vpu activities in the two viruses which are as yet unrecognised.

Vpr

Two sites are identified with swMutSel in Vpr. Strikingly, these are within the region in helix 3 (sites 60-68) to which HIV-1 M Vpr is known to interact with DCAF1 and recruit the E3 ubiquitin ligase complex (Zhao et al., 1994). While SIVcpz Vprs have not been tested, DCAF1 interaction has been observed in divergent virus lineages and suggests this is an ancestral trait common to most primate lentiviruses (Berger et al., 2015). Furthermore, human and chimpanzee DCAF1 are highly conserved (data not shown) and therefore it is surprising sites in this interaction domain would be identified. It may be that the nature of the DCAF1 interaction differs between the viruses, just as Vpx proteins from different SIVs bind the restriction factor SAMHD1 at different interfaces (Fregoso et al., 2013). However, they may be expe-

riencing divergent selective constraints due to host-specific interactions connected with another, possibly undiscovered, Vpr function.

It would be interesting to perform similar comparative analyses with lentivirus groups known to possess divergent Vpr functions. Berger et al. (2015) have found species specific interaction between Vpr proteins and the SLX4 complex (SLX4com), with the result that some SIVs are incapable of arresting the cell cycle in human cell lines, while maintaining this function in African green monkey cells. Testing for different selective constraints between those that do and do not bind SLX4com in human cells could therefore identify the regions of Vpr responsible for SLX4com engagement. However, we note that highly divergent Vprs have high numbers of insertions and deletions, making confident alignment very challenging; and only small numbers of sequences from SIVs infecting monkeys are publicly available, potentially limiting statistical power.

General Discussion

HIV-1 M is substantially better represented in our datasets than SIVcpz, with over 1000 sequences compared with 21. (SIVcpz isolates are rare because chimpanzees are endangered and samples must be collected in field expeditions to remote parts of Africa; e.g. Keele et al., 2006.) If one virus group is insufficiently defined due to lack of data, we would expect the alternative model (representing the hypothesis of non-homogeneous selective constraints) to fit the data little better than the null model (representing homogeneous constraints), in which case the null hypothesis cannot be rejected. Therefore our identification of sites is not an artifact of sampling bias, provided the available SIVcpz sequences are representative. The small number of SIVcpz sequences provides sufficient statistical power due to the high amount of sequence divergence (fig. 4.1).

In our analyses of both *vpu* and *vpr*, we found our initial phylogeny estimations by ML did not produce topologies where HIV-1 M subtypes were monophyletic, which suggested there was insufficient phylogenetic signal in these short alignments (147 and 189 nucleotides long, respectively). We therefore resorted to constraining the topologies so subtypes were monophyletic. The impact of inaccurate topologies on swMutSel results has not been determined, but we sought to mitigate the danger of inaccurate topologies introducing false positives by repeating the analysis with

several alternative topologies and found all (Vpr, table 4.5) or all but one (Vpu, table 4.4) of the sites were identified consistently, suggesting the results were not dependent on the topology used. As the sequences used were originally excised from whole virus genome sequences, in future work topologies could be estimated by expanding the alignment used to include nucleotide sites flanking the genes of interest. Selection analyses would still be confined to non-overlapping parts of the genes themselves, but the resulting topology itself may be more reliable.

We were also unable to compare our *vpu* and *vpr* swMutSel results with data from codeml models M7 and M8, as M8 consistently failed to converge with each topology used. In future work this could be addressed by repeating the model fittings with a range of starting parameter values, to allow the numerical optimisation routine to better explore the parameter space.

In conclusion, we have identified sites in *nef*, *vpu* and *vpr* experiencing divergent selective constraints in HIV-1 M and SIVcpz. In *vpu*, we identify several sites known to contribute to HIV-1 Vpu anti-tetherin function and several more which we hypothesise to be similarly involved. Surprisingly, in *nef* and *vpu* we identify sites in regions which bind host factors, in interactions expected to be conserved in these viruses. This suggests the mechanisms of action differ in host dependent ways and therefore human specific adaptation may have occurred at the identified sites during the establishment of pandemic HIV-1 M.

4.5 Methods

A dataset of 1333 HIV-1 and 21 SIVcpz full length sequences was obtained from the Los Alamos HIV Sequence Database, the 2012 curated dataset (`http://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html`). Known HIV-1 M inter-subtype recombinants and SIVcpz sequences known to be derived from the same molecular clone had been excluded. *nef*, *vpu* and *vpr* nucleotide sequences were excised and aligned manually (using the alignment editor SEAVIEW version 4.4.0; Gouy et al., 2010) as protein sequences and mapped back to produce codon alignments. As *nef* is positioned at the 3' end of the primate lentivirus genome, genome sequencing had had apparently terminated before the end of the coding sequence in many cases: 117 sequences had only 120 codon sites at the 5' end, and of these 84 had fewer than 36 codon sites at the 5' end. Several *nef* sites were omitted due to alignment uncer-

tainty (corresponding to sites 8-11, 22-25, 61-65, 124 and 204-206, inclusive) and we did not analyse sites 56 and 134 as they were universally conserved throughout the dataset; 188 sites were analysed in total. Sites involved in coding sequence overlap in *vpu* and *vpr* were also excluded.

Phylogenies were estimated by maximum likelihood, using RAxML 7.7.2 HPC-HYBRID-AVX (Stamatakis, 2006; Ott et al., 2007), with GTR model of nucleotide substitution and Gamma-distributed rates. Three HIV-1 M *nef* sequences were positioned in clades other than their classified subtype; since these may represent recombinant sequences they were excluded by pruning from the tree (sequence accessions: JF683759, JQ403079 and JF683760). RAxML permits a 'constraint tree' to be specified, which allows the search space to be confined to topologies conforming to the groupings in the constraint tree. For the *vpu* and *vpr* trees we constrained each HIV-1 M subtype and SIVcpz *Pts* to be monophyletic. No constraint was used for *nef*. Alternative tree topologies were obtained and tree figures produced as described in §3.5.1.

Selection analyses were performed as described in §3.5.2.

Chapter 5

Detecting Positive Selection in Single-Strand Overlapping Coding Sequences

5.1 Summary

In a wide range of organisms, multiple protein-coding sequences can be found overlapping at a single genetic locus, occupying different reading frames. Conventional selection analysis tools will produce false positives if applied in this situation, because they cannot account for selection acting on multiple frames. Indeed, several published studies have wrongly inferred positive selection in overlapping coding sequences for this reason. Previous attempts to produce adequate models for the purpose have been either heuristic or too computationally expensive for routine use. We have developed a new likelihood approach which infers selection pressure on overlapping coding sequences on a singe DNA strand using novel 'codon aware' nucleotide substitution models. We present four such models and have tested their ability to estimate parameters accurately with synthetic data. Only one of these was accurate under all conditions tested and had low false positive and false negative rates for identifying positive selection. From this we developed a mixture model to identify positive selection at a minority of codon sites. The mixture model was found to have a high false positive rate, and we suggest modifications to the underlying substitution model which may improve the specificity.

5.2 Introduction

In protein-coding nucleic acid sequences, amino acid sequences are represented by sets of contiguous codons, which, if they contain no stop codons, forms an open reading frame and can be translated into a polypeptide. If translation is initiated one or two nucleotides further along, the reading frame, or phase, changes, and a different set of codons can be obtained from the same nucleotide sequence. For example, the codons ATG, GGA and CCG if read directly translate to the amino acids methionine, glycine and proline. Starting translation from the second nucleotide, however, yields the complete codons TGG and GAC, which translate to cysteine and aspartate residues. As codons comprise triplets of nucleotides, there are three reading frames present in a single sequence, and a further 3 frames when considering the opposite strand in double stranded nucleic acids. Some organisms exploit these additional frames and allow multiple, overlapping, coding sequences to be present in the same nucleotide sequence.

Overlapping coding sequences were originally identified in the first DNA genome to be sequenced, the bacteriophage ΦX174 (Barrell et al., 1976), but have since been found in all domains of life. In bacteria, overlapping coding sequences are found across many diverse species and reportedly as many as a third of bacterial genes are engaged in overlap (Johnson and Chisholm, 2004). Most of these are short (85% are < 30 bp) and on the same strand, perhaps because close union of genes allows their transcription to be tightly coupled and produces consistent stoichiometries of gene products. Greater proportions of overlapping coding sequences are observed among bacterial species with very small genomes, suggesting genome compression may be a cause in some cases (Sakharkar et al., 2005).

Because eukaryotic genomes are large and apparently nonfunctional regions are abundant, it had been implicitly assumed overlapping coding sequences do not occur in these organisms (Mouilleron et al., 2016). But more recent surveys of diverse vertebrate genomes have identified instances of overlapping coding sequences from humans to zebra fish (Makalowska et al., 2005). A study of 34,604 human genes identified 57 overlapping coding sequences and a similar proportion was found among mouse genes (Veeramachaneni et al., 2004), while other work has found many overlaps are homologous in the two species, suggesting they persist through long periods of time (Sanna et al., 2008). (Still higher numbers of overlapping *genes* have been

observed, where exons of one gene are interspersed with those of another at the same locus, without overlap of coding exons themselves; these are not the interest of the present work.) The well studied examples of overlapping coding sequences in humans (e.g. Klemke et al., 2001; Nekrutenko et al., 2005; Bergeron et al., 2013) have found the encoded proteins interact with each other, suggesting the overlap confers the benefit of consistent co-expression for fulfilling the same function.

Coding overlap is still more common in viruses, with 75% of around 2000 discovered virus species found to exhibit at least some overlap (Chirico et al., 2010), including opposite strand overlap (e.g. Epstein-Barr virus) and even cases where all three reading frames on one strand are protein coding at a single locus (e.g. lentiviruses such as HIV-1). Explanations for the preponderance of overlap in viruses include: the need for compactness in relatively short genomes (both for genome packaging and rapid replication); controlled co-expression as proposed for bacteria; and even as a means to produce genetic innovation, as new genes are unlikely to be created by large genomic rearrangements (Brandes and Linial, 2006).

Techniques for identifying selection in canonical coding sequences (e.g. Nei and Gojobori, 1986; Goldman and Yang, 1994) have been widely applied in studies of protein coding sequences. The presence of more than one coding sequence in a single nucleotide sequence, however, gives rise to unusual evolutionary pressures which make these methods unsuitable. Synonymous codon substitutions are usually assumed to have a neutral effect on fitness, but a synonymous change in one reading frame is likely to be nonsynonymous in another. Because most approaches assume codon sites evolve independently, the interdependence of multiple coding frames cannot be accounted for. Following an example we presented in Monit et al. (2015), consider two overlapping coding sequences A and B as in figure 5.1 and a selection analysis of gene A. The first nucleotide position in each gene B codon corresponds to the third nucleotide position in gene A codons. We see from the universal genetic code that substitutions in the first nucleotide position of a codon are likely to be nonsynonymous in that frame, meaning purifying selection acting on gene B would reduce substitution rates at these sites. This corresponds to a decreased substitution rate at the third nucleotide position in gene A codons, which are likely to be synonymous in that frame, therefore inflating the d_N/d_S ratio for gene A and resulting in the mistaken impression of positive selection. Failing to account for selective constraints acting on alternative frames therefore renders conventional se-

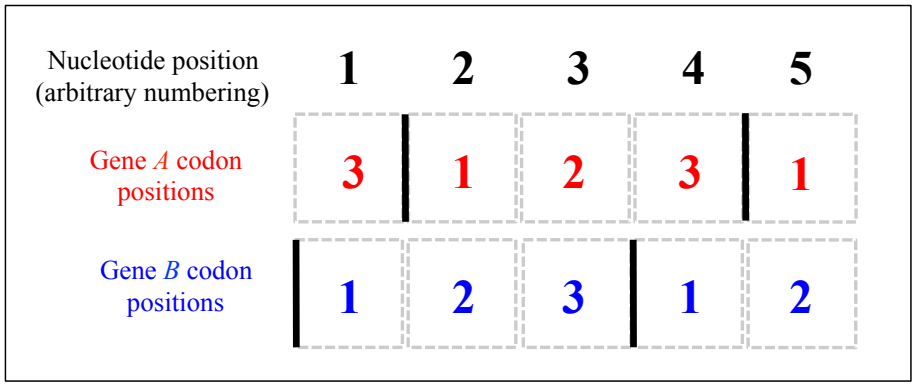

Figure 5.1: Overlapping codons of two hypothetical genes, A and B. Boxes represent nucleotide positions and solid black lines show the boundaries between codons. Numbers within boxes represent nucleotide positions within each codon triplet. It can be seen that the nucleotide occupying the first position in gene B codons is the same as the nucleotide at the third codon position in gene A. Adapted from Monit et al. (2015), with permission.

lection analysis tools invalid. Many published studies (e.g. Obenauer et al., 2006; Snoeck et al., 2011; Roy et al., 2015) have reported finding positive selection selection when studying overlapping coding sequences, and it is highly likely that these are false positives resulting from this model misspecification.

Several strategies for modelling selection in overlapping reading frames have been attempted previously. The first (Hein and Stovlbaek, 1995) generalised to overlapping coding sequences the method of Li et al. (1985), a heuristic counting method for estimating selection pressure. This relies on assigning nucleotide sites within codons to degeneracy classes based on whether substitutions introduce nonsynonymous codon changes. A single sequence in the dataset is used to determine the classes and it is assumed they do not change through time. While the method offers the advantages of speed and simplicity, it is approximate only. Similarly, Wei and Zhang (2015) extended the method developed by Nei and Gojobori (1986) to accommodate overlapping coding sequences, but as this is also a heuristic counting method.

Pedersen and Jensen (2001) tackled directly the fundamental problem of being unable to assume independence between sites by developing a probabilistic model in which substitutions between whole sequences are considered. The substitution process is treated as a reversible, homogeneous Markov process with parameters common to codon substitution models such as the transition/transversion rate ratio (κ) and a d_N/d_S ratio for each gene. However, the size of the state-space forces them to use a Monte Carlo simulation to compute transition probabilities. The procedure

is computationally expensive and is therefore impractical for analysing more than two sequences.

Sabath et al. (2008) noted the disadvantages of the approximate and whole-sequence approaches and proposed a model in which the unit of evolution was a sextet, comprising two adjacent codons in one frame and an overlapping codon in another frame. They construct a sextet substitution matrix for each gene, including d_N/d_S parameters for both genes, from which they derive codon transition probability matrices by amalgamating substitution rates for sextets sharing the same central codon. Both the probability matrices are used to compute the probability of observing the sequence data, and these values are then multiplied to give the total likelihood of the model. The approach therefore treats evolution of the two overlapping genes as independent events, despite their being in the same sequence. Motivated to identify selection on coding overlaps for genome-wide screens rather than selection analysis, Chung et al. (2007) developed a codon model dealing with uncertainty about flanking states affecting an overlapping gene by marginalising; but like Sabath et al. (2008) substitutions in the two genes are ultimately treated as independent processes.

The presence of overlapping coding sequences in diverse cellular organisms and their prevalence in clinically relevant viruses makes understanding their evolution a matter of importance. The following qualities would be desirable from an approach to studying selection in overlapping coding sequences: (1) adequate separation of effects on the different genes, so the selective constraints acting on each can be identified; (2) ability to test hypotheses about the nature of the selective constraints; (3) computationally tractable, so datasets of more than a handful of sequences can be analysed; and (4) preferably modelling substitutions as a Markov process, to take advantage of the well developed mathematical framework used in phylogenetics (Felsenstein, 1981). None of the selection analysis methods developed to date satisfy all of these and so, as yet, there is no appropriate tool suitable for studying these genes routinely.

In this work we present a new approach to modelling selection in single-strand overlapping coding sequences. We have developed four nucleotide substitution models which account for protein-level selection. In tests with synthetic data, only one of these (the pentamer model) estimated parameter values accurately; summarised in table 5.1. To identify varied selection pressure across overlapping genes, the pen-

Sub. Model	Theory	Results	Outcome
Genetic code weighting model	§5.3.2.1; p120	§5.4.1.1; p132	Overestimated w in some cases and, when used to detect positive selection, had high false positive rate
Codon weighting model	§5.3.2.2; p122	§5.4.1.3; p135	Overestimated w in some cases; suggests high false positive rate
Frame Independence Model	§5.3.2.3; p123	§5.4.1.4; p138	Overestimated w in some cases; suggests high false positive rate
Pentamer model	§5.3.2.4; p125	§5.4.1.5; p138	Accurately estimated w and had low false positive rate. Used this to develop mixture model

Table 5.1: Summary of the four substitution ('sub.') models developed and the outcome of tests with synthetic data generated with uniform $\omega = d_N/dS$. Sections (§) and page numbers (p) describing them are indicated.

tamer model was used to implement a mixture model inspired by a commonly used codon mixture model (Nielsen and Yang, 1998). We found the mixture model had an unacceptably high false positive rate, using both the likelihood ratio test and parametric bootstrapping. We suggest modifications to the underlying model which may improve the specificity.

5.3 Theory

5.3.1 Premise

5.3.1.1 Aims

We seek to develop substitution models applicable to overlapping coding sequences which account for protein-level selection using the conventional $d_N/d_S = \omega$ ratio (Goldman and Yang, 1994). Computing transition probability matrices, the likelihood of a model can then be computed with the traditional approach (Felsenstein, 1981). That framework in place, we can test the hypothesis of positive selection being active in one of the overlapping genes using the likelihood ratio test (LRT).

As discussed above, codon-based Markov models are not applicable to overlapping coding sequences because sites are assumed to evolve independently, while codons in alternative reading frames span boundaries between codons in the frame considered. We aim to salvage the assumption of site independence by modelling codon-level selection at the nucleotide level and including parameters representing selection on amino acid changes. To bridge the gap between nucleotide and codon

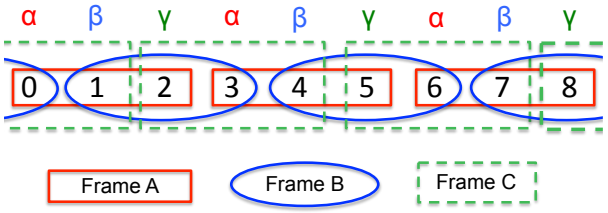

Figure 5.2: Overlapping reading frames and site types.

substitutions, we weight these selection parameters in accordance with our uncertainty about whether the given nucleotide change causes a nonsynonymous codon substitution, which is dependent on the site's (unknown) sequence context.

We assume that overlapping coding sequences are evolving independently of each other and that the substitution process is the same in overlapping and non-overlapping parts of a gene.

5.3.1.2 Genetic Layout

A nucleic acid strand contains three overlapping reading frames (fig. 5.2), which we name A, B and C. Nucleotide substitutions at the three positions within a codon do not have equal chances of resulting in nonsynonymous change in codon: in the universal genetic code, we see most (95.5%) changes at the first position, all changes at the second position and few (28.4%) at the third position will result in nonsynonymous codon change. Any nucleotide site is simultaneously a first, second and third codon position depending on the frame. This gives rise to three types of site (hereafter *site types*), which we name α, β and γ. α sites are defined as codon position 1 in frame A, β sites are codon position 1 in frame B and γ sites are codon position 1 in frame C (fig. 5.2).

Furthermore, if we consider a stretch of the genome containing several genes, we will have *zones* (or alignment partitions; sets of contiguous sites) in which the gene in each frame is consistent. Moving from one end of the sequence to another, a new zone begins with each beginning or end of a gene in any frame. The HIV-1 genome, for example, can be divided into 18 zones (fig. 5.3).

We model nucleotide substitutions at each site type in each partition as distinct Markov processes with their own rate ($\boldsymbol{Q}$) matrices (in total $3N_z$ matrices, if N_z is the number of zones), and parameter values are estimated by maximum likelihood (ML). We call these models codon-aware nucleotide (CAN) substitution models.

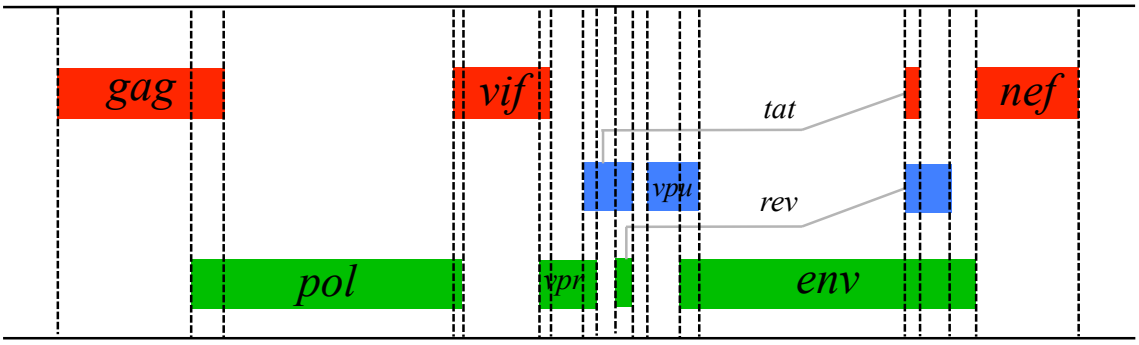

Figure 5.3: Coding zones (partitions) in the HIV-1 genome. Protein-coding open reading frames A, B and C shown top, middle and bottom, respectively (coloured as in fig. 5.2). Genes *tat* and *rev* are both made up of two exons. Gene lengths are not to scale.

5.3.2 Substitution Models

5.3.2.1 Genetic Code Weighting Model

Our simplest model extends the HKY85 nucleotide substitution model (Hasegawa et al., 1985) to include protein-level selection in each frame.

Let p_n be the probability of a nucleotide substitution within a codon causing a nonsynonymous codon substitution, while the probability of being synonymous $p_s = 1 - p_n$. In the first instance, p_n can be derived from the fraction of nucleotide substitutions at the given codon position (first, second or third) which cause non-synonymous codon changes, with reference to the genetic code. That is, over all combinations of codons we count the number of codon pairs which are nonsynonymous at the codon position (C_n) and the number which are synonymous (C_s). For the first, second and third codon positions,

$$C_n^1 = \sum_x^4 \sum_y^4 \sum_z^4 \sum_{x' \neq x}^3 g(xyz, x'yz) \tag{5.1}$$

$$C_n^2 = \sum_x^4 \sum_y^4 \sum_z^4 \sum_{y' \neq y}^3 g(xyz, xy'z)) \tag{5.2}$$

$$C_n^3 = \sum_x^4 \sum_y^4 \sum_z^4 \sum_{z' \neq z}^3 g(xyz, xyz') \tag{5.3}$$

where x, y and z denote nucleotide states making up codons, and $g(xyz, x'y'z')$ returns 1 if the codons are both nonsynonymous and neither are stop codons, and

0 otherwise. C_s is defined similarly, but counting synonymous differences. Then,

$$p_n = \frac{C_n}{C_n + C_s}. \tag{5.4}$$

From the universal genetic code, p_n is 0.954, 1.0 and 0.284 for the three codon positions, respectively. (Alternative values could be computed for other genetic codes, such as mitochondrial.)

Which value of p_n applies to a nucleotide substitution then depends on the site type and the frame of interest. For example, the probability of a nonsynonymous nucleotide substitution in frame A at an α site is 0.954, as α sites are defined as codon position 1 in frame A; if considering frame C at a β site, p_n is 0.284, as this is the third codon position (fig. 5.2).

We then define a weighting function f, specific to frame F and site type s:

$$f_{F,s}(\omega_G) = p_n\omega_G + p_s \tag{5.5}$$

where ω_G is the d_N/d_S ratio for gene G which resides in frame F, and the value of p_n is determined by F and s. Since $\omega = 1$ is appropriate for a synonymous substitution, the function returns ω_G value weighted by the probability of the substitution type, summed over the probabilities that the change is nonsynonymous or synonymous.

The HKY85 substitution matrix is defined

$$q_{ij}^{\mathrm{HKY}} = \nu\pi_j K(i,j) \tag{5.6}$$

and

$$q_{ii}^{\mathrm{HKY}} = -\sum_{i\neq j} \nu\pi_j K(i,j) \tag{5.7}$$

where $K(i,j)$ returns the transition/transversion rate ratio κ if i and j are both purines or pyrimidines, but 1 otherwise. ν is a scaling factor which ensures the mean substitution rate across sites is equal to 1:

$$\nu = \frac{1}{-\sum_i \pi_i q_{ii}^{\mathrm{HKY}}}. \tag{5.8}$$

We apply the weighting function accounting for selection as a scalar multiplication:

$$q_{ij,s} = \nu\pi_j K(i,j) f_F^s(\omega_G) c \tag{5.9}$$

and

$$q_{ii,s} = -\sum_{i \neq j} \nu\pi_j K(i,j) f_F^s(\omega_G) c. \tag{5.10}$$

As the selection term will alter the overall substitution rate, we include a free parameter c shared among all sites which allows the matrices to be scaled such that the overall substitution rate matches the branch length units.

Extending the model to 3-fold gene overlap, we define an ω for each gene (named 1, 2 and 3 below, occupying frames A, B and C respectively) present at this site (which is of site type s) and introduce a weighting function for each frame.

$$q_{ij,s} = \nu\pi_j K(i,j) f_{A,s}(\omega_1) f_{B,s}(\omega_2) f_{C,s}(\omega_3) c. \tag{5.11}$$

This yields a substitution matrix $\boldsymbol{Q}$ for every site type s within each zone. If a frame in a given zone is noncoding, ω for that frame is fixed to 1. Corresponding transition probability matrices and site log-likelihoods are computed as with standard phylogenetic models (Felsenstein, 1981), and log-likelihoods are summed across sites to give the total log-likelihood. We use branch lengths estimated separately with the HKY85 substitution model (Methods §5.6.1). In all, the model has $N_G + 5$ free parameters: κ, π_T, π_C, π_A (with $\pi_G = 1 - \pi_T - \pi_C - \pi_A$), c and one ω for each of N_G genes.

5.3.2.2 Codon Weighting Model

The codon table weighting model assumes all sense codons are equally probable. It is well known, however, that codon probabilities are uneven in most organisms, for example in response to unequal concentrations of synonymous tRNA (reviewed in Plotkin and Kudla, 2011). We therefore modify the above to allow for codon biases. If we have a normalised codon probability distribution, where a codon xyz has probability (frequency) P_{xyz}, then we can extend equation 5.1 above:

$$C_n^1 = \sum_x^4 \sum_y^4 \sum_z^4 \sum_{x' \neq x}^3 g(xyz, x'yz) P_{xyz} \tag{5.12}$$

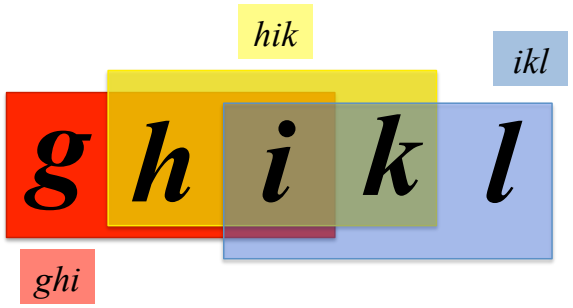

Figure 5.4: Pentamer containing three overlapping codons. As we model substitutions at the central nucleotide site only (occupied by state i and then j following a substitution), we are ignorant of nucleotide states g, h, k and l. Note that the central site occupies a different codon position (first, second or third) depending on the codon's frame: for example, if this were an α site, codon ikl is in frame A, ghi in frame B and hik in frame C (see fig. 5.2).

with C_n^2, C_n^3 (equation 5.2 and 5.3) and the synonymous counts similarly modified. That is, we weight each putative codon transition by the probability of the starting codon. p_n is then computed as in equation 5.4. We assume codon probabilities in non-coding frames are all 1/64.

