Modelling the Evolution of HIV-1 Protein-Coding Sequences with Particular Focus on the Early Stages of Infection

Natasha Thandi Wood

Thesis presented for the Degree of

Doctor of Philosophy

in the Division of Molecular and Cell Biology
University of Cape Town

Supervisor: Professor Cathal Seoighe

Co-supervisor: Professor Carolyn Williamson

May 21, 2010

Declaration

I, Natasha Thandi Wood, declare that this thesis is my own, unaided work (except where acknowledgements indicate otherwise). Neither the whole work, nor part thereof has been, is being, or is to be submitted for any degree or examination at any other university. I empower the University of Cape Town to reproduce for the purpose of research either the whole or any part of the contents of this thesis, in any manner whatsoever.

Signature of candidate:_____

Signed on the _____ day of _____, 2010.

Abstract

Modelling the Evolution of HIV-1 Protein-Coding Sequences with Particular Focus on the Early Stages of Infection
Natasha Thandi Wood, ***February 2010***

The evolution of the viral genome sequence over the course of HIV-1 infection is of interest for vaccine and drug design, and for the development of effective treatment strategies. Characteristics of the transmitted viral genome that could render the virus more sensitive to host immune responses, are of particular interest for vaccine studies. However, sequence samples from the earliest phase of HIV infection are scarce, and inferences about the nature of the infecting virus and its evolution during the course of early infection are often made from samples isolated from later stages, or from chronic infections.

To establish in detail the adaptive changes that occur in early infection, an investigation was carried out on a large dataset consisting of sequences isolated from individuals in early infection. The majority of these infections were inferred to have resulted from transmission of a single virion or virally infected cell, which permitted a detailed investigation of HIV-1 diversification in early infection for the first time. Comparing viral diversification across multiple patients, it was possible to identify specific evolutionary patterns in the HIV-1 genome that occur frequently during the earliest stages of infection. The analyses revealed that APOBEC-mediated hypermutation has an important role in early viral diversification and may enable rapid escape from the first wave of host immune responses. Several mutations in early infection that were likely to result in immune escape were identified, some of which have subsequently been confirmed experimentally. In general, experimental verification of model-based inferences is necessary, but can be expensive and time-consuming. To reduce the costs involved, it is essential that the evolutionary methods produce accurate results. Simulation results presented in this thesis show that inferences made about viral evolution can be subject to bias when key aspects of viral biology are not accounted for by the models used. In particular, some previous comparisons between sequence groups that share genealogical histories, positive selection studies that fail to account for recombination, and research on HIV covariation, may need to be revisited, using more accurate evolutionary models.

The results presented in this thesis demonstrate the importance of accurate evolutionary models to understand the selection pressures acting on the virus during various stages of infection. Furthermore, using a phylogenetic model it was possible to identify sites in the HIV genome that were evolving adaptively and are implicated in CTL immune escape during early infection. Characterising escape mutations in the transmitted virus may lead to novel approaches to develop vaccines and antiviral drugs.

List of Peer Reviewed Publications

Treurnicht, F. K.; Seoighe, C.; Martin, D. P.; **Wood, N.**; Abrahams, M.-R.; de Assis Rosa, D.; Bredell, H.; Woodman, Z.; Hide, W.; Mlisana, K.; Karim, S. A.; Gray, C. M. & Williamson, C. Adaptive changes in HIV-1 subtype C proteins during early infection are driven by changes in HLA-associated immune pressure. *Virology*, 2010, 396(2), 213-225

Harkins, G.; Delport, W.; Duffy, S.; **Wood, N.**; Monjane, A.; Owor, B.; Donaldson, L.; Saumtally, S.; Triton, G.; Briddon, R.; Shepherd, D.; Rybicki, E.; Martin, D. & Varsani, A. Experimental evidence indicating that mastreviruses probably did not co-diverge with their hosts. *Virology Journal*, 2009, 6, 104

Wood, N.; Bhattacharya, T.; Keele, B.F.; Giorgi, E.; Liu, M.; Gaschen, B.; Daniels, M.; Ferrari, G.; Haynes, B.F.; McMichael, A.; Shaw, G.M.; Hahn, B.H.; Korber, B.; Seoighe, C. HIV evolution in early infection: selection pressures, patterns of insertion and deletion, and the impact of Apobec. *PLOS Pathogens* 2009, 5:e1000414

Abrahams, M.-R.; Anderson, J. A.; Giorgi, E. E.; Seoighe, C.; Mlisana, K.; Ping, L.-H.; Athreya, G. S.; Treurnicht, F. K.; Keele, B. F.; **Wood, N**.; Salazar-Gonzalez, J. F.; Bhattacharya, T.; Chu, H.; Hoffman, I.; Galvin, S.; Mapanje, C.; Kazembe, P.; Thebus, R.; Fiscus, S.; Hide, W.; Cohen, M. S.; Karim, S. A.; Haynes, B. F.; Shaw, G. M.; Hahn, B. H.; Korber, B. T.; Swanstrom, R.; Williamson, C.; for the CAPRISA Acute Infection Study Team, the Center for HIV-AIDS Vaccine Immunology Consortium. Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-poisson distribution of transmitted variants. *J Virol*, 2009, 83, 3556-3567

Keele, B. F.; Giorgi, E. E.; Salazar-Gonzalez, J. F.; Decker, J. M.; Pham, K. T.; Salazar, M. G.; Sun, C.; Grayson, T.; Wang, S.; Li, H.; Wei, X.; Jiang, C.; Kirchherr, J. L.; Gao, F.; Anderson, J. A.; Ping, L.-H.; Swanstrom, R.; Tomaras, G. D.; Blattner, W. A.; Goepfert, P. A.; Kilby, J. M.; Saag, M. S.; Delwart, E. L.; Busch, M. P.; Cohen, M. S.; Montefiori, D. C.; Haynes, B. F.; Gaschen, B.; Athreya, G. S.; Lee, H. Y.; **Wood, N.**; Seoighe, C.; Perelson, A. S.; Bhattacharya, T.; Korber, B. T.; Hahn, B. H. & Shaw, G. M. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *PLoS Pathogens*, 2008, 105, 7552-7557

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

Acknowledgements

I would like to express immense gratitude to my supervisor, Professor Cathal Seoighe, his sound advice, academic guidance, and critical reviewing of my thesis have been invaluable to me. I am extremely grateful for the opportunities he provided over the years, and for the patience during the times when my motivation was faltering.

I would also like to thank my co-supervisor, Professor Carolyn Williamson, for the useful discussions throughout my PhD studies, and for the wonderful opportunity to be part of the larger HIV research community.

I would like to extend my sincere thanks to Dr Konrad Scheffler, Dr Wayne Delport, and Professor Sergei Kosakovsky Pond for their advice on evolutionary modelling and/or HyPhy coding, and to Dr Darren Martin, Dr Zenda Woodman, Dr Nobubelo Ngandu, Gama Bandawe, and Melissa-Rose Abrahams for sharing their thoughts on HIV and their willingness to consider any question I asked them. I am also grateful for the conversations with numerous international collaborators and the opportunity to learn from experienced researchers.

To our computer and technical expert, Rodger Duffet, not only for his unfailing efficiency as our systems administrator, but also for his endless support throughout my PhD and for going out of his way to keep my simulations alive. I can not thank him enough. He will never be disestablished in our world.

I am extremely grateful to Jeremy Baxter and Miguel Lacerda who provided me with statistical advice at all times of need.

I greatly appreciate the many e-mails from André Faure, Anthony Smith, and Francois Malan; thank you for inadvertently learning so much about my research.

To my many student colleagues for providing an enjoyable environment, and to my Bioinformatics associates for being an accessible knowledge base, I am tremendously grateful. I would specifically like to thank André Faure (and everything ranunculaceous), Graham Poulter, Ryan Goosen, and Joël Ravelomanantsoa-Ratsimihah for their unwavering support

during the core periods of my research. I am further thankful to Dr Nicola Mulder, Venu Vuppa, Bukiwe Lupindo, Dr Victoria Nembaware, Sachin Somers, Renaud Gaujoux, Gustavo Adolfo Salazar Orejuela, Cashifa Karriem, Ayton Meintjes, Elizabeth Kelly, Kenneth Opap, Gerrit Botha, Jean-Michel Safari Serufuri, Athlee Maclear, James Dominy, Amanda Gillespie, Claudia Harrison, Dane Kennedy, Pieter Burger, Beverly van Rooyen, and John Parathyras, for reasons as diverse as those ridiculously early morning runs, offering advice on how not to write acknowledgements, and for willingly jumping out of an aeroplane with me.

For the informal support and encouragement of many friends and partners in crime over the years, particularly Kylie Wolhuter, I thank her for reminding me how precious memories are and how important it is to adventure ahead to make new ones. I acknowledge Gerhardt Vorster, Jacobus Brink, Ulrich Schoeman, Brett Wordon, Tamsin Olsen, Sean Olsen, Donovan Tose, Gary Birch, Mark Vismer, Karl Schneeberger, Eon van Zyl, and Réza de Greeff; I have so much appreciation for their unconventional encouragement, loyalty, and patience. I am extremely fortunate to have spent time as one of Dr Eugeny Grigoriev's students (in a discipline completely different to science), and I thank him for teaching me how to make a different mistake, so that I can learn the right way on my own.

For their endless support and millions of virtual hugs, thanks to Kristy Meyer, Philip Law, and Christa Oosthuizen.

I am forever indebted to Jeanine Engelbrecht, Wendy Kröger, Tony Chang, and André Faure, for sharing their experiences, motivating me when I needed it most, and for PhD retreating over many margarita jugs, mojitos and sushi; I can not imagine completing my studies without them. To all the ACE people for taking the mac adapter hostage and for the ice-cream invites, I thank them for making me smile when the basement got too lonely.

I have received so many words of encouragement at times when no one could understand how much it meant to me; I appreciate all the moments, particularly the weekends away amongst the trees, where I could escape from work and relax unconditionally.

I fail to find the words that convey my appreciation and gratitude for my best friend, my love, Nicholas Young, who had more faith in me than I deserve. Thank you for being my biggest support, for indulging my science twitterings, and for gently carrying me back to reality when the work became overwhelming. I thank him for always making the most of the times we did spend together, but never asking why it was taking so long.

I am thankful to my brother, Christiaan, for looking after me on all my Florida visits like only a big brother can; I can not express in words the wonderful example and support he has been throughout my life, there is no substitute for a sibling who is ever-missing in presence, and yet so near when I need him.

Finally, I am forever grateful to my parents, Elmien and Malcolm, for creating the environment and providing the opportunities that have made the journey seem so natural. I would

not have made it this far without their emotional and financial support; I am immensely grateful for their continued interest, concern and love.

I also acknowledge my funders for the scholarships and conference subsidies: the South African AIDS Vaccine Initiative (SAAVI), National Bioinformatics Network (NBN), Centre for High Performance Computing (CHPC), Center for HIV/AIDS Vaccine Immunology (CHAVI), and the Centre for the AIDS Programme of Research in South Africa (CAPRISA).

Contents

List of Abbreviations

APOBEC3G	APOlipoprotein B mRNA editing Enzyme, Catalytic polypeptide-like 3G
BEAST	Bayesian Evolutionary Analysis Sampling Trees
BEB	Bayes Empirical Bayes
BSP	Bayesian Skyline Plot
CHAVI	Center for HIV/AIDS Vaccine Immunology
CI	Confidence Interval
CRF	Circulating Recombinant Form
CTL	Cytotoxic T Lymphocyte
DC	Dendritic Cell
EIA	Enzyme ImmunoAssay
ELISpot	Enzyme-Linked Immunosorbent Spot
ESS	Effective Sample Size
FEL	Fixed-Effects Likelihood
GARD	Genetic Algorithms for Recombination Detection
GASP	Gapped Ancestral Sequence Prediction

GTR	General Time Reversible
HIV	Human Immunodeficiency Virus
HLA	Human Leukocyte Antigen
HPD	Highest Posterior Density
Indels	Insertions and Deletions
INI	Integrase Inhibitor
Lowess	Locally-weighted polynomial regression
LRT	Likelihood-Ratio Test
LTR	Long Terminal Repeats
MCMC	Markov Chain Monte Carlo
MHC-I	Major Histocompatibility Complex class one
ML	Maximum Likelihood
MP	Maximum Parsimony
MRCA	Most Recent Common Ancestor
Nab	Neutralising antibody
NEB	Naive Empirical Bayes
NJ	Neighbour-joining
NK	Natural Killer
NNRTI	Non-Nucleoside Reverse Transcriptase Inhibitor
NRTI	Nucleoside Reverse Transcriptase Inhibitor
NtRTI	Nucleotide Reverse Transcriptase Inhibitor

PAMP	Pathogen-Associated Molecular Pattern
PARRIS	PARtitioning for the Robust Inference of Selection
PBMC	Peripheral Blood Mononuclear Cell
PI	Protease Inhibitor
PNGS	Potential N-linked Glycosylation Site
PRR	Pattern Recognition Receptor
REL	Random-Effects Likelihood
RT	Reverse Transcriptase
SGA	Single Genome Amplification
SIV	Simian Immunodeficiency Virus
SLR	Sitewise Likelihood-Ratio
SSP	Sequence-Specific Primer
STCL	Short-Term Cell Line
TH	T Helper
TLR	Toll-Like Receptor
tMRCA	time to Most Recent Common Ancestor
UPGMA	Unweighted Pair Group Method with Arithmic mean

List of Figures

List of Tables

Introduction

The human immunodeficiency virus (HIV) genome is extremely variable, and high levels of recombination further contribute to the number of different forms of the virus currently circulating in the human population. The degree of sequence heterogeneity is of interest for drug and vaccine design since anti-HIV drugs or vaccine-induced immune responses commonly recognise particular amino acid patterns., and broad effectivity depends on the sequence similarity across multiple viral subtypes. The characteristics of the virus around the time of transmission are of particular interest for vaccine and drug development. However, investigations of the sequences of transmitted viruses have been hampered by the lack of available sequence data. HIV samples from the later stages of infection are, therefore, investigated more frequently, and various computational methods have been developed to identify specific characteristics that play a role in viral fitness. These evolutionary approaches are often aimed at modelling the changes that occur along the phylogenetic tree relating the sequences of viral variants. Models of sequence evolution rely on a number of assumptions, for example the absence of recombination, and violation of these assumptions can lead to unreliable results. This thesis explores the diversification of HIV, with particular focus on the early stages of infection and the context under which various evolutionary models are applied.

Chapter 1 describes the general background of both HIV and models of adaptive sequence evolution, and is aimed at presenting an overview of viral features and modelling approaches that are further investigated in the subsequent chapters. The first section provides a background to evolutionary analysis and to models of protein-coding sequence evolution. Common approaches for detecting positive Darwinian selection, and caveats that may bias the

inference of selection, are also discussed. The second half of the chapter describes the HIV genome, the viral life cycle and the different phases of infection. It also covers HIV genome features that are targeted by anti-HIV drugs and the human immune response to HIV infection, illustrating the challenges for effective drug and vaccine design.

Chapter 2 presents an approach to identify the founder virus responsible for productive HIV infection in a new host. The work carried out in Chapter 2 and 3 forms part of a collaboration through CHAVI (Center for HIV/AIDS Vaccine Immunology) . These international collaborative studies were aimed at describing the transmitted virus and identifying the selection pressures present during the earliest stages of infection. In Chapter 2 coalescent-based methods and the use of Bayesian coalescent simulations together with a relaxed molecular clock to estimate various population parameters, are first described. The software package BEAST (Bayesian Evolutionary Analysis Sampling Trees) is then applied to a large collection of HIV datasets derived from sequencing samples isolated from patients in the earliest stages of infection. The aim was to estimate the time to most recent common ancestor (tMRCA) for each individual patient, and, thereby, to classify HIV-1 infections as consistent with homogeneous (if the tMRCA was shorter than the infection time estimated from clinical data) or heterogeneous infection (if the tMRCA was longer than the clinically estimated duration of infection).

In Chapter 3, sequences from the HIV-1 subtype B infections that were inferred to represent homogeneous infections were further investigated. The availability of multiple homogeneous early infection datasets created the opportunity to evaluate the selection pressures present at the point of transmission and during early infection. A method that incorporates the data from all homogeneous infections is described, and was used to establish whether the virus evolves adaptively in early infection. In this way, specific amino acid sites that are evolving under positive selection can be identified, and through experimental evaluation it was possible to establish whether these sites are implicated in viral escape from the immune system. Furthermore, the impact of APOBEC-mediated hypermutation is explored, and the extent to which the data relates to the notion that reduced Env variable loop lengths are a feature of transmission, is discussed.

Chapter 4 expands upon the hypothesis that the length of the envelope variable loops is implicated in a transmission bottleneck. The degree to which sequences with shared genetic history may bias the experiments aimed at describing the association between Env loop lengths and transmission, is investigated by carrying out a simulation study. A novel method that accounts for the shared sequence history is then described. This approach is used to re-evaluate the results from a previous study that applied a method dependent on the phylogenetic tree relating the sequences.

Chapter 5 and chapter 6 investigate the application of codon models to HIV-1 evolution more generally, and not necessarily in the context of early infection. The focus of these chapters is on biases that can affect phylogenetic models of HIV-1 evolution and on the reanalysis of previously published studies, using methods to correct these biases. The extent to which recombination has affected the inference of positive selection in HIV sequence datasets is re-evaluated. Chapter 5 describes features of HIV-1 evolution , for example recombination and overlapping coding regions, that can cause a high rate of false positive inference of positive Darwinian selection. The results of a number of previous studies of adaptive evolution in HIV-1 protein-coding genes are re-evaluated using methods that account for these biases.

Chapter 6 describes the covarion model of sequence evolution where switching between rate classes is allowed. Simulation studies are carried out to determine whether sequence datasets that fit the covarion model of sequence evolution better than the M2a null model, are an artifact of under-fitting of the distribution of rate classes in the null model. The last section of the chapter focuses on the zoonotic event where SIV crossed the species barrier to humans. Ancestral sequences of a collection of SIV and HIV-1 group M sequences are reconstructed, and the sites where switching between rate classes are implicated, are identified. The premise for carrying out this analyses is that sites where selective pressures change along the branch leading to HIV-1 descendent sequences, may be associated with specific adaptation within the new host.

Chapter 1

Background

1.1 Methods of Evolutionary Analysis

1.1.1 Introduction to Molecular Evolution and Phylogenetics

The theory of evolution has enjoyed renewed attention with the 200th celebration of Charles Darwin's birthday and the 150th anniversary of Darwin's book *On the Origin of Species* (Darwin, 1859). Since the initial understanding of evolution, which included ideas regarding the role of pure chance and natural selection (Koonin, 2009), evolutionary analysis has progressed substantially. The progress in computer design and subsequent availability of processing power, the abundance of diverse genomic data, and the development of powerful tools has allowed scientists to re-create lost genetic information and explain the forces that dictate the complex life-forms that exist today.

Researchers in the field of molecular evolution endeavour to understand the changes that render descendants functionally and morphologically different from their predecessors. They are interested in the underlying mechanisms that gave rise to the physical features which define the variation between and within species (Pevsner, 2003). These efforts were benefited significantly by the development of PCR (Mullis et al., 1986), which assured an abundance of

easily-obtainable samples of nucleotide sequence data, and DNA sequencing, which provided means to compare the organisation of genomes within and between species.

Phylogenetic trees are commonly used to represent the shared history of a set of DNA sequences. Genes, made from the four-character DNA alphabet (nitrogen bases: adenine "A", cytosine "C", guanine "G", and thymine "T"), are transmitted from one generation to the next, during which they can undergo numerous changes in their DNA architecture. These changes are reflected in the resulting genealogy, where coalescent events occur at the nodes and sequence changes occur along the branches of the tree. The relationship and difference (distance) between two sequences is apparent from the tree structure. Long branches represent large differences in base composition, whereas short branches indicate limited variation between the parental gene copy and that of its offspring (Galtier et al., 2005).

Different tree-building methods have been developed to illustrate the relationship between different sequences. Two of the main approaches were distance- and character-based methods. Distance-based methods are fast and straightforward, and include UPGMA (unweighted pair group method with arithmic mean) and the Neighbour-joining (NJ) methods. Character based methods include maximum parsimony (MP) and maximum likelihood (ML) methods. Parsimony analysis provides the most parsimonious tree, which describes the sequence relationships with the minimum number of character changes that would result in the observed differences. Maximum likelihood methods on the other hand, produce the tree with the highest probability of resulting in the observed differences (Pevsner, 2003). Maximum likelihood methods were considered as early as the 1960s, although at the time the calculation was too computationally complex, and simpler, approximate methods were used (Huelsenbeck and Crandall, 1997).

Developments in computing power have allowed researchers to use maximum likelihood approaches to implement models of sequence evolution. In the genealogical sense, the likelihood of a phylogeny can be defined as the probability of observing the data, given the specified model parameters. In the maximum likelihood framework, the values that maximise the

probability of the data are chosen (Huelsenbeck and Crandall, 1997). Although maximum likelihood methods are computationally intensive, their flexibility and usefulness for hypothesis testing, together with computational developments allowing faster run times, have ensured that they are commonly used in evolutionary analysis.

Ancient molecular data is extremely sparse and researchers generally only have information from existing organisms; therefore statistical inference is a central component of evolutionary analysis (Nielsen, 2005). Various models of molecular sequence change exist, each with their own underlying assumptions and applicable to specific genomes or genomic regions (Galtier et al., 2005). These models attempt to describe the substitution process that occurred along the branches of a phylogenetic tree. Evolutionary changes are commonly modelled as Markov chains that are stochastic and memoryless, that is past states are irrelevant and future processes are dependent only on the present state. Together with this mathematical quality, a range of different biological constraints are also taken into consideration, resulting in several distinct Markov models of sequence evolution (Yang, 2006; Galtier et al., 2005).

1.1.2 Probabilistic Models of DNA Substitution

The first DNA model was proposed by Jukes and Cantor in 1969 (Jukes and Cantor, 1969). This one parameter model specifies equal base frequencies, i.e. $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$, and assumes that each base has an equal probability of changing to any of the other three bases. However, this assumption is unrealistic since substitutions between bases do not occur at the same rate. Two base classes are distinguished: the double ring purines A and G, and the single ring pyrimidines C and T. Changes within purines or within pyrimidines, called transitions, are more likely to occur than transversions, which refer to changes between purines and pyrimidines (Klug and Cummings, 1999). Kimura's subsequent two-parameter model took the transition/transversion rate bias into account (Kimura, 1980). Following these basic models, parameters that account for unequal base frequencies, models F81 and HKY for example, were also included. (Felsenstein, 1981a; Hasegawa et al., 1985). The most general reversible model that has 9 free parameters, is the GTR (General Time Reversible)

or REV model (Tavaré, 1986). Two of the commonly used models are the HKY and GTR models; the decision in favour of implementing either model depends largely on the size of the dataset and not the specific biological attributes (Delport et al., 2009). As an example, the GTR model of DNA substitution is shown below. The substitution rate matrix Q, with stationary base frequencies $\Pi = (\pi_1,\pi_2,\pi_3,\pi_4)$, is

$$Q = \{q_{ij}\} = \begin{array}{cccc} * & x_1 & x_2 & x_3 \\ \frac{\pi_1 x_1}{\pi_2} & * & x_4 & x_5 \\ \frac{\pi_1 x_2}{\pi_3} & \frac{\pi_2 x_4}{\pi_3} & * & x_6 \\ \frac{\pi_1 x_3}{\pi_4} & \frac{\pi_2 x_5}{\pi_4} & \frac{\pi_3 x_6}{\pi_4} & * \end{array}$$

where the diagonal " * " characters represent the negative sum of the remaining values in each row. The probability of nucleotide i being exchanged for nucleotide j, is the equal to the probability of j being exchanged for i, and it follows that $q_{ij}\pi_i = q_{ji}\pi_j$. There are 9 free parameters in this model - 6 exchangeability parameters ($x_1..x_6$) and 3 free frequency parameters (π_i, where $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$). The HKY model is a special case of the GTR model with a single transition exchangeability parameter, and a separate transversion exchangeability parameter. The HKY model therefore has five parameters - 2 exchangeability and 3 frequency parameters (Tavaré, 1986; Galtier et al., 2005). For smaller datasets the HKY model, with fewer parameters, is more suitable, whereas larger datasets have a better fit to the GTR model (Delport et al., 2009).

The early models assumed that all sites along a sequence are independent and identical, which is unrealistic as spatial variation in the evolutionary rate occurs in almost all sequence data (Yang, 1993, 1994). An alternative is to draw, at random, the value of the rate for each site from a statistical distribution, a practice first introduced by Yang in 1993 (Yang, 1993). Variable rates are typically drawn from a discrete gamma distribution with a specified number of categories, in order to allow for feasible computational times (Galtier et al., 2005; Yang, 1994), and it is also common to allow for a proportion of sites in the sequence to remain invariable. Generally, when the HKY model of DNA substitution is used together

with a certain proportion of invariable sites and the remainder of rates drawn from a gamma distribution, the model is indicated by HKY + I + Γ, where "I" refers to invariable sites and "Γ" to the gamma distribution.

1.1.3 Models of Coding Sequence Evolution

The advances made in DNA and protein (the latter not discussed) substitution models allowed for the further development of codon models that provide means of describing which nucleotide mutations cause changes in protein expression. A DNA mutation does not necessarily reflect a change in the structure or function of the associated protein since, due to the degeneracy of the genetic code, two or more codons can express the same amino acid. Codons 'CGC' and 'CGA' for example both encode Arginine, and a C↔A substitution at the third codon position has no effect on the encoded protein. These changes are called synonymous substitutions, as they are silent at the protein level. If, however, there is a C→A substitution at the first position of codon 'CGC', the resulting codon ('AGC') encodes a different amino acid (Serine), which could have a marked effect on gene expression. These changes are referred to as non-synonymous substitutions. Positive Darwinian selection occurs when alleles that are adaptively advantageous are selectively favoured. Positive selection is often inferred when the non-synonymous substitution rate is significantly greater than the synonymous substitution rate. Purifying selection refers to the case where deleterious mutations are selectively removed from the population, and can be inferred when the synonymous substitution rate is significantly higher than the non-synonymous substitution rate (Yang et al., 2000; Macbeth and Collinson, 2002; Hughes, 1999; Larson, 2006). Coding DNA sequence data therefore provide additional information about evolution that is not contained in non-coding DNA or protein sequence data.

The first models of codon evolution were described by Muse and Gaut (1994), and Goldman and Yang (1994). These models account for the triplet nature of DNA and incorporated parameters to address non-independence between sites (nucleotides within the same codon). Codon models have an alphabet size of 61 ($4 \times (\text{number of nucleotides})^{3\,(\text{word size})} = 64$,

less the 3 stop codons) and are described by an instantaneous transition rate matrix, which contains elements indicating the rate of change from codon i to codon j :

$$q_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ by more than one change,} \\ \pi_j & \text{for a synonymous transversion,} \\ \kappa\pi_j & \text{for a synonymous transition,} \\ \omega\pi_j & \text{for a nonsynonymous transversion,} \\ \omega\kappa\pi_j & \text{for a nonsynonymous transition.} \end{cases}$$

where π_j is the equilibrium frequency of either the target nucleotide or target codon, κ is the transition/transversion rate ratio, and ω is the non-synonymous/synonymous substitution rate ratio ($^{dN}/_{dS}$).

1.1.4 Identifying Positive Darwinian Selection

Omega (ω), the ratio of non-synonymous (dN) to synonymous (dS) substitution rates, provides an important measure of the selection pressures acting on a protein-coding gene. The (non)synonymous substitution rate is defined as the number of (non)synonymous substitutions per (non)synonymous site. If a gene is evolving neutrally, then $\omega = 1$, dN and dS are equal, and all substitutions are fixed at the same rate. If $\omega < 1$, the non-synonymous substitutions are deleterious and there is a reduced fixation rate due to purifying selection acting on the gene. A $\omega > 1$ suggests that the non-synonymous substitutions are favoured, which means that the non-synonymous substitutions are fixed at a higher rate than are the synonymous substitutions ($dN > dS$). Positive Darwinian selection can be inferred when the ratio of non-synonymous to synonymous substitution rates is significantly greater than one (Martin et al., 1998; Yang, 1998; Anisimova et al., 2001; Suzuki and Nei, 2001; Galtier et al., 2005; Yang, 2006).

Two types of codon models are commonly used, empirical and mechanistic codon models. Empirical codon models do not consider differences in amino acid replacement rates that

influence protein evolution beyond the distinction between synonymous and nonsynonymous mutations and differences in nucleotide substitution rates. Since these matrices are large and general evolutionary patterns are thought to be applicable across various datasets, the parameter estimates are often determined once for a large dataset and subsequently used again for further analysis of different data (Kosiol et al., 2007; Anisimova and Kosiol, 2009). The parameters of mechanistic codon models are however estimated for each dataset independently. These models account for a range of evolutionary features, for example, the frequencies of the observed nucleotides or codons (Schneider et al., 2005; Kosiol et al., 2007; Anisimova and Kosiol, 2009).

The original codon models were used to estimate an average ω for all the amino acids in the sequence over the entire evolutionary tree. Positive selection could therefore only be inferred if on average $\omega > 1$; an unlikely scenario given that large regions of protein sequences tend to be subject to purifying selection due to functional and structural constraints, and positive selection is likely to affect only a handful of amino acids at a limited number of time points. The statistical power of the first codon models was therefore inadequate for detecting positive selection in a wide-range of datasets (Yang et al., 2000). Yang et al. (2000) extended the neutral and positive selection models described in 1998 (Nielsen and Yang, 1998) to allow for heterogeneous ω ratios across sites (Yang et al., 2000). Various models were presented, each assuming a unique distribution for ω.

The neutral model (M1) described by Yang et al. (2000) includes two rate classes, $\omega_0 = 0$ and $\omega_1 = 1$, with proportions p_0 and $p_1 = p_0$ - 1, respectively. The selection model (M2) allows for an additional class where $\omega_2 > 1$, with proportion p_2 and requirement for $p_0 + p_1 + p_2 = 1$ (Yang et al., 2000; Nielsen and Yang, 1998; Wong et al., 2004). These models were slightly adapted so as not to constrain $\omega_0 = 0$ but rather allow it to vary between 0 and 1, therefore $0 < \omega_0 < 1$. The modified models (M1a and M2a) were shown to perform better than the original M1 and M2 models (Wong et al., 2004).

Model Comparison

The neutral and selection models can be used to detect whether protein-coding genes are evolving under positive selection. Since the neutral model is a special case of the selection model, a likelihood-ratio test (LRT) statistic can be used to determine which model provides a better fit for the data. As with the M1a and M2a models, many of the evolutionary models are nested within or special cases of another, more complex model; they differ only in the number of estimated parameters. Likelihood-based methods provide means for comparing the fit of two nested models. The LRT considers the likelihood (L_0) of a dataset given the null model (M_0), compared to the likelihood (L_1) of the same dataset given the alternative hypothesis (M_1). M_0 is nested in M_1 and the LRT is defined as

$$LRT = -2\,log\left(\frac{max[L_0(M_0|Data)]}{max[L_1(M_1|Data)]}\right)$$

This statistic is asymptotically X^2 distributed under the null hypothesis, with the number of degrees of freedom equal to the difference in number of free parameters between the two models. Since the null model is a special case of the alternative model L_0 must be less than or equal to L_1. The LRT permits evaluation of the log-likelihood increase brought about by adding additional parameters to a model. A significant increase in log-likelihood would result in the rejection of the null hypothesis. The general critical rejection level is 5%, or a p-value of less than 0.05, which signifies that assuming the null hypothesis is true, the probability of obtaining a result as extreme as the current one is less than 5%. Therefore, the smaller the p-value, the safer it is to reject the null hypothesis in favour of the alternative hypothesis (Buschbom and von Haeseler, 2005; Galtier et al., 2005).

Site-Specific Analysis of Selection

The likelihood-ratio test provides a means to evaluate whether there is evidence of sites evolving under positive selection in a gene or other coding sequences. It does not, however,

provide information on the selection pressure at specific sites in the alignment. This information can be obtained by estimating posterior probabilities for each site belonging to a site class, for example $\omega < 1$, $\omega = 1$ and $\omega > 1$ in the case of M2a. If the posterior probability for a site belonging to the $\omega > 1$ class is high, the site is considered to be evolving under positive selection. One approach to determine the selective pressures at particular sites is to use the maximum likelihood estimates of the proportion of sites in each site class as prior probabilities; these estimates are then used to calculate the posterior probabilities of a site belonging to a specific site class, given the data at that site. This is called the Naive Empirical Bayes (NEB) approach. Since uncertainties in the estimated maximum likelihood values are not accounted for, the subsequent posterior probabilities are dependent on the accuracy of these parameter estimates. The NEB approach therefore relies heavily on the reliability of the maximum likelihood estimates (Delport et al., 2009; Bielawski and Yang, 2005; Kosakovsky-Pond and Muse, 2005). Further fully Bayesian alternatives to this approach include a full Bayesian method using Markov Chain Monte Carlo (MCMC) to approximate the posterior distribution (Huelsenbeck and Dyer, 2004; Delport et al., 2009; Scheffler and Seoighe, 2005), and a method which takes uncertainties in the maximum likelihood estimates into account by integrating over assigned priors for these parameters (hierarchical Bayes or Bayes Empirical Bayes, BEB) (Yang et al., 2005; Delport et al., 2009).

The methods described above model site-to-site parameter variation using an approach known as the random-effects likelihood (REL) model. REL assumes that the substitution rates across sites are random values drawn from a given distribution (Llopart and Comeron, 2008; Delport et al., 2009). The alternative approach is to treat the parameters as fixed, and as belonging to specific classes with distinct parameters. This is referred to as the fixed-effects likelihood (FEL) model. Several site classes can be defined, even to the extent of allowing each site in the sequence to be associated with an independent value of ω. However, the latter scenario with the maximum number of site classes, requires the estimation of a large number of parameters and therefore renders the model susceptible to over-parameterisation (Delport et al., 2009). An advantage of the FEL approach is that the degree to which the branch length parameters, synonymous and non-synonymous substitution rates, κ values (transition/transversion rate ratios), as well as codon frequencies, are

shared across classes, can be defined prior to analysis. This saves a great deal of computational time and prevents potential convergence problems when estimating the maximum likelihood values (Kosakovsky-Pond and Frost, 2005; Delport et al., 2009).

1.1.5 Violation of the Assumptions of Phylogenetic Models for Detecting Positive Selection

All models, regardless of the complexity, input data, or computational requirements, make certain assumptions. If these assumptions are violated the results are often meaningless. One assumption that most phylogenetic models of sequence evolution make is the absence of recombination. This means that a single phylogenetic tree is considered to represent the relationship between all the taxa. However, when the linkage between sites is disrupted by recombination, the single tree assumption does not hold since this results in several distinct phylogenetic trees describing the relationships between the taxa in different parts of the sequence (Shriner et al., 2003). Recombination is known to occur in many organisms, including viruses, and neglecting to account for recombination can result in a false positive detection of positive selection (Anisimova et al., 2003; Shriner et al., 2003). Since sites found to evolve under positive selection are thought to be important for viral escape from the immune system, misidentification of adaptively evolving sites could have a negative impact on drug design and vaccine development. In 2006, a robust method to detect positive Darwinian selection called PARRIS (PARtitioning for the Robust Inference of Selection) was developed that can be applied to recombining sequences (Scheffler et al., 2006). Preliminary analyses indicated that the sites inferred to be evolving under positive selection in certain genes differed greatly when recombination was taken into account, and the authors recommend that a method which screens for recombination should be mandatory in analyses of datasets, such as HIV, where recombination is known to occur (Scheffler et al., 2006). Chapter 5 covers the reanalysis, using PARRIS, of a collection of published HIV datasets for which positive selection has been described previously.

The site-to-site synonymous substitution rate is often treated as a single parameter; the

same, constant value is therefore shared over all the sites in the sequence (Delport et al., 2009). Substantial synonymous substitution rate variation has, however, been shown to occur in RNA viruses as well as other coding sequences, including sperm lysin, β-globin, and primate COXI (Hanada et al., 2004; Kosakovsky-Pond and Muse, 2005). Synonymous substitution rate variation has specifically been identified in HIV (Lemey et al., 2005; Kosakovsky-Pond and Muse, 2005; Ngandu et al., 2008). This feature could reflect a variation in the mutation rate if the synonymous substitution rate is believed to approximate the neutral rate of evolution, or could also indicate nucleotide-level selection pressures acting to retain essential viral functions (Ngandu et al., 2008). Regardless of the underlying reason for the synonymous substitution rate variation, if it is not accounted for by the model of coding sequence evolution, then the comparison of dN and dS is invalid and the inference of selection may be incorrect (Kosakovsky-Pond and Muse, 2005; Ngandu et al., 2008). Allowing site-to-site synonymous substitution rate variation is therefore an additional consideration during selection analysis of protein-coding datasets.

A further source of bias is the non-independence between nucleotide sites. The non-independence relates to both within codon nucleotides, where the triplet nature of the genetic code means that individual sites do not evolve independently, as well as to nucleotides spread across a genome. The latter case is harder to account for than within codon non-independence. For HIV in particular, overlapping coding regions, cis-regulatory regions, potential N-linked glycosylation sites and specific three-dimensional protein folding, contribute to the non-independence between nucleotide sites (Stern and Pupko, 2006; Poon et al., 2007; Mayrose et al., 2007), and these features violate evolutionary model assumptions in many cases. Sequence data from various sources are unlikely to fit any one evolutionary model perfectly. Deciding on the most appropriate model, or adapting an existing model, is important for obtaining meaningful selection results from evolutionary modelling.

1.1.6 Covarion Models of Sequence Evolution

In certain contexts, for example when recombination has occurred between sequences, positive selection can be very hard to detect accurately (Scheffler et al., 2006). The general comparison of the likelihoods between a model allowing sites to evolve under positive selection (as well as neutrally and under purifying selection) and a model that only allows neutral or purifying selection, is often not sufficient. The covarion approach is one extension to the general method that allows for the detection of sites that may have been selected for in only one part of the phylogeny, and where the signal for selection may be too weak and would be missed if tested with the common approach (Huelsenbeck, 2002; Galtier, 2001).

As described, the traditional models of codon evolution allow different selective processes at all the sites along a sequence (Nielsen and Yang, 1998; Yang et al., 2000). This allowed for the development of evolutionary models that provided a more realistic representation of sequence evolution. With these models, a site is found to be evolving under positive selection if the average ω is greater than one over all lineages. Positive selection would therefore not be detected if adaptive evolution took place in a short branch of the tree and only affected a couple of protein sites. Nevertheless, averaging the variation in rate over lineages was thought to be less of a concern than averaging the rates across sites. It is, however, unrealistic to assume that the selection regime at a given position along the sequence remains constant through time (Yang and Nielsen, 2002).

Covarion models of sequence evolution were developed to allow site-specific rate variation among lineages. The first formal covarion model was described by Tuffley and Steel, where a site could either vary ("on" state) or not ("off" state) (Tuffley and Steel, 1998). The model was later extended by Huelsenbeck to allow among-site rate variation; each site was therefore either invariable, or evolving at a rate drawn from a discrete gamma distribution with a given number of rate classes (Figure 1.1, Huelsenbeck, 2002).

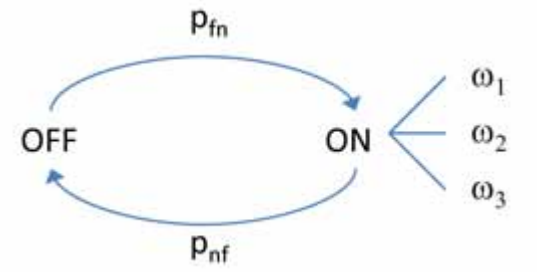


Figure 1.1: Rate switching in the Huelsenbeck model. A site is either invariable (OFF) or is allowed to vary according to a number of rate classes, $\omega_1 - \omega_3$. The probabilities of switching from the ON to OFF state is p_{nf}, and from the OFF to ON state p_{fn}.

Another adaptation to the original model was implemented by Galtier in 2001. Here, a proportion of sites evolve under the covarion model whereas the remainder do not (Figure 1.2). The site-specific rates of the latter sites (those not evolving under covarion evolution) are drawn from a discrete gamma distribution and do not change over time. The rates for the covarying sites are also drawn from this gamma distribution and fall into a given number of rate classes, and covary, that is switch between rate classes, at a constant rate (Wang et al., 2007).

Figure 1.2: Rate switching in the Galtier model. A proportion of sites are invariable (OFF) and the remaining proportion is allowed to vary (ON). Site that fall within the ON class, are allowed switch between a number of rate classes, $\omega_1 - \omega_3$, with equal probabilities.

These two models supplied means of obtaining an estimate of which, and how many, sites are undergoing episodic adaptive evolution. Huelsenbeck and Galtier's analyses provided evidence that covariation was taking place and argued that the conventional models were underestimating the degree and importance of positive selection (Huelsenbeck, 2002; Galtier, 2001). Despite the apparent value of covarion models of sequence evolution, they have been

applied to only a limited extent in studies of molecular sequence evolution.

1.2 HIV/AIDS

1.2.1 A Brief History and Background of HIV/AIDS

Almost twenty-eight years after Michael Gottlieb described the symptoms of this immunocompromising disease amongst homosexual men (Gottlieb et al., 1981), the human immunodeficiency virus (HIV) continues to challenge scientists worldwide. An estimated 33 million people were living with HIV in 2007; of this total, 50% were infected before they turned 25, and a significant majority of these individuals will die, or have died, before the age of 35 (UNAIDS, 2008). A global view of the HIV adult prevalence rates reveals how the low- and middle-income countries are the hardest struck by this pandemic (Figure 1.3, UNAIDS, 2008).

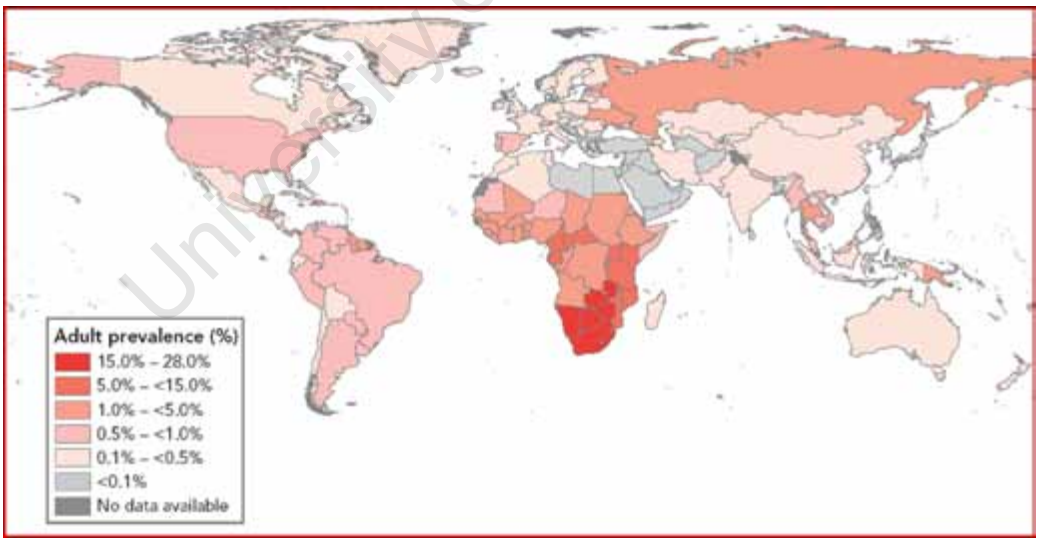


Figure 1.3: A global view of HIV infection in 2007.

1.2.2 The HIV Genome

HIV is part of the genus *Lentivirus* (of the *Retroviridae* family), and the RNA genome of HIV-1 consists of 9 genes: *gag*, *pol*, *env*, *tat*, *rev*, *vif*, *vpr*, *vpu*, and *nef* (Levy, 2007). HIV-2 does not have *vpu*, but a further gene, *vpx*, is present. The main functions of each of the 9 HIV-1 proteins are shown in Table 1.1 (Levy, 2007; Kuiken et al., 2008). HIV-1 is thought to have been transmitted from chimpanzees to humans, whereas HIV-2 is believed to have crossed the species barrier from sooty mangabeys, found in Western Africa (Gao et al., 1999; Newman et al., 2006). HIV-2 is less pathogenic and less prevalent than HIV-1 (Reeves and Doms, 2002). Both HIV-1 and HIV-2 share tropism for T lymphocytes (key cells of the human immune system), predominantly infecting CD4 T lymphocytes (helper cells) and CD8 T lymphocytes (killer cells) respectively (Brites et al., 2009).

Table 1.1: The nine HIV-1 proteins and the main function(s) of each.

Proteins	Main Functions
Gag	
CA	Capsid protein, structural protein
MA	Matrix protein, membrane anchoring, Env interaction
NC	Nucleocapsid, binds RNA
Pol	
Protease (PR)	Post-translational processing, Gag/Pol cleavage and maturation
Integrase (IN)	Proviral cDNA integration
Reverse Transcriptase (RT)	Reverse transcription
Envelope (Env)	Envelope surface (gp120) and transmembrane (gp41) proteins, bind to CD4 cells and secondary receptors
Tat	Plays a role in transcriptional transactivation
Rev	Regulation of viral RNA transport and stability
Nef	CD4 downregulation
Vif	Increases the viral infectivity, cell-to-cell transmission, and maturation
Vpr	Promotes nuclear localisation of the preintegration complex
Vpu	Plays a role in virus release

The genome size of HIV-1 is approximately 10kb, and a single virion is about 100nm in diameter (Burmester et al., 2003). HIV-1 is divided into four groups, M, N, O, and P (Plantier et al., 2009); group M consists of several different viral subtypes, including A1, A2, B, C, D, F1, F2, G, H, J, and K. There are also continually increasing numbers of circulating recombinant forms (CRFs), which are intersubtype recombinant viruses that have formed a significant epidemic distribution (Robertson et al., 2000). The combined variability of CRFs

add to the complexity of developing a broadly effective vaccine.. Furthermore, the genetic differences between group M, N and O viruses can extend up to 30% in *gag*, and 50% in *env* (Gao et al., 2005).

1.2.3 The HIV Life Cycle and Disease Progression

There are a number of steps that need to take place for HIV to infect a cell and reproduce. The life cycle includes attachment of the virus to a CD4 receptor of an immune cell, fusion in a pH independent manner, entry of the viral RNA and reverse transcription to produce a DNA copy (cDNA) of the viral genome, integration of the cDNA into the host chromosomal DNA, production of viral mRNA transcripts, translation to produce HIV proteins, assembly of the viral proteins near the cell membrane, and finally budding of the new mature HIV virus, which then ultimately attaches to another cell surface receptor, and so the cycle is repeated. Through this process, millions of virions can be produced per day, destroying the human immune system (Levy, 2007; Klimas et al., 2008).

HIV-1 infection, progression to disease, and finally AIDS follows a well established pattern (Figure 1.4[1]). Upon initial infection there is a significant drop in CD4 T cells, which are the primary target of HIV-1 infection, until immune responses are able to control viral replication and stabilisation occurs during which the number of viral RNA copies and CD4 cell count remains relatively stable. The asymptomatic period (Clinical Latency phase in Figure 1.4) during which the viral load maintains a relatively steady state, is often referred to as the viral set-point (Mei et al., 2008; Levy, 2007).

[1]Figure from http://en.wikipedia.org/wiki/HIV_infection

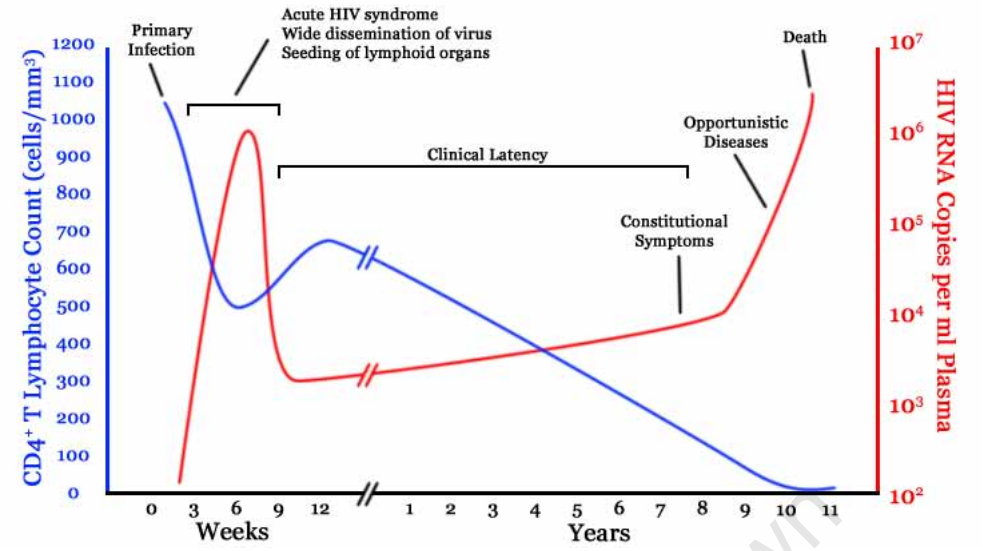


Figure 1.4: HIV disease progression indicating the decline in CD4 T cell counts and steady increase in HIV RNA copies.

The period between infection and AIDS differs vastly amongst infected individuals. Certain individuals remain symptomless for over 20 years, whereas others can die within 2 years of the original infection. Regardless of the duration of clinical latency during which an infected person can appear completely healthy, once the immune response is no longer able to keep viral replication in check, the progression to AIDS is inevitable without drug treatment. Antiretroviral therapy is typically introduced when the CD4 count is less than 200 cells/mm^3 (Klimas et al., 2008; Levy, 2007).

A classification method, which characterises the stage of HIV infection, is important for inter-study comparative analysis and for identifying recently infected individuals for research purposes. Such a classification method that uses different viral markers to define disease stage, was described by Fiebig et al. in 2003. The stages are distinguished based on laboratory evidence of the progressive appearance of viral markers, including viral RNA, the presence of p24 antigen, and detecting anti-HIV antibodies (seroconversion), using enzyme immunoassays and Western blot. Data was obtained from serial plasma donors with a median interval between donations of 4 days; the reactivity to the assays were highly consistent among donors, which resulted in the classification of six primary HIV-1 infection stages

(Table 1.2[2]). As seen in Table 1.2, the antibody enzyme immunoassays (EIAs) were carried out with both the second-generation indirect sandwich EIA, which is not IgM-sensitive, as well as the IgM-sensitive third-generation HIV-1/HIV-2 antigen sandwich EIA (Fiebig et al., 2003). A parametric Markov model was applied to the data in order to obtain estimates of the duration of each stage, including the 95% confidence intervals (CIs) for the estimates (Table 1.2).

Table 1.2: Classification of primary HIV infection based on laboratory assays, first described by Fiebig et al. (2003). Positive (+) and negative (-) tests to the various assays are shown in red.

			Antibody Enzyme Immunoassay			Duration In Days (95% CI)	
Stage	RNA	p24 Antigen	Not IgM-Sensitive	IgM-Sensitive	Western blot	Individual	Cumulative
I	+	–	–	–	–	5.0 (3.1, 8.1)	5.0 (3.1, 8.1)
II	+	+	–	–	–	5.3 (3.7, 7.7)	10.3 (7.1, 13.5)
III	+	+	–	+	–	3.2 (2.1, 4.8)	13.5 (10.0, 17.0)
IV	+	+/–	–	+	Indeterminate	5.6 (3.8, 8.1)	19.1 (15.3, 22.9)
V	+	+/–	+/–	+	+ (without p31 band)	69.5 (39.7, 121.7)	88.6 (47.4, 129.8)
VI	+	+/–	+	+	+	Open-ended	Open-ended

The first four primary stages are very short and, on average, stage V is reached within a month of the date of infection. No cut-off was defined for stage VI, which stretches from recent to early chronic infection. These staging criteria are useful for characterising HIV primary infections and thereby allowing comparisons between different studies where the infections share viral assay profiles (Fiebig et al., 2003). Several studies have used Fiebig classification for, for example, staging primary HIV infections, and/or estimating the time of HIV transmission (Novitsky et al., 2009; Abrahams et al., 2009; Haaland et al., 2009; Keele et al., 2008).

Disease progression has been shown to be impacted by HIV dual infection, where two divergent strains establish HIV infection in an individual (Chohan et al., 2005b; Gottlieb et al., 2004; Grobler et al., 2004). Furthermore, extensive inter- and intra-subtype recombination has been observed in HIV-1 infected patients (Robertson et al., 2000; Rousseau et

[2]Table adapted from Fiebig et al. (2003).

al., 2007; Abecasis et al., 2007). The implications of dual infections and the added recombination events are widespread (Gottlieb et al., 2004; Grobler et al., 2004); for example, dual infection is associated with a rapid progression to disease, and also has a large impact on epidemiological assessment (Steain et al., 2004). Moreover, phylogenetic analysis of recombinant sequences can lead to incorrect estimates of the rate of HIV evolution, an overestimation of positively selected sites during selection analysis, as well as impact the inference of the most recent common ancestor (MRCA) (Rousseau et al., 2007; Anisimova et al., 2003). Drug resistance and greater viral fitness as result of recombination of two or more pre-existing viral strains is an additional important consideration for broad-based vaccine development (Gottlieb et al., 2004). Mutations that result in escape from the human immune system further impact HIV-1 disease progression (Chopera et al., 2008). Escape mutants may have lower fitness compared to the wild-type, and this has been linked to a slower rate of disease progression (Chopera et al., 2008; Goulder et al., 1996)

1.2.4 Drug and Vaccine Targets

The HIV life cycle presents a number of potential vaccine and drug targets; 25 anti-HIV compounds have been developed and approved since the 1987 when zidovudine was first licensed (de Clercq, 2009; Stürmer et al., 2009). Anti-HIV agents fall into different groups depending on their mechanism of action: nucleoside reverse transcriptase inhibitors (NRTIs), nucleotide reverse transcriptase inhibitors (NtRTIs), non-nucleoside reverse transcriptase inhibitors (NNRTIs), protease inhibitors (PIs), fusion and co-receptor inhibitors, and integrase inhibitors (INIs) (Levinson, 2006; de Clercq, 2009; Stürmer et al., 2009). In 2006 it was estimated that anti-HIV compounds had collectively saved 3 million years of life in the United States (Walensky et al., 2006)

Antiviral drugs thus far target the viral reverse transcriptase, protease and integrase, which are encoded by the *pol* gene, as well as Env, which is targeted by the fusion inhibitor and co-receptor inhibitors (de Clercq, 2009; Stürmer et al., 2009). Anti-HIV drug and vaccine research covers a broad field of study, but particular focus is given to *env*, as it is responsible

for the initial interaction with the host cells and is thus essential for transmission (Doms and Moore, 2000).

