Bayesian Phylogenetic Approaches to Retroviral Evolution: Recombination, Cross-Species Transmission, and Immune Escape



N.C. Kist

Lincoln College

University of Oxford

Supervised by Dr Astrid Iversen

Submitted in partial fulfillment of the requirements for the degree of:

Doctor of Philosophy (Zoology)

Trinity 2017

"I have nothing to offer but blood, toil, tears, and sweat."

— Winston Churchill

Acknowledgements

Where to begin but at the beginning? I would like to thank my parents for introducing me to computing, biology, and medicine; for always encouraging my interests and believing in my abilities; and for demonstrating — every day of the week — how the good life is lived.

I would also like to thank the following for their assistance, without which this thesis could never have been written:

- My supervisor, Dr Astrid Iversen, for her mentorship, for the brilliant biological intuition that led to this project, and especially for her boundless enthusiasm even when things weren't going to plan. There are no better supervisors.
- My departmental supervisor, Prof Adrian Thomas, for facilitating the move to Dr Iversen's group and supporting the final phases of my DPhil.
- Dr Daniel Lunn for his patient teaching, for his statistical support of Chapters One, Two, Three, and Four, and especially for the formula used in the change-point analysis.
- Mr Ben Lambert for our discussions, in particular for his help with the multiple response case in Chapter One and the hierarchical model and posterior predictive checks in Chapter Three.
- Prof Philippe Lemey for our discussions about Chapters One, Two, and Three; and especially for his help calculating the Bayes Factors in Chapter Three.
- Prof Mark Pagel for our meeting about Chapters One, Two, and Three.
- Dr Chih-Hao Chang for supplying the sequence data used in Chapters Four and Five.
- Mr Daniel Rowe, Mr Kevin Ray, Dr Leah Rand, Ms Katie Allan, Mr Ben Lambert, and Mr Sam Campbell for proofreading.
- The women of the Pumwani cohort for their samples.
- The Scatcherd Foundation, EPSRC, Royal Society, and the Hendrik Muller Vaderlandsch Fonds for their financial support.

One does not write a thesis with academic support alone. I would like to thank Mrs Heather Green for shepherding me through the seemingly endless GSO forms, going along with my get-funding-quick schemes, and for many games of tennis; Mrs Carmella Gaston for her assistance in all matters; Mr Simon Faulkner and Ms Laura Perrin for many half-pints, and for their general sense of sunny optimism.

This leaves rather little space for a personal note to my friends, who have supported me throughout this DPhil. It has been the privilege of my life to spend my time in Oxford with you. All my friends at Lincoln, the University, and beyond: I salute you!

Abstract

Since human immunodeficiency virus type 1 (HIV-1) was discovered over thirty years ago, much effort has been expended developing a vaccine. Unfortunately, all randomised clinical trials have failed except one, which may have shown a slight beneficial result. In this thesis, I study the CD8+ cytotoxic T Lymphocyte (CTL) immune response against p24 group-specific antigen (Gag), a prime candidate for vaccine boosting. This immune response is associated with long-term survival in some patients with beneficial HLA variants. Independently of HLA variant, the number of Gag epitopes targeted by the patient's CTL response is associated with lower viral loads.

In the first three chapters, I considered the HIV-1M subtype-specific sites; the current paradigm is that these differences between the consensus sequences of HIV-1M subtypes are the result of founder effects. However, I use Chapter Two to show that these sites are strongly associated with the patients' HLA variants, suggesting a role for CTL selection in subtype-diversification through epitope processing escape.

In Chapter Three, I show that as the United States HIV-infected population became increasingly African-American over the last two decades, the amino acids in the subtype-specific sites became less typical of HIV-1M subtype B, and appeared to become increasingly similar to HIV sequences found in Africa. Together with the results from Chapter Two this suggests that viral adaptation to population HLA frequencies may drive subtype diversification.

In Chapter Four, I show that patients with beneficial HLA variants select for the outgrowth of HIV-1 recombinants, possibly because recombination allows for large fitness jumps and the large-scale, rapid spreading of escape and compensatory mutations. In Chapter Five, I show how repeated recombinations could lead to HIV-1 outgrowths in the presence of beneficial HLA types. If viral recombination can abrogate the natural CTL response, it is likely that a vaccine-boosted immune response can also be evaded. Because recombination over large phylogenetic distances is a likely avenue towards HIV vaccine escape, I recommend that patients exposed to many strains should be monitored closely.

In Chapter Six I use a clade of betaretroviruses, which are related to HIV, to study viral cross-species transmission between mammals over the last 40 million years. I found strong evidence for transmissions between mammals separated by large geographical and phylogenetic distances, and evidence suggesting that rodents play a central role in the transmission of betaretroviruses, and that host phylogenetic relatedness is associated with cross-species transmission in this clade.

Contents

Introduction 1
Evolutionary History of HIV . . . 3
HIV-Host Interaction . . . 5
Recombination . . . 12
Phylogenetics . . . 14
This Thesis . . . 21

1 Validating the Multiple Response Bayesian Phylogenetic Logistic Regression Using Simulations 23
1.1 Introduction . . . 23
1.2 Methods . . . 26
1.2.1 Notation and terminology . . . 26
1.2.2 Obtaining the correlation matrix . . . 28
1.2.3 Simulations . . . 29
1.2.4 Fitting models to the simulated data . . . 29
1.3 Results . . . 30
1.3.1 Uncorrected Linear Regression is Not Anti-Conservative when both x and y are Phylogenetically Informed . . . 30
1.3.2 Bernoulli Model with Phylogenetic Random Effect . . . 31
1.3.3 Allowing for Multiple Amino Acid Positions . . . 32
1.4 Conclusions & Discussion . . . 36
1.5 Model Code . . . 39
1.5.1 Gaussian Phylogenetic Regression Stan Model . . . 39
1.5.2 Binomial Phylogenetic Regression Stan Model . . . 40
1.5.3 Binomial Phylogenetic Regression Stan Model, Multiple Response . . . 41

2 HLA-Associated CTL Selection on HIV-1M Subtype-Specific Sites in Gag p24 42
2.1 Introduction . . . 42
2.2 Methods . . . 43
2.2.1 Obtaining the Data . . . 43
2.2.2 Exploratory Principal Component Analysis . . . 44
2.2.3 Exploratory Amino Acid Multiple Correspondence Analysis . . . 45
2.2.4 Modelling the Subtype-Specific Sites . . . 45
2.2.5 Generating the Phylogeny and the Phylogenetic Variance-Covariance Matrix 49
2.2.6 Cross-Validation . . . 49
2.2.7 Parsimonious Model . . . 51
2.2.8 Goodness of Fit Testing . . . 51
2.3 Results . . . 52
2.3.1 Principal Component Analysis . . . 52
2.3.2 Amino Acid Multiple Correspondence Analysis. . . . 54

2.3.3 Phylogenetic Trees . . . 54
2.3.4 Bayesian Regressions . . . 58
2.4 Discussion . . . 71

3 Ethno-Demographic Changes in the American HIV-Infected Population Associated with Changing Patterns of CTL Escape in Subtype-Specific Sites 75
3.1 Introduction . . . 75
3.2 Methods . . . 77
3.2.1 Building a Posterior Distribution of Phylogenies . . . 77
3.2.2 Bayesian Model Comparison with BEAST . . . 77
3.2.3 Calculating HLA Frequencies in the American HIV-infected Population Using a Simple Population Model . . . 78
3.2.4 Regression of the HIV-infected Population HLA Frequencies on Time . . . 79
3.2.5 Bayesian Multi-Level Model Incorporating Phylogenetic Information . . . 80
3.2.6 Estimating the Phylogenetic Variance-Covariance Matrix **C** . . . 83
3.2.7 Model Validation using a Posterior Predictive Check . . . 83
3.2.8 Population-Level Estimates of Subtype B Consensus Amino Acid Frequencies 84
3.3 Results . . . 84
3.3.1 BEAST Phylogeny . . . 84
3.3.2 Beast Model Comparison . . . 86
3.3.3 Bayesian Regressions . . . 87
3.3.4 Half-Normal Plot . . . 88
3.3.5 Posterior Predictive Plots . . . 88
3.4 Discussion . . . 93

4 HIV-infected Sex Workers with Beneficial HLA-Variants are Potential Hubs for Selection of HIV-1 Recombinants that may Affect Disease Progression 95
4.1 Introduction . . . 95
4.2 Methods . . . 97
4.2.1 Study Participants . . . 97
4.2.2 Generation of the HIV-1 Sequences . . . 97
4.2.3 Assignment of HLA Variants as "Beneficial", "Detrimental" and "Neutral" . 98
4.2.4 HIV-1 Subtyping and Analyses of Recombinant HIV-1. . . . 98
4.2.5 Subsequent Statistical Analyses . . . 101
4.3 Results . . . 103
4.3.1 HIV Subtype Distribution and Recombination Analyses . . . 103
4.3.2 Overrepresentation of Beneficial HLA Variants in URF infected women . . 105
4.3.3 Reduced Protective Effect from Beneficial HLA variants in URF-Infected Women . . . 108
4.4 Discussion . . . 110
4.5 Data Tables . . . 112

5 Repeated Recombination and Decades-Long Absence of Superinfection in Two Dually HIV-Infected High-Risk Female Commercial Sex Workers 117
5.1 Introduction . . . 117
5.2 Methods . . . 118
5.2.1 HIV-1 Hypermutation and Subtype Analyses . . . 119

5.3 Results . . . 120
5.3.1 Patient ML1812 . . . 120
5.3.2 Patient ML900 . . . 123
5.3.3 Epitope Evolution in ML900 and ML1812 . . . 126
5.4 Discussion . . . 127
5.4.1 Repeated Outgrowth of Identical Recombinant Forms of Different Lineages 127
5.4.2 Absence of Further Superinfection Beyond Initial Dual-Infection . . . 128

6 Paleovirology of Mammalian Betaretroviruses Reveals Ancient and Intercontinental Viral Flow 129
6.1 Introduction . . . 129
6.2 Results & Discussion . . . 132
6.2.1 Does Host Phylogenetic Relatedness Affect Cross-Species Transmission? . 133
6.2.2 Inferring Betaretroviral Cross-Species Transmissions Between Mammals . 135
6.2.3 Centrality of the Rodents . . . 137
6.2.4 Absence of Bats from the Viral Network . . . 138
6.2.5 Intercontinental Viral Flow . . . 139
6.3 Concluding Remarks . . . 144
6.4 Methods . . . 145
6.4.1 Querying the WGS Database . . . 145
6.4.2 Detection of Orthologous Sequences . . . 149
6.5 Viral Networks . . . 149
6.6 Beast's Generalized Linear Model . . . 150
6.7 Dating Insertions . . . 151

Conclusions 153
Subtype-specific sites . . . 153
Implications for the Evolutionary History of HIV . . . 154
Recombination and Vaccine Escape . . . 155
Recommendations for Vaccine Development . . . 156
Future work . . . 157

Citations 160

List of Abbreviations

AIDS	Acquired Immune Deficiency Syndrome
ART	Anti-Retroviral Therapy
AUC	Area Under the Curve
BCI	Bayesian Credible Interval
CCR5	C-C Chemokine Receptor type 5
CD4	Cluster of Differentiation 4
CD8	Cluster of Differentiation 8
CRF	Circulating Recombinant Form
CTL	Cytotoxic T Lymphocyte
FPR	False Positive Rate
GLM	Generalised Linear Model
GTR	Generalised Time-Reversible
GWAS	Genome-Wide Association-Study
HAART	Highly Active Retroviral Therapy
HIV-1M	Humman Immunodeficiency Virus 1 (Group M)
HLA	Human Leukocyte Antigen
HPD	Highest Posterior Density
i.i.d	Independent and Identically Distributed
IDU	Intravenous Drug User
K80	Kimura 1980
LANL	Los Alamos National Laboratory
LCA	Last Common Ancestor
LOESS	LOcal regrESSion
LTNP	Long-Term Non-Progressor
LTS	Long-Term Survivor
MRCA	Most Recent Common Ancestor
MSM	Men who have Sex with Men
nr	NCBI Protein Database
nt	NCBI Nucleotide Database
PrEP	Pre-Exposure Prophylaxis
ROC	Receiver-Operator Curve
RP	Rapid Progressor
SIV	Simian Immunodeficiency Virus
SLP	Slow Progressor
SP	Standard Progressor

URF	Unique Recombinant Form
WGS	Whole Genome Sequencing
WHO	World Health Organisation
ZIKAV	Zika Virus

Introduction

The pandemic strain of human immunodeficiency virus type 1, Major (M) group (HIV-1M), first emerged in what is now the Democratic Republic of the Congo in the early twentieth century (Salemi et al. 2001; Worobey et al. 2008; Faria et al. 2014). Since then HIV-1M has undergone a major expansion and spread throughout the world, becoming a major public health challenge. No vaccine currently exists, but highly effective antiretroviral therapy (HAART) can turn a patient's HIV infection from a terminal and deadly disease into a chronic illness. However, today only 46% of the 37 million people infected with HIV are receiving antiretroviral therapy (ART) (WHO 2016). Delivering ART is challenging because it requires a sophisticated healthcare infrastructure which is often missing in the politically and economically unstable regions most affected by the epidemic. This instability is damaging because fully suppressing HIV using these drugs requires excellent adherence. In addition, consistently delivering ART requires an economy that allows for the acquisition of the expensive drugs. Even in a resource-rich nation like the United States where 86% of HIV-infected individuals are diagnosed, only 30% of HIV-infected individuals are able to manage their care to the point where their viral load is undetectable (Kay, Batey and Mugavero 2016).

Recently pre-exposure prophylactic treatments (PrEP) have started to become available (Karim and Karim 2011). PrEP is prescribed to uninfected individuals at high risk of acquiring HIV, such as particularly active Men who have Sex with Men (MSM), intravenous drug users (IDU), and the uninfected partners of HIV-infected patients (World Health Organization 2012). PrEP treatment is stopped when individuals transition out of these risk groups, whether due to settling down, quitting drugs, leaving their HIV-infected partner, or when they become infected in spite of the treatment. These drugs are more expensive than conventional antiretroviral therapy (ART) and must be prescribed to healthy people, making them an unattractive choice for developing countries, especially those where heterosexual women make up the majority of newly infected patients (Ramjee et al.

2013). Beyond the expense, adherence has been a problem in some PrEP studies (Baeten et al. 2014). One such study found only 12% of patients were able to maintain drug concentrations consistent with adherence, and also that 'liking the pill colour' was positively associated with adherence (OR: 2.93, CI: 1.18 – 7.27) (Corneli et al. 2014).

Governments, non-governmental organisations, and civil society have achieved tremendous success fighting other diseases using vaccinations, which can be administered once and do not require nearly as much infrastructure as pharmacological compounds which must be taken daily. Vaccinations providing long-term immunity do not have issues with treatment adherence and are much more robust to changes in the availability of medical care and disrupted supply lines, which can occur when civil strive or emerging diseases like Ebola deteriorate existing medical infrastructure (Parpia et al. 2016). Beyond a traditional prophylactic vaccine, some have suggested that the creation of a therapeutic vaccine which would replace the need for daily antiretroviral therapy (ART) or even sterilise HIV in infected patients (Walensky et al. 2004). Whether prophylactic or therapeutic, it is apparent that a vaccine is necessary to end the HIV pandemic in a reasonable time-frame (Fauci, Folkers and Marston 2014; Fauci and Marston 2014). Developing a safe and effective HIV vaccine remains a high public health priority.

In the over thirty years since HIV was discovered to be the etiological agent of Acquired Immune Deficiency Syndrome (AIDS) (Gottlieb et al. 1981; Barré-Sinoussi et al. 1983), millions of dollars have been spent on developing a vaccine with little success (McMichael and Haynes 2012). AIDSVAX was the first vaccine to make it to trial. This recombinant HIV-1M envelope glycoprotein subunit was meant to induce an antibody response, but when tested in randomised controlled clinical trials it had no clinical effect in Thailand (Pitisuttithum et al. 2006) or in America (Flynn et al. 2005), though it did boost the antibody response (Gilbert et al. 2005). The RV144 trial in Thailand, which vaccinated subjects with the ALVAC-HIV canarypox vector vaccine combined with the previously ineffective AIDSVAX B/E vaccine, showed a borderline significant effect only after deviating from the pre-trial protocol (Rerks-Ngarm et al. 2009). After this deviation, the analysis yielded a small vaccine efficacy of 31.2% (p=0.04). By comparison, voluntary circumcision reduces the infection rate by 60% (Gray et al. 2012). Furthermore, a Bayesian reinterpretation of this trial showed there was a 22% posterior probability of no vaccine efficacy (Gilbert et al. 2011). However, this

Bayesian re-analysis involved assuming a prior probability of a 50% chance of vaccine effectiveness (for good or ill), more conservative priors further reduce the posterior probability that the vaccine was successful. The inefficiency of these vaccines is perhaps not surprising, as the envelope protein targeted evolves rapidly, and is not subject to severe functional constraints.

The STEP study in the Americas tested a vaccine that took another approach and attempted to boost CD4+ (cluster of differentiation 4) and CD8+ (cluster of differentiation 8) T cell response against Gag, Pol and Nef using an adenovirus type 5 vector. These proteins may make better targets given that they are under significant functional constraint compared to the envelope protein. This study considered two efficacy endpoints: whether the vaccine prevented HIV infection and whether the vaccine reduced the viral load in patients that did get infected. Unfortunately, patients who received this vaccine were more likely to become infected with HIV-1M (Duerr et al. 2012). A companion trial in South Africa was aborted because of the potential harm associated with this vaccine (Gray et al. 2012). After these failures, the HVTN 505 trial was held in the United States. This trial tested a vaccine which used a prime-boost approach where the patient was first primed with a DNA vaccine and then boosted with an Adenovirus 5 booster vaccination (using a different adenovirus carrier and more viral fragments than the STEP trial). Unfortunately, this vaccine was also shown to be ineffective (Hammer et al. 2013).

These failed vaccines show that more work is needed so that we can understand the fundamental biology of the system before engaging in further large and expensive trials. In this thesis, we explore the effects CD8+ cytotoxic T Lymphocyte (CTL) responses against the highly constrained Gag protein have on viral evolution, the outgrowth of recombinants, and HIV-1M subtype diversification. CTL responses against Gag are uniquely associated with delayed disease progression (Kiepiela et al. 2007).

This information might help inform CTL based approaches to HIV-1M vaccine design.

Evolutionary History of HIV

The pandemic strain of HIV, HIV-1M, crossed into humans from chimpanzees (Gao et al. 1999) at the beginning of the twentieth century (Korber et al. 2000). Since then it has undergone adaptation

to humans (Wain et al. 2007; Bibollet-Ruche et al. 2012) and spread through the population. Phylogenies of the SIVs and HIVs have shown that four lineages of HIV-1 and eight lineages of HIV-2 crossed into humans, while only HIV-1M became a pandemic. The four HIV-1 lineages are known as HIV-1 group M (HIV-1M), HIV-1 group Non-M, Non-O (N), HIV-1 group Outlier (O), and HIV-1 group "Pending the identification of further human cases" (P), each of which represents a single independent cross-species transmission from primates, like the chimpanzee (Hahn et al. 2000; Sharp and Hahn 2011) (group M and group N), the western lowland gorilla (D'arc et al. 2015) (group O), it is not clear whether group P came from gorillas or chimpanzees (Plantier et al. 2009; Sharp and Hahn 2011). However, not all have adapted as successfully to their new hosts. HIV-2 is present almost exclusively in West Africa, where it has infected ~1 million people (Silva, Cotten and Rowland-Jones 2008; Silva et al. 2012). HIV-1 group N and HIV-1 group O are present mostly in Cameroon and Gabon, where they are associated with a small minority of HIV infections (Vallari et al. 2010; Liegeois et al. 2013). HIV-1 group P has only been described in two patients thus far (Plantier et al. 2009; Vallari et al. 2011).

The natural history of the HIV/SIV clade of viruses raises the question of why 12 HIV strains crossed into humans in the last century, while humans and SIV infected non-human primates have been interacting in Africa for many hundreds of thousands of years before then (Peeters et al. 2002). The fact that modern bushmeat hunters are exposed to and infected with spillover retroviruses from non-human primates (Wolfe et al. 2004; Kalish et al. 2005) suggests that there must have been many spillover events before the successful spread of HIV-1M. The virus only started spreading when socio-economic changes such as colonisation and the construction of roads allowed the virus to spread beyond the bushmeat hunters and their villages, and when urbanisation and changes in sexual behaviour caused the basic reproductive number of HIV in humans to increase dramatically (Morse 1995). Indeed, in Guinea-Bissau, the population dynamics of HIV-2 are similar to those of HIV-1, which suggests host behaviour is more important than biological differences between the viruses (Silva et al. 2012). In contrast, in Kinshasa, the HIV-1M and HIV-1 group O lineages experienced a similar epidemiological trajectory until the nineteen sixties when HIV-1M underwent explosive growth and began to outpace HIV-1 group O (Faria et al. 2014). Therefore, the exact reason why HIV-1M became the pandemic strain rather than any of the others remains unclear.

The worldwide spread of HIV-1M caused the diversification of this group into several subtypes, which are approximately phylogenetically equidistant and prevalent in different parts of the world. Patients in the United States, Europe, and Australia are primarily infected with subtype B; Subtype A is common in East Africa, West Africa, Central Africa, Eastern Europe, and Central Asia. Subtype C is prevalent in India, Ethiopia, and Southern Africa. Subtype D frequently occurs in North Africa/the Middle East and circulates in East and Central Africa, and CRF01_AE circulates in South-East Asia (McCutchan 2006; Hemelaar et al. 2006). HIV-1M subtypes can be distinguished by their subtype-specific sites: sites where the amino acid consensus sequence differs between subtypes. The current paradigm is that differences between these subtypes arose through genetic drift during the rapid expansion of the HIV pandemic (Holmes et al. 1995; Peeters and Sharp 2000). However, there is also evidence for population-level viral adaptation to the frequency of certain Human Leukocyte Antigen (HLA) types in each host population (Carlson et al. 2016). This is best illustrated by subtype B, which evolved in populations of mostly European ancestry, and evolved subtype-C-like Gag subtype-specific sites when it was transmitted into Afro-Caribbean populations (Tenzer et al. 2014). In Chapters Two and Three we study the role immune selection plays in determining the amino acids on subtype-specific sites.

HIV-Host Interaction

HIV-1 is a retrovirus of the lentiviral family and consists of approximately 10,000 basepairs of RNA, which code for three polyproteins (Gag, Pol, and Env) and several accessory proteins (fig. 1). HIV-1 *gag* codes for several structural proteins, *pol* contains the reverse transcription machinery and integrase, and *env* codes for the highly mutable envelope protein, which provides an outside layer to the virion. Retroviruses can be transmitted through the transfer of bodily fluids such as blood, semen or breastmilk, and also spread from mother to child during pregnancy and labour. The majority of newly HIV-infected patients today acquire the disease through heterosexual intercourse, even though only 1 in 80 heterosexual coital acts lead to infection, even during the most infectious phase (Wawer et al. 2005). Depending on infectee sex, concurrent genital infections, and mode of transmission, HIV can experience a strong bottleneck (Zhu et al. 1996) — 78% of subtype C heterosexual infections were shown to have started with only a single effective virus (Keele et al. 2008; Abrahams et al.

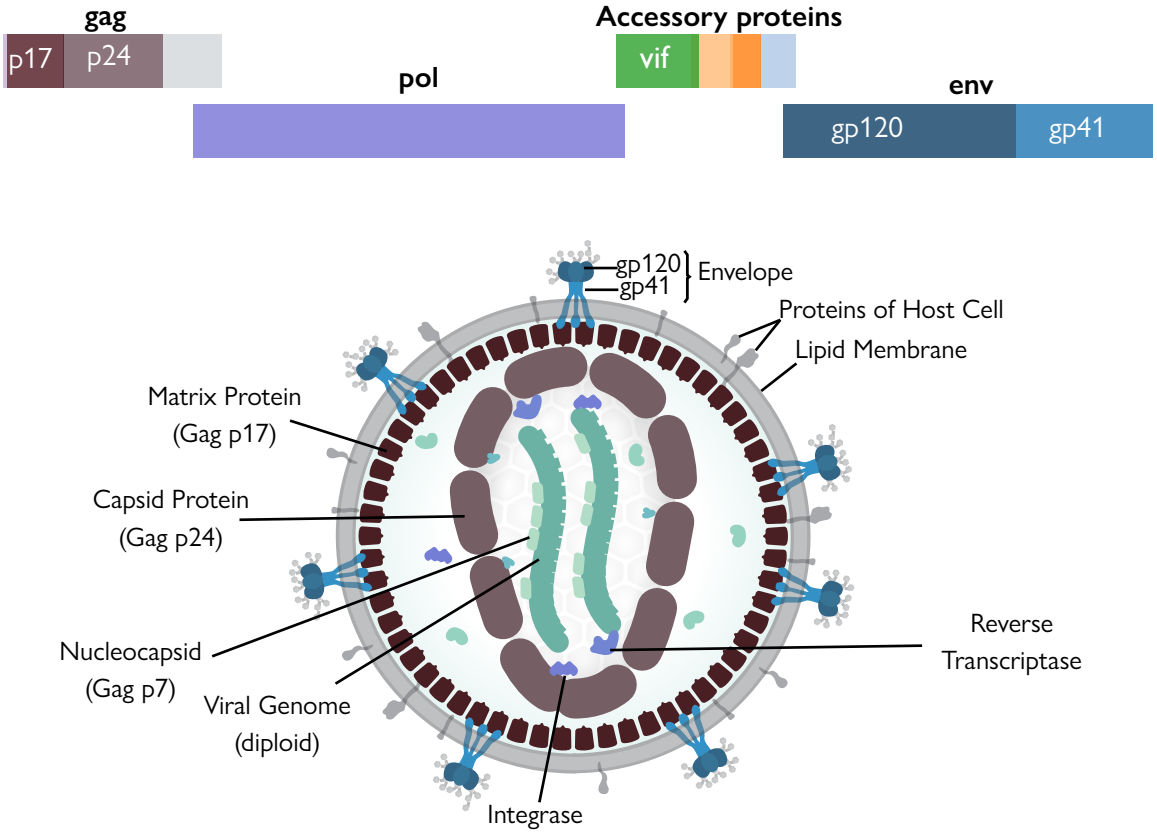


Figure 1: **The HIV genome and virion structure**
The HIV genome is 10kb long and consists of three large genes (*gag*, *pol*, *env*) coding for polyproteins as well as several smaller accessory proteins. Figure adapted with permission from www.scistyle.com

2009). The mucosal barrier may be in part responsible for this bottleneck, with animal experiments showing that intravenous transmission of SIV is 2,000-20,000 times more efficient than intra-rectal transmission (Keele et al. 2009). When HIV-infected patients' mucosal barrier is disturbed by an inflammatory infection, patients are more likely to become co-infected with additional viral lineages unrelated to their original infection. This shows that the effectiveness of the mucosal barrier can be compromised (Haaland et al. 2009). An analysis of 171 subtypes B and C transmission events showed that the count of virions starting an infection does not follow a Poisson distribution; therefore the transmission of individual virions cannot be seen as a series of events occurring at some low probability (Abrahams et al. 2009). The non-Poissonness of the distribution could imply that there is some disparity between patients who get infected with only one virion and patients in whom a breakdown of the mucosal barrier allows for infection with multiple virions.

Some have suggested that activities leading to high infection risks such as intravenous drug use and anal receptive sex are more likely to result in multiple infections due to their circumvention of cervicovaginal mucosal barriers (Li et al. 2010). However, another study has shown that MSM, heterosexuals, and IDUs (Intravenous Drug Users) are equally likely to become multiply infected (Tully et al. 2016). A study of patients infected by the IDU route, which circumvents the mucosal barriers, showed IDU patients were also likely to have been infected with only one virus (Masharsky et al. 2010), however, this could have been due to the small inoculum compared to sexual transmission. In another cohort of IDU patients, 60% were co-infected with multiple strains of HIV, but this could have been due to them acquiring an additional HIV infection (superinfection) (Bar et al. 2010).

The life-cycle of HIV consists of a virion infecting a $CD4+$ T-cell by binding with the cell's $CD4$ and $CCR5$ (chemokine receptor type 5) surface proteins. $CD4$ is found on immune cells such as T helper cells, monocytes/macrophages and dendritic cells, but T helper cells are the prime target for HIV-1M. These cells are central in the orchestration of both cellular and humoral immune responses as they help activate CTL and macrophages, and stimulate B cells to secrete antibodies. If the binding is successful, the viral membrane will merge with the host cell membrane and the virion's RNA and replication machinery payload will enter the cell. The RNA is then reverse-transcribed into cDNA and integrated into the host DNA. From there, the integrated provirus is expressed, and the host

cell's machinery is recruited to construct new virions. HIV virions have a half-life of only 2.1 ± 0.4 days (Ho et al. 1995), which means an active HIV infection causes a severe depletion of CD_4+ cells. This depletion disrupts the proper functioning of the immune system, which eventually leads to AIDS (Mohri et al. 2001). Though this molecular process is unique to HIV, other retroviruses have been infecting mammals in similar ways for millions of years (see Chapter Six).

APOBEC3G

The association between retroviruses and their vertebrate hosts throughout the last 450 million years (Aiewsakun and Katzourakis 2017) has resulted in the evolution of several immune pathways which vertebrates use to suppresses viral replication, as well as counter-adaptations which the virus uses to escape suppression. Primates carry the gene APOBEC$_3$G, a member of the APOBEC family of host restriction factors (Harris and Liddament 2004), which associates with the viral RNA (Khan et al. 2005) and is incorporated in the budding virion. The cDNA product of reverse transcriptase is then deaminated in the cytoplasm of the target cell inducing CC->CU hypermutation of the cDNA (Harris et al. 2003; Zhang and Webb 2004), which when translated appears as a GG->GA mutation in the viral RNA (Yu et al. 2004). APOBEC$_3$G is the only member of the APOBEC family that preferentially causes GG to GA mutations. This makes it particularly effective as it can turn tryptophan codons (TGG) into stop codons (TGA). In humans, this restriction factor is effective against the retroviruses HIV (Zhang et al. 2003), HTLV-1 (Strebel 2005; Fan et al. 2010) and HERV-K (Lee, Malim and Bieniasz 2008; Armitage et al. 2008). As a counter to this form of innate immunity, the primate lentiviral lineage (including HIV) evolved the *vif* gene (Sheehy et al. 2002). The protein encoded by this gene binds to APOBEC$_3$G and targets it for proteasomal degradation (Conticello, Harris and Neuberger 2003) thereby excluding APOBEC$_3$G particles from the budding virion (Mariani et al. 2003). Studies on the population level have shown that patients carrying hypermutated HIV sequences have reduced viral loads (Pace et al. 2006), which is a key measure of HIV disease progression. The long-term co-evolution between primates and retroviruses has resulted in strong positive selection on the APOBEC$_3$G and other DNA editing genes (Sawyer, Emerman and Malik 2004).

Cytotoxic T Lymphocyte Response

One of the most important adaptive immune responses to HIV infection is CTL lysing of infected $CD4+$ cells. An infected $CD4+$ cell will present peptide epitopes in complex with HLA proteins to signal to CTLs that it has been infected. Following recognition of the infected cell, CTLs will release cytokines that lyse the infected $CD4+$ cell, which prevents it from shedding more HIV virions. This immune response builds during the acute infection and causes the reduction in viraemia (Gómez, Smaill and Rosenthal 1994) associated with clinical latency: macaques in which the CTL response has been deactivated due to CD8+ T cell depletion are unable to control the acute infection (Matano et al. 1998). Furthermore, the first HIV-specific CTL cells are detected during primary infection, and the increase of their numbers is paired with the decrease in plasma virus load (Borrow et al. 1994; Koup et al. 1994; Ritchie et al. 2014). CTLs can detect infected cells that exhibit HIV epitopes on their HLA molecules (fig. 2).

However, the CTL-induced selective pressure favours the outgrowth of HIV with CTL escape-mutations within or next to the target epitope (Allen et al. 2004). These escape-mutations can limit or abolish CTL recognition by limiting or abolishing binding of the epitopes to HLA, decreasing or abolishing epitope production by the cellular antigen-processing machinery or by interfering with T cell receptor recognition. This leads to a striking difference between the regions coding for epitopes in archived proviral sequences as compared to blood-borne virions (Bailey et al. 2006). Some patients maintain this immune response beyond the acute phase of the infection and use it to control HIV infection for much longer times than typical patients (Rinaldo et al. 1995; Migueles et al. 2000). Many of these patients have the relatively rare human leukocyte antigens B*27:05, B*57:01, B*57:03 — though not all patients with these HLA variants can control HIV (Dinges et al. 2010; Goulder and Walker 2012).

CTLs can only recognise and lyse HIV-infected host cells when a host cell is presenting HLA molecules in complex with short virus-derived protein fragments called epitopes (Goulder and Walker 2012). Epitopes are produced when viral proteins are digested in the cytosolic proteasome and transported by the Transporter associated with Antigen Processing (TAP) into the endoplasmatic reticulum (ER) where they are trimmed by endoplasmic reticulum aminopeptidase 1 (ERAP1) and

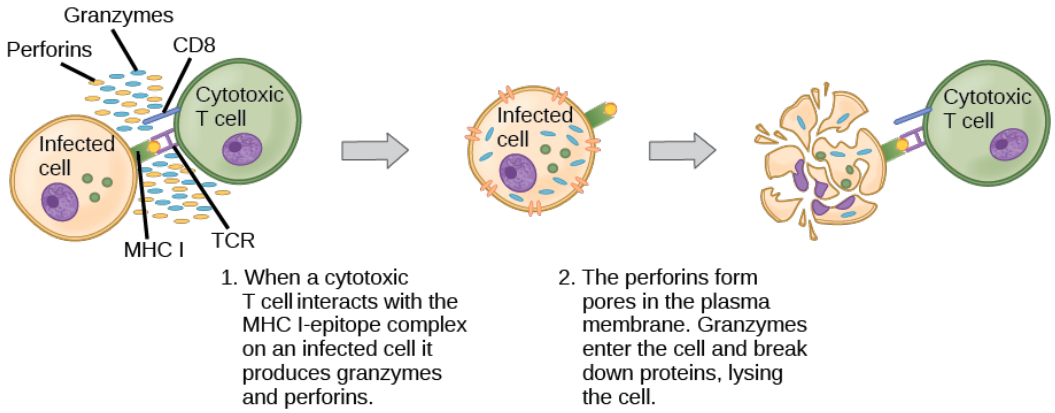


Figure 2: **CTL recognises infected cell's HLA-epitope complex** using its T-cell receptor (TCR) and lyses it using perforins and granzymes. Adapted with permission from bio.libretexts.org.

ERAP2. The trimmed epitope is then loaded into the groove of compatible HLA molecules (Tenzer et al. 2009). Theoretically one might assume that all protein fragments are equally likely to become an epitope. However, at every step of epitope processing potential effective epitopes are lost in non-random ways due to amino acid recognition preferences (Price et al. 1997; Draenert et al. 2004; Iversen et al. 2006; Tenzer et al. 2009, 2014; Bronke 2013). Amino acid recognition preferences in the epitope processing machinery limit the ability of the host cell to process certain protein fragments. These biases in the epitope processing machinery provide a mechanism for immune escape, whereby the virus changes its amino acid sequences in functionally unconstrained regions to exploit these biases. This is harder in functionally constrained regions, such as those coding for internal hydrophobic protein components which cannot escape immune selection without a fitness cost. This process leads to the clustering of produced epitopes in hydrophobic regions of the viral proteins (https://www.hiv.lanl.gov/content/immunology/tables/tables.html). Furthermore, of the epitopes which survive this process, only a few become immunodominant. Strikingly, several clusters of subdominant and dominant epitopes are found near subtype-specific sites in Gag, and a role for subtype-specific sites in the escape from epitope processing has been previously demonstrated (Tenzer et al. 2009, 2014).

The epitopes that can be presented by cells are determined the patient's HLA variants. Because the HLA region is the most diverse in the human genome, this means there are considerable differences between the epitopes each patient can present. Patients carrying HLA-B*27:05, -B*57:01, -B*57:03, or -B*58:01 (Goulder and Walker 2012) are much more likely to be long-term non-progressors (LT-

NPs) than other patients because they can present epitopes in highly conserved regions in p24 Gag. The CTL pressures then force the virus to adopt non-optimal amino acids in these crucial highly conserved regions (Streeck et al. 2007; Dinges et al. 2010; Dahirel et al. 2011). The virus thereby incurs a replicative fitness cost that leads to better patient outcomes. Beyond these relatively rare cases of HIV control, the number of CTL responses against Gag is uniquely associated with reduced viral load independent of HLA variant (Kiepiela et al. 2007). This creates strong evolutionary pressures on the virus to escape CTL recognition, which could lead to the spread of mutations that interfere with epitope processing. In Chapters Two and Three we study five subtype-specific sites in a part of the Gag best associated with clinically efficient CTL responses. Mutations in these subtype-specific sites strongly affect how cellular proteasomes process the protein region (Tenzer et al. 2014), making it plausible that mutations in these subtype-specific sites could be a road to immune escape. In Chapter Two we will investigate whether changes in the amino acids at the five Gag subtype-specific sites correlate with the patient's HLA variants. In Chapter Three, we will study how changes at these subtype-specific sites over time may correlate with changes in the HIV-infected population's HLA frequencies.

The high polymorphism in the HLA region of the genome is thought to result from selective pressures induced by pathogens (Zinkernagel and Doherty 1974; Doherty and Zinkernagel 1975; Spurgin and Richardson 2010). After accounting for the history of human migration, areas with a high pathogen diversity are associated with high HLA diversity - an effect which is strongest for HLA B (Prugnolle et al. 2005). Similar associations are also found on longer evolutionary time-scales and in other mammals: rodent species with higher helminth parasite diversity have more MHC class II polymorphisms (Goüy De Bellocq, Charbonnel and Morand 2008). HLA variants obtained through admixture from Neanderthals and Denisovans also appear to be highly overrepresented in modern out-of-Africa populations (Abi-Rached et al. 2011). There is statistical evidence for archaic introgression in Africa (Hammer et al. 2011), but thus far no archaic hominins from Africa have been sequenced. About 4,000-5,000 years ago the Bantu agriculturalist expansion into Eastern and Southern Africa led them to encounter previously unfamiliar pathogens, which imposed a selection that caused the enriched admixture of HLA genes obtained from local rainforest hunter-gatherers. HLA genes common in those rainforest dwelling populations were likely to be well-adapted to local

pathogens (Patin et al. 2017). Selective effects in humans have also been observed on shorter, historical time-scales: African MHCs (rather than European or Amerindian alleles) were preferentially selected for in the Mexican population (Zhou, Zhao and Guan 2016). This selection may have been caused by a historical and long-lasting epidemic called 'heuy cocoliztli', which disproportionally affected native people and was symptomatically different from the Old World epidemics at the time. Medical historians believe this epidemic may have been caused by an arenavirus associated haemorrhagic fever spread by rodents; this epidemic contributed to the 90% reduction of the native population between 1545 and 1815 (Marr and Kiracofe 2000; Acuna-Soto, Calderon Romero and Maguire 2000).

The strength of the selective effects HIV can have on this region was shown by a GWAS (Genome-Wide Association Study) study of 974 HIV controllers and 2,648 HIV progressors which found over 300 candidate SNPs associated with LTNPs in the HLA region (particularly in the epitope binding groove of HLA) but none elsewhere (Pereyra et al. 2010). This study showed that mutations in this region could have large fitness effects in HIV-infected patients by preventing or delaying their progression to AIDS.

Recombination

HIV virions are diploid, containing two RNA molecules between which the reverse transcriptase complex can cross over to create a DNA product which is a mosaic hybrid of the two RNA sequences encapsulated in the virion (Hu and Temin 1990). Hybridization experiments have shown this happens on average three times per genome per replication (Jetzt et al. 2000; Zhuang et al. 2002). These recombination breakpoints are not randomly distributed over the length of the HIV genome, but instead are concentrated in certain areas and nearly absent in others — even before selection takes place (Smyth et al. 2014). Because most infected $CD4+$ cells have only one integrated provirus (Josefsson et al. 2013), the vast majority of two co-packaged RNA molecules will share that provirus as a common ancestor. In cells productively infected by multiple virions, this can lead to recombination across much larger phylogenetic distances. However, recombination over large phylogenetic distances has been shown to be less likely than recombination between closely related sequences *in vitro* (Chin et al. 2005). In HIV-1M, recombination is easily detected between

the different subtypes, using techniques such as bootscanning (Salminen et al. 1995) or Hidden Markov Models (Schultz et al. 2006). Intrasubtype recombination is harder to detect due to the small phylogenetic distance between the two parent sequences, but it has been shown to have played a major role in the evolution HIV-1M subtype C, and is likely to be important in other subtypes as well (Rousseau et al. 2007).

Several evolutionary advantages of recombination have been proposed. First, because a slightly deleterious mutation is much more likely than a back-mutation even under selection, viruses will always acquire many small deleterious mutations which would collectively make the virus non-viable. This process is called Muller's ratchet (Muller 1964) and could be avoided by recombination breaking the linkage between deleterious mutations and wild-type DNA stretches. Further, recombination has also been associated with fitness jumps and cross-species transmissions (Burke 1997), such as when SIVrcm (red-capped mangabey) and an SIV from a clade infecting Cercocebus monkeys recombined and expanded into chimpanzees to become SIVcpz, which would later become HIV-1M (Bailes et al. 2003; Keele et al. 2006; Parrish et al. 2008; Sharp and Hahn 2011). Recombinants could also allow for the faster combination of mutations that confer an evolutionary advantage, which would lead to the outgrowth of the recombinant sequences: such as when recombinant forms gain multi-drug resistance (Kellam and Larder 1995; Kemal et al. 2012), or when a recombinant's descendants are better able to escape a CTL response (Ritchie et al. 2014). Other research suggests that recombination is the mechanistic by-product of other evolutionary pressures rather than an adaptation to gain fitness more quickly or discard deleterious mutations (Simon-Loriere and Holmes 2011). However, some recombinants have had such evolutionary success that some now consider all recombinants 'hopeful monsters', because although they are usually unfit, they can be capable of tremendous leaps through fitness space (Worobey and Holmes 1999), showing the great evolutionary potential of recombination.

Recombination can lead to intersubtype recombinants if a patient is infected with two different HIV-1 subtypes. Some of these are generated de novo in individual patients and do not spread beyond their original host; these are called Unique Recombinant Forms (URFs) (Delgado et al. 2008; Wang et al. 2008; Leye et al. 2013). However, some intersubtype recombinants are examples of successful 'hopeful monsters', these are recombinants forms which spread within a population and

are called Circulating Recombinant Forms (CRFs). CRFs are defined as intersubtype recombinants that infect at least three people that are not directly epidemiologically related (Robertson et al. 2000). Despite extensive sequencing efforts, there are currently only 68 recognised HIV-1 CRFs in the LANL (Los Alamos National Laboratory) database, representing ~16% of worldwide HIV-1 infections while URFs represent only 4% of infections (Hemelaar et al. 2011). Since only intersubtype recombinants are considered CRFs and intrasubtype recombinant forms are hard to detect, this gives no information about the outgrowth of intrasubtype recombinants. Given the high RT crossover rate and the tendency towards recombination with more homologous sequences, outgrowths of intrasubtype recombinants are likely to be much more frequent than outgrowths of intersubtype recombinants.

Phylogenetics

Phylogenetics is the science which seeks to explain the great variety of life today by reconstructing the ancestral trees of living things. Its ultimate aim is the piecing together of the entire genealogy of life, going back over two billion years. In practice, this involves making phylogenetic trees, which are models showing the relatedness of living things by modelling their descent along common ancestors (fig. 3). In this thesis I explore a small part of this tree, starting with the cross-species transmission of HIV-1M in the Belgian Congo a hundred years ago, and following into the present day, where HIV-1M continues to wreak havoc.

Phylogenetic methods work by finding the tree which best explain the data. Today the data are invariably aligned sequence data, and the tree is a bifurcating structure, where each node has one parent node and two child nodes, unless it is a leaf node for which the sequence is known (fig. 3). In no case can a node be older than its direct ancestor. While mathematically succinct and conducive to computational analysis, this definition of a tree does not grasp the full complexity of life, as biological concepts such recombination and introgression cannot be contained in its structure. Nevertheless, it remains the best tool we have and has yet to be superseded by other approaches such as phylogenetic networks. Several ways of inferring trees from sequence data have been developed, but today most practitioners will use maximum likelihood or Bayesian trees — both of which are likelihood based models.

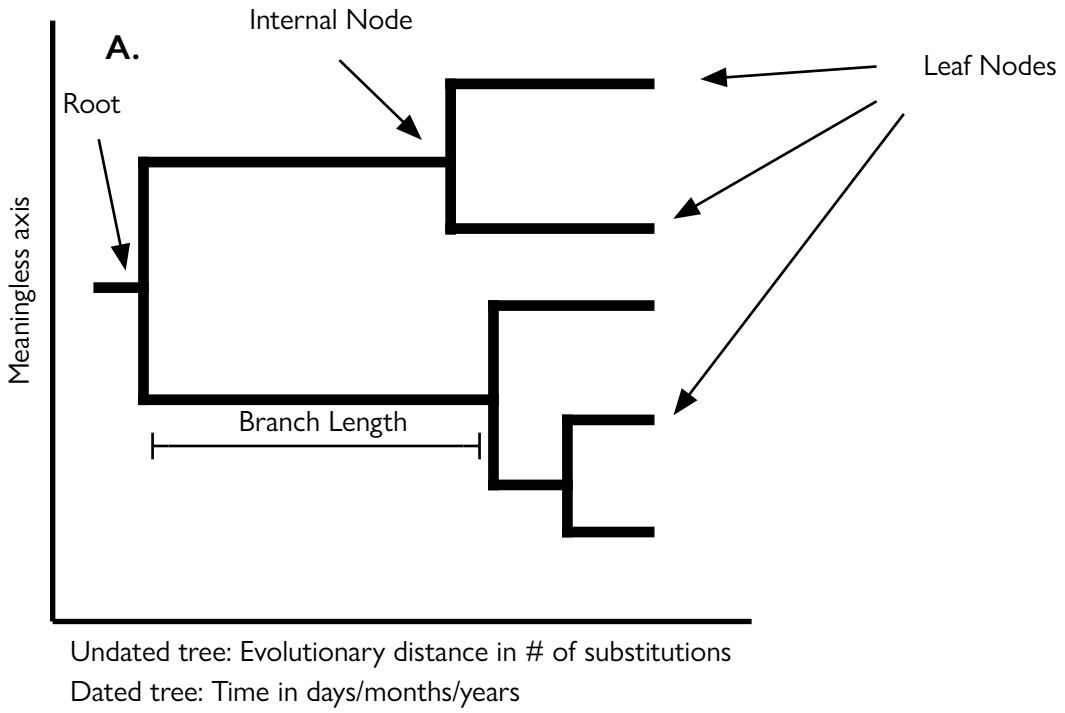


Figure 3: **Anatomy of a tree** annotated with the names describing its different parts. The x-axis can be in terms of evolutionary distance (usually the number of substitutions), or in case of a dated tree, in terms of time. The y-axis is always meaningless.

The Likelihood

Likelihood-based models assume that the observed sequence alignment arose through a Markov model along the branches of the tree. Markov models require a transition rate (not to be confused with nucleotide transitions vs transversions), so additional parameters are necessary to link the sequence alignment data to the tree. In phylogenetics, this is often called the model of sequence evolution (such as K80, GTR, or HKY), this contains the transition rates of the Markov model (fig. 4). Models of sequence evolution vary in the number of parameters, from simple models which consider all nucleic acid changes as occurring with equal rates, to complicated models which fit a different parameter for every possible change. Multiple Markov rate parameters are often necessary because some evolutionary changes may be more biologically likely than others, e.g., nucleic acid transversions are less likely than nucleic acid transitions.

The tree likelihood, like all likelihoods, is a slightly confusing concept. Rather than being the probability a tree is true given the data; it is the reverse, the probability of seeing the observed data given

a particular model, rendered mathematically as $P(\mathbf{D} \mid \mathcal{T}, \Omega)$. Briefly, this likelihood can be calculated by taking each column from the sequence alignment independently and then calculating the probability separately for each possible internal character at each internal node. While this works in toy examples, in bigger trees it is not feasible due to the explosive growth of the number of internal nodes, in which all possible combinations of nucleotides must be evaluated. Due to the Markov property, where a nucleotide at a node depends only on the nucleotide of its parent node and the branch length, the likelihood can be calculated separately for each branch given the nucleotide probabilities of the parent branch, reducing the complexity somewhat. However, it remains infeasible computationally unless a dynamic programming approach called Felsenstein's pruning algorithm is used, as it is in virtually all modern tree-builders — this is what makes maximum likelihood and Bayesian approaches feasible.

It is important to be aware of the dual meaning of transition; transition in the Markovian sense means any possible change between the four nucleotides, while transition in the nucleotide sense means any changes between nucleotides which are not the transversions A $\leftrightarrow$ G or C $\leftrightarrow$ G). To give an example of a model of sequence evolution which accounts for the different Markov transition rates, consider the K80 model shown below. Here the Markov transition rate for nucleotide transitions is set to 1, and the nucleotide transition/nucleotide transversion ratio is set to κ. The maximum likelihood procedure then finds the appropriate value for κ, allowing the model to fit different substitution rates for transitions and transversions.

$$\Omega = \begin{pmatrix} * & \kappa & 1 & 1 \\ \kappa & * & 1 & 1 \\ 1 & 1 & * & \kappa \\ 1 & 1 & \kappa & * \end{pmatrix}$$

Maximum likelihood tree-builders try out small changes to the model of sequence evolution parameters Ω and the tree $\mathcal{T}$, such as adding or subtracting a small random number to a parameter in the model of sequence evolution, lengthening or shortening a branch in the tree, or swapping around two nodes in a tree. The likelihood is then re-evaluated, and a hill climbing algorithm (such as expectation maximisation (Dempster, Laird and Rubin 1977)) is used to repeat these two steps until

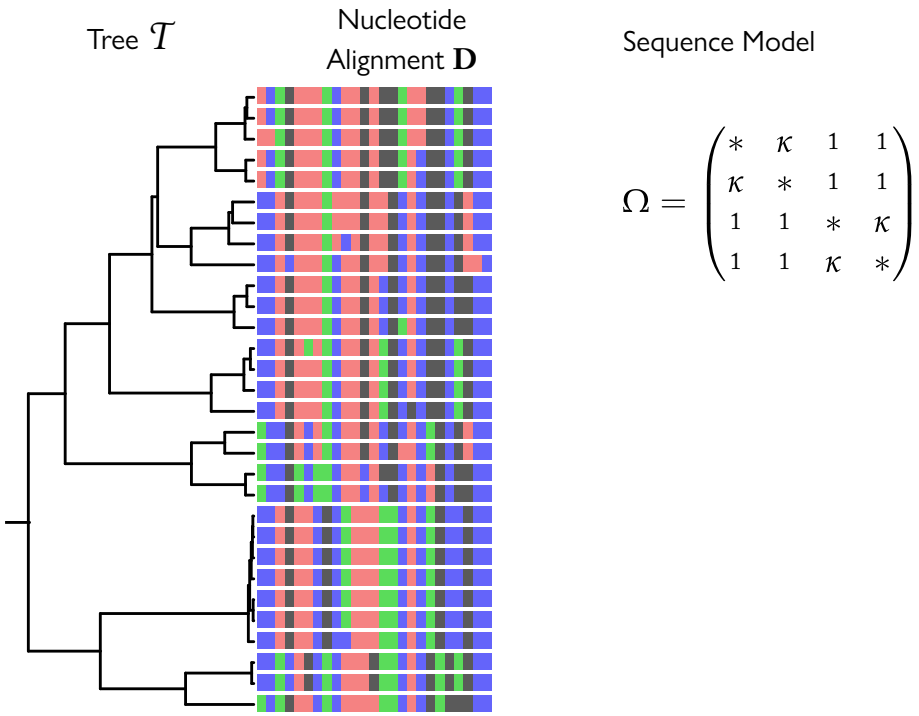


Figure 4: **The three components of a likelihood model**: the phylogenetic tree $\mathcal{T}$, the nucleotide alignment $\mathbf{D}$, and the model of sequence evolution (the K80 model is shown here).

(hopefully) the best possible tree (or at least the tree with the highest likelihood) is found (Felsenstein 1981). This procedure yields a point-estimate with no understanding of the uncertainty around the tree or model of sequence evolution parameters. Understanding this uncertainty requires additional procedures such as the bootstrap, which generates many maximum likelihood trees using different subsets of the data, to develop an understanding of model uncertainty (Felsenstein 1985a).