Codon probabilities for many organisms have been estimated and made available in the Codon Usage Database (http://www.kazusa.or.jp/codon/; Nakamura et al., 2000), which estimates frequencies from sequence data from the 2007 NCBI-GenBank release. We use human and HIV-1 codon probability estimates from this database, as well as a uniform distribution (each sense codon having probability 1/61).

5.3.2.3 Frame Independence Model

A more complex model deals with the uncertainty about sequence context explicitly. We imagine the nucleotide site of interest resides at the centre of a pentamer $ghikl$, which contains one complete codon in each reading frame (fig. 5.4). $i \rightarrow j$ substitution at the central site will cause each codon in the pentamer to change. The instantaneous substitution rate at the central nucleotide site is assumed to be dependent on the codon probabilities of the original and target codons in each frame, as well as selection acting on each gene present. As we are considering substitutions at nucleotide sites independently, we remain ignorant of the nucleotide states at the neighbouring two sites on either side. We therefore marginalise over nucleotide triplets (i.e. codon probabilities), with a term for each codon comprising

the pentamer. For an α site,

$$q_{ij,\alpha} = \nu K(i,j) \times \frac{\sum_k^4 \sum_l^4 \Omega_A(ikl,jkl) P_{ikl} P_{jkl}}{\sum_{k'}^4 \sum_{l'}^4 P_{ik'l'}} \times \frac{\sum_g^4 \sum_h^4 \Omega_B(ghi,ghj) P_{ghi} P_{ghj}}{\sum_{g'}^4 \sum_{h'}^4 P_{g'h'i}} \times \frac{\sum_h^4 \sum_k^4 \Omega_C(hik,hjk) P_{hik} P_{hjk}}{\sum_{h'}^4 \sum_{k'}^4 P_{h'ik'}} \tag{5.13}$$

where

$$\Omega_F(xyz, x'y'z') = \begin{cases} \omega_F, & \text{if codons } xyz \text{ and } x'y'z' \text{ are nonsynonymous} \\ 1, & \text{otherwise} \end{cases} \tag{5.14}$$

and F being the frame in question and ω_F the d_N/d_S ratio for the gene present in frame F at this site. Indices g-l represent nucleotide states. Substitution to or from a stop codon is not counted in the summations since we assume the codon probability P for each stop is 0. β and γ sites are similarly defined, the only difference being a rearrangement of codon arguments in the Ω functions.

Each term is divided by a sum over the starting state codon so that q_{ij} represents an instantaneous rate of transition from i to j, rather than flux. Ensuring the model is reversible (satisfying detailed balance: $\pi_i q_{ij} = \pi_j q_{ji}$) we define the equilibrium frequency of nucleotide i:

$$\pi_i = \left(\sum_k^4 \sum_l^4 P_{ikl} \right) \left(\sum_h^4 \sum_k^4 P_{hik} \right) \left(\sum_g^4 \sum_h^4 P_{ghi} \right). \tag{5.15}$$

This model has $N_G + 1$ free parameters: κ and one ω for each of N_G genes.

Rather than normalising the matrix first by estimating a compensating scaling factor as above, for this model we undertook to normalise each rate matrix by computing the scalar ν exactly:

$$\nu = \frac{\sum_z^{N_z} \sum_s^3 n_{zs}}{-\sum_z^{N_z} \sum_s^3 n_{zs} \sum_i^4 \pi_i q_{ii,zs}} \tag{5.16}$$

where n_{zs} is the number of sites in zone z and site type $s \in \{\alpha, \beta, \gamma\}$, N_z is the number of zones, i is nucleotide state and $q_{ii,zs}$ is an element of $\boldsymbol{Q}_{[zs]}$, which is specific for sites of type s in zone z. The numerator is equal to the total number of

nucleotide sites in the alignment.

5.3.2.4 Pentamer Model

The frame independence model treats a substitution as an independent process in each frame. In fact, nucleotide substitutions in overlapping codons are not independent, since a substitution affects all three frames simultaneously. We assume, however, that the probabilities of overlapping codons are independent of each other. We can therefore define the probability of a nucleotide pentamer $ghikl$, comprising three complete codons, as

$$P_{ghikl} = \frac{P_{ghi}P_{hik}P_{ikl}}{\sum_{g'}^{4}\sum_{h'}^{4}\sum_{i'}^{4}\sum_{k'}^{4}\sum_{l'}^{4} P_{g'h'i'}P_{h'i'k'}P_{i'k'l'}} \tag{5.17}$$

where codon probabilities P_{xyz} are taken from counts in sequence databases, as described above. A derivation explaining the relationship between codon and pentamer probabilities is presented in appendix F. For an α site,

$$q_{ij,\alpha} = \frac{\nu K(i,j)\sum_{g}^{4}\sum_{h}^{4}\sum_{k}^{4}\sum_{l}^{4} P_{ghikl}\Omega_A(ikl,jkl)\Omega_B(ghi,ghj)\Omega_C(hik,hjk)P_{ghjkl}}{\sum_{g'}^{4}\sum_{h'}^{4}\sum_{k'}^{4}\sum_{l'}^{4} P_{g'h'ik'l'}} \tag{5.18}$$

where Ω functions are defined as in equation 5.14 and the matrix scaling factor ν is computed as in equation 5.16. Nucleotide equilibrium frequencies are defined

$$\pi_i = \sum_{g}^{4}\sum_{h}^{4}\sum_{k}^{4}\sum_{l}^{4} P_{ghikl} \tag{5.19}$$

and it can be seen that detailed balance is satisfied. This model has the same free parameters as the frame independence model.

5.3.3 Test for Positive Selection

5.3.3.1 Uniform Selective Constraints

Using the models above one can test for positive selection (i.e. test if $\omega > 1$) in a chosen gene using the LRT. A null model is fitted where the gene 1's ω is fixed to 1, while ω values of other genes are freely optimised. An alternative model is separately fitted without constraint. The LRT with 1 degree of freedom then provides statistical significance for non-neutral evolution, and a maximum likelihood estimate (MLE)

$\omega > 1$ indicates positive selection.

5.3.3.2 Mixture Models For Identifying Varying Selective Constraints

The models above will estimate a single ω value for each gene present in the dataset. It is unlikely, however, that selective constraints will be uniform across a coding sequence. Nielsen and Yang (1998) defined mixture models of codon evolution for a single gene, permitting negative, neutral and positive selection selection by defining $\omega_0 < 0$, $\omega_1 = 1$ (fixed) and $\omega_2 > 1$ which populate three separate codon substitution matrices. A site class probability distribution comprising p_0, p_1 and $p_2 = 1 - p_0 - p_1$ is estimated and a site's probability is given by a sum over site classes, in which the probability of the site class is multiplied by the probability of the site, conditional on that site class' ω. Two mixture models are defined: the null model has two site classes ($\omega_0 < 0$, $\omega_1 = 1$) and the alternative has the positive selection class ($\omega > 1$) in addition. The LRT is used to assess significance of a superior fit of the positive selection model.

Similarly, we define two or three site classes with the same constrained ω parameters and site class probability distributions, but for each individual gene. For the positive selection (alternative) model we define a rate matrix $\boldsymbol{Q}_{[zsm_Am_Bm_C]}$ which is specific to the two or three site classes m for the genes in frames A, B and C in a site of type s in zone z. Again we normalise with scalar ν so the average substitution rate across all matrices is 1:

$$\nu = \frac{\sum_z^{N_z} \sum_s^3 n_{zs}}{-\sum_z^{N_z} \sum_s^3 n_{zs} \sum_{m_A}^{N_m} \sum_{m_B}^{N_m} \sum_{m_C}^{N_m} p_{m_A}^A p_{m_B}^B p_{m_C}^C \sum_i^4 \pi_i q_{ii}^{[zsm_Am_Bm_C]}} \tag{5.20}$$

where m_A, m_B and m_C are site classes (of which there are up to N_m; i.e. 2 for the null model and 3 for alternative) and $p_{m_A}^A$ is the probability that the gene in frame A of zone z is in site class m_A at this site, etc.

Having computed transition probability matrices, the likelihood for nucleotide alignment column X_k is then given by

$$L(X_k) = \sum_{m_A}^{N_m} \sum_{m_B}^{N_m} \sum_{m_C}^{N_m} p_{m_A}^A p_{m_B}^B p_{m_C}^C P(X_k | \omega_{m_A}^A, \omega_{m_B}^B, \omega_{m_C}^C), \tag{5.21}$$

where P defines the probability of column X_k given the combination of ω parameters for each frame (together defining $\boldsymbol{Q}_{[zsm_Am_Bm_C]}$), computed by the pruning algorithm

(Felsenstein, 1981). The total likelihood for an alignment is then a product over individual sites, or, in practice, a sum of log-likelihoods.

For the alternative model, the free parameters are: κ and for each gene ω_0, ω_2 ($\omega_1 = 1$), and also for each gene p_0 and p_1 ($p_2 = 1 - p_0 - p_1$). The null model differs in having no ω_2 and only p_0 is free ($p_1 = 1 - p_0$) for the gene of interest. In practice we fix κ to the MLE from HKY when estimating branch lengths.

Positive selection is inferred to occur at some sites in the gene of interest if the alternative model has significantly better likelihood than the null, assessed with the LRT and two degrees of freedom. This mixture model scheme could be applied to any of the models outline above, though we have done so with the pentamer model only.

5.3.3.3 Empirical Bayes Procedure

Continuing in the vein of Nielsen and Yang (1998), we can then identify sites with statistical support for undergoing positive selection. Maximum likelihood estimates (MLEs) for parameter values are used in an empirical Bayes procedure to compute the posterior probability that an alignment column belongs to a particular site class, conditional on the data observed at that site.

The probability of *nucleotide* alignment column X_k belonging to site class x in the gene residing in frame A at this site is

$$P\big(\omega^{A(k)} = \omega^A_{m_A} | X_k\big) = \frac{p^A_{m_A} \sum_{m_B} \sum_{m_C} p^B_{m_B} p^C_{m_C} P(X_k | \omega^A_{m_A}, \omega^B_{m_B}, \omega^C_{m_C})}{\sum_{m'_A} \sum_{m'_B} \sum_{m'_C} p^A_{m'_A} p^B_{m'_B} p^C_{m'_C} P(X_k | \omega^A_{m'_A}, \omega^B_{m'_B}, \omega^C_{m'_C})}, \quad (5.22)$$

where each ω^F_m is the MLE ω of site class m belonging to the gene in frame F and p^F_m is the MLE probability of site class m for the gene in frame F.

However, the site class of an individual nucleotide site is not biologically meaningful, as we assume selection acts at the level of codons rather than nucleotides. Indeed, neighbouring nucleotide sites comprising a single codon in a particular frame could be ascribed different site classes, implying paradoxically that a codon can experience apposing selective constraints simultaneously. We therefore want to calculate the probability that a *codon* in any frame belongs to a particular site class.

A single codon V made up of nucleotide sites V_1, V_2 and V_3 partially overlaps with four other codons W, X, Y and Z (fig. 5.5). W and X are neighbouring codons in the same frame and may belong to the same gene (and the same for Y

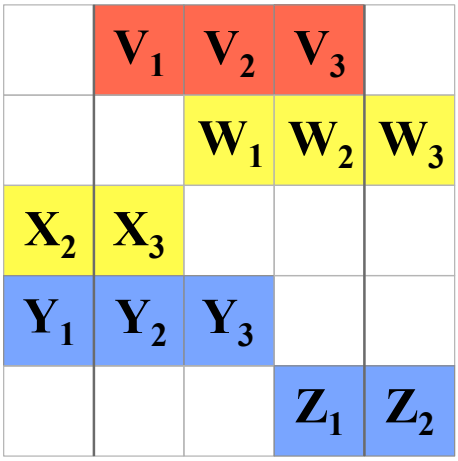

Figure 5.5: A single codon partially overlaps with four other codons in the two remaining reading frames. Columns in the grid represent nucleotide sites comprising a pentamer, at the centre of which lies codon V. Positions within each codon are shown as subscripts and the boundaries of codon V are shown with bold lines. Each lower row contains a codon partially overlapping with V. Codons X (of which only positions 2 and 3 are shown) and W are neighbours in the same frame (yellow); likewise codons Y and Z (of which positions 1 and 2 are shown) are in the same frame (blue). If V spans a zone boundary, W/Y will be in a different gene (or noncoding region) to X/Z.

and Z), but considering them separately allows for the situation where V spans a border between zones and the neighbouring codons belong to separate genes, or one is non-coding. As defined here, frames V, WX and YZ could represent frames A, B and C in any combination; e.g. if V_1 is an α site then V is in frame A, XW is in frame B and YZ is in frame C. (We do not calculate probabilities for codons which span splice boundaries (as in the HIV-1 genome, fig. 5.3) since concatenating terminal sites to form the bridging codon in the frame of interest, as would happen in splicing, gives rise to non-existent codons in the other frames.)

The probability of codon V belonging to site class i given data at nucleotide sites V_1, V_2 and V_3 is equal to the (prior) probability of site class i for the gene to which codon V belongs (p_i^V) multiplied by the likelihood of nucleotide sites V_1, V_2 and V_3 conditional on the ω for the V gene being of class i (ω_i^V), normalised by the unconditional likelihood of the data at nucleotide sites V_1, V_2 and V_3:

$$P(\omega^V = \omega_i^V | X_V) = \frac{p_i^V P(V_1, V_2, V_3 | \omega_i^V)}{P(V_1, V_2, V_3)} =$$

$$\frac{p_i^V \sum_{j,k,l,m} [p_j^W p_k^X p_l^Y p_m^Z] P(X_{V1} | \omega_i^V, \omega_k^X, \omega_l^Y) P(X_{V2} | \omega_i^V, \omega_j^W, \omega_l^Y) P(X_{V3} | \omega_i^V, \omega_j^W, \omega_k^X)}{\sum_{i',j',k',l',m'} [p_{i'}^V p_{j'}^W p_{k'}^X p_{l'}^Y p_{m'}^Z] P(X_{V1} | \omega_{i'}^V, \omega_{k'}^X, \omega_{l'}^Y) P(X_{V2} | \omega_{i'}^V, \omega_{j'}^W, \omega_{l'}^Y) P(X_{V3} | \omega_{i'}^V, \omega_{j'}^W, \omega_{k'}^X)} \tag{5.23}$$

where, i, j, k, l, m and the same variables with primes represent site classes for the genes residing in codons V, W, X, Y and Z, respectively. As in equation 5.21, the likelihood functions involve constructing rate matrices from the sets of ω parameters, which will depend on the site types of V_1, V_2 and V_3.

5.3.4 Synthetic Data for Model Testing

5.3.4.1 Uniform Selective Constraints (LSD Simulation)

We want to test our nucleotide substitution models with synthetic data involving overlapping genes, generated by a process where selection has acted at the codon level. We use a kinetic Monte Carlo simulation which makes nucleotide substitutions at rates dependent on states at both the changing site and its neighbours. We call this the local site dependence (LSD) simulation.

Each nucleotide site is at the centre of a pentamer (fig. 5.4) and has a rate of substitution to a specific target nucleotide dependent on the probabilities for the new pentamer's three codons and the ω value for each relevant gene, if the substitution causes a nonsynonymous codon change in that frame. For example, if site x is of type α and is in the centre of pentamer which has nucleotide states $ghikl$, then the rate of substitution $i \to j$ is

$$r_x(i,j|g,h,k,l) = \nu K(i,j)\Omega_A(ikl,jkl)\Omega_B(ghi,ghj)\Omega_C(hik,hjk)P_{jkl}P_{ghj}P_{hjk}, \quad (5.24)$$

the components of which are defined as in §5.3.2. The formulation for β and γ sites is the same, but with alternative combinations of codons and Ω functions, as the codons belong to different genes. (ν is a scaling factor which scales the rates so branch lengths can represent 1 substitution per site per unit length, which we define below.)

The overall rate of substitution across the sequence is then

$$R = \sum_{x}^{N_{\text{sites}}} \sum_{j \neq i}^{3} r_x(i,j|g,h,k,l) \quad (5.25)$$

where N_{sites} is the number of nucleotide sites and again $ghikl$ are the nucleotide states of pentamer centred on site x.

Substitutions arise as a Poisson process and therefore, evolving along a branch of length T, intervals between substitutions δt are drawn from an exponential distribution parameterised by R, so that

$$P(\delta t|R) = Re^{-R\delta t}. \quad (5.26)$$

A nucleotide site x and target state j are chosen with probability

$$\frac{r_x(i,j|g,h,k,l)}{R} \tag{5.27}$$

and an $i \rightarrow j$ substitution occurs at site x. This is repeated until the sum of δt values equals or would otherwise exceed T (at which point we have reached the end of the branch). As we can compute substitution rates only for sites at the centre of a pentamer, we do not permit substitutions at the first and last two nucleotide sites of the sequence. Evolution can be simulated in this way along each branch in a tree topology.

The root or starting sequence for the simulation is created by firstly choosing each site's state from a uniform distribution, and any stop codon (TAA, TGA or TAG) identified in any frame has its leading T replaced with a C. This sequence is then 'evolved' along a branch length of 10.0 substitutions per site using the procedure described above and the specified parameter values, so that the root sequence is in equilibrium at the start of the simulation.

The scaling factor ν is computed by numerical simulation. The product of a branch length t and overall substitution rate R gives the number of substitutions to have occurred; we wish to scale R such that $t = 1$ corresponds to 1 expected substitution per site. Evolution of a single sequence (starting from a random sequence corrected for stop codons, as above) is performed with the above procedure (but where $\nu = 1$ in equation 5.24) until a specified number of substitutions have occurred (in practice $N_{\text{subs}} = 1000$). With each substitution a δt is drawn from an exponential distribution, as in equation 5.26. If the sum of these is T, then

$$\nu = \frac{T}{N_{\text{subs}}/N_{\text{sites}}}, \tag{5.28}$$

where N_{sites} is the number of sites in the sequence. Scaling each $r_x(i,j|g,h,k,l)$ as in equation 5.24 results in

$$R = \frac{R_0 T}{N_{\text{subs}}/N_{\text{sites}}} \tag{5.29}$$

(where R_0 is the unscaled value), causing $R \approx N_{\text{sites}}$, because the other terms cancel. Therefore if $t = 1$, then $tR = N_{\text{subs}} \approx N_{\text{sites}}$, meaning there is one expected substitution per site. For each simulation, we estimate 10 ν values independently with different starting sequences and use the mean of these in the simulation which generates sequences for analysis.

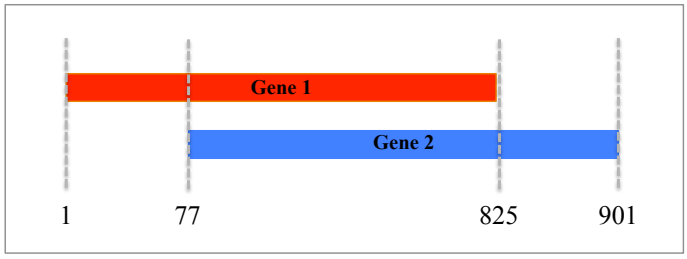

Figure 5.6: Layout of genes in synthetic data simulations and used for testing substitution models. Genes 1 and 2 were in frames A and B, respectively, and each 275 codons long. The total length of the sequences was 901 nucleotides, and the nucleotide site numbers for the zone boundaries are indicated. Not drawn to scale.

5.3.4.2 Variable Selective Constraints (vLSD Simulation)

Equation 5.24 assumes each gene has a single ω value. For testing the effectiveness of our mixture models at identifying codon sites undergoing different types of selection, we modify the simulation algorithm so two or three ω values can be specified for each gene, and call this the variable local site dependence (vLSD) simulation. Prior to the simulation, each codon in a gene is dedicated to one of three site classes, according to probabilities p_0, p_1 and p_2, the values of which are specified for each gene. We then specify values for parameters $\omega_0 < 1$, $\omega_1 = 1$ and $\omega_2 > 1$ for each gene. Then, 5.24 becomes:

$$r_x(i,j|g,h,k,l) = \nu K(i,j)\Omega_A(ikl,jkl|x)\Omega_B(ghi,ghj|x)\Omega_C(hik,hjk|x)P_{jkl}P_{ghj}P_{hjk}, \tag{5.30}$$

where the Ω functions are defined as before (equation 5.14), but now return the ω associated with the site class of the codon to which nucleotide site x belongs. All other aspects of the simulation remain the same.

5.4 Results

5.4.1 Testing Substitution Models

We tested the substitution models described above under varying conditions of ω values, total branch length and overlap proportion. We generated five datasets for each test condition, using the local site-dependence simulation (LSD). Every dataset comprised 64 sequences simulated using a bifurcating tree topology and contained two genes (numbered 1 and 2) and, unless varying overlap proportion itself, with most of their lengths overlapping (fig. 5.6; see Methods §5.6.3). These were analysed with each of the models described to assess accuracy of ω estimation, particularly for gene 1.

5.4.1.1 Genetic Code Weighting Model Over-Estimates ω

Overlap proportion and sequence divergence

We tested accuracy of ω estimation by the genetic code weighting model (Theory §5.3.2.1) over increasing proportions of gene overlap in the local site dependence (LSD) simulation and found no clear correlation between inaccuracy and overlap proportion (appendix G.1). However, it was clear that accurate ω estimation is not possible when sequences are entirely overlapping, and so subsequent tests were done with a proportion of 5/6 overlap to provide challenging but not impossible test cases (fig. 5.6). Similarly, testing ω estimation with data generated using a range of input branch lengths found no correlation between inaccuracy and divergence within a biologically relevant range (appendix G.2). In both these sets of tests, over-estimation of ω was seen when the input $\omega > 1$ for the gene of interest and when using HIV-1 codon probabilities.

Varied input ω values

We then tested the model's ability to estimate ω for the gene of interest across varying input ω values using data generated with the LSD simulation, where the background gene (gene 2) $\omega \in \{0.7, 1.7\}$ (fig. 5.7). Using equal codon probabilities (panels A and B) there is slight overestimation of ω at high values (> 2) but approximately accurate estimation of the ω for gene 2. Using human codon probabilities (fig. 5.7 panels C and D) the MLE value is accurate except when the input $\omega = 3$. Using HIV-1 codon probabilities (panels E and F), however, shows overestimation of ω for each gene in almost every condition.

False positive rate

We sought to assess the rate at which the model falsely identifies positive selection. We simulated 100 sequence datasets for each of the three sets of codon probabilities used above, for both conditions where the gene 2 input ω is 0.7 or 1.7 and where gene 1 input ω is 0.4, 0.6, 0.8 or 1.0 (i.e. $100 \times 3 \times 2 \times 4 = 2400$ datasets in total). We then fitted the genetic code weighting model as above to each dataset and a null model where the ω parameter is fixed at 1.0. As the null model is nested within the alternative with a difference of one free parameter, we can use the likelihood ratio test (LRT) with 1 degree of freedom to assess the statistical significance of the alternative model's better fit to the data. We define a false positive as the LRT wrongly rejecting the null model and $\hat{\omega} > 1+\epsilon$, where $\hat{\omega}$ is the MLE value from the alternative model and ϵ is a tolerance threshold, which we set to 0.2.

Table 5.2 shows the false positive counts. There were no false positives in any condition

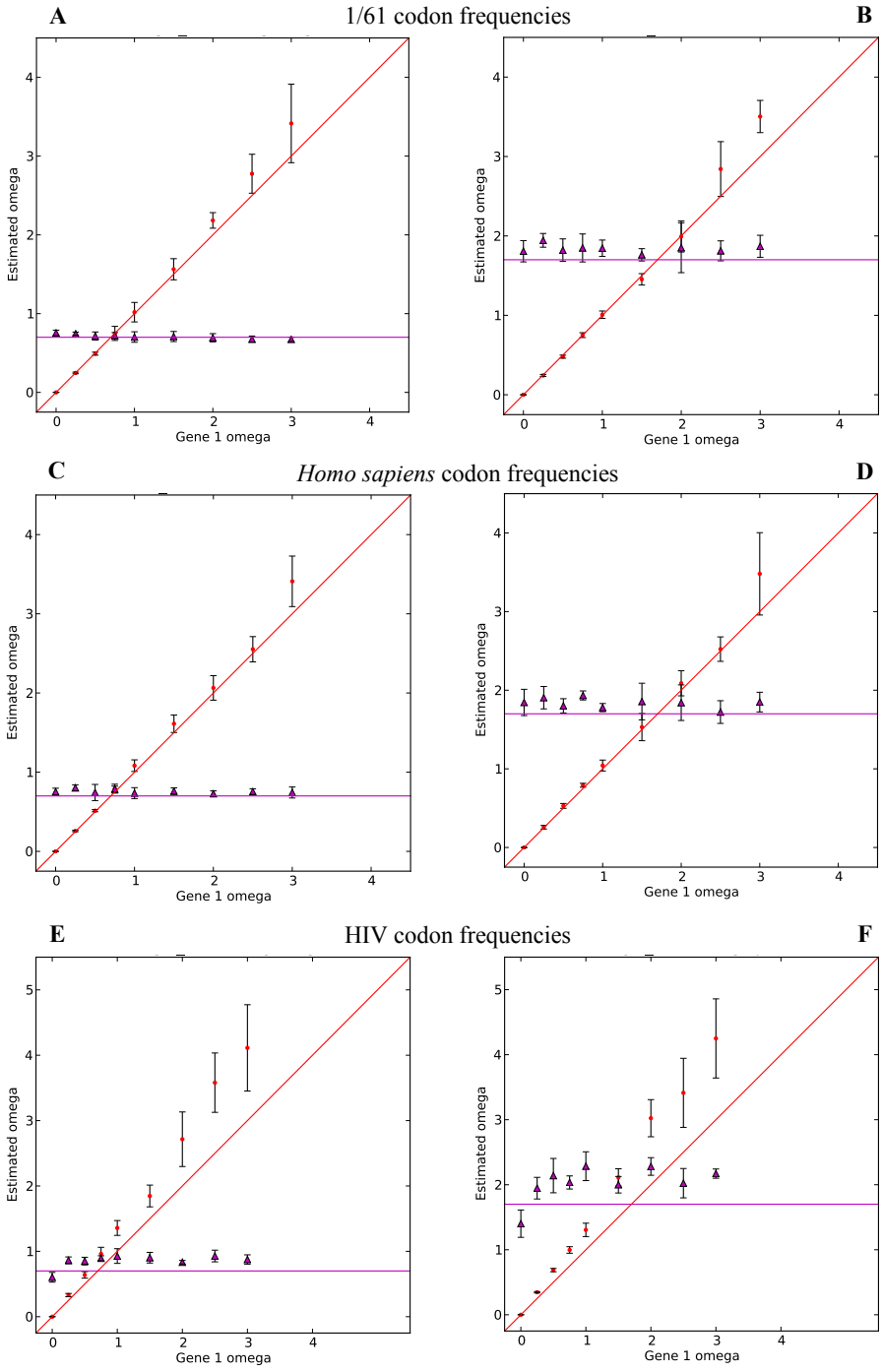

Figure 5.7: Maximum likelihood estimates of ω values estimated by the genetic code weighting model as a function of input ω value for gene 1 in simulated data. While the gene 1 ω value was varied across simulated data sets (red points, with solid red line showing the true value), gene 2's ω was fixed at either 0.7 or 1.7 (purple triangles, with solid purple line showing true value). Each point is the mean estimate from 5 simulated datasets and error bars show 1 standard deviation unit. NB the scale for the y-axis is different in panels E and F compared with others.