Env is a trimeric glycoprotein composed of three gp120 and gp41 heterodimers that are exposed on the viral surface and are anchored to the viral particle by the transmembrane domain of the gp41 subunit. Viral cell entry is initiated by the binding of gp120 to the CD4 receptor, followed by conformational changes that allow CCR5 or CXCR4 coreceptor binding (Wyatt and Sodroski, 1998), which then induces further conformational changes in gp120 that trigger the activation of gp41 (Blish et al., 2007; Hollier and Dimmock, 2005; Sattentau, 1998). Activated gp41 leads to fusion of the viral and cell membranes, followed by release of viral RNA into the target cell (Hollier and Dimmock, 2005; Root and Hamer, 2003).[3]

The development of vaccines capable of eliciting antibodies against Env that can block viral entry is complicated by the vast sequence variability between viruses, which in *env* can differ up to 50% between viral groups (Hollier and Dimmock, 2005; Johnston and Fauci, 2007; Gao et al., 2005). Furthermore, adding to the complexity are escape variants arising through the high mutation and recombination rates coupled with immune selection, as well as by the large number of N-linked glycans in gp120 that provide the virus with a protective coat (Derdeyn et al., 2004). In addition, gp120 consists of five highly variable regions (V1-V5) interspersed among more conserved regions (C1-C5) (Burton et al., 2005); of these, V1, V2, V4 and V5 are notable for rapid shifts in length, number and localisation of glycosylation sites (Blay et al., 2006; Zhang et al., 2004). The variable loops shield more conserved receptor binding regions (Wyatt et al., 1998), and only when the CD4-gp120 binding takes place, do structural changes expose previously masked epitopes and surfaces (Zhang et al., 1999; Rits-Volloch et al., 2006). [4]

The HIV genome diversity within and between subtypes, together with the presence of circulating recombinant forms (CRFs) arising due to multiple infection and recombination, lends further support to the argument that an anti-HIV vaccine needs to be effective against

[3]Text adapted from Wood et al., 2009.
[4]Text adapted from Wood et al., 2009.

a very wide range of HIV variants (Chohan et al., 2005b; McBurney and Ross, 2008).

1.2.5 The Immune Response to HIV Infection

1.2.5.1 Innate Immune Characteristics During HIV Infection

The specificity of the immune response, as well as the strength of the attack, against HIV varies depending on the stage of infection. Innate immunity, which provides the initial, rapid, defense against any foreign invading organism, is also the first point of prevention, besides surface barriers, against HIV infection. These defenses include cytokines (signalling proteins), dendritic cells (DCs), macrophages, natural killer (NK) cells, and soluble components, for example mannose-binding lectins and complement (Levy, 2007; Vander et al., 2001a; Goepfert, 2003).

The NK cells and DCs constitute two of the fundamental cellular factors of the innate immune system (Fortis and Poli, 2005); NK cells have inhibitory as well as activating cell-surface receptors that through interactions regulate, for example, NK cell-mediated cytotoxicity of virus-infected cells. DCs initiate various innate as well as adaptive immune responses and, together with macrophages, have important roles as antigen-presenting cells (Levy, 2007). DCs and other mucosal surface cells, including Langerhans cells and macrophages, recognise HIV as a foreign organism and secrete cytokines, and these interactions may influence the replicative capacity of the virus in the genital tract and other mucosa (Lehner, 2003; Levy, 2007). Interestingly, because DCs are present in the mucosal epithelia in the genital tract, and are able to bind Env, it has been suggested that these cells may be the first cells to take up HIV (Levy, 2007; Piguet and Steinman, 2007; Lama and Planelles, 2007).

Since the innate immune cells recognise pathogen-associated molecular patterns (PAMPs) and not specific epitopes, the innate response to invading pathogens is rapid (Levy, 2007). Specific pattern recognition receptors (PRRs), including the Toll-like receptors (TLR), recognise PAMP and act as sensors. PRRs are found on many cells of the immune system, for

example macrophages, DCs, and NK cells. These pathogen sensors activate a response pathway that leads to cytokine and chemokine production, which in turn elicits a wide-range of immune-modulating processes (Levy, 2007; Katsikis et al., 2007). For example, the production of cytokines results in the further activation of natural killer (NK) cells and macrophages, and in turn this generates inflammatory responses and cytotoxic processes (Ekene, 2008; Levy, 2007). NK cells also produce cytokines that effect the adaptive immune response, and therefore play a role in both the primary as well as the secondary response to HIV infection. Similarly, DCs play a large role in regulating the initial adaptive immune response through signals resulting in B and T cell activation, resulting in further immune pathways which are discussed in section 1.2.5.3 (Levy, 2007).

The initial immune response against the invading virus can have a large effect on the ability of HIV to successfully infect an individual. The innate immune system can induce cell-lysis and destroy virus-infected cells, and can also activate important adaptive immune responses. Moreover, a strong innate immune response has been associated with a slower rate of progression to AIDS (Levy, 2007; Borrow and Bhardwaj, 2008).

1.2.5.2 Host-specific Factors Influencing the Course of HIV Infection

Although there are many host-specific factors that can alter the expected prognosis for an HIV infected individual (Lama and Planelles, 2007; Kumar et al., 2006), only a short overview is presented here, with particular focus given to the anti-HIV protein, APOBEC. Human genes that disrupt the infectivity of the virus are valuable to our knowledge of how HIV has evolved to evade destruction by the immune system. Several human factors have been identified that have an impact on HIV's replication capacity to different degrees (Lama and Planelles, 2007). A well-known polymorphism that leads to resistance to HIV-1 infection is the homozygous allelic variation CCR5Δ32. In individuals who have this form of the allele, the CCR5 coreceptor is mutated to such a degree that the virus can not enter the host cells via this route. HIV-1 strains that use the CXCR4 coreceptor can still infect these individuals and the resistance to infection is therefore not guaranteed (Lama and Planelles,

2007; Levy, 2007).

Certain allelic forms of, for example, TRIM5α, RANTES, and various other chemokines, have been associated with varying rates of disease progression (Lama and Planelles, 2007; Newman and Johnson, 2007). Furthermore, DC-SIGN, which is present on immature DCs and activated B cells, binds to HIV via particular carbohydrate moieties present on Env gp120, and although the exact mechanisms of action are not yet known, certain polymorphisms within the promoter region of DC-SIGN have been associated with increased susceptibility to HIV infection (Lama and Planelles, 2007; Hong et al., 2002). In contrast, Langerin, which is expressed on Langerhans cells and found throughout the mucosal surfaces where HIV-infection occurs, has been linked to prevention of HIV-transmission (Lama and Planelles, 2007). Human defensins, which are grouped into α- β- and Θ-defensins according to the distribution and size of their disulphide bridges, are also anti-HIV factors that combat HIV in various ways in order to prevent infection, although there have been reports that certain defensins can enhance infection (Ding et al., 2009). Certain α-defensins can act to down-modulate CD4 expression, or disrupt the viral membrane and thereby inactivate the virus (Lama and Planelles, 2007; Ding et al., 2009). A particular homozygous polymorphism in the DEFB1 gene that encodes the β-defensin, HBD-1, has been linked to exposed seronegative individuals, suggesting a degree of resistance in individuals harbouring this polymorphism (Zapata et al., 2008; Ding et al., 2009). It is evident that the genetic variation found from one infected individual to another has a considerable effect on HIV-transmission and manifestation of disease; research on how an individual's genetic background can influence the infectivity of HIV therefore plays a big role in understanding natural complete, or partial, immunity. A recently identified human factor that can also influence the course of infection is the anti-HIV cellular protein APOBEC3G (Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G) (Chiu and Greene, 2008; Williams and Burdo, 2009). This protein is from the Apobec superfamily that all have cytidine deamination capacity (Kobayashi et al., 2004). APOBEC3G induces cytosine to uracil (C-to-U) changes in the reverse transcript resulting in guanine to adenine (G-to-A) hypermutation in the sense-strand (Jern et al., 2009; Yu et al., 2003). These hypermutations change the genetic composition of HIV and in many cases render the virus non-viable (Zhang and Webb, 2004).

Shortly after APOBEC3G was discovered, another protein, APOBEC3F, was also shown to have similar activity (Zheng et al., 2004; Pillai et al., 2008). It is possible to approximate the effect of APOBEC by the extent of hypermutation observed, and thus allows a valuable measure of the role APOBEC3G/F has in conferring protection against HIV infection.

Notwithstanding this innate defense mechanism against infection, HIV evolved a mechanism to evade APOBEC-induced hypermutation. The viral accessory protein Vif (virion infectivity factor) causes degradation of the APOBEC3G/F proteins in the host cells, enabling viral replication (Goila-Gaur and Strebel, 2008; Jern et al., 2009). The interplay between viral *vif* and human APOBEC provides an additional target for potential antiviral compounds, since a strong inhibition of Vif would allow cytidine deaminase activity to control viral replication (Russell et al., 2009; Goila-Gaur and Strebel, 2008). A recent study suggests that such antiviral avenues require careful consideration. G-to-A hypermutations are frequently observed in drug resistant HIV strains, and it is therefore feasible that the virus uses such increases in genome variation to evolve resistance to both innate antiviral factors as well as drug treatments (Berkhout and de Ronde, 2004; Chiu and Greene, 2008). Drugs or vaccines mimicking APOBEC3G/F activity would have to be thorough and absolute since incomplete functionality could ultimately drive HIV evolution (Chiu and Greene, 2008).

1.2.5.3 The Adaptive Immune Response to HIV Infection

The adaptive immune response specifically recognises foreign invading organisms and acts through lymphocytes to eliminate the antigens (and therefore the pathogens) selectively. Two types of lymphocytes are distinguished: B cells (humoral response) and T cells (cell-mediated response) (Levy, 2007). Humoral immunity involves antigen-recognition and the subsequent maturation of B cells into antibody producing cells, as well as the establishment of a reservoir of memory cells. (Porth, 2007). Therefore, when B cells attach to a foreign antigen they develop into mature memory B cells or plasma cells. Memory B cells ensure that the future recognition and eliminating response against a specific pathogen is rapid, since their primary maturation was initiated by the same pathogen antigen. Plasma cells are

responsible for producing antibodies that specifically recognise the originally encountered antigen. Antibody production is mediated by T helper cells (described below), and the production of sufficient amounts of antibody against a specific invading organism can take several days after it is first recognised by B cells (Vander et al., 2001a; Levy, 2007). In HIV-infection, the envelope protein is the main target for inactivation through neutralisation by antibodies (Levy, 2007; Sattentau et al., 1999).

Neutralising Antibodies Targeting HIV

Research suggests that an effective HIV-vaccine would have to induce a broadly neutralising antibody (NAb) response as well as stimulate vigorous HIV-specific T-cell activity (Johnston and Fauci, 2008; Haynes and Montefiori, 2006; Hessell et al., 2009). Neutralising antibodies compete with CD4 cells to bind to Env gp120, and the ease at which the NAbs outcompete the CD4 cells determines the antibodies' neutralising efficiency (Prabakaran et al., 2006; Nabel, 2002; Korber et al., 2001). Given the high genetic diversity of HIV, the ability to elicit an antibody response that recognises a wide array of viruses is not an easy task, and only a few broadly effective neutralising antibodies have been identified (Scheid et al., 2009; Dhillon et al., 2007).

A recent screening of the neutralisation breadth of sera samples obtained from approximately 1800 HIV-1 infected individuals, lead to the identification of a donor that displayed potent antibody mediated neutralisation (Walker et al., 2009). Two antibodies, P9 and P16, that elicited a broadly reactive neutralising response were identified during the analysis of the sample from this HIV-1 subtype A infected donor. These antibodies primarily recognised conserved regions of the V2 and V3 loops, and displayed preferential binding to a subunit of the stable trimeric form of Env (Walker et al., 2009).

Previous studies have also focused on identifying monoclonal antibodies (mAbs) that bind to complete viral particles and complex epitopes, as opposed to soluble viral proteins (Moulard et al., 2002; Gorny et al., 2005). The first mAb identified in this way, was mAb 2909, which had potent neutralisation potential and recognised a composite epitope consisting of parts

of V2, V3 and the CD4 binding region (Gorny et al., 2005). However, the activity of mAb 2909 was specific to only certain virus strains (Gorny et al., 2005). These studies suggest that the V2 and V3 loops contain conserved regions that can serve as targets for neutralising antibodies, and in addition, antibodies that target these regions elicited a broadly reactive neutralisation response (Walker et al., 2009). These results are therefore very promising for future vaccine research, since vaccine-induced antibodies with similar breadth and potency may have the potential to inactivate HIV upon entry into the host (Walker et al., 2009).

Therefore focussing on the envelope gene, and establishing its transmission characteristics, adaptation to immune pressures, and evolutionary constraints, forms a critical part of the endeavor to develop a successful HIV vaccine (Derdeyn et al., 2004; Mahalanabis et al., 2009; Zhou et al., 2007).

The T Cell Response to HIV Infection

The second class of cells associated with the adaptive immune response, are T cells. These cells are also able to distinguish self from non-self cells, and are divided into three types: cytotoxic T cells, helper T cells, and regulatory T cells (not discussed here). Cytotoxic T lymphocytes (CTLs) express the CD8 surface protein and function by attacking and destroying cells displaying non-self antigens. CTL activity has also been observed for cells expressing the CD4 protein on their surface membranes, however, CD4+ cells predominantly have helper T cell activity (Levy, 2007; Vander et al., 2001a; Ye et al., 2008). T helper (TH) cells, similar to B cells and CTLs, bind to specific antigens, and subsequently undergo activation. Once activated, TH cells secrete cytokines which stimulate antigen-bound B cell and CTL functioning, as well as cytokines that mediate the inflammatory response (Vander et al., 2001a). T helper cell activity is essential for B cell and CTL functioning (Levy, 2007; Vander et al., 2001a; Ye et al., 2008).

1.2.5.4 Human Leukocyte Antigen Diversity and the Course of HIV Infection

Whereas neutralising antibodies prevent infection, CTLs are responsible for recognising and destroying cells that have been infected by a pathogen. When a cell is infected by HIV, peptide fragments (epitopes) of the invading pathogen are presented as antigens on the surface of the host cell, bound to the major histocompatibility complex class one (MHC-I) molecules. The genes encoding MHC molecules in humans, human leukocyte antigen (HLA) genes, are located on chromosome 6 and the HLA complex extends to almost four million DNA base pairs. The HLA class I antigens are encoded by different allelic forms of the three major HLA genes, HLA-A, -B, and -C, and the three minor genes HLA-E, -F, and -G. (Vander et al., 2001b; Marsh et al., 2000). In 2008, more than 230 HLA-A alleles and more than 300 HLA-B alleles were described (Holdsworth et al., 2009), and collectively, over 1500 class I molecules exist, demonstrating the highly polymorphic nature of HLA molecules present in the human population (Nielsen et al., 2008). The diversity of HLA allotypes, and hence peptide-binding specificity, dictates the range of epitopes that can be bound and presented to resist infection by pathogens (Trachtenberg et al., 2003).

Various host genetic factors, in addition to those mentioned in the innate immune section (section 1.2.5.2), have been associated with HIV pathogenesis and progression to AIDS (Kaur and Mehra, 2009; Roger, 1998; Singh and Spector, 2009). Specific HLAs are known to either increase (Roger, 1998; Qing et al., 2006), or decrease (Cornelissen et al., 2009; Crawford et al., 2009) disease progression. Individuals with a large degree of heterozygosity at HLA-I loci tend to produce broader and stronger CTL responses against invading pathogens (Carrington et al., 1999; Kimman, 2001). A relationship between the HLA-B*57 allele and long-term nonprogression has been described (Langford et al., 2007; Miura et al., 2009), and varying degrees of protection against disease progression have been associated with, for example, the HLA-B*5801 and HLA-B*51 alleles (Langford et al., 2007; Brumme et al., 2008; Delport et al., 2008; Chopera et al., 2008). CTL escape mutations that lead to viruses with lower reproductive fitness have also been described for individuals carrying either the HLA-B*57 or HLA-B*5801 alleles (Chopera et al., 2008; Martinez-Picado et al., 2006; Crawford et al., 2007). Due to the fitness cost incurred as result of these CTL escape mutations, the HLA

background can have a considerable impact on HIV disease progression (Chopera et al., 2008).

The host genetic background is currently the only factor that can consistently predict the long-term prognosis of infected individuals (Levy, 2007), and further exploitation of the known effective anti-HIV immune responses may be useful for developing approaches for controlling HIV infection and progression. Moreover, although host factors can not be altered, research on individual genetic differences and the corresponding variation in control of HIV-infection could provide a valuable resource for establishing patient-specific therapeutic guidelines (Langford et al., 2007). Individuals with natural complete immunity, or partial immunity, are a further important resource for studying the interactions between the infected host and virus, and investigating these interactions could lead to novel vaccine strategies (Kaur and Mehra, 2009; Langford et al., 2007).

Progress in developing an anti-HIV agent is reliant on an understanding of host genetic factors related to HIV-infection, HIV specific viral characteristics, and the dynamics of the host-virus interaction. An effective anti-HIV vaccine needs to recognise and remove the virus that is transmitted (Keele et al., 2008), before infection is established and viral evolution enables evasion of the vaccine response. It is therefore also important to identify viral features that are essential for successful transmission from one individual to another. HIV samples obtained from newly-infected individuals are ideal for studying transmission characteristics, since these sequences represent the HIV population capable of infecting new hosts and surviving the earliest immune responses elicited against them. Therefore, investigating the events that occur during transmission and the earliest stages of HIV-infection, together with the known host specific factors and viral dynamics, could provide significant insights into how the virus evolves to escape the immune response to establish productive infection. Ultimately, such research avenues may lead to novel strategies for vaccine and drug design.

Chapter 2

Use of Coalescent Simulations to Identify Homogeneous Acute and Early HIV-1 Infections

2.1 Introduction

Ample data from chronic infections have lead to the understanding that HIV is able to maintain extremely high levels of sequence diversity due to the error prone nature of the reverse transcriptase (Roberts et al., 1988; Taylor et al., 2008), and high replication (Julg and Goebel, 2005; Lynch et al., 2009) and recombination (Ramirez et al., 2008; Jetzt et al., 2000) rates. These characteristics have allowed drug resistance mutations to accumulate over the course of antiretroviral treatment, and have made the search for an effective treatment strategy very challenging (Clavel and Hance, 2004; Bennett et al., 2009; Okazaki et al., 2006).

Due to the logistical difficulties associated with identifying individuals immediately after HIV infection, the early stage of infection has been studied in detail in a relatively small number of patients (Lewis et al., 2008). Our understanding of viral transmission and the

establishment of new infections, as well as the mechanisms by which HIV evades the immune response during the earliest stages of infection, is also hampered by the difficulty of obtaining samples from acute and early HIV infection (Keele et al., 2008). Since the transmitted virus is the variant that an effective vaccine needs to target, obtaining samples from acute and early HIV-infected individuals is essential.

Remarkably, HIV infection can be caused by a small subset of the viruses present in the donor quasiespecies (Wolfs et al., 1992; McNearney et al., 1992; Zhang et al., 1993; Delwart et al., 2002). In 2004, a study by Derdeyn et al. (2004) on the neutralisation sensitivity of envelope genes after transmission between heterosexual transmission pairs, suggested that HIV transmission is associated with an extreme bottleneck. They further argued that HIV infection is likely to take place through the transmission of an individual sequence from the donor viral population, or from the expansion of a single virally infected cell. If productive infection is established through the transmission or outgrowth of a single virus, then characterising these sequences may lead to new ideas for preventing HIV transmission and infection.

A method to directly sequence uncloned DNA amplicons generated from single genome amplification (SGA) of plasma viral RNA has been described (Simmonds et al., 1990; Shriner et al., 2004; Palmer et al., 2005; Salazar-Gonzalez et al., 2008). Acute and early infection sequences obtained from such a method, provide an ideal source of data for identifying the transmitted, or founder, virus (Keele et al., 2008; Wood et al., 2009). Differentiating between infections caused by multiple viral strains and those caused by a single virus is interesting for various reasons. Apart from elucidating the extent of sequence diversity in transmitted HIV strains as well as describing the associated transmission bottleneck, analyses of single HIV transmission datasets allows us to describe the diversity of the virus caused solely by changes that occurred post-infection (Abrahams et al., 2009; Wood et al., 2009; Keele et al., 2008).

In addition to gaining a greater understanding of the diversification of HIV after infection has been established, it is also of interest to investigate the HIV population history of an

infected individual and to estimate the time frame in which the infection is likely to have taken place (Hué et al., 2005; Keele et al., 2008). Evolutionary analysis of a sample of HIV sequences obtained from an infected individual can provide information about the likely time of infection, and could therefore contribute valuable information on the evolutionary history of the virus from the point of transmission and subsequent manifestation of infection (Hué et al., 2005; Keele et al., 2008; Wood et al., 2009; Abrahams et al., 2009).

Establishing whether a sample from an infected individual consists of homogeneous sequences (derived from a single infecting viral strain) or heterogeneous sequences (derived from the transmission of two or more of the viral strains in the donor individual) is therefore a primary step for characterising the transmitted virus. Examining the sequence alignment, and determining the heterogeneity of the viral infection within a single patient, provides an initial insight into the number of viral strains that caused productive infection. If the diversity of the sample taken from an individual is higher than a given diversity threshold, it is likely that more than one virus was transmitted and multiple strains caused the infection (Keele et al., 2008). A model-based approach to further corroborate the primary findings of multiple virus infection, involves estimating the minimum time it would take for a viral population to obtain the observed level of diversity under the assumption of an initial homogeneous infection (Keele et al., 2008). The results of this model-based approach can then be compared to the clinical stage of infection, determined in the laboratory, for example using the clinical stages proposed by Fiebig et al. (2003).

2.1.1 The Coalescent

In population history, a coalescence describes the point where lineages join to form a common ancestor at some time in the past (Rosenberg and Nordborg, 2002; Fu and Li, 1999; Kuhner et al., 1995). As described in depth by Kingman in 1982, the coalescent therefore describes the ancestral relationship between a sample of sequences (Kingman, 1982a,b, 2000; Wakeley, 2008). An example of a coalescent genealogy is show in Figure 2.1. In this example, there are $n = 6$ individuals, and $n - 1$ coalescent events, indicated with blue circles (Figure 2.1).

Going backward in time from the present (T_0), each coalescent event reduces the number of lineages by one, until the most recent common ancestor (at the point T_{MRCA}) of the sample is reached. The coalescent is therefore a stochastic process by which genealogies are modelled (Nordborg, 2001; Rosenberg and Nordborg, 2002), which includes a rooted bifurcating tree, and $n-1$ coalescent events that occur at particular time intervals (Wakeley, 2008). Where classical population genetics describes divergence events where branching to newer forms take place, the coalescent process is concerned with merging events going back in time (Fu and Li, 1999; Stone et al., 2007).

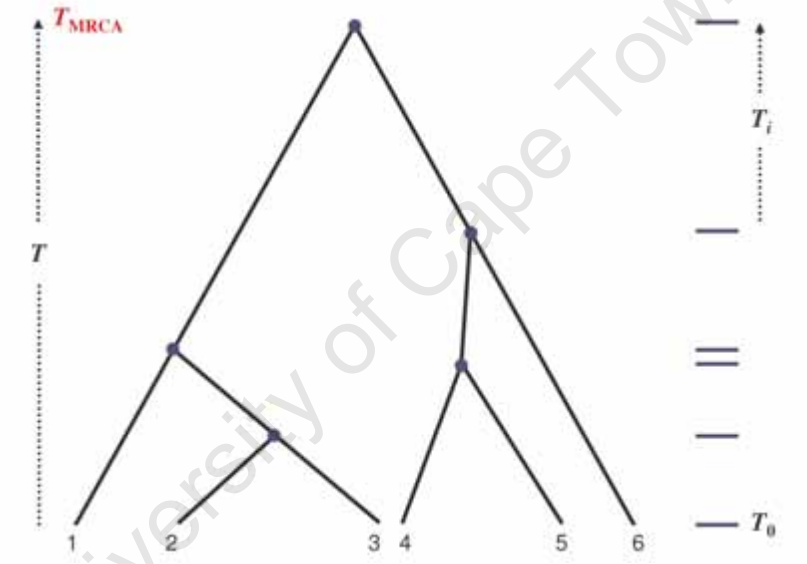


Figure 2.1: A coalescent genealogy of a sample with $n = 6$ individuals. The blue circles indicate the points where two lineages coalesce, and T_i illustrates the time interval from the point where only two lineages remain, to the final coalescent event when the most recent common ancestor is reached.

Coalescent theory provides a way to use the observed pattern of polymorphisms in samples of DNA sequences to make inferences about the genealogical history (Wakeley, 2008). Since the coalescent process is stochastic in nature, there exists a sample space that describes all the possible outcomes of a model (Wakeley, 2008). A population genetics model includes features such as mutation, recombination, and population size. These are examples of processes that need to be modelled based on the observations made from the polymorphisms

present in the sampled sequences (Hudson, 1991; Wakeley, 2008). Two of the popular and well-researched population genetics models, are the Wright-Fisher model (Kingman, 1982b; Fu and Li, 1999), and the Moran model (Moran, 1958; Itoh and Mahmoud, 2005; Wakeley and Sargsyan, 2009), which make different assumptions regarding the population parameters. Various extensions of these population models have been described, allowing for, for example, recombination (Hudson and Kaplan, 1988; Griffiths and Marjoram, 1996) and selection (Krone and Neuhauser, 1997; Neuhauser and Krone, 1997; Kaplan et al., 1988). Coalescent-based methods can therefore be very useful for determining, for example, the time a pathogen first entered the human population, or for estimating the rate of evolution of a pathogen in a new host (Rosenberg and Nordborg, 2002). In this study, a coalescent-based method was used to determine the most recent common ancestor for each of the viral samples isolated from multiple early infection patients.

2.1.2 Bayesian Evolutionary Analysis Sampling Trees (BEAST)

BEAST, Bayesian Evolutionary Analysis Sampling Trees, is a software package that was developed to offer a flexible platform for coalescent-based Bayesian statistical analysis of molecular sequence data using phylogenies (Drummond and Rambaut, 2007). Since a large number of parameters are modelled during the analysis, high dimensional probability distributions are generated. Markov chain Monte Carlo (MCMC) methods can be applied to integrate over the large dimensional space and to solve complex optimisation problems (Andrieu et al., 2003). MCMC is used for generating samples from a multidimensional distribution, as well as exploring and evaluating expectations based on the posterior probability distributions (Andrieu et al., 2003; Gilks et al., 1998; Kedem and Fokianos, 2002). The distribution resulting from the MCMC eventually converges to the posterior probability distribution of the parameters, and can therefore be treated as a representative sample from the distribution from which predictions and inferences can then be made (Qian et al., 2003; Mossel and Vigoda, 2005).

BEAST uses a Metropolis-Hastings MCMC algorithm to generate estimates of a large num-

ber of parameters. The algorithm works by sampling from a posterior distribution that was generated from a particular parameterisation of an evolutionary model relating an alignment of molecular sequences (Drummond and Rambaut, 2007). In this way, for each new sample, the Metropolis-Hastings algorithm generates a proposal distribution from the current states, which is accepted or rejected depending on values contained in the posterior probability density functions of the current as well as proposed states (Christensen et al., 2004; Larget and Simon, 1999; Chib and Greenberg, 1995). In practice, the MCMC is run for several thousand steps until convergence is reached, and sampling occurs at specified intervals. The ideal result is a set of of posterior probability distributions for the parameters of interest (Miller et al., 2004; Drummond and Rambaut, 2007). An evolutionary model in BEAST is comprised of five features, including the substitution model, among-site rate heterogeneity model, the rate model among branches, the tree, and tree prior. The user can specifically parameterise these five components of the evolutionary model for a given dataset of aligned DNA or protein sequences (Drummond and Rambaut, 2007).

An important consideration for Bayesian statistical analysis is how to set the prior probability distributions for each parameter. Specifying priors is an inherent part of Bayesian statistical analysis, and although including definitive information could lead to more accurate estimates of the posterior probability, deciding on relevant prior values can be very difficult. The posterior probability distribution for each parameter is dependent on the original incorporated prior distribution, and inaccurate priors could lead to distorted evolutionary parameter estimates (Drummond and Rambaut, 2007; Yang and Rannala, 2005). Tree priors are used to model population size changes over time, where calibration priors can also be added if dates or other units of time are known for specific nodes. Specifying a realistic model of demographic history, for example an exponential or logistic growth model, is particularly important for studies interested in estimating rates of evolution, since the tree prior can significantly bias the rate parameter of interest (Drummond et al., 2002, 2005). Despite the potential bias that may arise due to unsubstantiated priors, BEAST, as a flexible programme incorporating various model possibilities and extensibility, is an extremely useful tool for Bayesian evolutionary analysis.

2.1.3 Relaxed Molecular Clocks

The molecular clock hypothesis asserts that the rate of molecular evolution, for all genes (or proteins), is approximately constant over time (Zuckerkandl and Pauling, 1965; Nei and Kumar, 2000; Pevsner, 2003). The significance of the hypothesis is evident; if a gene evolves at a constant rate then the rate of nucleotide substitution and divergence times can be estimated (Pevsner, 2003). However, not all proteins evolve according to the molecular clock hypothesis, and various factors contribute to evolutionary rate heterogeneity, for example generation time and metabolic rate can vary across lineages. A strict uniform molecular clock is therefore not an accurate assumption in many cases (Pybus, 2006; Martin and Palumbi, 1993). The relaxed molecular clock model provides an alternative as it allows the evolutionary rate to vary among lineages (Pybus, 2006; Kumar, 2005; Sanderson, 1997).

Various approaches to relax and overcome the limitations resulting from the molecular clock hypothesis have been proposed (Aris-Brosou and Yang, 2002). One method suggests essentially ignoring the clock, and therefore allowing complete rate heterogeneity across all branches of the phylogeny. However, in most cases it is not possible to estimate dates of divergence or absolute evolutionary rates with such an approach, since time and divergence dates are linked (Ho et al., 2005). An alternative is to assign a unique local rate to different parts of the phylogenetic tree, the so called local molecular clock model (Aris-Brosou and Yang, 2002). This approach is however not always tractable when large phylogenies are analysed, since there can be multiple possibilities for assigning rates in different regions of the tree (Ho et al., 2005).

The relaxed molecular clock that allows rates to vary across branches but accounts for close and distant evolutionary relationships, provides an alternative to the above-mentioned methods. By allowing sequences that are closely related to share similar evolutionary rates, the rates are effectively inherited along the tree (Ho et al., 2005; Aris-Brosou and Yang, 2002). The Bayesian approach of this methodology allows priors to be selected for how the descendant rate is related to the ancestral rate (Thorne et al., 1998; Ho et al., 2005). The rates amongst all branches are therefore autocorrelated, and the rate for each branch

depends on the selected prior distribution, where the mean rate on the descendant branch is a function of the estimated rate on the ancestral branch (Thorne et al., 1998; Aris-Brosou and Yang, 2002; Ho et al., 2005; Drummond et al., 2006). A drawback of these relaxed molecular clock models is that the tree topology has to be specified; this is problematic when there are ambiguities in particular regions of the tree that, therefore, result in numerous possible tree topologies (Drummond et al., 2006).

Drummond et al. (2006) proposed a new method for incorporating the relaxed molecular clock, where both the divergence times and phylogeny are estimated. This approach therefore overcomes the drawback of the previous methods where a specific tree topology has to be provided. In addition, the authors describe a relaxed clock model which, as opposed to the *autocorrelated* model, does not, *a priori*, specify rate correlation among related branches on a tree. This *uncorrelated* relaxed clock model, avoids the requirement of having to specify a prior on the degree of autocorrelation (Drummond et al., 2006). The fundamental contribution of relaxed molecular clock models is that they allow for the estimation of divergence times and the rates of substitution from sequence data where different lineages do not share the same evolutionary rate (Sanderson, 2003). The relaxed molecular clock is therefore particularly useful for analysing HIV sequences, where the rate of evolution can vary substantially over the course of HIV infection.

The Bayesian MCMC method is implemented in BEAST, and presents an application to perform relaxed phylogenetics (Drummond et al., 2006). Here, BEAST was used to estimate the time to the most recent common ancestor of a set of HIV-1 sequences in early infection, and therefore, in the case of homogeneous infections, a lower bound on the time to infection. This approach was applied to a large sample of HIV-1 subtype B and C sequences isolated from early infections (Keele et al., 2008; Abrahams et al., 2009). The results obtained from this coalescent-based estimation were subsequently compared to both the clinically determined stage assigned to each sample, as well as to the estimates obtained from a mathematical model of sequence diversification. The analyses formed part of two international collaborative studies focusing on identifying the transmitted virus in HIV-1 subtype B (Keele et al., 2008) and subtype C (Abrahams et al., 2009) sequence datasets.

2.2 Methods

Coalescent analysis was carried out on 102 HIV-1 subtype B sequence alignments, and 69 HIV-1 subtype C alignments obtained from acute and early infected patients. These analyses formed part of two international collaborative projects (Keele et al., 2008; Abrahams et al., 2009); sequencing, as well as Fiebig clinical staging for each infected individual, was carried out by collaborators (Keele et al., 2008; Abrahams et al., 2009).

2.2.1 Clinical Staging

The stage of disease for each of the infections was clinically categorised based on an adaptation of the original Fiebig stages (Fiebig et al., 2003; Lee et al., 2009). One of the major changes compared to the original staging was the inclusion of an eclipse period. Keele et al. (2008) allowed a period of time, 10 days, during which plasma viral RNA is undetectable. This eclipse phase was estimated to last between 7 and 21 days (Table 2.1). A further 2 days were also added to Fiebig stage I to account for the greater sensitivity of the current assays as compared to those available in 2003, when Fiebig et al. published the original classification (Lee et al., 2009). Finally, the confidence intervals were recalculated using the quadrature approach, in which the errors for each stage duration are assumed to be uncorrelated (Lee et al., 2009; Drosg, 2007). Combining individual errors in quadrature means that the cumulative error n at Fiebig stage i is calculated as:

$$n_i = \sqrt{m_i^2 + n_{i-1}^2}$$

where m_i is the difference between the mean and the minimum (or -maximum) 95% CI estimate for Fiebig stage i, and n_{i-1} is the result of the previous calculation. If $i = 1$, then $n_{i-1} = 3$, which is the difference between the mean and minimum (or -maximum) 95% CI for the eclipse phase. The cumulative errors for both the minimum and maximum ranges were calculated with this method (Table 2.1). Combining individual errors in quadrature

results in cumulative error estimates that are smaller than the sum of the individual errors, but larger than the errors in either measurement (Drosg, 2007). The cumulative measures indicate the expected time interval spanned by each stage. For example, Fiebig stage I ranges from day 10, the end of the Eclipse phase, to day 17, and Fiebig stage II occurs between 17 and 22 days after infection. Throughout the chapter, where reference is made to Fiebig stage classification, the cumulative durations are the modified estimates, as shown in Table 2.1.

Table 2.1: Adapted Fiebig stage ranges for classification of HIV-1 primary infection. The individual durations of each phase as well as the cumulative durations are shown.

	Duration In Days (95% CI)					
Stage	**Individual**	**Min**	**Max**	**Cumulative**	**Min**	**Max**
Eclipse	10	7	21	**10**	**7**	**21**
I (vRNA+)	7	5	10	**17**	**13**	**28**
II (p24Ag+)	5	4	8	**22**	**18**	**34**
III (ELISA+)	3	2	5	**25**	**22**	**37**
IV (Western Blot ±)	6	4	8	**31**	**27**	**43**
V (Western Blot +, p31-)	70	40	122	**101**	**71**	**154**
VI (Western Blot +, p31+)	Open-ended					

2.2.2 Poisson Model of Replication and Diversification

A mathematical model, described in detail by Lee et al. (2009), was used to estimate the extent of diversification of each dataset (Keele et al., 2008; Abrahams et al., 2009). The model design and development, as well as the analysis performed using the model, was carried out by collaborators and is described here in brief for comparative purposes.

The model uses previously published estimates of parameters related to the HIV-1 life cycle; a generation time of two days (Markowitz et al., 2003), reproductive ratio, R_0, of 6 (Stafford et al., 2000), and a nucleotide substitution rate equal to 2.16×10^{-5} substitutions per site per generation (Mansky and Temin, 1995), were applied in an exponentially replicating model of sequence evolution that assumes no recombination or selection (Keele et al., 2008). This results in a star-like phylogeny relating the sequences with a frequency distribution of the

number of differences between any pair of sequences (Hamming Distance), described by a Poisson distribution. The mean of the distribution is linearly dependent on the number of generations between the sample and the founder virus (the root of the star phylogeny). This model therefore provides estimates of the number of generations (or time) required to produce the observed sequence diversity. See Lee et al. (2009) for a detailed justification of this model.

The coalescent analysis carried out using BEAST (described in the subsequent sections 2.2.3.1 and 2.2.3.2) provides an alternative estimate of the number of generations since the last common ancestor of the population, but this estimate is in the reverse direction - that is, the results reflect the number of generations required to reach a hypothetical most recent common ancestor from the present viral population, as opposed to the number of generations required to generate the present observed diversity starting from a single viral particle. The mathematical model described first by Keele et al. (2008) and later extensively discussed by Lee et al. (2009), therefore represents a forward evolution simulation, whereas BEAST simulates in the reverse direction (Drummond and Rambaut, 2007).

2.2.3 Applying a Coalescent Model of Evolution to HIV-1 Subtype B and C Sequences

2.2.3.1 Estimating the MRCA for 102 HIV-1 Subtype B Datasets

A Bayesian MCMC approach, implemented in BEAST v1.4.1 (Drummond et al., 2006; Drummond and Rambaut, 2007), was used to estimate the time, in days, to the MRCA for each of the 102 HIV-1 subtype B datasets. The analyses were carried out by using the general time reversible (GTR) substitution model with invariant sites and gamma-distributed rate heterogeneity (four gamma categories). Furthermore, the mean substitution rate was fixed (2.16×10^{-5} substitutions per site per generation), and the substitution and rate heterogeneity models were unlinked across codon positions. Exponential population growth and a relaxed (uncorrelated exponential) molecular clock was assumed. These model pa-

rameters were used for all the analyses, and the MCMC algorithm was run for at least 10^7 generations (logging every 1000 generations; burn-in was set to 10% of the original chain length). Additional runs were carried out if the effective sample size for the estimate was less than 100, as recommended by the documentation on BEAST and commonly used in the literature (for example Hughes et al., 2004, Edwards et al., 2006, and Bello et al., 2009). The results were visualised in TRACER (Rambaut and Drummond, 2007).

The analyses were repeated with fixed values for the GTR model instead of 5 free parameters. The fixed values were estimated with the HyPhy package (Kosakovsky-Pond and Muse, 2005) by using the combined data from all homogeneous early infection patients. Furthermore, two additional demographic and evolutionary models (relaxed uncorrelated molecular clock with logistic population growth, and strict molecular clock with exponential population growth) were used. Similar results (estimates and confidence intervals) for the MRCA times were obtained for the alternative relaxed clock models. However, the estimates were 25% lower when the strict molecular clock was used (data not shown).

2.2.3.2 Estimating the MRCA for 69 HIV-1 Subtype C Datasets

BEAST v1.4.7. (Drummond and Rambaut, 2007) was used to estimate the time to most recent common ancestor (tMRCA) of the 69 HIV-1 subtype C sequence datasets (Abrahams et al., 2009). Similar to the subtype B analysis, a GTR substitution model with gamma distributed rates (four categories) and a proportion of invariant sites was used to model rate heterogeneity among sites. The relative substitution rate priors of the GTR model were set to the empirical estimates obtained from analysis carried out on the entire acute homogeneous dataset using the HYPHY package (Kosakovsky-Pond and Muse, 2005). The models were unlinked across codon positions and the mean substitution rate was fixed to 2.16 x 10^{-5} substitutions per site per generation, following the analyses carried out on the HIV-1 subtype B sequences (Mansky and Temin, 1995; Keele et al., 2008). A relaxed (uncorrelated exponential) molecular clock, and both a constant piecewise model of population growth, where the demographic parameter estimates are obtained from a Bayesian skyline

plot (BSP), and an exponential population growth model were used to describe the population size changes over time (Drummond et al., 2005). The Bayes factor (Suchard et al., 2001) estimates did not provide evidence in favour of either of the demographic models, which suggests that the parametric (exponential) and non-parametric (constant piecewise derived from a BSP) models provide an equally good fit to the data; the tMRCA estimates and confidence intervals were also similar for each model (data not shown). Results from the constant piecewise BSP model are given because the the MCMC chain appeared to converge faster for all the datasets under this model of population growth. The MCMC was initially run for 10^7 steps (logging every 1000 steps, with the first 10% of the run discarded as burn-in) and subsequently repeated if the effective sample size for the estimate was below 100. As with the subtype B data, the results were visualised in TRACER (Rambaut and Drummond, 2007).

In both the HIV-1 subtype B and subtype C analyses the tMRCA estimate generated by BEAST represents the time in generations, since the substitution rate provided as input, 2.16 x 10^{-5}, is in units of substitutions per site per generation. The same generation time estimate as was used in the Poisson model was also applied here. The generation time was therefore multiplied by 2 (Markowitz et al., 2003) in order to obtain the final time, in days, to the MRCA.

2.3 Results and Discussion

2.3.1 BEAST Estimates of the tMRCA

2.3.1.1 First Round of Classification of Homogeneous Infection

The time to most recent common ancestor (tMRCA) for 102 HIV-1 subtype B and 69 HIV-1 subtype C datasets was determined using BEAST. The MCMC chain was run until the effective sample size (ESS) for the tMRCA estimate was above 100. The BEAST estimates

and highest posterior densities present coalescent-based estimates of the evolutionary time needed to reach the MRCA of each individual patient sample. In this study, a patient is believed to harbour a homogeneous infection if the BEAST estimate of the tMRCA, as well as the results from the Poisson evolution model, are consistent with the estimates to the time of infection based on the clinical stage. If the BEAST tMRCA estimate for a patient does not correspond to the clinical diagnosis, further investigation is warranted. The concept is illustrated in Figure 2.2.

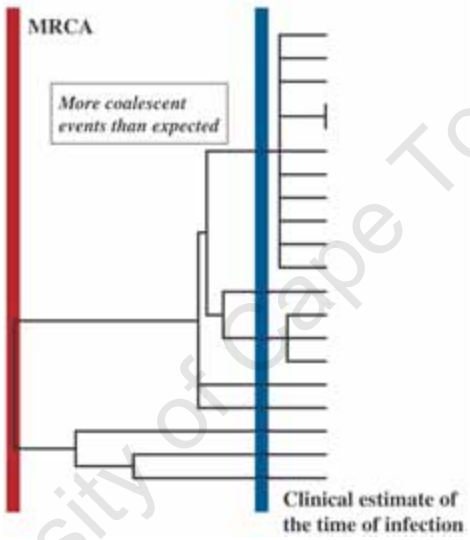


Figure 2.2: Phylogenetic tree illustrating a situation where the sequence diversity present in the sample, suggests that the MRCA existed at a time before the infection took place. The time of infection, estimated from the Fiebig clinical stage, is shown in blue.

In this example, at the time when infection is believed to have taken place based on the Fiebig clinical stage (shown in blue), the diversity of the sample is greater than would be expected if the infection resulted from transmission of a single virus. The sequence diversity therefore suggests that the MRCA existed further back in time than the transmission event, implying that some of the diversity in donor individual was transmitted to the recipient (Figure 2.2).

The coalescent analysis was carried out on all the HIV-1 subtype B (n = 102) and C (n = 69) acute and early sequences, and the BEAST values, in units of days, were compared to

the clinically determined Fiebig stage estimates assigned to each patient (Keele et al., 2008; Abrahams et al., 2009). The comparisons resulted in three different groups, and patients were categorised depending on the degree of overlap between the BEAST and Fiebig stage estimates. The three scenarios are illustrated by specific examples below. Examples are from HIV-1 subtype B infections, however, the same analysis steps were performed on both subtype B and C patient datasets. One situation resulting from the comparison between the clinical and coalescent-based estimates, is when the BEAST and Fiebig stage estimates overlap. This situation is depicted in Figure 2.3.

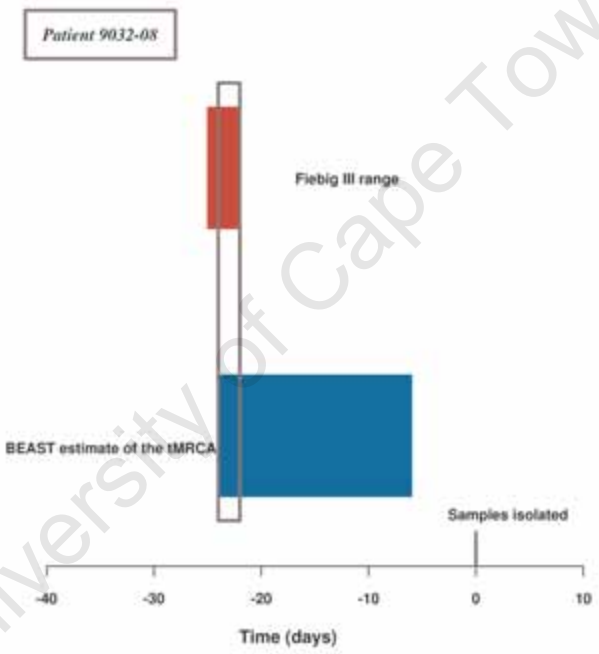


Figure 2.3: Timing estimates for Patient 9032-08. The BEAST tMRCA estimate and Fiebig stage range overlap, suggestive of a homogeneous infection.

Patient 9032-08 was clinically categorised as Fiebig III, which suggests that infection took place between 22 and 25 days prior to isolation of the samples (Fiebig et al., 2003). The BEAST estimate indicated that the time needed for the required number of coalescent events to take place in order to reach the MRCA, was between 6 and 24 days. The results are consistent with infection by a single HIV-1 strain. The Poisson evolution model provided similar estimates, between 6 and 13 days, reinforcing the belief that the viral population

isolated from this patient represents an homogeneous infection.

A further scenario that suggests the presence of a homogeneous infection, is show in Figure 2.4. The BEAST estimated range for patient SC11, between 2 and 8 days, suggests a younger MRCA than the clinical Fiebig II range (between 17 and 22 days). It is likely that patients displaying post-infection diversity, as is represented in Figure 2.4, are also harbouring homogeneous infections. In these cases the overall *env* diversity is extremely low, and with a fixed substitution rate of 2.16 x 10^{-5} together with the other assumptions of the evolutionary model, the upper bound on the time it takes for the diversity to reduce to a single sequence through coalescent events, is lower than the minimum clinical estimated duration of infection. The evolutionary constraints acting on these *env* sequences, may therefore have limited the diversification of the virus.

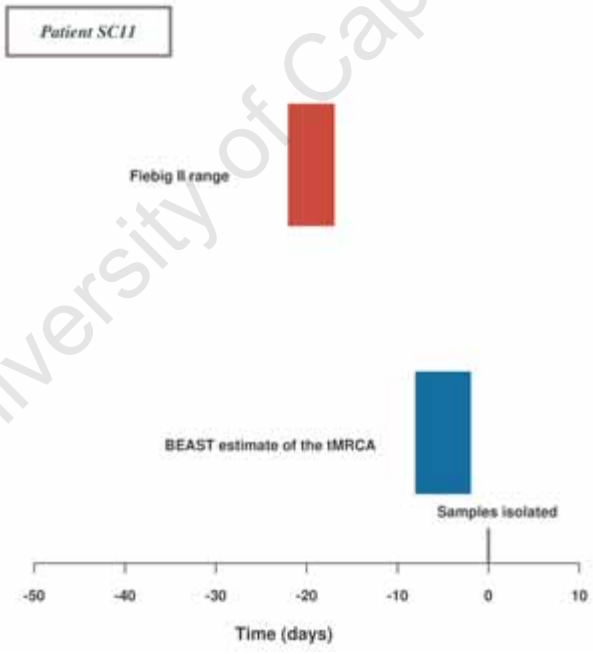


Figure 2.4: Timing estimates for Patient SC11. The BEAST estimate of the number of days to the MRCA is lower than the Fiebig stage range. This suggests that the sequences are less diverse than would be expected under the provided model of sequence evolution.

The final situation occurs when the BEAST tMRCA estimate suggests that the infection took place at an earlier time point than the corresponding indication from the clinical esti-

mate. If the sequences are more diverse than expected under the presented model of sequence evolution, more coalescent events would need to take place in order to reach the MRCA of the sample. The BEAST estimated range for patient 1059-09 (Figure 2.5), between 43 and 90 days, places the infection stage as Fiebig IV or V, however the clinical tests resulted in a Fiebig III classification (between 22 and 25 days).

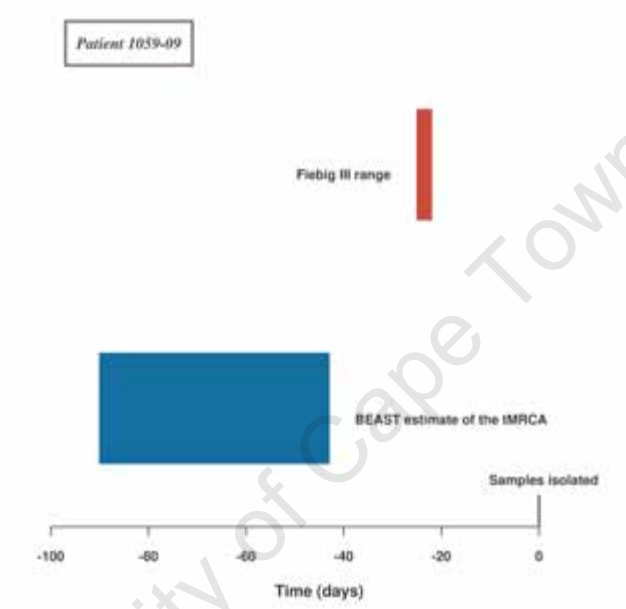
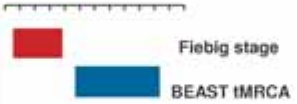
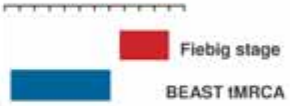


Figure 2.5: Timing estimates for Patient 1059-09. The BEAST tMRCA range is larger than that of the Fiebig stage classification, suggesting the presence of more sequence diversity than expected under the current model of evolution.

Patients in this class, where the comparison between the clinical and the BEAST estimates are not consistent, fit the heterogeneous infection category. The post-infection diversity in these cases may have been caused by infection of more than one viral strain (heterogeneous infection), or an external factor causing an increased rate of mutation. APOBEC-mediated hypermutation is an example of the last mentioned situation, where an inflated mutation rate could bias the BEAST results and lead to the classification of a later stage of infection, despite transmission of a single viral strain. The complete results for the comparisons between Fiebig stage ranges and BEAST estimated ranges, are shown in Table 2.2.

Table 2.2: Comparison between the BEAST tMRCA estimate and Fiebig stage range, for 102 HIV-1 subtype B, and 69 HIV-1 subtype C infection datasets.

	BEAST tMRCA & Fiebig stage Overlap 	Post-infection Diversity 	Pre-infection Diversity 	Total # Sequences
Subtype B	70	2	30	102
Subtype C	49	0	20	69

In total, 72 of 102 (~70%) subtype B, and 49 of 69 (~71%) subtype C, infections were labelled as homogeneous. This classification was based on the BEAST tMRCA and Fiebig stage range comparison, where the estimates either overlapped ($^{70}/_{72}$ subtype B and $^{49}/_{49}$ subtype C datasets) or the BEAST estimated range suggested a younger MRCA than the equivalent Fiebig range ($^{2}/_{72}$ subtype B and $^{0}/_{49}$ subtype C datasets, Table 2.2).

No infections were classified as homogeneous for the subtype C investigations that were not also within the homogeneous infection group of the published study (Abrahams et al., 2009). Further investigation, however, revealed that 1 of the subtype B infections that was labelled homogeneous, was infected with multiple viruses. The polymorphic pattern apparent from the *Highlighter* plot, as well as the structure of the phylogenetic tree relating the sequences, showed a clear distinction between subsets of sequences in the dataset (Figure 2.6).

