SELECTION ALONG THE HIV-1 GENOME THROUGH THE CTL MEDIATED IMMUNE RESPONSE

Duncan Palmer

St Cross College



Department of Statistics

Doctor of Philosophy, D.Phil

ABSTRACT

Selection along the HIV-1 genome through the CTL mediated immune response

Duncan Palmer
St Cross College
Doctor of Philosophy, D.Phil
Michaelmas 2014.

During human immunodeficiency virus 1 (HIV-1) infection, the viral population is in constant battle with the host immune system. The cytotoxic T-lymphocyte (CTL) response, a branch of the adaptive immune response, is implicated in viral control and can drive viral evolution in the infected host population. Endogenous viral peptides, or 'epitopes', are presented to CTLs by human leukocyte antigen (HLA) class I molecules on the surface of infected cells where they may be identified as non-self. Mutations in or proximal to a viral epitope can result in 'escape' from CTLs targeting that epitope. The repertoire of epitopes which may be presented is dependent upon host class I HLA types. As such, reversion may occur after transmission due to changes in viral fitness and selection in the context of a new HLA background. Thus, parameters describing the dynamics of CTL escape and reversion are key to understanding how CTL responses within individuals relate to HIV-1 sequence evolution in the infected host population.
Escape and reversion can be studied directly using biological assays and longitudinal viral sequence data, or indirectly by considering viral sequences across multiple hosts. Indirect approaches include tree based methods which detect associations between host HLA and viral sequence but do not estimate rates of escape and reversion, and ordinary differential equation (ODE) models which estimate these rates but do not consider the dependency structure inherent in viral sequence data. We introduce two models which estimate escape and reversion rates whilst accounting for the shared ancestry of viral sequence data.
For our first model, we lay out an integrated Bayesian approach which combines genealogical inference and an existing epidemiological model to inform escape and reversion rate estimates. Using this model, we find evidence for correlation between escape rate estimates across widely separated geographical regions. We also observe a non-linear negative correlation between *in vitro* replicative capacity and escape rate. Both findings suggest that epistasis does not play a strong role in the escape process.
Although our first model worked well, it had some key limitations which we address in our second method. Notably, by making a series of approximations, we are able account for recombination and analyse very large datasets which would be computationally infeasible under the first model. We verify our second approach through extensive simulations, and use the method to estimate both drug and HLA associated selection along portions of the HIV-1 genome. We test the results of the model using existing knowledge, and determine a collection of putative selected sites which warrant further investigation. Finally, we find evidence to support the notion that the CTL response played a role in HIV-1 subtype diversification.

ACKNOWLEDGEMENTS

I'd first like to thank my two supervisors Angela and Gil for their support, insight and helpful discussion. I couldn't have asked for a better pair of supervisors for my time here at Oxford. Thanks to everyone that's helped and made me laugh along the way; both in Zoology and up the hill. In particular, I want to thank Greg, Jess, Isaac, Struan, Ollie, Jonas and James. Without you guys the last five years wouldn't have been nearly as much fun. Finally, I want to thank all of my colleagues for putting up with my increasingly long hair and thesis beard. Time to crawl outside to get a haircut.

CONTENTS

1 INTRODUCTION 1

2 LITERATURE REVIEW 7

2.1 The Biology of HIV-1 8

2.1.1 Structure 8

2.1.2 Life cycle 9

2.1.3 Viral Components 11

2.2 Genetic variation in other infections 13

2.2.1 Measles 14

2.2.2 Influenza 14

2.2.3 Hepatitis C virus 15

2.2.4 Tuberculosis 15

2.2.5 Trypanosomes 16

2.3 Genetic variation in HIV-1 17

2.4 Summary of immune responses to HIV-1 19

2.5 Epitopes and escape 22

2.6 HIV-1 modelling 26

2.6.1 Within host ODE models of HIV-1 27

2.6.2 A model of CTL escape, transmission and reversion in the infected human population 31

2.7 The Coalescent, tree peeling, and model based inference 33

2.8 Incorporating the genealogy in parameter estimation 37

3 INTEGRATING GENEALOGICAL AND DYNAMICAL MODELLING 39

3.1 Introduction 39

3.2 Methods 44

3.3 Results 47

3.3.1 Simulations 47

3.3.2 Simulation results 49

3.3.3 Analysis of real data 62

3.4 Discussion 68

4 INTEGRATING GENEALOGICAL AND DYNAMICAL MODELLING: SUPPLEMENT 73

4.1 Introduction 73

4.2 Justification of combining the peeling algorithm and BEAST 74

4.3 Rewriting the sampled reconstructed process 76

4.4 Detailed Methods 80

4.4.1 Data 81

4.4.2 Definition of escape mutation 82

4.4.3 Reference sequences 82

4.4.4 Alignment 83

4.4.5 Data trimming 83

4.4.6 Nucleotide substitution model 83

4.4.7 Sampling from the posterior tree distribution 84

4.4.8 Peeling algorithm 84

4.4.9 Checking agreement between effective infected population and infected population 88

4.4.10 Integrating over trees 89

4.4.11 Credible regions 93

4.4.12 Mahalanobis distance 93

4.5 Equivalence of ODE model and composite likelihood 93

4.6 Justification of validity of hypothesis testing 94

4.7 Connections between tree likelihoods 95

4.8 Derivation of lineage survival in a deterministic model 98

5 MEASURING FITNESS ACROSS TECHNIQUES AND POPULATIONS 101

5.1 Introduction 101

5.2 Methods 102

5.2.1 Cohorts 102

5.2.2 Estimating escape and reversion rates 105

5.2.3 *In vitro* fitness assays 106

5.3 Results 109

5.3.1 Correlation between MAP and ODE estimates within populations 112
5.3.2 Correlation between rates across cohorts 114
5.3.3 Escape and reversion rates within cohorts vs *in vitro* replicative capacity estimates 118
5.3.4 Association between *in vitro* estimates of replicative capacity 124
5.4 Discussion 124
6 HYPOTHESIS FREE ESTIMATION OF HLA ASSOCIATED SELECTION 137
6.1 Introduction 137
6.2 Description of the model 140
6.2.1 Summary 140
6.2.2 Filling in the detail 142
6.2.3 Part 1: Approximating the coalescent with recombination 143
6.2.4 Part 2: Emission probabilities - A codon model of substitution 148
6.3 Inference 155
6.3.1 MCMC moves 157
6.3.2 Window Moves 159
6.4 Code and Memory Constraints 163
6.4.1 Restricting $\mathbf{D}_B$ 164
6.5 Simulations 165
6.6 Discussion and Future Work 183
7 HYPOTHESIS FREE ESTIMATION OF HLA ASSOCIATED SELECTION: REAL DATA 187
7.1 Introduction 187
7.2 Drug resistance study 188
7.2.1 Data 189
7.2.2 Data preparation and performing inference via MCMC 190
7.2.3 Results 191
7.3 HLA associated selection 204
7.3.1 Sources used for our reference datasets 204

7.3.2 Summaries of paired host HLA type and viral sequence data used for query datasets 205
7.3.3 Data preparation 206
7.3.4 Checking for potential sources of bias 207
7.3.5 Results 208
7.3.6 Famous epitopes 210
7.3.7 Putative epitopes 218
7.3.8 Drug resistance mutations 225
7.3.9 Differences between HIV-1 subtypes 227
7.4 Discussion 233
8 DISCUSSION AND FUTURE WORK 239
8.1 Discussion 239
8.2 Future work 244
A HYPOTHESIS FREE ESTIMATION OF HLA ASSOCIATED SELECTION 249
A.1 Expected divergence to time to coalescent tree 249
A.2 Simulations using the gold standard restriction of $\mathbf{D}_B$ 250
B HYPOTHESIS FREE ESTIMATION OF HLA ASSOCIATED SELECTION: REAL DATA 255
B.1 Detailed description of query and reference dataset preparation 255
B.1.1 Drug associated selection analyis 255
B.1.2 HLA associated selection analysis 256
B.2 Supplementary figures and tables 260
BIBLIOGRAPHY 275

LIST OF FIGURES

Figure 2.1 Diagram of HIV-1 virion. 9

Figure 2.2 HIV genome diagram. 11

Figure 2.3 Summary of the early immune response to HIV-1. 19

Figure 2.4 Shoulder region. 29

Figure 3.1 Summary of the inference problem. 43

Figure 3.2 Simulation results. 50

Figure 3.3 Four simulations with fixed underlying escape and reversion rates: $(\lambda_{esc}, \lambda_{rev}) = (2, 0.5)\ years^{-1}$ 51

Figure 3.4 Four simulations with fixed underlying escape and reversion rates: $(\lambda_{esc}, \lambda_{rev}) = (0.5, 0.5)\ years^{-1}$ 52

Figure 3.5 Four simulations with fixed underlying escape and reversion rates: $(\lambda_{esc}, \lambda_{rev}) = (0.5, 0.05)\ years^{-1}$ 53

Figure 3.6 Four simulations with fixed underlying escape and reversion rates: $(\lambda_{esc}, \lambda_{rev}) = (0.05, 0.05)\ years^{-1}$ 54

Figure 3.7 Four simulations with fixed underlying escape and reversion rates: $(\lambda_{esc}, \lambda_{rev}) = (0.05, 0.01)\ years^{-1}$ 55

Figure 3.8 Histograms of MAP estimates obtained using our peeling algorithm and ODE point estimates: 60 sampled tips. 56

Figure 3.9 Histograms of MAP estimates obtained using our peeling algorithm and ODE point estimates: 100 sampled tips. 57

Figure 3.10 Histograms of MAP estimates obtained using our peeling algorithm and ODE point estimates: 200 sampled tips. 58

Figure 3.11 Histograms of MAP estimates obtained using our peeling algorithm and ODE point estimates: 500 sampled tips. 59

Figure 3.12 Shuffling simulation. 62

Figure 3.13 Impact of altering mutation rate upon escape and reversion rate estimates. 63

Figure 3.14 Real data: Maximum clade credibility tree examples and credible regions. 65
Figure 4.1 Summary of escape mutation definition. 82
Figure 4.2 Unseen transmission events between nodes. 87
Figure 4.3 Determining the MAP with simple hill-climbing. 90
Figure 4.4 Bivariate normal approximation to the likelihood surface. 91
Figure 5.1 Site directed mutagenesis followed by replicative capacity assay. 108
Figure 5.2 Replicative capacity assay applied to viral samples taken from hosts 110
Figure 5.3 Scatterplot of escape proportions in HLA matched and HLA mismatched hosts, coloured by population. 111
Figure 5.4 Correlation between the MAP and ODE estimates of escape and reversion rates in different populations. 113
Figure 5.5 Correlation between time to escape estimated from SSITT data and time to escape estimated from the HOMER cohort. 116
Figure 5.6 Correlation between time to escape estimated from SSITT/HOMER data and the Bloemfontein portion of the South African dataset. 117
Figure 5.7 Correlation between time to escape/reversion estimated from Bloemfontein data using the integrated method, and from all available African data using a naive approach. 117
Figure 5.8 Correlation between time to escape estimated from SSITT/HOMER data and all available African data using a naive approach. 118
Figure 5.9 Correlation between time to reversion estimated from SSITT data and time to reversion estimated from the HOMER cohort. 119
Figure 5.10 Correlation between time to reversion estimated from SSITT/HOMER data and the Bloemfontein portion of the South African dataset. 120

Figure 5.11 Scatterplots of time to escape and reversion estimates against *in vitro* viral replicative capacity. 122

Figure 5.12 Scatterplots of time to escape and reversion estimates against *in vitro* viral replicative capacity: Bloemfontein. 123

Figure 5.13 Correlation between time to escape/reversion estimated from all African data using naive ODE approach and replicative capacity as measured by *in vitro* fitness assays. 123

Figure 5.14 Correlation between the average fitness of strains containing an escape mutation in South African replicative capacity data, and estimated replicative capacity cost determined by site directed mutagenesis. 125

Figure 5.15 Summary of all associations between expected time to escape and *in vitro* replicative capacity, according to the Spearman rank correlation coefficient. Significant associations (p<0.05) are coloured according to the Colombo, non-significant associations are shown in white. 128

Figure 6.1 Overview of the model. 141

Figure 6.2 The product of approximate conditionals (PAC) likelihood. 144

Figure 6.3 HLA associated selection profiles 166

Figure 6.4 Hamming distance of protease sequence data in $\mathbf{D}$ from $\mathbf{D}_B$. 167

Figure 6.5 Simulations to evaluate HLA associated selection across a region. $r = 0$ 169

Figure 6.6 Simulations to evaluate HLA associated selection across a region. $r = 0$ 170

Figure 6.7 Simulations to evaluate HLA associated selection across a region. $r = 0$ 171

Figure 6.8 Estimating scaling of selection due to reversion (ζ) and recombination probability across a region; $r = 0$. 172

Figure 6.9 Simulations to evaluate HLA associated selection across a region. $r = 0.01$ 173

Figure 6.10 Simulations to evaluate HLA associated selection across a region. $r = 0.01$ 174

Figure 6.11 Simulations to evaluate HLA associated selection across a region. $r = 0.01$ 175

Figure 6.12 Estimating scaling of selection due to reversion (ζ) and recombination probability across a region; $r = 0.01$. 176

Figure 6.13 Simulations to evaluate HLA associated selection across a region. $r = 0.05$ 177

Figure 6.14 Simulations to evaluate HLA associated selection across a region. $r = 0.05$ 178

Figure 6.15 Simulations to evaluate HLA associated selection across a region. $r = 0.05$ 179

Figure 6.16 Estimating scaling of selection due to reversion (ζ) and recombination probability across a region; $r = 0.05$. 180

Figure 6.17 Examples of estimates of ω across the region 181

Figure 6.18 Effect of having true sequences copied from in the 100 closest sequences 182

Figure 7.1 Results of an MCMC run applied to paired drug regime and protease sequence data taken from the Stanford drug resistance database. 194

Figure 7.2 Results of an MCMC run applied to paired drug regime and protease sequence data taken from the Stanford drug resistance database (continued). 195

Figure 7.3 Results of an MCMC run applied to paired drug regime and protease sequence data taken from the Stanford drug resistance database: PIs (continued). 196

Figure 7.4 Results of an MCMC run applied to paired drug regime and reverse transcriptase sequence data taken from the Stanford drug resistance database: NRTIs. 197

Figure 7.5 Results of an MCMC run applied to paired drug regime and reverse transcriptase sequence data taken from the Stanford drug resistance database: NRTIs (continued). 198

Figure 7.6 Results of an MCMC run applied to paired drug regime and reverse transcriptase sequence data taken from the Stanford drug resistance database: NRTIs (continued). 199
Figure 7.7 Results of an MCMC run applied to paired drug regime and reverse transcriptase sequence data taken from the Stanford drug resistance database: NNRTIs. 200
Figure 7.8 Results of an MCMC run applied to paired drug regime and reverse transcriptase sequence data taken from the Stanford drug resistance database: NNRTIs (continued), and ntRTIs. 201
Figure 7.9 Subtype of closest sequences by Hamming distance 209
Figure 7.10 B*51 associated selection within TAFTIPSI. 211
Figure 7.11 Summary of HLA-B associated selection in reverse transcriptase. 214
Figure 7.12 Summary of HLA-B associated selection in protease. 215
Figure 7.13 A-list and B-list epitopes across reverse transcriptase: HLA-B alleles 219
Figure 7.14 A-list and B-list epitopes across protease: HLA-B alleles 220
Figure 7.15 Motif scan across reverse transcriptase: HLA-B alleles 223
Figure 7.16 Distributions of HLA-B types of subtype B infected hosts and subtype C infected hosts. 227
Figure 7.17 Flow diagram: Is HIV-1 diversification at the level of subtype partially driven by host HLA? 228
Figure 7.18 Codon distributions at sites which distinguish subtype B and subtype C consensus at the amino acid level: protease 230
Figure 7.19 Scatter-plot showing selection away from subtype B consensus conditional on HLA distribution: Protease, HLA B. 231
Figure A.1 Estimating HLA associated selection across a region using sequences copied from. $r = 0.05$ 251

Figure A.2 Estimating HLA associated selection across a region using sequences copied from. $r = 0.05$ 252
Figure A.3 Estimating HLA associated selection across a region using sequences copied from. $r = 0.05$ 253
Figure A.4 Estimating scaling of selection due to reversion and recombination probability across a region; $r = 0.05$. 254
Figure B.1 Results are consistent across independent MCMC runs 260
Figure B.2 Summary of HLA-A associated selection in protease. 261
Figure B.3 Summary of HLA-C associated selection in protease. 262
Figure B.4 Summary of HLA-A associated selection in reverse transcriptase. 263
Figure B.5 Summary of HLA-C associated selection in reverse transcriptase. 264
Figure B.6 A-list and B-list epitopes across protease: HLA-A alleles 265
Figure B.7 A-list and B-list epitopes across protease: HLA-C alleles 266
Figure B.8 A-list and B-list epitopes across reverse transcriptase: HLA-A alleles 267
Figure B.9 A-list and B-list epitopes across reverse transcriptase: HLA-C alleles 268
Figure B.10 Distribution of HLA-A and HLA-C types across subtype B and subtype C populations 268
Figure B.11 Scatterplots showing selection away from a subtype B consensus conditional on HLA frequencies: protease, HLA-A and HLA-C 269
Figure B.12 Scatterplots showing selection away from subtype C consensus conditional on HLA frequencies: protease 270
Figure B.13 Scatterplots showing selection away from a subtype B consensus conditional on HLA frequencies: reverse transcriptase. 271

Figure B.14 Scatterplots showing selection away from subtype C consensus conditional on HLA frequencies: reverse transcriptase. 272

Figure B.15 Codon distribution at subtype B and subtype C seqgregating sites: reverse transcriptase 273

LIST OF TABLES

Table 2.1 Variables and parameters for model in [85]. 33

Table 3.1 Summary of escape mutations analysed. 67

Table 4.1 Summary of the notation used in the detailed methods. 80

Table 4.1 Summary of the notation used in the detailed methods. 81

Table 5.1 Summary of three studies from which *gag* sequence data was available 104

Table 5.2 Fitness costs of epitopes in Gag. MAP and ODE *in vivo* estimates of escape and reversion rates for the collection of epitopes in [30], obtained using Swiss based subtype B SSITT cohort data. 130

Table 5.2 Summary of escape and reversion rate estimates for known epitopes in Gag: Estimated using SSITT data. 131

Table 5.3 Summary of escape and reversion rate estimates for known epitopes in Gag: Estimated using Bloemfontein portion of the South African dataset. 132

Table 5.3 Summary of escape and reversion rate estimates for known epitopes in Gag: Estimated using Bloemfontein portion of the South African dataset. 133

Table 5.4 Summary of escape and reversion rate estimates for known epitopes in Gag: Estimated using sequence data from a cohort based in British Columbia, Canada. 134

Table 5.4 Summary of escape and reversion rate estimates for known epitopes in Gag: Estimated using sequence data from a cohort based in British Columbia, Canada. 135

Table 6.1 Priors on parameters of the model 167

Table 7.1 Estimates of sensitivity to detect major drug resistance sites 193

Table 7.2 Summary of significant unknown drug associated mutations in protease 202

Table 7.3 Summary of significant unknown drug associated mutations in reverse transcriptase 203

Table 7.4 Famous epitopes 212

Table 7.5 p-values for overlap using Test 2 224

Table 7.6 Summary of significant unknown HLA associated mutations in protease and reverse transcriptase. 226

Table 7.7 p-values for observing a value of A at least as great as that seen in the data. 232

Table 7.8 Top-tier and second-tier candidate sites for HLA associated selection. 234

Table B.1 Significant correlations between HLA type occurences. 266

Table B.2 Significant correlations between HLA type occurences. 269

Table B.3 Top-tier and second-tier candidate sites for HLA-A associated selection. 271

Table B.4 Top-tier and second-tier candidate sites for HLA-C associated selection. 274

1

INTRODUCTION

This thesis is about the evolution of human immunodeficiency virus-1 (HIV-1). It particularly focuses upon evolutionary escape from host immunity mediated by cytotoxic T-lymphocytes (CTLs). To do this, we examine the interaction between HIV-1 viral populations and the human leukocyte antigen (HLA) class I molecules within collections of hosts.

In HIV infected individuals there is a constant arms race between the virus on one side and the host immune system on the other. The CTL response is part of the adaptive immune system. Within most nucleated cells bits of protein residing in the cytoplasm are chopped up and presented on the outside of the cell by complexes known as HLA class I molecules. Upon presentation, CTLs may recognise these bits of protein as foreign intruders and destroy the presenting cell. Mutations within or close to these small strings of amino acids (known as epitopes) can result in the loss of the ability of HLAs to present them, or in CTLs no longer recognising them as non-self upon presentation. This phenomenon is known as CTL escape. Importantly, the repertoire of viral epitopes that may be presented is dependent upon the 'type' of the HLA class I molecule. If an epitope can be presented, it is said to be 'restricted' by that HLA type. There are many different HLA types. In fact, the HLA genes which encode this epitope presenting cell machinery are the most variable in the entire human genome. This diversity serves the human population well - it ensures that immune evasion by a viral population residing in one host does not lead to immune evasion across all potential hosts. Add to this the fact that mutations which result in CTL immune escape often incur a fitness cost to the virus, and we have a hugely interesting collection of complex processes across time and length scales to make sense of.

At first, publications in the field were generally restricted to individual case stud-

ies. In these initial investigations, patients' blood was sampled longitudinally and tested to confirm if escape from a particular HLA type had occurred *in vivo*. Such studies were groundbreaking as they brought CTL escape to the attention of the wider HIV community. They also posed a series of questions:

- Could CTL escape mutations be transmitted?
- Would they spread between hosts and propagate through the entire viral population?
- Would future HIV-1 strains be pre-adapted to the CTL response in large portions of the human population?

Research papers in which larger numbers of patients were analysed arrived later and sought to answer these questions. It was found that it is indeed possible for CTL escape mutations to be transmitted between hosts, but reversion can occur if the recipient's HLA types do not match those of the transmitter. This reversion suggested a replicative fitness cost imposed by escape and helped explain why the process could be readily observed *in vivo*: escape mutations do not simply fix in within host viral populations and then spread throughout the infected host population, there are multiple processes at play.

To ascertain whether an escape mutation is likely to hit a high frequency in the HIV-1 infected host population Fryer *et al.* [85] used a differential equation model to estimate trajectories of CTL escape mutation frequencies in the host population through time. The combination of the three integral processes: transmission, escape and reversion could be written down as a system of just five differential equations. Fitting these equations to real data also allowed the authors to estimate the average rate of escape and reversion of escape mutations. However, in order to do so they made a series of simplifying assumptions that need to be addressed. Perhaps the most important of these is that observations of viral sequence data taken from a collection of hosts in the infected population are assumed to be independent. This is not true. Cross-sectional viral sequences have a dependency structure: their shared history. This presents a collection of obvious questions:

- Can CTL escape and reversion rates be estimated using a model that correctly accounts for the shared ancestry of viral sequences?
- Can such a model reliably recapitulate known escape and reversion rates from simulated data?
- How does accounting for this dependency structure change estimates of escape and reversion rates when using real data?

We address these questions in Chapters 3 and 4 by constructing and implementing a model that incorporates the shared history of viral sequence samples. This is achieved by estimating the viral genealogy, and taking it into consideration when performing rate estimation. We perform a large number of simulation studies to test our approach, and compare our results to those of the model that makes the independence assumption. We then use our model to estimate population escape and reversion rates for a series of well known HLA/epitope pairs.

HIV-1 is endemic across the globe. In different populations throughout the world, the distribution of both HLA class I genotypes and viral sequences differ. This presents a further question:

- Are rates of escape and reversion broadly similar in different populations?

Answering this question would shed some light on the importance of epistasis in the processes of CTL escape and reversion in HIV-1. We address this question in Chapter 5.

As stated above, it is known that CTL escape mutations can impose a fitness cost on the virus and as a result reversion can occur within a host not possessing the relevant restricting HLA type. Differences in fitness costs associated to particular escape mutations may be indicative of the rate of reversion and/or escape of that mutation. In Chapter 5 we also ask:

- Is the rate of reversion of a CTL escape mutation positively correlated to its replicative fitness cost?
- Is the rate of escape of a CTL escape mutation negatively correlated to its replicative fitness cost?

Replicative fitness cost is as estimated by *in vitro* replicative capacity assays.

In order to take the non-independence of viral isolates into consideration, we estimated the viral genealogy of the analysed sequence samples as part of our inference. We use existing software to perform MCMC on genealogies [65]. Doing so is notoriously difficult and computationally intensive. In fact, we find that for cross-sectional HIV-1 sequence datasets consisting of more than a few hundred sequences the methods become infeasible. This presented a problem. Power to detect escape and reversion rates comes from coalescence events in the viral genealogy (if we assume that the genealogy may be used as a surrogate for the underlying transmission history of cross-sectional viral sequences). If the effective infected host population size is approximately constant then many coalescence events take place close to the tips of the genealogy. However, growth of the infected population through time (until recently the reality for the HIV-1 epidemic) leads to long terminal branches when the sampling proportion is small: genealogies appear star-like. As a result, the dependency structure is difficult to estimate and often has little impact upon escape and reversion rate estimates. Thus, we require more sequence data to gain recent coalescence events, but then sampling the genealogy becomes computationally infeasible. What is more, for our purposes the genealogy is a nuisance parameter which we are not interested in. We require a new strategy.

Before creating a new modelling and inference framework, we take a step back and ask what other desirable properties we would want our model to have. Current methods in the field neglect recombination, and tend to parameterise a nucleotide substitution model (due to the computational constraints of amino-acid and codon models). In Chapters 3, 4 and 5 we had to rely on prior biological findings which identified viral epitopes and associated restricting HLA types. We were also confined to the use of HIV-1 sequence data for which host HLA information was available. The vast majority of HIV-1 sequence data does not have associated host HLA genotype data. In Chapter 6 we ask:

- Can we create an inference regime for detecting HLA associated selection that:
    1. Accounts for the process of recombination?

2. Makes use of the very large databases of viral sequences from hosts with unknown HLA profiles?
3. Uses a realistic codon model of viral sequence evolution?
4. Does not require *a priori* knowledge of location and HLA restriction of epitopes?

We take the existing work of Wilson and McVean [304] as a basis for our modelling and inference strategy and build upon it to incorporate HLA associated selection. We also make a series of approximations to allow for speedy inference. We test the method under varying amounts of recombination using simulation studies and ask:

- Do we successfully recapitulate HLA associated selection parameters used to create simulated datasets under our model, even in the presence of high recombination rates?

The approach that we lay out generalises to other problems and may be extended still further. In fact, any research question for which the researcher wishes to estimate the effect of a host variable upon pathogen sequence evolution could be tackled using a similar procedure. Indeed, in Chapter 7 we apply our inference framework to estimate drug associated selection along the viral sequence. In order to test our method, we ask:

- Can our method detect known drug resistance mutations?

We then estimate HLA class I associated selection in the presence of recombination along regions of the HIV-1 genome without prior knowledge of selected sites and ask the following questions:

- Are well known HLA genotype/epitope pairings detected using our model?
- How sensitive is the method?
- Are unknown HLA/epitope pairings detected?

It is accepted that the HIV-1 subtypes seen today emerged as a result of founder effects as HIV-1 spread worldwide. It has also been hypothesised that differences

in HLA distributions between populations also played a role in the divergence of HIV-1 subtypes. As our final question in Chapter 7 we ask:

- Is there evidence that HLA frequencies help shape the differences we observe in consensus viral sequences between populations?

We conclude by summarising our collection of answers to the questions posed in this introduction and suggest a series of extensions and further questions for the future.

2

LITERATURE REVIEW

The acquired immunodeficiency syndrome (AIDS) pandemic is one of the greatest threats to human health in history. It is currently the sixth largest cause of death on the planet [208]. Human immunodeficiency virus (HIV), the virus that causes AIDS is so far responsible for seventy five million cases worldwide, a number which is currently expanding at a rate of 2.5 million per year [286]. First reports of what was later termed acquired immunodeficiency syndrome were made in California and New York in 1981 [43, 42]. Some two years later, HIV-1 was found to be the cause [18, 86]. The regions of the world most badly affected by HIV-1 are in sub-Saharan Africa, where the epidemic is known to have originated [258, 309]. A recent publication by Faria *et al.* [74] traced the earliest origins of HIV-1 to 1920s Kinshasa, before the subsequent spread of the virus through populations in the subcontinent.

Early optism for a vaccine was thwarted due to the virus's incredible ability to evade the hosts immune system. There is no cure. The infection cannot be cleared by therapy or the host's immune system and is therefore, lifelong [78]. Combination antiretroviral therapy has proved extremely successful, but is required throughout life to keep the virus at bay and delay onset to AIDS.

The cytotoxic T-lymphocyte (CTL) response, a branch of the adaptive immune response, is known to play a major role in driving viral diversity both within-host and at the level of the host population [102]. There is high variability in vulnerability to infection and viral control if infection takes hold. Certain human leukocyte antigen (HLA) class I genotypes (which encode a complex that plays a central role in the CTL response) are correlated with rapid progression [101], whilst others are correlated to slow progression [101] and lower levels of latently infected cells; the so called 'latent reservoir' [39, 33]. As such, it is important to

understand the biological processes which underlie these correlations if we are to make progress in informing any treatments or a potential cure.

The successful immune evasion that the virus exhibits is as a consequence of the high viral turnover [117] and error-prone method by which it replicates [50]. The resultant high diversity found in HIV-1 populations make study of processes governing relatedness of viral sequences highly amenable to population genetics and phylogenetic methodologies. In this chapter, we provide a brief overview of the biology of HIV-1, summarise immune responses to the pathogen, and further highlight the importance of the CTL response. We then describe a small selection of mathematical models of HIV-1. By doing so, we demonstrate the flexibility of simple ordinary differential equation (ODE) based representations of epidemics and viral infection. Amongst these examples we highlight two models of CTL escape. One purely within-host, and the other modelling both within-host and between-host viral dynamics. We wish to create models and perform inference to learn about CTL escape and reversion in the host population. Compartment based ODE methods allow us to parameterise such a model, but implicitly assume that observed viral sequence data evolved independently. This is not the case. There is a structure of relatedness governed by viral/host dynamics which leads to the data we see. The purpose of this work is to develop methods to use this dependency structure to inform CTL escape and reversion parameter estimation. This goal motivates the remainder of the chapter. We outline population genetic and phylogenetic approaches used to incorporate the viral genealogy into parameter inference.

2.1 THE BIOLOGY OF HIV-1

2.1.1 *Structure*

HIV-1 is a member of the *Lentivirus* genus which form part of the *Retroviridae* family. HIV-1 particles are roughly spherical, and each virus particle, or virion, is around 10^{-7}m in diameter [126]. The outer coating of the virus, known as the

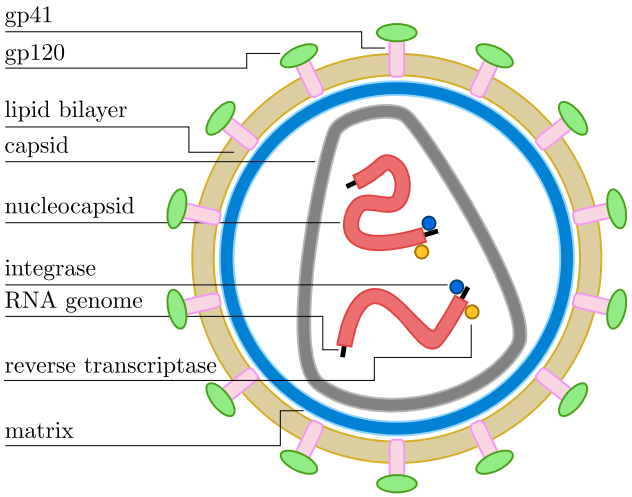


Figure 2.1: Simple cartoon representation of the HIV-1 virion. Key structural proteins and enzymes are labelled.

envelope is a lipid bilayer and is covered with around 14 spikes [272, 316] formed from stalks made from gp41 which project from the envelope and connect to a gp120 head. The morphology of each spike is thought to resemble a tripod, with on average 3 gp41 splayed legs connecting to a head domain made from a trimer of gp120 [317]. Immediately beneath the envelope is the matrix layer, constructed from the protein p17. This matrix encloses the viral core, or capsid, a protein shell which contains the genetic material of the virus. In HIV-1, the capsid resembles a truncated cone and is comprised of p24. In each virion the capsid contains the genetic material of the virus: two RNA strands of around 9700 nucleotides in length, which are tightly bound to two nucleocapsid proteins, p6 and p7. Also present in the viral core are the proteins Vif, Vpr and Nef as well as a collection of enzymes essential for development of the viral particle and replication: integrase, protease and reverse transcriptase (often abbreviated RT) [79, 233].

2.1.2 *Life cycle*

HIV-1 is a retrovirus. This means that genetic information undergoes the extra step of RNA $\rightarrow$ DNA through reverse transcription before being integrated into the host's genome whereupon it may then undergo the standard transcription translation process.

We begin viewing the HIV-1 life cycle at the stage where an HIV-1 virion is

within a host cell and is being transported to the nucleus. Reverse transcriptase creates a single strand anti-sense DNA copy of the RNA, known as cDNA. The reverse transcription process is very error prone as reverse transcriptase has no proof-reading capacity [224]. For this reason, viruses which replicate via reverse transcription have high-mutation rates and can therefore acquire drug resistant mutations very quickly. In fact, the error rate of HIV-1 reverse transcriptase has been shown to be high even among reverse transcriptases of other viruses [240, 224]. After the cDNA strand has been reverse transcribed, reverse transcriptase then degrades the now copied RNA and through its DNA polymerase activity creates a corresponding sense strand for the anti-sense cDNA. This double stranded DNA can then be incorporated into the host cell's genome via integrase which acts by catalysing two reactions: firstly a DNA processing step which must take place before the viral genome can be integrated, and the strand transfer reaction where this processed DNA is covalently bound to the host genome. The HIV-1 genetic information is now a provirus (a viral genome which has been integrated into the genome of a host cell), and infection of the cell is permanent. Infected cells either lie dormant (at low frequency, but sometimes for a very long period of time, up to the order of years) or become active and replicate. Cell machinery of the host cell allows viral genes to be expressed, specifically cell transcription factors, the most significant of which is NF-κB [79]. Transcription of the viral sequence begins when viral transcripts are expressed from the 5′ end of the long terminal repeat (LTR) section of the virus. NF-κB is up-regulated when T-cells become active [116], it follows that cells fighting infection are themselves the most likely to be infected. Indeed, HIV-1 specific $CD4^+$ cells are depleted in the early stages of infection [182].

The replication process then proceeds through another series of steps. DNA provirus is transcribed into messenger RNA (mRNA) which is then spliced and transported out of the nucleus and into the cytoplasm. Translation of the regulatory proteins Tat (*t*rans-*a*ctivator of *t*ranscription) and Rev (*re*gulator of *v*irion expression) then takes place. One of the roles of Tat is to increase the amount of transcription of the provirus, leading to a positive feedback loop and allowing for an explosion in virions after a threshold of Tat is reached. Rev allows un-

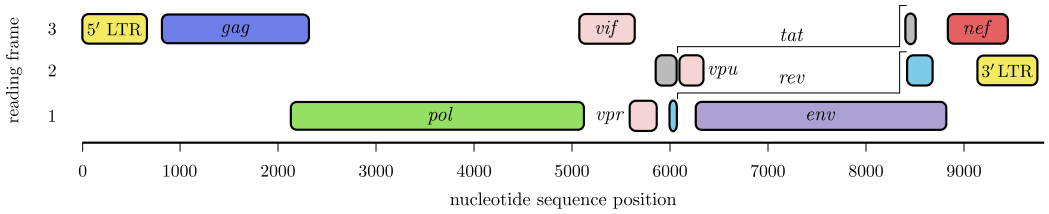


Figure 2.2: HIV-1 genome, open reading frames are displayed as rectangles. Note that *tat* and *rev* are spliced.

spliced transcripts to be released from the nucleus into the cytoplasm including the entire RNA genome which can be packaged after the larger viral components *gag* and *env* are translated. Gag binds the RNA genome and is later joined by Env near the inner surface of the host cell where budding begins. During the process of budding from the host cell, the particle undergoes morphological changes and becomes a mature virion which has the potential to infect other host cells.

But how does the virion invade a host cell? gp120, the glycoprotein which forms the head of each spike protruding from the surface of the viral envelope, targets and interacts with CD4 (cluster of differentiation 4) and chemokine receptors (usually CCR5 or CXCR4) via binding domains. CD4 is a glycoprotein expressed on the surface of some T-cells, dendritic cells, monocytes and macrophages. gp120 first binds CD4, promoting a conformational change which then provides a binding domain for chemokine receptors. Once the virion has attached to the new host cell, gp41 undergoes conformational changes which mediate fusion of the viral cell membrane with the target cell membrane. Upon entry into the host cell, the virion core loses its coat and the resulting viral nucleoprotein complex is taken to the target cell nucleus via microtubule transport, completing the cycle.

2.1.3 *Viral Components*

The genes *gag*, *pol* and *env* are common to all retroviruses. They encode structural proteins and enzymes that are essential for replication of viral particles. Lentiviruses are defined by long incubation periods (*lenti* is latin for 'slow'). HIV-1 has additional accessory genes *tat*, *nef*, *vif*, *vpr* and *vpu* which aid invasion and replication during HIV-1 infection. The genetic structure of HIV-1 is shown in

Figure 2.2. LTRs are not part of the viral genome, but help regulatory genes control expression of *gag*, *pol* and *env*. Of the accessory genes, *nef*, *vif*, *vpr* and *vpu* are known as auxiliary genes as their inactivation does not completely prevent the replication cycle *in vitro*. We now describe each of the genes, the proteins they code for, and a selection of the processes in which they are involved.

- ***Gag*** (*g*roup specific *a*nti*g*en) encodes the core structural components of the virion. Initially a p55 precursor is created, before being processed to p17, p24, p7 and p6 by protease (itself encoded by the *pol* gene). p17 forms the protective matrix in a scaffold assembly, p24 makes up the viral capsid, and p7 the nucleocapsid [99]. Gag is moved to the plasma membrane for assembly into the virion.

- ***Pol*** encodes a collection of enzymes essential in the HIV-1 replication cycle outlined above: protease, reverse transcriptase and integrase.

- ***Vif*** (*v*iral *i*nfectivity *f*actor) disrupts antiviral activity of the enzyme APOBEC3G (part of our innate immunity against retroviruses which works by interfering with viral replication) by ubiquitination (which provides a marker for destruction) or cellular degredation [313].

- ***Vpu*** (*v*iral *p*rotein *U*) is associated with budding from the host cell. It down-regulates CD4 through targeting of the protein in the endopasmic reticulum (ER). Sometimes, after interaction with CD4, gp160 (a precursor of gp120 and gp41) is transported there. Degradation of CD4 within these complexes in the ER results in gp160 moving to the cell membrane where viral assembly takes place. Thus vpu is associated with increased virion release. It also down-regulates major histocompatibility complex (MHC) class I molecules (complexes which display fragments of antigens from within the cell which could either be self or non-self, to $CD8^+$ cells), which helps protect infected host cells from detection by cytotoxic T-lymphocytes (CTLs) and whose significance is discussed in Section 2.5.

- ***Vpr*** (*v*iral *p*rotein *R*) is related to HIV-1's ability to infect new host cells. Transmission of viral RNA from the cytoplasm to the nucleus is mediated

by vpr, as it has a nuclear localisation signal (NLS) (a tag used to target a protein to the cell nucleus).

- ***Tat*** (*t*rans-*a*ctivator of *t*ranscription) is a regulatory factor of gene expression essential for replication. It up-regulates mRNA production, thereby increasing transcription of the provirus. Under some conditions it is also responsible for increasing transcription initiation.

- ***Rev*** (*re*gulator of *v*irion expression), is initially found mainly in the nucleus, though around 24 hours after infection of the host cell, it starts to oscillate between the nucleus and the cytoplasm. In the nucleus it binds to mRNA at the Rev binding site RRE (Rev response element). This binding allows partially spliced or unspliced transcripts to be exported into the cytoplasm. This process is fundamental in the life cycle of HIV-1 as it means that the larger transcripts can be translated enabling the virion particle to be constructed [151, 281].

- ***Env*** (*env*elope) codes for proteins in the viral envelope of HIV-1. First a precursor gp160 is translated, before being processed into a complex of gp120 and gp41 which form the spikes found on surface of the virion, bound together by non-covalent interactions.

- ***Nef*** (*ne*gative regulatory *f*actor) is a virulence factor which manipulates the host's cell machinery allowing infection, survival and replication of the pathogen. Like vpu, it down-regulates CD_4 and MHC class I. In addition it helps increase infectivity and alters the activation state of cells.

2.2 GENETIC VARIATION IN OTHER INFECTIONS

Before focusing in on the genetic variation of HIV-1, we consider a collection of disease causing agents and give a summary of the variation they exhibit. By doing so, I hope to convey the wide assortment of methods by which agents are transmitted and the variety of tactics they employ to evade the immune response of their hosts. We then move on to HIV-1: summarising the genetic diversity and

host immune responses to HIV-1 both innate and adaptive. Finally, we provide an expanded discussion of the CTL response, the branch of the immune response most frequently implicated in slow progression and subject of this work.

2.2.1 *Measles*

A *Paramyxovirus* and member of the *Morbillivirus* genus of negative sense single stranded RNA viruses, the measles virus can be split into distinct genetic groups using the haemagglutinin (HA), nucleoprotein (N) sequences and the use of phylogenetic methods. HA and N are the most highly variable regions of the viral genome, varying by up to 8% in exons [248]. The majority of the genotypes are distributed worldwide, though some are restricted to particular geographical regions [20]. The most variable region in the viral genome is a 450 base sequence which codes for the COOH terminus of N [248]. The mutation rate of measles is approximately equal to that of other RNA viruses *in vitro* [255]. However, during outbreaks, sampled sequences have displayed virtually no sequence variation, even in the most variable regions [206, 247], though sporadic outbreaks show heterogeneity [247]. Despite having multiple genotypes, measles has just one serotype. This explains the success of vaccination: even viruses exhibiting the largest sequence divergence from the vaccine virus are neutralised by antiserum.

2.2.2 *Influenza*

An RNA virus of the *Orthomyxoviridae* family, influenza is split into three types: A, B and C [276]. Type A displays the greatest diversity and can be split into a subtypes distinguished by combinations of Haematglutinin (HA) and Neuraminidase (NA) surface glycoproteins. There are 16 known HA types and 9 NA types [283]. Of the 144 possible combinations, 103 have been found in wild birds [120]. These avian influenza viruses (AIVs) can sometimes undergo a zoonotic transfer which may result in isolated outbreaks or on occasion a global pandemic.

Human influenza type A consists of 8 segmented negative stranded RNAs, each encoding 1-2 genes. Of the 8 segments, 5 do not display reassortment with viruses from other hosts. HA and NA (and PB1, a polymerase protein) subtypes are known to reassort, so that HA and NA subtypes in AIVs can find their way into human influenza viruses. Indeed, such reassortment events have marked pandemics. HA subtype H2 and NA subtype N2 reassortments created the H2N2 strain which led to the 1957 Asian flu, and the reassortment of H3 resulted in the H3N2 strain and the Hong Kong flu of 1968 [120]. Most recently, the 'swine flu' epidemic of 2009 was the result of a complex sequence of reassortment events between the eight viral segments in bird, swine and human lineages [266].

2.2.3 *Hepatitis C virus*

The hepatitis C virus (HCV) is a single-stranded positive sense RNA virus and member of the *Flaviviridae* family. HCV is split into seven major genotypes based on its genetic diversity, and further subcategorised into subtypes [120]. There is $\sim 30\% - 35\%$ nucleotide diversity between HCV genotypes, and $\sim 20\% - 25\%$ between subtypes within genotypes. Variation in viral sequence data can provide insights into the transmission history. For example, correlations between viral sequence and geography [261]. Diversity is also present within-host, these 'quasispecies' allow the pathogen to respond more rapidly to the immune response of the host [261]. HCV is far older than HIV, and though the exact timescale is unknown, it has been circulating in human populations for thousands of years. The highest genetic diversity in HCV is found in African and Asian populations [120, 119].

2.2.4 *Tuberculosis*

Tuberculosis is caused by members of the *Mycobacteria*, most commonly *Mycobacterium tuberculosis*; a small non motile *bacillus*. *M. Tuberculosis* has very low vari-

ability in its genome [252, 69] due to its lack of plasmids and very low rate of horizontal gene transfer [69]. Inspite of having such a highly conserved genome, antibiotic drug resistance of the bacteria is increasing, as point mutations and deletions can lead to high phenotypic variation [193, 252]. A small number of sequence changes can affect epidemiological properties such as virulence, drug resistance and can even impact diagnostic test results. Indeed, mutations and deletions in intergenic regions and catalyst regions are known to confer resistance to the drugs isoniazid and rifampicin respectively [69]. Such mutations act by altering protein binding sites or enzyme activity.

2.2.5 *Trypanosomes*

Trypanosoma brucei is a protozoan parasite transmitted between mammalian hosts by the tsetse fly, resulting in African trypanosomiasis in humans, also known as sleeping sickness. To evade the immune response within the bloodstream of the mammalian host, *T brucei* can alter its glycoprotein coat, known as a variant surface glycoprotein (VSG) [60, 278]. *T brucei* employs a variety of methods to express a given VSG, the most common of these is known as gene conversion. Expressed VSG genes are found near telomeres at so-called expression sites, gene conversion simply switches out the VSG at the expression site with a VSG found elsewhere in the genome [28, 60]. Around 1000 VSG genes are found across the parasite genome [60] and swaps between immunologically distinct glycoprotein coats occurs at a rate of between 10^{-7} and 10^{-2} per doubling time of 5-10 hours [285, 284], resulting in cycles of infection within chronically infected individuals as the parasite evades the immune response after switching coats. This is completely different to HIV variability. The variability is already encoded within the genome, and is then exploited during its life cycle.

2.3 GENETIC VARIATION IN HIV-1

HIV-1 strains are currently split into four types: M (main), O (outlier), N (non M, non O) and P (putative - now confirmed [287]) based on their appearance on four distinct lineages in a phylogenetic trees built from sequence data. The four lineages are each the result of an independent cross-species transmission event [258, 88]. Lineages M and N can be traced back to independent zoonotic transfers from chimpanzees (*Pan troglodytes troglodytes*) [89, 139]. Type P viruses are thought to have originated from western gorillas (*Gorilla gorilla gorilla*), due to their clustering with gorilla simian immunodeficiency virus (SIVgor) [288] in viral phylogenies [287, 219]. By contrast, the source of group O is unknown; it shows no strong relatedness to SIV viruses taken from either chimpanzee or gorilla populations, though the origin is thought to be one of these two species [258]. The pandemic lineage, group M causes the majority of HIV-1 infections in humans (over 90%). The type M clade can be further subdivided into genetically distinct viral subtypes (A, B, C, D, F, G, H, J and K [241]) and circulating recombinant forms (CRFs). Subtypes can be grouped by geography and are believed to have been formed as a result of founder effects [233]. Subtype C represents the highest proportion of all the M subtypes. It is currently responsible for just under half of all HIV-1 infections (48% [112]) and is predominantly found in Africa and Asia. Subtype A (which makes up ∼12% of global infections) is mostly found in Africa, Central Asia and Eastern Europe [112]. The majority of research into HIV-1 is concentrated on subtype B (representing around 11% of all global infections [112]), the subtype which causes the majority of HIV-1 infections in Europe and the Americas [217]. The greatest diversity of HIV-1 subtypes is found in central Africa, where the pandemic is thought to have originated [233, 290]. Every subtype currently defined can be found here [256]. Subtypes G and D correspond to ∼5% and ∼2% of worldwide infections respectively, the remaining subtypes together cause less than 1% of cases globally [112]. The most recent common ancestor (MRCA) of group M is thought to have arisen between 1884 and 1924 [309], and subsequently spread along transport networks from what is now the Democratic Republic of Congo [74, 258]. The remaining three groups are rare.

Group O is restricted to individuals in or epidemiologically linked to hosts in West-Central Africa [111, 214, 280]. All confirmed cases resulting from Group N viruses (since the group was identified in 1998) are restricted to Cameroon [262, 245, 16]. The final group, P, is thought to have derived from SIVgor via a gorilla to human transmission [219]. Group P was confirmed to be circulating in human populations in 2011 [287].

HIV-1 has very high variability in its genome compared to other retroviruses. This genetic variation is not constant across the viral genome. The greatest genetic diversity is found in *env*; which codes for the envelope glycoproteins [233]. A high mutation rate (sum of point mutations, insertions and deletions) of around $\sim 10^{-5}$ per replication cycle, per nucleotide site, (due to the lack of a proof reading mechanism in reverse transcription [224]), coupled with a high virion turnover of $\sim 10^{10}$ per day per individual [215] results in the huge amount of diversity in the viral population. This diversity may affect the transmissibility, virulence and immunogenicity of strains [136]. HIV-1 is diploid (two RNA strands are found in the nucleus of each virion), as a result, host cells infected with two or more viral particles have their genetic information packaged in the same virion during the replication and recombination may arise through strand-switching during reverse transcription [207]. The process of recombination increases diversity still further. Indeed, CRFs (which emerge when two or more subtypes co-infect a host cell at the same time and recombine) are now responsible for around 16% [112] infections. The diversity found in HIV-1 presents clear challenges to the development of treatments; selection pressures exerted by the host and antiretroviral drugs accelerate evolution of the virus [141, 115].

Within any given host, the viral population of HIV-1 is composed of mutant and recombinant genomes known as a quasi-species [147]. It is suggested that on average, infection can be tracked back to a single founder virus [140, 5]. This bottleneck is thought to be due to both the mucosal barrier (backed up by the finding that up to 16 virions have been found responsible for infection via needle sharing in drug users [17]) and the selection bias for viruses with increased fitness in the newly infected, immunologically naive recipient [37].

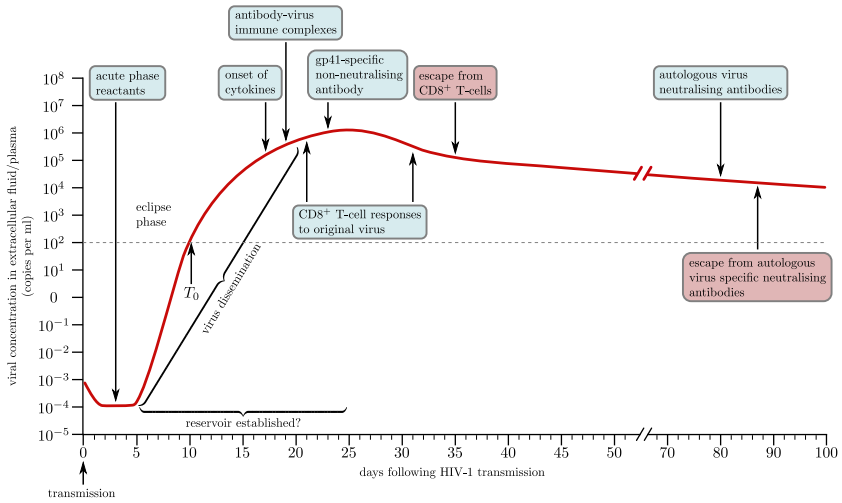


Figure 2.3: Summary of key events in the early immune response to HIV-1, taken from [182]. T_0 denotes the time at which the virus can first be detected in the blood of the host, and marks the end of the so-called eclipse phase. CTL escape from initial immune responses takes place quickly, as does the viral response to the first neutralising antibodies; with escape taking place in under a week.

2.4 SUMMARY OF IMMUNE RESPONSES TO HIV-1

HIV-1 is characterised by its destruction of the host's immune system through depletion of $CD4^+$ T-cells, leading to AIDS. The immune system provides protection to the host against infection through a three pronged defence of increasing specificity. The first line of defence is surface barriers to keep pathogens out. If this is broken, innate immunity and adaptive immunity provide non-specific and specific responses to the pathogen. A summary of some of the key events in the early interaction between HIV-1 and the host [182] is shown in Figure 2.3.

HIV-1 experiences the same order of immune response as other viruses, the innate immune response occurs first, very early in infection, preceding adaptive B-cell and T-cell responses which are detectable after a few weeks when the infection has been established.

A variety of cells play a part in mounting the innate immune response to HIV-1. Langerhans cells, a type of dendritic cell found in the epithelial layer of skin and mucosa, act against the pathogen by internalising virions for degredation after they are captured by langerin on their surface [57]. Natural killer cells (NK) and Natural killer T-cells (often abbreviated NKT, these are highly specialised lymphocytes expressing a restricted T-cell receptor) are activated during acute infec-

tion. These cells are central to the control of replication of invading pathogens via the fast production of cytokines and chemokines, as well as osmotic lysis of infected cells [8].

An important example of an innate intracellular response towards HIV-1 is the innate intracellular antiviral factor APOBEC3G (also known as A3G and CEM15). It functions by altering HIV-1 transcripts just after the viral DNA has been transcribed by deaminating cytidine into uridine leaving them non-functional [259]. A3G also restricts HIV-1 infection in resting $CD4^+$ T-cells through another mechanism taking place before transcription occurs [48]. The *vif* gene counteracts these mechanisms by not allowing A3G to enter the virion and inducing its degredation [177]. Another protein TRIM5α has been shown to restrict HIV-1 though the underlying mechanism is not fully understood [311]. Although innate immunity plays an important role in a host's defence against HIV-1 infection, it is not yet apparent how that response can be utilised in a potential treatment.