Bayesian Phylogenetics

Bayesian approaches, which posit that probability is a subjective belief about which one can be uncertain, has become a popular approach to phylogenetics. In addition to the likelihood methods described above, this requires specifying prior beliefs on our parameters. These take the form of probability distributions representing where we think the true value may lie. We then use Bayes' Theorem to update our prior beliefs with the data. In the phylogenetic case this is rendered as follows:

$$P(\mathcal{T}, \Omega \mid \mathbf{D}) = \frac{P(\mathbf{D} \mid \mathcal{T}, \Omega) \times P(\mathcal{T}, \Omega)}{P(\mathbf{D})}$$

The prior $P(\mathcal{T}, \Omega)$ consists of two elements. In a basic Bayesian phylogenetic analysis all trees are assumed to be equally likely, so $P(\mathcal{T}) = 1$; the prior on Ω is calculated by assuming an independent prior on all elements of Ω. In our example using a K80 model, one might expect transversions to be 50% less likely than transitions, but one might not be particularly sure, so one might set the prior on the κ parameter to a normal distribution with a mean of 0.5 and a standard deviation of 0.25, bounded at 0 because a negative rate is impossible. If one believes that transversions are never more common than nucleotide transitions one might instead bound the prior between 0 and 1.

Statements such as $P(\mathbf{D} \mid \mathcal{T}, \Omega)$ are high-dimensional objects unamenable to analytical work. Instead, we use an MCMC sampling method such as Metropolis-Hastings (Metropolis et al. 1953; Hastings 1970) to sample from the posterior. Each sample from the posterior is associated with a particular tree and particular parameter values — together many of these samples will hopefully make up a good representation of the posterior distribution. Care must be taken to ensure one has sampled from the posterior, as this can be difficult in practice. Because the probability of the

data $P(D)$ is a normalising constant, we can ignore it in the sampling procedure, and consider the posterior probability as being proportional to the prior times the likelihood.

$$P(\mathcal{T}, \Omega \mid \mathbf{D}) \propto P(\mathbf{D} \mid \mathcal{T}, \Omega) \times P(\mathcal{T}, \Omega)$$

The first sample is obtained by taking a random tree and random values for all other parameters and then obtaining the posterior associated with those parameters by multiplying the likelihood and the prior, each of which is calculated as described above. We then propose a small modification to the current tree or the current model of sequence evolution (called the proposal), in the same way as we did in the maximum likelihood model, and recalculate the posterior. In the case of Metropolis-Hastings, the proposal is immediately accepted if the associated posterior is higher than the current sample's posterior, and the proposed $\mathcal{T}$ and Ω are kept and used as the base for the next proposal. Metropolis-Hastings will also randomly accept proposals whose posteriors are lower than the current sample, in proportion to how much lower they are. This makes Metropolis-Hastings sample from the posterior when certain conditions are met, allowing us to identify both the phylogenetic uncertainty and the uncertainty around the other parameters.

Of course, the practitioner need not worry too much about these implementation details, as user-friendly programs such as MrBayes (Ronquist et al. 2012) and BEAST (Drummond and Rambaut 2007) are available. While these programs successfully abstract away much of the detail, some intuitive understanding of the internals is beneficial when considering which extension to the phylogenetic method is most applicable to a particular biological question.

Extensions of the Bayesian Phylogenetic Model

Almost all elements of the basic phylogenetic model described above can be replaced by more complicated versions which allow additional inference. One of the most important innovations in phylogenetics has been the dated tree, where the branch lengths represent time rather than evolutionary distance. Essentially Ω is modified to fit an additional parameter, the clock rate, which converts evolutionary distance to time. This clock rate is usually denominated as substitutions per year — in which case branch lengths now represent years (Li, Tanimura and Sharp 1988; Rambaut 2000). The

leaf nodes (also known as tips) for sequences which are associated with a date are then fixed at these dates, i.e., no proposals are generated which move these tips, ensuring they always remain at the same position in time. The clock rate parameter is then inferred by the Bayesian procedure along with the tree and the sequence model parameters, which hopefully results in the internal nodes being placed at approximately the right times. This model can be further extended by allowing different clock rates on each branch, and imposing a hierarchical prior on these individual clock rates to permit partial pooling (Drummond et al. 2006). Other extensions, which allow the use of fossil evidence to calibrate internal nodes, have also been described (Heath, Huelsenbeck and Stadler 2014).

Another important and commonly used extension is the coalescent (Kingman 1982), which links the phylogeny to a vast body of statistical genetics literature by imposing a prior on the tree (Li, Tanimura and Sharp 1988; Pybus, Rambaut and Harvey 2000). This relies on the fundamental insight that the shape of the tree can be related mathematically to the population genetic processes that gave rise to it, e.g., a tree generated from a population undergoing exponential growth looks different than a tree generated from a population under heavy selection. These differences can be especially stark in pathogens (Grenfell et al. 2004). Essentially, the tree prior is a second order likelihood function, which relates the shape of the tree to parameters describing the underlying population genetics process. Such parameters can include historical population sizes (Drummond et al. 2005), the basic reproductive number (the number of secondary cases a single infected individual is expected to produce in an entirely susceptible population) of a pathogen (Stadler et al. 2012), or even a SIR model of infectious disease (Popinga et al. 2015).

Further extensions to the phylogenetic method have also been used to track the migration of genes through space, called phylogeography. Such methods have been based on maximum likelihood (Pagel 1999) and parsimony (Parker, Rambaut and Pybus 2008). Similar to phylodynamics most studies incorporating phylogeography are done in a Bayesian framework to allow for the incorporation of uncertainty, usually implemented in the program BEAST (Pagel, Meade and Barker 2004; Lemey et al. 2009; Faria et al. 2013). Most notably, phylogeographic methods were used to show that the pandemic HIV-1M strain first emerged in Kinshasa, the capital city of what was then the Belgian Congo (now the Democratic Republic of the Congo). The virus probably travelled by waterways and rail-roads (Faria et al. 2014); hitch-hiking on journeys similar to those taken by Joseph Conrad

when he travelled the Congo basin some twenty years before the initial expansion of HIV-1M, a journey which inspired his 1899 novella *Heart of Darkness*. The Faria study also used skyline plots to demonstrate that national independence was associated with faster growth of the epidemic, possibly due to reduced medical care or behavioural changes, though this did not apply to HIV-1M group O. The phylogeographic model was later extended with a generalised linear model (GLM) which predicts transmission rates between locations based on potential explanatory variables associated with every possible transmission (Lemey, Rambaut and Bedford 2012; Faria et al. 2013). This method was used to show that air travel routes are the major predictor of global H_3N_2 influenza transmission dynamics (Lemey et al. 2014).

This Thesis

It is likely that a vaccine is necessary to end the HIV-1M pandemic (Fauci, Folkers and Marston 2014). CTL responses against Gag are uniquely associated with delayed disease progression (Kiepiela et al. 2007), making Gag a prime vaccination target. However, in light of recent vaccine failures (McMichael and Haynes 2012), a better understanding of the fundamental biology is necessary. HIV-1M immune escape of the CTL response has been well-documented, but little is known of how these selective forces have affected, and are affecting, the evolution of HIV-1M. In this thesis, we use phylogenetic and statistical methods to ask questions relating to the adaptation of HIV-1M to the CTL immune response, both at a patient and a population level. In Chapter One, we implement the phylogenetic logistic regression, which accounts for phylogenetic covariance which might otherwise lead to type II error. We are the first to extend the phylogenetic regression to allow for multiple response variables which have evolved independently along the same phylogeny. In Chapter Two, we use our improved phylogenetic regression to study if certain HLA variants are associated with the particular amino acids found at HIV-1M Gag subtype-specific sites; we demonstrate an association between HLA variants and the amino acid on the Gag subtype specific positions, which could be evidence of CTL immune escape. In Chapter Three, we again use our phylogenetic regression to ask whether the changing ethnic demographics of the HIV-infected population, and corresponding changes in HLA frequencies, has had a selective effect on the amino acids on Gag subtype-specific positions. We show that as the HIV-infected population in the United States has

become more African-American, the resulting change in HLA variants encountered by the virus may have led to population-level immune escape resulting in a HIV-1M subtype B virus that has some functional similarities to HIV-1M subtypes in African countries. In Chapter Four, we use a cohort of Commercial Sex Workers (CSWs) from Nairobi Kenya to study whether effective CTL responses are associated with increased recombination, probably due to increased immune pressure driving the necessity for bigger fitness steps. We find that it has. Chapter Five presents a case-study of two of these CSWs, who are co-infected with multiple strains of HIV-1M. We study these patients in detail and use them as an example of the broader patterns in the previous chapters. Finally, in Chapter Six, we study retroviral cross-species transmission by analysing a clade of betaretroviruses. We sought to determine which host-factors predict cross-species transmission, and found that host phylogenetic relatedness is a predictor of cross-species transmission and that rodents play a major role in the cross-species transmission of betaretroviruses.

1 Validating the Multiple Response Bayesian Phylogenetic Logistic Regression Using Simulations

1.1 Introduction

The necessity of correcting for phylogenetic correlations when performing regressions on evolving traits has long been recognised (Felsenstein 1985b; Grafen 1989; Harvey and Purvis 1991). Phylogenetic relatedness between measurements violates the assumption of statistical independence, as measurements taken at the tip of a phylogeny cannot be assumed to be independent of one another. This phylogenetic covariance is a potential source of type I error (false positives, FP) in comparative biology. In this chapter, we implement a phylogenetic logistic regression in a flexible Bayesian framework and subsequently extend this method to allow for multiple traits evolving independently along the same phylogeny, which was necessary to model the evolution of five subtype-specific amino acids in HIV-1M Gag.

Several methods have been developed to perform phylogenetic regressions, mostly in the context of Gaussian explanatory and response variables. These methods include Felsenstein's phylogenetically independent contrasts (Felsenstein 1985b) and phylogenetic generalised least squares (PGLS, alternatively known as the phylogenetic regression) (Grafen 1989). Felsenstein's phylogenetically independent contrasts and PGLS were later shown to be equivalent when evolution is modelled as Brownian Motion (Garland, Jr., and Ives 2000; Rohlf 2001; Blomberg et al. 2012), and have been extensively reviewed (Harvey and Purvis 1991; Freckleton, Harvey and Pagel 2002). It would take another decade before the first methods that could model binary responses became available (Ives and Garland 2010). Soon afterwards a threshold model was developed which used a latent variable to model binary traits (Felsenstein 2012). Usually, these methods assume a Brownian Motion model

of evolution, which can be visualised as a random walk along branch lengths, however, others have argued for an Ornstein-Uhlenbeck model of evolution, which can incorporate selection and drift (Hansen 1997; Butler and King 2004). Unfortunately, no good way exists of testing whether the Brownian or Ornstein-Uhlenbeck model is more appropriate (Cooper et al. 2016).

These methods are all grounded in frequentist statistics, and a variety of packages are required to fit these models. Recent developments in Bayesian statistics and the increased computational feasibility of MCMC has made fitting Bayesian models an attractive alternative. Not only do real credible intervals become available rather than the artificial confidence intervals used in frequentist statistics, but it's also a much more stimulating environment for learning statistics. Studying frequentist statistics can be a disjoint experience where increasing the complexity of a model in one way or another often relies on packages that are mutually exclusive and often require different syntax. For instance, one might want to do something as simple as adding a random effect to a GLM (Generalised Linear Model). This requires fitting the model differently than a regular GLM, which is implemented in several commonly used R packages (amongst others: lme4 (Bates et al. 2015), MASS (Venables and Ripley 2013), mcmcGLMM (Hadfield and Nakagawa 2010)), each requiring a slightly different syntax. In the face of this syntactical complexity, Bayesian statistics, as implemented in BUGS, JAGS and Stan, becomes an attractive alternative. Here a statistical model is defined in a programming language, and one can add and remove model complexity at will by changing a few lines of code rather than installing entirely new packages. Bayesian models implemented in BUGS, JAGS or Stan also have the advantage of being self-documenting. This more holistic approach to modelling comes at the price of increased computational requirements, occasional complications in achieving convergence, and the requirement that the practitioner is already comfortable with programming languages.

There are two commonly used Bayesian approaches to phylogenetic regressions, of which MCMCglmm (Hadfield and Nakagawa 2010) is currently most popular. It allows for several common likelihood functions (Gaussian, Poisson, Binomial, allowing for continuous, count and binary/binomial data respectively). However, it has many of the same disadvantages as frequentist R methods in the sense that it is a predefined package with its own unique syntax and somewhat lacking documentation. Recently, de Villemereuil published a simpler model for phylogenetic Gaussian regressions

in JAGS (Villemereuil et al. 2012), which also allows for sampling from a posterior distribution of phylogenies to incorporate phylogenetic uncertainty into the regression. Here, we extend the de Villemereuil approach to the binary response case (corresponding to the 5 dichotomised subtype-specific positions), using a phylogenetic random effect; this ends up being similar conceptually to the basic MCMCglmm model, though it is implemented in an entirely different way. However, it does not include the further optimisations of the MCMCglmm model, which involves partitioning the correlation structure into a component that can be integrated deterministically and another component which must be integrated numerically (Hadfield 2015). We trade in these optimisations for the flexibility inherent to a Stan model and exploit this flexibility to extend the model to allow for multiple responses that evolve independently along the same phylogeny.

Our goal in developing this method is to investigate potential associations between HIV-1M amino acids and patient HLA variants, which was first attempted without correcting for phylogeny (Moore et al. 2002), though later work did apply a phylogenetic correction using a likelihood ratio test (Bhattacharya et al. 2007). However, both those papers combined all subtypes in one analysis, which is likely to lead to false positives due to associations between HIV-1M subtype distribution and HLA frequencies. Subtype C is common in South African blacks while subtype B is common in Americans, however, South African blacks, African-Americans and American Caucasians are populations with different HLA frequencies. These differences could lead to spurious associations — this is referred to as a founder effect in Bhattacharya et al. (2007). Though it is possible phylogenetic corrections could account for that, it is probably better to analyse different subtypes separately to prevent subtype-effects, as we have done in Chapters Two and Three; and only rely on the phylogenetic correction to make adjustments for very closely related viruses, such as those found in transmission-pairs.

In this chapter we seek to describe a statistical method which is appropriate to study the evolution of multiple HIV-1M subtype-specific sites. We developed Bayesian phylogenetic models using Stan, and have used simulations to show that our models are sensitive to true effects while preventing phylogenetic covariance induced type II error. We further show that — when using response variables derived from phylogenetic data — it is appropriate to use a newly described model which incorporates an independent phylogenetic component for each amino acid position. These simula-

tions show that the methods developed herein can be safely used to study the evolution of HIV-1M subtype-specific sites without worrying about phylogenetic covariance related type II error.

1.2 Methods

1.2.1 Notation and terminology

Here we use the term patient for each phylogenetic leaf node and its associated measurements x and y. However, this is a matter of terminology, and this method could be applied without further modification to species instead of patients — for instance, if one wanted to use this method in comparative biology. Because we will use this method to model viral amino acids we consider patients to be related to each other through the phylogeny of their infecting viruses, i.e. we do not concern ourselves with possible familial relationships between the patients themselves as could be described by a genealogy.

The basic model for the Gaussian case is a multivariate normal likelihood function, where the mean is calculated by multiplying the explanatory variable x and β, while the phylogenetic covariance is accounted for by the likelihood function's variance-covariance matrix $\sigma\mathbf{C}$, where $\mathbf{C}$ is a correlation matrix derived from the virus phylogeny:

$$y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma\mathbf{C})$$

Because we are interested in using a binary response representing the presence/absence of amino acids, which we model using a Bernoulli likelihood function, it was not possible to keep the variance-covariance matrix in the likelihood function. Instead, we move the phylogenetic relatedness to the patient parameter; in frequentist terms, this would be called a random intercept.

$$y_i \sim \text{Bernoulli}(p_i)$$

$$\text{logit}(p_i) = \alpha + \beta x_i + u_i$$

$$u_* \sim \mathcal{N}(0, \sigma\mathbf{C})$$

This leaves us with a σ parameter that is intended to scale the phylogenetic correlation matrix $\mathbf{C}$ to a variance-covariance matrix on the scale of the response. However, the shape of the logistic function allows this parameter to grow explosively, which causes the model to become non-identifiable. We were able to resolve this by fixing the σ parameter to 1, meaning that we can use the phylogenetic correlation matrix as the variance-covariance matrix. This is a sensible choice in the context of the inverse logit function, which rapidly approaches 0 and 1 outside of the range [-2.71, 2.71]. It is important to normalise x, so that it is on the same scale as the logit and the patient parameter u_i.

$$u_i \sim \mathcal{N}(0, \mathbf{C})$$

This model could be extended to work with multiple covariates by substituting the following formula:

$$\text{logit}(p_i) = \mathbf{X}\beta + u_i$$

Where $\mathbf{X}$ is a design matrix, β is a column vector, β_1 is the intercept and the other betas are the coefficients. A variation on this is used in Chapter Two, but in the rest of this chapter we study only the single covariate case.

Extending to the Multiple Response Case

We now consider the hitherto unstudied case of multiple binary responses which have evolved along the same phylogeny, such as different amino acids j sampled from HIV-1M sequences obtained from patient i. In this example, we consider only one explanatory variable x_i, sampled at the patient level, but we do allow different intercept and coefficient parameters for every j. This parameterisation would be especially powerful when combined with hierarchical priors on the intercept and coefficient parameters, such as in the case where it is expected that the amino acid positions are influenced in similar ways by a covariate.

$$y_{ij} \sim \text{Bernoulli}(p_{ij})$$

$$\text{logit}(p_{ij}) = \alpha_j + \beta_j x_i + u_{ij}$$

$$u_{*j} \sim \mathcal{N}(0, \sigma \mathbf{C})$$

We have separated the patient parameter u_i into a separate patient/subtype parameter for each j, allowing each amino acid to evolve along the phylogeny separately rather than as a block.

Priors

In the Gaussian regressions we used a $\mathcal{N}(0, 2)$ prior on α and β, and a half-normal prior with location 0 and scale 1 on the σ. In the binomial model for one amino acid we used a $\mathcal{N}(0, 1)$ prior on α and β. For the binomial regression on multiple simulated amino acid responses, Gelman's suggested Cauchy prior with a scale of 2.5 was adopted. We did not use hierarchical priors in the multiple response case simulations in order to show that the reduction in false positive rates (FPR) was obtained by the separate multivariate normal prior for each position, rather than by partial pooling.

We fitted the models using Stan, which was run for 5000 iterations for each simulated dataset. Some caution is advised if the reader wishes to reimplement this model, as we encountered some numerical problems in JAGS generally and also in Stan when using the multivariate normal parameterised with the precision matrix but not when using the variance-covariance parameterisation in Stan.

1.2.2 Obtaining the correlation matrix

Because we elected to use a random walk model the variance-covariance matrix was calculated as follows: the variances (diagonals) of the phylogenetic variance-covariance matrix were calculated by taking the phylogenetic distance (the sum of branch lengths) between the root node and the leaf node (a leaf node represents a simulated virus sample). The covariances (off-diagonals) were calculated by taking the shared phylogenetic history of two leaf nodes, e.g. the distance between the root and the Most Recent Common Ancestor (MRCA) of the two leaf nodes (Felsenstein 1985b; Freckleton, Harvey and Pagel 2002). The resulting phylogenetic variance-covariance matrix was then transformed into a phylogenetic correlation matrix.

1.2.3 Simulations

All simulations were done using R, the `rtree` function from the ape package was used to generate a random tree with 100 leaf 'patient' nodes for each simulation. The ape package was also used to calculate phylogenetic correlation matrices (used to generate the data with multivariate normal), and to perform the Markov simulations of molecular evolution using an equal rates model. All datasets were randomly generated 2,000 times.

First, we used Gaussian regressions to examine whether accounting for phylogenetic variance is necessary when only one of x and y are phylogenetically informed. We chose to use Gaussian regressions rather than Bernoulli regressions because Gaussian regressions are more vulnerable to phylogenetic type I error and any such behaviour in the Gaussian case can be expected to also occur in the generalised linear models.

We then simulated several different kinds of datasets in the absence of any true relationship between x and y, where both variables are independently generated from various distributions (Table I). This was to examine whether our models were vulnerable to phylogenetic covariance-induced false positives. We also compared our phylogenetic models to frequentist models without a phylogenetic correction fit to the simulated data. We then examined whether our model is sensitive enough to pick up a β when data were simulated with a real effect, and also whether our model was robust to varying the σ on the simulated random effect.

1.2.4 Fitting models to the simulated data

Uncorrected frequentist Gaussian regressions were fit using `lm`; uncorrected frequentist Bernoulli regressions were fit using `glm`. Both phylogenetically corrected and uncorrected Bayesian regressions were fit using Stan (code in supplement); each Stan model was run for 5,000 MCMC iterations.

1.3 Results

1.3.1 Uncorrected Linear Regression is Not Anti-Conservative when both x and y are Phylogenetically Informed

We simulated x and y either from a normal distribution of location 0 and scale 1, or from a multivariate normal distribution informed by a phylogenetic correlation matrix. We then fit a frequentist LM (Linear Model) to this simulated data, uncorrected for any phylogenetic relatedness. We find that this naive approach is only anti-conservative (i.e. a Type I error rate greater than 0.05 when using a p threshold of 0.05) when both x and y are informed by the phylogeny.

We then fitted the phylogenetic model to the datasets where both x and y were simulated from a multivariate normal: this resulted in a type I error rate indistinguishable from 0.05. This means that although either the response or explanatory variable was not i.i.d. (as measurements from that variable were related to one another through the phylogeny), the phylogenetically uncorrected model was not anti-conservative. This suggests that, pragmatically, a phylogenetic correction can be omitted when only one parameter is phylogenetically informed. However, when both x and y are phylogenetically informed, one should always use a phylogenetic regression as this gives an increased FPR under the non-corrected model.

Table 1.1: Results from fitting an uncorrected linear model to various kinds of simulated datasets. Phylogenetic correction is only necessary when both x and y are simulated from a multivariate normal distribution informed by the same phylogenetic correlation matrix. Simulations where the False Positive Rate (FPR) differed significantly from the expected 0.05 according to a Binomial exact test (p<0.05) are marked with *.

x	y	Type I error (FPR)
$\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$	0.061
$\mathcal{N}(0,1)$	MVN$(0,\mathbf{C})$	0.056
MVN$(0,\mathbf{C})$	$\mathcal{N}(0,1)$	0.040
MVN$(0,\mathbf{C})$	MVN$(0,\mathbf{C})$	*0.397

1.3.2 Bernoulli Model with Phylogenetic Random Effect

Quantifying the Type 1 Error (False Positive Rate)

We simulated a phylogenetically informed binary distribution by either drawing from a Bernoulli distribution where p is the logit of a multivariate normal draw informed by a phylogenetic correlation matrix, or by using a Markov Model of molecular evolution with various rates. Phylogenetically informed continuous distributions were simulated by drawing from the phylogenetically informed multivariate normal directly. In almost all cases, the FPR of the frequentist GLM was too high, except when both the explanatory and the response variables were binary (presumably owing to a want of power), or when the rate on the Markov chain was too high (rates higher than 1 minimised the phylogenetic signal given the simulated trees). Adding the phylogenetic random effect makes the model more conservative; statistically indistinguishable from a perfect model when the explanatory variable is continuous and distributed according to the model's assumptions (Binomial Exact Test, $p > 0.05$ in all cases). It is slightly over-conservative when using a binary explanatory variable, simulated by drawing from Bernoulli distributions where p is determined by the inverse logit of a multivariate normal informed by the phylogenetic correlation matrix. To further challenge our model with data that does not match the model assumptions we considered an x drawn from a multivariate normal distribution informed by the phylogenetic correlation matrix, and a y calculated by allowing a Markov model of molecular evolution to evolve along the simulated tree with various rates. We found that though the model was not perfect under these conditions, it did considerably reduce the FPR as compared to a frequentist GLM. Because the data were simulated using a different model than was used for inference, we did not expect a perfect FPR of 0.05 given a 95% BCI.

Table 1.2: Results from fitting an uncorrected classical Bernoulli model and the phylogenetic random effect Bernoulli model to various kinds of simulated datasets. Simulations where the False Positive Rate (FPR) differed significantly from the expected 0.05 according to a Binomial exact test ($p<0.05$) are marked with *. The phylogenetic model does not reach a perfect FPR of 0.05 every time because the data are not simulated following the model, however, in each case the phylogenetic model is better than the uncorrected model.

x Simulation	y Simulation	Frequentist GLM FPR	Phylogenetic RE FPR
MVN$(0, \mathbf{C})$	Bernoulli_logit(MVN$(0, \mathbf{C})$)	$*0.126$	0.048
Bernoulli_logit(MVN$(0, \mathbf{C})$)	Bernoulli_logit(MVN$(0, \mathbf{C})$)	0.065	$*0.034$

x Simulation	y Simulation	Frequentist GLM FPR	Phylogenetic RE FPR
MVN(0, **C**)	Markov Model, rate=0.1	∗0.268	∗0.136
MVN(0, **C**)	Markov Model, rate=0.2	∗0.309	∗0.146
MVN(0, **C**)	Markov Model, rate=0.5	∗0.233	∗0.097
MVN(0, **C**)	Markov Model, rate=1	∗0.140	∗0.065
MVN(0, **C**)	Markov Model, rate=2	∗0.070	0.044
MVN(0, **C**)	Markov Model, rate=3	0.054	∗0.031

Showing the Model Sensitivity

Beyond limiting Type I error it is important that the model can accurately recover effects when they exist, which means the model avoids type II error (false negative results). We tested this by simulating datasets where we knew there was a real effect, which we then recovered using our model (fig. 1.3). We found that our model accurately recovered the simulated coefficient β, with a small conservative bias towards 0. This conservative bias is probably due in part to the prior on β.

Because it was necessary to fix the model's σ to 1 to prevent non-identifiability and non-convergence we wanted to investigate whether this would limit our ability to recover true signal if the simulated σ was greater than 1. To this end we evaluated nine different simulation scenarios with the simulated sigma between $[0, 4]$ (fig. 1.4). We found that the estimates remained reliable in spite of the increasing differences between the simulated data and the model as σ is inflated.

1.3.3 Allowing for Multiple Amino Acid Positions

Because we are interested in modelling more than one amino acid at a time, it was necessary to evaluate whether the model could handle more than one amino acid (y) evolving independently along a tree. We simulated each amino acid separately using a Bernoulli distribution where p is a draw from the multivariate normal distribution using a phylogenetic variance-covariance matrix and the logistic link function; x was a draw from a multivariate Gaussian distribution using the same phylogenetic variance-covariance matrix. We varied the number of amino acids simulated between

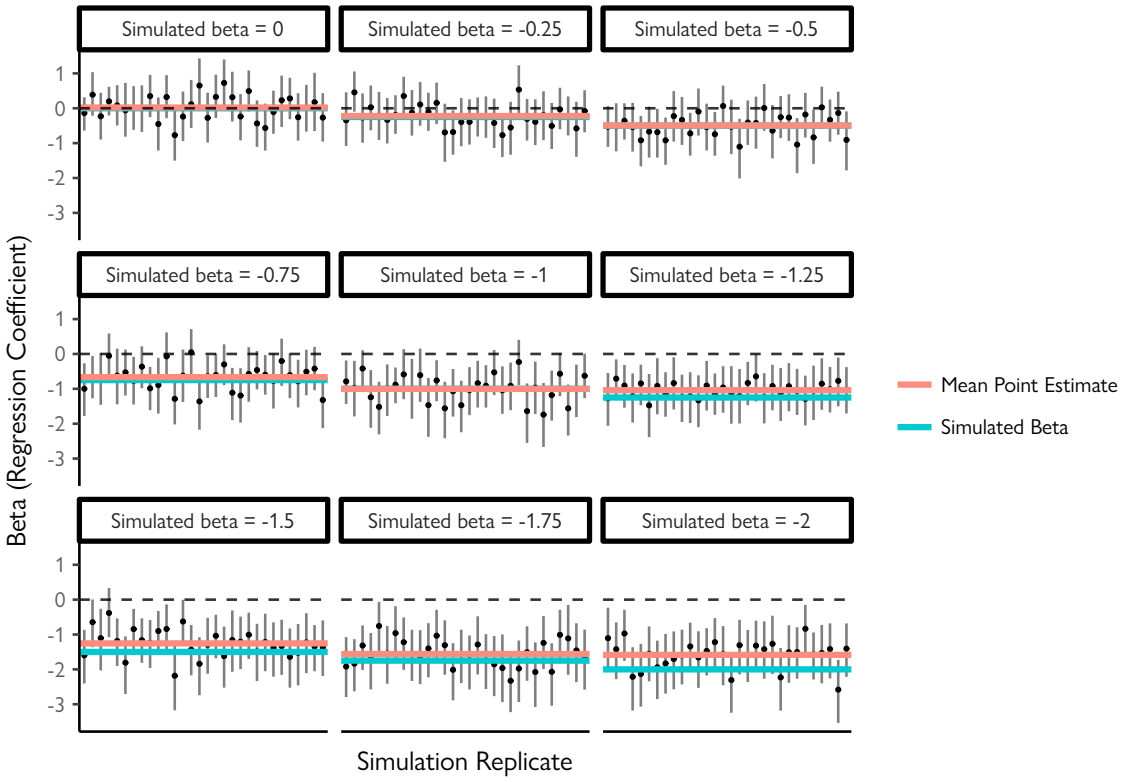


Figure 1.3: **Fitting model to simulated datasets with varying coefficients**
The 95% BCI is shown by the vertical lines, with the median coefficient shown as a point. In every simulation setting the mean is taken of the coefficient β point estimates and shown as the red line. The original simulated beta is shown as the green line. Our model accurately recovers the simulated coefficient β, with a small conservative bias towards 0.

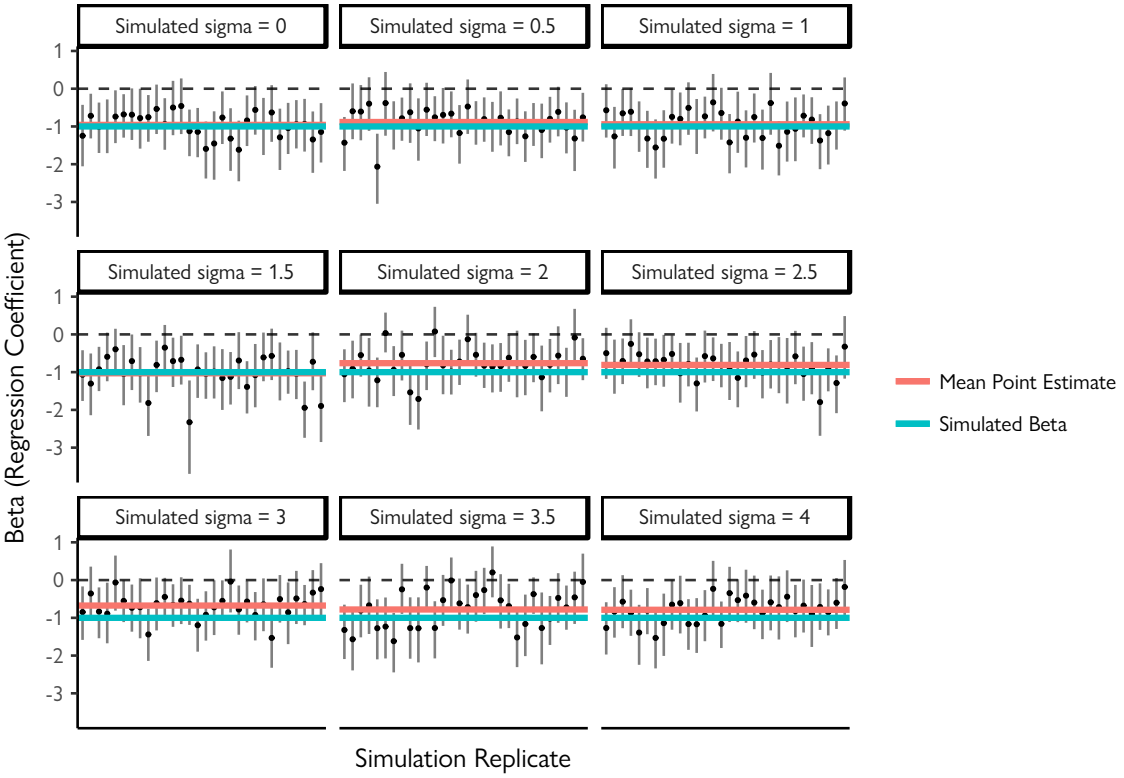


Figure 1.4: **Fitting model to simulated datasets with different random effect scale parameters**
Here we vary the simulation's random effect scale parameter while holding the rest of the parameters constant. This shows that the model is robust to stronger phylogenetic influence in y, even though we fixed the model's random effect scale parameter to 1.

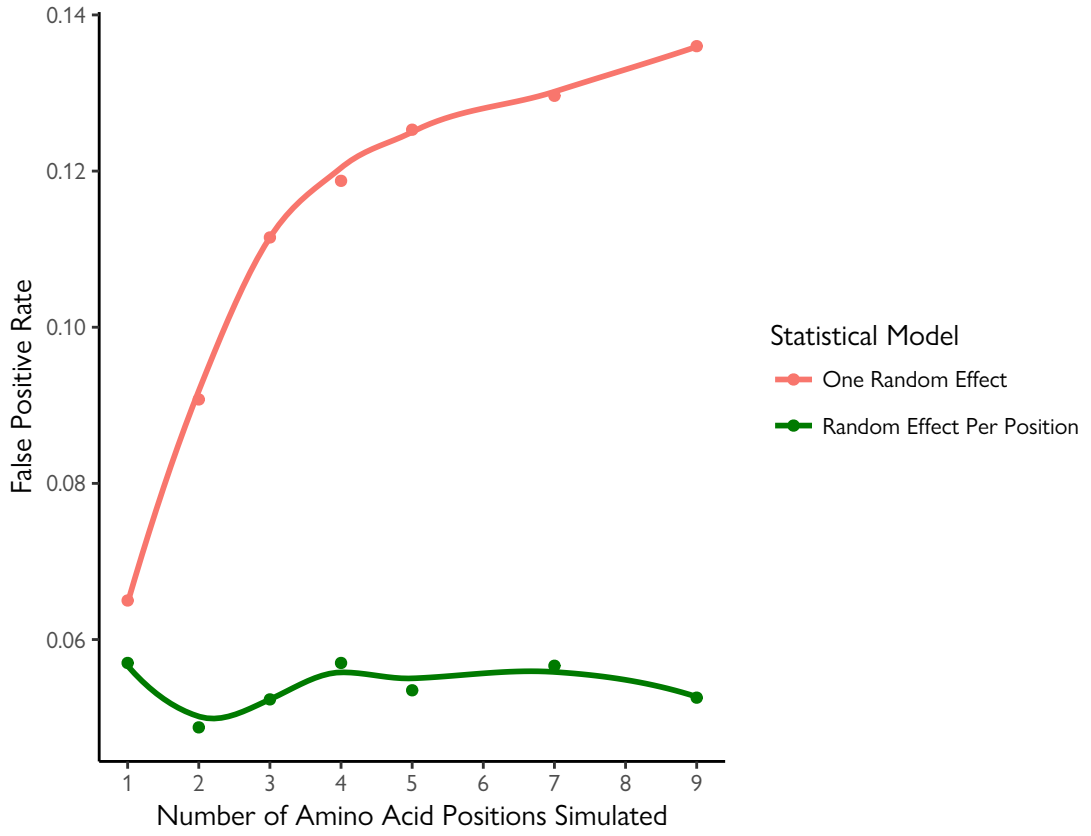


Figure 1.5: **Fitting models to datasets with varying numbers of independently evolving amino acids**
Because amino acids are assumed to evolve independently in a phylogenetic framework it is important to model them with independent phylogenetic random effects (e.g. to use a separate multivariate normal distribution informed by the phylogenetic correlation matrix for every amino acid). If one models this using a single multivariate normal (equivalent to a random effect multiple measures model with a correlation matrix, see the red line) the FPR increases as more amino acid positions are simulated, if instead, one uses a separate multivariate normal for each amino acid position the FPR stays the same. Each point represents the FPR calculated from 2000 simulations of patients with the number of amino acids. These were simulated for each position separately using the Bernoulli distribution on the multivariate normal distribution draw using the logistic link function in the absence of any true correlation to the multivariate Gaussian x. The trend lines were fitted using LOESS.

1 and 9 and generated 2,000 simulated datasets for each amino acid count. We fitted a model with a single phylogenetic multivariate Gaussian distributed random effect shared between all amino acids in a patient, as well as a model with a separate phylogenetic multivariate Gaussian distributed prior for each amino acid. We found that the model with the shared multivariate Gaussian was prone to increased FPRs and that the FPR was higher the more amino acids were simulated (red line in fig. 1.5). The FPR in the model with a separate phylogenetical multivariate normal distribution for each amino acid was stable (green line in fig. 1.5).

1.4 Conclusions & Discussion

Here we have implemented a phylogenetic logistic regression and extended this to allow for multiple binary responses, each of which evolves independently along the phylogeny. We have shown that this method is appropriate for analysing data derived from simulated models of molecular evolution. This allowed us to build powerful and extendible phylogenetic models, for instance by incorporating hierarchical priors on the model parameters, which correct for phylogenetic relatedness.

We also have shown that, from a pragmatic perspective, it is not necessary to use a phylogenetic regression when only either the explanatory or response variable is phylogenetically informed. However, because this still represents a violation of the i.i.d., it could be argued that it is necessary to do so in any case (Felsenstein 1985b; Grafen 1989; Harvey and Purvis 1991; Freckleton, Harvey and Pagel 2002), though some suggest providing both the phylogenetically corrected and the uncorrected analyses would be appropriate (Gaston, Blackburn and Lawton 1997; Schluter 2000). For the sake of conservatism and mathematical correctness we have used the phylogenetic regression, which was always the most conservative, whenever possible in this thesis.

This method is in some ways similar in aim to the discrete trait framework implemented in BEAST (Lemey et al. 2009), which also allows traits to evolve along a phylogeny independently. However, that method was unsuited for our actual data because the phylogenies we obtained had very long terminal branches. Furthermore, our implementation allowed for explanatory variables to be associated directly to the leaf node (patient), while the BEAST GLM extension (Faria et al. 2013) requires explanatory variables to correspond to the discrete state the Markov chain is in. This latter method was used in Chapter Six to study what factors drove the cross-species transmission of betaretroviruses over the last 30 million years, since in that case we had covariates associated with the discrete state, i.e., the species of animal. In Chapters Two and Three, however, we were trying to predict the amino acid state using the patient's HLA, which means the covariates (HLA variants) were associated with the patient rather than with the discrete state, and this cannot be modelled inside the BEAST framework.

One limitation of our methodology is that we do not consider phylogenetic uncertainty; instead we treat the phylogeny as fixed. This contrasts with the approach used in de Villemereuil, which does

incorporate phylogenetic uncertainty. Early experiments with incorporating phylogenetic variation through sampling from the posterior distribution of trees tended to result in the MCMC chain getting 'stuck' on one particular tree, which was also reported as a potential problem in Villemereuil et al. (2012). This could perhaps be mitigated by extending Stan to always accept proposals to switch trees, as the BEAST MCMC sampler does when provided with a pre-existing posterior distribution of phylogenies. However, because this would require modifying Stan itself, and because even then we would not have been entirely assured of success, we opted to use a point estimate of the Brownian motion correlation matrix as described above. Our point estimate should give a better result than an analysis that fixated on a single tree, but is not as robust to phylogenetic uncertainty as an analysis that properly samples from the posterior distribution. This approach is easily adaptable to other ways of calculating the phylogenetic variance-covariance matrix from which the correlation matrix is derived, for instance, the Ornstein-Uhlenbeck model could be used instead of the Brownian motion model (Hansen 1997; Butler and King 2004). However, we have shown that our method holds up well in the face of binary traits simulated using a Markov Model, in spite of the large differences between the Markov model of molecular evolution used to simulate the data and the regression model used to fit the simulated data.

Our model is highly parameterised, as it must fit a parameter for every single patient or every single patient/amino acid combination in the multiple response case. This is necessary as no multivariate Bernoulli function exists. Therefore the phylogenetic correlations must be accounted for in this manner, essentially imposing a modified multiple measures model where no multiple measures exist. However, our simulations show that this possible overparameterization does not interfere with the ability of the model to prevent type I error and to properly recover real effects.

The phylogenetic GLM presented here should easily adapt to a Poisson or Binomial function, if there is a reason to believe that the latent variable (p for binomial, or λ for Poisson) is phylogenetically influenced. Some caution is needed in this case, since we found that the binomial was not appropriate for multiple sequences sampled from the same patient (using the total number of sequences as the Binomial size parameter). It would be trivial to add Pagel's lambda parameter to the off-diagonal elements in the phylogenetic random effect; thereby scaling the covariances downward (making the phylogenies more starlike) and increasing the statistical power of the analysis when the phylogenetic

effect is small (Pagel 1999; see Villemereuil et al. 2012 for an example in a similar analysis). We have not done so here because our HIV phylogenies were already quite starlike, and not much statistical power could have been gained by increasing this star-likeness further. This existing star-likeness in the phylogenies of our actual data also motivated the choice for a regression framework rather than a BEAST discrete trait analysis as the long terminal branches made it impossible to use the discrete trait analysis.

The method developed here allows us to ask a tightly defined biologically informed question of the data, which gives us additional power to detect effects over GWAS-like approaches that consider all sites in the genome and therefore requires considerable multiple testing correction (Palmer et al. 2017). This comes at the cost of having to pre-select sites of interest, rather than evaluating all sites in the viral genome. However, some of these methods do compare favourably to ours by explicitly modelling recombination — our method could be expanded to allow for recombination when recombination-aware phylogenetic tree inference programs become available, as it would only require a recombination-aware estimate of the phylogenetic correlation matrix to be substituted in. Indeed, in the case of our model which allows for multiple amino acids to evolve independently, it would be trivial to model phylogenetic relatedness in different sections separately, simply by including a different phylogenetic correlation matrix for each position.

1.5 Model Code

1.5.1 Gaussian Phylogenetic Regression Stan Model

```
data {
    int<lower=1> N_leaves;
    int<lower=1> N;
    corr_matrix[N_leaves] cor_mat;

    real x[N];
    real y[N];
    int leaf[N];   ## Group membership vector.
}
transformed data{
    vector[N_leaves] zero_vector;
    for (j in 1:N_leaves){
        zero_vector[j]=0;
    }
}
parameters {
    real alpha;
    real beta;
    real<lower=0> sigma;
    real<lower=0> re_sigma;
  vector[N_leaves] random_effect;
}
transformed parameters{
    cov_matrix[N_leaves] re_var_covar_matrix;
    re_var_covar_matrix = re_sigma^2*cor_mat;
}
model{
    real mu[N];
    alpha ~ normal(0,2);
    beta ~ normal(0,2);
    sigma ~ normal(0,2);   # Halfnormal through parameter definition
    re_sigma ~ normal(0,2);    # Halfnormal through parameter definition

    random_effect[1:N_leaves] ~ multi_normal(zero_vector[],
            re_var_covar_matrix);

    for (i in 1:N){
        mu[i] = alpha + beta * x[i] + random_effect[leaf[i]];
    }
    y~normal(mu,sigma^2);
}
```

1.5.2 Binomial Phylogenetic Regression Stan Model

```
data {
    int<lower=1> N_patients;
    corr_matrix[N_patients] cor_mat;
    real x[N_patients];
    int<lower=0> y[N_patients];
    int patient[N_patients];   ## Group membership vector.
}
parameters {
  real alpha;
    real beta;
  vector[N_patients] random_effect;
}
transformed parameters{
    cov_matrix[N_patients] re_var_covar_matrix;
    re_var_covar_matrix = 1*cor_mat;
}
model{
    alpha                        ~ normal(0,1);
      beta                           ~ normal(0,1);

    random_effect[1:N_patients] ~ multi_normal(rep_vector(0,N_patients),
            re_var_covar_matrix);

    for (i in 1:N_patients){
        y[i]~bernoulli_logit(alpha +
          beta * x[i] +
          random_effect[patient[i]]);
    }
}
```

1.5.3 Binomial Phylogenetic Regression Stan Model, Multiple Response

```
data {
    int<lower=1> N_patients;
    int<lower=1> N_positions;
    int<lower=1> N;
    corr_matrix[N_patients] re_matrix;
    real x[N];
    int<lower=0> y[N];
    int patient[N];   ## Group membership vector.
    int position[N];   ## 'amino acid/position' each y is taken from.
}
parameters {
    vector[N_positions] alpha;
    vector[N_positions] beta;
    matrix[N_patients, N_positions] random_effect;
}
transformed parameters{
    cov_matrix[N_patients] re_var_covar_matrix;
    re_var_covar_matrix = 1*re_matrix;
}
model{
    alpha                          ~ student_t(1,0,2.5);
    beta                           ~ student_t(1,0,2.5);

    for(j in 1:N_positions){
        random_effect[1:N_patients,j] ~ multi_normal(rep_vector(0,N_patients),
                re_var_covar_matrix);
    }
    for (i in 1:N){
        y[i]~bernoulli_logit(alpha[position[i]] +
          beta[position[i]] * x[i] +
          random_effect[patient[i], position[i]]);
    }
}
```

2 HLA-Associated CTL Selection on HIV-1M Subtype-Specific Sites in Gag p24

2.1 Introduction

HIV-1M subtypes are viral lineages which all descend from the same HIV-1M cross-species transmission event. Subtypes are defined using a combination of phylogenetic and distance analyses (Kuiken, Korber and Shafer 2003); the genetic distance between subtypes varies by region. Within HIV-1M protein coding regions, certain amino acid positions vary consistently between subtypes; these are called subtype-specific sites (fig. 2.1). Subtype-specific sites are easiest to observe when comparing an alignment of within-subtype consensus sequences. The amino acids in subtype-specific positions often toggle between two or three different amino acids. Therefore, it is possible for a sequence which is phylogenetically a member of one subtype to have amino acids more common in another subtype on several of its subtype-specific sites.

Selective pressures on HIV-1 amino acids within the CTL epitopes that a given patient can present have long been associated with immune escape (Borrow et al. 1997). CTLs recognise infected cells because infected cells present viral epitopes bound to HLA on the cell surface. The viral peptide fragments which have the potential to become epitopes are typically found in clusters in hydrophobic parts of HIV-1 proteins (Tenzer et al. 2014). Interestingly, these clusters are often separated by subtype-specific amino acids. Previous studies have shown that the amino acid present in a subtype-specific position can affect epitope processing of all epitopes in an epitope cluster (Tenzer et al. 2009). Briefly, epitope processing is the process by which cellular proteasomes digest viral proteins, the resulting peptide fragments (epitopes) are then transported to the endoplasmatic reticulum where they bind to HLA. The resulting HLA-epitope complex is then transported to the cell surface for presentation to immune cells. The amino acid occupying a subtype-specific position can enhance or abrogate

epitope production, which is of importance to the success of the subsequent immune response, as HIV-1 epitope abundances are positively associated with the strength of the CTL response (Tenzer et al. 2009, 2014).

HIV-1M subtype diversification is believed to be the result of selectively neutral processes such as founder effects (Pybus and Rambaut 2009), where different lineages of HIV-1M were introduced into different regions and established localised epidemics. Here we advance another hypothesis, where the subtype-specific sites are under CTL selective pressure, and that these pressures drive subtype diversification.

We test this hypothesis using publicly available data by searching for associations between HLA variants and HIV-1M *gag* subtype-specific sites using the phylogenetic logistic regression developed in Chapter One. Such associations would indicate that these subtype-specific sites are under CTL selective pressure. Given the strong associations between HLA variants and ethnicity — which has described previously (Gragert et al. 2013) and is again shown here — this would suggest a role for HLA-driven subtype diversification as the virus spread through the world.

2.2 Methods

2.2.1 Obtaining the Data

We searched the LANL database for HLA-typed patients from whom HIV-1M *gag* had been sequenced. For each patient, we obtained the majority rule consensus for five *gag* subtype-specific positions (fig. 2.1). LANL HLA annotations were manually curated — only HLA class I annotations with four or more digits were used to generate an HLA presence-absence matrix, except in specific cases where a four-digit HLA variant could be inferred. To perform these inferences, we searched for American patients annotated with HLA B*27, and recategorised these as HLA B*27:05 if they had other HLA variants typical of Caucasians, or, if they had no HLA annotation other than HLA B*27 if they had the HLA B*27:05-KK10 typical escape mutations R264K, R264T, or L268M. American patients annotated with HLA B*57 in the LANL patients were reclassified as HLA B*57:01 based on the title of a study (Migueles et al. 2003); two other patients were classified as HLA B*57:01 carriers based on their other Caucasian HLA variants. We reassigned 6 American

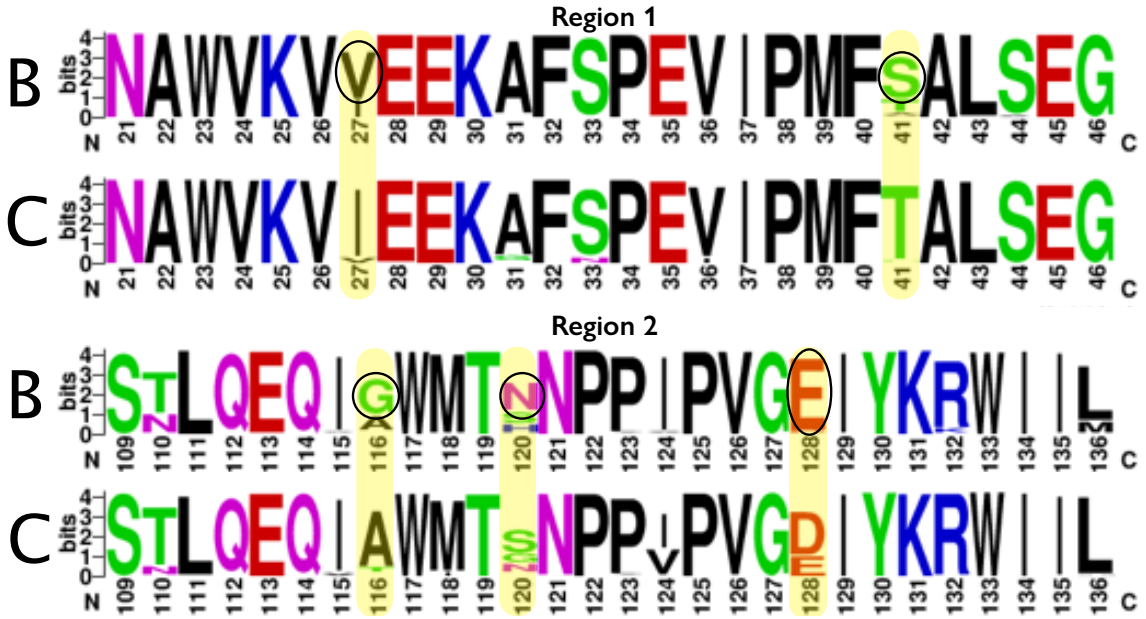


Figure 2.1: **The HIV Gag Regions of Interest**
Showing the subtype-specific sites 27, 41, 116, 120, and 128, the letters representing amino acids are sized according to their associated frequencies. Subtype B position 128 and subtype C positions 27, 41, and 116 are almost always occupied by the consensus amino acids and cannot be used in the analysis. The subtype B consensus amino acids are circled. This figure was generated using a single representative sequence from each HLA annotated patient.

patients annotated with HLA B*35 to B*35:01 because their other HLA variants suggested they were African-American.

We limited ourselves to subtypes B and C, the only subtypes for which there were enough sequences in the LANL database. We discarded position 128 in subtype B because over 90% of patients infected with subtype B were dominated by the subtype B consensus amino acid (Valine), giving us insufficient power to detect any HLAs that might be driving selection on this position. We discarded positions 27, 41, and 116 from the subtype C analysis for the same reason (see fig. 2.1).

2.2.2 Exploratory Principal Component Analysis

To confirm that the structure of the HLA presence-absence matrix was representative of expected ethnic and haplotype patterns in HLA frequencies, a principal component analysis (PCA) was performed. PCA has long been used in Quantitative Genetics to study population structure (Menozzi, Piazza and Cavalli-Sforza 1978). For this analysis, we used the binary HLA matrix where the rows represent patients, and the columns represent HLA variants. The entries of this matrix were 1 when

a patient was annotated with an HLA variant in the LANL database, and 0 otherwise. Because this high-dimensional matrix of binaries is uninterpretable when visualised directly, we instead used the PCA procedure to convert this matrix into a series of orthogonal vectors called principal components. Each principal component explains the greatest possible amount of variance left in the data after the variance explained by the previous principal components has been taken out. Therefore each subsequent principal component explains a smaller proportion of the HLA variance than the last, with the first principal component explaining the greatest proportion of the variance. Each principal components consist of a real number for every patient, which is much easier to visualise and understand than the high-dimensional HLA matrix; for instance by plotting the first two principal components (fig. 2.2).