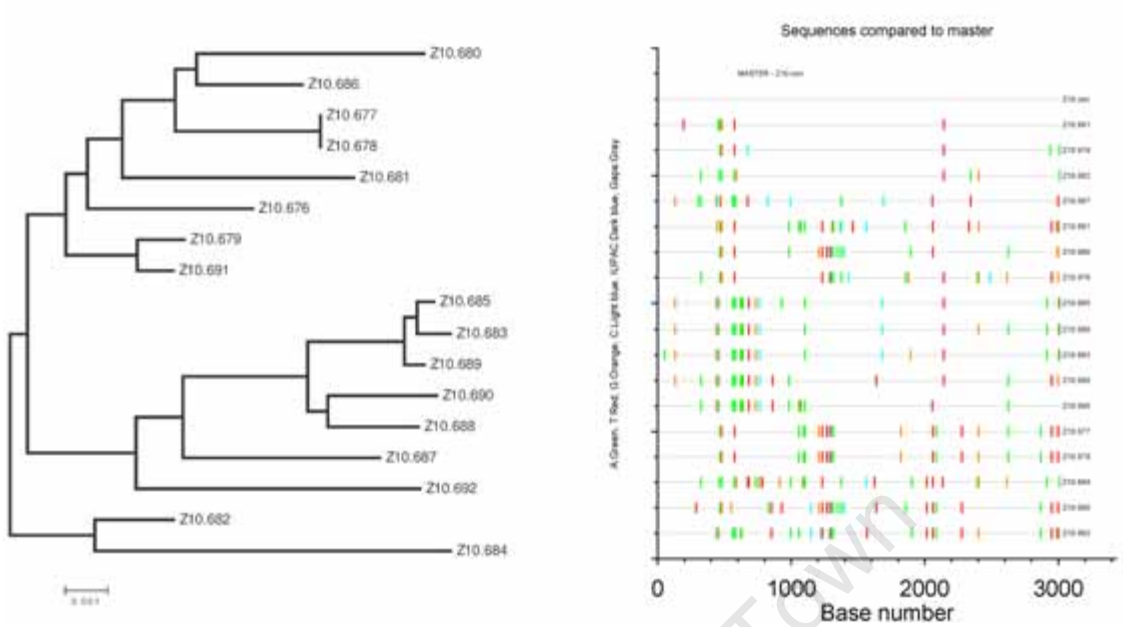


Figure 2.6: Neighbour-joining tree and *Highlighter* plot of *env* sequences illustrating a multiple virus infection for patient Z10.

Patient Z10 (Figure 2.6) was classified as Fiebig VI, which occurs from 101 days post-infection through to chronic infection. Since a Fiebig VI classification provides an open-ended maximum duration, BEAST tMRCA estimates will always be within, or smaller than, the corresponding clinical estimate. Comparing Fiebig VI estimates to model-based estimates as a means to classify homogeneous infections is therefore not suitable. In general it is possible that viral samples do not represent the entire viral diversity of the population within the infected individual, and that the coalescent time is therefore shorter than the corresponding clinical time estimate. Another scenario is where the diversity within *env* reached some threshold beyond which the MRCA is difficult to estimate accurately. Selection is also likely to occur in viral populations from Fiebig stages V and VI (Lee et al., 2009), which BEAST does not explicitly model, and in such cases the higher diversity due to positive selection may therefore lead to a tMRCA further back in time than expected under neutral evolution.

2.3.1.2 Addressing APOBEG3G Hypermutation

APOBEC3G hypermutation is one of the external factors that can affect the BEAST estimate of the tMRCA. Since APOBEC3G causes an excess of A-to-G mutations (see section 1.2.5.2), and is not included in the model, the amount of diversity in sequences affected by APOBEC hypermutation will be greater than expected for a given time to infection. The viral population isolated from patient 1059-09 provides an example of a hypermutated dataset (Figure 2.7). For this patient, the APOBEC3G signature was found to be scattered across the individual sequences, as indicated by the blue stars (Figure 2.7). The BEAST estimate originally suggested that the MRCA existed between 43 and 90 days in history, despite the Fiebig III clinical classification suggesting that infection took place at an earlier time point (22 - 25 days back in time, Figure 2.8). After removal of the hypermutated sequences, the BEAST tMRCA estimate was decreased to between 17 and 50 days (Figure 2.8), which corresponds to the Fiebig III range. Patient 1059-09 was therefore classified as harbouring a homogeneous infection after the APOBEC3G-induced hypermutated sequences were removed.

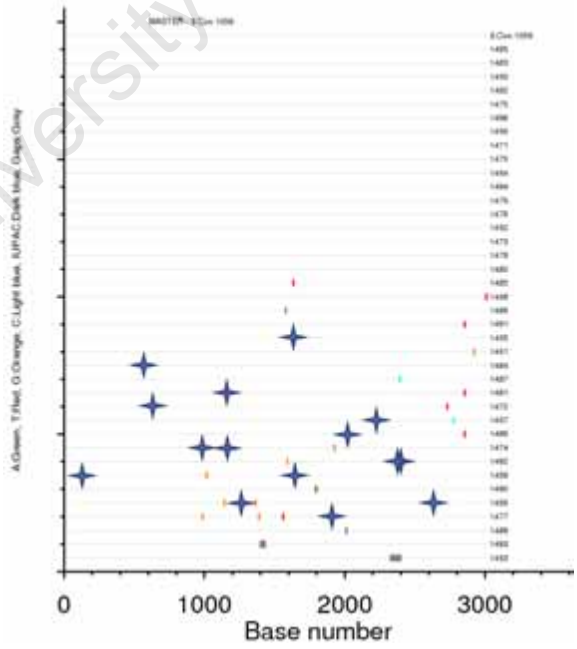


Figure 2.7: Sequence alignment for Patient 1059-09. The blue stars indicate sites that are in the APOBEC3G context, and the tick marks represent other substitutions.

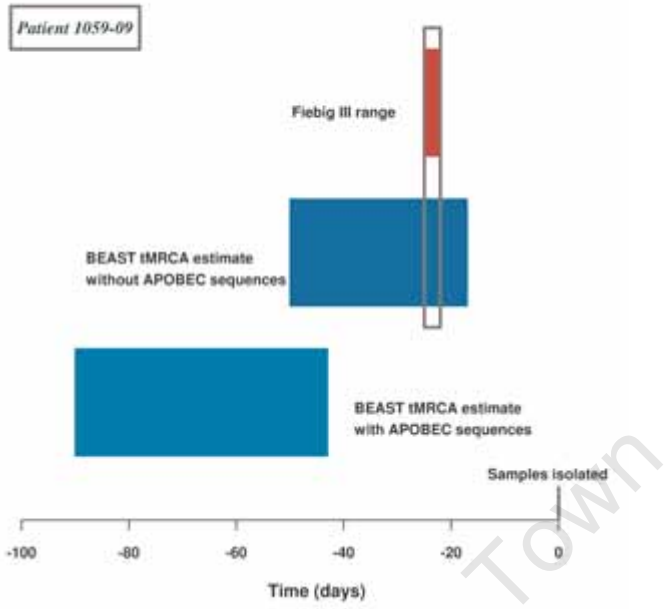


Figure 2.8: Timing estimates for Patient 1059-09 with APOBEC hypermutated sequences included as well as removed. The BEAST tMRCA range for the dataset with hypermutated sequences is larger than that of the Fiebig stage classification, suggesting the presence of more sequence diversity than expected under the current model of evolution.

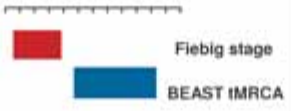
The subtype B investigation carried out by collaborators (Keele et al., 2008), revealed that 13 datasets contained significant hypermutation, while 2 others also harboured hypermutated sequences, but the prevalence of hypermutation was not high enough to be labelled as significant. Among the 13 patients with sequences that were significantly hypermutated, there were six where the evidence of G-to-A hypermutation was isolated to a single sequence. In the remaining seven datasets, APOBEC-mediated hypermutation was spread across multiple sequences (Keele et al., 2008). In the current study, five of the hypermutated datasets were initially not classified as homogeneous infections due to APOBEC-induced hypermutation resulting in an increase in the diversity of the sequences. The APOBEC-mediated mutations were excluded for all 15 hypermutated datasets, and reanalysis was carried out. Where the BEAST estimate originally suggested a MRCA further back in time than the clinical estimate, removal of mutations likely to be the consequence of APOBEC hypermutation resulted in estimates that overlapped in all cases. In the current study, the five

infections that were originally determined to be heterogeneous due to the excess mutations, were therefore classified as homogeneous after the reanalysis.

Similar results were obtained for the HIV-1 subtype C analysis. Eight datasets contained APOBEC3G hypermutated sequences, and 5 of the 8 infections were originally not classified as homogeneous. After removal of the hypermutated sequences, the BEAST tMRCA and Fiebig stage range for all five of these overlapped. The hypermutated sequences were discarded, and all 8 infections were labelled as homogeneous.

Once the impact of APOBEC3G hypermutation was addressed, the classification of homogeneous infections could be updated (Table 2.3). Twenty-five HIV-1 subtype B and 15 subtype C datasets were not classified as homogeneous infections (Table 2.3). The level of diversity for these patients was higher than expected under the given model of evolution, which suggested that alternative factors, beyond APOBEC-mediated hypermutation, are responsible. It is likely that a large number of infections in this group were caused by multiple infections, and therefore can be classified as heterogeneous infections. However, further investigation is warranted for these patients because a higher observed diversity does not necessarily imply that an infection was caused by multiple viruses, as is clear from the APOBEC hypermutated examples.

Table 2.3: Comparison between the BEAST tMRCA estimate and Fiebig stage range, after removal of the APOBEC hypermutated sequences. One subtype B overlapping time range was assumed to be representative of homogeneous infections, but were later shown to be infected by multiple viruses (indicated in parenthesis).

BEAST tMRCA & Fiebig stage Overlap 	Post-infection Diversity 	Pre-infection Diversity 	Total # Sequences
71 (and 1 heterogeneous)	5	25	102
54	0	15	69

2.3.1.3 Categorising the Difficult Cases

In most cases described, the overall diversity observed in a dataset corresponds to the number of viruses transmitted during infection. Coalescent-based estimates of the tMRCA will therefore resemble the true time since infection. However, in certain situations transmission of two closely related viruses that have low overall diversity, or transmission of a single virus strain where the sequence diversity is higher than expected due to external influences, for instance APOBEC, the observed diversity will not correspond the number of infecting viruses. This type of infection is harder to classify and requires further investigation.

An example of a patient sample that was difficult to characterise, is shown in Figure 2.9. The sequence diversity of the viral *envs* isolated from patient 6247-08, who was classified as Fiebig stage II, was consistent for an infection from this clinical stage. The BEAST estimated range (2 - 25 days) indicated a MRCA that overlapped the Fiebig II range (17 - 22 days). The sequences diversity was visualised by drawing a Neighbour-joining tree and creating a *Highlighter* (http://www.hiv.lanl.gov/) plot (analysis carried out by collaborators, Keele et al., 2008).

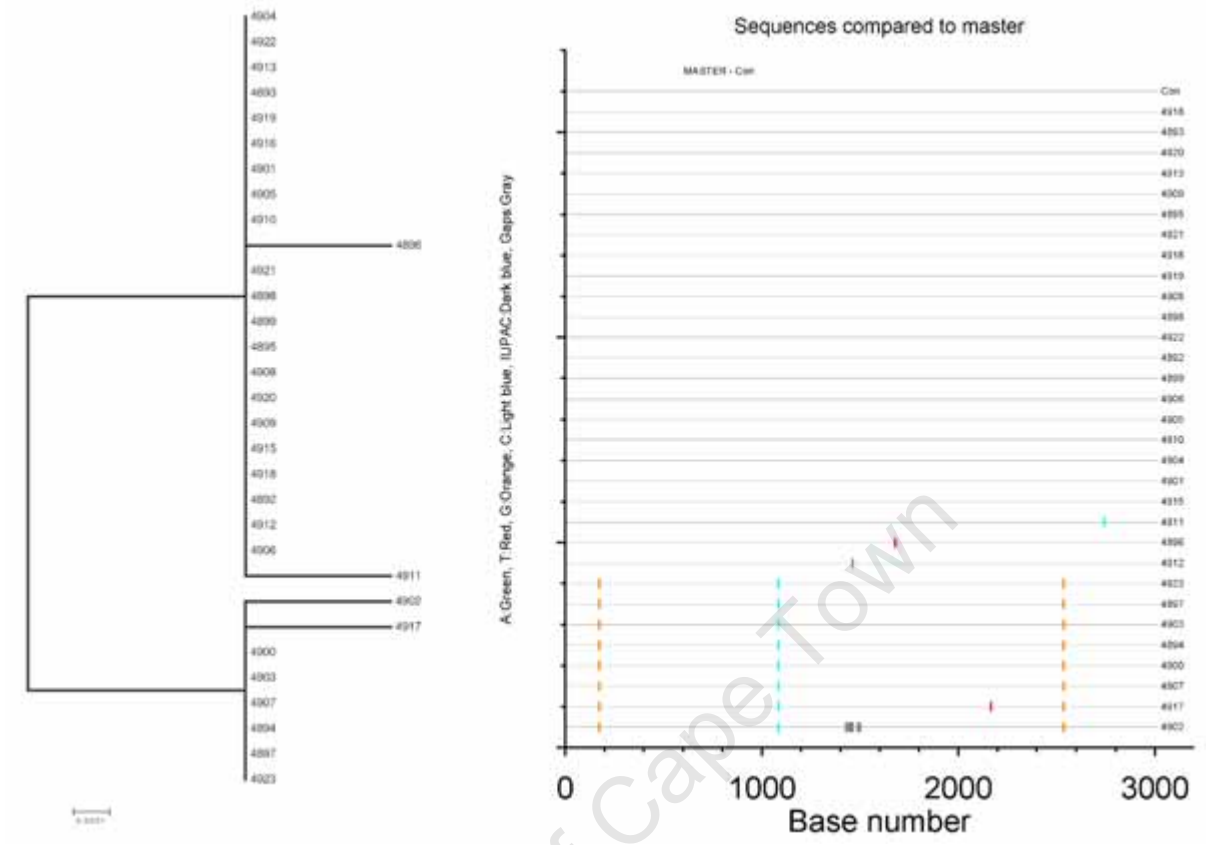


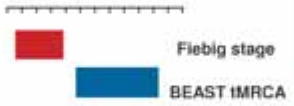
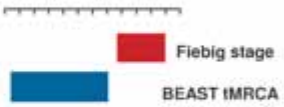
Figure 2.9: Neighbour-joining phylogeny and *Highlighter* plot for Patient 6247-08. Evidence of infection by two closely related viral strains can be observed in both figures. There are 3 polymorphisms that clearly distinguish the two strains, and two distinct populations are evident from the neighbour-joining tree.

The *Highlighter* plot revealed 3 apparent differences between two subsets of sequences, as is clear from Figure 2.9. Furthermore, the phylogenetic tree relating the sequences indicated that two distinct lineages were discernible (Figure 2.9). One group of sequences resembled the consensus sequence, where the second contained 3 mutations that differentiated them from the rest. Patient 6247-08 was therefore classified as infected with two very closely related viruses (Keele et al., 2008). One patient in the subtype C study displayed similar characteristics to patient 6247-08, and although both were initially categorised as homogeneous infections after comparing the BEAST tMRCA and clinical estimate, a more accurate explanation resulted from further investigation. These two patients were therefore classified as infected by two closely related virus strains.

The BEAST estimates for the infections that were not classified as homogeneous at this stage, suggested that pre-infection diversity was present in these datasets (Table 2.3). The 25 subtype B infections that were not labelled as homogeneous, were investigated further. Three of these were identified as homogeneous infections by other methods, for example low overall nucleotide diversity, and near overlap between the clinical estimated date of infections and the BEAST tMRCA estimate. For the remainder of the patients ($^{22}/_{25}$) no alternative explanation for the high overall levels of diversity contained within these datasets could be described, and these patients were classified as harbouring heterogeneous infections. Further investigations were also carried out on the uncharacterised subtype C datasets, and one patient was classified as homogeneous due to the near overlap between the clinical and BEAST estimate. The remainder of the infections that displayed pre-infection diversity ($^{14}/_{15}$) were classified as heterogeneous due to the high levels of diversity and the presence of more than one transmitted variants (Abrahams et al., 2009).

The number of individuals identified as harbouring homogeneous infections by comparing the BEAST estimated tMRCA and the Fiebig stage classification (after accounting for APOBEC3G-induced hypermutation as well as categorising the difficult cases), was $^{77}/_{102}$ for the subtype B analysis, and $^{54}/_{69}$ for subtype C (Table 2.4). For 2 subtype B infections, and 1 for subtype C, overlapping time ranges were assumed to be representative of homogeneous infections, but were later shown to be infected by multiple viruses. Furthermore, the near overlap between the BEAST and clinical estimates for 3 subtype B infections, and 1 subtype C infection, suggested that these individuals were infected with a single virus strains (Table 2.4, indicated in parenthesis). The overall results were similar to the numbers determined for the publications, $^{78}/_{102}$ for subtype B (Keele et al., 2008), and $^{54}/_{69}$ for subtype C (Abrahams et al., 2009). Furthermore, the comparative approach described here, resulted in a total of only 3 misclassifications of homogeneous cases where the individuals were later shown to be infected by multiple strains (shown in Table 2.4).

Table 2.4: Final comparison between the BEAST tMRCA estimate and Fiebig stage range. The misclassifications are indicated in parenthesis.

	BEAST tMRCA & Fiebig stage Overlap	Post-infection Diversity	Pre-infection Diversity	Total # Sequences
				
Subtype B	70 (and 2 heterogeneous)	5	22 (and 3 homogeneous)	102
Subtype C	53 (and 1 heterogeneous)	0	14 (and 1 homogeneous)	69

A recent publication has indicated that the comparative approach for identifying homogeneous infections, where model-based estimates and clinical estimates of the time since infection are compared, is only valid for patients in Fiebig I and II, during which time the viruses are evolving neutrally (Lee et al., 2009). Since selection is likely to influence model-based estimates of the tMRCA, the comparative approach on its own is not suitable for classifying homogeneous infections in patients within the later Fiebig stages (III-VI). Similarly, if the BEAST estimates for datasets from later Fiebig stages (III-VI) suggest a shorter time from infection than the clinical estimates, which would erroneously place the infection within the homogeneous group, it is possible that purifying selection has taken place in a heterogeneous infection (Lee et al., 2009).

Nevertheless, BEAST, as a coalescent-based approach for determining the MRCA of a collection of sequences from a population, provides a very useful tool for evolutionary analysis. In the present research, if only BEAST estimates and Fiebig stage ranges were available, only $^{2}/_{77}$ infections for the subtype B, and $^{1}/_{54}$ infections for the subtype C analyses, would be classified as homogeneous, although further investigation revealed that the infections were caused by more than one viral strain (Keele et al., 2008; Abrahams et al., 2009). Furthermore, only 4 infections were misclassified as heterogeneous instead of homogeneous. These

results suggest that BEAST is a very valuable resource for studies aimed at estimating the time to most common recent ancestor of a group of sequences, and comparing the estimates to laboratory staging criteria for classifying homogeneous and heterogeneous infections.

2.3.2 Comparison Between the Poisson Evolution Model and BEAST

The results from the Poisson evolution model and those obtained from the coalescent analysis performed with BEAST were compared. A total of 164 estimates were available, and the BEAST estimate was smaller than the corresponding Poisson model estimate in only $^{4}/_{164}$ cases. In the remainder of cases the estimates from the Poisson model were smaller than the BEAST estimate. The more diverse the underlying sequence alignment, the greater the difference between the BEAST and Poisson estimates.

The mean tMRCA estimates from the BEAST and Poisson models, for all the homogeneous infections, were compared. The scatterplots in Figures 2.10 and 2.11 illustrate the relationship between the tMRCA estimates from BEAST and the Poisson models for HIV-1 subtype B and subtype C respectively. In both cases the results indicated a strong positive correlation; for the subtype B analysis, $R^2 = 0.955$ and the gradient for the linear fit is 0.701 (Figure 2.10). The subtype C analysis resulted in a $R^2 = 0.968$ and a slope parameter in the linear model of 0.504 (Figure 2.11). If the two models produced the same estimates, the gradient of the linear model is expected to be equal to one. A comparison was carried out with the aim to establish which model perform better at predicting homogeneous infections.

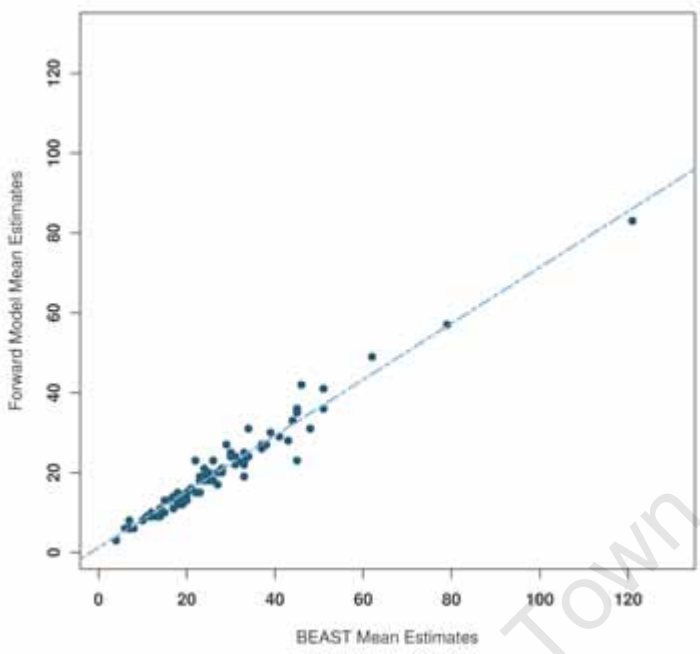


Figure 2.10: Scatterplot of the HIV-1 subtype B tMRCA estimates from the Poisson and BEAST models. The linear model is shown as a blue dashed line.

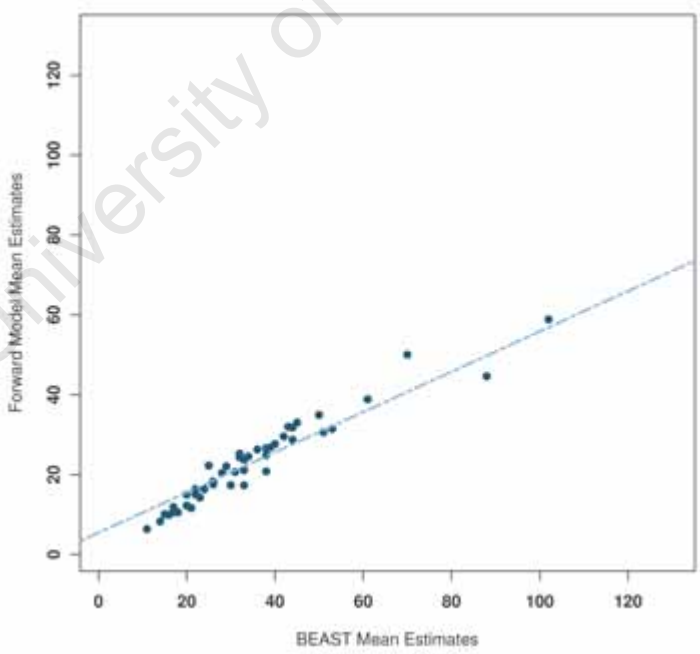


Figure 2.11: Scatterplot, and linear fit, of the HIV-1 subtype C tMRCA estimates from the Poisson and BEAST models.

The performances of the Poisson and BEAST models to classify homogeneous subtype B[1] infections, were compared. However, the Poisson model was not applied to all the patient datasets due to the complexity introduced by hypermutation or multiple recombination events (Keele et al., 2008), and the comparison was therefore carried out on the the patients that were classified as homogeneous infections by Keele et al., 2008. The number of individuals for which the predicted tMRCAs postdated the clinically estimated time of infection was three for both models. These patients would therefore have been classified as harbouring heterogeneous infections if further investigation was not carried out. The similarity between the prediction results is not entirely unexpected considering that the estimates from the Poisson model were consistently smaller than the BEAST estimates (as the gradient of the linear model in Figure 2.10 illustrates) in all but four cases. It is likely that the Poisson model would therefore do no worse than the BEAST model at classifying homogeneous infections as homogeneous, as in the present comparison.

2.3.3 Relating the HIV-1 Subtype B and C Analysis

A comparison between the results obtained from the HIV-1 subtype B and C analysis may reveal interesting subtype-specific transmission characteristics. In the studies carried out by Keele et al. (2008) and Abrahams et al. (2009), homogeneous infections were identified in 78 of 102 (76.5%) HIV-1 subtype B infected individuals, whereas 54 of 69 (78.3%) subtype C infections were labelled homogeneous. These results are very similar and lead to the hypothesis that 3 out of 4 HIV-1 infections are caused by the transmission of a single viral strain, regardless of the subtype.

Previous reports have suggested the existence of an HIV-1 transmission bottleneck (for example Derdeyn et al., 2004 and Sagar et al., 2004), however the subtype B and C studies compared here, have provided a quantification of this bottleneck. Furthermore, multiple strain infections were considered to be caused by either subsequent individual transmission events or simultaneous transmission of multiple viruses (Sagar et al., 2004; Gottlieb et al.,

[1]The analysis was not carried out on the subtype C data since the CIs for the mean estimates from the Poisson model were not available.

2004; Ritola et al., 2004). However, the low probability of sexual transmission (Ritola et al., 2004), together with the support from the results obtained from Keele et al. (2008) and Abrahams et al. (2009), suggest that multiple infections occur simultaneously and are not as frequent as homogeneous infections.

2.4 Conclusions

The results described in this chapter illustrate how a coalescent-based approach, for example BEAST, can be used to estimate the time to the most common recent ancestor of a group of sequences. If only the BEAST estimates and Fiebig stage durations were available, only 4 patients that were classified as harbouring homogeneous infections in the final published cases, would not be identified here. Furthermore, only $^{2}/_{77}$ subtype B and $^{1}/_{54}$ subtype C infections were classified as homogeneous infections and later shown to be caused by transmission of more than one virus. Therefore, only ~4% ($^{7}/_{171}$) of cases were misclassified. Comparing the BEAST tMRCA and modified Fiebig stage estimates to identify homogeneous and heterogeneous infections therefore corresponded to the published classification for ~96% of the infections (Abrahams et al., 2009; Keele et al., 2008).

Furthermore, the results from this study suggest that a single strain is transmitted from an infected individual to a new host in 3 out of 4 transmission events, providing a quantification to the transmission bottleneck previously described (for example Derdeyn et al., 2004). The knowledge that a remarkably high proportion of infections are caused by the transmission of a single viral strain, has a large impact on future early infection studies. If infection is caused by the transmission or outgrowth of a single virus, then the diversity observed in the new host represents the changes that occured after infection, and these changes are likely to play a role in viral adaptation to the new host. Identifying the mutations that occur after infection may therefore lead to new research avenues for developing effective prevention strategies. Furthermore, understanding the selective pressures acting on the virus during the earliest stages of infection may present novel approaches for drug and vaccine design.

This objective, to identify specific sites in the HIV genome that evolve adaptively during early infection, was the aim of the next chapter (Chapter 3).

Chapter 3

Evaluation of Selection Pressures and the Impact of APOBEC in Early Infection

3.1 Introduction

The work described in this chapter forms part of an international collaborative project, which was published in PLoS Pathogens during May 2009 (Wood et al., 2009). Not all of the content contained in the publication is included here, since parts of the research was carried out by collaborators. Parts of this chapter that are based on the work of collaborators and this is indicated clearly in the text. These sections were included to maintain the clarity and provide the context of this work. Although some changes to the originally published text have therefore been made, the results and eventual conclusions are the same.

Adaptation of HIV-1 to new hosts, in particular to the initial immune response, is likely to influence the establishment and progression of HIV infection. Mutations that enable escape from the host's immune responses will come under selective pressure, in an order that reflects the timing of the immune responses, underlying mutational rates, and the relative

fitness costs of mutations. In addition, escape mutations that occurred in the infecting host, particularly those resulting in a reduction of viral fitness (Chopera et al., 2008), will come under selective pressure to revert to wild-type in the newly infected individual (Friedrich et al., 2004). If transmission itself is a selective process, such that, for example, viral variants that are adapted to replication in mucosal cell types are selectively transmitted, the virus may experience selection to acquire forms that are adapted to replication in other host tissues. Since the viral envelope glycoprotein (Env) plays a vital role in transmission, is an early target of immune responses, and a key determinant of target cell tropism, the early evolution of the HIV-1 *env* gene is of particular interest (See section 1.2.4 for a description of Env).

The transmission of HIV is associated with a severe population bottleneck, such that, in most cases, new infections result from transmission or expansion of a single virus (Keele et al., 2008). As described in Chapter 2, it has recently been shown that the sequence of the transmitted virus can be identified accurately through direct sequencing of the uncloned DNA amplicons derived from single genome amplification (SGA) of plasma viral RNA sampled during acute and early infection (Salazar-Gonzalez et al., 2008; Keele et al., 2008). Keele et al. (2008) classified each of 102 HIV-1 subtype B primary infections into one of six clinical stages according to patterns of laboratory assay reactivity defined by Fiebig et al. (2003). Following the analysis described in Chapter 2, of the 102 subtype B patients, there was evidence of infection by or rapid outgrowth of a single virus (or in three cases possibly two or more very closely related viruses) for 81 of these individuals (Keele et al., 2008). In the present study a detailed analysis of the early evolution and divergence of the virus from the ancestral infecting strain in these 81 homogeneous infections was carried out. The study was restricted to the homogeneous infections because in these patients the observed sequence diversity is expected to be due mainly, or entirely (in the case of transmission of a single viral variant), to mutations that have occurred following transmission. The data analysed in this study consisted of sets of between 10 and 67 sequences per patient, generated previously using SGA from single time-points in acute or early infection (Keele et al., 2008).

Phylogenetic trees of the sequences from individual patients had short branch lengths and

few internal branches, reflecting the short time since transmission and rapid expansion in the new host. The individual sequence datasets contain very little information, and standard phylogenetic models are therefore not sufficient to gauge whether it is selection pressures or stochastic changes that are resulting in the diversification of the virus following transmission. However, by considering all patients simultaneously, it was possible to determine whether similar patterns of sequence diversification were repeated either in multiple patients or within the same patient. Shared patterns of rapid diversification at the same codons in early infection, can reveal sites that are typically involved in viral adaptation to the new host at the earliest stage of infection. In cases where these sites are associated with specific host immune responses, they shed light on the nature, extent, and timing of the immune responses that most frequently exert selective pressure on the transmitted virus during acute and early infection. However, there is also the possibility that shared mutations across multiple patients may be the result of deviation from the expected mutation rates, so that particular sites with high mutation rates, for example resulting from APOBEC-mediated substitutions, might be mutated in an unexpectedly high proportion of patients.

In this study, HyPhy (Kosakovsky-Pond and Muse, 2005), an interpreted computer programming environment for fitting evolutionary models, was used to develop a method which can identify sites in *env* that tend to diversify most rapidly across recently infected patients. This allowed for a further exploration of the reasons for rapid diversification at these particular sites. The main findings from the research described here, were that there is evidence of positive selection in early HIV-1 infection, which appears to be driven in many cases by escape from early cytotoxic T lymphocyte (CTL) responses; and furthermore, in several cases CTL escape occurred via mutations in the APOBEC sequence context, suggesting a role for APOBEC in determining the pathway of immune escape. These results provide the most detailed view yet of the process of diversification of the homogeneous viral population following transmission.

3.2 Methods

3.2.1 Dataset Assessment: Addressing Recombination and Hypermutation

Envelope sequences (n = 3476) from 102 individuals infected with HIV-1 subtype B, were generated using Single Genome Amplification (SGA), as part of a previous study (Keele et al., 2008). Among these individuals, 81 were found to be likely to have been productively infected by just a single virion (or infected cell), or two or more closely related viruses, while 21 were multiply infected since the level of diversity seen in these cases was greater than could have accrued since the time of infection. The 21 heterogeneous cases were not included in the analysis of selection pressure. Intra-patient alignments were generated through an iterative process, which included the alignment of all patient consensus sequences, manual editing, and re-alignment of each patients' sequence data to the associated consensus, as described in the Supporting Information of Keele et al. (2008).

Recombination analysis was carried out with GARD (Kosakovsky-Pond et al., 2006b) and evidence of recombination was found in just two of the homogeneous patients. A further 14 from the heterogeneous group (analysed previously Keele et al., 2008) showed evidence of recombination that was confirmed by either GARD (Kosakovsky-Pond et al., 2006b) or Recco (Maydt and Lengauer, 2006). For our previous study, acutely infected patients were staged using clinical criteria described by Fiebig et al. (2003), into six clinical stages from the earliest, Fiebig stage I, to the latest stage of early infection, Fiebig stage VI. All regions of the alignment that are translated in more than one frame, as well as all regions that were ambiguously aligned, were masked for the analysis of selection pressure.

In addition, for the model-based inference of selection, described in the subsequent section (section 3.2.2), individual sequences that showed statistically significant evidence of APOBEC-mediated hypermutation using Hypermut 2.0 (http://www.hiv.lanl.gov/content/-/sequence/HYPERMUT/hypermut.html) were discarded. Among the patients, 15 had a

generalized increase in the rate of G-to-A mutations in the context of the sequence motif associated with APOBEC hypermutation. In these cases there was an excess of G-to-A mutations in the APOBEC sequence context across all sequences from the patients and not concentrated in individual severely hypermutated sequences (Keele et al., 2008). Since in these cases APOBEC mediated mutations may be occurring at an elevated rate, and this context-specific rate would deviate from the evolutionary model, a separate analysis of selection excluding these 15 patients was carried out. On the other hand, recurrent G-to-A mutations in the context of APOBEC substitution motifs are enriched overall in these data, suggesting APOBEC may have a role in HIV-1 evolution in early infection, therefore, in parallel the full set of 81 subjects was analysed.

The alignments used for these analyses were from our previous study Keele et al. (2008) and are available at http://www.hiv.lanl.gov/content/sequence/hiv/user_alignments/keele.

3.2.2 Maximum Likelihood Model for Detecting Selection

HyPhy (Kosakovsky-Pond and Muse, 2005) was used to assess the evidence for a subset of codon sites in the *env* gene evolving under the influence of positive Darwinian selection (Kosakovsky-Pond et al., 2008) in early infection. Essentially, this method works by fitting standard codon models (Nielsen and Yang, 1998) of sequence evolution to multiple sequence alignments from all of the patients. The critical parameter of the codon models is ω, the ratio of nonsynonymous to synonymous substitution rates. If synonymous substitutions are assumed to be generally neutral, then sites for which ω is significantly greater than one are inferred to be evolving under positive Darwinian selection (Yang et al., 2003; Delport et al., 2009). However, deviation from the evolutionary model as a result of enhanced mutation rates at specific sites, for example, those embedded in APOBEC signature motifs, can also contribute in some cases to an elevated ω (see section 3.3).

The random effects likelihood approach (Yang and Nielsen, 2000) and a model of site-to-site variation in ω equivalent to Model M2a from Wong et al. (2004), was used. The selection

parameter, ω, was modelled using a discrete distribution with three classes: a purifying selection class for which ω, constrained to be less than one, is estimated from the data; a neutral class for which $\omega = 1$, and a positive selection class for which ω is estimate from the data and constrained to be greater than one. The fit of this model is compared, using the likelihood ratio test, to a model equivalent to M1a from Wong et al. (2004), in which the proportion of sites in the positive selection category is fixed at zero. Both of these are mixture models. In the standard implementation of these models the likelihood of the data is given by:

$$L(D|\theta, T) = \prod_i \left(\sum_j L(D_i|\omega_i = j, T, \theta) p(\omega_i = j|\theta) \right)$$

where the product is taken over sites of the alignment, the sum is over the discrete ω classes $(j_1..j_3)$, and D_i, T and θ are the data at site i in the alignment, the phylogenetic tree and the parameters of the model, respectively. Here, this is extended to data from multiple patients by writing the likelihood as:

$$\prod_i \sum_j \left(\prod_k \Big(L\big(D_i^k|\omega_i = j, T_k, \theta\big) \Big) p\big(\omega_i = j|\theta\big) \right)$$

where ${D_i}^k$, and T_k are the data and tree from patient k, respectively. In this formulation ω has the same value at a given site across all within-patient sequence alignments.

Following likelihood parameter estimation, posterior probabilities of belonging to each of the site classes of ω were estimated for each site using the naïve empirical Bayes method (Yang et al., 2005). For all of the models, mutation rates were estimated separately for each pair of nucleotides, thus assuming a general time-reversible model of nucleotide substitution, and the Goldman and Yang method (Goldman and Yang, 1994) was used to account for biased codon usage.

3.2.3 ELIspot and Intracellular Cytokine Assay

The work described in this section (section 3.2.3) was carried out by collaborators (Wood et al., 2009). The results obtained from these laboratory experiments, are relevant to the conclusions drawn, and the description of the methods, provided by these collaborators, is therefore included here.

3.2.3.1 ELISpot assays

Ex-vivo IFN-γ ELISpot assays:

Eighteen- and 9-mer peptides were synthesised (MRC Human Immunology Unit, WIMM, Oxford UK) to match sequences from patients MEMI, Z13, Z33 and Z36. Cryopreserved Peripheral Blood Mononuclear Cell (PBMC) samples were defrosted, incubated for 2 hours and placed in the Enzyme-linked immunosorbent spot (ELISpot) plates at 1 x 105 cells/well. Peptides in the peptide plates were mixed with 1:1 with PBMC in the ELISpot plate to a final concentration of 2 µg/ml and incubated for 20 hours at 37°C, 5% CO_2. PBMCs and peptides were tested in R-10 medium (10% fetal bovine serum, 86% RPMI 1640, 2 mM L-glutamine, 1 x penicillin-streptomycin solution, 10 mM HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) buffer and 1 mM sodium pyruvate, all Sigma). Coating, development and reading of ELISpot plates have been described previously (Goonetilleke et al., 2006). A definition of positive responses was applied as: $\geq$ 50 SFU/million, $>$ 4 x background. For all assays 6 negative control wells (media only) and 2 positive control wells (10 µg/ml PHA, Sigma) were used.

Cultured IFN-γ ELISpot assays:

This assay has been previously described in Goonetilleke et al. (2006). In brief, short-term cell lines (STCLs) were set up whereby 2 x 106 cell/ml of PBMCs were stimulated with a

pool of peptides specific to the patient at 2 µg/ml in the presence of 25 ng/ml IL-7 (R&D Systems). Cells were cultured at 37°C, 5% CO_2, for 10 days in RAB-10 (10% human AB serum, 86% RPMI 1640, 2 mM L-glutamine, 1x penicillin-streptomycin solution, 10 mM HEPES buffer and 1 mM sodium pyruvate, all Sigma) and were stimulated with 1800 U/ml IL-2 (Novartis) on days 3 and 7. On day 10, the STCLs were washed three times with PBS (Sigma) and incubated for 30 hours in RAB-10 at 37°C, 5% CO_2. On day 11, 4 x 104 cells/well were placed into an IFN-γ ELISpot as described above and stimulated by each individual peptide in triplicate at 2 µg/ml in the presence of RAB-10.

HLA typing:

HLA typing was performed by T. Rostron (WIMM, Oxford UK) using the Sequence-Specific Primer (SSP) method adapted from Bunce (2003), that uses allele-specific primer combination in PCR amplification to provide absolute HLA resolution to 2-digits and high probability resolution to 4-digits.

3.2.3.2 Intracellular Cytokine Assay

Cryopreserved PBMCs were defrosted, incubated overnight and stimulated for 6 hours at 37°C, 5% CO_2, with 2 µg/ml of individual HIV-1 peptides representing the wild-type and escape sequences of the epitopic regions. The stimulation was conducted in the presence of anti-CD28 mAb (1 mg/ml; L293; BD Biosciences), anti-CD49d mAb (1 mg/ml; L25; BD Biosciences), anti-CD107a-Alexa 680 (H4A3; BD Biosciences), 5 µg/ml Brefeldin A (Sigma) and 1 µg/ml Golgi Stop (Pharmingen). After stimulation, the PBMCs were stained with LIVE/DEAD Fixable Violet Dead Cell Stain (InVitrogen, Eugene, OR), anti-CD14-Cascade Blue (M5E2), anti-CD19-Cascade Blue (HIB 19) (to exclude non-viable cells), anti-CD4-Cy5.5-phycoerythrin (PE) (M-T477), anti-CD8-QD705 (RPA-T8), anti-CD27-Cy5-PE (M-T271), anti-CD57-QD605) (NK-1), anti-CD45RO-PE-Texas Red (all from BD Biosciences, except CD45RO, which was from BeckmanCoulter) for 20 minutes at room temperature. PBMCs were then fixed and permeabilised with Cytofix/Cytoperm (Pharmingen). PBMCs

were stained with anti-CD3-Cy7-APC (SK7), anti-IFN-g-fluorescein isothiocyanate (FITC) (B27), anti-IL-2-Allophycocyanin (APC) (MQ1-17H12), anti-TNF-a-Cy7-PE (MAb11) and anti-MIP-1b-PE (D21-1351) (all BD Biosciences) for 45 min at 4°C. After washing and fixation, samples were run on a custom built LSRII (BD Bioscences). A minimum of 300,000 total events was acquired using a custom made LSR II flow cytometer (BD Bioscience, San Jose, CA). Data analysis was performed using FlowJo 8.8.2 software (TreeStar), Pestle for background subtraction, and Spice for frequency analysis. (Pestle and Spice provided by Dr. M. Roederer Vaccine Research Center, NIH, Bethesda). 'Memory' CD8 T-cells are defined as live / lymphocytes / CD3 / CD8 / excluding the CD27hiCD45RO-negative population. Individual and total cytokine responses were considered as positive if the frequency of the total responding memory CD8+ subset was greater than 0.05% after background subtraction.

3.3 Results and Discussion

The availability of data from a large number of individuals in acute and early infection that harbour a homogeneous set of viruses, leads to an opportunity to study the pattern of diversification of HIV-1 upon transmission to a new host. Since 78 of these patients are believed to have been productively infected by single viruses (three others are likely to have been infected by two or more closely related viruses), the observed diversity can be inferred to have arisen in the newly infected patient and not in the individual from whom the infection was acquired (Keele et al., 2008). The size of the sample and the quality of the sequence data provide the first opportunity to assess whether there are consistent patterns of evolution that tend to recur across multiple patients in early infection.

3.3.1 Model-based Inference of Selection in Early Infection

HyPhy (Kosakovsky-Pond and Muse, 2005) was used to assess the selective pressure acting on viruses in early infection. Models of codon sequence evolution were fitted to 2,207 *env* coding sequences from early stages of HIV-1 infection as described in section 3.2.1. In order to

accumulate evidence of selection from multiple patients, the models were implemented such that at a given site the value of ω (the ratio of nonsynonymous to synonymous substitution rates) was constrained to be the same in all patients. This enabled us to use standard model comparison techniques to investigate whether there was evidence that the mean value of ω across patients was significantly greater than one at a subset of sites in the alignment (see section 3.2.2). It was found that, as expected, the majority of codon sites evolve under the influence of purifying selection (with $\omega < 1$) in early infection, and the mean value of ω across all sites in all patients, was 0.67.

Three overlapping groups of patients were considered, consisting of individuals in the earliest clinical stages (Fiebig stages I and II), earliest through intermediate stages (Fiebig stages I through III) and all patients in early infection (Fiebig stages I through V). For the data from Fiebig stages I and II the model allowing a subset of sites to evolve under positive selection did not provide a significantly better fit to the data than the model in which positive selection was not permitted (p = 0.52; Table 3.1). For the other two datasets, however, the selection model (M2a) provided a significantly better fit to the data than the model in which selection was excluded (M1a). M1a was rejected against M2a with p-values 0.006 and 0.0001 in the case of the datasets from Fiebig stages I-III and stages I-V, respectively (Table 3.1).

Table 3.1: Parameter estimation for the neutral and selection model applied to *env* sequences from different Fiebig stage datasets.

Fiebig Stage	dN/dS: Neutral Model (M1a)	dN/dS: Selection Model (M2a)	2 * Delta Log Likelihood	P-value (M1a vs M2a)
I – II	0.6912	0.6995	1.2833	0.5264
I – III	0.6687	0.7016	10.0741	0.0065
I – V	0.6307	0.7130	18.3186	0.0001

Before this analysis was carried out, all individual sequences with evidence of APOBEC-

mediated hypermutation were removed using Hypermut 2.0 (www.hiv.lanl.gov). For fifteen of the patients the *env* sequences showed an overall elevated rate of G-to-A mutations in a sequence context associated with APOBEC hypermutation (using the methodology described in the supporting information of Keele et al., 2008, see Figure 3.1), even though no individual sequences showed evidence of hypermutation. The *Highlighter* plots (http://www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter.html) in Figure 3.1 illustrate the different arrangements of hypermutation. Statistical tests of hypermutation were performed using Hypermut (http://www.hiv.lanl.gov/content/sequence/HYPERMUT/hypermut.html). As described in Chapter 2, most of the patient datasets displayed no evidence of hypermutation (Figure 3.1 A). A further scenario is where a single sequence within a patient sample is significantly hypermutated ($p < 0.1$; Figure 3.1 B). A patient with an overall elevated rate of hypermutation is shown in Figure 3.1 C. Although no individual sequence showed significant evidence of hypermutation in this patient, after compressing the mutations into a single sequence (using a tool available from the *Highlighter* website, http://www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter.html), and re-running Hypermut with the compressed sequence, there was a significant elevation in the rate of G-to-A mutations in the APOBEC sequence context ($p = 9 \times 10^{-6}$).

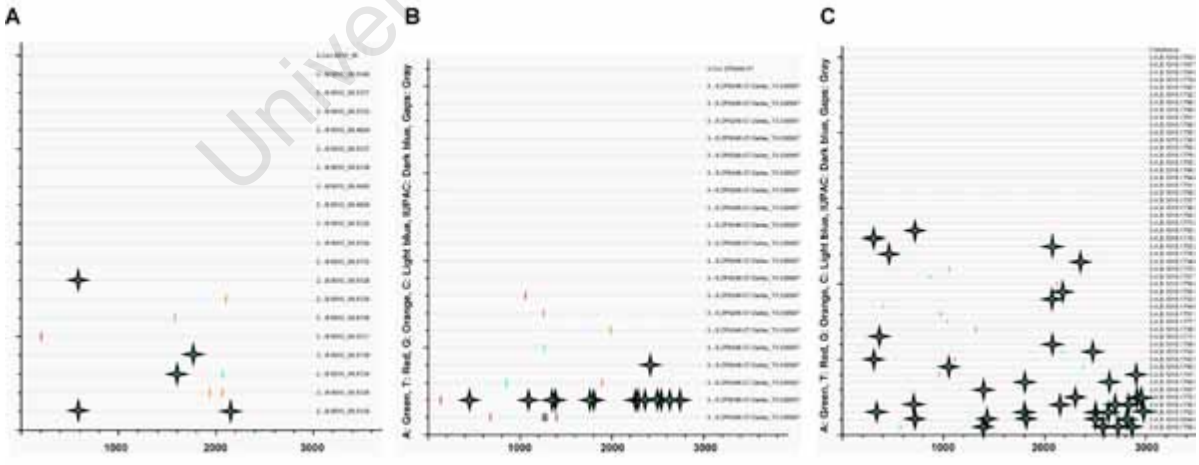


Figure 3.1: *Highlighter* plots illustrating, with green stars, different patterns of hypermutation. A) Sequences from a patient showing no evidence of hypermutation. B) A patient with a single significantly hypermutated sequence. C) Sequences from a patient showing an overall elevated rate of hypermutation.

Hotspots of mutation could create false positive signatures of positive selection. This is particularly true if the same hotspots are repeated across multiple patients, as would be expected for mutation hotspots resulting from APOBEC hypermutation. To limit the impact of hypermutation on our results, these patients were removed from the dataset, leaving 66 subjects. Interestingly, even with these precautions, it was found that sites that were placed in the set of sites with $\omega > 1$ (positive selection site class) are enriched for the APOBEC-motif (see Table 3.2). Given the potential biological relevance of these mutations, a second analysis including all 81 homogeneous patients was performed, bearing in mind the caveat that the model may be affected by the higher rate of G-to-A mutation in the APOBEC context. Therefore, some of the sites identified from this dataset may reflect mutational hotspots rather than selection acting on the amino acid sequence.

A statistical model comparison was used to make the claim that there is evidence of adaptive evolution in HIV-1. To do this, the fit to the data of a model in which no sites in the alignment were allowed to evolve with $\omega > 1$, was compared to the fit of a model in which a subset of sites were allowed to have $\omega > 1$. The latter model had a significantly better fit to the data for the sequences from Fiebig stages I-III and Fiebig stages I-V, suggesting that the virus evolves adaptively in these clinical stages. However, there are two important issues to consider here. First, the fact that the positive selection model did not have a significantly better fit to the data in Fiebig stages I-II does not prove that viral diversification is not shaped by positive selection at this stage of infection. Sequences from stages I-II were highly homogeneous, thus a very small number of mutations from the infecting virus were observed in each patient, severely limiting the power to detect evidence of selection. Secondly, a more general issue is that models of codon evolution used to detect positive selection rely upon several assumptions about the way the sequences are evolving, including the absence of recombination and independence between nucleotide sites, and deviations from these assumptions could lead to false positive results (Suzuki and Nei, 2002).

APOBEC induced mutations give rise to hotspots of amino acid change which have a mutational rather than a selective cause. Although an attempt was made to control for the effects of hypermutation by removing sequences and patients with evidence of an elevated

rate of mutation in the APOBEC3 sequence context (Figure 3.1), the enrichment of mutations in this context among sites placed in the positive selection site class, suggests that hypermutation may have had an effect on the model comparison results.

It is of interest to test for the evidence of directional change in *env* in early infection, since this could impact the understanding of the nature of the transmitted virus. As is clear from several studies (Keele et al., 2008; Derdeyn et al., 2004; Sagar et al., 2006; Edwards et al., 2006), there is a severe viral population bottleneck during HIV transmission, with newly transmitted infections resulting frequently from transmission or outgrowth of a single virion. If this extreme bottleneck is non-random and if viruses that are best adapted for transmission from one host to the next tend to be less well adapted for replication in the new host, then there may be selection for reversion of the characteristics associated with transmission in the newly infected host. If, further, these characteristics are shared across multiple transmitted viruses, then such reversions in early infection could potentially be observed through our approach. Here, evidence for purifying selection was found, indicating that the transmitted virus was a reasonably fit form prior to the initial immune response imposing new selection constraints. Furthermore, relatively little support for characteristic patterns of reversion at the population level in this cross-sectional view was identified.

3.3.2 Analysis of Rapidly Evolving Sites

Since the sequence datasets from early infection are very homogeneous and generally well described by a star-like phylogeny, it was rare to find multiple independent mutations within the same patient, and consequently, primarily sites that tended to diversify rapidly in many patients were identified. Twenty-four sites were found in *env* with posterior probability of greater than 0.5 of belonging to the positive selection site class based on the 66 subjects discussed above (Table 3.2).

Table 3.2: Positive selection results obtained using HyPhy from a dataset excluding individuals with sequences enriched for APOBEC hypermutation as well as from the complete dataset including the hypermutated sequences (the latter sites are indicated with +). The location, timing, and mutational patterns observed are provided. The sites for which CTL testing was carried out or not are shown in blue font.

HXB2	Data	Location	Posterior Probability	Number of subjects with variation	Contribution of APOBEC3	Number of subjects with a mutational pattern
62	79	gp120 C1	0.765	5	4 of 5	2 D to N, 1 D to Y, 2 E to K
64	81	gp120 C1	0.505	2	No	*2 E to G*
66	83	gp120 C1	0.979	4	No	*4 H to Y*
175	226	gp120 V2	0.983	7	No	3 L to P, 3 L to F, 1 N to S
176	227	gp120 V2	0.655	2	No	1 F to S, 1 F to V
232	306	gp120 C2	0.853	4	No	1 K to E, 1 K to R, 1 T to A, 1 T to M
242	316	gp120 C2	0.652	3	No	2 V to I, 1 V to L
274	349	gp120 C2	0.722	3	No	1 S to T, 1 S to F, 1 S to P
322 +	396	gp120 V3	0.510	4	2 of 4	2 E to K, 1 E to G, 1D to N
337	412*	gp120 C3	0.932	5	2 of 5	1 K to E, 1 K to R, 1 E to K, 1 D to N, *1 D to A*
344 +	419	gp120 C3	0.588	3	No	*1 K to M*, 1 Q to L, 1 R to G
347 +	422	gp120 C3	0.910	5	4 of 5	2 R to K, 1 K to R, 1 G to R, *1 G to E*
354	431	gp120 C3	0.895	3	2 of 3	1 E to K, 1 G to E, *1 K to R/T*
360	439	gp120 C3	0.897	3	No	1 V to G, 1 V to A, *1V to G/A*
460	566	gp120 C4	0.870	2	No	1 N to K, *1 D to N/K/T*
482 +	601	gp120 C5	0.669	5	5 of 5	*5 E to K*
509	634	gp120 C5	0.904	7	7 of 7**	7 E to K
513	638	gp41	0.728	2	No	1 V to S, 1 V to G
518	646	gp41	0.999	9	No	2 M to I, 7 M to V
587	719	gp41	0.770	2	No	2 L to I
588	720	gp41	0.546	3	2 of 3	2 R to K, 1 K to R
612	744	gp41	0.841	2	No	*1 I to T/S/N*, 1 T to I
651	789	gp41	0.598	3	No	*2 N to S*, 1 S to G
696	834	gp41	0.635	7	6 of 7	*6 R to K*, 1 R to S
700	838*	gp41	0.535	3	No	*1 A to V*, 1 A to T,1 T to S
702 +	840	gp41	0.533	3	No	2 L to F, 1 L to P
703 +	841	gp41	0.541	3	No	1 S to A, 1 S to F, 1 S to P
831	979	gp41	0.664	2	2 of 2	*2 E to K****
833	981	gp41	0.993	2	No	1 V to A, *1 V to A/G*
841	989	gp41	0.772	4	No	1 L to H or gap, 1 L to P, 1 L to F, 1 I to T

Key for Table 3.2:
HXB2: Coordinates listed according to HXB2 numbering (http://www.hiv.lanl.gov/content/sequence/LOCATE/locate.html).
Data: Coordinates listed according to the protein alignment of all 81 early infected subjects.
Location: Region in Envelope.
Posterior Probability: Sites identified in HyPhy with a posterior probability > 0.5 are included in this table.
Number of subjects with variation: Out of the 81 subjects, the number that had any variation in this site.
Contribution of APOBEC: The number of subjects, among those that vary in a given position, that have a G-to-A change in the context of an APOBEC3 motif.
Number of subjects with a mutational pattern: This summarises all individuals with changes found in this position in the data. *Italics* means a change was found more than once in at least one individual with the pattern.
Example 1: HXB2 Site 651: *2 N to S*, 1 S to G means: Two people had N as the most common amino acid, but S was present. PRB931 had 17 N and 2 S, PRB956 had 25 N and 1 S. Because there was more than one S in one of them, it is in *italics*. One individual, 700010077, is noted as S to G, and had 51 S and 1 G.
continues...