Codon Prob.	Input G. 1 ω	Input G. 2 ω	False Positives (of 100 tests)
1/61	0.4	0.7	0
1/61	0.6	0.7	0
1/61	0.8	0.7	0
1/61	1.0	0.7	0
1/61	0.4	1.7	0
1/61	0.6	1.7	0
1/61	0.8	1.7	0
1/61	1.0	1.7	0
H. sapiens	0.4	0.7	0
H. sapiens	0.6	0.7	0
H. sapiens	0.8	0.7	0
H. sapiens	1.0	0.7	0
H. sapiens	0.4	1.7	0
H. sapiens	0.6	1.7	0
H. sapiens	0.8	1.7	0
H. sapiens	1.0	1.7	1
HIV-1	0.4	0.7	0
HIV-1	0.6	0.7	0
HIV-1	0.8	0.7	4
HIV-1	1.0	0.7	89
HIV-1	0.4	1.7	0
HIV-1	0.6	1.7	0
HIV-1	0.8	1.7	16
HIV-1	1.0	1.7	98

Table 5.2: Genetic Code Weighting Model False Positives. Codon Prob., codon probabilities set; G., gene.

where the input ω for gene 1 was 0.4 or 0.6, and there were none when using equal codon probabilities for any input ω value. There was a single false positive when using human codon probabilities and the input ω value was 1.0. But when using HIV-1 codon probabilities, the number of false positives was unacceptably high, rising to 98% when input $\omega = 1$ and gene 2 $\omega = 1.7$. This is consistent with the other simulation test data (fig. 5.7) showing overestimation of ω at high input values, particularly when using HIV-1 codon probabilities.

5.4.1.2 High Inequality of Synonymous Codon Probabilities Affects ω estimation

The very high false positive rate found when using HIV-1 codon probabilities contrasted markedly with the very low rate found when using equal or human codon probabilities. Comparing the sets of probabilities directly, we observed that the distribution of codon probabilities within a set of codons coding for the same residue appears more uneven in HIV-1 probabilities than in human, as measured by the Gini coefficient, a measure of inequality (fig. 5.8; Ceriani and Verme, 2012). In the LSD simulation, the substitution

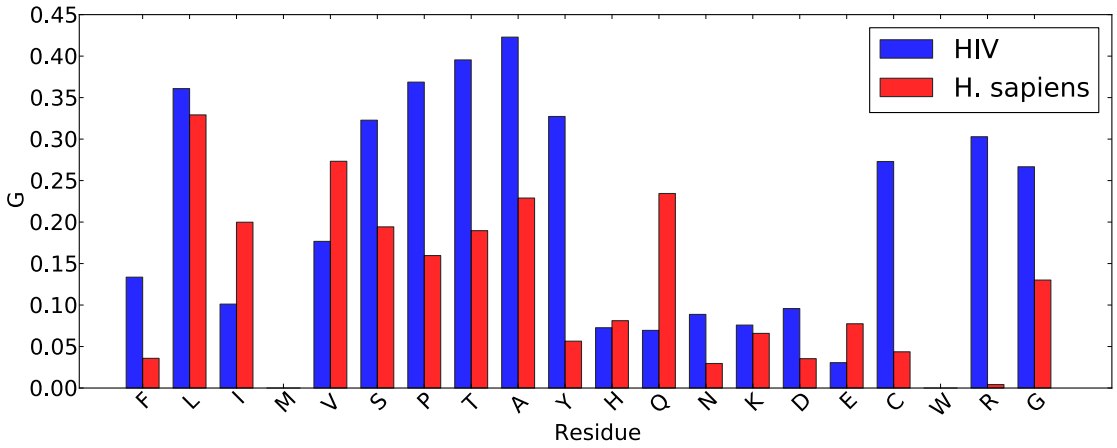

Figure 5.8: Gini coefficient (G) for each set of synonymous codon probabilities (denoted by the letter code for the amino acid they represent), comparing those of *Homo sapiens* and Human Immunodeficiency Virus type 1, from the Codon Usage Database.

rate at a nonsynonymous nucleotide site is given by a product of the ω value and the target codon probability (eq. 5.24). Therefore substitution from a codon of high probability to a synonymous codon of much lower probability could have a lower rate than substitution to a nonsynonymous codon, even if $\omega < 1$.

To test whether uneven probabilities within synonymous codon groups are responsible for the inaccurate ω estimation, we generated new sets of simulated data as before but using human and HIV-1 codon probabilities averaged across synonymous groups. Fitting the genetic code weighting model to each of these datasets, we found accuracy of ω estimation was much improved (fig. 5.9). This suggested the accuracy of the model was sensitive to codon bias and we therefore considered adjustments to the model to account for it.

5.4.1.3 Codon Weighting Model Over-Estimates ω

We tested the ability of the codon weighting model (Theory §5.3.2.2) to accurately estimate ω using the same simulated datasets (fig. 5.10). As would be expected, there is little difference in the accuracy of ω estimation when using codon probabilities of 1/61, since the effect is only to take account of the three stop codons (eq. 5.12). There was little difference in accuracy when using human codon probabilities and over-estimation of ω was actually worse when using HIV-1 probabilities. It is therefore likely the false positive rate would be unacceptably high.

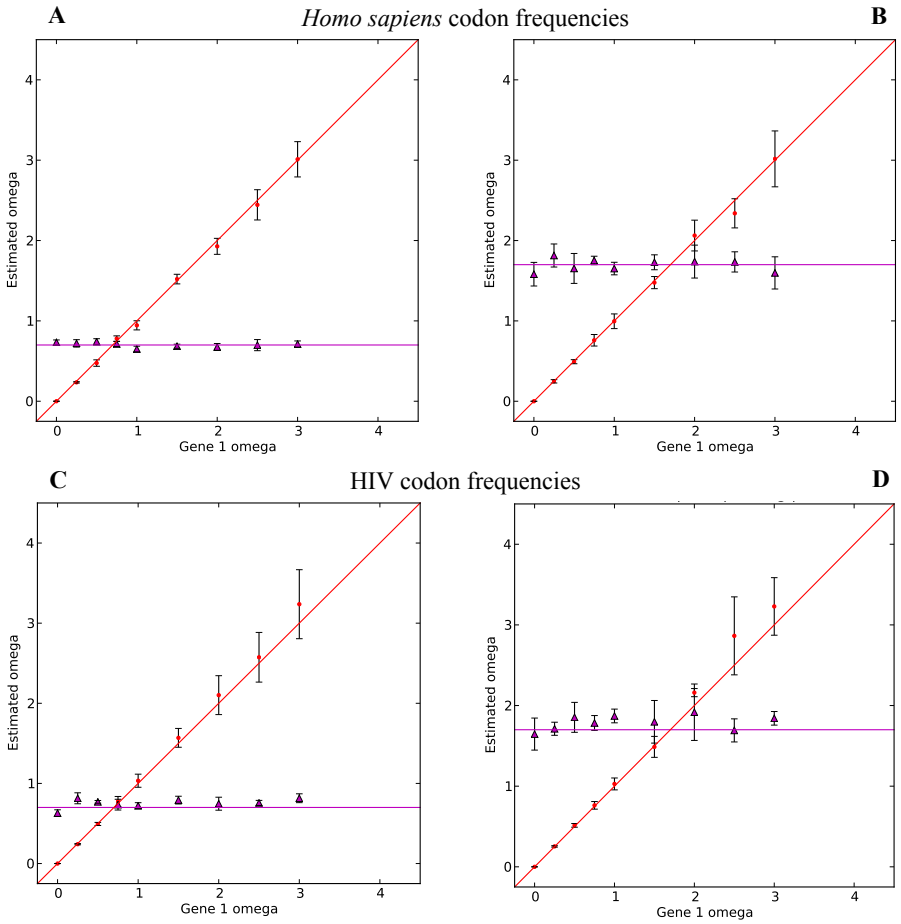

Figure 5.9: Codon weighting model fitted to data simulated using averaged human and HIV-1 synonymous codon probabilities. While the gene 1 ω value was varied across simulated data sets (red points, with solid red line showing the true value), gene 2's ω was fixed at either 0.7 or 1.7 (purple triangles, with solid purple line showing true value). Each point is the mean estimate from 5 simulated datasets and error bars show 1 standard deviation unit. Note that all panels are on the same scale.

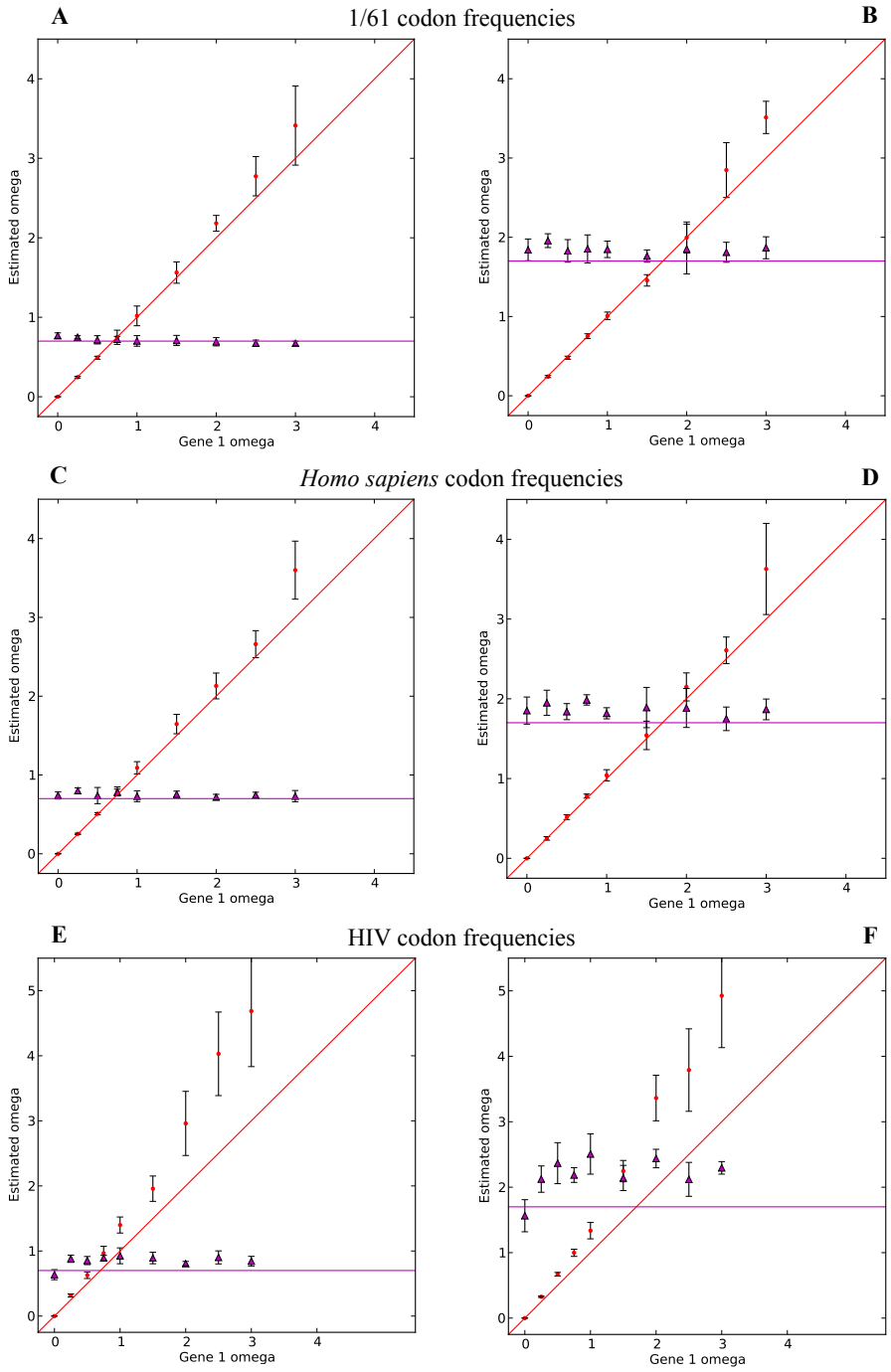

Figure 5.10: Codon weighting model and varied ω values, presented as in fig. 5.7. While the gene 1 ω value was varied across simulated data sets (red points, with solid red line showing the true value), gene 2's ω was fixed at either 0.7 or 1.7 (purple triangles, with solid purple line showing true value). Each point is the mean estimate from 5 simulated datasets and error bars show 1 standard deviation unit. NB panels E and F are on a different scale.

5.4.1.4 Frame Independence Model Over-Estimates ω

We sought to develop a new model which accounted for both codon biases and uncertainty about sequence context explicitly. The frame independence model (Theory, §5.3.2.3) was assessed for its ability to accurately determine ω values by fitting to the same simulated datasets (fig. 5.11). These data show almost entirely accurate estimation of ω when using equal and human codon probabilities. There is an improvement in accuracy when using HIV-1 codon probabilities compared with the models presented above, but there is still a considerable over-estimation. We anticipate, as with the genetic code weighting model, the over-estimation of ω would produce a high false positive rate.

5.4.1.5 Pentamer Model Accurately Estimates ω Values and Has Low False Positive Rate

ω Estimation

Reflecting on the previous approach, we noted that the frame independence model considers codon transitions in each frame as independent events (equation 5.13), when nucleotide transitions involving codon biases may be better represented by considering a set of overlapping codons as a single unit.

We therefore developed the pentamer model (Theory, §5.3.2.4) and tested its ability to estimate ω values as above, by fitting the model to the same simulated datasets. There was accurate estimation and lower variance under all conditions compared to other models (fig. 5.12).

We next tested the effect on ω estimation of increasing the proportion of overlap using the same simulated data as analysed with the genetic code weighting model (appendix. G.3). We found ω values were accurately estimated in most conditions, with the exception of the extreme case of near total overlap (6/6), where there was over-estimation and much higher variance of the estimate. We then tested the effect of increasing branch lengths on ω estimation, again using the same data used to test the genetic code weighting model, and found mostly accurate estimates within a biologically relevant range of divergence (appendix G.4).

False Positive and Negative Rates

We assessed the false positive rate in the same approach taken in testing the genetic code weighting model, using the same simulated datasets. There were no false positives in any condition when using equal or human codon probabilities. Using HIV-1 codon probabilities, there were 2 false positives when gene 1's input $\omega = 1.0$ and gene 2 input

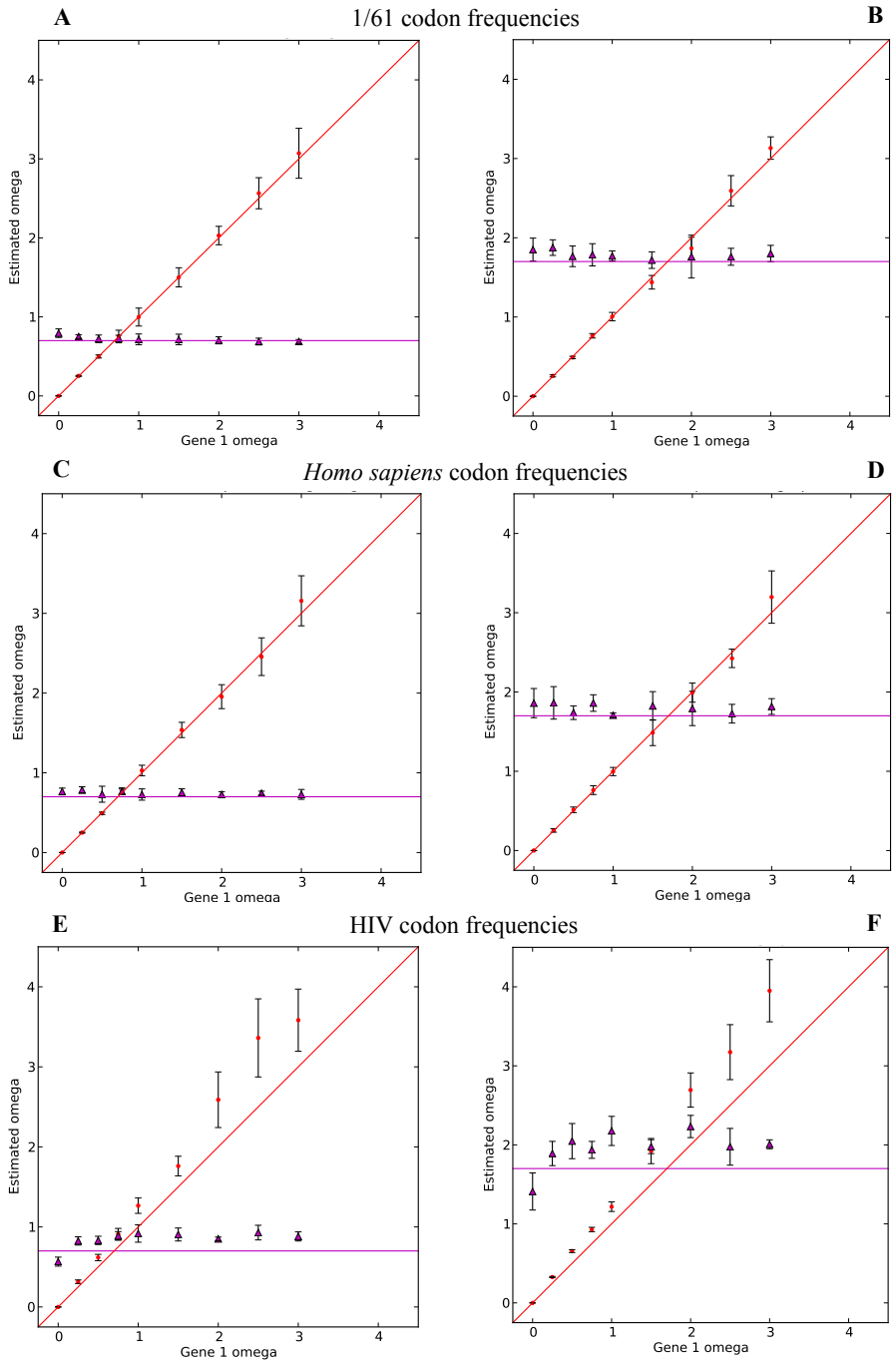

Figure 5.11: Frame Independence Model and Varied ω Values, presented as in fig. 5.7. While the gene 1 ω value was varied across simulated data sets (red points, with solid red line showing the true value), gene 2's ω was fixed at either 0.7 or 1.7 (purple triangles, with solid purple line showing true value). Each point is the mean estimate from 5 simulated datasets and error bars show 1 standard deviation unit. NB unlike in some previous figures, all panels are on the same scale.

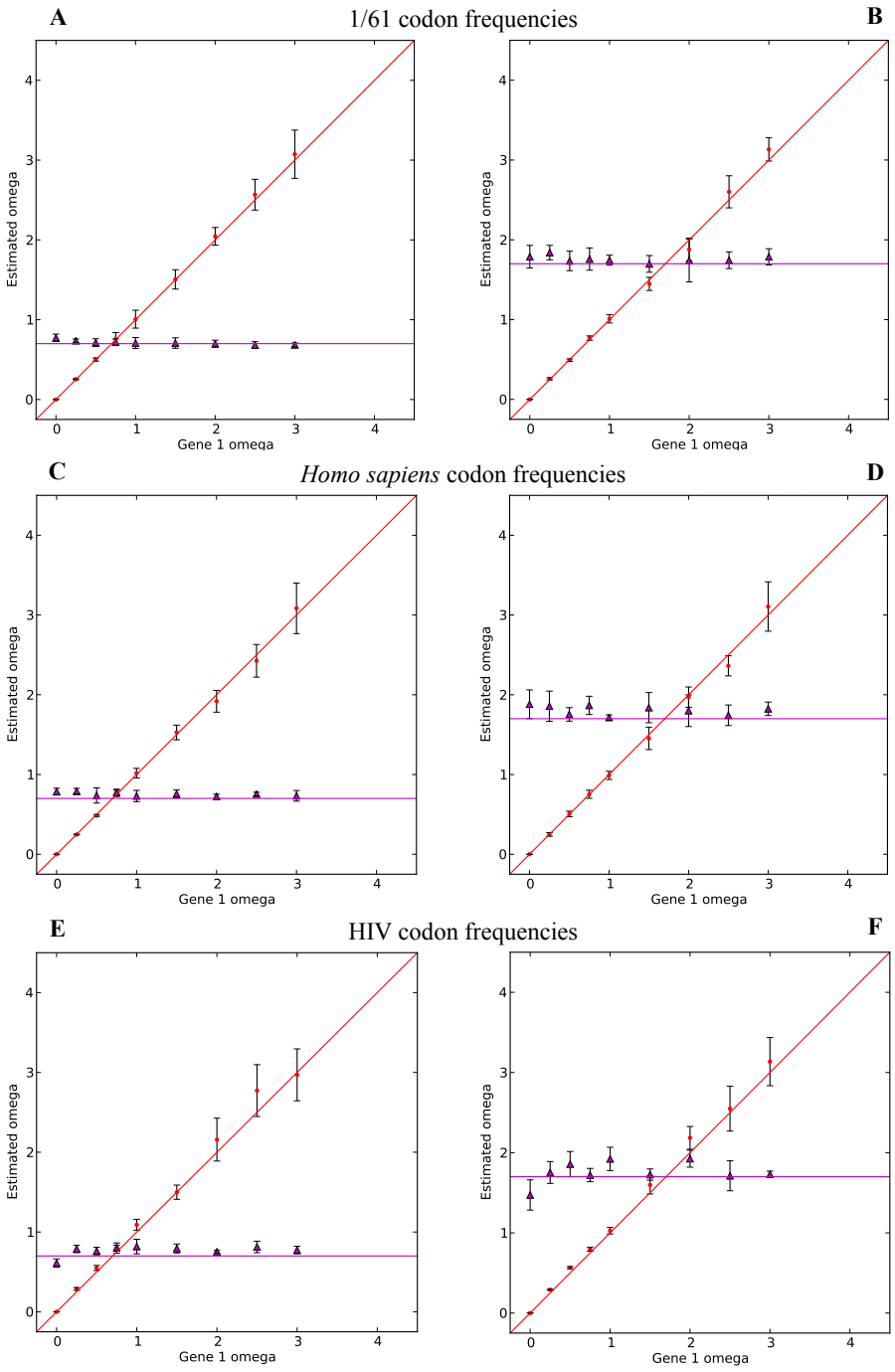

Figure 5.12: Pentamer model and varied ω values, presented as in fig. 5.7. While the gene 1 ω value was varied across simulated data sets (red points, with solid red line showing the true value), gene 2's ω was fixed at either 0.7 or 1.7 (purple triangles, with solid purple line showing true value). Each point is the mean estimate from 5 simulated datasets and error bars show 1 standard deviation unit. NB unlike in some previous figures, all panels are on the same scale.

Codon Prob.	Input G. 1 ω	Input G. 2 ω	False Negatives (of 100 tests)
1/61	1.2	0.7	12
1/61	1.4	0.7	0
1/61	1.6	0.7	0
1/61	1.2	1.7	8
1/61	1.4	1.7	0
1/61	1.6	1.7	0
H. sapiens	1.2	0.7	17
H. sapiens	1.4	0.7	1
H. sapiens	1.6	0.7	0
H. sapiens	1.2	1.7	16
H. sapiens	1.4	1.7	0
H. sapiens	1.6	1.7	0
HIV-1	1.2	0.7	0
HIV-1	1.4	0.7	0
HIV-1	1.6	0.7	0
HIV-1	1.2	1.7	0
HIV-1	1.4	1.7	0
HIV-1	1.6	1.7	0

Table 5.3: Pentamer Model False Negatives. Codon prob., codon probabilities set; G., gene.

$\omega = 0.7$, while there were 3 false positives when the gene of interest's input $\omega = 1.0$ and the gene 2's input $\omega = 1.7$, each out of 100 datasets.

We then assessed the false negative rate, following the same procedure but simulating data where gene 1 was under mild positive selection, with input ω equal to 1.2, 1.4 or 1.6 (table 5.3). Out of 100 tests, we found the highest number of false negatives was 17, when the input $\omega = 1.2$, using human codon probabilities and when gene 2's input $\omega = 0.7$; most conditions had no false negatives. Together, these data showed the pentamer model had acceptably low false positive and false negative rates.

5.4.2 Testing Pentamer Mixture Model

Having found the pentamer model to be accurate in simulation studies where genes experience uniform selective constraints across sites, we constructed a mixture model based on the pentamer model to account for more varied constraints (Theory, §5.3.3.2). We tested the model's false positive rate by fitting to datasets of 64 synthetic sequences generated with the vLSD simulation, again with two genes in frames A and B, with 300 codons in each and similar overlap proportion as in figure 5.6. We used each set of codon probabilities used previously and multiple input site class distributions (i.e. $\{p_0, p_1, p_2\}$; Methods §5.6.3 and table 5.5) for gene 1, but all with $p_2 = 0$, i.e. where there is no positive selection. All density was on the negative selection site class for gene 2 ($p_0 = 1$) and for both genes

$\omega_0 = 0.4$. For each combination of input parameters, we simulated three independent datasets.

5.4.2.1 Likelihood Ratio Test Shows High False Positive Rate

We fitted the null and alternative models to the datasets generated with the vLSD. Initial tests showed model optimisation was sensitive to starting values (data not shown) and we therefore repeated each model fitting using six different start values for the site class distributions to guard against optimisation errors (listed in appendix H.1). The log-likelihood values were mostly consistent across the different start conditions used (data not shown), but where there were differences we selected the highest log-likelihood result, summarised in table 5.4.

Despite there having been no positive selection in the generation of these 27 datasets, the LRT indicates significantly superior fit of the alternative model for 14 them. There were two false positives among the datasets generated with equal codon probabilities, both generated with input $p_1 = 1$. One dataset (generated with equal codon probabilities and $p_0 = 1$) saw a higher null log-likelihood than the best alternative model, indicating the alternative model had failed to find the global optimum. There were no false positives for any codon probabilities when input $p_0 = 1$. However, for datasets generated with HIV-1 or human codon probabilities, there were false positives from all datasets generated with input $\{p_0, p_1, p_2\} = \{0, 1, 0\}$ or $\{0.5, 0.5, 0\}$.

5.4.2.2 Parametric Bootstrapping Shows High False Positive Rate

In our simulation tests, the process generating the data (vLSD simulation) is not identical to the models which we fit to it (the pentamer mixture models, null or alternative). Having found a high false positive rate with the LRT, we considered whether the significantly large differences between the null and alternative log-likelihoods could be exaggerated by this model misspecification, in which case the null distribution used in the LRT (a chi-squared, here with 2 degrees of freedom) is somewhat inappropriate. We therefore investigated whether the outcome is consistent if using a precise null distribution, created by parametric bootstrapping.