2.2.3 Exploratory Amino Acid Multiple Correspondence Analysis

To explore the diversity of the subtype-specific amino acids, we used a Multiple Correspondence Analysis (MCA) trained on the five subtype-specific sites of the consensus sequences for each subtype and common CRFs obtained from LANL. The MCA is similar to the PCA but allowed for categorical variables (Tenenhaus and Young 1985). By training the MCA on the subtype consensus sequences, we can later calculate the two MCA components for each patient's majority rule subtype-specific amino acids. The MCA converts each patient's HIV subtype-specific position amino acid profile into two dimensions, allowing us to visually assess the degree to which the variability of subtype-specific amino acids is explained by subtype.

2.2.4 Modelling the Subtype-Specific Sites

Fitting a Bayesian Multilevel Model with a Patient Parameter

We built two different kinds of Bayesian models to find relationships between the patients' HLA variants and the majority-rule consensus amino acid on each subtype-specific site of the virus infecting each patient. These two models both account for different potential sources of variance. First, we built the Bayesian equivalent of a classical random-effect model, where an additional parameter (u_i) is fitted for each patient as follows:

n is the number of patients.

k is the number of amino acid positions.

l_j is the number of HLA variants used as explanatory variables for subtype-specific position j.

$\mathbf{Y}$ is an $n \times k$ matrix of binaries representing whether the majority rule amino acid in a patient is the same as the subtype B majority rule consensus amino acid (fig. 2.1).

$\mathbf{X}$ is a $n \times k \times l_j$ array of binaries representing the selected HLA variants for the jth subtype-specific position. This somewhat awkward 3-dimensional construction is necessary because we use different sets of HLA variants as explanatory variables for each amino acid. Hence a separate 2-dimensional HLA matrix is necessary for each position.

$\boldsymbol{\beta}$ is a $k \times \max(l_j)$ matrix representing the HLA coefficients for the K different amino acids. Vectors $\boldsymbol{\beta}_{j*}$ are not used in the model beyond index l_j. These unused parameters are necessary because ragged arrays are not available in Stan.

We define $\mathbf{P}$ as an $n \times k$ matrix containing the inferred probabilities p_{ij} of amino acid j in patient i being B-like.

$$\begin{aligned} \operatorname{logit}(p_{ij}) &= \alpha_j + \boldsymbol{\beta}_{j,1:l_j} \cdot \mathbf{X}_{i,j,1:l_j} + u_i \\ u_i &\sim \mathcal{N}_n(0, \sigma^2) \\ y_{ij} &\sim \text{Bernoulli}(p_{ij}) \end{aligned} \tag{2.1}$$

The patient parameter u in this model helps us account for variance introduced by the multilevel structure of the data, which includes multiple amino acid positions sequenced from the virus infecting each patient.

This parameterisation of the HLA regression ($\boldsymbol{\beta}_{j,1:l_j} \cdot \mathbf{X}_{i,j,1:l_j}$) was chosen instead of the more conventional parameterisation using dummy variable interaction-adjustment terms because there is little overlap between the HLA variants selected for each amino acid position. If we had performed this analysis with interaction terms, we would have had to include approximately $4 \times 8 \times 4 = 128$ HLA parameters for subtype B (4×8 reference level parameters, and the 3 additional interaction adjustment terms for the other positions). The alternative parameterisation used here reduced the number

of HLA parameters to 32. The l_j variable is necessary because different numbers of HLA variants are appropriate for each subtype-specific position.

Gelman's recommendation (Gelman et al. 2008) for weakly informative priors appropriately scaled for logistic regression was adopted; namely Cauchy distributions with a location parameter of 0 and a scale parameter of 2.5, equivalently known as t-distributions with 1 degree of freedom. Weakly informative priors aim to set the correct scale, rather than provide all possible prior information.

$$\alpha \sim \mathcal{T}(1, 0, 2.5)$$

$$\beta \sim \mathcal{T}(1, 0, 2.5)$$

$$\sigma \sim \text{Half-Normal}(0, 1)$$

All regressions were conducted on each subtype separately. Because we are studying viral amino acids, we consider patients to be related to one another through their virus's phylogeny. We do not consider host familial relationships in this model. We also do not consider relationships between HLA variants, which might be more or less similar to one another, because even closely related HLA variants can have very different effects on HIV evolution, e.g., HLA B*58:01 and HLA B*58:02 (Lazaryan et al. 2006).

Accounting for Phylogenetic Variance

We have shown in Chapter 1 that it is only anti-conservative to ignore possible phylogenetic correlations when both the explanatory and response variables may have plausibly evolved along the same phylogeny (See Chapter One). While the HIV-1M amino acids response variables evolved along the HIV phylogeny, it may appear implausible for there to be a phylogenetic signal in the HLA presence-absence matrix. This is because the HLA profiles of modern human populations evolved over (at least) the last 500,000 years while the HIV phylogenies only cover the last 30 years. However, covariance in the HLA dataset could be explained by the virus phylogeny if, for instance, there are antenatal transmissions of HIV in the dataset. Antenatal transmission and other transmissions between closely related individuals would lead to closely related viruses being present in related indi-

viduals (which have similar HLA variants) which could lead to virus phylogeny informed covariance in the explanatory HLA variables. Even in the absence of such transmissions between closely related individuals in the dataset, it would also be more mathematically precise to correct for phylogeny because the i.i.d. assumption is not met.

The phylogenetic correction was achieved by changing our patient parameter (u_i) to a parameter for patient/position combination (u_{ij}). Every column in the patient/position matrix (representing the phylogenetic effect for one subtype-specific position) was distributed according to a multivariate normal distribution over patients using a phylogenetically informed covariance matrix, thereby representing the similarity between patients through the relatedness of their viruses. We chose to use a separate phylogenetically-informed multivariate normal distribution for each subtype-specific position because using just a single multivariate normal (i.e., a random effect) would not have been able to accurately reflect two positions evolving independently along the same phylogeny (see Chapter One). Such a model with only one shared random effect would only work when all five sites are subject to the same evolutionary process, e.g. as tag-along mutations, a scenario which is adequately considered by the single random effect model (patient parameter model).

The model now looks as follows:

$$\text{logit}(p_{ij}) = \alpha_j + \boldsymbol{\beta}_{j,1:l_j} \cdot \mathbf{X}_{i,j,1:l_j} + u_{ij}$$

$$u_{*j} \sim \mathcal{N}_n(0, \sigma^2 \mathbf{C})$$

$$y_{ij} \sim \text{Bernoulli}(p_{ij})$$

Because this analysis is highly parameterised (one u_{ij} per data point), allowing σ to vary leads to the rapid growth of σ and the u_{ij} parameters until the inverse logit function collapses to 0 or 1 for all data points. In such circumstances it no longer matters exactly how much larger the u_{ij} parameters become, leading to instability of the σ parameter and non-convergence due to non-identifiability. To prevent this we set $\sigma = 1$. This choice of parameter is very reasonable in the context of a logit link function, which rapidly approaches 0 and 1 outside the range $[-2.71, 2.71]$.

2.2.5 Generating the Phylogeny and the Phylogenetic Variance-Covariance Matrix

For each subtype, we built manually curated alignments with a single representative sequence per patient and used the LANL HYPERMUT tool to remove hypermutated sequences. We used this alignment to generate a figure showing the amino acid frequencies in the regions of interest (fig. 2.1). To avoid biasing the analysis, we removed the five subtype-specific columns we used as response variables in the multilevel model, as well as columns subject to the following CTL escape mutations: A146P, A163G, S165N, T242N, G248X/A248X, R264K, and L268M, as these have been shown to bias phylogenies (Matthews et al. 2009). BEAST (Drummond et al. 2012) was used to sample from the posterior distributions of phylogenies by running ten chains for 200M steps with an HKY+G model of nucleotide evolution, a lognormal clock, and a logistic growth coalescent. Appropriate burn-ins were discarded (50M for subtype B; 100M for subtype C, except one subtype C chain for which we discarded a 150M burn-in). Convergence was confirmed using CODA (Plummer et al. 2006), with all ESS values on continuous parameters $\geq$ 200. We used TreeAnnotator to build a Maximum Clade Credibility (MCC) tree for each subtype, keeping the heights of the *maximum a posteriori* tree to prevent visualisation artefacts.

For every tree sampled from the posterior, the variances (diagonals) of the phylogenetic variance-covariance matrix were calculated by taking the phylogenetic distance (the sum of branch lengths) between the root node and each virus leaf node. The covariances (off-diagonals) were calculated by taking the shared phylogenetic history of two viruses, e.g. the distance between the root and the Most Recent Common Ancestor (MRCA) of the representative virus sequences from the two patients (Felsenstein 1985b; Freckleton, Harvey and Pagel 2002). By averaging over all trees sampled from the posterior, we generated an estimate of the posterior phylogenetic variance-covariance matrix scaled in years. The phylogenetic variance-covariance matrix was then transformed into a phylogenetic correlation matrix before being used in Stan.

2.2.6 Cross-Validation

To avoid overfitting the data, which could occur because we must consider a high number of HLA variants relative to the number of patients (subtype B: 26 HLAs / 292 patients; subtype C: 45

HLAs / 687 patients) we performed a 5-fold Cross-Validation (CV) procedure incorporating both feature selection and model fitting, which allowed us to optimise the bias-variance trade-off. The bias-variance trade-off is a balancing act between including too few parameters in the model or including too many parameters. Building models with too many parameters risks fitting the noise rather than the data (variance), while building models with too few parameters are likely to be biased. Because of the sparsity of our data and the computational feasibility compared to more complex methods, it was appropriate to use a simple feature selection regime. Within each training set, we calculated the correlation coefficients between each subtype-specific position and all HLA variants; we then constructed Bayesian models with the l best-correlating HLA variants for every subtype specific position and used the inferred parameters to predict the excluded validation sets. From these predictions, we calculated the Area Under the Curve (AUC) of the Receiver Operator Curves (ROC) values for each subtype-specific position. The ROC is a plot of the true positive rate (number of true positives / total number of positives) versus the true negative rate (number of true negatives / total number of negatives); the area under the ROC is a common metric of model accuracy for Bernoulli models. We chose this metric because it is invariant to the calibration of the latent probabilities in the Bernoulli model. Some people use the entire dataset to both fit the model and then to calculate the ROC, which leads to much higher AUC values but is highly sensitive to overfitting. Using CV is much more likely to lead to an unbiased estimate. The training and validation sets were constructed on the patient level to avoid upwardly biasing the AUC (Saeb et al. 2016).

This procedure was repeated for every l until all HLA variants were included in the model. In order to reduce CV error, after ten CV replications a LOESS (LOcally Estimated Scatterplot Smoothing) curve of the results was calculated for each subtype-specific site (fig. 2.6, fig. 2.7). In the CV model fitting procedure on the training set all l_js where identical. The optimum number of best-correlating HLA variants to include for each subtype-specific position was then used as the starting point for the parsimonious model (l_j in the actual analysis).

CV was performed on subtype B and C separately for both the patient-parameter and the phylogenetic model.

2.2.7 Parsimonious Model

We built parsimonious models by first taking the l_j HLA variants for each position which had the highest correlations with the amino acids, this time calculating the correlations using the entire dataset, where l_j was the optimum number of HLA variants for a subtype-specific position j according to the CV. We then fitted a multilevel Bayesian model using these predictors and refitted it after removing any HLA variants where the 80% BCI interval on the odds ratio failed to exclude 1.

We obtained posterior distributions from the model using the Stan MCMC sampler (Carpenter et al. 2016). We ran four chains for 2000 generations each, discarded the default 50% burn-in and confirmed convergence by ESS > 500 for all parameters and $\hat{R} < 1.01$; Gelman diagnostic plots were generated and found to be satisfactory (Gelman and Rubin 1992). Pairs plots of the beta posteriors were examined to check for collinearity between the HLA variants, which could arise if two or more included HLA variants were part of a common haplotype. Pair plots of the MCMC traces were examined at every step to search for potential interactions, which would point to colinearity between HLA variants, this could arise if two or more HLA variants were part of a common haplotype.

2.2.8 Goodness of Fit Testing

A half-normal quantile-quantile plot (q-q plot) (Collett 2002: 129–30) was used to evaluate the quality of the fitted Bernoulli model. Briefly, this works as follows:

1. For the latent p_{ij} parameters we took the median fitted value from the model for every amino acid j in every patient i, and used this to calculate Pearson residuals, which we then used to construct a q-q plot against the corresponding half-normal quantiles. These true residuals are shown as points.
2. We used a Bernoulli random number generator with the original estimates of the p to create 20 simulated data sets with newly generated amino acids but with the same explanatory variables as the true dataset. We then refit the model to each simulated data set and calculated the absolute values of those Pearson residuals.
3. On the half-normal q-q plot, we plotted lines representing the means of the simulated resid-

uals, along with the 95% envelope (the minimum and maximum from the 20 simulations).

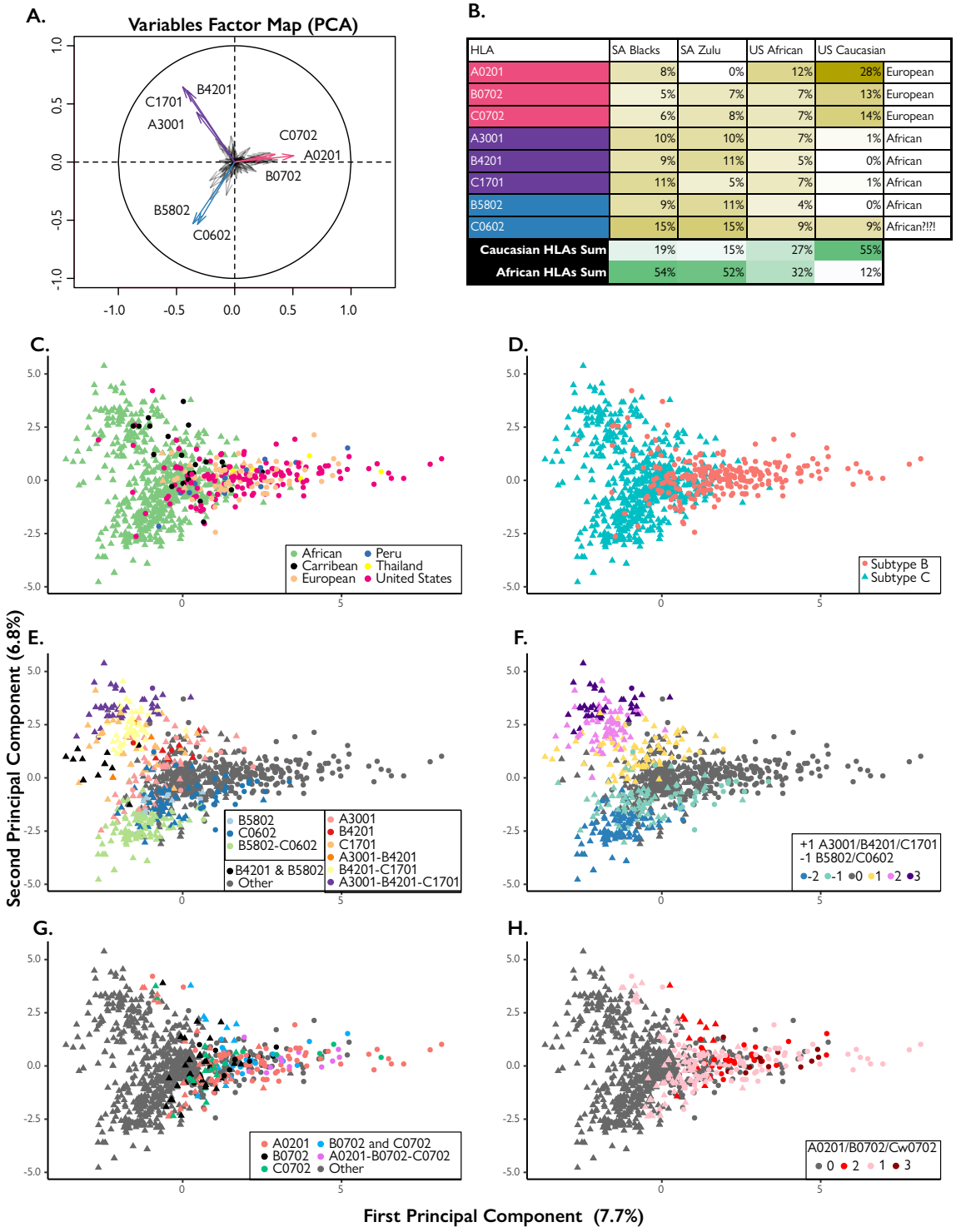
In this way, a typical half-normal q-q plot line was produced from a model which was genuinely Bernoulli, along with a 95% envelope; thus allowing the quality of the fitted model to be visually assessed.

2.3 Results

2.3.1 Principal Component Analysis

We used the principal component analysis to reduce the binary HLA presence-absence matrix to two real vectors of length n (number of patients), which can be visualised and understood. The first two principal component (PC1 & PC2) explain 7.7% and 6.8% of the variance respectively. Therefore fig. 2.2 visually represents 14.5% of the variance of the HLA presence-absence matrix. The first 12 principal components explain 50% of the variance. The variables factor map for the PCA is shown in fig. 2.2A, showing the HLA variants which contribute most to the first two principal components. European HLA variants are associated with higher values on the first principal component, while African HLA variants are associated with lower values (fig. 2.2C-F). The first principal component (x-axis in fig. 2.2) distinguishes between patients of African and European descent, as can be seen from the table in fig. 2.2B, where the HLA variants which most strongly influence the PCA have different frequencies in African and European populations (and presumably other Out-of-Africa populations). Note that the Caribbean patients are situated near the centre of the first principal component, which could reflect European and Native American admixture in these otherwise West-African populations (Benn-Torres et al. 2008).

The second principal component (y-axis in fig. 2.2) separates HLA variants belonging to the two most common South African haplotypes (A*30:01-B*42:01-C*17:01 and B*58:02-C*06:02), showing that these common haplotypes are the cause of considerable structure in this dataset (fig. 2.2). Although these HLA variants can also occur in other combinations, the linkage disequilibrium is almost complete (Paximadis et al. 2012). Therefore, this predominant structure is important to take into account in the subsequent regression, as including multiple HLA variants from the same haplotype could reduce statistical power due to collinearity.



HLA	SA Blacks	SA Zulu	US African	US Caucasian	
A0201	8%	0%	12%	28%	European
B0702	5%	7%	7%	13%	European
C0702	6%	8%	7%	14%	European
A3001	10%	10%	7%	1%	African
B4201	9%	11%	5%	0%	African
C1701	11%	5%	7%	1%	African
B5802	9%	11%	4%	0%	African
C0602	15%	15%	9%	9%	African?!?!
Caucasian HLAs Sum	19%	15%	27%	55%	
African HLAs Sum	54%	52%	32%	12%	

Figure 2.2: **Principal Component Analysis of Patient's HLA Profile**
The PCA transforms each patient's HLA profile into two dimensions (represented by a point in plots B-H); **A)** The Variables Factor Map shows which HLA variants contribute most to the first two principal components. **B)** table showing the population frequencies of the HLA variants which contribute most to the first two principal components. **C)** points representing each patient coloured by patient's country; PC1 (x-axis) separates patients of African and European origin. **D)** PCA coloured by virus subtype. **E)** The PC2 (y-axis) separates the two most common haplotypes in South Africa: B*58:02-C*06:02 and A*30:01-B*42:01-C*17:01. **F)** Patient's haplotype score is incremented by one for each allele from haplotype A*30:01-B*42:01-C*17:01 and reduced by one for each allele from haplotype B*58:02-C*06:02, showing near-complete stratification of PC2 due to these haplotypes. **G)** The common European haplotype A*02:01-B*07:02-C*07:02 **H)** patients are scored by the number of alleles they carry from the haplotype A*02:01-B*07:02-C*07:02. Showing less stratification than for the common African haplotypes, because PC1 separates Africans and Europeans generally.

2.3.2 Amino Acid Multiple Correspondence Analysis.

We used the MCA to reduce the patient's five HIV-1M subtype-specific amino acids to two real vectors of length n. These were then visualised in fig. 2.3, and coloured by the patient's subtype. This analysis revealed that the subtype-specific site variability is not entirely determined by subtype. Though many patient's HIV subtype-specific amino acids do cluster near a subtype consensus, many do not. Some subtype B viruses have the same subtype-specific amino acids as some subtype C viruses.

2.3.3 Phylogenetic Trees

To further address the question of whether the patient parameter model (which does not consider phylogeny) or the phylogenetic model (which does) is appropriate we examined the phylogenetic trees annotated with the subtype-specific positions (fig. 2.4, fig. 2.5). Every tip on the phylogeny represents a patient, and the four lines next to the tip show whether the majority rule amino acid in each of the four eligible subtype-specific positions was identical to the subtype B consensus amino acid. This visualisation allowed us to ascertain what kind of model is appropriate: if the pattern is horizontal we can assume that a single patient's subtype-specific amino acids change as a block (which can be modelled by a single patient parameter), if the pattern is more vertical than horizontal a model which corrects for each subtypes-specific position independently is necessary (requiring a separate multivariate normal random effect for each position, see Chapter One). Ignoring phylogenetic variance is only anti-conservative when both the explanatory (HLA variants) and dependent variables (subtypes-specific positions) are dependent on the phylogeny. In our case the amino acids on the subtype-specific positions are dependent on the phylogeny; however, the HLA variants in humans did not evolve along the HIV phylogeny. Indeed, any phylogenetic signal in the HLA variants could only be caused by mother-child transmission (even then the patients would share only half their HLA variants) and similar transmission events between closely related hosts. Because such transmission events would necessarily result in very closely related viruses, we can use our phylogenies to show that this happens only rarely in our dataset. However, not all closely related viruses are the result of transmission between closely related hosts. Most closely related viruses will be present in unrelated hosts, having been transmitted via the IDU or sexual routes. The vast majority of branches

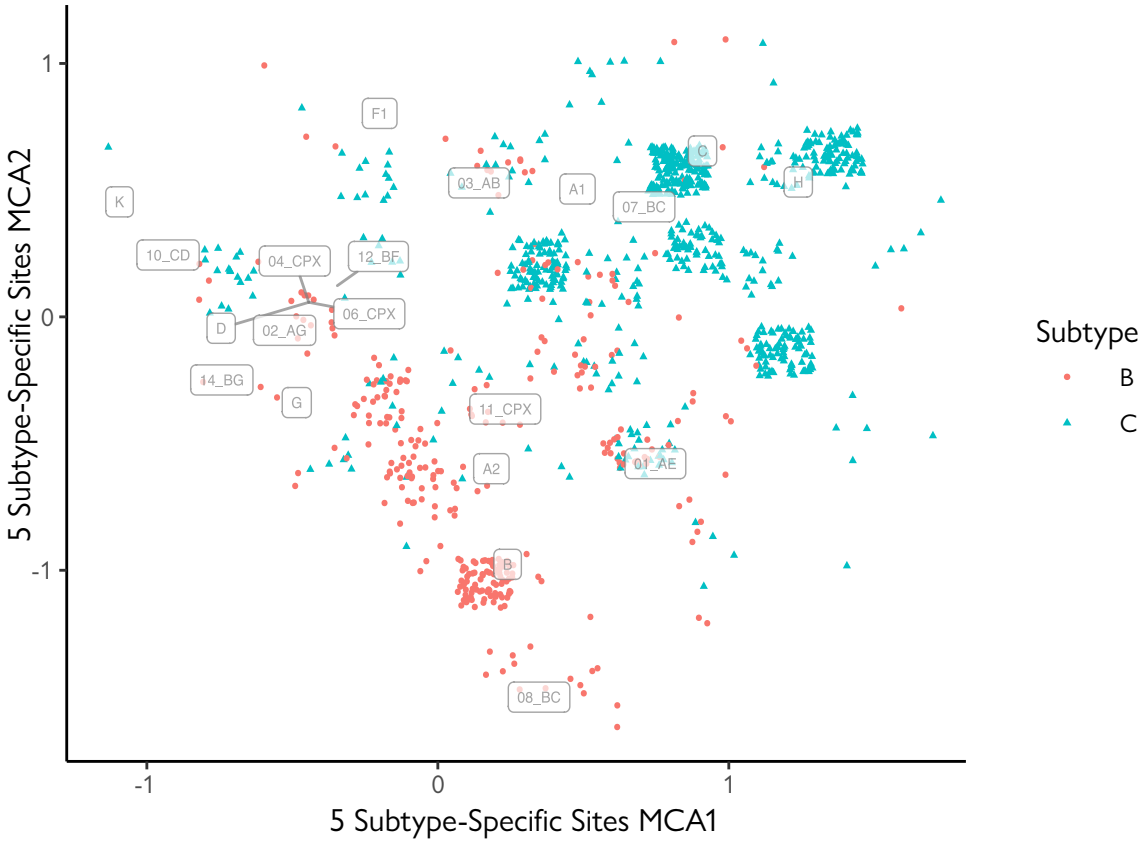


Figure 2.3: **MCA of Patient's HIV Subtype-Specific Amino Acids**
The MCA converts each patient's HIV subtype-specific position amino acid profile into two dimensions. This MCA was trained using the subtype consensus sequences (labels), the MCA was then used to project the patient's HIV subtype-specific amino acids onto the two-dimensional space. Points were jittered to prevent them obfuscating one another. While many patients subtype-specific amino acids are identical to the subtype consensus (see large clusters around the B and C labelled consensus sequences), some other subtype B and subtype C sequences are identical in the five Gag subtype-specific sites studied here (e.g., around the labelled 01_AE consensus).

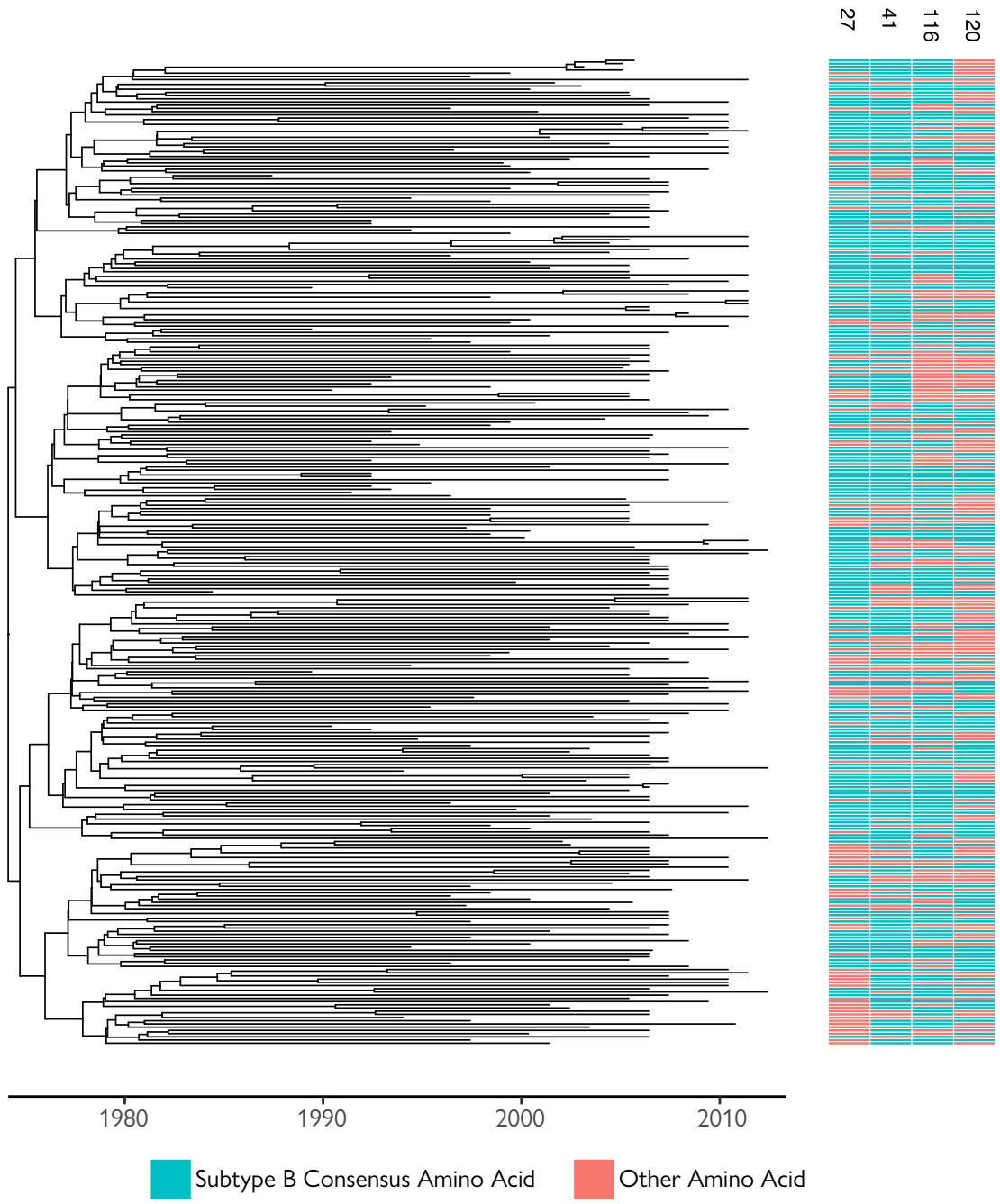


Figure 2.4: **Subtype B Phylogeny**
The subtype B phylogeny (BEAST Maximum Clade Credibility) annotated with the subtype-specific positions 27, 41, 116, and 120. Every leaf-node on the phylogeny represents a patient, and the four lines next to the leaf node show whether the majority rule amino acid in each of the four eligible subtype-specific positions was identical to the subtype B consensus amino acid. The amino acid patterns shown here follow the phylogeny (i.e. continuous patches of red are more vertical than horizontal, and are therefore shared between closely related sequences rather than shared between multiple positions on the same patient), and show the necessity of using a multiple response random effect model where each amino acid is allowed to evolve independently (Chapter One). If the colour pattern had been more horizontal that would have meant a single patient's subtype-specific amino acids tended to change as a block, in which case a single patient parameter (random effect model) would have been more appropriate than the multiple random effect phylogenetic model developed in Chapter One.

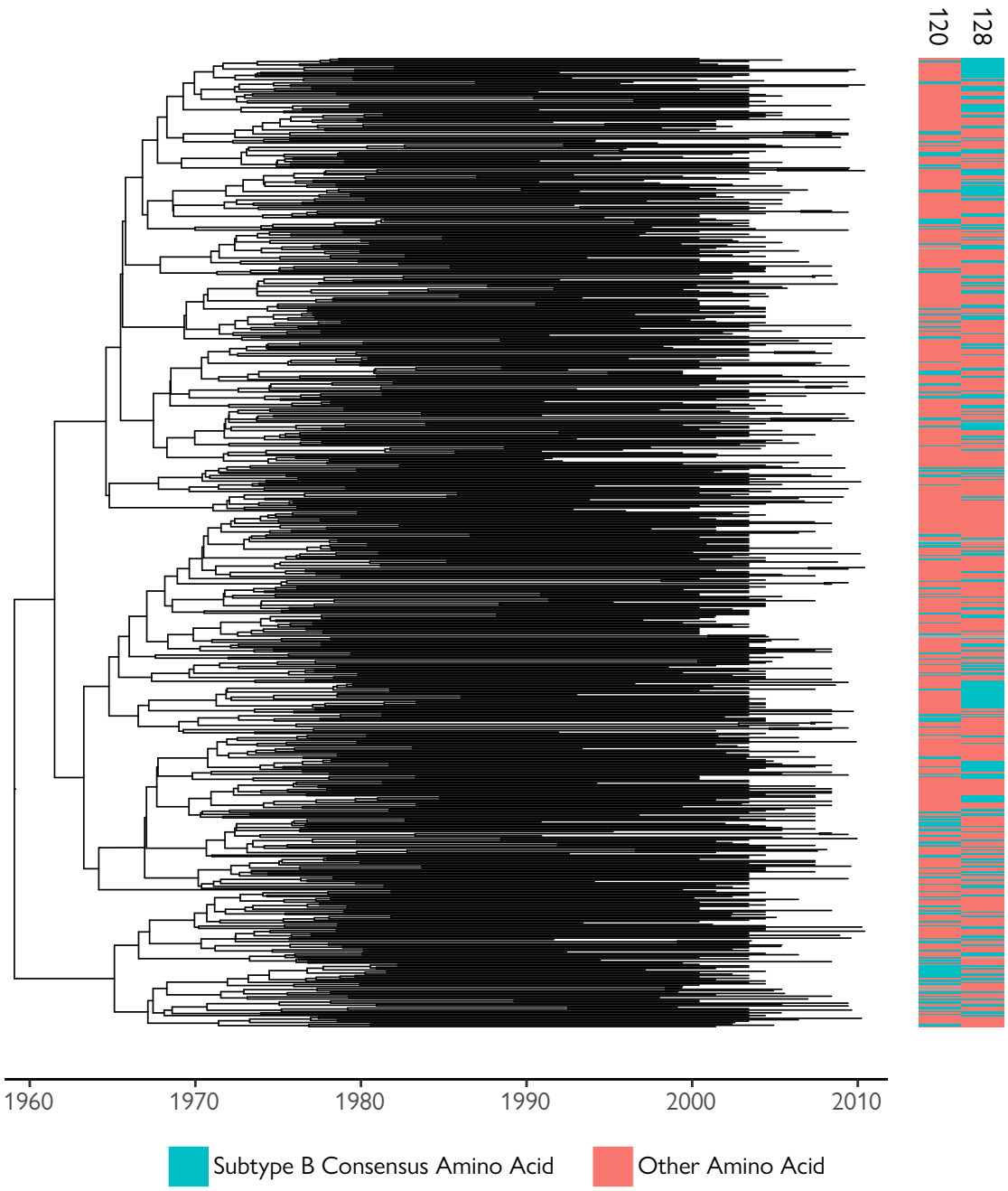


Figure 2.5: **Subtype C Phylogeny**
The subtype C Maximum Clade Credibility phylogeny annotated with the subtype-specific positions 120 and 128. Patients, where the majority rule amino acid is the same as the subtype B consensus (see fig. 2.1), are coloured blue; any other majority rule amino acid is coloured red.

coalesce early in the epidemic, as one would expect when sampling a few hundred patients from a many million strong epidemic which has undergone recent exponential growth.

2.3.4 Bayesian Regressions

We fitted two different models to the data. The first was a patient parameter model accounting for patient-level variance using a multilevel model which posits different amino acids sequences from the HIV infecting the same patient move together (e.g., other subtype-specific positions are more likely to be B-like or C-like if one of them is). The second was a phylogenetic model where we accounted for phylogenetic variance using multivariate normal distributions informed by a phylogenetic variance-covariance matrix (this model allowed each position to evolve independently according to the phylogeny calculated using BEAST). Unfortunately, model comparison is difficult in Stan (Carpenter et al. 2016), in part because Stan doesn't distinguish between likelihoods and priors making it challenging to calculate Bayes Factors or Bayesian Information Criteria using techniques such as path sampling. Bayes Factors have the additional disadvantage of being sensitive to prior specification. Instead, we compared the models using the same CV approach we used to determine the appropriate number of HLA variants to include in the analysis (fig. 2.6, fig. 2.7). We find that both models are almost identical in their prediction accuracy. Any apparent differences between the phylogenetic and non-phylogenetic accuracy curves are more likely to be the result of CV error rather than real differences. The high computational cost of the phylogenetic model prevents us from running additional replicates to further reduce the CV error.

Besides serving as a tool for model comparison the CV allowed us to find a balance in the bias-variance trade-off; thereby preventing overfitting this highly parameterised problem by finding the point at which adding additional HLA variants ceased to improve our predictions for a particular subtype-specific position (fig. 2.6, fig. 2.7). In the CV adding additional HLA variants was always beneficial for subtype B position 27, as the CV prediction accuracy kept improving. In the non-phylogenetic model, for subtype B positions 41 and position 116, the optimum number of HLA variants was 9 (12 and 14 HLA variants respectively for positions 41 and 116 in the phylogenetic model). The predictions for subtype B position 120 are never more accurate than guessing randomly; however, it does appear that the predictions improve as more HLA variants are added. Still, the fact

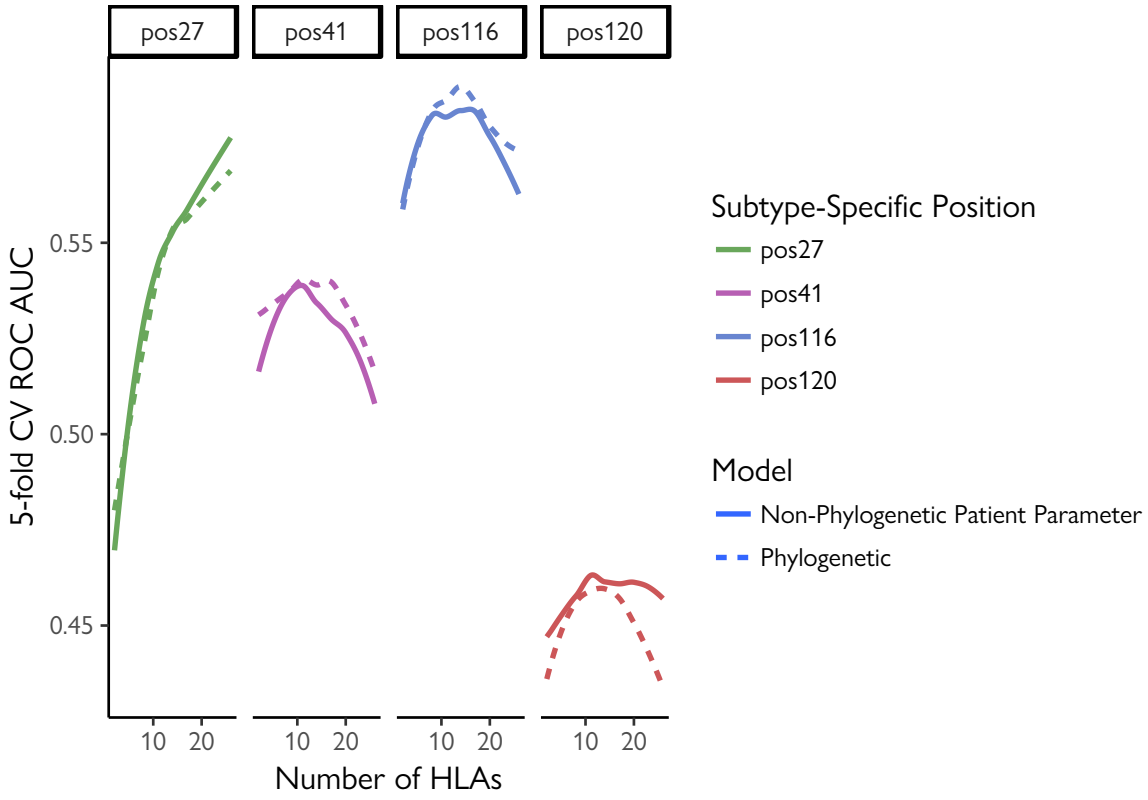


Figure 2.6: **Subtype B: LOESS Curve of Cross-Validation Prediction Accuracy**
Predictions for subtype B positions 27, 41, and 116 are made more accurate by including more HLA variants, position 27 gets better irrespective of how many HLA variants are added, while the performance of the model declines for positions 41 and 116 after adding the best 9 HLA variants (12 and 14 HLA variants for positions 41 and 116 respectively in the phylogenetic model). The predictions for position 120 never go above the ROC AUC of 0.5, which is the expected value when guessing randomly.

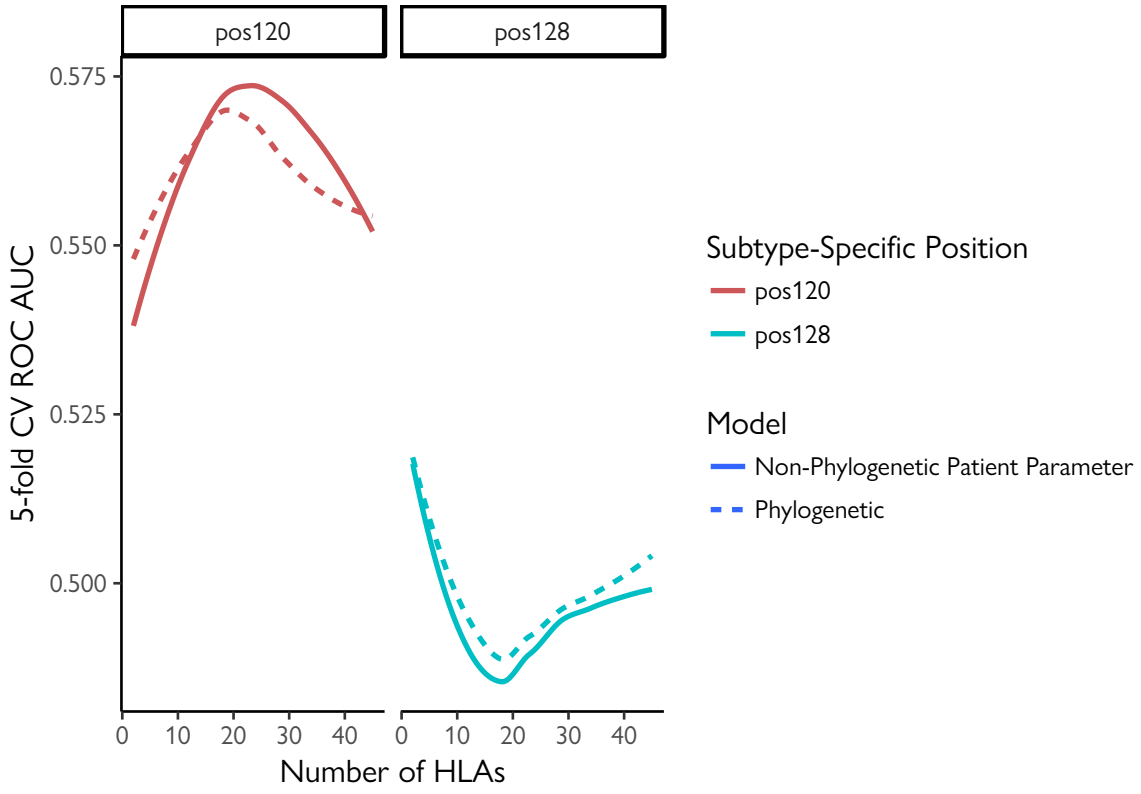


Figure 2.7: **Subtype C: LOESS Curve of Cross-Validation Prediction Accuracy**
Prediction accuracy for subtype C position 120 peaks after 22 HLA variants are added in cross-validation (18 HLA variants in the phylogenetic model), while the model accuracy for position 128 is only as good as guessing randomly when 2 HLA variants are included, and deteriorates if more than 2 HLA variants are included.

that the model has no predictive accuracy for position 120 may have implications for the reproducibility of this position. For subtype C the results are similar: in the non-phylogenetic model, in position 120, the CV accuracy peaks at 18 HLA variants (22 HLA variants for the non-phylogenetic model), and position 128 peaks at 2 HLA variants (same in phylogenetic model). The number of HLA variants which gave the highest accuracy in the CV were used as the starting point for the construction of the parsimonious model. It should be noted that the predictions for position 128 are only slightly better than random. Because of these slight differences in where the CV accuracy peaks between the non-phylogenetic and phylogenetic models, some HLA variants were included in only one of the two analyses. Furthermore, during the construction of the parsimonious model, some HLA variants which were on the borderline of the inclusion threshold did not make the cut in one of the models. Together, this has resulted in slightly disjoint sets of HLA variants included in the phylogenetic and non-phylogenetic models.

The Bayesian framework has the advantage of providing probability distributions for the parameters, so that true probability statements such as credible intervals become available, rather than artificial constructs such as classical confidence intervals. We summarise the parameters by specifying 95% Bayesian Credible Intervals (BCI); an equal-tailed interval was chosen because it is invariant under transformation between the β and odds ratio scale (another commonly used summary, the 95% Highest Posterior Density is sensitive to this transformation, which makes it poorly suited for use in Binomial regressions). In light of the similarity observed in the CV, it is perhaps unsurprising that the posterior densities over the odds ratios are very similar between the patient-parameter and phylogenetic models (fig. 2.10, fig. 2.12). The probability distributions on the odds ratios in the patient-parameter model are slightly further away from 1, probably because this model fits only a single parameter per patient, while the power for the phylogenetic model is reduced by fitting a parameter for every amino acid in every patient. This over-parameterisation is necessary to allow each subtype-specific position to evolve independently following the phylogeny, but it does result in the half-normal plots for the parsimonious phylogenetic models being less satisfactory owing to overfitting. This overfitting occurs because there is only one data point per patient/amino acid combination, which is insufficient data to allow the model to converge to the true simulated value, resulting in 95% intervals (dotted lines in half-normal plots) which exclude the true residuals. In

Selection Order	**Position 27**	**Position 41**	**Position 116**	**Position 120**
1	C1601	B2705	B5703	C0304
2	C0304	C0401	A0201	A0201
3	B3501	A0301	B0702	C0303
4	A2301	A2301	B2705	C1203
5	B0801	B0702	B5701	B1501
6	B5101	A1101	B1501	B2705
7	A2402	A3101	C0602	A2402
8	B5703	B4403	C0401	C1601
9	B2705	B3501	C0702	A0101
10	C0102	C0702	B3501	B5101
11	B1501	C1203	C1203	B5701
12	B1801	B5703	B4402	C0602
13	A0301	B0801	A3101	B4403
14	A3101	C0701	A0301	B4402
15	A1101	A2402	C0701	B3501
16	C0602	B1801	B5101	C0702
17	B5701	A0101	B1801	A3101
18	B0702	B5101	A1101	A0301
19	B4403	C1601	A2402	A2301
20	A0101	C0602	A0101	B0702
21	C0702	B5701	B4403	C0701
22	C0701	B4402	A2301	C0401
23	B4402	A0201	C1601	C0102
24	C1203	C0102	C0102	B1801
25	A0201	B1501	C0304	B5703
26	C0401	C0304	B0801	B0801

Colours:
Included in Both Models
Only Non-Phylogenetic Model
Only Phylogenetic Model
Excluded from Both Models

Figure 2.8: Subtype B: The order in which HLA variants were selected using CV. The differences between the non-phylogenetic model and the phylogenetic model are due to the slightly different peaks in fig. 2.6

Selection Order	**Position 120**	**Position 128**
1	A6802	B4202
2	B5801	B1510
3	B4202	B5301
4	A0101	A3001
5	C0210	A0301
6	A0205	C0702
7	B1503	A3004
8	A3001	B4501
9	B3910	C0704
10	B1510	C0602
11	C0802	A4301
12	A4301	B5802
13	C0202	C0802
14	C1801	B8101
15	B4403	B1503
16	B5802	B3910
17	C0602	A6801
18	A2301	B4201
19	B8101	B5801
20	A3402	B0702
21	C0801	C0801
22	C0704	C0210
23	C0304	C0304
24	A2902	A3402
25	B5301	A0201
26	B4501	A6802
27	C1601	C0401
28	A6801	B0801
29	A0301	C1701
30	B1401	B1401
31	C0701	C0202
32	B5703	B5703
33	A7401	A6601
34	A3004	A0205
35	A3002	A7401
36	C0702	A3002
37	B0702	A2402
38	A0201	A0101
39	C1701	A0202
40	A2402	B4403
41	A6601	C0701
42	B0801	C1801
43	A0202	A2301
44	C0401	C1601
45	B4201	A2902

Colours:
Included in Both Models
Only Non-Phylogenetic Model
Only Phylogenetic Model
Excluded from Both Models

Figure 2.9: Subtype C: The order in which HLA variants were selected using CV. The differences between the non-phylogenetic model and the phylogenetic model are due to the slightly different peaks in fig. 2.7

contrast, the residuals from the patient-parameter model fell mostly inside the envelope (fig. 2.14, fig. 2.15), showing that this problem is not caused by the entirely satisfactory Bernoulli link function, but by the unavoidable overparameterization of the phylogenetic model. It should be noted that although this overparameterisation negatively affects the half-normal q-q plot model validation, it only mildly changes the inferred odds ratios and does not affect the CV prediction accuracy.

Because the results from the phylogenetic analyses are generally more conservative, we report the medians and 95% BCI from these rather than the patient-parameter analyses. It should be noted here that owing to our Bayesian approach we are not attempting to reject a null hypothesis but are instead quantifying uncertainty directly. Therefore, our results are fit for interpretation even if the 95% BCI over the odds ratios does not exclude 1.

Subtype B

We identified the HLA variants that excluded 1 from the 95% BCI interval on the odds ratios (OR) in positions 27, 41, and 116 (fig. 2.10). Surprisingly there is one from HLA-A, two from HLA-B and four from HLA-C, rather primarily from HLA-B; most previously described associations with virus genotypes implicated HLA-B (Kiepiela et al. 2004). We find many of the HLA variants associated with CTL pressure on the subtype-specific sites are derived from Neanderthals or Denisovans (fig. 2.11); this contrasts with the non-Denisovan/Neanderthal HLA variants driving the evolution of subtype-specific sites in subtype C (figs. 2.13, 2.12).

We examined the pairs plots for the betas associated with each position in the parsimonious model, and only found an interaction between A*03:01, B*07:02 and C*07:02 in subtype B position 27. These HLA variants are part of a common haplotype composed of HLA variants with opposing selective effects; B*07:02 selects for the subtype B consensus amino acid (V) and A*03:01 or C*07:02 (or both) select for the subtype C consensus amino acid (I). Leaving out either A*03:01 or C*07:02 from the regression turns B*07:02 into a proxy variable for the entire haplotype and moves the B*07:02 beta closer to the odds ratio of 1 (no selective effect). Removing A*03:01 and C*07:02 from the regression moves the B*07:02 beta even closer an odds ratio of 1. Similarly, leaving out B*07:02 weakens the effects associated with A*03:01 and C*07:02.

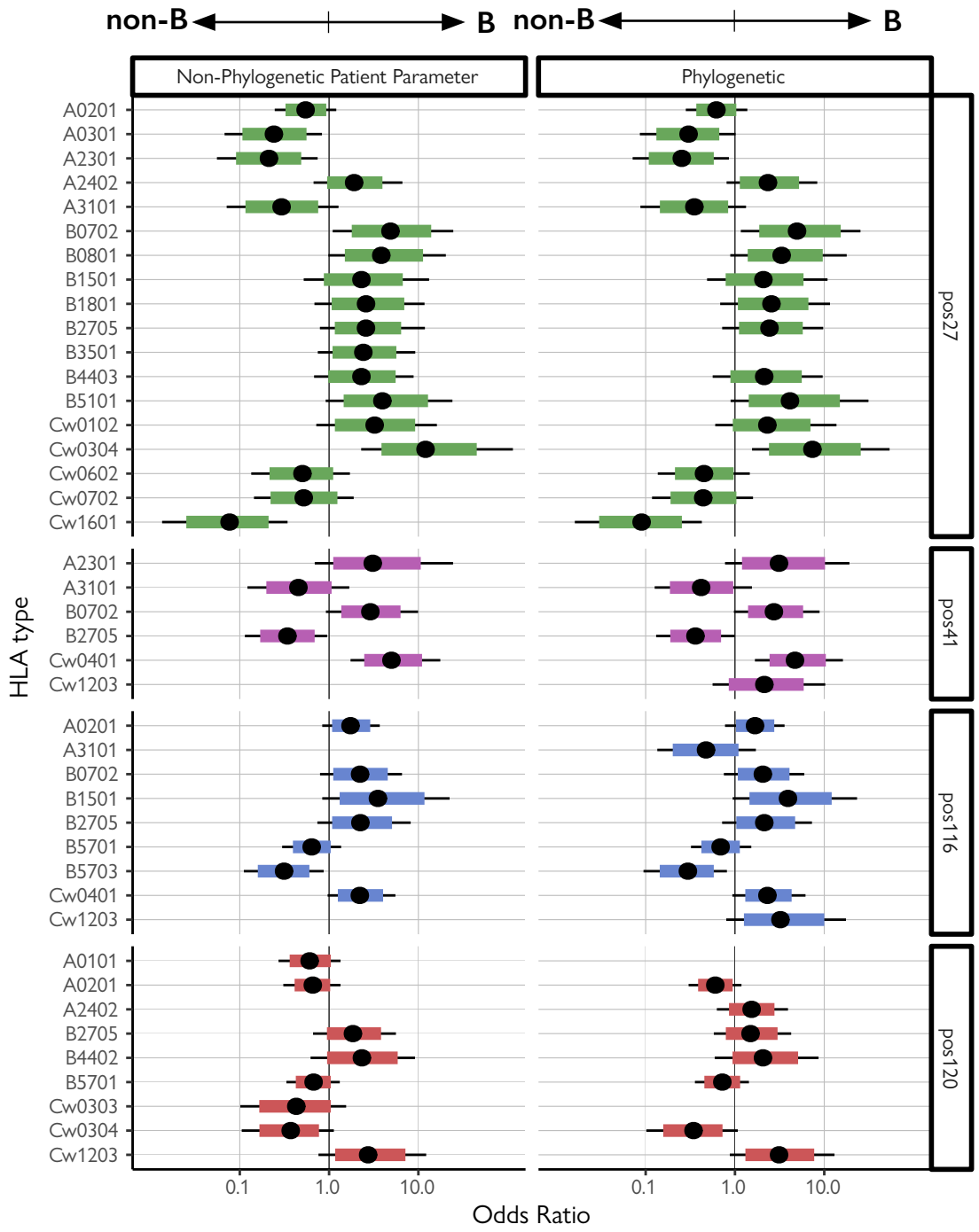


Figure 2.10: **Subtype B: The Posterior Probability Distributions on HLA Associated Odds Ratios**
The intervals on the left are derived from the non-phylogenetic patient parameter model (Bayesian random effect model), and the intervals on the right are derived from the phylogenetic model with separate phylogenetic multivariate normal distributions for each subtype-specific site. Odds Ratios greater than one represent the belief that an HLA variant is associated with the subtype B consensus amino acid in that position for the parsimonious models. The central point represents the median, the shaded region is the 80% BCI interval, and the whiskers represent the 95% BCI interval. In some cases an HLA variant is missing from one or the other because different numbers of HLA variants were selected by the CV, but generally, the results are very similar.