Table 3.2 continued:

Example 2: HXB2 Site 518: 2 M to I, 7 M to V means: Nine people had M as the most common amino acid, and each of the nine had a single variant among their sequences; the variant was I in two, and V in seven subjects. For example, Patient 1012 had 42 sequences, with 41 M and 1 V.

*In Keele et al. (2008), it was noted patients bearing these changes might have been infected with more than one closely related form. Alternatively, early selection or maintenance of a very early mutational event might be giving rise to the pattern.

**This site was difficult to align in a few patients due to a frameshifting insertion of an A in a string of As in the primary sequences, thus a number of the 7 noted E to K changes were actually due to a frameshifting indel.

***In an additional subject there was an ambiguous base call in this position.

Thirteen of the selected site occurred in gp120 and eleven in gp41 (Figure 3.2).

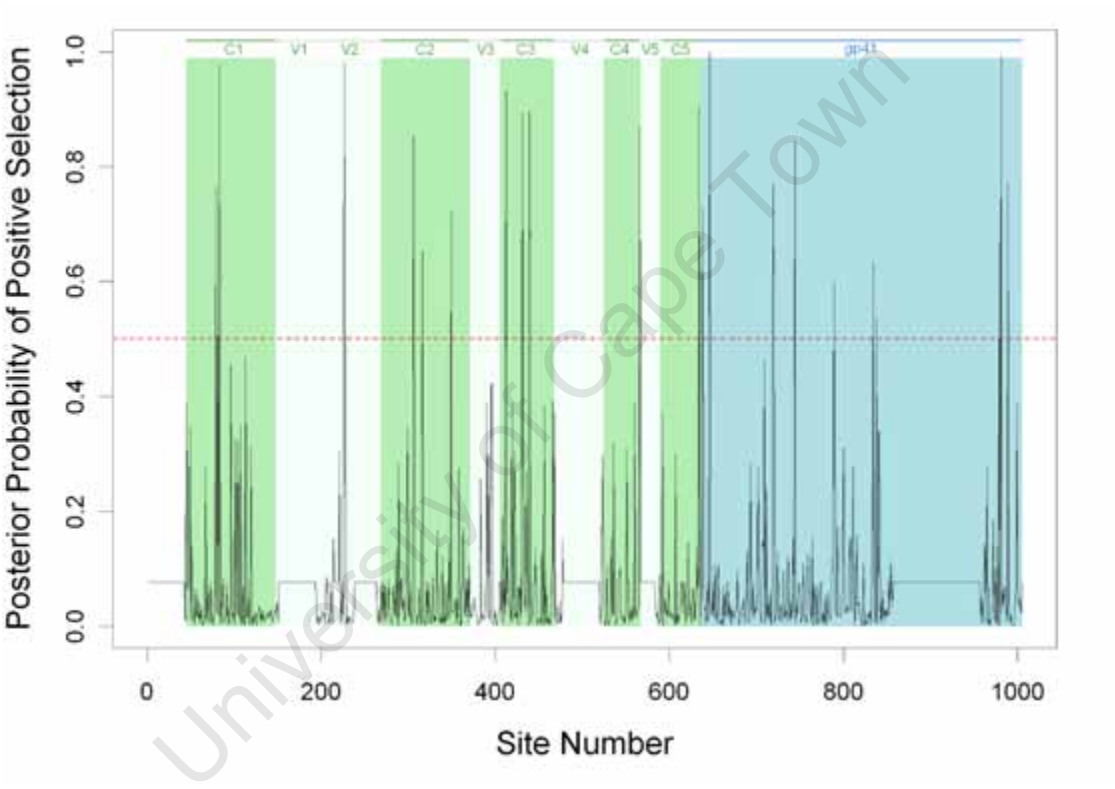


Figure 3.2: Posterior probabilities of belonging to the positive selection class ($\omega > 1$) for all sites in gp160. The dashed line indicates the 0.5 posterior probability, which was used as a threshold to assign sites to the selection site class. Flat parts of the graph correspond to sequence regions that were masked either because they were poorly aligned, or coding in more than one frame.

The three-dimensional structure context of the selected sites in gp120 is shown in Figure 3.3 A. There is a cluster of sites in the C3 region of gp120 (Table 3.2) that are in the positive selection class. In general, sites evolving under positive selection appear to be at least par-

tially solvent exposed (Figure 3.3 A). Six additional sites were identified by including all 81 subjects in the analysis (Table 3.2, and Figure 3.3 B); 4 were in gp120.

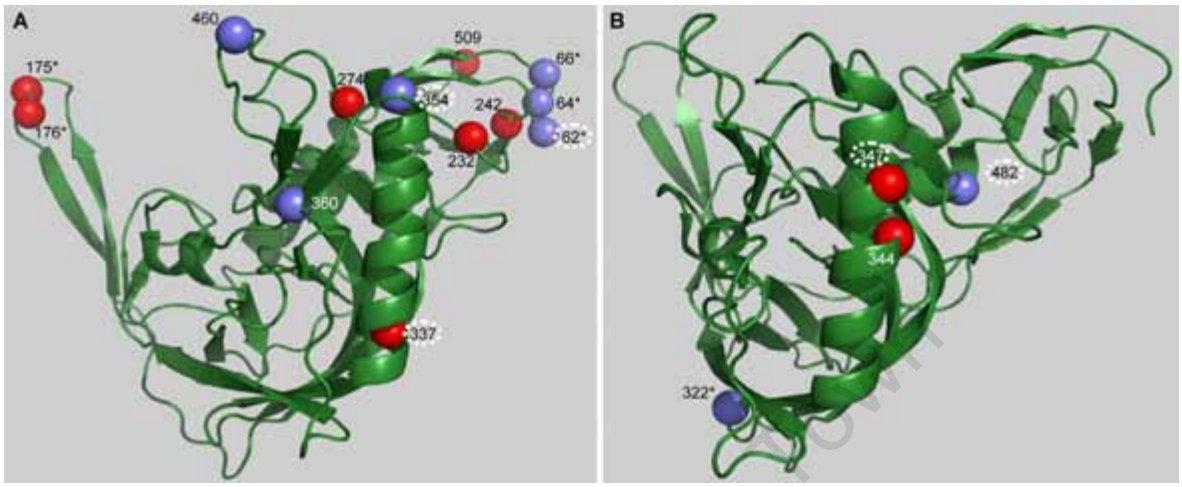


Figure 3.3: Three-dimensional structure context of the selected sites identified in gp120. A) 13 sites from the dataset excluding sequences with evidence of APOBEC-mediated hypermutation, and B) 4 additional sites identified analyzing the complete dataset. The sites depicted in blue represent those sites that are embedded in a known or potential CTL epitope and circled site numbers are potentially affected by APOBEC hypermutation. Sites marked with an asterisk occur in a region for which there is no available structure, therefore the positions are shown in proximity to their actual locations.

For all of the sites identified, the sequence alignment was considered carefully to assess the possibility that the evidence of positive selection could be resulting, for example, from APOBEC-mediated mutations or alignment errors associated with indel hotspots (Table 6.4 in the Appendix provides a detailed list of the observed mutations and their sequence context). For one site (at position 509 on the HXB2 sequence), alignment error introduced by an indel hotspot provides a better explanation of the data than selection. This site occurs at the beginning of a long sequence of Adenines that could serve as an indel hotspot (Pathak and Temin, 1990) and mutations at this site could, in several cases, be more parsimoniously explained as indels.

For 7 of the 24 identified sites (indicated in Table 3.2) some or all of the mutations involve

G-to-A in the APOBEC3G or APOBEC3F hypermutation motif (the motif is GRD, where R is the IUPAC code for G or A, and D for G, A, or T). A baseline of 12.3% of the bases in the patients' transmitted sequences corresponded to a G embedded in an APOBEC3 motif, while 26.8% (25/93) of the distinct mutations found in the 24 sites identified as evolving under positive selection involved a G in an APOBEC3 motif. When all 81 patients were included, 12.3% of the bases in transmitted sequences were G's in an APOBEC3 motif, while 28.6% (32/112) of the observed changes in selected sites involved a G in an APOBEC3 motif. In contrast, G's not embedded in an APOBEC motif were not enriched among sites under positive selection relative to the transmitted virus. Indeed, the proportion of G's embedded in an APOBEC3 motif was significantly higher for the selected sites than for the remainder of the virus in both the 66 patient subset (Fisher's exact test $p = 0.004$), and the full 81 patient dataset ($p = 0.0002$) . This suggests that a proportion of the sites are identified as belonging to the positive selection class due to the elevated APOBEC3 mediated mutation. Since the waiting time for G-to-A mutations in the APOBEC3 context may be shorter than for other mutations, it is possible that selection acting on these mutations may be a characteristic of early immune escape.

3.3.3 Investigating Potential CTL Escape Mutations

There were 9 positively selected sites that were located in 5 potential CTL epitopes. These selected sites were investigated as potential CTL epitopes because each was embedded in a localised region with multiple mutations in one of the study subjects (Figure 3.4), and some of the variants were repeated multiple times within the subject (Figure 3.4), two features providing corroborative evidence for ongoing positive selection for immune escape.

Specifically, in Figure 3.4, A to E) are based on i) evidence for selection from the model implemented in HyPhy, ii) evidence of selection from locally concentrated regional evidence for selection relative to extremely homogeneous early infection alignments, and iii) HLA appropriate known epitopes or anchor motifs found using the Los Alamos HIV immunology database epitope location finder, ELF. The set of mutations shown in Figure 3.4 F, includes

the only other sites in the entire data set of 81 subjects where mutations tended to cluster over a small region in one person. No site in this region had statistical support for positive selection, but this cluster of mutations was embedded in a potential epitope.

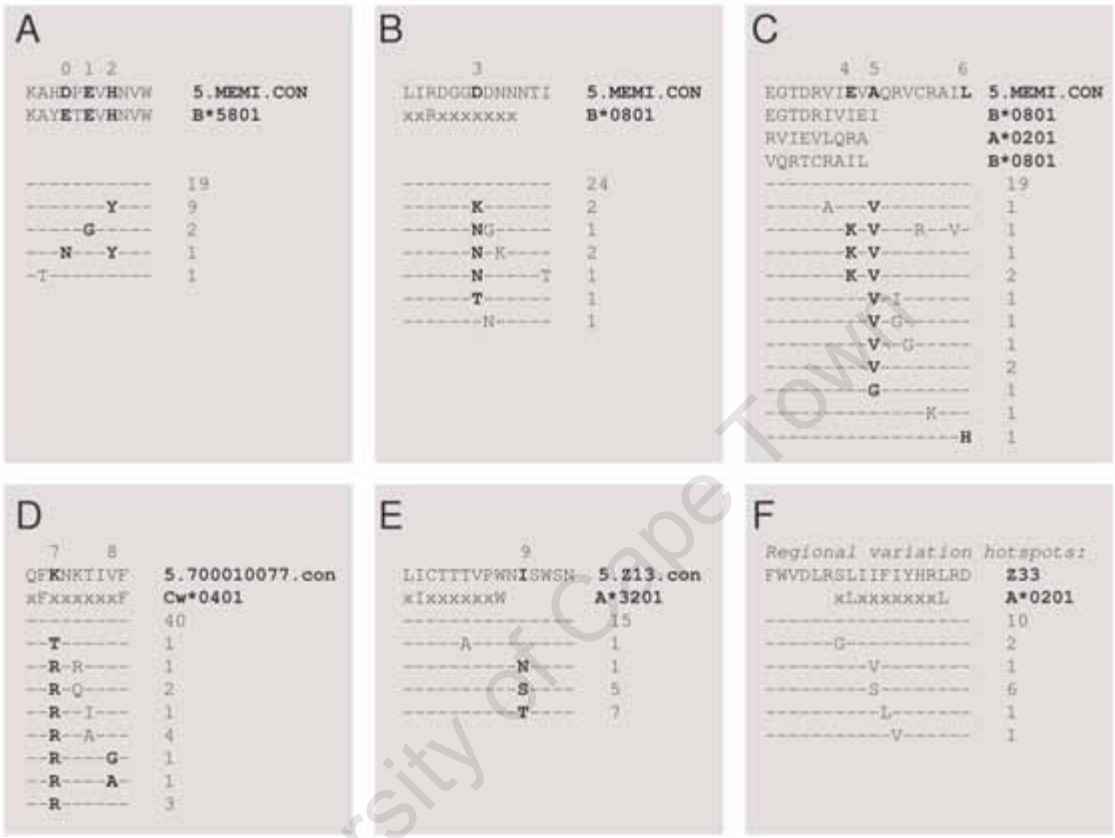


Figure 3.4: Selected sites that are embedded in potential CTL epitopes. The patient consensus sequence is shown at the top of each alignment, with Fiebig stage, patient ID, and CON for consensus indicated. The proposed epitope is shown beneath the patient consensus, followed by the HLA. Previously reported epitopes are provided in full, predicted epitopes are written with uppercase letters representing the anchor motif embedded in a string of x's.

Key to patient HLA's:

MEMI: HLA A01/A02, B08/B58

700010077: HLA A*0205/A*0205, B*5301/B*5701, Cw*0401/Cw*1801

Z33: HLA A02/A68, B45/B53

Z13: HLA A01/A32, B08/B44

Key to selected sites found in the alignment (HXB2/Alignment Coordinate):

0: site 62 / 79 1: site 64 / 81 2: site 66 / 83

3: site 460 / 566 4: site 831 / 979 5: site 833 / 981

6: site 841 / 989 7: site 354 / 431 8: site 360 / 439 9: site 612 / 744

In each of these five cases, the cross-sectional sample appeared to capture a moment of transition in the viral population, with the virus exploring different escape options. In each of these cases, patient HLA data were available, and the clusters of mutations were found to occur either within a known CTL epitope from the literature, or within anchor motifs appropriate for the HLA type of the infected individual (all potential epitopes are shown in Figure 3.4 and summarised in Table 3.3, identified using the tool ELF http://www.hiv.lanl.gov/content/sequence/ELF/epitope_analyzer.html). As this was a cross-sectional study, the sequences from which selective pressure is inferred are from the moment of transition, thus only direct experiments can determine which of the variants represent the susceptible form.

In order to determine whether there indeed was a T-cell response to these putative epitopes, 18 amino acid long peptides representing the most common and second most common form of the regional variants encompassing the variable sites, as well as the predicted epitopes shown in Figure 3.4, were synthesised and tested using either ELISpot or ICS (see section 3.2.3). For example, three sites (79, 81 and 83 in the alignment), showing evidence of adaptive evolution (Table 3.2) were mutated in patient MEMI. These sites were embedded within a previously described B*5801 epitope, and MEMI carried HLA B*5801 (Figure 3.4 A). Upon testing, MEMI in fact did make a T-cell response to the most common form of the B*5801 epitope (KAHDPEV**H**NVW) and a diminished response to the variant KAHDPEV***Y***NVW (Table 3.3).

It was possible to detect a T-cell response to 4 of the 5 proposed epitopes tested (Table 3.3). The lack of the response to the other proposed epitope could imply an alternative cause for the selection at this site (for example, antibody escape or reversion). However, it is also possible that all of the observed forms of the peptide are escape variants, and that the susceptible form has decreased in frequency to the extent that it does not appear in the sample and was therefore not tested. Further possible explanations include a transient T-cell response driving the escape that was already beginning to subside due to immune escape and reduced stimulation, or limited sample viability (this was a retrospective study, and the condition of several of the samples was not ideal).

Table 3.3: CTL results indicating the five regions tested as well as the specific epitope sequences. Sites that occur in the APOBEC3 context are shown in *italics*.

Alignment coordinate *[APOBEC3]*	Patient (Fiebig stage)	Sequences Tested (Database Analog)	Epitope Annotation	EliSpot SFU/10^6 PBMCs	Cultured Elispot SFU/10^6 PBMCs	ICS % Total CD8+ memory T-cells	Summary
79, 81, 83	MEMI (V)	KAHDPEVHNVW (B*5801)	Putative epitope	487	15144	N/A	Positive
		KAHDPEV**Y**NVW	Common variant - escape?	222	10227	N/A	Diminished
		ASDAKAHDPEVHNVWATH	18mer bound selected area	133	1655	N/A	Positive
		ASDAKAHDPEV**Y**NVWATH	18mer common transmitted variant	Negative	2086	N/A	Diminished
431, 439	700010077 (V)	QFRNKTIVF (Cw*0401)	Putative epitope	477	N/A	N/A	Positive
		REQFRNKTIVFNHSSGGD	18mer bound selected area	107	N/A	0.71 (32d)	Positive
		REQF**K**NKTIVFNHSSGGD	18mer common transmitted variant	Negative	N/A	N/A	Negative
		LSHVVDKLREQF**K**NKTIV	18mer common transmitted variant	Negative	N/A	N/A	Negative
979, 981, 989	MEMI (V)	RVIEVAQRV (A*0201)	Putative epitope	Negative	9930	N/A	Positive
		RVI**K**V**V**QRV	Common variant - escape?	Negative	2468	N/A	Diminished
		EGTDRVIEVAQRVCRAIL	18mer bound selected area	Negative	Negative	N/A	Negative
		EGTDRVI**K**V**V**QRVCRAIL	18mer common transmitted variant	Negative	Negative	N/A	Negative
		EGTDRVIEVA (B*0801)	Putative epitope	Negative	Negative	N/A	Negative
		EGTDRVI**K**V**V**	Common variant - escape?	Negative	Negative	N/A	Negative
		AQRVCRAIL (B*0801)	Putative epitope	Negative	Negative	N/A	Negative
		VQRVCRAIL	Common variant - escape?	Negative	Negative	N/A	Negative
744	Z13 (V)	LICTTTVPW (A*3201)	Putative epitope	N/A	N/A	Negative	Negative
		GKLICTTTVPWNISWSNK	18mer bound selected area	572	N/A	N/A	Positive
		GKLICTTTVPWN**T**SWSNK	18mer common transmitted variant	N/A	N/A	Negative	Negative
566	MEMI (V)	LIRDGGDDNN (B*0801)	Putative epitope	Negative	Negative	N/A	Negative
		LIRDGGDN**N**K	Common variant - escape?	Negative	Negative	N/A	Negative
		LIRDGGDDNNNTIEIFRF	18mer bound selected area	Negative	Negative	N/A	Negative
		LIRDGGDN**N**K**N**TIEIFRF	18mer common transmitted variant	Negative	Negative	N/A	Negative

Interestingly, 3 of the 4 validated T-cell escape forms included a G-to-A mutation embedded in APOBEC3G or 3F motif, a further indication that an APOBEC-mediated enhanced mutation rate may facilitate early CTL immune escape in some cases (Table 3.3). Clusters of mutations that co-occur in a patient may represent an alternate way in which immune escape can be inferred and is consistent with previously observed patterns of escape from an immune response early in infection (Jones et al., 2004). Again patient MEMI provides an example with at least three independent nonsynonymous substitutions at site 566, which was embedded in a proposed, but not validated, epitope for one of the HLA alleles of this patient (B*0801). Collectively the results suggest that there is strong evidence for selection preferentially occurring within CTL epitope regions relatively early in infection; all potential CTL escape variants were found among Fiebig stage V cases in this study (Table 3.3).

As described, many of the sites identified coincide with experimentally confirmed CTL responses, indicating the importance of CTL escape early in infection. However, the sites that were detected are likely to represent only a fraction of the mutations that have been driven to detectable frequency as a consequence of immune escape. For example, since the applied method relies on shared patterns of mutation across multiple patients, it is unlikely that the method has the power to detect escape from immune responses mediated by rare HLA alleles. Additionally, since parts of *env*, particularly in the variable loops, were masked from the analysis due to ambiguous sequence alignment, sites that confer resistance to neutralising antibody escape may have been overlooked. Although the results underscore the importance of CTL escape in early infection, they do not provide a basis to discount the potential contribution of escape from antibody responses to the evolution of *env* early in HIV-1 infection.

The majority of the rapidly evolving sites that recurred within individuals were clustered and appeared to be under selection for escape from cellular immune responses, and several of these mutations are G-to-A substitutions in APOBEC3 motifs. A highly significant enrichment overall for APOBEC3 motifs among the rapidly diversifying sites identified using HyPhy was observed, even when stringent precautions were taken to exclude the 15 subjects that had enrichment of G-to-A mutations within the APOBEC3 motif. This suggests the possibility that APOBEC3 and imperfect Vif-functioning can play an important biological role in the patterns of mutation found in early infection, and may contribute to determining the first immune escape mutations. Thus, base mutation rate may in some cases be making a contribution to determining which immune escape events happen first, along with fitness (Asquith and McLean, 2007) and possibly T-cell potency (Leslie et al., 2004). CTL immune responses have been shown to play a key role in controlling viral replication in early infection (Goulder et al., 1997; McMichael and Rowland-Jones, 2001; Schmitz et al., 1999; Turnbull et al., 2006) and, in animal models, positively selected CTL escape variants have been reported within weeks of infection (Fernandez et al., 2005; Loh et al., 2007, 2008). Escape from CTL responses has been observed in humans both early and later in infection (Allen et al., 2000; Goulder et al., 1997) and escape from successive immune responses is an important feature of HIV disease progression.

In the analysed dataset several of the rapidly evolving sites are associated with clusters of mutations that occur in close proximity to one another in the same patient (Figure 3.4). In some cases there are multiple escape mutations at the same site. Given the short times since infection in these patients and the otherwise highly homogeneous sequence alignments, the multiplicity of alternative CTL escape pathways that are already beginning to be explored by the virus in early infection is remarkable. Several studies have also shown that upon transmission some HIV-1 escape mutations can revert to the consensus in order to re-establish effective replication efficiency within the new host (Friedrich et al., 2004; Li et al., 2007; Allen et al., 2005; Herbeck et al., 2006). These reversion mutations frequently occur in structurally conserved regions and often reside within defined CD8 epitopes that are no longer subject to the restrictions of the original host's HLAs (Friedrich et al., 2004; Li et al., 2007; Allen et al., 2005; Peyerl et al., 2004). Envelope evolution towards the wild-type sequence state after transmission is thought to occur in the absence of specific CTL mediated immune pressure (Leslie et al., 2004; Allen et al., 2004). Even when the newly infected host has the HLA allele associated with the response, reversion can occur followed by development of a CTL response targeting the wild-type epitope. Given that essentially all examples of positive selection, which also exhibited recurrent and localised concentrated change in these 81 individuals, occurred within one of 4 experimentally validated CTL epitopes (or in one of 2 putative epitopes based on the individuals HLA), in this cross-sectional study the earliest changes seemed to result more commonly from escape rather than from reversion. However, only in cases where CTL escape was confirmed experimentally can the conclusion be made with certainty that the mutations are associated with escape.

3.3.4 Further Examination of Positively Selected Sites

The remaining sites that were identified cannot be explained by APOBEC mediation and do not occur within mutation clusters. These sites were therefore not investigated as possible CTL epitopes. Two sites (412 and 838 in the alignment) are identified largely because of mutations in patients Z31 and TT31P, where infection may have resulted from multiple closely related strains of the virus, rather than a single viral strain (Keele et al., 2008).

For all but two of the remaining sites, the observed mutation occurred on a single terminal branch of the within-patient phylogenetic tree and mutations were consequently observed only once per patient. These sites were identified as under positive selection because the site was mutated in multiple patients. Thus, unlike the substitution patterns in the putative CTL epitopes described above, there was no evidence at this stage of infection that these mutations were spreading in the intra-patient viral populations. As a result, these sites may be mutational hotspots, albeit hotspots that cannot be explained by known mechanisms of hypermutation in HIV-1.

Consistent Replacement of a Site by Another Specific Amino Acid

A visual inspection was carried out to determine whether there was evidence of directional evolution in any of the 30 rapidly evolving sites identified by the HyPhy selection analysis. Therefore, the goal was to ascertain whether a consistent trend towards replacement of one amino acid by another amino acid across multiple patients, could be observed. Position 646 (518 in HXB2), which occurs in the fusion peptide of gp41, showed consistent replacement of Methionine with either Valine or Isoleucine (the observed substitutions relative to the within patient consensus are summarised for each selected site in the Appendix, Table A1). This trend is remarkable when the data from all patients are considered. Fewer than half (38 of 81) of the patients had Methionine at this position in the inferred infecting viruses, yet all nine of the patients with a nonsynonymous mutation at this site had Methionine in the patient consensus sequence. The probability of this occurring by chance is 0.001. Such strong directional selection in early infection suggests the possibility of reversion of amino acids associated with a selective transmission bottleneck, but could also be explained by reversion of an escape mutation associated with an immune response in the transmitting virus. The apparent directionality could also be explained by a sequence-specific mutational hotspot if the ATG codon, encoding Methionine, results in a sequence motif that promotes mutation. This is suggested by the fact that the mutation from M to V or I is found in just a single sequence in each of the patients in which it occurs and thus does not appear to show evidence of spreading in the intra-patient viral population. However, none of the mutations

at this site are consistent with APOBEC3G/F mutation, and thus rapid evolution at this site cannot be explained by known mechanisms of hypermutation.

Therefore, discounting sites that may have been affected by APOBEC3G/F hypermutation, only a single site with significant evidence for directional change was identified. At site 646 (518 in HXB2), as described, there is only a single mutant sequence observable in each of the nine patients in which the mutation is found, and this seems inconsistent with a mutation that is coming under selection for increased frequency in multiple patients. Furthermore, no increased frequency of Methionine in early infection sequences compared with chronic sequences at this site was found (data not shown), as would be expected if the consistent mutations from Methionine to Isoleucine or Valine represented reversion associated with selective viral transmission. A single site was observed to undergo recurrent change in a single subject in Fiebig stage II, and as this is prior to the immune response it may be indicative of rapid reversion. Alternatively, the observation could be explained instead by transmission of two highly related forms, together with recombination either in the donor or soon after transmission.

Potential Enrichment of Mutations within Particular Sequence Patterns

Neutralising antibody responses drive the evolution of viral escape and are therefore thought to play an important role in shaping the evolutionary changes observed in the envelope gene. Mutations, insertions and deletions, as well as the distribution of glycosylation sites along the envelope gene mediate the ability of the virus to escape from neutralising antibodies, although the relative contribution of each of these is unknown (Frost et al., 2005). The role of N-linked glycosylation sites in the *env* gene in maintaining the survival of HIV-1 during the earliest stages of infection, is of particular interest as the carbohydrate surface establishes the first means of contact between the host and the virus. A previous study (Frost et al., 2005) compared the rate of phenotypic escape of HIV-1 in recently infected individuals from neutralising antibodies to various characteristics of the envelope gene, and found that escape from neutralising antibodies can occur regardless of the variation in glycosylation or rate of

insertions and deletions in *env*.

Therefore, since the number of potential N-linked glycosylation sites (PNGSs) in *env* may affect the fitness of the virus by rendering it more or less sensitive to neutralising antibodies, a test for enrichment of mutations within PNGSs was carried out. The observed number of mutations within PNGSs ($NX_1[S|T]X_2$ where X is any amino acid other than Proline) was not significantly different than the expectation under a model of random mutation. Therefore, no evidence of a tendency for mutations to disrupt PNGSs, and therefore no evidence of accelerated diversification, in early infection ($p = 0.736$; Fisher's exact test) was found. This suggests that if escape from neutralising antibodies is among the major driving forces of HIV-1 evolution at this stage, the mechanism is unlikely to be primarily through single base substitutions in glycosylation sites, at least in the portion of *env* that could be aligned with high confidence.

A similar approach was applied to determine whether there was an overall enrichment in mutations within epitope-dense regions. In this case, the comparison is between the number of mutations that occur within epitope rich regions (coordinates based on the best defined CTL/CD8+ epitope summary Frahm et al., 2008), and the total number of mutations found outside the epitope regions. There was no evidence of enrichment of mutations occurring within a CTL epitope-dense region as opposed to less dense regions ($p = 0.933$, Fisher's exact test). However, because all mutations were considered in this case, the lack of such evidence may result from the high level of noise in the data; that is, the fact that the majority of the observed mutations are likely to be neutral or under purifying selection, rather than positively selected. When only the 24 rapidly evolving sites are considered, a weak trend ($p = 0.27$) towards enrichment within epitope rich regions (odds ratio: 1.6) is found.

The Impact of Recombination

The phylogenetic codon model that was applied using HyPhy does not account for recombination, which is known to cause false positive inference of selection (Anisimova et al., 2003; Shriner et al., 2003). With the highly homogeneous intra-patient datasets analysed here,

recombination is not likely to have a large effect, because the sequences are, in most cases, well described by a star phylogeny with little internal structure in the tree. However, in the case of the later Fiebig stage patients there is some tree structure and recombination in these patients may have the potential to cause false inference of positive selection.

A test was therefore carried out on all the homogeneous intra-patient sequence data sets for evidence of recombination using GARD (Kosakovsky-Pond et al., 2006b,a). Evidence of a single recombination breakpoint in the sequence alignments from each of two patients (Z31 and MEMI) was found. In both cases the breakpoint is located near the end of the sequence alignment (just 3% and 10% of the alignment lies to the right of the breakpoint for MEMI and Z31 respectively), and therefore a single tree topology applies for most of the alignment. Only in the case of MEMI do nucleotide mutations occur to the right of the recombination breakpoint and at inferred adaptively evolving sites (979, 981 and 989). Since these mutations are also clustered and two of the three are confirmed CTL escape mutations (Table 3.3), it appears unlikely that these sites are false positives.

3.4 Conclusions

This cross-sectional study has provided a glimpse into the nature of the diversification of HIV-1 following transmission in a population of 81 HIV-1 subtype B infected individuals. As expected, purifying selection acts on single base mutations at the majority of sites, but there is also evidence of a subset of sites that diversify more rapidly than expected under a model of neutral evolution. The results presented here suggest that APOBEC3 may play a role in shaping the dynamics of early T-cell immune escape. APOBEC3 motifs are highly enriched among the rapidly diversifying sites. This could be the result of an anomalously high mutation rate giving the appearance of selection pressure when there is none, or a higher mutation rate promoting a particular route of immune escape; the two are not mutually exclusive. Finally early T-cell escape appears to be much more common than reversion. This study provides new evidence supporting the idea that concentrated regions of change

within an individual during early infection (Jones et al., 2004) are frequently indicative of early escape from T-cell responses.

Chapter 4

A Phylogeny Aware Method to Compare Sequences from Early and Chronic HIV-1 Infections

4.1 Introduction

As described in Chapter 3, understanding the selection pressures acting on HIV during the acute and early stages of infection is crucial towards the development of a successful vaccine. A potential vaccine must be effective against the virus that is actually transmitted and knowledge of constraints associated with transmission, if they exist, could help in the design of more targeted vaccines. Several previous studies have focused on transmission signatures associated with the envelope gene, due to its importance for infection and for antibody neutralisation. Neutralising antibodies can appear within two months after the detection of HIV-specific antibodies in acutely infected individuals and are so restrictive that the original viruses are entirely superseded by escape mutant populations (Wei et al., 2003).

In HIV-1 subtypes A and C the variable loops of the *env* gene have been reported to be

shorter and to have fewer potential N-linked glycosylation sites (PNGSs) in sequences taken during acute and early infection compared to sequences isolated later in infection (Chohan et al., 2005a; Derdeyn et al., 2004; Sagar et al., 2006; Repits et al., 2008). The number of N-linked glycans presents a crucial balance between efficient protein folding and function of Env gp120, and the escape from recognition by neutralising antibodies (Li et al., 2009). The association between resistant viruses and N-linked glycosylation sites has led to the premise of an evolving glycan shield, where the carbohydrate coat shields the underlying viral binding regions from antibody detection (Wei et al., 2003; Bunnik et al., 2008). During the investigation carried out on selected sites in acute and early infection patients, an increased mutation rate within PNGSs for subtype B viruses was not found (Chapter 3, section 3.3.4). However, shorter variable loops and fewer PNGSs are thought to be the result of outgrowth of fitter, but neutralisation sensitive, viruses in newly infected hosts (Chohan et al., 2005a; Derdeyn et al., 2004; Sagar et al., 2006; Repits et al., 2008) in subtype A and C viruses. In addition, the viral population in long-term non-progressors, who are said to be able to maintain an effective neutralising antibody response (Zhang et al., 1997; Wang et al., 2008), has been shown to have increased V2 lengths compared to rapid progressors (Masciotra et al., 2002).

In a recent longitudinal study, reversion was shown to occur at amino acid sites driven to evolve early resistance due to neutralising antibody pressure, reflecting the fitness cost sacrificed to ensure initial survival (Bunnik et al., 2008). The characteristics associated with Env loop length and number of PNGSs of the transmitted virus have been inferred from comparisons of linked donor and recipient sequences in small studies of discordant couples, where the one partner was HIV-uninfected and the other HIV-positive (Derdeyn et al., 2004), as well as from comparisons of unlinked sequences from individuals in early and chronic infection (Chohan et al., 2005a; Repits et al., 2008).

Although epidemiologically linked sequences from studies of discordant couples provide an ideal source of data to investigate the selective transmission of viruses with specific characteristics, there have been relatively few studies that have acquired a large enough number of individuals to afford acceptable levels of statistical power to identify signatures of viral

transmission. Sample sizes in studies of unlinked individuals are much larger, and therefore the idea of selective processes influencing the spread of the carbohydrate moieties and thereby affecting the sensitivity of the virus to neutralisation, can be revisited. However, published comparisons of the characteristics of viruses from different stages of HIV-1 infection are potentially subject to biases due to the phylogenetic relatedness of the sequences. In particular, the presence of sub-clades of sequences (that is, groups of sequences that are clustered on the phylogenetic tree) from the same stage of infection causes violation of the key assumption of sample independence, which underlies the standard statistical tests that have been applied to these data.

In this chapter, to continue the investigation on the transmission characteristics and unique features of the early infection envelope sequences, the focus was to carry out a comparison between viruses isolated from early and chronic infections. Therefore, instead of finding sequence properties that are shared across a large group of very early infection datasets as described in Chapter 3, a comparative approach was followed using early and chronic infection datasets.

To this aim, a simulation study was carried out to determine whether false positive results are likely to occur if the phylogenetic structure of an unlinked sequence dataset is not taken into account. The focus of these analyses was on a published comparison (Chohan et al., 2005a) of *env* variable loop lengths where the authors used the associated phylogenetic conformation as support for their findings. Furthermore, a novel method that can be used to compare variable loop lengths or other characteristics of viruses using epidemiologically unlinked sequences, is proposed. This method was applied to evaluate the evidence of a reduced length of the *env* V1-V2 region in sequences from a previous study (Chohan et al., 2005a).

4.2 Methods

4.2.1 Simulations

In order to evaluate if the degree of shared phylogenetic history of the early and chronic sequences from the study by Chohan et al. (2005a) was large enough to create a significant deviation away from actual difference between the variable loop lengths of the two groups, a sequence simulation study was carried out. MySSP (Rosenberg, 2005) was used to simulate 1000 sequences evolving with insertions and deletions along the tree inferred from the *env* V1-V2 sequences used in the above study, which included 35 sequences sampled from early infection and 51 sequences isolated from chronic infection Chohan et al. (2005a). The insertion and deletion rate (on average one insertion and one deletion every 70 substitutions) and size parameters (Poisson distributed indel length with mean = 24bp), that were used in these simulations, were chosen such that the simulated sequences qualitatively resembled those from the original study.

This simulation study allows us to calculate the number of times a significant difference between loop length would be observed by random chance. If the difference is significant in more than, for instance, 5% of the cases, the test may not be sensitive to bias in the phylogeny and the results should be interpreted with caution. The statistical programming environment, R (http://www.r-project.org/), was used to conduct the statistical analysis (one-sided Wilcoxon rank-sum test) of variable loop lengths carried out on all the simulated early and chronic sequences.

4.2.2 Phylogeny Aware Method to Compare Variable Loop Lengths in Early and Chronic Sequences

If the shared history of a significantly sized phylogenetic cluster of early or chronic sequences is not accounted for during comparative tests, it is reasonable to expect some degree of

statistical bias towards that characteristic, for example Env variable loop length, which the cluster favours.

Here a method is proposed that estimates the variable loop lengths at the ancestral nodes of the phylogenetic tree by averaging over the lengths of the V1-V2 loops of the descendant sequences, weighted by the reciprocals of the respective branch lengths (Felsenstein, 1981b). The difference between the loop length of a virus and its immediate ancestor on the phylogenetic tree is then used as the variable to be compared between the early and chronic groups (Figure 4.1).

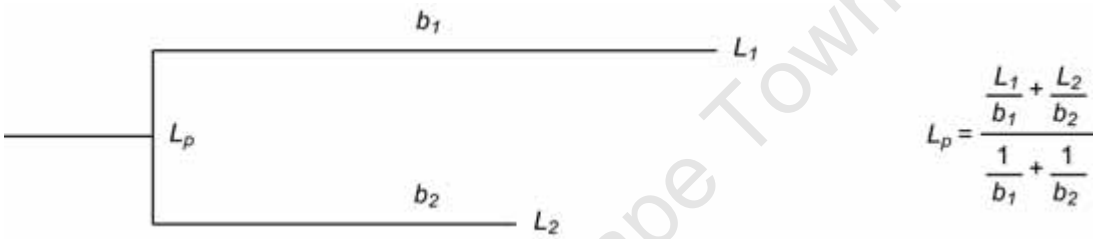


Figure 4.1: Estimation of the parent V1-V2 length (L_p) from the descendant V1-V2 length (L_1 and L_2) and branch length (b_1 and b_2) data.

4.2.3 Recombination Testing

Recombination is known to cause false positive results in studies that rely on a phylogenetic tree that reflects the true relationship between sequences (Anisimova et al., 2003; Shriner et al., 2003; Scheffler et al., 2006). A recombination event is likely to bias the results since it implies that a single tree can not reliably represent the sequence alignment, and any method that does not account for recombination can produce potentially spurious results.

A genetic algorithm for recombination detection, GARD (Kosakovsky-Pond et al., 2006b), was used to eliminate this as source of bias. This method was applied to the aligned V1-V2 region from the original study (Chohan et al., 2005a). GARD implements a maximum likelihood model comparison approach to determine whether there is evidence of recombination in a dataset of aligned sequences.

4.2.4 Correlation Tests to Identify Potential Relationships between the Variable Loop Lengths and Further Specific Disease Stage Determinants

Once an investigation on the potential sources of bias was carried out, and a method that accounts for phylogenetic relatedness was formulated, further correlation studies were explored. A possible extension to the notion that shorter variable loop lengths are associated with the transmitted virus, is to compare the V1-V2 lengths to other disease stage determinants. If the V1-V2 variable loop length in early and chronic infections varies significantly, then it is possible to determine whether the association holds during subsequent related comparisons. CD4 counts are related to disease stage, and since substantial patient CD4 count information together with the associated viral sequence data, and therefore V1-V2 loop lengths, are available, tests can be carried out to determine whether Env loop length is correlated with the particular defined stage of disease.

HIV-1 clinical and sequence data were downloaded from the Los Alamos HIV Sequence Database (www.hiv.lanl.gov). CD4 count data and the associated Env sequence data of 268 patients were retrieved; 107 sequences were HIV-1 subtype B, 97 subtype C, and 41 were the recombinant subtype AE. The statistical analyses, correlations tests (Spearman's rank correlation test) and conditional inferences (Hothorn et al., 2006) as well as tests for significant differences (Wilcoxon rank-sum tests), were carried out in R (http://www.r-project.org/).

4.2.5 Identification of Sites with Different Amino Acid Profiles in Early and Chronic Infection [1]

In a separate analysis, an additional phylogenetic method to identify sites that differ between two specific groups, for example an early and chronic dataset, was developed. The method

[1] This section was carried out with the assistance of Graham Poulter; the idea for using, and coding of, the multivariate hypergeometric distribution is attributed to him.

takes the phylogenetic history, and hence the non-independence between sites, into account. For each site, a "loss", "gain" or "no change" of a site was determined, depicted with -1, +1, and 0 respectively.

GASP (Gapped Ancestral Sequence Prediction; Edwards and Shields, 2004) was used to generate the ancestral sequences of a given input dataset containing sequences from two groups. An example tree is shown in Figure 4.2, indicating three points on the phylogeny (A, B, and C) to illustrate how the gain, loss or no change of a site was defined. A, B, and C each refer to an amino acid at a given site in the alignment. The site at A is the most recent common ancestor of the entire dataset, C is the terminal amino acid, and B is the immediate ancestor of C.

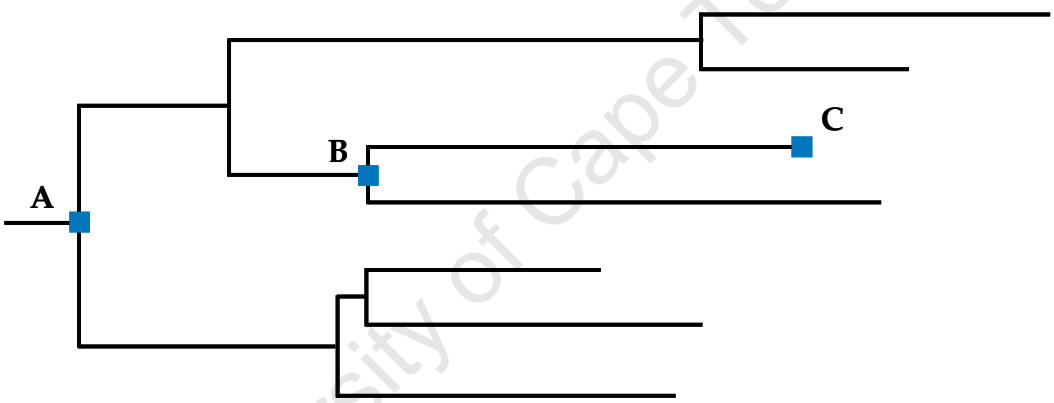


Figure 4.2: Tree topology reflecting the relationship between sites A, B, and C.

The outcome R depends on the relationship between the amino acids at sites A, B, and C, defined by:

$$R = \begin{cases} +1 & \text{if A} = \text{C and A} \neq \text{B}, \\ -1 & \text{if A} \neq \text{C and A} = \text{B}, \\ 0 & \text{Otherwise.} \end{cases}$$

In this way, a distribution of +1 (gain), -1 (loss), and 0 (no change) is determined for each site in the alignment. The distribution can be likened to an urn with three different colour

balls in it, x_1, x_2, and x_3. If n number of balls are taken from the urn without replacement, then

$$X\,(f\,,\,n) = (x_1,\,x_2,\,x_3)$$

follows a multivariate hypergeometric distribution. A population comprised of three types of outcomes, as in this case, and where there are N_1, N_2, and N_3, of each type. If n samples are taken from the population at random, then the probability of selecting f_1 objects from N_1, f_2 objects from N_2, and f_3 objects from N_3, is:

$$P(X_1 = f_1,\, X_2 = f_2\,, X_3 = f_3)\, \frac{\binom{N_1}{f_1}\binom{N_2}{f_2}\binom{N_3}{f_3}}{\binom{N_1+N_2+N_3}{n}}$$

where $f_1 + f_2 + f_3 = n\,;\;\; 0 \leq f_1 \leq N_1;\;\; 0 \leq f_2 \leq N_2$; and $\;\; 0 \leq f_3 \leq N_3\;$.

The values assigned to gains and losses for each site, represent the weights, w_1, w_2, and w_3, for each category. And the weighted sum of the n random draws is given by,

$$Y\; = -1\,X_1\; +\; 0\,X_2\; +1\,X_3\,,$$

which represents the net gain or loss at a site. The distribution of Y can then be determined by enumerating X for each of the possible values for f_1, f_2, and f_3; and for each outcome it is possible to calculate the the weighted sum, t. The values of f_1 , f_2 , and f_3 are results of random draws from X, and thus the sum $t = -f_1 + f_3$ has the distribution of $Y = -X_1 + X_3$.

Under the null hypothesis, the sample n is a random selection from this weighted sum of the multivariate hypergeometric distribution. If t is the weighted sum of the observed values at a given early infection site, then for that value of t, a test can be carried out to determine

what the probability is that the observed sum is a random sample. If the probability of drawing a result at least as extreme as the sum which was observed is low, then it is unlikely that t is a random sample. The p-value is determined by summing over all the discrete probabilities in the calculated null distribution of Y.

This method was applied to the same dataset which was analysed for determining V1-V2 loop length differences. The method finds sites in the acute group, which have undergone gains and/or losses that are significantly different to that of the chronic group. In effect the method is comparing the distribution of the number of changes on the terminal branches, conditioned on the predicted common ancestor, in one group to that of another. In the present study the groups are the early and chronic sequence datasets.

4.3 Results and Discussion

4.3.1 Simulation Analysis

The potential of sub-clades of sequences from the same stage of infection, in the data analysed by Chohan et al. (2005a), to cause false positive inferences of differences between the variable loop lengths of the early and chronic sequences was assessed using the phylogenetic tree inferred from the sequences from the original study (Figure 4.3[2]).

[2]Duplicate sequences were included in the original paper and these samples, referred to as AF069670-AF069670_A and M62420-M62430 in this study, were retained in the current analysis for comparative purposes.

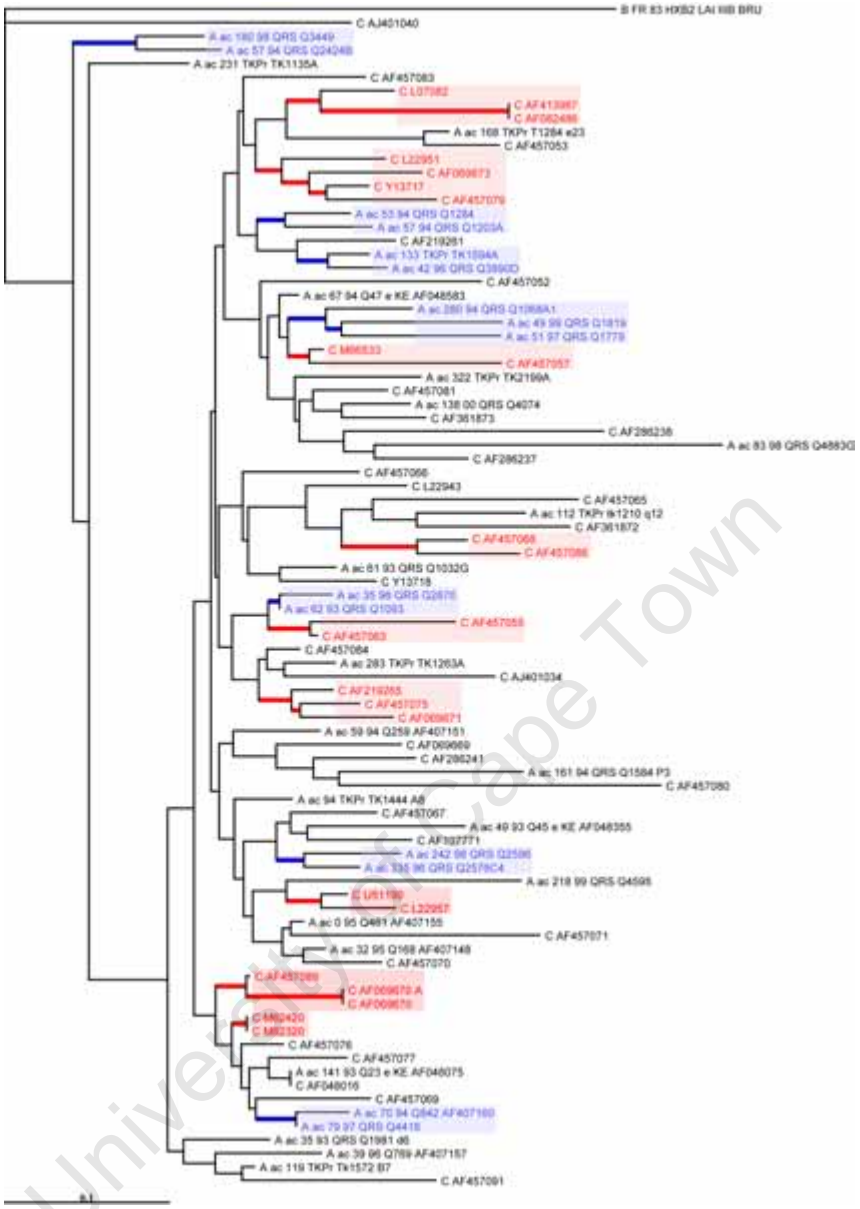


Figure 4.3: Neighbour-joining tree inferred from amino acid sequences of the V1-V2 region. Branches shown leading to sub-clades consisting only of early sequences, beginning with "A" and are coloured in blue, or only of chronic sequences, beginning with "C" and coloured in red, respectively.

If the phylogenetic relationships are neglected and sequences treated as though they were sampled independently, any insertions or deletions that occurred along the blue branches (leading to sub-clades of early sequences) or red branches (leading to sub-clades of sequences

from chronic infection) are effectively counted multiple times. To test whether the observed clustering of early and chronic sequences was sufficient to cause a significant bias in the comparison of variable loop lengths between the early and chronic groups, 1000 sequences were simulated evolving with insertions and deletions along the tree using MySSP (Rosenberg, 2005).

Following the analysis of the original study (Chohan et al., 2005a), the length of the V1-V2 loops of the simulated early and chronic sequences were compared using a one-sided Wilcoxon rank-sum test. The null hypothesis for the one-sided test (variable loops not shorter in early sequences than in chronic sequences) was rejected 109 times (10.9%) at the 5% significance level in 1000 simulated datasets. This represents a significant excess over the expected 5% false positive rate. This analysis demonstrates that failure to account for phylogenetic relationships between sequences can cause a bias in comparisons of the variable loop lengths of the early and chronic sequences.

The tree in Figure 4.3 is just one estimate of the possible phylogenetic relationships between the sequences and, in most cases, the sub-clades present on the tree are not well supported as the sequences used are short and highly variable (mean length = 69.7 amino acids). However, our argument does not depend on the robustness of the inferred sub-clades as there is no reason to believe that the degree of clustering of early sequences with early sequences or chronic sequences with chronic sequences is any less in the true tree describing the relationships between the sequences. Instead, it is the assumption that each sequence can be treated as an independent sample that requires justification, particularly if it is not valid using the tree inferred from the sequences. It may be interesting to further investigate the affect of the underlying tree by repeating the analysis using a range of phylogenetic trees, each with different measures of support. In this way it may be possible to illustrate the extent of bias as result of various inferred phylogenetic relationships.

4.3.2 Comparison of Env Variable Loop Length Differences Observed in Early and Chronic Infection

While the optimal data for investigating genetic characteristics associated with transmission are sequences derived from transmission pairs, these are limited due to small sample sizes. To address the problem of lack of independence of sequences due to phylogenetic relationships (Felsenstein, 1981b), a novel tree-based method was developed that can be applied to compare the lengths of the variable loops between the early and chronic sequences (see section 4.2.2 for details). The method described here accounts for the non-independence of phylogenetically related samples and can therefore be used in cross-sectional comparisons of viruses from different infection stages. When this method was applied to the simulated datasets (using phylogenetic trees inferred separately from each simulated dataset) 36 false positive results (3.6%) at the 5% significance level were obtained, suggesting that the method can prevent false positive inferences resulting from the sub-clades of early and chronic sequences in the data. The fact that the method performs well with the trees inferred from the sequences (rather than the trees that were used to produce the sequences in the simulations) suggests that the method is robust to uncertainty in the inferred trees.

This phylogenetic method was applied to re-evaluate the evidence for a reduced length of the V1-V2 loop region in early infection sequences from HIV-1 subtype A (Chohan et al., 2005a). The original publication reports a p-value of 0.008 (one-sided Wilcoxon rank-sum test) for the comparison of the V1-V2 length between early [n=35] and database [n=51] sequences. The results from our method applied to the same dataset confirm the original finding though with a somewhat more marginal estimate of statistical significance ($p = 0.02$). This suggests that the length difference is unlikely to be the result of insertions and deletions shared by sub-clades of early and chronic sequences.

The influence of an underlying population structure on the identification of associated evolutionary patterns, is widely accepted (Carlson et al., 2007; Kang et al., 2008; Pagel, 1994). Recently, Carlson et al. (2007) evaluated two methods designed to find correlations between discrete data while accounting for potential shared phylogenetic histories. The first is aimed

at identifying discrete traits that have co-evolved and assumes that a change in either trait can induce a change in the other. The second method is an extension of the approach described by Bhattacharya et al. (2007), where a single discrete variable evolves along the branches of the tree and the associated trait affects only the terminal branches of the tree (Carlson et al., 2007). This method was applied to investigate the effect of a phylogeny on identifying amino acids associated with specific HLAs (human leukocyte antigens), where the amino acids evolve across the phylogenetic tree and the HLAs only have an influence at the tips of the tree (Bhattacharya et al., 2007). Their observed results were inconsistent with previously described findings, and these differences were accounted to non-independence between genetic lineages.

Population based studies aimed at identifying correlated characteristics could be particularly useful for finding vaccine targets that would be broadly effective. Therefore, considering the potential impact of such a vaccine, false positives or false negatives should be minimised as far as possible. The phylogeny aware investigation presented here, together with our simulation study and previously reported findings (Bhattacharya et al., 2007; Carlson et al., 2007; Kang et al., 2008), highlights the importance of accounting for shared phylogenetic patterns when the tree structure forms a fundamental part of the analysis.