As demonstrative examples, we took the MLE parameters from null models fitted to one dataset generated with equal, HIV-1 or human codon probabilities and input $p_1 = 1$ (i.e. 3 datasets in total; table 5.4, repeat index 1 in each case), for which there had been false positives observed originally. We used these to generate 100 datasets as before using the vLSD simulation, giving 300 datasets in total (summarised in fig. 5.13). We fitted the null and alternative models to each of these and computed each log-likelihood difference

Codon Freq.	$\{p_0, p_1, p_2\}$	Rep.	Null lnL	Alt. lnL	D	p	Alt. $\hat{\omega}_2$
1/61	$\{0, 1, 0\}$	1	-15166.217	-15162.658	7.118	0.028	1.281
1/61	$\{0, 1, 0\}$	2	-15279.973	-15275.446	9.054	0.011	1.511
1/61	$\{0, 1, 0\}$	3	-14677.330	-14674.988	4.683	0.096	1.231
1/61	$\{0.5, 0.5, 0\}$	1	-14990.498	-14988.325	4.345	0.114	1.694
1/61	$\{0.5, 0.5, 0\}$	2	-15416.263	-15416.249	0.028	0.986	1.012
1/61	$\{0.5, 0.5, 0\}$	3	-15074.770	-15074.687	0.166	0.920	1.475
1/61	$\{1, 0, 0\}$	1	-14611.263	-14613.908	-5.290	n/a	1.000
1/61	$\{1, 0, 0\}$	2	-15022.424	-15022.424	0.000	1.000	1.000
1/61	$\{1, 0, 0\}$	3	-14963.210	-14963.210	0.000	1.000	1.000
H. sapiens	$\{0, 1, 0\}$	1	-15254.079	-15239.219	29.720	0.000	2.275
H. sapiens	$\{0, 1, 0\}$	2	-15666.062	-15651.901	28.323	0.000	1.726
H. sapiens	$\{0, 1, 0\}$	3	-15543.447	-15532.750	21.393	0.000	1.365
H. sapiens	$\{0.5, 0.5, 0\}$	1	-14967.938	-14962.193	11.488	0.003	1.405
H. sapiens	$\{0.5, 0.5, 0\}$	2	-15214.590	-15210.196	8.789	0.012	2.149
H. sapiens	$\{0.5, 0.5, 0\}$	3	-15327.023	-15320.347	13.353	0.001	2.171
H. sapiens	$\{1, 0, 0\}$	1	-14934.784	-14934.784	0.000	1.000	1.000
H. sapiens	$\{1, 0, 0\}$	2	-15066.507	-15066.507	0.000	1.000	1.000
H. sapiens	$\{1, 0, 0\}$	3	-15307.179	-15307.179	0.000	1.000	1.000
HIV-1	$\{0, 1, 0\}$	1	-15840.412	-15804.508	71.807	0.000	2.309
HIV-1	$\{0, 1, 0\}$	2	-16494.878	-16463.533	62.688	0.000	2.145
HIV-1	$\{0, 1, 0\}$	3	-16460.042	-16419.475	81.134	0.000	2.051
HIV-1	$\{0.5, 0.5, 0\}$	1	-16080.608	-16048.353	64.511	0.000	2.342
HIV-1	$\{0.5, 0.5, 0\}$	2	-15563.937	-15538.949	49.976	0.000	2.299
HIV-1	$\{0.5, 0.5, 0\}$	3	-15817.324	-15796.941	40.766	0.000	2.035
HIV-1	$\{1, 0, 0\}$	1	-15948.315	-15946.565	3.499	0.174	2.401
HIV-1	$\{1, 0, 0\}$	2	-15998.759	-15996.573	4.373	0.112	1.383
HIV-1	$\{1, 0, 0\}$	3	-16122.190	-16121.468	1.445	0.485	2.148

Table 5.4: Log-likelihoods (natural logarithms) and likelihood ratio test statistics from analyses of datasets generated with the variable local site dependence (vLSD) simulation under a range of simulation conditions. Null and alternative models were fitted with 6 sets of starting values (see table H.1) and the best-fitting result for each dataset is shown. Columns: Codon freqs., codon probability set; $\{p_0, p_1, p_2\}$, input site class probability distribution in vLSD; Rep., arbitrary repeat dataset index (all generated with identical simulation conditions); Null lnL, null model log-likelihood; Alt. lnL, alternative model log-likelihood; D, test statistic for likelihood ratio test (LRT); p, p value computed with LRT, 2 degrees of freedom; Alt. $\hat{\omega}_2$, maximum likelihood estimate for ω_2 parameter by alternative model. A negative D indicates optimisation failure of the null model, meaning the LRT is not applicable (n/a).

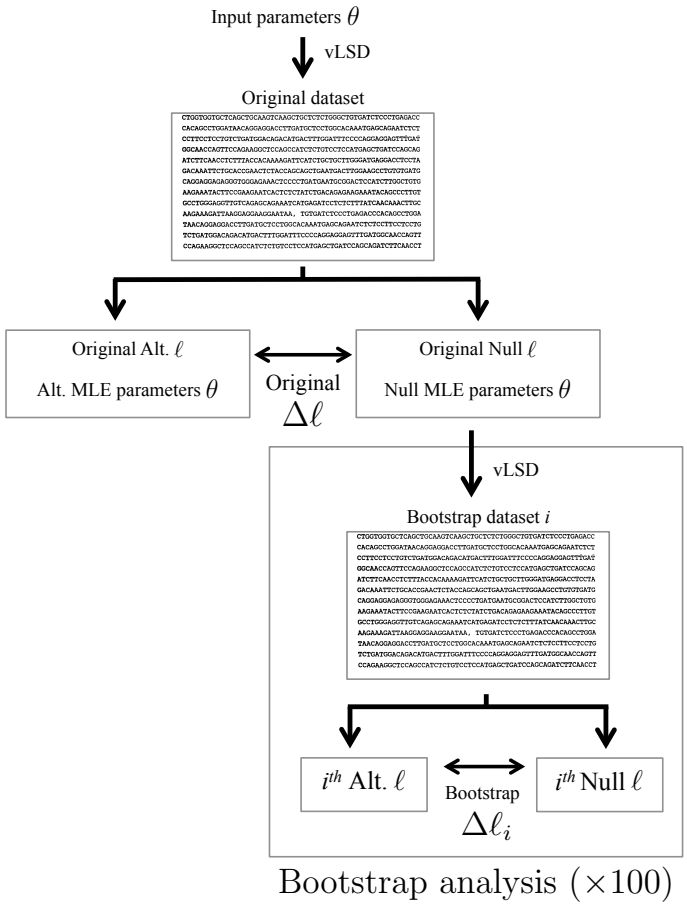

Figure 5.13: Summary of the parametric bootstrap procedure. Chosen input parameters ($\theta = \{\omega_0, \omega_2, p_0, p_1, \text{etc.}\}$) were used to generate the original sequence datasets with the variable local site dependence (vLSD) simulation. The alternative (Alt.) and null pentamer mixture models were fitted to these datasets (table 5.4), and the difference in log-likelihoods ($\Delta\ell$) was noted. The maximum likelihood estimate (MLE) parameter values from the null model were used to generate 100 bootstrap datasets, again with the vLSD. The alternative and null models were fitted to each of these and the bootstrap $\Delta\ell$ was noted. The set of bootstrap $\Delta\ell$ constitutes a null distribution. For example, if the original $\Delta\ell$ is greater than the 5th highest bootstrap $\Delta\ell$, then $p < 0.05$.

($\Delta\ell$; fig. 5.14). This was compared with the $\Delta\ell$ observed with the original datasets; these were 3.559, 35.903 and 14.860 for equal, HIV-1 and human codon probabilities, respectively.

The largest value in the distribution associated with equal codon probabilities (fig. 5.14a) was 3.651, slightly higher than the $\Delta\ell$ observed in the original analysis, which would indicate the null hypothesis could not be rejected with $p < 0.01$. However, this high value was itself an extreme outlier, and may result from an optimisation failure of the null model. (Limited computing resources prevented us trying different starting values for each bootstrap dataset.) The highest $\Delta\ell$ observed in the HIV-1 and human codon probability distributions were approximately 1×10^{-7} and 6×10^{-7}, respectively, considerably lower

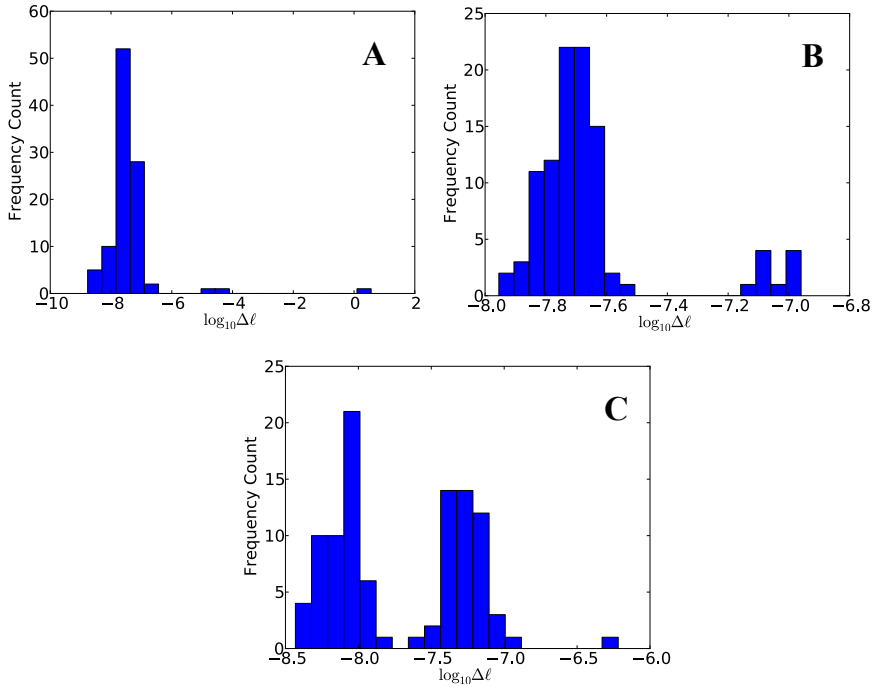

Figure 5.14: Frequency histograms of $\Delta\ell$, found by fitting the null and alternative models to 100 datasets generated with the variable local site dependence simulation (vLSD) using the maximum likelihood estimate parameter values, themselves found by fitting the same null model to data originally generated by the vLSD simulation with: (A) equal codon probabilities,; (B) HIV-1 codon probabilities; or (C) human codon probabilities; all where input $p_1 = 1$. Note that different scales are used in each panel.

than the $\Delta\ell$ observed with the original datasets (35.903 and 14.860, respectively).

Together, these results indicated that the null hypothesis could be comfortably rejected with $p < 0.05$ in each instance. The original model fitting results found to be significant with the LRT (table 5.4) therefore probably all constitute real false positives, despite the possible distortion caused by model misspecification.

5.5 Discussion

Multiple attempts have been made to study selection of overlapping coding sequences (Hein and Stovlbaek, 1995; Pedersen and Jensen, 2001; Sabath et al., 2008; Wei and Zhang, 2015), but to date a precise, mathematically consistent and tractable approach to modelling their evolution has not been developed. We have designed several substitution models intended for routine selection analysis of single-strand overlapping coding sequences.

The novelty of our approach is to restore the assumption of site independence by modelling selection at the level of nucleotides, taking into account uncertainty about the effect of nucleotide substitutions on the amino acid sequence. Applying standard nucleotide

substitution models to overlapping genes in the hepatitis B virus genome, Yang et al. (1995) were able estimate which overlapping genes were experiencing more stringent purifying selection, based on comparing the estimated nucleotide substitution rates associated with each gene's predominantly synonymous or nonsynonymous codon positions. However, from rate estimates alone it is difficult to infer the type of selection experienced by individual overlapping genes: for example, if two genes have their respective codon positions 1 and 2 at the same nucleotide sites, a somewhat elevated substitution rate is consistent with both (i) positive selection in one gene and purifying selection in the other, or (ii) relaxed purifying selection in both, etc. We have therefore invoked parameters (ω) representing explicitly the propensity for amino acid diversification over conservation for each gene. This offers the added advantage of being conceptually familiar to investigators routinely using standard codon models for identifying positive selection in non-overlapping coding sequences. Applying the mixture model approach with constrained ω parameters allows us to identify specific codon sites under positive selection (eq. 5.23), as is done with conventional codon models (e.g. Nielsen and Yang, 1998; Yang et al., 2005).

The simplistic genetic code weighting model (Theory §5.3.2.1), was found to be unable to estimate ω values accurately (fig. 5.7, 5.10), and was associated with an unacceptably high false positive rate (table 5.2). We note that the model formulation assumes both equal codon frequencies (eq. 5.1-5.3) and unequal nucleotide frequencies (eq. 5.6) and that this inconsistency may partly account for the poor parameter estimation. Similarly, the slightly modified codon weighting model (Theory §5.3.2.2), was also unable to estimate ω values accurately (fig. 5.7, 5.10) and relies on predefined codon frequencies (eq. 5.12), while also estimating nucleotide frequency parameters. The more sophisticated frame independence model (Theory §5.3.2.3) offered an improvement in accuracy but still over-estimated ω values (fig. 5.11).

The pentamer model (Theory §5.3.2.4) was found to be suitably accurate in estimating ω for all input values tested (fig. 5.12), to have acceptable false positive and false negative rates and to be accurate over several of gene overlap proportions and sequence divergence (appendix G.3, G.4). Based on the pentamer model substitution matrix, we implemented a mixture model which would allow us to investigate selective constraints varying across sites in overlapping coding sequences, without resorting to a overly parameter rich site-specific model (Nielsen and Yang, 1998). Using an empirical Bayes procedure we would then seek to identify codon sites under positive selection, while still modelling nucleotide substitutions.

Based on a small number of samples, we found the mixture model test for positive selection has an unacceptably high false positive rate both with the LRT (table 5.4)

and parametric bootstrapping (fig. 5.14). This was apparently associated only with data generated with all or half of the input probability density on the neutral selection class ($p_1 \in \{1, 0.5\}$). While the ω over-estimation and high false positive rate found with the non-mixture models were dependent on the codon probabilities used, this problem with the pentamer-based mixture model affected all codon probability sets tested.

It may be necessary to adjust the functional form used in the substitution model to better represent the nucleotide substitution process in overlapping coding sequences, and in particular it may be useful to better define codon probabilities. Here in the first instance, we have assumed codon probabilities for all genes under study can be represented by their frequencies in sequence databases for each organism; however, codon frequencies will be expected to differ for individual genes in real data. Indeed, our assumption of identical codon probabilities for overlapping coding sequences (eq. 5.17) may be inappropriately simplistic, since certain combinations of overlapping codons are effectively impossible. For example, methionine has a single codon (ATG) and a pentamer cannot contain more than one instance; therefore a high probability of ATG in one frame necessarily reduces the probability in other frames. Moreover, in this work we have fixed codon probabilities to the same values used in the vLSD simulation (which we might expect to unduly enhance accuracy) but this may prevent the model accounting for the stochastic variability in codons actually present in different datasets; a higher nonsynonymous substitution rate may therefore be favoured to account for the discrepancy between the codon probabilities provided and the frequencies observed. This could be addressed by several means:

1. For each gene in each dataset, use the observed codon counts as the probability distribution for that gene.

2. Estimate nucleotide equilibrium frequencies (probabilities) with a standard nucleotide substitution model and assume codon probabilities are the product of the probabilities of constituent nucleotides, as is done for codon models implemented in codeml (Yang, 2007). For example, one could allow a set of nucleotide frequencies for each site type in each zone; however, this would introduce $9N_z$ free parameters for N_z zones.

3. Estimate codon fitnesses more precisely. The pentamer model assumes pentamer probabilities are a function of codon fitnesses, which we assume underly the frequencies of codons in sequence databases (explained in appendix F). But codon fitnesses could be estimated from each dataset using codon substitution models which incorporate fixation probabilities (Halpern and Bruno, 1998; Yang and Nielsen, 2008; Tamuri et al., 2012). Codon models necessarily ignore overlapping coding sequences,

but this model misspecification may be acceptable if it provides a more accurate representation of codon probabilities.

Separately, some technical improvements could be made to the implementation. As we found there was sensitivity to parameter start values in the mixture model optimisations (data not shown), different numerical algorithms may identify likelihood peaks more reliably; e.g. the conjugate gradient method, which computes derivatives numerically. Instead or in addition, more sophisticated strategies for avoiding local optima could be used, such as multiple optimisations where output values from one optimisation course are manipulated in a non-deterministic way and then used as starting values for a subsequent course. We have conditioned on branch lengths estimated separately with the HKY85 substitution model for the sake of speed, but one would preferably estimate these alongside other model parameters.

We have not tested our implementation of an empirical Bayes approach to computing probabilities for codons belonging to particular site classes (Theory §5.3.3.3) and naturally this would be one of the first things to consider if the false positive rate is addressed successfully. It has been argued that this 'naive' empirical Bayes approach does not account for inadequate sampling and that a method incorporating data at all sites is more appropriate (Yang et al., 2005); similar extensions could be developed in this instance.

Our models' performances have been tested using data generated with a single genetic layout, namely two genes in frame A and B, respectively. It will be necessary to address whether the models are suited to multiple genetic layouts, as the various possible pairings (first codon position in A being the same as the third codon position in B, etc.; fig. 5.2) could have different evolutionary dynamics.

In conclusion, modelling selection in overlapping coding sequences has been attempted several times (Hein and Stovlbaek, 1995; Pedersen and Jensen, 2001; Sabath et al., 2008; Wei and Zhang, 2015), but still no precise method suited to routine analysis has been described. We have developed an approach which has initially shown encouraging results. However, improvements to the form of the substitution model will be required in order to extend it so variable selective constraints are accurately accounted for, and individual sites under positive selection can be identified.

5.6 Methods

5.6.1 Implementation

Each of the models described above have been implemented in a single program written in Java, with the working title 'YesWeCAN'. This uses several third-party class libraries: Phylogenetic Analysis Library for phylogenetics specific data structures (Drummond and Strimmer, 2001); Apache Commons Math for linear algebra and function optimisation routines (Apache-Commons, 2015); classes from swMutSel for representing optimisable model parameters (described by Tamuri et al., 2012); jCommander for command line interface (Beust, 2015); and opencsv for reading comma-separated value files (Smith and Conway, 2016).

We use tree branch lengths estimated by the HKY85 substitution model without rate heterogeneity (as implemented in baseml in the PAML package, version 4.7b, Yang, 2007) and the same topology as used in the simulations (see below).

5.6.2 Function Optimisation

5.6.2.1 Algorithms

The maximum likelihood optimisation of all non-mixture models used the Nelder-Mead simplex algorithm (Nelder and Mead, 1965), implemented in Apache Commons Math (Apache-Commons, 2015). This algorithm performed poorly in tests of the mixture model (data not shown) and subsequently Michael J. D. Powell's 'Bound Optimization BY Quadratic Approximation' (BOBYQA; Powell, 2009) algorithm was used in the mixture model, also implemented in Apache Commons Math.

5.6.2.2 Optimisation starting values

For the non-mixture models, starting parameter values were: $\kappa = 1$, $\omega = 1$ for each gene and, where relevant, $c = 1$. For the mixture model, we used repeated analyses using six combinations of start values for the site class probability distributions for each dataset analysed, detailed in appendix H.1. We used starting values of $\omega_0 = 0.5$, $\omega_2 = 1.5$, fixed κ to the MLE found with HKY85 while optimising branch lengths.

	p_0	p_1	p_2
a	0.0	1.0	0.0
b	0.5	0.5	0.0
c	1.0	0.0	0.0

Table 5.5: Site class probability distributions used for gene 1 in the variable local site dependence (vLSD) simulation to generate datasets exhibiting heterogeneous selection pressure. Three site classes (0, 1, and 2) were used, representing negative, neutral and positive selection respectively. Each distribution is named by a letter (first column) for reference.

5.6.3 Simulated Data for Testing Models

5.6.3.1 Testing Non-Mixture Models

For testing the non-mixture models, all synthetic datasets were generated with the LSD simulation, involving two overlapping genes in frames A and B respectively (fig. 5.6). Except when varying overlap proportion explicitly, there were 76 nucleotide sites containing either gene alone and 749 nucleotide sites of overlap. The tree used in the simulation was a symmetric 64 taxa bifurcating tree, with all branch lengths equal to 0.02; when testing branch length explicitly, all branch lengths were set to 0.002, 0.02, 0.2 or 2.0. Branch lengths were then estimated with HKY85 for use in the model fitting (see Methods §5.6.1). Input codon probabilities were equal (1/61 for sense codons), HIV-1 or human, from the Codon Usage database (see above). The gene 2's input ω was either 0.7 or 1.7 and input $\kappa = 2.0$. We tested 5 overlap proportions from $\frac{1}{6}$ (149 nucleotide sites) to $\frac{6}{6}$ (901 nucleotide sites), 4 branch lengths ($\{0.002, 0.02, 0.2, 2.0\}$) or 9 ω values for gene 1 ($\{0.0, 0.25, 0.5, 0.75, 1.0, 1.5, 2.0, 2.5, 3.0\}$). We generated 5 repeat datasets for each combination of parameter input values.

5.6.3.2 Testing Mixture Models

Synthetic datasets were generated with the vLSD simulation with varied ω applied across sites. 64 sequences were evolved along a bifurcating tree with each branch length equal to 0.02 substitutions per site, as for the non-mixture models. Two genes were present in frames A and B respectively, each comprising 300 codons, with 824 nucleotide sequences in the overlap region and equal lengths without overlap for each gene. Gene 2 (in frame B) had all probability density on the negative selection class ($p_0 = 1$), with $\omega_0 = 0.4$. For gene 1, three site class distributions were used (table 5.5). Gene 1 $\omega_0 = 0.4$. We also used each set of codon probabilities tested previously: equal (1/61), HIV-1 or human probabilities. For each of these $3(3 + [4 \times 2]) = 33$ conditions, we simulated five separate datasets with the same parameters.

5.6.4 Parametric Bootstrapping

Bootstrap datasets were generated with the vLSD simulation using an identical genetic layout as the original simulation (two genes with the same amount of overlap), the branch length and κ values estimated by HKY85 and the MLEs from the null model estimated from the original datasets. The large number of datasets precluded using multiple different start values when fitting models, so the start values used for each set of bootstrap datasets was the start values yielding the highest log-likelihood when analysing the original data.

Appendices

Appendix A

SAMHD1: List of Species and Sequence Accession Numbers

Table A.1: Species represented in the SAMHD1 analyses described and Genbank accession numbers for DNA sequences. The Tasmanian devil sequence was found to be split between two records (see SAMHD1 methods). M., mammals; *Cet.*, *Cetartiodactyla*; *Pr.*, *Primates*.

Analysis group	Common name	Binomial name	Sequence accession
M. *Carnivora*	amur tiger	*Panthera tigris altaica*	XM_007077693.1
M. *Carnivora*	cat	*Felis catus*	XM_003983547.2
M. *Carnivora*	dog	*Canis lupus familiaris*	XM_542986.4
M. *Carnivora*	ferret	*Mustela putorius furo*	XM_004746473.1
M. *Carnivora*	giant panda	*Ailuropoda melanoleuca*	XM_002915170.2
M. *Carnivora*	polar bear	*Ursus maritimus*	XM_008698122.1
M. *Carnivora*	walrus	*Odobenus rosmarus divergens*	XM_004393158.1
M. *Carnivora*	weddell seal	*Leptonychotes weddellii*	XM_006733646.1
M. *Cet.*	alpaca	*Vicugna pacos*	XM_006202728.1
M. *Cet.*	bactrian camel	*Camelus bactrianus*	XM_010960760.1
M. *Cet.*	baiji	*Lipotes vexillifer*	XM_007447732.1
M. *Cet.*	common bottlenose dolphin	*Tursiops truncatus*	XM_004326928.1
M. *Cet.*	cow	*Bos taurus*	NM_001075861.1
M. *Cet.*	dromedary camel	*Camelus dromedarius*	XM_010975213.1
M. *Cet.*	goat	*Capra hircus*	XM_013968902.1
M. *Cet.*	minke whale	*Balaenoptera acutorostrata scammoni*	XM_007193161.1
M. *Cet.*	mouflon or sheep	*Ovis aries musimon*	XM_012114778.1
M. *Cet.*	orca	*Orcinus orca*	XM_004272978.2
M. *Cet.*	plains or american bison	*Bison bison bison*	XM_010829204.1
M. *Cet.*	sheep	*Ovis aries*	XM_012189160.1
M. *Cet.*	sperm whale	*Physeter catodon*	XM_007116777.1
M. *Cet.*	tibetan antelope	*Pantholops hodgsonii*	XM_005968060.1
M. *Cet.*	water buffalo	*Bubalus bubalis*	XM_006051717.1
M. *Cet.*	wild bactrian camel	*Camelus ferus*	XM_006188650.1
M. *Cet.*	wild boar	*Sus scrofa*	NM_001292105.1
M. *Cet.*	yak	*Bos mutus*	XM_005896009.1
M. *Chiroptera*	big brown bat	*Eptesicus fuscus*	XM_008149758.1
M. *Chiroptera*	black flying fox	*Pteropus alecto*	XM_006921792.1
M. *Chiroptera*	brandts bat	*Myotis brandtii*	XM_005868463.1
M. *Chiroptera*	davids myotis bat	*Myotis davidii*	XM_006764508.1
M. *Chiroptera*	large flying fox	*Pteropus vampyrus*	XM_011366586.1
M. *Chiroptera*	little brown bat	*Myotis lucifugus*	XM_006089575.1
M. *Glires*	brown rat	*Rattus norvegicus*	NM_001191743.1
M. *Glires*	chinese hamster	*Cricetulus griseus*	XM_007625631.1
M. *Glires*	damara mole rat	*Fukomys damarensis*	XM_010640022.1
M. *Glires*	deer mouse	*Peromyscus maniculatus bairdii*	XM_006984164.1
M. *Glires*	degu	*Octodon degus*	XM_004630995.1
M. *Glires*	european rabbit	*Oryctolagus cuniculus*	XM_008256128.1
M. *Glires*	golden hamster	*Mesocricetus auratus*	XM_005084653.1
M. *Glires*	guinea pig	*Cavia porcellus*	XM_003467744.2
M. *Glires*	house mouse	*Mus musculus*	NM_018851.3
M. *Glires*	lesser egyptian jerboa	*Jaculus jaculus*	XM_004666629.1
M. *Glires*	long tailed chinchilla	*Chinchilla lanigera*	XM_005384981.1