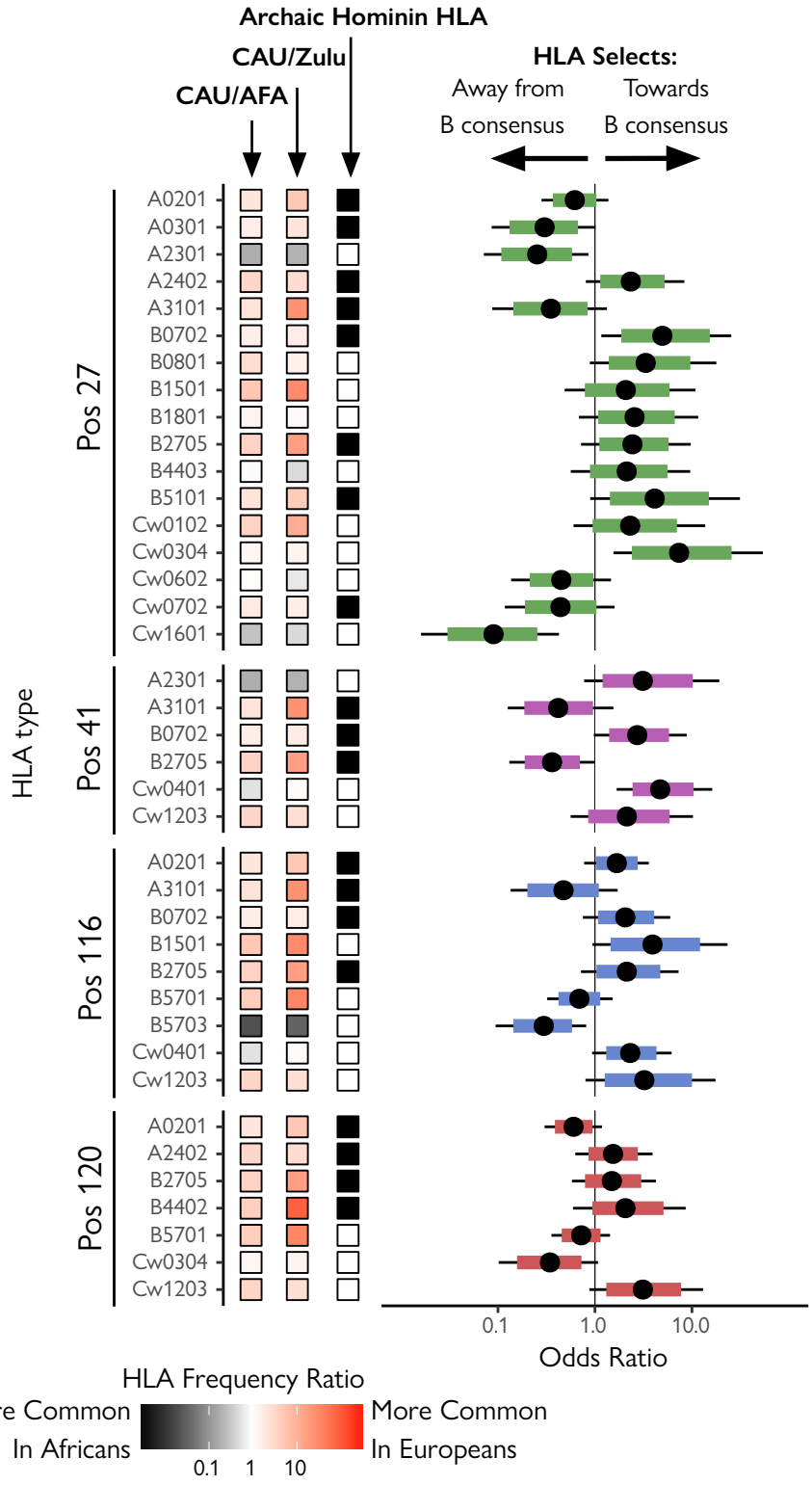


Figure 2.11: **Subtype B: Odds Ratios From Phylogenetic Model Annotated with HLA Frequency ratios** in African Americans (AFA), Zulus, and US Caucasians (CAU). HLA variants which had an origin in Neanderthals or Denisovans are shown under 'archaic hominin HLA'. Half of the HLA variants associated with CTL pressures in the subtype B analysis originate in Neanderthals/Denisovans, making for a very different selective environment.

We did not attempt a regression for position 128, as there were too few non-subtype B consensus amino acids in that subtype-specific position (fig. 2.1). We did attempt a regression for position 120, but the CV showed that we were unable to build a robust model. This inability to build a robust model is probably owing to the high amino acid variability in that position, and our limited sample size.

Position 27

Most of the 95% credible effects are found for position 27, which accounts for 3/8 HLA variants which credibly exclude the 95% BCI in the subtype B results (fig. 2.10). The strong results from the parsimonious model for position 27 correspond with the results from the subtype B CV, where the prediction accuracy for position 27 continuously increases the more HLA variants are added to the model (fig. 2.6). We find that the HLA variants A*23:01, Cw*03:04 and C*16:01 exclude the 95% BCI, showing that these HLA variants are exerting selective pressure on position 27. A*23:01 (OR median: 0.25, 95% BCI 0.07 – 0.86), and C*16:01 (OR median: 0.09, 95% BCI: 0.02 – 0.43) are selecting for the subtype C consensus amino acid (I), while C*03:04 (OR median: 7.36, 95% BCI: 1.55 – 53.86) is selecting (possibly very strongly) for the subtype B consensus amino acid (V).

Only in position 27 did we find correlations in the β trace scatterplots, suggesting HLA colinearity. Interestingly we found a negative correlation between the A*03:01/C*07:02 β and the B*07:02 β (and no correlation between the betas representing A*03:01 and C*07:02, possibly suggesting that they are both independently selecting for the subtype C consensus amino acid). This negative correlation is the opposite of what we were expecting to find in the pairwise scatterplots for all HLA variants associated with a position, which was a way of guarding against collinearity obscuring the effect of HLA variants that form part of a common haplotype. Instead, the negative correlation suggests that these HLA variants are strengthening each other's effects. We are only likely to pick up such multivariate effects in subtype B position 27, where every HLA variant is included because the univariate feature selection regime would be unlikely to include these HLA variants because it is blind to their interaction. This insensitivity to other potential interactions is the price we pay for knowing that we have avoided overfitting this highly parameterised problem.

Position 41

The HLA variant B*27:05 (OR median: 0.36, 95% BCI: 0.13 – 0.99) selects for a non-subtype B consensus amino acid. B*27:05 is known to be associated with an intra-epitope mutation R264X (position 132 in fig. 2.1) in Gag p24 (Kelleher et al. 2001). However, that mutation is located over 100 amino acids away from position 27. Previously amino acid covariation between R264X and position 41 has been shown to be due to tertiary structure (Carlson et al. 2008; Crawford et al. 2011). Two positions might be separated by 100 amino acids, but if those positions are near one another in the folded protein, they may covary.

The HLA variant C*04:01 (OR median: 4.70, 95% BCI: 1.67 – 16.17) select for the subtype B consensus amino acid serine (S). The 95% BCI interval for B*07:02 doesn't quite exclude 1, but it does appear to have a potentially strong selective effect for the subtype B consensus serine (OR median: 2.73, 95% BCI: 0.97 – 8.86).

Position 116

HLA variant B*57:03 (OR median: 0.30, 95% BCI: 0.09 – 0.81) is associated with the subtype C consensus amino acid (alanine) on this position. There are potentially other effects in the opposite direction from A*02:01, B*07:02, B*15:01, C*04:01 and C*12:03 but these do not fully exclude 1 from the 95% BCI.

Position 120

The C*03:04 credible interval (OR median: 0.34, 95% BCI: 0.10 – 1.08) nearly excludes 1 showing that it may select against the subtype B consensus amino acid (asparagine, abbreviation: N). The variability in this position (there are three common amino acids) make it difficult to make a definitive statement about which amino acids are being selected for. The low prevalence of the subtype B consensus amino acid may have resulted in the CV results below an AUC of 0.5, which reduces our confidence that the position 120 results will be reproducible.

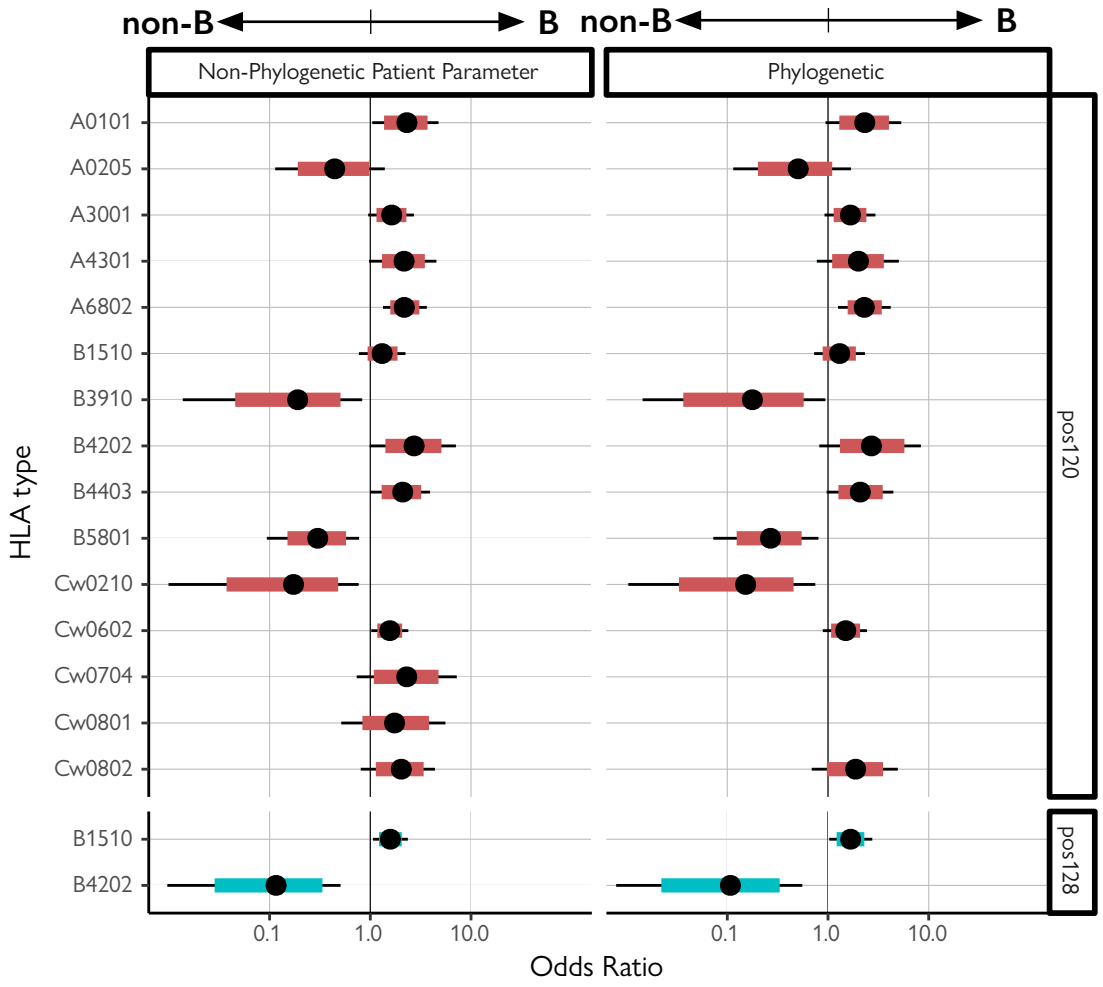


Figure 2.12: **Subtype C: The Posterior Probability Distributions on HLA Associated Odds Ratios**
The intervals on the left are derived from the non-phylogenetic patient parameter model (Bayesian random effect model), and the intervals on the right are derived from the phylogenetic model with separate phylogenetic multivariate normal distributions for each subtype-specific site. Odds Ratios greater than one represent the belief that an HLA variant is associated with the subtype B consensus amino acid in that position for the parsimonious models. The central point represents the median, the shaded region is the 80% BCI interval, and the whiskers represent the 95% BCI interval. In some cases an HLA variant is missing from one or the other because different numbers of HLA variants were selected by the CV, but generally, the results are very similar.

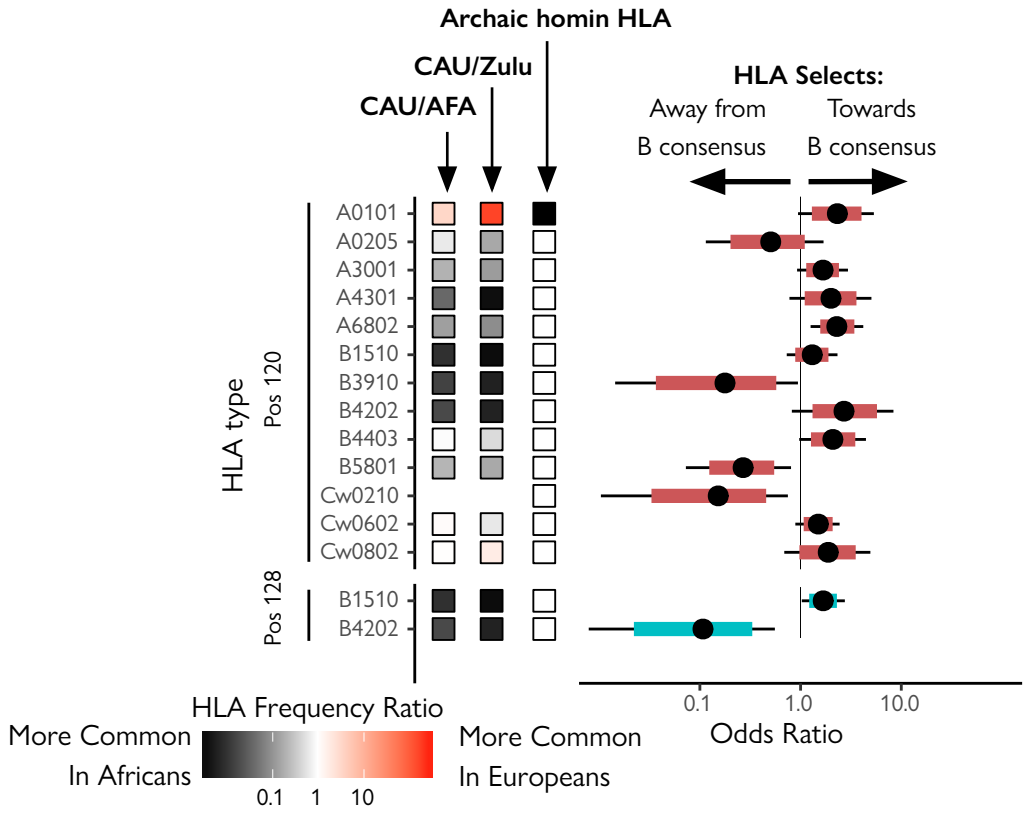


Figure 2.13: **Subtype C: Odds Ratios from phylogenetic model Annotated with HLA Frequency ratios** in African Americans (AFA), Zulus, and US Caucasians (CAU). HLA variants which had an origin in Neanderthals or Denisovans are shown under 'archaic hominin HLA'. None of the HLA variants associated with CTL selective pressures on the subtype-specific sites in the subtype C analysis originates in Neanderthals or Denisovans.

Subtype C

We did not attempt to do a regression for subtype C positions 27, 41 and 116 as there was not enough amino acid variability in those positions. Only the regression for subtype C position 120 worked well in CV.

Position 120

A*68:02 (OR median: 2.30 95% BCI: 1.30 – 3.89) are selecting for the subtype B consensus amino acid. The 95% BCI for HLA variants A*01:01, A*30:01, B*44:03 and C*06:02 only barely include 1, showing that these HLA variants may be selecting for the subtype B consensus amino acid.

The HLA variants B*39:10 (OR median: 0.20, 95% BCI: 0.02 – 0.92), B*58:01 (OR median: 0.28, 95% BCI 0.08 – 0.71) and C*02:10 (OR median: 0.16, 95% BCI: 0.01 – 0.69) are selecting against the subtype B consensus amino acid.

Position 128

B*42:02 (OR median: 0.12, 95% BCI: 0.01 – 0.59) is selecting against the subtype B consensus amino acid (E). B*15:10 (OR median: 1.72, 95%BCI: 1.03 – 2.62) is selecting for the subtype B consensus amino acid. In this position, we are worried about replication, as the CV revealed our model is not predictive for this site, and hence we are likely to be overfitting the data.

2.4 Discussion

Here we have shown that the HIV subtype-specific positions are associated with patient HLA types, suggesting that the subtype-specific positions are under CTL selection. Modern HLA frequencies stem from an ongoing process of pathogen-driven selection (Doherty and Zinkernagel 1975; Prugnolle et al. 2005; Spurgin and Richardson 2010), this selection also resulted in the enriched introgression of the HLA genes from archaic hominins encountered by populations of modern humans venturing into Eurasia (Abi-Rached et al. 2011; Ségurel and Quintana-Murci 2014; Patin et al. 2017). Introgression from archaic hominins essentially allowed for additional years of adaptation to Eurasian pathogens (i.e. Neanderthals left Africa at least 300,000 years ago (Meyer et al. 2014),

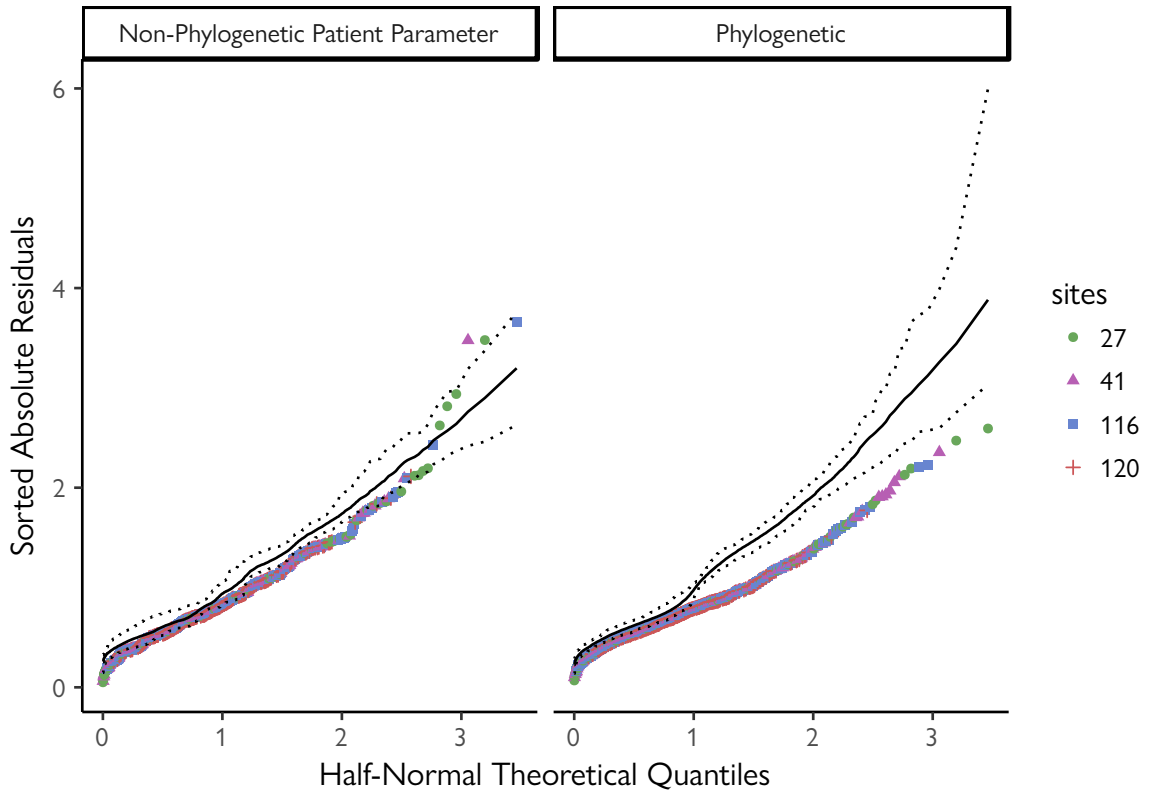


Figure 2.14: **The Half-Normal Quantile-Quantile Plots for Subtype B**
showing the sorted absolute residuals from the real models as points, with the 95% envelope of sorted absolute residuals from simulated data assuming the fitted model was true shown as lines. The non-phylogenetic model fits better than the phylogenetic model, probably because there is insufficient power to recover the many additional parameters in the phylogenetic model. This overparameterisation is the price we pay for phylogenetic correction.

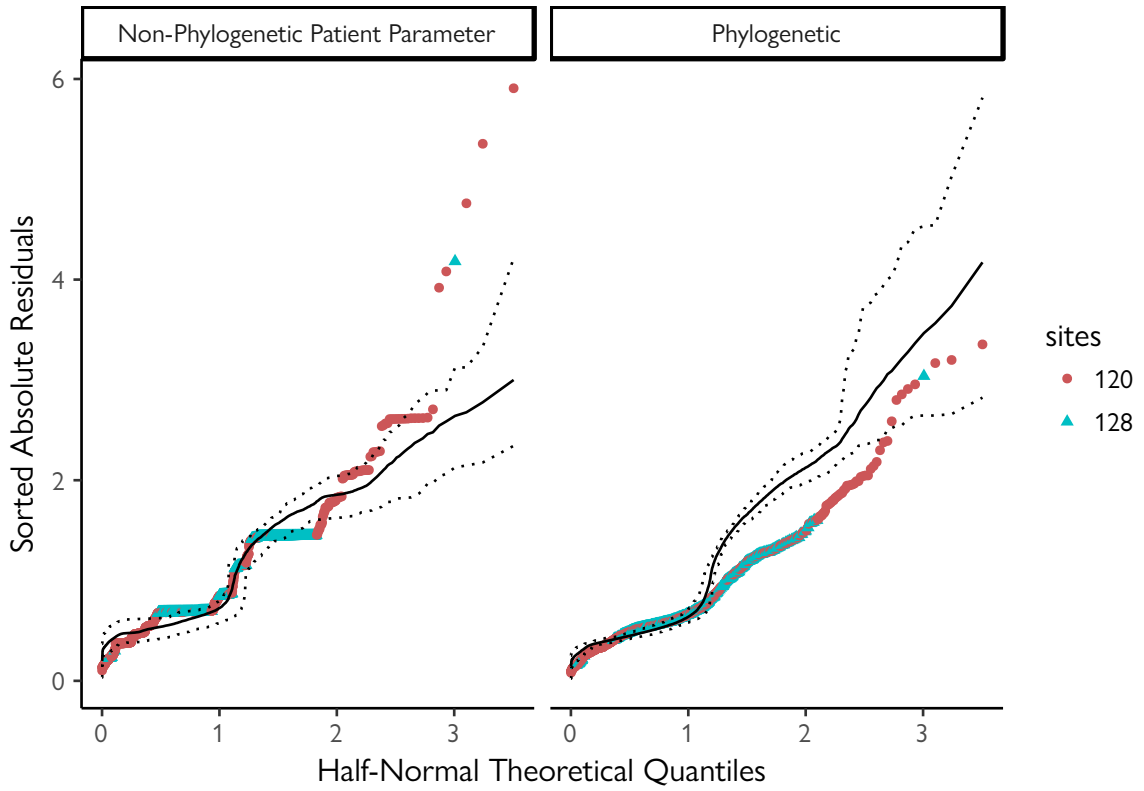


Figure 2.15: **The Half-Normal Quantile-Quantile Plots for Subtype C** showing the sorted absolute residuals from the real models as points, with the 95% envelope of sorted absolute residuals from simulated data assuming the fitted model was true shown as lines. The non-phylogenetic model fits better than the phylogenetic model, probably because there is insufficient power to recover the many additional parameters in the phylogenetic model. This overparameterisation is the price we pay for phylogenetic correction.

compared to the modern humans who left 56,000 years ago (Fu et al. 2013)). These great differences between HLA frequencies in human populations (Gragert et al. 2013), combined with the fact that subtypes tend to be confined to certain ethnic groups, suggests a role for CTL selection in the creation and maintenance of subtype-specific sites. Indeed, it suggests that subtype-specific amino acid differences are the result of a long and ongoing selection process resulting in the adaptation of the virus to the most common HLA variants in each population in order to abolish or minimise CTL epitope production.

We considered implementing several different kinds of negative control. However, this was difficult for several reasons. One obvious negative control would be a non-subtype-specific site. Unfortunately, these sites had insufficient intra-subtype variance in the studied p24 Gag region to be analysed (fig. 2.1). We then considered using a different HIV-1 protein, but the LANL immunology database showed CTL epitopes could be generated from all HIV-1 proteins. The potential CTL responses and the subsequent viral escape could lead to true associations between amino acids from those proteins and patient HLA variants. Finally, we considered synonymous sites. Unfortunately, preliminary analyses showed that the majority of synonymous sites with sufficient intra-subtype variation were nucleotide mutations typical of those induced by APOBEC$_3$G and APOBEC$_3$F (host immune factors which cause hypermutation of the virus). These sites are not appropriate for a negative control because they could be influenced by CTL responses against Vif, an HIV protein which binds APOBEC$_3$G and prevents it from being incorporated into nascent virions and hypermutating HIV-1. Effective CTL responses against Vif are likely to cause escape mutations which reduce the efficiency of Vif, thereby leading to associations between hypermutated sites and host HLA variants (Armitage et al. 2008, 2012).

Our findings contrast considerably with the current paradigm in which subtype-specific sites are the result of selectively neutral processes such as founder effects (Pybus and Rambaut 2009). However, the selection favouring individuals which carry HLAs well-adapted to local diseases is an accepted phenomenon (Prugnolle et al. 2005; Spurgin and Richardson 2010). In that light, the adaptation of pathogens, especially the fast-evolving ones like HIV, to local HLA frequencies may seem plausible.

3 Ethno-Demographic Changes in the American HIV-Infected Population Associated with Changing Patterns of CTL Escape in Subtype-Specific Sites

3.1 Introduction

HIV-1M subtype diversification is currently thought to be the result of selective neutral processes such as founder effects whereby random HIV-1M lineages were exported to new geographic regions and established localised epidemics. Under the neutral model, the consensus sequence of localised epidemics is largely determined by the founder sequence. However, in the previous chapter, we have shown that CTL selection affects the subtype-specific positions and that this selection is mediated by the patient's HLA variants. HLA is the most diverse region in the human genome. This diversity is likely caused in part by pathogen-driven evolution (Doherty and Zinkernagel 1975; Prugnolle et al. 2005; Spurgin and Richardson 2010). Other human genes have also been influenced by pathogen selection: e.g., the sickle cell genes which in heterozygotes convey protection against malaria, or the CCR5 – Δ32 mutation in Europeans which may provide protection against the plague. These powerful pathogen-driven selective pressures on the HLA region have resulted in the localised adaptation of (archaic) humans to their environments. This adaptation increased chances of survival so much that invading populations of modern humans show signs of enriched introgression of HLA variants pre-adapted to prevailing diseases in that region. Increased introgression of HLA variants has been observed in populations of modern humans that left Africa and admixed with archaic humans in Eurasia, in modern Eurasian populations over 50% of HLA variants have been obtained through archaic admixture (Ségurel and Quintana-Murci 2014). More recently, preferred introgression of HLA variants obtained from indigenous Khoi-San and rainforest hunger-gatherers has been

observed in the Bantu people as they expanded into southern Africa (Patin et al. 2017). Due to the climate, no archaic hominins from Africa have been sequenced. Therefore detecting archaic admixture directly in Africans has not been possible. However, statistical analyses show some evidence that this may have occurred (Hammer et al. 2011). Echoes of these processes can be seen in modern HLA frequencies, which vary considerably between ethnicities (Gragert et al. 2013).

Interestingly, the distribution of HIV subtypes coincides with the distribution of ethnic groups. In spite of considerable sequencing efforts, no native subtype B sequences have ever been detected in Africa (excluding back migrations from the west). Evidence which seems to suggest that subtype-specific positions could be sensitive to changes in the ethnic composition of a HIV-infected population. In this context, the American HIV epidemic presents a natural experiment: HIV-1M subtype B was introduced to the USA around 1971 (Worobey et al. 2016), where it initially circulated amongst Caucasian MSM (Curran et al. 1988). Today, substantial structural inequalities in access to healthcare and ineffective or absent HIV prevention programs (Beatty, Wheeler and Gaiter 2004) have resulted in over 50% of new infections occurring in African-Americans (Hall et al. 2008). Specifically, African-American MSM and Intravenous Drug Users (IDUs) are considerably more likely to be infected with HIV than their Caucasian counterparts (Prejean et al. 2011).

Therefore, we hypothesised that the CTL selection on the subtype-specific sites should be affected by the demographic change in the American HIV-infected population. Specifically, that the proportion of subtype B consensus amino acids would lessen as the African-Americans proportion in the HIV-infected population increased and selective pressures shift to become more similar to the pressures experienced by African HIV subtypes. If true, this would suggest that HIV subtype diversification is the result of CTL selection rather than neutral evolution.

To establish whether there was a relationship between the amino acids on the subtype-specific positions and the changing ethnic demographics of the HIV-infected population in the United States we implemented the phylogenetic regression described in Chapter 1. We used the amino acid on a patient's HIV sequence's subtype-specific position as the response, and the expected prevalence of HLA B*58:01 in the HIV-infected population as the explanatory variable. HLA B*58:01 was chosen because it is common in African-Americans. If we find that the increased prevalence of HLA B*58:01 results in a lower probability that the subtype-specific sites are occupied by subtype B

consensus amino acids, this would suggest that the American HIV epidemic is changing away from the subtype B consensus as the proportion of African-Americans in the HIV-infected population in the United States increases.

3.2 Methods

All analyses are performed exclusively on four subtypes-specific sites in Gag (chapter two), using subtype B sequences from the American epidemic. We excluded position 128 because it was too invariant.

3.2.1 Building a Posterior Distribution of Phylogenies

We searched LANL for all dated subtype B sequences sampled in America with data between HXB2 nucleotide coordinates 1252 and 1569 (fragment of *gag* p24) that were annotated with a patient ID. This limitation was necessary to avoid pseudo-replication by the inclusion of multiple sequences from the same patient as unique data points. Unlike the previous chapter, we also included patients without HLA information and allowed for multiple sequences to be used per patient rather than taking the majority rule amino acid.

We ran 10 BEAST chains for 100,000,000 steps using a single representative sequence per patient and a GTR+G model of nucleotide evolution, using a log-normal clock and a logistic coalescent; this has previously been shown to be appropriate for similar data (Worobey et al. 2016). We fully incorporated ambiguous sites in this model and did not treat them as missing. An appropriate 20% burn-in was removed, and ESS$>$ 200 was confirmed for all continuous parameters in CODA. We used TreeAnnotator to calculate a Maximum Clade Credibility summary tree, using the 'Keep target heights' option to prevent negative branch lengths.

3.2.2 Bayesian Model Comparison with BEAST[1]

We also used the posterior distribution of trees to perform some exploratory analyses. As a consequence of computational and methodological limitations, it was not possible to use BEAST to

[1]We thank Professor Lemey, Rega Institute, University of Leuven for showing us how to do model comparison in BEAST.

analyse multiple sequences per patient. Instead, we limited our analysis to whichever sequence was most recently sampled from the patient, to allow the virus the most time to adapt to the patient. We treat the amino acids as binary traits, which can be either positive if the amino acid in patient's subtype-specific position is the same as the subtype B consensus, or negative if it is different. We fitted four models of Discrete Trait evolution, varying the parameter configuration. Because we believe most transitions likely would occur from the subtype B consensus amino acids to different amino acids we wanted to compare a unidirectional model (which did not allow the reverse transition), to a bidirectional model. Because we believe that these four subtype-specific positions are undergoing roughly the same process we also decided to evaluate a hierarchical prior on their transition rates.

Implementing the hierarchical prior was trivial in the unidirectional case, as there we could set the B→C transition rates to 1, and the C→B transition rates to 0, and put a hierarchical prior on the clock rates. The bidirectional models were parameterised by setting the B→C transitions to 1 and allowing the C→B transitions to vary as a relative rate; this model was made hierarchical by allowing the clocks to vary freely and independently, while hierarchical priors were placed on the relative rate for each subtype-specific site.

We used BEAST's stepping stone sampling to obtain estimates of the marginal likelihood for each of the four models. These were then converted to Bayes Factors for model comparison by taking the difference between the marginal likelihoods of two models on the log scale. (Baele et al. 2012, 2013).

3.2.3 Calculating HLA Frequencies in the American HIV-infected Population Using a Simple Population Model

To estimate the ethnic demographics of HIV-infected people in the United States we converted estimates of the ethnic demographics of HIV incidence (new infections) to prevalence (number of living hosts infected). We obtained incidence data from the literature (Hall et al. 2008; Prejean et al. 2011) and made the following assumptions:

1. At the beginning of the epidemic, when no antiretroviral drugs were available, survival time was ten years.
2. From 1996 the availability of combination therapy prolonged life to 25 years

3. In 2002 further improvement in the standard of care extended this to 32 years.

These assumptions were the same for all ethnic groups.

We then calculated the HLA frequencies by multiplying these estimated ethnic demographics of HIV prevalence by the HLA frequencies in each ethnic group (Gragert et al. 2013).

3.2.4 Regression of the HIV-infected Population HLA Frequencies on Time

To smooth the data and to incorporate uncertainty in our estimates we performed a Bayesian regression of time on the estimated HLA frequencies in the HIV-infected population. Because all calculated HLA frequencies are linear transformations of the ethnic demographic data with known constants representing the HLA frequencies per ethnicity it is appropriate to use a single HLA variant as the proxy for all other HLA variants — we used HLA B*58:01 as the prototypical 'African HLA variant'. This covariate was normalised, and to prevent it asymptoting at one, which made it very similar to the intercept covariate (which is always one) we subtracted 1 thereby making it asymptote at 0. Initial observations revealed a change between growth and decay of the projected HLA B*58:01 frequency in around 1998; to model this change accurately a Bayesian change-point analysis was needed. To prevent discontinuity at the change-point, it was required that the function be continuous with a continuous first derivative. Therefore we chose the following form[2]:

$$f(x) = \alpha_0 + \alpha_1(x-p) + \alpha_2(x-p)^2 + \alpha_3(x-p)^3 + \alpha_4(x-p)^4$$

$$g(x) = c + de^{\gamma(x-p)}$$

$$y = \begin{cases} f(x) & \text{if } x \leq p \\ g(x) & \text{otherwise.} \end{cases}$$

where:

$$\alpha_0 = c + d$$

[2]We are grateful to Dr Daniel Lunn, Dept of Statistics, University of Oxford, for providing the function preventing discontinuity at the change-point.

$$\alpha_1 = d\gamma$$

We used normal priors of mean 0 and standard deviation 10 on all parameters, except the p change-point parameter (normal with mean 1998 and standard deviation 1, because the approximate position of the change-point is between 1996 and 2000), and the likelihood function precision parameter (gamma with shape=0.001 and rate=0.001).

To stabilise the variance (and thereby ensure homoscedasticity), a regression model was fitted which was weighted by the total number of patients alive according to the simple population model; this had the further advantage of allowing some uncertainty around the earlier time-points. The calculated HLA B*58:01 frequencies were standardised, and 1980 was taken as year 0. Because the model implicitly uses binary variables to distinguish between the f and g functions we used the JAGS Gibbs sampler (Plummer 2003) to keep from having to marginalise out the binary variables, which would have been necessary had we chosen to use the Stan NUTS sampler. Because of high correlations between the parameters (due to collinearity of the $(x - p)$ polynomials), it was necessary to run JAGS for 100,000 steps. 100 samples were drawn from the posterior and used to draw red lines, to visualise the uncertainty in the model (fig. 3.3).

The median fitted values from this regression were used as the explanatory variable in the subsequent Bayesian multi-level model.

3.2.5 Bayesian Multi-Level Model Incorporating Phylogenetic Information

Here we will be regressing the year in which a sequence was sampled on that sequence's subtype-specific positions — sampling dates are by definition strongly associated with dated trees like those we generated in BEAST. Furthermore, in our preliminary analyses, we found a phylogenetic signal in both the year parameter and the subtype-specific sites. Therefore, we accounted for the phylogenetic signal in much the same way as in the previous chapter (fitting parameters for every position/patient combination). These parameters were sampled from a single multivariate normal distribution for every subtype-specific position, using the phylogenetic variance-covariance matrix to relate patients according to the evolutionary history of the virus.

Mathematically this is rendered as follows:

n is the number of sequences.

m is the number of patients.

k is the number of subtype-specific positions.

$\mathbf{Y}$ is a $n \times k$ matrix of binaries which are 1 when the amino acid in that position is identical to the subtype B consensus.

x is a vector of length n associating every sequence with the inferred HLA B*58:01 frequency in that year. This inferred frequency is taken from the fitted values in the regression of HLA frequency in time.

z is a vector of length n associating every sequence with a patient ID between 1 and m; these patient IDs also act as the indices in the $\mathbf{C}$ correlation matrix.

$\mathbf{C}$ is an $m \times m$ correlation matrix calculated using BEAST, representing the phylogenetic relatedness of the virus infecting the patients. The indices are related to the sequences by z.

α is a vector of length k containing the intercept terms for each subtype-specific position.

λ is the top level slope term, around which all βs are distributed.

β is a vector of length k, containing the slope terms for each subtype-specific position. These are normally distributed around λ with variance σ^2.

We define p_{ij} as the entries in an $n \times k$ matrix representing the probability subtype-specific position j in sequence i is the subtype B consensus amino acid.

$$y_{ij} \sim \text{Bernoulli}(p_{ij})$$

$$\text{logit}(p_{ij}) = \alpha_j + \beta_j x_i + u_{z_i j}$$

The (hyper-)priors are as follows:

$$\alpha \sim \mathcal{N}(0, 2)$$

$$\beta \sim \mathcal{N}(\lambda, \sigma^2)$$

$$\lambda \sim \mathcal{N}(0, 2)$$

$$\sigma \sim \text{Half-Normal}(0, 1)$$

$$u_{*j} \sim \mathcal{N}_m(0, \mathbf{C})$$

This creates a multi-level model in which each data point informs both the estimated probability of HIV B-likeness for a subtype-specific position in an individual patient (through the u_{ij} parameters) as well as the population-level estimates (through $\beta_j x_i$). Additionally, we used a hierarchical prior on the slopes to relate the four subtype-specific sites to one another, creating what would be in classical statistical terms a random slope model, which allows the partial pooling of information to further protect against phylogenetic covariance and giving us additional power. This construction is justified by our prior belief that these four positions are special in the same way, and by our BEAST model comparison explorations of the dataset. These phylogenetic explorations favoured a hierarchical prior on the transition rates, which is the phylogenetic analogue to the hierarchical prior on the regression coefficients $\hat{\beta}$ implemented here. More generally, a hierarchical prior on β can also be seen as a conservative choice due to the protection it confers against multiple hypothesis testing (Gelman, Hill and Yajima 2012). To prevent collinearity with the intercept, we reparameterised the fitted values from the HLA regression by subtracting 1, which made the standardised HLA frequencies regression curve asymptote around 0 rather than around 1.

As in Chapter Two, we use the correlation matrix as a variance-covariance matrix to prevent non-identifiability, implicitly setting the σ^2 in $\sigma^2\mathbf{C}$ to 1 (not shown in the multivariate normal distribution above to prevent confusion with the β variance σ^2).

This model was fitted using Stan and ran for 5,000 steps. We encountered convergence problems due to the high correlations between λ and β resulting in warnings about divergent transition due to highly curved sections of posterior space (Papaspiliopoulos and Roberts 2003; Betancourt and Girolami 2013). These warnings were resolved by reimplementing the model using a non-centred parameterisation (sometimes referred to as the 'Matt-trick'), after which the MCMC chains converged on the posterior distributions (all $\hat{R} \leq 1.01$) without further warnings.

3.2.6 Estimating the Phylogenetic Variance-Covariance Matrix C

We used the sample of posterior phylogenies obtained using BEAST to estimate the variance-covariance matrix. Briefly, the variances (diagonals) of this matrix were calculated by taking the phylogenetic distance (the sum of branch lengths) between the root node and each virus leaf node. The covariances (off-diagonals) were calculated by taking the shared phylogenetic history of two viruses, e.g., the distance between the root and the MRCA of the representative virus sequences from the two patients (Felsenstein 1985b; Freckleton, Harvey and Pagel 2002). We obtained an accurate point-estimate of the posterior variance-covariance matrix by performing this procedure for all sampled trees and taking the element-wise mean over all generated matrices.

Because this variance-covariance matrix is scaled in years, we divide it by its diagonal to obtain the correlation matrix $\mathbf{C}$. However, because we are unable to rescale it in our model due to non-identifiability, this is also our variance-covariance matrix, implicitly setting the variance to 1 as described above and in Chapters One and Two.

This method of generating the phylogenetic variance-covariance matrix is theoretically justified for a normally distributed response (Felsenstein 1985b). We use the logistic regression to apply this method to a binary response, and show this is not anti-conservative using simulations in Chapter One.

3.2.7 Model Validation using a Posterior Predictive Check

To further validate the generative properties of the model we performed a posterior predictive test. Like the half-normal plot, this is a way of comparing real data to simulated data. But, rather than performing refits as in the half-normal plot, we took samples from the posterior and used them to simulate data. We took a summary statistic of the simulated data and compared that to the same summary statistic of the real data. By plotting the summary statistic for the real data, surrounded by a credible interval over the summary statistic for data simulated from the posterior samples, we confirm that the real data could have been generated by our model. Due to the binary nature of our data, it was not wholly trivial to construct an appropriate summary statistic: we took the proportion of subtype B consensus amino acids in each position for every patient in each year and then calculated

the yearly average for each subtype-specific position separately.

3.2.8 Population-Level Estimates of Subtype B Consensus Amino Acid Frequencies

Due to the multilevel nature of our model, it is possible for us to reconstruct the population-level proportions of the subtype B consensus amino acid over time. We simply set x' as the inferred HLA frequency in every year between 1982-2009; we then used the following formula that only included the population-level parameters in the multilevel model and left out the patient-specific adjustments to the intercept:

$$\text{logit}(q_{ij}) = \alpha_j + \beta_j x'_i$$

By plotting these population-level estimates (q_{ij}) against time, we can recover the underlying pattern from the data (fig. 3.5), which in this case is the proportion of American HIV-infected patients infected with a majority subtype B consensus amino acid at position 27. Furthermore, by using parameters sampled from the posterior distribution, we can reconstruct the uncertainty around this population-level estimate. These estimates differ from the posterior predictive check since it incorporates less uncertainty (the 95% credible intervals represent the uncertainty around the population mean, which does not take the patient parameters u_{ij} into account). The purpose of these population-level estimates is to separate the signal from the noise, whereas in the posterior predictive checks we are regenerating noisy data from the model.

3.3 Results

3.3.1 BEAST Phylogeny

The MCC tree derived from the American HIV-infected population is rooted in 1952 (95% BCI: 1944-1960), which is similar to previous estimates of the HIV-1M subtype B root (Gilbert et al. 2007). In fig. 3.1 we show this phylogeny together with the American HIV-infected ethnic demographic data, which reveals that most splits predate 1982, the first year for which we have data describing the ethnic demographics of HIV-infected patients in the United States.

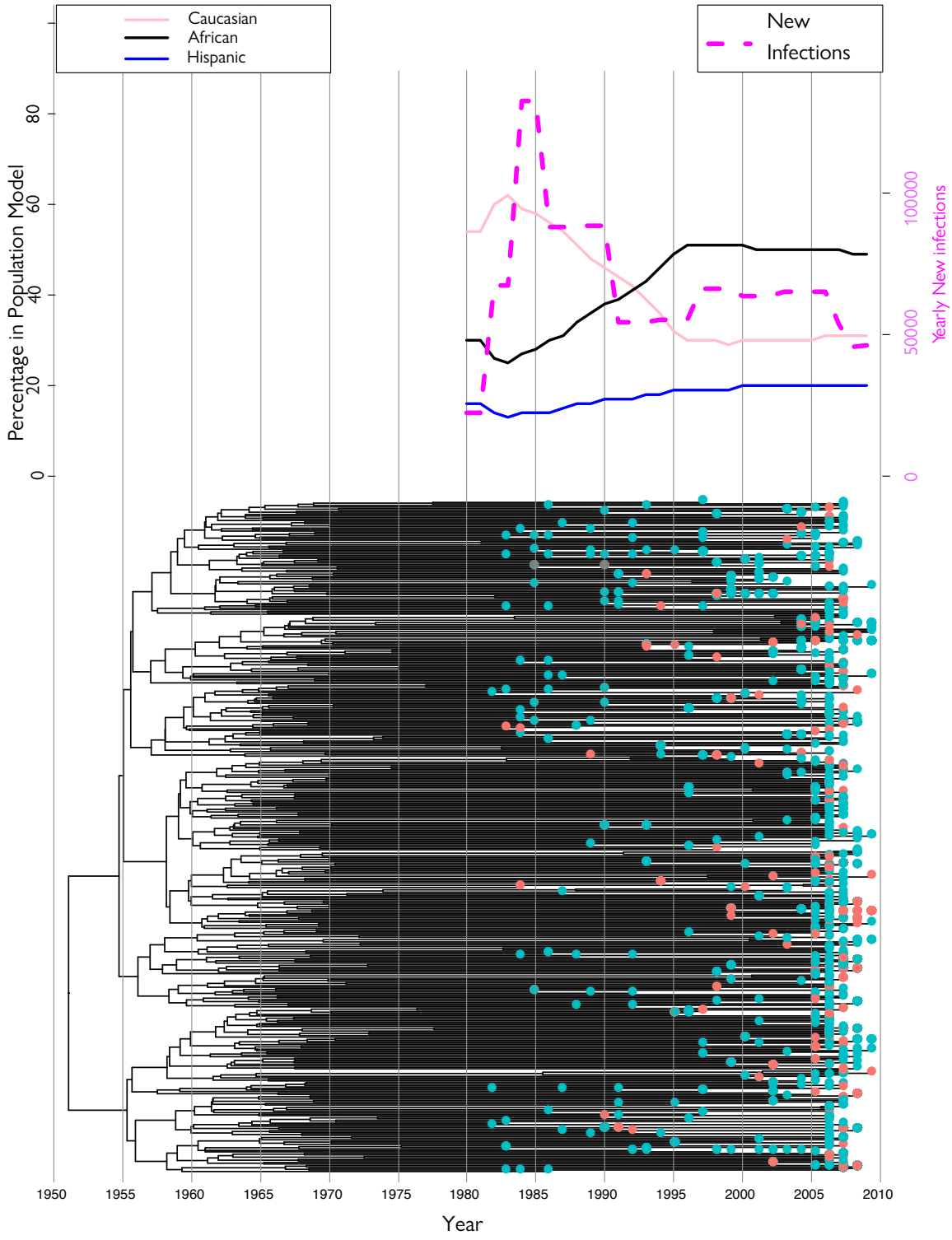


Figure 3.1: **Phylogeny of American Subtype B Epidemic Combined With Demographic Data**
The Maximum Clade Credibility tree, annotated with individual sequence data for position 27 (top, blue if a sequence is identical to the subtype B consensus amino acid; red otherwise). The phylogeny was generated using one sequence per patient, but position 27 amino acid data from multiple sequences sampled from the same patient have been superimposed on the phylogeny. These annotations show the gradual increase of non-subtype B consensus amino acids on position 27 as the African-American proportion of HIV-infected individuals in the United States increases (black line, top figure).

3.3.2 Beast Model Comparison

Model	Marginal Likelihood
Bidirectional Separate	−1277
Unidirectional Separate	−1258
Bidirectional Hierarchical	−1237
Unidirectional Hierarchical	−1026

We find that both hierarchical models give a better model fit than either separate model, showing that partial pooling allows BEAST to better estimate the transition rates of the discrete traits model. Furthermore, we find 'very strong' evidence (Kass and Raftery 1995; Baele et al. 2012) that the unidirectional model without pooling is better than the bidirectional model without pooling (log BF= 19). This was expected since there were very few opportunities for BEAST to infer transitions from non-B consensus to the subtype B consensus, so it is appropriate to set this non-B consensus→B consensus transition rate to 0. We find that the unidirectional hierarchical model is the best by a large margin (log BF= 211 with next best model, Bidirectional Hierarchical). This exploratory analysis suggests we should use hierarchical models when considering the four amino acid specific sites (position 128 is too invariant to be used in the regression), and that we should not be concerned with back-mutations to the subtype B consensus amino acid. We were able to adopt the former suggestion in our regression models, but because the regression models do not consider directionality we were unable to adopt the second recommendation in the final analysis.

We considered further analyses using BEAST, including an epoch based analysis and a latent parameter analysis, however, since this phylogeny was derived from only 444 virus samples from an HIV epidemic which has undergone exponential growth the terminal branches are very long, making it difficult to use phylogenetic tools. Indeed, exploratory analyses in which we counted Markov jumps revealed that the vast majority of jumps happened on terminal branches, which means that it is not possible to confidently assign them to a particular time on that branch. However, the fact that these long terminal branches dominate the phylogeny also means that phylogenetic covariance is unlikely to be a confounder, as the vast majority of leaf nodes in this analysis have their most recent parent

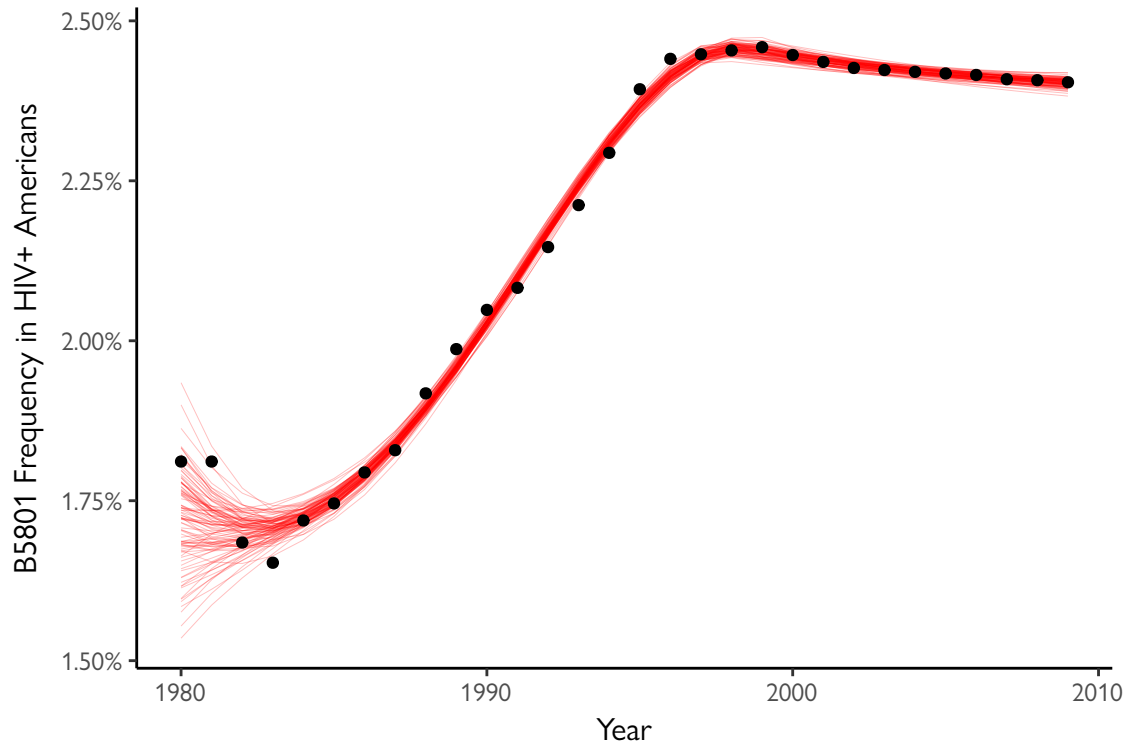


Figure 3.3: **HLA Frequencies over Time**
HLA frequencies calculated by multiplying the ethnic demographic proportions in the simple population model with the HLA Frequencies per ethnicity from Gragert et al. (2013) are shown as points. Samples drawn from the posterior distribution of the change-point model are shown as lines. Note the increased uncertainty around 1980.

node before demographic data is available.