4.3.3 Evidence of Recombination

HIV-1 is known to recombine (Jetzt et al., 2000) and it is possible that the phylogenetic method proposed here, which assumes that a single phylogenetic tree represents the relationships between the samples along the entire sequence, is compromised as it does not control for recombination. To address this issue, GARD (Kosakovsky-Pond et al., 2006b) was used to investigate the evidence for recombination in the V1-V2 region from the subtype A dataset (Chohan et al., 2005a). Although GARD found no evidence of recombination in the V1-V2 region, it is possible that there are undetectable recombination events in the data or other errors in the phylogenetic tree and this can be considered as a caveat of the phylogenetic method.

4.3.4 Correlation between V1-V2 Length and CD4 Count

It is believed that V1-V2 loop lengths differ significantly between early and chronic disease stages, and since CD4 counts are commonly used for classification of disease progression, the relationship between V1-V2 loop lengths and CD4 counts could present interesting results. As CD4 cell counts generally decline over the course of HIV infection (Arnaout et al., 1999; Levy, 2007) it is possible to use a larger dataset as an additional, indirect, means to test whether there is a relationship between V1-V2 length and disease stage. If there is a relationship between disease stage and V1-V2 length, then it would be predicted that V1-V2 length would correlate with CD4 count. In this case the weaker claim that V1-V2 length is related to disease stage is being tested, and not whether this relationship is necessarily the result of a difference between the transmitted virus and the virus from the chronic infection stage. However, if the length of the V1-V2 loops is reduced in the transmitted virus compared to the chronic stage, then, assuming a steady state (that is, that there is not an ongoing reduction in HIV-1 V1-V2 length over time) the variable loops would be expected to tend to increase in length, on average, over the course of infection.

A highly significant negative correlation between V1-V2 length and CD4 count (ρ = -0.207; p = 0.0007; n = 268; Figure 4.4) was found, consistent with lengthening of the V1-V2 region over the course of infection. This correlation remained highly significant (p = 0.0004) when it was conditioned on viral subtype (Hothorn et al., 2006) and also when this phylogenetic method was applied to correct for phylogenetic relationships between sequences (p = 0.005).

These data were derived primarily from subtypes B, C and the circulating recombinant form (CRF) AE. Data from subtypes B and C alone showed marginally significant correlation between CD4 count and V1-V2 length (ρ = -0.18; p = 0.06; n = 107 and ρ = -0.19; p = 0.07; n = 91, for subtypes B and C, respectively), however the correlation for AE alone was highly significant (ρ = -0.46; p = 0.004; n = 41). Combining the data from subtype B and C and conditioning on subtype, the relationship between CD4 count and V1-V2 length was significant (ρ = -0.17; p = 0.01; n = 198).

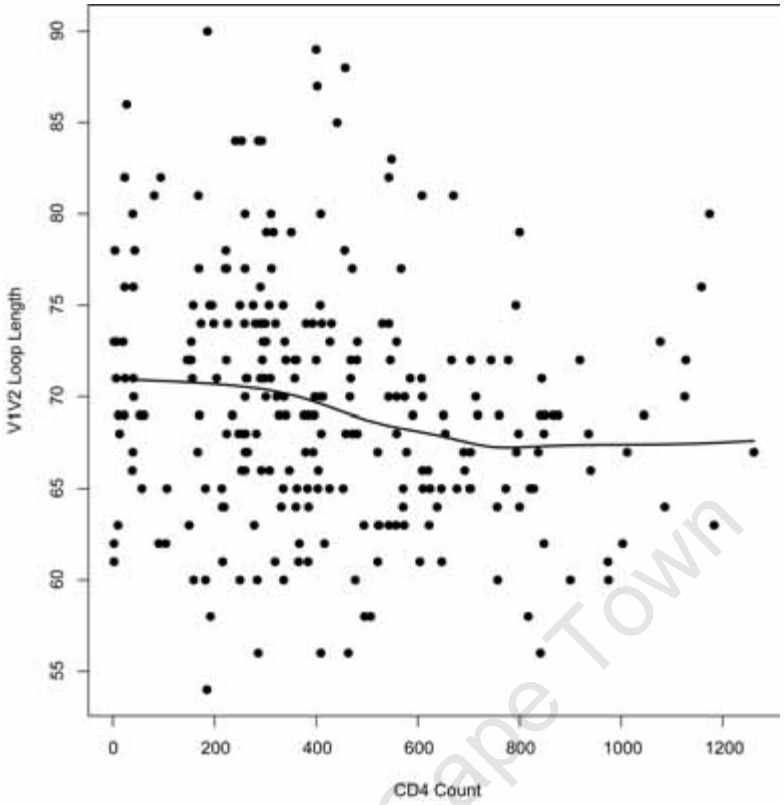


Figure 4.4: Correlation between CD4 count and V1-V2 length. The line represents a locally-weighted polynomial regression (lowess) curve.

Therefore, although the strength of the relationship between CD4 count and V1-V2 length appears to vary among subtypes, these results are suggestive of a link between V1-V2 length and the stage of disease, and consistent with a shorter V1-V2 in the transmitted virus.

To investigate the relationship between CD4 count and V1-V2 lengths further, the data were divided into three CD4 categories based on the CDC classification system for HIV infection (Castro et al., 1993) and the median lengths for each group were compared. There was no significant difference in V1-V2 lengths between sequences from individuals with CD4 counts less than 200 cells/μl (AIDS defining) and individuals with CD4 counts between 200 and 499 cells/μl. However the V1-V2 loop length was significantly lower in sequences from individuals with CD4 counts greater than 499 cells/μl compared to either of the other two categories ($p = 0.008$ and $p = 0.0008$ for the comparison with the <200 cells/μl and the 200-499 cells/μl groups respectively, using a two-sided Wilcoxon rank-sum test; Figure

4.5). The difference remained significant after the phylogenetic method was applied to correct for evolutionary relationships between the sequences (p = 0.01 and p = 0.004 for the comparison with the <200 cells/μl and the 200-499 cells/μl groups respectively, using a two-sided Wilcoxon rank-sum test).

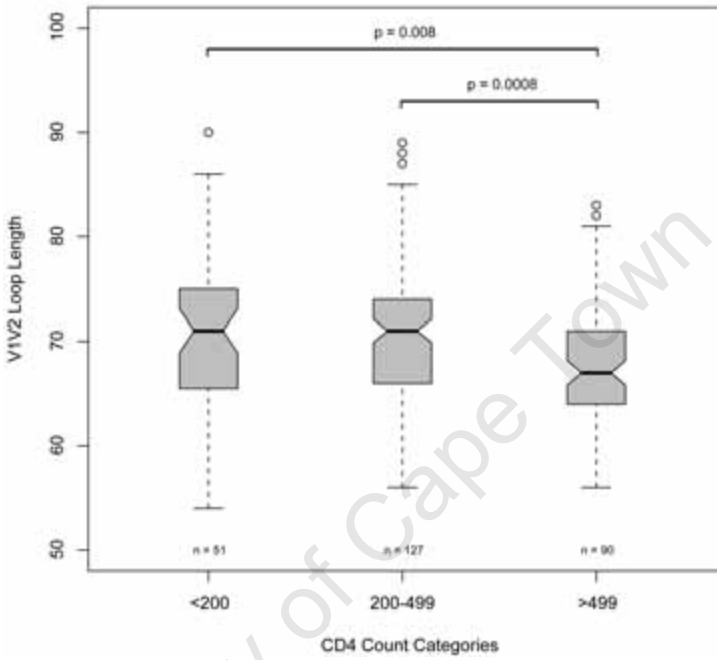


Figure 4.5: Notched boxplot of V1-V2 lengths in different CD4 count categories.

The results illustrate that variable loop lengths are correlated with the immunological state and that, on average, individuals whose CD4 counts have not been depleted, many of whom are likely to have been sampled early in infection, have shorter V1-V2 variable loops.

Together with the findings from the phylogenetic method (see section 4.3.2), the observation that CD4 count is significantly negatively correlated with V1-V2 length in database samples lends further support to a relationship between V1-V2 length and disease stage. However, the strength of this relationship appears to depend on viral subtype, with a far weaker and marginally significant effect observed in subtypes B and C and a strong effect observed for recombinant AE viruses. Nonetheless, the findings provide evidence for a gradual continuous change in V1-V2 over the course of infection.

The phylogenetic method designed to identify sites that differ significantly between two groups was used to extend the analysis of the V1-V2 loop length variation results discussed in section 4.3.2. The method predicts the ancestral amino acids from a phylogenetic tree, and determines whether the mutational patterns observed in two groups of sequences, at particular sites along the alignment, are significantly different. This method has been used in another study where the authors aimed to locate amino acids in HIV-1 subtype C sequences that differ significantly between samples obtained from early and chronic infection (Treurnicht et al., 2009). The authors found 18 amino acid sites along the entire genome that differed significantly ($p < 0.05$, uncorrected for multiple testing), and 17 of these (9 in *env*), were located within known CTL epitopes. These results illustrate the potential value of the method for identifying sites that are selectively maintained in a defined subset of sequences.

When the dataset from Chohan et al. (2005a) was analysed with this method only one site, 88 in the alignment, was identified ($p = 0.049$). Interestingly, it forms part of a PNGS, which is implicated in early viral escape from the immune system. The result should however be interpreted with consideration of the short sequence lengths, poorly aligned regions including large gaps, and lack of a correction for multiple testing. Multiple testing corrections account for the occurrence of false positives when multiple statistical tests are carried out. The occurrence and frequency of false positives can be controlled using adjusted p-values (Westfall and Young, 1993). This was not applied here, as the analysis was exploratory in nature. Because the phylogenetic method identified a site considered to play a role in early immune escape, and given the observations from Treurnicht et al. (2009), the results may nevertheless point to broader application of the method for finding different mutational patterns in subsets of aligned sequences.

4.4 Conclusions

The results of the simulation study presented in this chapter indicate that inferring a difference in the variable loop lengths between viruses from different stages of infection, without taking phylogenetic relationships into account, is subject to bias. While the optimal data for investigating genetic characteristics associated with transmission are sequences derived from transmission pairs, these are limited due to small sample sizes. The method described here accounts for the non-independence of phylogenetically related samples and can therefore be used in cross-sectional comparisons of viruses from different infection stages.

The findings presented in this chapter are consistent with the premise that an increase in PNGSs and longer V1-V2 loop lengths play an important role during viral evasion of host immune pressures (Chohan et al., 2005a; Derdeyn et al., 2004; Sagar et al., 2006). The re-evaluation of an HIV-1 subtype A dataset (Chohan et al., 2005a) with the described phylogeny aware method supports the conclusions of the original paper. The further observation that CD4 count is significantly negatively correlated with V1-V2 length in database samples, provides additional support for a relationship between V1-V2 length and disease stage. Collectively, these results indicate that the changes in variable loop lengths during HIV infection are associated with the stage of disease and are therefore likely to be relevant to our understanding of the immune pressures that drive viral adaptation in the host.

Chapter 5

Re-evaluation of the Evidence for Positive Selection in HIV Coding Sequences Using a Robust Method

5.1 Introduction

The simulation study carried out in Chapter 4 illustrated that the phylogenetic relatedness of samples can significantly bias the findings obtained from methods dependent on the underlying tree structure. Any features that violate the fundamental assumptions of the methods used to identify particular characteristics, have to be accounted for prior to the research. Alternatively, the methodology can be adapted or extended to accommodate the source of bias. Following on from the previous chapter, where shared genetic history was addressed, here the goal was to establish the extent of bias present in positive selection studies where recombination is likely to have taken place.

Adaptive molecular evolution is likely to shape the genome of HIV-1 on an ongoing basis and describing positive selection in, for example, early and late infection, drug-naive and drug-treated individuals, or in geographically distinct viral sequence datasets, would provide

a useful characterisation of the evolutionary changes in the HIV genome when faced with various challenges. Identifying the sites in a viral genome that are selected during viral escape from immune responses can help guide the selection of immunogens for vaccines, and understanding the evolution of resistance to antiviral drugs can help in the design of effective treatment programmes. Mutations resulting in sequence changes that lead to fitter individuals are more likely to become fixed in a population than mutations that provide no additional benefit to the organism (Yang, 2006). Identifying these sites in HIV and understanding their value to viral reproductive fitness is an important goal in the broader search for effective therapeutics to treat infected individuals.

Methods to infer selection from protein-coding sequences have been applied in a wide variety of species and genes, however these methods typically assume a single phylogenetic tree describing the relationships between the taxa. This assumption does not hold when linkage between sites is disrupted through recombination, as this results in several distinct phylogenetic tree topologies describing the relationships between the taxa in different parts of the sequence (Shriner et al., 2003). The high rates of recombination found in many pathogens, especially viruses, have been shown to cause substantial false positive inference of positive selection (Anisimova et al., 2003). The task of finding sites under selection, taking account of the extent to which recombination occurs throughout the HIV genome, has not been addressed thoroughly.

HIV has two copies of genomic RNA, and the virus readily recombines (Burke, 1997). During reverse transcription of viral RNA to single stranded DNA, the RT (reverse transcriptase) can switch between the two RNA strands, often resulting in mosaic HIV genomes (Ramirez et al., 2008). Numerous circulating recombinant forms (CRFs) have been described, which originate as result of dual infection with two divergent RNA sequences (Chan, 2004; Requejo, 2006). Intersubtype as well as intra-subtype recombinants are widespread, and recombination in HIV is one of the central ways the virus ensures genetic variability (Requejo, 2006; Rousseau et al., 2007).

The effect of recombination on the detection of positive selection has been evaluated ex-

tensively in recent years. An evolutionary model that does not account for recombination, treats homoplasies resulting from recombining sequences as mutational events, and the resulting tree topologies are inaccurate. This leads to a false representation of highly variable sites containing multiple non-synonymous changes, and these sites may therefore incorrectly appear to be evolving under positive selection (Shriner et al., 2003). In some cases the false positive rate has been shown to be as high as 90% (Anisimova et al., 2003). High rates of false positive inference of selection raise serious questions as to the validity of published examples of positive selection in recombining sequences (Shriner et al., 2003; Anisimova et al., 2003).

Scheffler et al. (2006) developed a novel and robust maximum likelihood method for inferring positive selection from recombining coding sequences. The method, PARRIS (PARtitioning for the Robust Inference of Selection), detects recombination breakpoints and then determines tree topologies and branch lengths for each partition resulting from the detected breakpoints, independently. The synonymous substitution rate at each site is also modelled in such a way to allow for variation across all sites, which brought about additional improvements to the model for detecting selection. PARRIS analysis, together with other statistical models for detecting signatures of selection, can be carried out online at http://www.datamonkey.org/ .

For the present analysis, PARRIS was used to re-analyse published HIV datasets in which positive selection was previously reported. Many of these results were published before the realisation that recombination causes false inference of positive selection using phylogenetic methods. The development of a robust method to infer selection creates an opportunity to review the findings of these published results and to provide a more accurate account of the contribution of positive selection to the evolution of HIV-1 coding sequences.

5.2 Methods

5.2.1 Novel Approach for Detecting Selection

Previous results of positive selection were generated by standard methods, which commonly assume that the topology, relative branch length, and total tree length over all sites remains constant. In the present study a method proposed by Scheffler et al. (2006), was applied to collected published datasets.

In the proposed method, PARRIS, recombination is taken into account and individual sequence partitions are defined based on where recombination breakpoints are found to exist. The topology and relative branch lengths are assumed constant over all sites within each partition, but may vary between partitions (Figure 5.1). However, the total tree length is allowed to vary from site to site regardless of partitioning. This method was implemented using the batch language of the HyPhy package (Kosakovsky-Pond and Muse, 2005).

5.2.2 Detecting Recombination Breakpoints and Partitioning

For the present analysis, breakpoint positions were estimated using a genetic algorithm implemented in the localGARD.bf module within the HyPhy distribution (see Kosakovsky-Pond et al., 2006b, for a description of the algorithm). The HKY85 nucleotide substitution model (Hasegawa et al., 1985), with no rate variation between sites, was used for breakpoint detection.

The detected recombination breakpoints were used to define the boundaries of the partitions. Since the phylogenetic information (topologies and branch lengths) is dependent on the length of the sequences used to produce the phylogeny, a limit was set on the number of partitions used for selection analysis. When more than the maximum number of partitions ($N = 20$ in the present study) were determined, only the N largest segments were used in the subsequent analysis. It is likely that the results following inclusion of short segment

parameter estimates can be highly unreliable, and the loss of information resulting from discarding these short segments is negligible compared to the bias they could cause (Scheffler et al., 2006).

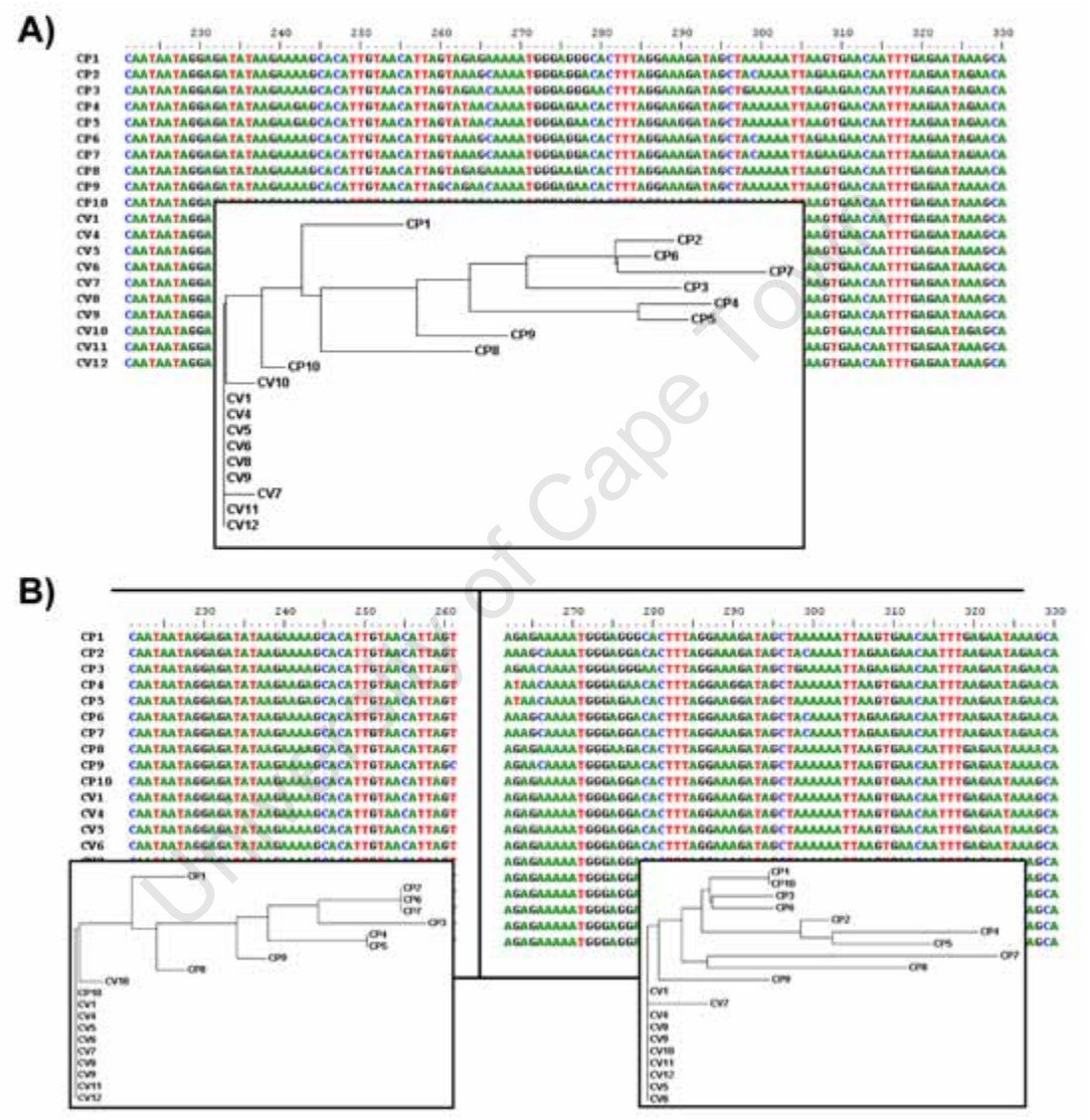


Figure 5.1: A) Standard methods assume the entire sequence is described by a single tree topology and set of branch lengths; B) PARRIS incorporates partitioning where each of the N largest segments, that contain no recombination breakpoints, are modelled using a separate topology and set of branch lengths.

5.2.3 Topology and Branch Length Estimation

The topology and branch lengths for each aligned partition were computed independently, whereas the remaining model parameters are shared across all segments. Positive selection was detected by comparing the discrete nearly neutral and selection models: M1a and M2a of Wong et al. (2004). Maximum likelihood topologies were estimated with the PAUP* program (Swafford, 2002) under the HKY85 model (Hasegawa et al., 1985). Branch length estimations were calculated with the M0 (single rate) model (Yang et al., 2000). A sequence is reported to be under positive selection at the 5% or 1% level if model M2a provides a significantly better fit than model M1a as measured by a likelihood ratio test with the appropriate significance level (Scheffler et al., 2006).

In order for the method to allow for synonymous rate variation, another parameter is added to the algorithm. The synonymous rate s, is treated as belonging to one of a number of discrete rate classes, similar to the treatment of the non-synonymous/synonymous rate ratio ω, and the instantaneous substitution rate from codon i to codon j at site h can be expressed by:

$$q_{ij}^{(h)} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ by more than one change,} \\ \pi_j s^{(h)} & \text{for a synonymous transversion,} \\ \kappa\pi_j s^{(h)} & \text{for a synonymous transition,} \\ \omega^{(h)}\pi_j s^{(h)} & \text{for a non-synonymous transversion,} \\ \omega^{(h)}\kappa\pi_j s^{(h)} & \text{for a non-synonymous transition.} \end{cases}$$

as described by Scheffler et al. (2006). Kappa, κ, is the transition/transversion rate ratio, and π_j is the codon equilibrium frequency of codon j. The non-synonymous/synonymous rate ratio at site h is denoted by $\omega^{(h)}$, and the synonymous rate at h by $s^{(h)}$. Similar to the ω distribution used for the discrete M3 model of Yang et al. (2003), the synonymous rate is drawn from a discrete distribution with three rate categories. However, unlike the ω distribution, the synonymous rates are scaled in such a way that the average synonymous

rate over all sites is equal to 1. Therefore, each site belongs to a synonymous rate category as well as belonging to one of the ω categories (Scheffler et al., 2006). In this method, separate parametric models are applied to the distribution of synonymous rates and of selective strengths, whereas Kosakovsky-Pond and Muse (2005) have applied the same parametric models to the distributions of synonymous and non-synonymous rates (Scheffler et al., 2006).

5.2.4 Datasets Re-analysed

The analysis included in this chapter focuses on the re-evaluation of a set of published studies in which positive selection has been inferred in HIV. Several published datasets of HIV coding sequences reported to contain sites undergoing adaptive evolution were downloaded or requested from the authors for re-evaluation (de Oliveira et al., 2004; Zanotto et al., 1999; Massingham and Goldman, 2005; Lemey et al., 2005; Yang et al., 2003).

(de Oliveira et al., 2004) analysed both subtype B and C sequences and selection was found in all 9 HIV genes. Only sequences that were classified as nonrecombinant in the Los Alamos HIV database (http://www.hiv.lanl.gov) were included in the study. For the PARRIS re-analysis, a comparison between the *gag*, *rev* and *tat* genes was carried out (de Oliveira et al., 2004). Separate analysis was done for the *gag* datasets where, in the one case the overlapping coding regions were removed, and the other where the overlapping regions were retained. Reanalysis of a *nef* dataset (47 sequences; 618bp) from a hemophiliac patient infected with HIV-1 subtype B (Plikat et al., 1997) was also carried out. The sequences were collected over a period of 30 months; fifteen sequences from 11 months, 16 from 25 months, and 16 from 41 months post-infection. Zanotto et al. (1999) applied a maximum likelihood method to this *nef* gene datasets and compared the selection results to previous studies where contrasting selective pressures were found.

Yang et al. (2003) investigated the selective pressure in 5 individual HIV-1 gene datasets, which included 5, 7, and 14 sequences from subtypes A, B, and C respectively. Overlapping as well as *env* hypervariable regions were removed prior to the analysis and only nonrecom-

binant sequences, as defined in GenBank, were included. Here, reanalysis of the *env*, *gag*, and *pol* datasets from the original study was carried out. A second subtype B *pol* dataset for which positive selection has been shown to occur (Massingham and Goldman, 2005), was also included in the re-evaluation. This dataset was analysed with and without the overlapping coding regions contained in the alignment. The final dataset used for comparison comprised of 56 full-length coding sequences from HIV-1 group M subtypes (Lemey et al., 2005). A summary of the datasets re-analysed in this chapter is shown in Table 5.1.

Table 5.1: Summary of the sequence datasets that were re-evaluated in this study and for which selection has been previously described.

Gene / Region	Dataset	# Seq	# nt	Subtype	Dataset Reference
env	envABC	26	2082	A, B, C	Yang *et al.* (2003)
pol	polABC	26	2721	A, B, C	Yang *et al.* (2003)
	polB	23	2841	B	Massingham and Goldman (2004)
	polB_No_Overlap	23	2739	B	Massingham and Goldman (2004)
tat	tatB	26	319	B	de Oliveira *et al.* (2004)
	tatC	26	316	C	de Oliveira *et al.* (2004)
rev	revB	26	351	B	de Oliveira *et al.* (2004)
	revC	26	374	C	de Oliveira *et al.* (2004)
nef	11 mo	15	626	B	Zanotto *et al.* (1999)
	25 mo	16	621	B	Zanotto *et al.* (1999)
	41 mo	16	621	B	Zanotto *et al.* (1999)
	11+25 mo	31	626	B	Zanotto *et al.* (1999)
	25+41 mo	32	621	B	Zanotto *et al.* (1999)
gag	gagB	27	1620	B	de Oliveira *et al.* (2004)
	gagC	27	1583	C	de Oliveira *et al.* (2004)
	gagB_No_Overlap	27	1347	B	de Oliveira *et al.* (2004)
	gagC_No_Overlap	27	1305	C	de Oliveira *et al.* (2004)
	gagABC	26	1260	A, B, C	Yang *et al.* (2003)
Full Genome	FG	56	8667	Group M	Lemey *et al.* (2005)

The PARRIS output files include information on the partitioning results, likelihood estimates for each evolutionary model, the p-value for inference of positive selection, as well as the posterior probability estimates for individual sites evolving under the influence of positive selection. The results from these files were compared to the respective published findings.

5.3 Results and Discussion

Scheffler et al. (2006) used PARRIS to analyse two HIV-1 subtype C datasets for which recombination levels have been shown to be high enough to cause false inference of positive selection. In the case of an *env* dataset, the four different models (standard, synonymous rate variation, partitioning, and synonymous rate variation with partitioning) all detected positive selection at a high significance level. However, the partition model with synonymous rate variation, produced a far more conservative estimate of significance for the result, as well as in the magnitude of positive selection inferred under the M2a model.

The results for the analysis of the *gag* sequence dataset revealed evidence for positive selection only when the standard and partitioning models were used. When the synonymous rate was allowed to vary however, the results were no longer significant, and the significance levels dropped even lower when partitioning together with synonymous rate variation was applied to the data.

The simulation results from the original study suggest that by using partitioned data and allowing distinct tree topologies and branch lengths for individual segments, the false positive rate is dramatically reduced. The reduction in false positives, to within acceptable levels given a significance threshold, was associated with only a minor loss in power to detect selection. These simulation results suggest that by using PARRIS, the false positives rate is significantly reduced, and the method has the power to detect selection when the signal is present in an aligned dataset (Scheffler et al., 2006).

In consideration of these initial results, as well as the problems associated with identifying

selection in recombining coding sequences, PARRIS was used reanalyse a collection of HIV-1 datasets (Table 5.1). The aim of the study was to compare the findings from previously reported selection results to those obtained when a method accounting for recombination, such as PARRIS, is applied to the same published datasets.

5.3.1 Extent of Recombination in the Datasets Re-analysed

To evaluate the extent of recombination present in the datasets re-analysed in this study, GARD, within PARRIS, was applied to the aligned sequences. The number of recombination events detected is shown in Table 5.2. For the datasets analysed here, no more than 2 breakpoints were identified for sequences shorter than 1500 nucleotides.

Although these results are based on a limited number of HIV-1 datasets, it is evident from Table 5.2, that recombination occurs throughout the HIV genome. It is likely that the high number of breakpoints identified by Yang et al. (2003) for the envelope gene, is indicative of the extent of *env* sequence variation between subtypes, which can be as high as 30% between subtypes from group M HIV-1 viruses (Gao et al., 2004). Interestingly, the same conclusion can not be drawn from the intersubtype recombination analysis performed on the *gag* gene in this study. Only a single recombination event was identified, even though *gag* heterogeneity between group M subtypes can be as high as 15% (Levy, 2007). The analysis performed on the *gag* datasets from de Oliveira et al. (2004), provided an indication of the potential influence overlapping coding regions can have on the number of recombination events occurring within a gene.

The results in Table 5.2 illustrate the extent of recombination in the HIV-1 genome, and provide the underlying justification for re-analysing these datasets with a method that takes recombination into account.

Table 5.2: Summary of the sequence datasets together with the number of recombination breakpoints for each dataset, detected by PARRIS.

Gene / Region	Dataset	Subtype	Breakpoints	Partition			Dataset Reference
				Start	End	Length	
env	envABC	A, B, C	5	1	285	285	Yang *et al.* (2003)
				286	438	153	
				439	819	381	
				820	1203	384	
				1204	1648	445	
				1649	2082	434	
pol	polABC	A, B, C	3	1	492	492	Yang *et al.* (2003)
				493	1044	552	
				1045	2377	1333	
				2378	2721	344	
	polB	B	4	1	670	670	Massingham & Goldman (2004)
				671	1509	839	
				1510	1800	291	
				1801	2298	498	
				2299	2841	543	
	polB_No_Overlap	B	4	1	594	594	Massingham & Goldman (2004)
				595	1533	939	
				1534	1731	198	
				1732	2349	618	
				2350	2739	390	
tat	tatB	B	1	1	236	236	de Oliveira *et al.* (2004)
				237	319	83	
	tatC	C	1	1	212	212	de Oliveira *et al.* (2004)
				213	316	104	
rev	revB	B	2	1	158	158	de Oliveira *et al.* (2004)
				159	289	131	
				290	351	62	
	revC	C	1	1	257	257	de Oliveira *et al.* (2004)
				258	374	117	
nef	11 mo	B	2	1	69	69	Zanotto *et al.* (1999)
				70	573	504	
				574	626	53	
	25 mo	B	1	1	323	323	Zanotto *et al.* (1999)
				324	621	298	
	41 mo	B	1	1	49	49	Zanotto *et al.* (1999)
				50	621	572	
	11+25 mo	B	2	1	288	288	Zanotto *et al.* (1999)
				289	573	285	
				574	626	53	
	25+41 mo	B	1	1	16	16	Zanotto *et al.* (1999)
				17	621	605	

...continues

Table 5.2 *continued*:

Gene / Region	Dataset	Subtype	Breakpoints	Partition			Dataset Reference
				Start	End	Length	
gag	gagB	B	5	1	240	240	de Oliveira *et al.* (2004)
				241	409	169	
				410	996	587	
				997	1238	242	
				1239	1468	230	
				1469	1620	152	
	gagC	C	5	1	257	257	de Oliveira *et al.* (2004)
				258	449	192	
				450	651	202	
				652	950	299	
				951	1266	316	
				1267	1583	317	
	gagB_No_Overlap	B	3	1	228	228	de Oliveira *et al.* (2004)
				229	411	183	
				412	956	545	
				957	1347	391	
	gagC_No_Overlap	C	4	1	235	235	de Oliveira *et al.* (2004)
				236	495	260	
				496	891	396	
				892	1213	322	
				1214	1305	92	
	gagABC	A, B, C	1	1	302	302	Yang *et al.* (2003)
				303	1260	958	
Full Genome	FG	Group M	7	1	326	326	Lemey *et al.* (2005)
				327	1479	1153	
				1480	2439	960	
				2440	4976	2537	
				4977	6009	1033	
				6010	6884	875	
				6885	8170	1286	
				8171	8667	497	

5.3.2 Reanalysis of Specific Previously Published Datasets

PARRIS was used to re-evaluate the selection pressures in HIV-1 genes while taking recombination into account. The previously published studies reported widespread selection acting on the HIV-1 genome. Here the results for particular studies are shown, whereas comparisons between the results from different studies presented in the subsequent section.

The reanalysis of the *nef* gene dataset, originally tested for positive selection by Zanotto et al. (1999), produced contrasting results. PARRIS favoured the selection model for the analysis of the 11mo dataset ($p = 0.018$) only, whereas the original publication reported selection in the 25mo dataset ($p < 0.0001$), the 25mo in combination with the 11mo dataset ($p < 0.001$), as well as the 25mo in combination with the 41mo ($p < 0.001$) dataset. For the PARRIS analysis, 5 sites had a posterior probability of >95% of evolving under positive selection. The original publication reported a total of 17 sites falling into this class (Zanotto et al., 1999), albeit some overlap in the sites identified for different combinations of the contemporaneous datasets. GARD identified either 1 or 2 breakpoints for each dataset. Recombination is one possible explanation for the contrast between the PARRIS results and those from the original publication. An additional potential bias is that the *nef* gene overlaps with the 3' LTR within which, for example, important promoter and enhancer elements are found (Levy, 2007). More than half of the *nef* gene falls within the overlapping *nef*-3'LTR region; the selection analysis of this region could therefore provide spurious results as the dual functioning region can result in violation of assumptions made by codon models of sequence evolution. PARRIS allows for synonymous rate variation, which can account to some extent for selection acting at the nucleotide level to preserve the functions of the LTR, whereas the method applied by Zanotto et al. (1999) does not, which may have contributed to the contrasting results.

Reanalysis of the six datasets, and two additional *gag* datasets for which the overlapping coding regions were excluded, obtained from de Oliveira et al. (2004) produced similarly conflicting results. Where the original publication identified selection in subtype C *tat*, *rev*, and *gag*, the results from PARRIS suggest that selection took place in only subtype B *tat* ($p = 0.01$), and the subtype C *gag* ($p = 0.02$) dataset that contained no overlapping coding regions (Table 5.3). Regarding *rev* and *gag*, the comparison between the selection results for subtypes B and C showed the same trend in level of significance; smaller p-values were obtained for the subtype C genes than for the subtype B genes from both the PARRIS and the published analyses. The opposite was found for *tat*. de Oliveira et al. (2004) found evidence for selection in subtype C *tat* and not in subtype B, whereas the converse was true for the PARRIS results. Only one site, for subtype C *gag* (overlapping coding regions

excluded) could be inferred with confidence to be evolving under positive selection when recombination was taken into account using PARRIS.

Table 5.3: Results from de Oliveira et al. (2004) and PARRIS; for *gag*, the analysis was repeated for a dataset where the overlapping coding regions were removed.

	De Oliveira *et al.* (2004)				PARRIS			GARD
Gene	**Subtype**	**2 x Delta Log Likelihood**	**P-value**	**# Sites**	**2 x Delta Log Likelihood**	**P-value**	**# Sites**	**# Brkpts**
tat	B	106.68	$p < 0.001$	19	10.29	0.01	-	1
	C	75.6	$p < 0.001$	30	1.19	0.55	-	1
rev	B	75.68	$p < 0.001$	32	0.31	0.85	-	2
	C	48.66	$p < 0.001$	17	2.35	0.31	-	1
gag	B	52.86	$p < 0.001$	-	0.00	1.00	-	5
	C	216.08	$p < 0.001$	-	0.13	0.94	-	5
gag	B_No_Overlap	-	-	-	3.43	0.18	-	3
	C_No_Overlap	-	-	-	8.02	0.02	1	4

The analyses of *tat* and *rev*, upon which much of the focus was placed in the original publication (de Oliveira et al., 2004), may have been significantly biased due to their overlap with other coding regions. The entire *rev* gene shares its coding region with other genes, it partially overlaps with *tat* and partially with *env* (Figure 5.2 [1]). Similarly, *tat* completely overlaps with parts of *rev* and *env*.

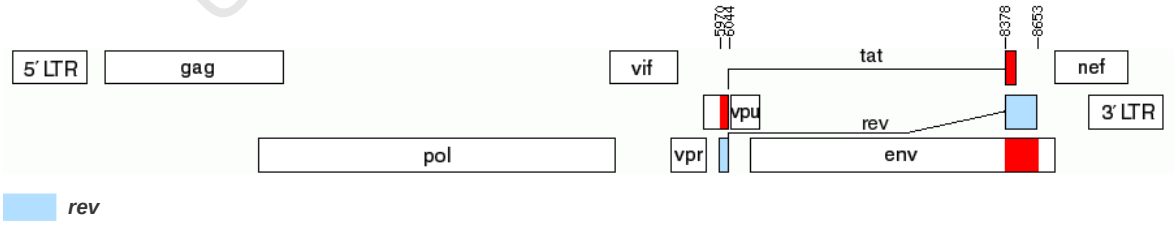


Figure 5.2: Schematic of the HIV-1 genome, showing the overlap of the *rev* gene with *tat* and *env*.

[1]Figure created with the Recombinant HIV-1 Drawing Tool http://www.hiv.lanl.gov/content/sequence/DRAW_CRF/recom_mapper.html

This means that *rev* (and likewise *tat*) does not evolve independently since it is affected by the evolutionary constraints affecting *tat* and *env*. In this way the nucleotide substitution rate for any gene within an overlapping coding region may be reduced due to the evolutionary limitations imposed by the overlapping genes (Suzuki, 2006), which in turn would affect the validity of tests of positive selection. Moreover, it has been shown that ignoring the evolutionary constraints resulting from overlapping genes could lead to an elevated estimate of positive selection (Sabath et al., 2008). Since PARRIS allows the synonymous substitution rates to vary across sequence positions, and therefore permits the rates in overlapping regions to differ according to the evolutionary constraints involved, the number of false positive results may be reduced. However, although PARRIS may fare better than the conventional methods that assume no synonymous rate variation, allowing the synonymous rate to differ across the sequence is not an effective way of addressing overlapping coding regions when dN/dS is used to infer positive selection. Only methods that allow the selection pressures to vary by considering all the reading frames in which proteins are encoded simultaneously, would sufficiently account for overlapping coding regions. However, it is a challenging task to develop such a model, and these methods suffer from a lack in power to identify selection (Pedersen and Jensen, 2001; McCauley et al., 2007). Neither PARRIS nor the approach used by de Oliveira et al. (2004) specifically address different selection pressures over all reading frames. The positive selection results obtained for *tat* and *rev* should be interpreted with caution, since the observed synonymous substitution rate may not be an accurate reflection of the neutral rate, and potential nucleotide-level selection pressures, or varying mutation rates, may bias the subsequent selection analyses of coding genes (Ngandu et al., 2008).

In the genome-wide adaptive evolution study by Yang et al. (2003), *tat*, *rev*, and *vpu* were excluded from the complete selection analyses due to the degree by which the genes overlap with different reading frames. They did however indicate that positive selection was found by the LRT for *tat* and *vpu*, although the non-overlapping sequence lengths of 39 and 58 codons respectively were too short to be included in the subsequent analysis. After removing the overlapping regions present in the remaining genes, as well as excluding the hypervariable regions of gp160 and sequence parts harbouring large indels, positive selection was found in all the other major HIV genes. The PARRIS analysis was carried out on *env*, *pol*, and *gag*.

No evidence of adaptive evolution was found for the *pol* and *gag* genes, and GARD detected 3 and 1 breakpoint/s respectively. However, the results for *env* were highly significant ($p = 0.0007$). Five breakpoints were detected by GARD, and one site (52 in the alignment) was identified as evolving under positive selection. The detected site maps to position 85 in the HXB2 reference sequence and corresponds to a site identified in the original analysis (Yang et al., 2003). It is likely that the PARRIS analysis is more conservative than the standard methods since Yang et al. (2003) classified more than fifty other sites in *env* as evolving under positive selection.

The results from the PARRIS analysis contradict one of the main findings from the original paper where genome-wide positive selection was detected (Yang et al., 2003). No evidence for positive selection was found by PARRIS for the *pol* and *gag* genes. Yang et al. (2003) investigated the influence that recombination may have had on their selection analysis by re-analysing the data with a star tree, which is expected to be wrong for all sites, instead of the gene tree, and similar results were obtained. Moreover, the authors argue that the positively selected sites were dispersed on the primary sequence and clustered on the tertiary structure, suggesting functional constraints in certain regions of the genome and potential hotspots of selection in other regions, for example, T-helper cell epitopes. Following from this they reason that if recombination had caused false positive inferences, the positively selected sites would cluster on the primary sequence since recombination is expected to affect a group of neighbouring nucleotides (Yang et al., 2003).

The above arguments were used to claim that recombination may not have seriously affected the positive selection results (Yang et al., 2003). Shriner et al. (2003), however, discussed the demerits of using the star phylogeny approach, which disregards the phylogenetic history of the sequences completely, as a possible alternative when recombination has affected the underlying sequence dataset. Since for a star phylogeny each mutational event is counted independently and represents an individual step in evolution, the estimate of the total number of mutations will generally be inflated, which may result in an overestimation of the number of sites evolving under positive selection (Shriner et al., 2003). Overestimation of the number of selected sites is therefore likely to occur in both cases where an incorrect phylogeny

is used for the analysis, the first as result of recombination and the second due to the star tree approach. Phylogenetic network models may be the most preferable method to use for this analysis (Huson and Bryant, 2006; Woolley et al., 2008), and further investigation of these methods is an interesting avenue for future analyses.

Furthermore, the argument that the selected sites would cluster on the primary sequence, and not the tertiary structure, if recombination biased the results, is not valid. There is no reason to believe that true positively selected sites should be uniformly distributed on the primary sequence or tertiary structure (Huzurbazar et al., 2010). Selection is also not expected to occur uniformly across the primary sequence; the presence of CTL epitopes, and linear antibody epitopes, among other causes, can result in selection hotspots. This would lead to a non-uniform distribution of positively selected sites that can not be attributed to recombination. Discounting the potential influence of recombination with these arguments, is therefore not well-grounded. Since PARRIS identifies fewer positively selected sites than the analysis carried out by Yang et al. (2003), and since simulation studies have shown that PARRIS has a lower rate of false positive inference without significant loss of power (Scheffler et al., 2006), it is likely that the PARRIS results contain a higher proportion of true positive estimates.

Massingham and Goldman (2005) applied a slightly different method to identify adaptive evolution in coding sequences. The test integrated the best existing features available at the time, including the substitution models of Nielsen and Yang (1998) and the site by site testing method of Suzuki and Gojobori (1999). This combined method, the sitewise likelihood-ratio (SLR) test, uses the entire alignment to determine parameter values common to all sites, but tests each site for neutrality independently. Another method was included in this study for comparison with the SLR test; in the second method, called SNY, the M8a model (Swanson et al., 2003) and M8b model (where $\omega \geq 1$; a restricted form of model M8 from Yang, 2000) were used (Massingham and Goldman, 2005).

Evidence for adaptive evolution was found by all three methods (Table 5.4). Three sites, 67, 347, and 478 in the alignment, were identified by PARRIS to be evolving under positive

selection (Table 5.4). Two of these, sites 67 and 347, were also identified by PARRIS for the analysis where the overlapping regions were removed. Furthermore, all the sites identified by PARRIS were also contained within the site-specific selection inference from the SNY and SLR method described by (Massingham and Goldman, 2005). Of the 13 sites identified by the SNY approach, sites 67, 347, and 478 were three of the six that had a posterior probability of >99%. The remaining 7 sites all had a posterior probability of >95%. Similarly, the SLR method found 22 sites, with the aforementioned three sites falling into the category of highest significance ($p < 0.05$). Additionally, sites 67 and 347 were the only two sites which retained significance after correcting for multiple comparisons (Massingham and Goldman, 2005). These were also the two sites identified by PARRIS for the analysis of the *pol* dataset for which the overlapping coding regions were removed, illustrating the effect dual coding regions can have on selection analysis. The support for positive selection acting on the gene was also not as significant for the dataset where the overlapping coding regions were retained ($p < 0.001$ for the dataset including, and $p = 0.02$ for the dataset excluding the overlapping coding regions, Table 5.4).

Table 5.4: PARRIS and published findings (Massingham and Goldman, 2005) showing evidence for selection in an HIV-1 *pol* dataset; the PARRIS analysis was applied to datasets including and excluding the overlapping coding regions.

Massingham *et al.* (2004)						PARRIS				GARD
Gene	**Subtype**	**Method**	**2 x Delta Log Likelihood**	**P-value**	**# Sites**	**Dataset - Overlapping Regions**	**2 x Delta Log Likelihood**	**P-value**	**# Sites**	**# Brkpts**
pol	B	SLR	N/A	N/A	22	Retained	21.6469	$p < 0.001$	3	4
		SNY	60.27	$p < 0.001$	13	Excluded	8.2794	0.02	2	4

Shriner et al. (2003) demonstrated the effect of recombination on sitewise methods designed to detect positive selection. Due to the reliance of the method on the parameters defining the phylogeny, factors influencing the phylogenetic reconstruction would affect the outcome of the selection analysis. Recombination would result in a longer phylogeny due to the erroneous assumption that multiple changes had occurred along the branches of the phy-

logeny, where in reality, the changes are false homoplasies caused by recombination events (Shriner et al., 2003). Recombination may therefore have affected the results obtained from the sitewise approach used by Massingham and Goldman (2005).

The final reanalysis was carried out on a full HIV-1 genome dataset, comprising of 56 group M sequences, previously used by Lemey et al. (2005). The authors applied the single rate model (M0) together with the discrete model (M3) to estimate the selective pressures acting on the viral genome (Yang et al., 2000; Lemey et al., 2005). PARRIS analysis was completed for the same alignment of 56 sequences (8667 nucleotides), and both methods found strong evidence for selection. The published parameter estimates indicated that the discrete model provided a better fit for the data and suggested that 9% of the entire non-overlapping genome was evolving under adaptive evolution, with $^{d_N}/_{d_S} = 2.51$ for the positive selection class (Lemey et al., 2005). The analogous results from PARRIS suggested that positive selection occurred at 6.5% of the genome with $^{d_N}/_{d_S} = 2.21$.

The authors drew a comparison of the results they found to those obtained in the study of Yang et al. (2003), which also forms part of the reanalysis carried out in this chapter. The collective analysis of all the HIV-1 genes analysed separately by Yang et al. (2003), suggest that approximately 10% of the entire genome was evolving selectively (Lemey et al., 2005). This is comparable to the estimate described by Lemey et al. (2005) (9%). Similar to the discussion on the results from Yang et al. (2003), it is likely that the estimates from the report by Lemey et al. (2005) also may have been biased by the presence of recombination. The PARRIS results here suggest a slightly lower percentage (6.5%) of sites were evolving adaptively, which again appears to suggest that this may represent a more conservative estimate with the site-specific results containing fewer false positives.

The PARRIS site-specific analysis identified 19 codon sites in the positive selection class with a posterior probability of greater than 0.95, suggesting that these sites were evolving adaptively. The distribution of posterior probabilities for the full-genome analysis is shown in Figure 5.3. Of the 19 positively selected sites, 9 were located within *env*, 1 in *gag*, 2 in *pol*, and the remaining 7 were clustered within the accessory genes: *vif*, *vpr*, *tat*, *vpu*, and

rev. These results correspond to the results from Lemey et al. (2005), where the highest density of positively selected sites was found in *env* and the accessory genes.

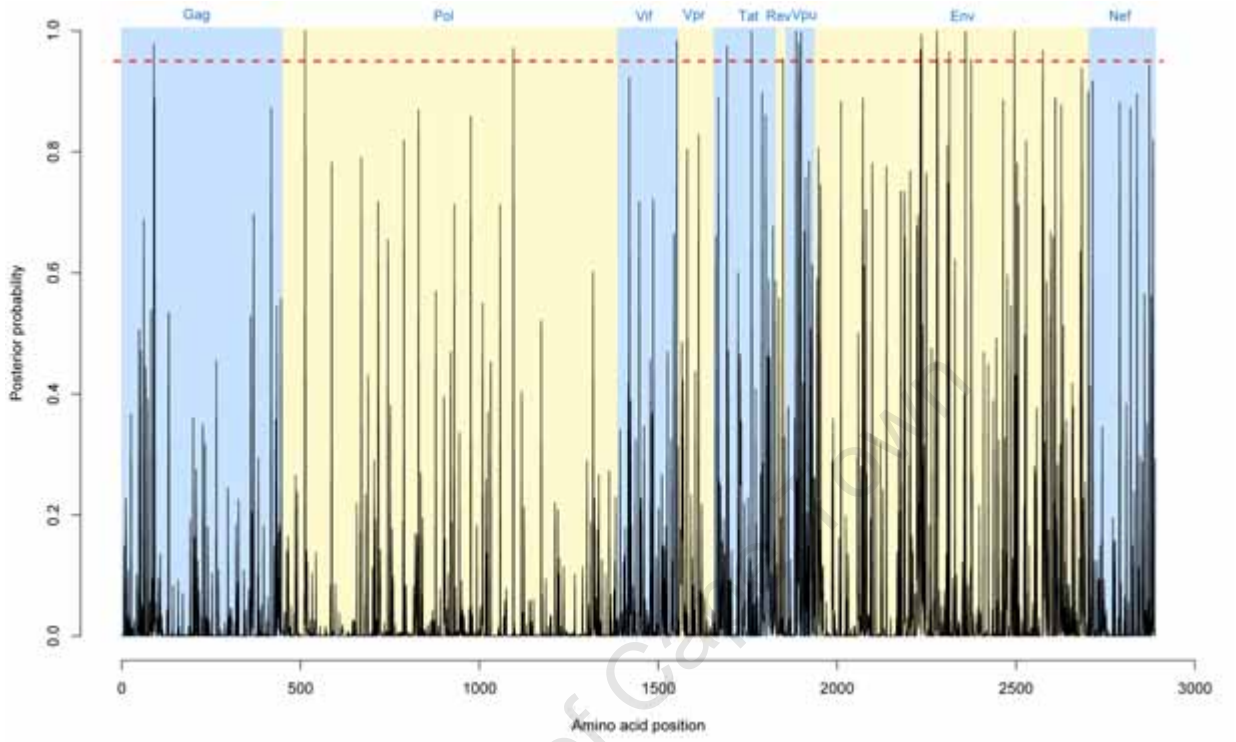


Figure 5.3: Distribution of posterior probabilities for the full HIV-1 genome from Lemey et al. (2005), estimated by PARRIS. The red dashed line indicates the > 0.95 cut-off of significance.

Interestingly, 9 of the positive selected sites mapped to known epitopes in the Los Alamos HIV Immunology Database (http://www.hiv.lanl.gov/content/sequence/ELF/epitope_analyzer.html), and 5 of these were in the accessory genes, *vif*, *vpr*, *tat*, *vpu*, and *rev*. Furthermore, one of the sites identified in *tat* (Lys at position 29, numbered according to the HXB2 reference sequence, starting from Tat exon 1) was directly adjacent to a site, Lys28, which has previously been associated with impaired viral replication when mutated to either Gln or Arg (Brès et al., 2002). Lastly, only 2 of the 19 sites, one in *gag* and one in *env* (within the Rev response element), appeared to be as result of APOBEC3G hypermutation.

5.3.3 Overall Comparison between Results from PARRIS and the Original Publications

Collectively, the reanalysis of previously published HIV datasets (Table 5.1) illustrates how selection results can differ when different evolutionary models or assumptions are applied to the data; moreover, the differences in the obtained results suggest that the original selection results may be unreliable, and that the conclusions drawn from the initial results need reconsideration. Incorrect model assumptions, for example when a model does not account for the possibility of recombination when analysing recombining sequence datasets, may lead to results that contain a large number of false positive inferences. In two of the five published cases, the authors only included sequences thought to be nonrecombinant (de Oliveira et al., 2004; Yang et al., 2003); however, none of the studies explicitly tested the sequences for recombination. Furthermore, only Yang et al. (2003) and Lemey et al. (2005) removed overlapping coding regions before the selection analysis. Although PARRIS allows synonymous rate variation and therefore provides a comparatively more suitable means to analyse overlapping coding regions, this feature is not sufficient to exclude the bias resulting from overlapping coding regions, and in all cases these areas should be removed prior to positive selection analysis. Modelling synonymous rate variation is also not unique to PARRIS, but yet an important feature of the model. PARRIS provides a more robust method for identifying selection in HIV by accounting for these two features, recombination and shared coding regions, common to the pathogen.

The disparities between results obtained from PARRIS and those from previously published reports, further demonstrate how recombination can significantly bias both the inference of positive selection from model comparison, as well as the site-specific selection inferences (Anisimova et al., 2003; Shriner et al., 2003; Scheffler et al., 2006). A high false discovery rate has been demonstrated when standard phylogenetic methods are applied to infer positive selection from recombining sequences (Anisimova et al., 2003; Shriner et al., 2003), and this could lead to inaccurate predictions and therefore irrelevant results.

The reanalysis carried out in this chapter, included two separate datasets, from different

publications, for each of the *pol* and *gag* genes. Although there is no reason to believe that selection will be present at the same sites in different datasets, it is nevertheless interesting to compare the results obtained. For instance, the strong support for positive selection (p = 0.00002) obtained from the PARRIS analysis for the *pol* gene originally included in the study by Massingham and Goldman (2005), contrasts with the PARRIS results for the second *pol* gene dataset, from Yang et al. (2003), where no significant evidence was found. A major difference between the two datasets used, is that, prior to analysis, Yang et al. (2003) excluded overlapping regions as well as indels, with a final dataset of sequence length 2721 bp, while Massingham and Goldman (2005) did not (sequence length 2841 bp). It is possible that the overlapping regions may have influenced the selection analysis, since the evolutionary pressures acting on the different reading frames encoding proteins are not accounted for by the method. An additional analysis was carried out on the *pol* dataset from Massingham and Goldman (2005)where the overlapping coding regions were removed, resulting in an alignment of 2739 bp in length. For this dataset, the evidence of positive selection found by PARRIS, was far less significant (p = 0.02), further illustrating the importance of removing dual coding regions prior to selection analysis.