M. *Glires*	naked mole rat	*Heterocephalus glaber*	XM_004888567.1
M. *Glires*	pika	*Ochotona princeps*	XM_004586040.1
M. *Glires*	prairie vole	*Microtus ochrogaster*	XM_005363152.1
M. *Glires*	thirteen lined ground squirrel	*Spermophilus tridecemlineatus*	XM_005329859.1
M. *Glires*	upper galilee mountains mole rat	*Nannospalax galili*	XM_008836546.1
M. *Pr.*	agile gibbon	*Hylobates agilis*	JQ231127.1
M. *Pr.*	allens swamp monkey	*Allenopithecus nigroviridis*	JN936900.1
M. *Pr.*	angola colobus	*Colobus angolensis palliatus*	JN936905.1
M. *Pr.*	angolan talapoin	*Miopithecus talapoin*	JN936901.1
M. *Pr.*	black capped squirrel monkey	*Saimiri boliviensis boliviensis*	XM_010342629.1
M. *Pr.*	black white ruffed lemur	*Varecia variegata variegata*	JN936913.1
M. *Pr.*	bonobo	*Pan paniscus*	NM_001279186.1
M. *Pr.*	bornean orangutan	*Pongo pygmaeus*	JN936888.1
M. *Pr.*	brown woolly monkey	*Lagothrix lagotricha*	JQ231150.1
M. *Pr.*	chimp	*Pan troglodytes*	NM_001280510.1
M. *Pr.*	collared mangabey	*Cercocebus torquatus*	JQ231133.1
M. *Pr.*	colobus	*Colobus guereza*	JQ231145.1
M. *Pr.*	common marmoset	*Callithrix jacchus*	JN936906.1
M. *Pr.*	common squirrel monkey	*Saimiri sciureus*	JN936909.1
M. *Pr.*	cotton tamarin	*Saguinus oedipus*	JN936908.1
M. *Pr.*	crab eating macaque	*Macaca fascicularis*	NM_001287721.1
M. *Pr.*	de brazzas monkey	*Cercopithecus neglectus*	JQ231141.1
M. *Pr.*	diana monkey	*Cercopithecus diana*	JN936902.1
M. *Pr.*	drill	*Mandrillus leucophaeus*	JQ231131.1
M. *Pr.*	francois langur	*Trachypithecus francoisi*	JN936904.1
M. *Pr.*	gelada	*Theropithecus gelada*	JN936896.1
M. *Pr.*	geoffroys spider monkey	*Ateles geoffroyi*	JN936911.1
M. *Pr.*	golden bellied mangabey	*Cercocebus chrysogaster*	JN936898.1
M. *Pr.*	golden snub nosed monkey	*Rhinopithecus roxellana*	XM_010355318.1
M. *Pr.*	gorilla	*Gorilla gorilla*	NM_001279619.1
M. *Pr.*	gray mouse lemur	*Microcebus murinus*	JN936914.1
M. *Pr.*	green monkey	*Chlorocebus sabaeus*	NM_001292080.1
M. *Pr.*	grivet	*Chlorocebus aethiops*	KF741041.1
M. *Pr.*	hamadryas baboon	*Papio hamadryas*	JN936890.1
M. *Pr.*	human	*Homo sapiens*	NM_015474.3
M. *Pr.*	lar gibbon	*Hylobates lar*	JN936889.1
M. *Pr.*	mandrill	*Mandrillus sphinx*	JN936897.1
M. *Pr.*	mantled howler	*Alouatta palliata*	JN936912.1
M. *Pr.*	nancy mas night monkey	*Aotus nancymaae*	XM_012455058.1
M. *Pr.*	northen greater galago	*Otolemur garnettii*	XM_003788255.1
M. *Pr.*	olive baboon	*Papio anubis*	NM_001279525.1
M. *Pr.*	patas monkey	*Erythrocebus patas*	JQ231138.1
M. *Pr.*	philippine tarsier	*Tarsius syrichta*	XM_008049665.1
M. *Pr.*	proboscis monkey	*Nasalis larvatus*	JQ231144.1
M. *Pr.*	pygmy marmoset	*Callithrix pygmaea*	JQ231146.1
M. *Pr.*	red bellied titi	*Callicebus moloch*	JQ231152.1
M. *Pr.*	red shanked douc	*Pygathrix nemaeus nemaeus*	JN936903.1
M. *Pr.*	rhesus macaque	*Macaca mulatta*	JQ231135.1
M. *Pr.*	siamang	*Symphalangus syndactylus*	JQ231128.1
M. *Pr.*	sooty mangabey	*Cercocebus atys*	JQ231132.1
M. *Pr.*	southern pig tailed macaque	*Macaca nemestrina*	XM_011767026.1
M. *Pr.*	sumatran orangutan	*Pongo abelii*	XM_002830274.3
M. *Pr.*	tantalus monkey	*Chlorocebus tantalus*	JN936892.1
M. *Pr.*	three striped night monkey	*Aotus trivirgatus*	JN936907.1
M. *Pr.*	tufted capuchin	*Cebus apella*	JN936910.1
M. *Pr.*	vervet monkey	*Chlorocebus pygerythrus*	JQ231137.1
M. *Pr.*	white cheeked gibbon	*Nomascus leucogenys*	NM_001280119.1
M. *Pr.*	white faced saki	*Pithecia pithecia*	JQ231151.1
M. *Pr.*	white lipped tamarin	*Saguinus labiatus*	JQ231147.1
M. *Pr.*	wolfs mona monkey	*Cercopithecus wolfi*	JQ231140.1
M. Remaining	aardvark	*Orycteropus afer afer*	XM_007934742.1
M. Remaining	african bush elephant	*Loxodonta africana*	XM_003411700.2
M. Remaining	armadillo	*Dasypus novemcinctus*	XM_004466604.2
M. Remaining	cape elephant shrew	*Elephantulus edwardii*	XM_006881709.1
M. Remaining	cape golden mole	*Chrysochloris asiatica*	XM_006875733.1
M. Remaining	common shrew	*Sorex araneus*	XM_004612642.1
M. Remaining	european hedgehog	*Erinaceus europaeus*	XM_007518471.1
M. Remaining	gray short tailed opposum	*Monodelphis domestica*	XM_001381548.3
M. Remaining	horse	*Equus caballus*	XM_001499498.3
M. Remaining	lesser hedgehog tenrec	*Echinops telfairi*	XM_004698253.1
M. Remaining	przewalskis horse	*Equus przewalskii*	XM_008541885.1
M. Remaining	star nosed mole	*Condylura cristata*	XM_004686953.1
M. Remaining	sunda flying lemur	*Galeopterus variegatus*	XM_008575366.1
M. Remaining	tasmanian devil (5')	*Sarcophilus harrisii*	XM_003758997.2
”	tasmanian devil (3')	”	XM_012553363.1
M. Remaining	treeshrew	*Tupaia chinensis*	XM_006144103.1
M. Remaining	west indian manatee	*Trichechus manatus latirostris*	XM_004370368.1
M. Remaining	white rhinoceros	*Ceratotherium simum simum*	XM_004430428.1
Birds	mallard	*Anas platyrhynchos*	XM_005018014.1
Birds	chimney swift	*Chaetura pelagica*	XM_009999860.1
Birds	rhinoceros hornbill	*Buceros rhinoceros silvestris*	XM_010146105.1
Birds	chuck wills widow	*Caprimulgus carolinensis*	XM_010173576.1
Birds	killdeer	*Charadrius vociferus*	XM_009886339.1
Birds	yellow throated sandgrouse	*Pterocles gutturalis*	XM_010087934.1
Birds	speckled mousebird	*Colius striatus*	XM_010208317.1

Birds	rock dove	*Columba livia*	XM_005512218.1
Birds	cuckoo roller	*Leptosomus discolor*	XM_009957655.1
Birds	northern carmine beeeater	*Merops nubicus*	XM_008938233.1
Birds	common cuckoo	*Cuculus canorus*	XM_009565365.1
Birds	golden eagle	*Aquila chrysaetos canadensis*	XM_011597780.1
Birds	white tailed eagle	*Haliaeetus albicilla*	XM_009919832.1
Birds	bald eagle	*Haliaeetus leucocephalus*	XM_010564547.1
Birds	peregrine falcon	*Falco peregrinus*	XM_005229325.1
Birds	chicken	*Gallus gallus*	NM_001030845.1
Birds	wild turkey	*Meleagris gallopavo*	XM_010722256.1
Birds	red throated diver	*Gavia stellata*	XM_009822020.1
Birds	red legged seriema	*Cariama cristata*	XM_009702527.1
Birds	sunbittern	*Eurypyga helias*	XM_010157504.1
Birds	grey crowned crane	*Balearica regulorum gibbericeps*	XM_010310362.1
Birds	brown mesite	*Mesitornis unicolor*	XM_010190900.1
Birds	macqueens bustard	*Chlamydotis macqueenii*	XM_010128452.1
Birds	red crested turaco	*Tauraco erythrolophus*	XM_009985674.1
Birds	hoatzin	*Opisthocomus hoazin*	XM_009935818.1
Birds	rifleman	*Acanthisitta chloris*	XM_009076070.1
Birds	american crow	*Corvus brachyrhynchos*	XM_008643497.1
Birds	hooded crow	*Corvus cornix cornix*	XM_010413192.1
Birds	zebra finch	*Taeniopygia guttata*	XM_002195540.2
Birds	atlantic canary	*Serinus canaria*	XM_009098810.1
Birds	white throated sparrow	*Zonotrichia albicollis*	XM_005495259.1
Birds	collared flycatcher	*Ficedula albicollis*	XM_005056959.1
Birds	ground tit	*Pseudopodoces humilis*	XM_005532216.1
Birds	golden collared manakin	*Manacus vitellinus*	XM_008934650.1
Birds	medium ground finch	*Geospiza fortis*	XM_005428399.1
Birds	little egret	*Egretta garzetta*	XM_009637821.1
Birds	dalmatian pelican	*Pelecanus crispus*	XM_009486761.1
Birds	white tailed tropicbird	*Phaethon lepturus*	XM_010296384.1
Birds	great cormorant	*Phalacrocorax carbo*	XM_009514424.1
Birds	crested ibis	*Nipponia nippon*	XM_009473373.1
Birds	downy woodpecker	*Picoides pubescens*	XM_009903354.1
Birds	northen fulmar	*Fulmarus glacialis*	XM_009575330.1
Birds	kia	*Nestor notabilis*	XM_010010583.1
Birds	budgerigar	*Melopsittacus undulatus*	XM_005146833.1
Birds	emperor penguin	*Aptenodytes forsteri*	XM_009283269.1
Birds	adelie penguin	*Pygoscelis adeliae*	XM_009329356.1
Birds	barn owl	*Tyto alba*	XM_009970142.1
Birds	ostrich	*Struthio camelus australis*	XM_009679283.1
Birds	white throated tinamou	*Tinamus guttatus*	XM_010212581.1
Birds	annas hummingbird	*Calypte anna*	XM_008506368.1
Birds	bar tailed trogon	*Apaloderma vittatum*	XM_009878222.1
Reptiles	american alligator	*Alligator mississippiensis*	XM_006278901.1
Reptiles	chinese alligator	*Alligator sinensis*	XM_006034284.1
Reptiles	carolina anole	*Anolis carolinensis*	XM_003220542.2
Reptiles	burmese python	*Python bivittatus*	XM_007419878.1
Reptiles	green sea turtle	*Chelonia mydas*	XM_007053931.1
Reptiles	painted turtle	*Chrysemys picta bellii*	XM_005295427.2
Reptiles	chinese softshell turtle	*Pelodiscus sinensis*	XM_006113201.1
Amphibians	african clawed frog	*Xenopus laevis*	BC072238.1
Amphibians	western clawed frog	*Xenopus tropicalis*	XM_002932885.2
Fish	japanese rice fish	*Oryzias latipes*	XM_004070493.2
Fish	mexican tetra	*Astyanax mexicanus*	XM_007239426.1
Fish	astatotilapia burtoni cichlid	*Haplochromis burtoni*	XM_005940268.1
Fish	zebra mbuna cichlid	*Maylandia zebra*	XM_004546443.1
Fish	neolamprologus brichardi cichlid	*Neolamprologus brichardi*	XM_006804309.1
Fish	haplochromis nyererei cichlid	*Pundamilia nyererei*	XM_005740532.1
Fish	zebrafish	*Danio rerio*	NM_001159933.1
Fish	amazon molly	*Poecilia formosa*	XM_007544174.1
Fish	guppy	*Poecilia reticulata*	XM_008413867.1
Fish	southern platyfish	*Xiphophorus maculatus*	XM_005811227.1
Fish	northern pike	*Esox lucius*	XM_010868248.1
Fish	bicolor damselfish	*Stegastes partitus*	XM_008293652.1
Fish	large yellow croaker	*Larimichthys crocea*	XM_010741581.1
Fish	tonguefish	*Cynoglossus semilaevis*	XM_008319362.1
Fish	spotted gar	*Lepisosteus oculatus*	XM_006639574.1
Fish	japanese pufferfish	*Takifugu rubripes*	XM_011621666.1
Fish	australian ghostshark	*Callorhinchus milii*	XM_007908660.1
Fish	lancelet	*Branchiostoma floridae*	XM_002604911.1
Fish	west indian ocean coelacanth	*Latimeria chalumnae*	XM_006003809.1

Appendix B

SAMHD1: Log-Likelihoods, Test Statistics and p Values

Dataset	**Top.**	**Null**	**Alt.**	**Null ℓ**	**Alt. ℓ**	D	**d.f.**	p	**Codon freq.**
All mammals	ML	M1a	M2a	-39160.68	-38954.72	411.93	2	3.55E-90	F1X4
Mammals ex. primates	ML	M1a	M2a	-33185.82	-33034.90	301.84	2	2.86E-66	F1X4
$T1$ (mammals)	T_1	M1a	M2a	-39239.19	-39029.14	420.10	2	5.98E-92	F1X4
$T2$ (mammals)	T_2	M1a	M2a	-39233.30	-39035.09	396.42	2	8.29E-87	F1X4
$T3$ (mammals)	T_3	M1a	M2a	-39176.35	-38971.45	409.80	2	1.03E-89	F1X4
Carnivora	ML	M1a	M2a	-4683.41	-4672.52	21.78	2	1.86E-05	F3X4
Chiroptera	ML	M1a	M2a	-4733.32	-4716.64	33.36	2	5.70E-08	F1X4
Glires	ML	M1a	M2a	-12913.41	-12901.97	22.88	2	1.08E-05	F1X4
Primates	ML	M1a	M2a	-8298.97	-8282.41	33.12	2	6.42E-08	F1X4
Cetartiodactyla	ML	M1a	M2a	-5957.73	-5928.79	57.87	2	2.72E-13	F1X4
Remaining mammals	ML	M1a	M2a	-13740.77	-13723.18	35.17	2	2.31E-08	F1X4
Carnivora	ML	M7	M8	-4684.21	-4672.55	23.32	2	8.62E-06	F3X4
Chiroptera	ML	M7	M8	-4704.80	-4666.78	76.02	2	3.10E-17	F3X4
Glires	ML	M7	M8	-12902.26	-12884.58	35.36	2	2.09E-08	F1X4
Primates	ML	M7	M8	-8280.43	-8242.52	75.83	2	3.42E-17	F3X4
Cetartiodactyla	ML	M7	M8	-5936.35	-5890.21	92.27	2	9.19E-21	F3X4
Remaining mammals	ML	M7	M8	-13672.56	-13613.78	117.57	2	2.94E-26	F1X4
Carnivora f.g.	ML	BS ($\omega = 1$)	BS	-39178.68	-39170.47	16.42	1	5.06E-05	F1X4
Chiroptera f.g.	ML	BS ($\omega = 1$)	BS	-39170.64	-39151.43	38.43	1	5.67E-10	F1X4
Glires f.g.	ML	BS ($\omega = 1$)	BS	-39170.88	-39115.75	8.54	1	8.54E-26	F1X4
Primates f.g.	ML	BS ($\omega = 1$)	BS	-39165.57	-39151.13	28.88	1	7.71E-08	F1X4
Cetartiodactyla f.g.	ML	BS ($\omega = 1$)	BS	-39172.92	-39139.65	3.44	1	3.44E-16	F1X4
Remaining mammals f.g.	ML	BS ($\omega = 1$)	BS	-39143.62	-39127.72	31.80	1	1.71E-08	F1X4

Table B.1: Log-likelihood values, test statistics and p values for analyses of mammalian SAMHD1. Each model fitting was done with 5 different initial parameter values (see methods) and the best fitting models are shown. Abbreviations: Top., tree topology; **ML**, maximum likelihood tree topology; T_i, alternative topology; **Null/Alt.**, null/alternative model; ℓ, log-likelihood; D, test statistic for likelihood ratio test; **d.f.**, degrees of freedom for chi square distribution (the difference in number of free parameters between null and alternative models); **codon freq.**, model of codon frequency used (see methods); **BS**, branch-site specific model with $\omega \geq 1$ on foreground branches; **BS** ($\omega = 1$), branch-site model with ω fixed at 1 on foreground branches; **N/A**, not applicable; **f.g.**, branches foreground.

Dataset	Top.	Null	Alt.	Null ℓ	Alt. ℓ	D	d.f.	p	Codon freq.
All chordates	ML	M1a	M2a	-79753.78	-79003.87	1499.83	2	0.00E+00	F1X4
Non-mammal chordates	ML	M1a	M2a	-41835.10	-41373.66	922.89	2	3.95E-201	F1X4
Birds	ML	M1a	M2a	-17309.67	-17232.54	154.25	2	3.20E-34	F1X4
Reptiles	ML	M1a	M2a	-6596.07	-6590.81	10.51	2	0.00522886	F3X4
Amphibians	ML	M1a	M2a	-3146.08	-3145.21	1.74	2	0.419378678	F3X4
Fish	ML	M1a	M2a	-18076.91	-17942.88	268.06	2	6.17E-59	F1X4
Birds	ML	M7	M8	-17241.20	-17241.20	0.00	2	1.00	F1X4
Reptiles	ML	M7	M8	-6590.51	-6586.12	8.77	2	0.012454728	F3X4
Amphibians	ML	M7	M8	-3146.31	-3145.20	1.74	2	0.419378678	F3X4
Fish	ML	M7	M8	-17852.35	No convergence	N/A	2	N/A	F1X4
Mammals f.g.	ML	BS ($\omega = 1$)	BS	-79679.34	-79517.17	324.34	1	1.64E-72	F1X4
Birds f.g.	ML	BS ($\omega = 1$)	BS	-79686.80	-79559.98	253.65	1	4.15E-57	F1X4
Reptiles f.g.	ML	BS ($\omega = 1$)	BS	-79766.32	-79762.18	8.28	1	0.004000139	F1X4
Amphibians f.g.	ML	BS ($\omega = 1$)	BS	-79784.43	-79753.66	61.53	1	4.36E-15	F1X4
Fish f.g.	ML	BS ($\omega = 1$)	BS	-79600.92	-79457.17	287.49	1	1.75E-64	F1X4
All chordates	T_1	M1a	M2a	-79913.61	-79135.79	1555.65	2	0.00	F1X4
Mammals f.g.	T_1	BS ($\omega = 1$)	BS	-79838.07	-79604.54	467.06	1	1.40E-103	F1X4
Birds f.g.	T_1	BS ($\omega = 1$)	BS	-79829.34	-79711.88	234.91	1	5.07E-53	F1X4
Reptiles f.g.	T_1	BS ($\omega = 1$)	BS	-79926.85	-79924.25	5.19	1	0.022725348	F1X4
Amphibians f.g.	T_1	BS ($\omega = 1$)	BS	-79944.48	-79914.77	59.43	1	1.27E-14	F1X4
Fish f.g.	T_1	BS ($\omega = 1$)	BS	-79758.01	-79604.50	307.02	1	9.75E-69	F1X4
All chordates	T_2	M1a	M2a	-79945.95	-79175.64	318243.18	2	0.00	F1X4
Mammals f.g.	T_2	BS ($\omega = 1$)	BS	-79865.57	-79714.20	302.73	1	8.36E-68	F1X4
Birds f.g.	T_2	BS ($\omega = 1$)	BS	-79857.52	-79742.58	229.90	1	6.27E-52	F1X4
Reptiles f.g.	T_2	BS ($\omega = 1$)	BS	-79958.27	-79953.90	8.68	1	0.003214404	F1X4
Amphibians f.g.	T_2	BS ($\omega = 1$)	BS	-79976.10	-79946.22	59.76	1	1.07E-14	F1X4
Fish f.g.	T_2	BS ($\omega = 1$)	BS	-79792.71	-79672.00	241.42	1	1.93E-54	F1X4
All chordates	T_3	M1a	M2a	-79893.35	-79123.37	1539.96	2	0.00	F1X4
Mammals f.g.	T_3	BS ($\omega = 1$)	BS	-79819.68	-79663.34	312.68	1	5.69E-70	F1X4
Birds f.g.	T_3	BS ($\omega = 1$)	BS	-79808.87	-79697.46	222.84	1	2.18E-50	F1X4
Reptiles f.g.	T_3	BS ($\omega = 1$)	BS	-79906.59	-79902.76	7.66	1	0.005637377	F1X4
Amphibians f.g.	T_3	BS ($\omega = 1$)	BS	-79922.89	-79892.60	60.59	1	7.04E-15	F1X4
Fish f.g.	T_3	BS ($\omega = 1$)	BS	-79737.18	-79593.12	288.13	1	1.27E-64	F1X4

Table B.2: Log-likelihood values, test statistics and p values for analyses of chordate SAMHD1 ('master' dataset). Each model fitting was done with 5 different initial parameter values (see methods) and the best fitting models are shown. See table B.1 for key to abbreviations.

Appendix C

SAMHD1: Annotated Mammal Trees

Figure C.1: Maximum likelihood tree for mammalian SAMDH1, coloured by residues in each sequence at site 566. Topology and node support identical to figure 2.5. Sequence names contain: analysis group prefix, species common name, sequence accession and letter code for residue. Group prefixes: 'pr', *Primates*; 'gl', *Glires*; 'wu', *Cetartiodactyla*; 'cn', *Carnivora*; 'cp', *Chiroptera*; 'ma', remaining mammals. Branch lengths are nucleotide substitutions per site.

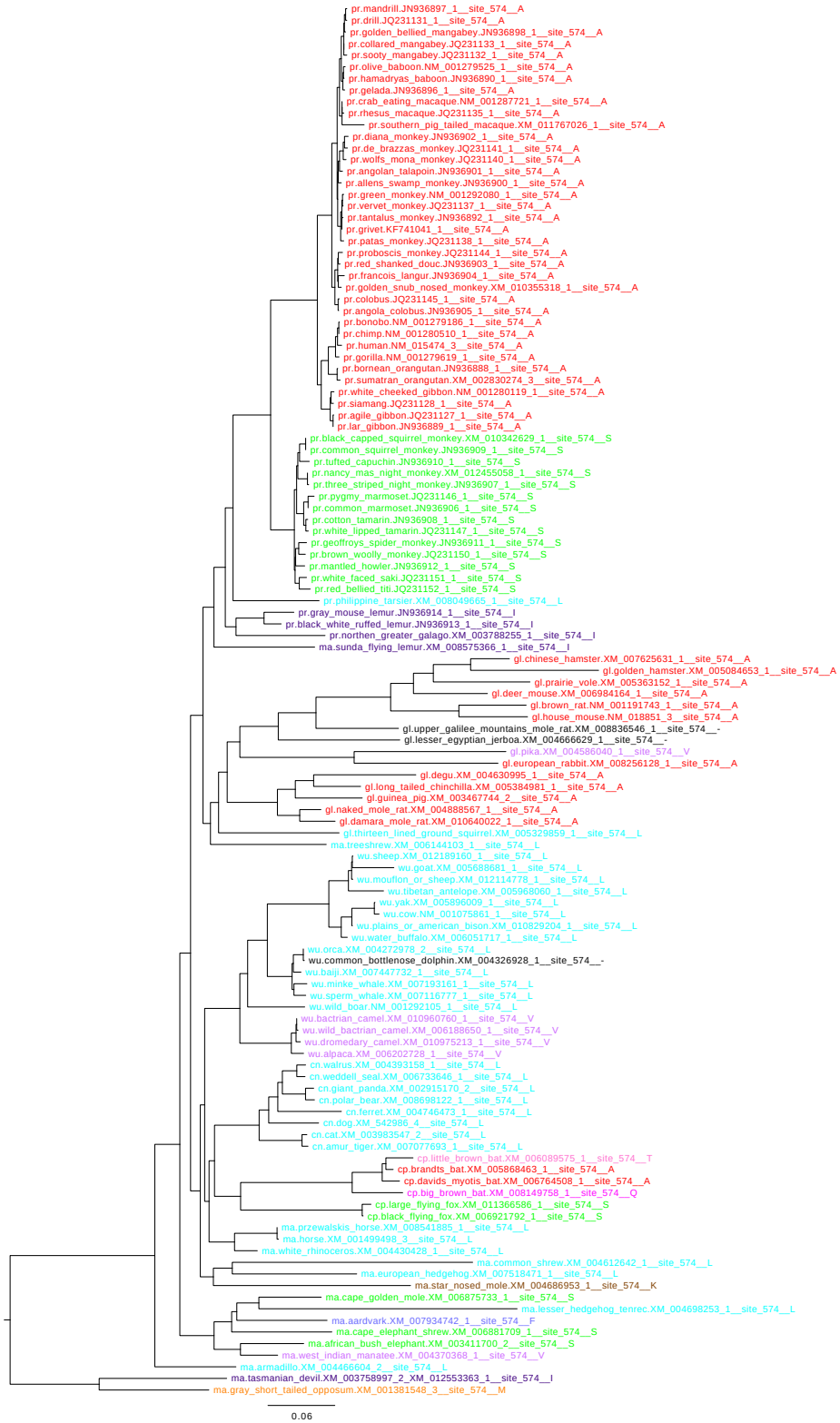

Figure C.2: The same tree as in C.1, showing residues for site 574.

Figure C.3: The same tree as in C.1, showing residues for site 594.

Figure C.4: The same tree as in C.1, showing residues for site 596.

Appendix D

Capsid: swMutsel and M8 Statistics

CA Site	Null lnL	Alt. lnL	D	N	p (LRT)	p Adj.	M8 BEB if > 0.5
5	-382.427	-369.912	25.029	5	$< 10^{-3}$	0.001	-
13	-637.358	-626.377	21.963	7	0.001	0.014	-
14	-1703.045	-1679.461	47.168	9	$< 10^{-3}$	$< 10^{-3}$	1.0
15	-1556.981	-1535.878	42.207	9	$< 10^{-3}$	$< 10^{-3}$	-
31	-914.086	-900.244	27.682	7	$< 10^{-3}$	0.002	-
41	-809.025	-797.965	22.119	6	$< 10^{-3}$	0.007	-
47	-306.044	-297.577	16.934	6	0.005	0.045	
50	-826.978	-824.376	5.204	11	0.877	0.999	1.0
58	-535.445	-520.496	29.897	7	$< 10^{-3}$	0.001	-
68	-161.088	-147.145	27.885	4	$< 10^{-3}$	$< 10^{-3}$	-
83	-1668.036	-1652.526	31.019	10	$< 10^{-3}$	0.004	1.0
86	-1352.182	-1334.430	35.504	10	$< 10^{-3}$	0.001	-
87	-949.131	-926.638	44.988	6	$< 10^{-3}$	$< 10^{-3}$	-
91	-1766.896	-1746.538	40.715	14	$< 10^{-3}$	0.002	1.0
94	-445.594	-436.168	18.852	5	$< 10^{-3}$	0.010	-
98	-682.367	-673.707	17.319	5	0.002	0.018	-
110	-852.800	-843.726	18.149	6	0.003	0.028	-
116	-1379.594	-1359.572	40.043	13	$< 10^{-3}$	0.001	1.0
120	n/a	n/a	n/a	10	n/a	n/a	1.0
131	-294.859	-284.180	21.358	5	$< 10^{-3}$	0.004	-
141	-111.847	-102.098	19.497	3	$< 10^{-3}$	0.001	-
171	-960.372	-942.834	35.075	6	$< 10^{-3}$	$< 10^{-3}$	-
177	-638.185	-623.433	29.504	5	$< 10^{-3}$	$< 10^{-3}$	-
194	-1179.364	-1169.241	20.246	5	$< 10^{-3}$	0.006	-
204	-420.406	-413.413	13.986	3	$< 10^{-3}$	0.011	-
225	-1564.469	-1562.007	4.924	6	0.425	0.999	1.0

Table D.1: Test statistics from selection analyses of HIV-1 M and SIVcpz capsid (CA) using maximum likelihood tree topology. Null and Alt. lnL, null and alternative swMutsel model log-likelihoods; D, test statistic for likelihood ratio test (LRT); N, number of residues observed at the site; p (LRT), p value from LRT with $N - 1$ degrees of freedom; p Adj., p values adjusted for multiple hypothesis testing by controlling false discovery rate to 5% (Benjamini and Hochberg, 1995); M8 BEB, probabilities computed by the model M8 Bayes empirical Bayes procedure for belonging to a positive selection site class. Shown are all sites found to be significant (p adj. < 0.05) with swMutsel or with BEB probability > 0.5 with M8.