Adaptive immunity can be subdivided into humoral and cell-mediated immunity, which are controlled by B-cells and T-cells respectively. The humoral response includes antibody production by B-cells, of which there are millions of forms; each having different antigen binding sites. A degree of hope was pinned on the humoral response as a potential option for vaccination against HIV-1 due to its successful use in vaccination against other diseases such as measles, mumps and rubella [194]. Sterilising immunity could only be achieved if neutralising antibodies specific to a wide range of strains were present within the host at the time of transmission. Although antibodies with this capacity have been found in a few individuals [295], their occurrence is rare [105, 272] and occurs years into infection [187]. However, the mere fact that they exist shows that the viral envelope does have a few chinks in its armour which can potentially be exploited [152], despite its numerous effective defense mechanisms [34]. Without such a breadth in specificity, the relatively slow creation of neutralising antibodies, which appear around 12 weeks after transmission [302], compared to the rapidly mutating regions of *env* mean that the humoral immune response is constantly a step behind. Indeed, in HIV-1 subtype C, initial neutralising antibodies are effective over only a small selection of epitopes [190], resulting in fast

immune escape by the virus [238].

In addition to the highly variable regions in *env*, HIV-1 has evolved a series of mechanisms to defend itself against antibody binding. The *env* gene is expressed on the surface of viral particles in the form of the spike complexes of gp120 and gp41. These apparently vulnerable envelope proteins are thus the prime target for the humoral response, leading to selection of an envelope complex with low immunogenicity whilst still performing its various tasks in the replication cycle [132]. HIV-1 has a coating of very poorly or non-immunogenic glycans through 24 glycosylation sites on gp120 and between 3 and 4 on gp41, prohibiting access to the peptide structure beneath [34]. Changes to these sites can lead to epitopes becoming visible to antibodies [209]. A large portion of the gp120/gp41 complex is not accessible to binding by antibodies, so any conserved regions which could result in neutralisation are occluded. Early HIV-1 antibodies are specific for gp41 and gp120 (found around 23 and 38 days after transmission respectively) [7, 58]. However, despite their high reactivity in Western Blots to the two proteins, antibodies produced early in acute infection exhibited little or no reaction with the complex, suggesting that in this form on the viral surface it is virtually unrecognisable to the very antibodies which are generated against its constituent parts [132]. The underlying mechanism is unknown, though it has been suggested that non-functional virions or fragments of lysed virions could drive the antibody response to Env [213].

During the adaptive immune response, $CD4^+$ T-cells, also known as T-helper cells, play a fundamental role, directing the acquired immune response of the host. In the humoral response they are essential to antibody class switching, whereby production the 'constant' portion of the heavy chain of is switched from one isotype to another by means of signalling cytokines [8]. T-helper cells are fundamental in the cell-mediated response as they recognise virions and activate the CTL response, recruiting CTLs [8]. Upon arrival of a CTL, in order to trigger a response an infected cell's MHC class I molecules (known as human leukocyte antigen (HLA) in humans, MHC class I corresponds to HLA A, B and C) must present antigen. CTLs recognise short peptides of around 9-11 amino acids, known as epitopes. Upon recognition, the CTLs bind with the T-cell re-

ceptor (TCR) and lyse the cell. Exogenous antigens are recognised by $CD4^+$ cells via HLA class II molecules on antigen presenting cells (APCs). In this case, epitopes are longer at around 11-20 amino acids in length. The majority of antigens bound to HLA class I molecules are created within the cell, so usually a $CD8^+$ T-cell response is to endogenous antigens. This twin response inside and outside the cell allows targeting of the virus during its entire replication cycle.

Certain HLA types known to result in both slower and accelerated progression to AIDS, which we discuss in Section 2.5. *In vivo* studies of chronically infected rhesus macaque monkeys (a species in which SIV progresses to sAIDS) demonstrated the importance of CTLs in SIV progression. CTLs were purposely depleted, and increased viral replication was observed. In addition, in some monkeys the CTLs were regained and the virus was kept under control [254]. Temporal correlation found between viral load and HIV-1 specific CTLs [26] lead researchers to believe that CTLs are expected to play as strong a role in control of HIV-1 replication. However, SIV is model of HIV-1. Using a mathematical model to estimate rates of escape and reversion in SIV and HIV-1, Asquith and McLean [14] discovered a significant differences between rates of escape in HIV-1 and SIV, and rates of reversion (as measured from the emergence of an escape variant) in HIV-1. This finding suggested that it may be easier to elicit a protective CTL response in macaques than in humans and that vaccine studies conducted in macaques should be interpreted with caution.

2.5 EPITOPES AND ESCAPE

As described in the introduction, the CTL response is thought to play a fundamental role in viral control during the course of HIV-1 infection. During the CTL response, proteins floating in the cytoplasm are digested into peptides by the proteasome and taken to the endoplasmic reticulum by the transporter associated processing (TAP) complex. Once there, ERAP trimming takes place; the epitope precursors are cut to a length of $\sim$9-11 amino acids and loaded onto HLA class I molecules. These HLA complexes then present the epitopes on the outside of

the cell [8]. Epitope presentation takes place at the surface of almost all nucleated cells in the body. Currently, over 5000 alleles of these three genes have been found [242], making them the most genetically variable loci in humans.

The set of peptides that may be presented by HLA class I molecules is dictated by the collection of HLA genotypes the host possesses. CTLs restricted by certain HLA class I molecules are associated with viral control, whilst others are associated with disease progression [101]. Homozygosity at any one of the three HLA class I loci is associated with a faster loss of viral control [40]. Each of these findings imply that the CTL response plays a pivotal role during infection. However, the relationship between HLA and disease outcome was found to be complex: no clear relationship between magnitude of response and progression exists [102].

The first findings that CTL escape can take place was made in 1991 by Philips *et al.* [218]. Subsequent small cohort studies provided further evidence of CTL escape *in vivo* (e.g. [226, 27]). The initial viewpoint was that any escape must take place during acute infection due to the rapid turnover of the virus [301], coupled with an error-prone replication mechanism [50] and rapid recombination rate [207]. Surely then, if any mutation existed it would be found almost instantaneously and exploited by the virus. This thought process neglects the potential fitness costs imposed by such variants. Indeed, whilst viral escape can often take place during acute infection, and is sometimes associated with viral control e.g. the B*57 restricted epitope `TSTLQEQIGW`, in other cases escape can take place far into chronic infection and precede the onset of AIDS [103]. The discovery of CTL escape mutations emerging during chronic infection was supported by experiments conducted using the macaque model: SIV experiments generated evidence suggesting a correlation between escape mutations and disease progression as well as vaccine failure [72]. Again though, the exact relationship is unclear; some monkeys showed rapid progression to sAIDS without undergoing any known escape mutations.

In the literature that followed, examples of escape variants were found in almost all proteins encoded in the HIV-1 genome [9, 27, 61, 141, 154, 153, 183, 226, 312]. These mutations, which can occur close to or within epitopes can impair HLA binding, affect T-cell receptor contact sites, or interfere with antigen processing

steps [315, 61, 129, 101]. CTL responses are shown to display strong correlation with such escape mutations during acute and chronic infection. Variation in viral set-point, CD_4 count and, by inference, rate of disease progression is strongly associated with particular HLA-B allele expression [145, 90].

HLA-B*57 and B*27 are examples of alleles associated with LTNPs [82, 135]. The B*27 restricted CTL response to a conserved epitope in the *gag* gene, `KRWIILGLNK`, is well documented [305, 200, 141, 100]. Mutation at position 2 in this epitope results in reduced binding affinity to the B*27 class I molecule and incurs a fitness cost, though this cost is mitigated by further nearby mutations. This escape mutation is expected to take years to arise (*Fryer et al.* [85] estimate approximately 8 years using their ODE model) and is coincident with loss of viral control in these individuals. The time taken for such escape mutations to arise is not fully understood, though changes in the host immune response over time coupled with structural constraints imposed by the viral amino acid sequence have been implicated as a partial explanation [204]. The 'protective' HLA type B*57 restricts the *gag* epitope `TSTLQEQIGW` [11]. Escape in this epitope is known to revert upon transmission to a host not possessing the B*57 allele. This suggests that the trade-off between viral fitness costs and CTL escape is key in understanding the expected time to escape/reversion and why certain alleles appear to be protective in some people. Indeed, the most effective CTL responses select for escapes that revert in HLA mismatched hosts [180]. Another collection of HLA types are known to be associated with an increased rate of disease progression, examples include HLA-B*35-Px and B*5802 [39, 40, 157, 145, 90].

The region of the viral genome in which epitopes reside is associated with the level of viraemia in hosts [146, 299]. Responses targeting Gag, such as those conferred by HLA-B*27 and B*57 are associated with reduced viraemia, whereas responses specific to Env and accessory/regulatory proteins are associated with increased viral loads. This result helped to explain the fact that CTL responses are detectable in individuals progressing to AIDS: a CTL response does not necessarily result in reduced viraemia and control of the pathogen. One suggestion is that Env-specific responses are simply less effective at controlling viraemia than those targeted to Gag [146]. Various other studies also associate Gag spe-

cific responses with reduced viraemia via positive selection of escape mutations which result in a reduction in viral fitness [163]. As previously hinted at, we therefore expect that the most effective immune responses are likely to be those that effectively target regions of the viral genome where escape results in a dramatic drop in fitness. Such a reduction is likely to result in decreased viral load and an increase in the expected time until onset of AIDS [144].

We do not need to conduct lengthy longitudinal studies or perform difficult biological assays to detect the footprints of CTL dependent selection in viral sequences. Markers of HLA association can be seen in population studies through association between HLA type and sequence polymorphisms in the HIV-1 genome [189]. However, care should be taken when determining such correlations due to linkage disequilibrium between HLA types and founder effects. Indeed, results of Moore *et al* [189] were called into question by Bhattacharya *et al.* [22]. Their study considered the viral phylogeny and took founder effects into consideration. They determined that less than 10% of the relationships found in [189] remained, and suggest that the frequency at which escape events take place is lower than was previously suggested.

The importance of the HLA class I genes in viral control has also been highlighted through a series of genome wide association studies (GWAS) [275, 75, 168, 54, 41]. Single nucleotide polymorphisms (SNPs) in the region were found to have the greatest association with set point viral load, a strong predictor of disease progression [184]. The first GWAS to look at correlates of disease progression, performed by Fellay *et al*. [13] found two SNPs that reached genome wide significance: one was in strong LD with HLA-B*57; a known protective allele for HIV-1, and another relating to HLA-C expression. A subsequent GWAS found over 300 SNPs associated with viral load, all of them found in the HLA region [275]. Differential HLA-C expression was tested directed by Apps *et al.* [13]. They found that HLA-C expression is positively correlated with viral control *in vivo*, and that higher HLA-C expression leads to an increased likelihood of CTL responses.

If an escape variant reaches high frequency in one host, the mutation may be transmitted and thus has the potential to spread through the infected host popu-

lation [9, 81], or revert due selection within hosts who do not have the restricting HLA [82, 163]. The potential transmission of HIV-1 variants both horizontally and vertically [253, 81, 163, 83, 104] has clear implications for the evolution of HIV-1 at the population level. Dependent upon the success of a variant within an HLA mismatched host, escape mutations may revert back to the wildtype at that epitope [144] or compensatory mutations may occur. Clearly if a mutation results in fast reversion in HLA mismatched hosts we expect that it will only be retained at low frequency within the population. However, some mutations are known to persist in HLA-mismatched hosts either because the loss of viral fitness is very small, or because compensatory mutations regain the escape mutation's original fitness cost [49], for example KRWIILGLNK. Such variants have the possibility of reaching high prevalence in the population. In some cases, HLA mismatched hosts infected with escaped strains result in lower viraemia and higher $CD4^+$ counts [49], though these examples revert quickly so cannot establish themselves at the population level. The existence of stable escape mutations has unsurprisingly resulted in concern, if these mutations propagate through the infected population they would lead to a reduction in the absolute number of currently known HIV-1 CTL epitopes recognised across the collection of HLA types, and thus could lead to increased virulence in the future if such mutations accumulate, though so far no such observation has been made [113, 282].

2.6 HIV-1 MODELLING

In the chapters following this literature review we seek to improve our understanding of HIV CTL escape. To do this, we restate the processes of CTL escape and subsequent reversion in a simplified theoretical framework. Using a simple model we may estimate parameters using biological data, relate these estimates back to reality, and make testable predictions. Mathematical modelling and statistical inference in the field of HIV-1 research has a long history. A wide collection of models exist which attempt to capture important features of biological processes, both within-host and between hosts. In this section, we provide exam-

ples of within-host models of infection. We then describe a model which captures both within-host and between-host dynamics by combining the processes of CTL escape, transmission and viral reversion. Using this model, the authors aimed to determine if escape variants may propagate through the infected population, a concern raised in Section 2.5. However, they make a fundamental assumption that should be relaxed, namely that observations of viral sequence data taken from a population are independent. In Section 2.7 we describe a well studied process, the coalescent, used to describe shared ancestry in populations. We then outline methods to infer this dependency structure using sequence data. Combining epidemiological and genealogical methods to account for non-independence of viral sequence data is known as phylodynamics. We describe some examples in which epidemiological parameters of interest (e.g. R_0) were estimated using phylodynamics in Section 2.8. The success of such methods motivates Chapter 3, in which we integrate the underlying dependency structure into a model of CTL escape and reversion in the HIV-1 infected population.

2.6.1 *Within host ODE models of HIV-1*

Within-host models of HIV-1 have been used to examine viral dynamics [301] during infection. Some early models were used to answer the important questions: How long does a productively infected cell live and what is the rate of viral turnover? In a classic within host model of HIV-1 infection, a system of just three differential equations is used. Uninfected cells; x, infected cells; y, and viral particles; v were modelled by:

$$\textit{Uninfected cells} \qquad \frac{dx}{dt} = \lambda - \delta x - \beta xv; \tag{1}$$

$$\textit{Infected cells} \qquad \frac{dy}{dt} = \beta xv - \alpha y; \tag{2}$$

$$\textit{Virions} \qquad \frac{dv}{dt} = ky - uv. \tag{3}$$

Uninfected cells are produced at a constant rate λ, and die at rate δ. Newly infected cells are produced from infected cells and free virus at rate βxv. Infected

cells are depleted at rate α, and virions are created at rate k per infected cell, and removed at rate u. This model was used to estimate the rate of decrease of virally infected cells and thus provide an estimate for the average lifespan of an HIV-1 infected cell, the rate of increase of virions, and the expected number of virions an infected cell is expected to contain before it is destroyed. The use of reverse transcriptase inhibitors was modelled by setting $\beta = 0$ (assuming complete efficacy of the drug). This allows the system to be solved analytically.

$$y(t) = y(0) \exp(-\alpha t), \tag{4}$$

$$v(t) = v(0) \frac{u \exp(-\alpha t) - \alpha \exp(-u t)}{u - \alpha}, \tag{5}$$

where $u \neq \alpha$. Using this simple model (and approximations to it), a half life of 2 ± 0.9 days and 1.5 ± 0.5 days was estimated for the higher and lower of the two rates α and u [301]. However, α and u could not be distinguished as only viral levels in the plasma could be measured, and the analytic solution for $v(t)$ is symmetric in α and u. The dynamics of the free virus v first encounter a shoulder phase before an exponential drop, as shown in Figure 2.4. Note that Equations (4)−(5) only resemble the true dynamics if reverse transcriptase inhibitors are 100% efficient.

Modelling of complete protease inhibition was also carried out, using a slightly modified group of equations [307]:

$$\textit{Uninfected cells} \qquad \frac{dx}{dt} = \lambda - dx - \beta x v; \tag{6}$$

$$\textit{Infected cells} \qquad \frac{dy}{dt} = \beta x v - \alpha y, \tag{7}$$

$$\textit{Viral Particles} \qquad \frac{dv}{dt} = -uv, \tag{8}$$

$$\textit{Non-infectious virions} \qquad \frac{dw}{dt} = ky - uw. \tag{9}$$

The additional state w denotes a population of non-infectious virions created by protease inhibition. This state is required in addition to v, as although protease inhibition is assumed to halt the production of infectious virions, there is an initial pool of infectious virions yet to be depleted at the onset of treatment ($v(0) \neq 0$). Assuming that the uninfected cell population, x, remains constant

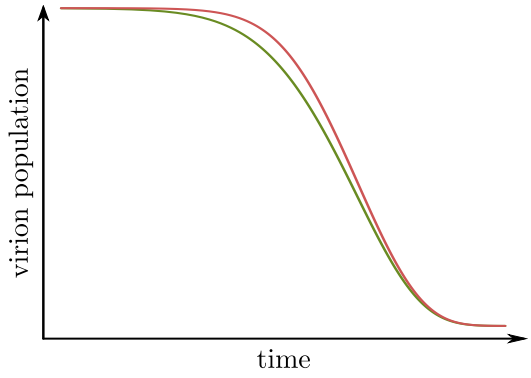


Figure 2.4: Illustration of the shoulder region for Equation (5), in green, and Equation (10) in red.

over the period of time of interest, analytical results for the response of the virion population can be derived:

$$v(t)+w(t)=v(0)\left(\exp(-ut)+\left(\frac{\exp(-\alpha t)-\exp(-ut)}{\alpha-u}u+\alpha t\exp(-ut)\right)\frac{u}{\alpha-u}\right). \tag{10}$$

Similar to the previous expression for the dynamics of the virion population, there is a shoulder region followed by an exponential decay (shown in Figure 2.4). However, in this case the virion population dynamics are no longer symmetric in α and u. This distinction allowed separate estimates of the virion and infected cell declines to be made. Models also consider longer term dynamics after treatment and its implication for decline in virions and infected cells. Estimates using the above model predict a half-life of infected cells $\sim$1-3 days [301, 117, 215, 216, 50]. This has since been revised down to $\sim \frac{1}{2}$ day [15]. The half-life of free virions is estimated to be $\sim$1-3 hours [236].

However, eradicating HIV-1 infection through antiretroviral treatment is not possible as long-lived latently infected cells (not included in the models described by Equations (1)-3 or (6)-(9)) maintain a viral reservoir. This result has led to the investigation of the CTL response, known to be important in fighting infection [26, 38, 27]. We now present two ODE models to investigate the importance of CTL responses. The first is an extension of the above framework, focusing on the dynamics of the CTL response itself [306], whilst the second explores the importance of CTL escape by examining competition within the host [15].

$CD4^+$ T-cells provide signals which lead to an immune response, yet HIV-1 viri-

ons have a strong affinity for CD4 via gp120 [126, 79, 45], so the cells helping to instigate an immune response are destroyed by the virus. A simplification of the CTL response can be obtained by splitting the T-cell population into T-helpers; x, precursors (cells which are responsible for the development of immune memory); w, and effectors; z [306, 307]. Precursors are assumed to proliferate a rate proportional to the product of the number of infected cells; y, and the number of uninfected helper cells; x. CTL precursors can differentiate into CTL effectors at rate $cqyw$ (this process is not assumed to require the presence of uninfected $CD4^+$ cells). Effector cells are then able to destroy infected cells; y, at rate pyz.

$$\textit{Helpers} \qquad \frac{dx}{dt} = \lambda - dx - \beta xv, \tag{11}$$

$$\textit{Infected cells} \qquad \frac{dy}{dt} = \beta xv - \alpha y, \tag{12}$$

$$\textit{Free virus} \qquad \frac{dv}{dt} = ky - uv, \tag{13}$$

$$\textit{Precursors} \qquad \frac{dw}{dt} = cxyw - cqyw - bw, \tag{14}$$

$$\textit{Effectors} \qquad \frac{dz}{dt} = cqyw - hz. \tag{15}$$

The mechanism described allows proliferation of precursors to occur at a lower rate than differentiation into effector cells for sufficiently low x. This results in a loss of CTL precursors and therefore the CTL response. The model provided some interesting outcomes depending on the rate of viral replication k. Low k results in the population z able to establish itself and take control of the infection. High k leads to extinction of the CTL precursors, resulting in a loss of the CTL response. Between these two cases, the equilibrium reached is dependent on initial conditions. The authors assume that HIV-1 lies in this region of the parameter space. Experiments in macaques [175] confirm the authors' conclusions, suggesting that the CTL response plays a fundamental role in immune response to HIV-1, and that memory arises as predicted by the model.

Although Equations (11)−(15) nicely recreate a simplified version of the CTL response, the authors recognise that they do not consider the potential for evolution of CTL escape mutations, which then have the potential to out-compete an unescaped 'wildtype' strain and overcome the hosts CTL response. A simple model incorporating strain competition within-host was put forward to attempt

to infer the impact of the CTL response from estimates of the rate at which an escape mutants out-compete the wildtype [15].

$$\textit{Wildtype} \qquad \frac{dy}{dt} = (a - b - c)y, \tag{16}$$

$$\textit{Escaped strain} \qquad \frac{dx}{dt} = (a' - b)x. \tag{17}$$

Wildtype and escape variant virus replicate at rates a and a' respectively. Depletion of the wildtype through CTL recognition of the epitope under consideration is assumed to occur at rate c, all other depletion is assumed to take place at rate b. Thus, the selective advantage gained from escape is summarised in the outgrowth of x compared to y, $(a' - b) - (a - b - c) = a' - a + c$. Assuming mutation is not limiting, this is the escape rate of the variant. If an escape variant is transmitted to a host in which it does not have a selective advantage, $c \to 0$ and thus the difference in fitness between the wildtype and variant results in a reversion rate of $a - a'$. This simple model helped highlight important differences between HIV-1 and SIV. CTL response in macaques destroys HIV-1 infected cells significantly more quickly that the corresponding CTL response in humans [14]. Differences between the two rates were approximately 0.22 day^{-1} per CTL response: a large discrepancy in CD8$^+$ killing of infected cells. In addition, it was determined that escape mutations in SIV lead to a higher fitness cost than in HIV-1. These findings demonstrate the need for understanding underlying differences in animal models of disease and suggest that CTL escape variants will pose a greater challenge to the immune response in humans.

2.6.2 *A model of CTL escape, transmission and reversion in the infected human population*

In order to consider selection of escape mutations at the level of the population, Fryer *et al.* [85] introduce a demographic model which includes transmission between hosts using the standard *SI* framework, but also allows CTL escape in HLA matched hosts and viral reversion of CTL escape mutants in HLA mismatched hosts. Two strains are considered: an 'escaped' strain and a 'wildtype'

strain, with distinct forces of infection. The force of infection exerted by strain i is $\Lambda_i(t) = \frac{\beta c}{N(t)}\left(Y_i^0(t) + Y_i^1(t)\right)$, where $N(t) = \sum_{n=1}^{2}\sum_{m=1}^{2}\left(\frac{1}{2}X^m(t) + Y_n^m(t)\right)$ is the size of the population at time t. Infected individuals have an increased mortality rate, specified by α. All variables and parameters of the model are summarised in Table 2.1.

Mismatched susceptibles

$$\frac{dX^0(t)}{dt} = (1-p)B - \left(\Lambda_0(t) + \Lambda_1(t) + m\right)X^0(t), \tag{18}$$

Matched susceptibles

$$\frac{dX^1(t)}{dt} = pB - \left(\Lambda_0(t) + \Lambda_1(t) + m\right)X^1(t), \tag{19}$$

Mismatched infecteds with wildtype

$$\frac{dY_0^0(t)}{dt} = \Lambda_0(t)X^0(t) + \lambda_{\text{rev}}Y_1^0(t) - (m+\alpha)Y_0^0(t), \tag{20}$$

Mismatched infecteds with escape

$$\frac{dY_1^0(t)}{dt} = \Lambda_1(t)X^0(t) - \lambda_{\text{rev}}Y_1^0(t) - (m+\alpha)Y_1^0(t), \tag{21}$$

Matched infecteds with wildtype

$$\frac{dY_0^1(t)}{dt} = \Lambda_0(t)X^1(t) - \lambda_{\text{esc}}Y_1^0(t) - (m+\alpha)Y_0^1(t), \tag{22}$$

Matched infecteds with escape

$$\frac{dY_1^1(t)}{dt} = \Lambda_1(t)X^1(t) + \lambda_{\text{esc}}Y_0^1(t) - (m+\alpha)Y_1^1(t). \tag{23}$$

Based upon cross-sectional population prevalences of the four infected classes, the authors were able to infer estimates of λ_{esc} and λ_{rev} for a collection of 26 epitopes. These estimates were then validated against longitudinal data where escape and reversion events can be observed *in vivo*. A strong correlation between the predicted rates and longitudinal data was observed. Estimates also revealed that despite individual studies demonstrating very fast escape and reversion at a number of epitopes [98, 225] the rate of escape and reversion at the population level is in fact far lower.

Variables	
t	Time
h	Host type (0 = HLA mismatched, 1 = HLA matched)
v	Virus type (0 = wildtype at epitope, 1 = escape mutant at epitope)
$X^h(t)$	Number of susceptible hosts at time t with host type h
$Y_v^h(t)$	Number of infected h hosts with virus type v
$\Lambda_v(t)$	Force of infection of virus type v at time t
Parameters	
β	Transmission probability per partnership
c	Contact rate
p	Proportion of hosts in the population with the relevant HLA type
λ_{esc}	Escape rate
λ_{rev}	Reversion rate
m	Mortality rate
α	Disease related mortality rate
B	Birth rate

Table 2.1: Variables and parameters in the Equations (18)-(23).

Implicit in the Equations (18) - (23) is the assumption that cross-sectional viral sequences are independent. This assumption is clearly violated. There is a dependency structure present in the viral genealogy that should be considered. Methods to incorporate this non-independence are discussed in the following sections.

2.7 THE COALESCENT, FELSENSTEIN'S TREE PEELING, AND MODEL BASED INFERENCE

When sequence data is considered and we look at presence and absence of a particular mutation taken from a sample in a population, it is important to realise that these samples are not independent due to shared ancestry. Population genetics helps us address this issue of independence by providing a framework in which we can use information in the viral sequence data and a series of assumptions to make statements about the process generating the underlying genealogy. Combining both pathogen dynamics and approaches from population genetics,

we hope to obtain more accurate estimates of parameters relevant to disease.

Coalescent theory is central to modern population genetics inference. The standard coalescent process with fixed population size describes of the ancestral history of a small sample from a large constant population that is evolving according to a neutral model of sequence evolution [149]. The basic idea is that in the absence of selection, one can generate the ancestral history backwards in time by letting individuals (viruses in our case) pick their parents at random from the population. When two individuals pick the same parent, a coalescence event occurs and the number of lineages to keep track of decreases, until it eventually reaches one, at the most recent common ancestor. The coalescent framework can be extended and adapted to model changes in effective population size [108, 230, 274, 64, 188], selection [197], recombination [124] or population structure [294, 246, 202].

Once we have a model governing the genealogy of a collection of viral sequences, we are interested in the probability of the data (sequence or otherwise) when presented with a particular genealogy. Felsenstein's peeling algorithm [76] allows us to do exactly this. The idea goes as follows: suppose we knew the sequence history at each internal node. We have some parameters governing sequence evolution, and so can evaluate the probabilities of observing the sequence evolution seen along each of the branches. To determine the likelihood of the whole genealogy, we just take the product of all of these probabilities. We don't know the sequences at the internal nodes. However, it turns out that we can integrate over all of the possibilities at each internal node in a computationally efficient manner.

If investigators are interested in the underlying evolutionary and epidemiological processes which generated the true underlying genealogy, but not the genealogy itself, then it makes sense to use Bayesian approach to take the genealogy into consideration. We will use the coalescent as a prior on tree structure and perform MCMC, but note that we may have chosen other priors. For example, birth-death processes can be used to describe distributions of genealogies. The generalised birth-death process was first formalised by Kendall in 1948 [142]. Distributions have been derived for the full process [279, 196, 95] and in the presence

of incomplete sampling [196, 310, 267]. More recently, distributions have been described to allow for time-stamped sampling of sequence data and variable birth and death rates [270, 268]. The coalescent and birth-death process opens the door for estimation of population genetics model parameters from molecular sequences. Using priors such as the coalescent or birth-death process are important for phylodynamic approaches. They both enforce a molecular clock. This is a requirement when using a genealogy to estimate parameters of a dynamical model as it forces branch lengths to be measured in units of time. Using a coalescent prior requires a further parameter to perform phylodynamic inference if epidemiological parameters are to be estimated in meaningful units (*weeks*, *months*, *years* etc). The MCMC scheme introduced by Drummond *et al.* [65, 66] co-estimates a scaling parameter to switch from 'coalescent time' to 'calendar time'. This extra parameter is estimated by considering the difference in sampling time between data at the tips (if we know how long a year was on the tree and we know how long has elapsed in coalescent time we may create a scaling to move between the two) and using an extension of the coalescent prior which allows for non-contemporaneous tip dates, known as the serially-sampled coalescent [243, 2]. MCMC exploration of the space of genealogies and evolutionary parameters is performed using a set of carefully constructed MCMC moves. The MCMC scheme allows parameters relevant to epidemiology for a wide variety of pathogens to be estimated with the aid of sequence information. Some examples are provided in Section 2.8.

Sometimes the assumption of a molecular clock is strongly violated, particularly in viruses [130] and thus rate variation present within lineages can alter conclusions we made from the analysis. There is much debate over whether this is the case for HIV-1 [171, 159, 222, 130]. The introduction of 'relaxed clock' models [67] represents middle ground between a 'strict' molecular clock in which substitutions occur at some fixed rate independent of lineage, and phylogenetic approaches such as that initially put forward by Felsenstein [76] where lengths of lineages within the unrooted tree are systematically altered in order to obtain a maximum likelihood tree. However, such an approach means that rate of substitution and time cannot be separated. So how do relaxed clocks differ?

Rather than assuming a constant rate across all lineages, relaxed clocks allow for different rates sampled from some distribution. They come in two flavours, firstly we may suggest that there is some correlation between the rates found in the parental lineage and those found in the daughter lineages. This is implemented by letting the mean of the distribution from which the rate is sampled be a function of the rate on the parental branch. Alternatively, we may assume that there is no correlation between branches and simply sample the rate along each branch from some distribution: uncorrelated relaxed clocks. A collection of relaxed clock models are implemented in BEAST [67].

For the models that we present in Chapters 3 and 6, we will take a population genetics approach. In Chapter 3 we will use BEAST to integrate over genealogies (G) and a set of evolutionary parameters (Θ) to perform inference on parameters relevant to disease. In Chapter 6 we avoid sampling genealogies entirely, using an approximation by Li and Stephens [167] to directly estimate the integral over G. The motivation for such an approximation is simple: performing statistical inference within a coalescent framework is difficult. Li and Stephens made a simple observation; the coalescent is itself an approximation to the true process. They suggested that if we create a model that retains some useful properties of the coalescent but has an easily computable likelihood, then we can perform inference on this simpler object much more rapidly, even allowing for recombination. Their idea worked. Li and Stephens created an approximate likelihood which is amenable to inference, even in the presence of recombination. And, despite the simplifying assumptions, the approximation to the coalescent performs well in practice. The model we present is itself based on the model of Wilson and McVean [304], which uses the Li and Stephens approximation to infer both selection and recombination along sequences.

2.8 EFFORTS TO INCORPORATE THE UNDERLYING GENEALOGY IN PARAMETER ESTIMATION

In the next Chapter we will describe the method we use to combine coalescent theory with epidemic modelling to attempt to pin down parameters of interest. The use of tree building as a method to investigate the dynamics of an epidemic was first used in [118] to track viral transmission of HCV and HIV-1 via the branching structure in phylogenetic trees. Each lineage of the tree was assumed to be present in a single individual, and a plot of lineages over the length of the tree was used to make qualitative statements about the history of the epidemic. Later, coalescent theory was used to estimate the HIV-1 viral generation time *in vivo* (it was estimated to be 1.2 days per generation, similar to estimates of around 1.8 days per generation using ODE models) [243] under an assumption of constant virion population size within the host.

Using a more complicated model of population structure over time (not assuming constant or exponential growth of the infected population $N(t)$ over time), the first paper to explicitly link coalescent and epidemiological based modelling was used to estimate the basic reproduction ratio (R_0) of the hepatitis C virus (HCV) [232] in the Egyptian epidemic. A model of HCV spread was considered and maximum likelihood approaches were used in order to parameterise it and thus estimate R_0. This same approach was followed in [131] to look for differences in R_0 between HCV subtypes, and to determine an estimate of R_0 for HIV-1.

Bayesian coalescent analysis has been used to look for geographical structure present in the genealogy of a recombinant rabies virus (RRV) found in raccoons. Spatial data was laid over the maximum clade credibility (MCC) tree. R_0 was then inferred at distinct phases of the epidemic [23]. Bayesian coalescent analysis has also been used to determine the time to the most recent common ancestor (T_{MRCA}) in the 2009 H1N1 pandemic, the MRCA was estimated to be at 12 January 2009. Authors assumed the epidemic was in the exponential growth phase and used their estimate of the T_{MRCA} as an upper bound [80]. A similar setup is used to estimate growth rates of HIV-1 subtypes [296]. The coalescent is

generated under simple models of population growth, these are compared, and parameters for the growth of the infected population are then inferred.

Moving on to applications of the Li and Stephens approximation [167], highly cited examples include measuring haplotype sharing across human populations in the HapMap project [51], imputing missing genotypes as part of data preparation in a set GWA studies used to look for causal SNPs of common complex human diseases [3], and to create a fine scale map of recombination across the human genome [291]. The model can be extended in various ways. Recently it has used to impute HLA types using SNP data [59], which in turn is highly valuable when looking for risk factors in autoimmune and infectious diseases. Wilson and McVean extended the model to allow for co-estimation of recombination rates and selection along sequence data [304]; which in turn has numerous applications. So far these include detecting balancing selection in Mexican salamanders *Ambystoma mexicanum* as part of an investigation into the species being characterised as immunodeficient, and investigating selection in malaria [25].

The total number of papers which include models combining genealogical and epidemiological models is small. This lack of knowledge in the link between disease evolution and epidemiological processes is highlighted in [107]. More recently, papers defining a formal framework for the relationship between coalescent theory and the classic ordinary differential equation models of epidemics (SI, SIS, SIR) have been published [292, 84, 155, 270], which will hopefully lead to increased activity in this exciting area.

3

INTEGRATING GENEALOGICAL AND DYNAMICAL MODELLING TO INFER ESCAPE AND REVERSION RATES IN HIV-1 EPITOPES

The material in this Chapter is published as 'Integrating genealogical and dynamical modelling to infer escape and reversion rates in HIV epitopes' Palmer *et el.* in Proceedings of the Royal Society B: Biological Sciences, Vol. 280, No. 1762.

3.1 INTRODUCTION

The rates of escape and reversion in response to selection pressure arising from the host immune system, notably the cytotoxic T-lymphocyte (CTL) response, are key factors determining the evolution of HIV-1. Existing methods for estimating these parameters from cross-sectional population data using ordinary differential equations (ODE) ignore information about the genealogy of sampled HIV-1 sequences, which has the potential to cause systematic bias and over-estimate certainty. Here, we describe an integrated approach, validated through extensive simulations, which combines genealogical inference and epidemiological modelling, to estimate rates of CTL escape and reversion in HIV-1 epitopes. We show that there is substantial uncertainty about rates of viral escape and reversion from cross-sectional data, which arises from the inherent stochasticity in the evolutionary process. By application to empirical data, we find that point estimates of rates from a previously published ODE model and the integrated approach presented here are often similar, but can also differ several-fold depending on the structure of the genealogy. The model-based approach we apply provides a framework for the statistical analysis and hypothesis testing of escape and rever-

sion rates in population data and highlights the need for longitudinal and denser cross-sectional sampling to enable accurate estimate of these key parameters.

Cytotoxic T-lymphocytes (CTLs) are implicated in the control of human immunodeficiency virus 1 (HIV-1). In fact, they are thought to be the most important mediators in reducing viraemia in individuals able to control HIV-1 infection, and certain HLA-types and high levels of CTL activity are associated with repression of viral replication in long-term non-progressors (LTNPs) [186, 239, 221]. Epitopes are presented to CTLs by human leukocyte antigen (HLA) class I proteins at the surface of almost all nucleated cells in the body. The collection of epitopes which may be presented by the HLA class I molecules is determined by an individual's combination of alleles at these highly variable loci. Mutations in or close to epitopes in the viral sequence can result in alterations to the binding affinity to the HLA class I, reduce CTL recognition, or abrogate T-cell receptor (TCR) binding. Such mutations are known as escape mutations. Examples of escape mutations have been described in almost all proteins encoded in the HIV-1 genome [9, 27, 61, 141, 154, 153, 183, 225, 312], with the strongest signal of association with host HLA type found at the HLA-B locus [145, 101]. After an escape event takes place, viruses with escape mutations may be transmitted between individuals and thus have the potential to spread across the infected population [9, 81], or they may revert through selection pressure within hosts whose immune responses do not drive escape in a given epitope [82, 163]. Associations between HLA type and putative CTL escapes have been demonstrated statistically in population studies [189], though these results have been called into question [22], and it is suggested that the frequency at which escape events take place is lower than previously thought. More recent studies [98, 87, 169] have shown that there is a large variation in time to escape observed across epitopes, ranging from days to years. There is strong interest in understanding the selective pressures applied to the virus at the level of the population as there are clear implications for any putative vaccine. To date, simple ordinary differential equation (ODE) based models have been used to estimate the expected time to escape and reversion by using cross-sectional data across hosts. Such estimates make use of only a small portion of the available data, namely presence or absence of an escape

mutation and the HLA type of the sampled hosts (we denote this collection of pairs E), and disregard any remaining sequence information. These methods also make assumptions about the independence of the sampled data which could potentially lead to bias in parameter estimates. Furthermore, deterministic models only provide point estimates and thus cannot provide meaningful confidence regions which account for phylogenetic uncertainty.

Figure 3.1 illustrates the inference problem that we are addressing. We wish to infer three rates; the rate of viral escape (switching from dark green to dark red in Figure 3.1a), the rate of viral reversion (switching from pink to light green in Figure 3.1a), and the transmission rate. If the underlying transmission tree was known, the problem would be straightforward. However, we only have a collection of tip labellings (sequence data and HLA information) which are the culmination of an embedded subtree of the full process. To make statements about parameters of the full transmission tree, we must reconstruct the subtree together with the dynamic processes occurring along its lineages through time.

We combine dynamic programming (Felsenstein's peeling algorithm [76]) with existing software (BEAST [65]) to merge statistical phylogenetic approaches with well studied ODE based modelling. This allows us to incorporate viral sequence data into our estimates. By doing so, we generate more informed estimates and credible regions of population level escape and reversion rates, $(\lambda_{\text{esc}}, \lambda_{\text{rev}})$, which incorporate the underlying dependency structure present in the viral genealogy. There are four key steps in our inference:

1. Make the mild assumption that the genealogy; G, and HLA/escape information; E, are conditionally independent given the sequence information with the epitope removed; X.

2. Genealogies are sampled from the posterior conditional on X using the program BEAST [68], with a coalescent prior for an exponentially growing population whose rate parameter is sampled in the MCMC.

3. For each sampled genealogy, we evaluate the posterior density for $(\lambda_{\text{esc}}, \lambda_{\text{rev}})$ using a modification of Felsenstein's peeling algorithm [76] based on the processes in Figure 3.1a.

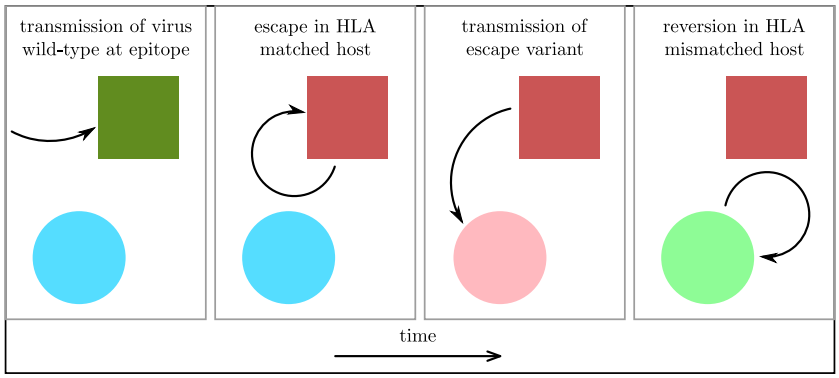
4. Tree specific posteriors are averaged to produce a final posterior density for $(\lambda_{esc}, \lambda_{rev})$.

We envisage similar methodologies being applied to a wide range of problems. Our model represents an addition to the highly active area of phylodynamics, in which both stochastic and deterministic approaches are being developed [155, 237, 84].

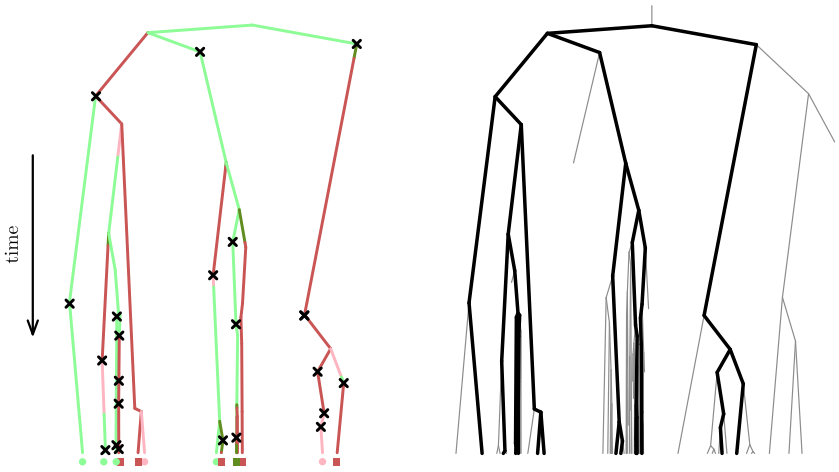
To test the robustness of our integrated method, we perform a series of simulations and compare the results to those of an existing ODE based dynamical model [85]. Notably, our method does not require assumptions about the start time of the epidemic, or transmission rate during the exponential growth phase, as these are estimated by the model. When nucleotide substitution rates are fixed at estimates generated from empirical sequence data, we find that in simulation studies our model successfully estimates escape and reversion rates. By altering the nucleotide substitution rate, we find that a lack of information about the genealogy (through lower substitution rates) can dramatically affect escape and reversion rate estimates using our integrated approach, though we find that the rates of substitution found in HIV-1 are sufficiently large for this effect to be considered negligible.

The integrated approach is then applied to estimate escape and reversion rates in four previously identified epitopes. The four epitopes were chosen in order to explore as much of the space of escape and reversion rates as possible on the basis of previous estimates generated using population level data. Again, we compare our integrated method to the ODE method which generated these estimates [85]. We illustrate the benefit of setting our approach in a model-based framework by demonstrating some simple hypothesis tests.

Our model provides evidence for the hypothesis that rates of escape and reversion within-host are slower than published estimates generated from individual case studies [98], and highlights the large amount of uncertainty inherent in estimates that make use of cross-sectional population data.



(a)



(b)

Figure 3.1: The inference problem: A reconstructed sampled birth-death process with an overlaid Poisson process. The cartoon in Figure 3.1a displays the dynamic processes which may occur along a branch within the transmission tree. Time increases from left to right. A susceptible individual is shown in light blue. An individual infected with a virus which is wild-type at the epitope under consideration is green, and individuals with a viral strain with the escape mutation are red. HLA matched hosts are darker rectangles, HLA mismatched hosts are lighter circles. From left to right, transmission of a variant which is wild-type at the epitope under investigation occurs within the population. This strain may escape within an HLA matched host. Transmission of an escaped viral strain to an individual who is HLA mismatched can occur. This strain may then revert in this HLA mismatched host. Viruses exist within these environments over their evolutionary history. Thus, associated to a collection of individuals sampled at the present (shown at the tips in Figure 3.1b) is a colour coded transmission tree, illustrated in Figure 3.1b. A transmission event is associated with each coalescence, but due to incomplete sampling, unseen transmission events also occur. These are shown by black crosses. This sampled transmission tree is embedded in the full transmission tree, with the unseen portions of the tree shown in grey. We have sequence data and colourings at the tips of a sampled transmission tree. Using these data we hope to reconstruct the embedded subtree in Figure 3.1b, and use this reconstruction to make inferences about the unknown full transmission process.

3.2 METHODS

We wish to estimate escape and reversion rates within host, taking into account the dependency structure between sampled individuals arising from the phylogenetic tree. Throughout, we define hosts with an HLA type known to bind the epitope under consideration as HLA matched, and those without such an HLA type as HLA mismatched. We have the HLA types and cross-sectional viral sequences from a collection of hosts, taken from the Swiss-Spanish intermittent treatment trial (SSITT) [73, 82]: a collection of 79, 67, and 53 viral sequences from HLA typed hosts for the genes *gag*, *pol* and *nef* respectively. Epitopes and associated HLA types are considered one at a time, independent of other epitopes. We consider four previously defined epitopes. These epitopes were chosen in order to test our method across a wide range of escape and reversion rates - based on previous rate estimates [85]. The chosen epitopes are shown in Table 3.1, column 1. Throughout, we abbreviate these epitopes by their first three amino acids (e.g. TST). By removing the epitope under consideration from sequences and determining the presence or lack of an escape mutation, we consider data from two processes.

1. The sequence data with epitope removed, X, allows us to perform inference on the genealogy; G.
2. The combination of HLA type and presence or lack of escape, E (which we refer to as HLA/escape information), provides information about the dynamical processes shown in Figure 3.1a, which occur along the lineages of G over time.

By assuming escape information is uninformative about G; G and E are conditionally independent given X ($P(G|X,E) \approx P(G|X)$). This allows us to consider these two processes separately, with the second conditional on the first. Full details are provided in Section 4.2. Adding a collection of time-stamped data taken from the Los Alamos HIV sequence database [256] to these HLA typed cross-sectional sequence data, we create a DNA multiple sequence alignment [91] and perform some data trimming. These time-stamped data are required for BEAST

to estimate a scaling from units of time measured in *generations* to *years*. After data-trimming, the resulting number of HLA typed SSITT sequences are 55, 54 and 48 for *gag*, *pol* and *nef* respectively. Given the alignment, genealogies are sampled using BEAST with a standard coalescent prior for an exponentially-growing population whose rate parameter is estimated in the MCMC, which is then used to estimate the transmission rate (this is possible since the effective infected population size is proportional to the number of infected individuals under our model, see Section 4.4.9). Taking a sample of genealogies from the BEAST output, we consider the embedded subtree for which we have HLA and escape information, and determine the likelihood of parameters $(\lambda_{\text{esc}}, \lambda_{\text{rev}})$ using a modification of Felsenstein's peeling algorithm [76]. Our peeling algorithm is based on the transitions shown in Figure 3.1a. Defining $(a, b) : a, b \in \{0, 1\}$, where $a \in \{0, 1\}$ denotes {HLA mismatch, HLA match}, and $b \in \{0, 1\}$ denotes {no escape mutation, escape mutation} in the epitope under investigation.

$$\mathbf{Q} = \begin{array}{c} \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \begin{pmatrix} -\hat{\lambda}(t)q & 0 & \hat{\lambda}(t)q & 0 \\ \lambda_{\text{rev}} & -\lambda_{\text{rev}} - \hat{\lambda}(t)q & 0 & \hat{\lambda}(t)q \\ \hat{\lambda}(t)(1-q) & 0 & -\lambda_{\text{esc}} - \hat{\lambda}(t)(1-q) & \lambda_{\text{esc}} \\ 0 & \hat{\lambda}(t)(1-q) & 0 & -\hat{\lambda}(t)(1-q) \end{pmatrix} \end{array}. \tag{24}$$

In our model, newly infected individuals are generated at rate λ, individuals are assumed to be HLA matched with probability (i.e. frequency in the population) q. Transitions between the states in the continuous time Markov chain may be described by the instantaneous rate matrix $\mathbf{Q}$ in Equation (24), where $Q_{i,j}$ describes the transition from state i to state j. $\hat{\lambda}(t) = \lambda p_0(T_{\text{MRCA}} - t)$, where t increases towards the present. T_{MRCA} is the time before the present of the most recent common ancestor (MRCA), when $t = 0$. $p_0(t)$ is the probability that a lineage at time t in the past does not have any sampled descendents [310], and depends on the sampling proportion at the present, ρ. Details of the derivation of $p_0(t)$ excluded from the original paper are provided in Section 4.3. The $p_0(T - t)$ scaling

is required to avoid double counting of transmission events. We assume the state at the root node does not have the escape mutation. In our peeling algorithm we must account for the fact that each internal node in the sampled genealogy corresponds to exactly one transmission event. For each sampled genealogy, we evaluate the two-dimensional posterior surface for the parameters of interest $(\lambda_{esc}, \lambda_{rev})$ using our peeling algorithm on a 50×50 lattice of parameter space which encompasses high probability density (within a factor of 1000 of the maximum density as determined by hill-climbing), conditional on E with a uniform prior over $[10^{-20}, 10^{3}] \times [10^{-20}, 10^{3}]$. We then integrate over 1000 samples from the BEAST output to produce a final posterior density for $(\lambda_{esc}, \lambda_{rev})$, and define credible regions for our estimates. A description of the data and full details of the method are provided in Section 4.4.

As in ODE models of escape and reversion [85], we make a variety of assumptions. We assume homogeneous mixing in the infected population. We suppose that the infected population is in exponential growth and assume constant rates of escape in HLA matched hosts, and constant rates of reversion in HLA mismatched hosts. We ignore variation within individuals' viral populations, by assuming that 100% of each host's viral population either has an escape variant in the epitope under consideration, or has the wild type (as shown in Figure 3.1a). All escape mutations within a given epitope are assumed to occur at the same rate, and HLA types are considered to two digits. It is assumed that a single individual seeded the epidemic, and we suppose individuals with the corresponding HLA type are always able to make an immune response, and drive escape at the same rate. Finally, recombination is not considered in our estimation of the genealogy; G.

3.3 RESULTS

3.3.1 *Simulations*

Under our model, assuming the population of infected individuals is in exponential growth, a birth-death process [142] generates the full transmission history, and after sampling, a time-inhomogeneous Poisson process governed by $\mathbf{Q}$ generates states in $\{(0,0), (0,1), (1,0), (1,1)\}$ at the tips. The birth rate (of infected hosts) is λ, and the death rate, μ, is equal to the rate of becoming non-infectious (through death or otherwise). Throughout our simulations, $q = 0.15$, $\lambda = 0.45534\ years^{-1}$ (established from the average of a BEAST run on *gag* data), $\mu = 0.1\ years^{-1}$ [85, 191]. Where required, we sample 500 genealogies from BEAST output. We find that this is sufficient due to the low variance in *maximum a posteriori* (MAP) estimates across sampled trees. This is a consequence of the uncertainty in the genealogy being concentrated at deep nodes, which have little impact upon the likelihood of observing data at the tips. We find the most recent coalescent events, which have far more power to inform estimates, are more consistent across sampled genealogies. There are three steps in our data simulation:

1. Generate a full birth-death tree forwards through time with an MRCA of 25 years.

2. Simulate human leukocyte antigen (HLA) class I and escape information along lineages of this birth-death tree. Set the sampling proportion at the present, ρ, such that the expected number of present day tips is 200, and sample extinct tips at rate ν such that the expected number of historically sampled tips is 50. For step 3 we consider the embedded subtree defined by the sampling and discard HLA/escape information at historically sampled tips.

3. Simulate sequence information at the tips, X, each 500 nucleotides long, using mutation parameters, Θ (set at the average of the required parameters from a BEAST run on SSITT *gag* sequences) and a $GTR + I + \Gamma$ model

of substitution (see Section 4.4.6 for a description). Historical sequences are required by BEAST to estimate a scaling from time measured in *generations* to *years*.

3.3.1.1 *Testing the integrated method and comparison to existing approach*

We test and compare our model to a differential equation approach previously described [85]. It can be shown that the dynamics of our model when the sampling proportion is zero match the relative proportions through time of the ordinary differential equation (ODE) model during exponential growth, and that this is equivalent to assuming a completely star-like genealogy in which all lineages emanate from the most recent common ancestor (MRCA), see Section 4.5. We perform three simulation studies:

a) For each of the $(\lambda_{esc}, \lambda_{rev})$ parameter sets $\{(2, 0.5), (0.5, 0.5), (0.5, 0.05), (0.05, 0.05), (0.05, 0.01)\}$ $years^{-1}$, we apply steps 1 - 3. Using HLA/escape information at the present day tips, E and sequence information X, we apply the integrated method and ODE model. This process is repeated four times for each $(\lambda_{esc}, \lambda_{rev})$, and we compare the results. The ODE model requires an estimate of the transmission rate, death rate and the initiation time of the epidemic. We fix these parameters at their true values. We bootstrap E 10,000 times to provide an estimate of sampling uncertainty under the ODE model. Estimates of μ and ρ are required under the coalescent tree prior in an exponentially growing population, which we set at the true value. In order to define an approximation to confidence regions for sampling under the ODE method, we use the Mahalanobis distance measure [176] (see Section 4.4.12) widely used in cluster analysis, which takes covariance between escape and reversion rates into consideration. Two instances of estimating the parameter set $(0.5, 0.05)$ $years^{-1}$ are shown in Figure 3.2a. Rate estimates based on the four simulations given the five parameter sets are shown in Figures 3.3 - 3.7.

b) We investigate further by running steps 1 - 2 1000 times for each of the parameter sets in Figure 3.1a comparing the MAP estimate under the integrated model on the true tree obtained in step 1 to the ODE point estimate. Results

are shown in Figures 3.2b and 3.10. We also check how both methods vary with the number of sampled tips by running this simulation on samples of size 60, 100, 200 and 500 in Figures 3.8 - 3.11 respectively. This approximation to the full integrated approach is reasonable as we find point estimates for the MAP under the full method are similarly spread (e.g. compare Figure 3.2c to Figure 3.8 row 1). The number of samples in the empirical data in the Swiss-Spainish intermittent treatment trial (SSITT) for the genes *gag*, *pol* and *nef* is 55, 54, and 48 respectively.

c) We examine the ability of the integrated method to estimate true underlying parameters over a large number of simulations. Setting the truth at $(\lambda_{esc}, \lambda_{rev}) = (2, 0.5)$ $years^{-1}$, we run steps 1 - 3 100 times and, for each, apply our full integrated method. Results are shown in Figure 3.2c.