3.3.3 Bayesian Regressions

Using our change-point model we find that the frequencies of HLA variants more common in African-Americans than Caucasians increased in the HIV-infected population until the switchover in 1998 (95% BCI: 1997-1999) when the HLA frequencies stabilise.

We find credible evidence that the increased proportion of African-American HLA variants in the HIV-infected population is associated with changes away from subtype B consensus amino acids in all subtype-specific positions studied here (fig. 3.4, all 95% BCI intervals exclude 0). The population-level probability estimates that a subtype-specific amino acid matches the subtype B consensus decays following the increased frequency of African HLA variants, plotted in fig. 3.5. Interestingly, the differences between subtype-specific positions in the population-level probability estimates appear

much stronger than the posterior probabilities over the β would suggest. This is the result of the curvature of the logistic function combined with the different intercept (α) estimates; the $\text{logit}(p) = \theta$ requires ever bigger steps in θ as p approaches 1; which is why we see more dramatic change in those sites where the population-level probability estimate was lower in 1982.

There were not enough patients with a non-consensus amino acid on position 128 to include this position in the analysis.

Because β is multiplied with the standardised HLA B*58:01 frequency, the β parameters are not immediately interpretable in a quantitative sense. One cannot easily say that a percentage point increase in HLA B*58:01 is associated with a certain odds ratio, and it is unclear whether such an interpretation would be meaningful. However, the fact that the credible intervals on these parameters exclude 0 represents our justified belief that the subtype-specific sites are influenced by changing HLA frequencies in the HIV-infected population.

3.3.4 Half-Normal Plot

The model fit only comes close to being satisfactory according to the half-normal plot of Pearson residuals (fig. 3.6), as recommended by Collett (2002) for assessing the goodness-of-fit of a Bernoulli generalised linear model. The pattern where the simulated absolute Pearson's residuals were higher than the model's actual residuals is identical to a similar analysis in chapter 2; this is the trade-off for correcting for possible phylogenetic covariance.

3.3.5 Posterior Predictive Plots

We find the posterior predictive plots are as expected, with the majority of points within the 95% posterior predictive interval. Each point represents the average proportion of subtype B consensus sequences in a particular year. This average proportion is calculated by finding the proportion of subtype B consensus sequences in every patient in every year, and then taking the average of those numbers in every year. The posterior predictive interval (grey in fig. 3.7) is then calculated by generating realisations of the Bernoulli distribution using the probability associated with every subtype-specific position in every sequence in the original data, and calculating the same yearly patientwise averages as we did for the points. We would have preferred for the posterior predictive intervals to

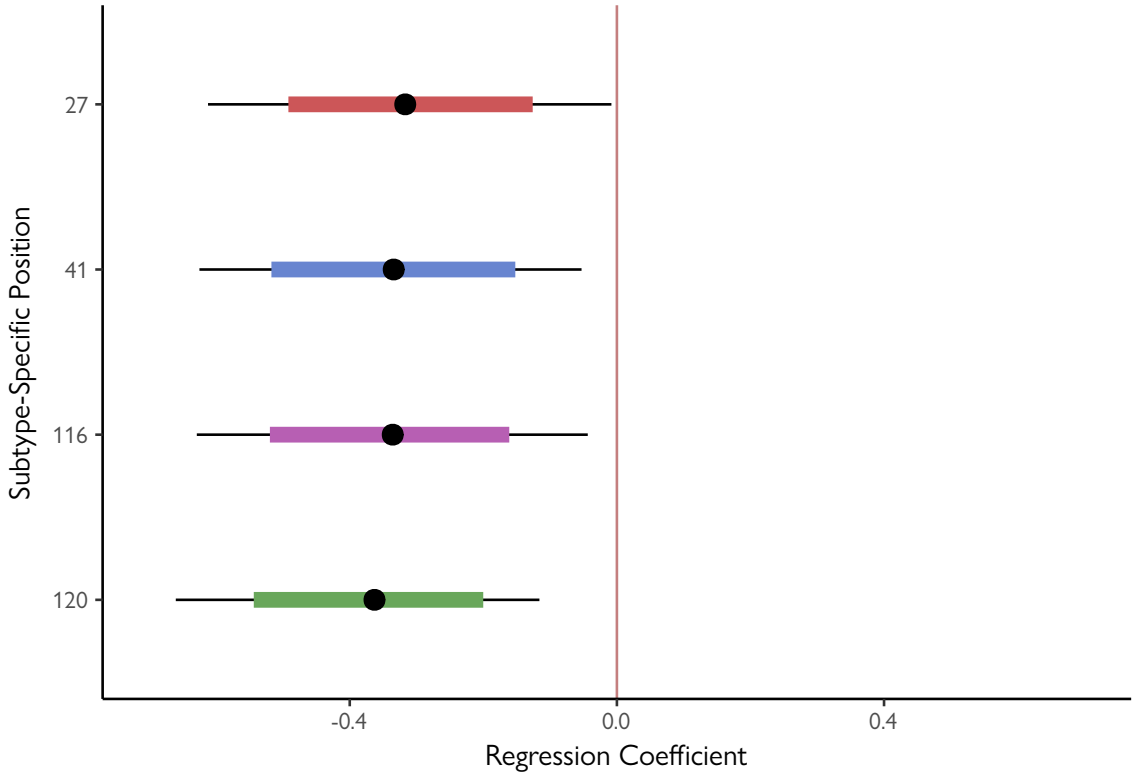


Figure 3.4: **Posterior Distributions over the Regression Coefficients of the Phylogenetic Model**
The central point represents the median, the shaded region is the 80% BCI interval, and the whiskers represent the 95% BCI interval. The 95% BCI intervals (black lines) exclude 0 for positions 27, 41, 116, and 120. This signifies a strong belief that the probability a site is occupied by an HIV subtype B consensus amino acid is inversely correlated with the inferred HLA B*58:01 frequency in the year the sequence was sampled.

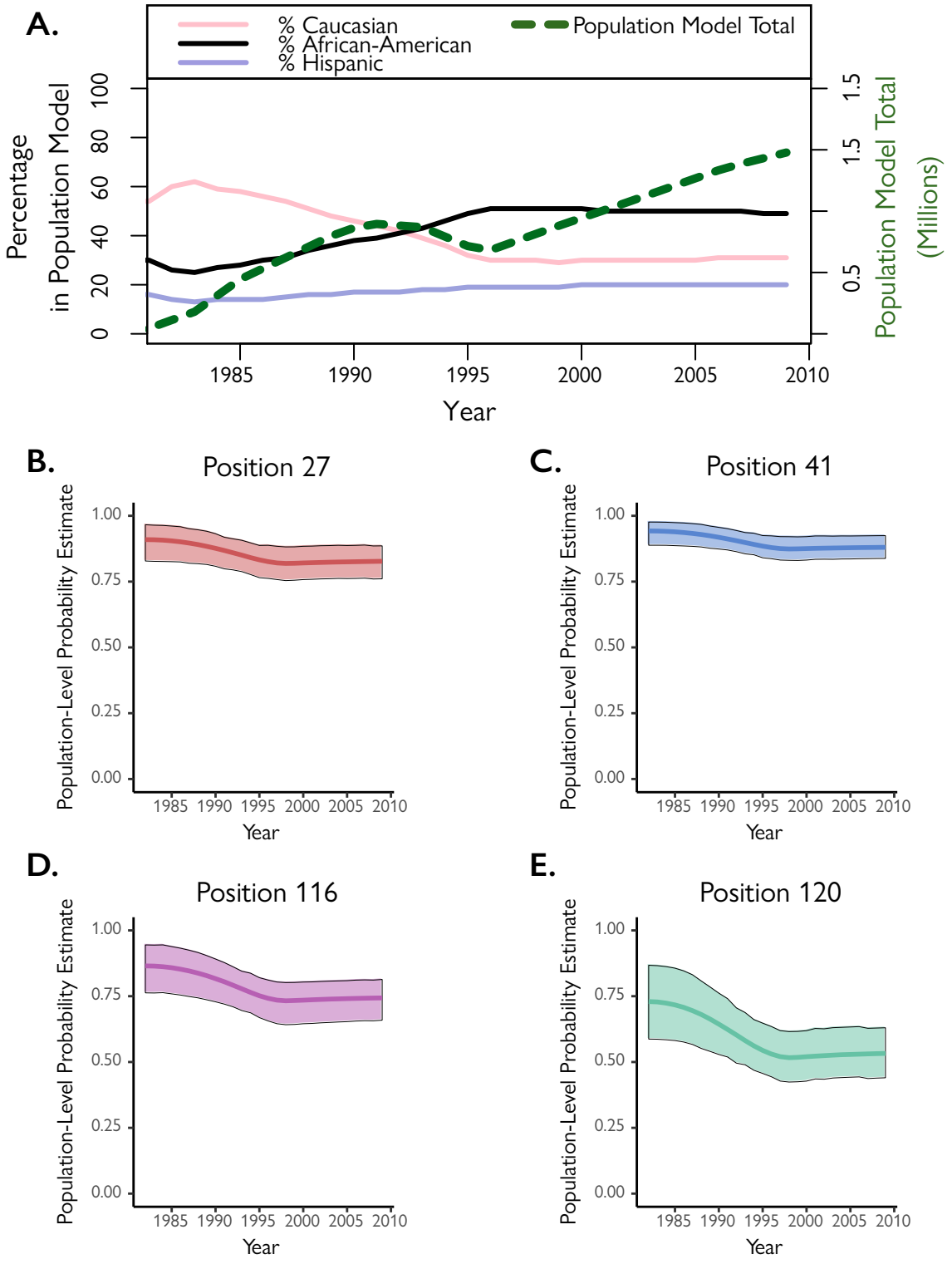


Figure 3.5: **A. The ethnic demographics of HIV-infected patients in the United States** (adapted from Hall et al. 2008; Prejean et al. 2011 using the simple population model); **B-E. Population-level estimates** of the probability a subtype-specific site is occupied by an HIV subtype B consensus amino acid as explained by the inferred HLA B*58:01 frequency plotted against time (with the 95% BCI). The coefficients estimates (fig. 3.4) are all similar, but their curves differ due to their different intercepts (α) and the logistic function.

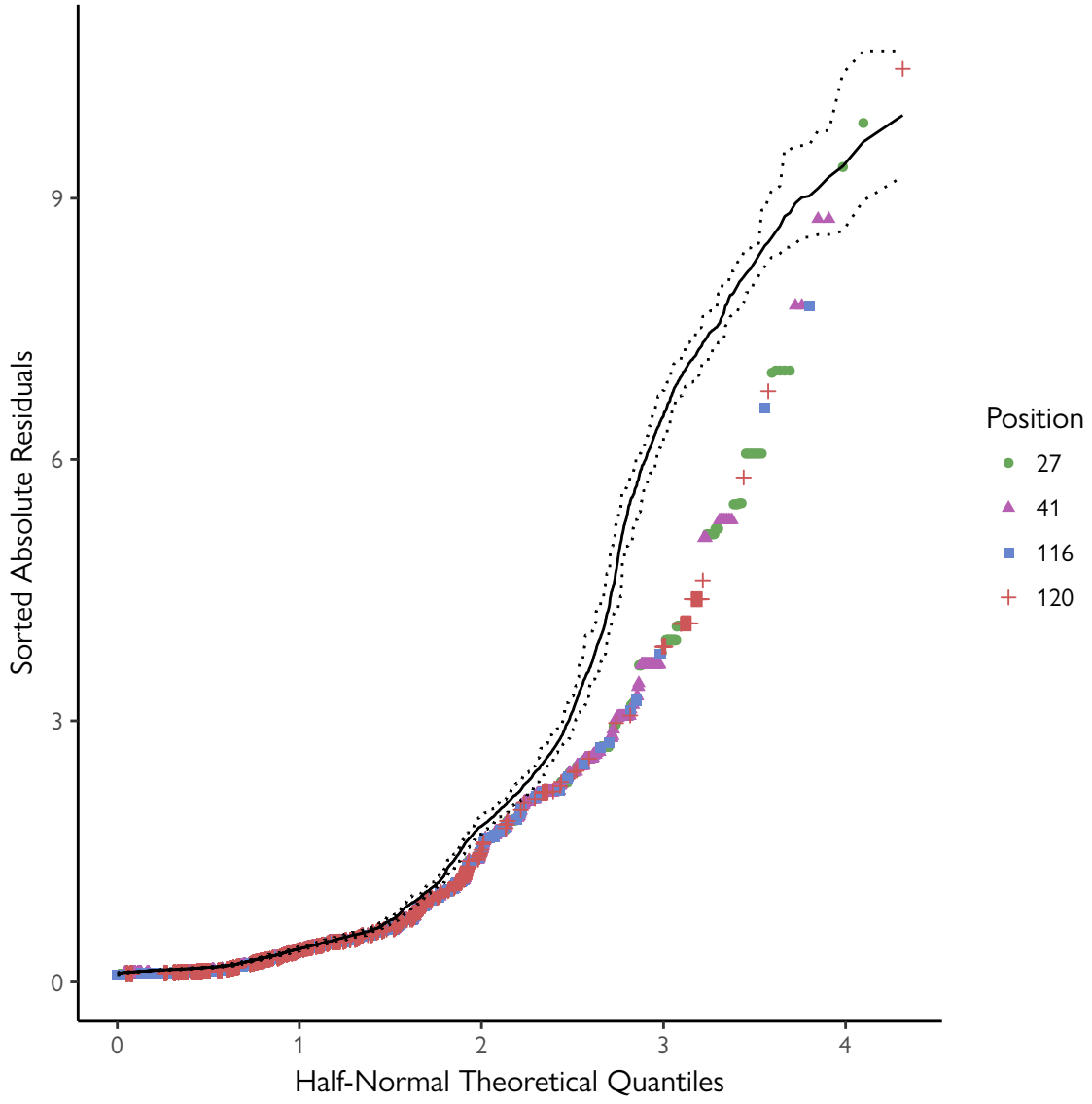


Figure 3.6: **Half-Normal Plot**
The points represent sorted absolute Pearson's residuals against half-normal quantiles following Collett (2002). The lines represent the median and 95% envelope generated by simulating data as if the model were true. The model fit is not wholly satisfactory due to the overparameterization necessary in the phylogenetic model.

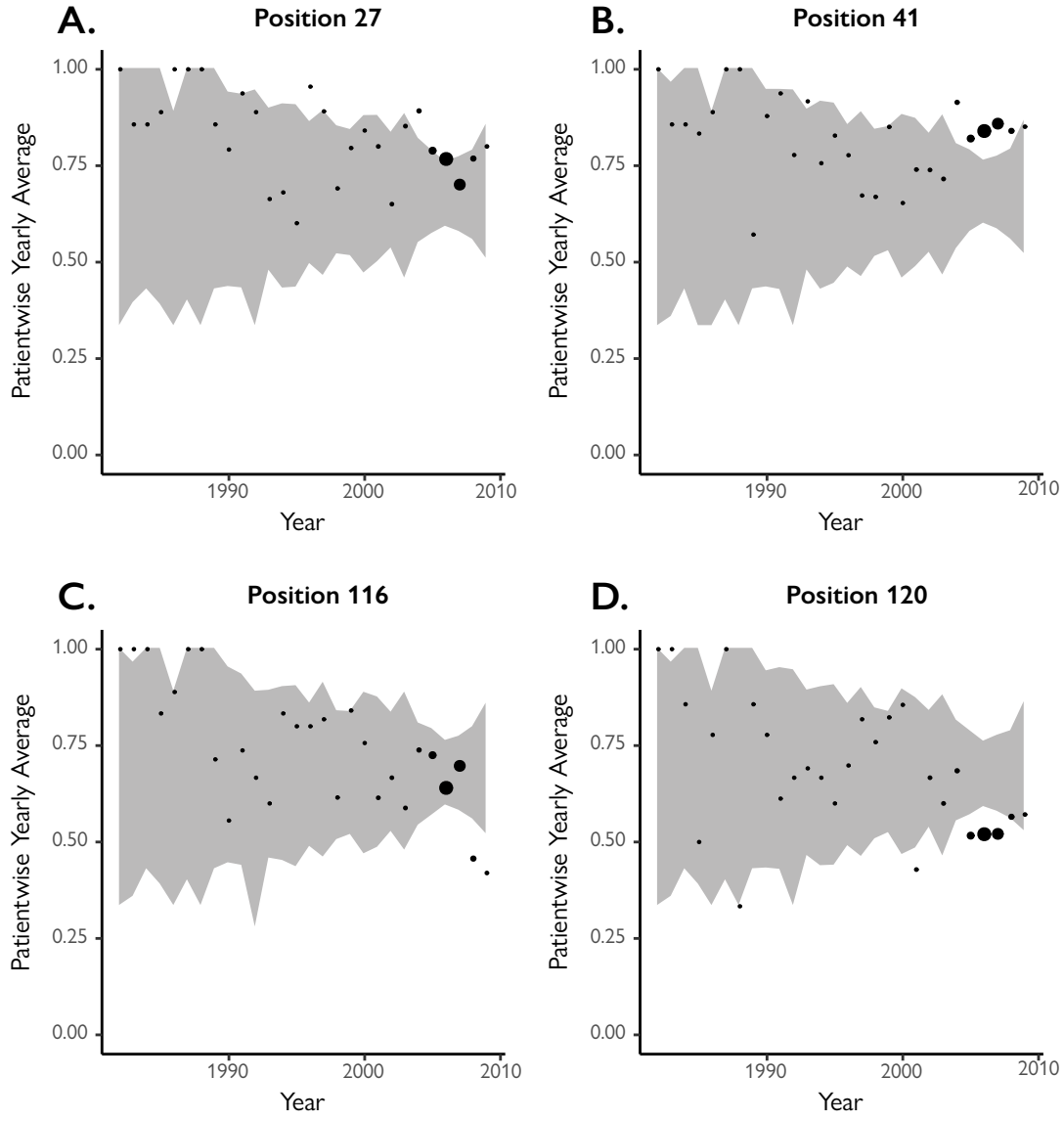


Figure 3.7: **Posterior Predictive Checks**
For both the real data (points) and simulated data (shaded 95% BCI) we calculate the proportion of sequences in each patient with the subtype B consensus on each subtype-specific site. We then take the average over all patients sampled in each year. Most points are within the 95% BCI interval, showing the model has good generative properties.

be narrower in the early years. However, this greater than expected variance was not the result of any uncertainty in the model. Instead, it was caused by taking the average of very few realisations of the Bernoulli distribution, a process that is inevitably laden with uncertainty. Indeed, the first year in which more than ten sequences were available was 1990, which is when this uncertainty starts to be reduced as more sequences become available. When comparing the shaded area in these noisy posterior predictive plots in fig. 3.7 to the much cleaner population-level probability plots in fig. 3.5 it is important to note that the posterior predictive plots show the noise, while the signal in the population-level probability plots only becomes clear when this noise is stripped away by plotting only the 'fixed-effect' component of the model.

3.4 Discussion

At present, HIV subtype diversification is thought to be the result of a selectively neutral process such as founder effects. In the previous chapter, we showed that subtype-specific sites are subject to CTL selection pressures; previous research has shown that differences in epitope processing are the molecular mechanism leading to these effects (Tenzer et al. 2009, 2014). In this chapter, we have extended that research and shown that selective pressures within individuals can work collectively on a population level to change the amino acid prevalence in certain subtype-specific sites in HIV-1 Gag. Specifically, we have shown that as HLA frequencies in the United States HIV-infected population have altered, the subtype-specific amino acids have changed with it. This is not entirely dissimilar to previous observations in other — non-subtype specific — sites, where CTL selective forces within patients made the virus adapt at the population level. HLA B*51, for example, was initially associated with delayed disease progression in Japan, but when a non-reverting I135X escape mutation increased in prevalence to 75%, the protective effect of B*51 was lost (Kawashima et al. 2009).

In our study, we have demonstrated that a relatively small change in ethnic demographics can still lead to an observable change in amino acid frequency. This allows us to predict that this process will continue in the future, as African Americans continue to suffer a disproportionate number of new HIV-1M infections we expect the frequencies of non-subtype B consensus amino acids to continue to increase. Because HLA-specific polymorphisms have been shown to be associated with viral load

(Moore et al. 2002) population-level adaptation to HLA variants common in African-Americans could have clinical consequences. Given the already steep disparities in health outcomes, the continued adaptation to African-American HLA variants is tragic, as the viral adaptation to CTL selective pressures associated with African HLA variants might make the virus more adapted to one of the most vulnerable ethnic groups in the United States.

In addition to the regression model with a phylogenetic multivariate normal parameter implemented here, we also considered a regression within a BEAST phylogenetic model, where the Markov rate changes followed the HLA frequency in the population. Essentially this would have allowed us to directly predict the rate of amino acid transition events using the ethnic demographics of the American HIV-infected population. However, this was not feasible due to the abundance of long terminal branches, which made it impossible to accurately date transition events.

The great diversity of HLA variants and enriched introgression of HLA variants by invaders from indigenous populations (Abi-Rached et al. 2011; Ségurel and Quintana-Murci 2014; Patin et al. 2017) suggest the geographically granular adaptation of humans to endemic disease. However, similar geographically granularity of pathogen adaptations to man has not previously been observed. Here we suggest that this does occur as HIV-1M has adapted to different human populations. Our model can also explain why a relatively high diversity of HIV-1M subtypes is found in Africa, while no significant phylogenetic differences have arisen in the HIV-1B strain, in spite of it circulating in Caucasians for approximately 30 years. To compare, HIV-1M has been circulating in Humans for 100 years and has diversified into many subtypes (Faria et al. 2014).

4 HIV-infected Sex Workers with Beneficial HLA-Variants are Potential Hubs for Selection of HIV-1 Recombinants that may Affect Disease Progression[1]

4.1 Introduction

In Chapters Two and Three we have shown that the selective pressures exerted by the CTL response can drive the evolution of the subtype-specific sites in HIV-1M p24 Gag. Because these intra-epitope escape mutations are spread throughout Gag (fig. 4.1), we hypothesised it might be beneficial for the virus to recombine and join different segments containing these and other mutations which convey fitness benefits. In HIV-1M, recombination has been implicated in the creation of the strain which would later be transmitted to humans. This virus, SIVcpz, was the product of a recombination event between ancestors of SIV lineages which today infect red-capped mangabeys and Cercopithecus monkeys (Bailes et al. 2003; Keele et al. 2006; Sharp and Hahn 2011). Recombination has also been shown to be a contributing factor to the evolution of multidrug-resistance (Kellam and Larder 1995; Kemal et al. 2012) and to allow escape from the CTL response in primary infection (Ritchie et al. 2014).

Recombination most often occurs between closely related sequences, but can occur between distantly related sequences when a cell has been coinfected by two or more distinct viral lineages of the same (Moutouh, Corbeil and Richman 1996) or different subtypes (Chin et al. 2005). Recombination between closely related sequences is difficult to detect and is often ignored in phylogenetic analysis; mostly because of methodological and conceptual constraints, despite recombination being an im-

[1] This work has previously appeared as Chang & Kist et al. HIV-infected Sex Workers with Beneficial HLA-variants are Potential Hubs for Selection of HIV-1 Recombinants that may Affect Disease Progression. Scientific Reports (2015). The work in this chapter was performed under the joint supervision of Dr Iversen and Dr Katzourakis.

portant mechanism of viral evolution. However, because the subtypes of HIV-1M have been on distinct evolutionary trajectories for the last ~50 years, robust methods for detecting inter-subtype recombinants exist (Salminen et al. 1995; Schultz et al. 2006). In HIV-1M we distinguish between HIV-1 homogenous subtypes/subsubtypes (HHS), Circulating Recombinant Forms (CRFs) and Unique Recombinant Forms (URFs). HHSs are HIV sequences consisting of only one subtype. CRFs are existing inter-subtype recombinant HIV strains that circulate in a patient population and are defined as recombinant forms found in at least three epidemiologically unlinked patients (Robertson et al. 2000; Delgado et al. 2008; Wang et al. 2008; Leye et al. 2013). URFs are the result of an inter-subtype recombination that (to the best of our knowledge) occurred in a patient.

Here we studied a cohort of female Commercial Sex Workers (CSWs) from the Pumwani slum in Nairobi, Kenya, for whom we have clinical progression data, HLA variants (Simonsen et al. 1990), and HIV *gag* sequences (Chih-Hao Chang, Iversen Group).

We hypothesise that patients capable of mounting strong CTL responses against HIV Gag epitopes were more likely to be infected with a URF, and that the beneficial effect of HLA types might be mitigated by a successful immune-escaped URF outgrowth. The Pumwani cohort is well-suited to address these questions because of the high subtype diversity in Kenya, where subtypes A, C, and D co-circulate (Burns et al. 2002). The Pumwani women are highly exposed to this HIV diversity as they have on average ~4 contacts per day (Peterson 2014).

If our hypothesis is correct, and an effective CTL response drives recombination, we should expect that patients with beneficial HLA variants are more likely to be infected with URFs than with CRFs/HHS. Because viruses which have recombined within a patient are likely to have alleviated some immune pressures, we further hypothesised that patients infected with URFs should experience a reduced protective effect from beneficial HLA variants.

To determine whether patients with beneficial HLA variants were more likely to be infected with a URF, it was necessary to distinguish between URFs and local CRFs, for which no tools existed. We used existing methods to separate recombinant sequences into their homogenous subtype parts (Schultz et al. 2009) and developed a novel algorithm to distinguish between likely URFs and CRFs. We then use this algorithm to separate the Pumwani patients into two categories: those infected with

a HHS/CRF and those infected with a URF. These categories were then used in regressions to show associations with beneficial/neutral/detrimental HLA variants and disease progression.

4.2 Methods

4.2.1 Study Participants

The study participants were members of the Pumwani cohort, a group of CSWs operating out of the Pumwani slum in Nairobi, Kenya (Peterson et al. 2013). Upon enrolment, all patients were subjected to clinical examination, were interviewed and completed a questionnaire. Blood was drawn for HIV-1 testing, lymphocyte counts, and HLA typing (Simonsen et al. 1990). The women gave written and oral informed consent. Patients were followed up biannually for blood samples and lymphocyte counts. All patients were enrolled between 1985 and 1995, and peripheral blood mononuclear cell (PBMC) samples were taken between 1997 and 2004, before antiretroviral drugs were available to these patients.

4.2.2 Generation of the HIV-1 Sequences[2]

HIV-1 p17-p24 *gag* sequences were obtained from 113 HIV infected patients in this cohort by Dr Chang, a previous student of Dr Iversen. He amplified approximately 1200 bases using a hot-start nested polymerase chain reaction (PCR). Because of limited sample availability, he initially performed bulk DNA amplification to optimise primers and identify patients with mixed sequence reads. Subsequently, he used the PBMC DNA samples from patients who appeared to be infected with multiple subtypes to perform nested single-genome PCR amplification of 5-10 genomes per patient, subject to DNA availability. This allowed us to be sure that the observed inter-subtype recombinant sequences were true recombinants rather than the result of PCR hybridisation.

The Pumwani cohort studies were conducted according to the principles expressed in the Declaration of Helsinki, the National AIDS Committee. The National Ethics and Scientific Review Committee of Kenyatta National and the University of Manitoba Use of Human Subjects in Research review committee have approved studies using this cohort.

[2]We are grateful to Dr Chang (Iversen Group) for providing the sequences.

4.2.3 Assignment of HLA Variants as "Beneficial", "Detrimental" and "Neutral"[3]

Dr Iversen assigned HLA-variants as "beneficial," "detrimental," and "neutral" with regard to disease progression based on previously published reports (Gao et al. 2005; Matthews et al. 2011; Goulder and Walker 2012; Kloverpris et al. 2012; Peterson et al. 2013). Beneficial HLA-variants included HLA-A*74:01, HLA-B*13:02, HLA-B*14/-C*8, HLA-B*57:01, HLA-B*57:02, HLA-B*57:03, HLA-B*58:01 and HLA-B*81:01. Detrimental HLA-variants comprised HLA-B*07, HLA-B*15:10, HLA-B*18:01, HLA-B*35, HLA-B*53 and HLA-B*58:02. All other HLA-variants were labelled as "neutral". One patient (ML1317) carried two beneficial (HLA-A*74:01 and HLA-B*58:01) and one detrimental (HLA-B*35) HLA-variant, and was classified as having an overall beneficial HLA profile (tbl. 4.4), because we estimated that the beneficial effect was stronger based on the HLA-distribution results in the group with HHS/CRF, the co-operative additive effect of beneficial HLA-variants (Leslie et al. 2010; Salgado et al. 2011; Matthews et al. 2012) and previous reports (Gao et al. 2005; Matthews et al. 2011; Goulder and Walker 2012; Kloverpris et al. 2012; Peterson et al. 2013; Martin and Carrington 2013). Two patients carried one beneficial, and one detrimental HLA-variant (ML293: HLA-B*81:01/HLA-B*58:02, ML2000: HLA-B*74:01/HLA-B*58:02) and both carried HLA-C*06:02. Because HLA-B*58:02 was not associated with detrimental effects in this cohort and HLA-C*06:02 was significantly associated with decreased risk of seroconversion, although not with slower disease progression (Peterson et al. 2013), we estimated that the overall HLA profiles were beneficial (tbl. 4.4).

The scoring of patients' HLA variants was done by Dr Iversen so that I could be blinded to the patients' HLA status while devising the method for the URF/CRF calling of recombinant HIV sequences.

4.2.4 HIV-1 Subtyping and Analyses of Recombinant HIV-1.

All HIV-1 sequences were subtyped, and each section of the recombinant sequences was annotated (fig. 4.2) with breakpoints and its subtype using jpHMM (Schultz et al. 2006). The recombinant subtype compositions were also examined using REGA v3.0 (RegaSubtyping/stanford-hiv/typingtool; as recommended in Pineda-Peña et al. (2013)), which gave similar results. Because REGA does not

[3]We thank Dr Iversen for scoring the patients' HLA variants.

explicitly infer breakpoints, we used the jpHMM results in subsequent analyses. Two patients carried recombinants with unique breakpoints (1700 and 2033) and were classified as URFs. In the case of patient ML1700, the breakpoint was outside the p17 and p24 border and combined a homogeneous subtype A sequence with a CRF_D HIV. Similar CRF_D HIV fragments were found in other recombinants (patients ML1317, ML1852 and ML199); however, in these cases, the recombinant breakpoints were found in the p17-p24 border region. In the case of patient 2033, we observed two distinct recombinant HIV variants composed of HIV subtype A and HIV subtype C sequence segments. These variants had multiple additional breakpoints outside the p17-p24 border. These patterns were unique to this patient, and the presence of two related, but distinct, recombinants forms of HIV subtype A and HIV subtype C suggested that the recombination events had occurred in this patient. In the other patients, the breakpoints between subtypes in the conserved p17-p24 border region were too close to classify them as URFs based on the position of the recombination breakpoint (shaded area in fig. 4.2a), such 'breakpoint hotspots' are common in HIV (Smyth et al. 2014).

To identify women in whom selection had favoured the outgrowth of URFs, we developed a novel phylogenetic algorithm that did not rely on breakpoint patterns to distinguish between URFs and any CRFs that had infected the women by chance. We only used single recombinants (i.e., sequences with one breakpoint) because we needed the longest alignments possible to ensure adequate power in our phylogenetic analyses. Briefly, we constructed separate alignments for each subtype; these alignments consisted of 1) the subtype sections of the recombinant sequences, 2) matching sections of the HHS sequences from our cohort, and, 3) matching sections of the most similar sequences in the HIV database (www.hiv.lanl.gov), as determined by BLASTn (Altschul, Gish and Miller 1990) using 1) and 2) as query sequences (fig. 4.2b, tbl. 4.5, tbl. 4.6). Because HIV-1 recombinants with HIV-C had two distinct breakpoints (see ML1765 and ML1812), we constructed two alignments for this subtype. It was necessary to split the recombinant sequences, and analyse each subtype-section in the context of sequences from the same subtype because each subtype-section in a given recombinant has a separate evolutionary history, which is highly likely to confound the analysis if the recombinant sequence is included as a whole.

We used the Bayesian phylogenetic tree-builder MrBayes (Huelsenbeck et al. 2001) to build separate

posterior distributions of phylogenies from each of the subtype alignments (HIV-A1 and HIV-C: model K80+G; HIV-D: model GTR + G). We ran four replicates per analysis for 50M steps each and confirmed convergence in TRACER. Because the sequences were short and relatively conserved, we were unable to construct maximum clade credibility trees from the posterior trees with sufficient support values on the internal nodes to make URF/CRF calls. Instead, we used a novel algorithm to infer the posterior probability that a patient was infected with a URF directly from the trees sampled from the posterior tree distribution.

We assumed that an inter-subtype recombinant HIV would not recombine again to become an HHS. From this assumption follows that the parent node of any HHS sequence must also be an HHS sequence. Therefore, any tree in which a recombinant sequence has an HHS sister sequence supports the hypothesis that the recombinant sequence is a URF because it is derived from an HHS parent node (fig. 4.2c), which would occur if a patient was infected with a homogenous subtype that then recombined within the host to form the URF. Conversely, a recombinant sequence that has another patient's recombinant sequence as its sister sequence is evidence that both those recombinant sequences are CRFs (fig. 4.2d).

Thus, for each tree sampled from the posterior tree distribution derived from a subtype-section alignment (consisting of subtype-sections of recombinant sequences, matching sections from homogenous sequences, and matching sequences from the LANL database) we examined whether the tree supported the URF hypothesis or the CRF hypothesis for each patient. For every patient infected with a recombinant, we identified the common ancestor of the patient's subtype-section sequences in the tree. If the sequences from that patient were monophyletic (i.e., the common ancestor of the patient's sequences had descendants from that patient exclusively), then we selected the sequences' common ancestor's parent node to assess the recombinant form. If the patient's sequences were paraphyletic (i.e., the common ancestor of the patient's sequences had sequences from other patients amongst its descendants), then we selected the common ancestor of the patient's subtype-section sequences. If the selected node had an HHS from another patient amongst its descendants, then we interpreted the tree as evidence that an HHS had infected the patient and that this HHS had subsequently recombined in the patient to form a URF (fig. 4.2c). Conversely, if the selected node only had recombinant sequences amongst its descendants, then we interpreted the tree as evidence

that the patient had been infected by a CRF (fig. 4.2d).

To ensure that rooting effects did not confound the analysis, we rooted each tree sampled from the posterior distribution on a random sequence from a random HHS-infected patient and excluded patients for that tree whose sequences were too close to the root, i.e., when the selected node had more than four different patient/virus combinations as its descendants. For each patient, we calculated the probability that their sequences in this subtype-section alignment were part of a URF (URF-probability) by dividing the number of trees supporting the hypothesis that the patient was infected with a URF (fig. 4.2c) by the number of trees that were not excluded due to rooting effects. Thus, we were able to interrogate the data directly and to integrate over the phylogenetic uncertainty that prevented us from building interpretable trees.

We implemented this algorithm in a Python script using the ete2 tree manipulation module (Huerta-Cepas, Dopazo and Gabaldón 2010), discarding the first 25% of trees as a burn-in. All analyses were performed separately for each subtype-specific alignment, with the results for each patient's two subtype-sections only brought together immediately before doing the downstream statistical analysis. URFs can be composed of a fragment derived from a CRF virus joined to a fragment derived from an HHS virus, in which case only the subtype-specific analysis containing the HHS-derived fragment would report the appropriate URF probability. Therefore we only used the higher of the two URF probabilities for every patient when combining the results from the two subtype section analyses performed for that patient. Because no previous studies of this sort exist, it was not possible to use an informative prior. The results were expressed as the Bayesian posterior probability that the recombinant HIV-1M sequences in a given patient was a URF (URF maximum probability (MP) in tbl. 4.4). If Fisher's exact test required bifurcating the patients into two classes 50% was used as the threshold URF MP for classification into patients infected with HHSs or CRFs (tbl. 4.4).

4.2.5 Subsequent Statistical Analyses[4]

Equality of proportion was tested with Fisher's exact test, but this approach required neglecting the phylogenetic uncertainty incorporated in the posterior probabilities and bifurcating the patients into two classes. To incorporate the phylogenetic uncertainty we used binary GLMs with disease pro-

[4]We thank Dr Daniel Lunn for his assistance with the downstream statistical analysis.

gression as the response variable, and the posterior probability log(odds) as the explanatory variable. Taking the log(odds) of the posterior probabilities was necessary to ensure homoscedasticity. As described above, these explanatory variables were calculated by taking the highest Bayesian posterior probabilities of the subtype-sections of a given recombinant HIV-1 (the "URF MP", tbl. 4.4). The two recombinant sequences with unique breakpoint patterns were given a URF probability of 100%. Binary GLMs were also used for comparing the frequency of beneficial, neutral and detrimental HLA variants in DP1 and DP2.

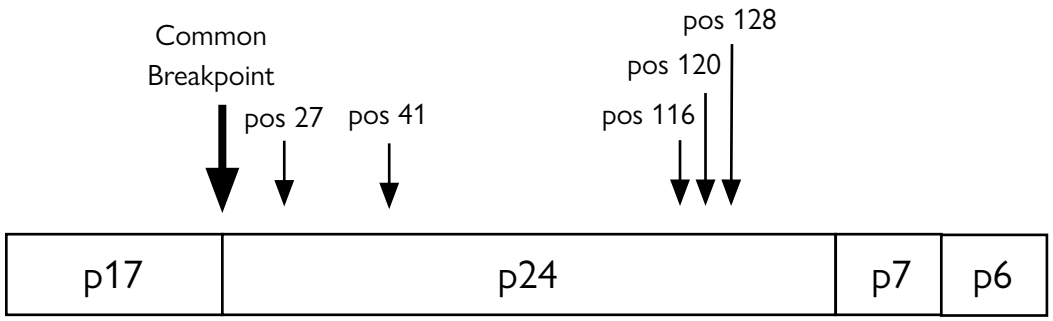


Figure 4.1: **HIV Gag** annotated with constituent proteins, subtype-specific positions considered in chapters 2 and 3 and the most common recombinant breakpoint between p17 and p24. (fig. 4.2)

4.3 Results

4.3.1 HIV Subtype Distribution and Recombination Analyses

Several HIV-1 subtypes were prevalent in our cohort; HIV-A1 61% (69/113), HIV-C 6% (7/113), and HIV-D 18% (20/113) (tbl. 4.5), consistent with other studies in Kenya (Neilson et al. 1999; Dowling et al. 2002). In addition, 17 women carried inter-subtype recombinant HIV-1 (15%). These recombinant viruses were mixtures of HIV-A1 and HIV-C (n= 5) and HIV-A1 and HIV-D (n= 12); three patients carried additional recombinant forms (HIV-A2/HIV-D (ML900), HIV-A1/HIV-A2/HIV-C (ML1765), and, HIV-C/HIV-F (ML1812)) and two carried recombinants with unique breakpoint patterns (ML1700 and ML2033, fig. 4.2).

To identify women in whom selection had favoured the outgrowth of URFs, we developed an algorithm to distinguish between URFs and any CRFs that had infected the women by chance. This analysis is based on an examination of subtype-section-specific phylogenetic trees made from alignments of sequences from this cohort and the HIV database (as illustrated in fig. 4.2B-D and described in Methods). Briefly, we examined each tree sampled from the posterior distribution for every patient with recombinant HIV-1. For our assessment of whether a given subtype-section derived from an HHS or a CRF, we assumed that a URF would not recombine again to become an HHS. Therefore, if a patient's subtype-section descended directly from a phylogenetic tree node that also had HHS descendants, we estimated that the patient was infected with a URF (fig. 4.2C) because the parent node of an HHS sequence must have been an HHS as well. Conversely, if a

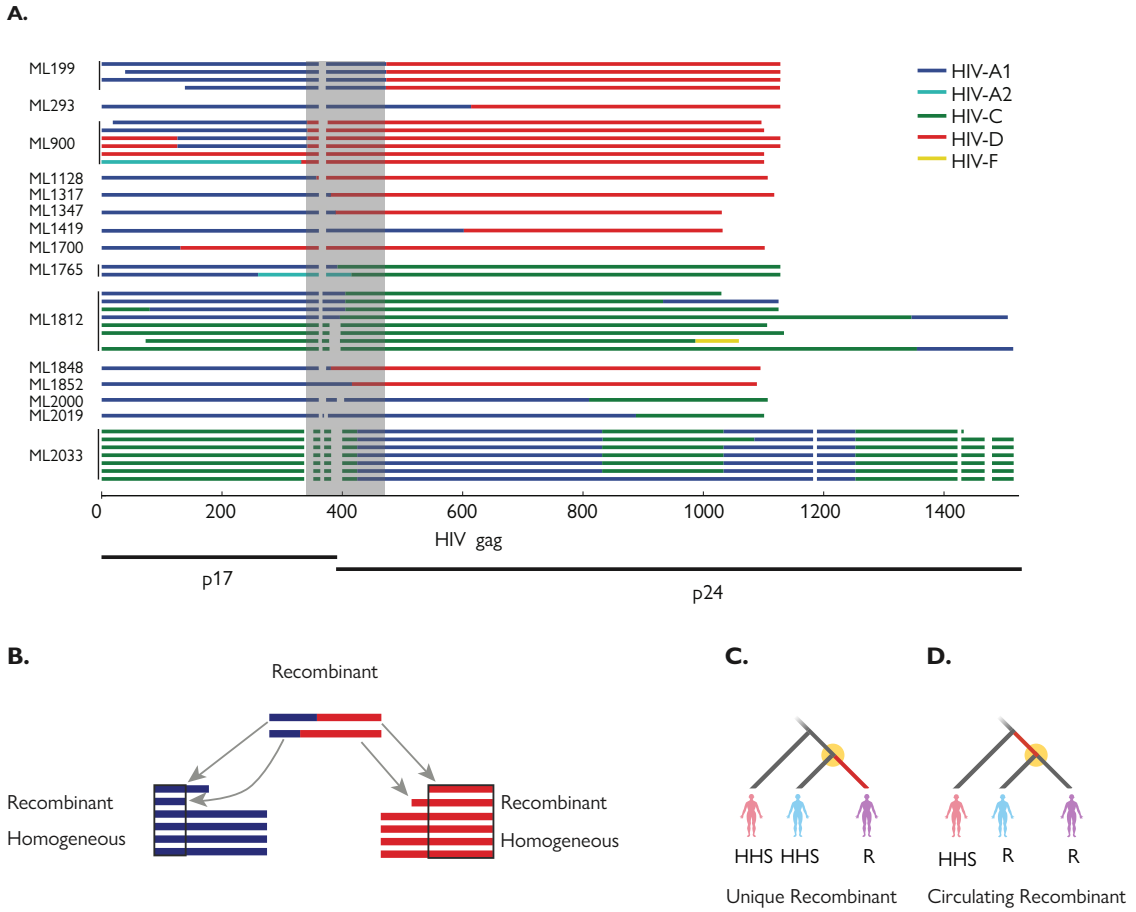


Figure 4.2: **Outline of all unique recombinant HIV-1 gag forms**. a. Ten women carried one predominant URF, whereas five women carried several distinct URFs (ML199, ML900, ML1765, ML1812, ML2033); in three cases these URFs incorporated sections from additional HIV-1 subtypes (ML900, ML1765, ML1812). Blue=HIV subtype A1 (HIV-A); cyan=HIV subtype A2 (HIV-A2); green=HIV subtype C (HIV-C); red=HIV subtype D (HIV-D); yellow=HIV subtype F (HIV-F); and white=gap. A unique breakpoint pattern was found in patients ML1700 and ML2033. b. Outline of the construction of sequence alignments. Subtype-specific alignments (labelled X and Y) were created by dividing the recombinant sequences into subtype-specific sections and using the longest shared subtype-section and matching sections of the HHS sequences from the rest of the women in the cohort and sequences from the HIV database. c,d. Outline of the Bayesian phylogenetic analysis based on the assumption that a URF will not recombine again to become an HHS. Any recombinant patient sequence on the tree whose parent node (yellow circle) has an HHS descendant likely is derived from an HHS, which means that the recombinant sequence is likely to be a URF (c). If the parent node only has recombinant descendants, then the patient is likely to be infected with a CRF (d). The branch on which the recombination event is suspected to have occurred is shown in red.

subtype-section's parent node had only other recombinant sequences amongst its descendants, we estimated that the patient was infected with a CRF (fig. 4.2D). By performing this analysis on each tree sampled from the posterior distribution of trees, and bringing together the analyses for both segments, we were able to calculate the Bayesian posterior probability that a patient was infected with a URF.

Using the results of the phylogenetic tree and recombinant breakpoint analyses, we estimated that the patient most likely carried a URF if 1) both recombinant sequence-sections derived from HHSs, 2) one derived from an HHS and one from a CRF, or, 3) the recombinant sequence had a unique breakpoint pattern. Based on this, 15/17 women (88%) likely carried a URF, and 2/17 carried the same CRF (12%)(tbl. 4.4). This CRF was related to a previously reported recombinant HIV-1 in the HIV database that derived from Pumwani, and the CRF may thus have been generated and spread locally.

We next examined the composition of each URF and found that 6/15 (40%) of the URFs contained sections of two HHSs, and 9/15 (60%) contained one section of a local CRF and one section of an HHS (tbl. 4.4). URFs consisting of segments derived from CRF viruses have previously been reported (Visawapoka et al. 2006).

The overrepresentation of CRF-sections in HIV recombinants in the URF group suggests that local CRFs may have a distinct survival advantage in women with beneficial HLA variants, possibly because the CRFs arise from URFs generated in this patient group.

This overrepresentation of CRF sections in the URF group is highly unlikely to occur by chance given the frequency of CRFs in the cohort (2/98 (2%) versus 9/15 (60%), Fisher's exact test, $p < 0.0001$, odds ratio 9.8). Thus, our results indicate that segments derived from local CRFs may convey a selective advantage when incorporated in a URF.

4.3.2 Overrepresentation of Beneficial HLA Variants in URF infected women

We hypothesised that efficient CTL responses would lead to increased recombination, as the increased selective pressures associated with a strong CTL response would favour larger steps in fitness space (Worobey and Holmes 1999; Ritchie et al. 2014). Unfortunately there were insufficient

PBMC samples to obtain direct estimates of CTL response strength, so instead, we used HLA variants associated with delayed disease progression as a proxy. There has been ample research showing a link between HLA variants, CTL responses, and delayed disease progression, and HLA variants have previously been used as proxies for CTL selection (Gao et al. 2005). Specifically, patients carrying the HLA variants HLA-A*74:01, HLA-B*13:02, HLA-B*14/-Cw8, HLA-B*57:01, HLA-B*57:02, HLA-B*57:03, HLA-B*58:01 and HLA-B*81:01 (Matthews et al. 2011; Goulder and Walker 2012; Kloverpris et al. 2012; Peterson et al. 2013) were designated as having beneficial HLA variants. We then tested whether patients carrying beneficial HLA variants were more common in the group found to be infected with URFs versus the group with HHS/CRFs. We found women with beneficial HLA-variants (12/52) were more likely to carry URFs compared to those without beneficial HLA variants (3/61) (Fisher's exact test, $p < 0.0055$; odds ratio 5.7).

Because Fischer's exact test requires a definitive classification of all patients into the CRF or URF categories (based on whether or not the patient's URF MP $>$ 50%), it does not incorporate the phylogenetic uncertainty captured by our estimates of the Bayesian posterior probabilities. We incorporated this uncertainty into the analysis by expressing the URF probability as log(odds) and used these as explanatory variables in fitting binary Generalised Linear Models (GLMs) with patient HLA group (beneficial vs detrimental/neutral) as the response variable. This more appropriate approach incorporating phylogenetic uncertainty confirmed that women with beneficial HLA variants were more likely to carry URFs than those without (binary GLM, log(odds) unit change odds ratio= 1.25, p= 0.0107). However, the disadvantage of this method is that no readily interpretable odds ratios can be calculated. Any effort in that direction would require quoting odds ratios in terms of a unit-change in the log-odds of a single patient, which is not as meaningful as the odds ratio between two groups. A log(odds) unit change odds ratio corresponds to a change in URF posterior probability of a single patient by 75%; odds-ratios corresponding to a posterior probability change of 90% can be approximated by squaring the quoted odds ratio, and cubing it would correspond to a patient's change in posterior probability of 95%.

These results suggest that the efficient CTL responses associated with beneficial HLA variants select for the outgrowth of HIV-1 recombinants

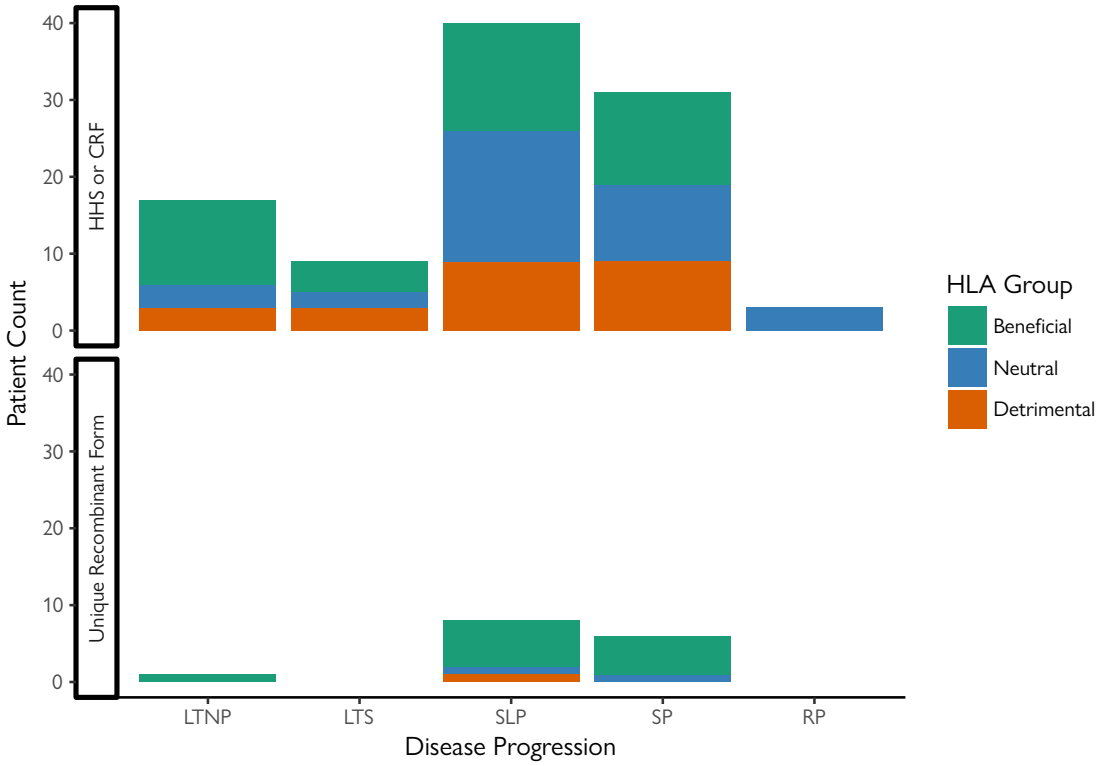


Figure 4.3: **Disease Progression, HLA groups, and Unique Recombinant Form Status**. There are fewer patients infected with URFs than with CRFs/HHSs. Patients infected with a URF were over five times as likely to carry beneficial HLA variants. In HHS/CRF patients, beneficial HLA variants were associated with delayed disease progression — no hint of this pattern could be recovered from the URF patients, which suggests that URFs expand in patients with beneficial HLA variants because they can escape the strong CTL response. Disease Progression Group (DP1) consists of Long-Term Non-Progressors (LTNPs) and Long-Term Survivors (LTS), while DP2 consisted of Slow Progressors (SLP) and Standard Progressors (SP). Rapid Progressors (RP) were not included in either group.

4.3.3 Reduced Protective Effect from Beneficial HLA variants in URF-Infected Women

Having shown that efficient CTL responses can lead to the outgrowth of HIV-1 recombinants, we next examined if the protective effect conveyed by beneficial HLA variants was maintained in women infected with URFs, compared to women infected with HHSs or CRFs (fig. 4.3). We used the biannual CD_4-counts to classify the patients according to clinical progression rate and examined the distribution of beneficial, detrimental and neutral HLA-variants in each group. We used the following classification: Long-Term Non-Progressors (LTNPs, HIV-infected for over 7 years and stable CD_4 counts over 400/µl), long-term survivors (LTS, HIV-infected for over 7 years and stable CD_4 counts between 200/µl and 400/µl), slow-progressors (SLP, HIV-infected for 5-7 years and unstable CD_4 counts between 200/µl and 400/µl, standard-progressors (SP, HIV-infected for 3-5 years and CD_4 counts over 200/µl) and rapid-progressors (RP, HIV-infected for under 3 years and CD_4 counts under 200/µL).