Appendix E

Accessory Genes: swMutsel and M8 Statistics

Nef site	Null lnL	Alt. lnL	N	D	p	p Adj.	M8 BEB
14	-1634.171	-1601.654	13	65.033	$< 10^{-3}$	$< 10^{-3}$	1.0
15	-2437.847	-2425.835	17	24.026	0.089	0.236	1.0
16	-1184.301	-1175.739	4	17.124	$< 10^{-3}$	0.005	-
21	-1492.448	-1470.787	15	43.322	$< 10^{-3}$	$< 10^{-3}$	-
24	-1903.716	-1885.054	13	37.324	$< 10^{-3}$	0.002	1.0
39	-1486.826	-1479.344	10	14.964	0.092	0.24	1.0
45	-1182.828	-1170.461	9	24.733	0.002	0.011	-
47	-534.094	-521.987	8	24.213	0.001	0.008	-
48	-794.783	-781.428	12	26.711	0.005	0.029	-
49	-1825.979	-1816.103	11	19.753	0.032	0.108	1.0
50	-2096.867	-2082.195	16	29.343	0.015	0.068	1.0
53	-1042.444	-1023.316	12	38.257	$< 10^{-3}$	$< 10^{-3}$	-
54	-1303.755	-1249.834	9	107.842	$< 10^{-3}$	$< 10^{-3}$	-
76	-865.779	-854.694	8	22.171	0.002	0.014	-
83	-1496.943	-1492.159	12	9.568	0.57	0.871	1.0
85	-2683.493	-2655.699	13	55.588	$< 10^{-3}$	$< 10^{-3}$	1.0
87	-795.247	-786.326	5	17.841	0.001	0.009	-
100	-853.279	-836.8	4	32.957	$< 10^{-3}$	$< 10^{-3}$	-
102	-1297.009	-1297.009	6	0	1	1	1.0
113	-14.069	-10.229	2	7.681	0.006	0.031	-
116	-910.114	-896.945	5	26.339	$< 10^{-3}$	$< 10^{-3}$	-
120	-1379.792	-1372.918	6	13.75	0.017	0.077	1.0
125	-767.065	-745.765	6	42.6	$< 10^{-3}$	$< 10^{-3}$	-
126	-734.069	-713.058	7	42.021	$< 10^{-3}$	$< 10^{-3}$	-
133	-1814.695	-1793.588	7	42.215	$< 10^{-3}$	$< 10^{-3}$	1.0
135	-1678.934	-1678.934	7	0	1	1	1.0
137	-598.33	-590.148	4	16.364	$< 10^{-3}$	0.007	-
138	-492.009	-471.894	5	40.23	$< 10^{-3}$	$< 10^{-3}$	-
142	-404.903	-382.43	5	44.947	$< 10^{-3}$	$< 10^{-3}$	-
143	-752.225	-739.09	6	26.27	$< 10^{-3}$	$< 10^{-3}$	-
151	-1936.835	-1921.871	13	29.929	0.003	0.017	1.0
158	-1541.862	-1533.748	13	16.228	0.181	0.382	1.0
163	-1911.919	-1891.606	16	40.626	$< 10^{-3}$	0.003	-
168	-1465.518	-1451.924	14	27.189	0.012	0.061	1.0
169	-972.743	-959.93	7	25.626	$< 10^{-3}$	0.002	-
173	-1150.187	-1134.769	13	30.835	0.002	0.013	0.999
176	-2039.646	-2026.724	14	25.843	0.018	0.077	1.0
178	-1329.437	-1325.312	6	8.25	0.143	0.33	1.0
180	-882.347	-864.55	8	35.595	$< 10^{-3}$	$< 10^{-3}$	-
182	-2248.378	-2227.062	10	42.633	$< 10^{-3}$	$< 10^{-3}$	1.0
184	-1405.963	-1382.001	9	47.925	$< 10^{-3}$	$< 10^{-3}$	-
188	-1861.416	-1842.999	16	36.834	0.001	0.009	1.0
194	-2282.652	-2260.835	14	43.634	$< 10^{-3}$	$< 10^{-3}$	1.0
196	-1118.536	-1088.872	9	59.327	$< 10^{-3}$	$< 10^{-3}$	-
198	-2093.148	-2076.96	13	32.376	0.001	0.009	1.0
202	-1327.213	-1293.057	6	68.312	$< 10^{-3}$	$< 10^{-3}$	1.0

Table E.1: Test statistics from selection analyses of HIV-1 M and SIVcpz Nef using maximum likelihood tree topology. Null and Alt. lnL, null and alternative swMutSel model log-likelihoods; D, test statistic for likelihood ratio test (LRT); N, number of residues observed at the site; p, p value from LRT with $N-1$ degrees of freedom; p Adj., p values adjusted for multiple hypothesis testing by controlling false discovery rate to 5% (Benjamini and Hochberg, 1995); M8 BEB, probabilities computed by the model M8 Bayes empirical Bayes procedure for belonging to a positive selection site class. Shown are all sites found to be significant (p adj. < 0.05) with swMutSel or with BEB probability > 0.95 with M8.

Vpu Site	Null lnL	Alt. lnL	D	N	p (LRT)	p Adj.
6	-1983.587	-1970.257	26.660	12	0.005	0.025
10	-692.669	-679.756	25.826	9	0.001	0.011
14	-979.955	-965.636	28.638	11	0.001	0.012
15	-1724.487	-1712.124	24.727	11	0.006	0.025
23	-1480.134	-1469.130	22.007	9	0.005	0.025
25	-1021.676	-1012.731	17.890	7	0.007	0.025
27	-782.581	-766.084	32.994	10	$< 10^{-3}$	0.007
29	-345.664	-331.376	28.574	12	0.003	0.018
34	-1026.178	-1013.073	26.210	12	0.006	0.025
37	-928.667	-913.237	30.860	11	$< 10^{-3}$	0.010
42	-734.522	-719.880	29.283	9	$< 10^{-3}$	0.007
43	-1386.587	-1373.312	26.550	11	0.003	0.019
46	-1241.981	-1226.374	31.215	12	0.001	0.011

Table E.2: Test statistics from selection analyses of HIV-1 M and SIVcpz Vpu using maximum likelihood tree topology. Null and Alt. lnL, null and alternative swMutsel model log-likelihoods; D, test statistic for likelihood ratio test (LRT); N, number of residues observed at the site; p, p value from LRT with $N-1$ degrees of freedom; p Adj., p values adjusted for multiple hypothesis testing by controlling false discovery rate to 5% (Benjamini and Hochberg, 1995). Shown are all sites found to be significant (p adj. < 0.05) with swMutsel.

Vpr Site	Null lnL	Alt. lnL	D	N	p (LRT)	p Adj.
61	-946.812	-931.096	31.433	8	$< 10^{-3}$	0.002
63	-1743.501	-1721.757	43.489	11	$< 10^{-3}$	$< 10^{-3}$

Table E.3: Test statistics from selection analyses of HIV-1 M and SIVcpz Vpr using maximum likelihood tree topology. Null and Alt. lnL, null and alternative swMutsel model log-likelihoods; D, test statistic for likelihood ratio test (LRT); N, number of residues observed at the site; p, p value from LRT with $N-1$ degrees of freedom; p Adj., p values adjusted for multiple hypothesis testing by controlling false discovery rate to 5% (Benjamini and Hochberg, 1995). Shown are all sites found to be significant (p adj. < 0.05) with swMutsel.

Appendix F

Pentamer Probability

From Kimura (1962) and Iwasa (1988), we can describe the probability for a codon (nucleotide triplet) xyz as

$$P_{xzy} = \frac{(P^0_{xyz})(e^{F_{xyz}})}{\sum_{a,b,c}(P^0_{abc})(e^{F_{abc}})} \tag{F.1}$$

where F_{xyz} is the fitness of xyz scaled by the effective chromosomal number and P^0_{xyz} is its probability in the absence of selection. It follows that the probability of a nucleotide pentamer is

$$P_{ghikl} = \frac{(P^0_{ghikl})(e^{F_{ghikl}})}{\sum_{a,b,c,d,e}(P^0_{abcde})(e^{F_{abcde}})}. \tag{F.2}$$

We assume the fitnesses of each constituent codon are independent of one another. Therefore,

$$\begin{aligned} P_{ghikl} &= \frac{(P^0_{ghikl})(e^{F_{ghikl}})}{\sum_{a,b,c,d,e}(P^0_{abcde})(e^{F_{abcde}})} \\ &\approx \frac{(P^0_{ghikl})(e^{F_{ghi}+F_{hik}+F_{ikl}})}{\sum_{a,b,c,d,e}(P^0_{abcde})(e^{F_{abc}+F_{bcd}+F_{cde}})} \\ &= \frac{(P^0_{ghikl})(e^{F_{ghi}})(e^{F_{hik}})(e^{F_{ikl}})}{\sum_{a,b,c,d,e}(P^0_{abcde})(e^{F_{abc}})(e^{F_{bcd}})(e^{F_{cde}})} \end{aligned} \tag{F.3}$$

We also assume $P^0_{xyz} = \frac{1}{64}$ for all xyz, and therefore $(P^0_{ghi})(P^0_{hik})(P^0_{ikl}) \propto \frac{1}{64^3}$.

We then assume P_{xyz} is equivalent to the frequency G of xyz observed in large databases of codon sequences; i.e.

$$P_{xyz} \propto \frac{e^{F_{xyz}}}{\sum_{abc} e^{F_{abc}}} \equiv G_{xyz}. \tag{F.4}$$

Therefore,

$$P_{ghikl} \propto \frac{P_{ghi}P_{hik}P_{ikl}}{\sum_{a,b,c,d,e} P_{abc}P_{bcd}P_{cde}} \equiv \frac{G_{ghi}G_{hik}G_{ikl}}{\sum_{a,b,c,d,e} G_{abc}G_{bcd}G_{cde}}. \tag{F.5}$$

Appendix G

Dependence of ω Estimation on Divergence and Overlap Proportion

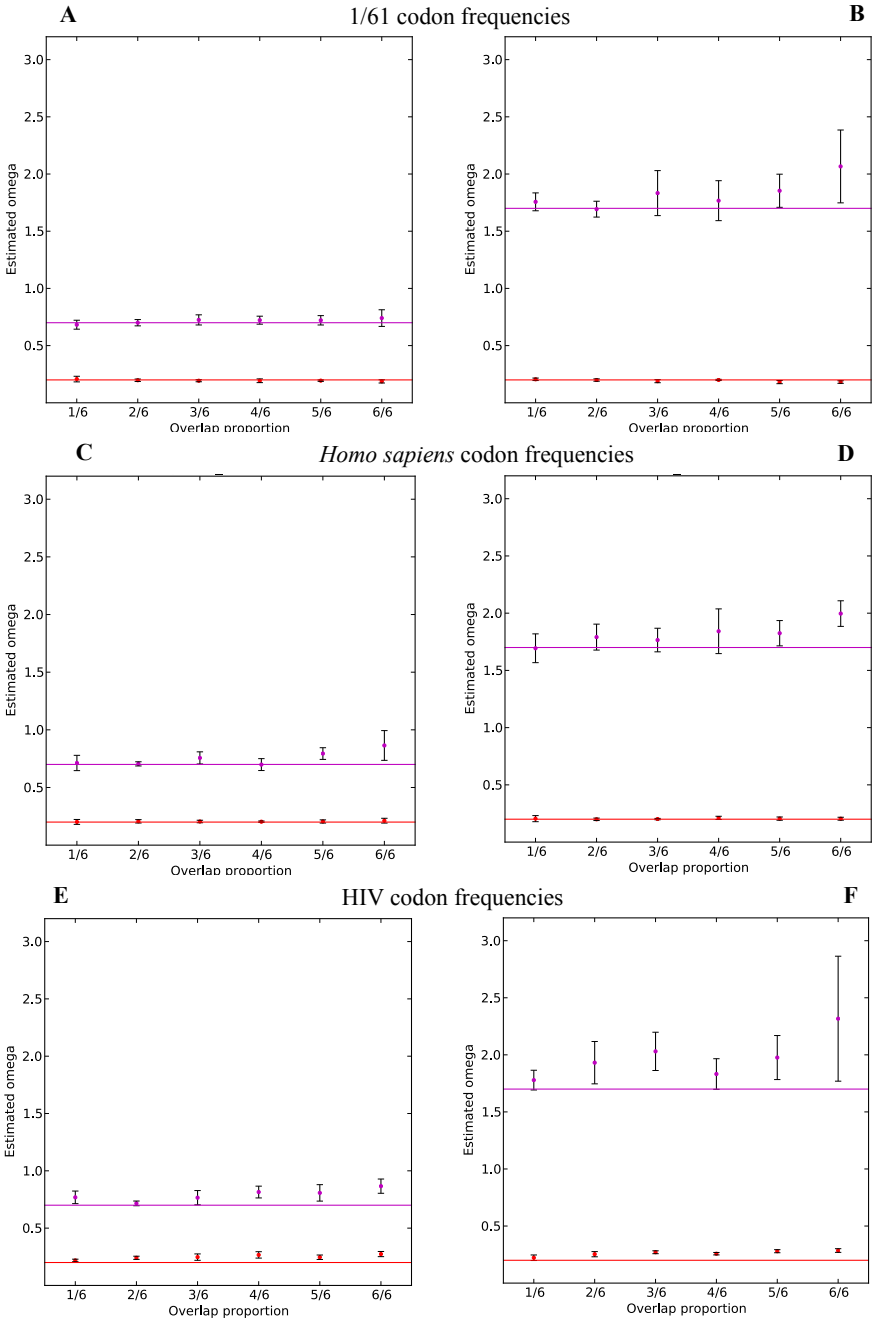

Figure G.1: Genetic code weighting model and overlap proportion. Maximum likelihood estimates of ω values estimated by the genetic code weighting model, as a function of overlap proportion between two genes in simulated data. 64 sequences were simulated, each 901 nucleotide sites long. Each point is the mean ω estimate across separate model fitting to five independently simulated datasets and error bars represent one standard deviation unit. The true ω values used in the simulation are shown as solid lines; these were 0.7 or 1.7 for the gene of interest (purple) and 0.5 for the background gene. The same procedure was done with equal codon probabilities (A and B), human codon probabilities (C and D) and HIV codon probabilities (E and F).

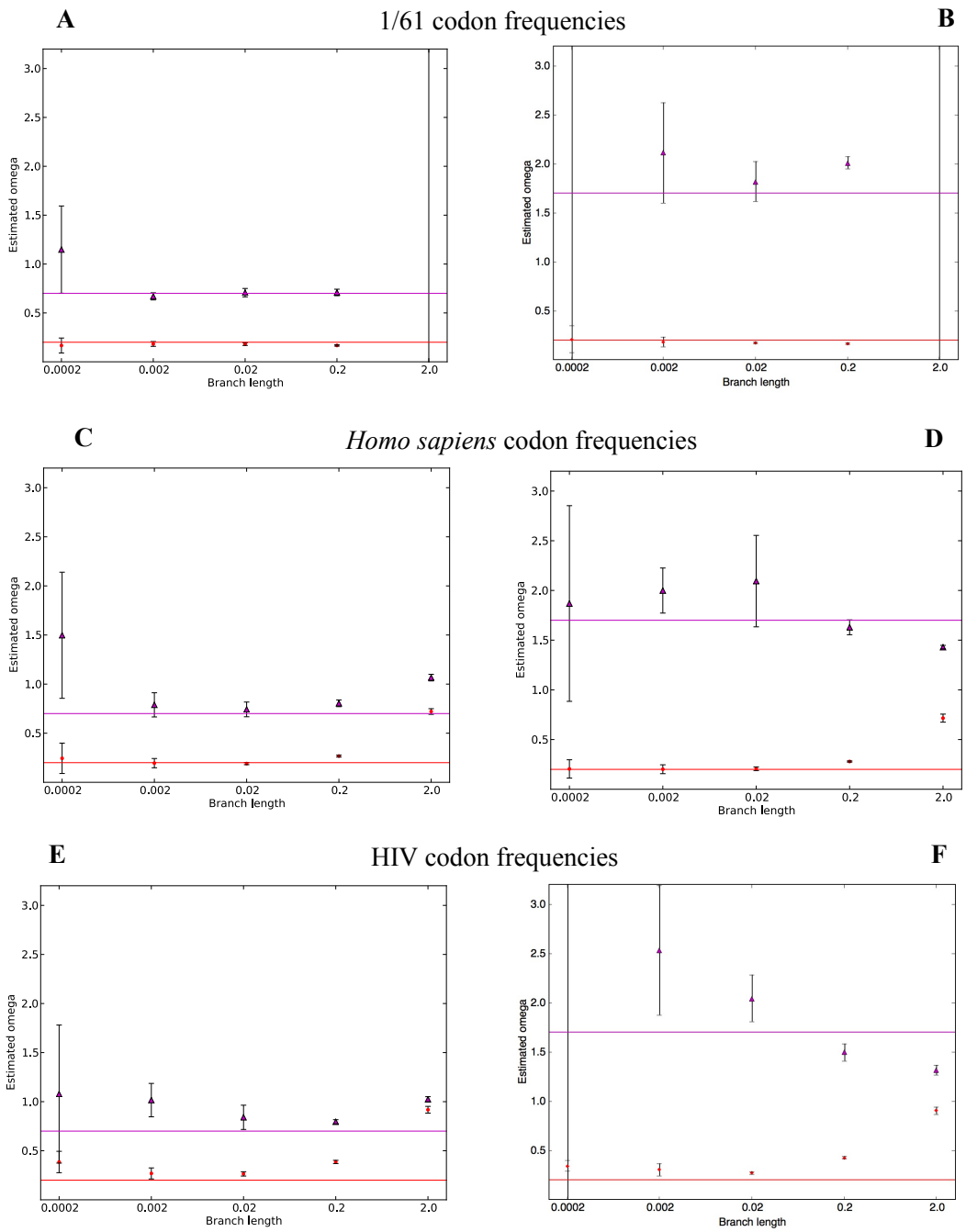

Figure G.2: Genetic code weighting model and branch length. Maximum likelihood estimates of ω values estimated by the genetic code weighting model as a function of branch length used in simulated data, presented as in figure G.1. 64 sequences were simulated along a bifurcating tree where each branch length (in units of 1 expected substitution per site) was specified by the value on the x-axis. The black vertical lines at extreme branch lengths are large error bars which extend beyond the scale of the plot.

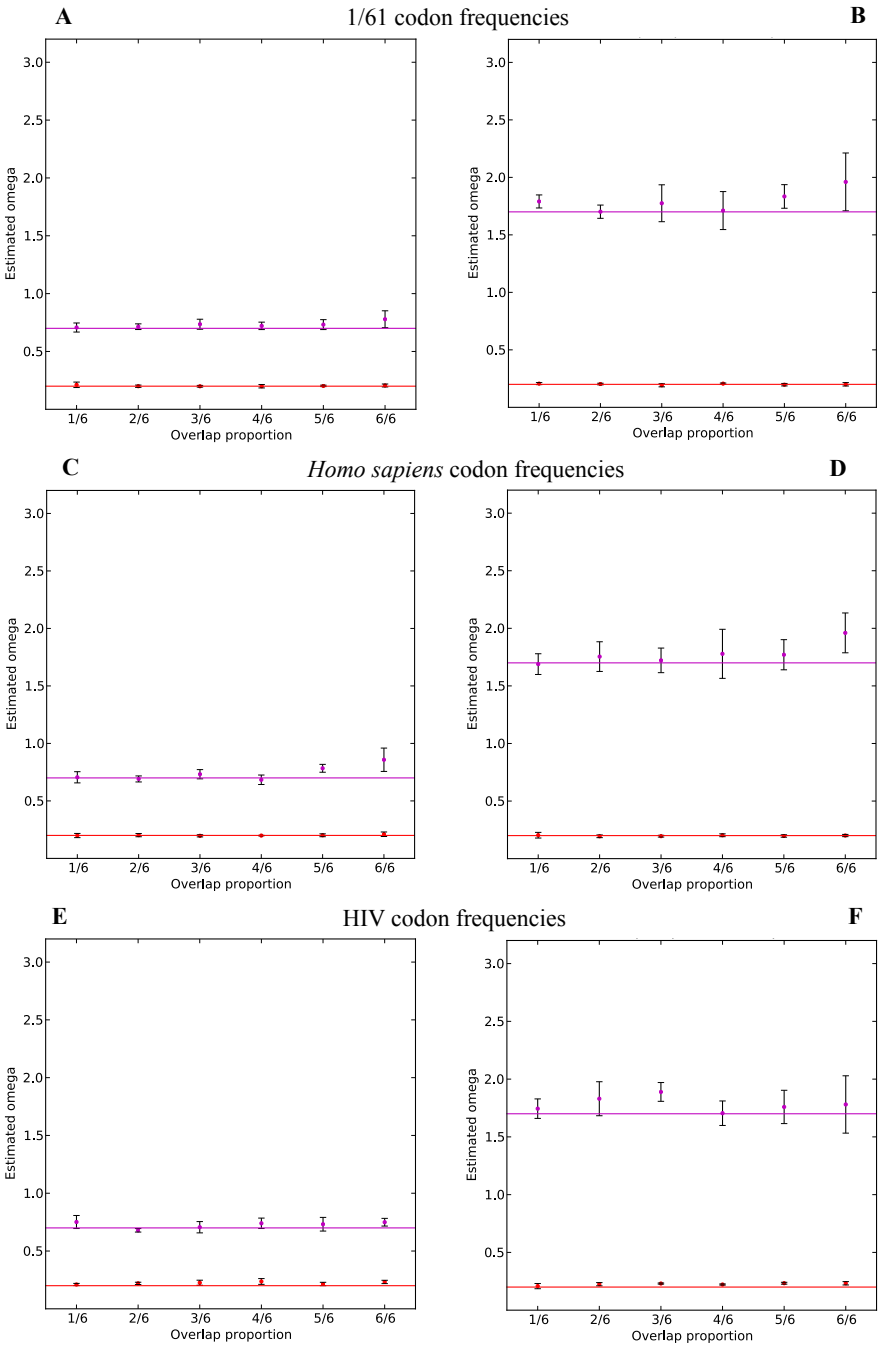

Figure G.3: Maximum likelihood estimates of ω values estimated by the pentamer model, as a function of overlap proportion between two genes in the same simulated data analysed with the genetic code weighted model. Presented as in fig. G.1.

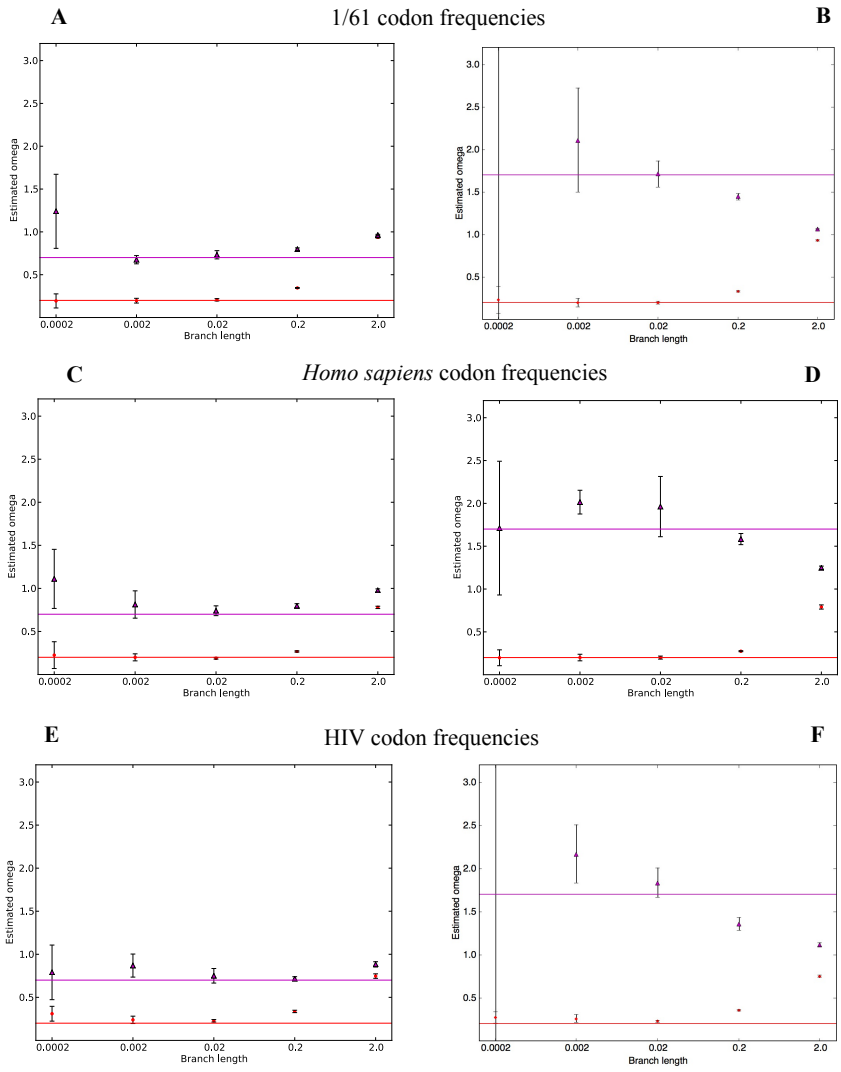

Figure G.4: Maximum likelihood estimates of ω values estimated by the pentamer model as a function of branch length (nucleotide substitutions per site) used in simulated data, presented as in figure G.2. The black vertical line in panels B and F are large error bars which extend beyond the scale of the plot.

Appendix H

Mixture Model Optimisation

ID	Gene 1 (Alt.)			Gene 1 (Null)			Gene 2 (All models)		
	p_0	p_1	p_2	p_0	p_1	p_2	p_0	p_1	p_2
0	0.6	0.2	0.2	0.75	0.25	0	$0.\dot{3}$	$0.\dot{3}$	$0.\dot{3}$
1	0.2	0.6	0.2	0.25	0.75	0	$0.\dot{3}$	$0.\dot{3}$	$0.\dot{3}$
2	0.2	0.2	0.6	0.5	0.5	0	$0.\dot{3}$	$0.\dot{3}$	$0.\dot{3}$
3	$0.\dot{3}$	$0.\dot{3}$	$0.\dot{3}$	0.5	0.5	0	0.6	0.2	0.2
4	$0.\dot{3}$	$0.\dot{3}$	$0.\dot{3}$	0.5	0.5	0	0.2	0.6	0.2
5	$0.\dot{3}$	$0.\dot{3}$	$0.\dot{3}$	0.5	0.5	0	0.2	0.2	0.6

Table H.1: Different combinations of starting values used for the site class probability distribution in tests using data generated with the variable local site dependence simulation (vLSD). Abbreviations: ID, identifier used in the text; Alt., alternative mixture model; Null, null model; $0.\dot{3} \approx \frac{1}{3}$

References

Aggarwal, A., McAllery, S., and Turville, S. G. (2013). Revising the Role of Myeloid cells in HIV Pathogenesis. *Current HIV/AIDS reports*, 10(1):3–11.

Aiken, C., Konner, J., Landau, N. R., Lenburg, M. E., and Trono, D. (1994). Nef induces CD4 endocytosis: Requirement for a critical dileucine motif in the membrane-proximal CD4 cytoplasmic domain. *Cell*, 76(5):853–864.

Apache-Commons (2015). Commons Math: The Apache Commons Mathematics Library, version 3.3.2. *URL: http://commons.apache.org/proper/commons-math/ (Last accessed June 2016)*.