3.3.1.2 *Robustness to tree topology and impact of mutation rate*

In order to estimate the impact of uncertainty in the tree topology we performed two tests. Firstly we run steps 1 - 2, permute tip labellings E in the true tree 250 times, re-estimating $(\lambda_{esc}, \lambda_{rev})$ each time using our peeling algorithm. Setting the true values at $(\lambda_{esc}, \lambda_{rev}) = (0.6, 0.1)$ $years^{-1}$. $(\lambda_{esc}, \lambda_{rev})$ is set at an intermediate escape and reversion rate as such rates lead to both diversity and clustering in the true tree. Secondly we apply steps 1 - 3, but multiply each substitution rate in step 3 by a factor of 5 and 0.2, before applying the integrated method. The truth is set at $(\lambda_{esc}, \lambda_{rev}) = (2, 0.5)$ $years^{-1}$.

3.3.2 *Simulation results*

3.3.2.1 *Simulations with known underlying tree (simulation study b))*

By determining MAP of 1000 instances of HLA and escape data, supposing we know the true underlying genealogy, we find our integrated method best estimates the true rates when the truth lies in the centre of our range of parameter simulations. Results are shown in Figures 3.2b and 3.10. The proportion of MAP

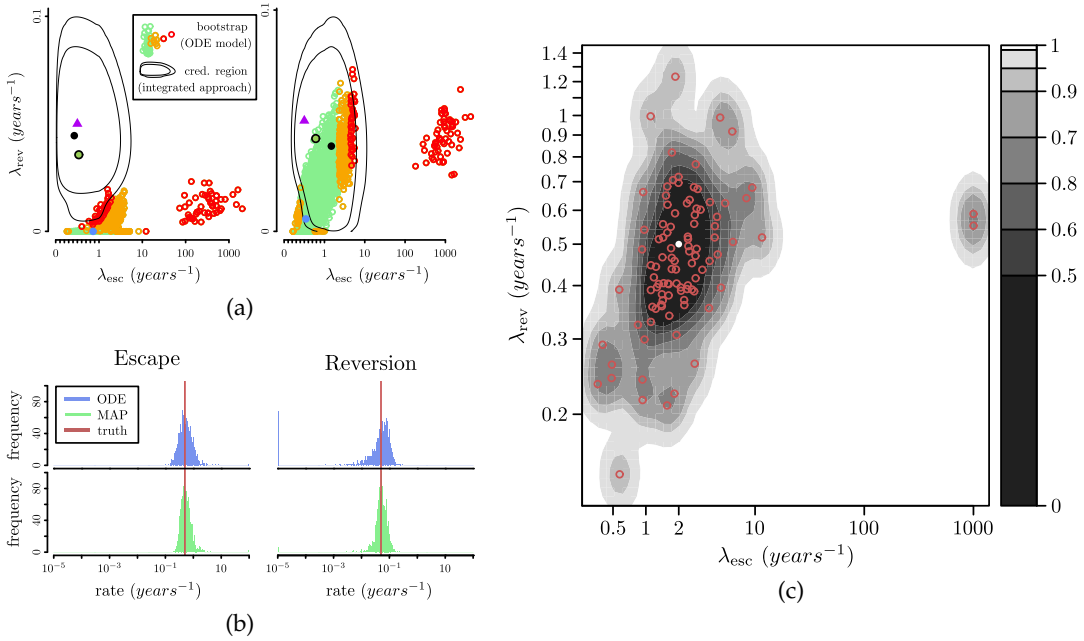


Figure 3.2: Summary of simulation results. For each simulation, we generate a full birth-death process forwards through time and set the sampling proportion; ρ, such that the expected number of sampled tips is 200. Figure 3.2a shows two instances of simulations under the integrated approach and the ODE model, the truth is set at $(\lambda_{esc}, \lambda_{rev}) = (0.5, 0.05)$ $years^{-1}$. Estimates of 10,000 bootstraps of the data under the ODE model are shown as dots, coloured by Mahalanobis distance (the furthest 1% and 1 - 5% are coloured red and orange respectively, and the remainder green - see methods). The ODE point estimate based on the tip data is the blue dot. 95% and 99% credible regions of the integrated method are in black. The MAP is the black dot. The MAP if the genealogy is known is the circled green dot. The purple triangle is the truth. Minimum and maximum escape and reversion rates in the 95% credible region are $(0.0269, 3.14)$ and $(0.0177, 0.0859)$ $years^{-1}$ respectively for the first plot, and $(0.256, 6.21)$ and $(1.90 \times 10^{-4}, 0.0834)$ $years^{-1}$ for the second. Figure 3.2b shows MAP and ODE point estimates of 1000 simulations assuming the true tree is known. Histograms of ODE point estimates and MAP estimates are shown in blue and green respectively. The truth is overlaid in red. Rates $< 10^{-5}$ and $> 10^{2}$ are grouped. Figure 3.2c shows the results of 100 simulations of the integrated model applied to data generated from underlying rates $(\lambda_{esc}, \lambda_{rev}) = (2, 0.5)$ $years^{-1}$, shown as a white dot. 100 MAP estimates are shown in red. Contours are coloured using a 2D kernel density estimate [297, 298].

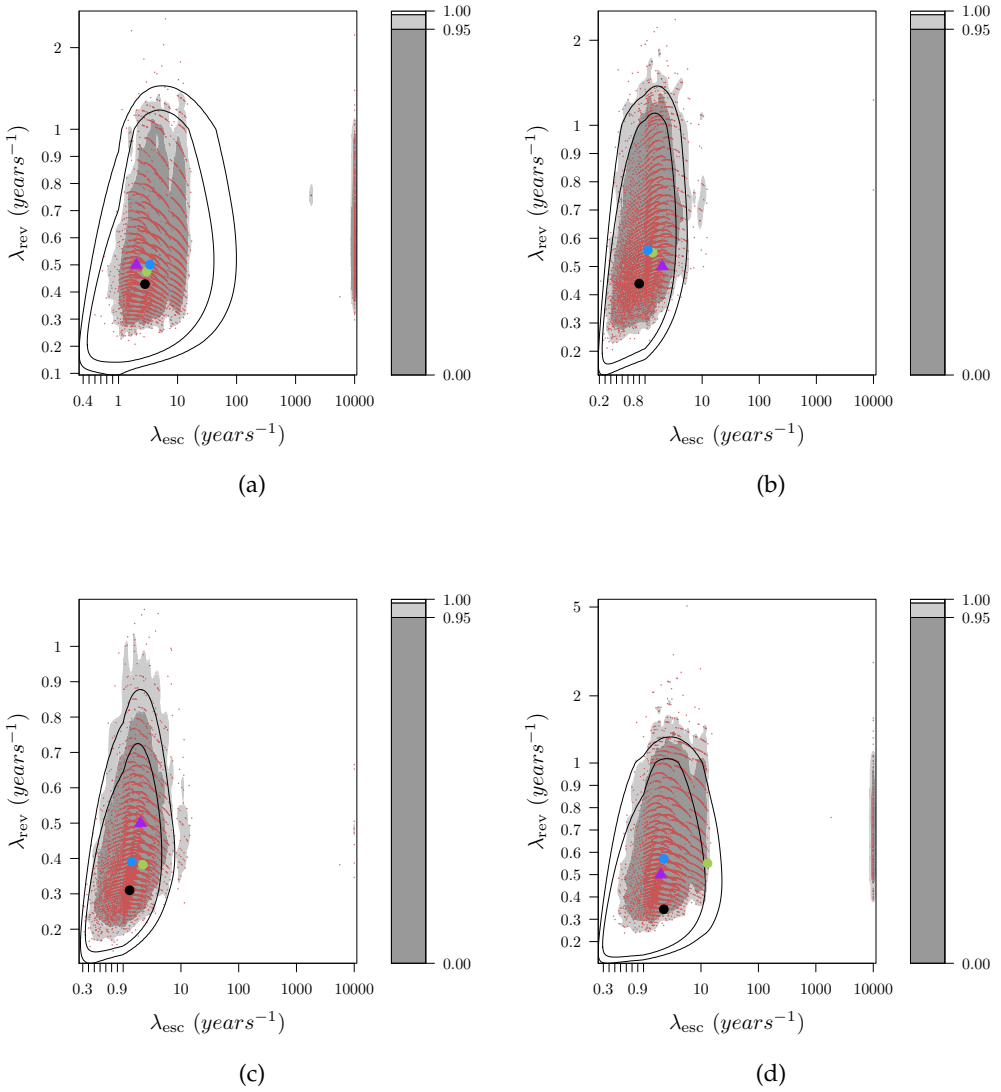


Figure 3.3: Four simulations with fixed escape and reversion parameters. True underlying rate $(\lambda_{esc}, \lambda_{rev}) = (2, 0.5)$ $years^{-1}$. 95% and 99% credible regions are shown in black. 10,000 ODE point estimates are shown in red, and contours coloured using a 2D kernel density estimate [297, 298] according to the key. The true rates are shown as a purple triangle, the MAP estimate is shown in black, and ODE point estimate in blue. The MAP based on the true underlying transmission tree is shown as a green outlined dot.

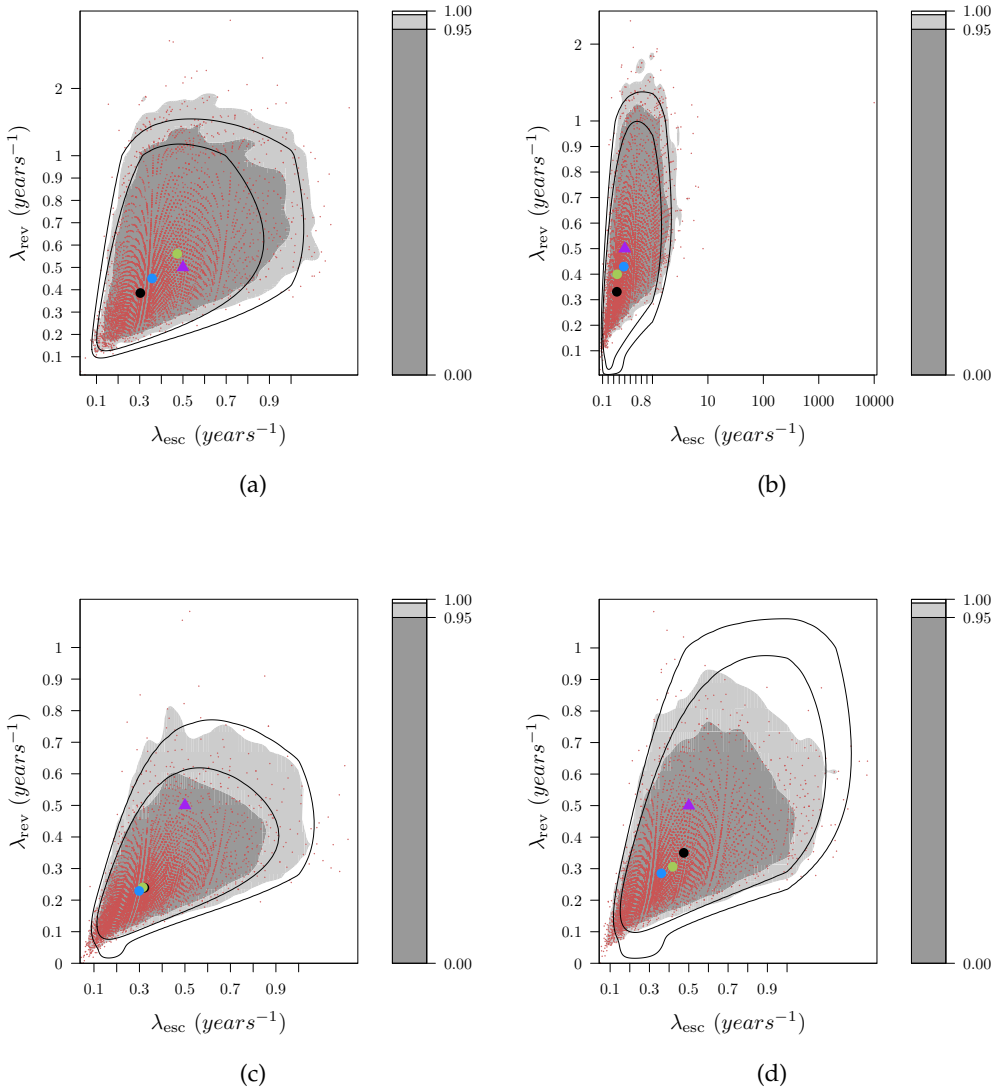


Figure 3.4: Four simulations with fixed escape and reversion parameters. True underlying rate $(\lambda_{esc}, \lambda_{rev}) = (0.5, 0.5)$ $years^{-1}$. 95% and 99% credible regions are shown in black. 10,000 ODE point estimates are shown in red, and contours coloured using a 2D kernel density estimate [297, 298] according to the key. The true rates are shown as a purple triangle, the MAP estimate is shown in black, and ODE point estimate in blue. The MAP based on the true underlying transmission tree is shown as a green outlined dot.

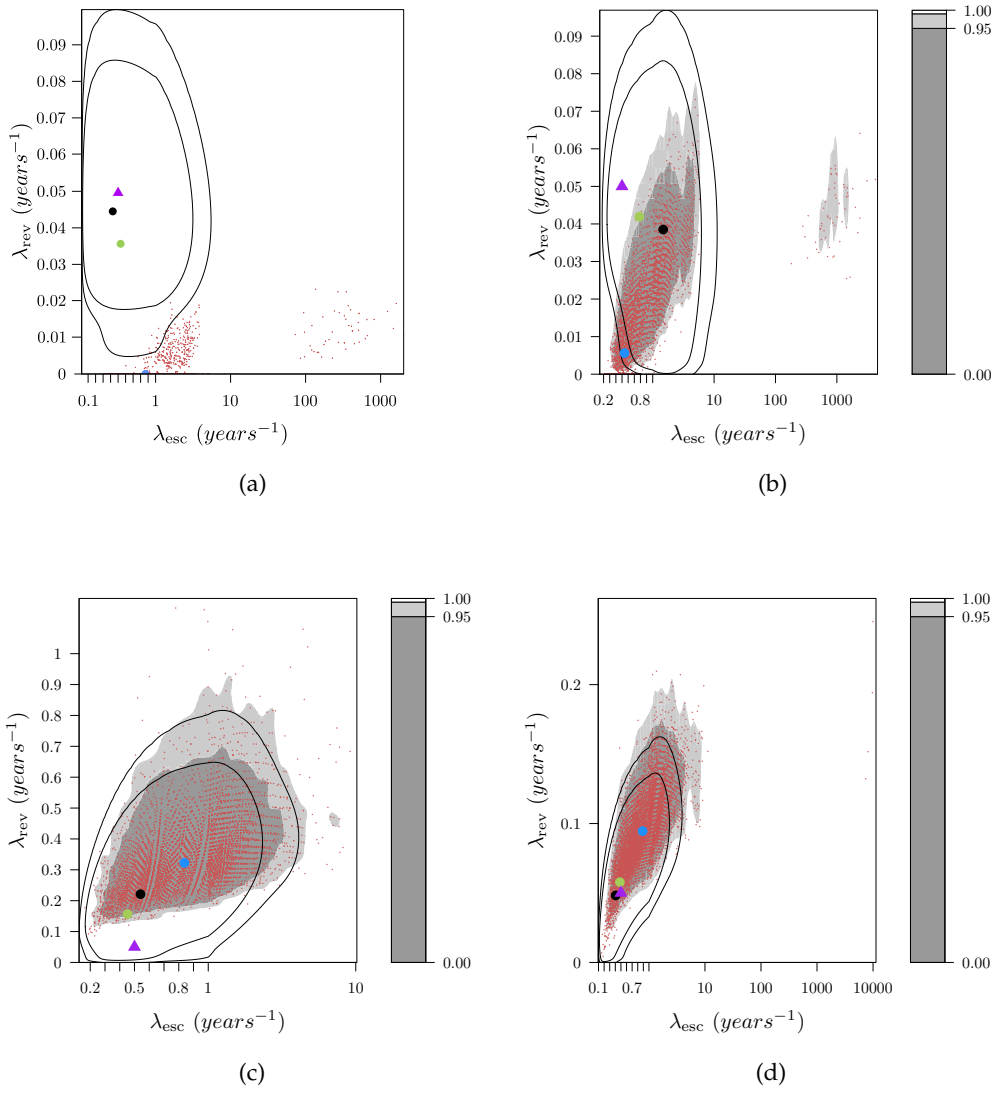


Figure 3.5: Four simulations with fixed escape and reversion parameters. True underlying rate $(\lambda_{esc}, \lambda_{rev}) = (0.5, 0.05)$ $years^{-1}$. 95% and 99% credible regions are shown in black. 10,000 ODE point estimates are shown in red, and contours coloured using a 2D kernel density estimate [297, 298] according to the key. The true rates are shown as a purple triangle, the MAP estimate is shown in black, and ODE point estimate in blue. The MAP based on the true underlying transmission tree is shown as a green outlined dot.
$(\lambda_{esc}, \lambda_{rev}) = (0.5, 0.05)$ $years^{-1}$. An estimate of the density for Figure 3.5a could not be established due to the concentration of point estimates along $\lambda_{rev} = 0$.

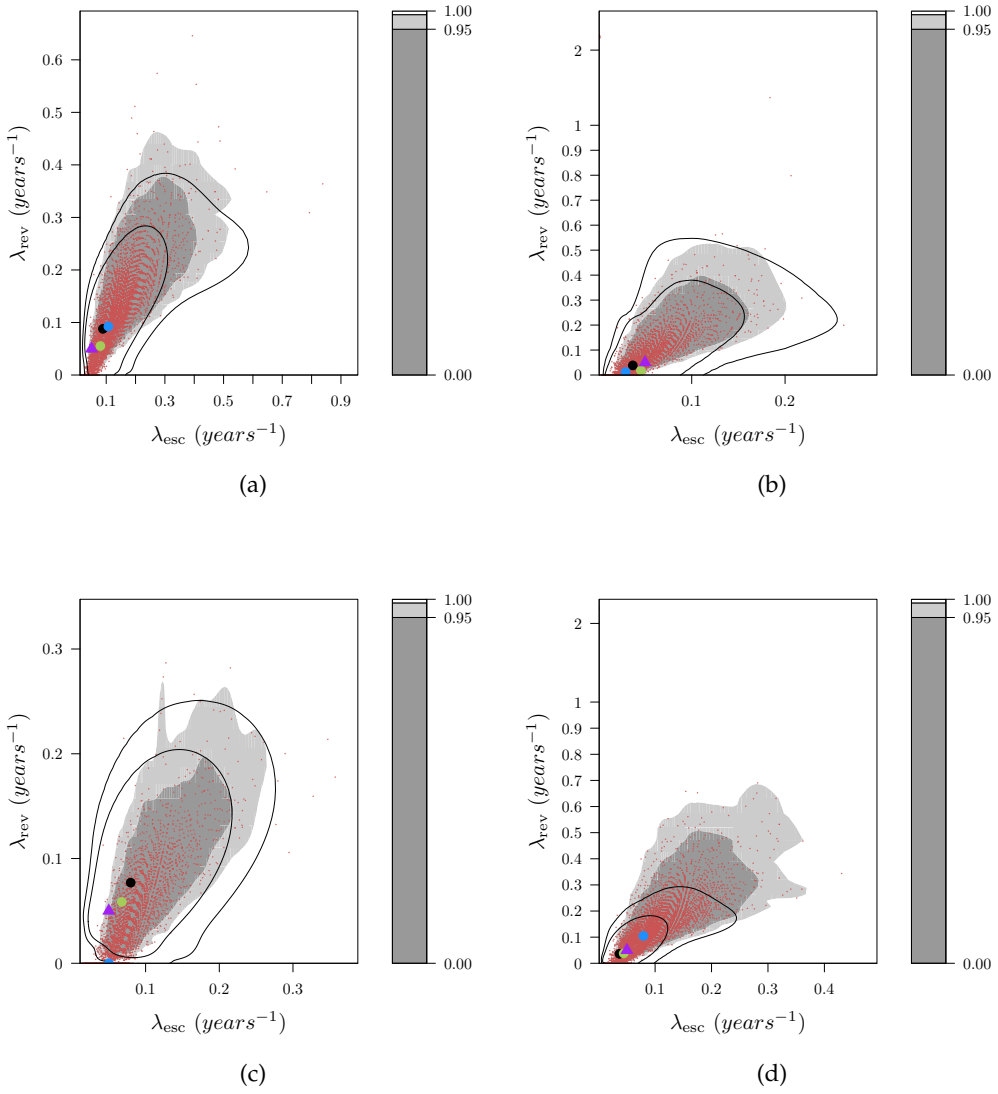


Figure 3.6: Four simulations with fixed escape and reversion parameters. True underlying rate $(\lambda_{esc}, \lambda_{rev}) = (0.05, 0.05)$ $years^{-1}$. 95% and 99% credible regions are shown in black. 10,000 ODE point estimates are shown in red, and contours coloured using a 2D kernel density estimate [297, 298] according to the key. The true rates are shown as a purple triangle, the MAP estimate is shown in black, and ODE point estimate in blue. The MAP based on the true underlying transmission tree is shown as a green outlined dot.

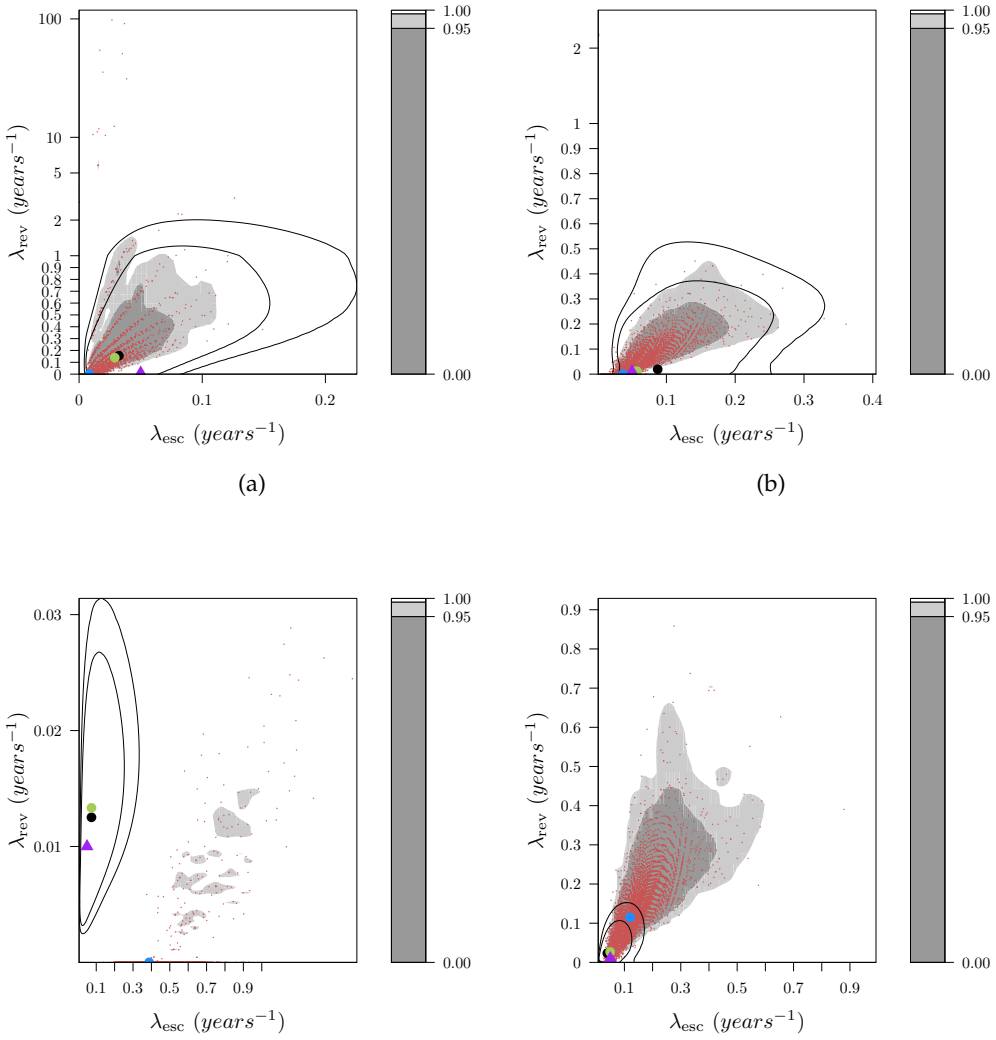


Figure 3.7: Four simulations with fixed escape and reversion parameters. True underlying rate $(\lambda_{esc}, \lambda_{rev}) = (0.05, 0.01)$ $years^{-1}$. 95% and 99% credible regions are shown in black. 10,000 ODE point estimates are shown in red, and contours coloured using a 2D kernel density estimate [297, 298] according to the key. The true rates are shown as a purple triangle, the MAP estimate is shown in black, and ODE point estimate in blue. The MAP based on the true underlying transmission tree is shown as a green outlined dot. An estimate of the density for Figure 3.7c could not be established due to the concentration of ODE point estimates along $\lambda_{rev} = 0$.

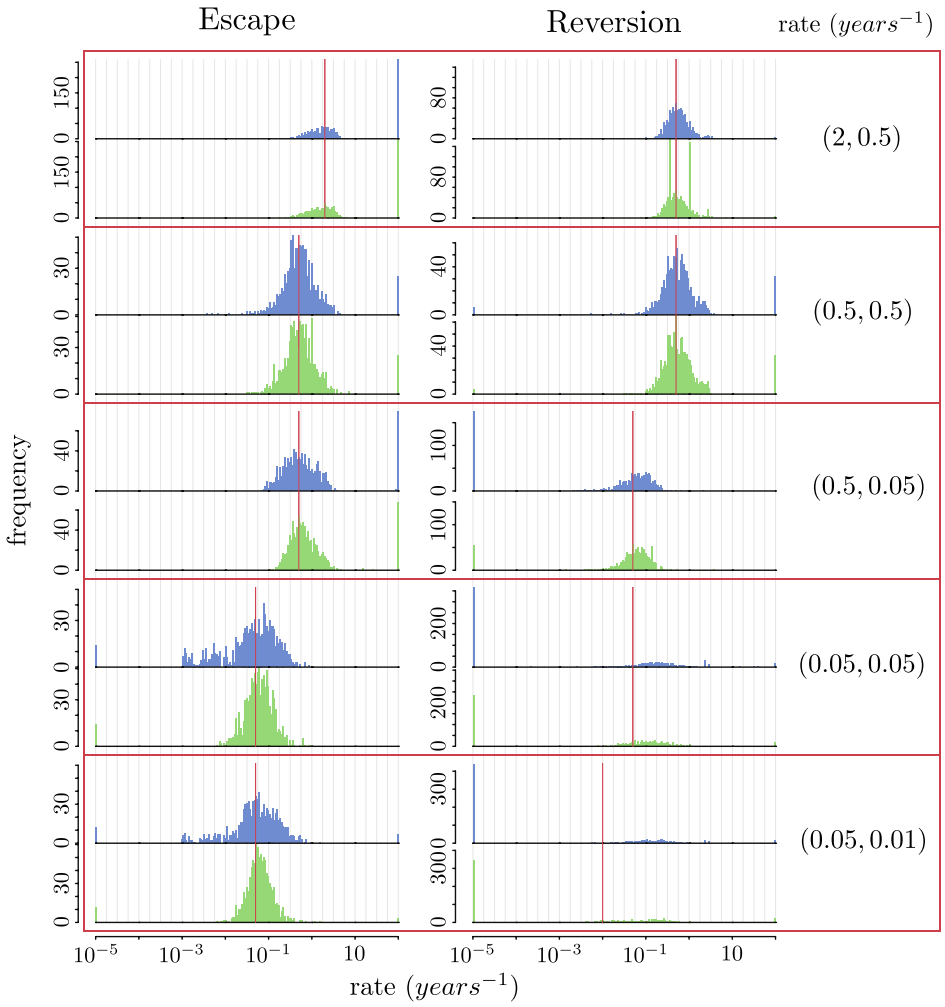


Figure 3.8: Histograms of MAPs using peeling algorithm and ODE point estimates for 60 sampled tips and 5 parameter sets, $(\lambda_{\text{esc}}, \lambda_{\text{rev}}) = \{(2, 0.5), (0.5, 0.5), (0.5, 0.05), (0.05, 0.05), (0.05, 0.01)\}$ respectively. True rates for each parameter decrease down the figure. Histograms of the ODE point estimates are in blue, histograms of the MAP under the peeling algorithm are in green. Rates $< 10^{-5}$ and $> 10^2$ are grouped in the histograms. In each plot, the truth is plotted in red.

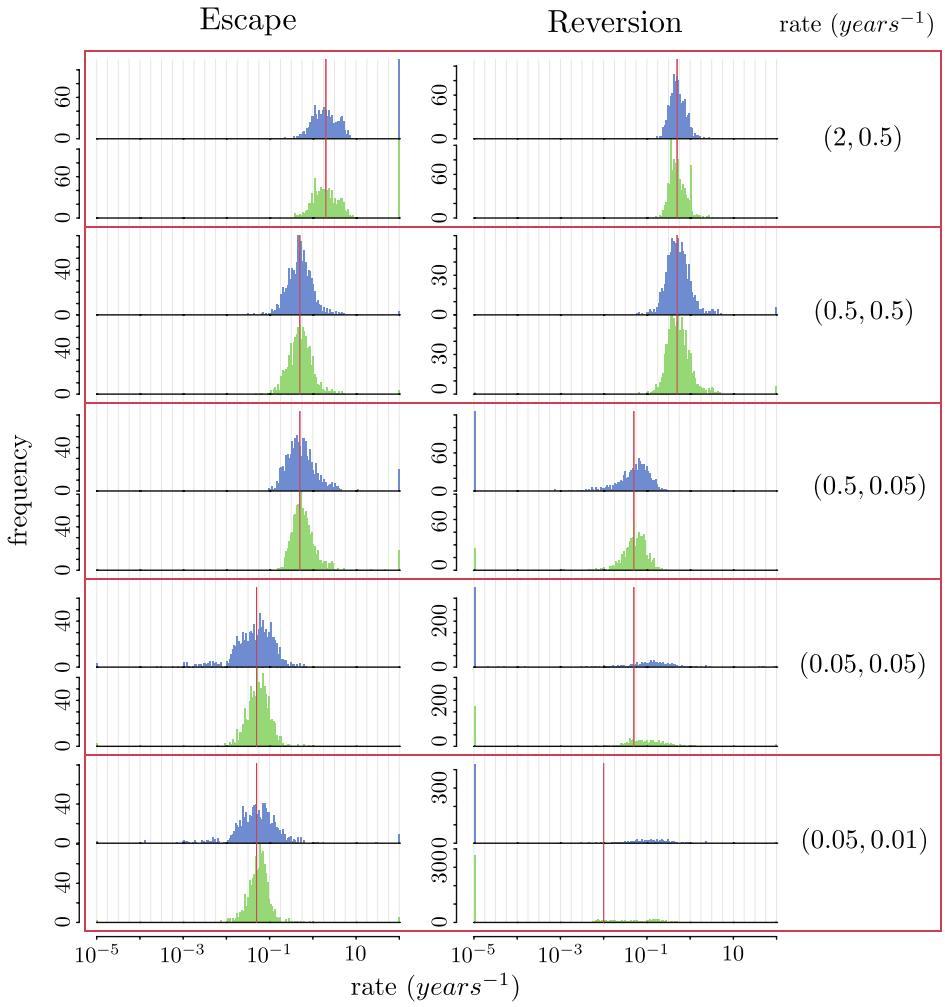


Figure 3.9: Histograms of MAPs using peeling algorithm and ODE point estimates for 100 sampled tips and 5 parameter sets, $(\lambda_{esc}, \lambda_{rev}) = \{(2, 0.5), (0.5, 0.5), (0.5, 0.05), (0.05, 0.05), (0.05, 0.01)\}$ respectively. True rates for each parameter decrease down the figure. Histograms of the ODE point estimates are in blue, histograms of the MAP under the peeling algorithm are in green. Rates $< 10^{-5}$ and $> 10^{2}$ are grouped in the histograms. In each plot, the truth is plotted in red.

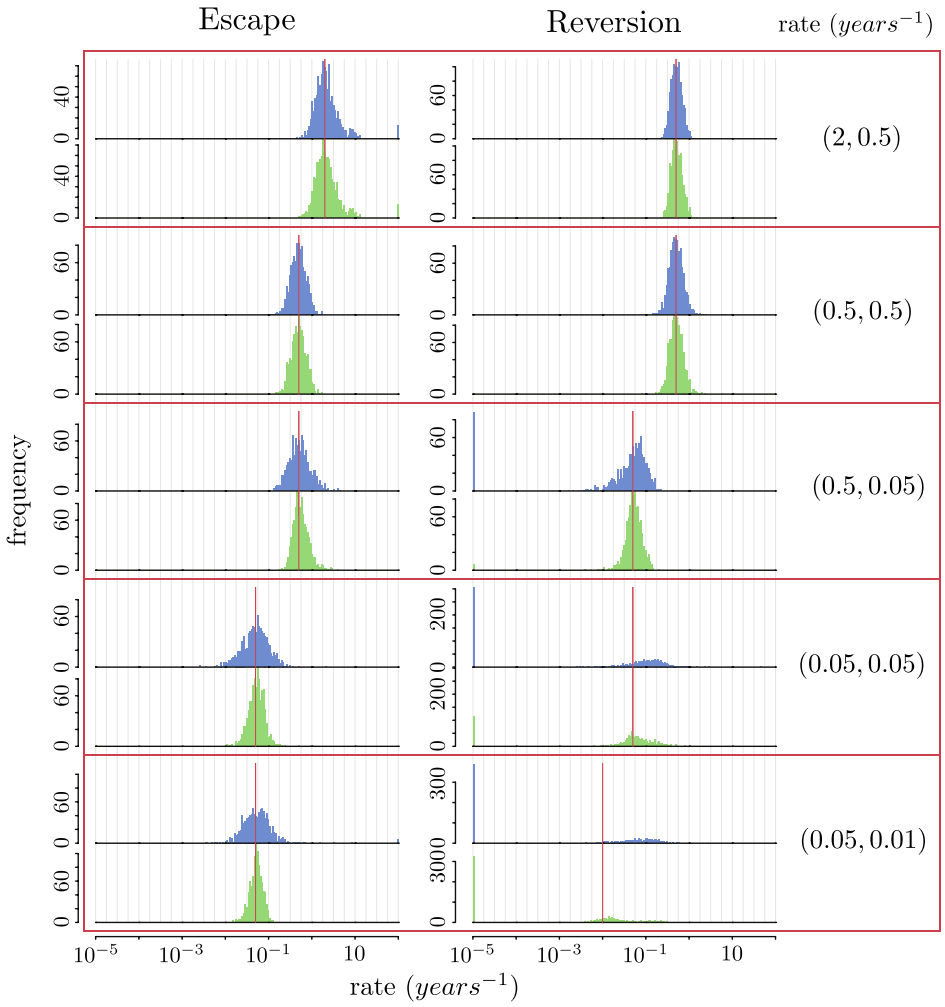


Figure 3.10: Histograms of MAPs using peeling algorithm and ODE point estimates for 200 sampled tips and 5 parameter sets, $(\lambda_{esc}, \lambda_{rev}) = \{(2, 0.5), (0.5, 0.5), (0.5, 0.05), (0.05, 0.05), (0.05, 0.01)\}$ respectively. True rates for each parameter decrease down the figure. Histograms of the ODE point estimates are in blue, histograms of the MAP under the peeling algorithm are in green. Rates $< 10^{-5}$ and $> 10^{2}$ are grouped in the histograms. In each plot, the truth is plotted in red.

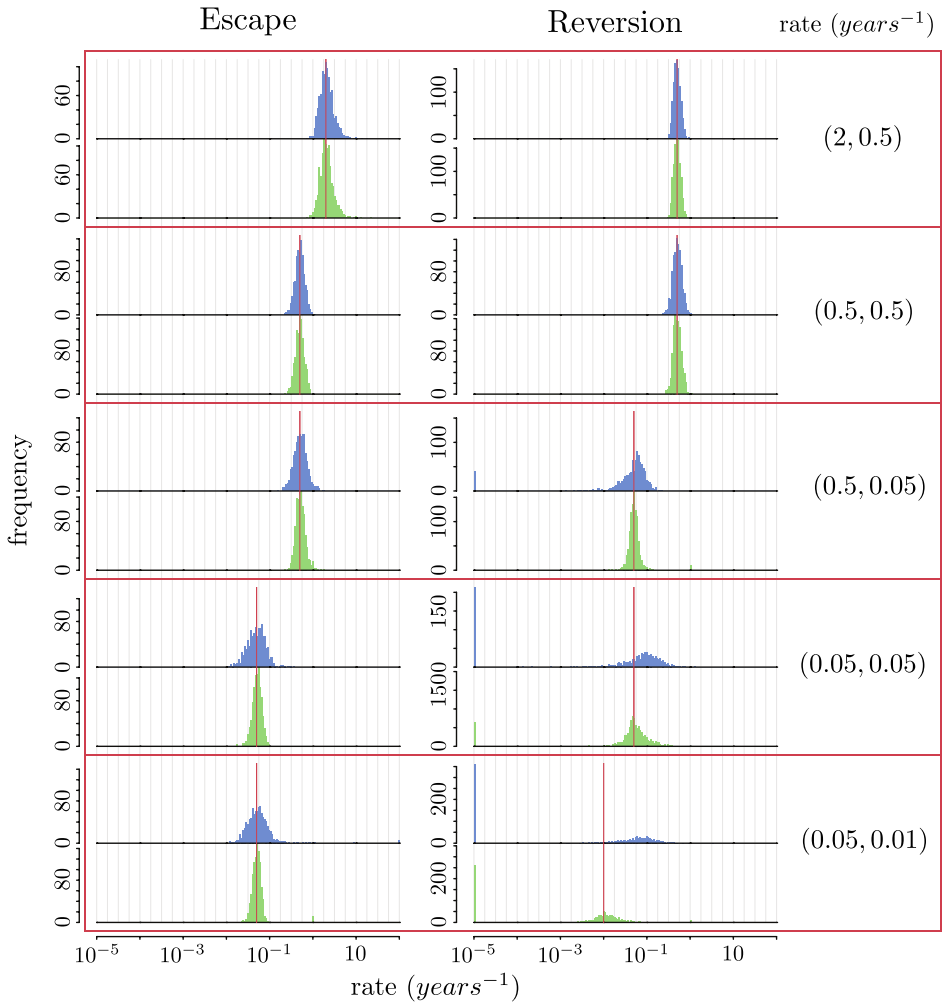


Figure 3.11: Histograms of MAPs using peeling algorithm and ODE point estimates for 500 sampled tips and 5 parameter sets, $(\lambda_{esc}, \lambda_{rev}) = \{(2, 0.5), (0.5, 0.5), (0.5, 0.05), (0.05, 0.05), (0.05, 0.01)\}$ respectively. True rates for each parameter decrease down the figure. Histograms of the ODE point estimates are in blue, histograms of the MAP under the peeling algorithm are in green. Rates $< 10^{-5}$ and $> 10^2$ are grouped in the histograms. In each plot, the truth is plotted in red.

estimates within a factor of 2 of the truth were $\{0.771, 0.825, 0.809, 0.468, 0.268\}$ for parameter sets $(\lambda_{esc}, \lambda_{rev}) = \{(2, 0.5), (0.5, 0.5), (0.5, 0.05), (0.05, 0.05), (0.05, 0.01)\}$. This makes sense: very high or low escape rates lead to a lack information to discern from either an infinite rate, or a rate of 0. Such estimates will result by chance under non-zero rates, the extreme example is data in which all individuals show escape. Our integrated approach substantially outperforms the ODE method when escape and reversion rates are low. As we increase the number of sampled tips in genealogies (Figures 3.8 - 3.11, sampling 60, 100, 200 and 500 tips respectively) we see increased accuracy under both methods, most notable at fast and slow rates. We find that the integrated method consistently outperforms the ODE method, with a tighter distribution of the MAP about the true value even when the number of sampled tips is low. This distinction displays the extra information present in the genealogy.

3.3.2.2 *Simulations with unknown underlying tree (simulation studies a) and c))*

We conduct four simulations over the five parameter sets of escape and reversion rates $(\lambda_{esc}, \lambda_{rev}) = \{(2, 0.5), (0.5, 0.5), (0.5, 0.05), (0.05, 0.05), (0.05, 0.01)\}$ shown in Figures 3.2a, 3.3 - 3.7, and a further 100 simulations for $(2, 0.5)$ $years^{-1}$ shown in Figure 3.2c. We find large variation in the size of credible regions across genealogies, particularly for low underlying rates. In the large simulation shown in Figure 3.2c, true parameters lie within the 50, 90 and 95% credible regions 79, 92 and 95 times out of 100.

3.3.2.3 *Comparison to ODE method*

We find that in general, the ODE method estimates the truth well, particularly when escape and reversion rates are high. This makes sense: along branches culminating in tips, fast escape and reversion leads to convergence to the equilibrium distribution, which is independent of the tree. For slower rates however, our integrated method performs better. This can be seen in single simulation runs in Figure 3.2a and 3.3 - 3.7, and in rate estimates assuming the true tree is

known (Figures 3.2b and 3.8 - 3.11). Here, the integrated approach performs far better, with a tighter distribution about the truth in all parameter sets. However, as would be expected, the signal begins to drop under both models as underlying rates are reduced further. For the underlying parameter sets $\{(2, 0.5), (0.5, 0.5), (0.5, 0.05), (0.05, 0.05), (0.05, 0.01)\}$, the proportion of ODE point estimates and MAP estimates within a factor of 2 were $\{0.768, 0.823, 0.570, 0.214, 0.055\}$ and $\{0.771, 0.825, 0.809, 0.468, 0.268\}$ respectively (the corresponding values within a factor of 10 were $\{0.982, 0.999, 0.913, 0.679, 0.356\}$ and $\{0.982, 0.999, 0.990, 0.850, 0.515\}$ respectively). Knowledge of the transmission tree increases accuracy of rate estimates across the parameter space of $(\lambda_{esc}, \lambda_{rev})$, particularly when escape and reversion rates are low.

3.3.2.4 *Robustness to tree topology and impact of mutation rate*

By shuffling tip labellings we investigate robustness of estimates to the tree topology. We find, in addition to an increase in variance of estimates, a systematic bias towards higher rate estimates, shown in Figure 3.12. This makes intuitive sense: Consider an example where the true genealogy contains a strong amount of clustering of escaped hosts due to slow reversion compared to escape. If sequences do not provide sufficient information to reconstruct the underlying genealogy, this is equivalent to being unable to label the tips with the sequences, so any shuffling of the tips would be equally likely. This shuffling removes the clustering present in the true tree, and subsequent estimates of λ_{esc} and λ_{rev} will be based on a pattern at the leaves which, on average, resembles that of higher reversion and escape rates (where we would expect far less clustering, due to fast within host escape or reversion). In summary, reduction in knowledge of tip labellings will act to randomise any clustering present in the true tree of escaped and wild-type strains at the epitope under consideration, leading to a forced increase in the lower bound of the number of escape and reversion events in the tree, increasing rate estimates.

To investigate the effect of mutation rate on estimates, we multiply and divide substitution rates by a factor of 5. This is displayed in Figure 3.13. As the mutation rate is increased, we see a reduction in variance and increase in accuracy of

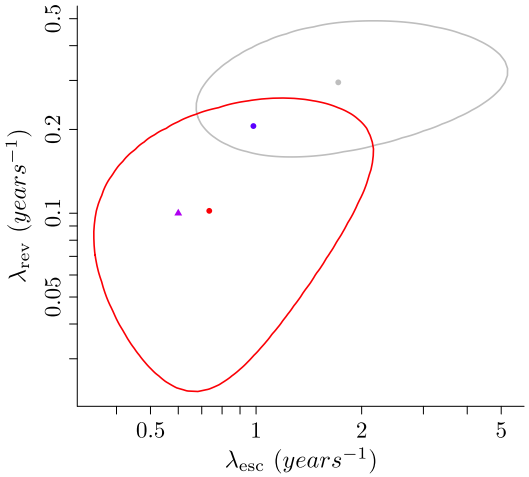


Figure 3.12: The true underlying sampled birth-death tree is shuffled 250 times, and our peeling algorithm applied, leading to the 90% credible region shown in grey, with MAP shown as a grey dot. The red credible region is the result of applying our peeling algorithm to the true tip labellings. The purple triangle is the truth (0.6,0.1) and the blue dot is the point estimate from the ODE method.

estimates as we would expect. It is important that this observation is considered in pathogens in which mutation rates are far lower and phylodynamic methods are beginning to be applied. Figure 3.12 and 3.13 demonstrate that a lack of knowledge of the underlying genealogy can seriously impact any parameter estimations leading to potentially spurious results.

3.3.3 *Analysis of real data*

The result of applying the integrated method to the available SSITT data is displayed in Figure 3.14 and summarised in Table 3.1. The cross-sectional proportion sampled, ρ, is set at 0.003, based on incidence data [300]. The first aspect to notice, which was also present in many of our simulations, is the similarity between the simple ODE method and the MAP from the integrated approach. This is not surprising as the purely dynamical model [85] can be written as a composite likelihood, which results from the assumption that all lineages are independent and of equal weight. The success of composite likelihoods is reflected in the similarity seen. However, looking more closely at the `TAF` and `RPM` epitopes (see Table 3.1), we see that the estimate of the underlying genealogy is playing a

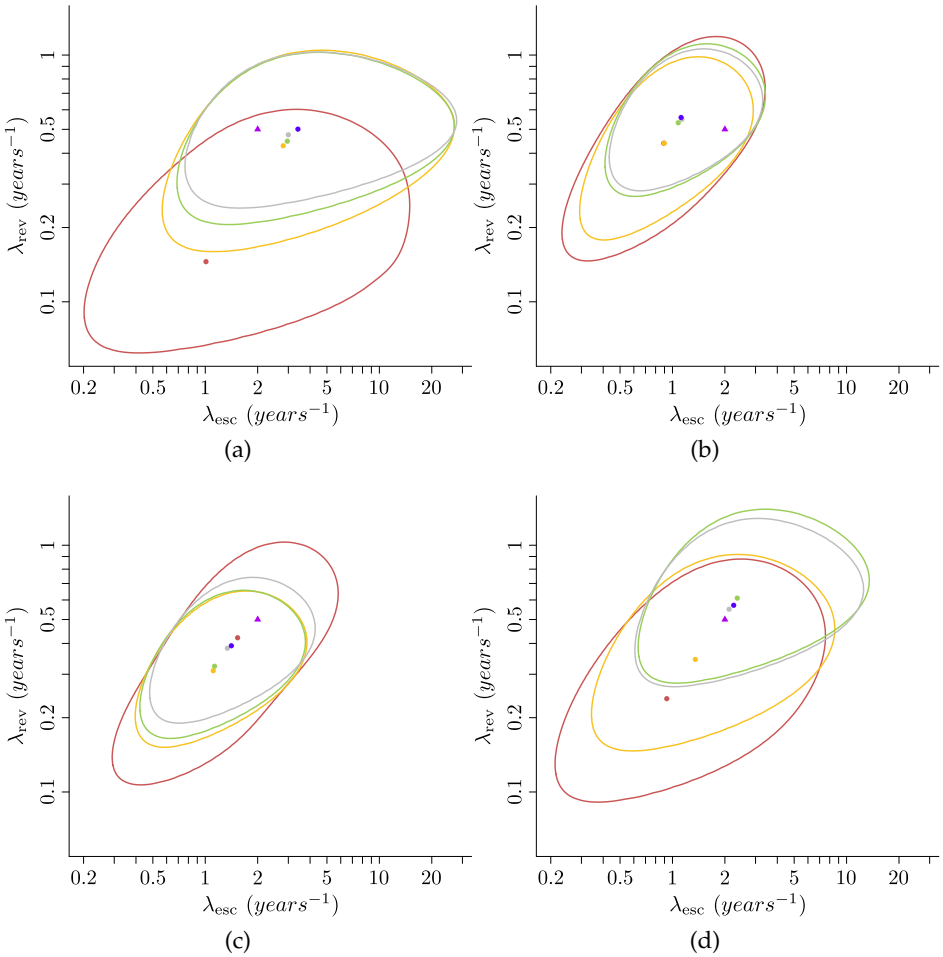


Figure 3.13: Four instances of simulations with reduced and accelerated substitution rates. In each, a 90% credible region is drawn for substitution rates at 0.2, 1 and 5 times the average of BEAST output for the SSITT *gag* sequence data under the $GTR + \Gamma + I$ model. Regions are coloured red, orange and green respectively. The 90% credible region assuming the true tree was known is shown in grey. The truth, $(\lambda_{esc}, \lambda_{rev}) = (2, 0.5)$ $years^{-1}$ is shown as a purple triangle and MAPs for the corresponding credible regions are shown in the same colour. The ODE point estimate is displayed in blue in each simulation. Note the strong agreement between the ODE point estimate and MAP under these high $(\lambda_{esc}, \lambda_{rev})$ rates.

strong role. The maximum clade credibility trees for the BEAST runs of TAF and RPM using TreeAnnotator [68] are shown in Figure 3.14a. We see that in the case of TAF, the escape rate is ~ 6 times lower in the MAP estimate than the ODE estimate. This is caused by clustering seen in the tree. Of the 24 individuals who have a consensus escape, 13 occur in clusters of 2 or more. Moreover, singleton escaped lineages coalesce deep into the tree. These combine to reveal the existence of a lower escape rate than that seen in the ODE approximation, combined with transmission of escapes indicative of a very low reversion rate. Contrast this to the rate approximation in RPM. Recent coalescent events in which one lineage has the escape variant whilst the other doesn't leads to faster escape and reversion estimates under the integrated approach. Secondly, we notice the large amount of uncertainty arising through the evolutionary process in our estimates. This can be explained through two main factors: the lack of seen transmissions close to the present, and uncertainty in the states deep into the ancestry. Improvement in each will result from increased sampling. An increase at the present will increase the number of recent coalescent events, and increased historical sampling will add confidence to inferred states deep in the genealogy. Throughout, we have considered the coalescent in an exponentially growing population as our prior on the genealogy. This is because a birth-death prior with historical sampling and sampling at the present is currently untested in BEAST. Use of a birth-death prior would increase uncertainty in the time to the most recent common ancestor due to its stochastic nature. Indeed, the deterministic nature of the coalescent underestimates uncertainty in T_{MRCA} in an epidemic setting. However, using a sampled birth-death prior allows us to sample large portions of the infected population completely correctly, increasing our power to infer dynamical parameters. Unfortunately, the long terminal branches indicative of exponential growth mean that it will always be difficult to estimate parameters of such dynamical models. Major advantages of our model-based framework are that we obtain meaningful credible intervals for our parameter estimates, and gain a statistical framework in which hypotheses about these parameters can be tested. For example, consider the null hypothesis that escape and reversion rates are common across the four epitopes, with the alternative that they each

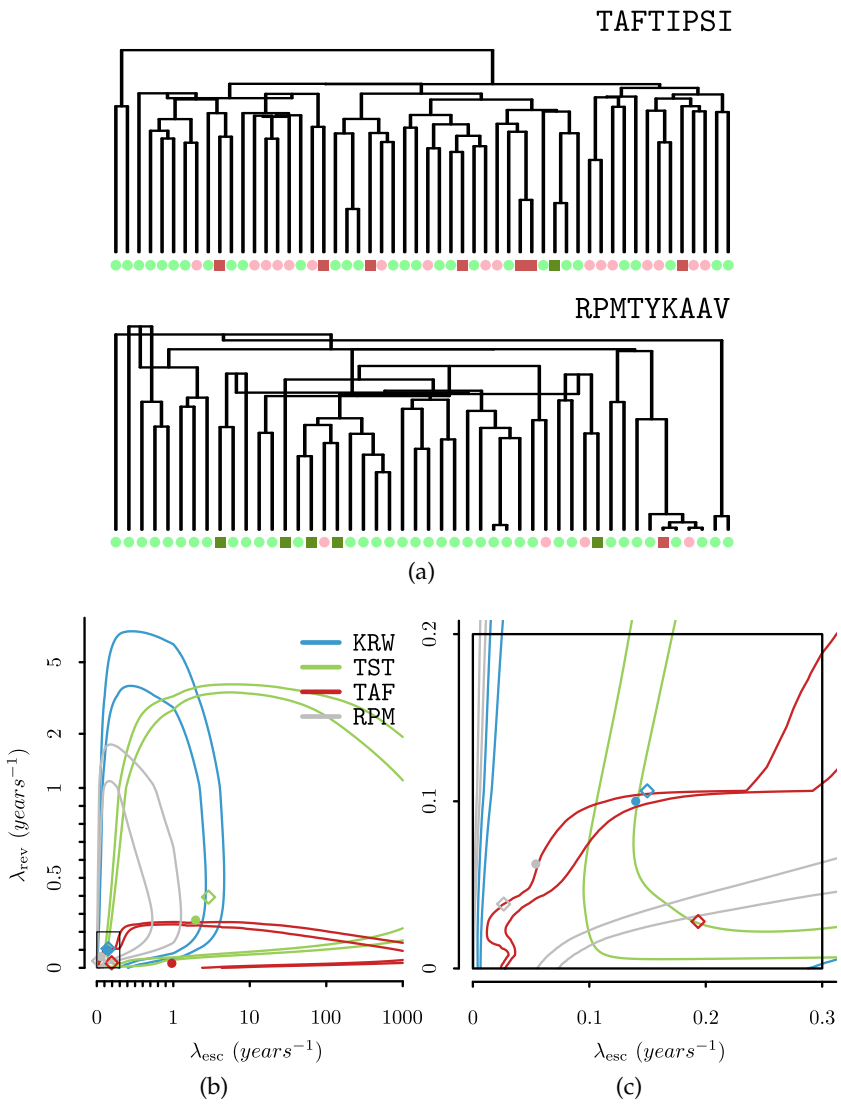


Figure 3.14: Figure 3.14a displays the maximum clade credibility tree for cross-sectional data for epitopes TAFTIPSI and RPMTYKAAV, with tips coloured as in Figure 3.1. Figures 3.14b and 3.14c show 95% and 99% credible regions under our integrated method. Figure 3.14c is a zoom in of the rectangle in Figure 3.14b. Axes are linear on $[0, 1]$ and on a log scale for values > 1. Colourings are shown in upper right figure legend of 3.14b. Coloured diamonds are MAP estimates of $(\lambda_{esc}, \lambda_{rev})$, ODE method estimates are filled circles.

have distinct rates. Using a likelihood-ratio test, we reject the null hypothesis ($p = 0.00014$). Testing the difference in escape and reversion rates between `RPM` and `KRW`, we find the data cannot reject the null hypothesis that there is a common escape and reversion rate across these two epitopes ($p = 0.412$). Validation of our use of the likelihood-ratio test is given in Section 4.6.