The PARRIS results for the three original *gag* gene datasets, two from (de Oliveira et al., 2004) and one from Yang et al. (2003), provided no evidence of positive selection, despite the publications reporting evidence for selection in each case. However, for the additional analysis of the de Oliveira et al. (2004) subtype C *gag* dataset, where the overlapping coding regions were excluded, PARRIS did report evidence for adaptive evolution acting on the gene, and one site was identified during the site-specific selection inference.

A total of nineteen datasets, from five separate studies (de Oliveira et al., 2004; Zanotto et al., 1999; Massingham and Goldman, 2005; Lemey et al., 2005; Yang et al., 2003), for which positive selection has been described, were re-analysed in this chapter. PARRIS only identified positive selection in 7 of the 19 datasets, and the number of positively selected sites were reported by the respective publications for five of these seven cases. The number of positively selected sites found by PARRIS was compared to the corresponding number provided by the original publication (Table 5.5).

Table 5.5: The number of positively selected sites identified by PARRIS and the numbers from four original publications for which the values were available.

Publication	Gene	Dataset / Method	Number of Selected Sites		
			Published	PARRIS	Overlap
Zanotto *et al.* (1999)	*nef*	11mo	17*	5	4
de Oliveira *et al.* (2004)	*tat*	tatB	19	0	0
Yang *et al.* (2003)	*env*	envABC	61	1	1
Massingham *et al.* (2004)	*pol*	SLR	22	3	3
		SNY	13	3	3
Massingham *et al.* (2004) [excluding overlapping regions]	*pol*	SLR	22	3	2
		SNY	13	3	2

* Number of sites identified from the combined datasets (11 and 25 months, and 25 and 41 months), values for the individual time-point-datasets were not provided.

It is clear from Table 5.5 that PARRIS identifies a far lower proportion of sites evolving under positive selection than the corresponding original reports. Interestingly, all but one of the sites identified by PARRIS, were also found by the published reports. This is remarkable consistency since not all the published studies applied the same methods to identify selection. Furthermore, PARRIS identified no more than 25% of the sites found for the published cases. It is therefore evident that the results PARRIS provides, includes only a fraction of the sites that were previously believed to be evolving under positive selection. Given that PARRIS has previously been shown to have a reduced false positive rate but without much reduction in power (Scheffler et al., 2006), the majority of sites previously reported to be evolving adaptively, may be false positives. This further illustrates the potential bias recombination may have on positive selection inference studies and the importance of taking recombination into account in investigations of positive selection.

5.4 Conclusions

The preliminary findings reported by Scheffler et al. (2006), as well as the PARRIS results described in this chapter, illustrate the importance of considering the assumptions made by the evolutionary model prior to selection analysis. High recombination rates have been identified for HIV (Burke, 1997), and this is associated with large false positive rates, suggesting that a method that takes recombination into account is very useful for HIV selection studies. Highly disparate results were obtained in the majority of cases when PARRIS was applied to previously published datasets reported to contain sites undergoing adaptive evolution. Since PARRIS accounts for both recombination and synonymous rate variation, the difference in the observed result could be due to the inclusion of either of these features.

Numerous studies have focused on identifying selection in HIV-1 coding regions, and sites likely to be evolving under positive selection have been listed extensively (Frost et al., 2001; Yang et al., 2003; Zanotto et al., 1999; Lemey et al., 2005; de Oliveira et al., 2004). It is likely that many of these sites that are thought to be selected for in HIV-1, are false positives. False positive rates provide a measure of confidence in a specific method to produce accurate results, with fewer Type 1 errors reflecting greater reliability in the method. Compared to the previously published results discussed in this chapter, PARRIS gives a much more conservative estimate for the evidence of positive selection, and far fewer sites are identified in the site-specific selection analysis. It is also very encouraging that all but one of the sites identified by PARRIS were within the originally reported set of positively selected sites. With less sites identified, more focused analysis can be carried out on specific sites, for example experimentally verifying the function or effects of mutations on viral fitness. A dramatic reduction in false positive rates has been shown with PARRIS, and although this was associated with a slight loss in power to detect selection, the drop in power was not significant (Scheffler et al., 2006). PARRIS may therefore be a very promising method for accurately identifying sites undergoing positive selection. Therefore, the analysis carried out, and the results described, in this chapter are an important step towards accurately identifying adaptively evolving sites in the HIV-1 genome.

Chapter 6

Evaluation of Covarion Models of Codon Evolution and Application to SIVcpz / HIV-1 Zoonosis

6.1 Introduction

The method discussed in Chapter 5, PARRIS, identifies adaptively evolving sites in coding sequences while taking recombination into account. Furthermore, the method also allows synonymous substitution rate variation across all sites along the sequence. This is particularly important for studies focusing on HIV evolution, as site-to-site synonymous rate variation has been shown to be a prominent feature of RNA viruses, and specifically so for HIV (Hanada et al., 2004; Kosakovsky-Pond and Muse, 2005; Ngandu et al., 2008). In this chapter, the idea of site-to-site rate variation is further explored, although in this case not specifically focussing on the variation along a sequence, but rather variation between different lineages of a phylogenetic tree.

The covarion model of molecular evolution was developed to account for within-site substitution rate changes along the lineages of a phylogenetic tree (Tuffley and Steel, 1998;

Galtier, 2001; Huelsenbeck, 2002). According to this model, a site can evolve rapidly along one branch and more slowly along others (Dorman, 2007). This can be used to search for specific amino acids that are responsible for functional changes between divergent branches of a phylogeny (Gaucher et al., 2001; Knudsen and Miyamoto, 2001). Covariation occurs when, for example, selective pressures or functional constraints either favour or prevent change at different time points in evolutionary history (Allman and Rhodes, 2009). This could take place during HIV infection where viral evolution is driven by immune-mediated selection, and therefore fluctuates with the strength of the immune response (Moore et al., 2002; Ross and Rodrigo, 2002; Guindon et al., 2004). Covariation may also take place after gene duplication, where functional divergence of the gene families may result in adaptive evolution (Ohta, 1993; Wang and Gu, 2001; Holmes, 2003).

Covarying evolution could also be described for zoonotic events, as well as subtype divergence, where functional changes result in distinct organisms with separate evolutionary pathways. During zoonosis, the evolutionary constraints acting on a transmitted pathogen may change due to the varying immune responses of the infecting-species and that of the new host (Reid et al., 1999; Furuse et al., 2009). The adaptive sequence changes will be present as covarying sites along the phylogenetic tree relating sequences obtained prior to and after the zoonotic event. Numerous studies have shown that site-specific switching between selection classes across lineages is likely to occur (for example Ohta, 1993; Pupko and Galtier, 2002; Clark et al., 2003; Guindon et al., 2004; Penn et al., 2008). A recent study focussing on HIV-1 group M sequences, reported that the rate of intersubtype HIV-1 evolution varied across a phylogenetic tree, reflecting the change in protein function for each subtype (Penn et al., 2008). This study provided evidence for covariation in all HIV-1 genes, and the authors suggested that differing functional constraints were affecting the course of evolution and sequence variability in HIV-1 group M subtypes (Penn et al., 2008).

The M2a model of Wong et al. (2004) has been used as a null model in studies inferring covarying evolution (Guindon et al., 2004; Penn et al., 2008). However, a random-effects likelihood (REL) implementation of the M2a codon model, where the site-to-site substitution rate variations are random variables drawn from a distribution with 3 discrete categories,

may not provide an appropriate null model. Consider, for instance, a fit of the M2a model to a long sequence alignment, resulting in three ω categories, one with strong purifying selection (for example $\omega = 0.1$), another corresponding to neutral evolution ($\omega = 1$), and a selection site class with $\omega > 1$. If the true value of ω at a site is 0.5, then a covarion version of M2a may provide a better fit at this site, fitting the intermediate value of ω by switching between the two discrete classes. Therefore, the site would appear to covary since the covarion model would provide a relatively better fit for the data than the non-switching model. In this case, a model specifying multiple rate classes would be a more appropriate model for describing the evolution of the sequences.

One of the aims of the current study was to test whether inferred covariation (switching between rate classes) may, in some cases, be an artifact of under-fitting resulting from the use of too few rate classes for ω. Simulation experiments were carried out to estimate the proportion of false positive inferences of covarying evolution in the absence of rate switching. An alternative approach is proposed to overcome the limitations of the REL implementation of the covarion model. The expectation is that a fixed-effects likelihood (FEL) approach would not be susceptible to false inference of switching due to this underfitting problem since ω is estimated separately at each site. A FEL version of the covarion model was implemented to test this hypothesis.

The further aim was to investigate covariation in an HIV - SIV (simian immunodeficiency virus) dataset. HIV-1 is thought to have originated through a cross-species (zoonotic) infection from chimpanzees to humans (Gao et al., 1999), and SIVcpz (infecting chimpanzees) and HIV-1 evolved independently after the zoonosis event. Although the human and chimpanzee genomes are similar at the DNA sequence level, substantial differences between the human immune system and that of other primates exist (Gagneux and Varki, 2001; Kehrer-Sawatzki and Cooper, 2007; Perry et al., 2008; Danilova, 2008). These differences may have driven the evolution of the human and simian form of the virus to varying degrees. In the current study, the goal was to identify sites that switched between rate classes at the point where the cross-species jump occurred, since such sites may reflect changes in selective constraints at positions dependent on species specific immune responses (Moore et al., 2002;

Ross and Rodrigo, 2002; Guindon et al., 2004). Furthermore, the identified sites may reveal sequence regions responsible for functional shifts of HIV proteins, which previously has been reported for protein comparisons between different species (Gaucher et al., 2001; Knudsen and Miyamoto, 2001).

6.2 Methods

6.2.1 The Covarion Model of Sequence Evolution

HyPhy was used to implement the covarion model in a maximum likelihood framework. The model parameterisation was previously described by Guindon et al. (2004), where the authors extended the M2a codon-based model of DNA substitution allowing flexibility in the evolutionary rate over time (Guindon et al., 2004). The same instantaneous transition rate matrix as described in section 1.1.3, was used. Additionally, switches between the three rate classes (the purifying selection class: $\omega < 1$, neutral class: $\omega = 1$, and the positive selection class for which $\omega > 1$) were allowed. Figure 6.1 illustrates the switching between selection classes.

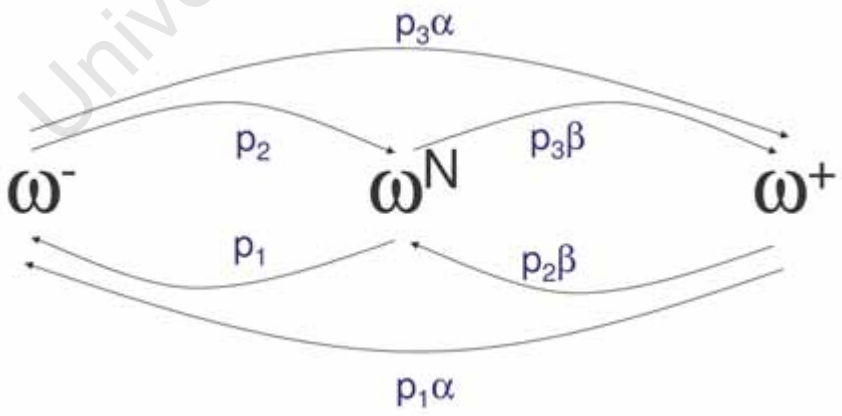


Figure 6.1: Probabilities of switching between the three rate classes, where ω^- is the purifying selection class, ω^N depicts the neutral class, and ω^+ represents the positive selection class. The relative rates α and β are also shown.

A rate of interchange parameter, δ, is applicable to all switching events, and p_x is the equilibrium frequency of each selection class (x= 1..3). The relative rate parameters, α and β, apply to the switching events between the purifying and positive selection classes, and the neutral and positive selection classes, respectively (Figure 6.1). See Guindon et al. (2004) for detailed model specifications. Both a REL as well as a FEL approach to model rate variation were used. The REL model approximates the values of ω across all site with a distribution, whereas the FEL model determines separate ω^-, ω^N, and ω^+ for each site.

Compared to the M2a selection model, the covarion model contains only three extra parameters (δ, α, and β) for the REL implementation. For the FEL approach, the M2a model has a single ω value because there is no need to model a ω distribution for a single site. In this case there are therefore 5 extra parameters (two extra ω values, δ, α, and β) in the covarion model. For both approaches, if $\delta = 0$, switching is prohibited and the covarion model reduces to the traditional selection model (Guindon et al., 2004). The models are therefore nested and a likelihood ratio tests (LRTs) can be used to determine whether the data fit the covarion model significantly better than model M2a. Huelsenbeck (2002) noted that the χ^2 distribution with the appropriate number of degrees of freedom does not hold for the LRT between the M2a and covarion models. The reason for this is that the simpler model (no switching occurs between selection classes) has parameters on the boundaries of the parameter space, and in the current REL example, a mixture of χ^2_0 and χ^2_1 distributions provide a more appropriate null distribution (Whelan and Goldman, 1999; Huelsenbeck, 2002; Ota et al., 2000; Self and Liang, 1987). Huelsenbeck (2002), however, further made the point that the 95% critical value for the general χ^2 distribution with appropriate number of degrees of freedom, is larger than the critical value when a mixture of χ^2 distributions is used, which means that the former, simpler, calculation provides a more conservative estimate (Ota et al., 2000; Huelsenbeck, 2002). In the current study, a χ^2 distribution with 3 (REL implementation) and 5 (FEL implementation) degrees of freedom was used as conservative tests, with a further aim to to avoid potential errors as result of small sample sizes (Ota et al., 2000; Huelsenbeck, 2002; Penn et al., 2008).

6.2.2 Simulations

To test whether false inference of covariation was taking place, HyPhy was used to simulate datasets under a codon model that specifies four rate classes: $\omega_1 = 0.05$, $\omega_2 = 0.5$, $\omega_3 = 1.5$ and $\omega_4 = 3$. Two subsets of sequences (30 and 100 taxa) from an HIV *gag* dataset for which covariation was previously inferred (Penn et al., 2008), were used for the initial likelihood estimation. The existing likelihood function for each dataset was then used to simulate 100 datasets of the same dimensions (sequence number and length), and with the same values of parameters (transition/transversion rate ratio and codon equilibrium frequencies) as those in the current likelihood function. Branch lengths and substitution rate bias parameters were approximated with values derived from the nucleotide model. The simulated datasets were then analysed with both the REL, and FEL, implementation of the M2a and the covarion models. For the REL implementation, the fit of each simulated alignment to each model was compared with a LRT. The number of simulated datasets that fit the covarion model significantly better than the M2a model, formed the false positive estimate for each group of simulations.

For the FEL implementation a random sample from each of the 100 simulated alignments (for both the 30 taxon and 100 taxon datasets) was drawn, and the fit of the randomly selected site to the covarion and M2a models was compared using a LRT. The false positive rate for the FEL implementation, was estimated by counting the number of sites inferred to fit the covarion model significantly better than the M2a model. Although these analyses do not present a direct comparison of the rate of false positive inference of covariation between the REL and the FEL models, the estimates nonetheless provide an indication of the potential false positive rate when gene- (REL) and site-level (FEL) inference of covariation are reported.

6.2.3 Switching Between Rate Classes Associated with Zoonosis

Thirty-two HIV-1 group M and 9 SIV complete nonrecombinant genome sequences were downloaded from the Los Alamos HIV sequence database (http://www.hiv.lanl.gov/). The Genbank accession numbers are AB253421, AB253429, AF004885, AF005494, AF005496, AF061641, AF075703, AF077336, AF082394, AF082395, AF084936, AF103818, AF190127, AF190128, AF286237, AF286238, AF377956, AF447763, AJ249235, AJ249238, AJ249239, AY169968, AY173951, AY253311, AY331295, AY612637, AY772699, DQ373063, DQ373064, DQ373065, DQ373066, DQ676872, DQ853463, K03454, K03455, U42720, U46016, U52953, U88824, U88826, and X52154.[1] Two SIVcpzPts sequences, ANT (Genbank accession U42720) and TAN1 (Genbank accession AF447763), which are commonly used as outgroups for HIV and SIV alignments (Holzmayer et al., 2009; Van Heuverswyn et al., 2007; Wu et al., 2007; Santiago et al., 2003; Huang et al., 2003; Santiago et al., 2002; Vanden Haesevelde et al., 1996), formed the outgroup for all phylogenetic trees in the current study. A table indicating the sequence names and corresponding Genbank accession numbers is provided in the Appendix (Table A2).

The analysis carried out here was focussed on identifying specific sites associated with switching between rate classes along the branch representing the zoonotic event that lead to HIV-1 group M sequences, since these are the viruses that account for the largest percentage of HIV-1 infected individuals (Levy, 2007; Butler et al., 2007; Brennan, 2007; Kandathil et al., 2005). Furthermore, the analysis follows on from the research carried out in the previous chapters that were aimed at characterising the most common subtypes, which form part of the group M viruses.

The genome alignment was separated into individual genes, and the overlapping codon regions as well as the 3' and 5' LTR regions, were removed. The genes and sequence length, of each dataset used for the analysis, is shown in Table 6.1. The *rev* gene was excluded from the analysis due to the complete overlap with other coding regions.

[1] Dr. Nobubelo Ngandu carried out the initial sequence alignment.

Table 6.1: Genes and sequence lengths used to determine the sites where switching between rate classes, at the point of zoonosis, occurred.

Gene	Sequence Length (Nucleotides)
gag	1272
pol	2748
vif	459
vpr	210
tat	117
vpu	189
env	2133
nef	288

The REL implementation of the covarion model was used for the initial likelihood optimisation of the 8 sequence datasets. The ancestral sequences were then reconstructed using a method available in HyPhy (Kosakovsky-Pond and Muse, 2005), which resulted in an alignment that included the rate class as an extra "site" at each codon position. The additional fourth "site" represents either purifying selection ("A"), neutral selection ("C"), or positive selection ("T").

6.3 Results and Discussion

6.3.1 Simulation Study

Simulated datasets were generated to estimate the false positive rate of inference of covariation. The motivation for performing the test was that underfitting through the use of too few ω rate classes could result in an improved fit of the covarion model versus, for example, a 3 rate class null model. In the case of the M2a model, there are the following three rate classes: a purifying selection class ($\omega < 1$), a neutral class ($\omega = 1$) and a positive selection class ($\omega > 1$). If the true ω value for a given site falls between two of these rate classes, then the covarion model would provide a relatively better fit for the data by switching between

the omega categories. Therefore, if underfitting of the null model occurs as in the case described above, it is likely that false positive inference of covariation may take place.

The REL implementation of the models provides gene-level likelihoods, whereas the FEL implementation estimates the likelihoods on a site-by-site basis. Two sets of simulations, a 30 taxon dataset and a 100 taxon dataset, were evaluated with both models. The results are shown in Table 6.2. For the REL implementation, the null hypothesis, that no switching between rate classes occurred, was rejected 57 times (57%) and 100 times (100%) at the 5% significance level in the simulated datasets containing 30 taxa and 100 taxa respectively (Table 6.2). This represents a significant excess over the expected 5% false positive rate. With the FEL implementation, the null hypothesis was rejected 5 times (0.05%) at the 5% significance level in the 100 taxon simulated datasets. There were no false positive inferences of covariation for the FEL analysis of the 30 taxon simulated datasets (Table 6.2). Although the FEL analysis was carried out on individual sampled sites (one from each of the simulated datasets), the lack of false positive inferences of covariation is in stark contrast with the clearly unreliable results obtained from the REL analysis.

Table 6.2: Number of false positives present in the two sets (30 and 100 taxa) of 100 simulated datasets for the REL and FEL implementations of the M2a and covarion models.

Implementation	Number of False Positives present in 100 Simulated Datasets	
	30 Taxa	100 Taxa
REL	57	100
FEL	0	5

The analyses demonstrate that underfitting, potentially through the use of too few rate classes in the null model, can lead to a significant bias in the inference of covariation using a REL approach. However, specifying an adequate number of rate classes for the null model *a priori* is a difficult task, and could lead to overfitting of the data (Yang et al., 2000; Yang and Nielsen, 2002). The FEL approach provides an alternative method for estimating covariation; the false positive rate for the analyses on both the smaller (30 taxa) and the

100 taxon datasets, was within the accepted range at the 5% significance level. However, the model is computationally expensive and therefore difficult to apply, in practice. The power of the FEL model to detect covariation at the gene-level is likely to be very low, since the number of degrees of freedom scaled up from a site specific estimation is dependent on the sequence length. Even for a relatively small dataset the LRT will be very conservative, resulting in few false positives, but also lacking power to detect covariation.

The simulation study presents good evidence that false inference of switching may occur when the REL approach is used to infer temporal inhomogeneity in evolutionary rate, and the number of rate classes is inadequate in the null model. The validity of published studies (Guindon et al., 2004; Penn et al., 2008) may therefore be brought into question, since the frequency of switching may not be as high as previously thought. The results appear to be significant particularly for datasets containing more than 30 taxa, which is less than what is commonly used for maximum likelihood covarion evolutionary analysis (n = 182 for the dataset used in Penn et al., 2008 and between 87 and 160 taxa for the datasets from Guindon et al., 2004) .

The number of discrete rate classes specified in the null model is an important consideration during studies of covarying evolution, since the REL implementation may be susceptible to false positive inference of covarying evolution. The false positive rate is significantly reduced when the FEL approach is implemented and, given sufficient computational resources, it may provide an effective alternative for the REL model.

6.3.2 Evaluating Switching Between Rate Classes Associated with HIV-1 Zoonosis

The REL implementation of the covarion model was applied to 8 non-overlapping gene alignments (*env*, *pol*, *gag*, *tat*, *vpu*, *vpr*, *vif*, and *nef*) containing HIV (n = 32) and SIV (n = 9) sequences. In section 6.3.1 it was shown that this method can produce a high level of false inference of rate switching at the gene level. It is possible that the site-specific inference

of rate switching is also prone to false positive inference using the REL implementation; however the FEL approach is computationally intensive and was not applied to the real data due to time constraints.

The ancestral sequences were reconstructed by applying a method available in the HyPhy package. The rate class at each site could be read directly from the reconstructed ancestral sequences, since the rate classes were incorporated into the sequence as an extra character at each site, that is, a fourth position where "A", "C", and "G" represent ω^{-} (purifying selection), ω^{N} (neutral class), and ω^{+} (positive selection), respectively. Similarly, the most probable rate class at each site for the terminal sequences were available from the resulting alignment containing both the ancestral and descendant sequences. With these data, as well as a phylogenetic tree depicting the position along the tree where zoonosis occurred, it was possible to determine which sites switched between rate classes at the point of zoonosis.

As an example, the analysis carried out on the *gag* gene is shown. The phylogenetic tree relating the *gag* sequences is shown in Figure 6.2. The branch between node "Z" and "X" (Figure 6.2) represents the time at which the zoonotic event took place. Because the aim was to identify HIV-1 group M sites that switched between rate classes upon entering the human host, the focus was on the changes that occurred along this branch. Thirty-five rate switching sites were found along this branch for *gag*; for 6 sites the results indicated a shift from the purifying to the neutral selection class, and for 29 sites a rate switch from the neutral to the purifying selection class. The complete results for all 8 HIV-1 genes included in the study, is shown in Table 6.3.

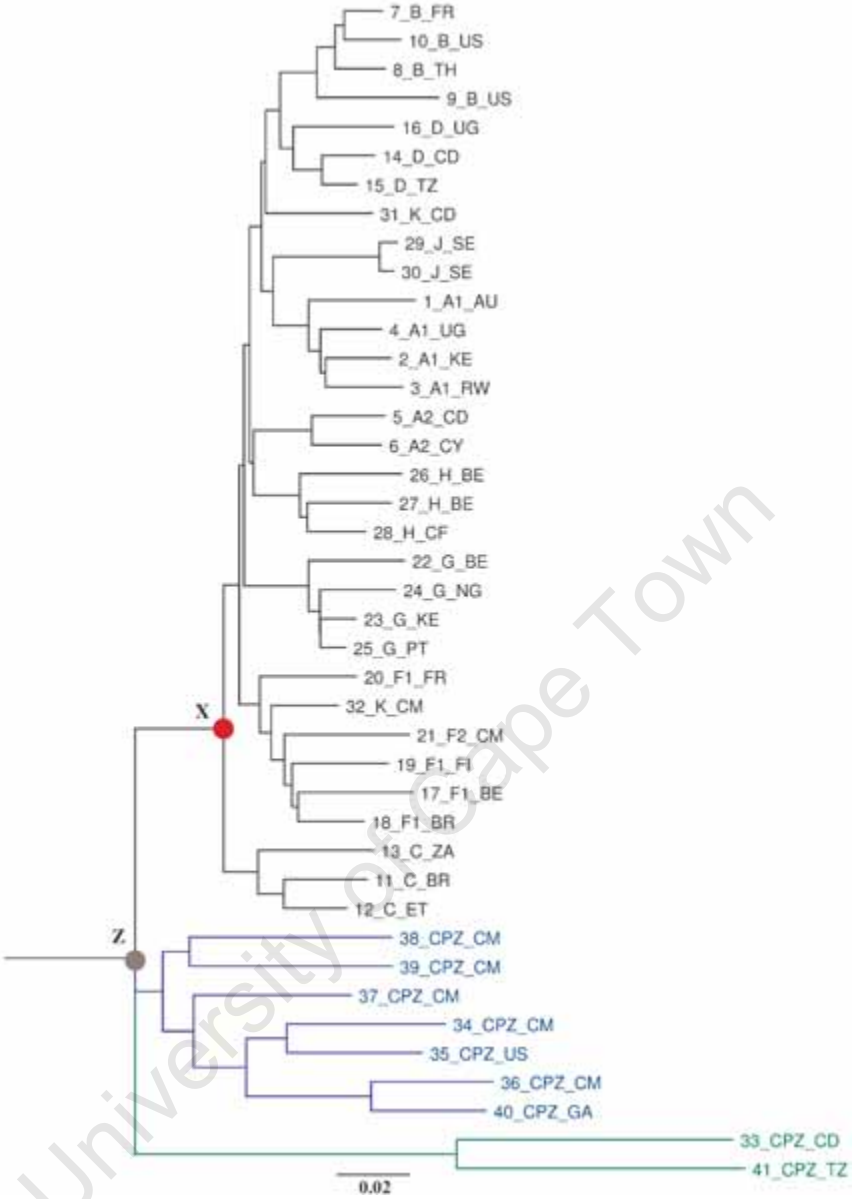


Figure 6.2: Neighbour-joining phylogenetic tree relating the HIV and SIV *gag* gene sequences. The branch between the nodes marked "Z" and "X" indicates where cross-species transmission is thought to have occurred. The black branches lead to HIV sequences, and the blue and green branches to SIV sequences. The green branch represents the outgroup (33_CPZ_CD and 41_CPZ_TZ).

Table 6.3: Number of sites inferred to have switched between rate classes along the branch on which zoonosis is inferred.

Gene	ω^- to ω^N	ω^N to ω^-	ω^+ to ω^N	ω^N to ω^+	TOTAL
gag	6	29	-	-	**35**
pol	3	30	-	-	**33**
vif	-	2	-	-	**2**
vpr	-	1	1	2	**4**
tat	-	2	-	-	**2**
vpu	1	11	-	3	**15**
env	8	29	-	-	**37**
nef	-	1	-	-	**1**
TOTAL	**18**	**105**	**1**	**5**	**129**

A total of 129 sites were inferred to have switched between rate classes along the branch of the tree leading from SIV to the HIV-1 group M sequences. The predominant observed change was from neutral (ω^N) to purifying (ω^-) selection, which accounted for over 80% ($^{105}/_{129}$) of the switches. The next most frequently observed change was in the opposite direction, from neutral to purifying selection ($^{18}/_{129}$). Switching from positive (ω^+) to neutral selection ($^{1}/_{129}$), and from neutral to positive selection ($^{5}/_{129}$) constituted the remainder of the switching sites. No rate switches were observed between the purifying and positive selection rate classes.

It is interesting that all the sites inferred to switch between the positive and neutral selection classes were from two of the accessory HIV-1 genes, *vpr* and *vpu*. The change from neutral to positive selection upon transmission to the human host may reflect adaptation to new immune pressures, for instance, modification to escape CTL responses. For *vpu*, all three ω^N-to-ω^+ switching sites were within known CTL epitopes, which fits the hypothesis of immune escape. In *vpr*, only one of the two sites inferred to switch from ω^N to ω^+, and the single ω^+-to-ω^N switching site, were within known CTL epitopes. This may illustrate varying CTL immune pressures between the two species. However, in the *vpr* and *tat* phylogenetic trees the HIV-1 sequences were not monophyletic, and thus it was not possible to localise the zoonosis to a specific branch (Figure 6.3). This could be the result of the short sequence lengths (210 nt for *vpr* and 117 nt for *tat*), or alternatively for *vpr*, the interspersed

HIV and SIV sequences could potentially be due to the conservation of the *vpr* gene across all the primate lentiviruses (Stivahtis et al., 1997). The numbers of switching sites (Table 6.3) for these two genes were obtained by counting the changes along the branch that lead to the majority of HIV-1 sequences, even though these clades also contained SIV sequences (Figure 6.3). The results for these genes may therefore not present an accurate picture of the change in selective pressure associated with zoonosis.

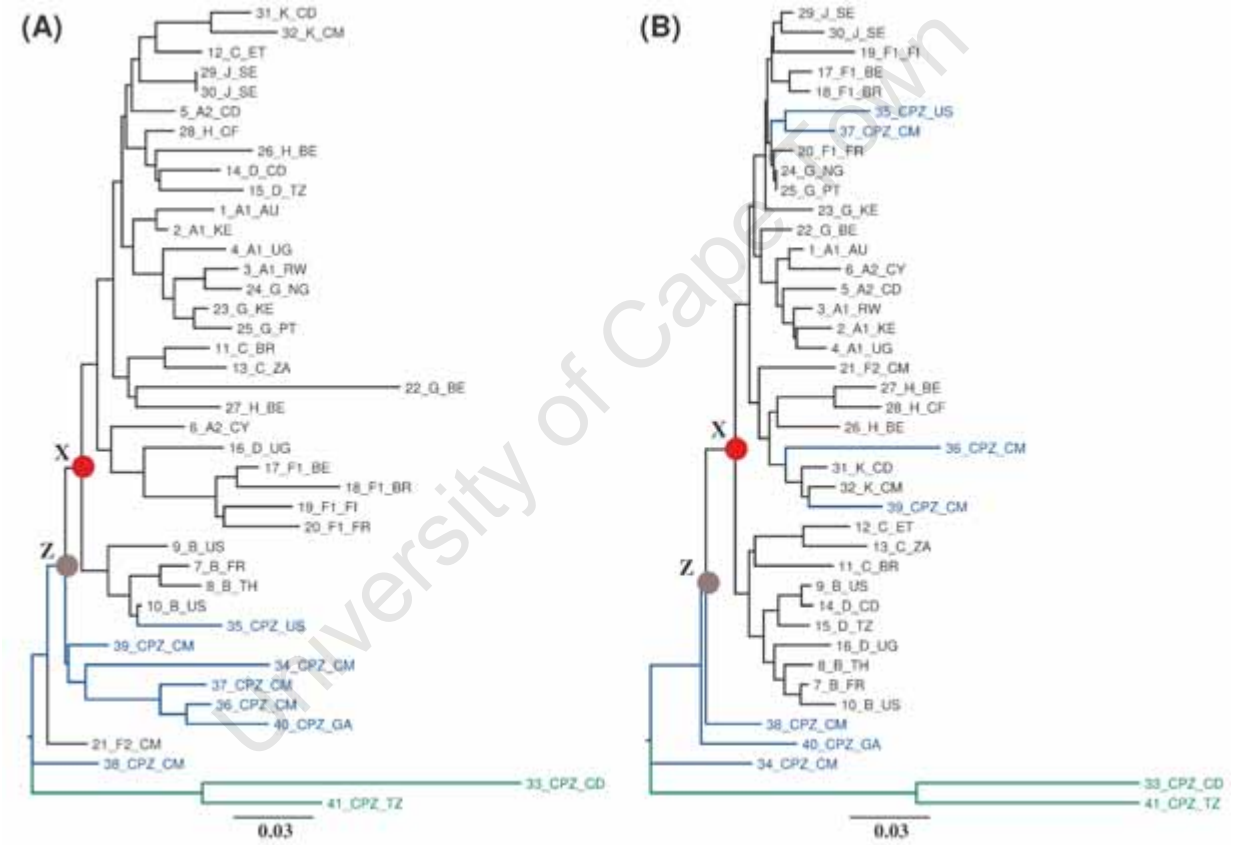


Figure 6.3: Neighbour-joining phylogenetic trees for (A) *tat* and (B) *vpr*. The branches leading to SIV sequences are labelled in blue (or green for the outgroup). The zoonotic event could not be resolved clearly from these trees, and the switches along the branches between points marked "Z" and "X" were therefore counted.

Since recombination can affect positive selection results (Anisimova et al., 2003; Shriner et al., 2003), GARD (Kosakovsky-Pond et al., 2006b) was used to test whether any evidence of

recombination was present in the gene alignments. GARD found evidence for recombination in 7 of the 8 datasets (no recombination breakpoints were found for *tat*). The covarion model assumes that a single phylogenetic tree represents the sequence history of the entire alignment; recombination results in a violation of this assumption and may therefore have affected the results. Although recombination has not specifically been shown to cause false positive inference of switching, it remains a caveat of the covarion, as well as any other, phylogenetic models.

In order to further characterise the rate shifts associated with zoonosis, the proportion of switching sites per gene alignment used in this study (excluding the overlapping regions) was calculated (Table 6.4). The largest percentage of gene-sequence sites inferred to switch between rate classes, was found in the *vpu* gene. Out of 15 switching sites, 11 were inferred to switch from neutral to purifying, 1 from purifying to neutral, and 3 from neutral to positive selection.

Table 6.4: Percentage of inferred switching sites in each of the 8 HIV-1 group M gene sequence datasets analysed.

Gene	Number of Switching Sites	Sequence Length	Percentage of Sequence Sites that Switch
gag	35	1272	2.75%
pol	33	2748	1.20%
vif	2	459	0.44%
vpr	4	210	1.90%
tat	2	117	1.71%
vpu	15	189	7.94%
env	37	2133	1.73%
nef	1	288	0.35%

Among the roles Vpu has during HIV-1 infection, is to down-regulate CD4 glycoprotein levels and increase viral loads (Li et al., 1995; Bour and Strebel, 2003). Furthermore, a study on pig-tailed macaques revealed that defective *vpu* (mutated start codon) reverted to functional form within 16 weeks post-infection, and that this reversion was associated with higher Env and Nef diversity (McCormick-Davis et al., 1998; Bour and Strebel, 2003). McNatt et al. (2009) recently reported a further function of *vpu* as an antagonist of the primate antiviral

protein, tetherin. Tetherin inhibits the release of emerging viruses from HIV-1 infected cells; however, Vpu interferes with the localisation of tetherin to these emerging HIV-1 virions, thereby counteracting the antiviral activity (McNatt et al., 2009). Interestingly, the authors also show that Vpu has adapted independently to retain functionality against the human and chimpanzee forms of tetherin (McNatt et al., 2009).

Collectively, these studies suggest that *vpu* plays a very important part during HIV proliferation within a new host. It is therefore likely that, following a zoonotic episode, the stability of a functional form of this gene is essential for initial survival. However, notwithstanding the potential role of *vpu* after SIV cross-species transmission, the SIV sequences for the *vpu* gene used in this study displayed high variability and could not be aligned with confidence; the poorly aligned areas may therefore have affected the identification of rate switching sites.

The observation that the large majority (~80%) of switching sites involved a change from the neutral to purifying selection class, may illustrate increased evolutionary constraints acting on the virus upon transmission to humans. To evaluate this hypothesis, the proportion of sites within each rate class was tabulated (Table 6.5). The percentage of sites estimated to belong to the purifying selection class was larger in the HIV-1 MRCA than the SIV sequence ancestral to the HIV-1 MRCA, for all genes. This suggests an overall shift towards purifying selection, similar to the findings shown in Table 6.3. Further investigation revealed that the increase in number of sites estimated to be evolving under purifying selection persisted along the branches leading to the HIV descendant sequences (data not shown). These results suggest that, upon entering the human host, HIV-1 protein-coding regions evolved, on average, under greater selective constraints, which resulted in an increase in the proportion of sites evolving under purifying selection.

Table 6.5: Percentage of sites in the HIV-1 MRCA sequence as well as the SIV sequence ancestral to the HIV-1 MRCA (node Z in Figure 6.2), inferred to belong to each rate class.

	Percentage of Sites estimated to belong to each Rate Class (%)					
	Purifying Selection		Neutral Selection		Positive Selection	
Gene	HIV-1 MRCA	Node Z*	HIV-1 MRCA	Node Z*	HIV-1 MRCA	Node Z*
gag	71.70	66.27	28.30	33.73	0	0
pol	82.00	79.32	18.00	20.68	0	0
vif	62.09	60.78	37.91	39.22	0	0
vpr	81.43	80.00	15.71	18.57	2.86	1.43
tat	69.23	64.10	30.77	35.90	0	0
vpu	38.10	22.22	57.14	77.78	4.76	0
env	67.65	64.70	32.35	35.30	0	0
nef	63.54	62.50	36.46	37.50	0	0

*Node Z refers to the node ancestral to the MRCA of the HIV-1 sequences (see Figure 6.2).

The different evolutionary constraints acting on the virus within humans and chimpanzees may be due to the specific immune pressures present in each species. A recent study indicated that CTL responses in humans, directed by specific HLAs, caused adaptation of the virus upon initial cross-species transmission (Ngandu et al., 2009). However, the results obtained in the present study do not provide evidence of sites switching to positive selection. There does, however, appear to be a shift towards a higher rate of purifying selection in humans. The observation that the large majority of rate shifts are towards the purifying selection rate class, further suggests that the evolutionary constraints acting on HIV-1 protein-coding sequences are, on average, greater in humans compared to the constraints affecting the evolution of the SIV ancestral viruses in chimpanzees. This may potentially be due to a narrower, more focussed, immune response in humans than that of chimpanzees. Future work could include simulation studies to estimate the false positive rate of switching between rate classes when a method that incorporates ancestral sequence reconstruction is used.

6.4 Conclusions

The covarion model of sequence evolution is a valuable extension to the more general codon selection models, since identifying sites that switch between different rate classes over time may reflect important adaptations to deviating external pressures. However, using simulations it is shown that apparent covariation (switching between rate classes) may, in some cases, be an artifact of under-fitting of the distribution of rate classes with a discrete null model. If the number of discrete rate classes does not adequately fit the ω distribution, the covarion model of evolution may provide a statistically better fit for the data even though no shifts between rate classes have occurred. A site-by-site model of covarion evolution is shown to overcome the high false positive rate associated with the gene-level model; however, due to the computational burden of the site-specific model, it does not perform well as a feasible alternative method to detect selection class switches. Future work could contribute to the development of a covarion model that provides a practical approach for carrying out gene-level inference of covariation. Future analysis could further include testing the power of the various methods to detect rate shifts, which would provide an indication of the strengths and weaknesses of each method.

Despite the problems with false positive inference of covariation, it was possible to characterise the sites implicated in rate switching after the cross-species jump of SIV to humans. A total of 129 switching-sites were identified, and the most prevalent change was from the neutral to the purifying selection class. The largest proportion of switching sites were identified in the HIV-1 accessory gene, *vpu*, suggesting that the selective pressures acting on this protein differed upon transmission to the new host. The overall proportion of viral sites evolving under purifying selection was higher in humans than chimpanzees, which suggests that greater evolutionary constraints, potentially due to a narrower immune response, acted on the virus after cross-species transmission. Future laboratory research could reveal whether minor changes to the inferred founder HIV sequence affects the fitness of the virus and/or viral replication and proliferation. These results may provide further insight into the evolutionary history of HIV-1 and the sequence characteristics that allowed SIV to infect

humans.

Concluding Remarks

The work carried out in this thesis focussed on the broad topic of HIV evolution, but particular attention was given to the early stages of infection. The sequence changes that occur during this phase, during which the virus first encounters the host immune system are important for the virus' initial survival in the newly infected host. Several mutations, as well as insertions and deletions, may contribute to the overall fitness of the virus. Identifying these changes is important for vaccine and prophylactic drug design, since anti-HIV agents that are effective during the earliest stage of infection could prevent further viral replication and proliferation. Describing the characteristics of early HIV infection is therefore a valuable and important step in the process of establishing effective treatment strategies.

In Chapter 2 a Bayesian method that samples phylogenetic trees relating a sequence alignment, was used to estimate the time to the most recent common ancestor of multiple early infection sequence datasets. The estimated time needed for all the sequences to coalesce to a single founder was compared to the clinical estimate of the time since infection. The comparison between the coalescent estimate and the laboratory staging allowed for classification of 171 early infection isolates into either homogeneous or heterogeneous infections. Identifying the actual virus that is transmitted provides valuable insight into the number of viruses that successfully enter a new host and cause productive infection. Furthermore, identifying the founder allows for further investigation of the sequence characteristics of the virus that overcomes the barriers encountered during transmission. These sequence features may be essential for initial viral survival and may present important targets for vaccine and drug design. This was the aim of Chapter 3, where the HIV subtype B homogeneous

infections (n = 81) identified in the first chapter were combined to form a single dataset representative of the earliest stages of infection. If similar mutations are present across a large sample of virus's isolated during early infection, then the changes may illustrate characteristic evolutionary patterns essential for early survival of the virus. Positive selection analysis was carried out on the homogeneous dataset and 24 adaptively evolving sites were identified. Further investigation suggested that early evolution of HIV frequently selects for changes that lead to CTL immune escape, and that APOBEC3 is likely to play a role in the dynamics of this early immune evasion. The results from this chapter provide supporting evidence that concentrated regions of diversification are associated with CTL escape during early infection.

Chapter 4 was aimed at describing further sequence characteristics of HIV that are specifically features of early HIV infection. Previous publications have reported that the Env variable loop lengths of viruses isolated from early infection are shorter and contain fewer potential N-linked glycosylation sites that those present in chronic infection. However, the shared history of the sequences was often not accounted for and this is shown to be a potential source of bias. A method that accounts for the phylogenetic relatedness of the sequences was developed in this chapter. The method was applied to a subtype B datasets previously reported to contain significant differences between the variable loop length of early and chronic infection sequences. Although the results were confirmed with the method that accounts for the shared history of the viruses, the statistical evidence for the results suggesting a significant difference, was more marginal than the originally reported result. Further analysis revealed an association between the variable loop length and stage of disease. The results from this chapter support previous reports illustrating that changes in the *env* variable loop length and number of PNGSs are a characteristic of early HIV infection.

The focus of Chapter 5 was to evaluate the impact of recombination on previously published analyses of adaptive evolution in HIV-1 protein-coding sequences. Reanalysis was carried out using a method that takes recombination into account, since recombination can cause a significant bias in positive selection results. If recombining sequences are present in a dataset, a single bifurcating phylogenetic tree cannot be used to represent the evo-

lutionary history of the sequences. This is often not accounted for by the evolutionary models used to infer positive selection in coding sequence, which may result in a high rate of false positive inference of adaptive evolution. In this chapter a method that accounts for recombination (PARRIS) was applied to several previously analysed HIV-1 datasets and the results suggested that recombination has a marked effect on positive selection results; PARRIS provided more conservative estimates for the evidence of positive selection, and far fewer sites were identified in the site-specific analysis. PARRIS provides results that include fewer false positives than methods that do not account for recombination; furthermore, the results obtained in this chapter included only a proportion of the sites originally reported to be evolving adaptively. Collectively, these findings suggest that the site-specific results estimated by PARRIS present a more accurate set of sites undergoing positive selection than those reported in the original publications. These results are very promising for future positive selection analysis of recombining coding sequences.

The aim of Chapter 6 was twofold; the covarion model of sequence evolution was evaluated to investigate the extent of false positive inference of rate class switching, and the covariation of sites associated with the zoonotic episode, during which SIV successfully infected the human species, was estimated. High false positive results were found for the gene-level approach to model covarion evolution. Since many previously published studies focussing on selection class switches have used the same covarion model, the validity of the reported results may be questionable, and the conclusions may need revisiting. To address the second aim of the chapter, ancestral sequences representing the evolutionary history of a group of SIVcpz and HIV-1 group M sequences, were reconstructed. The implementation of the covarion evolutionary model allowed for the identification of sites that shifted between selection classes along the branch leading from SIV to the most recent common HIV-1 ancestor. The main switch occurred between the neutral and purifying selection class, which accounted for more than 80% of the individual site switches. Furthermore, an increased proportion of viral sites was estimated to be evolving under purifying selection after SIVcpz transmission to humans. These results suggest that the evolutionary constraints acting on the virus present in chimpanzees and humans are different, and that these constraints are, on average, greater in humans.

The analysis carried out in this thesis showed that, although extremely useful for evolutionary and comparative research, computational models that fail to account for key features of HIV-1 evolution may lead to invalid conclusions. Several previous studies that aimed to describe HIV evolution therefore need revisiting. These include phylogenetic models of positive selection that neglected the effects of recombination, comparisons of sets of sequences (for example isolated from early versus chronically infected patients) that did not take into account the effects of phylogenetic linkage, as well as published reports on HIV covariation. Methods that account for these features could lead to more accurate results, thereby providing a more focussed set of amino acid sites potentially implicated in HIV adaptation and escape from antiviral drugs.

The high degree of HIV diversity is currently one of the main obstacles to developing a broadly-effective vaccine. Characterising the transmitted virus during the earliest stages of infection is therefore an important step towards designing effective antiviral compounds. In this thesis it was possible to identify infections caused by a single viral strain in a large group of recently infected individuals by using a Bayesian approach together with available clinical data. This amalgamated homogeneous dataset provided, for the first time, the opportunity to characterise the selective pressures present during the earliest period of infection. A detailed analysis of this founder virus dataset showed that HIV adapts to evade CTL immune responses, and that APOBEC3 may be implicated during early immune escape. These results further our understanding of the nature of HIV evolution, and provide valuable insight into future endeavours to develop vaccines and antiviral drugs.