Arnold, L. H., Groom, H. C. T., Kunzelmann, S., Schwefel, D., Caswell, S. J., Ordonez, P., Mann, M. C., Rueschenbaum, S., Goldstone, D. C., Pennell, S., Howell, S. A., Stoye, J. P., Webb, M., Taylor, I. A., and Bishop, K. N. (2015). Phospho-dependent Regulation of SAMHD1 Oligomerisation Couples Catalysis and Restriction. *PLoS Pathogens*, 11:e1005194.

Baker, M. L., Schountz, T., and Wang, L. F. (2013). Antiviral Immune Responses of Bats: A Review. *Zoonoses and Public Health*, 60(1):104–116.

Baldauf, H.-M., Pan, X., Erikson, E., Schmidt, S., Daddacha, W., Burggraf, M., Schenkova, K., Ambiel, I., Wabnitz, G., Gramberg, T., Panitz, S., Flory, E., Landau, N. R., Sertel, S., Rutsch, F., Lasitschka, F., Kim, B., König, R., Fackler, O. T., and Keppler, O. T. (2012). SAMHD1 restricts HIV-1 infection in resting CD4+ T cells. *Nature Medicine*, 18:1682–1689.

Ballana, E. and Esté, J. A. (2015). SAMHD1: At the Crossroads of Cell Proliferation, Immune Responses, and Virus Restriction. *Trends in Microbiology*, 23(11):680–692.

Barre-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamarat, S., Gruest, J., Dauget, C., Axler-Blin, C., Vezinet-Brun, F., Rouzioux, C., Rozenbaum, W., and Montagnier, L. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, 220(4599):868–71.

Barré-Sinoussi, F., Ross, A. L., and Delfraissy, J.-F. (2013). Past, present and future: 30 years of HIV research. *Nature reviews. Microbiology*, 11(12):877–83.

Barrell, B. G., Air, G. M., and Hutchison III, C. A. (1976). Overlapping genes in bacteriophage phi X174. *Nature*, 264:34–41.

Bartee, E., McCormack, A., and Früh, K. (2006). Quantitative membrane proteomics reveals new cellular targets of viral immune modulators. *PLoS pathogens*, 2(10):e107.

Baum, L. L. (2010). Role of humoral immunity in host defense against HIV.

Belzile, J. P., Duisit, G., Rougeau, N., Mercier, J., Finzi, A., and Cohen, r. A. (2007). HIV-1 Vpr-mediated G2 arrest involves the DDB1-CUL4AVPRBP E3 ubiquitin ligase. *PLoS Pathogens*, 3(7):0882–0893.

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*, 57:289–300.

Bergamaschi, A., Ayinde, D., David, A., Le Rouzic, E., Morel, M., Collin, G., Descamps, D., Damond, F., Brun-Vezinet, F., Nisole, S., Margottin-Goguet, F., Pancino, G., and Transy, C. (2009). The human immunodeficiency virus type 2 Vpx protein usurps the CUL4A-DDB1 DCAF1 ubiquitin ligase to overcome a postentry block in macrophage infection. *Journal of virology*, 83(10):4854–4860.

Berger, A., Sommer, A. F. R., Zwarg, J., Hamdorf, M., Welzel, K., Esly, N., Panitz, S., Reuter, A., Ramos, I., Jatiani, A., Mulder, L. C. F., Fernandez-Sesma, A., Rutsch, F., Simon, V., König, R., and Flory, E. (2011). SAMHD1-deficient CD14+ cells from individuals with Aicardi-Goutières syndrome are highly susceptible to HIV-1 infection. *PLoS Pathogens*, 7(12):1–12.

Berger, G., Lawrence, M., Hué, S., and Neil, S. J. D. (2015). G2/M cell cycle arrest correlates with primate lentiviral Vpr interaction with the SLX4 complex. *Journal of virology*, 89(1):230–40.

Bergeron, D., Lapointe, C., Bissonnette, C., Tremblay, G., Motard, J., and Roucou, X. (2013). An out-of-frame overlapping reading frame in the ataxin-1 coding sequence encodes a novel ataxin-1 interacting protein. *Journal of Biological Chemistry*, 288(30):21824–21835.

Beust, C. (2015). JCommander version 1.32: Because life is too short to parse command line parameters. *URL: http://jcommander.org/ (Last accessed June 2016)*.

Bosco, D. a., Eisenmesser, E. Z., Pochapsky, S., Sundquist, W. I., and Kern, D. (2002). Catalysis of cis/trans isomerization in native HIV-1 capsid by human cyclophilin A. *Proceedings of the National Academy of Sciences of the United States of America*, 99(8):5247–52.

Brandes, N. and Linial, M. (2006). Gene overlapping and size constraints in the viral world. *Biology Direct*, 11(1):1–15.

Campbell, E. M. and Hope, T. J. (2015). HIV-1 capsid: the multifaceted key player in HIV-1 infection. *Nature Reviews Microbiology*, 13(8):471–483.

Cavalieri, E., Florido, C., Leal, E., Machado, D. M., Camargo, M., Diaz, R. S., and Janini, L. M. (2009). Intrahost and interhost variability of the HIV type 1 nef gene in Brazilian children. *AIDS research and human retroviruses*, 25:1129–1140.

Ceriani, L. and Verme, P. (2012). The origins of the Gini index: Extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *Journal of Economic Inequality*, 10(3):421–443.

Chaudhuri, R., Mattera, R., Lindwasser, O. W., Robinson, M. S., and Bonifacino, J. S. (2009). A basic patch on alpha-adaptin is required for binding of human immunodeficiency virus type 1 Nef and cooperative assembly of a CD4-Nef-AP-2 complex. *Journal of virology*, 83:2518–2530.

Chen, L., Perlina, A., and Lee, C. J. (2004). Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. *Journal of virology*, 78:3722–3732.

Chirico, N., Vianelli, A., and Belshaw, R. (2010). Why genes overlap in viruses. *Proceedings. Biological sciences / The Royal Society*, 277(July):3809–3817.

Choi, J., Ryoo, J., Oh, C., Hwang, S., and Ahn, K. (2015). SAMHD1 specifically restricts retroviruses through its RNase activity. *Retrovirology*, 12(1):46.

Chung, W. Y., Wadhawan, S., Szklarczyk, R., Pond, S. K., and Nekrutenko, A. (2007). A first look at ARFome: Dual-coding genes in mammalian genomes. *PLoS Computational Biology*, 3(5):0855–0861.

Clavel, F. and Hance, A. J. (2004). HIV drug resistance. *The New England journal of medicine*, 350:1023–1035.

Cohen, E. A., Dehni, G., Sodroski, J. G., and Haseltine, W. A. (1990). Human Immunodeficiency Virus vpr Product Is a Virion-Associated Regulatory Protein. *Journal of Virology*, 64(6):3097–3099.

Collins, K. L., Chen, B. K., Kalams, S. A., Walker, B. D., and Baltimore, D. (1998). HIV-1 Nef protein protects infected primary cells against killing by cytotoxic T lymphocytes. *Nature*, 391:397–401.

Compton, A. A., Hirsch, V. M., and Emerman, M. (2012). The host restriction factor APOBEC3G and retroviral Vif protein coevolve due to ongoing genetic conflict. *Cell Host and Microbe*, 11:91–98.

Cribier, A., Descours, B., Valadão, A., Laguette, N., and Benkirane, M. (2013). Phosphorylation of SAMHD1 by Cyclin A2/CDK1 Regulates Its Restriction Activity toward HIV-1. *Cell Reports*, 3(4):1036–1043.

Crow, M. K. and Wohlgemuth, J. (2003). Microarray analysis of gene expression in lupus. *Arthritis research & therapy*, 5(6):279–87.

Crow, Y. J., Hayward, B. E., Parmar, R., Robins, P., Leitch, A., Ali, M., Black, D. N., van Bokhoven, H., Brunner, H. G., Hamel, B. C., Corry, P. C., Cowan, F. M., Frints, S. G., Klepper, J., Livingston, J. H., Lynch, S. A., Massey, R. F., Meritet, J. F., Michaud, J. L., Ponsot, G., Voit, T., Lebon, P., Bonthron, D. T., Jackson, A. P., Barnes, D. E., and Lindahl, T. (2006). Mutations in the gene encoding the 3'-5' DNA exonuclease TREX1 cause Aicardi-Goutières syndrome at the AGS1 locus. *Nature genetics*, 38(8):917–920.

Cuevas, J. M., Geller, R., Garijo, R., López-Aldeguer, J., and Sanjuán, R. (2015). Extremely High Mutation Rate of HIV-1 In Vivo. *PLoS biology*, 13(9):e1002251.

D'arc, M., Ayouba, A., Esteban, A., Learn, G. H., Boué, V., Liegeois, F., Etienne, L., Tagg, N., Leendertz, F. H., Boesch, C., Madinda, N. F., Robbins, M. M., Gray, M., Cournil, A., Ooms, M., Letko, M., Simon, V. A., Sharp, P. M., Hahn, B. H., Delaporte, E., Mpoudi Ngole, E., and Peeters, M. (2015). Origin of the HIV-1 group O epidemic in western lowland gorillas. *Proceedings of the National Academy of Sciences of the United States of America*, 112(11):E1343–52.

de Oliveira, T., Salemi, M., Gordon, M., Vandamme, A.-M., van Rensburg, E. J., Engelbrecht, S., Coovadia, H. M., and 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.

Deeks, S. G., Overbaugh, J., Phillips, A., and Buchbinder, S. (2015). HIV infection. *Nature Reviews Disease Primers*, 1:15035.

DeHart, J. L., Zimmerman, E. S., Ardon, O., Monteiro-Filho, C. M. R., Argañaraz, E. R., and Planelles, V. (2007). HIV-1 Vpr activates the G2 checkpoint through manipulation of the ubiquitin proteasome system. *Virology journal*, 4(1):57.

Delsuc, F., Brinkmann, H., Chourrout, D., and Philippe, H. (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, 439(7079):965–968.

Dixon, S. (2002). The impact of HIV and AIDS on Africa's economic development. *Bmj*, 324(7331):232–234.

Domenech, R. and Neira, J. L. (2013). The HIV-1 capsid protein as a drug target: recent advances and future prospects. *Current protein & peptide science*, 14(8):658–68.

Drummond, A. and Strimmer, K. (2001). PAL: an object-oriented programming library for molecular evolution and phylogenetics. *Bioinformatics*, 17(7):662–663.

Duggal, N. K. and Emerman, M. (2012). Evolutionary conflicts between viruses and restriction factors shape immunity. *Nature Reviews Immunology*, 12(10):687–695.

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

Erhouma, E., Guiguen, F., Chebloune, Y., Gauthier, D., Lakhal, L. M., Greenland, T., Mornex, J. F., Leroux, C., and Alogninouwa, T. (2008). Small ruminant lentivirus proviral sequences from wild ibexes in contact with domestic goats. *Journal of General Virology*, 89(6):1478–1484.

Esparza, J. (2013). A brief history of the global effort to develop a preventive HIV vaccine.

Evans, D. T., O'Connor, D. H., Jing, P., Dzuris, J. L., Sidney, J., da Silva, J., Allen, T. M., Horton, H., Venham, J. E., Rudersdorf, R. A., Vogel, T., Pauza, C. D., Bontrop, R. E., DeMars, R., Sette, A., Hughes, A. L., and Watkins, D. I. (1999). Virus-specific cytotoxic T-lymphocyte responses select for amino-acid variation in simian immunodeficiency virus Env and Nef. *Nature medicine*, 5:1270–1276.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–76.

Foster, J. L., Denial, S. J., Temple, B. R. S., and Garcia, J. V. (2011). Mechanisms of HIV-1 Nef function and intracellular signaling. *Journal of Neuroimmune Pharmacology*, 6:230–246.

Franzolin, E., Pontarin, G., Rampazzo, C., Miazzi, C., Ferraro, P., Palumbo, E., Reichard, P., and Bianchi, V. (2013). The deoxynucleotide triphosphohydrolase SAMHD1 is a major regulator of DNA precursor pools in mammalian cells. *Proceedings of the National Academy of Sciences*, 110(35):14272–14277.

Fregoso, O. I., Ahn, J., Wang, C., Mehrens, J., Skowronski, J., and Emerman, M. (2013). Evolutionary Toggling of Vpx/Vpr Specificity Results in Divergent Recognition of the Restriction Factor SAMHD1. *PLoS Pathogens*, 9(7).

Gamble, T. R., Vajdos, F. F., Yoo, S., Worthylake, D. K., Houseweart, M., Sundquist, W. I., and Hill, C. P. (1996). Crystal structure of human cyclophilin A bound to the amino-terminal domain of HIV-1 capsid. *Cell*, 87(7):1285–1294.

Ganser, B. K., Li, S., Klishko, V. Y., Finch, J. T., and Sundquist, W. I. (1999). Assembly and analysis of conical models for the HIV-1 core. *Science (New York, N.Y.)*, 283(5398):80–83.

Gao, D., Wu, J., Wu, Y.-T., Du, F., Aroh, C., Yan, N., Sun, L., and Chen, Z. J. (2013). Cyclic GMP-AMP Synthase Is an Innate Immune Sensor of HIV and Other Retroviruses. *Science*, 341(6148):903–906.

Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., Cummins, L. B., Arthur, L. O., Peeters, M., Shaw, G. M., Sharp, P. M., and Hahn, B. H. (1999). Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature*, 397(6718):436–41.

Garcia, J. V. and Miller, A. D. (1991). Serine phosphorylation-independent downregulation of cell-surface CD4 by nef. *Nature*, 350:508–511.

Geyer, M., Fackler, O. T., and Peterlin, B. M. (2001). Structure–function relationships in HIV-1 Nef. *EMBO reports*, 2(7):580–585.

Geyer, M. and Peterlin, B. M. (2001). Domain assembly, surface accessibility and sequence conservation in full length HIV-1 Nef. *FEBS Letters*, 496:91–95.

Gharib, W. H. and Robinson-Rechavi, M. (2013). The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. *Molecular Biology and Evolution*, 30(7):1675–1686.

Gifford, R. J. (2012). Viral evolution in deep time: Lentiviruses and mammals.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular biology and evolution*, 11(5):725–36.

Goldstone, D. C., Ennis-Adeniran, V., Hedden, J. J., Groom, H. C. T., Rice, G. I., Christodoulou, E., Walker, P. A., Kelly, G., Haire, L. F., Yap, M. W., de Carvalho, L. P. S., Stoye, J. P., Crow, Y. J., Taylor, I. A., and Webb, M. (2011). HIV-1 restriction factor SAMHD1 is a deoxynucleoside triphosphate triphosphohydrolase. *Nature*, 480(7377):379–382.

Goncalves, A., Karayel, E., Rice, G. I., Bennett, K. L., Crow, Y. J., Superti-Furga, G., and Bürckstümmer, T. (2012). SAMHD1 is a nucleic-acid binding protein that is mislocalized due to aicardi-goutières syndrome-associated mutations. *Human Mutation*, 33(7):1116–1122.

Gottlieb, M., Schroff, R., Schanker, H., Weisman, J., Fan, P., Wolf, R., and Saxon, A. (1981). Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. *The New England Journal of Medicine*, 305(24):1425–1431.

Goujon, C., Arfi, V., Pertel, T., Luban, J., Lienard, J., Rigal, D., Darlix, J.-L., and Cimarelli, A. (2008). Characterization of simian immunodeficiency virus SIVSM/human immunodeficiency virus type 2 Vpx function in human myeloid cells. *Journal of virology*, 82(24):12335–12345.

Goujon, C., Jarrosson-Wuillème, L., Bernaud, J., Rigal, D., Darlix, J.-L., and Cimarelli,

a. (2006). With a little help from a friend: increasing HIV transduction of monocyte-derived dendritic cells with virion-like particles of SIV(MAC). *Gene therapy*, 13(12):991–994.

Goujon, C., Rivière, L., Jarrosson-Wuilleme, L., Bernaud, J., Rigal, D., Darlix, J.-L., and Cimarelli, A. (2007). SIVSM/HIV-2 Vpx proteins promote retroviral escape from a proteasome-dependent restriction pathway present in human dendritic cells. *Retrovirology*, 4:2.

Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular biology and evolution*, 27:221–224.

Gramberg, T., Kahle, T., Bloch, N., Wittmann, S., Müllers, E., Daddacha, W., Hofmann, H., Kim, B., Lindemann, D., and Landau, N. R. (2013). Restriction of diverse retroviruses by SAMHD1. *Retrovirology*, 10:26.

Greenberg, M., DeTulleo, L., Rapoport, I., Skowronski, J., and Kirchhausen, T. (1998). A dileucine motif in HIV-1 Nef is essential for sorting into clathrin-coated pits and for downregulation of CD4. *Current Biology*, 8(22):1239–S3.

Gres, A. T., Kirby, K. A., KewalRamani, V. N., Tanner, J. J., Pornillos, O., and Sarafianos, S. G. (2015). X-ray crystal structures of native HIV-1 capsid protein reveal conformational variability. *Science (New York, N.Y.)*, 349(6243):99–104.

Gupta, R. K., Hué, S., Schaller, T., Verschoor, E., Pillay, D., and Towers, G. J. (2009). Mutation of a single residue renders human tetherin resistant to HIV-1 Vpu-mediated depletion. *PLoS Pathogens*, 5.

Halpern, a. L. and Bruno, W. J. (1998). Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular biology and evolution*, 15(7):910–7.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating the human-ape split by a molecular clock of mitochondrial DNA. *Evolution*, 22:160–174.

Hauser, H., Lopez, L. A., Yang, S. J., Oldenburg, J. E., Exline, C. M., Guatelli, J. C., and Cannon, P. M. (2010). HIV-1 Vpu and HIV-2 Env counteract BST-2/tetherin by sequestration in a perinuclear compartment. *Retrovirology*, 7:51.

Hein, J. and Stovlbaek, J. (1995). A maximum-likelihood approach to analyzing nonoverlapping and overlapping reading frames. *Journal of Molecular Evolution*, 40:181–189.

Hirsch, V. M., Olmsted, R. a., Murphey-Corb, M., Purcell, R. H., and Johnson, P. R. (1989). An African primate lentivirus (SIVsm) closely related to HIV-2. *Nature*, 339(6223):389–392.

Hollenbaugh, J. a., Gee, P., Baker, J., Daly, M. B., Amie, S. M., Tate, J., Kasai, N., Kanemura, Y., Kim, D.-H., Ward, B. M., Koyanagi, Y., and Kim, B. (2013). Host factor SAMHD1 restricts DNA viruses in non-dividing myeloid cells. *PLoS pathogens*, 9(6):e1003481.

Hrecka, K., Gierszewska, M., Srivastava, S., Kozaczkiewicz, L., Swanson, S. K., Florens, L., Washburn, M. P., and Skowronski, J. (2007). Lentiviral Vpr usurps Cul4-DDB1[VprBP] E3 ubiquitin ligase to modulate cell cycle. *Proceedings of the National Academy of Sciences of the United States of America*, 104(28):11778–83.

Hrecka, K., Hao, C., Gierszewska, M., Swanson, S. K., Kesik-Brodacka, M., Srivastava, S., Florens, L., Washburn, M. P., and Skowronski, J. (2011). Vpx relieves inhibition of HIV-1 infection of macrophages mediated by the SAMHD1 protein. *Nature*, 474(7353):658–661.

Hu, S., Li, J., Xu, F., Mei, S., Le Duff, Y., Yin, L., Pang, X., Cen, S., Jin, Q., Liang, C., and Guo, F. (2015). SAMHD1 Inhibits LINE-1 Retrotransposition by Promoting Stress Granule Formation. *PLoS genetics*, 11(7):e1005367.

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

Iwabu, Y., Fujita, H., Kinomoto, M., Kaneko, K., Ishizaka, Y., Tanaka, Y., Sata, T.,

and Tokunaga, K. (2009). HIV-1 accessory protein Vpu internalizes cell-surface BST-2/tetherin through transmembrane interactions leading to lysosomes. *Journal of Biological Chemistry*, 284:35060–35072.

Iwasa, Y. (1988). Free fitness that always increases in evolution. *Journal of Theoretical Biology*, 135(3):265–281.

Jacques, D. A., McEwan, W. A., Hilditch, L., Price, A. J., Towers, G. J., and James, L. C. (2016). HIV-1 uses dynamic capsid pores to import nucleotides and fuel encapsidated DNA synthesis. *Nature*, 536:349–353.

Ji, X., Tang, C., Zhao, Q., Wang, W., and Xiong, Y. (2014). Structural basis of cellular dNTP regulation by SAMHD1. *Proceedings of the National Academy of Sciences of the United States of America*, 111(41):E4305–E4314.

Ji, X., Wu, Y., Yan, J., Mehrens, J., Yang, H., DeLucia, M., Hao, C., Gronenborn, A. M., Skowronski, J., Ahn, J., and Xiong, Y. (2013). Mechanism of allosteric activation of SAMHD1 by dGTP. *Nature structural & molecular biology*, 20(11):1304–9.

Johnson, Z. I. and Chisholm, S. W. (2004). Properties of overlapping genes are conserved across microbial genomes. *Genome Research*, 14(11):2268–2272.

Joos, B., Fischer, M., Schweizer, A., Kuster, H., Böni, J., Wong, J. K., Weber, R., Trkola, A., and Günthard, H. F. (2007). Positive in vivo selection of the HIV-1 envelope protein gp120 occurs at surface-exposed regions. *The Journal of infectious diseases*, 196:313–320.

Kaletsky, R. L., Francica, J. R., Agrawal-Gamse, C., and Bates, P. (2009). Tetherin-mediated restriction of filovirus budding is antagonized by the Ebola glycoprotein. *Proceedings of the National Academy of Sciences of the United States of America*, 106(8):2886–2891.

Kataoka, N., Bachorik, J. L., and Dreyfuss, G. (1999). Transportin-SR, a nuclear import receptor for SR proteins. *Journal of Cell Biology*, 145(6):1145–1152.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–80.

Katzourakis, A., Tristem, M., Pybus, O. G., and Gifford, R. J. (2007). Discovery and analysis of the first endogenous lentivirus. *Proceedings of the National Academy of Sciences of the United States of America*, 104(15):6261–5.

Keele, B. F., Van Heuverswyn, F., Li, Y., Bailes, E., Takehisa, J., Santiago, M. L., Bibollet-Ruche, F., Chen, Y., Wain, L. V., Liegeois, F., Loul, S., Ngole, E. M., Bienvenue, Y., Delaporte, E., Brookfield, J. F. Y., Sharp, P. M., Shaw, G. M., Peeters, M., and Hahn, B. H. (2006). Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science (New York, N.Y.)*, 313(5786):523–6.

Kestler, H. W., Ringler, D. J., Mori, K., Panicali, D. L., Sehgal, P. K., Daniel, M. D., and Desrosiers, R. C. (1991). Importance of the nef gene for maintenance of high virus loads and for development of AIDS. *Cell*, 65:651–662.

Kim, E. T., White, T. E., Brandariz-Núñez, A., Diaz-Griffero, F., and Weitzman, M. D. (2013). SAMHD1 restricts herpes simplex virus 1 in macrophages by limiting DNA replication. *Journal of virology*, 87(23):12949–56.

Kimura, M. (1962). On the Probability of Fixation of Mutant Genes in a Population . *Genetics*, 47(6):713–719.

Kirchhoff, F. (2010). Immune evasion and counteraction of restriction factors by HIV-1 and other primate lentiviruses. *Cell host & microbe*, 8(1):55–67.

Kirchhoff, F., Greenough, T. C., Brettler, D. B., Sullivan, J. L., and Desrosiers, R. C. (1995). Absence of Intact nef Sequences in a Long-Term Survivor with Nonprogressive HIV-1 Infection. *New England Journal of Medicine*, 332(4):228–232.

Klemke, M., Kehlenbach, R. H., and Huttner, W. B. (2001). Two overlapping reading frames in a single exon encode interacting proteins - A novel way of gene usage. *EMBO Journal*, 20(14):3849–3860.

Kluge, S., Mack, K., Iyer, S., Pujol, F., Heigele, A., Learn, G., Usmani, S., Sauter, D., Joas, S., Hotter, D., Bibollet-Ruche, F., Plenderleith, L., Peeters, M., Geyer, M., Sharp, P., Fackler, O., Hahn, B., and Kirchhoff, F. (2014). Nef Proteins of Epidemic HIV-1 Group O Strains Antagonize Human Tetherin. *Cell Host & Microbe*, 16(5):639–650.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, a., Hahn, B. H., Wolinsky, S., and Bhattacharya, T. (2000). Timing the ancestor of the HIV-1 pandemic strains. *Science (New York, N.Y.)*, 288(5472):1789–96.

Laguette, N., Bregnard, C., Hue, P., Basbous, J., Yatim, A., Larroque, M., Kirchhoff, F., Constantinou, A., Sobhian, B., and Benkirane, M. (2014). Premature activation of the slx4 complex by vpr promotes g2/m arrest and escape from innate immune sensing. *Cell*, 156(1-2):134–145.

Laguette, N., Rahm, N., Sobhian, B., Chable-Bessia, C., Münch, J., Snoeck, J., Sauter, D., Switzer, W. M., Heneine, W., Kirchhoff, F., Delsuc, F., Telenti, A., and Benkirane, M. (2012). Evolutionary and functional analyses of the interaction between the myeloid restriction factor SAMHD1 and the lentiviral Vpx protein. *Cell Host and Microbe*, 11(2):205–217.

Laguette, N., Sobhian, B., Casartelli, N., Ringeard, M., Chable-Bessia, C., Ségéral, E., Yatim, A., Emiliani, S., Schwartz, O., and Benkirane, M. (2011). SAMHD1 is the dendritic- and myeloid-cell-specific HIV-1 restriction factor counteracted by Vpx. *Nature*, 474(7353):654–657.

Lahouassa, H., Daddacha, W., Hofmann, H., Ayinde, D., Logue, E. C., Dragin, L., Bloch, N., Maudet, C., Bertrand, M., Gramberg, T., Pancino, G., Priet, S., Canard, B., Laguette, N., Benkirane, M., Transy, C., Landau, N. R., Kim, B., and Margottin-Goguet, F. (2012). SAMHD1 restricts the replication of human immunodeficiency virus type 1 by depleting the intracellular pool of deoxynucleoside triphosphates. *Nature Immunology*, 13(6):621–621.

Lecossier, D., Bouchonnet, F., Clavel, F., and Hance, A. J. (2003). Hypermutation of HIV-1 DNA in the Absence of the Vif Protein. *Science*, 300(5622):1112.

Lee, B. (2013). HIV provides ample PAMPs for innate immune sensing. *Proceedings of the National Academy of Sciences*, 110(48):19183–19184.

Lee, K., Ambrose, Z., Martin, T. D., Oztop, I., Mulky, A., Julias, J. G., Vandegraaff, N., Baumann, J. G., Wang, R., Yuen, W., Takemura, T., Shelton, K., Taniuchi, I., Li, Y., Sodroski, J., Littman, D. R., Coffin, J. M., Hughes, S. H., Unutmaz, D., Engelman, A., and KewalRamani, V. N. (2010). Flexible Use of Nuclear Import Pathways by HIV-1. *Cell Host and Microbe*, 7(3):221–233.

Lenzi, G. M., Domaoal, R. A., Kim, D. H., Schinazi, R. F., and Kim, B. (2015). Mechanistic and kinetic differences between reverse transcriptases of Vpx coding and non-coding lentiviruses. *Journal of Biological Chemistry*, 290(50):30078–30086.