Epitope	Gene	HLA	Escape	q	Escape rate (λ_{esc})		Reversion rate (λ_{rev})	
					MAP	ODE	MAP	ODE
TSTLQEQIGW	*gag* (108-117)	B*57, B*58	T3N	0.096	2.89	2.83 (0.138, 1000)	0.393	0.383 (0.0220, 3.40)
KRWIILGLNK	*gag* (131-140)	B*27	R2K, R2G, R2Q	0.073	0.150	0.200 (0.00662, 2.66)	0.106	0.154 (0, 3.69)
TAFTIPSI	*pol* (128-135)	B*51	I8T	0.126	0.194	1.18 (0.0274, 1000)	0.0281	0.0562 (0, 0.242)
RPMTYKAAV	*nef* (77-85)	B*07	T4S, Y5F	0.166	0.0265	0.0144 $(6.06 \times 10^{-4}, 0.726)$	0.0385	0.0226 (0, 1.09)

Table 3.1: Summary table of escape mutations analysed. Location is defined in parentheses by the HXB2 B-clade reference sequence. q is the HLA proportion in Caucasians [178]. Rates are in $years^{-1}$. Minimal and maximal rates of escape and reversion rate within the 95% credible regions for MAP estimates are in parentheses.

3.4 DISCUSSION

We have combined a dynamical modelling approach with cross-sectional sequence data to infer escape and reversion rates at CTL epitopes within hosts whilst taking the underlying genealogy into account. We compared our integrated approach to an ODE based method through simulations and parameter estimates using cross-sectional data from the SSITT study cohort [85]. Our model is set in a statistical framework and makes use of the information present in sequence data for parameter estimates. We gain meaningful credible regions which consider uncertainty in the true underlying genealogy. A strong benefit of our integrated method is that it provides a probabilistic framework in which hypotheses can be tested. Our most striking conclusion is the large amount of uncertainty present in rates estimates utilising cross-sectional sequence data. Great care must therefore be taken before strong conclusions are made on the basis of such estimations.

Under the ODE approach outlined in [85], sequence information outside the epitope under consideration is ignored. Cross-sectional data are considered to have arisen independently, up to initial conditions. We show that this is equivalent to the assumption of a completely star-like phylogeny in Section 4.5. Given this major assumption of the ODE method, we expected our integrated method to perform more accurately in simulations. We find that this is indeed the case, the presented model consistently outperforms the simpler ODE based approach, particularly when strong clustering of escape mutations is more likely to be present in the true tree. Despite this improvement, we find that the ODE method recaptures escape and reversion rates relatively well in simulations, in spite of its simplifying assumptions.

In order to incorporate the underlying dependency structure present in the genealogy of our samples, and assuming exponential growth, there are two major processes to decide between: the coalescent in an exponentially growing population [148, 264, 108] (which we chose) and the sampled birth-death process [310, 267]. In the case of the birth-death process, a large number of likelihoods describing the process have been constructed [279, 196, 310, 95], differing only

in what the authors decide to condition on. We summarise the links between each likelihood in Section 4.7. Our choice between exponential coalescent and sampled birth-death process determines the prior on the genealogy. One major distinction is that the sampled birth-death process considers a subpopulation within a larger stochastically growing population, and the exponential coalescent assumes that the subpopulation is a tiny subset within a very large deterministically growing population. There are benefits and drawbacks to each. Under a birth-death prior, the sum of seen transmissions informed by the prior and unseen transmissions given by the matrix $\mathbf{Q}$ is the overall transmission rate over time - a desirable property which the coalescent lacks [267]. The birth-death process incorporates early stochasticity, which more accurately represents the truth in an epidemic setting. Additionally, no coalescent approximation is required, making the prior suitable for datasets in which the number of samples is comparable to the total population size. This is becoming increasingly relevant as such datasets are becoming more commonplace [133]. Under the exponential coalescent, inclusion of time-stamped datasets is straightforward. In contrast, under the birth-death process, assumptions about the sampling rate of these historical events must be made [268], which are often invalid for many datasets. However, both are only priors on tree shape and if the data is strong, the distribution from which the trees are sampled will be near identical. We show an interesting link between the two processes in Section 4.8.

In our estimates we are fundamentally restricted by coalescence events in the genealogy. Long terminal branches are indicative of exponential growth, yet the greatest power to inform our parameter estimates comes from coalescence events occurring in the recent history of the virus. Thus, obtaining extra information from the genealogy is intrinsically difficult. Greater sampling at the present will increase the occurrence of recent coalescence events, and provides greater power to distinguish high and low rates from infinity and zero respectively. However, the inclusion of more and more sequence data calls the coalescent approximation into question as the sampled infected population becomes a larger portion of the infected population. Dense sampling can also lead to a breakdown of the assumed connection between the genealogy and the transmission tree due

to lineage sorting [231]. Including time-stamped data with with tip information would allow estimation of ancestral states with greater confidence, and thus increase our ability to infer escape and reversion rates. If longitudinal and cross-sectional sequence data could be combined, this would dramatically increase power to estimate rates. Unfortunately, current methods cannot support this extension due to the use of the genealogy as a proxy for the transmission tree. It is clear that greatest power to estimate these parameters lies in longitudinal sampling within cohorts of hosts, but here we create another collection of issues: recombination plays a large role within hosts [249], and we would require longitudinal sequences across a large number of individuals in order to make any meaningful statements about rate estimates across the infected population.

Models which attempt to integrate the underlying genealogy are currently being developed to incorporate epidemiological dynamics outside the exponential growth phase [292] assumed under this model. Another potential improvement would be to co-estimate escape and reversion rates within the MCMC scheme.

Our model also makes many assumptions about the underlying biological processes. For example, overlapping epitopes which are prevalent across the HIV-1 genome [127] mean that mutations conferred by one HLA type could be incorrectly inferred to be the result of selection due to another HLA type. Violation of this assumption of selectively neutral sites outside the epitope of interest will affect branch lengths more than topology, so should not greatly alter our estimates and their ordering within genes. It is through changes in topology in the recent history of the data that estimates will be most drastically altered.

We assume that transmission occurs at rate λ, independent of consensus viral sequence. HLA types are considered to two digits and escape mutations within a given epitope are grouped together due to the relatively small dataset. All individuals with the restricting HLA type are assumed to be capable of making a response which drives selection at the epitope under investigation. With larger datasets, such complications could be included in a similar model in the same framework with more parameters.

Throughout, we assume homogeneous mixing. This assumption may affect our estimates and will have the greatest impact if the HLA distributions are strongly

segregated across the host population. Simultaneous analysis of individuals from different continents, for example, would be highly inappropriate. HLA distributions across Europe are relatively homogeneous, so we feel that this assumption is reasonable for the SSITT dataset to which it is here applied.

We have neglected recombination in our analysis. This is clearly a factor in HIV-1 evolution, however this certainly does not discredit the use of trees in our investigations. We are assuming that the HIV-1 epidemic is in the exponential growth phase. This leads to star-like phylogenies in which long external branches coalesce high up the tree at short internal branches. Just by probability then, the chance that recombination strongly affects the topology of the tree is lower than would be the case in a constant population tree. This is evidenced by only small differences in the genealogies estimated when a study assumed unlinked evolutionary trees across genes compared to a single tree for all genes in HIV-1 [296]. Building an accurate tree across hosts is made difficult by the occurrence of recombinants within hosts due to infection and recombination of multiple variants either through multiple invasions by different variants - super-infection, or through infection by a combination of variants during a single infection event. Circulating recombinant forms certainly exist, however, rates of super-infection are reported to be low (around 5% within 6-12 months of initial infection for subtype B super-infection [265]), and the majority of instances of their occurrence are restricted to the small proportion of the population that have the greatest risk of exposure to the pathogen. It has also been shown that there is a strong viral population bottleneck upon transmission to a new host [58, 179]. Given a collection of sequences which have undergone recombination, it will appear that variation in rates across sites is higher than it actually is and the topology of phylogenies will be biased towards the star-like structure indicative of exponential growth [160]. It is therefore important that parameters estimated are compared where possible to epidemiological data to check our assumptions. Recombination is also known to contribute to rejection of a strict molecular clock [160].

Finally, we assume that sampled genealogies provide a reasonable approximation of the underlying transmission process. A variety of other hypotheses could be tested by extending the model in the presence of extra data. For example,

given disease outcome information, HLA typing and sequence information, is it possible to discern that escape mutations are associated with faster disease progression?

In summary, we have constructed a model which integrates sequence data and considers the evolutionary history, transmission, and set of dynamical processes together. The model was created using existing techniques, and we use it to address pressing practical questions. Whilst we created the presented model to answer a specific question, we believe that similar integrated models making use of epidemiological and viral sequence data may be more broadly applied. Such models can be used to estimate various parameters of interest more accurately with the help of sequence data. Indeed, we use the model introduced in Chapter 6 to detect drug associated selection.

4

INTEGRATING GENEALOGICAL AND DYNAMICAL MODELLING TO INFER ESCAPE AND REVERSION RATES IN HIV-1 EPITOPES: SUPPLEMENTARY MATERIAL

The majority of the material in this Chapter is contained in the supplementary materials of 'Integrating genealogical and dynamical modelling to infer escape and reversion rates in HIV epitopes' Palmer *et el.* published in Proceedings of the Royal Society B: Biological Sciences, Vol. 280, No. 1762.

4.1 INTRODUCTION

In this Chapter we provide further discussion and elaboration of the topics and methods of Chapter 3. The Chapter is split into the following sections:

- **Section 4.2: Justification of combining the peeling algorithm and BEAST**. We provide a short derivation of our sampling scheme, detailing our assumptions.
- **Section 4.3: Rewriting the sampled reconstructed process as a generalised pure birth process**. We carefully derive the sampled reconstructed birth-death process as a pure birth process with a variable birth rate following the arguments of Nee *et al.* [196] (in which a fully sampled birth-death process is rewritten as a pure birth process).
- **Section 4.4: Detailed Methods**. Full details of the model outlined in Chapter 3, describing the data, our definition of escape, modelling assumptions and methods for alignment, quality control, tree sampling, and a description of our Peeling algorithm. We also demonstrate the linear association between effective infected population size and infected population size by

following the arguments of Frost and Volz [84], thus validating our sampling of the transmission rate λ from the MCMC on trees. Finally, we describe the methods used to integrate over trees and generate credible regions for our estimates.

- **Section 4.5: Equivalence of ODE model and composite likelihood**. We determine the implicit assumption about tree topology when using the ODE based method of Fryer *et al.* [85].
- **Section 4.6: Justification of validity of hypothesis testing**. By manipulating the likelihood we demonstrate the validity of hypothesis testing within our framework.
- **Section 4.7. Connections between formulations of reconstructed tree likelihoods**. We discuss the connections between a collection of derived likelihoods in the literature: Thompson [279], Nee *et al.* [196], Yang and Rannala [310] and Gernhard [95].
- **Section 4.8. Derivation of surviving lineages through time within a deterministic model of transmission**. We demonstrate connections between a deterministic approximation of the sampled birth-death process and that derived using the methods of Kendall [142].

4.2 JUSTIFICATION OF COMBINING THE PEELING ALGORITHM AND BEAST

To create the sequences observed at the tips of the true genealogy together with the HLA and escape status of the host, three processes occur concurrently. We have a collection of epidemiological parameters Ψ which govern the structure of true underlying genealogy G, mutation parameters Θ which determine the distribution of sequences at the tips (excluding the epitope of interest), and finally some escape parameters $\boldsymbol{\lambda}$ governing the distribution of presence of escape and HLA type of the host at the tips. Suppose we have the following collection of priors and observed data:

- Priors f, g, h on Ψ, Θ and $\boldsymbol{\lambda}$ respectively.

- Observed sequences at the tips X (excluding epitope) and observed presence or lack of escape at the tips and HLA information, E.

We wish to determine

$$P(\boldsymbol{\lambda}|X, E, f, g, h). \tag{25}$$

BEAST provides samples from $P(G, \Theta, \Psi|X, f, g)$. To sample from $P(\boldsymbol{\lambda}|X, E, f, g, h)$, we marginalise over G, Θ and Ψ in $P(\boldsymbol{\lambda}, G, \Theta, \Psi|X, E, f, g, h)$ by considering:

$$P(\boldsymbol{\lambda}, G, \Theta, \Psi|X, E, f, g, h) = P(G, \Theta, \Psi|X, E, f, g, h)\ P(\boldsymbol{\lambda}|X, E, f, g, h, G, \Theta, \Psi) \tag{26}$$

$$= \frac{P(G,\Theta,\Psi|X,E,f,g,h)}{P(G,\Theta,\Psi|X,f,g,h)}\ P(G, \Theta, \Psi|X, f, g, h)\ P(\boldsymbol{\lambda}|E, h, G, \Psi). \tag{27}$$

If we assume that

$$\frac{P(G, \Theta, \Psi|X, E, f, g, h)}{P(G, \Theta, \Psi|X, f, g, h)} \approx 1 \tag{28}$$

then,

$$P(\boldsymbol{\lambda}, G, \Theta, \Psi|X, E, f, g, h) \approx P(G, \Theta, \Psi|X, f, g, h)\ P(\boldsymbol{\lambda}|E, h, G, \Psi). \tag{29}$$

Note that Equation (28) is equivalent to assuming that escape within the epitope of interest does not have any impact on our inference of the underlying tree, the neutral mutation parameters or parameters governing the dynamics of the number of infecteds in the epidemic. By sampling from $P(G, \Theta, \Psi|X, f, g)$ (the output from BEAST), and further sampling $P(\boldsymbol{\lambda}|E, h, G, \Psi)$ from these BEAST samples, we are averaging $P(\boldsymbol{\lambda}, G, \Theta, \Psi|X, E, f, g, h)$ over G, Θ, Ψ and thus sampling from the desired distribution $P(\boldsymbol{\lambda}|X, E, f, g, h)$.

4.3 REWRITING THE SAMPLED RECONSTRUCTED PROCESS AS A GENERALISED PURE BIRTH PROCESS

In Nee et al. [196], it is shown that a birth-death process with complete sampling at the present can be rewritten as a generalised birth process with variable rate $\lambda P(t,T)$, where $P(t,T)$ defines the probability that a lineage existing at t survives to time T. In a similar manner, we show that the inclusion of sampling a proportion ρ of the extant lineages at the present can also be rewritten as a pure birth process. This, though not explicitly stated, is what was done between Equations (2) and (3) in Yang and Rannala [310].

Using a Dirac delta function at the present in order to include a mass extinction of $(1-\rho)$ of the population which is equivalent to sampling a proportion ρ at the present, Nee et al. [196] consider a generalised birth-death process with constant birth rate λ and death rate $\mu(t,T) = \mu - \delta(t,T)\log(\rho)$. Using Kendall's derivations [142], under this new process

$$P_s(t,T) = \frac{\rho(\lambda-\mu)}{\rho\lambda + (\lambda(1-\rho)-\mu)\exp(-(\lambda-\mu)(T-t))}, \tag{30}$$

where the subscripted s denotes sampling, and thus $P_s(t,T)$ represents the probability that a lineage existing at time t survives to time T and is sampled. Consider a pure birth process with speciation rate $\lambda P_s(t,T)$. Using Kendall's notation [142], the probability of i lineages existing at time t is

$$\hat{p}_i(t) = (1-\eta_{t,T})\eta_{t,T}^{i-1} \quad i > 0, \tag{31}$$

$$\text{where} \quad \eta_{t,T} = \frac{\lambda\rho(1-\exp(-(\lambda-\mu)t))}{\lambda\rho - (\mu-\lambda(1-\rho))\exp(-(\lambda-\mu)T)}. \tag{32}$$

We now show that this is equivalent to considering a full birth-death process, sampling, and reconstructing the transmission history.

Applying the same steps as in Nee et al. [196], the probability that a birth-death

process with birth rate λ and death rate $\mu - \delta(t,T)\log(\rho)$ surviving to t with i descendants survives to time T is

$$z_i := \frac{(1-u_t)u_t^{i-1}(1-(1-P_s(t,T))^i)}{\sum_{j=1}^{\infty}(1-u_t)u_t^{j-1}(1-(1-P_s(t,T))^j)} \tag{33}$$

where

$$u_t = \frac{\lambda(1-\exp(-(\lambda-\mu)t))}{\lambda-\mu\exp(-(\lambda-\mu)t)}. \tag{34}$$

We now create the reconstructed process with sampling. Of k lineages existing in the past, at time t, we require those ancestries which result in i lineages at the present, time T.

$$P(l=i,t,T) = \sum_{k=1}^{\infty} \frac{z_k\binom{k}{i}\left(P_s(t,T)\right)^i\left(1-P_s(t,T)\right)^{k-i}}{1-(1-P_s(t,T))^k}. \tag{35}$$

This simplifies to

$$\left(1-u_t\frac{P_s(0,T)}{P(0,t)}\right)\left(u_t\frac{P_s(0,T)}{P(0,t)}\right)^{i-1}, \tag{36}$$

Finally, $u_t P_s(0,T)/P(0,t) = \eta_{t,T}$ and we are done.

To derive the simplification from Equation (35) to Equation (36), consider the probability generating function for (35):

$$g(s) = \sum_{i=1}^{\infty}\sum_{k=1}^{\infty} \frac{z_k\binom{k}{i}\left(P_s(t,T)\right)^i\left(1-P_s(t,T)\right)^{k-i}s^i}{1-\left(1-P_s(t,T)\right)^k} \tag{37}$$

$$= \sum_{i=1}^{\infty}\left(\frac{sP_s(t,T)}{1-P_s(t,T)}\right)^i \sum_{k=1}^{\infty}\frac{z_k\binom{k}{i}\left(1-P_s(t,T)\right)^k}{1-(1-P_s(t,T))^k}. \tag{38}$$

z_k simplifies (using the sum of a geometric series) to

$$\frac{u_t^{k-1}\left(1-(1-P_s(t,T))^k\right)}{\frac{1}{1-u_t}-\frac{1-P_s(t,T)}{1-u_t(1-P_s(t,T))}} = \frac{(1-u_t)(1-u_t(1-P_s(t,T))u_t^{k-1}\left(1-(1-P_s(t,T))^k\right)}{P_s(t,T)} \tag{39}$$

Thus,

$$g(s) = \frac{(1-u_t)\left(1-u_t\left(1-P_s(t,T)\right)\right)}{P_s(t,T)u_t}\sum_{i=1}^{\infty}\left(\frac{sP_s(t,T)}{1-P_s(t,T)}\right)\sum_{k=1}^{\infty}\left(u_t(1-P_s(t,T))\right)^k\binom{k}{i} \tag{40}$$

Using the identity $\sum_{k=l}^{\infty} y^k\binom{k}{l} = \frac{y^l}{(1-y)^{l+1}}$ we have

$$g(s) = \frac{1-u_t}{P_s(t,T)u_t}\sum_{i=1}^{\infty}\left(\frac{sP_s(t,T)}{1-P_s(t,T)}\right)^i\left(\frac{u_t\left(1-P_s(t,T)\right)}{1-u_t\left(1-P_s(t,T)\right)}\right)^i \tag{41}$$

$$= \frac{(1-u_t)\,s}{1-u_t+P_s(t,T)u_t\,(1-s)} \tag{42}$$

We now check (35) = (39) using $\frac{g^i(0)}{i!} = P(l=i,t,T)$. First, for $l=1$ we require

$$g'(0) = \left(1-u_t\frac{P_s(0,T)}{P(0,t)}\right). \tag{43}$$

Differentiating with respect to s, and setting $s=0$,

$$g'(0) = \frac{(1-u_t)}{(1-u_t(1-P(t,T)))} \tag{44}$$

Thus, it remains to prove

$$\frac{P_s(t,T)}{1-u_t(1-P_s(t,T))} = \frac{P_s(0,T)}{P(0,t)}. \tag{45}$$

Expanding and simplifying the denominator of the left hand side of (45),

$$1-u_t\left(1-P_s(t,T)\right) = \frac{(\lambda-\mu)\left(\lambda\rho+\left(\lambda\left(1-\rho\right)-\mu\right)\right)e^{-(\lambda-\mu)T}}{\left(\lambda-\mu e^{-(\lambda-\mu)t}\right)\left(\lambda\rho+\left(\lambda\left(1-\rho\right)-\mu\right)e^{-(\lambda-\mu)(T-t)}\right)} \tag{46}$$

$$\Rightarrow \frac{P_s(t,T)}{1-u_t\left(1-P_s(t,T)\right)} = \frac{\lambda-\mu\exp\left(-\left(\lambda-\mu\right)t\right)}{\lambda\rho+\left(\lambda\left(1-\rho\right)-\mu\right)\exp\left(-\left(\lambda-\mu\right)T\right)} \tag{47}$$

$$= \frac{P_s(0,T)}{P(0,t)} \tag{48}$$

Finally, we demonstrate that the probability generating function

$$g(s) = \frac{(1-u_t)s}{1-u_t+P_s(t,T)u_t(1-s)} \tag{49}$$

is the probability generating function of a geometric distribution with parameter

$$\theta = \frac{u_t P_s(t,T)}{1-u_t(1-P_s(t,T))}. \tag{50}$$

$$\begin{aligned}
\tilde{g}(s) &= \left(1 - \frac{u_t P_s(t,T)}{1-u_t\left(1-P_s(t,T)\right)}\right) \sum_{i=1}^{\infty} \left(\frac{u_t P_s(t,T)}{1-u_t(1-P_s(t,T))}\right)^{i-1} s^i && (51)\\
&= \left(1 - \frac{u_t P_s(t,T)}{1-u_t\left(1-P_s(t,T)\right)}\right) \frac{s}{1-\frac{u_t P(t,T) s}{1-u_t(1-P(t,T))}} && (52)\\
&= \frac{(1-u_t)\, s}{1-u_t+P_s(t,T)u_t\,(1-s)} && (53)\\
&= g(s) && (54)
\end{aligned}$$

As required. Using the birth-death process with Dirac delta function described above, it can be derived that the probability a lineage at time t leaves exactly one descendant is

$$\begin{aligned}
p_1(t) :=& \frac{P_s(t,T)\,(\lambda-\mu)}{\lambda(1-\rho)-\mu+\lambda\rho\exp((\lambda-\mu)(T-t))} && (55)\\
=& \frac{1}{\rho}\exp(-(\lambda-\mu)(T-t))(P_s(t,T))^2, && (56)
\end{aligned}$$

as stated in Yang and Rannala [310] (note, this is $P_s(t,T)\,(1-\eta_{T-t,T-t})$). Using exactly the same arguments as above, the joint distribution of node times $\mathbf{s}$ (measured backwards from the present), conditional on survival, n tips, a root time t_1 ago and sampling at rate ρ can be written as

$$\begin{aligned}
f(\mathbf{s}|n,t_1,\lambda,\mu,\rho,\text{survival}) &= (n-2)!\prod_{i=2}^{n-1}\lambda\frac{p_1(t_i)}{\nu_{t_1}}, && (57)\\
\text{where}\quad \nu_t &= 1-\frac{1}{\rho}P_s(0,t)\exp(-(\lambda-\mu)t). && (58)
\end{aligned}$$

To make this more clear, the two lineages from the initial split make it into the sample with single progeny (themselves), the probability that this occurs is $p_1(t_1)^2$. Births of new lineages occur at rate $n(t)\lambda P_s(t,T)$, and the probability of each leaving exactly 1 sampled descendant is $p_1(t)$. In order to condition on

n individuals in the sample, and root time t_1, we must make sure that from the two initial lineages there are n descendants, this amounts to division by

$$(n-1)\left(P_s(0,t)\left(1-\eta_{T-t_1,T-t_1}\right)\right)^2 \eta_{T-t_1,T-t_1}^{n-2} \tag{59}$$

$$= (n-1)\left(P_s(0,t)\left(1-\frac{u_t P_s(0,T)}{P(0,T)}\right)\right)^2 \left(\frac{u_t P_s(0,T)}{P(0,T)}\right)^{n-2} \tag{60}$$

$$= (n-1)p_1(t)^2\left(1-\frac{p_1(t)}{P_s(0,t)}\right)^{n-2}. \tag{61}$$

Finally consider unseen transmissions. These events along a given lineage occur independently of seen events and thus may be placed independently at rate $\lambda p_0(t)$. This observation means that we may lay unseen transmission events over a variable pure birth tree.

4.4 DETAILED METHODS

The full collection of variables and parameters are summarised in Table 4.1.

Notation summary	
G	The genealogy.
E	HLA data at the leaves of G.
X	Sequence data at the leaves of G, excluding the epitope under consideration.
t	Time, measured forwards from the time of the most recent common ancestor, T_{MRCA}.
Θ	Collection of mutation parameters which determine the distribution of sequences at the tips of G.
Ψ	Collection of epidemiological parameters governing the structure of G.
$\boldsymbol{\lambda}$	Collection of parameters that model escape and reversion within host, $\boldsymbol{\lambda} := \{\lambda_{\text{esc}}, \lambda_{\text{rev}}, q\}$.
$S(t)$	Number of susceptible individuals at time t.
$N(t)$	Number of individuals at time t.
$Y_b^a(t)$	Number of individuals with host type a and virus type b at time t.
$\mathbf{Y}(t)$	Vector of each of the $Y_b^a(t)$s. $\mathbf{Y}(t) = \left(Y_0^0(t), Y_1^0(t), Y_0^1(t), Y_1^1(t)\right)$.
$Y(t)$	Number of infected individuals in the populations. $Y(t) = Y_0^0(t) + Y_1^0(t) + Y_0^1(t) + Y_1^1(t)$.
B	Birth rate into the susceptible population.

Table 4.1: Summary of the notation used in the detailed methods.

m	Death rate in the susceptible population.
c	Contact rate.
β	Probability of transmission given a contact.
λ	Birth rate of the birth-death process - the rate of becoming infectious.
λ_{esc}	Escape rate.
λ_{rev}	Reversion rate.
q	Proportion of hosts in the population with the relevant HLA type
f, g, h	Priors on $\Psi, \Theta, \boldsymbol{\lambda}$ respectively.
a	Host type (0 = HLA mismatched, 1 = HLA matched).
b	Virus type (0 = wildtype at epitope, 1 = escape mutant at epitope).
$\mathcal{S}$	State space of the combination of host and virus type (a, b). $\mathcal{S} = \{(0,0), (0,1), (1,0), (1,1)\}$.
$\mathbf{Q}(t)$	Instantaneous rate matrix governing transitions between states in $\mathcal{S}$.
$\mathbf{P}(t)$	Probability of observing each state in $\mathcal{S}$ at time t, $\left(\mathbf{P}(t) = \left(P_{(0,0)}(t), P_{(0,1)}(t), P_{(1,0)}(t), P_{(1,1)}(t)\right)\right)$.
s_k	State of the host virus pair at node k in the genealogy G.
$L_s(k)$	Likelihood of the data below node k, conditional on it having state s.
L	Likelihood of the tree G.
π_{s_0}	Prior distribution at the root.
μ	Rate of becoming non-infectious.
ρ	Proportion of the infected population sampled at the present.
γ	Rate of growth of the infected population.
$\hat{\lambda}(t)$	Rate of placement of unseen transmissions along lineages existing in G at time t.
$p_0(t)$	Probability that a lineage existing at time t in the full birth-death process has no sampled lineages existing at the present.

Table 4.1: (continued) Summary of the notation used in the detailed methods.

4.4.1 *Data*

The Swiss-Spanish intermittent treatment trial (SSITT) was a study of structured treatment interruptions in chronically HIV-1 infected patients, described in detail elsewhere [73]. A consensus sequence (obtained from cycle sequencing) from each of the individuals is sampled at a single time-point [82]. We consider the HLA typed subset of the cohort for our analyses, and remove sequences in which >10% of nucleotides could not be called. As not all patients were sequenced in

each gene this led to a collection of 79, 67, and 53 sequences for *gag*, *pol* and *nef* with associated HLA types respectively. q was estimated from population data taken from [178].

4.4.2 *Definition of escape mutation*

All amino acid changes which occur at the same point(s) as a previously defined escape mutations validated *in vitro* are considered to confer escape. Examples are shown in Figure 4.1. This is clearly an imperfect definition, but provides a sensible compromise between the clear overestimation of allowing all mutations within an epitope to be defined as escape, and only considering mutations validated as escapes by *in vitro* assays.

RPMTYKAAV	consensus
RPMSYKAAV	verified as escape
RPMTFKAAV	verified as escape
RPMQYKAAV	defined as escape
RTMTYKAAV	not defined as escape

Figure 4.1: Summary of escape mutation definition, any mutation at a site at which a mutation has been verified as conferring escape, we define as an escape mutation.

4.4.3 *Reference sequences*

For the later peeling step in our method, sequences sampled over a collection of dates are required to estimate a scaling from 'coalescent time' (measured in units of N_e generations) to 'calendar time' (which we measure in *years*). This is needed to obtain our rate estimates for λ_{esc} and λ_{rev} in units of $years^{-1}$. We added dated European B clade sequences with full gene coverage before the year 2000 from the Los Alamos HIV sequence database [256]. These consisted of 52, 14 and 68 *gag*, *pol* and *nef* sequences respectively.

4.4.4 *Alignment*

We use HIValign [91] which uses a hidden Markov model (HMM) to align viral sequence data, followed by manual editing of the resulting alignment.

4.4.5 *Data trimming*

Non-B clade sequences were identified using recombinant identification program (RIP) [229] and were removed if confirmed as non-B clade or could not be identified at the 90% significance level. The epitope under investigation was removed from the alignment, as this is considered to be under selection, and could therefore lead to clustering of escape mutations and thus affect rate estimates. We do not remove all epitopes under selection as this would involve removing the majority of our sequence information. This led to 55, 54 and 48 *gag*, *pol* and *nef* sequences.

4.4.6 *Nucleotide substitution model*

We used jmodeltest [223] Akaike information criterion (AIC) scores to determine the nucleotide substitution model to use. We found that the generalised time-reversible model with a Gamma distributed substitution rate across sites and a proportion of invariant sites ($GTR + \Gamma + I$) was chosen for each gene. The GTR model is the most general neutral, independent, finite-sites, time-reversible model possible. $GTR + \Gamma + I$ allows for an estimated portion of invariant sites along the sequence, and lets among site rate variation be Γ distributed.

4.4.7 *Sampling from the posterior tree distribution*

We ran BEAST [68] on our sequence data with the epitope removed, X, to generate samples from the posterior distribution of trees given an exponential coalescent tree prior, together with the $GTR + \Gamma + I$ model of nucleotide substitution. We did not use a model which considers codon position as we found that this significantly reduced estimated sample sizes (ESS). Chains are 10^8 updates long and we allow a burn-in of 10^7 states. Chains converged for each run.

4.4.8 *Peeling algorithm*

After sampling a genealogy from the posterior output by BEAST, we use extra information, E, in order to determine estimates of $(\lambda_{\text{esc}}, \lambda_{\text{rev}})$. Our algorithm is based on that first proposed in [76], replacing the four nucleotide state space $\mathcal{N} := \{A, C, G, T\}$ with $\mathcal{S} := \{(0,0), (0,1), (1,0), (1,1)\}$. We denote the pair (a, b); $a, b \in \{0, 1\}$ such that $\{0, 1\}$ in the first entry denotes {HLA mismatch, HLA match}, and $\{0, 1\}$ in the second entry denotes {no escape mutation, escape mutation}. Transmission is assumed to take place at rate λ and the probability that a given individual is HLA matched at the epitope under investigation is q. We assume q is constant in the population over time, as any selection on hosts would occur on far larger timescales [53]. We may then obtain the probability of a given configuration of the data E at the tips, conditional on the genealogy G, a collection of epidemiological parameters Θ and a collection of parameters $\boldsymbol{\lambda} = \{\lambda_{\text{esc}}, \lambda_{\text{rev}}, q\}$, which model escape and reversion down each lineage governed by

an instantaneous rate matrix $\mathbf{Q}$ which we specify. The instantaneous rate matrix is based on the following collection of differential equations:

Susceptibles:

$$\frac{dS(t)}{dt} = B - \left(\frac{\beta c}{N(t)} Y(t) + m\right) S(t) \tag{62}$$

HLA mismatched hosts infected with viruses without escape:

$$\frac{dY_0^0(t)}{dt} = \frac{\beta c}{N(t)}(1-q)S(t)\left(Y_0^0(t) + Y_0^1(t)\right) + \lambda_{\text{rev}} Y_1^0(t) - \mu Y_0^0(t) \tag{63}$$

HLA mismatched hosts infected with viruses with escape:

$$\frac{dY_1^0(t)}{dt} = \frac{\beta c}{N(t)}(1-q)S(t)\left(Y_1^0(t) + Y_1^1(t)\right) - \lambda_{\text{rev}} Y_1^0(t) - \mu Y_1^0(t) \tag{64}$$

HLA matched hosts infected with viruses without escape:

$$\frac{dY_0^1(t)}{dt} = \frac{\beta c}{N(t)} q S(t)\left(Y_0^0(t) + Y_0^1(t)\right) - \lambda_{\text{esc}} Y_0^1(t) - \mu Y_0^1(t) \tag{65}$$

HLA matched hosts infected with viruses with escape:

$$\frac{dY_1^1(t)}{dt} = \frac{\beta c}{N(t)} q S(t)\left(Y_1^0(t) + Y_1^1(t)\right) + \lambda_{\text{esc}} Y_0^1(t) - \mu Y_1^1(t) \tag{66}$$

where $Y(t) = Y_0^0(t) + Y_1^0(t) + Y_0^1(t) + Y_1^1(t)$ and $N(t) = Y(t) + S(t)$. During exponential growth, $S(t) \approx N(t)$, thus $\frac{\beta c S(t)}{N(t)} \approx \beta c$, which we define as our transmission rate λ. As in [85], we suppose that escape may only occur from the $(1,0)$ state at rate λ_{esc} and reversion may only occur from the $(0,1)$ state at rate λ_{rev} (Figure 3.1a). Thus, the instantaneous rate matrix $\mathbf{Q}$ may be defined

$$\mathbf{Q} = \begin{array}{c} \\ (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{array} \begin{array}{c} \begin{array}{cccc} (0,0) & (0,1) & (1,0) & (1,1) \end{array} \\ \begin{pmatrix} -\hat{\lambda}(t)q & 0 & \hat{\lambda}(t)q & 0 \\ \lambda_{\text{rev}} & -\lambda_{\text{rev}} - \hat{\lambda}(t)q & 0 & \hat{\lambda}(t)q \\ \hat{\lambda}(t)(1-q) & 0 & -\lambda_{\text{esc}} - \hat{\lambda}(t)(1-q) & \lambda_{\text{esc}} \\ 0 & \hat{\lambda}(t)(1-q) & 0 & -\hat{\lambda}(t)(1-q) \end{pmatrix} \end{array}, \tag{67}$$

where $\hat{\lambda}(t) = \lambda p_0(T_{\text{MRCA}} - t)$, with time increasing towards the present. T_{MRCA} is the time before the present of the most recent common ancestor (MRCA), when $t = 0$. We use the convention that Q_{ij} refers to the transition from state i to state

j. $p_0(t)$ is the probability that a lineage at time t in the past does not have any sampled descendants.

$$p_0(t) = 1 - \frac{\rho(\lambda - \mu)}{\rho\lambda + (\lambda(1-\rho) - \mu)\exp(-(\lambda - \mu)t)}, \tag{68}$$

where μ is the rate of becoming non-infectious, and ρ is the sampled proportion at the present which we estimate separately [310]. This extra factor must be included to avoid double counting of transmission events, as we assume transmission events occur at coalescences within the tree. Indeed, these internal nodes are precisely those transmission events which result in a descendant which is sampled at the present, which occur with probability $1 - p_0(T_{MRCA} - t)$. $p_0(t)$ for a sampled birth-death process was originally derived in [310] and details of the derivation, excluded from the original paper, are given in the Section 4.3. We wish to determine the probability of a branch beginning in state i and ending in state j in order to then evaluate the likelihood over the entire tree. The time dependency in $\hat{\lambda}(t)$ means that this no longer has a simple closed form (as is the case for constant coefficients). If we define $\mathbf{P}(t) = (P_{(0,0)}(t), P_{(0,1)}(t), P_{(1,0)}(t), P_{(1,1)}(t))$ to be the vector of probabilities of observing each state in $\mathcal{S}$ at time t, we must numerically evaluate the solution to

$$\dot{\mathbf{P}}(t) = \mathbf{P}(t)\mathbf{Q}(t) \tag{69}$$

at the end of each branch within the tree, $\mathbf{P}(t_i)$ where t_i is the time of node i, subject to the initial condition $\mathbf{P}(t_k)$ being one of the four states with certainty at the start of the branch, where node k is the parent of node i. This amounts to evaluating the probability of starting in any one of the four states and ending in any one of the four states for each branch within the tree, which we define as $P_{s_k s_j}$ for $s_k, s_j \in \mathcal{S}$. The assumption of exactly one transmission at each internal node requires an alteration to the original peeling algorithm illustrated in Figure 4.2. We make the following change: pick with equal probability the branch which exists as a result of the transmission (as we are using the phylogeny as a proxy for the transmission tree, and assume that exactly one transmission takes place immediately after each internal node), the other branch represents virus remaining

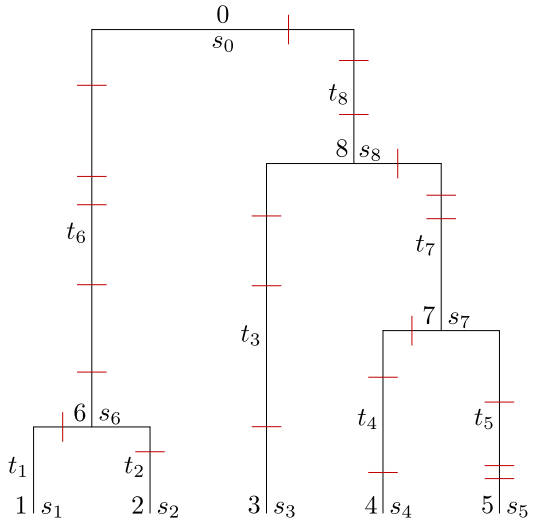


Figure 4.2: Each node is numbered from left to right, moving back through the generations, with the root node labelled 0. s_i at each node denotes the state at that node. Each t_i defines the lengths of the segment in the tree ending at node i. Transmission events between nodes are labelled by horizontal red segments, transmissions immediately after an internal node are displayed by the red vertical segments.

in the transmitting host. As we assume this transmission has taken place at the internal node, no change in the escape state can occur. We therefore sum over the two HLA states, keeping the escape state fixed at node k. Adopting the notation in [76], let $s_k \in \mathcal{S}$ denote the state of the host/virus pair at node k. To incorporate the sum over HLA types, we include the following additional notation: define $s_{k_{\text{HLA}}}$ as the state with the same escape state as node k, and first entry HLA. That is, $s_{k_{\text{HLA}}} := (\text{HLA}, s_k(2))$ where $\text{HLA} \in \{0, 1\}$. We may now write our process in terms of conditional likelihoods. $L_s(k)$ is defined as the likelihood of the data below node k conditional on it having state s. Node k has daughter nodes i and j,

$$L_{s_k}(k) = \frac{1}{2}\left(\left(\sum_{s_i \in \mathcal{S}}\left((1-q)P_{s_{k_0}s_i}(t_i) + qP_{s_{k_1}s_i}(t_i)\right)L_{s_i}(i)\right)\left(\sum_{s_j \in \mathcal{S}} P_{s_k s_j}(t_j)L_{s_j}(j)\right)\right.$$
$$\left. + \left(\sum_{s_i \in \mathcal{S}} P_{s_k s_i}(t_i)L_{s_i}(i)\right)\left(\sum_{s_j \in \mathcal{S}}\left((1-q)P_{s_{k_0}s_j}(t_j) + qP_{s_{k_1}s_j}(t_j)\right)L_{s_j}(j)\right)\right). \tag{70}$$

Finally considering the prior distribution at the root node (defined as node 0, with distribution of states π_{s_0}),

$$L = \sum_{s_0 \in \mathcal{S}} \pi_{s_0} L_{s_0}(0). \tag{71}$$

Assuming the initial state s_0 will not display the escape mutation and have the restricting HLA allele with probability q, $\pi_{s_0} = (1-q, 0, q, 0)$. Part of the output of BEAST is used to determine parameters in the peeling algorithm. Under the exponential coalescent prior, λ is equal to $\gamma + \mu$ where γ is the exponential growth rate of the infected population sampled in the MCMC [84]. μ is the estimated rate of becoming non-infectious, which we set at 0.1 $years^{-1}$ [85, 191], this emerges as the coalescence rate scales with the total infected number of individuals in the case of an exponentially growing population, see Section 4.4.9 and [84]. q is estimated from data in the HLA FactsBook [178]. Code to evaluate the likelihood of a genealogy conditional on some parameter set and tip data was coded in C++ using boost libraries and the odeint package [6], and checked against R code.

4.4.9 *Checking agreement between effective infected population and infected population*

We wish to validate that the effective infected population size is proportional to the infected population size. To do this, we follow the notation and arguments of Frost and Volz [84]. From Griffiths and Tavaré [108], we may incorporate fluctuations in population size by adjusting the time scale accordingly with $1/\nu(t)$ where $\nu(t)$ is the effective population size time t ago. Thus the coalescent in a deterministically changing population becomes a non-homogeneous Poisson process dependent upon the number of existing lineages through time. Therefore, the pdf of the time spent with i lineages (starting at time t say) is

$$P(x|t) = \frac{\binom{i}{2}}{\nu(t+x)} \exp\left(-\binom{i}{2}\int_t^{t+x} \frac{1}{\nu(\tau)} d\tau\right). \tag{72}$$

It was noticed [84] that we may then consider a deterministic version of the number of lineages based on these pdfs. Approximating i by a continuous variable, which we call $A(t)$, the rate of decrease of lineages measuring backwards through time is simply $\binom{A(t)}{2}/\nu(t)$, thus

$$\frac{dA(t)}{dt} = -\binom{A(t)}{2}\frac{1}{\nu(t)}. \tag{73}$$

Now, for our estimates to make sense, we require that $\nu(t)$ is proportional to $Y(t) = Y_0^0(t) + Y_1^0(t) + Y_0^1(t) + Y_1^1(t)$. In the case of SI models, the number of ancestral lineages over time can be modelled using the differential equation

$$\frac{dA(t)}{dt} = -\binom{A(s)}{2}\frac{f_{SY}}{\binom{Y(t)}{2}} \tag{74}$$

where f_{SY} is the rate at which susceptibles become infected [292]. In a simple frequency dependent SI model, $f_{SY} = \frac{\beta c S(t) Y(t)}{N(t)}$. Noting that the coalescent assumption assumes that $A(t)$ is small relative to $Y(t)$ over the course of the process. Approximating $\binom{Y(t)}{2}$ by $Y(t)^2/2$,

$$\frac{dA(t)}{dt} = -\binom{A(t)}{2}\frac{2\beta c S(t)}{Y(t)N(t)}. \tag{75}$$

Finally, during exponential growth, $S(t)/N(t) \approx 1$. Thus

$$\frac{dA(t)}{dt} \approx -\binom{A(t)}{2}\frac{2\beta c}{Y(t)} \tag{76}$$

and therefore $Y(t) \propto \nu(t)$ as required. It is important to recognise that this is not true in general.

4.4.10 *Integrating over trees*

We assume HLA and escape information is uninformative about the tree topology (see Section 4.2). From the BEAST output sample from the posterior, we subsample 1000 trees. *A priori* it is unknown where the majority of probability mass of $P(\boldsymbol{\lambda}|E, h, G, \Psi)$ lies (where h is our prior on $\boldsymbol{\lambda}$), so we perform simple hill climb-

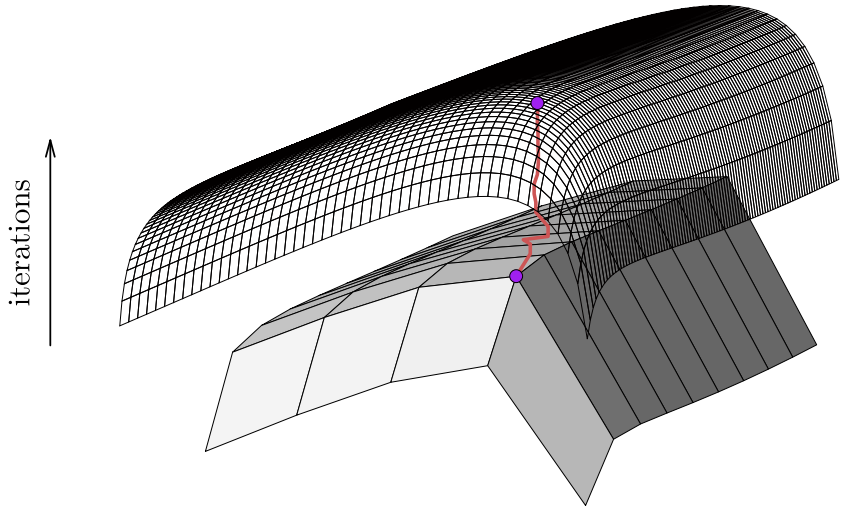
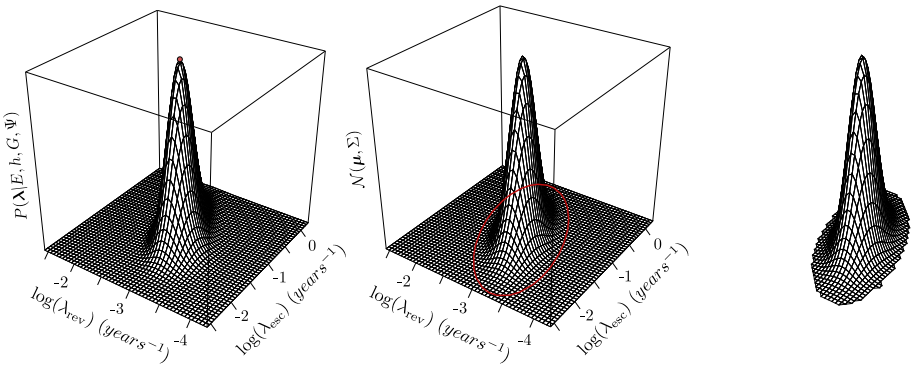


Figure 4.3: Here we consider a low resolution surface (shaded grey) and a higher resolution surface shown as a wireframe above. In each, the approximate MAP is coloured purple, and the red line shows the output of the hill climbing algorithm over time, displaying a high level of accuracy after only a few iterations.

ing to determine its maximum, $P_G^{\max}$ as shown in Figure 4.3. We may then approximate the surface by a bivariate normal distribution. We define an x% error ellipse as the ellipse enclosing x% of the density of the bivariate normal approximation. We evaluate $P(\boldsymbol{\lambda}|E,h,G,\Psi)$ for each sampled $\{G,\Psi\}$ pair on the smallest 50×50 lattice containing the 99.9% error ellipse, and enforce a $U[10^{-20},10^{3}]$ prior on λ_{esc} and λ_{rev} (Figure 4.4). To ensure that the normal approximation encompasses the majority of the mass, we check $P(\boldsymbol{\lambda}|E,h,G,\Psi)$ on the edge of the 99.9% error ellipse and evaluate over the entirety of $[10^{-20},10^{3}]\times[10^{-20},10^{3}]$ if at any point it lies above the threshold of $P_G^{\max}/1000$. We then normalise these probability densities over the domain on which they are evaluated.

4.4.10.1 *Details of the bivariate normal approximation*

For a given tree sampled from the posterior by BEAST, we may generate a likelihood surface for $(\lambda_{\text{esc}},\lambda_{\text{rev}})$. However, *a priori* we do not know what the surface will look like. We therefore need some scheme whereby the majority of density can be extracted. The method that we have chosen is to approximate the tree specific posteriors by a bivariate normal distribution (after taking logs of the λ_{esc} and λ_{rev} axes). The mean approximation $\hat{\boldsymbol{\mu}}=(\hat{\mu}_x,\hat{\mu}_y)$ is simply the MAP. It is important to note that the integral of the approximate tree specific posterior ob-



(a) (b) (c)

Figure 4.4: An example of the bivariate normal approximation. Figure 4.4a displays a true likelihood surface obtained using the peeling algorithm. Note the log transformation has been applied to the x and y axes. The MAP is marked red, the is the first parameter required for the normal approximation, next the curvature and value at the MAP are used to generate the normal approximation shown in Figure 4.4b. We mark an ellipse containing 99.9% of the probability density, to obtain the portion of the true likelihood surface shown in Figure 4.4c.

tained by applying the peeling algorithm over the chosen values of $(\lambda_{\text{esc}}, \lambda_{\text{rev}})$ is not 1, we therefore need some scaling (A, say) to be incorporated before making the approximation. To allow for this we suppose that tree specific posteriors may be approximated by a surface of the form:

$$P(\boldsymbol{\lambda}|E,h,G,\Psi) \approx A\mathcal{N}(\boldsymbol{\mu},\Sigma). \tag{77}$$

We determine the covariance matrix Σ of the bivariate normal approximation by making the following observation, consider the curvature of $P(\boldsymbol{\lambda}|E,h,G,\Psi)$ at the MAP, $\boldsymbol{\mu}$:

$$C(\boldsymbol{\mu}) = \left(\begin{array}{cc} \frac{\partial^2 P}{\partial {\lambda_{\text{esc}}}^2} & \frac{\partial^2 P}{\partial \lambda_{\text{esc}} \lambda_{\text{rev}}} \\ \frac{\partial^2 P}{\partial \lambda_{\text{rev}} \lambda_{\text{esc}}} & \frac{\partial^2 P}{\partial {\lambda_{\text{rev}}}^2} \end{array} \right) \Bigg|_{(\mu_{\lambda_{\text{esc}}}, \mu_{\lambda_{\text{rev}}})} \tag{78}$$

$$= \frac{A}{2\pi\sigma_{\lambda_{\text{esc}}}\sigma_{\lambda_{\text{rev}}}(1-\rho^2_{\lambda_{\text{esc}}\lambda_{\text{rev}}})^{3/2}} \left(\begin{array}{cc} \frac{-1}{\sigma^2_{\lambda_{\text{esc}}}} & \frac{\rho_{\lambda_{\text{esc}}\lambda_{\text{rev}}}}{\sigma_{\lambda_{\text{esc}}}\sigma_{\lambda_{\text{rev}}}} \\ \frac{\rho_{\lambda_{\text{esc}}\lambda_{\text{rev}}}}{\sigma_{\lambda_{\text{esc}}}\sigma_{\lambda_{\text{rev}}}} & \frac{-1}{\sigma^2_{\lambda_{\text{rev}}}} \end{array} \right). \tag{79}$$

Using this information we may determine the correlation between λ_{esc} and λ_{esc}, $\rho_{\lambda_{\text{esc}}\lambda_{\text{rev}}}$.

$$\rho_{\lambda_{\text{esc}}\lambda_{\text{rev}}} = \text{sgn}\left(C_{1,2}(\boldsymbol{\mu})\right)\sqrt{\frac{C_{1,2}(\boldsymbol{\mu})^2}{C_{1,1}(\boldsymbol{\mu})C_{2,2}(\boldsymbol{\mu})}}. \tag{80}$$

Since $\frac{\sigma_{\lambda_{\text{rev}}}}{\sigma_{\lambda_{\text{esc}}}} = -\rho_{\lambda_{\text{esc}}\lambda_{\text{rev}}}\frac{C_{1,1}(\boldsymbol{\mu})}{C_{1,2}(\boldsymbol{\mu})}$ and $P(\boldsymbol{\mu}|E,h,G,\Psi) = \frac{A}{2\pi\sigma_{\lambda_{\text{esc}}}\sigma_{\lambda_{\text{rev}}}\sqrt{1-\rho^2_{\lambda_{\text{esc}}\lambda_{\text{rev}}}}}$,

$$A = \frac{\text{sgn}(P(\boldsymbol{\mu}|E,h,G,\Psi))2\pi\rho_{\lambda_{\text{esc}}\lambda_{\text{rev}}}P^2(\boldsymbol{\mu}|E,h,G,\Psi)}{C_{1,2}(\boldsymbol{\mu})\sqrt{1-\rho^2_{\lambda_{\text{esc}}\lambda_{\text{rev}}}}}, \tag{81}$$

$$\sigma_{\lambda_{\text{esc}}} = \left(\frac{-P(\boldsymbol{\mu}|E,h,G,\Psi)}{C_{1,1}(\boldsymbol{\mu})(1-\rho^2_{\lambda_{\text{esc}}\lambda_{\text{rev}}})}\right)^{\frac{1}{2}}, \tag{82}$$

$$\sigma_{\lambda_{\text{rev}}} = -\rho_{\lambda_{\text{esc}}\lambda_{\text{rev}}}\frac{C_{1,1}(\boldsymbol{\mu})}{C_{1,2}(\boldsymbol{\mu})}\left(\frac{-P(\boldsymbol{\mu}|E,h,G,\Psi)}{C_{1,1}(\boldsymbol{\mu})(1-\rho^2_{\lambda_{\text{esc}}\lambda_{\text{rev}}})}\right)^{\frac{1}{2}}. \tag{83}$$

It follows that from just $P(\boldsymbol{\mu}|E,h,G,\Psi)$ and $C(\boldsymbol{\mu})$, we may determine a normal approximation of each tree specific posterior. Contours of a bivariate normal are described by concentric ellipses, the direction of the semi-major and semi-minor axes is given by the eigenvectors of Σ. The corresponding eigenvalues λ_{maj} and λ_{min} determine the squared relative lengths of the semi-major and semi-minor axes of an ellipse which form a contour on the surface. The ellipse with semi-major and semi-minor axis lengths equal to $\chi^2_2(\alpha)\sqrt{\lambda_{\text{maj}}}$ and $\chi^2_2(\alpha)\sqrt{\lambda_{\text{min}}}$ (where $\chi^2_2(\alpha)$ is the quantile function for probability α of the chi-squared distribution with 2 degrees of freedom) defines the level set which encloses α of the density. We will use this to define a region on which we evaluate the true likelihood. An example of this approximation is displayed in Figure 4.4. In order to integrate over trees, we must define a common lattice over which to sum. This is achieved by defining an 800×800 grid on the minimum region of parameter space which encloses the union of domains defined for each tree specific posterior. We interpolate each surface, before summing over all surfaces obtained. The normalisation of this marginalisation defines our estimate of $P(\boldsymbol{\lambda}|X,E,f,g,h)$, where f and g are priors on epidemiological parameters and substitution parameters respectively.