Bibliography

Abecasis AB, Lemey P, Vidal N, de Oliveira T, Peeters M, Camacho R, Shapiro B, Rambaut A, Vandamme AM (2007) Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: Subtype g is a circulating recombinant form. *J. Virol.* 81:8543–8551. 22

Abrahams MR, Anderson JA, Giorgi EE, Seoighe C, Mlisana K, Ping LH, Athreya GS, Treurnicht FK, Keele BF, Wood N, Salazar-Gonzalez JF, Bhattacharya T, Chu H, Hoffman I, Galvin S, Mapanje C, Kazembe P, Thebus R, Fiscus S, Hide W, Cohen MS, Karim SA, Haynes BF, Shaw GM, Hahn BH, Korber BT, Swanstrom R, Williamson C, for the CAPRISA Acute Infection Study Team, the Center for HIV-AIDS Vaccine Immunology Consortium (2009) Quantitating the multiplicity of infection with hiv-1 subtype c reveals a non-poisson distribution of transmitted variants. *J Virol* 83:3556–3576. 21, 33, 34, 39, 40, 41, 43, 46, 49, 56, 57, 60, 61

Allen TM, O'Connor DH, Jing P, Dzuris JL, Mothé BR, Vogel TU, Dunphy E, Liebl ME, Emerson C, Wilson N, Kunstman KJ, Wang X, Allison DB, Hughes AL, Desrosiers RC, Altman JD, Wolinsky SM, Sette A, Watkins DI (2000) Tat-specific cytotoxic t lymphocytes select for siv escape variants during resolution of primary viraemia. *Nature* 407:386–390. 83

Allen TM, Altfeld M, Geer SC, Kalife ET, Moore C, O'sullivan KM, Desouza I, Feeney ME, Eldridge RL, Maier EL, Kaufmann DE, Lahaie MP, Reyor L, Tanzi G, Johnston MN, Brander C, Draenert R, Rockstroh JK, Jessen H, Rosenberg ES, Mallal SA, Walker BD (2005) Selective escape from cd8+ t-cell responses represents a major driving force of human immunodeficiency virus type 1 (hiv-1) sequence diversity and reveals constraints on hiv-1 evolution. *J Virol* 79:13239–13249. 84

Allen TM, Altfeld M, Yu XG, O'Sullivan KM, Lichterfeld M, Gall SL, John M, Mothe BR, Lee PK, Kalife ET, Cohen DE, Freedberg KA, Strick DA, Johnston MN, Sette A, Rosenberg ES, Mallal SA, Goulder PJR, Brander C, Walker BD (2004) Selection, transmission, and reversion of an antigen-processing cytotoxic t-lymphocyte escape mutation in human immunodeficiency virus type 1 infection. *J Virol* 78:7069–7078. 84

Allman ES, Rhodes JA (2009) The identifiability of covarion models in phylogenetics. *IEEE/ACM Trans Comput Biol Bioinform* 6:76–88. 133

Andrieu C, de Freitas N, Doucet A, Jordan MI (2003) An introduction to mcmc for machine learning. *Machine Learning* 50:5–43. 36

Anisimova M, Bielawski JP, Yang Z (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol* 18:1585–1592. 9

Anisimova M, Kosiol C (2009) Investigating Protein-Coding Sequence Evolution with Probabilistic Codon Substitution Models. *Mol Biol Evol* 26:255–271. 10

Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164:1229–1236. 13, 22, 87, 94, 109, 110, 128, 145

Aris-Brosou S, Yang Z (2002) Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18s ribosomal rna phylogeny. *Systematic Biology,* 51:703–714. 38, 39

Arnaout RA, Lloyd AL, O'Brien TR, Goedert JJ, Leonard JM, Nowak MA (1999) A simple relationship between viral load and survival time in hiv-1 infection. *Proc Natl Acad Sci U S A* 96:11549–11553. 103

Asquith B, McLean AR (2007) In vivo cd8+ t cell control of immunodeficiency virus infection in humans and macaques. *Proc Natl Acad Sci U S A* 104:6365–6370. 83

Bello G, Guimaraes ML, Passaes CP, Almeida SEM, Veloso VG, Morgado MG (2009) Short communication: Evidences of recent decline in the expansion rate of the hiv type 1 subtype c and crf31-bc epidemics in southern brazil. *AIDS Research and Human Retroviruses* 25:1065–1069 PMID: 19895209. 43

Bennett DE, Camacho RJ, Otelea D, Kuritzkes DR, Fleury H, Kiuchi M, Heneine W, Kantor R, Jordan MR, Schapiro JM, Vandamme AM, Sandstrom P, Boucher CAB, van de Vijver D, Rhee SY, Liu TF, Pillay D, Shafer RW (2009) Drug resistance mutations for surveillance of transmitted hiv-1 drug-resistance: 2009 update. *PLoS One* 4:e4724. 32

Berkhout B, de Ronde A (2004) Apobec3g versus reverse transcriptase in the generation of hiv-1 drug-resistance mutations. *AIDS* 18:1861–1863. 27

Bhattacharya T, Daniels M, Heckerman D, Foley B, Frahm N, Kadie C, Carlson J, Yusim K, McMahon B, Gaschen B, Mallal S, Mullins JI, Nickle DC, Herbeck J, Rousseau C, Learn GH, Miura T, Brander C, Walker B, Korber B (2007) Founder effects in the assessment of hiv polymorphisms and hla allele associations. *Science* 315:1583–1586. 102

Bielawski JP, Yang Z (2005) *Maximum Likelihood methods for Detecting Adaptive Protein Evolution*, chapter 5, pp. 103–124 Springer. 12

Blay WM, Gnanakaran S, Foley B, Doria-Rose NA, Korber BT, Haigwood NL (2006) Consistent patterns of change during the divergence of human immunodeficiency virus type 1 envelope from that of the inoculated virus in simian/human immunodeficiency virus-infected macaques. *J Virol* 80:999–1014. 23

Blish CA, Nedellec R, Mandaliya K, Mosier DE, Overbaugh J (2007) Hiv-1 subtype a envelope variants from early in infection have variable sensitivity to neutralization and to inhibitors of viral entry. *AIDS* 21:693–702. 23

Borrow P, Bhardwaj N (2008) Innate immune responses in primary hiv-1 infection. *Curr Opin HIV AIDS* 3:36–44. 25

Bour S, Strebel K (2003) The hiv-1 vpu protein: a multifunctional enhancer of viral particle release. *Microbes and Infection* 5:1029 – 1039. 146

Brennan CA (2007) Review of status of hiv strain diversity in the united states. *Journal of Medical Virology* 79:S27–S31. 138

Brites C, Sampalo J, Oliveira A (2009) Hiv/human t-cell lymphotropic virus coinfection revisited: impact on aids progression. *AIDS Rev* 11:8–16. 18

Brès V, Kiernan R, Emiliani S, Benkirane M (2002) Tat acetyl-acceptor lysines are important for human immunodeficiency virus type-1 replication. *J Biol Chem* 277:22215–22221. 127

Brumme ZL, Brumme CJ, Carlson J, Streeck H, John M, Eichbaum Q, Block BL, Baker B, Kadie C, Markowitz M, Jessen H, Kelleher AD, Rosenberg E, Kaldor J, Yuki Y, Carrington M, Allen TM, Mallal S, Altfeld M, Heckerman D, Walker BD (2008) Marked epitope- and allele-specific differences in rates of mutation in human immunodeficiency type 1 (hiv-1) gag, pol, and nef cytotoxic t-lymphocyte epitopes in acute/early hiv-1 infection. *J Virol* 82:9216–9227. 30

Bunce M (2003) Pcr-sequence-specific primer typing of hla class i and class ii alleles. *Methods Mol Biol* 210:143–171. 70

Bunnik EM, Pisas L, van Nuenen AC, Schuitemaker H (2008) Autologous neutralizing humoral immunity and evolution of the viral envelope in the course of subtype b human immunodeficiency virus type 1 infection. *J Virol* 82:7932–7941. 91

Burke DS (1997) Recombination in hiv: an important viral evolutionary strategy. *Emerg Infect Dis* 3:253–259. 109, 131

Burmester GR, Pezzutto A, Ulrichs T, Aicher A (2003) *Color atlas of immunology* Thieme. 18

Burton DR, Stanfield RL, Wilson IA (2005) Antibody vs. hiv in a clash of evolutionary titans. *Proc Natl Acad Sci U S A* 102:14943–14948. 23

Buschbom J, von Haeseler A (2005) *Introduction to Applications of the Likelihood Function in Molecular Evolution*, chapter 2, pp. 25–44 Springer. 11

Butler IF, Pandrea I, Marx PA, Apetrei C (2007) Hiv genetic diversity: biological and public health consequences. *Curr HIV Res* 5:23–45. 138

Carlson J, Kadie C, Mallal S, Heckerman D (2007) Leveraging hierarchical population structure in discrete association studies. *PLoS One* 2:e591. 101, 102

Carrington M, Nelson GW, Martin MP, Kissner T, Vlahov D, Goedert JJ, Kaslow R, Buchbinder S, Hoots K, O'Brien SJ (1999) Hla and hiv-1: heterozygote advantage and b*35-cw*04 disadvantage. *Science* 283:1748–1752. 30

Castro KG, Ward JW, Slutsker L, Buehler JW, Jaffe HW, Berkelman RL (1993) Revised classification system for hiv infection and expanded surveillance case definition for aids among adolescents and adults. *Morbid Mortal Wkly Rep* 41:1–19. 104

Chan DJ (2004) Hiv-1 superinfection: evidence and impact. *Curr HIV Res* 2:271–274. 109

Chib S, Greenberg E (1995) Understanding the metropolis-hastings algorithm. *The American Statistician* 49:327–335. 37

Chiu YL, Greene WC (2008) The apobec3 cytidine deaminases: an innate defensive network opposing exogenous retroviruses and endogenous retroelements. *Annu Rev Immunol* 26:317–353. 26, 27

Chohan B, Lang D, Sagar M, Korber B, Lavreys L, Richardson B, Overbaugh J (2005a) Selection for human immunodeficiency virus type 1 envelope glycosylation variants with shorter v1-v2 loop sequences occurs during transmission of certain genetic subtypes and may impact viral rna levels. *J Virol* 79:6528–6531. 91, 92, 93, 94, 98, 100, 101, 102, 106, 107

Chohan B, Lavreys L, Rainwater SMJ, Overbaugh J (2005b) Evidence for frequent reinfection with human immunodeficiency virus type 1 of a different subtype. *J Virol* 79:10701–10708. 21, 24

Chopera DR, Woodman Z, Mlisana K, Mlotshwa M, Martin DP, Seoighe C, Treurnicht F, de Rosa DA, Hide W, Karim SA, Gray CM, Williamson C, Team CAPRISAS (2008) Transmission of hiv-1 ctl escape variants provides hla-mismatched recipients with a survival advantage. *PLoS Pathog* 4:e1000033. 22, 30, 31, 64

Christensen N, Dupuis RJ, Woan G, Meyer R (2004) Metropolis-hastings algorithm for extracting periodic gravitational wave signals from laser interferometric detector data. *Phys. Rev. D* 70:022001. 37

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, Ferriera S, Wang G, Zheng X, White TJ, Sninsky JJ, Adams MD, Cargill M (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302:1960–1963. 133

Clavel F, Hance AJ (2004) Hiv drug resistance. *N Engl J Med* 350:1023–1035. 32

Cornelissen M, Hoogland F, Back N, Jurriaans S, Zorgdrager F, Bakker M, Brinkman K, Prins M, van der Kuyl A (2009) Multiple transmissions of a stable human leucocyte antigen-b27 cytotoxic t-cell-escape strain of hiv-1 in the netherlands. *AIDS* . 30

Crawford H, Lumm W, Leslie A, Schaefer M, Boeras D, Prado JG, Tang J, Farmer P, Ndung'u T, Lakhi S, Gilmour J, Goepfert P, Walker BD, Kaslow R, Mulenga J, Allen S, Goulder PJR, Hunter E (2009) Evolution of hla-b*5703 hiv-1 escape mutations in hla-b*5703-positive individuals and their transmission recipients. *J Exp Med* 206:909–921. 30

Crawford H, Prado JG, Leslie A, Hué S, Honeyborne I, Reddy S, van der Stok M, Mncube Z, Brander C, Rousseau C, Mullins JI, Kaslow R, Goepfert P, Allen S, Hunter E, Mulenga J, Kiepiela P, Walker BD, Goulder PJR (2007) Compensatory mutation partially restores fitness and delays reversion of escape mutation within the immunodominant hla-b*5703-restricted gag epitope in chronic human immunodeficiency virus type 1 infection. *J Virol* 81:8346–8351. 30

Danilova N (2008) *Encyclopedia of Life Sciences (ELS)*, chapter Evolution of the Human Immune System John Wiley & Sons, Ltd: Chichester. 134

Darwin C (1859) *On the Origin of Species* John Murray, London. 4

de Clercq E (2009) Anti-hiv drugs: 25 compounds approved within 25 years after the discovery of hiv. *Int J Antimicrob Agents* 33:307–320. 22

de Oliveira T, Salemi M, Gordon M, Vandamme AM, van Rensburg EJ, Engelbrecht S, Coovadia HM, Cassol S (2004) Mapping sites of positive selection and amino acid diversification in the hiv genome: an alternative approach to vaccine design? *Genetics* 167:1047–1058. xxi, 114, 117, 120, 121, 122, 128, 129, 131

Delport W, Scheffler K, Seoighe C (2008) Frequent toggling between alternative amino acids is driven by selection in hiv-1. *PLoS Pathog* 4:e1000242. 30

Delport W, Scheffler K, Seoighe C (2009) Models of coding sequence evolution. *Brief Bioinform* 10:97–109. 7, 12, 13, 14, 67

Delwart E, Magierowska M, Royz M, Foley B, Peddada L, Smith R, Heldebrant C, Conrad A, Busch M (2002) Homogeneous quasispecies in 16 out of 17 individuals during very early hiv-1 primary infection. *AIDS* 16:189–195. 33

Derdeyn CA, Decker JM, Bibollet-Ruche F, Mokili JL, Muldoon M, Denham SA, Heil ML, Kasolo F, Musonda R, Hahn BH, Shaw GM, Korber BT, Allen S, Hunter E (2004) Envelope-constrained neutralization-sensitive hiv-1 after heterosexual transmission. *Science* 303:2019–2022. 23, 29, 33, 60, 61, 75, 91, 107

Dhillon AK, Donners H, Pantophlet R, Johnson WE, Decker JM, Shaw GM, Lee FH, Richman DD, Doms RW, Vanham G, Burton DR (2007) Dissecting the neutralizing antibody specificities of broadly neutralizing sera from human immunodeficiency virus type 1-infected donors. *J Virol* 81:6548–6562. 28

Ding J, Chou YY, Chang T (2009) Defensins in viral infections. *J Innate Immun* 1:413–420. 26

Doms RW, Moore JP (2000) Hiv-1 membrane fusion: targets of opportunity. *J Cell Biol* 151:F9–14. 23

Dorman KS (2007) Identifying dramatic selection shifts in phylogenetic trees. *BMC Evol Biol* 7 Suppl 1:S10. 133

Drosg M (2007) *Dealing with uncertainties: a guide to error analysis* Springer. 40, 41

Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* 22:1185–1192. 37, 44

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4:e88. 39, 42

Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161:1307–1320. 37

Drummond AJ, Rambaut A (2007) Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7:214. 36, 37, 42, 43

Edwards CTT, Holmes EC, Wilson DJ, Viscidi RP, Abrams EJ, Phillips RE, Drummond AJ (2006) Population genetic estimation of the loss of genetic diversity during horizontal transmission of hiv-1. *BMC Evol Biol* 6:28. 43, 75

Edwards RJ, Shields DC (2004) Gasp: Gapped ancestral sequence prediction for proteins. *BMC Bioinformatics* 5:123. 96

Ekene OC (2008) Cytokines and hiv/aids: a critical look at the existing relationship between them. *Roum Arch Microbiol Immunol* 67:67–80. 25

Felsenstein J (1981a) Evolutionary trees from dna sequences: a maximum likelihood approach. *J Mol Evol* 17:368–376. 6

Felsenstein J (1981b) Evolutionary trees from gene frequencies and quantitative characters: Finding maximum likelihood estimates. *Evolution* 35:1229–1242. 94, 101

Fernandez CS, Stratov I, Rose RD, Walsh K, Dale CJ, Smith MZ, Agy MB, Hu SL, Krebs K, Watkins DI, O'connor DH, Davenport MP, Kent SJ (2005) Rapid viral escape at an immunodominant simian-human immunodeficiency virus cytotoxic t-lymphocyte epitope exacts a dramatic fitness cost. *J Virol* 79:5721–5731. 83

Fiebig EW, Wright DJ, Rawal BD, Garrett PE, Schumacher RT, Peddada L, Heldebrant C, Smith R, Conrad A, Kleinman SH, Busch MP (2003) Dynamics of hiv viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary hiv infection. *AIDS* 17:1871–1879. xx, 20, 21, 34, 40, 46, 64, 66

Fortis C, Poli G (2005) Dendritic cells and natural killer cells in the pathogenesis of hiv infection. *Immunol Res* 33:1–21. 24

Frahm N, Baker B, Brander C (2008) Identification and optimal definition of hiv-derived cytotoxic t-lymphocyte (ctl) epitopes for the study of ctl escape, functional avidity and viral evolution In BT K, C B, BF H, R K, JP M, BD W, DI W, editors, *HIV Molecular Immunology 2008.*, pp. 3–24. Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico. LA-UR 08-05096. 87

Friedrich TC, Dodds EJ, Yant LJ, Vojnov L, Rudersdorf R, Cullen C, Evans DT, Desrosiers RC, Mothé BR, Sidney J, Sette A, Kunstman K, Wolinsky S, Piatak M, Lifson J, Hughes AL, Wilson N, O'Connor DH, Watkins DI (2004) Reversion of ctl escape-variant immunodeficiency viruses in vivo. *Nat Med* 10:275–281. 64, 84

Frost LS, Leplae R, Summers AO, Toussaint A (2005) Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol* 3:722–732. 86

Frost SD, Günthard HF, Wong JK, Havlir D, Richman DD, Brown AJL (2001) Evidence for positive selection driving the evolution of hiv-1 env under potent antiviral therapy. *Virology* 284:250–258. 131

Fu YX, Li WH (1999) Coalescing into the 21st century: An overview and prospects of coalescent theory. *Theor Popul Biol* 56:1–10. 34, 35, 36

Furuse Y, Suzuki A, Kamigaki T, Oshitani H (2009) Evolution of the m gene of the influenza a virus in different host species: large-scale sequence analysis. *Virology Journal* 6:67. 133

Gagneux P, Varki A (2001) Genetic differences between humans and great apes. *Molecular Phylogenetics and Evolution* 18:2 – 13. 134

Galtier N (2001) Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol Biol Evol* 18:866–873. 15, 16, 133

Galtier N, Gascuel O, Jean-Marie A (2005) *Statistical Methods in Molecular Evolution*, chapter Markov Models in Molecular Evolution, pp. 3–44 Springer. 5, 6, 7, 9, 11

Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, Cummins LB, Arthur LO, Peeters M, Shaw GM, Sharp PM, Hahn BH (1999) Origin of hiv-1 in the chimpanzee pan troglodytes troglodytes. *Nature* 397:436–441. 18, 134

Gao F, Korber BT, Weaver E, Liao HX, Hahn BH, Haynes BF (2004) Centralized immunogens as a vaccine strategy to overcome hiv-1 diversity. *Expert Rev Vaccines* 3:S161–S168. 117

Gao F, Weaver EA, Lu Z, Li Y, Liao HX, Ma B, Alam SM, Scearce RM, Sutherland LL, Yu JS, Decker JM, Shaw GM, Montefiori DC, Korber BT, Hahn BH, Haynes BF (2005) Antigenicity and immunogenicity of a synthetic human immunodeficiency virus type 1 group m consensus envelope glycoprotein. *J Virol* 79:1154–1163. 19, 23

Gaucher EA, Miyamoto MM, Benner SA (2001) Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. *Proceedings of the National Academy of Sciences of the United States of America* 98:548–552. 133, 135

Gilks WR, Roberts GO, Sahu SK (1998) Adaptive markov chain monte carlo through regeneration. *Journal of the American Statistical Association* 93:1045–1054. 36

Goepfert PA (2003) Making sense of the hiv immune response. *Top HIV Med* 11:4–8. 24

Goila-Gaur R, Strebel K (2008) Hiv-1 vif, apobec, and intrinsic immunity. *Retrovirology* 5:51. 27

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding dna sequences. *Mol Biol Evol* 11:725–736. 8, 68

Goonetilleke N, Moore S, Dally L, Winstone N, Cebere I, Mahmoud A, Pinheiro S, Gillespie G, Brown D, Loach V, Roberts J, Guimaraes-Walker A, Hayes P, Loughran K, Smith C, Bont JD, Verlinde C, Vooijs D, Schmidt C, Boaz M, Gilmour J, Fast P, Dorrell L, Hanke T, McMichael AJ (2006) Induction of multifunctional human immunodeficiency virus type 1 (hiv-1)-specific t cells capable of proliferation in healthy subjects by using a prime-boost regimen of dna- and modified vaccinia virus ankara-vectored vaccines expressing hiv-1 gag coupled to cd8+ t-cell epitopes. *J Virol* 80:4717–4728. 69

Gorny MK, Stamatatos L, Volsky B, Revesz K, Williams C, Wang XH, Cohen S, Staudinger R, Zolla-Pazner S (2005) Identification of a new quaternary neutralizing epitope on human immunodeficiency virus type 1 virus particles. *J Virol* 79:5232–5237. 28, 29

Gottlieb GS, Nickle DC, Jensen MA, Wong KG, Grobler J, Li F, Liu SL, Rademeyer C, Learn GH, Karim SSA, Williamson C, Corey L, Margolick JB, Mullins JI (2004) Dual hiv-1 infection associated with rapid disease progression. *Lancet* 363:619–622. 21, 22, 60

Gottlieb MS, Schroff R, Schanker HM, Weisman JD, Fan PT, Wolf RA, Saxon A (1981) Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. *N Engl J Med* 305:1425–1431. 17

Goulder P, Price D, Nowak M, Rowland-Jones S, Phillips R, McMichael A (1997) Co-evolution of human immunodeficiency virus and cytotoxic t-lymphocyte responses. *Immunol Rev* 159:17–29. 83

Goulder PJ, Bunce M, Krausa P, McIntyre K, Crowley S, Morgan B, Edwards A, Giangrande P, Phillips RE, McMichael AJ (1996) Novel, cross-restricted, conserved, and immunodominant cytotoxic t lymphocyte epitopes in slow progressors in hiv type 1 infection. *AIDS Res Hum Retroviruses* 12:1691–1698. 22

Goulder PJ, Phillips RE, Colbert RA, McAdam S, Ogg G, Nowak MA, Giangrande P, Luzzi G, Morgan B, Edwards A, McMichael AJ, Rowland-Jones S (1997) Late escape from an immunodominant cytotoxic t-lymphocyte response associated with progression to aids. *Nat Med* 3:212–217. 83

Griffiths RC, Marjoram P (1996) Ancestral inference from samples of dna sequences with recombination. *Journal of Computational Biology* 3:479–502. 36

Grobler J, Gray CM, Rademeyer C, Seoighe C, Ramjee G, Karim SA, Morris L, Williamson C (2004) Incidence of hiv-1 dual infection and its association with increased viral load set point in a cohort of hiv-1 subtype c-infected female sex workers. *J Infect Dis* 190:1355–1359. 21, 22

Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP (2004) Modeling the site-specific variation of selection patterns along lineages. *Proc Natl Acad Sci U S A* 101:12957–12962. 133, 135, 136, 141

Haaland RE, Hawkins PA, Salazar-Gonzalez J, Johnson A, Tichacek A, Karita E, Manigart O, Mulenga J, Keele BF, Shaw GM, Hahn BH, Allen SA, Derdeyn CA, Hunter E (2009) Inflammatory genital infections mitigate a severe genetic bottleneck in heterosexual transmission of subtype a and c hiv-1. *PLoS Pathog* 5:e1000274. 21

Hanada K, Suzuki Y, Gojobori T (2004) A large variation in the rates of synonymous substitution for rna viruses and its relationship to a diversity of viral infection and transmission modes. *Mol Biol Evol* 21:1074–1080. 14, 132

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *J Mol Evol* 22:160–174. 6, 111, 113

Haynes BF, Montefiori DC (2006) Aiming to induce broadly reactive neutralizing antibody responses with hiv-1 vaccine candidates. *Expert Rev Vaccines* 5:579–595. 28

Herbeck JT, Nickle DC, Learn GH, Gottlieb GS, Curlin ME, Heath L, Mullins JI (2006) Human immunodeficiency virus type 1 env evolves toward ancestral states upon transmission to a new host. *J Virol* 80:1637–1644. 84

Hessell AJ, Rakasz EG, Poignard P, Hangartner L, Landucci G, Forthal DN, Koff WC, Watkins DI, Burton DR (2009) Broadly neutralizing human anti-hiv antibody 2g12 is effective in protection against mucosal shiv challenge even at low serum neutralizing titers. *PLoS Pathog* 5:e1000433. 28

Ho SYW, Phillips MJ, Drummond AJ, Cooper A (2005) Accuracy of rate estimation using relaxed-clock models with a critical focus on the early metazoan radiation. *Mol Biol Evol* 22:1355–1363. 38, 39

Holdsworth R, Hurley CK, Marsh SGE, Lau M, Noreen HJ, Kempenich JH, Setterholm M, Maiers M (2009) The hla dictionary 2008: a summary of hla-a, -b, -c, -drb1/3/4/5, and -dqb1 alleles and their association with serologically defined hla-a, -b, -c, -dr, and -dq antigens. *Tissue Antigens* 73:95–170. 30

Hollier MJ, Dimmock NJ (2005) The c-terminal tail of the gp41 transmembrane envelope glycoprotein of hiv-1 clades a, b, c, and d may exist in two conformations: an analysis of sequence, structure, and function. *Virology* 337:284–296. 23

Holmes EC (2003) Molecular clocks and the puzzle of rna virus origins. *J Virol* 77:3893–3897. 133

Holzmayer V, Aitken C, Skinner C, Ryall L, Devare SG, Hackett J (2009) Characterization of genetically diverse hiv type 1 from a london cohort: near full-length genomic analysis of a subtype h strain. *AIDS Res Hum Retroviruses* 25:721–726. 138

Hong PWP, Flummerfelt KB, de Parseval A, Gurney K, Elder JH, Lee B (2002) Human immunodeficiency virus envelope (gp120) binding to dc-sign and primary dendritic cells is carbohydrate dependent but does not involve 2g12 or cyanovirin binding sites: implications for structural analyses of gp120-dc-sign binding. *J Virol* 76:12855–12865. 26

Hothorn T, Hornik K, van de Wiel M, Zeileis A (2006) A lego system for conditional inference. *American Statistician* 60:257 – 263. 95, 103

Hué S, Pillay D, Clewley JP, Pybus OG (2005) Genetic analysis reveals the complex structure of hiv-1 transmission within defined risk groups. *Proc Natl Acad Sci U S A* 102:4425–4429. 34

Huang DD, Giesler TA, Bremer JW (2003) Sequence characterization of the protease and partial reverse transcriptase proteins of the ned panel, an international hiv type 1 subtype reference and standards panel. *AIDS Research and Human Retroviruses* 19:321–328. 138

Hudson RR, Kaplan NL (1988) The coalescent process in models with selection and recombination. *Genetics* 120:831–840. 36

Hudson RR (1991) *Oxford Surveys in Evolutionary Biology*, Vol. 7, chapter Gene Genealogies and the Coalescent Process, pp. 1–44 Oxford University Press. 36

Huelsenbeck JP (2002) Testing a covariotide model of dna substitution. *Mol Biol Evol* 19:698–707. 15, 16, 133, 136

Huelsenbeck JP, Crandall KA (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu Rev Evol Syst* 28:437–466. 5, 6

Huelsenbeck JP, Dyer KA (2004) Bayesian estimation of positively selected sites. *J Mol Evol* 58:661–672. 12

Hughes AL (1999) *Adaptive Evolution of Genes and Genomes* Oxford University Press. 8

Hughes GJ, Paez A, Boshell J, Rupprecht CE (2004) A phylogenetic reconstruction of the epidemiological history of canine rabies virus variants in colombia. *Infection, Genetics and Evolution* 4:45 – 51. 43

Huson DH, Bryant D (2006) Application of Phylogenetic Networks in Evolutionary Studies. *Mol Biol Evol* 23:254–267. 124

Huzurbazar S, Kolesov G, Massey SE, Harris KC, Churbanov A, Liberles DA (2010) Lineage-specific differences in the amino acid substitution process. *Journal of Molecular Biology* 396:1410 – 1421. 124

Itoh Y, Mahmoud HM (2005) Age statistics in the moran population model. *Statistics & Probability Letters* 74:21 – 30. 36

Jern P, Russell RA, Pathak VK, Coffin JM (2009) Likely role of apobec3g-mediated g-to-a mutations in hiv-1 evolution and drug resistance. *PLoS Pathog* 5:e1000367. 26, 27

Jetzt AE, Yu H, Klarmann GJ, Ron Y, Preston BD, Dougherty JP (2000) High rate of recombination throughout the human immunodeficiency virus type 1 genome. *J Virol* 74:1234–1240. 32, 102

Johnston MI, Fauci AS (2007) An hiv vaccine–evolving concepts. *N Engl J Med* 356:2073–2081. 23

Johnston MI, Fauci AS (2008) An hiv vaccine–challenges and prospects. *N Engl J Med* 359:888–890. 28

Jones NA, Wei X, Flower DR, Wong M, Michor F, Saag MS, Hahn BH, Nowak MA, Shaw GM, Borrow P (2004) Determinants of human immunodeficiency virus type 1 escape from the primary cd8+ cytotoxic t lymphocyte response. *J Exp Med* 200:1243–1256. 82, 89

Jukes T, Cantor C (1969) *Mammalian Protein Metabolism*, chapter Evolution of protein molecules., pp. 21–132 Academic Press, New York. 6

Julg B, Goebel FD (2005) Hiv genetic diversity: any implications for drug resistance? *Infection* 33:299–301. 32

Kandathil AJ, Ramalingam S, Kannangai R, David S, Sridharan G (2005) Molecular epidemiology of hiv. *Indian J Med Res* 121:333–344. 138

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E (2008) Efficient control of population structure in model organism association mapping. *Genetics* 178:1709–1723. 101, 102

Kaplan NL, Darden T, Hudson RR (1988) The coalescent process in models with selection. *Genetics* 120:819–829. 36

Katsikis PD, Pulendran B, Schoenberger SP, editors (2007) *Crossroads between innate and adaptive immunity*, Vol. 590 of *Advances in Experimental Medicine and Biology* Springer. 25

Kaur G, Mehra N (2009) Genetic determinants of hiv-1 infection and progression to aids: susceptibility to hiv infection. *Tissue Antigens* 73:289–301. 30, 31

Kedem B, Fokianos K (2002) *Regression models for time series analysis* John Wiley & Sons, Inc. 36

Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, Sun C, Grayson T, Wang S, Li H, Wei X, Jiang C, Kirchherr JL, Gao F, Anderson JA, Ping LH, Swanstrom R, Tomaras GD, Blattner WA, Goepfert PA, Kilby JM, Saag MS, Delwart EL, Busch MP, Cohen MS, Montefiori DC, Haynes BF, Gaschen B, Athreya GS, Lee HY, Wood N, Seoighe C, Perelson AS, Bhattacharya T, Korber BT, Hahn BH, Shaw GM (2008) Identification and characterization of transmitted and early founder virus envelopes in primary hiv-1 infection. *Proc Natl Acad Sci U S A* 105:7552–7557. 21, 31, 33, 34, 39, 40, 41, 42, 43, 46, 52, 54, 55, 56, 57, 60, 61, 64, 66, 67, 71, 73, 75, 77, 84

Kehrer-Sawatzki H, Cooper DN (2007) Understanding the recent evolution of the human genome: insights from human-chimpanzee genome comparisons. *Hum Mutat* 28:99–130. 134

Kimman TG (2001) *Genetics of infectious disease susceptibility* Kluwer Academic Publishers. 30

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16:111–120. 6

Kingman JFC (1982a) The coalescent. *Stochastic Processes and their Applications,* 13:235–248. 34

Kingman JFC (1982b) On the genealogy of large populations. *J. Appl. Probab.* 19:27–43. 34, 36

Kingman JFC (2000) Origins of the coalescent. 1974-1982. *Genetics* 156:1461–1463. 34

Klimas N, Koneru AO, Fletcher MA (2008) Overview of hiv. *Psychosom Med* 70:523–530. 19, 20

Klug WS, Cummings MR (1999) *Concepts of Genetics* Prentice Hall, 6 edition. 6

Knudsen B, Miyamoto MM (2001) A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proceedings of the National Academy of Sciences of the United States of America* 98:14512–14517. 133, 135

Kobayashi M, Takaori-Kondo A, Shindo K, Abudu A, Fukunaga K, Uchiyama T (2004) Apobec3g targets specific virus species. *J Virol* 78:8238–8244. 26

Koonin EV (2009) Towards a postmodern synthesis of evolutionary biology. *Cell Cycle* 8:799–800. 4

Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C, Detours V (2001) Evolutionary and immunological implications of contemporary hiv-1 variation. *Br Med Bull* 58:19–42. 28

Kosakovsky-Pond SL, Poon AFY, Zárate S, Smith DM, Little SJ, Pillai SK, Ellis RJ, Wong JK, Brown AJL, Richman DD, Frost SDW (2008) Estimating selection pressures on hiv-1 using phylogenetic likelihood models. *Stat Med* 27:4779–4789. 67

Kosakovsky-Pond S, Muse SV (2005) Site-to-site variation of synonymous substitution rates. *Mol Biol Evol* 22:2375–2385. 14, 132

Kosakovsky-Pond SL, Frost SDW (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol* 22:1208–1222. 13

Kosakovsky-Pond SL, Muse SV (2005) *HyPhy: Hypothesis Testing Using Phylogenies*, chapter 6, pp. 125–181 Springer. 12, 43, 65, 67, 71, 111, 114, 139

Kosakovsky-Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW (2006a) Automated phylogenetic detection of recombination using a genetic algorithm. *Mol Biol Evol* 23:1891–1901. 88

Kosakovsky-Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW (2006b) Gard: a genetic algorithm for recombination detection. *Bioinformatics* 22:3096–3098. 66, 88, 94, 102, 111, 145

Kosiol C, Holmes I, Goldman N (2007) An Empirical Codon Model for Protein Sequence Evolution. *Mol Biol Evol* 24:1464–1479. 10

Krone, Neuhauser (1997) Ancestral processes with selection. *Theor Popul Biol* 51:210–237. 36

Kuhner MK, Yamato J, Felsenstein J (1995) Estimating effective population size and mutation rate from sequence data using metropolis-hastings sampling. *Genetics* 140:1421–1430. 34

Kuiken C, Leitner T, Foley B, Hahn B, Marx P, McCutchan F, Mellors J, Wolinsky S, Korber B, editors (2008) *HIV Sequence Compendium 2008* Theoretical Biology and Biophysics Group. 18

Kumar S (2005) Molecular clocks: four decades of evolution. *Nat Rev Genet* 6:654–662. 38

Kumar V, Prakash O, Manpreet S, Sumedh G, Medhi B (2006) Genetic basis of hiv-1 resistance and susceptibility: an approach to understand correlation between human genes and hiv-1 infection. *Indian J Exp Biol* 44:683–692. 25

Lama J, Planelles V (2007) Host factors influencing susceptibility to hiv infection and aids progression. *Retrovirology* 4:52. 24, 25, 26

Langford SE, Ananworanich J, Cooper DA (2007) Predictors of disease progression in hiv infection: a review. *AIDS Res Ther* 4:11. 30, 31

Larget B, Simon D (1999) Markov chasin monte carlo algorithms for the bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* 16:750–759. 37

Larson RS, editor (2006) *Methods in Molecular Biology 316: Bioinformatics and Drug Discovery* Humana Press. 8

Lee HY, Giorgi EE, Keele BF, Gaschen B, Athreya GS, Salazar-Gonzalez JF, Pham KT, Goepfert PA, Kilby JM, Saag MS, Delwart EL, Busch MP, Hahn BH, Shaw GM, Korber BT, Bhattacharya T, Perelson AS (2009) Modeling sequence evolution in acute hiv-1 infection. *J Theor Biol* 261:341–360. 40, 41, 42, 50, 57

Lehner T (2003) Innate and adaptive mucosal immunity in protection against hiv infection. *Vaccine* 21 Suppl 2:S68–S76. 24

Lemey P, Dooren SV, Vandamme AM (2005) Evolutionary dynamics of human retroviruses investigated through full-genome scanning. *Mol Biol Evol* 22:942–951. xviii, 14, 114, 115, 126, 127, 128, 129, 131

Leslie AJ, Pfafferott KJ, Chetty P, Draenert R, Addo MM, Feeney M, Tang Y, Holmes EC, Allen T, Prado JG, Altfeld M, Brander C, Dixon C, Ramduth D, Jeena P, Thomas SA, John AS, Roach TA, Kupfer B, Luzzi G, Edwards A, Taylor G, Lyall H, Tudor-Williams G, Novelli V, Martinez-Picado J, Kiepiela P, Walker BD, Goulder PJR (2004) Hiv evolution: Ctl escape mutation and reversion after transmission. *Nat Med* 10:282–289. 83, 84

Levinson W (2006) *Review of medical microbiology and immunology* McGraw-Hill. 22

Levy JA (2007) *HIV and the Pathogenesis of AIDS* ASM Press. 18, 19, 20, 24, 25, 26, 27, 28, 29, 31, 103, 117, 120, 138

Lewis F, Hughes GJ, Rambaut A, Pozniak A, Brown AJL (2008) Episodic sexual transmission of hiv revealed by molecular phylodynamics. *PLoS Med* 5:e50. 32

Li B, Gladden AD, Altfeld M, Kaldor JM, Cooper DA, Kelleher AD, Allen TM (2007) Rapid reversion of sequence polymorphisms dominates early human immunodeficiency virus type 1 evolution. *J Virol* 81:193–201. 84

Li H, Xu CF, Blais S, Wan Q, Zhang HT, Landry SJ, Hioe CE (2009) Proximal glycans outside of the epitopes regulate the presentation of hiv-1 envelope gp120 helper epitopes. *J Immunol* 182:6369–6378. 91

Li J, Halloran M, Lord C, Watson A, Ranchalis J, Fung M, Letvin N, Sodroski J (1995) Persistent infection of macaques with simian-human immunodeficiency viruses. *J. Virol.* 69:7061–7067. 146

Llopart A, Comeron JM (2008) Recurrent events of positive selection in independent drosophila lineages at the spermatogenesis gene roughex. *Genetics* 179:1009–1020. 12

Loh L, Batten CJ, Petravic J, Davenport MP, Kent SJ (2007) In vivo fitness costs of different gag cd8 t-cell escape mutant simian-human immunodeficiency viruses for macaques. *J Virol* 81:5418–5422. 83

Loh L, Petravic J, Batten CJ, Davenport MP, Kent SJ (2008) Vaccination and timing influence siv immune escape viral dynamics in vivo. *PLoS Pathog* 4:e12. 83

Lynch RM, Shen T, Gnanakaran S, Derdeyn CA (2009) Appreciating hiv type 1 diversity: subtype differences in env. *AIDS Res Hum Retroviruses* 25:237–248. 32

Macbeth HM, Collinson P, editors (2002) *Human Population Dynamics: Cross-disciplinary Perspectives* Cambridge University Press. 8

Mahalanabis M, Jayaraman P, Miura T, Pereyra F, Chester EM, Richardson B, Walker B, Haigwood NL (2009) Continuous viral escape and selection by autologous neutralizing antibodies in drug-naive human immunodeficiency virus controllers. *J Virol* 83:662–672. 29

Mansky LM, Temin HM (1995) Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J Virol* 69:5087–5094. 41, 43

Markowitz M, Louie M, Hurley A, Sun E, Mascio MD, Perelson AS, Ho DD (2003) A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and t-cell decay in vivo. *J Virol* 77:5037–5038. 41, 44

Marsh SGE, Parham P, Barber LD (2000) *The HLA Facts Book* Academic Press. 30

Martin AP, Palumbi SR (1993) Body size, metabolic rate, generation time, and the molecular clock. *Proc Natl Acad Sci U S A* 90:4087–4091. 38

Martin MJ, Nunez JI, Sobrino F, Dopazo J (1998) A procedure for detecting selection in highly variable viral genomes: evidence of positive selection in antigenic regions of capsid protein vp1 of foot-and-mouth disease virus. *Journal of Virological Methods* 74:215–221. 9

Martinez-Picado J, Prado JG, Fry EE, Pfafferott K, Leslie A, Chetty S, Thobakgale C, Honeyborne I, Crawford H, Matthews P, Pillay T, Rousseau C, Mullins JI, Brander C, Walker BD, Stuart DI, Kiepiela P, Goulder P (2006) Fitness cost of escape mutations in p24 gag in association with control of human immunodeficiency virus type 1. *Journal of Virology* 80:3617–3623. 30

Masciotra S, Owen SM, Rudolph D, Yang C, Wang B, Saksena N, Spira T, Dhawan S, Lal RB (2002) Temporal relationship between v1v2 variation, macrophage replication, and coreceptor adaptation during hiv-1 disease progression. *AIDS* 16:1887–1898. 91

Massingham T, Goldman N (2005) Detecting amino acid sites under positive selection and purifying selection. *Genetics* 169:1753–1762. xxi, 114, 115, 124, 125, 126, 129

Maydt J, Lengauer T (2006) Recco: recombination analysis using cost optimization. *Bioinformatics* 22:1064–1071. 66

Mayrose I, Doron-Faigenboim A, Bacharach E, Pupko T (2007) Towards realistic codon models: among site variability and dependency of synonymous and non-synonymous rates. *Bioinformatics* 23:319–327. 14

McBurney SP, Ross TM (2008) Viral sequence diversity: challenges for aids vaccine designs. *Expert Rev Vaccines* 7:1405–1417. 24

McCauley S, de Groot S, Mailund T, Hein J (2007) Annotation of selection strengths in viral genomes. *Bioinformatics* 23:2978–2986. 122

McCormick-Davis C, Zhao LJ, Mukherjee S, Leung K, Sheffer D, Joag SV, Narayan O, Stephens EB (1998) Chronology of genetic changes in the vpu, env, and nef genes of chimeric simian-human immunodeficiency virus (strain hxb2) during acquisition of virulence for pig-tailed macaques. *Virology* 248:275–283. 146

McMichael AJ, Rowland-Jones SL (2001) Cellular immune responses to hiv. *Nature* 410:980–987. 83

McNatt MW, Zang T, Hatziioannou T, Bartlett M, Fofana IB, Johnson WE, Neil SJD, Bieniasz PD (2009) Species-specific activity of hiv-1 vpu and positive selection of tetherin transmembrane domain variants. *PLoS Pathog* 5:e1000300. 146, 147

McNearney T, Hornickova Z, Markham R, Birdwell A, Arens M, Saah A, Ratner L (1992) Relationship of human immunodeficiency virus type 1 sequence heterogeneity to stage of disease. *Proceedings of the National Academy of Sciences of the United States of America* 89:10247–10251. 33

Mei Y, Wang L, Holte SE (2008) A comparison of methods for determining hiv viral set point. *Stat Med* 27:121–139. 19

Miller RE, McDonald JA, Manos PS (2004) Systematics of ipomoea subgenus quamoclit (convolvulaceae) based on its sequence data and a bayesian phylogenetic analysis. *Am. J. Bot.* 91:1208–1218. 37

Miura T, Brockman MA, Schneidewind A, Lobritz M, Pereyra F, Rathod A, Block BL, Brumme ZL, Brumme CJ, Baker B, Rothchild AC, Li B, Trocha A, Cutrell E, Frahm N, Brander C, Toth I, Arts EJ, Allen TM, Walker BD (2009) Hla-b57/b*5801 human immunodeficiency virus type 1 elite controllers select for rare gag variants associated with reduced viral replication capacity and strong cytotoxic t-lymphocyte [corrected] recognition. *J Virol* 83:2743–2755. 30

Moore CB, John M, James IR, Christiansen FT, Witt CS, Mallal SA (2002) Evidence of hiv-1 adaptation to hla-restricted immune responses at a population level. *Science* 296:1439–1443. 133, 134

Moran PAP (1958) Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society* 54:60–71. 36

Mossel E, Vigoda E (2005) Phylogenetic mcmc algorithms are misleading on mixtures of trees. *Science* 309:2207–2209. 36

Moulard M, Phogat SK, Shu Y, Labrijn AF, Xiao X, Binley JM, Zhang MY, Sidorov IA, Broder CC, Robinson J, Parren PWHI, Burton DR, Dimitrov DS (2002) Broadly cross-reactive hiv-1-neutralizing human monoclonal fab selected for binding to gp120-cd4-ccr5 complexes. *Proc Natl Acad Sci U S A* 99:6913–6918. 28

Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H (1986) Specific enzymatic amplification of dna in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol* 51 Pt 1:263–273. 4

Muse S, Gaut B (1994) A likelihood approach for comparing synonynous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol* 11:715–724. 8

Nabel GJ (2002) Hiv vaccine strategies. *Vaccine* 20:1945–1947. 28

Nei M, Kumar S (2000) *Molecular Evolution and Phylogenetics* Oxford University Press. 38

Neuhauser C, Krone SM (1997) The genealogy of samples in models with selection. *Genetics* 145:519–534. 36

Newman RM, Hall L, Connole M, Chen GL, Sato S, Yuste E, Diehl W, Hunter E, Kaur A, Miller GM, Johnson WE (2006) Balancing selection and the evolution of functional polymorphism in old world monkey trim5alpha. *Proc Natl Acad Sci U S A* 103:19134–19139. 18

Newman RM, Johnson WE (2007) A brief history of trim5alpha. *AIDS Rev* 9:114–125. 26

Ngandu N, Seoighe C, Scheffler K (2009) Evidence of hiv-1 adaptation to host hla alleles following chimp-to-human transmission. *Virology Journal* 6:164. 148

Ngandu NK, Scheffler K, Moore P, Woodman Z, Martin D, Seoighe C (2008) Extensive purifying selection acting on synonymous sites in hiv-1 group m sequences. *Virol J* 5:160. 14, 122, 132

Nielsen M, Lundegaard C, Blicher T, Peters B, Sette A, Justesen S, Buus S, Lund O (2008) Quantitative predictions of peptide binding to any hla-dr molecule of known sequence: Netmhciipan. *PLoS Comput Biol* 4:e1000107. 30

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the hiv-1 envelope gene. *Genetics* 148:929–936. 10, 15, 67, 124

Nielsen R (2005) *Statistical Methods in Molecular Evolution* Springer. 6

Nordborg M (2001) *Handbook of Statistical Genetics*, chapter Coalescent Theory, pp. 179–212 John Wiley & Sons. 35

Novitsky V, Woldegabriel E, Kebaabetswe L, Rossenkhan R, Mlotshwa B, Bonney C, Finucane M, Musonda R, Moyo S, Wester C, van Widenfelt E, Makhema J, Lagakos S, Essex M (2009) Viral load and cd4+ t-cell dynamics in primary hiv-1 subtype c infection. *J Acquir Immune Defic Syndr* 50:65–76. 21

Ohta T (1993) Pattern of nucleotide substitutions in growth hormone-prolactin gene family: a paradigm for evolution by gene duplication. *Genetics* 134:1271–1276. 133

Okazaki T, Terabe M, Catanzaro AT, Pendleton CD, Yarchoan R, Berzofsky JA (2006) Possible therapeutic vaccine strategy against human immunodeficiency virus escape from reverse transcriptase inhibitors studied in hla-a2 transgenic mice. *J Virol* 80:10645–10651. 32

Ota R, Waddell PJ, Hasegawa M, Shimodaira H, Kishino H (2000) Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Mol Biol Evol* 17:798–803. 136

Pagel M (1994) Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings: Biological Sciences* 255:37–45. 101

Palmer S, Kearney M, Maldarelli F, Halvas EK, Bixby CJ, Bazmi H, Rock D, Falloon J, Davey RT, Dewar RL, Metcalf JA, Hammer S, Mellors JW, Coffin JM (2005) Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *J Clin Microbiol* 43:406–413. 33

Pathak VK, Temin HM (1990) Broad spectrum of in vivo forward mutations, hypermutations, and mutational hotspots in a retroviral shuttle vector after a single replication cycle: deletions and deletions with insertions. *Proc Natl Acad Sci U S A* 87:6024–6028. 78

Pedersen AMK, Jensen JL (2001) A Dependent-Rates Model and an MCMC-Based Methodology for the Maximum-Likelihood Analysis of Sequences with Overlapping Reading Frames. *Mol Biol Evol* 18:763–776. 122

Penn O, Stern A, Rubinstein ND, Dutheil J, Bacharach E, Galtier N, Pupko T (2008) Evolutionary modeling of rate shifts reveals specificity determinants in hiv-1 subtypes. *PLoS Comput Biol* 4:e1000214. 133, 136, 137, 141

Perry GH, Yang F, Marques-Bonet T, Murphy C, Fitzgerald T, Lee AS, Hyland C, Stone AC, Hurles ME, Tyler-Smith C, Eichler EE, Carter NP, Lee C, Redon R (2008) Copy number variation and evolution in humans and chimpanzees. *Genome Res* 18:1698–1710. 134

Pevsner J (2003) *Bioinformatics and Functional Genomics* John Whiley and Sons, Inc., New Jersey. 4, 5, 38

Peyerl FW, Bazick HS, Newberg MH, Barouch DH, Sodroski J, Letvin NL (2004) Fitness costs limit viral escape from cytotoxic t lymphocytes at a structurally constrained epitope. *J Virol* 78:13901–13910. 84

Piguet V, Steinman RM (2007) The interaction of hiv with dendritic cells: outcomes and pathways. *Trends Immunol* 28:503–510. 24

Pillai SK, Wong JK, Barbour JD (2008) Turning up the volume on mutational pressure: is more of a good thing always better? (a case study of hiv-1 vif and apobec3). *Retrovirology* 5:26. 27

Plantier JC, Leoz M, Dickerson JE, Oliveira FD, Cordonnier F, Lemee V, Damond F, Robertson DL, Simon F (2009) A new human immunodeficiency virus derived from gorillas. *Nature Medicine* 15:871–872. 18

Plikat U, Nieselt-Struwe K, Meyerhans A (1997) Genetic drift can dominate short-term human immunodeficiency virus type 1 nef quasispecies evolution in vivo. *J Virol* 71:4233–4240. 114

Poon AFY, Lewis FI, Pond SLK, Frost SDW (2007) An evolutionary-network model reveals stratified interactions in the v3 loop of the hiv-1 envelope. *PLoS Comput Biol* 3:e231. 14

Porth C (2007) *Essentials of pathophysiology: concepts of altered health states* Lippincott Williams & Wilkins. 27

Prabakaran P, Gan J, Wu YQ, Zhang MY, Dimitrov DS, Ji X (2006) Structural mimicry of cd4 by a cross-reactive hiv-1 neutralizing antibody with cdr-h2 and h3 containing unique motifs. *J Mol Biol* 357:82–99. 28

Pupko T, Galtier N (2002) A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. *Proc Biol Sci* 269:1313–1316. 133

Pybus OG (2006) Model selection and the molecular clock. *PLoS Biol* 4:e151. 38

Qian SS, Stow CA, Borsuk ME (2003) On monte carlo methods for bayesian inference. *Ecological Modelling* 159:269 – 277. 36

Qing M, Li T, Han Y, Qiu Z, Jiao Y (2006) Accelerating effect of human leukocyte antigen-bw6 homozygosity on disease progression in chinese hiv-1-infected patients. *J Acquir Immune Defic Syndr* 41:137–139. 30

Rambaut A, Drummond AJ (2007) Tracer v1.4, available from http://beast.bio.ed.ac.uk/tracer. 43, 44

Ramirez BC, Simon-Loriere E, Galetto R, Negroni M (2008) Implications of recombination for hiv diversity. *Virus Res* 134:64–73. 32, 109

Reeves JD, Doms RW (2002) Human immunodeficiency virus type 2. *J Gen Virol* 83:1253–1265. 18

Reid AH, Fanning TG, Hultin JV, Taubenberger JK (1999) Origin and evolution of the 1918 spanish influenza virus hemagglutinin gene. *Proceedings of the National Academy of Sciences of the United States of America* 96:1651–1656. 133

Repits J, Sterjovski J, Badia-Martinez D, Mild M, Gray L, Churchill MJ, Purcell DFJ, Karlsson A, Albert J, Fenyo EM, Achour A, Gorry PR, Jansson M (2008) Primary hiv-1 r5 isolates from end-stage disease display enhanced viral fitness in parallel with increased gp120 net charge. *Virology* 379:125–134. 91

Requejo HIZ (2006) Worldwide molecular epidemiology of hiv. *Rev Saude Publica* 40:331–345. 109

Ritola K, Pilcher CD, Fiscus SA, Hoffman NG, Nelson JAE, Kitrinos KM, Hicks CB, Eron JJ, Swanstrom R (2004) Multiple v1/v2 env variants are frequently present during primary infection with human immunodeficiency virus type 1. *J Virol* 78:11208–11218. 61

Rits-Volloch S, Frey G, Harrison SC, Chen B (2006) Restraining the conformation of hiv-1 gp120 by removing a flexible loop. *EMBO J* 25:5026–5035. 23

Roberts JD, Bebenek K, Kunkel TA (1988) The accuracy of reverse transcriptase from hiv-1. *Science* 242:1171–1173. 32

Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B, Funkhouser RK, Gao F, Hahn BH, Kalish ML, Kuiken C, Learn GH, Leitner T, McCutchan F, Osmanov S, Peeters M, Pieniazek D, Salminen M, Sharp PM, Wolinsky S, Korber B (2000) Hiv-1 nomenclature proposal. *Science* 288:55–56. 18, 21

Roger M (1998) Influence of host genes on hiv-1 disease progression. *FASEB J* 12:625–632. 30

Root MJ, Hamer DH (2003) Targeting therapeutics to an exposed and conserved binding element of the hiv-1 fusion protein. *Proc Natl Acad Sci U S A* 100:5016–5021. 23

Rosenberg MS (2005) Myssp: Non-stationary evolutionary sequence simulation, including indels. *Evol Bioinform Online* 1:81–83. 93, 100

Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat Rev Genet* 3:380–390. 34, 35, 36

Ross HA, Rodrigo AG (2002) Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. *J Virol* 76:11715–11720. 133, 135

Rousseau CM, Learn GH, Bhattacharya T, Nickle DC, Heckerman D, Chetty S, Brander C, Goulder PJR, Walker BD, Kiepiela P, Korber BT, Mullins JI (2007) Extensive intrasubtype recombination in south african human immunodeficiency virus type 1 subtype c infections. *J Virol* 81:4492–4500. 21, 22, 109

Russell RA, Moore MD, Hu WS, Pathak VK (2009) Apobec3g induces a hypermutation gradient: purifying selection at multiple steps during hiv-1 replication results in levels of g-to-a mutations that are high in dna, intermediate in cellular viral rna, and low in virion rna. *Retrovirology* 6:16. 27

Sabath N, Landan G, Graur D (2008) A method for the simultaneous estimation of selection intensities in overlapping genes. *PLoS One* 3:e3996. 122

Sagar M, Kirkegaard E, Long EM, Celum C, Buchbinder S, Daar ES, Overbaugh J (2004) Human immunodeficiency virus type 1 (hiv-1) diversity at time of infection is not restricted to certain risk groups or specific hiv-1 subtypes. *J Virol* 78:7279–7283. 60

Sagar M, Wu X, Lee S, Overbaugh J (2006) Human immunodeficiency virus type 1 v1-v2 envelope loop sequences expand and add glycosylation sites over the course of infection, and these modifications affect antibody neutralization sensitivity. *J Virol* 80:9586–9598. 75, 91, 107

Salazar-Gonzalez JF, Bailes E, Pham KT, Salazar MG, Guffey MB, Keele BF, Derdeyn CA, Farmer P, Hunter E, Allen S, Manigart O, Mulenga J, Anderson JA, Swanstrom R, Haynes BF, Athreya GS, Korber BTM, Sharp PM, Shaw GM, Hahn BH (2008) Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J Virol* 82:3952–3970. 33, 64

Sanderson MJ (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19:301–302. 39

Sanderson M (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol Biol Evol* 14:1218–1231. 38

Santiago ML, Bibollet-Ruche F, Bailes E, Kamenya S, Muller MN, Lukasik M, Pusey AE, Collins DA, Wrangham RW, Goodall J, Shaw GM, Sharp PM, Hahn BH (2003) Amplification of a complete simian immunodeficiency virus genome from fecal rna of a wild chimpanzee. *J Virol* 77:2233–2242. 138

Santiago ML, Rodenburg CM, Kamenya S, Bibollet-Ruche F, Gao F, Bailes E, Meleth S, Soong SJ, Kilby JM, Moldoveanu Z, Fahey B, Muller MN, Ayouba A, Nerrienet E, McClure HM, Heeney JL, Pusey AE, Collins DA, Boesch C, Wrangham RW, Goodall J, Sharp PM, Shaw GM, Hahn BH (2002) Sivcpz in wild chimpanzees. *Science* 295:465–465. 138

Sattentau QJ (1998) Hiv gp120: double lock strategy foils host defences. *Structure* 6:945–949. 23

Sattentau QJ, Moulard M, Brivet B, Botto F, Guillemot JC, Mondor I, Poignard P, Ugolini S (1999) Antibody neutralization of hiv-1 and the potential for vaccine design. *Immunol Lett* 66:143–149. 28

Scheffler K, Seoighe C (2005) A bayesian model comparison approach to inferring positive selection. *Mol Biol Evol* 22:2531–2540. 12

Scheffler K, Martin DP, Seoighe C (2006) Robust inference of positive selection from recombining coding sequences. *Bioinformatics* 22:2493–2499. 13, 15, 94, 110, 111, 112, 113, 114, 116, 124, 128, 130, 131

Scheid JF, Mouquet H, Feldhahn N, Seaman MS, Velinzon K, Pietzsch J, Ott RG, Anthony RM, Zebroski H, Hurley A, Phogat A, Chakrabarti B, Li Y, Connors M, Pereyra F, Walker BD, Wardemann H, Ho D, Wyatt RT, Mascola JR, Ravetch JV, Nussenzweig MC (2009) Broad diversity of neutralizing antibodies isolated from memory b cells in hiv-infected individuals. *Nature* 458:636–640. 28

Schmitz JE, Kuroda MJ, Santra S, Sasseville VG, Simon MA, Lifton MA, Racz P, Tenner-Racz K, Dalesandro M, Scallon BJ, Ghrayeb J, Forman MA, Montefiori DC, Rieber EP, Letvin NL, Reimann KA (1999) Control of viremia in simian immunodeficiency virus infection by cd8+ lymphocytes. *Science* 283:857–860. 83

Schneider A, Cannarozzi G, Gonnet G (2005) Empirical codon substitution matrix. *BMC Bioinformatics* 6:134. 10

Self S, Liang K (1987) Asymptotic properties of maximum-likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* 82:605–610. 136

Shriner D, Nickle DC, Jensen MA, Mullins JI (2003) Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genet Res* 81:115–121. 13, 87, 94, 109, 110, 123, 125, 126, 128, 145

Shriner D, Rodrigo AG, Nickle DC, Mullins JI (2004) Pervasive genomic recombination of hiv-1 in vivo. *Genetics* 167:1573–1583. 33

Simmonds P, Balfe P, Ludlam CA, Bishop JO, Brown AJ (1990) Analysis of sequence diversity in hypervariable regions of the external glycoprotein of human immunodeficiency virus type 1. *J Virol* 64:5840–5850. 33

Singh K, Spector S (2009) Host genetic determinants of hiv infection and disease progression in children. *Pediatr Res* . 30

Stafford MA, Corey L, Cao Y, Daar ES, Ho DD, Perelson AS (2000) Modeling plasma virus concentration during primary hiv infection. *J Theor Biol* 203:285–301. 41

Steain MC, Wang B, Dwyer DE, Saksena NK (2004) Hiv-1 co-infection, superinfection and recombination. *Sex Health* 1:239–250. 22

Stern A, Pupko T (2006) An Evolutionary Space-Time Model with Varying Among-Site Dependencies. *Mol Biol Evol* 23:392–400. 14

Stivahtis GL, Soares MA, Vodicka MA, Hahn BH, Emerman M (1997) Conservation and host specificity of vpr-mediated cell cycle arrest suggest a fundamental role in primate lentivirus evolution and biology. *J Virol* 71:4331–4338. 145

Stone L, Lurquin PF, Cavalli-Sforza LL (2007) *Genes, culture, and human evolution: a synthesis* Blackwell Publishing. 35

Stürmer M, Doerr H, Gürtler L (2009) Human immunodeficiency virus: 25 years of diagnostic and therapeutic strategies and their impact on hepatitis b and c virus. *Med Microbiol Immunol* . 22

Suchard MA, Weiss RE, Sinsheimer JS (2001) Bayesian selection of continuous-time markov chain evolutionary models. *Mol Biol Evol* 18:1001–1013. 44