Li, N., Zhang, W., and Cao, X. (2000). Identification of human homologue of mouse IFN-γ induced protein from human dendritic cells. *Immunology Letters*, 74(3):221–224.

Li, W.-H., Wu, C.-I., and Luo, C.-C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution*, 2(2):150–174.

Lim, E. S., Fregoso, O. I., McCoy, C. O., Matsen, F. A., Malik, H. S., and Emerman, M. (2012). The ability of primate lentiviruses to degrade the monocyte restriction factor SAMHD1 preceded the birth of the viral accessory protein Vpx. *Cell Host and Microbe*, 11:194–204.

Lindwasser, O. W., Smith, W. J., Chaudhuri, R., Yang, P., Hurley, J. H., and Bonifacino, J. S. (2008). A diacidic motif in human immunodeficiency virus type 1 Nef is a novel determinant of binding to AP-2. *Journal of virology*, 82:1166–1174.

Llano, A., Williams, A., Olvera, A., Silva-Arrieta, S., and Brander, C. (2013). *Best-Characterized HIV-1 CTL Epitopes: The 2013 Update*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos.

Luis, A. D., Hayman, D. T. S., O'Shea, T. J., Cryan, P. M., Gilbert, A. T., Pulliam, J. R. C., Mills, J. N., Timonin, M. E., Willis, C. K. R., Cunningham, A. a., Fooks, A. R., Rupprecht, C. E., Wood, J. L. N., and Webb, C. T. (2013). A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special? *Proceedings. Biological sciences / The Royal Society*, 280(1756):20122753.

Makalowska, I., Lin, C. F., and Makalowski, W. (2005). Overlapping genes in vertebrate genomes. *Computational Biology and Chemistry*, 29(1):1–12.

Maldarelli, F., Chen, M. Y., Willey, R. L., and Strebel, K. (1993). Human immunodeficiency virus type 1 Vpu protein is an oligomeric type I integral membrane protein. *J Virol*, 67(8):5056–5061.

Malim, M. H. and Emerman, M. (2008). HIV-1 Accessory Proteins-Ensuring Viral Survival in a Hostile Environment. *Cell Host and Microbe*, 3(6):388–398.

Mangeat, B., Turelli, P., Caron, G., Friedli, M., Perrin, L., and Trono, D. (2004). Broad Antiretroviral Defence by Human APOBEC3G Through Lethal Editing of Nascent Reverse Transcripts. *Nature*, 424(6944):99–103.

Matreyek, K. A., Yücel, S. S., Li, X., and Engelman, A. (2013). Nucleoporin NUP153 Phenylalanine-Glycine Motifs Engage a Common Binding Pocket within the HIV-1 Capsid Protein to Mediate Lentiviral Infectivity. *PLoS Pathogens*, 9(10).

McNatt, M. W., Zang, T., and Bieniasz, P. D. (2013). Vpu Binds Directly to Tetherin and Displaces It from Nascent Virions. *PLoS Pathogens*, 9(4).

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

Meredith, R. W., Janečka, J. E., Gatesy, J., Ryder, O. a., Fisher, C. a., Teeling, E. C., Goodbla, A., Eizirik, E., Simão, T. L. L., Stadler, T., Rabosky, D. L., Honeycutt, R. L., Flynn, J. J., Ingram, C. M., Steiner, C., Williams, T. L., Robinson, T. J., Burk-Herrick, A., Westerman, M., Ayoub, N. a., Springer, M. S., and Murphy, W. J. (2011). Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science (New York, N.Y.)*, 334(6055):521–4.

Monit, C., Goldstein, R. A., Towers, G., and Hué, S. (2015). Positive Selection Analysis of Overlapping Reading Frames Is Invalid. *AIDS Research and Human Retroviruses*, 31(10):947–947.

Morellet, N., Bouaziz, S., Petitjean, P., and Roques, B. P. (2003). NMR structure of the HIV-1 regulatory protein VPR. *Journal of Molecular Biology*, 327:215–227.

Mouilleron, H., Delcourt, V., and Roucou, X. (2016). Death of a dogma: Eukaryotic mRNAs can code for more than one protein. *Nucleic Acids Research*, 44(1):14–23.

Murrell, B., de Oliveira, T., Seebregts, C., Kosakovsky Pond, S. L., and Scheffler, K. (2012). Modeling HIV-1 drug resistance as episodic directional selection. *PLoS Computational Biology*, 8(5).

Nakamura, Y., Gojobori, T., and Ikemura, T. (2000). Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res*, 28(1):292.

Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular biology and evolution*, 3(5):418–26.

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

Nekrutenko, A., Wadhawan, S., Goetting-Minesky, P., and Makova, K. D. (2005). Oscillating evolution of a mammalian locus with overlapping reading frames: An XL-alpha-s/ALEX relay. *PLoS Genetics*, 1(2):0197–0204.

Nelder, J. A. and Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, 7:308–313.

Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino

acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148:929–936.

Obenauer, J. C., Denson, J., Mehta, P. K., Su, X., Mukatira, S., Finkelstein, D. B., Xu, X., Wang, J., Ma, J., Fan, Y., Rakestraw, K. M., Webster, R. G., Hoffmann, E., Krauss, S., Zheng, J., Zhang, Z., and Naeve, C. W. (2006). Large-scale sequence analysis of avian influenza isolates. *Science (New York, N.Y.)*, 311(5767):1576–80.

Ostell, J. and McEntyre, J. (2007). The NCBI Handbook. *URL: www.ncbi.nlm.nih.gov/books/NBK21101/ (Accessed November 2015)*.

Ott, M., Zola, J., Stamatakis, A., and Aluru, S. (2007). Large-scale maximum likelihood-based phylogenetic analysis on the IBM BlueGene/L. *Proceedings of the 2007 ACM/IEEE Conference on Supercomputing (SC '07)*.

Park, S. H., Mrse, A. A., Nevzorov, A. A., Mesleh, M. F., Oblatt-Montal, M., Montal, M., and Opella, S. J. (2003). Three-dimensional structure of the channel-forming transmembrane domain of virus protein "u" (Vpu) from HIV-1. *Journal of Molecular Biology*, 333(2):409–424.

Pauls, E., Ruiz, A., Badia, R., Permanyer, M., Gubern, A., Riveira-Muñoz, E., Torres-Torronteras, J., Alvarez, M., Mothe, B., Brander, C., Crespo, M., Menéndez-Arias, L., Clotet, B., Keppler, O. T., Martí, R., Posas, F., Ballana, E., and Esté, J. a. (2014). Cell cycle control and HIV-1 susceptibility are linked by CDK6-dependent CDK2 phosphorylation of SAMHD1 in myeloid and lymphoid cells. *Journal of immunology (Baltimore, Md. : 1950)*, 193(4):1988–97.

Pedersen, a. M. and Jensen, J. L. (2001). A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Molecular biology and evolution*, 18(5):763–76.

Phillips, R. E., Rowland-Jones, S., Nixon, D. F., Gotch, F. M., Edwards, J. P., Ogunlesi, a. O., Elvin, J. G., Rothbard, J. a., Bangham, C. R., and Rizza, C. R. (1991). Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. *Nature*, 354(6353):453–9.

Planelles, V., Jowett, J. B., Li, Q. X., Xie, Y., Hahn, B., and Chen, I. S. (1996). Vpr-induced cell cycle arrest is conserved among primate lentiviruses. *Journal of virology*, 70:2516–2524.

Plantier, J.-C., Leoz, M., Dickerson, J. E., De Oliveira, F., Cordonnier, F., Lemée, V., Damond, F., Robertson, D. L., and Simon, F. (2009). A new human immunodeficiency virus derived from gorillas. *Nature medicine*, 15(8):871–2.

Plotkin, J. B. and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. TL - 12. *Nature reviews. Genetics*, 12 VN - r(1):32–42.

Pornillos, O., Ganser-Pornillos, B. K., and Yeager, M. (2011). Atomic-level modelling of the HIV capsid. *Nature*, 469(7330):424–7.

Powell, M. J. D. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives (unpublished). *URL: http://www.damtp.cam.ac.uk/user/na/NA_papers/NA2009_06.pdf (last accessed July 2016)*.

Powell, R. D., Holland, P. J., Hollis, T., and Perrino, F. W. (2011). Aicardi-Goutières syndrome gene and HIV-1 restriction factor SAMHD1 is a dGTP-regulated deoxynucleotide triphosphohydrolase. *Journal of Biological Chemistry*, 286(51):43596–43600.

Price, A. J., Fletcher, A. J., Schaller, T., Elliott, T., Lee, K. E., KewalRamani, V. N., Chin, J. W., Towers, G. J., and James, L. C. (2012). CPSF6 Defines a Conserved Capsid Interface that Modulates HIV-1 Replication. *PLoS Pathogens*, 8(8).

Price, D. A., Goulder, P. J., Klenerman, P., Sewell, A. K., Easterbrook, P. J., Troop, M., Bangham, C. R., and Phillips, R. E. (1997). Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. *Proceedings of the National Academy of Sciences of the United States of America*, 94:1890–1895.

Qiao, F. and Bowie, J. U. (2005). The many faces of SAM. *Science's STKE : signal transduction knowledge environment*, 2005(286):re7.

Rambaut, A. (2006). FigTreee, version 1.3.1. Available at: http://tree.bio.ed.ac.uk/software/figtree/. Last accessed January 2015.

Rasaiyaah, J., Tan, C. P., Fletcher, A. J., Price, A. J., Blondeau, C., Hilditch, L., Jacques, D. A., Selwood, D. L., James, L. C., Noursadeghi, M., and Towers, G. J. (2013). HIV-1 evades innate immune recognition through specific cofactor recruitment. *Nature*, 503(7476):402–405.

Re, F., Braaten, D., Franke, E. K., and Luban, J. (1995). Human immunodeficiency virus type 1 Vpr arrests the cell cycle in G2 by inhibiting the activation of p34cdc2-cyclin B. *Journal of virology*, 69(11):6859–64.

Ren, X., Park, S. Y., Bonifacino, J. S., and Hurley, J. H. (2014). How HIV-1 Nef hijacks the AP-2 clathrin adaptor to downregulate CD4. *eLife*, 2014(3).

Rice, G. I., Bond, J., Asipu, A., Brunette, R. L., Manfield, I. W., Carr, I. M., Fuller, J. C., Jackson, R. M., Lamb, T., Briggs, T. A., Ali, M., Gornall, H., Couthard, L. R., Aeby, A., Attard-Montalto, S. P., Bertini, E., Bodemer, C., Brockmann, K., Brueton, L. A., Corry, P. C., Desguerre, I., Fazzi, E., Cazorla, A. G., Gener, B., Hamel, B. C. J., Heiberg, A., Hunter, M., van der Knaap, M. S., Kumar, R., Lagae, L., Landrieu, P. G., Lourenco, C. M., Marom, D., McDermott, M. F., van der Merwe, W., Orcesi, S., Prendiville, J. S., Rasmussen, M., Shalev, S. A., Soler, D. M., Shinawi, M., Spiegel, R., Tan, T. Y., Vanderver, A., Wakeling, E. L., Wassmer, E., Whittaker, E., Lebon, P., Stetson, D. B., Bonthron, D. T., and Crow, Y. J. (2009). Mutations involved in Aicardi-Goutières syndrome implicate SAMHD1 as regulator of the innate immune response. *Nature genetics*, 41(7):829–832.

Roy, C., Khandaker, I., and Oshitani, H. (2015). Intersubtype Genetic Variation of HIV-1 Tat Exon 1. *AIDS Research and Human Retroviruses*, 31(6):641–648.

Ruepp, M.-D., Aringhieri, C., Vivarelli, S., Cardinale, S., Paro, S., Schümperli, D., and Barabino, S. M. L. (2009). Mammalian pre-mRNA 3' end processing factor CF I m 68 functions in mRNA export. *Molecular biology of the cell*, 20(24):5211–23.

Ruffin, N., Brezar, V., Ayinde, D., Lefebvre, C., Wiesch, J. S. Z., van Lunzen, J., Bockhorn, M., Schwartz, O., Hocini, H., Lelievre, J.-D., Banchereau, J., Levy, Y., and Seddiki, N. (2015). Low SAMHD1 expression following T-cell activation and proliferation renders CD4+ T cells susceptible to HIV-1. *AIDS*, 29(5):1.

Ryoo, J., Choi, J., Oh, C., Kim, S., Seo, M., Kim, S.-Y., Seo, D., Kim, J., White, T. E., Brandariz-Nuñez, A., Diaz-Griffero, F., Yun, C.-H., Hollenbaugh, J. a., Kim, B., Baek, D., and Ahn, K. (2014). The ribonuclease activity of SAMHD1 is required for HIV-1 restriction. *Nature Medicine*, 20:936–941.

Sabath, N., Landan, G., and Graur, D. (2008). A method for the simultaneous estimation of selection intensities in overlapping genes. *PloS one*, 3(12):e3996.

Sakharkar, K. R., Sakharkar, M. K., Verma, C., and Chow, V. T. K. (2005). Comparative study of overlapping genes in bacteria, with special reference to Rickettsia prowazekii and Rickettsia conorii. *International Journal of Systematic and Evolutionary Microbiology*, 55(3):1205–1209.

Saksela, K., Cheng, G., and Baltimore, D. (1995). Proline-rich (PxxP) motifs in HIV-1 Nef bind to SH3 domains of a subset of Src kinases and are required for the enhanced growth of Nef+ viruses but not for down-regulation of CD4. *The EMBO Journal*, 14(3):484–491.

Sanna, C. R., Li, W.-H., and Zhang, L. (2008). Overlapping genes in the human and mouse genomes. *BMC genomics*, 9:169.

Sauter, D., Schindler, M., Specht, A., Landford, W. N., Münch, J., Kim, K.-A., Votteler, J., Schubert, U., Bibollet-Ruche, F., Keele, B. F., Takehisa, J., Ogando, Y., Ochsenbauer, C., Kappes, J. C., Ayouba, A., Peeters, M., Learn, G. H., Shaw, G., Sharp, P. M., Bieniasz, P., Hahn, B. H., Hatziioannou, T., and Kirchhoff, F. (2009). Tetherin-driven adaptation of Vpu and Nef function and the evolution of pandemic and nonpandemic HIV-1 strains. *Cell host & microbe*, 6(5):409–21.

Sawyer, S. L., Emerman, M., and Malik, H. S. (2004). Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS Biology*, 2.

Sawyer, S. L., Wu, L. I., Emerman, M., and Malik, H. S. (2005). Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. *Proceedings of the National Academy of Sciences of the United States of America*, 102:2832–2837.

Schaller, T., Ocwieja, K. E., Rasaiyaah, J., Price, A. J., Brady, T. L., Roth, S. L., Hue, S., Fletcher, A. J., Lee, K., KewalRamani, V. N., Noursadeghi, M., Jenner, R. G., James, L. C., Bushman, F. D., and Towers, G. J. (2011). HIV-1 capsid-cyclophilin interactions determine nuclear import pathway, integration targeting and replication efficiency. *PLoS Pathogens*, 7(12).

Schmidt, S., Schenkova, K., Adam, T., Erikson, E., Lehmann-koch, J., Sertel, S., Verhasselt, B., Fackler, O. T., Lasitschka, F., and Keppler, O. T. (2015). SAMHD1's protein expression profile in humans. *Journal of Leukocyte Biology*, 97(June):1–10.

Schrödinger, L. (2010). The PyMOL Molecular Graphics System (unpublished). *URL: https://www.pymol.org/ (Accessed July 2014)*.

Schwartz, O., Maréchal, V., Le Gall, S., Lemonnier, F., and Heard, J. M. (1996). Endocytosis of major histocompatibility complex class I molecules is induced by the HIV-1 Nef protein. *Nature medicine*, 2:338–342.

Schwefel, D., Boucherit, V. C., Christodoulou, E., Walker, P. A., Stoye, J. P., Bishop, K. N., and Taylor, I. A. (2015). Molecular determinants for recognition of divergent SAMHD1 proteins by the lentiviral accessory protein Vpx. *Cell host & microbe*, 17(4):489–99.

Schwefel, D., Groom, H. C. T., Boucherit, V. C., Christodoulou, E., Walker, P. A., Stoye, J. P., Bishop, K. N., and Taylor, I. A. (2014). Structural basis of lentiviral subversion of a cellular protein degradation pathway. *Nature*, 505(7482):234–8.

Seamon, K. J., Sun, Z., Shlyakhtenko, L. S., Lyubchenko, Y. L., and Stivers, J. T. (2015). SAMHD1 is a single-stranded nucleic acid binding protein with no active site-associated nuclease activity. *Nucleic acids research*, 43(13):6486–99.

Sharp, P. M. and Hahn, B. H. (2011). Origins of HIV and the AIDS pandemic. *Cold Spring Harbor perspectives in medicine*, 1(1):a006841.

Sheehy, A. M., Gaddis, N. C., Choi, J. D., and Malim, M. H. (2002). Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature*, 418(6898):646–650.

Sheehy, A. M., Gaddis, N. C., and Malim, M. H. (2003). The antiretroviral enzyme APOBEC3G is degraded by the proteasome in response to HIV-1 Vif. *Nat Med*, 9(11):1404–1407.

Smith, G. and Conway, S. (2016). opencsv version 3.7. *URL: http://opencsv.sourceforge.net/ (last accessed June 2016)*.

Snoeck, J., Fellay, J., Bartha, I., Douek, D. C., and Telenti, A. (2011). Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints. *Retrovirology*, 8:87.

Soares, A. E. R., Soares, M. A., and Schrago, C. G. (2008). Positive selection on HIV accessory proteins and the analysis of molecular adaptation after interspecies transmission. *Journal of Molecular Evolution*, 66:598–604.

Spragg, C. J. and Emerman, M. (2013). Antagonism of SAMHD1 is actively maintained in natural infections of simian immunodeficiency virus. *Proceedings of the National Academy of Sciences of the United States of America*, 110(52):21136–21141.

Srivastava, S., Swanson, S. K., Manel, N., Florens, L., Washburn, M. P., and Skowronski, J. (2008). Lentiviral Vpx accessory factor targets VprBP/DCAF1 substrate adaptor for cullin 4 E3 ubiquitin ligase to enable macrophage infection. *PLoS Pathog*, 4(5):e1000059.

St Gelais, C., de Silva, S., Amie, S. M., Coleman, C. M., Hoy, H., Hollenbaugh, J. A., Kim, B., and Wu, L. (2012). SAMHD1 restricts HIV-1 infection in dendritic cells (DCs) by

dNTP depletion, but its expression in DCs and primary CD4+ T-lymphocytes cannot be upregulated by interferons. *Retrovirology*, 9(105).

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

Stewart, S. A., Poon, B., Jowett, J. B., and Chen, I. S. (1997). Human immunodeficiency virus type 1 Vpr induces apoptosis following cell cycle arrest. *Journal of virology*, 71(7):5579–92.

Stivahtis, G. L., Soares, M. a., Vodicka, M. a., Hahn, B. H., and Emerman, M. (1997). Conservation and host specificity of Vpr-mediated cell cycle arrest suggest a fundamental role in primate lentivirus evolution and biology. *Journal of virology*, 71(6):4331–4338.

Strebel, K. (2014). HIV-1 Vpu an ion channel in search of a job. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1838(4):1074–1081.

Talevich, E., Invergo, B. M., Cock, P. J., and Chapman, B. A. (2012). Bio.Phylo: A unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics*, 13:209.

Tamuri, A. U., dos Reis, M., and Goldstein, R. a. (2012). Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics*, 190(3):1101–15.

Tamuri, A. U., Dos Reis, M., Hay, A. J., and Goldstein, R. a. (2009). Identifying changes in selective constraints: host shifts in influenza. *PLoS computational biology*, 5(11):e1000564.

Tang, C., Ji, X., Wu, L., and Xiong, Y. (2015). Impaired dNTPase Activity of SAMHD1 by Phosphomimetic Mutation of T592. *The Journal of biological chemistry*, 290(44):26352–26359.

Tavare, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. In *American Mathematical Society: Lectures on Mathematics in the Life Sciences*, volume 17, pages 57–86. Applied Math.

Thali, M., Bukovsky, A., Kondo, E., Rosenwirth, B., Walsh, C. T., Sodroski, J., and Gottlinger, H. G. (1994). Functional association of cyclophilin A with HIV-1 virions. *Nature*, 372(6504):363–365.

Thiltgen, G., dos Reis, M., and Goldstein, R. A. (2016). Finding Direction in the Search for Selection. *Journal of Molecular Evolution*, pages 1–12.

Trapero-Bertran, M. and Oliva-Moreno, J. (2014). Economic impact of HIV/AIDS: a systematic review in five European countries. *Health economics review*, 4(1):15.

Trautz, B., Pierini, V., Wombacher, R., Stolp, B., Chase, A. J., Pizzato, M., and Fackler, O. T. (2016). The Antagonism of HIV-1 Nef to SERINC5 Particle Infectivity Restriction Involves the Counteraction of Virion-Associated Pools of the Restriction Factor. *Journal of Virology*, 90(23):10915–10927.

Tristem, M., Marshall, C., Karpas, A., Petrik, J., and Hill, F. (1990). Origin of vpx in lentiviruses. *Nature*, 347:341–342.

UNAIDS (2013). *GLOBAL REPORT: UNAIDS report on the global AIDS epidemic 2013*. UNAIDS.

UNAIDS (2016). Global AIDS Update.

Usami, Y., Wu, Y., and Gottlinger, H. G. (2015). SERINC3 and SERINC5 restrict HIV-1 infectivity and are counteracted by Nef. *Nature*, 526(7572):218–223.

Van Damme, N., Goff, D., Katsura, C., Jorgenson, R. L., Mitchell, R., Johnson, M. C., Stephens, E. B., and Guatelli, J. (2008). The Interferon Induced Protein BST-2 Restricts HIV-1 Release and Is Downregulated from the Cell Surface by the Viral Vpu Protein. *Cell Host and Microbe*, 3:245–252.

Van Heuverswyn, F., Li, Y., Neel, C., Bailes, E., Keele, B. F., Liu, W., Loul, S., Butel, C., Liegeois, F., Bienvenue, Y., Ngolle, E. M., Sharp, P. M., Shaw, G. M., Delaporte, E., Hahn, B. H., and Peeters, M. (2006). Human immunodeficiency viruses: SIV infection in wild gorillas. *Nature*, 444(7116):164.

Van Valen, L. (1973). Van Valen.1973 - Evol Theor.pdf. *Evolutionary Theory*, 1:1–30.

VandeWoude, S. and Apetrei, C. (2006). Going wild: Lessons from naturally occurring T-lymphotropic lentiviruses.

Veeramachaneni, V., Makałowski, W., Galdzicki, M., Sood, R., and Makałowska, I. (2004). Mammalian overlapping genes: The comparative perspective. *Genome Research*, 14(2):280–286.

Vigan, R. and Neil, S. J. D. (2010). Determinants of tetherin antagonism in the transmembrane domain of the human immunodeficiency virus type 1 Vpu protein. *Journal of virology*, 84:12958–12970.

Villet, S., Bouzar, B. A., Morin, T., Verdier, G., Legras, C., and Chebloune, Y. (2003). Maedi-visna virus and caprine arthritis encephalitis virus genomes encode a Vpr-like but no Tat protein. *Journal of virology*, 77(17):9632–9638.

Wei, X. and Zhang, J. (2015). A simple method for estimating the strength of natural selection on overlapping genes. *Genome biology and evolution*, 7(1):381–90.

Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution*, 18(5):691–699.

White, T. E., Brandariz-Nuñez, A., Valle-Casuso, J. C., Amie, S., Nguyen, L. A., Kim, B., Tuzova, M., and Diaz-Griffero, F. (2013). The retroviral restriction ability of SAMHD1, but not its deoxynucleotide triphosphohydrolase activity, is regulated by phosphorylation. *Cell Host and Microbe*, 13(4):441–451.

Willey, R. L., Maldarelli, F., Martin, M. A., and Strebel, K. (1992). Human immunodeficiency virus type 1 Vpu protein induces rapid degradation of CD4. *Journal of virology*, 66(12):7193–200.

Wittlich, M., Koenig, B. W., Stoldt, M., Schmidt, H., and Willbold, D. (2009). NMR structural characterization of HIV-1 virus protein U cytoplasmic domain in the presence of dodecylphosphatidylcholine micelles. *FEBS Journal*, 276(22):6560–6575.

Wong, W. S. W., Yang, Z., Goldman, N., and Nielsen, R. (2004). Accuracy and Power of Statistical Methods for Detecting Adaptive Evolution in Protein Coding Sequences and for Identifying Positively Selected Sites 10.1534/genetics.104.031153. *Genetics*, 168(2):1041–1051.

Yan, J., Hao, C., DeLucia, M., Swanson, S., Florens, L., Washburn, M. P., Ahn, J., and Skowronski, J. (2015). CyclinA2-Cyclin-dependent Kinase Regulates SAMHD1 Protein Phosphohydrolase Domain. *The Journal of biological chemistry*, 290(21):13279–92.

Yan, J., Kaur, S., DeLucia, M., Hao, C., Mehrens, J., Wang, C., Golczak, M., Palczewski, K., Gronenborn, A. M., Ahn, J., and Skowronsk, S. (2013). Tetramerization of SAMHD1 is required for biological activity and inhibition of HIV infection. *Journal of Biological Chemistry*, 288(15):10406–10417.

Yang, W., Bielawski, J. P., and Yang, Z. (2003). Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. *Journal of Molecular Evolution*, 57:212–221.

Yang, Z. (2006). *Computational molecular evolution*. Oxford University Press, Oxford.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–91.

Yang, Z. and Dos Reis, M. (2011). Statistical properties of the branch-site test of positive selection. *Molecular Biology and Evolution*, 28(3):1217–1228.

Yang, Z., Lauder, I. J., and Lin, H. J. (1995). Molecular evolution of the hepatitis B virus genome. *J Mol Evol*, 41(5):587–596.

Yang, Z. and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular biology and evolution*, 19:908–917.

Yang, Z. and Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular Biology and Evolution*,

25:568–579.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, a. M. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1):431–449.

Yang, Z., Wong, W. S. W., and Nielsen, R. (2005). Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution*, 22:1107–1118.

Yu, X., Yu, Y., Liu, B., Luo, K., Kong, W., Mao, P., and Yu, X.-F. (2003). Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex. *Science*, 302(5647):1056–1060.

Zhang, J., Nielsen, R., and Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution*, 22:2472–2479.

Zhao, G., Perilla, J. R., Yufenyuy, E. L., Meng, X., Chen, B., Ning, J., Ahn, J., Gronenborn, A. M., Schulten, K., Aiken, C., and Zhang, P. (2013a). Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature*, 497(7451):643–646.

Zhao, K., Du, J., Han, X., Goodier, J. L., Li, P., Zhou, X., Wei, W., Evans, S. L., Li, L., Zhang, W., Cheung, L. E., Wang, G., Kazazian, H. H., and Yu, X. F. (2013b). Modulation of LINE-1 and Alu/SVA Retrotransposition by Aicardi-Goutières Syndrome-Related SAMHD1. *Cell Reports*, 4(6):1108–1115.

Zhao, L. J., Mukherjee, S., and Narayan, O. (1994). Biochemical mechanism of HIV-I Vpr function. Specific interaction with a cellular protein. *Journal of Biological Chemistry*, 269(22):15577–15582.

Zhou, L., Sokolskaja, E., Jolly, C., James, W., Cowley, S. A., and Fassati, A. (2011). Transportin 3 promotes a nuclear maturation step required for efficient HIV-1 integration. *PLoS Pathogens*, 7(8).