4.4.11 *Credible regions*

These are defined in $P(\boldsymbol{\lambda}|X, E, f, g, h)$ by determining the contour level enclosing α% of the probability mass, where α is the percentage credible region required.

4.4.12 *Mahalanobis distance*

The distance of a vector $x = (x_1, x_2, \ldots, x_n)$ from a group of values with mean $\eta = (\eta_1, \eta_2, \ldots, \eta_n)$ and covariance matrix Σ is $D(x) = \sqrt{(x-\eta)\Sigma^{-1}(x-\eta)^T}$ [176].

4.5 EQUIVALENCE OF ODE MODEL AND COMPOSITE LIKELIHOOD

The ODE system of [85] may be written as $\dot{\mathbf{Y}} = \tilde{\mathbf{Q}}\mathbf{Y}$ where $\tilde{\mathbf{Q}} = \mathbf{Q} + \mathbf{D}$, $\mathbf{D}$ the diagonal matrix with $D_{ii} = \lambda - \mu$, $D_{ij} = 0$; $i \neq j$, and $\mathbf{Y}$ is the vector of the number of infected individuals in each class. The general solution of $\dot{\mathbf{X}} = \mathbf{A}\mathbf{X}$, where $\mathbf{A}$ has constant entries is $\mathbf{X} = \sum_i c_i \boldsymbol{\nu}_i e^{\omega_i t}$ where $\boldsymbol{\nu}_i$ and ω_i are the eigenvectors and corresponding eigenvalues of $\mathbf{X}$. Letting ω and $\boldsymbol{\nu}$ be an eigenvalue and corresponding eigenvector of $\mathbf{Q}$,

$$\mathbf{Q}\nu = \omega\boldsymbol{\nu} \quad \Rightarrow \quad (\mathbf{Q} + \mathbf{D})\boldsymbol{\nu} = \omega\boldsymbol{\nu} + \mathbf{D}\boldsymbol{\nu}. \tag{84}$$

$\mathbf{D}$ diagonal $\Rightarrow (\mathbf{Q} + \mathbf{D})\boldsymbol{\nu} = (\omega + \lambda - \mu)\boldsymbol{\nu}$. Therefore, solutions of $\dot{\mathbf{Y}} = \tilde{\mathbf{Q}}\mathbf{Y}$ may be written as $\mathbf{Y} = e^{(\lambda-\mu)t}\mathbf{P}$ where $\mathbf{P}$ is the solution to $\dot{\mathbf{P}} = \mathbf{Q}\mathbf{P}$. $\mathbf{P}$ is simply a rescaling of $\mathbf{Y}$ by total population size through time. Therefore, using the ODE approach is equivalent to assuming a multinomial distribution over the four possible states, subject to some initial condition. Thus, we may consider the likelihood of the parameters data under the ODE model as a composite likelihood $L_c(\boldsymbol{\lambda}|E) = (1-q)\left(\prod_{j=1}^{m} P_{(0,0)s_j}(T)\right) + q\left(\prod_{j=1}^{m} P_{(1,0)s_j}(T)\right)$, where m is the total number of

cross-sectional samples. Note the implicit assumption that λ is not a function of time, hence the ODE model is equivalent to the assumption that all transmission events are unseen ($\rho = 0$), and therefore the genealogy is completely star-like and all lineages are independent, with the same initial condition.

4.6 JUSTIFICATION OF VALIDITY OF HYPOTHESIS TESTING

In order to test simple hypotheses, we need to compute the likelihood under the various models, and perform a likelihood ratio test. To do this, we make the following observation:

First, note $P(E|G,\Theta,\Psi,\boldsymbol{\lambda}) = P(E|G,\Psi,\boldsymbol{\lambda})$. Now, letting

$$L(\boldsymbol{\lambda}^*|X,E) := L(\boldsymbol{\lambda} = \boldsymbol{\lambda}^*|X,E), \tag{85}$$

$$L(\boldsymbol{\lambda}^*|X,E) \propto \int\int\int P(X,E|G,\Theta,\Psi,\boldsymbol{\lambda}^*)P(G,\Theta,\Psi,\boldsymbol{\lambda}^*)\, dG d\Psi d\Theta \tag{86}$$

Thus,

$$\begin{aligned}
L(\boldsymbol{\lambda}^*|X,E) &= C\int\int\int P(X,E|G,\Theta,\Psi,\boldsymbol{\lambda}^*)P(G,\Theta,\Psi,\boldsymbol{\lambda}^*)\, dG d\Psi d\Theta && (87)\\
&= \widetilde{C}\int\int\int \tfrac{P(X|G,\Theta,\Psi)P(G,\Theta,\Psi)}{P(X)} P(E|G,\Theta,\Psi,\boldsymbol{\lambda}^*,X)\, d\Theta d\Psi dG && (88)\\
&= \widetilde{C}\int\int\int P(G,\Theta,\Psi|X)P(E|G,\Theta,\Psi,\boldsymbol{\lambda}^*,X)\, d\Theta d\Psi dG && (89)\\
&\approx \frac{1}{M}\sum_{i=1}^{M} P(E|G_i,\Theta_i,\Psi_i,\boldsymbol{\lambda}^*,X) && (90)\\
&\approx \frac{1}{M}\sum_{i=1}^{M} P(E|G_i,\Psi_i,\boldsymbol{\lambda}^*) && (91)
\end{aligned}$$

where

$$\widetilde{C} = CP(X) \qquad G_i,\Theta_i,\Psi_i \sim P(G,\Theta,\Psi|X), \tag{92}$$

and M is the number of samples from BEAST. Thus, by sampling from BEAST and computing the mean of the probability of observing E under the parameters

conditional on the genealogy and epidemiological parameters, we approximate the likelihood.

4.7 CONNECTIONS BETWEEN FORMULATIONS OF RECONSTRUCTED TREE LIKELIHOODS

For what follows, we need the following probabilities, derived by Kendall [142]:

$$P(\tilde{t}, t) = \frac{\lambda - \mu}{\lambda - \mu \exp(-(\lambda - \mu)(t - \tilde{t}))}, \tag{93}$$

$$p_0(t) = 1 - P(0, t), \tag{94}$$

$$p_i(t) = P(0, t)(1 - u_t)u_t^{i-1} \quad i > 0, \tag{95}$$

where

$$u_t = \frac{\lambda(1 - \exp(-(\lambda - \mu)t))}{\lambda - \mu \exp(-(\lambda - \mu)t)}. \tag{96}$$

$P(\tilde{t}, t)$ is the probability that a single lineage alive at time $\tilde{t}$, has at least one descendant at the later time t, $p_0(t)$ is the probability that a lineage starting at time 0 has no descendants, some time t later. Finally $p_i(t)$ is the probability that a lineage starting at time 0 has i descendants after time t.

In [279], the likelihood of a reconstructed birth-death tree with speciation times $\mathbf{s} = \{s_1, s_2, \ldots, s_{n-2}\}$ (where s_i is the i^{th} speciation event measured and indexed backwards in time), n leaves, and some labelled history F, given a transmission rate λ, a death rate μ and time of the first transmission event in the tree (the root), $s_{n-1} := t$ is derived as

$$f_{\text{Thompson}} := f(\mathbf{s}, n, F|\lambda, \mu, t) = \frac{2^{n-1}}{n!}\lambda^{n-2}(p_1(t))^2 \prod_{j=1}^{n-2} p_1(s_j). \tag{97}$$

If we condition on survival of the tree and labellings, then we can obtain the likelihood in [196], f_{Nee}. Conditioning on survival amounts to dividing by the probability that the two lineages at the root survive to the present, $P(0, t)^2 =$

$(1-p_0(t))^2$. The number of possible labellings is $|F| = 2^{n-1}/(n!(n-1)!)$. Therefore,

$$f_{\text{Nee}} := f(\mathbf{s}, n|F, \lambda, \mu, t, \text{survival}) \tag{98}$$

$$= \frac{f_{\text{Thompson}}}{|F|P(0,t)^2} \tag{99}$$

$$= \frac{n!(n-1)!}{2^{n-1}(1-p_0(t))^2}\left(\frac{2^{n-1}}{n!}\lambda^{n-2}(p_1(t))^2\prod_{j=1}^{n-2}p_1(s_j)\right) \tag{100}$$

$$= \frac{(n-1)!}{P(0,t)^2}\left(\lambda^{n-2}((1-u_t)P(0,t))^2\prod_{j=1}^{n-2}p_1(s_j)\right) \tag{101}$$

$$= (n-1)!\lambda^{n-2}(1-u_t)^2\prod_{j=1}^{n-2}p_1(s_j). \tag{102}$$

Rewritten, with time increasing forwards and the present set at time T, 0 the time of the initial infected individual, and t_i the time of the $(i-1)^{\text{th}}$ speciation event, indexed forwards through time,

$$f_{\text{Nee}} = (n-1)!\lambda^{n-2}(1-u_t)^2\prod_{j=1}^{n-2}p_1(s_j) \tag{103}$$

$$= (n-1)!\lambda^{n-2}(1-u_{(T-t_2)})^2\prod_{j=3}^{n}(1-u_{(T-t_j)})P(t_j,T). \tag{104}$$

This is the likelihood in equation 20 of [196]. In order to obtain the likelihood in [310] for $\rho = 1$ (complete sampling), we must condition on n lineages surviving to the present. Given that the initial birth event is fixed at time t ago (or $T-t_2$ ago in the notation of Nee et al [196]), we divide f_{Nee} by

$$(n-1)(P(0,t)(1-u_{(T-t_2)}))^2u_{(T-t_2)}^{n-2}. \tag{105}$$

Here, we are conditioning that two lineages must survive to the present and that their combined number of descendants is n. The two lineages are independent, thus for any combination of numbers of progeny $(k, n-k)$ with $k > 0$ for the two lineages, we obtain $P(0,t)(1-u_{(T-t_2)})u_{(T-t_2)}^{k-1}P(0,t)(1-u_{(T-t_2)})u_{(T-t_2)}^{n-k-1} =$

$(P(0,t)(1-u_{(T-t_2)}))^2 u_{(T-t_2)}^{n-2}$. There are $(n-1)$ such combinations. Division of f_{Nee} by Equation (105) gives

$$f_{\text{Y\&R}} := f(\mathbf{s}|n,F,\lambda,\mu,t)(n-2)!\lambda^{n-2}\prod_{j=3}^{n}\frac{(1-u_{(T-t_j)})P(t_j,T)}{u_{(T-t_2)}}. \tag{106}$$

This is $(n-2)!\lambda^{n-2}\prod_{j=1}^{n-2}\frac{p_1(s_j)}{u_{s_{n-1}}}$ in the notation of Thompson [279]. Letting time increase into the past, but indexing forwards and defining the deepest birth to have occurred at time t_1, we can obtain the likelihood of Yang and Rannala [310]. This change amounts to a relabelling of nodes in the Thompson notation $s_j \to t_{n-j}$:

$$f_{\text{Y\&R}} = (n-2)!\lambda^{n-2}\prod_{j=2}^{n-1}\frac{p_1(t_j)}{u_{t_1}}. \tag{107}$$

Noting that $u_t = 1 - P(0,t)\exp(-(\lambda-\mu)t)$, which is v_t in Yang and Rannala [310],

$$f_{\text{Y\&R}} = (n-2)!\prod_{j=2}^{n-1}\lambda\frac{p_1(t_j)}{v_{t_1}}. \tag{108}$$

Finally, to obtain Gernhard's likelihood [95], f_{Gern}, we note that

$$u_t = \frac{\lambda}{\mu}\frac{\mu(1-\exp(-(\lambda-\mu)t))}{\lambda-\mu\exp(-(\lambda-\mu)t)} = \frac{\lambda}{\mu}p_0(t) \tag{109}$$

$$\Rightarrow \quad f_{\text{Y\&R}} = (n-2)!\prod_{j=2}^{n-1}\mu\frac{p_1(t_j)}{p_0(t_1)}. \tag{110}$$

Removing the ordering on the node times, and using the lemma [95] that any collection of unordered internal node times $\{s_2, s_3, \ldots, s_{n-1}\}$ is equally likely (and that there are $(n-2)!$ such orderings), we obtain f_{Gern}:

$$f_{\text{Gern}} = \prod_{j=2}^{n-1}\mu\frac{p_1(s_j)}{p_0(t_1)}. \tag{111}$$

Thus everything is nicely connected when we consider full sampling. To consider partial sampling lineages at the present, we must return to Kendall's derivations [142]. Nee et al. [196] looked to Kendall for their derivation of $P_s(\tilde{t},t)$. Analogous to derivations carried out in [196], a reconstructed sampled birth-death process

can be shown to be equivalent to a pure birth process with variable rate $\lambda P_s(t,T)$. Conditioning in exactly the same way as the full sampling case yields $f_{\text{Y\&R}}$ for $\rho \in (0,1]$, which is also written in [267].

4.8 DERIVATION OF SURVIVING LINEAGES THROUGH TIME WITHIN A DETERMINISTIC MODEL OF TRANSMISSION

Let $N(t)$ be the total number of infected individuals in the population, with t measured backwards in time and initiated at the present. $M(t)$ is the number of individuals in the sampled subtree, where sampling only occurs at $t = 0$. Finally, $\tau(t)$ is the cumulative number of transmissions present in the sampled subtree in $[0,t]$.

$$\frac{dN}{dt} = \mu N - \lambda N, \tag{112}$$

$$\frac{dM}{dt} = -\lambda N \left(\frac{M}{N}\right)^2, \tag{113}$$

$$\frac{d\tau}{dt} = \lambda N \left(2\left(\frac{M}{N}\left(1-\frac{M}{N}\right)\right)\frac{1}{2}\right) + \lambda N \left(\frac{M}{N}\right)^2 = \lambda M. \tag{114}$$

Note that as in [292, 84], N and M are assumed to be large such that $M-1 \approx M$ and $N-1 \approx N$. Solving for $N(t)$ and $M(t)$, we obtain

$$N(t) = N(0)\exp((\mu-\lambda)t), \tag{115}$$

$$M(t) = \frac{(\mu-\lambda)\,N(0)}{\lambda + \frac{N(0)}{M(0)}(\mu-\lambda) - \lambda\exp(t(\lambda-\mu))}. \tag{116}$$

We wish to find what proportion of the total transmissions per lineage are unseen transmissions through time. Note that $\frac{d\tau}{dt}$ can be split into seen transmissions and unseen transmissions, and considering transmissions per lineage we

divide through by $M(t)$. Thus, the rate of change of unseen transmissions as a proportion of the total transmissions in a subtree is

$$1 - \frac{M}{N} = 1 - \frac{(\mu - \lambda)}{\exp\left(-(\lambda - \mu)\, t\right)\left(\lambda + \frac{N(0)}{M(0)}(\mu - \lambda)\right) - \lambda} \tag{117}$$

$$= 1 - \frac{\frac{M(0)}{N(0)}(\lambda - \mu)}{\lambda\frac{M(0)}{N(0)} + \exp\left(-(\lambda - \mu)\, t\right)\left(\lambda\left(1 - \frac{M(0)}{N(0)}\right) - \mu\right)}. \tag{118}$$

This is the same as $1 - P_s(0, t)$ previously derived, the probability that a lineage existing t ago survives into a sample at the present.

5

MEASURING FITNESS ACROSS TECHNIQUES AND POPULATIONS

5.1 INTRODUCTION

We wish to know how fast escape mutations arise and revert within individual infected hosts. By carefully incorporating the dependency structure present within the viral genealogy of a collection of cross-sectional HIV-1 sequences, we are better able to estimate these average escape and reversion rates for known epitopes. Certain HLA genotypes are associated with viral control [186, 239, 221], and understanding the mechanism underlying this observation is of great interest. For example, are viruses sampled from individuals with the HLA B*57 genotype replicating at a lower rate than average? Such questions can be addressed experimentally using *in vitro* assays that measure viral replicative capacity: essentially the exponential *in vitro* growth rate of an isolate relative to some standard. A lower replicative capacity would suggest that the HLA response has constrained the viral population to a region of sequence space with lower average replicative capacity. Assuming all escape mutations have the same impact upon CTL immune evasion, we would expect CTL escape mutations that are more costly *in vitro* (i.e. have a lower viral replicative capacity) to escape more slowly in HLA matched hosts and revert more quickly in HLA mismatched hosts. However, this has never been tested. Though the assumption makes sense intuitively, it is clearly a vast simplification. The high dimensional sequence space in which the HIV-1 proteome resides implies that the path of least resistance to restore viral fitness in a new host may not be as trivial as a simple flip of an amino acid back to 'wildtype' at the escape position in question. For the same reason, apparently crippling mutations (as measured by fitness after site directed mutagenesis) may

be restored through compensatory mutations.

In this chapter, we investigate whether replicative capacity *in vitro* is related to either escape or reversion rates *in vivo* as defined and estimated by our integrated model and an existing ODE based model [85]. By doing so, we will determine whether the assumption that times to escape or reversion are related to viral fitness as measured *in vitro* is justified. The following questions also naturally arise:

1. Do ODE and MAP estimates of escape and reversion rates broadly agree within cohorts?
2. Do our escape and reversion rate estimates agree across populations with distinct HLA genotype and viral sequence distributions?

Finally, we examine whether two methods of measuring the impact of an escape mutation on viral fitness (site directed mutagenesis applied to a reference strain, versus the average fitness of a viral strain containing the escape variant) are correlated to determine if the two assays give consistent results.

5.2 METHODS

5.2.1 *Cohorts*

For our analysis, we consider paired viral sequence and host HLA genotype data taken from studies carried out in three different populations. We briefly outline the studies and portions of datasets used in Table 5.1. Full details are provided in the respective publications.

1. Switzerland: The Swiss portion of the Swiss-Spanish intermittent treatment trial (SSITT) [73, 82].
2. British Columbia, Canada: A portion of the HAART observational medical evaluation and research (HOMER) cohort [308].

3. A cohort recruited from Bloemfontein, the Free State, South Africa [123, 122].

For each dataset, we restrict our attention to *gag* and to the subset of sequences with the majority subtype in each population as assessed using RIP [229] software in which sequences are compared to subtype representatives in sliding windows along the viral sequence. Our reasons are threefold. Firstly, non-synonymous mutations in *gag* are likely to have a detrimental fitness effect as they encode structural proteins that are among the most highly conserved in the viral proteome [166]. Secondly, the most protective HLA alleles are associated with CTL responses to Gag proteins [146, 318, 203] Finally, site directed mutagenesis followed by replicative capacity assays have been carried out for a large number of known escape variants located in the *gag* gene [30].

Due to issues of confidentiality, we were unable to obtain the year of sequencing for the analysed portion of the HOMER dataset. However, this is not required as we do not require a rescaling to calendar time (measured in $years^{-1}$) from substitutions per site to determine the existence of correlation between the datasets. The 184 viral sequences which we analyse are taken from the year between 1996 and 1999 inclusive in which the largest number of *gag* sequences were obtained. In the case of the SSITT dataset, separate collections of sequences were available for the p17 and p24 proteins in the *gag* gene. There was not a strong overlap between the patient identifiers of these two sequence sets and the number of sequences in the SSITT dataset was the lowest of our three cohorts, so we chose not to combine them.

Dataset	SSITT	HOMER	Bloemfontein
Population analysed	Switzerland	British Columbia, Canada	Bloemfontein, South Africa
Cohort size	$n = 133$ Zürich ($n = 29$), Geneva ($n = 26$), Bern ($n = 11$), Basel ($n = 11$) St Gall ($n = 8$), Lugano ($n = 7$) Lausanne ($n = 5$)	$n = 3116$	$n = 884$, plasma samples taken from $n = 278$
Sampling date	2000	Single year in 1996-2006	February-September 2006
Sequences analysed	p17 $n = 38$, p24 $n = 55$	$n = 184$	$n = 198$
Treatment	HAART	ART naive on recruitment, prescribed ART between August 1996 and June 2006	ART naive
Study requirements	Undetectable VL for >6 months, $CD4^+$ count $> 300\mu l^{-1}$. no history of non-nucleoside reverse transcriptase inhibitors	≥ 3 antiretroviral drugs	Chronic, plasma taken by $CD4^+$, low and high favoured: 96 high ($> 500\mu l^{-1}$), 18 medium ($200 - 400\mu l^{-1}$), 164 low ($< 100\mu l^{-1}$)

Table 5.1: Summary of three studies from which *gag* sequence data was available. In each case, the number of sequences analysed is lower than the cohort size through a combination of lack of sequencing of *gag*, and data cleaning.

5.2.2 *Estimating escape and reversion rates*

For both of the methods below, we estimate HLA prevalence using data from the HLA factsbook [178]. For the Canadian and Swiss data, we set the prevalence at the Caucasian estimate. For the South African data, we set the prevalence at the Black estimate.

5.2.2.1 *MAP estimate using integrating method*

Full details of the method are provided in Chapter 3 and Section 4.4. Briefly, we combine Felsenstein's tree peeling algorithm with existing phylogenetic software to merge statistical phylogenetic techniques with an ODE model. By doing so, we are able to use the information contained within the viral genealogy to infer better estimates of escape and reversion rates ($\lambda_{\text{esc}}, \lambda_{\text{rev}}$). There are four key steps in our inference:

1. Make the mild assumption that the genealogy and HLA/escape information are conditionally independent given the sequence information with the epitope removed.

2. Sample genealogies from the posterior conditional on the sequence data with the epitope removed using the program BEAST [68], with a coalescent prior in an exponentially growing population. The growth rate parameter is sampled within the MCMC.

3. For each sampled viral genealogy, determine the posterior density for $(\lambda_{\text{esc}}, \lambda_{\text{rev}})$. This is achieved through a modification of Felsenstein's peeling algorithm [76].

4. Averaging over tree specific posteriors, generate a consensus posterior density for the parameters of interest; $(\lambda_{\text{esc}}, \lambda_{\text{rev}})$.

In Section 5.3, we report the *maximum a posteriori* (MAP) estimate of escape and reversion in each epitope of the Gag proteins for which both estimates of fitness costs were reported in Boutwell *et al.* [30] and sequence data encoding the epitope was available to us. When an escape mutation residing in an epitope

reaches consensus in the population, the variant peptide is known as a negatope. Negatopes reported in Boutwell *et al.* [30] are highlighted in Tables 5.2-5.4 by asterisks.

5.2.2.2 *Purely ODE based model [85]*

For each escape variant under consideration, viral sequence/HLA genotype pairs are split into four classes: HLA mismatched without escape, HLA mismatched with escape, HLA matched without escape and HLA matched with escape. Escape is defined as in Section 4.4.2. Given these four proportions for an escape site, we then determine a best fit for λ_{esc} and λ_{rev} under the model described by Equations (18)-(23) during the exponential growth phase. Best estimates are determined using BFGS [195], an approximation to Newton's method of hill climbing. Time of the start of the epidemic is set at the T_{MRCA} of a BEAST run on the *gag* sequence data for each dataset.

5.2.3 In vitro *fitness assays*

5.2.3.1 *Site directed mutagenesis*

In Section 5.3.3 we determine whether there is an association between estimates of escape and reversion rates from populations of infected individuals and *in vitro* viral fitness assays. As in [30], we approximate the intrinsic fitness of an escape variant as the replicative capacity of the escape mutant if it is inserted into the 'wildtype' subtype B consensus strain, in the absence of any other sequence variation. Fitness assays measurements were taken from [30], in which full details of the procedure are provided. Summarising the method (shown in Figure 5.1), HIV-1 subtype B recombinant plasmid clones were created using *p*NL4-3, from which a *p*NL4-3 *gag* subclone was cleaved out and ligated into *p*UC19 (a cloning vector). The subclone and full length *p*NL4-3 were then propagated. The desired mutations were then attained through site directed mutagenesis. Double mutants were created through multiple rounds of mutation. *p*NL4-3 was used as

a 'backbone' for the variant viruses. In order to determine that the desired base changes were generated, viruses were sequenced before and after the mutation step. In order to determine a measure of replicative capacity, CD8 depleted peripheral blood mononuclear cells (PBMCs) were inoculated in triplicate for each modified virus at a multiplicity of infection (MOI) of 0.001. For each viral mutant population, experiments were performed at least twice, independently, in PBMCs from at least two donors. For each assay, wells with the 'parental' NL4-3 virus were included as controls. Starting at day 2 and for each day until day 7, $50\mu l$ of supernatant was sampled with replacement and stored. Viral RNA was then isolated to determine the number of copies of viral RNA each day. The growth rate was then estimated using linear regression on semi-log transformed data. Impact of of the escape mutation was defined as the relative growth rate of the escaped strain to the 'wildtype' NL4-3 strain.

5.2.3.2 *Average fitness of an escape variant*

We also consider a second estimate of change in replicative capacity imposed by an escape variant - the average replicative capacity of viral populations in which the consensus sequence (as measured by population sequencing) has the escape variant (or non-wildtype amino acid) at the site(s) in question. There are estimates of replicative capacity of viral populations drawn from individuals in numerous geographical locations. Unfortunately, due to slight differences between assay protocols, datasets could not be amalgamated. We consider replicative capacity assay measurements from Bloemfontein [122], Durban [250] and Gaborone [138] datasets.

We briefly describe the method of Huang *et. al* [122] (the other methods are similar), the protocol is summarised in Figure 5.2.

Chimeric viruses were constructed by recombining *gag-protease* deleted pNL4-3 HIV-1 plasmid (pNL4-3Δ*gag-protease*) with autologous *gag-protease* amplified from virus extracted from patient plasma. By doing so, any fitness cost/benefit will be associated to variation within the *gag-protease* region of the viral genome. For transfection, pNL4-3Δ*gag-protease* was linearised and co-transfected with patient derived *gag-protease* PCR product in tat driven GFP reporter T-cells (Figure

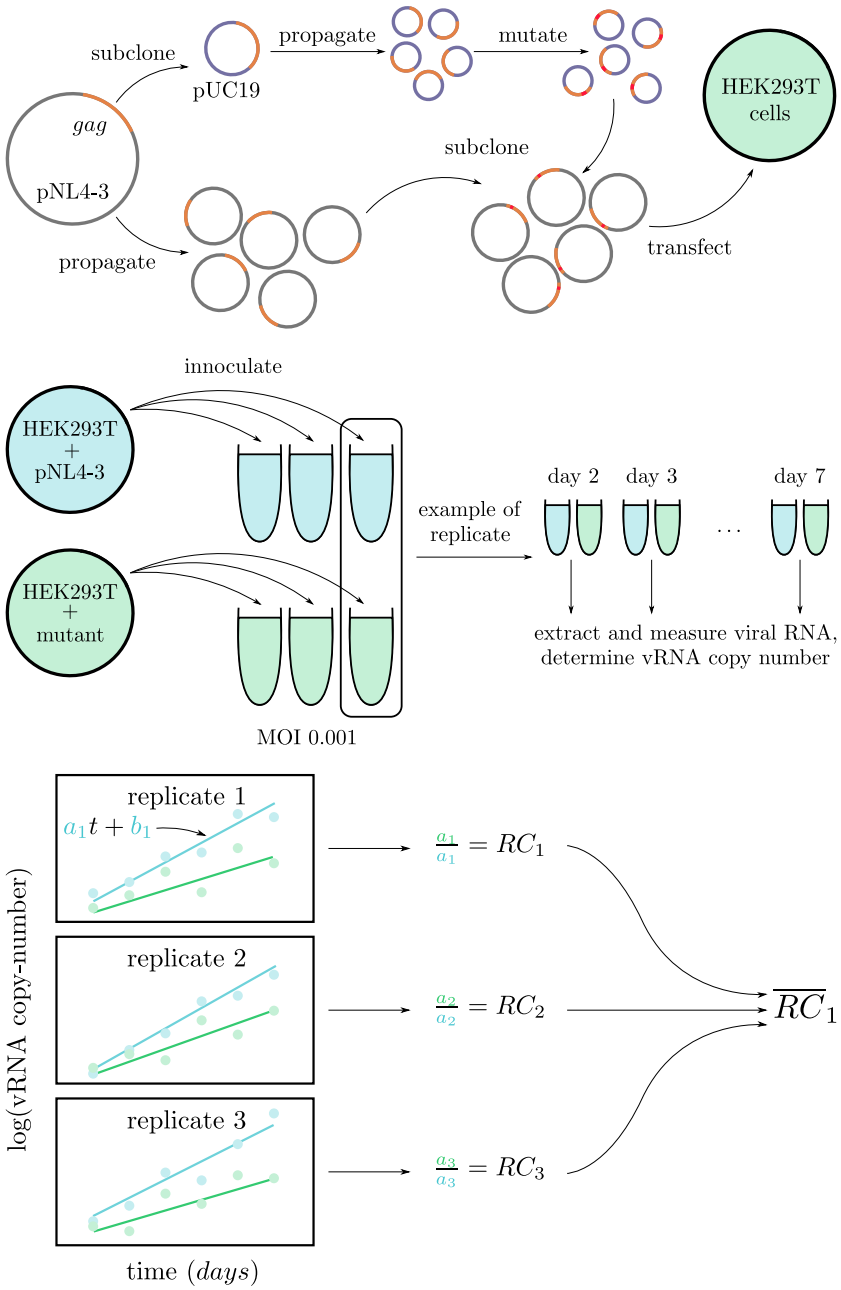


Figure 5.1: Site directed mutagenesis followed by replicative capacity assay.
Top row: Subtype B plasmid clones were created from *p*NL4-3, a *gag* subclone is cleaved out and ligated into *p*UC19. This subclone and *p*NL4-3 are then propagated. Site directed mutagenesis was applied to generate the required mutants. The mutated region is then replaced into *p*NL4-3 by subcloning, and viral stocks were generated by transfection into HEK293T cells.
Second row: CD8 depleted PBMCs were inoculated in triplicate for each modified virus. For each triplicate experiment a control was also carried out in triplicate in which the 'parental' NL4-3 was inoculated into CD8 depleted PBMCs. For day 2 to 7, $50\mu l$ of supernatent was sampled with replacement and then stored at $-80°$C.
Third row: Estimates of the growth rate were carried out using regression by plotting log(vRNA copy-number) against time and then normalising the resulting estimated gradient to that estimated for the growth the 'parental' NL4-3 virus, and averaging over the replicates in that assay (so for the n^{th} triplicate assay that would be $\overline{RC}_n$. The final replicative capacity reported was then the average of these $\overline{RC}_n$s over the number of true independent replicates.

5.2. After 5 days, GFP expression was measured every 1-2 days by fluorescence activated cell sorting (FACS). As soon as $\geq$ 15% of cells were positive for GFP (positive was defined as a GFP signal >10 times the mean fluorescent intensity of uninfected cells), the viral supernatent was harvested and stored. Viral supernatent was added to 1×10^6 GFP reporter T-cells at an MOI of 0.003. Replicative capacity was then estimated by considering the proportion of GFP-reporter T-cells that were positive for GFP over time. This proportion was measured each day, at least in duplicate, between days 2 and 7. Growth was then calculated by performing a regression on the natural log of this proportion through time. Replicative capacity was estimated by normalising to the growth rate of the NL4-3 lab strain.

5.3 RESULTS

We apply our integrated model, and purely differential equation based model [85] to the SSITT, HOMER and Bloemfontein cohorts. Full details of the results together with estimated replicative capacity of escape variants as previously estimated [30] are shown in Tables 5.2, 5.3 and 5.4 respectively. Figure 5.3 summarises the data for all three regions. We plot the proportion of HLA matched individuals with consensus escape against the proportion of HLA mismatched individuals with consensus escape for each escape variant detailed in Tables 5.2 - 5.4. Most of the points lie below $y = x$, suggesting that selection for escape is taking place for the majority of the HLA/epitope pairings. Also, note that many of the epitopes cluster together, independent of the population from which samples where drawn.

Throughout our results, estimates of time to escape and time to reversion for the HOMER dataset are given in units of substitutions per site. This is because the date of sequencing was not available due to issues of confidentiality, and thus a scaling from substitutions per site to *years* could not be estimated. For all plots in which the scaling of time is the same, we plot $y = x$ to show where points would lie if there were perfect agreement between rate estimates.

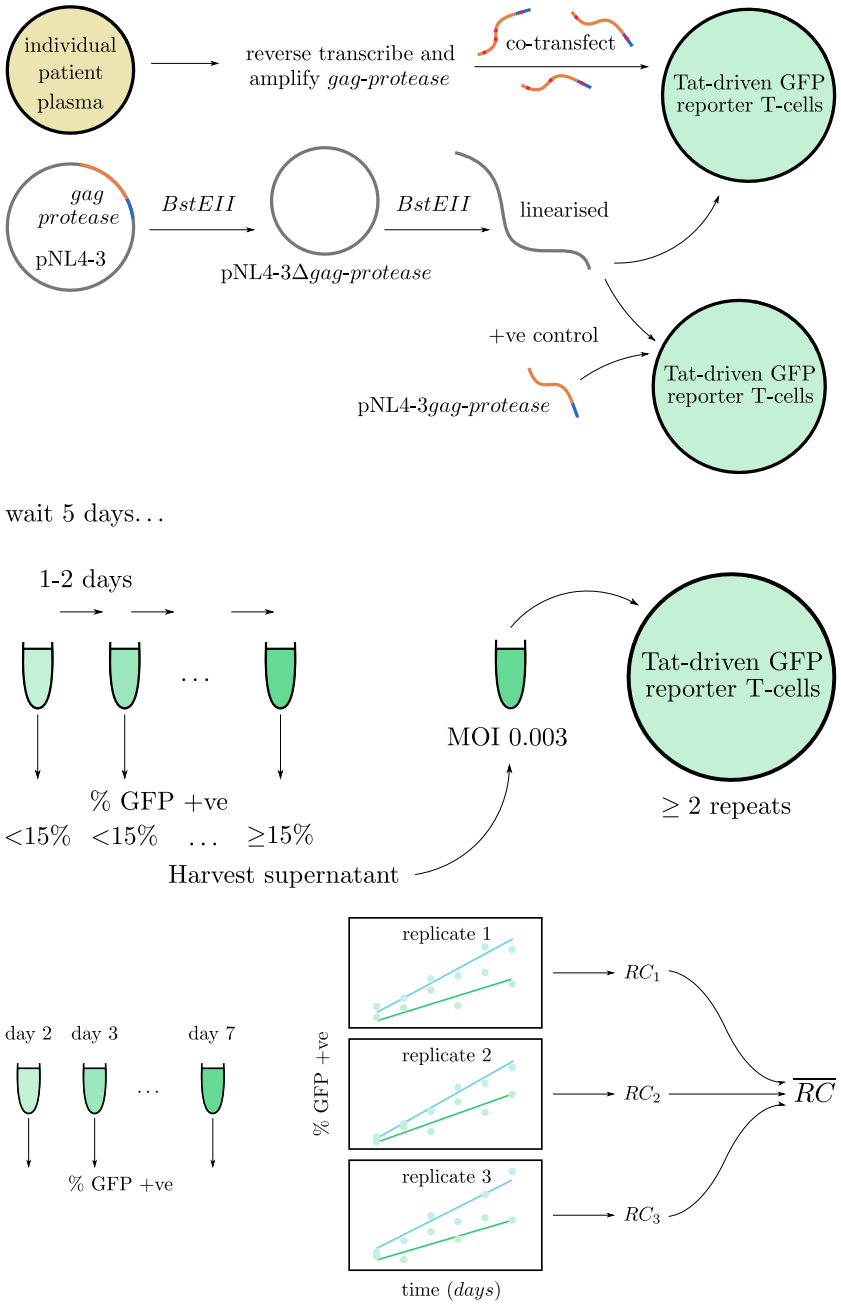


Figure 5.2: Replicative capacity assay applied to viral samples taken from hosts.
Top row: *gag-protease* RNA was reverse-transcribed and amplified from patient plasma. *p*NL4-3 was mutated so that BstEII (a restriction enzyme) digest would delete *gag-protease* and ligate the ends. This *p*NL4-3Δ*gag-protease* was then linearised and co-transfected with patient *gag-protease* into tat-driven GFP reported T-cells. As a control, *p*NL4-3Δ*gag-protease* was transfected with lab-strain *gag-protease*.
Second row: Samples were left for 5 days and GFP expression was determined every 1-2 days. As soon as $\geq$ 15% cells were positive, the supernatant was removed and stored at $-80°$C. When all samples were ready, recombinant supernatent was then added to tat-driven GFP reporter cells at a MOI of 0.003, so that replicative capacity could be determined.
Third row: Each day between days 2 and 7, the % GFP positive cells was determined at least in duplicate. The growth rate was estimated using regression, by plotting log(%GFP +ve cells) against time. This was then normalised by the growth rate of the control. and averaged over replicates to generate the reported replicative capacity; $\overline{RC}$.

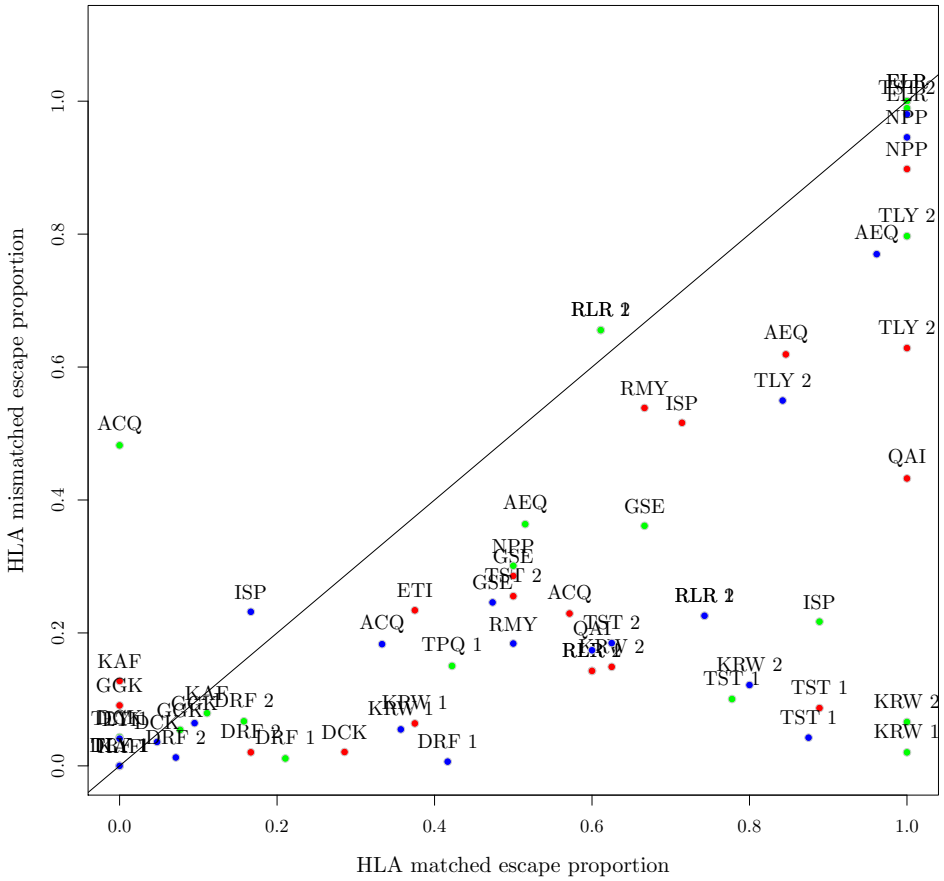
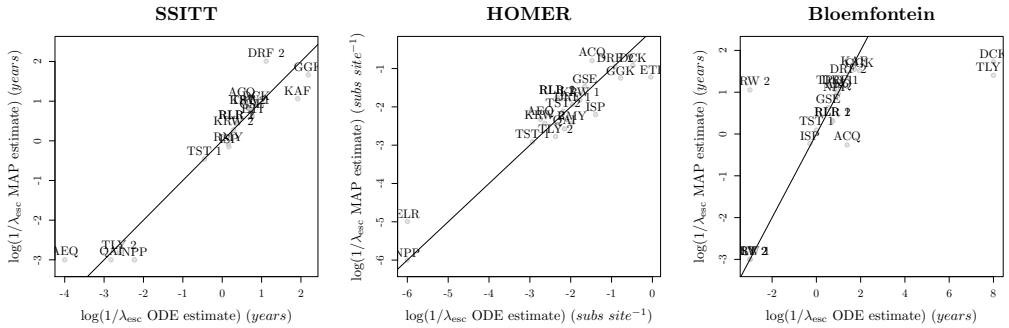


Figure 5.3: Scatterplot of escape proportions in HLA matched and HLA mismatched hosts, coloured by population. Data points in which no individuals in a population are HLA matched (0/0 escape proportion in HLA matched hosts) were excluded. SSITT, Bloemfontein and HOMER data are shown in red, green and blue respectively. Numbers after the three letter abbreviation of each epitope refers to the ordering in Tables 5.2-5.4, as some epitopes have more than one escape mutation and/or associated restricting HLA.

5.3.1 *Correlation between MAP and ODE estimates within populations*

We first examine whether ODE and MAP estimates of reversion rates agree within cohorts. Note that here we are comparing $\sim$ 20 epitopes with their associated restricting HLA types, far more than the 4 investigated in Chapter 3. We consider escape sites and associated restricting HLAs separately for these analyses when estimating escape and reversion rates, see Table 5.2-5.4 columns 1 and 2. Given the relatively small cohort sizes and growth of the infected population within each region, we expect estimates to be broadly similar under the two models. This is because small sample sizes will lead to long terminal branches in the genealogy, and so independence assumption of the ODE model (equivalent to assuming a completely star-like genealogy) will approximate our integrated method (see Section 3.4 and 4.5). For escape rate estimates, we find that this is indeed the case. Throughout this chapter, we denote r as the Pearson product moment correlation coefficient and ρ as the Spearman rank correlation coefficient. We observe a strong correlation between the two estimates in each population, see Figure 5.4a-5.4c. This was not the case for reversion. It was only with the SSITT dataset that reversion estimates from the two methods were significantly correlated ($r = 0.50$; $p = 0.029$, $\rho = 0.73$; $p = 0.00037$). However, we note that both models (but particularly the ODE estimate) can dramatically overestimate time to reversion (and escape to a lesser extent) when the true rate is low, and the number of sampled individuals is small (see Section 3.3.1). Reversion rate estimation is particularly difficult as transmission of an escape variant is required. By definition, this is rarer than transmission of wildtype at a given site. Potential for overestimation of time until reversion is most striking in ODE estimates based on the HOMER dataset, though further data is required to determine if this is the reason for discrepency between reversion estimates. Overall, our results agree with the simulation studies of Section 3.3.1. Despite simplifying assumptions, the ODE model performs similarly to the integrated model, particularly when estimating high rates.



(a) $r = 0.96, p = 9.4 \times 10^{-11}$. $\rho = 0.94, p = 1.4 \times 10^{-9}$.

(b) $r = 0.94, p = 9.0 \times 10^{-10}$. $\rho = 0.83, p = 7.5 \times 10^{-6}$.

(c) $r = 0.67, p = 0.0017$. $\rho = 0.81, p = 3/1 \times 10^{-5}$.

(d) $r = 0.50, p = 0.029$. $\rho = 0.73, p = 0.00037$.

(e) $r = 0.38, p = 0.10$. $\rho = 0.44, p = 0.050$.

(f) $r = 0.074, p = 0.76$. $\rho = 0.46, p = 0.046$

Figure 5.4: Correlation between the MAP estimates under our integrated model and ODE estimates of escape and reversion rates in different populations. Figures 5.4a, 5.4b and 5.4c show scatterplots of estimates for time until escape for the two models using data from the SSITT, HOMER and Bloemfontein cohort respectively. Corresponding estimates of the reversion rates using the two methods are shown in Figures 5.4d-5.4f. Pearson product-moment correlation coefficients (r) and Spearman rank correlation coefficients (ρ) with associated p-values are displayed. Numbers after the three letter abbreviation of each epitope refers to the ordering in Tables 5.2-5.4, as some epitopes have more than one escape mutation and/or associated restricting HLA.

5.3.2 *Correlation between rates across cohorts*

5.3.2.1 *Escape rates across cohorts*

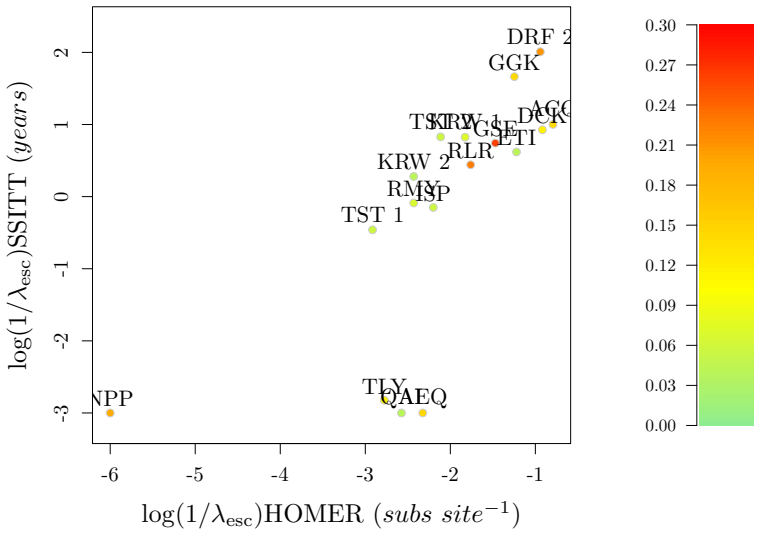
The viral backgrounds from any given two of the three populations we analyse are diverged. In the case of escape, if epistasis played a strong role, we would expect to observe differences in times to escape at the various epitopes. Conversely, a correlation between escape rates estimated from very different populations would suggest that, on average, the effect of an escape mutation on viral fitness (in the context of the HLA background) swamps epistatic effects driven by differences in the viral sequence distributions of the two populations. Figures 5.5, 5.6a and 5.6b show scatterplots of time to escape estimates for the different combinations of populations - HOMER against SSITT, HOMER against Bloemfontein, and SSITT against Bloemfontein respectively. We find a remarkable correlation between escape rate estimates for the SSITT and HOMER cohorts ($r = 0.72$; $p = 0.0012$, $\rho = 0.86$; $p = 9.9 \times 10^{-6}$). We do not find clustering of HLAs of similar prevalence in any portion of the space of times to escape suggesting that the observed correlation is not an artefact of the HLA prevalence in the two populations. We do not observe any correlation between the SSITT or HOMER escape rate estimates and the Bloemfontein dataset escape rate estimates. A potential source of error in our Bloemfontein escape rate estimates was the low number of individuals in the Bloemfontein dataset with the restricting HLA (shown in Table 5.3), due to the rarity of these HLA genotypes in African populations coupled with the limited size of the dataset. Additionally, some escape mutations are found at 100% prevalence or are not present at all in the Bloemfontein dataset. 7/20 rate estimates are obtained from data for which just one individual has the restricting HLA, or 0%/100% of HLA matched hosts have the escape mutation. Such escape proportions in HLA matched hosts result in a lack of power to estimate escape rates and leads to spurious rate estimates (instantaneous escape or $\sim$ infinite time to escape, see Section 3.3.1). To attempt to correct for this source of error we obtained a far larger dataset of African sequences coupled with host HLA genotype data [145, 138, 250, 164, 181, 123, 257, 180]. Unfortu-

nately, such a large dataset was computationally intractable using our integrated model. However, we note the strong correlation between our integrated method and the ODE based approach when the sampling proportion is small, discussed in Section 3.4 and shown in Section 5.3.1. We therefore apply the ODE based model to the entirety of the available African data. This requires the additional naive approximation that all of the available sequences are cross-sectional. To check that this approximation is robust, we determine the correlation between MAP estimates of time to escape using the Bloemfontein dataset and ODE based estimates when applied to all of the African dataset 5.7. Despite an outlier (possibly where our integrated method performs particularly poorly), we find a strong correlation when estimates could be obtained using both methods. Scatterplots of estimates of time to escape using the naive ODE method on the large African dataset against the two subtype B datasets are shown in Figures 5.8a and 5.8b. In both cases we obtain significant $p-$values associated to the Spearman rank correlation coefficient ($\rho = 0.73$; $p = 0.00094$ and $\rho = 0.62$; $p = 0.010$ for African against HOMER and African against SSITT time to escape estimates respectively, though the Pearson product moment correlation coefficients did not reach signifcance: $r = 0.44$; $p = 0.075$, $r = 0.36$; $p = 0.17$ respectively). We also note that the strongest disagreement seen in the plots occurs when the escape mutation is at high prevalence in one or other of the populations.

Overall, these results suggest that, at least within subtype, epistasis does not play a strong role in governing rates of escape. We note that it is interesting to see such a strong correlation between SSITT and HOMER escape rate estimates, even when the corresponding escape site data points do not cluster together in Figure 5.3.

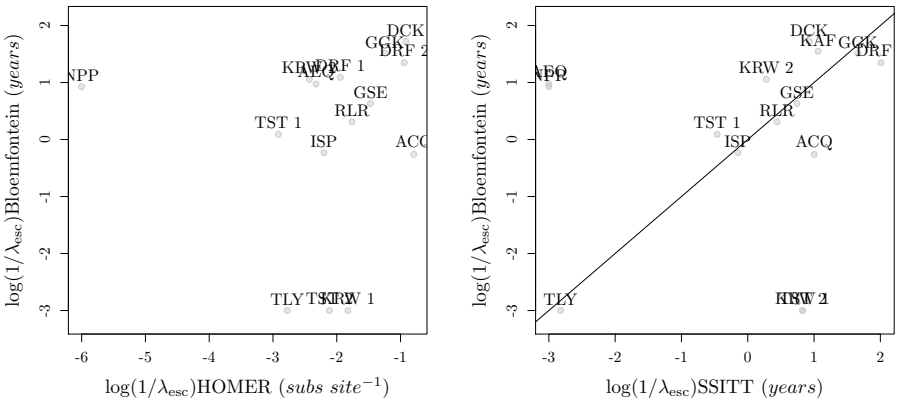
5.3.2.2 *Reversion rates across cohorts*

Figures 5.9, 5.10 and 5.7b show the corresponding scatterplots of estimates of time to reversion. We do not find a signal of association between reversion rate estimates across any pairwise combination of populations. This lack of correlation could be explained by lack of signal due to the inherent difficulty of estimating reversion rates, or more interestingly, by epistatic effects. The diverged



$$r = 0.72, p = 0.0012$$
$$\rho = 0.86, p = 9.9 \times 10^{-6}$$

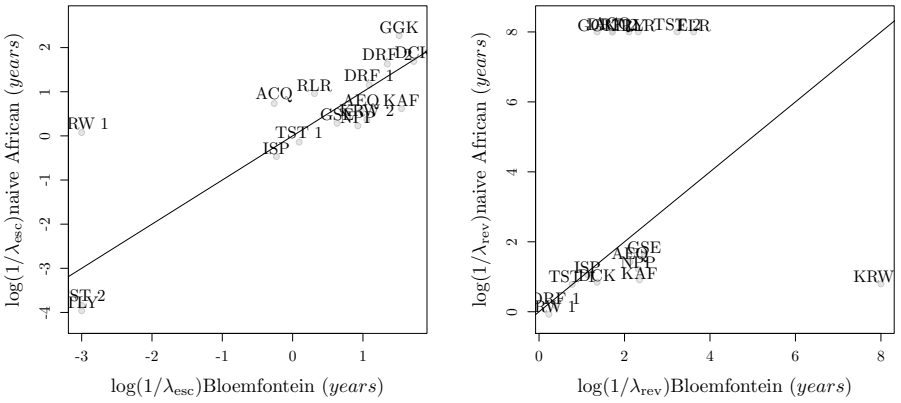
Figure 5.5: Correlation between time to escape estimated from the SSITT cohort and time to escape estimated from the HOMER cohort. Pearson product-moment correlation coefficient and Spearman rank correlation coefficient with associated p-values are displayed. Numbers after the three letter abbreviation of each epitope refers to the ordering in Tables 5.2-5.4, as some epitopes have more than one escape mutation and/or associated restricting HLA. The scale bar indicates the HLA prevalence taken from the HLA handbook [178] for each of the restricting HLA types associated to escape in Caucasian populations. We find a scattering of colours across the plot, suggesting that HLA prevalence is not generating the association.



(a) $r = 0.076, p = 0.79$.
$\rho = 0.30, p = 0.28$.

(b) $r = 0.18, p = 0.53$.
$\rho = 0.35, p = 0.20$.