Patients were divided into two disease progression groups based on their biannual CD_4-counts: Disease Progression group 1 (DP1) included patients with long-term control of HIV infection (LTNPs and LTS), DP2 were those patients with slow or standard progression (SP and SLP). This bifurcation of patients allowed the use of Fischer's exact test and logistic regression. We excluded RPs in the primary analysis; this was conservative because rapid progressors (n=2) were only found in the HHS/CRFs group. Standard and slow progressors were consolidated because standard progressors often had CD_4 counts around 400/µL, while slow progressor counts were unstable but occasionally approached 400/µL (McKinnon et al. 2012).

Table 4.4: **Patients with recombinant HIV-1** Pt ID = unique cohort number; URF MP = URF Maximum Probability; Gag composition = hA, hC, hD = homogeneous HIV-A1, HIV-C and HIV-D sections; CRF_A, CRF_D = CRF HIV-A-1 and CRF HIV-D sections, respectively; CRF_AD = local HIV-A-1/HIV-D CRF, URP = unique recombinant pattern, the HIV-A1 sections of ML900 were paraphyletic and could not be analysed using our algorithm; RF = recombinant form; HLA group = beneficial (B), neutral (N), or, detrimental (D). DP = disease progression classification.

Pt ID	URF MP	Composition	RF	HLA profile	G	DP
ML199	77	hA/CRF_D	URF	A0201/3001, B4201/5703,C0701/1701	B	LTNP
ML293	99	CRF_A/hD	URF	A0302/3201, B5802/8101,C0602/0804	B	SLP
ML900	100	HIV-A/hD	URF	A3001/6802, B4201/4501, C1601/1701	N	SLP
ML1128	91	hA/hD	URF	A0301/2902, B1302/4728, C0802/0401	B	SLP

Pt ID	URF MP	Composition	RF	HLA profile	G	DP
ML1317	72	hA/CRF_D	URF	A6802/7401, B3503/5801, C0701/0401	B	SP
ML1347	99	CRF_A/hD	URF	A0301/3002, B1402/4428, C0802/0401	B	SLP
ML1419	93	hA/CRF_D	URF	A0101/7401, B4201/5702, C1701/1801	B	SLP
ML1443	0	CRF_AD	CRF	A0202/0301, B1503/4901, C0210/0701	N	RP
ML1452	0	CRF_AD	CRF	A0202/2301, B2703/5703, C0202/0701	B	SP
ML1700	100	hA,/CRF_D, URP	URF	A0205/0205, B5702/5703, C0701/0701	B	SP
ML1765	100	hA/hC	URF	A2301/3402, B4901/1801, C0210/0401	D	SLP
ML1812	100	hA/hC	URF	A0201/0205, B1503/5801, C0701/0401	B	SLP
ML1848	57	CRF_A/ hD	URF	A0205/0240, B1503/4504, C0210/1601	N	SP
ML1852	94	hA/CRF_D	URF	A0102/0201, B5701/5801, C0701/0302	B	SP
ML2000	100	hA/hC	URF	A0202/7401, B5802/1503, C0202/0602	B	SP
ML2019	98	hA/hC	URF	A6801/7401, B1503/1507, C0210/0701	B	SLP
ML2033	100	HIV-A1/C, URP	URF	A3002/3303, B5801/5702, C0701/3402	B	SP

To examine if beneficial HLA variants lead to delayed disease progression in CRF/HHS patients, we compared the frequency of beneficial, neutral, and detrimental HLA variants in DP1 and DP2 using a binary GLM with neutral as the reference category. We found that beneficial HLA variants predominated in DP1 (binary regression, p=0.0449, odds ratio = 3.24), whereas no difference was observed between the frequencies of neutral and detrimental HLA variants (p=0.3858) (fig. 4.3, tbl. 4.5). Thus, in women with HHSs/CRFs, those with beneficial HLA variants are over three times more likely to be long-term controllers of HIV-1M than those with neutral or detrimental HLA variants. If we had included the rapid-progressors in DP2, we would have obtained slightly stronger results (binary regression, p=0.0328, odds ratio = 3.48). There were not enough women infected with URFs to perform the same analysis on them.

Next, we tested whether infection with either URFs or HHSs/CRFs affected disease progression, i.e., if disease progression was predicted by URF vs HHS/CRF status. We found that patients with URFs were slightly more likely to be in DP2 (standard/slow progression) and patients with HHSs/CRFs more likely to be in DP1 (LTNPs and LTS) (binary GLM, odds-ratio=1.04, p=0.0377, tbl. 4.5). Second, we tested if the frequency of beneficial HLA variants was similar in patients with either URF or HHSs/CRFs in DP2 (SP and SLP); we observed a significantly higher frequency of beneficial HLA-variants in the URF group (binary GLM, odds ratio = 1.26, p=0.0072).

This shows that the protective effect of beneficial HLA variants is less in patients carrying URFs than in those carrying HHSs/CRFs, and suggests that the HHS/CRF viruses infecting patients

with beneficial HLA variants might escape the CTL response by recombining to form a URF.

4.4 Discussion

Here we have shown that URF-infected patients were over five times as likely to carry beneficial HLA variants. Beneficial HLA variants were associated with delayed disease progression in HHS/CRF patients — this was not the case in the URF patients, suggesting that URF expansion occurs in patients with beneficial HLA variants because the novel recombinants enable the virus to escape immune control (Ritchie et al. 2014). This selection for URFs in patients with effective CTL responses may lead to the eventual domination of immune-escaped recombinants. These viral fitness gains may explain why HIV-1 containment fails in approximately 70% of patients with beneficial HLA variants within 16 years after infection (Carrington and O'Brien 2003).

In addition, we have shown that there are fewer patients infected with URFs than with CRFs/HHSs, which is consistent with previous reports from Kenya (Neilson et al. 1999; Dowling et al. 2002; Lihana et al. 2006; Peterson et al. 2013). Even women with beneficial HLA variants carried *gag* URFs only 23% of the time. This is likely to be near the absolute upper bound for URFs in p17-p24 Gag in any cohort, as the combination of high-risk individuals exposed to a large variety of subtypes in the Pumwani cohort creates conditions for inter-subtype recombination that are rare in the rest of the world. However, since we only considered the *gag* region we were unable to observe recombinations in *pol* or *env*, and our result is, therefore, an underestimate of full-length genome URF frequency. Previous work on near full-length sequences in Nairobi found recombinants in 16/41(39%) of patients; of these 3/41(7%) appeared to be CRFs and 13/41(32%) could potentially be URFs (Dowling et al. 2002).

It is important to note that we cannot distinguish statistically between URFs and *de novo* recombinants. URFs are sequences which appear unique to us, meaning that no other sequences have been obtained which are closely related to the URF over the entire sequence. However, finding a URF within a patient does not guarantee it is the product of a recombination event within that patient. Conversely, a true *de novo* recombinant sequence is the product of a recombination event within a patient, and will always be a URF. Nor do we make any attempt at detecting intra-subtype recombin-

ants; this is difficult due to the reduced phylogenetic signal compared to inter-subtype recombinants (Rousseau et al. 2007). Therefore it is possible that the actual effects of efficient CTL responses on selection for unique recombinants are greater than the estimates obtained here, as intra-subtype URFs were not accounted for.

Our results suggest that patients with protective HLA variants and high-risk behaviour could be 'hotspots' for the outgrowth of recombinants which have escaped immune control; similar non-random distributions of HIV breakpoints have previously been described (Smyth et al. 2014), but the selective mechanism for their outgrowth within patients with beneficial HLA variants has not. This is especially relevant in the context of a vaccine, which would boost the CTL response in individuals with ordinary HLA variants to at least the strength of patients with beneficial HLA variants. Strong vaccine-induced immune responses could therefore potentially lead to vaccine escape in high-risk patients, especially — but not only — those exposed to many different subtypes like the Pumwani CSWs. When a vaccine is implemented, it will be important to monitor high-risk populations for signs of novel vaccine-escaped HIV-1 recombinants and to use anti-retroviral drugs to prevent the spread of such strains. Vaccine escape has been documented in Hepatitis A (Pérez-Sautu et al. 2011) and Hepatitis B (Huang et al. 2004). A recombinant vaccine-escape has also contributed to the outgrowth and rising incidence of a *Streptococcus pneumoniae* strain in the United States (Brueggemann et al. 2007). Therefore, monitoring high-risk populations, especially those exposed to a large variety of HIV-1, should be considered a public health priority when large-scale vaccination programs are implemented.

4.5 Data Tables

Table 4.5: **HIV subtype, HLA group and disease progression data for patients with HHS or CRFs.** HLA group: beneficial (B), detrimental (D) or neutral (N) HLA variants. DP = disease progression; long-term non-progressor (LTNP), long-term survivor (LTS), slow progressor (SLP), standard progressor (SP) and rapid progressor (RP).

ID	HIV-1 subtype	HLA group	DP
5	C	B	SLP
13	A1	D	SP
16	A1	D	SLP
24	C	N	SP
58	D	B	LTS
59	A1	B	SP
115	A1	N	SLP
142	A1	B	LTS
157	D	B	LTNP
249	D	B	LTNP
264	A1	D	LTS
265	A1	N	LTNP
274	A1	B	LTNP
300	A1	N	SP
313	A1	N	SLP
330	A1	N	SP
406	A1	B	LTNP
412	A1	B	LTNP
415	A1	D	LTNP
495	A1	N	SLP
525	A1	B	LTNP
579	A1	N	LTNP
589	A1	D	SLP
590	A1	D	SLP
602	A1	B	SP
646	A1	B	LTNP
657	A1	N	SLP
672	A1	B	LTNP
810	A1	N	SLP
874	A1	N	SLP
890	A1	D	SLP
915	D	B	LTS
959	A1	N	SLP
960	A1	N	LTS
977	D	N	SP
995	A1	B	SLP
1032	A1	B	SLP
1051	D	D	SLP
1102	D	N	SP
1125	A1	B	LTNP

ID	HIV-1 subtype	HLA group	DP
1129	A1	N	SLP
1197	A1	B	SLP
1227	A1	B	SLP
1292	D	B	SLP
1295	A1	B	SP
1296	A1	D	SP
1318	A1	N	SP
1337	D	N	RP
1346	A1	N	SLP
1349	D	B	SLP
1387	C	B	SP
1390	A1	B	SLP
1401	A1	D	LTS
1404	A1	D	SLP
1423	D	B	SLP
1425	A1	D	SLP
1443	CRF_A/D	N	RP
1450	A1	N	SLP
1452	CRF_A/D	B	SP
1474	A1	D	SP
1497	A1	N	SLP
1514	D	B	SLP
1535	A1	D	SLP
1540	C	B	SLP
1560	A1	N	SP
1561	D	B	SP
1564	A1	B	SLP
1594	A1	N	SLP
1596	C	N	SLP
1611	A1	D	SP
1625	A1	B	LTNP
1649	D	B	SLP
1654	A1	B	LTNP
1670	A1	N	LTS
1725	A1	N	LTNP
1731	A1	B	SLP
1740	A1	D	LTS
1761	D	N	SLP
1770	A1	N	SLP
1771	D	N	SLP
1772	A1	D	SLP
1777	A1	B	SP
1802	C	B	LTS
1811	D	N	SLP
1825	C	N	SP
1834	A1	N	SP

ID	HIV-1 subtype	HLA group	DP
1859	D	B	SP
1862	A1	D	LTNP
1932	A1	B	SP
1947	A1	D	SP
1970	A1	D	SP
1974	D	B	SP
2001	A1	B	SP
2028	D	D	SP
2042	A1	D	LTNP
2052	A1	D	SP
2137	A1	D	SP
2204	A1	N	SP

Table 4.6: **LANL sequences found using blastn with Pumwani query sequences** These sequences (or their matching sequence segments if recombinants) were added into the alignments.

Sequence ID
HIV-A1
A.KE.x.K29.L11770
A1.KE.99.clone_2609.GQ432473
A1.SE.94.SE7253.AF069670
A1D.KE.97.clone_2405.GQ432269_A_frag_1_602
A1.KE.95.clone_2819.GQ432683
A1D.TZ.03.H495.FJ853568_A_frag_1_388
A1.KE.96.clone_1.GQ429865
A1.UG.x.UG031.AB098331
A1.KE.96.clone_2851.GQ432715
A1.KE.98.clone_2547.GQ432411
A1.RW.92.92RW025A.AB287377
AD.KE.98.M381gag.AY772972_A_frag_1_342
A1.TZ.03.H628.FJ853591
A1.KE.00.NKU3005.AF457089
A1.KE.06.06KECst_005.FJ623481
A1.KE.95.clone_1919.GQ431783
01A1.CY.06.CY178.FJ388953
A1D.KE.91.clone_1530.GQ431394_A_frag_1_410
A1D.KE.95.clone_1170.GQ431034_A_frag_1_443
A1.TZ.03.H564.FJ853577
A.KE.99.M439gag.AY772985
A1.KE.95.clone_2101.GQ431965
A1.KE.95.clone_519.GQ430383
A1.KE.95.clone_936.GQ430800

Sequence ID
A1D.KE.97.clone_1676.GQ431540_A_frag_1_590
A1.KE.95.clone_2698.GQ432562
A1.UG.03.03_9538NG.AY803390
A1C.KE.96.clone_256.GQ430120_A_frag_1_387
A1.KE.97.ML013_2.AY322185
A.KE.x.K112.L11768
A1.KE.95.clone_2169.GQ432033
A1.KE.92.clone_429.GQ430293
A1.KE.97.ML752.AY322193
A.KE.x.K98.L11775
A1.KE.95.clone_583.GQ430447
A1.KE.95.clone_2518.GQ432382
A1.KE.87.clone_2437.GQ432301
A1.KE.98.clone_2883.GQ432747
A.KE.98.M428gag.AY772986
A1.KE.88.clone_2286.GQ432150
A1.UG.92.92UG037.AB253429
A.KE.98.M369gag.AY772971
A1.KE.01.clone_2639.GQ432503
A1.KE.06.06KECst_007.FJ623476
A1.KE.99.KNH1237.DQ367272
A1.KE.95.clone_41.GQ429905
A1.KE.99.clone_999.GQ430863
A1.KE.02.clone_1129.GQ430993
A.KE.97.M104gag.AY772953
A1.KE.98.clone_1561.GQ431425
A.KE.98.M178gag.AY772958
A1.KE.96.clone_2730.GQ432594
A1.KE.01.clone_220.GQ430084
A1.KE.01.ML1945.EU110088
A1.KE.99.KNH1240.DQ367273
HIV-C
C.CD.x.VI313.L11787
A1C.KE.96.clone_256.GQ430120_C_frag_387_1337
A1C.KE.99.KNH1097.AF457064_C_frag_375_1018
C.TZ.03.H402.FJ853559
C.BR.04.04BR038.AY727524
C.IL.00.00ET669_154_114.AY255817
C.ZM.01.01ZM541036.DQ792998
C.KE.95.clone_1750.GQ431614
C.SO.89.SO145_SM145_SM1451.L11803
C.BW.96.96BW0402.AF110962
HIV-D

Sequence ID
D.KE.87.clone_347.GQ430211
D.UG.94.94UG114.U88824
D.KE.88.clone_461.GQ430325
D.KE.89.clone_1093.GQ430957
D.UG.93.93UG_065.AY713418

5 Repeated Recombination and Decades-Long Absence of Superinfection in Two Dually HIV-Infected High-Risk Female Commercial Sex Workers[1]

5.1 Introduction

In most studies of HIV-1 infection, a distinction is drawn between 1) dual infection, which is any patient found to be infected with two strains of HIV, 2) observed superinfection, when an already HIV-infected patient encounters an additional strain of HIV and thereby becomes dually (or triply) infected, and 3) coinfection, when a patient is found to be infected with two strains of HIV simultaneously and goes directly from uninfected to dually infected (Jost et al. 2002). Dual infection is much rarer than single infection (Keele et al. 2008).

Determining the multiplicity of infection is easier if the lineages infecting a patient derive from different subtypes, as the large phylogenetic distances between the subtypes can be detected using standard tools (Schultz et al. 2006; Pineda-Peña et al. 2013). If patients are infected with multiple sequences of the same subtype, phylogenetic analyses must be used to distinguish between them. Phylogenetically, different lineages infecting a single patient can be identified because their sequences will be polyphyletic (each lineage will have their own common ancestor, these common ancestors will be dispersed through the tree) in a tree derived from a well-sampled population (Ou et al. 1992; Hillis and Huelsenbeck 1994; Sajantila 2014). Dated trees can also be used to determine infection histories because most HIV-1 infections are established by a single virion (Keele et al. 2008) the time to the Most Recent Common Ancestor (MRCA) of the sampled patient sequences should reflect the time that the patient was exposed to the infecting virion. If a patient is infected with multiple

[1]We thank Dr Katzourakis for our discussions regarding the outline of the analyses of these patient sequences.

lineages, those HIV-1 lineages should be separated, or time to the root node of that lineage should be used instead.

In the previous chapters, we have studied how CTL escape may drive subtype formation and recombination at the population level. In this chapter we studied two patients from the Pumwani cohort over approximately a decade, to examine how the CTL immune response may drive Gag immune escape and recombination. Because these patients had an average of 3.8 clients per day over at least 10 (patient ML1812) and 18 years (patient ML900), they should have been highly exposed to a range of HIV-1M subtypes (Simonsen et al. 1990; Peterson et al. 2013).

We used sequence data from both patients to perform exploratory analyses. Firstly, we asked whether dually infected patients get infected with additional HIV-1 lineages or subtypes during their years as CSWs, or whether they became multiply infected at once or immediately after their initial infection. To do this we found a *gag* segment which was the same subtype in all of a patient's sequences and used this segment to build a dated phylogeny. We used this phylogeny together with the observed subtype patterns at each timepoint to show that one patient likely became multiply infected at the time of her first infection or immediately afterward — and as a basis to discuss possible scenarios for the other patient.

5.2 Methods

We used the HIV-1M sequence data and HLA types gathered by Dr Chang and Dr Iversen from longitudinal PBMC samples as described in the previous chapter. Four samples spanning 1991 to 2002 were available for ML900, and eight samples spanning 1996 to 2004 were obtained from patient ML1812. All samples were collected before ART was available to this cohort, and both patients were asymptomatic.

The Pumwani cohort studies were conducted according to the principles expressed in the Declaration of Helsinki, the National AIDS Committee. The National Ethics and Scientific Review Committee of Kenyatta National and the University of Manitoba Use of Human Subjects in Research review committee have approved studies using this cohort.

Here we use subtype-segment to mean sequences which align to a particular HIV region and belong

to one subtype exclusively. Creating separate alignments for each subtype-section was necessary, as mixing sequences of different subtypes would lead to a common ancestor around the initial outbreak of HIV-1M, which is not informative when studying the within-patient evolutionary history (Faria et al. 2014).

5.2.1 HIV-1 Hypermutation and Subtype Analyses

LANL's HYPERMUT program (Rose & Korber 2000) was used to detect hypermutated sequences in both dually infected patients. During reverse transcription in the target cell's cytoplasm, APOBEC3 (especially APOBEC3F and APOBEC3G) proteins induce cytidine-to-uridine (C-to-U) deamination of single-stranded minus-strand HIV-1 DNA, which generated guanosine-to-adenosine (G-to-A) mutations in plus-strand DNA (Armitage et al. 2008). These G-to-A mutations occur in APOBEC3F- and APOBEC3G-favored motifs (GGnn-to-AGnn and GAnn-to-AAnn contexts, respectively). Because hypermutation can lead to long branches and skewed trees in reconstructed phylogenies it was necessary to repair potentially hypermutated sites: nucleotides in adjacent alignment columns containing both GG and AG or both GA and AA were repaired to NG and NA respectively (Armitage et al. 2008, 2012). The repaired sequences were subtyped using jpHMM (Schultz et al. 2009), subtyping was verified using REGA (Pineda-Peña et al. 2013), and alignments of each subtype-section were created in AliView (Larsson 2014) for both patients.

To examine if patients were infected with recombinant forms of the same subtype-pattern but different subtype-segment lineages we used MrBayes v3.2.3 (Ronquist et al. 2012) to build phylogenies from alignments for each patient's subtypes separately, including: 1. the patient's subtype-segment (e.g., subtype C p17 sequences, or subtype A p24 sequences), 2. matching subtype-segments from all sequenced patients in the Pumwani cohort, 3. the best match sequences from the LANL database (as described in the previous chapter). We constructed undated phylogenies for each subtype-segment of both dually infected patients, using the GTR + Gamma model of nucleotide evolution for both patients' p24 alignments, and the K80 + Gamma model for each patient's two p17 alignments. In each case MrBayes was run for 50M steps, discarding a 25% burnin.

We used Path-O-Gen to examine clock-likeness for both patients, and used BEAST v1.8.2

(Drummond et al. 2005, 2006, 2012) to generate dated trees for each patient's p24 subtype-segment (p24 subtype D for ML900 and most of p24 subtype C for ML1812), using the GTR+G model of nucleotide evolution, a skyline coalescent tree prior (Drummond et al. 2005), and a lognormal clock with an exponential prior of mean 4.7E-6 substitutions / site / day (Alizon and Fraser 2013). We ran 10 BEAST chains for 250M steps each, discarded an appropriate 25% burnin and confirmed convergence in TRACER. Maximum Clade Credibility trees were generated using TreeAnnotator, which was visualised using FigTree and ete2 (Huerta-Cepas, Dopazo and Gabaldón 2010). We annotated these trees with the translated nucleotide sequences coding for epitopes presentable by our patients' HLA variants, which were extracted from the LANL epitope table (http://www.hiv.lanl.gov/content/immunology/tables/ctl_summary.html). We also included three amino acids on either side of the epitope in these sequences, as neighbouring amino acids can cause CTL escape by interfering with endoplasmic reticulum aminopeptidase 1 and 2 trimming or epitope processing (Draenert et al. 2004; Tenzer et al. 2014).

Due to the absence of a clock-like signal in any of the p17 alignments we were only able to construct undated phylogenies for the p24 subtype-segments. Possibly because the repeated outgrowths of novel p17 recombinants preclude the gradual temporal variation that can be interpreted as clock-like.

5.3 Results

5.3.1 Patient ML1812

This patient's HLA variants were A*02:01/02:05, B*15:03/58:01, C*07:01/04:01. The subtype analysis (fig. 5.1, fig. 5.2A), detected three forms of *gag*: C *gag*, AC *gag* (subtype A p17 and subtype C p24), and ACA *gag* (like AC *gag* but with an additional subtype A segment for the last 200bp of p24). jpHMM also reported one singleton recombinant form sampled on day 1975 which could not be confirmed by REGA (1812_GAG_AC_c_g_12_1975), similar to AC *gag* but with a short subtype C fragment at the beginning of p17. To determine that there were no indistinguishable dual infections of the same recombinant pattern, we made undated phylogenies including sequences from other patients. In these three undated phylogenies (subtype A p17, subtype C p17, subtype

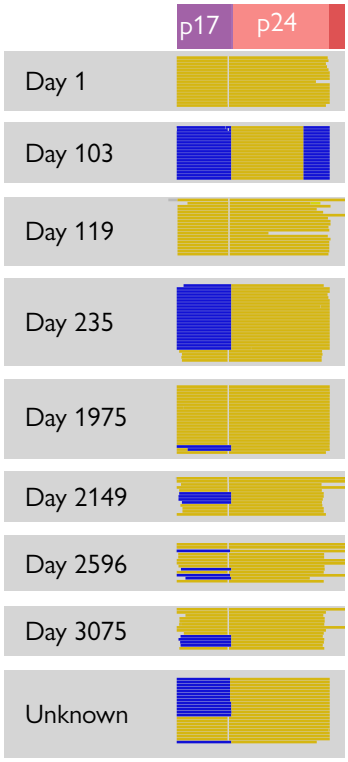


Figure 5.1: **ML1812 subtype analysis**
Subtype A is shown in yellow, and subtype C is shown in blue. Three HIV-1M *gag* recombinant patterns were observed; pure C, AC, and ACA.

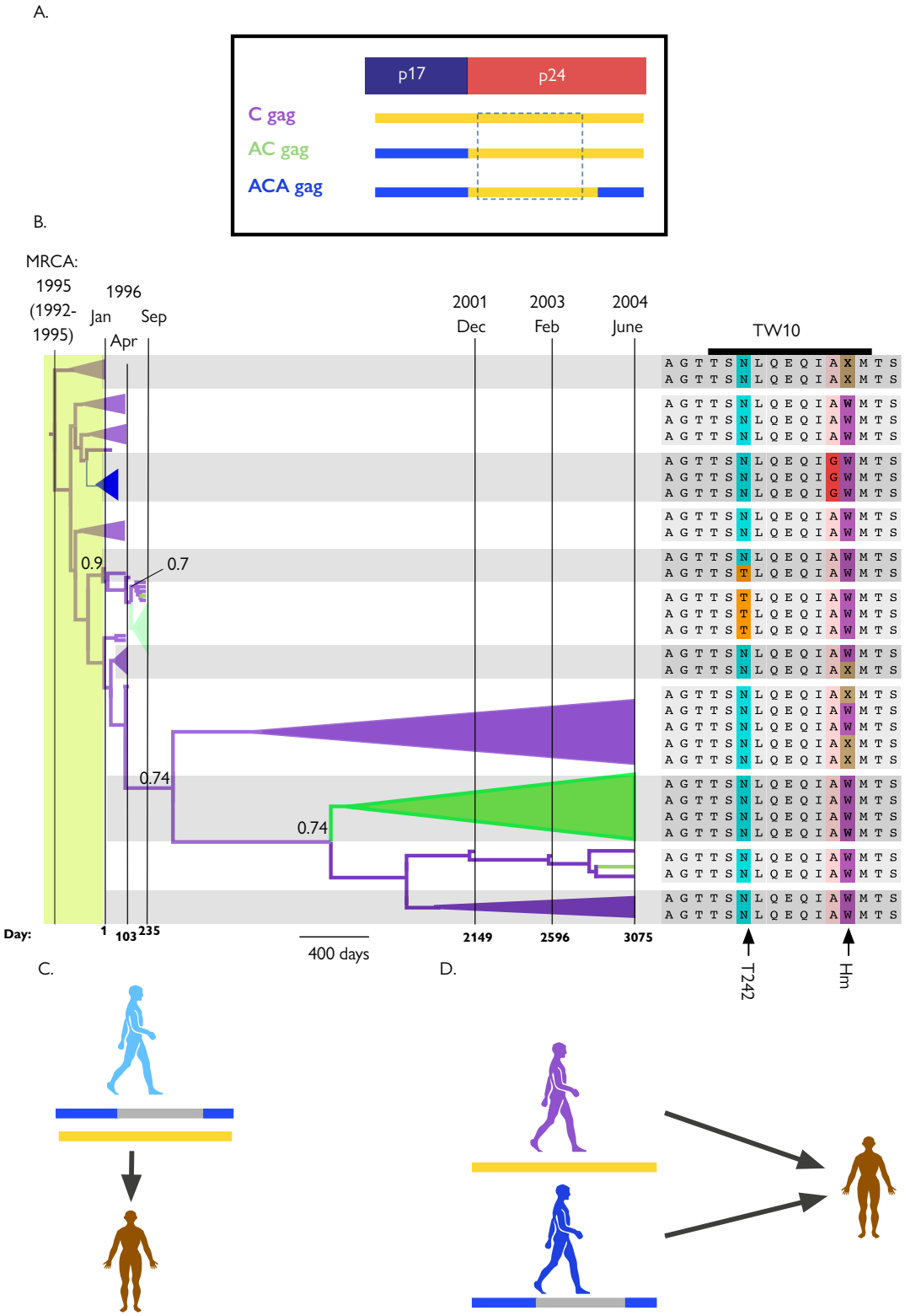


Figure 5.2: **Patient ML1812** A) diagram of recombinant forms B) simplified phylogeny annotated with most common TW10 sequences, Hm = hypermutant, T242N = a known B*58:01 escape mutant. The late AC clade (dark green) is not a direct descendant of the early AC clade (light green) C) coinfection: one donor simultaneously infects ML1812 with two or more recombinant forms. D) Early Superinfection: the patient is first infected with one recombinant pattern by the original donor, and superinfected shortly afterwards with another recombinant pattern by another donor. Superinfection must have happened before day 103, as the patient was already dual infected then.

C p24) the sequences derived from ML1812 were monophyletic, showing that ML1812 was not multiply infected beyond the two p17s (subtypes A & C) and one p24 (subtype C) observed in the subtype analysis. The origin of the two p17 fragments and the subtype A p24 fragment is unknown and could be the result of either coinfection or early superinfection (fig. 5.2C&D).

Due to the absence of a clock-like signal in either p17 alignment we were only able to construct a dated phylogeny (fig. 5.2B) representing the within-host evolution for the subtype C p24 segment (see dotted line in fig. 5.2A). The MRCA of this phylogeny occurred approximately two years before the first time point (95% HPD interval: 0-4 years), and represents one of the infecting virions (the only infecting virion from which subtype C p24 segments descend). The recombinant form composition varies considerably over time: on day one, the sample appears to consist exclusively of subtype C *gag*, but this is replaced with a majority ACA *gag* lineage by day 103 (blue clade in fig. 5.2B; ACA *gag* is not observed after this point). On day 235 the patient is predominantly infected with AC *gag* (light green clade in fig. 5.2 B) but there are still some circulating subtype C *gag* sequences. At the later time points (1975, 2149, 2596 and 3075) there is a varying composition of subtype C and AC *gag*. Unexpectedly, the dated phylogeny shows that the AC *gag* sequences from later time points (dark green clade in fig. 5.2B) are not descendants from the AC *gag* at time point 235 (light green clade in fig. 5.2B) but that patient ML1812 carried two phylogenetically distinct strains of AC *gag* (separated by two nodes with posterior probabilities 0.7 and 0.9, which both have subtype C *gag* descendants), as well as one singleton AC *gag* sequence (green branch in fig. 5.2B) embedded in a subtype C clade with good support (posterior probability is 0.7). This suggests that there were at least two recombination events (one fewer than the number of clades to allow for the original dual infection) resulting in the formation of AC *gag* in this patient, in addition to one recombination event leading to ACA *gag*. The undated phylogeny of subtype A p17 was insufficiently resolved to determine whether the p17 A segment in the late AC clade was derived from the ACA sequences or the early AC clade.

5.3.2 Patient ML900

This patient's HLA type was HLA-A*30:01/*68:02, HLA-B*42:01/*45:01, HLA-C*16:01/*17:01. The patient ML900 *gag* sequences occurred in two forms (fig. 5.4, fig. 5.3A): 1) pure subtype D *gag*

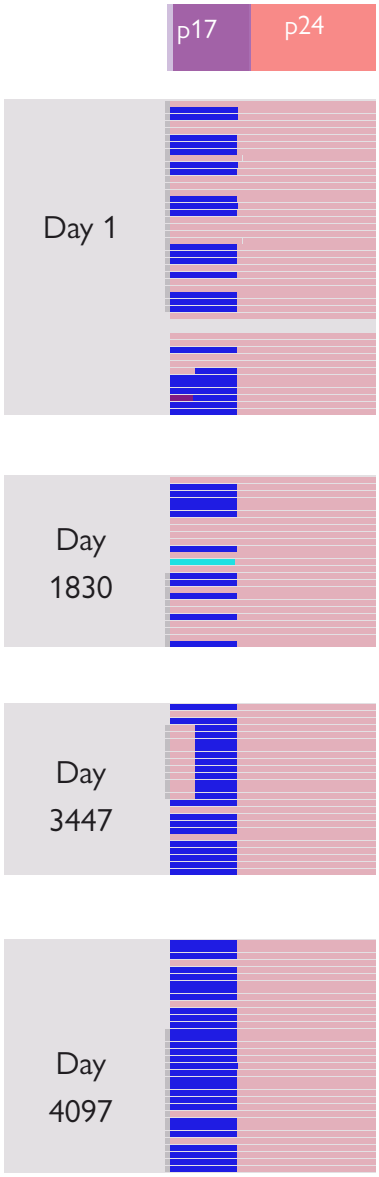


Figure 5.3: **ML900 Subtype Analysis**
Subtype A is shown in blue and subtype D in pink. This patient has two *gag* forms: pure D and an AD recombinant. We also found a singleton sequence with a p17 segment which was called as K by jpHMM, but this was not replicated in REGA. In this analysis using a newer version of jpHMM we did not find the A2 sequence mentioned in Chapter Four.

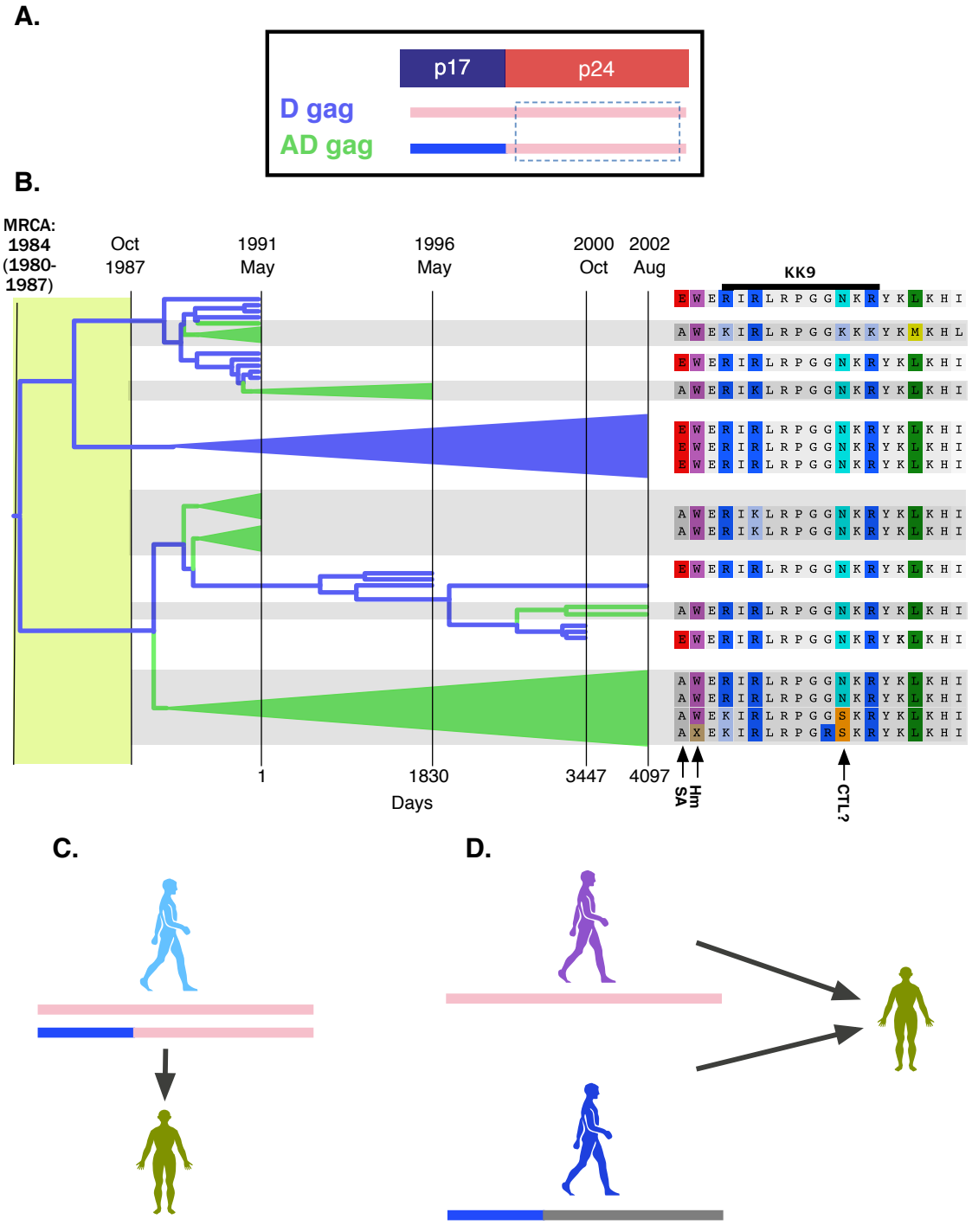


Figure 5.4: **Patient ML900** A) recombinant patterns present in this patient. B) A dated phylogeny of Patient ML900, with the region around the HLA A*30:01 K9R epitope illustrating the three different column categories. SA=subtype associated polymorphism, Hm = hypermutant, CTL? = possibly CTL escape mutation. C) coinfection hypothesis, where both recombinant patterns come from the same donor. The p24 segment in both transmitted virions is closely related D) superinfection hypothesis, where two donors each transmit one recombinant form to patient ML900, with one of the two p24 segments going extinct before the first sampled time point.

('D *gag*'), and 2) subtype A1 p17 and subtype D p24 ('AD *gag*'). In this analysis using a different version of jpHMM we did not find the A2 sequence mentioned in Chapter Four. The recombinant form composition in ML900 was much more stable over time than ML1812: both D *gag* and AD *gag* were present at all time-points (fig. 5.4B). However, like ML1812 the AD *gag* recombinant forms appears to have repeatedly evolved, and at least two independent recombination events are supported by the phylogeny, allowing for one well-supported clade as the original AD infection.

The undated MrBayes phylogenies including sequences from other patients were too polytomous to be informative.

The MRCA of this patient's p24 *gag* sequences is estimated to be from 1984 (95% HPD interval 1980-1987), which coincides with the beginning of the HIV-1 epidemic in Pumwani. The patient was enrolled three years later in 1987, by which time she was already seropositive; the first time point when samples were available for sequencing was in 1991. Thus, the long branches leading to the two D clades and the AD clade could be an artefact of late sampling or an indication that this patient was coinfected by two virions from the same donor (fig. 5.4C). In the latter case, the p24 MRCA occurred not in patient ML900 but in the donor. Alternatively, the patient could have been coinfected with two unrelated virions or experienced early superinfection followed by recombination which preserved the p17 segment while the other p24 subtype-segment went extinct before the first sequenced samples (fig. 5.4D).

5.3.3 Epitope Evolution in ML900 and ML1812

We translated the *gag* sequences from ML1812 and ML900 and examined the CTL epitopes that were potentially presentable in each patient. Ignoring rare polymorphisms, we found three major drivers of amino acid variation in these sequences: 1) hypermutation, 2) subtype associated polymorphisms, 3) potential CTL escape mutations (see fig. 5.2B & fig. 5.4B for examples of each). In the patient ML1812 epitopes we found 12 alignment columns driven by hypermutation, 14 columns with subtype associated polymorphisms (11 in p17 and 3 in p24, due to the subtype A fragment at the end of ACA *gag*), and five columns which appeared to be directly CTL driven. Also, we found three columns which appear to show both subtype associated and directly CTL driven polymorphisms. We observed a similar pattern within the ML900 sequences. Collectively, our results suggest

that p17 may be more exposed to CTL targeting than p24 as the epitopes in p24 appeared to carry fixed escape mutations (i.e., p24 was "fully escaped" or completely adapted to escape the patients CTL responses throughout) in contrast to the continuously changing epitopes in p17. This difference in fixed CTL escape mutation patterns might explain the repeated recombination events and the preservation of the p24 *gag* sequence.

5.4 Discussion

5.4.1 Repeated Outgrowth of Identical Recombinant Forms of Different Lineages

In both patients, we found evidence of multiple recombination events between p17 and p24, where a pure subtype p24 *gag* was linked to a p17 *gag* from another subtype to create distinct and repeated outgrowths of HIV-1M recombinants with the same overall recombination pattern. Recombination has previously been shown to allow viruses to rapidly adapt to the CTL response in one HIV-infected patient during the acute and early infection stages (Ritchie et al. 2014). The fact that a new subtype A p17 fragment is linked to a recombination-wise pre-adapted subtype C/D p24 fragment probably makes survival of the new recombinants more likely. Such recombinations are similar to our finding that segments derived from CRFs are overrepresented in URFs sampled from the Pumwani Cohort (Chapter 4). However, this does not explain why the new recombinant would have a selective advantage. Our sequences carry footprints of continuously acquiring new intra-epitope escape mutations in p17, but not p24, suggesting an on-going CTL response against this part of HIV-1. Therefore, new intra-epitope CTL escape mutations in p17 might carry an advantage compared to old escape mutations whereas p24 is already fully adapted to escape the patient's immune response. In addition, previous research has shown that epitope processing is highly sensitive to subtype-specific differences (Tenzer et al. 2014), which could lead to host cells being able to make epitopes for one subtype but not for another, which may favour one subtype in p17 and one in p24 in these patients. The continuous recombination events suggests that a p17 *gag* sequence with a combination of escape-mutations may spread more readily through the population if inherited "en-block", especially as both patients carry beneficial HLA variants (HLA B*58:01 and HLA B*42:01) in line with the effect of strong CTL responses in Chapter Four. Previous work has shown that the CTL pressure on p24 is greater than on p17 (Kiepiela et al. 2007), which might explain why p24 has escaped fully while p17 has

not. This could explain why there's only one p24 and two p17s, and also why it would be beneficial for subtype A p17 sequences to obtain the latest escape mutations to the p24 CTL responses (see fig. 5.2B: the later AC clade has the T242N B*58:01 CTL escape mutation in p24 that the early AC clade lacks).

5.4.2 Absence of Further Superinfection Beyond Initial Dual-Infection

We did not observe further superinfection in spite of the reportedly high degree of exposure these patients endured, possibly because HIV-infected patients mount an immune response against their existing infection which in most cases prevents superinfection. We naively calculated the expected number of superinfections for the period the patient was active (measured from the MRCA until the final sample). We used the mean number of contacts (3.8 contacts / per day (Simonsen et al. 1990)), condom usage (0% up to 1990 (Simonsen et al. 1990), 80% usage by 2013 (Peterson et al. 2013), linear interpolation for intermediate years) and the risk of seroconversion per sexual act (ranging from 0.4/1000 unprotected contacts in 1985 to 0.075 / 1000 unprotected contacts by 2000 (Kimani et al. 2008)). We then reduced this number by 25% to allow for the patients' vacations in the countryside. We were surprised to find that patient ML900 was expected to have been superinfected with only 3.4 viruses over her 18 years as a CSW, and patient ML1812 – active for only ten years in a period with a lower risk of infection – was expected to be superinfected with only 0.6 viruses. This calculation does not consider a recent study suggesting superinfection is approximately 50% less likely after HIV-specific immune responses have matured during the first six months after initial infection (Ronen et al. 2013). Allowing for this result would almost half the expected number of superinfections for both patients. These counterintuitive results show that no anti-superinfection CTL response is necessary to explain the degree of multiple infections found in either of these patients, not even the reported halving of superinfection risk after 6 months.

6 Paleovirology of Mammalian Betaretroviruses Reveals Ancient and Intercontinental Viral Flow[1]

6.1 Introduction

Cross-species transmission is the source of many human diseases, including some of those causing the greatest social and economic costs such as influenza (Kuiken et al. 2006), dengue (Rudnick, A. & Lim 1986) and HIV (Gao et al. 1999). Research on cross-species transmission has mostly focused on host-species of particular interest, such as the transmission of pathogens from other animals to humans or livestock (Weiss 2001; Guan et al. 2003; Li et al. 2005; Epstein et al. 2006; Vasilakis et al. 2007; Parrish et al. 2008; Levitt et al. 2013; Allison et al. 2013; LeBreton et al. 2014). In this paper, we take a less anthropocentric approach, by quantitatively studying all detectable cross-species transmissions within a small clade of viruses over millions of years rather than focusing on host species of particular interest. To accomplish this, we analyse endogenous retroviruses (ERVs), retroviruses that have inserted into their host's germ line and are now transmitted vertically from parent to offspring as Mendelian alleles (Weiss 2006). These viral sequences are referred to as 'fossil viruses' because they stopped evolving at the rapid viral rate of evolution after integrating into the germ line and are now changing at their host's rate of neutral evolution. These sequences contain information about the evolutionary history of their ancestor viruses. We can use this information to explicitly model cross-species transmission based on the idea that closely related viruses sequences in diverged hosts are the result of cross-species transmissions.

The earliest incarnations of these methods consisted of simple DNA-hybridization experiments, such as work showing the great similarity between an endogenous virus from the Asian mouse and

[1]The work in this chapter was supervised by Dr Katzourakis.

an exogenous virus from the distantly related gibbon (Lieber et al. 1975), both gammaretroviruses closely related to the currently endogenising Koala Retrovirus (KoRV) (Martin et al. 1999; Tarlinton, Meers and Young 2006). More recent data has shown that ERVs in the laboratory mouse and ERVs from the Asian mouse are equally closely related to the Koala retrovirus – which only arrived in Australia recently – supporting the hypothesis that European rodents may have spread the virus to Australia in colonial times (Martin et al. 1999; Hayward, Grabherr and Jern 2013). Similar methods have been applied with great success to HIV-1 and HIV-2, which are found dispersed in the Simian Immunodeficiency Virus (SIV) phylogeny as 12 distinct monophyletic clades, showing that HIV-1 and HIV-2 have crossed over into humans at least four and eight times respectively (Sharp and Hahn 2011).

Not all viruses can successfully cross between species: the foamy viruses are a clade of retroviruses that appear to have co-evolved with mammals for the last 100 million years with little evidence of cross-species transmission during that time (Katzourakis et al. 2009). Similarly the rhabdoviruses only rarely jump between their vertebrate and arthropod host species (Longdon et al. 2014). However, foamy virus spill-over infections in humans have been recorded in several populations that frequently interact with non-human primates such as lab workers (Brooks et al. 2002) and bushmeat hunters (Wolfe et al. 2004; Switzer et al. 2012). No subsequent outbreaks of foamy viruses in humans are known to literature, despite extensive surveillance including a study of African pygmies in which two strains of HTLV were shared between family members, but SFV had not transmitted beyond the spillover host (Calattini et al. 2011).

Unlike foamy viruses, the betaretroviral phylogeny is discordant with the mammalian phylogeny (Gifford et al. 2005), which provides evidence for an evolutionary history containing many cross-species transmission events. Endogenous members of this clade have been described in a great variety of mammals (Herniou et al. 1998; Baillie et al. 2004) which makes this an appropriate clade for the quantitative analysis of cross-species transmission. The exact clade we use is a phylogenetically well-supported group incorporating two defunct taxa of exogenous betaretroviruses: type B betaretroviruses (such as Mouse Mammary Tumor Virus (MMTV)) and type D betaretroviruses (such as Mason Pfizer Monkey Virus (MPMV) and Jaagsiekte Sheep Retrovirus (JSRV)), as well as ERVs homologous to these viruses, such as the intracisternal A-type particles (IAPs) and Simian ERV

(SERV) (Herniou et al. 1998; Gifford and Tristem 2003). The oldest endogenous representatives of this clade inserted into mammalian genomes approximately 40 million years ago (Katzourakis et al. 2014), which is long after the mammalian radiation (Bininda-Emonds et al. 2007; O'Leary et al. 2013).

Cross-species transmission is a relatively new problem in evolutionary biology but is conceptually similar to the study of migration. Migration has long been a feature of theoretical biology (Wright 1990) and has been further generalised as the evolution of a discrete state (such as a geographical location, host-species or any other discrete morphological or ecological trait) along a phylogeny. Some methods statistically test the hypothesis of association between phylogeny and a discrete state using parsimony (Parker, Rambaut and Pybus 2008). Others explicitly model the evolution of discrete states in maximum likelihood (Pagel 1999) and Bayesian frameworks (Pagel, Meade and Barker 2004; Lemey et al. 2009; Faria et al. 2013) to estimate the ancestral states of internal nodes as well as the transition rates between discrete states. BEAST is the most popular implementation of discrete phylogeography in a Bayesian framework. Recent improvements to this implementation include a generalised linear model that can reveal the factors driving discrete state transitions (Lemey, Rambaut and Bedford 2012) as well as a way of inferring the number of transitions between each pair of discrete states (Minin and Suchard 2008; Talbi et al. 2010). These transitions between states can also be conceptualised as the movement of viral lineages locations (Faria et al. 2014). Cross-species transmission is often studied using phylogeographic methods, which can be used without further modifications to infer transmissions between hosts rather than migrations between places.

Previous studies of virus cross-species transmission using BEAST's discrete phylogeographic model exclusively considered exogenous viruses, such as rabies in bats (Faria et al. 2013), genotype VI avian paramyxovirus serotype 1 in birds (Chong et al. 2013), SIV/HIV in primates (Zhang et al. 2012) and parvovirus in carnivores (Allison et al. 2013). Only one of these has used the generalized linear model and Markov counting enhancements to the discrete phylogeographic model (Faria et al. 2013). Conversely, studies on cross-species transmission in paleovirological data have lacked a quantitative analysis (Baillie et al. 2004; Hayward, Grabherr and Jern 2013).

In this chapter we attempt to understand which mammalian species have transmitted more viruses to one another over the evolutionary history of this this clade. We also seek to understand what factors

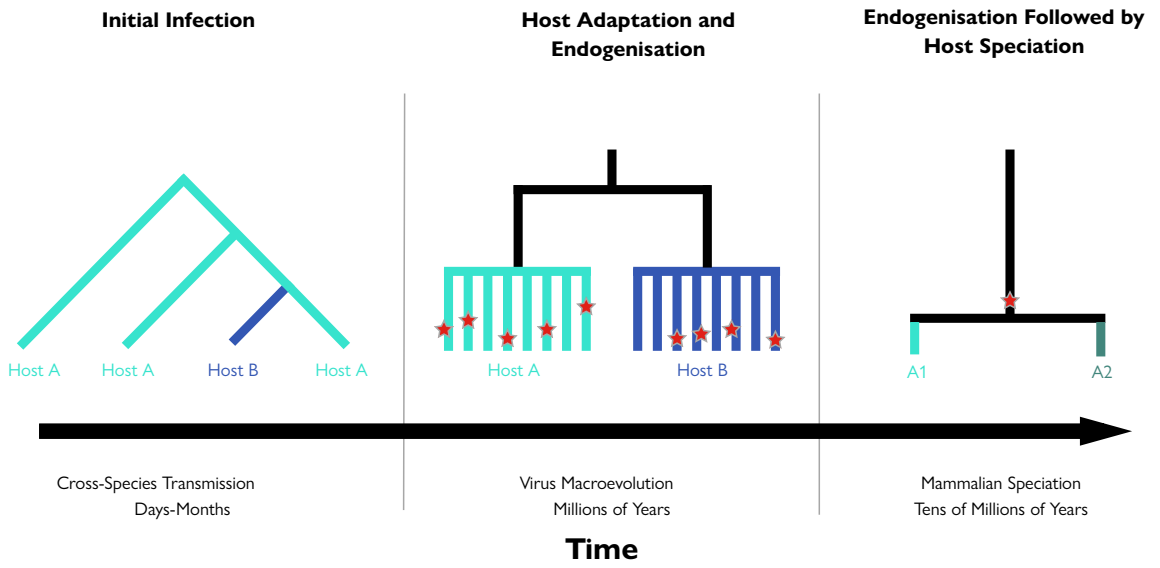


Figure 6.1: Virus phylogenies at different time scales following a cross-species transmission event. The first panel shows a hypothetical phylogeny of exogenous viruses immediately after a primary cross-species transmission event from host A to host B. The second panel shows two closely related exogenous viruses together, with integration events shown as red stars. The directionality of the cross-species transmission cannot be inferred from such a phylogeny. The third panel shows what a phylogeny of endogenous viruses look like if integration is followed by host speciation, showing sequence similarity due to orthology rather than cross-species transmission.

(host relatedness, geographic overlap) makes species more or less likely to transmit to one another. To accomplish this we apply BEAST's discrete phylogeographic method and its recent enhancements to paleovirological data - allowing us to construct a 'viral network' between host-species, as well as to draw more general conclusions about what host characteristics drive transmission by using a GLM to predict transmission rates between species using host characteristics.

6.2 Results & Discussion

By explicitly modelling cross-species transmission we have revealed ancient routes of viral flow between the ancestors of modern mammals, and shown that rodents are an important carrier for betaretroviruses (fig. 6.2). Though rodents have previously been implicated in the spread of beta and gammaretroviruses (Baillie et al. 2004; Hayward, Grabherr and Jern 2013), we show for the first time that more species share a statistically supported connection with the mouse and rat than any other animal. Bats have been implicated in the cross-species transmission of many viruses (Leroy et al. 2005; Calisher et al. 2006; Ge et al. 2014), and several bat species were included in the analyses. However, we did not recover a central position for bats in the viral network, showing that bats do

not play a universal role in the cross-species transmission of viruses.

In addition, we used the generalised linear model extension of the phylogeographic model to show that host phylogenetic relatedness drives cross-species transmission in this betaretroviral clade. Similar effects have previously been shown on much smaller phylogenetic scales, observationally using sequence-based methods in bats infected with different rabies viruses (Streicker et al. 2010; Faria et al. 2013), and experimentally in Drosophila species cross-infected with species-specific Drosophila sigma viruses (Longdon et al. 2011). We have now expanded these results over much larger phylogenetic and temporal scales; showing that this effect is persistent over a much more phylogenetically diverse clade of viruses, between the entire mammalian taxonomy rather than within bats or Drosophila.