Suzuki Y, Gojobori T (1999) A method for detecting positive selection at single amino acid sites. *Mol Biol Evol* 16:1315–1328. 124

Suzuki Y (2006) Natural selection on the influenza virus genome. *Mol Biol Evol* 23:1902–1911. 122

Suzuki Y, Nei M (2001) Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. *Mol Biol Evol* 18:2179–2185. 9

Suzuki Y, Nei M (2002) Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. *Mol Biol Evol* 19:1865–1869. 74

Swafford D (2002) Paup*. phylogenetic analysis using parsimony (*and other methods). version 4. 113

Swanson WJ, Nielsen R, Yang Q (2003) Pervasive adaptive evolution in mammalian fertilization proteins. *Mol Biol Evol* 20:18–20. 124

Tavaré S (1986) Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on Mathematics in the Life Sciences* 17:57–86. 7

Taylor BS, Sobieszczyk ME, McCutchan FE, Hammer SM (2008) The challenge of hiv-1 subtype diversity. *N Engl J Med* 358:1590–1602. 32

Thorne JL, Kishino H, Painter IS (1998) Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol* 15:1647–1657. 38, 39

Trachtenberg E, Korber B, Sollars C, Kepler TB, Hraber PT, Hayes E, Funkhouser R, Fugate M, Theiler J, Hsu YS, Kunstman K, Wu S, Phair J, Erlich H, Wolinsky S (2003) Advantage of rare hla supertype in hiv disease progression. *Nat Med* 9:928–935. 30

Treurnicht FK, Seoighe C, Martin DP, Wood N, Abrahams MR, de Assis Rosa D, Bredell H, Woodman Z, Hide W, Mlisana K, Karim SA, Gray CM, Williamson C (2009) Adaptive changes in hiv-1 subtype c proteins during early infection are driven by changes in hla-associated immune pressure. *Virology* 396:213–225. 106

Tuffley C, Steel M (1998) Modeling the covarion hypothesis of nucleotide substitution. *Math Biosci* 147:63–91. 15, 132

Turnbull EL, Lopes AR, Jones NA, Cornforth D, Newton P, Aldam D, Pellegrino P, Turner J, Williams I, Wilson CM, Goepfert PA, Maini MK, Borrow P (2006) Hiv-1 epitope-specific cd8+ t cell responses strongly associated with delayed disease progression cross-recognize epitope variants efficiently. *J Immunol* 176:6130–6146. 83

UNAIDS (2008) Report on the global hiv/aids epidemic 2008: executive summary. Technical report, UNAIDS. 17

Van Heuverswyn F, Li Y, Bailes E, Neel C, Lafay B, Keele BF, Shaw KS, Takehisa J, Kraus MH, Loul S, Butel C, Liegeois F, Yangda B, Sharp PM, Mpoudi-Ngole E, Delaporte E, Hahn BH, Peeters M (2007) Genetic diversity and phylogeographic clustering of sivcpzptt in wild chimpanzees in cameroon. *Virology* 368:155 – 171. 138

Vanden Haesevelde MM, Peeters M, Jannes G, Janssens W, Van der Groen G, Sharp P, Saman E (1996) Sequence analysis of a highly divergent hiv-1-related lentivirus isolated from a wild captured chimpanzee. *Virology* 221:346 – 350. 138

Vander, Sherman, Luciano (2001a) *Human Physiology: The Mechanisms of Body Function* McGraw-Hill, 8 edition. 24, 28, 29

Vander A, Sherman J, Luciano D (2001b) *Human Physiology, The Mechanisms of Body Function* McGraw-Hill. 30

Wakeley J, editor (2008) *Coalescent Theory: An Introduction* Roberts & Company. 34, 35, 36

Wakeley J, Sargsyan O (2009) Extensions of the coalescent effective population size. *Genetics* 181:341–345. 36

Walensky RP, Paltiel AD, Losina E, Mercincavage LM, Schackman BR, Sax PE, Weinstein MC, Freedberg KA (2006) The survival benefits of aids treatment in the united states. *J Infect Dis* 194:11–19. 22

Walker LM, Phogat SK, Chan-Hui PY, Wagner D, Phung P, Goss JL, Wrin T, Simek MD, Fling S, Mitcham JL, Lehrman JK, Priddy FH, Olsen OA, Frey SM, Hammond PW, Investigators PGP, Kaminsky S, Zamb T, Moyle M, Koff WC, Poignard P, Burton DR (2009) Broad and potent neutralizing antibodies from an african donor reveal a new hiv-1 vaccine target. *Science* 326:285–289. 28, 29

Wang HC, Spencer M, Susko E, Roger AJ (2007) Testing for covarion-like evolution in protein sequences. *Mol Biol Evol* 24:294–305. 16

Wang Q, Shang H, Han X, Zhang Z, Jiang Y, Wang Y, Dai D, Diao Y (2008) High level serum neutralizing antibody against hiv-1 in chinese long-term non-progressors. *Microbiol Immunol* 52:209–215. 91

Wang Y, Gu X (2001) Functional divergence in the caspase gene family and altered functional constraints: Statistical analysis and prediction. *Genetics* 158:1311–1320. 133

Wei X, Decker JM, Wang S, Hui H, Kappes JC, Wu X, Salazar-Gonzalez JF, Salazar MG, Kilby JM, Saag MS, Komarova NL, Nowak MA, Hahn BH, Kwong PD, Shaw GM (2003) Antibody neutralization and escape by hiv-1. *Nature* 422:307–312. 90, 91

Westfall PH, Young S (1993) *Resampling-based multiple testing: examples and methods for P-value adjustment* John Wiley and Sons, Inc. 106

Whelan S, Goldman N (1999) Distributions of Statistics Used for the Comparison of Models of Sequence Evolution in Phylogenetics. *Mol Biol Evol* 16:1292–1299. 136

Williams KC, Burdo TH (2009) Hiv and siv infection: the role of cellular restriction and immune responses in viral replication and pathogenesis. *APMIS* 117:400–412. 26

Wolfs TF, Zwart G, Bakker M, Goudsmit J (1992) Hiv-1 genomic rna diversification following sexual and parenteral virus transmission. *Virology* 189:103 – 110. 33

Wong WSW, Yang Z, Goldman N, Nielsen R (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* 168:1041–1051. 10, 67, 68, 113, 133

Wood N, Bhattacharya T, Keele BF, Giorgi E, Liu M, Gaschen B, Daniels M, Ferrari G, Haynes BF, McMichael A, Shaw GM, Hahn BH, Korber B, Seoighe C (2009) Hiv evolution in early infection: selection pressures, patterns of insertion and deletion, and the impact of apobec. *PLoS Pathogens* 5:e1000414. 23, 33, 34, 63, 69

Woolley SM, Posada D, Crandall KA (2008) A comparison of phylogenetic network methods using computer simulation. *PLoS ONE* 3:e1913. 124

Wu X, Cai Z, Wan XF, Hoang T, Goebel R, Lin G (2007) Nucleotide composition string selection in hiv-1 subtyping using whole genomes. *Bioinformatics* 23:1744–1752. 138

Wyatt R, Kwong PD, Desjardins E, Sweet RW, Robinson J, Hendrickson WA, Sodroski JG (1998) The antigenic structure of the hiv gp120 envelope glycoprotein. *Nature* 393:705–711. 23

Wyatt R, Sodroski J (1998) The hiv-1 envelope glycoproteins: fusogens, antigens, and immunogens. *Science* 280:1884–1888. 23

Yang W, Bielawski JP, Yang Z (2003) Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. *J Mol Evol* 57:212–221. 67, 113, 114, 117, 122, 123, 124, 126, 128, 129, 131

Yang Z (1993) Maximum-likelihood estimation of phylogeny from dna sequences when substitution rates differ over sites. *Mol Biol Evol* 10:1396–1401. 7

Yang Z (1994) Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. *J Mol Evol* 39:306–314. 7

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* 15:568–573. 9

Yang Z (2000) Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus a. *J Mol Evol* 51:423–432. 124

Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* 17:32–43. 67

Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449. 8, 10, 15, 113, 126, 140

Yang Z (2006) *Computational Molecular Evolution* Oxford University Press, Oxford. 6, 9, 109

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* 19:908–917. 15, 140

Yang Z, Rannala B (2005) Branch-length prior influences bayesian posterior probability of phylogeny. *Syst Biol* 54:455–470. 37

Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 22:1107–1118. 12, 68

Ye Z, Ahmed KA, Hao S, Zhang X, Xie Y, Munegowda MA, Meng Q, Chibbar R, Xiang J (2008) Active cd4+ helper t cells directly stimulate cd8+ cytotoxic t lymphocyte responses in wild-type and mhc ii gene knockout c57bl/6 mice and transgenic rip-mova mice expressing islet beta-cell ovalbumin antigen leading to diabetes. *Autoimmunity* 41:501–511. 29

Yu X, Yu Y, Liu B, Luo K, Kong W, Mao P, Yu XF (2003) Induction of apobec3g ubiquitination and degradation by an hiv-1 vif-cul5-scf complex. *Science* 302:1056–1060. 26

Zanotto PM, Kallas EG, de Souza RF, Holmes EC (1999) Genealogical evidence for positive selection in the nef gene of hiv-1. *Genetics* 153:1077–1089. 114, 120, 129, 131

Zapata W, Rodriguez B, Weber J, Estrada H, Quiñones-Mateu ME, Zimermman PA, Lederman MM, Rugeles MT (2008) Increased levels of human beta-defensins mrna in sexually hiv-1 exposed but uninfected individuals. *Curr HIV Res* 6:531–538. 26

Zhang J, Webb DM (2004) Rapid evolution of primate antiviral enzyme apobec3g. *Hum Mol Genet* 13:1785–1791. 26

Zhang LQ, MacKenzie P, Cleland A, Holmes EC, Brown AJ, Simmonds P (1993) Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. *J Virol* 67:3345–3356. 33

Zhang M, Gaschen B, Blay W, Foley B, Haigwood N, Kuiken C, Korber B (2004) Tracking global patterns of n-linked glycosylation site variation in highly variable viral glycoproteins: Hiv, siv, and hcv envelopes and influenza hemagglutinin. *Glycobiology* 14:1229–1246. 23

Zhang W, Canziani G, Plugariu C, Wyatt R, Sodroski J, Sweet R, Kwong P, Hendrickson W, Chaiken I (1999) Conformational changes of gp120 in epitopes near the ccr5 binding site are induced by cd4 and a cd4 miniprotein mimetic. *Biochemistry* 38:9405–9416. 23

Zhang YJ, Fracasso C, Fiore JR, Bjorndal A, Angarano G, Gringeri A, Fenyo EM (1997) Augmented serum neutralizing activity against primary human immunodeficiency virus type 1 (hiv-1) isolates in two groups of hiv-1-infected long-term nonprogressors. *J Infect Dis* 176:1180–1187. 91

Zheng YH, Irwin D, Kurosu T, Tokunaga K, Sata T, Peterlin BM (2004) Human apobec3f is another host factor that blocks human immunodeficiency virus type 1 replication. *J Virol* 78:6073–6076. 27

Zhou T, Xu L, Dey B, Hessell AJ, Ryk DV, Xiang SH, Yang X, Zhang MY, Zwick MB, Arthos J, Burton DR, Dimitrov DS, Sodroski J, Wyatt R, Nabel GJ, Kwong PD (2007) Structural definition of a conserved neutralization epitope on hiv-1 gp120. *Nature* 445:732–737. 29

Zuckerkandl E, Pauling L (1965) *Evolving Genes and Proteins*, chapter Evolutionary Divergence and Convergence in Proteins, pp. 97–166 New York Academic Press. 38

Appendix

Table A1: Observed mutations, sequence context within specific patients, and possible alternative explanations besides selection.

HXB2 #	Site #	Sequence	Patient ID	Consensus ID	Amino Acid in Consensus	Amino Acid in Patient	Codon in Consensus	Codon in Patient	9 nucleotides of the consensus (centred on the codon of interest)	9 nucleotides of the patient (centred on the codon of interest)	Possible explanation besides selection
62	79	5.C.B.MEM4948.UAB.012803.3229	MEM4948	5.Con.MEM4948	D	N	gat	aat	catgatcca	catAatcca	APOBEC3G
62	79	2.A.B.SUMA.UAB.051391.4988	SUMA	2.Con.SUMA	D	Y	gat	tat	tatgatcca	tatAatcca	
62	79	2.A.B.Z31.NC-Duke.120901.928	Z31	2.Con.Z31	E	K	gag	aag	tatgagaaa	tatAagaaa	APOBEC3G
62	79	2.A.B.Z31.NC-Duke.120901.921	Z31	2.Con.Z31	E	K	gag	aag	tatgagaaa	tatAagaaa	APOBEC3G
62	79	2.-.B.ZP12007-04.102299.21.3935	ZP12007-04	2.Con.ZP12007-04	E	K	gag	aag	tatgagaca	tatAagaca	APOBEC3G
64	81	3.-.B.PRB931_06.Milwaukee_WI.091195.5170	PRB931_06	3.Con.PRB931_06	E	G	gag	gga	acagagaca	acaggaca	
64	81	5.C.B.MEM4948.UAB.012803.3235	MEM4948	5.Con.MEM4948	E	G	gaa	gga	cagagata	caggata	
64	81	5.C.B.MEM4948.UAB.012803.4103	MEM4948	5.Con.MEM4948	E	G	gaa	gga	cagagata	caggata	
66	83	2.A.B.1056.NAgusta_SC.011498.1820	1056	2.Con.1056	H	Y	cac	tac	gtatatcat	gtatAtcat	APOBEC1
66	83	5.C.B.MEM4948.UAB.012803.3224	MEM4948	5.Con.MEM4948	H	Y	cac	tac	gtatatcat	gtatAtcat	APOBEC1
66	83	5.C.B.MEM4948.UAB.012803.3237	MEM4948	5.Con.MEM4948	H	Y	cac	tac	gtatatcat	gtatAtcat	APOBEC1
66	83	5.C.B.MEM4948.UAB.012803.3222	MEM4948	5.Con.MEM4948	H	Y	cac	tac	gtatatcat	gtatAtcat	APOBEC1
66	83	5.C.B.MEM4948.UAB.012803.3238	MEM4948	5.Con.MEM4948	H	Y	cac	tac	gtatatcat	gtatAtcat	APOBEC1
66	83	5.C.B.MEM4948.UAB.012803.4097	MEM4948	5.Con.MEM4948	H	Y	cac	tac	gtatatcat	gtatAtcat	APOBEC1
66	83	5.C.B.MEM4948.UAB.012803.3235	MEM4948	5.Con.MEM4948	H	Y	cac	tac	gtatatcat	gtatAtcat	APOBEC1
66	83	5.C.B.MEM4948.UAB.012803.3220	MEM4948	5.Con.MEM4948	H	Y	cac	tac	gtatatcat	gtatAtcat	APOBEC1
66	83	5.C.B.MEM4948.UAB.012803.3234	MEM4948	5.Con.MEM4948	H	Y	cac	tac	gtatatcat	gtatAtcat	APOBEC1
66	83	5.C.B.MEM4948.UAB.012803.4102	MEM4948	5.Con.MEM4948	H	Y	cac	tac	gtatatcat	gtatAtcat	APOBEC1
66	83	5.C.B.MEM4948.UAB.012803.3229	MEM4948	5.Con.MEM4948	H	Y	cac	tac	gtatatcat	gtatAtcat	APOBEC1
66	83	5.A.B.Z02.NC-Duke.052798.1369	Z02	5.Con.Z02	H	Y	cac	tac	gtatatcat	gtatAtcat	APOBEC1
66	83	5.A.B.Z02.NC-Duke.052798.1350	Z02	5.Con.Z02	H	Y	cac	tac	gtatatcat	gtatAtcat	APOBEC1
66	83	1.-.B.ZP62995-05.010397.4.5220	ZP62995-05	1.Con.ZP62995-05	H	Y	cac	tac	gtatatcat	gtatAtcat	APOBEC1
175	226	3.A.B.1012.LongBeach.CA.040797.3256	1012	3.Con.1012	L	F	ctt	ttt	gcatttttt	gcatttttt	APOBEC1 ?
175	226	2.-.B.9014_01.NAgusta_SC.111397.4771	9014_01	2.Con.9014_01	L	F	ctt	ttt	gcatttttt	gcatttttt	APOBEC1 ?
175	226	2.-.B.PRB959_02.Columbia_SC.111799.4350	PRB959_02	2.Con.PRB959_02	L	P	cct	cct	gtcctttt	gtccttttt	
175	226	2.A.B.WEAU.UAB.053090.5025	WEAU	2.Con.WEAU	L	P	cct	cct	gcatttttt	gcatttttt	
175	226	2.A.B.Z05.NC-Duke.092198.828	Z05	2.Con.Z05	L	P	cct	cct	gcatttttt	gcatttttt	
175	226	3.-.B.ZP6248-07.Dallas_TX.030997.3698	ZP6248-07	3.Con.ZP6248-07	L	F	ctt	ttt	gcatttttt	gcatttttt	APOBEC1 ?
175	226	3.-.B.ZP9022-09.Jacksonville_FL.083197.3724	ZP9022-09	3.Con.ZP9022-09	N	S	aat	agt	gcaagtttt	gcagttttt	
176	227	2.-.B.6240_08.Jacksonville_MS.112295.4633	6240_08	2.Con.6240_08	F	S	ttt	tct	cttttctat	cttcttcat	
176	227	2.A.B.1056.NAgusta_SC.011498.1822	1056	2.Con.1056	F	S	ttt	tct	cttttctat	cttcttcat	
176	227	3.A.B.1012.LongBeach.CA.040797.3282	1012	3.Con.1012	F	V	ttt	gtt	cttttctat	ctgttcat	
232	306	2.-.B.9015_07.NAgusta_SC.122797.4730	9015_07	2.Con.9015_07	K	E	aag	gag	aagaagttc	aagagttc	
232	306	3.-.B.9032_08.Mobile_AL.073098.4689	9032_08	3.Con.9032_08	T	A	aca	gca	aagatttc	aagcattc	
232	306	2.A.B.SC22.Trinidad.031694.2474	SC22	2.Con.SC22	K	R	aag	agg	aagaagttc	aaggattc	
232	306	5.-.B.ZP9019-03.NAugusta_SC.122897.5240	ZP9019-03	5.Con.ZP9019-03	T	M	acg	atg	aagatttc	aagattttc	APOBEC1
242	316	2.-.B.ZP9030-15.SanDiego_CA.100298.3863	ZP9030-15	2.Con.ZP9030-15	V	L	gtc	ctc	aatgcagc	aatccagc	
242	316	2.A.B.WITO4160.UAB.060400.2065	WITO4160	2.Con.WITO4160	V	I	gtc	atc	aatgcagt	aatcagt	
242	316	5.C.B.MEM4948.UAB.012803.3229	MEM4948	5.Con.MEM4948	V	I	gtc	atc	aatgcagc	aatcagc	
274	349	3.A.B.Z34.NC-Duke.121702.943	Z34	3.Con.Z34	S	P	tct	cct	aaatcagc	aaccagc	
274	349	2.A.B.SC45.Trinidad.011895.2624	SC45	2.Con.SC45	S	F	tct	ttt	agatcgaa	agttcgaa	
274	349	2.-.B.9015_07.NAgusta_SC.122797.4754	9015_07	2.Con.9015_07	S	T	tct	act	agatcgaa	agactgaa	
321	396	5.A.B.REJO4541.UAB.092801.1967	REJO4541	5.Con.REJO4541	E	K	gaa	aaa	ggagaata	ggaaaata	APOBEC3G
321	396	4.A.B.SC20.Trinidad.021794.2425	SC20	4.Con.SC20	E	K	gaa	aaa	ggagaata	ggaaaata	APOBEC3G
321	396	2.-.B.ZP63358-04.Charolette_NC.031097.4035	ZP63358-04	2.Con.ZP63358-04	D	N	gac	aac	ggagacta	ggaaacta	
321	396	5.A.B.Z02.NC-Duke.052798.1366	Z02	5.Con.Z02	E	G	gaa	gga	ggagaata	ggaggata	

HXB2 #	Site #	Sequence	Patient ID	Consensus ID	Amino Acid in Consensus	Amino Acid in Patient	Codon in Consensus	Codon in Patient	9 nucleotides of the consensus (centred on the codon of interest)	9 nucleotides of the patient (centred on the codon of interest)	Possible explanation besides selection
337	412	5.C.B.MEMI4948.UAB.012803.4095	MEMI4948	5.Con.MEMI4948	D	N	gat	aat	gcagattgg	gcaaattgg	APOBEC3G
337	412	3.A.B.1012.LongBeach,CA.040797.3284	1012	3.Con.1012	K	R	aaa	aga	tcaaaatgg	tcaagatgg	
337	412	3.-.B.PRB931_06.Milwukee_WI.091195.5178	PRB931_06	3.Con.PRB931_06	K	E	aaa	gaa	ggaaaatgg	ggagaatgg	
337	412	2.A.B.Z31.NC-Duke.120901.990	Z31	2.Con.Z31	D	A	gat	gct	aaagattgg	aaagcttgg	
337	412	2.A.B.Z31.NC-Duke.120901.991	Z31	2.Con.Z31	D	A	gat	gct	aaagattgg	aaagcttgg	
337	412	2.A.B.Z31.NC-Duke.120901.917	Z31	2.Con.Z31	D	A	gat	gct	aaagattgg	aaagcttgg	
337	412	2.A.B.Z31.NC-Duke.120901.924	Z31	2.Con.Z31	D	A	gat	gct	aaagattgg	aaagcttgg	
337	412	2.A.B.Z31.NC-Duke.120901.923	Z31	2.Con.Z31	D	A	gat	gct	aaagattgg	aaagcttgg	
337	412	2.A.B.Z31.NC-Duke.120901.922	Z31	2.Con.Z31	D	A	gat	gct	aaagattgg	aaagcttgg	
337	412	2.A.B.Z31.NC-Duke.120901.920	Z31	2.Con.Z31	D	A	gat	gct	aaagattgg	aaagcttgg	
337	412	2.A.B.Z31.NC-Duke.120901.919	Z31	2.Con.Z31	D	A	gat	gct	aaagattgg	aaagcttgg	
337	412	2.-.B.ZP9025-11.Portland_OR.081698.3802	ZP9025-11	2.Con.ZP9025-11	E	K	gaa	aaa	gaagaatgg	gaaaatgg	APOBEC3G
344	419	3.-.B.700010056.NC.083106.3387	700010056	3.Con.700010058	K	M	aag	atg	aagaagata	aagatgata	
344	419	3.-.B.700010058.NC.083106.4405	700010058	3.Con.700010058	K	M	aag	atg	aagaagata	aagatgata	
344	419	5.E.B.Z13.NC-Duke.051799.836	Z13	5.Con.Z13	Q	L	cag	ctg	agtcaggta	agtctggta	
344	419	3.-.B.ZP6248-07.Dallas_TX.030997.3707	ZP6248-07	3.Con.ZP6248-07	R	G	aga	gga	aaaagagta	aaaggagta	
347	422	4.A.B.1058.NAgusta_SC.031898.1538	1058	4.Con.1058	R	K	aga	aaa	gttagaaaa	gttaaaaaa	APOBEC3G
347	422	3.A.B.1059.LongBeach_CA.032698.1459	1059	3.Con.1059	R	K	aga	aaa	gctagaaga	gctaaaaga	APOBEC3G
347	422	2.-.B.9079_09.LongBeach_CA.121199.4546	9079_09	2.Con.9079_09	K	R	aaa	aga	gcaaaaaaa	gcaagaaaa	
347	422	3.A.B.Z34.NC-Duke.121702.942	Z34	3.Con.Z34	G	E	gga	gaa	gttggaaaa	gttgaaaaa	APOBEC3G
347	422	3.A.B.Z34.NC-Duke.121702.951	Z34	3.Con.Z34	G	E	gga	gaa	gttggaaaa	gttgaaaaa	APOBEC3G
347	422	3.A.B.Z34.NC-Duke.121702.943	Z34	3.Con.Z34	G	E	gga	gaa	gttggaaaa	gttgaaaaa	APOBEC3G
347	422	2.-.B.ZP9029-12.Fayetteville_NC.091098.5207	ZP9029-12	2.Con.ZP9029-12	G	R	gga	aga	gttggaata	gttagaata	
354	431	5.-.B.700010077.NC.090806.4414	700010077	5.Con.700010077	K	R	aag	agg	tttaagaat	tttaggaat	
354	431	5.-.B.700010077.NC.090806.4453	700010077	5.Con.700010077	K	R	aag	agg	tttaagaat	tttaggaat	
354	431	5.-.B.700010077.NC.090806.4430	700010077	5.Con.700010077	K	R	aag	agg	tttaagaat	tttaggaat	
354	431	5.-.B.700010077.NC.090806.4441	700010077	5.Con.700010077	K	T	aag	acg	tttaagaat	tttacgaat	
354	431	5.-.B.700010077.NC.090806.4432	700010077	5.Con.700010077	K	R	aag	agg	tttaagaat	tttaggaat	
354	431	5.-.B.700010077.NC.090806.4457	700010077	5.Con.700010077	K	R	aag	agg	tttaagaat	tttaggaat	
354	431	5.-.B.700010077.NC.090806.4433	700010077	5.Con.700010077	K	R	aag	agg	tttaagaat	tttaggaat	
354	431	5.-.B.700010077.NC.090806.4448	700010077	5.Con.700010077	K	R	aag	agg	tttaagaat	tttaggaat	
354	431	5.-.B.700010077.NC.090806.4426	700010077	5.Con.700010077	K	R	aag	agg	tttaagaat	tttaggaat	
354	431	5.-.B.700010077.NC.090806.4427	700010077	5.Con.700010077	K	R	aag	agg	tttaagaat	tttaggaat	
354	431	5.-.B.700010077.NC.090806.4446	700010077	5.Con.700010077	K	R	aag	agg	tttaagaat	tttaggaat	
354	431	5.-.B.700010077.NC.090806.4444	700010077	5.Con.700010077	K	R	aag	agg	tttaagaat	tttaggaat	
354	431	5.-.B.700010077.NC.090806.4455	700010077	5.Con.700010077	K	R	aag	agg	tttaagaat	tttaggaat	
354	431	5.-.B.700010077.NC.090806.3389	700010077	5.Con.700010077	K	R	aag	agg	tttaagaat	tttaggaat	
354	431	2.-.B.9014_01.NAgusta_SC.111397.4782	9014_01	2.Con.9014_01	E	K	gag	aag	tttgagaat	tttaagaat	APOBEC3G
354	431	4.A.B.SC20.Trinidad.021794.2445	SC20	4.Con.SC20	G	E	ggg	gag	gttgggaat	tttgagaat	APOBEC3G
360	439	2.A.B.TRJO4551.UAB.101001.2080	TRJO4551	2.Con.TRJO4551	V	G	gtc	ggc	atagtcttt	ataggcttt	
360	439	3.A.B.1012.LongBeach,CA.040797.3276	1012	3.Con.1012	V	I	gtc	atc	atagtcttt	ataatcttt	
360	439	5.-.B.700010077.NC.090806.4426	700010077	5.Con.700010077	V	G	gtc	ggc	atagtcttt	ataggcttt	
360	439	5.-.B.700010077.NC.090806.4444	700010077	5.Con.700010077	V	A	gtc	gcc	atagtcttt	atagccttt	
372	454	2.A.B.TT31P.Trinidad.100298.2802	TT31P	2.Con.TT31P	A	V	gca	gta	attgcaatg	attgtaatg	
372	454	2.A.B.TT31P.Trinidad.100298.2799	TT31P	2.Con.TT31P	A	V	gca	gta	attgcaatg	attgtaatg	
372	454	2.A.B.TT31P.Trinidad.100298.5049	TT31P	2.Con.TT31P	A	V	gca	gta	attgcaatg	attgtaatg	
372	454	2.A.B.TT31P.Trinidad.100298.5051	TT31P	2.Con.TT31P	A	V	gca	gta	attgcaatg	attgtaatg	
372	454	2.A.B.TT31P.Trinidad.100298.5048	TT31P	2.Con.TT31P	A	V	gca	gta	attgcaatg	attgtaatg	
372	454	2.A.B.TT31P.Trinidad.100298.5052	TT31P	2.Con.TT31P	A	V	gca	gta	attgcaatg	attgtaatg	
372	454	2.A.B.TT31P.Trinidad.100298.5047	TT31P	2.Con.TT31P	A	V	gca	gta	attgcaatg	attgtaatg	

HXB2 #	Site #	Sequence	Patient ID	Consensus ID	Amino Acid in Consensus	Amino Acid in Patient	Codon in Consensus	Codon in Patient	9 nucleotides of the consensus (centred on the codon of interest)	9 nucleotides of the patient (centred on the codon of interest)	Possible explanation besides selection
372	454	2.A.B.TT31P.Trinidad.100298.2792	TT31P	2.Con.TT31P	A	V	gca	gta	attgcaatg	attgtaatg	
372	454	2.A.B.TT31P.Trinidad.100298.5055	TT31P	2.Con.TT31P	A	V	gca	gta	attgcaatg	attgtaatg	
372	454	2.A.B.TT31P.Trinidad.100298.5050	TT31P	2.Con.TT31P	A	V	gca	gta	attgcaatg	attgtaatg	
372	454	2.A.B.TT31P.Trinidad.100298.5045	TT31P	2.Con.TT31P	A	V	gca	gta	attgcaatg	attgtaatg	
372	454	2.A.B.TT31P.Trinidad.100298.2793	TT31P	2.Con.TT31P	A	V	gca	gta	attgcaatg	attgtaatg	
372	454	2.A.B.TT31P.Trinidad.100298.2785	TT31P	2.Con.TT31P	A	V	gca	gta	attgcaatg	attgtaatg	
372	454	2.A.B.TT31P.Trinidad.100298.5053	TT31P	2.Con.TT31P	A	V	gca	gta	attgcaatg	attgtaatg	
372	454	2.A.B.TT31P.Trinidad.100298.2795	TT31P	2.Con.TT31P	A	V	gca	gta	attgcaatg	attgtaatg	
372	454	2.A.B.TT31P.Trinidad.100298.5046	TT31P	2.Con.TT31P	A	V	gca	gta	attgcaatg	attgtaatg	
372	454	2.A.B.TT31P.Trinidad.100298.5054	TT31P	2.Con.TT31P	A	V	gca	gta	attgcaatg	attgtaatg	
372	454	2.A.B.TT31P.Trinidad.100298.2798	TT31P	2.Con.TT31P	A	V	gca	gta	attgcaatg	attgtaatg	
381	463	3.-.B.ZP6248-07.Dallas_TX.030997.3703	ZP6248-07	3.Con.ZP6248-07	E	K	gaa	aaa	ggggaattt	gggaaattt	APOBEC3G
381	463	3.A.B.1018.Sumter_SC.062097.1756	1018	3.Con.1018	E	K	gaa	aaa	ggggaattt	gggaaattt	APOBEC3G
381	463	3.A.B.1001.StLouis_MO.021297.4151	1001	3.Con.1001	E	K	gaa	aaa	ggggaattt	gggaaattt	APOBEC3G
381	463	3.A.B.1059.LongBeach_CA.032698.1466	1059	3.Con.1059	E	K	gaa	aaa	ggggaattt	gggaaattt	APOBEC3G
460	566	5.C.B.MEMI4948.UAB.012803.3225	MEMI4948	5.Con.MEMI4948	D	N	gat	aat	ggtgatgat	ggtaatggt	APOBEC3G
460	566	5.C.B.MEMI4948.UAB.012803.3237	MEMI4948	5.Con.MEMI4948	D	N	gat	aat	ggtgatgat	ggtaatgat	APOBEC3G
460	566	5.C.B.MEMI4948.UAB.012803.3238	MEMI4948	5.Con.MEMI4948	D	N	gat	aat	ggtgatgat	ggtaatgat	APOBEC3G
460	566	5.C.B.MEMI4948.UAB.012803.4100	MEMI4948	5.Con.MEMI4948	D	K	gat	aag	ggtgatgat	ggtaaggat	APOBEC3G
460	566	5.C.B.MEMI4948.UAB.012803.3234	MEMI4948	5.Con.MEMI4948	D	T	gat	act	ggtgatgat	ggtactgat	
460	566	5.C.B.MEMI4948.UAB.012803.3221	MEMI4948	5.Con.MEMI4948	D	K	gat	aag	ggtgatgat	ggtaaggat	APOBEC3G
460	566	5.C.B.MEMI4948.UAB.012803.3239	MEMI4948	5.Con.MEMI4948	D	N	gat	aat	ggtgatgat	ggtaatgat	APOBEC3G
460	566	4.A.B.SC31.Trinidad.091594.2561	SC31	4.Con.SC31	N	K	aat	aag	ggtaataat	ggtaagaat	
482	601	3.A.B.1001.StLouis_MO.021297.2208	1001	3.Con.1001	E	K	gaa	aaa	agtgaatta	agtaaatta	APOBEC3G
482	601	3.A.B.1018.Sumter_SC.062097.1749	1018	3.Con.1018	E	K	gaa	aaa	agtgaatta	agtaaatta	APOBEC3G
482	601	4.A.B.1058.NAgusta_SC.031898.1538	1058	4.Con.1058	E	K	gaa	aaa	agtgaatta	agtaaatta	APOBEC3G
482	601	4.A.B.1058.NAgusta_SC.031898.1569	1058	4.Con.1058	E	K	gaa	aaa	agtgaatta	agtaaatta	APOBEC3G
482	601	5.-.B.700010077.NC.090806.4451	700010077	5.Con.700010077	E	K	gaa	aaa	agtgaatta	agtaaatta	APOBEC3G
482	601	2.-.B.ZP12007-04.102299.21.3928	ZP12007-04	2.Con.ZP12007-04	E	K	gaa	aaa	agtgaatta	agtaaatta	APOBEC3G
509	634	3.-.B.700010058.NC.083106.4381	700010058	3.Con.700010058	E	K	gaa	aaa	agagaaaaa	gagaaaaaa	Alignment Artifact
509	634	3.A.B.1059.LongBeach_CA.032698.1477	1059	3.Con.1059	E	K	gaa	aaa	agagaaaaa	agaaaaaaa	Alignment Artifact
509	634	2.-.B.62130_04.Youngstown_OH.090396.4766	62130_04	2.Con.62130_04	E	K	gaa	aaa	agagaaaaa	gagaaaaaa	Alignment Artifact
509	634	3.-.B.9032_08.Mobile_AL.073098.4715	9032_08	3.Con.9032_08	E	K	gaa	aaa	agagaaaaa	gagaaaaaa	Alignment Artifact
509	634	2.A.B.SC05.Trinidad.062893.2340	SC05	2.Con.SC05	E	K	gaa	aaa	agagaaaaa	gaaaaaaaa	Alignment Artifact
509	634	2.A.B.TT29P.Trinidad.032098.2781	TT29P	2.Con.TT29P	E	K	gaa	aaa	agagaaaaa	gagaaaaaa	Alignment Artifact
509	634	2.-.B.ZP12007-04.102299.21.5256	ZP12007-04	2.Con.ZP12007-04	E	K	gaa	aaa	agagaaaaa	agaaaaaaa	Alignment Artifact
513	638	1.-.B.ZP62995-05.010397.4.5218	ZP62995-05	1.Con.ZP62995-05	V	G	gtg	ggg	gcagtggga	gcaggggga	
513	638	2.A.B.WEAU.UAB.053090.2198	WEAU	2.Con.WEAU	V	S	gtg	agt	gcagtggga	agcagtgga	Alignment Artifact
518	646	3.A.B.1012.LongBeach.CA.040797.3285	1012	3.Con.1012	M	V	atg	gtg	gctatgttc	gctgtgttc	
518	646	2.-.B.9010_09.NAgusta_SC.112597.5125	9010_09	2.Con.9010_09	M	V	atg	gtg	gctatgttc	gctgtgttc	
518	646	2.-.B.9020_20.LongBeach_CA.061698.4612	9020_20	2.Con.9020_20	M	I	atg	ata	gctatgttc	gctatattc	
518	646	2.-.B.9077_12.Fayetteville_NC.120499.4651	9077_12	2.Con.9077_12	M	V	atg	gtg	gttatgttc	gttgtgttc	
518	646	3.-.B.PRB931_06.Milwukee_WI.091195.5171	PRB931_06	3.Con.PRB931_06	M	V	atg	gtg	gctatgttc	gctgtgttc	
518	646	3.-.B.PRB958_06.Jacksonville_FL.021000.4310	PRB958_06	3.Con.PRB958_06	M	V	atg	gtg	gctatgttc	gctgtgttc	
518	646	2.A.B.SC22.Trinidad.031694.2487	SC22	2.Con.SC22	M	V	atg	gtg	gctatgttc	gctgtgttc	
518	646	5.A.B.TT28P.Trinidad.012096.2737	TT28P	5.Con.TT28P	M	V	atg	gtg	gctatgttc	gctgtgttc	
518	646	5.E.B.Z27.NC-Duke.062001.887	Z27	5.Con.Z27	M	I	atg	ata	gctatgttc	gctatattc	
587	719	5.E.B.Z27.NC-Duke.062001.909	Z27	5.Con.Z27	L	I	cta	ata	tacctaaag	tacataaag	

HXB2 #	Site #	Sequence	Patient ID	Consensus ID	Amino Acid in Consensus	Amino Acid in Patient	Codon in Consensus	Codon in Patient	9 nucleotides of the consensus (centred on the codon of interest)	9 nucleotides of the patient (centred on the codon of interest)	Possible explanation besides selection
587	719	5.E.B.Z13.NC-Duke.051799.854	Z13	5.Con.Z13	L	I	cta	ata	tacctaaag	tacataaag	
588	720	3.-.B.700010058.NC.083106.4371	700010058	3.Con.700010058	R	K	agg	aag	ctaagggac	ctaaaggac	APOBEC3G
588	720	3.A.B.Z20.NC-Duke.110100.1374	Z20	3.Con.Z20	K	R	aag	agg	ctaaaggat	ctagggat	
588	720	4.A.B.SC31.Trinidad.091594.2540	SC31	4.Con.SC31	R	K	agg	aag	ctaagggat	ctaaaggat	APOBEC3G
612	744	5.C.B.MEMI4948.UAB.012803.3231	MEMI4948	5.Con.MEMI4948	T	I	acc	atc	aataccagt	aatatcagt	APOBEC1
612	744	5.E.B.Z13.NC-Duke.051799.850	Z13	5.Con.Z13	I	S	atc	agc	aatatcagt	aatagcagt	
612	744	5.E.B.Z13.NC-Duke.051799.862	Z13	5.Con.Z13	I	T	atc	acc	aatatcagt	aataccagt	
612	744	5.E.B.Z13.NC-Duke.051799.839	Z13	5.Con.Z13	I	T	atc	acc	aatatcagt	aataccagt	
612	744	5.E.B.Z13.NC-Duke.051799.858	Z13	5.Con.Z13	I	T	atc	acc	aatatcagt	aataccagt	
612	744	5.E.B.Z13.NC-Duke.051799.853	Z13	5.Con.Z13	I	T	atc	acc	aatatcagt	aataccagt	
612	744	5.E.B.Z13.NC-Duke.051799.854	Z13	5.Con.Z13	I	S	atc	agc	aatatcagt	aatagcagt	
612	744	5.E.B.Z13.NC-Duke.051799.847	Z13	5.Con.Z13	I	S	atc	agc	aatatcagt	aatagcagt	
612	744	5.E.B.Z13.NC-Duke.051799.851	Z13	5.Con.Z13	I	S	atc	agc	aatatcagt	aatagcagt	
612	744	5.E.B.Z13.NC-Duke.051799.842	Z13	5.Con.Z13	I	T	atc	acc	aatatcagt	aataccagt	
612	744	5.E.B.Z13.NC-Duke.051799.863	Z13	5.Con.Z13	I	T	atc	acc	aatatcagt	aataccagt	
612	744	5.E.B.Z13.NC-Duke.051799.856	Z13	5.Con.Z13	I	T	atc	acc	aatatcagt	aataccagt	
612	744	5.E.B.Z13.NC-Duke.051799.845	Z13	5.Con.Z13	I	S	atc	agc	aatatcagt	aatagcagt	
612	744	5.E.B.Z13.NC-Duke.051799.840	Z13	5.Con.Z13	I	N	atc	aac	aatatcagt	aataacagt	
632	768	3.A.B.1001.StLouis_MO.021297.4146	1001	3.Con.1001	E	K	gaa	aaa	tgggaaaaa	tggaaaaaa	APOBEC3G
632	768	5.A.B.RHPA4259.UAB.120500.1997	RHPA4259	5.Con.RHPA4259	E	K	gaa	aaa	tgggaaaga	tggaaaaga	APOBEC3G
632	768	3.A.B.1018.Sumter_SC.062097.1772	1018	3.Con.1018	E	K	gaa	aaa	tgggaaaga	tggaaaaga	APOBEC3G
632	768	3.A.B.1053.NAgusta_SC.120397.1682	1053	3.Con.1053	E	K	gaa	aaa	tgggaaagg	tggaaaaga	APOBEC3G
648	786	3.-.B.ZP6248-07.Dallas_TX.030997.3703	ZP6248-07	3.Con.ZP6248-07	E	K	gaa	aaa	gaggaatcg	gagaaatcg	APOBEC3G
648	786	3.A.B.1001.StLouis_MO.021297.2242	1001	3.Con.1001	E	K	gaa	aaa	gaagaatca	gaaaaatca	APOBEC3G
648	786	4.A.B.1058.NAgusta_SC.031898.1563	1058	4.Con.1058	E	K	gaa	aaa	ggagaatcg	ggaaaatcg	APOBEC3G
648	786	4.A.B.SC31.Trinidad.091594.2547	SC31	4.Con.SC31	D	N	gat	aat	gaagatgcg	gaaaatgcg	APOBEC3G
651	789	2.-.B.PRB956_04.Richmond_VA.081997.4245	PRB956_04	2.Con.PRB956_04	N	S	aac	agc	caaaaccaa	caaagccaa	
651	789	3.-.B.PRB931_06.Milwukee_WI.091195.5169	PRB931_06	3.Con.PRB931_06	N	S	aac	agc	caaaaccaa	caaagccaa	
651	789	3.-.B.PRB931_06.Milwukee_WI.091195.5177	PRB931_06	3.Con.PRB931_06	N	S	aac	agc	caaaaccaa	caaagccaa	
651	789	5.-.B.700010077.NC.090806.4452	700010077	5.Con.700010077	S	G	agc	ggc	caaagccag	caaggccag	
696	834	2.A.B.SC22.Trinidad.031694.2460	SC22	2.Con.SC22	R	K	aga	aaa	ttaagaata	ttaaaaata	APOBEC3G
696	834	5.-.B.700010077.NC.090806.4420	700010077	5.Con.700010077	R	K	aga	aaa	ttaagaata	ttaaaaata	APOBEC3G
696	834	5.A.B.THRO4156.UAB.080100.2027	THRO4156	5.Con.THRO4156	R	S	aga	agt	ttaagaata	ttaagtata	
696	834	5.E.B.Z13.NC-Duke.051799.845	Z13	5.Con.Z13	R	K	aga	aag	ttaagaata	ttaaagata	
700	838	2.A.B.TT31P.Trinidad.100298.2785	TT31P	2.Con.TT31P	A	V	gct	gtt	tttgctgtg	tttgttgtg	
700	838	2.A.B.TT31P.Trinidad.100298.2792	TT31P	2.Con.TT31P	A	V	gct	gtt	tttgctgtg	tttgttgtg	
700	838	2.A.B.TT31P.Trinidad.100298.2793	TT31P	2.Con.TT31P	A	V	gct	gtt	tttgctgtg	tttgttgtg	
700	838	2.A.B.TT31P.Trinidad.100298.2796	TT31P	2.Con.TT31P	A	V	gct	gtt	tttgctgtg	tttgttgtg	
700	838	2.A.B.TT31P.Trinidad.100298.2799	TT31P	2.Con.TT31P	A	V	gct	gtt	tttgctgtg	tttgttgtg	
700	838	2.A.B.TT31P.Trinidad.100298.2801	TT31P	2.Con.TT31P	A	V	gct	gtt	tttgctgtg	tttgttgtg	
700	838	2.A.B.TT31P.Trinidad.100298.2802	TT31P	2.Con.TT31P	A	V	gct	gtt	tttgctgtg	tttgttgtg	
700	838	2.A.B.TT31P.Trinidad.100298.5045	TT31P	2.Con.TT31P	A	V	gct	gtt	tttgctgtg	tttgttgtg	
700	838	2.A.B.TT31P.Trinidad.100298.5046	TT31P	2.Con.TT31P	A	V	gct	gtt	tttgctgtg	tttgttgtg	
700	838	2.A.B.TT31P.Trinidad.100298.5047	TT31P	2.Con.TT31P	A	V	gct	gtt	tttgctgtg	tttgttgtg	
700	838	2.A.B.TT31P.Trinidad.100298.5048	TT31P	2.Con.TT31P	A	V	gct	gtt	tttgctgtg	tttgttgtg	
700	838	2.A.B.TT31P.Trinidad.100298.5049	TT31P	2.Con.TT31P	A	V	gct	gtt	tttgctgtg	tttgttgtg	
700	838	2.A.B.TT31P.Trinidad.100298.5050	TT31P	2.Con.TT31P	A	V	gct	gtt	tttgctgtg	tttgttgtg	
700	838	2.A.B.TT31P.Trinidad.100298.5052	TT31P	2.Con.TT31P	A	V	gct	gtt	tttgctgtg	tttgttgtg	

HXB2 #	Site #	Sequence	Patient ID	Consensus ID	Amino Acid in Consensus	Amino Acid in Patient	Codon in Consensus	Codon in Patient	9 nucleotides of the consensus (centred on the codon of interest)	9 nucleotides of the patient (centred on the codon of interest)	Possible explanation besides selection
700	838	2.A.B.TT31P.Trinidad.100296.5053	TT31P	2.Con.TT31P	A	V	gct	gtt	tttgctgtg	tttgttgtg	
700	838	2.A.B.TT31P.Trinidad.100296.5054	TT31P	2.Con.TT31P	A	V	gct	gtt	tttgctgtg	tttgttgtg	
700	838	2.A.B.TT31P.Trinidad.100296.5055	TT31P	2.Con.TT31P	A	V	gct	gtt	tttgctgtg	tttgttgtg	
700	838	2.A.B.TT31P.Trinidad.100296.5057	TT31P	2.Con.TT31P	A	V	gct	gtt	tttgctgtg	tttgttgtg	
700	838	2.A.B.TT31P.Trinidad.100296.5074	TT31P	2.Con.TT31P	A	V	gct	gtt	tttgctgtg	tttgttgtg	
700	838	2.A.B.TT31P.Trinidad.100296.5081	TT31P	2.Con.TT31P	A	V	gct	gtt	tttgctgtg	tttgttgtg	
700	838	3.-.B.PRB931_06.Milwukee_WI.091195.5173	PRB931_06	3.Con.PRB931_06	T	S	act	tct	tttactgta	ttttctgta	
700	838	5.E.B.Z13.NC-Duke.051799.844	Z13	5.Con.Z13	A	T	gct	act	tttgctata	tttactata	
702	840	2.-.B.PRB959_02.Columbia_SC.111799.4350	PRB959_02	2.Con.PRB959_02	L	P	ctt	cct	gtgctttct	gtgccttct	
702	840	4.A.B.1058.NAgusta_SC.031898.1568	1058	4.Con.1058	L	F	ctt	ttt	gtactttct	gtattttct	
702	840	5.A.B.REJO4541.UAB.092801.1968	REJO4541	5.Con.REJO4541	L	F	ctt	ttt	gtactttct	gtattttct	
703	841	2.A.B.1056.NAgusta_SC.011498.1841	1056	2.Con.1056	S	F	tct	ttt	ctttctata	cttttata	
703	841	2.-.B.ZP63358-04.Charolette_NC.031097.4025	ZP63358-04	2.Con.ZP63358-04	S	P	tct	cct	ctttctata	cttcctata	
703	841	5.-.B.700010040.NC.072706.4516	700010040	5.Con.700010040	S	A	tct	gct	ctttctata	cttgctata	
817	965	2.-.B.9015_07.NAgusta_SC.122797.4750	9015_07	2.Con.9015_07	T	I	aca	ata	aatacaaca	aatataaca	APOBEC1
817	965	5.E.B.Z13.NC-Duke.051799.848	Z13	5.Con.Z13	A	T	gcc	acc	aacgccaca	aacaccaca	
831	979	2.A.B.TT31P.Trinidad.100296.5091	TT31P	2.Con.TT31P	E	K	gaa	aaa	gtagaagta	gtaaaagta	APOBEC3G
831	979	5.C.B.MEMI4948.UAB.012803.3221	MEMI4948	5.Con.MEMI4948	E	K	gaa	aaa	atagaagta	ataaaagta	APOBEC3G
831	979	5.C.B.MEMI4948.UAB.012803.3235	MEMI4948	5.Con.MEMI4948	E	K	gaa	aaa	atagaagta	ataaaagta	APOBEC3G
831	979	5.C.B.MEMI4948.UAB.012803.3238	MEMI4948	5.Con.MEMI4948	E	K	gaa	aaa	atagaagta	ataaaagta	APOBEC3G
831	979	5.C.B.MEMI4948.UAB.012803.4100	MEMI4948	5.Con.MEMI4948	E	K	gaa	aaa	atagaagta	ataaaagta	APOBEC3G
833	981	5.C.B.MEMI4948.UAB.012803.3224	MEMI4948	5.Con.MEMI4948	A	V	gca	gta	gtagcacaa	gtagtacaa	
833	981	5.C.B.MEMI4948.UAB.012803.3225	MEMI4948	5.Con.MEMI4948	A	V	gca	gta	gtagcacaa	gtagtacaa	
833	981	5.C.B.MEMI4948.UAB.012803.3238	MEMI4948	5.Con.MEMI4948	A	V	gca	gta	gtagcacaa	gtagtacaa	
833	981	5.C.B.MEMI4948.UAB.012803.3235	MEMI4948	5.Con.MEMI4948	A	V	gca	gta	gtagcacaa	gtagtacaa	
833	981	5.C.B.MEMI4948.UAB.012803.4100	MEMI4948	5.Con.MEMI4948	A	V	gca	gta	gtagcacaa	gtagtacaa	
833	981	5.C.B.MEMI4948.UAB.012803.3236	MEMI4948	5.Con.MEMI4948	A	V	gca	gta	gtagcacaa	gtagtacaa	
833	981	5.C.B.MEMI4948.UAB.012803.3234	MEMI4948	5.Con.MEMI4948	A	V	gca	gta	gtagcacaa	gtagtacaa	
833	981	5.C.B.MEMI4948.UAB.012803.4102	MEMI4948	5.Con.MEMI4948	A	V	gca	gta	gtagcacaa	gtagtacaa	
833	981	5.C.B.MEMI4948.UAB.012803.3241	MEMI4948	5.Con.MEMI4948	A	V	gca	gta	gtagcacaa	gtagtacaa	
833	981	5.C.B.MEMI4948.UAB.012803.3221	MEMI4948	5.Con.MEMI4948	A	V	gca	gta	gtagcacaa	gtagtacaa	
833	981	5.C.B.MEMI4948.UAB.012803.3231	MEMI4948	5.Con.MEMI4948	A	G	gca	gga	gtagcacaa	gtaggacaa	
833	981	5.A.B.TT34P.Trinidad.011398.2830	TT34P	5.Con.TT34P	A	V	gca	gta	gcggcacaa	gcggtacaa	
841	989	3.A.B.1018.Sumter_SC.062097.1767	1018	3.Con.1018	L	P	ctc	ccc	attctccac	attccccac	
841	989	5.C.B.MEMI4948.UAB.012803.4097	MEMI4948	5.Con.MEMI4948	L	H	ctc	cac	attctcaac	attcacaac	
841	989	2.-.B.ZP61792-03.Phoenix_TX.072996.3954	ZP61792-03	2.Con.ZP61792-03	I	T	atc	acc	attatccac	attacccac	
841	989	2.-.B.ZP9025-11.Portland_OR.081698.3793	ZP9025-11	2.Con.ZP9025-11	L	F	ctc	ttc	attctccac	attttccac	

Table A2: Sequence names and corresponding Genbank accession numbers for the HIV and SIV sequences used for the evaluation of zoonosis.

#	Sequence Name	Genbank Accession Number
1	1_A1_AU	DQ676872
2	2_A1_KE	AF004885
3	3_A1_RW	AB253421
4	4_A1_UG	AB253429
5	5_A2_CD	AF286238
6	6_A2_CY	AF286237
7	7_B_FR	K03455
8	8_B_TH	AY173951
9	9_B_US	AY331295
10	10_B_US	DQ853463
11	11_C_BR	U52953
12	12_C_ET	U46016
13	13_C_ZA	AY772699
14	14_D_CD	K03454
15	15_D_TZ	AY253311
16	16_D_UG	U88824
17	17_F1_BE	AF077336
18	18_F1_BR	AF005494
19	19_F1_FI	AF075703
20	20_F1_FR	AJ249238
21	21_F2_CM	AF377956
22	22_G_BE	AF084936
23	23_G_KE	AF061641
24	24_G_NG	U88826
25	25_G_PT	AY612637
26	26_H_BE	AF190127
27	27_H_BE	AF190128
28	28_H_CF	AF005496
29	29_J_SE	AF082394
30	30_J_SE	AF082395
31	31_K_CD	AJ249235
32	32_K_CM	AJ249239
33	33_CPZ_CD	U42720
34	34_CPZ_CM	DQ373066
35	35_CPZ_US	AF103818
36	36_CPZ_CM	AY169968
37	37_CPZ_CM	DQ373065
38	38_CPZ_CM	DQ373064
39	39_CPZ_CM	DQ373063
40	40_CPZ_GA	X52154
41	41_CPZ_TZ	AF447763