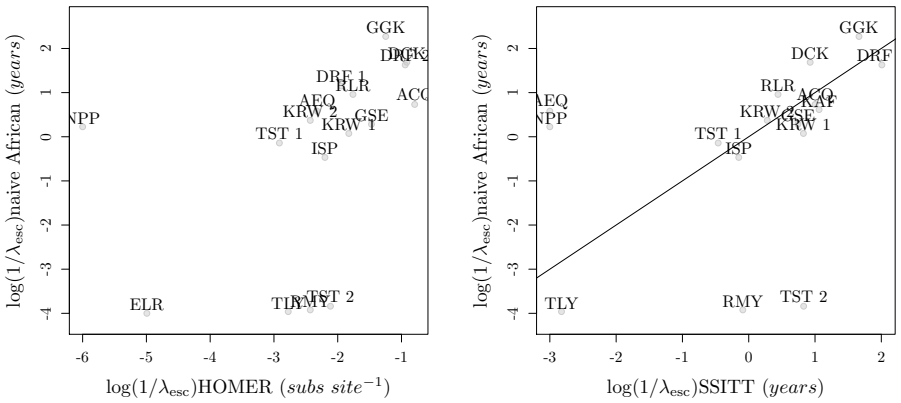
Figure 5.6: Correlation between time to escape estimated from the SSITT/HOMER data, and Bloemfontein data, taken from the South African cohort. Figures 5.6a - 5.6b show MAP estimates of time to escape for the Bloemfontein dataset plotted against SSITT and HOMER MAP estimates respectively. Spearman rank correlation coefficients (ρ) and Pearson product-moment correlation coefficients (r) with associated p-values are displayed. Numbers after the three letter abbreviation of each epitope refers to the ordering in Tables 5.2-5.4, as some epitopes have more than one escape mutation and/or associated restricting HLA.



(a) $r = 0.82, p = 9.3 \times 10^{-5}$.
$\rho = 0.79, p = 0.00029$.

(b) $r = 0.080, p = 0.76$.
$\rho = 0.43, p = 0.086$

Figure 5.7: Correlation between time to escape/reversion estimated from the Bloemfontein data using the integrated method and all available African data using a naive approach. Spearman rank correlation coefficients (ρ) and Pearson product-moment correlation coefficients (r) with associated p-values are displayed. Numbers after the three letter abbreviation of each epitope refers to the ordering in Tables 5.2-5.4, as some epitopes have more than one escape mutation and/or associated restricting HLA.



(a) $r = 0.44, p = 0.075$
$\rho = 0.73, p = 0.00094$.

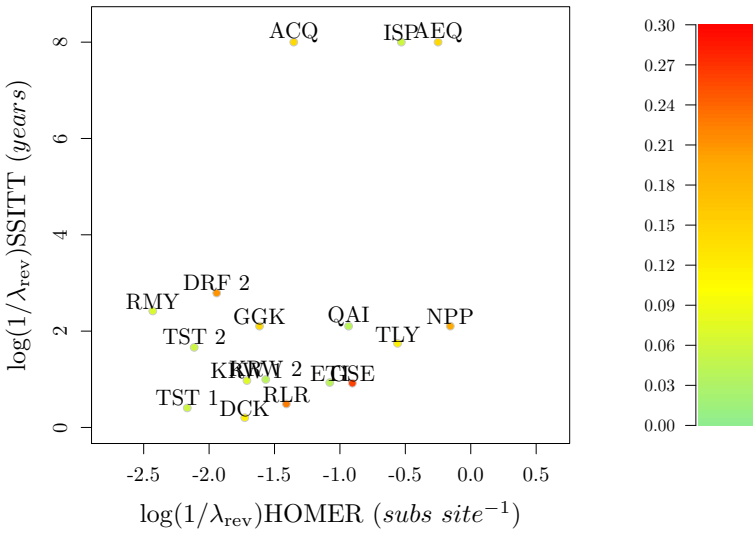
(b) $r = 0.36, p = 0.17$.
$\rho = 0.62, p = 0.010$.

Figure 5.8: Correlation between time to escape estimated from SSITT/HOMER data and all available African data using a naive approach. In Figures 5.8a - 5.8b SSITT and HOMER MAP estimates are plotted against a naive estimate of the escape rate. Spearman rank correlation coefficients (ρ) and Pearson product-moment correlation coefficients (r) with associated p-values are displayed. Numbers after the three letter abbreviation of each epitope refers to the ordering in Tables 5.2-5.4, as some epitopes have more than one escape mutation and/or associated restricting HLA.

viral backgrounds found in each population may lead to distinct favoured routes through sequence space to restore viral fitness (which may or may not include reversion at the escape site).

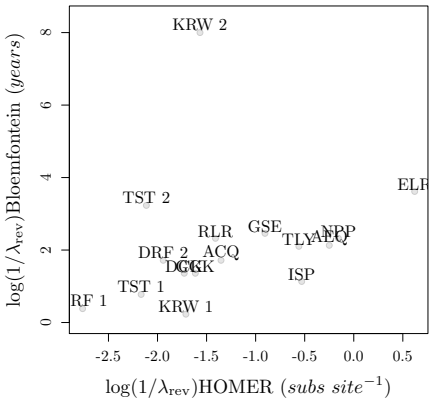
5.3.3 *Escape and reversion rates within cohorts vs* in vitro *replicative capacity estimates*

Moving to the main focus of the chapter, we wish to determine if reversion rate (or escape rate) may be used as a proxy for the viral fitness of an escape variant. To address this, we look for associations between our MAP estimate of time until escape/reversion and replicative capacity as measured by *in vitro* replicative assays after site directed mutagenesis has been applied to a subtype B lab strain. Implicit in this test of association is the assumption that such an *in vitro* measure of replicative capacity recapitulates viral fitness *in vivo*. We first examine the correlation between an *in vitro* measure of replicative capacity and

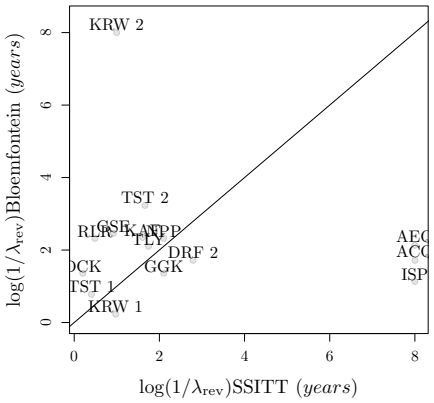


$$r = 0.42, p = 0.094$$
$$\rho = 0.32, p = 0.21$$

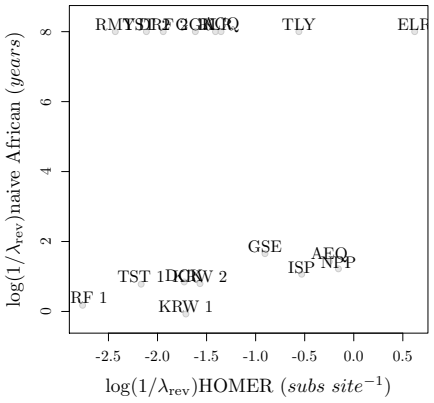
Figure 5.9: Correlation between time to reversion estimated from the SSITT cohort and time to reversion estimated from the HOMER cohort. The Pearson product-moment correlation coefficient and Spearman rank correlation coefficient with associated p-values are displayed. The scale bar indicates the HLA prevalence taken from the HLA handbook [178] for each of the restricting HLA types associated to escape in Caucasian populations. Numbers after the three letter abbreviation of each epitope refers to the ordering in Tables 5.2-5.4, as some epitopes have more than one escape mutation and/or associated restricting HLA.



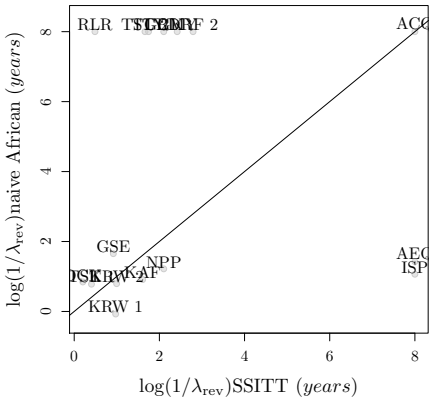
(a) $r = 0.22, p = 0.42$.
$\rho = 0.47, p = 0.068$.



(b) $r = -0.16, p = 0.57$.
$\rho = -0.038, p = 0.89$.



(c) $r = 0.032, p = 0.90$.
$\rho = 0.20, p = 0.45$.



(d) $r = 0.060, p = 0.82$.
$\rho = 0.45, p = 0.08$.

Figure 5.10: Correlation between time to reversion estimated from the SSITT/HOMER data, and Bloemfontein data, taken from the South African cohort. Figures 5.10a - 5.10b show MAP estimates of time to reversion for the Bloemfontein dataset plotted against SSITT and HOMER MAP estimates respectively. In Figures 5.10c - 5.10d SSITT and HOMER MAP estimates are plotted against a naive estimate of the reversion rate. Spearman rank correlation coefficients (ρ) and Pearson product-moment correlation coefficients (r) with associated p-values are displayed. Numbers after the three letter abbreviation of each epitope refers to the ordering in Tables 5.2-5.4, as some epitopes have more than one escape mutation and/or associated restricting HLA.

rate estimates in the two populations in which subtype B is the predominant subtype.

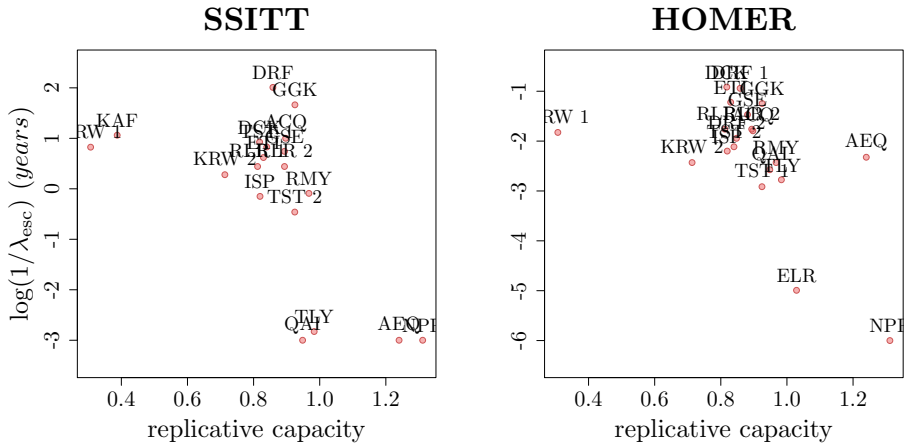
5.3.3.1 *Escape and reversion rate estimates in SSITT and HOMER datasets vs.* in vitro *replicative capacity estimates*

If we naively assume an absence of epistasis and that all escape mutations are equally beneficial to the virus, then we would expect to see escape rates negatively correlated with *in vitro* replicative capacity, and reversion rates positively correlated with *in vitro* replicative capacity.

Results are displayed in Figure 5.11. We find a negative correlation between time to escape and replicative capacity for the Spearman rank correlation coefficients ($p < 0.02$). Correlation measured by the Pearson product moment correlation coefficient was not significant, perhaps due to a non-linear relationship between replicative capacity and escape rate. We also observe a positive correlation between time to reversion and replicative capacity, though this association does not reach significance under either of the two measures of correlation considered. These findings again suggest that epistasis does not play a strong role in dictating escape rates, and add weight to the results of Section 5.3.2.1: replicative capacity of a given escape variant was determined for the variant applied to a lab strain background, thus any epistatic effects of pathways to escape were not captured in this assay.

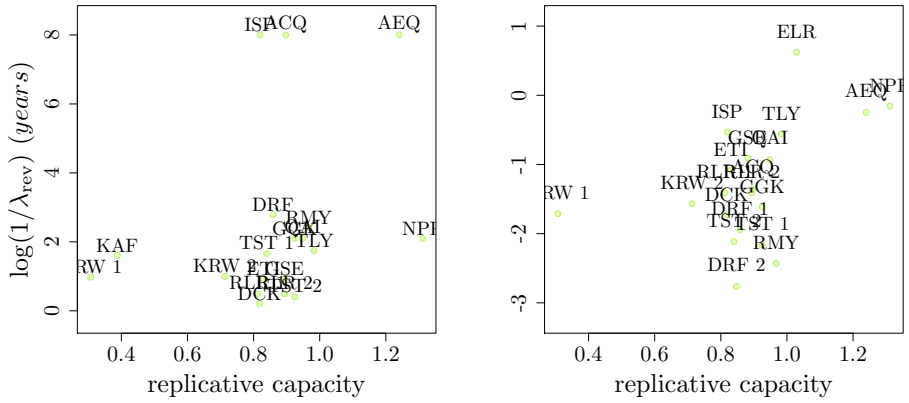
5.3.3.2 *Escape and reversion rate estimates in the Bloemfontein dataset vs.* in vitro *replicative capacity estimates*

Given the correlation observed between *in vitro* replicative capacity and escape rate in two datasets taken from populations where subtype B is the predominant HIV-1 subtype, we sought to determine if this correlation was replicated in the Bloemfontein dataset where subtype C predominates. We do not find any such correlation between estimated time to escape/reversion in Bloemfontein (and the full African dataset using the naive ODE method, see Figure 5.13) and *in vitro* replicative capacity.



(a) $r = -0.069, p = 0.78.$
$\rho = -0.54, p = 0.020.$

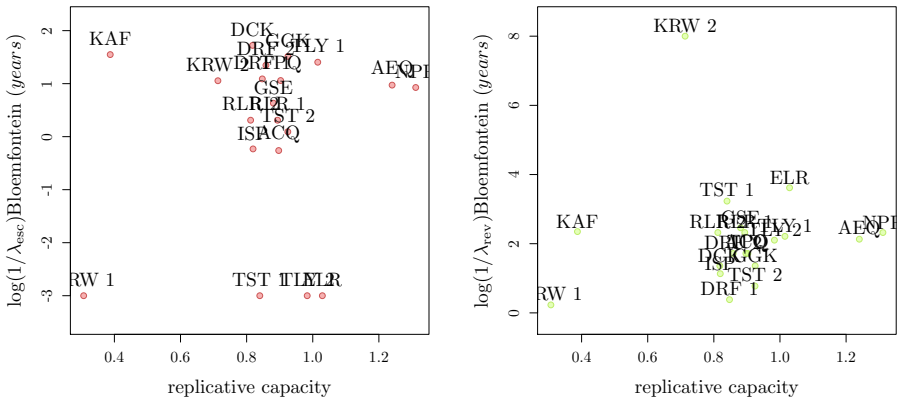
(b) $r = -0.18, p = 0.43.$
$\rho = -0.55, p = 0.012.$



(c) $r = 0.24, p = 0.32.$
$\rho = 0.44, p = 0.061.$

(d) $r = 0.30, p = 0.20.$
$\rho = 0.42, p = 0.069.$

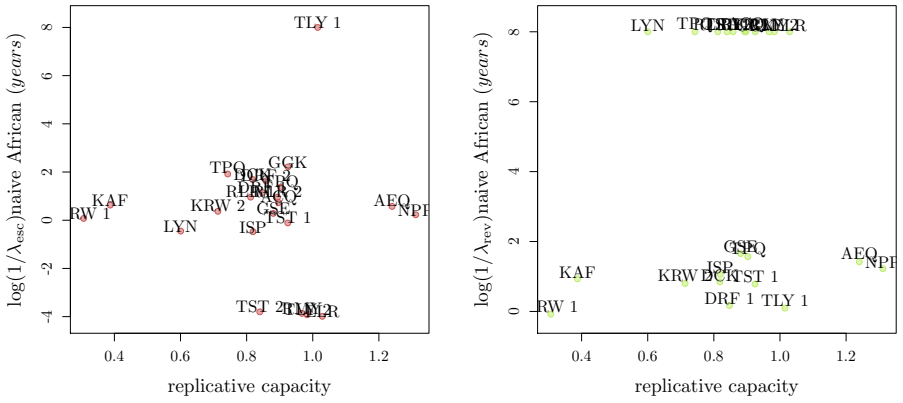
Figure 5.11: Replicative capacity plotted against the MAP estimates of time to escape is shown in Figures 5.11a-5.11b, and the MAP estimates of time to reversion in Figures 5.11c-5.11d. The first and second columns display estimates obtained using *in vivo* data from Switzerland (SSITT) and Canada (HOMER) respectively. Epitopes are labelled by the first three amino acids. Numbers after the three letter abbreviation of each epitope refers to the ordering in Tables 5.2-5.4, as some epitopes have more than one escape mutation and/or associated restricting HLA. Time to escape is measured in *years* - except in the case of the Canadian cohort in which it is measured in units of substitutions. Spearman rank correlation coefficients; ρ, with associated p-values are displayed on each plot.



(a) $r = -0.14, p = 0.55$.
$\rho = -0.073, p = 0.76$.

(b) $r = -0.16, p = 0.49$.
$\rho = 0.080, p = 0.74$.

Figure 5.12: Scatterplots of time to escape and reversion estimates against *in vitro* viral replicative capacity: Bloemfontein. Replicative capacity plotted against the MAP estimates of time to escape is shown in Figure 5.12a, and the MAP estimates of time to reversion in Figure 5.12b. Epitopes are labelled by the first three amino acids. Numbers after the three letter abbreviation of each epitope refers to the ordering in Tables 5.2-5.4, as some epitopes have more than one escape mutation and/or associated restricting HLA. Spearman rank correlation coefficients; ρ, and Pearson product moment correlation coefficients; r, with associated p-values are displayed on each plot.



(a) $r = -0.022, p = 0.92$.
$\rho = -0.10, p = 0.64$.

(b) $r = -0.092, p = 0.67$.
$\rho = 0.15, p = 0.48$.

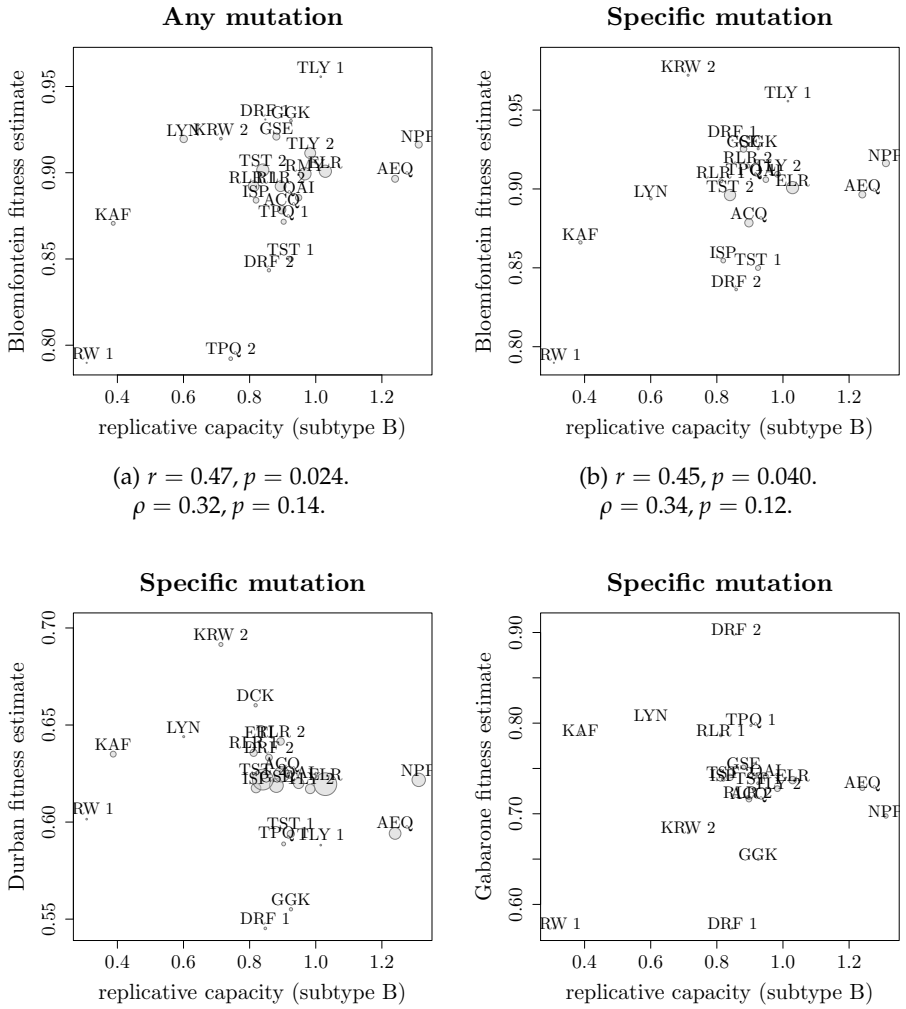
Figure 5.13: Correlation between time to escape/reversion estimated from all African data using naive ODE approach and replicative capacity as measured by *in vitro* fitness assays. Figure 5.13a shows a scatterplot of the respective escape rate estimates. Figure 5.13b shows the corresponding plot for reversion rate estimates. Epitopes are labelled by the first three amino acids. Numbers after the three letter abbreviation of each epitope refers to the ordering in Tables 5.2-5.4, as some epitopes have more than one escape mutation and/or associated restricting HLA. Spearman rank correlation coefficients and Pearson product-moment correlation coefficients with associated p values are shown.

5.3.4 *Association between* in vitro *estimates of replicative capacity*

So far we have used *in vitro* replicative capacity of escape mutations applied to an NL4-3 recombinant by site directed mutagenesis as a proxy for viral fitness *in vivo*. However, replication capacity experiments have also been carried out on viral samples taken from hosts [30] (for more detail, see Section 5.2). There are clear benefits and drawbacks to each approximation. The first method targets the mutation of interest, but does not take into consideration the possibility that other mutations may be required for a given escape mutation to take place (at a potentially lower fitness cost). However, the second is subject to large noise effects due to the high variability of HIV-1 sequences across hosts. Thus, a large number of assays are required to obtain a reasonable estimate if the effect of a particular escape mutation on replicative capacity is small. We determine whether there is any correlation between the two measures by considering results of replicative assays applied to a subset of the Bloemfontein [122], Durban [250] and Gabarone [138] patient samples. For the Bloemfontein fitness assay data we found an association between the two estimates (Figures 5.14c-5.14d). However, this correlation was not present when using the Durban or Gabarone fitness data (Figures 5.14c and 5.14d).

5.4 DISCUSSION

Perhaps surprisingly given the simplifying assumptions of our model and the use of replicative capacity assays to measure viral fitness, we find a negative correlation between replicative capacity and estimated time to escape in the two majority subtype B populations - Switzerland (SSITT) and Canada (HOMER). Similarly for the estimated time to reversion, though not significant, we observe a trend towards a positive correlation between estimated fitness cost *in vitro* and *in vivo* estimates of time until reversion using our integrated method. For the Bloemfontein dataset, no association was found for either escape or reversion estimates. This lack of correlation may be driven by two major factors.



(a) $r = 0.47, p = 0.024$.
$\rho = 0.32, p = 0.14$.

(b) $r = 0.45, p = 0.040$.
$\rho = 0.34, p = 0.12$.

(c) $r = -0.24, p = 0.28$.
$\rho = -0.51, p = 0.015$.

(d) $r = -0.092, p = 0.70$.
$\rho = -0.17, p = 0.47$.

Figure 5.14: Correlation between the average fitness of strains from African replicative capacity datasets and and the fitness cost associated to the given amino acid point mutation. the Bloemfontein dataset [122] containing a particular escape mutation, and the fitness cost associated to the given amino acid point mutation [30]. Figures 5.14a and 5.14b displays the estimated fitness of strains taken from the Bloemfontein dataset [123]. In Figure 5.14a we plot the average fitness of viral populations taken from hosts which have a non 'wildtype' amino acid at the escape site on the y-axis. Epitopes are labelled by the first three amino acids. Numbers after the three letter abbreviation of each epitope refers to the ordering in Tables 5.2-5.4, as some epitopes have more than one escape mutation and/or associated restricting HLA. In Figure 5.14b, the specific escape variant in [30] must be present. Figures 5.14c and 5.14d show the corresponding plot for the Durban [250] and Gabarone [138] datasets respectively. The area of of each dot corresponds to the number of sequences used to generate the average that the dot represents. Pearson product-moment correlation coefficients (r) with associated p-values are displayed.

- Due to the difference in HLA distributions in African and European populations, many of the escape estimates for the Bloemfontein dataset are calculated using few data points
- Point mutations were applied to a subtype B (an NL4-3 recombinant) strain, so it is possible that the complex nature of the HIV-1 fitness landscape would obscure any potential correlation (the predominant subtype in South Africa (and in Bloemfontein) is subtype C). Indeed, distinct HLA associated escape mutations have been observed in African populations, sometimes to subtype B consensus [19].

Interestingly, we do not observe fast escape rates for two of the putative 'negatopes' (a negatope is defined as an epitope for which the escape variant has reached consensus in the population, putative negatopes are labelled by asterisks in Tables 5.2-5.4) put forward in [30, 299, 32] in the African datasets (both the Bloemfontein dataset using our integrated method, or the full African dataset using the ODE method). One possible explanation is that those mutations resulted as a consequence of founder effects in North American and European populations. In the case of escape it appears that epistasis is overpowered by point mutation effects; if it weren't then no such correlation between *in vitro* replicative capacity and escape rate would be apparent. One possible explanation for the lack of correlation between replicative capacity and reversion rate (for which *a priori* we would expect to see the strongest correlation) is that epistasis plays a stronger role in the reversion process. Intuitively, this makes sense: on average a viral lineage spends a longer period of time within HLA mismatched hosts (because HLA diversity is so high). Selection to increase replicative capacity acts across the entirety of the viral genome both in the presence and absence of any HLA associated immune pressure (contrast this to CTL associated selective pressure, which is strongest in and around epitopes). The high dimensionality of sequence space would then lead us to expect multiple pathways to a restoration of viral fitness, not all of which may be associated to a simple reversion of the purported escape mutation. These conclusions are backed up by the correlations between estimates of time to escape, and lack of correlation between estimates of time to reversion that we see across populations. However, other explanations are available.

The lack of correlation between reversion rate estimates and *in vitro* replication capacity experiments will also be due in part to the lack of power to estimate reversion rates. Reliable reversion rate estimates require multiple instances of transmission of escape mutations in the population able to revert in an HLA mismatched host, an event which occurs less often than transmission of wildtype at the site(s) under consideration, particularly if the restricting HLA is rare, or the escape rate is relatively low.

We find no correlation between two *in vitro* estimates used to approximate *in vivo* viral fitness. This lack of correlation may be due to episasis and thus highlight the need for assays applied to viral sequences taken from hosts to assess the average fitness of escape strains, it may reflect the high amount of between host viral fitness variation, or simply display the high variability of estimates such *in vitro* assays provide. Either way, we emphasise the importance of multiple tests before drawing conclusions from the fitness data such as those shown in Figures 5.14c-5.14d.

The results of this chapter are summarised in Figure 5.15. With enough data, escape rates are consistent across three populations in three different continents and two viral subtypes. The Bloemfontein dataset does not conform to this pattern, possibly because it is too small. For subtype B viruses (SSITT and HOMER cohorts), *in vivo* escape rate estimates were negatively correlated with *in vitro* fitness assays. None of these patterns are apparent for reversion rates, possibly due to epistatic effects and compensatory mutations. However, we find a negative correlation between *in vivo* escape rate estimates and replicative capacity suggesting that escape variants that incur little or no fitness cost are selected rapidly in a host with the relevent restricting HLA. These observations suggest that there is often only one (or very few) route to escape, but a great many which restore viral fitness within host.

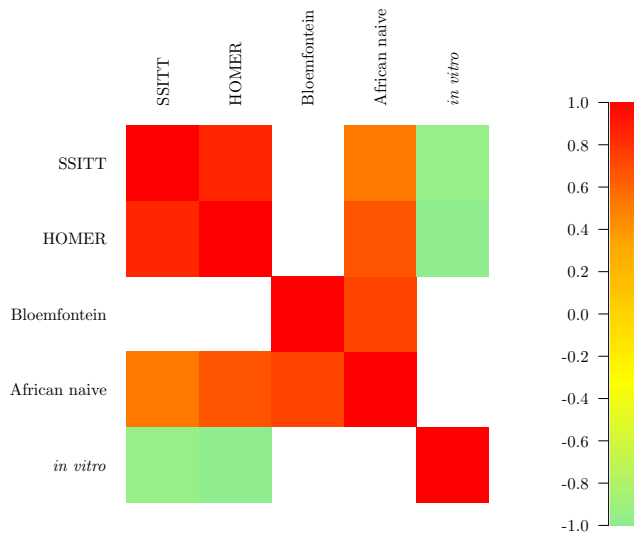


Figure 5.15: Summary of all associations between expected time to escape and *in vitro* replicative capacity, according to the Spearman rank correlation coefficient. Significant associations (p<0.05) are coloured according to the Colombo, non-significant associations are shown in white.

Epitope	HLA	HLA prop	with escape HLA +	with escape HLA −	t to escape ODE	t to escape MAP	t to reversion ODE	t to reversion MAP	fitness rank	fitness actual
G'SKKKYKLK	B*08	0.143	0/5	3/33	152.9	46.06	$> 10^8$	126.7	8	0.9250
RLRPGGKKK	A*03	0.224	6/10	4/28	2.296	2.765	2.841	3.073	20	0.8121
RLRPGGKKK	A*03	0.224	6/10	4/28	2.296	2.765	2.841	3.073	12	0.8935
JEWELRY'SY	A*01	0.262	5/10	8/28	5.653	5.523	8.065	8.291	13	0.8808
LYNTVATL	Cw*14	not available	–	–	–	–	–	–	23	0.6006
ELRSLYNTV*	B*08	0.143	5/5	33/33	–	–	–	–	3	1.029
[A]TLYCVHQR	A*11	0.116	0/3	0/35	–	–	–	–	4	1.0151
TLYCVHQR	A*11	0.116	3/3	22/35	0.0024	0.0015	33.75	55.55	5	0.9832
[A]ISPRTLNAW	B*57	0.057	5/7	16/31	1.481	0.7076	$> 10^8$	$> 10^8$	18	0.8194
QAISPRTLNAW	A*25	0.042	1/1	16/37	0.0015	*0.001*	90.15	126.7	7	0.9482
KAFSPEVIPMF	B*57	0.057	0/8	6/47	82.87	11.52	$> 10^8$	40.46	24	0.3874
TPQDLNTML	B*42	0.003	0/0	0/55	–	–	–	–	10	0.9029
TPQDLNTML	Cw*08	not available	–	–	–	–	–	–	21	0.7428
ETINEEAAEW	A*25	0.042	3/8	11/47	5.815	4.162	11.90	8.569	17	0.8298
TSTLQEQIGW	B*57	0.057	8/9	4/46	0.3539	0.3462	2.608	2.544	9	0.9245
TSTLQEQIGW	B*57	0.057	4/8	12/47	4.643	6.738	164.0	45.93	16	0.8398
NPPIPVGEIY*	B*35	0.196	6/6	44/49	0.0059	*0.001*	126.1	126.7	1	1.3114
KRWIILGLNK	B*27	0.073	3/8	3/47	5.003	6.679	6.500	9.413	25	0.3071

Table 5.2: Fitness costs of epitopes in Gag. MAP and ODE *in vivo* estimates of escape and reversion rates for the collection of epitopes in [30], obtained using Swiss based subtype B SSITT cohort data.

KRWIILGLNK	B*27	0.073	5/8	7/47	1.961	1.902	9.999	9.914	22	0.7131
RMYSPTSI	B*52	0.042	2/3	28/52	1.428	0.8149	$> 10^8$	261.7	6	0.9673
DRFYKTLRA	B*14	0.068	0/6	0/49	–	–	–	–	15	0.8474
DRFYKTLRA	B*14	0.068	1/6	1/49	13.07	102.0	4.389	618.0	14	0.8584
AEQASQEVKNW*	B*44	0.211	11/13	26/42	0.0001	*0.001*	$> 10^8$	$> 10^8$	2	1.240
DCKTILKAL	B*08	0.143	2/7	1/48	7.266	8.448	1.271	1.595	19	0.8181
ACQGVGGPGHK	A*11	0.116	4/7	11/48	3.147	10.02	13.22	$> 10^8$	11	0.8970

Table 5.2: (continued) Fitness costs of epitopes in Gag. MAP and ODE *in vivo* estimates of escape and reversion rates for the collection of epitopes in [30], obtained using Swiss based subtype B SSITT cohort data. HLA proportions are taken from the HLA factsbook [178]. HLA +ve hosts are those with the restricting HLA type. Time to escape and reversion are estimated using the model developed in [210]. Estimates in italic lie on the boundary of our strict prior of $[10^{-3}, 10^8] \times [10^{-3}, 10^8]$. The lack of an estimate is due to lack of knowledge of HLA type (individuals were typed at HLA A and B only), or if none/all individuals displayed the escape mutation. Putative negatopes are labelled with asterisks.

Epitope	HLA	HLA prop	with escape HLA +	with escape HLA −	t to escape ODE	t to escape MAP	t to reversion ODE	t to reversion MAP	fitness rank	fitness actual
GGKKKYKLK	B*08	0.094	1/13	10/185	89.03	32.68	$> 10^8$	22.87	8	0.9250
RLRPGGKKK	A*03	0.125	11/18	118/180	5.368	2.044	$> 10^8$	210.3	20	0.8121
RLRPGGKKK	A*03	0.125	11/18	118/180	5.368	2.044	$> 10^8$	210.3	12	0.8935
GSEELRSLY	A*01	0.095	12/18	65/180	3.401	4.286	46.75	288.8	13	0.8808
LYNTVATL	Cw*14	not available	–	–	–	–	–	–	23	0.6006
ELRSLYNTV*	B*08	0.094	13/13	185/185	–	–	–	–	3	1.029
[A]TLYCVHQR	A*11	0.029	0/1	8/197	$> 10^8$	25.37	1.383	163.2	4	1.0151
TLYCVHQR	A*11	0.029	1/1	157/197	*0.001*	*0.001*	$> 10^8$	126.7	5	0.9832
[A]ISPRTLNAW	B*57	0.078	8/9	41/189	0.5124	0.5874	11.86	13.58	18	0.8194
QAISPRTLNAW	A*25	0.007	0/0	54/198	–	–	–	–	7	0.9482
KAFSPEVIPMF	B*57	0.078	1/9	15/189	52.06	35.33	$> 10^8$	224.1	24	0.3874
TPQDLNTML	B*42	0.097	19/45	23/153	6.697	11.45	16.94	52.14	10	0.9029
TPQDLNTML	Cw*08	not available	–	–	–	–	–	–	21	0.7428
ETINEEAAEW	A*25	0.007	0/0	2/198	–	–	–	–	17	0.8298
TSTLQEQIGW	B*57	0.078	7/9	19/189	1.016	1.237	5.435	5.956	9	0.9245
TSTLQEQIGW	B*57	0.078	9/9	187/189	*0.001*	*0.001*	$> 10^8$	1704	16	0.8398
NPPIPVGEIY*	B*35	0.108	1/2	59/196	7.758	8.464	66.56	210.3	1	1.3114
KRWIILGLNK	B*27	0.029	1/1	4/197	*0.001*	*0.001*	2.041	1.699	25	0.3071

Table 5.3: Fitness costs of epitopes in Gag. MAP and ODE *in vivo* estimates of escape and reversion rates for the collection of epitopes in [30], obtained using sequence data from a cohort based in Bloemfontein, South Africa.

KRWIIL̲GLNK	B*27	0.029	1/1	13/197	*0.001*	11.35	6.960	$> 10^8$	22	0.7131
RMYSPT̲SI	B*52	0.033	0/0	197/198	–	–	–	–	6	0.9673
DRFYK̲TLRA	B*14	0.068	4/19	2/179	12.14	12.29	2.357	2.412	15	0.8474
DRFYKT̲LRA	B*14	0.068	3/19	12/179	27.73	22.22	46.46	52.14	14	0.8584
AEQASQE̲VKNW*	B*44	0.112	17/33	60/165	9.770	9.386	$> 10^8$	134.9	2	1.240
DCK̲TILKAL	B*08	0.094	0/13	8/185	$> 10^8$	52.70	0.8225	22.87	19	0.8181
ACQGVGGPG̲HK	A*11	0.029	0/1	95/197	24.51	0.5475	$> 10^8$	52.14	11	0.8970

Table 5.3: (...continued) Fitness costs of epitopes in Gag. MAP and ODE *in vivo* estimates of escape and reversion rates for the collection of epitopes in [30], obtained using sequence data from a cohort based in Bloemfontein, South Africa. HLA proportions are taken from the HLA factsbook [178]. HLA +ve hosts are those with the restricting HLA type. Time to escape and reversion are estimated using the model developed in [210]. Estimates in italic lie on the boundary of our strict prior of $[10^{-3}, 10^8] \times [10^{-3}, 10^8]$. The lack of an estimate is due to lack of knowledge of HLA type (individuals were typed at HLA A and B only), or if none/all individuals displayed the escape mutation. Putative negatopes are labelled with asterisks.

Epitope	HLA	HLA prop	with escape HLA +	with escape HLA −	t to escape ODE	t to escape MAP	t to reversion ODE	t to reversion MAP	fitness rank	fitness actual
GGKKKYKLK	B*08	0.143	2/21	9/140	0.1666	0.05665	$> 10^6$	0.02428	8	0.9250
RLRPGGKKK	A*03	0.224	26/35	28/124	0.004668	0.01734	0.01422	0.03897	20	0.8121
RLRPGGKKK	A*03	0.224	26/35	28/124	0.004668	0.01734	0.01422	0.03897	12	0.8935
GSEELRSLY	A*01	0.262	18/38	30/122	0.02217	0.03385	0.03297	0.1245	13	0.8808
LYNTVATL	Cw*14	not available	–	–	–	–	–	–	23	0.6006
ELRSLYNTV*	B*08	0.143	23/23	149/152	*0.000001*	0.00001017	$> 10^6$	4.176	3	1.029
[A]TLYCVHQR	A*11	0.116	0/23	0/153	–	–	–	–	4	1.0151
TLYCVHQR	A*11	0.116	16/19	72/131	0.004255	0.001675	9.082	0.2758	5	0.9832
[A]ISPRTLNAW	B*57	0.057	1/6	35/151	0.04080	0.006306	$> 10^6$	0.2947	18	0.8194
QAISPRTLNAW	A*25	0.042	3/5	28/161	0.007034	0.002664	0.2264	0.1165	7	0.9482
KAFSPEVIPMF	B*57	0.057	0/10	0/163	–	–	–	–	24	0.3874
TPQDLNTML	B*42	0.003	0/0	1/174	–	–	–	–	10	0.9029
TPQDLNTML	Cw*08	not available	–	–	–	–	–	–	21	0.7428
ETINEEAAEW	A*25	0.042	0/6	7/171	0.9160	0.06004	$> 10^6$	0.08364	17	0.8298
TSTLQEQIGW	B*57	0.057	7/8	7/165	0.001160	0.001217	0.006512	0.006811	9	0.9245
TSTLQEQIGW	B*57	0.057	5/8	29/157	0.006512	0.007693	0.08293	0.007693	16	0.8398
NPPIPVGEIY*	B*35	0.196	23/23	139/147	*0.000001*	*0.000001*	$> 10^6$	0.6975	1	1.3114
KRWIILGLNK	B*27	0.073	5/14	9/164	0.01669	0.01493	0.01844	0.01946	25	0.3071

Table 5.4: Fitness costs of epitopes in Gag. MAP and ODE *in vivo* estimates of escape and reversion rates for the collection of epitopes in [30], obtained using sequence data from the HOMER cohort based in British Columbia, Canada.

KRWIILGLNK	B*27	0.073	8/10	18/148	0.002316	0.003711	0.01804	0.02710	22	0.7131
RMYSPTSI	B*52	0.042	1/2	30/163	0.01067	0.003711	$> 10^6$	0.003711	6	0.9673
DRFYKTLRA	B*14	0.068	5/12	1/158	0.01111	0.01138	0.001652	0.001725	15	0.8474
DRFYKTLRA	B*14	0.068	1/14	2/159	0.1226	0.1144	0.02431	0.01143	14	0.8584
AEQASQEVKNW*	B*44	0.211	25/26	97/126	0.001812	0.004744	0.2703	0.5626	2	1.240
DCKTILKAL	B*08	0.143	1/21	5/139	0.3343	0.1211	$> 10^6$	0.01873	19	0.8181
ACQGVGGPGHK	A*11	0.116	7/21	26/142	0.03341	0.1610	0.5879	0.04437	11	0.8970

Table 5.4: (...continued) Fitness costs of epitopes in Gag. MAP and ODE *in vivo* estimates of escape and reversion rates for the collection of epitopes in [30], obtained using sequence data from the HOMER cohort based in British Columbia, Canada. HLA proportions are taken from the HLA factsbook [178]. HLA +ve hosts are those with the restricting HLA type. Time to escape and reversion are estimated using the model developed in [210]. Estimates in italic lie on the boundary of our strict prior of $[10^{-5}, 10^6] \times [10^{-5}, 10^6]$. the escape mutation. Putative negatopes are labelled with asterisks.

6

HYPOTHESIS FREE ESTIMATION OF HLA ASSOCIATED SELECTION IN THE PRESENCE OF RECOMBINATION USING LARGE DATASETS: THE GREAT ESCAPE

6.1 INTRODUCTION

In Chapters 3, 4 and 5 we have shown how integrating phylogenetic and epidemiological inference enables quantitative analysis of immune driven viral evolution that gives rise to detectable interactions between host HLA and viral sequence [211]. This work provided a proof of principle that cross-sectional HIV-1 sequence data in the context of host genotype provides information about the strength of such interactions. Our integrated method revealed the strength of using phylogenetic information, but also the inherent limitations of existing approaches. Notably, the computational demands limit performing MCMC on the space of trees limits cross-sectional sample sizes to hundreds of viral sequences. Our integrated method and other models in the literature [36, 85] require both viral sequence data and associated host HLA information. However, far larger HIV-1 sequence datasets exist without this extra host HLA information. If we can make use of this sequence data as well, we could expect to shorten terminal branches and so increase our power to detect HLA associated selection. To use the model presented in Chapters 3 - 5 we required prior knowledge of the location of epitopes and the associated restricting HLA class I molecule. We wish to identify and estimate HLA associated selection without this prior knowledge: hypothesis-free estimation of HLA associated selection in this sense. Finally, current methods often ignore recombination between HIV-1 genomes of the same subtype, which can lead to false signals of recurrent mutation [12, 260]. Put simply, if hypothesis-free discovery of meaningful host HLA-HIV interactions is to

be powerful, new statistical and computational methods need to be developed that can cope with datasets consisting of many thousands of samples, and account for recombination. There are national programmes in place for monitoring drug responses through the collection of detailed HIV-1 and patient data. Because of this, datasets consisting of viral sequences taken from tens of thousands of individuals are already available.

Examples of publicly available HIV-1 viral sequence data resources (of varying quality and with associated meta-data) include:

- **Los Alamos HIV sequence database [256]**

  `www.hiv.lanl.gov`. This database contains ∼4,000 complete genome sequences, which increases at the gene level to ∼14,000 *gag*, ∼6,000 *pol*, and ∼15,000 *nef* sequences. Corresponding numbers of sequences for protease and a 200 base pair region of reverse transcriptase are ∼130,000 and ∼160,000 respectively. Information on patient HLA information is available for $\sim 2,500$ *gag*, ∼1,500 *pol* and ∼3,000 *nef* sequence samples.

- **Stanford HIV drug resistance database [63]**

  `hivdb.stanford.edu`. The database currently contains ∼60,000 protease sequences, ∼60,000 reverse transcriptase sequences and ∼6,000 integrase sequences (the sequences coding for these enzymes are found in *pol*). A wealth of meta-data is publicly available for these sequences including geography, subtype and drug regimen (though not host HLA profile).

- **UCLA HIV positive selection mutation database [47, 212, 46]**

  `fold.doe – mbi.ucla.edu/HIV/`. This database contains ∼50,000 partial *pol* sequences. It was originally created to help identify drug resistance and positive fitness mutations in protease and reverse transcriptase.

There are also privately accessible HIV sequence data resources, including the UK HIV drug resistance database [55] and the Monogram Biosciences sequence dataset [24, 114]. Both require an approved project proposal before they can be accessed. Given the abundance of sequence data available in the public databases, we restrict our analysis in this chapter and Chapter 7 to these datasets.

In this chapter, we describe a new statistical method for identifying significant

associations between host HLA type and viral sequence evolution that accounts for recombination. We do so in a hypothesis-free manner that can make use of both sequence datasets with associated host HLA information *and* very large datasets (>100,000 sequences) without associated host HLA data. The chapter is split into the following sections:

- **Section 6.2: Description of the model**. We provide detailed descriptions of how we model recombination, and lay out our model of codon substitution which accounts for HLA associated selection. We describe how we combine existing sequence data for which the host HLA types are known, and far larger sequence datasets for which host HLA information is not present. Our methods are based on the work of Wilson and McVean [304], which we extend to incorporate HLA associated selection.

- **Section 6.3: Inference**. We use Markov chain Monte-Carlo to sample from our posterior distribution of parameters. We perform MCMC to sample:

  1. Recombination probabilities between each site.
  2. The non-synonymous/synonymous substitution ratio in the absence of HLA dependent selection at each site.
  3. Parameters relating to HLA associated selection away from the consensus viral amino acid sequence.
  4. HLA independent parameters relating to reversion back towards the consensus viral amino acid sequence.

  We describe the collection of MCMC moves used to explore the space of parameters defined in Section 6.2.

- **Section 6.4: Computer memory constraints**. Memory usage becomes an issue when dealing with very large viral sequence datasets. We describe some methods to selectively filter out portions of the sequence data in such a way that the majority of the likelihood is explained whilst maintaining computational tractability.

- **Section 6.5: Simulation studies**. We simulate sequence and HLA datasets under our model, and then re-estimate the parameters used to simulate

them. Using simulated data, we determine how our MCMC scheme performs under different recombination rates. We also perform a simple test to check that the method we described in Section 6.4 to reduce computational complexity is reliable.

- **Section 6.6: Discussion and motivations for future work**. We describe our findings and explore future applications. In Chapter 7 we apply the model to real data.

Our modelling framework is easily extended and lends itself well to the addition of further data. For example, longitudinal viral sequence data and viral sequences obtained from hosts in known transmission chains would be rich data to analyse.

Although our model is built specifically to address the problem of describing HLA associated selection upon viral sequence, we may consider a variety of applications in which we wish to probe the effect of a defined host property upon pathogen sequence. As an example of this, and a test of the model, we use the framework to estimate drug dependent selection on protease sequences taken from the Stanford drug resistance database in the Chapter 7. Throughout this chapter we refer to HLA associated selection, though this may be considered a synonym for 'property of the host' when used for other applications.

6.2 DESCRIPTION OF THE MODEL

6.2.1 *Summary*

We first summarise our model and assumptions as represented diagrammatically in Figure 6.1. Some HIV-1equence data has associated host HLA information; we call this the 'query' dataset $\mathbf{D}$. Much more HIV-1equence data is available without associated host HLA information; we call this the 'reference' dataset $\mathbf{D}_B$. Currently $\mathbf{D}_B$ is not used to inform HLA dependent selection. However, if we can combine $\mathbf{D}$ and $\mathbf{D}_B$ we will be able to obtain information about the recent sequence history of $\mathbf{D}$, and use this powerful extra information in parameter

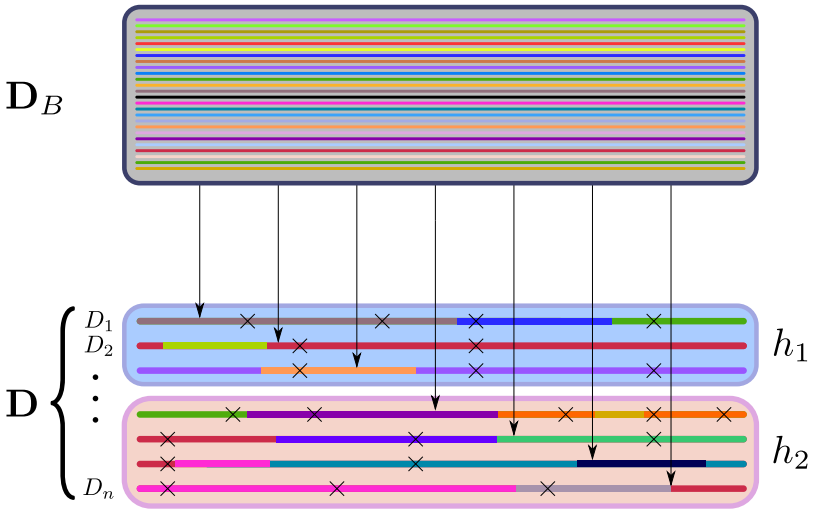


Figure 6.1: Diagram of the model. A smaller dataset $\mathbf{D}$ of sequences with associated host HLA information is generated from a large dataset $\mathbf{D}_B$ (without associated host HLA information). As $|\mathbf{D}_B| \gg 0$, we assume $\mathbf{D}_B$ represents the distribution of all possible sequences an individual can be infected with, and so members of $\mathbf{D}$ can be considered independent (shown by the arrows). Thus, members of $\mathbf{D}$ arise from $\mathbf{D}_B$ through the twin processes of recombination and mutation. The time until coalescence between any member of $\mathbf{D}$ and the reference sequence set $\mathbf{D}_B$ is small (since $\mathbf{D}_B$ is very large). We assume all selection along a lineage connecting each member of $\mathbf{D}$ to its closest neighbour in $\mathbf{D}_B$ (which can change as we track along the sequence) occurred within the host that the member of $\mathbf{D}$ was isolated from. Thus only HLA information for $\mathbf{D}$'s hosts are required. In this simple example there are two distinct HLA profiles h_1 and h_2 (three hosts have HLA profile h_1 and four hosts have HLA profile h_2), shown by the light blue and light pink backgrounds. Recombination in $\mathbf{D}_B$ results in the colourings of the sequences, and mutations are shown as black crosses.

estimates. As $|\mathbf{D}_B| \gg 0$, we assume $\mathbf{D}_B$ represents the distribution of all possible sequences an individual can be infected with. If recombination occurs in the population, then members of $\mathbf{D}$ arise from $\mathbf{D}_B$ through the twin processes of recombination and mutation. The time until coalescence between any member of $\mathbf{D}$ and the reference sequence set $\mathbf{D}_B$ is small since $\mathbf{D}_B$ is so large. Consider the lineage connecting a member of $\mathbf{D}$ to its nearest neighbour in $\mathbf{D}_B$. If we assume all selection along this lineage occurred within the member of $\mathbf{D}$, then only HLA information for $\mathbf{D}$ is required to write down a model of sequence evolution which incorporates HLA dependent selection.

6.2.2 *Filling in the detail*

In order to estimate the likelihood of the parameters Θ governing an evolutionary model given a collection of sequence data $\mathbf{D}$, it is important to account for the non-independence of samples by considering the underlying genealogy G. For our purposes, we are not interested in estimating G. We therefore treat the genealogy as a nuisance parameter and integrate over it.

$$P(\mathbf{D}|\Theta) = \int_{\Theta} P(\mathbf{D}|\Theta, G)P(G)dG. \tag{119}$$

Here, $P(G)$ is our prior on the genealogy G. $P(\mathbf{D}|\Theta, G)$ is the likelihood of the data under some parameter values Θ and genealogy G, which can be evaluated using Felsenstein's peeling algorithm [76]. Since the space of genealogies is very large, evaluating this integral is problematic. However, without recombination it is possible to estimate using a numerical summation approximation. We do precisely this in Chapters 3 - 5 using BEAST [65], but our analysis is limited to relatively small numbers of sequences. In the presence of recombination G becomes an ancestral recombination graph and so the space of G becomes far larger, making it impossible to use this approach in all but the simplest of models for very small sequence datasets.

The solution that we use is to estimate this integral directly using an approximation to the coalescent with recombination [167]. The model we present has two components:

1. An approximation of the coalescent with recombination.
2. A model of codon substitution: we first write down a baseline model of codon substitution, and then modify it to incorporate HLA associated selection.

By combining these two components, we are able to approximate the probability of a sequence emerging in the presence of HLA dependent selection and recombination. Throughout this chapter, we make the assumption that each viral sequence is from a distinct individual. In practice, when applying the model

to real data in Chapter 7, we perform a number of data filtering steps to ensure this assumption is reasonable.

6.2.3 *Part 1: Approximating the coalescent with recombination*

Recombination is known to occur frequently in HIV-1. During reverse transcription, recombination between two single-stranded RNA genomes occurs when switching between the two viral templates takes place. This 'strand switching' can occur up to 20 times per replication cycle [201]. Despite this, recombination is often ignored in HIV analysis e.g. [158, 125, 165, 269, 96, 56] backed up by the claims that co-infection is rare, and that a single founder virus is responsible for the viral population within a host [234, 140]. The majority viral subtype in a host population is analysed and circulating recombinant forms (CRFs) are removed from datasets. A clear example of recombination at the population level are CRFs, in which recombination between two or more subtypes has taken place and suggests that inter-subtype super-infection rates may be high in some populations. CRFs are the most easily identifiable recombinants due to the amount of divergence between the subtypes. Co-infection or multiple infection by viruses of the same subtype will necessarily outnumber co-infection by two distinct subtypes in regions where one viral subtype predominates. However, intra-subtype recombination is more difficult to detect and is less well studied as a consequence. In the few studies which attempt to investigate intra-subtype recombination there is disagreement. In some studies, estimated rates of between host recombination are high. For example, 84% of infections in female bar workers in Tanzania [150] underwent intra-subtype recombination. In another study intra-subtype recombination was detected in $\sim$47% of subtype C infections [249], a further study estimated rates of super-infection by subtype B viruses in subtype B infected hosts to be up to 15% of the infection rate [277]. By contrast, other investigations suggest that super-infection is rare [44, 97]. Chakraborty *et al.* [44] propose that chronic HIV-1 infection may protect against further infection by another strain. Lemey *et al.* [161] suggest that high between host recombination rate estimates

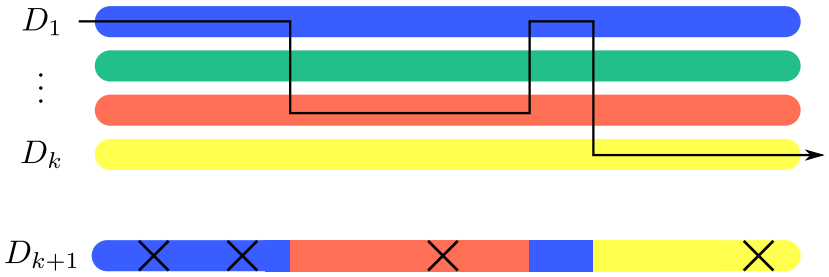


Figure 6.2: The PAC likelihood. Adapted from Li and Stephens [167], $\hat{\pi}(D_{k+1}|D_1, D_2, \ldots, D_k, \Theta)$ captures the probability that D_{k+1} has arisen as an imperfect mosaic of $D_1, D_2, \ldots D_k$ (in the example, the true path is shown by the arrow). Xs denote errors in copying the path followed by the arrow.

may be due to modelling assumptions and the impact of small networks of hosts with high super-infection rates (an observation acknowledged by Taylor and Korber [277]). Even if between host recombination rates are low on average, the existence of intra and inter-subtype recombinants suggests that recombination should not be discounted in analyses.