6.2.1 Does Host Phylogenetic Relatedness Affect Cross-Species Transmission?

We used BEAST's generalised linear model to show (Faria et al. 2013) that host phylogenetic relatedness drives the cross-species transmission of this clade of betaretroviruses. We found decisive evidence (BF$\geq$ 1000, (Jeffreys 1961) for a strong (median coefficient: -0.6; 95% HPD interval -0.5 to -0.7) effect of host-phylogenetic relatedness which was robust to the eight independent down-samplings from host-monophyletic clades. To prevent the possibility of geographic proximity - which correlates with phylogeny - from confounding the phylogenetic effect we also included the number of biogeographic zones in which both host species co-occur as a putative explanatory variable in the generalised linear model. We found a statistically unstable, weaker (at times nonexistent) effect for host-biogeographic overlap. The relative effect sizes and the statistical instability of the host-biogeographic effect suggest that the host-phylogenetic effect has not been confounded by host-geography, that there is no strong evidence for host-range driving cross-species transmission, and that more closely related species have higher rates of cross-species transmission between them. Additionally, the host-phylogenetic effect was reversed in three independent shuffles of the discrete states of the tip nodes (BF$\geq$ 1000, median coefficients 1.2, 95% HPD intervals 1 to 1.5), showing that the effect is not the result of an interaction between the prior and the number of times each host appears in the virus phylogeny.

It is unlikely that the host-phylogenetic effect is the result of orthology: most orthologues in our

dataset were easily recognized, as there were multiple copies in the ancestral genome which were inherited by both descendant species, creating clusters of pairs of orthologous viruses present in two closely related species, such as the mouse and rat. We also searched for orthologues in our dataset by pairwise blasting the contigs from related species on which closely related viruses were found. We found many orthologues in this clade, such as those between the mouse and rat, between the three *Myotis* species, and between the ungulates. Because of the difficulty of proving orthology in short sequences from incomplete assemblies we took a conservative approach. The oldest orthologues we found were between the *Tursiops truncatus* and *Lipotes vexillifer*, species that split approximately 40 million years BP. We collapsed the mammalian phylogeny at that time to prevent these and younger orthologues from influencing the analysis (fig. 6.5).

The strong effect of host-phylogenetic relatedness might be because phylogenetic relatedness between mammals correlates with immunological similarity (Collier and O'Brien 1985), which could make it easier for pathogens to jump between closely related hosts. The reduction in fitness a pathogen experiences when it first infects a new species has been conceptualised as a 'fitness valley' (Kuiken et al. 2006), it seems that for betaretroviruses this valley may be shallower between more closely related hosts. These findings are not entirely unprecedented: host phylogenetic relatedness driving cross-species transmission has previously been shown in bats (Streicker et al. 2010; Faria et al. 2013), primates (Davies and Pedersen 2008), and drosophila (Longdon et al. 2011), as well as in a broader analysis of mammal-virus associations (Olival et al. 2017). In drosophila, experimental evidence was found for two distinct effects of host phylogenetic relatedness: closely related species are more likely to be equally vulnerable to a virus, and species more closely related to a virus' original host are more vulnerable to that virus (Longdon et al. 2011). Because of the much larger timescales in our study we were unable to distinguish between these causes of the host-phylogenetic effect. The study on primates was the only one which did not find a statistically significant effect for host phylogenetic relatedness on the cross-species transmission of viruses (Davies and Pedersen 2008), though it did find such an effect for other parasites. This could be because the study used absence/presence data for each virus 'species', which is much less sensitive than determining the phylogenetic distance between virus sequences found in different hosts as we have done here. The work on bat rabies, which pioneered these phylogenetic methods, showed host phylogeny to be

predictive of cross-species transmission of rabies in bats (Streicker et al. 2010; Faria et al. 2013). All but one of these previous studies (Olival et al. 2017) considered much smaller groups of host organisms, as well as more tightly limited viral clades. We have shown that host-phylogenetic relatedness can drive cross-species transmission not just within clades of mammals, but between all mammals and over much larger timescales.

6.2.2 Inferring Betaretroviral Cross-Species Transmissions Between Mammals

Constructing viral networks with BEAST's phylogeographic model allows us to infer which pairs of species have – directly or through a hidden intermediate host –exchanged viruses in their evolutionary history. The connections within this network do not necessarily represent single transmission events but are an amalgamation of all phylogenetic evidence for cross-species transmissions between a particular species pair. BEAST's phylogeographic model yields both the statistical importance of every possible connection between the mammals included in the analysis, as well as the inferred number of transmission events associated with that connection. The importance of a species pair to the model is given as a Bayes Factor, representing the statistical support for that connection. We followed the standard nomenclature for Bayes Factors and only report connections (fig. 6.3) for which there was strong evidence (BF$\geq$ 10, (Jeffreys 1961)). To make the phylogeographic analysis computationally tractable it was necessary to downsample by selecting one sequence per host-monophyletic clade in the virus phylogeny (see methods), the Bayes Factors were robust to this (connections with BF$\geq$ 10 in runs derived from two maximally different sampling from the host-monophyletic clades had a Spearman's $\rho = 0.67$, $p \leq 10 - 7$).

Bayes Factors on connections in the viral network are not necessarily related to the number of transmission events in the phylogeny (Lemey et al. 2009; Faria et al. 2013): two closely related viruses with very short branch-lengths found in different species create a strongly supported connection in the viral network but are evidence for only one transmission event, whereas a mixed cluster of closely related viruses from two species can be evidence for many transmissions. To discover the number of actual cross-species transmission events between each species-pair in the posterior phylogenies we used the robust counting procedure, an enhancement to the BEAST phylogeographic model that generates a posterior estimate of the number of transitions between the discrete states on the phylo-

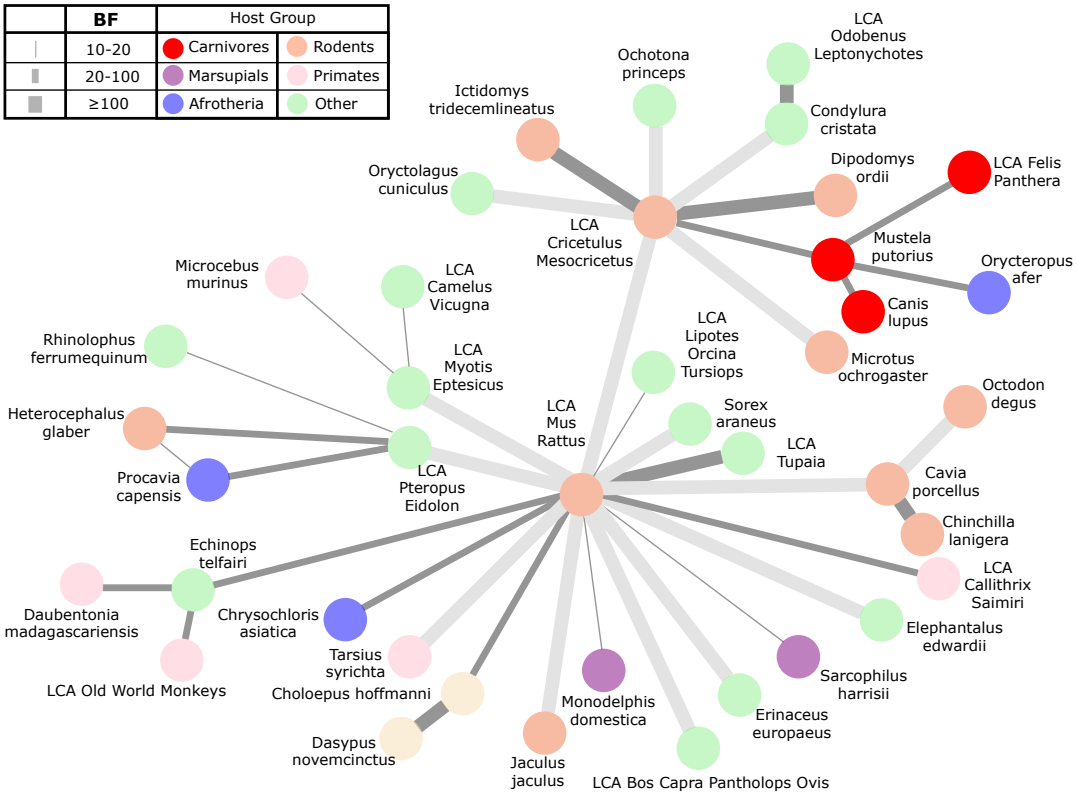


Figure 6.2: Nodes in this network represent mammals and the lines between them represent transmissions for which there is strong evidence (BF$\geq$ 10). Nodes labelled LCA represent the Last Common Ancestor of the named species. The width of an edge shows how strongly that particular transition is supported by the data but does not give any information regarding the rate of that transmission (Faria et al. 2013). Grey edges were also recovered from the tip-shuffled phylogeography with a BF$\geq$ 10, and may be due to an interaction of the prior and the number of times hosts are represented in the underlying virus phylogenies. Edge lengths are not meaningful.

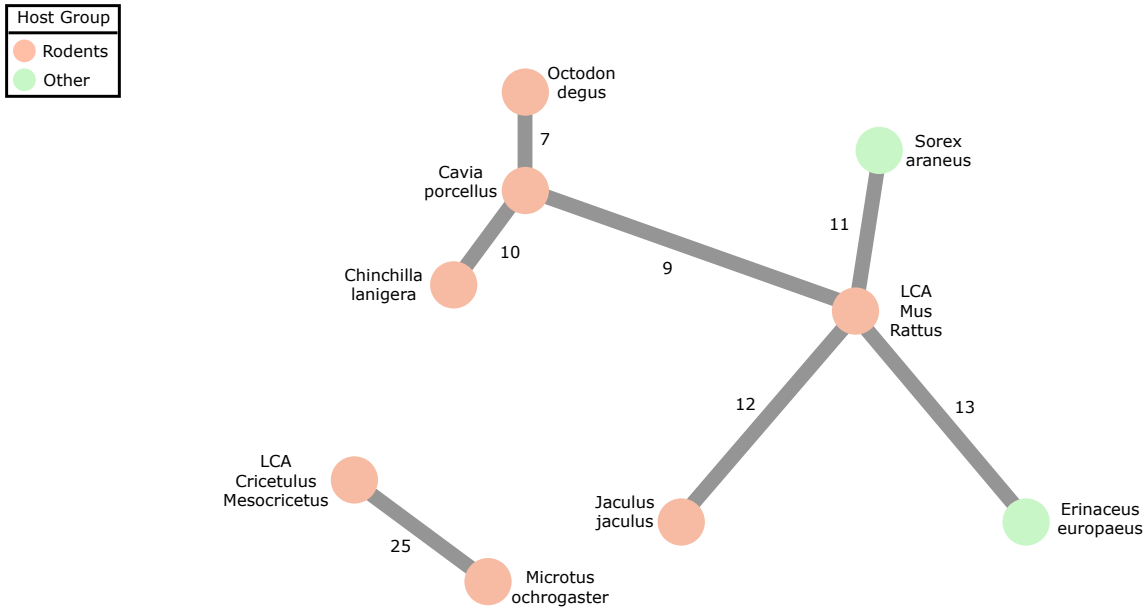


Figure 6.3: Nodes in this network represent mammals, and the lines between them represent robust counts of cross-species transmission events. All edges are annotated with the associated robust count according to BEAST. We only show robust counts where 0 is not included in the 95% HPD.

geny (Minin and Suchard 2008; O'Brien, Minin and Suchard 2009). This procedure recovered seven connections with 95% Highest Posterior Density (HPD) intervals excluding 0 (fig. 6.4), mostly between the rodents, confirming their important role in the transmission of betaretroviruses. However, many of the connections for which there is strong statistical support (fig. 6.2) have 95% HPDs that do not exclude 0. This could be because the phylogeographic approach is limited by the large number of host species that are highly interspersed in the viral phylogeny (fig. 6.6). This makes it difficult to consistently impute the host-state of internal nodes of the viral phylogeny, which in turn makes it difficult to consistently assign a state-change to one particular species-pair. This is a common limitation of the robust counting approach, which is reflected in the literature (Auguste et al. 2010; Faria et al. 2013, 2014; Nunes et al. 2014), where most papers report the posterior estimate without including the 95% HPD intervals.

6.2.3 Centrality of the Rodents

Rodents have long been thought to be important in the transmission of disease, having been implicated in zoonoses of hantavirus (Vapalahti et al. 2003), cowpox (Ninove et al. 2009) and the bubonic plague (Keeling and Gilligan 2000). Perhaps because they are a species-rich order, live in colonies

and have short lifespans, factors which could be related to disease transmission. Rodents have also been suggested as possibly important for the cross-species transmission of betaretroviruses (Martin, Kabat and Tristem 2003; Baillie et al. 2004) and gammaretroviruses (Hayward, Grabherr and Jern 2013). We use the viral network to statistically show that rodents are involved with the transmission of betaretroviruses to many mammals, not just to humans (fig. 6.2). The common ancestor of mice and rats, in particular, holds the most central position, with strong evidence (BF $\geq$ 10) for 18 links to other animals of various kingdoms, including bats, primates, artiodactyla, afrotheria, marsupials, several different rodents and various other mammals. The Last Common Ancestor of the Chinese hamster (*Cricetulus griseus*) and the golden hamster (*Mesocricetus auratus*) holds the second-most central position, with six connections to other mammals from assorted orders. Some of these connections have also been recovered from the tip-shuffled phylogeographic analysis, where we randomly shuffled the host assigned to each viral sequence, showing that those connections may be the result of an interaction between the prior and the number of times hosts are represented in the underlying virus phylogenies. However, even though some individual edges may be the result of such an interaction, even in this tip-shuffled viral network (fig. 6.7) the rodents retain a central position; probably due to the interspersed nature of rodent ERVs in the betaretroviral phylogeny.

6.2.4 Absence of Bats from the Viral Network

Bats have been shown to be natural reservoirs of SARS (Li et al. 2005), Ebola (Leroy et al. 2005), and rabies (Streicker et al. 2010; Faria et al. 2013). We found endogenous retroviruses in 9 bat species from 6 different genera in this clade of betaretroviruses. However, despite this extensive sampling bats do not occupy a central position in the betaretroviral network, indicating that bats have probably not been important reservoirs for the betaretroviruses over the last 40 million years. Although bats have been shown to be important hosts in the transmission of various viruses, and have received much attention as hosts of emerging diseases (Calisher et al. 2006; Leroy et al. 2009; Chan et al. 2013), they may not be universally important viral hosts.

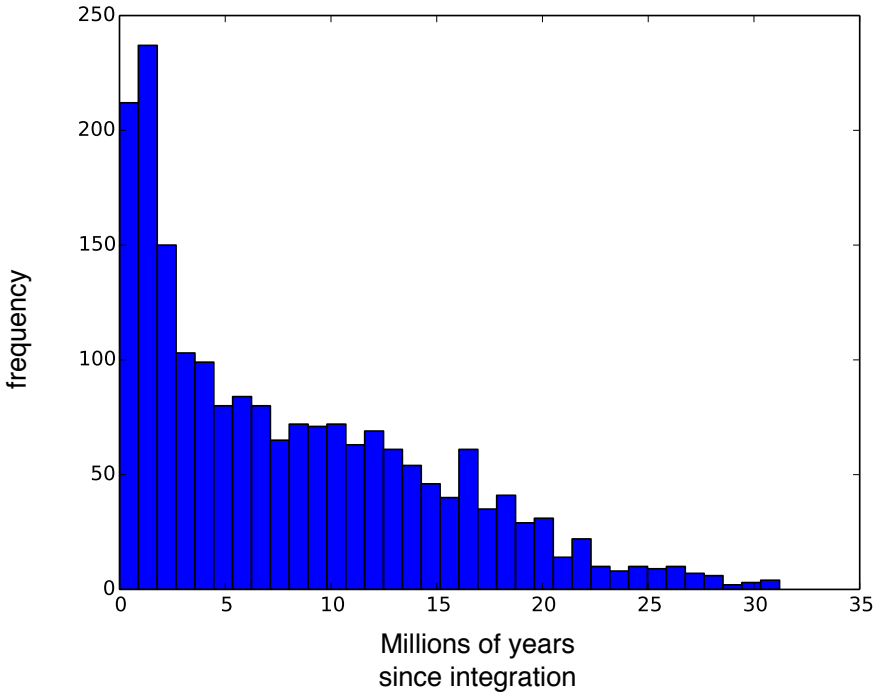


Figure 6.4: Histogram of LTR dates in our clade from all species, showing the majority of integrations happened before the evolution of humans.

6.2.5 Intercontinental Viral Flow

Our search revealed endogenous betaretroviruses in all continents except Antarctica, as well as cross-species transmissions between the ancestors of animals indigenous to different continents, such as the connection in the viral network between the Ictidomys tridecemlineatus (thirteen-lined squirrel, North America) and the last common ancestor of Cricetulus and Mesocricetus (hamsters, native to the Palearctic). Such non-anthropogenic long-range movement of viruses has been described in influenza (Koehler et al. 2008) and paramyxovirus (Ramey et al. 2013), probably mediated by migratory birds. Recently birds have been found to carry a group of recombinant ERVs that share some features with betaretroviruses (Henzy et al. 2014), which would provide a convenient explanation for the intercontinental flow of betaretroviruses. However, the TgERV recombinants have a *pro-pol* region clusters more closely related to the alpharetroviruses, such as the Avian Leukosis virus, than the betaretroviruses. Therefore these recombinants cannot be responsible for their intercontinental flow of this clade of these betaretroviruses, which have a typically betaretroviral *pol*. Polymerase genes from this betaretroviral clade are absent from the bird genomes currently in the wgs database,

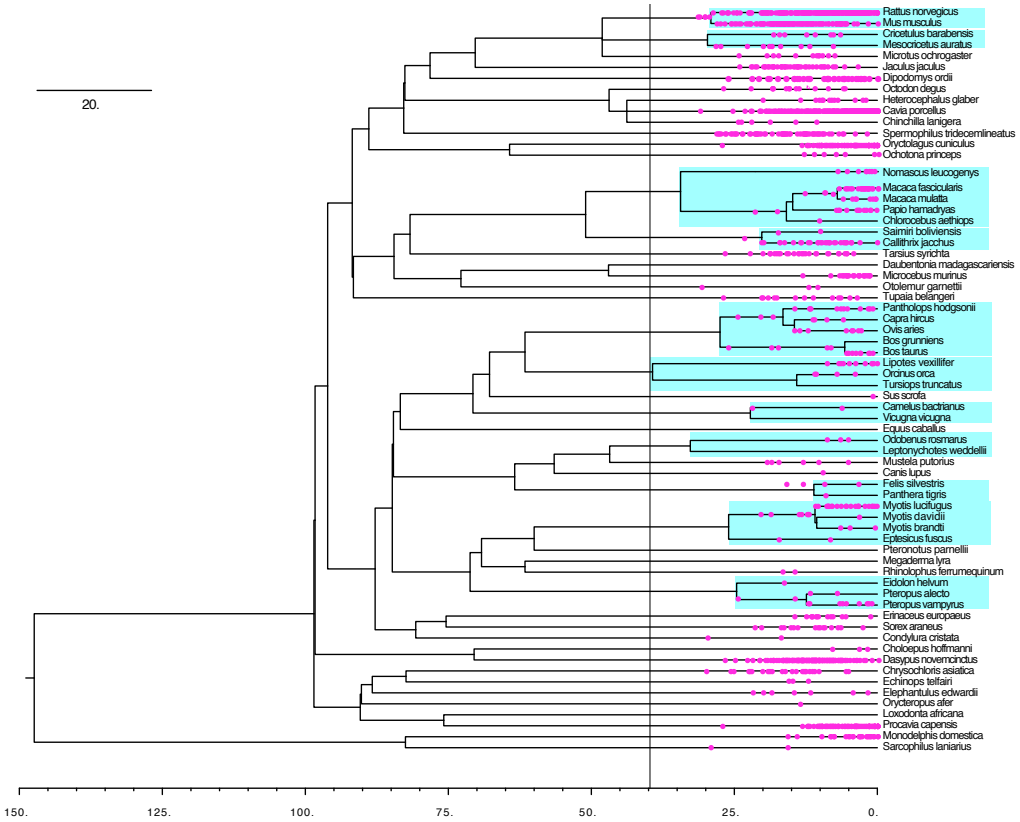


Figure 6.5: A phylogeny of mammals annotated with LTR dates (shown in magenta). Species that diverged less than 40MYA and may have orthologous ERVs are shown in blue, these species were folded into their last common ancestor in the Bayesian Analyses.

which means that birds are very unlikely to be responsible for the intercontinental spread. The influenza virus has also been reported in marine mammals with panarctic distributions (Fereidouni et al. 2014), which are in the wgs database, and could represent an animal pathway for intercontinental viral flow. However, none of the sequenced marine mammals occupies a central position in the viral network (fig. 6.2), as would be expected if they were the cause of the intercontinental flow of this betaretroviral clade.

More generally, influenza and paramyxoviruses transmit between hosts much more frequently than betaretroviruses like SRV, which can persist in its host for much longer periods (Montiel 2010). The much faster inter-host dynamics of influenza and paramyxoviruses could allow their viral lineages to move across larger phylogenetic and geographic distances with relative ease (Grenfell et al. 2004; Kuiken et al. 2006), perhaps because faster inter-host dynamics and higher reproductive rates are more common in airborne viruses. Viruses which cannot spread through aerosolization may be less capable of intercontinental movement, illustrated by the near-absence of Ebola beyond West Africa (Gire et al. 2014), and the limited flow of the retroviruses HIV-2 and HIV-1 non-M groups beyond Africa (McCutchan 2006), especially when compared to airborne viruses such as SARS (Min et al. 2004) and H1N1 swine-origin influenza (Smith et al. 2009). In contrast, retroviruses are often blood born (Snyder and Dodd 2001) and persist in their hosts (Feorino et al. 1985), which would appear to make them poorly suited to intercontinental movement – making their frequent intercontinental movement all the more interesting.

The connection in the viral network between the mouse and the Tasmanian devil is especially remarkable due to the history of the Australian continent, which has been isolated for the last 100Ma as it slowly drifted north from the remainder of Gondwanaland (Veevers 2006). Today the intercontinental transmission of viruses is often mediated by man, such as the introduction of the Rabbit Hemorrhagic Disease Virus (RHDV) to Australia and New Zealand (Kovaliski 1998). Anthropogenic means are also a likely explanation for the recent appearance of KoRV in Australia (Tarlinton, Meers and Young 2006), which may have crossed into Koala's in a zoo (Martin et al. 1999) but in any event started endogenising between 120 and 50.000 years ago (Ávila-Arcos et al. 2013; Ishida et al. 2014). A period that largely overlaps with the human presence on the Australian continent (O'Connell and Allen 2004). This contrasts sharply with the betaretroviruses in the Tasmanian

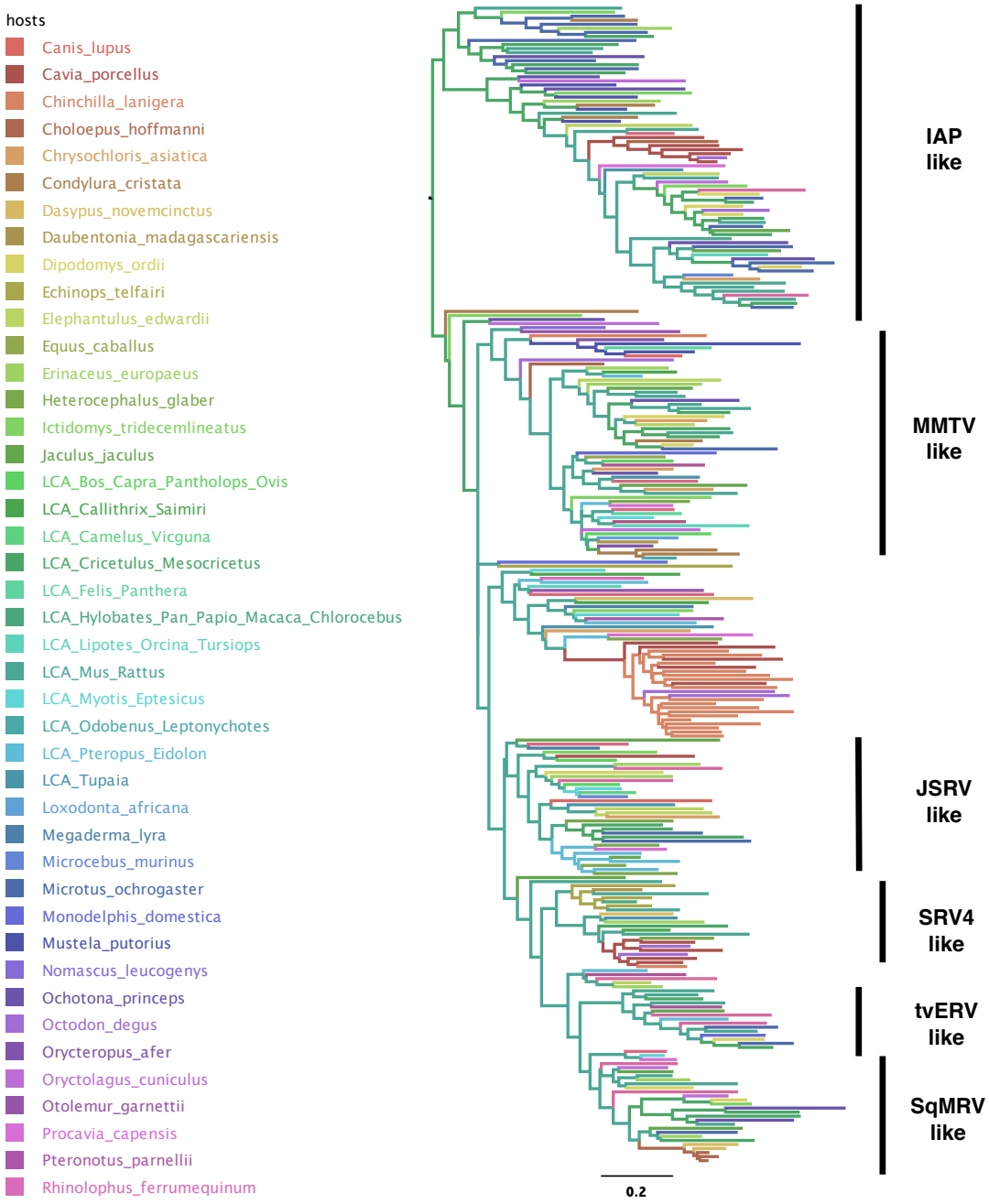


Figure 6.6: Phylogeny of betaretroviral clade colored by host species and annotated with most similar exogenous retrovirus per clade, showing the interspersed nature of the hosts throughout the entire phylogeny.

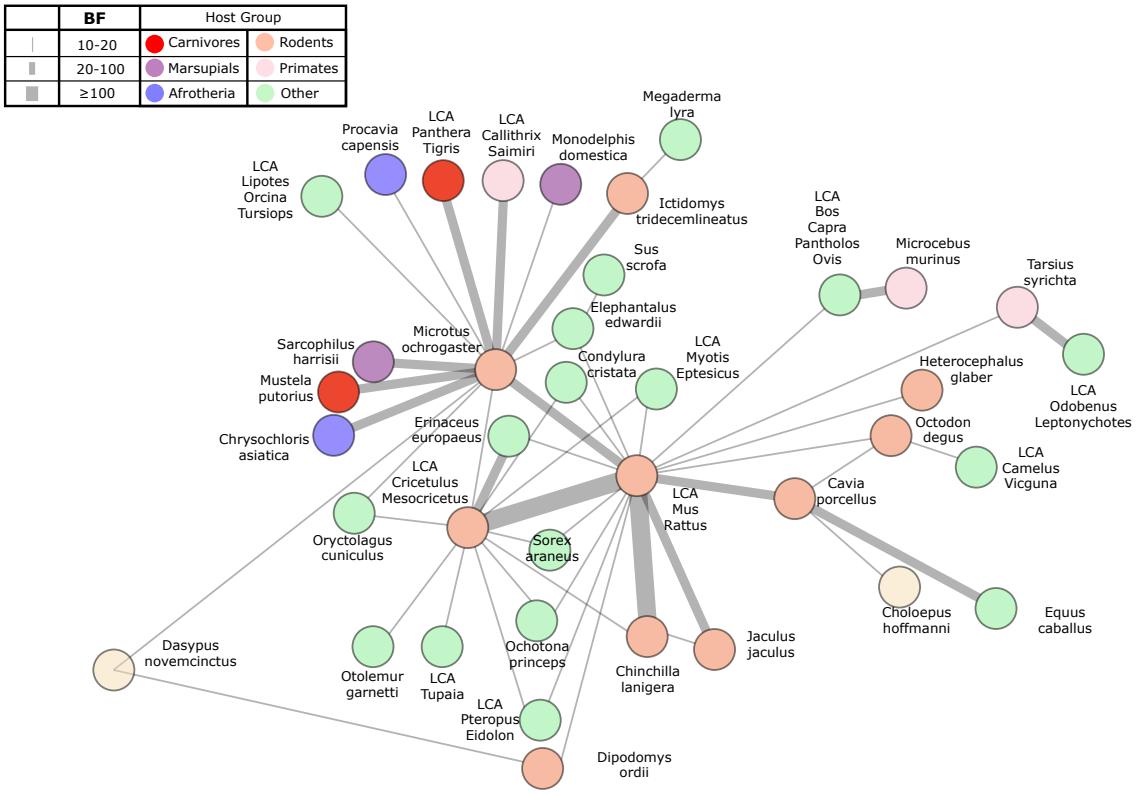


Figure 6.7: Viral network generated from a tip-shuffled dataset: the individual edges in this network are not correct, but the centrality of the rodents is maintained.

devil: the earliest dated endogenous betaretrovirus integrated approximately 30 MYA, long before the evolution of man. More generally, the vast majority of dated betaretroviral ERVs in all species integrated before the evolution of humans (fig. 6.6), which shows that the cosmopolitan nature of the betaretroviruses cannot be due to anthropogenic mechanisms alone, such as infected mice that were accidentally transported by ship or plane.

6.3 Concluding Remarks

We explicitly modelled cross-species transmission of betaretroviruses and revealed ancient routes of viral flow, showing that rodents occupy a central position in the viral network and demonstrating that host-phylogeny is an important driver of cross-species transmission. Limitations of the data mean that there may have been other connections that we have not been able to detect, such as those involving species that have not yet been sequenced, have gone extinct, or where the virus has not yet endogenised. One example of the latter is the squirrel monkey retrovirus, a virus which is closely related to the endogenous viruses in the sloth and the armadillo, but which has not yet endogenised in the squirrel monkey. Sometimes long branches on the phylogenetic tree can hint at the presence of hidden, unsequenced, intermediate hosts. One example is the interclass transmission of the reticuloendotheliosis virus from mammals to birds, where the diversity of mammalian gammaretroviruses makes it difficult to pinpoint the mammalian species where the virus originated (Martin et al. 1999). In contrast: very short branch lengths between viruses found in different species suggest that there were no intermediate hosts because short branches mean little time passed between the two integration events. This is exemplified by the transmission between the sloth and the armadillo, where the short branch lengths, coupled with the host's geographical overlap, suggesting that there was no intermediate host. Despite these limitations, the presence of closely related ERVs in two mammalian hosts means there must have been a cross-species transmission between those two mammals (Herniou et al. 1998; Martin et al. 1999).

In recent years much has been made of the danger of emerging zoonotic diseases (Jones et al. 2008; Cutler, Fooks and Van Der Poel 2010). When SARS first emerged in humans, several groups started a search for the animal reservoir, during which it was first detected in civet cats and racoons (Guan et al. 2003), before bats were found to be the likeliest reservoir species (Li et al. 2005). Discovering

the reservoir species of an emerging infection is crucial so measures can be taken to minimise further contact with humans and prevent additional zoonoses, such as discouraging the consumption of bushmeat in areas affected by Ebola (Frieden et al. 2014). The methods described in this paper could be used to inform the search for such an intermediate host. To give an example: the centrality of rodents in the betaretroviral network and the importance of host phylogenetic relatedness to betaretroviral cross-species transmission suggest two targets for surveillance should a pathogenic betaretrovirus emerge: rodents and species closely related to the newly infected host.

6.4 Methods

6.4.1 Querying the WGS Database

Because the rapid evolution of betaretroviruses results in a high amount of divergence between amino acid sequences of different betaretroviruses, homology-based search algorithms such as BLAST cannot find all ERVs with a single query. To ensure we captured all members of the diverse clade of interest we used a recursive strategy, which we seeded with JSRV (NC001494.1, starting from the PVW motif) as a tblastn query against the NCBI whole genome sequencing (WGS) database downloaded on October 20th 2013 (Benson et al. 2013). To avoid unnecessarily using multiple insertions of the same virus as tblastn queries in the next step of the search we extracted a single sequence for every clade in which all ERVs are from the same host-species (fig. 6.8) using Python and the ete2 phylogenetics library (Huerta-Cepas, Dopazo and Gabaldón 2010). We also performed a deep tblastn search using JSRV as a query (`NC001494.1`), allowing up to 50,000 hits.

During each iteration of the recursive search strategy, we extracted 8000 bp on both sides of the tblastn hits and used the exonerate protein2dna:bestfit function (Slater and Birney 2005) to pairwise align the unknown retroviral *pol* with each of the seven reference betaretrovirus *pol* sequences (tbl. 6.9). Exonerate forced the sequence into frame with each reference betaretrovirus *pol* sequence by removing nucleotides forming the incomplete codon responsible for the frameshift, and generated a score for every pairwise alignment. For each endogenous retroviral *pol* we only kept the frameshift-fixed sequence with the highest scoring alignment. We then translated the frameshift-fixed nucleotide sequence with BioPython (Cock et al. 2009), and made a pairwise amino acid alignment using

the Needleman-Wunsch algorithm between the translated endogenous *pol* sequence and the reference sequence with which it had the highest exonerate score. By discarding codons coding for amino acids that caused gaps to be inserted into the reference sequence, we can extend the original reference alignment indefinitely without introducing additional gaps. This method is necessary because multiple alignment algorithms, including muscle (Edgar 2004) and clustal omega (Sievers et al. 2011), introduced too many gaps and produced ambiguous alignments.

Exonerate's protein2dna:bestfit function can sometimes include non-homologous sequences in the extracted sequences, especially at the edges (exonerating a partially deleted ERV with an incomplete *pol* can leave host DNA from the insertion site in the alignment). An additional step was needed to remove these; non-homologous sequence sites are discordant with the rest of the alignment, and this discordancy can be recognised algorithmically. For every cell in the alignment matrix, we replaced the amino acid with the relative frequency of that amino acid in that column. This means that a discordant amino acid, which is much more likely to occur in a non-homologous sequence in an otherwise high-quality alignment, will be assigned a low value while an amino acid which frequently occurs in that column will be assigned a high value. We then normalised this discordancy matrix by dividing every cell by the mean value for its column and smoothed it using a moving average function on the rows (mirrored at the edges, window = 35). We sequentially removed amino acids starting from both edges of each sequence until we reach the first amino acid with a moving average above the manually chosen threshold value of 0.7. We further curated amino acid alignment of ERVs by manually removing highly variable columns. The resulting alignment was used to build a tree with FastTree. We extracted the clade containing JSRV, SqRV, SRV and the IAPs using ete2. We then extracted a single sequence per largest possible host-monophyletic clade (fig. 6.8) using a python script, and used these sequences as queries for the next iteration's tblastn step. Five iterations of this recursive search were performed.

We combined all the reference-aligned sequences found by the recursive searches and the deep search into a single FASTA file, excluding duplicates. Several animals represented in wgs have been sequenced multiple times, and for those, we chose the sequencing project with the highest N50 and discarded sequences derived from the others. We curated the combined amino acid alignment by manually removing highly variable columns, and applied the edge-cleaning algorithm in the same

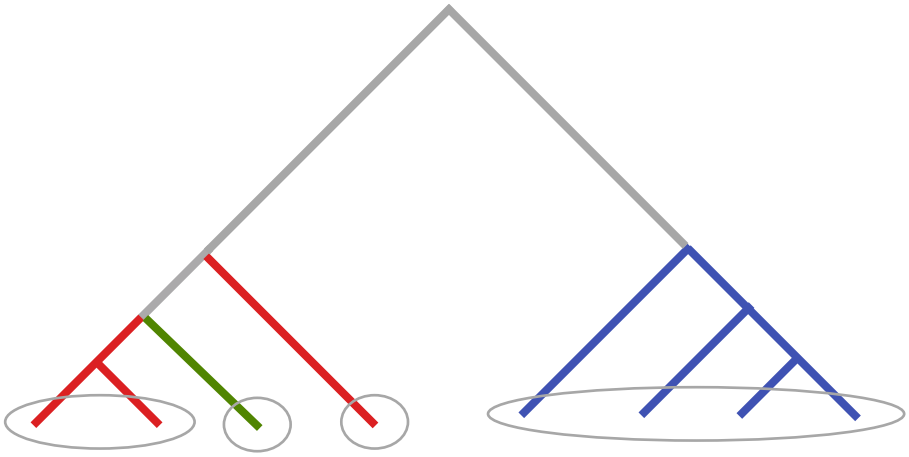


Figure 6.8: A hypothetical phylogeny of endogenous retroviruses with colours representing the species in which a virus was found. All largest possible host-monophyletic clades are circled, even those containing only one ERV.

way as during the search. Additionally, we assigned each sequence in the alignment a discordancy score by summing over the rows of the normalised discordancy matrix and removing the 10% most discordant sequences, which all contained non-homologous regions. We used FastTree to build a tree from this alignment and extracted the clade containing JSRV, MMTV, SRV and the IAPs using ete2. We then used FastTree to build a tree from this clade-only alignment and used ete2 to randomly select one sequence per host-monophyletic clade (fig. 6.8). We manually curated the resulting alignment by either removing sequences that had internal non-homologous regions, or replacing them with another sequence from the same host-monophyletic clade. As expected the vast majority of sequences with discordant regions were singletons. We also removed two sequences that were obvious contaminants: `AACZ03159228.1` from chimpanzee (*Pan troglodytes)* and `CABG01000502.1` from the blood fluke (*Schistosoma mansoni)*. The chimpanzee sequence consists of a short contig 99% identical to Mouse IAP, and is the only such sequence in the chimpanzee genome; IAPs have not been reported in chimpanzees otherwise. The *Schistosoma mansoni* sequence is derived from individuals cultured in *Mesocricetus auratus*, and is 99% identical at the nucleotide level to an IAP from that species. The resulting amino acid alignment was used for the phylogeographic analyses.

Table 6.9: Names and accessions of viral sequences added to the amino acid alignment after the search.

Virus Name	**Accession**
Mouse Mammary Tumor Virus	`AF228552.1`

Virus Name	**Accession**
Jaagsiekte Sheep Retrovirus	`AAK38686.1`
Squirrel Monkey Retrovirus	`P03364.2`
Simian Retrovirus 4	`YP_003864102.1`
Human Endogenous Retrovirus K (I)	`Q9BXR3`
Mouse IAP	`P11368.1`

6.4.2 Detection of Orthologous Sequences

The presence of orthologues in our dataset would cause false positive transmissions to be detected between closely related species. To prevent this, we searched for orthologous ERVs by performing a pairwise blastn alignment on closely related sequences from closely related hosts in the FastTree phylogeny of all sequences from the clade of interest. We limit our orthologues search to singletons and two sequence host-monophyletic clades because larger host-monophyletic clades are likely to be the result of repeated insertion or retrotranspositions than the result of host-divergence (fig. 6.1). We compare the blastn alignment with a blastx search of both sequences against viral sequences in nr, if the blastn aligned region isn't largely covered by retroviral blastx hits it may be an orthologue. The oldest orthologues we discovered are common to *Lipotes vexillifer* (Baiji) and *Orcinus orca* (killer whale), species that diverged approximately 40 MYA (`AUPI01035245.1` and `ANOL02030455.1`, amongst others). Because orthologues can be difficult to detect in short contigs from poorly assembled genomes, we took a conservative approach and assigned all viruses in hosts which diverged less than 40 MYA to their last common ancestor for the phylogeographic analyses (i.e., mouse and rat became LCA_Mus_Rattus).

6.5 Viral Networks

We inferred the evidence for and magnitude of cross-species transmissions using BEAST 1.8.0 and BEAGLE 2 (Drummond et al. 2012) and MrBayes 3.2.2 (Ronquist et al. 2012). BEAST's phylogeographic analysis models host species as a discrete state evolving along a phylogeny (Lemey et al. 2009) as a Continuous Time Markov Chain, and allows for the use of Bayesian Stochastic Search Variable Selection (BSSVS) to assign Bayes Factors to determine the evidence strength for putative cross-species transmissions, as well as robust counting to estimate the number of virus transmissions between two hosts (Minin and Suchard 2008; Talbi et al. 2010; Faria et al. 2013).

First, we used MrBayes to infer a posterior distribution of trees, performing three independent MCMC runs (using three hot chains each, the RTREV+F+G model, with the TBR operator turned off to avoid bug #1599) for at least 20M generations and confirmed convergence in TRACER. After discarding a 25% burnin, we down-sampled (1 tree per 25000 steps) the runs and combined them

into a single file containing 1800 trees, which was used as the empirical tree distribution for the BEAST discrete phylogeographic analysis with a symmetric rate matrix. To prevent the long tail of the host transition matrix prior from unduly influencing the analysis, it was necessary to run BEAST with a bounded prior $[0, 3]$ on the rates of the CTMC host transition matrix. To prevent the chains from getting stuck in local optima early in the run, we limited the rate-scaling operator to propose only one rate change to the host transition matrix. We verified whether the bounded transmission rates prior was unduly influencing the result by rerunning the analysis with a prior bounded between $[0, 6]$, which gave very similar results. We performed ten independent runs of the BEAST phylogeographic analysis, each for over 20M steps and confirmed convergence in TRACER. We used LOGCOMBINER to combine the log files and to remove 25% burnin per chain, we then calculated the Bayes factors resulting from the BSSVS procedure using an R script and visualized transmissions with a BF $\geq$ 10 (fig. 6.2) and transmissions with robust counts of which the 95% HPD did not include 0 (fig. 6.4) using the NetworkX library (Hagberg, Schult and Swart 2008). To test the robustness of the statistically strong connections in the viral network we performed a BEAST run using different sequences from each host-monophyletic clade where possible. As a sensitivity analysis, we shuffled the discrete states of the tips and generated a new viral network, allowing us to see which viral network connections may have been caused by interactions between the prior and the number of times a host occurs in the underlying phylogeny.

6.6 Beast's Generalized Linear Model

We also used BEAST's built-in generalised linear model (Lemey, Rambaut and Bedford 2012) to determine which characteristics drive the cross-species transmission of betaretroviruses. BEAST's generalised linear model works using matrices containing the putative explanatory variables, one for each pairwise combination of host-species. Similar to the normal phylogeographic analysis BEAST will assign a Bayes Factor using a BSSVS procedure similar to the non-GLM model. The Bayes Factor shows the strength of evidence and a posterior regression coefficient shows the importance of each explanatory variable. We use phylogenetic relatedness and host range overlap as explanatory variables. We approximate phylogenetic relatedness using time since divergence in MYA according to the mammalian phylogeny (Bininda-Emonds et al. 2007). We used the IUCN red list database

(IUCN 2008) to recover the native ranges of every host species, and used that to generate a predictor matrix containing the number of biogeographic realms (Olson et al. 2001) within which both species are present. For species that we collapsed into last common ancestor's to avoid orthologues, we used the superset of their native biogeographic zones. Because our dataset does not contain sufficient data to infer the direction of a transmission we use symmetric matrices for both our explanatory variables. Both explanatory variable matrices were standardised in natural log units according to the BEAST GLM tutorial (Nuno Faria, private correspondence). We ran three replicate BEAST chains for over 15M generations using the same empirical distribution of trees as in the regular phylogeographic analysis, used TRACER to confirm convergence, discard a 25% burnin and extract the results. To test the robustness of both putative explanatory variables we performed eight replicates of the generalised linear model phylogenetic analysis using independently sampled random representatives from each host-monophyletic clade. We also performed a sensitivity analysis by shuffling the discrete state of the tips three times and recalculating the Bayes Factors and coefficients after running the MCMC chains for at least 15M generations.

6.7 Dating Insertions

During the process of retroviral integration Long Terminal Repeats (LTRs) are created on either side of the integrating provirus. Because these LTRs are identical at the time of integration and proceed to independently accumulate mutations at the host rate of evolution their divergence can be used to date insertion events. We used LTRHarvest (Ellinghaus, Kurtz and Willhoeft 2008) to identify pairs of Long Terminal Repeats (LTRs) on sequences from our clade with an 8000bp buffer zone on either side of the *pol* blast hits. As an additional check, we only used LTR pairs when exonerate detected a *pol* between the two LTRs, if this was the case we excised the LTRs from the surrounding sequence and used them to make a Needleman-Wunsch pairwise alignment. We then calculated the evolutionary distance between the two LTRs using the K80 model (Kimura 1980) of nucleotide evolution, and converted the distance to MYA by halving it to compensate for the double counting of independent changes on both LTRs and dividing the halved distance by the host neutral rate of evolution (2.2×10^{-9} mutations per site per year) (Kumar and Subramanian 2002). The results of this analysis were then mapped onto a phylogeny of mammals (fig. 6.6) (Bininda-Emonds

et al. 2007).

Conclusions

Since HIV was discovered over thirty years ago (Gottlieb et al. 1981; Barré-Sinoussi et al. 1983), much effort has been spent developing a vaccine (McMichael and Haynes 2012). Unfortunately, all randomised clinical trials have failed except one, which may have shown a slight beneficial result (Flynn et al. 2005; Pitisuttithum et al. 2006; Rerks-Ngarm et al. 2009; Gilbert et al. 2011; Duerr et al. 2012; Gray et al. 2012; Hammer et al. 2013). More work appears to be needed on understanding the fundamental biology of the system before a successful vaccine can be designed.

In this thesis I have presented an infinitesimal fraction of that work, focussing on the CTL response against p24 Gag. This response is worth studying because it is associated with long-term non-progression to AIDS (Rinaldo et al. 1995; Migueles et al. 2000; Dinges et al. 2010; Goulder and Walker 2012). Furthermore, even in patients who are not long-term non-progressors, the number of targeted Gag epitopes is associated with lower viral loads and slower disease progression (Kiepiela et al. 2007).

Subtype-specific sites

In the first three chapters, I considered the HIV-1M subtype-specific sites; the current paradigm is that these differences between the consensus sequences of HIV-1M subtypes are the result of founder effects. In contrast, I used Chapter Two to show that these sites are strongly associated with the patients' HLA variants, and consequently, with the epitopes the infected cells can present, suggesting a role for CTL selection in subtype-diversification. Previous research suggests that viral escape of the epitope processing machinery is the molecular mechanism for these effects (Tenzer et al. 2009, 2014).

In Chapter Three, I went on to show that as the United States HIV-infected population became

increasingly African-American over the last two decades, the amino acids on the subtype-specific sites became less typical of HIV-1M subtype B, and appeared to become more similar to HIV sequences found in Africa. It is interesting how a relatively small change in the ethnic demographics could still lead to an observable change in amino acid frequency. Since African-Americans today continue to suffer a disproportionate number of new HIV infections, the frequencies of non-subtype B consensus amino acids are likely to continue increasing in the future.

This adaptation to African-American HLA variants might result in a less vigorous anti-HIV CTL response which could lead to higher viral loads and perhaps a more easily transmissible virus between individuals, as well as more rapid disease progression. These possible consequences are suggested by differences in CTL epitope recognition between Caucasians and African-Americans (Frahm et al. 2004). Such ethnic outcome discrepancies would not be unprecedented, as HLA-dependent CTL responses have been shown to vary considerably between ethnicities in influenza (Quinones-Parra et al. 2014). This continued adaptation to HLA variants common in African-Americans is doubly tragic given the already poor health outcomes in the African-American population.

Implications for the Evolutionary History of HIV

The current paradigm is that the different amino acids on HIV subtype-specific sites are the result of a selectively neutral process such as founder effects (Pybus and Rambaut 2009). Together the results from Chapter Two and Chapter Three suggest that subtype-diversification in Gag may have been influenced by CTL selection, possibly because the virus evades epitope processing (Tenzer et al. 2009, 2014).

The study reported in Chapters One, Two, and Three was limited by only considering five HIV-1M Gag subtype-specific sites, and also because these subtype-specific sites are similar between subtype B and subtype D. Subtypes B and D share a common ancestor around 1940 (Faria et al. 2014). A recent study has confirmed that HIV subtype B came to the Americas via Haiti. This paper reports on five expatriate Haitians who became infected in Haiti in the late 1970s, and were sequenced in the early 1980s. These patients already had B-like subtype-specific amino acids in their sequenced DNA, so it is unlikely that HLA variants common in Caucasians were the driving force in the 'creation' of

subtype B in the United States (Worobey et al. 2016).

Extensive sex tourism in Haiti has been reported throughout the 1970s (Girard 2005), which could mean that Worobey et al. (2016) sequenced five back migrations from the United States rather than indigenous Haitian sequences. However, there is no phylogenetic evidence for migration of strains beyond a single movement from Haiti to the United States, with the possible exception of B6.H6.83, which might be a back migration of HIV from the US to Haiti (Gilbert et al. 2007). Further sequencing efforts should give a clearer picture of the HIV-1M lineage exchange early in the epidemic. Efforts should be taken to ensure that sequences are obtained from asymptomatic patients, as CTL selection is released once the patient's immune system fails and AIDS develops (Iversen et al. 2006). The early Haitian sequences obtained by Worobey et al. (2016) are from AIDS patients, which suggests extra caution is appropriate in interpreting their subtype-specific sites.

Without more early sequence data from asymptomatic patients, it is impossible to know for sure what Gag subtype-specific sites were prevalent in Haiti during the early epidemic there. But given the current data, it is likely to have already been the subtype B consensus in a number of its subtype-specific positions as these are often shared between subtypes B and D. However, this does not diminish our results showing HLA associations with subtype-specific sites, and indeed raises the intriguing possibility that HIV subtype B was so successful in the Western world because it was 'pre-adapted' to HLA variants common in Caucasians. Furthermore, while there is at present little evidence of the HIV-1M subtype B virus being uniquely adapted to Caucasians, we have observed a tendency for subtype B to change away from the subtype-consensus when introduced to a population with African HLA variants. This was originally shown in the Carribean (Tenzer et al. 2014), and we have now shown it in the United States (Chapter Three).

Recombination and Vaccine Escape

In Chapter Four, I have shown that patients with beneficial HLA variants select for the outgrowth of HIV-1M recombinants, possibly because recombination allows for large fitness jumps and the large-scale spreading of escape and compensatory mutations. In Chapter Five, I showed how repeated recombinations could lead to HIV-1M outgrowths in the presence of beneficial HLA variants. If

viral recombination can abrogate the natural CTL response it is likely that a vaccine-boosted immune response can also be evaded. Vaccine escape has been documented in mumps (Greenland et al. 2012; Sane et al. 2014), hepatitis A (Pérez-Sautu et al. 2011) and hepatitis B (Huang et al. 2004), and recombination has been associated with vaccine escape in pneumococcus (Brueggemann et al. 2007).

Because recombination over large phylogenetic distances is a likely avenue towards HIV vaccine escape, patients exposed to many strains should be monitored closely as a vaccine-escaped HIV-1 strain could rapidly undo an otherwise effective vaccination campaign. Luckily, antiretroviral therapy could be used to suppress any vaccine-escaped HIV strains and keep them from spreading further. However, this will require an extensive monitoring program focused on high-risk individuals, as well as a reliable second-line care program.

Recommendations for Vaccine Development

My research suggests against vaccination using country-level consensus sequences. While these sequences are the most likely to match to most wild sequences, they are likely to be specially adapted to the HLA variants common in the host population. Therefore, they will have evolved to evade the epitope processing machinery (Tenzer et al. 2009, 2014). Similarly, any effort to immunize with natural HIV-1 sequences is unlikely to evoke effective CTL responses in the majority of the vaccinated population. Unfortunately, I cannot use the models to make a positive recommendation, as these have been exclusively trained on naturally occurring sequences that are well adapted to our immune systems.

Instead, we suggest experimenting with Gag-derived antigens while varying the amino acids on the subtype-specific positions (and other inter-epitope sites), to see which combinations generate the strongest immune response. Since subtype-specific amino acids are located between regions of high epitope density, but not usually within epitopes, any strong immune response produced by artificial sequences will also be triggered by naturally occurring sequences. This technique promises to be especially potent because much higher cell surface epitope densities are required to prime T cells and evoke an initial immune response (Faroudi et al. 2003) than are required for already primed T cells to recognise an infected cell (Purbhoo et al. 2004). Still, it will be key to examine processing

of naturally occurring HIV-1 sequences to ensure that the chosen epitopes are also presented, at least in low numbers, in the natural infection. Hopefully, this will allow an artifical vaccine insert that results in an unnaturally high epitope production to prime an epitope-specific CTL response that can recognize the low numbers of naturally occuring HLA-epitope complexes on the surface of infected cells if the vaccinee becomes infected with a naturally occuring HIV sequence.