We use the Li and Stephens [167] approximation, as it captures many of the properties of the coalescent with recombination in a computationally efficient manner. In this section, we carefully describe the details of the approximation and then explain how we apply it to our problem.

Let Θ denote a collection of parameters governing codon substitution and let $\mathbf{D} = \{D_1, D_2, \ldots, D_n\}$ denote a collection of sequences. The approximation follows from observing that the likelihood $P(D_1, D_2, \ldots, D_n|\Theta)$ may be written as a product of conditional likelihoods, each of which can then approximated by a function which we denote $\hat{\pi}$. I.e.:

$$P(D_1, D_2, \ldots, D_n|\Theta) = P(D_1|\Theta)P(D_2|D_1, \Theta)\ldots P(D_n|D_1, D_2, \ldots, D_{n-1}, \Theta) \tag{120}$$

$$\approx \hat{\pi}(D_1|\Theta)\hat{\pi}(D_2|D_1, \Theta)\ldots\hat{\pi}(D_n|D_1, D_2, \ldots, D_{n-1}, \Theta). \tag{121}$$

The right hand side of (121) is known as the product of approximate conditionals (PAC) likelihood as each conditional likelihood $P(D_{k+1}|D_1, D_2, \ldots, D_k, \Theta)$ is approximated using the function $\hat{\pi}$. Li and Stephens use a hidden Markov model (HMM) for $\hat{\pi}$ that captures a collection of desirable properties of the true likeli-

hood in a manner which is computationally tractable, namely that the $(k+1)^{\text{th}}$ sequence is an imperfect mosaic of the first k sequences due to mutation and recombination. For the $(k+1)^{\text{th}}$ sequence, the hidden states are the first k sequences. We denote $X_1, X_2, \ldots, X_m$ as the hidden Markov chain that emits the sequence D_{k+1}, where X_j is the sequence being copied from or 'SCF' at position j and m is the length of the sequence. Between each site, the SCF switches to a new sequence l with probability $q_{x,i}(j)$ (which may depend on the current SCF; x, the new SCF; i, and the current site; j). Given a hidden state at position j, the probability of observing the codon in the observed sequence D_{k+1} is generated by a model of codon substitution (we describe our model which incorporates HLA dependent selection in Section 6.2.4.1). To generate $\hat{\pi}(D_{k+1}|D_1, D_2, \ldots, D_k, \Theta)$ we integrate over all possible paths through sequences $D_1, D_2, \ldots, D_k$ which result in the sequence D_{k+1} through a combination of jumping between these k sequences (recombination) and error in copying (mutation), see Figure 6.2.

As this approximation to the process generating sequence D_k is an HMM, we can appeal to the forwards and backwards algorithms to evaluate $\hat{\pi}(D_{k+1}|D_1, D_2, \ldots, D_k, \Theta)$ for each k. The forwards algorithm proceeds as follows:

Let $D_{k+1}[1:j]$ denote the codons at the first j sites of sequence $k+1$ and let $f_{i,j}$ denote the probability of observing all codons in D_{k+1} up to site $j-1$ and copying the codon in position j from sequence i. Then $f_{i,j} = P(D_{k+1}[1:j], X_j = i|D_1, D_2, \ldots, D_k, \Theta)$. Let the probability of copying the codon at position j from sequence i be denoted $\varepsilon_{i,j}$. This is the probability of an observed state given a hidden state, and is known as an emission probability in the language of HMMs. Then $f_{i,1}$ is just an emission probability and easily computed if we have a uniform prior on the SCF at the first position. We then calculate $f_{i,j}$ for $j \geq 2$ recursively as follows:

$$f_{i,j} = \varepsilon_{i,j} \sum_{x=1}^{k} f_{x,j-1} q_{x,i}(j), \tag{122}$$

where as defined earlier, $q_{x,i}(j)$ is the probability of recombination and the destination sequence being sequence i. Note that we make the implicit assumption that recombination events may only occur between codons and not within them.

If we assume that all sequences are equally likely to recombine (as in [167] and [304]) then (122) may be simplified to

$$f_{i,j} = \varepsilon_{i,j}\left((1-r_j)f_{i,j-1} + \frac{r_j}{k}\sum_{x=1}^{k} f_{x,j-1}\right) \tag{123}$$

where r_j is the probability of recombination between the $(j-1)^{\text{th}}$ and j^{th} position.

Then, for the $(k+1)^{\text{th}}$ sequence,

$$\hat{\pi}(D_{k+1}|D_1, D_2, \ldots, D_k, \Theta) = \sum_{x=1}^{k} f_{x,m}. \tag{124}$$

We may similarly run the backwards algorithm, starting at the last site and evaluating iteratively towards the first site. Let $\mathbf{D}_1^k = \{D_1, D_2, \ldots, D_k\}$. $b_{i,j}$ is constructed such that

$$f_{i,j}b_{i,j} = P(X_j = i, D_{k+1}|\mathbf{D}_1^k, \Theta) \tag{125}$$

$$\Rightarrow \quad P(D_{k+1}[1:j], X_j = i|\mathbf{D}_1^k, \Theta)b_{i,j} = P(X_j = i, D_{k+1}|\mathbf{D}_1^k, \Theta) \tag{126}$$

$$\Rightarrow \quad b_{i,j} = P(D_{k+1}[j+1:m]|X_j = i, D_{k+1}[1:j], \mathbf{D}_1^k, \Theta) \tag{127}$$

$$\Rightarrow \quad b_{i,j} = P(D_{k+1}[j+1:m]|X_j = i, \mathbf{D}_1^k, \Theta) \tag{128}$$

where the last implication is a consequence of the Markov property. The $b_{i,j}$s for the $(k+1)^{\text{th}}$ sequence can then be generated iteratively as follows

$$b_{i,m} = 1 \tag{129}$$

$$b_{i,j} = \sum_{x=1}^{k} b_{x,j+1}\varepsilon_{x,j+1}q_{x,i}(j) \tag{130}$$

$$b_{i,j} = \varepsilon_{i,j+1}(1-r_j)b_{i,j+1} + \frac{r_{j+1}}{k}\sum_{x=1}^{k} b_{x,j+1}\varepsilon_{x,j+1}. \tag{131}$$

Hence we can calculate

$$\sum_{i=1}^{k} f_{i,j} b_{i,j} = \sum_{i=1}^{k} P(X_j = i, D_{k+1} | \mathbf{D}_1^k, \Theta) \tag{132}$$

$$= \hat{\pi}(D_{k+1} | \mathbf{D}_1^k, \Theta). \tag{133}$$

Using the forwards-backwards algorithm also allows us to evaluate the marginal distributions if required:

$$P(X_j = i | D_{k+1}, \mathbf{D}_1^k, \Theta) = \frac{P(X_j = i, D_{k+1} | \mathbf{D}_1^k, \Theta)}{P(D_{k+1} | \mathbf{D}_1^k, \Theta)} = \frac{f_{i,j} b_{i,j}}{\sum_{i=1}^{k} f_{i,j} b_{i,j}} = \frac{f_{i,j} b_{i,j}}{\hat{\pi}(D_{k+1} | \mathbf{D}_1^k, \Theta)}. \tag{134}$$

Note that the approximation of $P(D_1, D_2, \ldots, D_n | \Theta)$ using this HMM is dependent upon the ordering of $\{D_1, D_2, \ldots, D_k\}$. This is often resolved by randomly permuting the ordering some fixed number of times and determining an average [304], or by letting the order be an unknown parameter [52].

In our inference problem we have two datasets. One 'query' dataset $\mathbf{D}$ for which we have associated host HLA information and another larger 'reference' dataset $\mathbf{D}_B$ without host HLA information. Recall from Section 6.2.1 that we make two important assumptions:

1. All selection along the lineages connecting a member of $\mathbf{D}$ to $\mathbf{D}_B$ occurred within the host of the member of $\mathbf{D}$.
2. The reference dataset $\mathbf{D}_B$ is a good approximation of the distribution of viruses an individual can be infected with (which is reasonable when $|\mathbf{D}_B| \gg 0$).

These two assumptions allow us to neatly apply the Li and Stephens approximation to our inference problem. In the HMM we require emission probabilities. For our model these will be probabilities of codon substitutions in the presence of some HLA profile. Making assumption 1 allows us to approximate $P(\mathbf{D} | \mathbf{D}_B, \Theta)$. This is because we have HLA information for the members of

$\mathbf{D} = \{D_1, D_2, \ldots, D_n\}$ which we can use to determine the relevant emission probabilities. I.e. The first assumption allows us to use the sequence data in $\mathbf{D}_B$:

$$P(\mathbf{D}|\mathbf{D}_B, \Theta) = P(D_1|\mathbf{D}_B, \Theta)P(D_2|\mathbf{D}_B, D_1, \Theta)\ldots P(D_n|\mathbf{D}_B, D_1, D_2\ldots, D_{n-1}, \Theta). \tag{135}$$

Using assumption 2, we may write

$$P(\mathbf{D}|\mathbf{D}_B, \Theta) \approx P(D_1|\mathbf{D}_B, \Theta)P(D_2|\mathbf{D}_B, \Theta)\ldots P(D_n|\mathbf{D}_B, \Theta) \tag{136}$$

$$\approx \hat{\pi}(D_1|\mathbf{D}_B, \Theta)\hat{\pi}(D_2|\mathbf{D}_B, \Theta)\ldots \hat{\pi}(D_n|\mathbf{D}_B, \Theta). \tag{137}$$

In other words, assumption 2 results in the approximation that $\mathbf{D}$ is generated by independent realisations of recombination and mutation through $\mathbf{D}_B$. This avoids any need for averaging over orderings of $\mathbf{D}$ allowing rapid evaluation of the approximate likelihood.

6.2.4 *Part 2: Emission probabilities - A codon model of substitution*

In the methods of our integrated model used in Chapter 3 we used a nucleotide substitution model to describe sequence evolution. This was due to the computational restrictions imposed by performing MCMC on genealogies. A more realistic model of sequence evolution is a codon model. Importantly for our purposes, a codon model allows us to detect selection through non-synonymous nucleotide changes. At each site we are interested in detecting if these types of nucleotide changes away from some 'wildtype' sequence are enriched in a particular HLA background. We first write down the Nielsen and Yang (NY98) codon model (without HLA associated selection) [199] and then extend it to incorporate the HLA types of the host transmitted to (a given member of $\mathbf{D}$). This will allow us to infer HLA dependent selection along the viral sequence. Note that in the next two sections we consider codon substitutions at a single site (and drop any site subscript). However, when we perform inference we will allow our codon model to be parameterised by distinct collections of parameters at each site.

6.2.4.1 *Codon model of substitution: NY98 codon model*

In the NY98 codon model, the mutation rate from codon i to codon j is

$$q_{i,j} = \mu \begin{cases} \kappa & \text{if } i \text{ and } j \text{ differ by a synonymous transition} \\ 1 & \text{if } i \text{ and } j \text{ differ by a synonymous transversion} \\ \omega & \text{if } i \text{ and } j \text{ differ by a non-synonymous transition} \\ \kappa\omega & \text{if } i \text{ and } j \text{ differ by a non-synonymous transversion} \\ 0 & \text{otherwise} \end{cases} \tag{138}$$

where κ is the relative rate of transitions to transversions and ω is the ratio of non-synonymous base changes to synonymous base changes assuming equal codon usage. The original model [199] weighted by the frequency of the codon switched to at that position, to ensure reversibility. This is not a requirement for us. In fact, the process we want to model is not reversible, as we will include a boost in movement away from the consensus codon in the presence of a given HLA, and assume a distinct boost independent of HLA (to model reversion) towards consensus. Finally μ is the scaled mutation rate parameter for the rate of synonymous transversions measured in units of N generations (where N is the effective population size).

The exact probability of a substitution from codon i to codon j in time t cannot be solved analytically, though we may estimate it by numerically evaluating the matrix exponential of the 64×64 instantaneous rate matrix described by Equations (138). This is computationally expensive so we look for an approximation.

If the expected time until the first coalescence of a given lineage with all other sampled lineages t, is small (reasonable if the number of sampled sequences in $\mathbf{D}_B$ is very large), we can assume that the vast majority of codon changes occur through single base changes (as the probability of observing more than one nucleotide change in a codon is low). This allows us to avoid the matrix expo-

nentiation and reduce computational cost. Consider a switch from codon C_1 to codon C_2. The total rate of substitution out of codon C_1 is

$$(\kappa\beta_S + \beta_V + \omega(\kappa\alpha_S + \alpha_V))\,\mu =: \Lambda\mu \tag{139}$$

where β_S, β_V, α_S and α_V are the counts of the number of synonymous transitions, synonymous transversions, non-synonymous transitions, and non-synonymous transversions from C_1 respectively.
If the overall movement away from a particular codon C_1 is $\Lambda\mu$, then the probability a change has occurred in time t is

$$P(\text{leave } C_1|\Theta, t) = \int_0^t \Lambda\mu \exp(-\Lambda\mu t)\, dt = 1 - \exp(-\Lambda\mu t). \tag{140}$$

Here, Θ is the collection of parameters governing codon substitution. We can make an approximation by splitting up $P(C_1 \to C_2|\Theta, t)$. We consider two classes of moves from $C_1 \to C_2$: those which result from a single nucleotide change in time t (e.g. `AAA` $\to$ `AAG`) and those that result from multiple nucleotide changes in time t (e.g. `AAA` $\to$ `ATA` $\to$ `AAA` $\to$ `AAG`). For $C_1 \neq C_2$,

$$P(C_1 \to C_2|\Theta, t) = \sum_{i=1}^{\infty} P(C_1 \to C_2|i \text{ changes}, t) P(i \text{ changes}|\Theta, t) \tag{141}$$

$$= P(\text{leave } C_1|\Theta, t) \sum_{i=1}^{\infty} P(C_1 \to C_2|\text{leave } C_1, i \text{ changes}, t) P(i \text{ changes}|\Theta, t) \tag{142}$$

$$\approx (1 - \exp(-\Lambda\mu t))\,(P(C_1 \to C_2|\text{leave } C_1, 1 \text{ change}, t)\phi + P(C_1 \to C_2|\text{leave } C_1, > 1 \text{ change}, t)\,(1 - \phi)) \tag{143}$$

$$\approx (1 - \exp(-\Lambda\mu t))\,(\phi P(C_1 \to C_2|\text{leave } C_1, 1 \text{ change}, t) + (1 - \phi)\,\pi_{C_2}) \tag{144}$$

where ϕ is the probability of undergoing a 1 step change from C_1 to C_2 and π_{C_2} is the probability of observing codon C_2 at that position in the sequence.

Thus, for each codon C_1, we generate the following collection of probabilities for state changes at each site:

1 step non-synonymous transition

$$P(C_1 \to C_2|\Theta, t) = (1 - \exp(-\Lambda\mu t)) \left(\phi\frac{\kappa\omega}{\Lambda} + (1-\phi)\pi_{C_2}\right) \tag{145}$$

1 step synonymous transition

$$P(C_1 \to C_2|\Theta, t) = (1 - \exp(-\Lambda\mu t)) \left(\phi\frac{\kappa}{\Lambda} + (1-\phi)\pi_{C_2}\right) \tag{146}$$

1 step non-synonymous transversion

$$P(C_1 \to C_2|\Theta, t) = (1 - \exp(-\Lambda\mu t)) \left(\phi\frac{\omega}{\Lambda} + (1-\phi)\pi_{C_2}\right) \tag{147}$$

1 step synonymous transversion

$$P(C_1 \to C_2|\Theta, t) = (1 - \exp(-\Lambda\mu t)) \left(\phi\frac{1}{\Lambda} + (1-\phi)\pi_{C_2}\right) \tag{148}$$

>1 step change

$$P(C_1 \to C_2|\Theta, t) = (1 - \exp(-\Lambda\mu t)) \left((1-\phi)\pi_{C_2}\right) \tag{149}$$

if $C_1 \neq C_2$, and

No change in codon at the site $1 - ((145) + \ldots + (149)) = (150)$.

$$P(C_1 \to C_2|\Theta, t) = \exp(-\Lambda\mu t) + (1 - \exp(-\Lambda\mu t))\,(1-\phi)\pi_{C_2} \tag{150}$$

if $C_1 = C_2$.

We now check to ensure the sum over all possible substitutions is 1.

$$\sum_{C_2} P(C_1 \to C_2|\Theta, t) = P(C_1 \to C_1|\Theta, t) + \sum_{C_2 \neq C_1} P(C_1 \to C_2|\Theta, t). \tag{151}$$

$$\sum_{C_2 \neq C_1} P(C_1 \to C_2|\Theta, t) = (1 - \exp(-\Lambda\mu t)) \times \left(\alpha_S \phi \frac{\kappa\omega}{\Lambda} + \beta_S \phi \frac{\kappa}{\Lambda} + \alpha_V \phi \frac{\omega}{\Lambda} + \beta_V \phi \frac{1}{\Lambda} + \sum_{C_2 \neq C_1} (1 - \phi)\, \pi_{C_2} \right) \tag{152}$$

$$\Rightarrow \sum_{C_2} P(C_1 \to C_2|\Theta, t) = (1 - \exp(-\Lambda\mu t))\,(\phi + (1 - \pi_{C_1})\,(1 - \phi)) + P(C_1 \to C_1|\Theta, t) \tag{153}$$

$$= (1 - \exp(-\Lambda\mu t))\,(\phi + (1 - \pi_{C_1})\,(1 - \phi)) + \exp(-\Lambda\mu t) + (1 - \exp(-\Lambda\mu t))\,(1 - \phi)\,\pi_{C_1} \tag{154}$$

$$= 1 \tag{155}$$

as required.

We estimate π_C (for each codon C) and ϕ empirically using the Durban dataset [122, 145, 146] described in Section 7.3.2. The next step is to incorporate HLA associated selection into our codon model.

6.2.4.2 *Codon model of substitution: adding HLA associated selection*

In order to account for HLA associated selection we may choose to incorporate transmission and HLA proportions in the same way as the model previously described in Chapters 3 and 4 [85, 210]. Alternatively, we can consider a simpler model in which we are only concerned with the HLA type (h, say) of the sampled host in the smaller population. We explore the latter, as previously indicated in Section 6.2.1. For movement away from the consensus codon $\mathscr{C}$ at a given site, Λ is modified to Λ' in the presence of HLA type h:

$$\Lambda' = \kappa\beta_S + \beta_V + \omega\,(\kappa\alpha_S + \alpha_V)\,\gamma_h. \tag{156}$$

So, as an example, for a non-synonymous transition from the consensus (or 'wild-type'[1]) codon $\mathscr{C}$ to C_2, explicitly writing down the HLA dependence and using the approximation from 6.2.4.1,

$$P(\mathscr{C} \to C_2|h,t) = \left(1-\exp(-\Lambda'\mu t)\right)\left(\phi\frac{\kappa\omega\gamma_h}{\Lambda'} + (1-\phi)\pi_{\mathscr{C}}\right), \tag{157}$$

γ_h is a scaling of the non-synonymous/synonymous substitution ratio ω associated with HLA type h, away from $\mathscr{C}$. This weighting provides extra selection away from the consensus codon $\mathscr{C}$ if $\gamma_h > 1$ and less if $\gamma_h < 1$ to create a new total rate of movement out of the consensus codon $\mathscr{C}$ in the presence of HLA type h, Λ'. Thus an association between a particular HLA type and non-synonymous change from consensus at the site would result in $\gamma_h > 1$. Notice that no selection is applied to synonymous codon changes.

Extending to some HLA profile of a host $\mathbf{H} = \{h_1, h_2, \ldots, h_n\}$:

$$P(\mathscr{C} \to C_2 \mid \mathbf{H}, \Theta, t) = \left(1-\exp\left(-\Lambda'\mu t\right)\right)\left(\phi\frac{\kappa\omega\prod_{h\in\mathbf{H}}\gamma_h}{\Lambda'} + (1-\phi)\pi_{\mathscr{C}}\right), \tag{158}$$

We also wish to incorporate reversion into our codon model. Given the state of the codon under consideration, there is some set of nucleotide substitutions resulting in $\mathscr{C}$ at that position. It is these substitutions which we wish to scale through reversion parameters. We let ζ denote the scaling of selection due to reversion at a particular site. We assume that each ζ is independent of the host's HLA profile, $\mathbf{H}$. By doing so, we are assuming that in the absence of HLA pressure there is a selection towards the consensus codon $\mathscr{C}$ at each site.

Consider the case in which a non-synonymous transition results in the consensus codon, $\mathscr{C}$, from some codon C_1. Then Λ is modified to Λ':

$$\Lambda' = \kappa\beta_S + \beta_V + \omega\left(\kappa\left(\zeta + (\alpha_S - 1)\right) + \alpha_V\right). \tag{159}$$

1 The consensus codon $\mathscr{C}$ is predefined. In practice, we set it as either the consensus subtype B, or subtype C codon at the position.

Reversion is therefore modelled by

$$P(C_1 \to \mathscr{C} \mid \mathbf{H}, \Theta, t) = \left(1 - \exp(-\Lambda' \mu t)\right) \left(\phi \frac{\kappa \omega \zeta}{\Lambda'} + (1 - \phi)\pi_{\mathscr{C}}\right). \tag{160}$$

ζ is a scaling of the non-synonymous/synonymous substitution ratio ω towards $\mathscr{C}$. This weighting provides extra selection towards the consensus codon $\mathscr{C}$ if $\zeta > 1$ and less if $\zeta < 1$. We assume that the same ζ scales all the potential 1 step non-synonymous base changes to $\mathscr{C}$.

We have now created a codon model which accounts for reversion and HLA associated selection within a host, and can now determine emission probabilities $\varepsilon_{i,j}$. Recall that $\varepsilon_{i,j}$ is the probability of copying the codon at position j from sequence i. Therefore, by evaluating

$$P(C_1 \to C_2 | \Theta) = \int_0^\infty P(C_1 \to C_2 | t, \Theta) P(t) dt. \tag{161}$$

we can substitute in the relevant codon change ($C_1 \to C_2$) and determine $\varepsilon_{i,j}$. We integrate out t assuming the coalescent with fixed population size. In the standard coalescent with a fixed population size, the distribution of the divergence time between a sequence and its closest relative from among k other sequences may be approximated by an exponential distribution with rate k ($P(t) = k \exp(-kt)$) [304]. See Appendix A.1 for a derivation of the mean.

This is simply a case of substituting in and integrating. For example, for a non-synonymous transition from the consensus,

$$P(\mathscr{C} \to C_2 \mid \mathbf{H}, \Theta) = \left(\phi \frac{\kappa \omega \prod_{h \in \mathbf{H}} \gamma_h}{\Lambda'} + (1 - \phi)\pi_{C_2}\right) \left(1 - \frac{k}{k + \Lambda' \mu}\right). \tag{162}$$

Similarly, for a non-synonymous transition towards the consensus,

$$P(C_1 \to \mathscr{C} \mid \mathbf{H}, \Theta) = \left(\phi \frac{\kappa \omega \zeta}{\Lambda'} + (1 - \phi)\pi_{\mathscr{C}}\right) \left(1 - \frac{k}{k + \Lambda' \mu}\right). \tag{163}$$

Notice that in evaluating this integral we are focused on sequence evolution between a member of $\mathbf{D}$ and its nearest neighbour in $\mathbf{D_B}$. Only considering nearest neighbours in the genealogy has two clear advantages:

1. We only use the parts of the genealogy that contain the most information about HLA dependent selection within a sampled host: terminal branches.
2. The resulting reduction in computation time allows us to use far more sequence data. Using more sequence data will shorten terminal branches and so increase our power to detect HLA associated selection.

Using very large datasets is important for our method. As terminal branches shorten through increased sampling, assumption 1 (that selection occurs in the member of $\mathbf{D}$ along the lineage connecting each member of $\mathbf{D}$ to its nearest neighbour in $\mathbf{D}_B$) becomes more reasonable as we may assume that no or very few transmissions occur along the lineage connecting sequences in $\mathbf{D}$ to their nearest neighbour in $\mathbf{D}_B$.

In Equations (156) - (163) we considered a single host with an HLA type h or collection of HLA types $\mathbf{H}$. We can trivially extend this to obtain the desired approximation of the likelihood given in Equation (136). Let $\mathcal{H}$ be the collection of HLA profiles of the associated hosts for sequences in $\mathbf{D}$. Let $\mathbf{H}_i$ denote the HLA profile of host i. Then

$$P(\mathbf{D}|\mathbf{D}_B, \mathcal{H}, \Theta) \approx \prod_i \hat{\pi}(D_i|\mathbf{D}_B, \mathbf{H}_i, \Theta). \tag{164}$$

We have now created a model which accounts for recombination using the Li and Stephens [167] approximation and incorporates both reversion and HLA associated selection.

6.3 INFERENCE

The structure of the model we have presented lends itself very well to inference. We use MCMC to sample from the posterior. For our MCMC implementation, we use the Metropolis Hastings algorithm. We fix κ, ϕ and π_C for each codon C at empirical estimates. We perform MCMC moves on the following collection of parameters:

- The recombination probability r_j of switching the SCF between neighbouring sites j and $j+1$.
- The non-synonymous/synonymous substitution ratio in the absence of reversion or HLA associated selection at each site, ω_j.
- Scalings of selection due to reversion at each site, ζ_j.
- HLA associated scalings of selection at each site, $\gamma_{j,h}$.

In our sampling scheme we allow the ζ_js, and $\gamma_{j,h}$s to vary via piecewise constant functions across the codon sequence which we call selection windows. This is analagous to the block model used by Wilson and McVean [304]. Imposing this window structure allows information about HLA dependent selection (and reversion) to be combined across sites. This makes sense in a biological setting where sites close together in the coding sequence may result in amino-acids lying in the same epitope or close together in the protein structure. The window model will act to smooth across sites, but will be overwhelmed if a given site is under strong HLA dependent selection [304]. How strict the window model is depends on the expected number of windows, which is controlled by the parameter p_w (see the Merge and Split moves and Wilson and McVean [304]). If $p_w = 1$, the number of selection windows is equal to the number of sites, if $p_w = 0$ then there is just one selection window.

Using the backwards algorithm in combination with the forwards algorithm is useful for our inference regime. We can use $\sum_i f_{i,j} b_{i,j} = \hat{\pi}(D_k|\mathbf{D}_B, \mathbf{H}_k, \Theta)$ to quickly re-evaluate likelihood after altering a parameter which acts at a site (or window of sites). This is precisely the information required to determine if an MCMC move is accepted. Only after a move is accepted do we need to determine the entirety of $f_{i,j}$ and $b_{i,j}$. However, we note that there is a trade off between acceptance rate and computation time here.

6.3.1 *MCMC moves*

Within the MCMC, we have three main classes of move. Moves to alter a parameter at a single site or window of sites, moves to merge or split windows, and moves to expand and contract selection windows to push them around the genome. In Section 6.2.4.1 we considered separate parameters to scale selection away from consensus in the presence of each host HLA type at each site (ζ_j), and parameters to model reversion towards consensus at each site ($\gamma_{j,h}$). In order to increase speed we alter these parameters within selection windows in our MCMC scheme. The selection windows that we discuss are analogous to the selection windows and recombination windows described in [304]. We now introduce some notation:

Let $\mathcal{S} := \{s_1, s_2, \ldots, s_S\}$ be a collection of sites that split windows, and let $|\mathcal{S}| = S$. We may then alter parameters occurring within selection windows $[s_k, s_{k+1})$, $s_k \in \mathcal{S} \cup \{1, l\}$ (defining $s_0 = 1$, $s_{S+1} = l$). The number of selection windows is $S + 1$, and parameters are constant across these windows.

For recombination and selection, the parameters r_j and ω_j are estimated between each pair of neighbouring sites $(j-1, j)$, and at each site j respectively. For HLA associated selection and reversion parameter changes, moves are applied to a window of these parameters. Note, for each HLA type h we assign a distinct collection of selection windows.

6.3.1.1 *Selection parameters ω_j*

We only use a single type of MCMC move to alter ω_j across the codon sequence: changing ω_j at a given site j. We assume an exponential prior with parameter λ_ω for ω_j. Following the existing model [304], a new value ω_j' is chosen such

that $\omega_j' = \omega_j \exp(U)$, where $U \sim \text{Uniform}(-1,1)$ and the acceptance probability is

$$\alpha_\omega(\Theta \to \Theta') = \min\left\{1, \frac{L(\Theta'|\mathbf{D})}{L(\Theta|\mathbf{D})}\frac{P(\omega_j')}{P(\omega_j)}\frac{q(\omega_j' \to \omega_j)}{q(\omega_j \to \omega_j')}\right\} \tag{165}$$

$$\Rightarrow \alpha_\omega(\Theta \to \Theta') = \min\left\{1, \frac{P(\mathbf{D}|\Theta')}{P(\mathbf{D}|\Theta)}\exp\left(-\lambda_\omega\left(\omega_j' - \omega_j\right)\right)\frac{\omega_j'}{\omega_j}\right\}. \tag{166}$$

Where $L(\Theta|\mathbf{D})$ is the likelihood of Θ given the data, $P(x)$ is the prior on x, and $q(x \to x')$ is the proposal distribution. The ratio of proposals, $\frac{q(x' \to x)}{q(x \to x')}$ is known as the Hastings ratio. Note that we may have generated the Hastings ratio $\frac{\omega_j'}{\omega_j}$ by considering $u, u' \sim U(-1,1)$, and letting the mapping $h(\omega_j, u) = \omega_j \exp(u)$ determine a potential new value ω_j'. We need to consider moving back to ω_j via this mapping.

$$\text{Hastings ratio} = \frac{q(\omega_j' \to \omega_j)}{q(\omega_j \to \omega_j')} = \frac{g'(u')}{g(u)}|J|, \tag{167}$$

Equation (167) is known as Green's method. g and g' are densities to sample u and u' from, and h is a deterministic function mapping (u, ω_j) to ω_j'. J is the Jacobian of ω_j and u with respect to ω_j' and u'. Noting that $u = -u'$ (since $\omega_j' = \omega_j \exp(u)$, $\omega_j = \omega_j' \exp(u') \Rightarrow \exp(u) = 1/\exp(u') \Rightarrow u = -u'$), and that $g, g' = U(-1,1)$,

$$\frac{q(\omega_j' \to \omega_j)}{q(\omega_j \to \omega_j')} = \frac{1/2}{1/2}\begin{vmatrix} \frac{\partial \omega_j'}{\partial \omega_j} & \frac{\partial \omega_j'}{\partial u} \\ \frac{\partial u'}{\partial \omega_j} & \frac{\partial u'}{\partial u} \end{vmatrix} \tag{168}$$

$$= \beta^2 \begin{vmatrix} \exp(u) & \omega_j \exp(u) \\ 0 & -1 \end{vmatrix} \tag{169}$$

$$= \exp(u) \tag{170}$$

$$= \frac{\omega_j'}{\omega_j}. \tag{171}$$

The MCMC move to alter μ is identical.

6.3.1.2 *Recombination Probabilities r_j*

Assuming independence between sites, and a distinct recombination rate at each site, we must vary r_j for $j \in \{2, 3, \ldots, m\}$, where m is the number of codon sites in the sequence. We do this by perturbing the recombination rate at a given site in exactly the same way as for the selection parameters (ω) through a shift on the log scale. The recombination rate is related to the probability of recombination through the relation

$$r_j = 1 - \exp(-\nu_j / k), \tag{172}$$

where ν_j is the recombination rate between sites $j-1$ and j, and k is the number of sequences. We allow moves to alter the recombination rate in the same way as the selection parameters, ω_j. A new value ν' is chosen such that $\nu' = \nu \exp(U)$, where $U \sim \text{Uniform}(-1, 1)$. We have an exponential prior for ν with rate parameter λ_ν. The acceptance probability is identical to that of the selection parameters, with ω_j and λ_ω replaced with ν_j and λ_ν respectively.

6.3.2 *Window Moves*

For this section we consider only escape parameters (moves associated to reversion parameters are completely analogous) and for notational convenience we let $\gamma_{j,h} = \gamma_j$, and allow j to now enumerate windows rather than sites.

1. **Changes of selection parameters**

   Moves are proposed in the same way as Section 6.3.1, but across a window rather than at a specific site. The window is chosen uniformly at random from the current collection of windows. We use a log-normal distribution, $\log \mathcal{N} \sim (0, \sigma_\gamma^2)$ for the prior on HLA dependent scaling of selection parameters, and the resulting acceptance ratio is

$$\alpha_{w_1}(\Theta \to \Theta') = \min\left\{ \frac{P(\mathbf{D}|\Theta')}{P(\mathbf{D}|\Theta)} \exp\left(\frac{\log^2 \gamma_j - \log^2 \gamma_j'}{2\sigma_\gamma^2} \right) \frac{\gamma_j'}{\gamma_j} \right\}. \tag{173}$$

2. **Extension of a window in the $3'$ or $5'$ direction**

   Choose a block to extend uniformly at random, with $5'$ or $3'$ direction chosen with equal probability. Choose the number of sites to extend $g \in [1, \infty)$ from a geometric distribution. Any proposal that results in a block being merged is rejected, as is the proposal to extend the $5'$-most or $3'$-most block in the $5'$ or $3'$ direction respectively. Otherwise, the acceptance probability is

$$\alpha_{w_2}(\Theta \to \Theta') = \min\left\{1, \frac{P(\mathbf{D}|\Theta')}{P(\mathbf{D}|\Theta)}\right\}. \tag{174}$$

3. **Split a window**

   These final two moves are complementary reversible jump MCMC moves. A site s^* is chosen uniformly from the collection of available splitting sites, and with probability 1 splits an existing window, $[s_j, s_{j+1})$ say. New HLA dependent selection parameters within these new selection windows $[s_j, s^*)$ and $[s^*, s_{j+1})$ are defined as γ'_{j_1} and γ'_{j_2} where

$$\gamma'_{j_1}{}^{(s^*-s_j)}\gamma'_{j_2}{}^{(s_{j+1}-s^*)} = \gamma_j^{(s_{j+1}-s_j)}, \tag{175}$$

$$\frac{\gamma'_{j_2}}{\gamma'_{j_1}} = \frac{1-u_1}{u_1}; \quad u_1 \sim U[0,1]. \tag{176}$$

   It then follows that the acceptance ratio is

$$\alpha_{w_3}(\Theta \to \Theta') = \min\left\{1, \frac{P(\mathbf{D}|\Theta')}{P(\mathbf{D}|\Theta)} \frac{p_w}{1-p_w} \frac{\gamma_j}{\gamma'_{j_1}\gamma'_{j_2}\sqrt{2\pi}\sigma_\gamma} \times \exp\left(\frac{\log^2\gamma_j - \log^2\gamma'_{j_1} - \log^2\gamma'_{j_2}}{2\sigma_\gamma^2}\right) \frac{d_{S+1}}{c_S} \frac{m-S-1}{S+1} \frac{\left(\gamma'_{j_1}+\gamma'_{j_2}\right)^2}{\gamma_j}\right\} \tag{177}$$

   Steps to determine the acceptance ratio for the split move are shown in Section 6.3.2.1. Steps to evaluate the merge move acceptance ratio are similar and are omitted.

4. **Merge a window**. Pick a splitting site $s_{j+1},\ j \in \{0, 1, \ldots, S-1\}$ at random to remove, resulting in a merge of windows $[s_j, s_{j+1})$ and $[s_{j+1}, s_{j+2})$. Let

$$\gamma'^{(s_{j+2}-s_j)}_j = \gamma_j^{(s_{j+1}-s_j)} \gamma_{j+1}^{(s_{j+2}-s_{j+1})}, \tag{178}$$

and accept with probability

$$\alpha_{w_4}\left(\Theta' \to \Theta\right) = \min\left\{1, \frac{P(\mathbf{D}|\Theta')}{P(\mathbf{D}|\Theta)} \frac{(1-p_w)}{p_w} \frac{\gamma_j \gamma_{j+1} \sqrt{2\pi}\sigma_\gamma}{\gamma'_j} \times \exp\left(\frac{\log^2 \gamma_j + \log^2 \gamma_{j+1} - \log^2 \gamma'_j}{2\sigma_\gamma^2}\right) \frac{c_{S-1}}{d_S} \frac{S}{m-S} \frac{\gamma'_j}{\left(\gamma_j + \gamma_{j+1}\right)^2}\right\}. \tag{179}$$

6.3.2.1 *Determining the acceptance ratio of a window split move*

Following Green [106],

$$\alpha\left(\Theta \to \Theta'\right) = \min\left\{1, \text{likelihood ratio} \times \text{prior ratio} \times \text{proposal ratio} \times \left|\det\left(\text{Jacobian}\right)\right|\right\}. \tag{180}$$

To determine the prior ratio, let the prior on the number of window breakpoints $S \sim Bin\left(p_w, l-1\right)$ where l is the number of sites. Note that the ordering here is required, so we drop the binomial coefficient. Each selection coefficient $\gamma_j \sim \log \mathcal{N}(0, \sigma_\gamma^2)$. Thus the prior ratio is

$$\begin{aligned} \text{prior ratio} =& \frac{p_w^{S+1}\left(1-p_w\right)^{l-1-(S+1)}}{p_w^S\left(1-p_w\right)^{l-1-S}} \\ & \times \frac{\frac{1}{\gamma'_{j_1}\sqrt{2\pi}\sigma_\gamma} \frac{1}{\gamma'_{j_1}\sqrt{2\pi}\sigma_\gamma} \exp\left(-\frac{\log^2 \gamma'_{j_1} + \log^2 \gamma'_{j_2}}{2\sigma_\gamma^2}\right)}{\frac{1}{\gamma_j\sqrt{2\pi}\sigma_\gamma} \exp\left(-\frac{\log^2 \gamma_j}{2\sigma_\gamma^2}\right)} \qquad (181) \\ =& \frac{p_w}{1-p_w} \frac{\gamma_j}{\gamma'_{j_1}\gamma'_{j_2}\sqrt{2\pi}\sigma_\gamma} \exp\left(\frac{\log^2 \gamma_j - \log^2 \gamma'_{j_1} - \log^2 \gamma'_{j_2}}{2\sigma_\gamma^2}\right). \qquad (182) \end{aligned}$$

For the proposal ratio, define

$$c_S = \min\left\{1, \frac{P(S+1)}{P(S)}\right\}, \quad d_S = \min\left\{1, \frac{P(S-1)}{P(S)}\right\}. \tag{183}$$

as the probability of proposing a split and merge move respectively when the current number of splitting sites is S. Therefore the proposal ratio for a split move is

$$\text{proposal ratio} = \frac{d_{S+1}}{c_S}\frac{m-S-1}{S+1}. \tag{184}$$

Finally, we must evaluate the Jacobian to move between spaces. Splitting a window increases the dimensions of the space by 2: a site s^* of the split is inserted, an existing selection coefficient γ_j is removed, and two new HLA dependent selection parameters γ'_{j_1} and γ'_{j_2} are inserted. The fraction along the interval at which s^* is placed is generated through a uniform random variable $u_2 \sim U[0,1]$. Letting $\Delta := \frac{s^*-s_j}{s_{j+1}-s_j}$,

$$J^{-1} = \begin{pmatrix} \frac{\partial\gamma_j}{\partial\gamma'_{j_1}} & \frac{\partial\gamma_j}{\partial\gamma'_{j_2}} & \frac{\partial\gamma_j}{\partial\Delta} \\ \frac{\partial u_1}{\partial\gamma'_{j_1}} & \frac{\partial u_1}{\partial\gamma'_{j_2}} & \frac{\partial u_1}{\partial\Delta} \\ \frac{\partial u_2}{\partial\gamma'_{j_1}} & \frac{\partial u_2}{\partial\gamma'_{j_2}} & \frac{\partial u_2}{\partial\Delta} \end{pmatrix} \tag{185}$$

$$\frac{\partial\gamma_j}{\partial\gamma'_{j_1}} = \frac{\partial}{\partial\gamma'_{j_1}}\left(\gamma'_{j_1}{}^{\Delta}\gamma'_{j_2}{}^{1-\Delta}\right) = \Delta\gamma'^{\Delta-1}_{j_1}\gamma'^{1-\Delta}_{j_2} = \Delta\frac{\gamma_j}{\gamma'_{j_1}} \tag{186}$$

$$\frac{\partial\gamma_j}{\partial\gamma'_{j_2}} = \frac{\partial}{\partial\gamma'_{j_2}}\left(\gamma'_{j_1}{}^{\Delta}\gamma'_{j_2}{}^{1-\Delta}\right) = (1-\Delta)\,\gamma'^{\Delta}_{j_1}\gamma'^{-\Delta}_{j_2} = (1-\Delta)\,\frac{\gamma_j}{\gamma'_{j_2}} \tag{187}$$

$$\frac{\partial\gamma_j}{\partial\Delta} = \frac{\partial}{\partial\Delta}\left(\gamma'_{j_1}{}^{\Delta}\gamma'_{j_2}{}^{1-\Delta}\right) =: A \tag{188}$$

$$\frac{\partial u_1}{\partial\gamma'_{j_1}} = \frac{\partial}{\partial\gamma'_{j_1}}\left(\frac{\gamma'_{j_1}}{\gamma'_{j_1}+\gamma'_{j_2}}\right) = \frac{\gamma'_{j_2}}{\left(\gamma'_{j_1}+\gamma'_{j_2}\right)^2} \tag{189}$$

$$\frac{\partial u_1}{\partial\gamma'_{j_2}} = \frac{\partial}{\partial\gamma'_{j_2}}\left(\frac{\gamma'_{j_1}}{\gamma'_{j_1}+\gamma'_{j_2}}\right) = \frac{-\gamma'_{j_1}}{\left(\gamma'_{j_1}+\gamma'_{j_2}\right)^2} \tag{190}$$

$$\frac{\partial u_1}{\partial\Delta} = \frac{\partial}{\partial\Delta}\left(\frac{\gamma'_{j_1}}{\gamma'_{j_1}+\gamma'_{j_2}}\right) = 0 \tag{191}$$

$$\frac{\partial u_2}{\partial\gamma'_{j_1}} = \frac{\partial}{\partial\gamma'_{j_1}}(\Delta) = 0 \tag{192}$$

$$\frac{\partial u_2}{\partial\gamma'_{j_2}} = \frac{\partial}{\partial\gamma'_{j_2}}(\Delta) = 0 \tag{193}$$

$$\frac{\partial u_2}{\partial\Delta} = \frac{\partial}{\partial\Delta}(\Delta) = 1. \tag{194}$$

Thus,

$$|\det(J^{-1})| = \left| \det \begin{pmatrix} \Delta \frac{\gamma_j}{\gamma'_{j_1}} & (1-\Delta)\frac{\gamma_j}{\gamma'_{j_2}} & A \\ \frac{\gamma'_{j_2}}{\left(\gamma'_{j_1}+\gamma'_{j_2}\right)^2} & \frac{-\gamma'_{j_1}}{\left(\gamma'_{j_1}+\gamma'_{j_2}\right)^2} & 0 \\ 0 & 0 & 1 \end{pmatrix} \right| \tag{195}$$

$$= \left| \Delta \frac{\gamma_j}{\gamma'_{j_1}} \frac{-\gamma'_{j_1}}{\left(\gamma'_{j_1}+\gamma'_{j_2}\right)^2} - (1-\Delta)\frac{\gamma_j}{\gamma'_{j_2}} \frac{\gamma'_{j_2}}{\left(\gamma'_{j_1}+\gamma'_{j_2}\right)^2} \right| \tag{196}$$

$$= \frac{\gamma_j}{\left(\gamma'_{j_1}+\gamma'_{j_2}\right)^2}. \tag{197}$$

The Jacobian of the inverse mapping between the spaces $(\gamma_{j_1}, \gamma_{j_2}, \Delta)$ and (γ_j, u_1, u_2) (subject to renaming of variables) is J, so it follows that the determinant is $\frac{1}{\det(J^{-1})} = \frac{\left(\gamma'_{j_1}+\gamma'_{j_2}\right)^2}{\gamma_j}$.

6.4 CODE AND MEMORY CONSTRAINTS

In order to determine the posterior probability of a new state within the MCMC quickly, we keep track of both the forwards and backwards arrays $f_{i,j}$ and $b_{i,j}$ for each query sequence (as well as other quantities required to evaluate the likelihood). This requires storing two $|\text{codon sequence}| \times |\mathbf{D}| \times |\mathbf{D}_B|$ arrays of doubles. After multiple alterations to our C-code to reduce memory usage, when attempting to evaluate the likelihood of observing $\sim 2,000$ query sequences from a collection of $\sim 60,000$ reference protease sequences from the Stanford drug resistance database, over 300GB of RAM was required! We require a method to reduce the memory required to perform inference.

6.4.1 *Restricting* $\mathbf{D}_B$

The simple method that we use to reduce the required memory is to restrict $\mathbf{D}_B$ to a different subset $\mathbf{D}_{B_i}$ for each $D_i \in \mathbf{D}$. We make the approximation

$$P(\mathbf{D}|\mathbf{D}_B, \Theta) \approx \prod_i \hat{\pi}(D_i|\mathbf{D}_B, \Theta) \approx \prod_i \hat{\pi}(D_i|\mathbf{D}_{B_i}, \Theta). \tag{198}$$

If the sequences we choose for each $\mathbf{D}_{B_i}$ are similar to the true ancestors of each of the D_i then this will be a good approximation.

A simple way to restrict $\mathbf{D}_B$ is by considering the closest n sequences to D_i by Hamming distance (the Hamming distance between two strings is the total number of differences between the strings, giving each difference equal weighting). A drawback of using Hamming distance to restrict $\mathbf{D}_B$ is that in the presence of high recombination rates, one can imagine sequences which should be included in $\mathbf{D}_{B_i}$ but are excluded on the basis this metric.

Other potential approaches include using the likelihood of copying from a particular reference sequence to generate query sequences using the marginal distributions detailed in Equation (134), or using a Viterbi algorithm to reduce our inference to a smaller collection of sequences based on likely paths through the reference dataset $\mathbf{D}_B$.

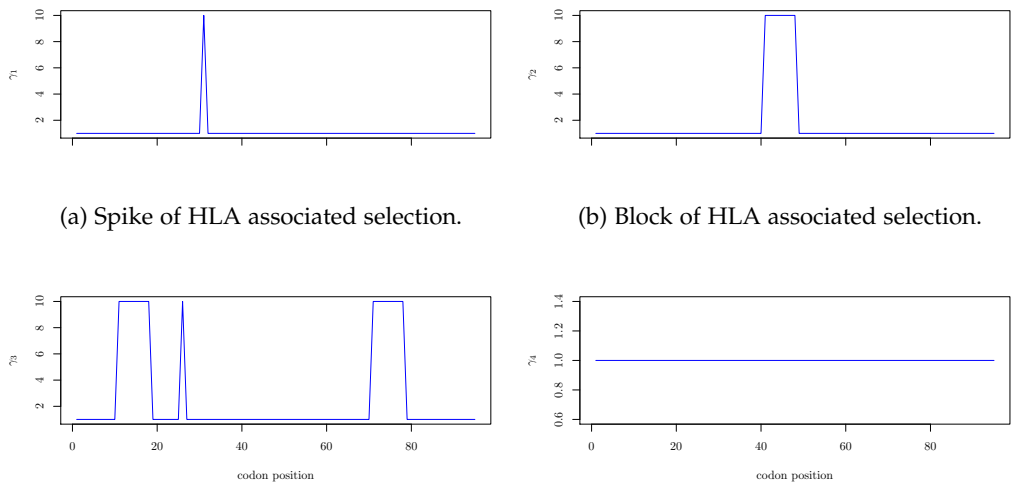
For the small amount of overhead, all of these methods dramatically reduce the computational time and memory footprint required by the program to the extent that a run of 2,000,000, states analysing $\sim$1000 query sequences over a codon sequences of length $\sim$100, considering the closest 100 reference sequences according to some metric may be performed in $\sim$24 hours on an Intel *i*7 desktop machine using $<$1GB of RAM. In the simulation studies that follow, we use the Hamming distance metric to restrict $\mathbf{D}_B$. In the next section we perform a simple test to determine the number of sequences required for a reliable restriction using Hamming distance.

6.5 SIMULATIONS

We now perform a collection of simulation studies to examine the performance of the model under differing recombination probabilities. In each case, query datasets are simulated under the model, from a reference dataset $\mathbf{D}_B$. $\mathbf{D}_B$ was created by combining HIV-1 protease sequence data from across the three major public databases (Los Alamos HIV sequence database [256], Stanford drug resistance database [63], and UCLA HIV positive selection mutation database [212]). Methods used to combine across datasets, align and avoid repetitions of sequences are described in detail in Section B.1.2. For our simulations, we consider HLA selection profiles designed to resemble the sorts of signal likely to exist in real data. Strong selection at a site, potentially induced by HLA associated selection at an anchor residue or site at which a compensatory mutation could occur is represented by a selection spike. HLA associated selection across a larger region; used as an approximation to selection over an epitope, is represented by a selection block. We consider a more complex case by combining spikes and blocks. Finally, we allow a complete lack of HLA associated selection across the region. Examples are displayed in Figure 6.3. For each profile we set any HLA associated selection to be $10\times$ the underlying value of ω across the sites. ω_i at each site is sampled from $Exp(1)$. We consider 6 HLA profiles for 4 common ($n = 100$) and 2 rare ($n = 30$) HLA types, resulting in a total of 460 simulated sequences:

- HLA 1: selection spike, $n = 100$.
- HLA 2: selection block, $n = 100$.
- HLA 3: complex selection, $n = 100$.
- HLA 4: no selection, $n = 100$.
- HLA 5: selection spike, $n = 30$.
- HLA 6: selection block, $n = 30$.

Note that we consider the simpler case that individuals have exactly one HLA type. κ and ϕ are set at 2.09 and 0.924 respectively (based on empirical estimates



(a) Spike of HLA associated selection.

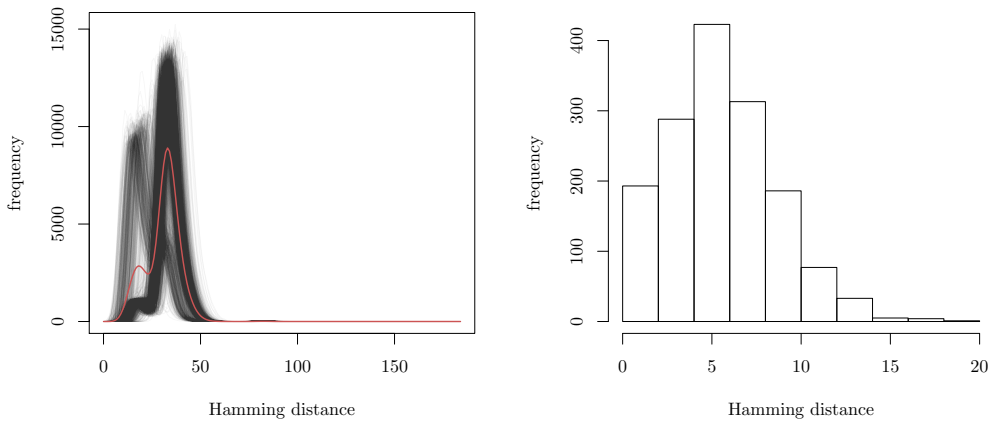
(b) Block of HLA associated selection.

(c) Complex HLA associated selection.

(d) No HLA associated selection.

Figure 6.3: HLA associated selection profiles. For our simulation studies, we consider 4 classes of selection profiles. Figure 6.3a is a spike in selection; to determine our power to estimate HLA associated change in selection over a small region, Figure 6.3b is a selection block; a simple approximation to an epitope, Figure 6.3c displays complex selection; a combination of spikes and blocks. Finally we consider HLA types that are not associated to any change in selection across the region in Figure 6.3d.

using a large African HIV-1 sequence dataset [250]), and μ is set at 50 such that the average distance to the nearest neighbour is $\sim$3 nucleotide differences; which is conservative given the results shown in Figure 6.4, where the mean Hamming distance to the nearest neighbour is $\sim$6 nucleotide differences. After simulating 100 sets of 460 viral sequences for each recombination probability, we perform MCMC on each simulated sequence dataset. We run each MCMC for 1,000,000 iterations, sampling every 200th state. Priors for each parameter are shown in Table 6.1. We first look to determine the number of sequences required for $\mathbf{D}_{B_i}$ if using Hamming distance as a metric to restrict $\mathbf{D}_B$ for each query sequence. We compare results of a run where we evaluate an estimate of the likelihood based on the closest 100 by Hamming distance to a run in which we estimate the likelihood based on the closest 10 by Hamming distance (results not shown), in the absence of recombination. Interestingly, the signal of selection in the absence of HLA dependent selection ω was conserved between datasets with 10 vs. 100 sequences, suggesting that even restricting to 10 sequences we have power to extract the true signal of ω. The signal of HLA associated selection was also very similar, again suggesting power to estimate parameters comes from a small subset of the



(a) Distribution of Hamming distance of $\mathbf{D}$ from $\mathbf{D}_B$ in protease.