Future work

Expanding on the Subtype-Specific Sites

While attempting to devise a negative control, preliminary analyses showed that the majority of synonymous sites with sufficient intra-subtype variation to be analysed carried nucleotide mutations typical of APOBEC$_3$G and APOBEC$_3$F. APOBEC$_3$G and APOBEC$_3$F are host immune factors which hypermutate the virus (Armitage et al. 2008, 2012). Effective CTL responses against Vif are likely to cause escape mutations which reduce the efficiency of vif, a process which could lead to associations between hypermutated sites and host HLA type. It would be an interesting exercise to build a model which attempts to distinguish between direct CTL selection and 'second-order' CTL selection driven by CTL escape mutations in *vif*.

Such a model should incorporate whether the amino acid codon includes an appropriate APOBEC substrate, because these substrates will all mutate in tandem. It then becomes a point of distinguishing between direct HLA effects and second order effects, where the fact that all the APOBEC sites mutate in tandem might be exploited (Armitage et al. 2012). There could also be a way to incorporate some kind of HIV sequence covariance model, as all hypermutations should, under this hypothesis, follow the same CTL escape mutations in Vif (Carlson et al. 2008; Poon et al. 2010).

One thing we have not done in this thesis, but which might be appropriate, is a dN/dS analysis (Hurst 2002). The dN/dS analysis allows for the detection of selection when the number of non-synonymous mutation is higher than expected given the number of synonymous mutations. Performing this on our Gag data (always limiting to one subtype) would be a useful extension of our research, and might give additional insights.

Understanding HIV Subtype Diversification

To truly understand subtype diversification it will be necessary to obtain early sequences. Two places of particular interest in the search for such sequences are Haiti and Belgium, as sequences from key time periods in the evolution of HIV might be available there. One challenge we faced in interpreting the data is that we do not know what the virus sequence was like before it was introduced to the United States. To resolve that it might be worth looking for archived samples in Haiti.

The Belgians were in the Congo from 1876 until 1960, a time when HIV-1 was starting to circulate (Faria et al. 2014). It is well known that the Belgians practiced all manner of unspeakable crimes, and it seems likely that some colonists ought to have become infected with HIV. The Belgians have several good hospitals specialising in tropical disease, and old samples may be available there. It would be worthwhile to search their archives and attempt to sequence HIV DNA from any tissue samples they may have stored, especially if clinical records associated with the sample are indicative of an HIV infection. The oldest currently known sequence was sampled in the Belgian Congo in 1959, and is located near the subtype B-subtype D ancestral node (Zhu et al. 1998).

It might also be worthwhile to search out patients of European ethnicity infected with non-subtype B HIV strains. We would predict that the subtype-specific sites in these patients would become more B-like over time.

Applications to Other Viruses

HIV-1M is not the only virus subject to immune selection, nor is it the only virus which can be categorised into subtypes. HIV-1M's high substitution rates and relatively recent cross-species transmissions a century ago followed by its introduction into the west 40 years ago, makes it an interesting model for studying virus subtype diversification and allows for the detection of immune driven subtype-diversification. Zika virus (ZIKAV), which has been circulating in humans since the 1950s has recently recently been introduced to the Americas (Wang et al. 2016). The South American epidemic shares a common ancestor, which dates the introduction of ZIKAV to South America between May and December 2013 (Faria et al. 2016). There is another strain of ZIKAV which is endemic to Africa. Like HIV, ZIKAV has presentable epitopes which trigger CTL lysing of infec-

ted host cells (Wen et al. 2017). It will be interesting to see if the geographically isolated strains will diversify, and whether the same CTL pressures on subtype-specific sites we observe in HIV also play a role in ZIKAV diversification. There are several projects gathering viral sequence data from South America, and it may be possible to extend one of those projects by gathering HLA data from the patients, which would enable us to apply the methods we have developed in Chapters One, Two, and Three to this emerging virus.

Citations

Abi-Rached L, Jobin M, Kulkarni S et al. The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science* 2011.

Abrahams M-R, Anderson JA, Giorgi EE et al. Quantitating the Multiplicity of Infection with Human Immunodeficiency Virus Type 1 Subtype C Reveals a Non-Poisson Distribution of Transmitted Variants. *Journal of Virology* 2009.

Acuna-Soto R, Calderon Romero L, Maguire JH. Large epidemics of hemorrhagic fevers in Mexico 1545-1815. *American Journal of Tropical Medicine and Hygiene* 2000.

Aiewsakun P, Katzourakis A. Marine origin of retroviruses in the early Palaeozoic Era. *Nature Communications* 2017.

Alizon S, Fraser C. Within-host and between-host evolutionary rates across the HIV-1 genome. *Retrovirology* 2013.

Allen TM, Altfeld M, Yu XG et al. Selection , Transmission , and Reversion of an Antigen-Processing Cytotoxic T-Lymphocyte Escape Mutation in Human Immunodeficiency Virus Type 1 Infection Selection , Transmission , and Reversion of an Antigen-Processing Cytotoxic T-Lymphocyte Escape Mutat. 2004.

Allison AB, Kohler DJ, Fox K a et al. Frequent Cross-Species Transmission of Parvoviruses among Diverse Carnivore Hosts. *Journal of Virology* 2013.

Altschul S, Gish W, Miller W. Basic Local Alignment Search Tool. *J Mol Biol* 1990.

Armitage AE, Deforche K, Chang C hao et al. APOBEC3G-induced hypermutation of human immunodeficiency virus type-1 is typically a discrete "all or nothing" phenomenon. *PLoS Genetics* 2012.

Armitage AE, Katzourakis A, Oliveira T de et al. Conserved footprints of APOBEC3G on Hypermutated human immunodeficiency virus type 1 and human endogenous retrovirus HERV-K(HML2) sequences. *Journal of virology* 2008.

Auguste AJ, Lemey P, Pybus OG et al. Yellow fever virus maintenance in Trinidad and its dispersal throughout the Americas. *Journal of virology* 2010.

Ávila-Arcos MC, Ho SYW, Ishida Y et al. One hundred twenty years of koala retrovirus evolution determined from museum Skins. *Molecular Biology and Evolution* 2013.

Baele G, Lemey P, Bedford T et al. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution*

2012.

Baele G, Li WLS, Drummond AJ et al. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Molecular Biology and Evolution* 2013.

Baeten JM, Haberer JE, Liu AY et al. Pre-exposure prophylaxis for HIV prevention: Where have we been and where are we going? *Journal of Acquired Immune Deficiency Syndrome* 2014.

Bailes E, Gao F, Bibollet-Ruche F et al. Hybrid origin of SIV in chimpanzees. *Science (New York, NY)* 2003.

Bailey JR, Williams TM, Siliciano RF et al. Maintenance of viral suppression in HIV-1-infected HLA-B*57+ elite suppressors despite CTL escape mutations. *The Journal of experimental medicine* 2006.

Baillie GJ, Lagemaat LN van de, Baust C et al. Multiple groups of endogenous betaretroviruses in mice, rats, and other mammals. *Journal of virology* 2004.

Bar KJ, Li H, Chamberland A et al. Wide variation in the multiplicity of HIV-1 infection among injection drug users. *Journal of virology* 2010.

Barré-Sinoussi F, Chermann JC, Rey F et al. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science (New York, NY)* 1983.

Bates D, Mächler M, Bolker B et al. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 2015.

Beatty LA, Wheeler D, Gaiter J. HIV Prevention Research for African Americans: Current and Future Directions. *Journal of Black Psychology* 2004.

Benn-Torres J, Bonilla C, Robbins CM et al. Admixture and population stratification in African Caribbean populations. *Annals of Human Genetics* 2008.

Benson Da, Cavanaugh M, Clark K et al. GenBank. *Nucleic Acids Research* 2013.

Betancourt MJ, Girolami M. Hamiltonian Monte Carlo for Hierarchical Models. *Arxiv preprint arXiv:13 120906* 2013.

Bhattacharya T, Daniels M, Heckerman D et al. Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science (New York, NY)* 2007.

Bibollet-Ruche F, Heigele A, Keele BF et al. Efficient SIVcpz replication in human lymphoid tissue requires viral matrix protein adaptation. *Journal of Clinical Investigation* 2012.

Bininda-Emonds ORP, Cardillo M, Jones KE et al. The delayed rise of present-day mammals. *Nature* 2007.

Blomberg SP, Lefevre JG, Wells JA et al. Independent contrasts and PGLS regression estimators are equivalent. 2012.

Borrow P, Lewicki H, Hahn BH et al. Virus-specific CD8+ cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. *Journal of virology* 1994.

Borrow P, Lewicki H, Wei X et al. Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus.

Nature Medicine 1997.

Bronke C. HIV escape mutations occur preferentially at HLA-binding sites of CD8 T-cell epitopes Corine. *AIDS* 2013.

Brooks JI, Rud EW, Pilon RG et al. Cross-species retroviral transmission from macaques to human beings Molecular diagnosis of syphilis : the Schaudinn-Hoffmann lymph-node For personal use . Only reproduce with permission from The Lancet Publishing Group . *Lancet* 2002.

Brueggemann AB, Pai R, Crook DW et al. Vaccine escape recombinants emerge after pneumococcal vaccination in the United States. *PLoS Pathogens* 2007.

Burke DS. Recombination in HIV: An Important Viral Evolutionary Strategy. *Emerging Infectious Diseases* 1997.

Burns CC, Gleason LM, Mozaffarian A et al. Sequence variability of the integrase protein from a diverse collection of HIV type 1 isolates representing several subtypes. *AIDS research and human retroviruses* 2002.

Butler MA, King AA. Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *The American Naturalist* 2004.

Calattini S, Betsem E, Bassot S et al. Multiple retroviral infection by HTLV type 1, 2, 3 and simian foamy virus in a family of Pygmies from Cameroon. *Virology* 2011.

Calisher CH, Childs JE, Field HE et al. Bats: Important reservoir hosts of emerging viruses. *Clinical Microbiology Reviews* 2006.

Carlson JM, Brumme ZL, Rousseau CM et al. Phylogenetic dependency networks: Inferring patterns of CTL escape and codon covariation in HIV-1 Gag. *PLoS Computational Biology* 2008.

Carlson JM, Du VY, Pfeifer N et al. Impact of pre-adapted HIV transmission. *Nature Medicine* 2016.

Carpenter B, Gelman A, Hoffman M et al. Stan: A Probabilistic Programming Language. *Journal of Statistical Software* 2016.

Carrington M, O'Brien SJ. The influence of HLA genotype on AIDS. *Annual review of medicine* 2003.

Chan JFW, To KKW, Tse H et al. Interspecies transmission and emergence of novel viruses: Lessons from bats and birds. *Trends in Microbiology* 2013.

Chin MPS, Rhodes TD, Chen J et al. Identification of a major restriction in HIV-1 intersubtype recombination. *Proceedings of the National Academy of Sciences of the United States of America* 2005.

Chong YL, Lam TTY, Kim O et al. Successful establishment and global dispersal of genotype VI avian paramyxovirus serotype 1 after cross species transmission. *Infection, Genetics and Evolution* 2013.

Cock PJ a, Antao T, Chang JT et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009.

Collett D. *Modelling binary data*. CRC press, 2002.

Collier G., O'Brien SJ. A molecular phylogeny of the Felidae: immunological distance. *Evolution*

1985.

Conticello SG, Harris RS, Neuberger MS. The Vif Protein of HIV Triggers Degradation of the Human Antiretroviral DNA Deaminase APOBEC3G. *Current Biology* 2003.

Cooper N, Thomas GH, Venditti C et al. A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. *Biological Journal of the Linnean Society* 2016.

Corneli AL, Deese J, Wang M et al. FEM-PrEP: adherence patterns and factors associated with adherence to a daily oral study product for pre-exposure prophylaxis. *Journal of acquired immune deficiency syndromes (1999)* 2014.

Crawford H, Matthews PC, Schaefer M et al. The hypervariable HIV-1 capsid protein residues comprise HLA-driven CD8+ T-cell escape mutations and covarying HLA-independent polymorphisms. *Journal of virology* 2011.

Curran JW, Jaffe HW, Hardy AM et al. Epidemiology of HIV infection and AIDS in the United States. *Science* 1988.

Cutler SJ, Fooks AR, Van Der Poel WHM. Public health threat of new, reemerging, and neglected zoonoses in the industrialized world. *Emerging Infectious Diseases* 2010.

Dahirel V, Shekhar K, Pereyra F et al. Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. *Proceedings of the National Academy of Sciences* 2011.

Davies TJ, Pedersen AB. Phylogeny and geography predict pathogen community similarity in wild primates and humans. *Proceedings Biological sciences / The Royal Society* 2008.

Delgado E, Ampofo WK, Sierra M et al. High prevalence of unique recombinant forms of HIV-1 in Ghana: molecular epidemiology from an antiretroviral resistance study. *Journal of acquired immune deficiency syndromes* 2008.

Dempster A, Laird N, Rubin D. *Maximum likelihood from incomplete data via the EM algorithm.*, 1977:1–38.

Dinges WL, Richardt J, Friedrich D et al. Virus-specific CD8+ T-cell responses better define HIV disease progression than HLA genotype. *Journal of virology* 2010.

Doherty PC, Zinkernagel RM. A biological role for the major histocompatibility antigens. *Lancet (London, England)* 1975.

Dowling WE, Kim B, Mason CJ et al. Forty-one near full-length HIV-1 sequences from Kenya reveal an epidemic of subtype A and A-containing recombinants. *AIDS (London, England)* 2002.

Draenert R, Le Gall S, Pfafferott KJ et al. Immune selection for altered antigen processing leads to cytotoxic T lymphocyte escape in chronic HIV-1 infection. *The Journal of experimental medicine* 2004.

Drummond AJ, Ho SYW, Phillips MJ et al. Relaxed phylogenetics and dating with confidence. *PLoS Biology* 2006.

Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology* 2007.

Drummond AJ, Rambaut A, Shapiro B et al. Bayesian coalescent inference of past population

dynamics from molecular sequences. *Molecular Biology and Evolution* 2005.

Drummond AJ, Suchard Ma, Xie D et al. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 2012.

Duerr A, Huang Y, Buchbinder S et al. Extended follow-up confirms early vaccine-enhanced risk of HIV acquisition and demonstrates waning effect over time among participants in a randomized trial of recombinant adenovirus HIV vaccine (Step Study). *Journal of Infectious Diseases* 2012.

D'arc M, Ayouba A, Esteban A et al. Origin of the HIV-1 group O epidemic in western lowland gorillas. *Proceedings of the National Academy of Sciences* 2015.

Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 2004.

Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC bioinformatics* 2008.

Epstein JH, Field HE, Luby S et al. Nipah virus: Impact, origins, and causes of emergence. *Current Infectious Disease Reports* 2006.

Fan J, Ma G, Nosaka K et al. APOBEC3G generates nonsense mutations in human T-cell leukemia virus type 1 proviral genomes in vivo. *Journal of virology* 2010.

Faria NR, Azevedo S, Kraemer MUG et al. Zika virus in the Americas : early epidemiological and genetic findings Summary. 2016.

Faria NR, Rambaut A, Suchard M a et al. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science* 2014.

Faria NR, Suchard M a, Rambaut A et al. Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 2013.

Faroudi M, Utzny C, Salio M et al. Lytic versus stimulatory synapse in cytotoxic T lymphocyte/target cell interaction: manifestation of a dual activation threshold. *Proceedings of the National Academy of Sciences of the United States of America* 2003.

Fauci AS, Folkers GK, Marston HD. Ending the global HIV/AIDS Pandemic: The critical role of an HIV vaccine. *Clinical Infectious Diseases* 2014.

Fauci AS, Marston HD. Ending AIDS — Is an HIV Vaccine Necessary? *New England Journal of Medicine* 2014.

Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* 1981.

Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 1985a.

Felsenstein J. Phylogenies and the Comparative Method. *The American Naturalist* 1985b.

Felsenstein J. A Comparative Method for Both Discrete and Continuous Characters Using the Threshold Model. *The American Naturalist* 2012.

Feorino PM, Jaffe HW, Palmer E et al. Transfusion-associated acquired immunodeficiency syn-

drome. Evidence for persistent infection in blood donors. *The New England journal of medicine* 1985.

Fereidouni S, Munoz O, Von Dobschuetz S et al. Influenza Virus Infection of Marine Mammals. *EcoHealth* 2014.

Flynn N, Forthal D, Harro C et al. Placebo-Controlled Phase 3 Trial of a Recombinant Glycoprotein 120 Vaccine to Prevent HIV-1 Infection. *Journal of Infectious Diseases* 2005.

Frahm N, Korber BT, Adams CM et al. Consistent cytotoxic-T-lymphocyte targeting of immunodominant regions in human immunodeficiency virus across multiple ethnicities. *Journal of virology* 2004.

Freckleton RP, Harvey PH, Pagel M. Phylogenetic Analysis and Comparative Data : *The American naturalist* 2002.

Frieden TR, Damon I, Bell B et al. Ebola 2014 - New Challenges, New Global Response and Responsibility. *New England Journal of Medicine* 2014.

Fu Q, Mittnik A, Johnson PL et al. A revised timescale for human evolution based on ancient mitochondrial genomes. *Current Biology* 2013.

Gao F, Bailes E, Robertson DL et al. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* 1999.

Gao X, Bashirova A, Iversen AKN et al. AIDS restriction HLA allotypes target distinct intervals of HIV-1 pathogenesis. *Nature Medicine* 2005.

Garland, Jr., T, Ives AR. Using the Past to Predict the Present: Confidence Intervals for Regression Equations in Phylogenetic Comparative Methods. *The American Naturalist* 2000.

Gaston KJ, Blackburn TM, Lawton JH. Interspecific abundance-range size relationships: An appraisal of mechanisms. *Journal of Animal Ecology* 1997.

Ge X-Y, Li J-L, Yang X-L et al. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* 2014.

Gelman A, Hill J, Yajima M. Why We (Usually) Don't Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness* 2012.

Gelman A, Jakulin A, Pittau MG et al. A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models. *The Annals of Applied Statistics* 2008.

Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* 1992.

Gifford R, Kabat P, Martin J et al. Evolution and distribution of class II-related endogenous retroviruses. *Journal of virology* 2005.

Gifford R, Tristem M. The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes* 2003.

Gilbert MTP, Rambaut A, Wlasiuk G et al. The emergence of HIV/AIDS in the Americas and beyond. *Proceedings of the National Academy of Sciences of the United States of America* 2007.

Gilbert PB, Berger JO, Stablein D et al. Statistical interpretation of the RV144 HIV vaccine efficacy trial in Thailand: A case study for statistical issues in efficacy trials. *Journal of Infectious*

Diseases 2011.

Gilbert PB, Peterson ML, Follmann D et al. Correlation between immunologic responses to a recombinant glycoprotein 120 vaccine and incidence of HIV-1 infection in a phase 3 HIV-1 preventive vaccine trial. *The Journal of infectious diseases* 2005.

Girard P. *Paradise Lost: Haiti's Tumultuous Journey from Pearl of the Caribbean to Third World Hotspot*. Palgrave, 2005:230.

Gire SK, Goba a, Andersen KG et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* 2014.

Gottlieb MS, Schroff R, Schanker HM et al. Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. *The New England journal of medicine* 1981.

Goulder PJR, Walker BD. HIV and HLA Class I: An Evolving Relationship. *Immunity* 2012.

Goüy De Bellocq J, Charbonnel N, Morand S. Coevolutionary relationship between helminth diversity and MHC class II polymorphism in rodents. *Journal of Evolutionary Biology* 2008.

Gómez a M, Smaill FM, Rosenthal KL. Inhibition of HIV replication by CD8+ T cells correlates with CD4 counts and clinical stage of disease. *Clinical and experimental immunology* 1994.

Grafen. The Phylogenetic Regression. *Philosophical Transactions of the Royal Soociety of London, Series B* 1989.

Gragert L, Madbouly A, Freeman J et al. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Human Immunology* 2013.

Gray R, Kigozi G, Makumbi F et al. The effectiveness of male circumcision for HIV prevention and effects on risk behaviors in a post-trial follow up study in Rakai, Uganda. *Aids* 2012.

Greenland K, Whelan J, Fanoy E et al. Mumps outbreak among vaccinated university students associated with a large party, the Netherlands, 2010. *Vaccine* 2012.

Grenfell BT, Pybus OG, Gog JR et al. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science (New York, NY)* 2004.

Guan Y, Zheng BJ, He YQ et al. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science (New York, NY)* 2003.

Haaland RE, Hawkins Pa, Salazar-Gonzalez J et al. Inflammatory genital infections mitigate a severe genetic bottleneck in heterosexual transmission of subtype A and C HIV-1. *PLoS Pathogens* 2009.

Hadfield JD. Increasing the efficiency of MCMC for hierarchical phylogenetic models of categorical traits using reduced mixed models. *Methods in Ecology and Evolution* 2015.

Hadfield JD, Nakagawa S. General quantitative genetic methods for comparative biology: Phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology* 2010.

Hagberg A a, Schult D a, Swart PJ. Exploring network structure, dynamics, and function using

NetworkX. *Network* 2008.

Hahn BH, Shaw GM, De Cock KM et al. AIDS as a zoonosis: scientific and public health implications. *Science (New York, NY)* 2000.

Hall HI, Song R, Rhodes P et al. Estimation of HIV incidence in the United States. *JAMA : the journal of the American Medical Association* 2008.

Hammer MF, Woerner AE, Mendez FL et al. Genetic evidence for archaic admixture in Africa. *Proceedings of the National Academy of Sciences* 2011.

Hammer SM, Sobieszczyk ME, Janes H et al. Efficacy Trial of a DNA/rAd5 HIV-1 Preventive Vaccine. *New England Journal of Medicine* 2013.

Hansen T. Stabilizing selection and the comparative analysis of adaptation. *Evolution* 1997.

Harris RS, Bishop KN, Sheehy AM et al. DNA deamination mediates innate immunity to retroviral infection. *Cell* 2003.

Harris RS, Liddament MT. Retroviral restriction by APOBEC proteins. *Nature reviews Immunology* 2004.

Harvey PH, Purvis A. Comparative methods for explaining adaptations. *Nature* 1991.

Hastings WK. Monte carlo sampling methods using Markov chains and their applications. *Biometrika* 1970.

Hayward A, Grabherr M, Jern P. Broad-scale phylogenomics provides insights into retrovirus-host evolution. *Proceedings of the National Academy of Sciences of the United States of America* 2013.

Heath TA, Huelsenbeck JP, Stadler T. The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences* 2014.

Hemelaar J, Gouws E, Ghys PD et al. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *AIDS (London, England)* 2006.

Hemelaar J, Gouws E, Ghys PD et al. Global trends in molecular epidemiology of HIV-1 during 2000-2007. *AIDS (London, England)* 2011.

Henzy JE, Gifford RJ, Johnson WE et al. A novel recombinant retrovirus in the genomes of modern birds combines features of avian and mammalian retroviruses. *Journal of virology* 2014.

Herniou E, Martin J, Miller K et al. Retroviral diversity and distribution in vertebrates. *Journal of virology* 1998.

Hillis DM, Huelsenbeck JP. Support for dental HIV transmission. *Nature* 1994.

Ho DD, Neumann a U, Perelson a S et al. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. 1995.

Holmes EC, Nee S, Rambaut a et al. Revealing the history of infectious disease epidemics through phylogenetic trees. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 1995.

Hu WS, Temin HM. Genetic consequences of packaging two RNA genomes in one retroviral particle: pseudodiploidy and high rate of genetic recombination. *Proceedings of the National Academy*

of Sciences of the United States of America 1990.

Huang X, Lu D, Ji G et al. Hepatitis B virus (HBV) vaccine-induced escape mutants of HBV S gene among children from Qidong area, China. *Virus Research* 2004.

Huelsenbeck JP, Ronquist F, Nielsen R et al. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science (New York, NY)* 2001.

Huerta-Cepas J, Dopazo J, Gabaldón T. ETE: a python Environment for Tree Exploration. *BMC bioinformatics* 2010.

Hurst LD. The Ka/Ks ratio: Diagnosing the form of sequence evolution. 2002.

Ishida Y, Zhao K, Greenwood A et al. Proliferation of Endogenous Retroviruses in the Early Stages of a Host Germ Line Invasion. *Molecular biology and evolution* 2014.

IUCN. An Analysis of Mammals on the 2008 IUCN Red List. <*http://wwwiucnredlistorg/mammals*> 2008.

Iversen AKN, Stewart-Jones G, Learn GH et al. Conflicting selective forces affect T cell receptor contacts in an immunodominant human immunodeficiency virus epitope. *Nature immunology* 2006.

Ives AR, Garland T. Phylogenetic logistic regression for binary dependent variables. *Systematic Biology* 2010.

Jeffreys H. *Theory of Probability.*, 1961.

Jetzt a E, Yu H, Klarmann GJ et al. High rate of recombination throughout the human immunodeficiency virus type 1 genome. *Journal of virology* 2000.

Jones KE, Patel NG, Levy M a et al. Global trends in emerging infectious diseases. *Nature* 2008.

Josefsson L, Palmer S, Faria NR et al. Single Cell Analysis of Lymph Node Tissue from HIV-1 Infected Patients Reveals that the Majority of CD4+ T-cells Contain One HIV-1 DNA Molecule. *PLoS Pathogens* 2013.

Jost S, Bernard M-C, Kaiser L et al. A Patient with HIV-1 Superinfection. *New England Journal of Medicine* 2002.

Kalish ML, Wolfe ND, Ndongmo CB et al. Central African hunters exposed to simian immunodeficiency virus. *Emerging Infectious Diseases* 2005.

Karim SSA, Karim QA. Antiretroviral prophylaxis: A defining moment in HIV control. *The Lancet* 2011.

Kass R, Raftery A. Bayes Factors. *J Amer Statist Assoc* 1995.

Katzourakis A, Gifford RJ, Tristem M et al. Macroevolution of complex retroviruses. *Science (New York, NY)* 2009.

Katzourakis A, Magiorkinis G, Lim AG et al. Larger Mammalian Body Size Leads to Lower Retroviral Activity. Maley C (ed.). *PLoS Pathogens* 2014.

Kawashima Y, Pfafferott K, Frater J et al. Adaptation of HIV-1 to human leukocyte antigen class I. *Nature* 2009.

Kay ES, Batey DS, Mugavero MJ. The HIV treatment cascade and care continuum: updates,

goals, and recommendations for the future. *AIDS Research and Therapy* 2016.

Keele BF, Giorgi EE, Salazar-Gonzalez JF et al. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proceedings of the National Academy of Sciences of the United States of America* 2008.

Keele BF, Li H, Learn GH et al. Low-dose rectal inoculation of rhesus macaques by SIVsmE660 or SIVmac251 recapitulates human mucosal infection by HIV-1. *The Journal of experimental medicine* 2009.

Keele BF, Van Heuverswyn F, Li Y et al. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science (New York, NY)* 2006.

Keeling M, Gilligan C. Metapopulation dynamics of bubonic plague. *Nature* 2000.

Kellam P, Larder B a. Retroviral recombination can lead to linkage of reverse transcriptase mutations that confer increased zidovudine resistance. *Journal of virology* 1995.

Kelleher a D, Long C, Holmes EC et al. Clustered mutations in HIV-1 gag are consistently required for escape from HLA-B27-restricted cytotoxic T lymphocyte responses. *The Journal of experimental medicine* 2001.

Kemal KS, Kitchen CMR, Burger H et al. Recombination Between Variants from Genital Tract and Plasma: Evolution of Multidrug-Resistant HIV Type 1. *AIDS Research and Human Retroviruses* 2012.

Khan M a, Kao S, Miyagi E et al. Viral RNA Is Required for the Association of APOBEC3G with Human Immunodeficiency Virus Type 1 Nucleoprotein Complexes Viral RNA Is Required for the Association of APOBEC3G with Human Immunodeficiency Virus Type 1 Nucleoprotein Complexes. *Journal of virology* 2005.

Kiepiela P, Leslie AJ, Honeyborne I et al. Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* 2004.

Kiepiela P, Ngumbela K, Thobakgale C et al. CD8+ T-cell responses to different HIV proteins have discordant associations with viral load. *Nature medicine* 2007.

Kimani J, Kaul R, Nagelkerke NJD et al. Reduced rates of HIV acquisition during unprotected sex by Kenyan female sex workers predating population declines in HIV prevalence. *AIDS (London, England)* 2008.

Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution* 1980.

Kingman J. The coalescent. *Stochastic Processes and their Applications* 1982.

Kloverpris HN, Stryhn A, Harndahl M et al. HLA-B*57 Micropolymorphism shapes HLA allele-specific epitope immunogenicity, selection pressure, and HIV immune control. *Journal of virology* 2012.

Koehler AV, Pearce JM, Flint PL et al. Genetic evidence of intercontinental movement of avian influenza in a migratory bird: The northern pintail (Anas acuta). *Molecular Ecology* 2008.

Korber B, Muldoon M, Theiler J et al. Timing the ancestor of the HIV-1 pandemic strains.

Science (New York, NY) 2000.

Koup R a, Safrit JT, Cao Y et al. responses with the initial control of viremia in Temporal Association of Cellular Immune Responses with the Initial Control of Viremia in Primary Human Immunodeficiency Virus Type 1 Syndrome. *J Virol* 1994.

Kovaliski J. Monitoring the spread of rabbit hemorrhagic disease virus as a new biological agent for control of wild European rabbits in Australia. *Journal of wildlife diseases* 1998.

Kuiken C, Korber B, Shafer RW. HIV sequence databases. *AIDS Reviews* 2003.

Kuiken T, Holmes EC, McCauley J et al. Host species barriers to influenza virus infections. *Science (New York, NY)* 2006.

Kumar S, Subramanian S. Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences of the United States of America* 2002.

Larsson A. AliView: a fast and lightweight alignment viewer and editor for large data sets. *Bioinformatics* 2014.

Lazaryan A, Lobashevsky E, Mulenga J et al. Human leukocyte antigen B58 supertype and human immunodeficiency virus type 1 infection in native Africans. *Journal of virology* 2006.

LeBreton M, Switzer WM, Djoko CF et al. A gorilla reservoir for human T-lymphotropic virus type 4. *Emerging microbes & infections* 2014.

Lee YN, Malim MH, Bieniasz PD. Hypermutation of an ancient human retrovirus by APOBEC$_3$G. *Journal of virology* 2008.

Lemey P, Rambaut A, Bedford T. The seasonal flight of influenza : a unified framework for spatiotemporal hypothesis testing. *arXiv preprint arXiv: ...* 2012.

Lemey P, Rambaut A, Bedford T et al. Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H$_3$N$_2$. *PLoS Pathogens* 2014.

Lemey P, Rambaut A, Drummond AJ et al. Bayesian phylogeography finds its roots. *PLoS Computational Biology* 2009.

Leroy EM, Epelboin A, Mondonge V et al. Human Ebola outbreak resulting from direct exposure to fruit bats in Luebo, Democratic Republic of Congo, 2007. *Vector borne and zoonotic diseases (Larchmont, NY)* 2009.

Leroy EM, Kumulungui B, Pourrut X et al. Fruit bats as reservoirs of Ebola virus. *Nature* 2005.

Leslie A, Matthews PC, Listgarten J et al. Additive contribution of HLA class I alleles in the immune control of HIV-1 infection. *Journal of virology* 2010.

Levitt AL, Singh R, Cox-Foster DL et al. Cross-species transmission of honey bee viruses in associated arthropods. *Virus Research* 2013.

Leye N, Vidal N, Ndiaye O et al. High frequency of HIV-1 infections with multiple HIV-1 strains in men having sex with men (MSM) in Senegal. *Infection, Genetics and Evolution* 2013.

Li H, Bar KJ, Wang S et al. High multiplicity infection by HIV-1 in men who have sex with men. *PLoS Pathogens* 2010.

Li W-H, Tanimura M, Sharp PM. Rates and Dates of Divergence between AIDS Virus Nucle-

otide Sequences. *Molecular Biology Evolution* 1988.

Li W, Shi Z, Yu M et al. Bats are natural reservoirs of SARS-like coronaviruses. *Science (New York, NY)* 2005.

Lieber MM, Sherr CJ, Todaro GJ et al. Isolation from the asian mouse Mus caroli of an endogenous type C virus related to infectious primate type C viruses. *Proceedings of the National Academy of Sciences of the United States of America* 1975.

Liegeois F, Butel C, Mouinga-onde A et al. HIV Type-1 Group O Infection in Gabon : Low Prevalence Rate But Circulation of Genetically. *AIDS research and human retroviruses* 2013.

Lihana RW, Khamadi SA, Kiptoo MK et al. HIV type 1 subtypes among STI patients in Nairobi: a genotypic study based on partial pol gene sequencing. *AIDS research and human retroviruses* 2006.

Longdon B, Brockhurst M a, Russell C a et al. The Evolution and Genetics of Virus Host Shifts. Hobman TC (ed.). *PLoS Pathogens* 2014.

Longdon B, Hadfield JD, Webster CL et al. Host phylogeny determines viral persistence and replication in novel hosts. *PLoS Pathogens* 2011.

Mariani R, Chen D, Schröfelbauer B et al. Species-specific exclusion of APOBEC3G from HIV-1 virions by Vif. *Cell* 2003.

Marr JS, Kiracofe JB. Was the Huey Cocoliztli a Haemorrhagic Fever? *Medical History* 2000.

Martin J, Herniou E, Cook J et al. Interclass transmission and phyletic host tracking in murine leukemia virus-related retroviruses. *Journal of virology* 1999.

Martin J, Kabat P, Tristem M. Cospeciation and horizontal transmission rates in the murine leukaemia-related retroviruses. *Tangled Trees: Phylogeny, Cospeciation and Coevolution* 2003.

Martin MP, Carrington M. Immunogenetics of HIV disease. *Immunological Reviews* 2013.

Masharsky AE, Dukhovlinova EN, Verevochkin SV et al. A substantial transmission bottleneck among newly and recently HIV-1-infected injection drug users in St Petersburg, Russia. *The Journal of infectious diseases* 2010.

Matano T, Shibata R, Siemon C et al. Administration of an anti-CD8 monoclonal antibody interferes with the clearance of chimeric simian/human immunodeficiency virus during primary infections of rhesus macaques. *Journal of virology* 1998.

Matthews PC, Adland E, Listgarten J et al. HLA-A*7401-mediated control of HIV viremia is independent of its linkage disequilibrium with HLA-B*5703. *Journal of immunology* 2011.

Matthews PC, Leslie AJ, Katzourakis A et al. HLA footprints on human immunodeficiency virus type 1 are associated with interclade polymorphisms and intraclade phylogenetic clustering. *Journal of virology* 2009.

Matthews PC, Listgarten J, Carlson JM et al. Co-Operative Additive Effects between HLA Alleles in Control of HIV-1. *PLoS ONE* 2012.

McCutchan FE. Global epidemiology of HIV. *Journal of Medical Virology* 2006.

McKinnon LR, Nagelkerke NJ, Kaul R et al. HIV-1 Clade D Is Associated with Increased

Rates of CD_4 Decline in a Kenyan Cohort. *PLoS ONE* 2012.

McMichael a J, Haynes BF. Lessons learned from HIV-1 vaccine trials: new priorities and directions. *Nature Immunology* 2012.

Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in Europeans. *Science* 1978.

Metropolis N, Rosenbluth AW, Rosenbluth MN et al. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* 1953.

Meyer M, Fu Q, Aximu-Petri A et al. A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature* 2014.

Migueles S a, Laborico AC, Imamichi H et al. The differential ability of HLA B*5701+ long-term nonprogressors and progressors to restrict human immunodeficiency virus replication is not caused by loss of recognition of autologous viral gag sequences. *Journal of virology* 2003.

Migueles SA, Sabbaghian MS, Shupert WL et al. HLA B*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors. *Proceedings of the National Academy of Sciences of the United States of America* 2000.

Min J, Yu D-w, Liang W-j et al. Molecular Evolution of the SARS Coronavirus During the Course of the SARS Epidemic in China. *Science* 2004.

Minin VN, Suchard Ma. Counting labeled transitions in continuous-time Markov models of evolution. *Journal of Mathematical Biology* 2008.

Mohri H, Perelson AS, Tung K et al. Increased turnover of T lymphocytes in HIV-1 infection and its reduction by antiretroviral therapy. *The Journal of experimental medicine* 2001.

Montiel Na. An updated review of simian betaretrovirus (SRV) in macaque hosts. *Journal of Medical Primatology* 2010.

Moore CB, John M, James IR et al. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science (New York, NY)* 2002.

Morse SS. Factors in the emergence of infectious diseases. *Emerging Infectious Diseases* 1995.

Moutouh L, Corbeil J, Richman DD. Recombination leads to the rapid emergence of HIV-1 dually resistant mutants under selective drug pressure. *Proceedings of the National Academy of Sciences of the United States of America* 1996.

Muller HJ. the Relation of Recombination To Mutational Advance. *Mutation research* 1964.

Neilson JR, John GC, Carr JK et al. Subtypes of human immunodeficiency virus type 1 and disease stage among women in Nairobi, Kenya. *Journal of virology* 1999.

Ninove L, Domart Y, Vervel C et al. Cowpox virus transmission from pet rats to humans, France. *Emerging infectious diseases* 2009.

Nunes MRT, Palacios G, Faria NR et al. Air Travel Is Associated with Intracontinental Spread of Dengue Virus Serotypes 1-3 in Brazil. *PLoS Neglected Tropical Diseases* 2014.

Olival KJ, Hosseini PR, Zambrana-Torrelio C et al. Host and viral traits predict zoonotic

spillover from mammals. *Nature* 2017.

Olson DM, Dinerstein E, Wikramanayake ED et al. Terrestrial Ecoregions of the World: A New Map of Life on Earth. *BioScience* 2001.

Ou CY, Ciesielski C a, Myers G et al. Molecular epidemiology of HIV transmission in a dental practice. *Science* 1992.

O'Brien JD, Minin VN, Suchard Ma. Learning to count: Robust estimates for labeled distances between molecular sequences. *Molecular Biology and Evolution* 2009.

O'Connell JF, Allen J. Dating the colonization of Sahul (Pleistocene Australia-New Guinea): A review of recent research. *Journal of Archaeological Science* 2004.

O'Leary M a, Bloch JI, Flynn JJ et al. The placental mammal ancestor and the post-K-Pg radiation of placentals. *Science (New York, NY)* 2013.

Pace C, Keller J, Nolan D et al. Population level analysis of human immunodeficiency virus type 1 hypermutation and its relationship with APOBEC3G and vif genetic variation. *Journal of virology* 2006.

Pagel M. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic biology* 1999.

Pagel M, Meade A, Barker D. Bayesian estimation of ancestral character states on phylogenies. *Systematic biology* 2004.

Palmer DS, Turner I, Fidler S et al. Mapping the drivers of within-host pathogen evolution using massive data sets. *bioRxiv* 2017.

Papaspiliopoulos O, Roberts GO. Non-Centered Parameterisations for Hierarchical Models and Data Augmentation. *Bayesian Statistics* 2003.

Parker J, Rambaut A, Pybus OG. Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty. *Infection, Genetics and Evolution* 2008.

Parpia AS, Ndeffo-Mbah ML, Wenzel NS et al. Effects of response to 2014-2015 ebola outbreak on deaths from malaria, HIV/AIDS, and tuberculosis, West Africa. *Emerging Infectious Diseases* 2016.

Parrish CR, Holmes EC, Morens DM et al. Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiology and molecular biology reviews : MMBR* 2008.

Patin E, Lopez M, Grollemund R et al. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science* 2017.

Paximadis M, Mathebula TY, Gentle NL et al. Human leukocyte antigen class I (A, B, C) and II (DRB1) diversity in the black and Caucasian South African population. *Human Immunology* 2012.

Peeters M, Courgnaud V, Abela B et al. Risk to human health from a plethora of Simian immunodeficiency viruses in primate bushmeat. *Emerging Infectious Diseases* 2002.

Peeters M, Sharp PM. Genetic diversity of HIV-1: the moving target. *AIDS (London, England)* 2000.

Pereyra F, Jia X, McLaren PJ et al. The Major Genetic Determinants of HIV-1 Control Affect

HLA Class I Peptide Presentation. *Science* 2010.

Peterson a T. Defining viral species: making taxonomy useful. *Virology journal* 2014.

Peterson Ta, Kimani J, Wachihi C et al. HLA class I associations with rates of HIV-1 seroconversion and disease progression in the Pumwani Sex Worker Cohort. *Tissue Antigens* 2013.

Pérez-Sautu U, Costafreda MI, Caylà J et al. Hepatitis A virus vaccine escape variants and potential new serotype emergence. *Emerging Infectious Diseases* 2011.

Pineda-Peña AC, Faria NR, Imbrechts S et al. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: Performance evaluation of the new REGA version 3 and seven other tools. *Infection, Genetics and Evolution* 2013.

Pitisuttithum P, Gilbert P, Gurwith M et al. Randomized, double-blind, placebo-controlled efficacy trial of a bivalent recombinant glycoprotein 120 HIV-1 vaccine among injection drug users in Bangkok, Thailand. *The Journal of infectious diseases* 2006.

Plantier J-C, Leoz M, Dickerson JE et al. A new human immunodeficiency virus derived from gorillas. *Nature medicine* 2009.

Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)* 2003.

Plummer M, Best N, Cowles K et al. CODA: convergence diagnosis and output analysis for MCMC. *R News* 2006.

Poon AFY, Swenson LC, Dong WWY et al. Phylogenetic analysis of population-based and deep sequencing data to identify coevolving sites in the nef gene of HIV-1. *Molecular Biology and Evolution* 2010.

Popinga A, Vaughan T, Stadler T et al. Inferring epidemiological dynamics with Bayesian coalescent inference: The merits of deterministic and stochastic models. *Genetics* 2015.

Prejean J, Song R, Hernandez A et al. Estimated HIV incidence in the United States, 2006-2009. *PLoS ONE* 2011.

Price DA, Goulder PJ, Klenerman P et al. 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* 1997.

Prugnolle F, Manica A, Charpentier M et al. Pathogen-driven selection and worldwide HLA class I diversity. *Current Biology* 2005.

Purbhoo MA, Irvine DJ, Huppa JB et al. T cell killing does not require the formation of a stable mature immunological synapse. *Nature Immunology* 2004.

Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. *Nature reviews Genetics* 2009.

Pybus OG, Rambaut A, Harvey PH. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* 2000.

Quinones-Parra S, Grant E, Loh L et al. Preexisting CD8+ T-cell immunity to the H7N9 influenza A virus varies across ethnicities. *Proceedings of the National Academy of Sciences* 2014.

Rambaut A. Estimating the rate of molecular evolution: incorporating non-contemporaneous se-

quences into maximum likelihood phylogenies. *Bioinformatics* 2000.

Ramey AM, Reeves AB, Ogawa H et al. Genetic diversity and mutation of avian paramyxovirus serotype 1 (Newcastle disease virus) in wild birds and evidence for intercontinental spread. *Archives of Virology* 2013.

Ramjee G, Daniels B, Ramjee G et al. Women and HIV in Sub-Saharan Africa. *AIDS Research and Therapy* 2013.

Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S et al. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *The New England journal of medicine* 2009.

Rinaldo C, Huang XL, Fan ZF et al. High levels of anti-human immunodeficiency virus type 1 (HIV-1) memory cytotoxic T-lymphocyte activity and low viral load are associated with lack of disease in HIV-1-infected long-term nonprogressors. *Journal of virology* 1995.

Ritchie AJ, Cai F, Smith NMG et al. Recombination-mediated escape from primary CD8 + T cells in acute HIV-1 infection. *Retrovirology* 2014.

Robertson DL, Anderson JP, Bradac J a et al. HIV-1 nomenclature proposal. *Science (New York, NY)* 2000.

Rohlf FJ. Comparative methods for the analysis of continuous variables: Geometric interpretations. *Evolution* 2001.

Ronen K, McCoy CO, Matsen Fa et al. HIV-1 Superinfection Occurs Less Frequently Than Initial Infection in a Cohort of High-Risk Kenyan Women. *PLoS Pathogens* 2013.

Ronquist F, Teslenko M, Van Der Mark P et al. Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* 2012.

Rousseau CM, Learn GH, Bhattacharya T et al. Extensive intrasubtype recombination in South African human immunodeficiency virus type 1 subtype C infections. *Journal of virology* 2007.

Rudnick, A. & Lim T. Dengue fever studies in Malaysia. *Bulletin from the Institute for Medical Research, Malaysia* 1986.

Saeb S, Lonini L, Jayaraman A et al. Voodoo Machine Learning for Clinical Predictions. *bioRxiv* 2016.

Sajantila A. Molecular clocks ticking in the court room. *Investigative genetics* 2014.

Salemi M, Strimmer K, Hall WW et al. Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 2001.

Salgado M, Simón A, Sanz-Minguela B et al. An Additive Effect of Protective Host Genetic Factors Correlates With HIV Nonprogression Status. *JAIDS Journal of Acquired Immune Deficiency Syndromes* 2011.

Salminen MO, Carr JK, Burke DS et al. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS research and human retroviruses* 1995.

Sane J, Gouma S, Koopmans M et al. Epidemic of mumps among vaccinated persons, the Neth-

erlands, 2009-2012. *Emerging Infectious Diseases* 2014.

Sawyer SL, Emerman M, Malik HS. Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS Biology* 2004.

Schluter D. The Ecology of Adaptive Radiation. 2000.

Schultz A-K, Zhang M, Leitner T et al. A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes. *BMC bioinformatics* 2006.

Schultz AK, Zhang M, Bulla I et al. jpHMM: Improving the reliability of recombination prediction in HIV-1. 2009.

Ségurel L, Quintana-Murci L. Preserving immune diversity through ancient inheritance and admixture. *Current Opinion in Immunology* 2014.

Sharp PM, Hahn BH. Origins of HIV and the AIDS pandemic. *Cold Spring Harbor Perspectives in Medicine* 2011.

Sheehy AM, Gaddis NC, Choi JD et al. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* 2002.

Sievers F, Wilm A, Dineen D et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 2011.

Silva TI de, Cotten M, Rowland-Jones SL. HIV-2: the forgotten AIDS virus. *Trends in Microbiology* 2008.

Silva TI de, Tienen C van, Onyango C et al. Population dynamics of HIV-2 in rural Guinea-Bissau. *Aids* 2012.

Simon-Loriere E, Holmes EC. Why do RNA viruses recombine? *Nature reviews Microbiology* 2011.

Simonsen JN, Plummer F a, Ngugi EN et al. HIV infection among lower socioeconomic strata prostitutes in Nairobi. 1990.

Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics* 2005.

Smith GJD, Vijaykrishna D, Bahl J et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* 2009.

Smyth RP, Schlub TE, Grimm AJ et al. Identifying Recombination Hot Spots in the HIV-1 Genome. *Journal of virology* 2014.

Snyder EL, Dodd RY. Reducing the risk of blood transfusion. *Hematology / the Education Program of the American Society of Hematology American Society of Hematology Education Program* 2001.

Spurgin LG, Richardson DS. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings Biological sciences / The Royal Society* 2010.

Stadler T, Kouyos R, VonWy V et al. Estimating the basic reproductive number from viral sequence data. *Molecular Biology and Evolution* 2012.

Strebel K. APOBEC3G & HTLV-1: inhibition without deamination. *Retrovirology* 2005.

Streeck H, Lichterfeld M, Alter G et al. Recognition of a Defined Region within p24 Gag

by CD8+ T Cells during Primary Human Immunodeficiency Virus Type 1 Infection in Individuals Expressing Protective HLA Class I Alleles. *Journal of Virology* 2007.

Streicker DG, Turmelle AS, Vonhof MJ et al. Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. *Science (New York, NY)* 2010.

Switzer WM, Tang S, Ahuka-Mundeke S et al. Novel simian foamy virus infections from multiple monkey species in women from the Democratic Republic of Congo. *Retrovirology* 2012.

Talbi C, Lemey P, Suchard Ma et al. Phylodynamics and Human-mediated dispersal of a zoonotic virus. *PLoS Pathogens* 2010.

Tarlinton RE, Meers J, Young PR. Retroviral invasion of the koala genome. *Nature* 2006.

Tenenhaus M, Young FW. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika* 1985.

Tenzer S, Crawford H, Pymm P et al. HIV-1 Adaptation to Antigen Processing Results in Population-Level Immune Evasion and Affects Subtype Diversification. *Cell Reports* 2014.

Tenzer S, Wee E, Burgevin A et al. Antigen processing influences HIV-specific cytotoxic T lymphocyte immunodominance. *Nature immunology* 2009.

Tully DC, Ogilvie CB, Batorsky RE et al. Differences in the Selection Bottleneck between Modes of Sexual Transmission Influence the Genetic Composition of the HIV-1 Founder Virus. *PLoS Pathogens* 2016.

Vallari A, Bodelle P, Ngansop C et al. Four new HIV-1 group N isolates from Cameroon: Prevalence continues to be low. *AIDS research and human retroviruses* 2010.

Vallari A, Holzmayer V, Harris B et al. Confirmation of putative HIV-1 group P in Cameroon. *Journal of virology* 2011.

Vapalahti O, Mustonen J, Lundkvist Å et al. Hantavirus infections in Europe. *Lancet Infectious Diseases* 2003.

Vasilakis N, Holmes EC, Fokam EB et al. Evolutionary processes among sylvatic dengue type 2 viruses. *Journal of virology* 2007.

Veevers JJ. Updated Gondwana (Permian-Cretaceous) earth history of Australia. *Gondwana Research* 2006.

Venables WN, Ripley BD. *Modern applied statistics with S-PLUS*. Springer Science & Business Media, 2013.

Villemereuil P de, Wells Ja, Edwards RD et al. Bayesian models for comparative analysis integrating phylogenetic uncertainty. *BMC Evolutionary Biology* 2012.

Visawapoka U, Tovanabutra S, Currier JR et al. Circulating and unique recombinant forms of HIV type 1 containing subsubtype A2. *AIDS research and human retroviruses* 2006.

Wain LV, Bailes E, Bibollet-Ruche F et al. Adaptation of HIV-1 to Its Human Host. *Molecular Biology and Evolution* 2007.

Walensky RP, Paltiel AD, Goldie SJ et al. A therapeutic HIV vaccine: How good is good

enough? *Vaccine* 2004.

Wang L, Valderramos SG, Wu A et al. From Mosquitos to Humans: Genetic Evolution of Zika Virus. *Cell Host and Microbe* 2016.

Wang W, Jiang S, Li S et al. Identification of subtype B, multiple circulating recombinant forms and unique recombinants of HIV type 1 in an MSM cohort in China. *AIDS research and human retroviruses* 2008.

Wawer MJ, Gray RH, Sewankambo NK et al. Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda. *The Journal of infectious diseases* 2005.

Weiss R a. The Leeuwenhoek Lecture 2001. Animal origins of human infectious disease. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 2001.

Weiss R a. The discovery of endogenous retroviruses. *Retrovirology* 2006.

Wen J, Tang WW, Sheets N et al. Identification of Zika virus epitopes reveals immunodominant and protective roles for dengue virus cross-reactive CD8+ T cells. *Nature Microbiology* 2017.

WHO. Progress Report 2016, prevent HIV, test and treat all. *Progress Report 2016, prevent HIV, test and treat all* 2016.

Wolfe ND, Switzer WM, Carr JK et al. Naturally acquired simian retrovirus infections in central African hunters. *Lancet* 2004.

World Health Organization. Guidance on pre-exposure oral prophylaxis (PrEP) for serodiscordant couples, men who have sex with men and transgender women at high risk of HIV in implementation research. 2012.

Worobey M, Gemmel M, Teuwen DE et al. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* 2008.

Worobey M, Holmes EC. Evolutionary aspects of recombination in RNA viruses. *Journal of General Virology* 1999.

Worobey M, Watts TD, McKay RA et al. 1970s and 'Patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America. *Nature* 2016.

Wright S. Evolution in Mendelian populations. *Bulletin of Mathematical Biology* 1990.

Yu Q, König R, Pillai S et al. Single-strand specificity of APOBEC3G accounts for minus-strand deamination of the HIV genome. *Nature structural & molecular biology* 2004.

Zhang C, Silva S de, Wang JH et al. Co-evolution of primate SAMHD1 and Lentivirus Vpx leads to the loss of the vpx gene in HIV-1 Ancestor. *PLoS ONE* 2012.

Zhang H, Yang B, Pomerantz RJ et al. The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA. *Nature* 2003.

Zhang J, Webb DM. Rapid evolution of primate antiviral enzyme APOBEC3G. *Human Molecular Genetics* 2004.

Zhou Q, Zhao L, Guan Y. Strong Selection at MHC in Mexicans since Admixture. *PLOS Genetics* 2016.

Zhu T, Korber BT, Nahmias a J et al. An African HIV-1 sequence from 1959 and implications

for the origin of the epidemic. *Nature* 1998.

Zhu T, Wang N, Carr a et al. Genetic characterization of human immunodeficiency virus type 1 in blood and genital secretions: evidence for viral compartmentalization and selection during sexual transmission. *Journal of virology* 1996.

Zhuang J, Zhuang J, Jetzt AE et al. Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. *Journal of virology* 2002.

Zinkernagel RM, Doherty PC. Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system. *Nature* 1974.