(b) Distribution of Hamming distance of closest sequence in $\mathbf{D}_B$ from members of $\mathbf{D}$.

Figure 6.4: Figure 6.4 shows the distribution of the Hamming distance to members of $\mathbf{D}_B$ from each of the members of $\mathbf{D}$ in protease sequences. Each distribution is plotted in grey, with the mean across all members of $\mathbf{D}$ plotted in red. Figure 6.4b shows the distribution of Hamming distances to the closest sequence in $\mathbf{D}_B$

Parameter	Prior
μ	$Exp(0.01)$
ν_i	$Exp(10)$
ω_i	$Exp(0.5)$
ζ_i	$\log\mathcal{N}(0,2)$
$\gamma_{i,h}$	$\log\mathcal{N}(0,2)$
window expansion	$Geom(0.1)$
number of windows	$Bin(m,0.2)$

Table 6.1: Priors on parameters of the model.

data. The approximation in Equation 198 works in the absence of recombination, even using a simple restriction of $\mathbf{D}_B$ by Hamming distance to as few as 10 sequences. The use of more sophisticated techniques to restrict $\mathbf{D}_B$ should allow us to more accurately estimate the important sequences most closely related to each query sequence, reducing the required size of each $\mathbf{D}_{B_i}$ and by extension, computation time. Restricting $\mathbf{D}_B$ is attractive in terms of computation time and can result in similar estimates even for the simplest approaches (Hamming distance restriction to 10 sequences in the absence of recombination). However, due to the relatively low computational cost of expanding to 100 sequences and the stronger signal of escape parameters this provides, we consider the closest 100 sequences by Hamming distance for the remainder of our analyses.

Next, we examine the effect of recombination upon estimates. We set a constant recombination probability between adjacent sites across the region and increase it from 0, to 0.01, to 0.05. Results for each of the three recombination rates are shown in Figures 6.5 - 6.17.

As expected, increased recombination results in less power to estimate HLA associated selection parameters. Coverage drops off and average estimates of HLA associated scalings are lower than their true values in regions under HLA associated selection. In addition, we find that as the rate of recombination is increased, both recombination and ω are increasingly underestimated over the entire region (not only at the edges of the region, as we would expect for recombination), as shown in Figures A.4, 6.12, 6.16 and 6.17. ζ is overestimated, see Figures A.4, 6.12 and 6.16. This is not surprising, as in order to detect reversion parameters we require non-consensus in the position of the reference dataset copying from, which is far rarer than observing consensus at the site (required to estimate $\gamma_{i,h}$). A potential source of some of the error in estimation is our use of the Hamming distance metric to define $\mathbf{D}_{B_i}$ for each sequence $i \in \mathbf{D}$. This is because it is a poor metric with which to restrict $\mathbf{D}_B$ for recombinant sequences. Consider the following extreme example: suppose we have a recombinant B/C sequence for which 60% of the sequence is subtype B. Then sequences that are close to it in $\mathbf{D}_B$ as measured by Hamming distance are likely to be subtype B sequences. Therefore $\mathbf{D}_{B_i}$ will contain no subtype C sequences and thus no recombination

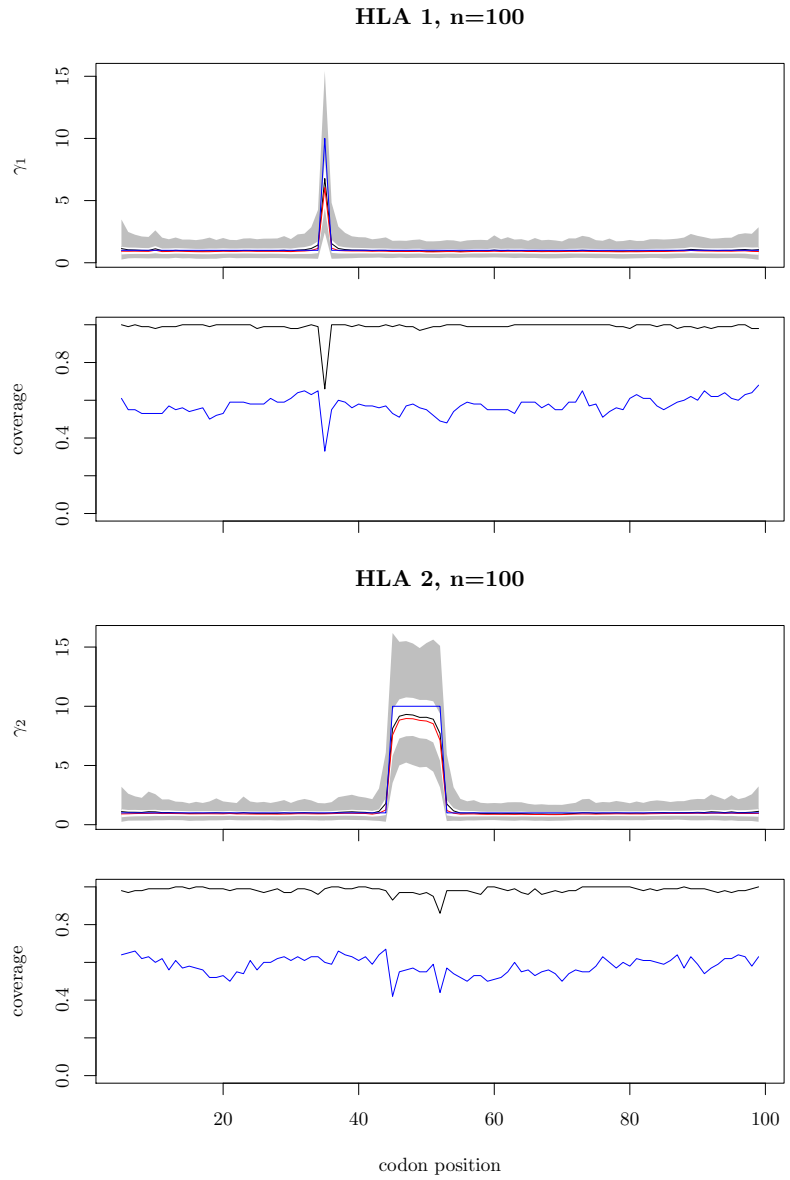


Figure 6.5: Simulations to evaluate HLA associated selection across a region; $\boldsymbol{r = 0}$ (no recombination). Odd rows show average estimates and credible intervals of HLA associated selection (γ_h) across the region for the first two HLA types of the simulated sequences. Averages are taken over 100 independent MCMC runs on independent simulated data with the same underlying parameters. The mean estimate is plotted in black, the median is plotted in red, and the true underlying parameter value is plotted in blue. The white band denotes the 50% credible interval within the 95% credible interval denoted by the grey band. Even rows display the coverage, the black line shows the proportion of the 100 estimates that lie within the 95% credible interval for each run, and the blue line shows the corresponding proportion lying within the 50% credible interval.

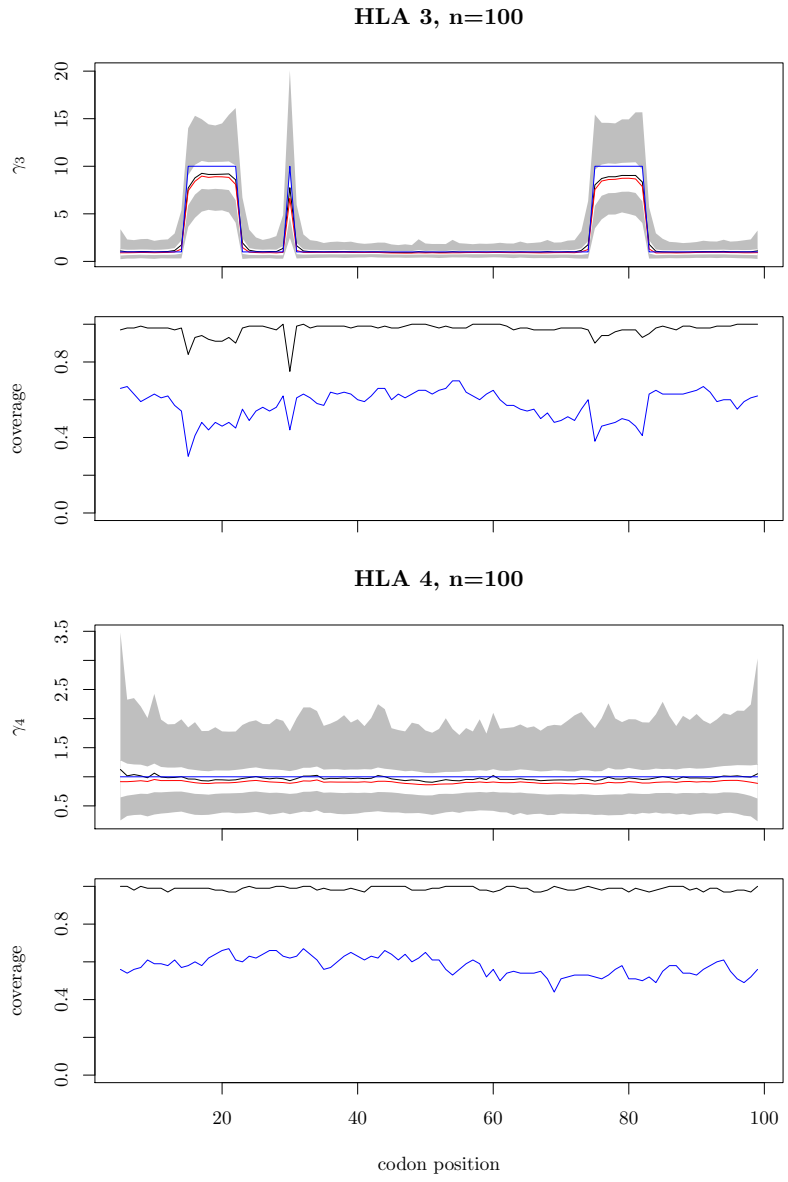


Figure 6.6: Simulations to evaluate HLA associated selection across a region; $\boldsymbol{r = 0}$ (no recombination). Plotting is as in Figure 6.5.

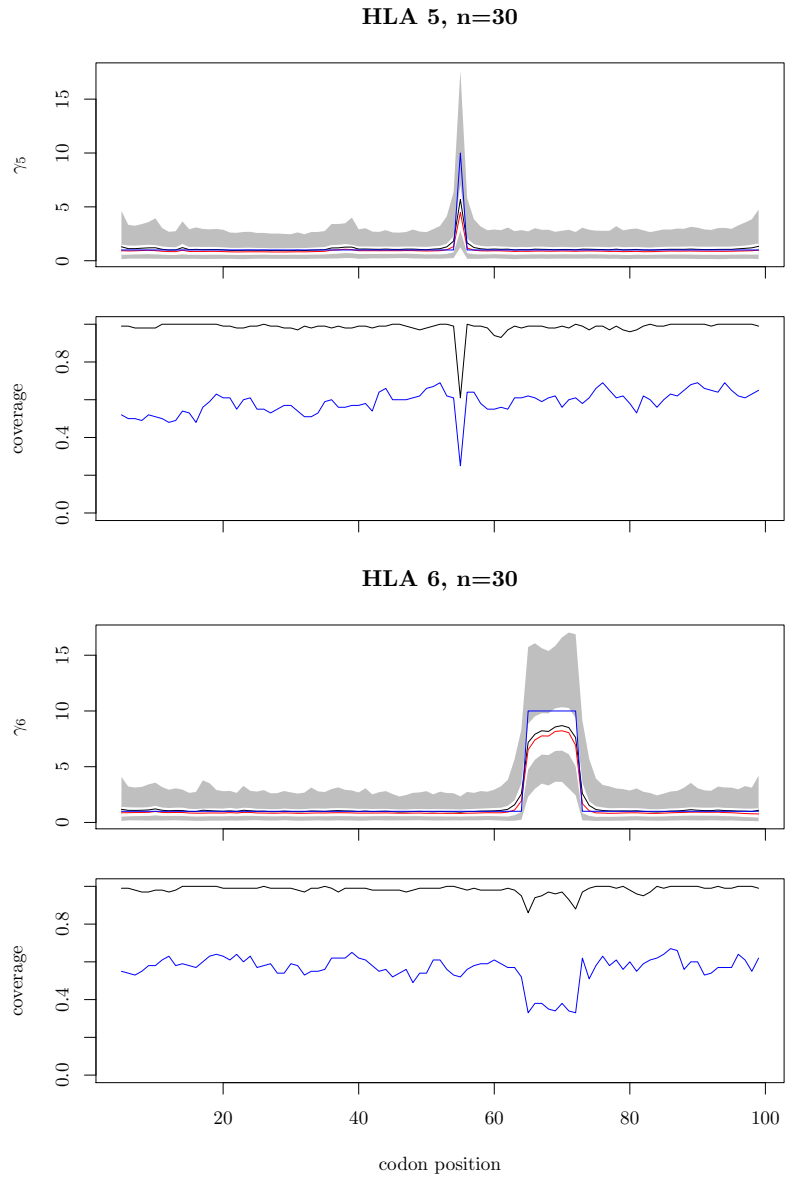


Figure 6.7: Simulations to evaluate HLA associated selection across a region; $\boldsymbol{r = 0}$ (no recombination). Plotting is as in Figure 6.5.

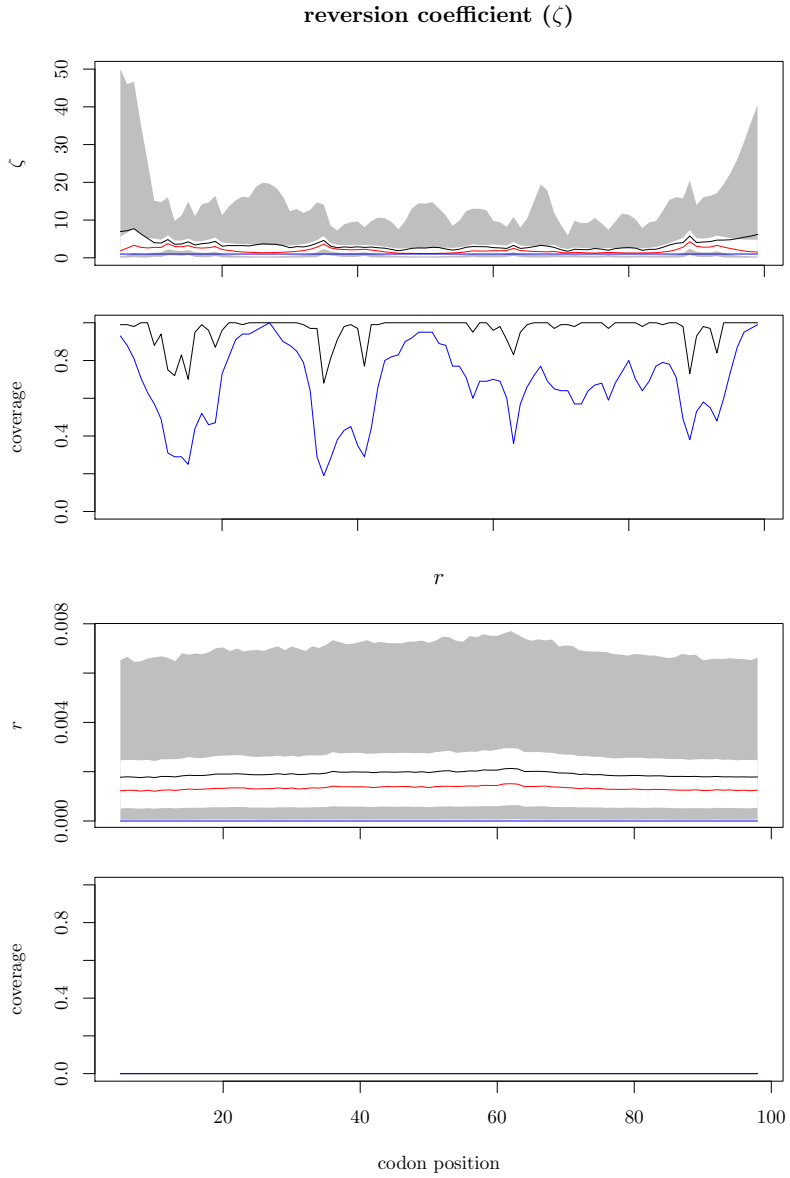


Figure 6.8: Estimating scaling of selection due to reversion (ζ) and recombination probability across a region; $\boldsymbol{r = 0}$ (no recombination). Odd rows show average estimates and credible intervals of parameters describing reversion (ζ) and recombination probability between adjacent sites (r). Averages are taken over 100 independent MCMC runs on independent simulated data with the same underlying parameters. The mean estimate is plotted in black, the median is plotted in red, and the true underlying parameter value is plotted in blue. The white band denotes the 50% credible interval within the 95% credible interval denoted by the grey band. Even rows display the coverage, the black line shows the proportion of the 100 estimates that lie within the 95% credible interval for each run, and the blue line shows the corresponding proportion lying within the 50% credible interval.

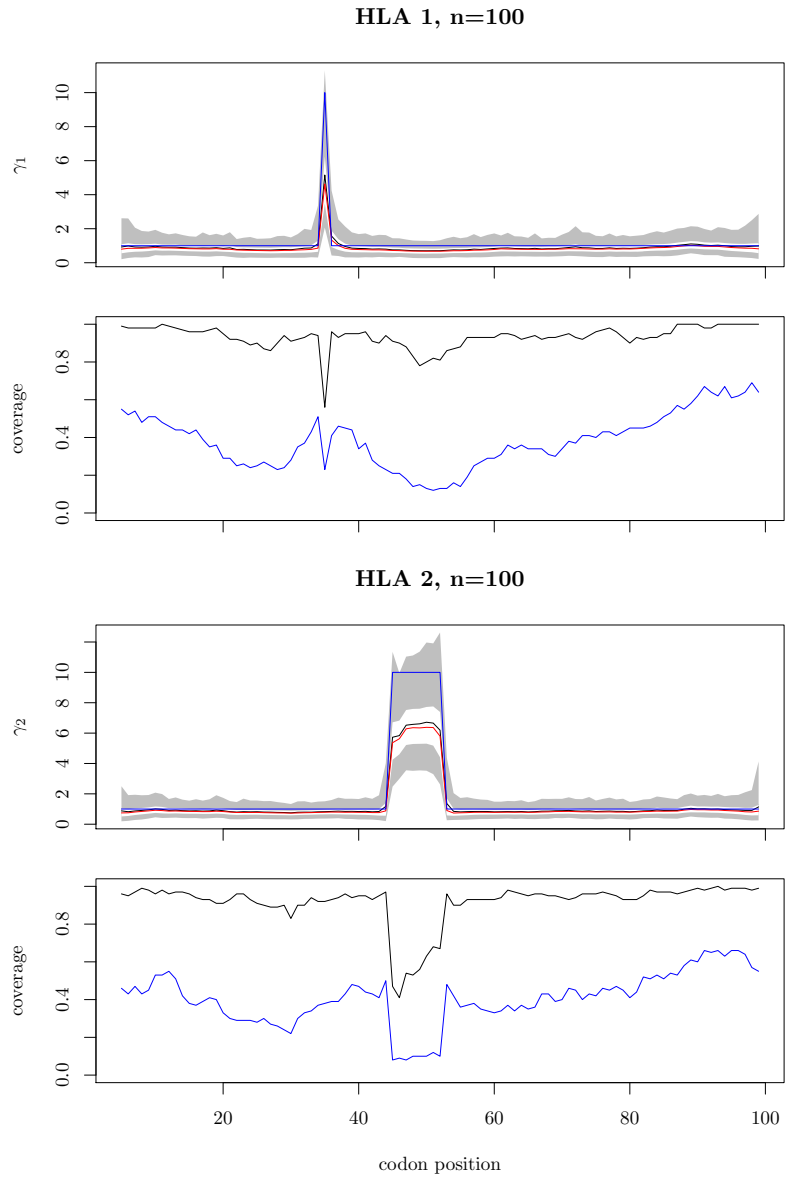


Figure 6.9: Simulations to evaluate HLA associated selection across a region; $r = 0.01$. Plotting is as in Figure 6.5.

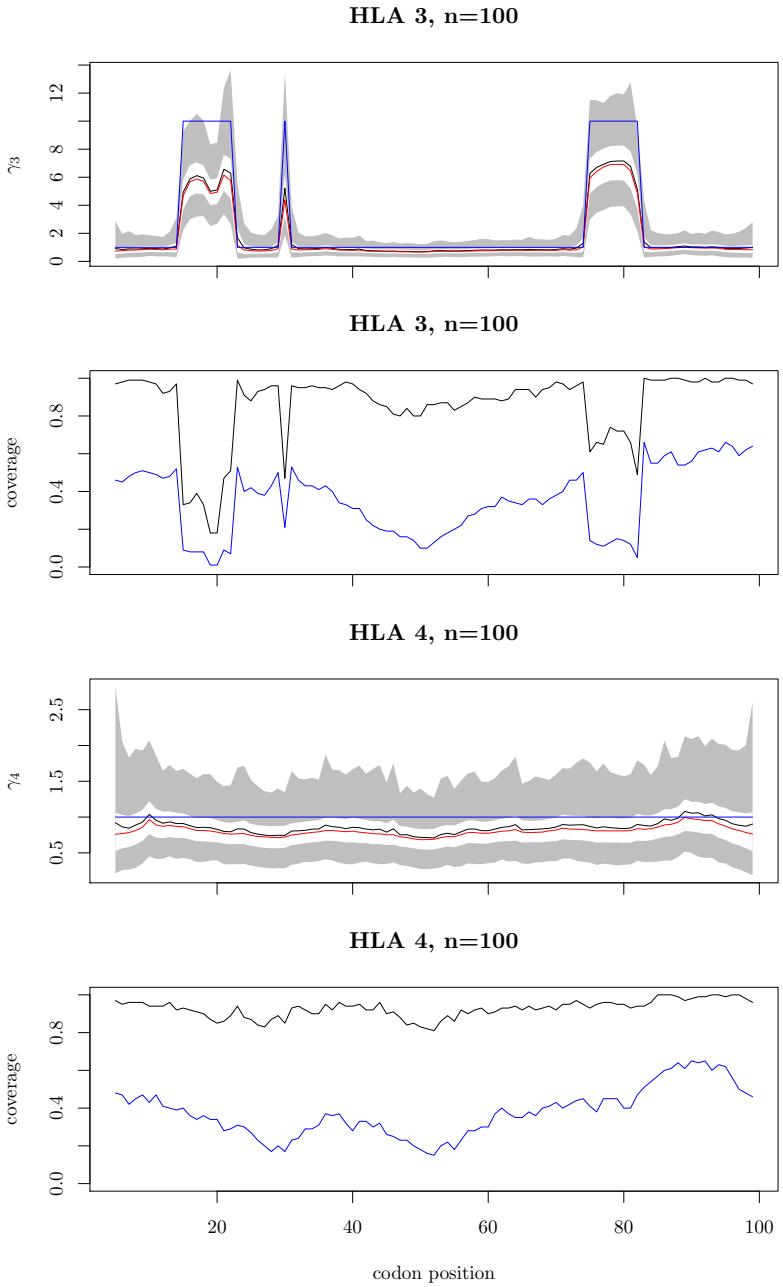


Figure 6.10: Simulations to evaluate HLA associated selection across a region; $r = 0.01$. Plotting is as in Figure 6.5.

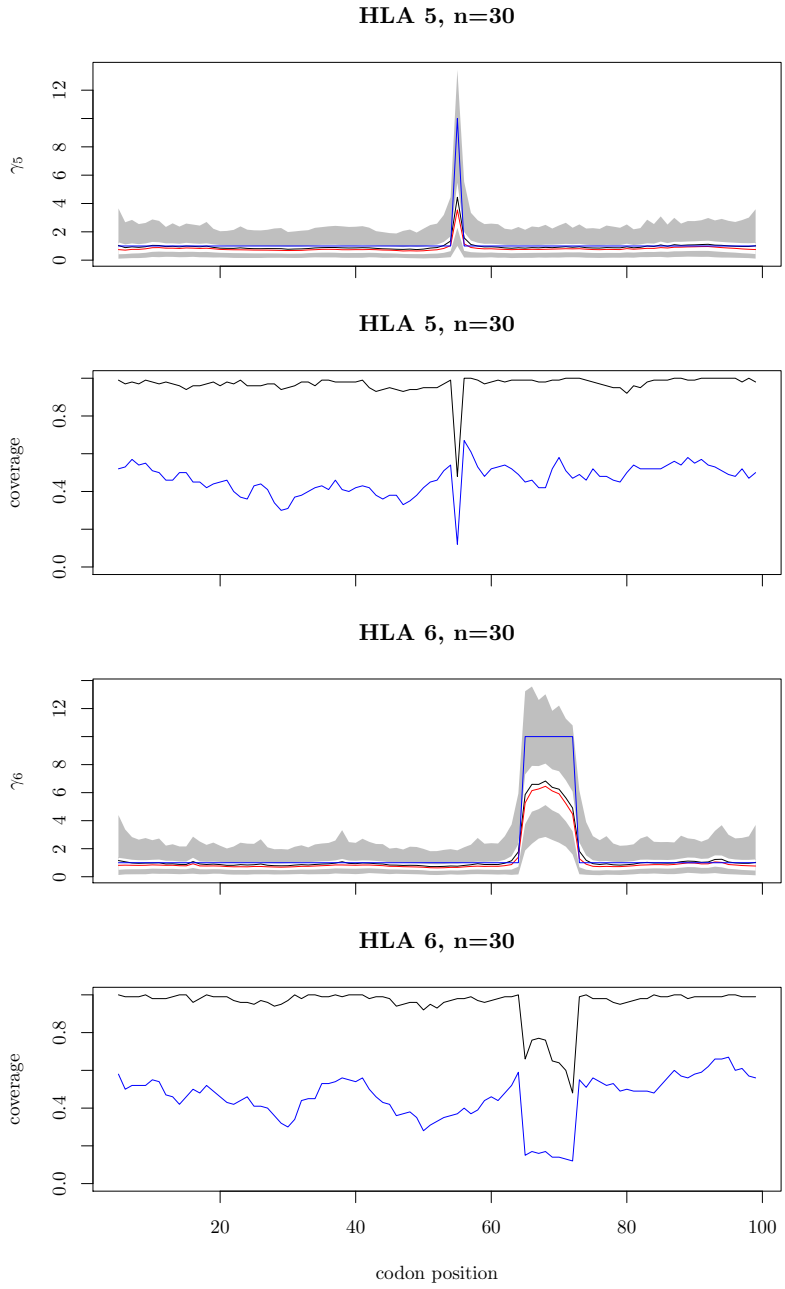


Figure 6.11: Simulations to evaluate HLA associated selection across a region; $r = 0.01$. Plotting is as in Figure 6.5.

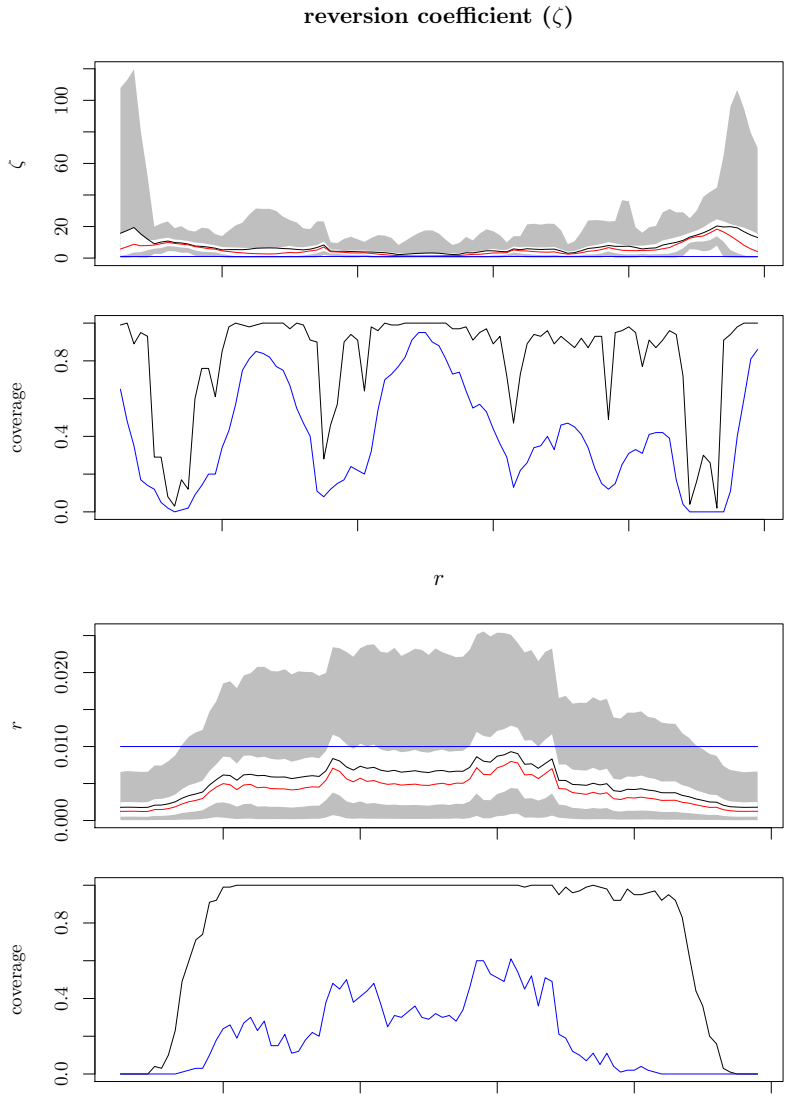


Figure 6.12: Estimating scaling of selection due to reversion (ζ) and recombination probability across a region; $\boldsymbol{r = 0.01}$. Plotting is as in Figure 6.8.

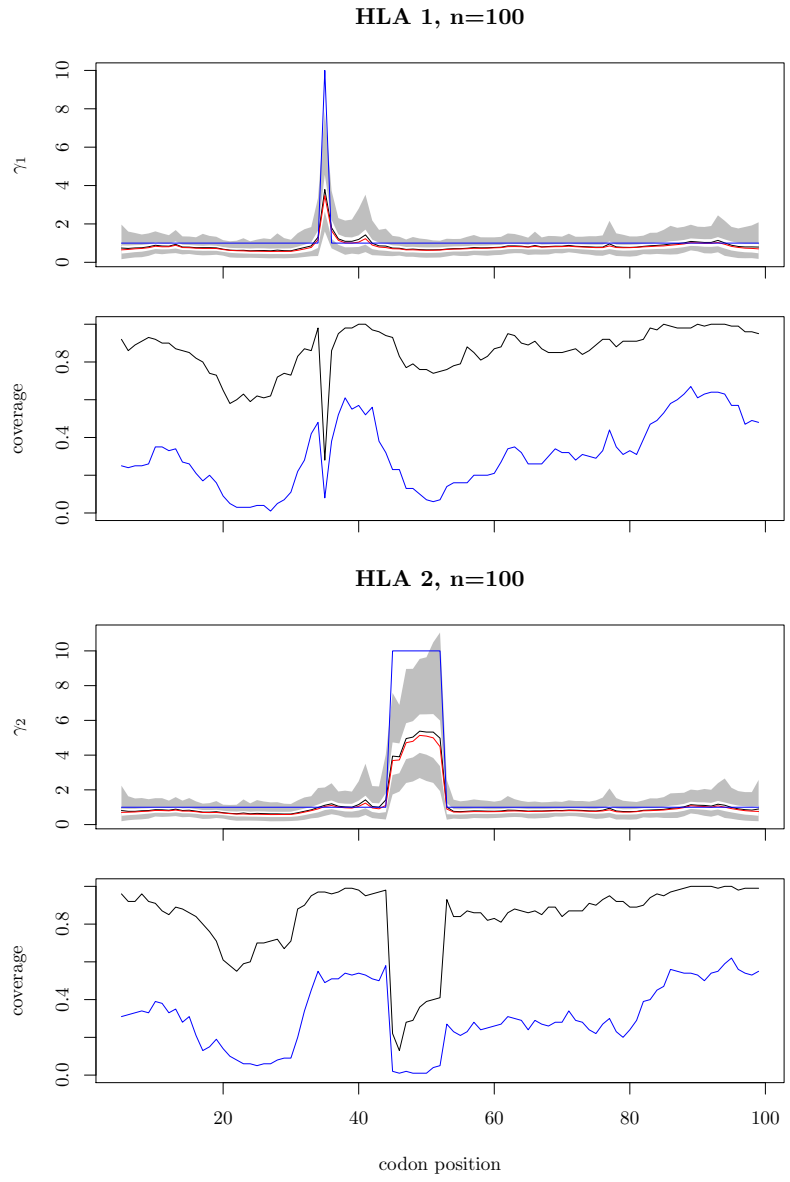


Figure 6.13: Simulations to evaluate HLA associated selection across a region; $r = 0.05$. Plotting is as in Figure 6.5.

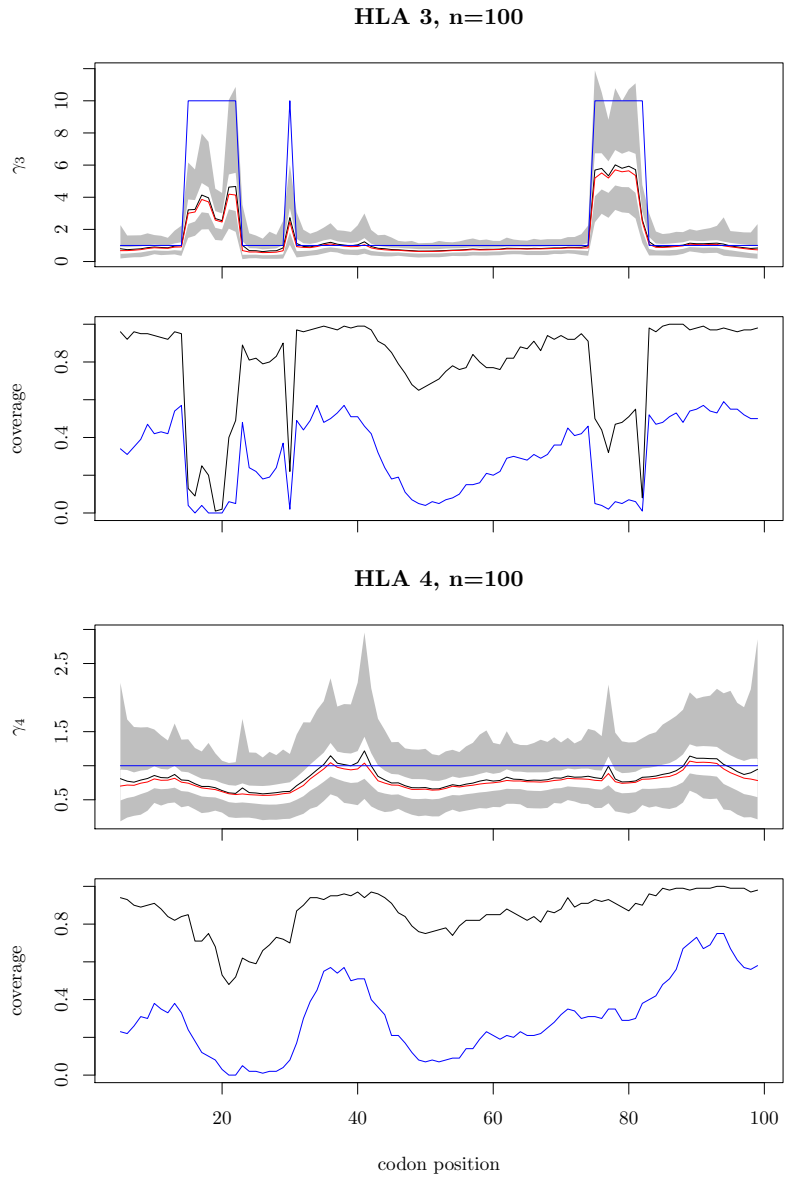


Figure 6.14: Simulations to evaluate HLA associated selection across a region; $r = 0.05$. Plotting is as in Figure 6.5.

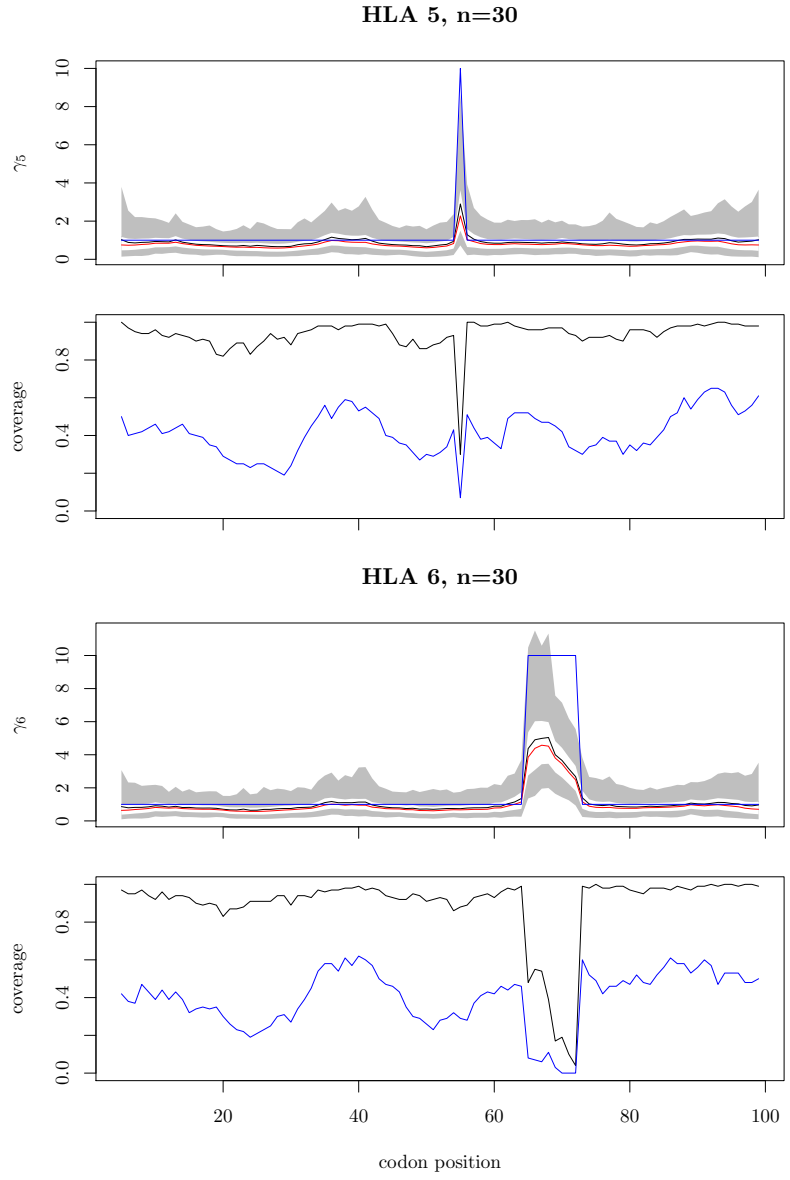


Figure 6.15: Simulations to evaluate HLA associated selection across a region; $r = 0.05$. Plotting is as in Figure 6.5.

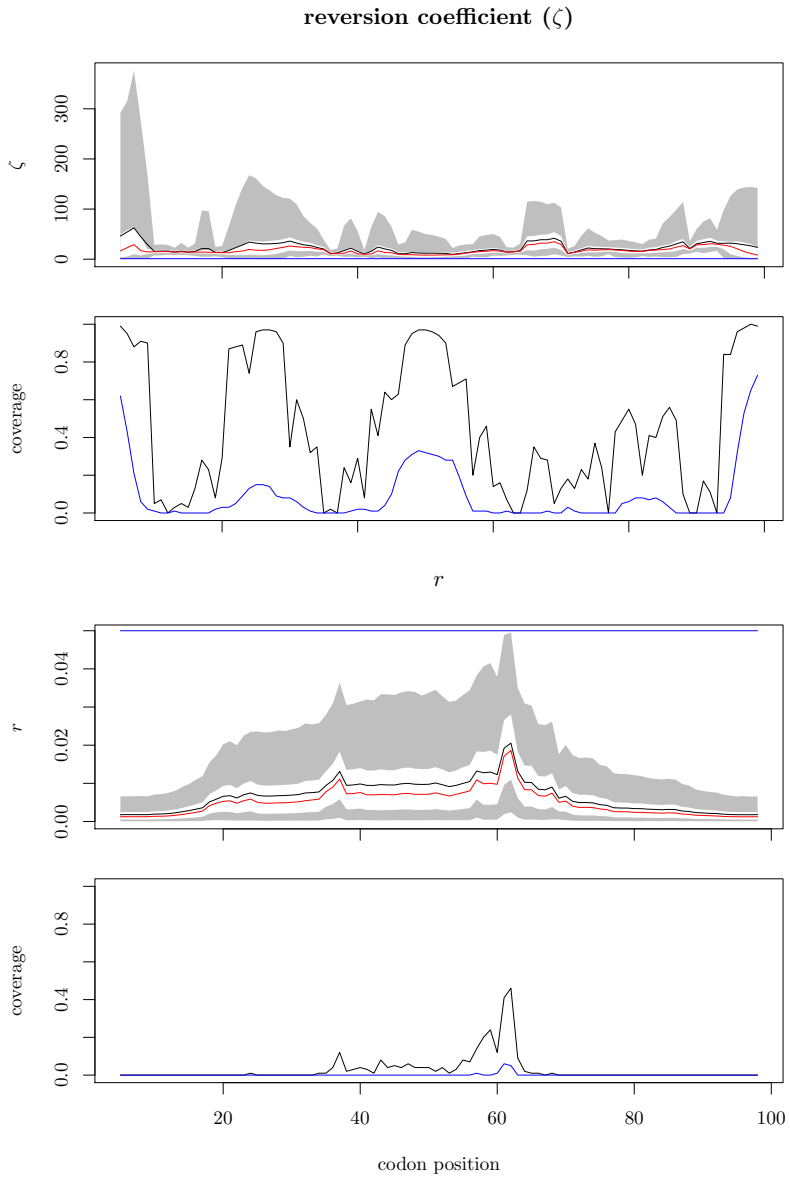


Figure 6.16: Estimating scaling of selection due to reversion (ζ) and recombination probability across a region; $\boldsymbol{r = 0.05}$. Plotting is as in Figure 6.8.

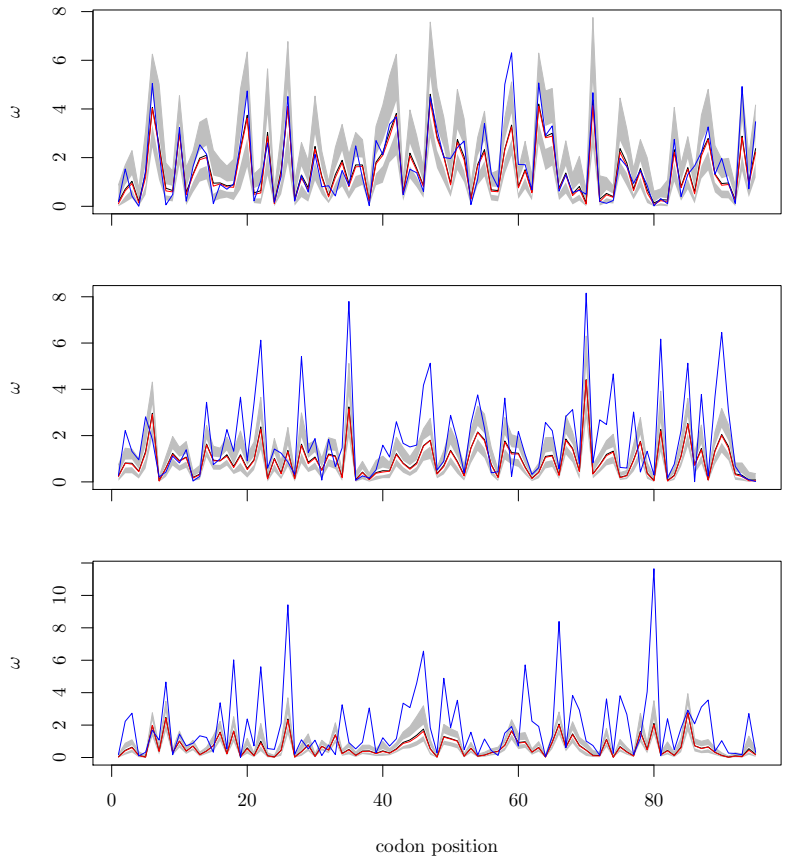


Figure 6.17: Examples of estimates of ω across the region. The first, second and third rows are estimates of ω from a single MCMC run when the underlying recombination probability between adjacent sites across the region was 0, 0.01 and 0.05 at each site, respectively. Plotting is as in Figure 6.5.

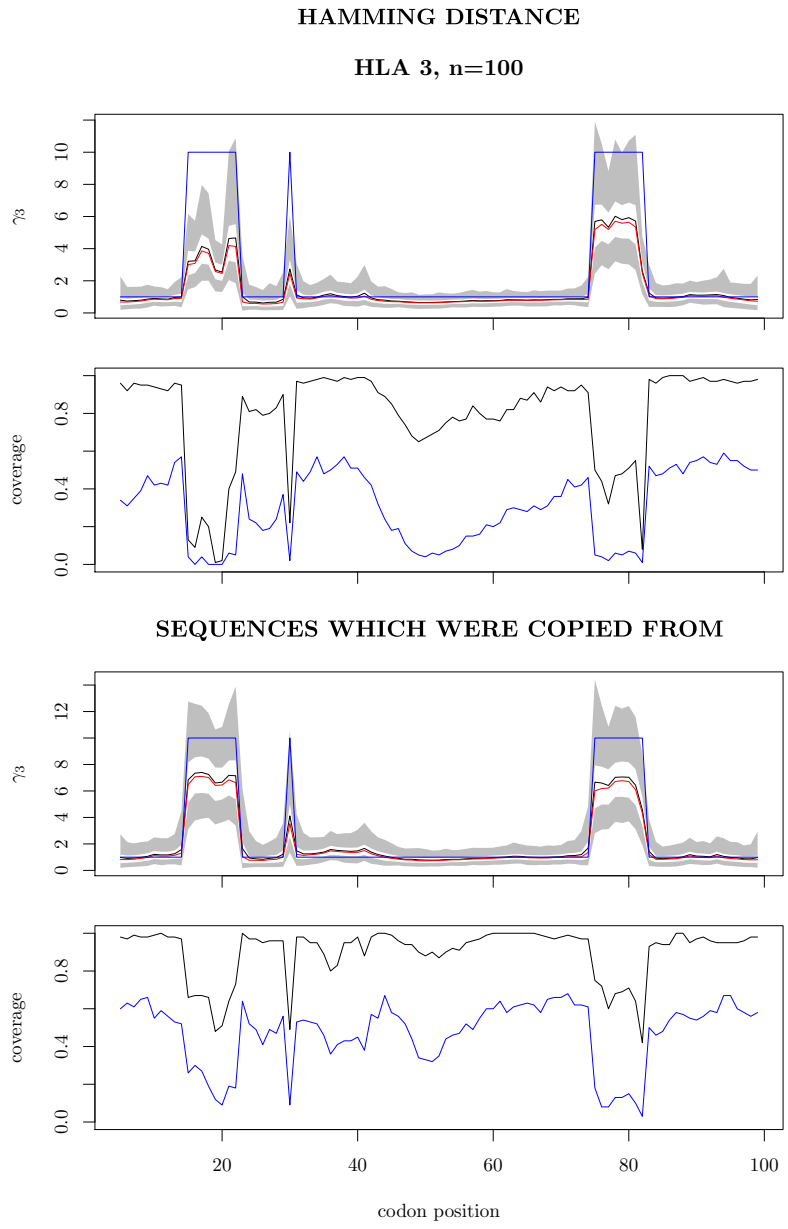


Figure 6.18: Plots of averages, credible intervals and coverage of estimates generated using the closest 100 sequences by Hamming distance in the first two plots, and the actual sequences copied from to generate simulated viral sequence datasets in the bottom two plots. Estimates for HLA type 3 are displayed. $r = 0.05$.

event will be detected. Instead a boost in reversion is used to explain the subtype C portion of the recombinant sequence instead. To determine the magnitude of this effect, we compare the restriction of $\mathbf{D}_B$ by Hamming distance to a restriction which contains the true SCFs for each member of $\mathbf{D}$. We construct each $\mathbf{D}_{B_i}$ for this 'gold standard' restriction as follows: we put all the SCFs (sequences that were copied from) to produce sequence D_i into $\mathbf{D}_{B_i}$, and fill the remainder of the 100 sequences with sequences closest to the SCFs by Hamming distance. The resulting estimates using this gold standard restriction of $\mathbf{D}_B$ are shown in Figures 6.18 and A.1-A.4. These parameters were estimated using simulated sequences for which the underlying recombination probability between adjacent sites across the region was 0.05.

We find that there is indeed an improvement in our estimates of HLA associated selection, estimates are closer to the truth and coverage is more accurate: see Figure 6.18 for a comparison of HLA 3 estimates. Also, recombination estimates are increased as shown in Figure A.4. However, reversion remains poorly estimated. Fortunately, this does not appear to affect our estimates of HLA dependent selection away from consensus. Given that we expect the true between host recombination rate to be small but non-zero over the HIV-1 genome due to low estimates of super-infection [97, 44, 161] we use the simple Hamming distance metric when applying our model to real data in Chapter 7 rather than incorporating a more sophisticated method to restrict $\mathbf{D}_B$ for each query sequence.

6.6 DISCUSSION AND FUTURE WORK

In this chapter we have developed a model that can as the question 'does this set of viral sequences from hosts with known HLA profiles indicate that hosts with distinct HLA profiles exert different selection pressures at particular sites in the viral genome?' To do so, we have introduced an extension to the model presented by Wilson and McVean [304] which incorporates HLA associated selection and approximates the NY98 codon model to avoid computationally expensive matrix exponentials. HLA associated selection at each codon is modelled by

parameters which modify the non-synonymous/synonymous ratio (ω_i) for any non-synonymous substitution away from the defined consensus sequence. We also model reversion back to the consensus sequence (independent of HLA type) at each codon through another collection of parameters ζ. These parameters scale ω_i for any non-synonymous substitution towards the defined consensus sequence. In this way we hope to detect increased ($\gamma_{i,h} > 1$) selection associated to host HLA types (though this may be extended to any host variable; be that HLA type, drug regime, or some other factor). In the model, we also estimate the recombination rate and ω across the sequence. To explore and sample from the posterior distribution, we use a number of MCMC moves (scaling moves and window based reversible-jump MCMC moves). In so doing, we hope to estimate HLA associated selection pressures whilst controlling for the potential confounding effects of recombination and selection.

To test the model we performed simulation studies to examine the effect of recombination on parameter estimates. In each case the resulting estimates of HLA associated selection were biased downwards. Despite this, even under very high recombination rates we have sufficient power to detect changes in HLA associated selection. There were also severe computational constraints which had to be overcome in order to perform inference. We found that using a simple threshold on Hamming distance to constrain the size of $\mathbf{D}_B$ has the ability to pick up the most important sequences and drastically increase our overall speed of computation. We note that more sophisticated approaches could restrict $\mathbf{D}_B$ still further, and lead to more accurate parameter estimates. Estimates of reversion parameters were poor, and recombination and ω were increasingly underestimated as the recombination rate was increased. Further investigation is required to pinpoint the source of this underestimation.

It is important to emphasise here that the model presented is a strong departure from many of the standard methods used to perform inference using pathogen sequence data. Existing analysis is predominantly centered around the genealogy of the sampled sequences, and uses maximum likelihood (e.g. Phylip [1], PhyML [110], RAxML [271] and FastTree [227, 228] among many others), and Bayesian frameworks (e.g. [68], [244]). Exploring the distribution of trees is noto-

riously difficult, though sampling schemes have been successfully implemented in programs such as BEAST [68] and MrBayes [244]. Models included in these programs can incorporate a number of parameterisations to include invariant sites, sites under strong selection, changes in effective population size (both parametric [267, 109] and non-parametric [64, 188, 270]), and phylogeography [162]. However, these methods and programs often have one thing in common - their computational intensity. In the case of HIV-1, and other viral epidemics, the effective population size has experienced strong growth, leading to long terminal branches in estimated/sampled genealogies. Thus, the ability to infer parameters of interest, using tree-pruning for example, requires yet more samples, leading to further computational time. E.g. using the program BEAST, we find that obtaining convergence of the MCMC for HIV-1 sequence datasets of size ~ 300 in p24 gag (~ 700 nucleotides) becomes infeasible. Add to this the issue of recombination, and we conclude that a different approach is required that can address questions that can only be answered by using very large sequence datasets.

The ability to analyse large datasets gives us access to the powerful information contained in recent coalescence events as we walk along the viral genome. The model we present here is fast and is amenable to extension - for example, we may consider transmission chains, or multiple within host sequences through simple modification of the likelihood. The model is general; we may analyse any putative host-pathogen interaction and look for selective effects. Obvious examples include other viruses such as hepatitis C, but we could also investigate other pathogens such as malaria.

There are a variety of assumptions and limitations of the model. As we noted earlier, the population is assumed to be fixed which is clearly incorrect as the dynamics of the HIV-1 infected population are highly complex. An assumption of our model is that given enough data, branches are sufficiently short that over this period of time the effective infected population size is approximately constant. Similar approaches to Frost and Volz [292, 84, 293] or Koelle and Rasmussen [155, 237] could be used to estimate the coalescent rate for a changing infected host population size based on epidemiological data, though we suggest checking the impact of such a change would have on parameter estimates using simulated

datasets. Throughout, we also assume homogeneous mixing.

Viruses lend themselves well to inference over short time-scales due to their rapid viral turnover and mutation rate. However, it also important to recognise that this fast mutation rate also presents new challenges. Consider a strong selective pressure for a particular variant associated to an HLA type. This could also be detected as a recombination event if a sufficiently large proportion of the population also have the variant at the site, particularly if the variant is close to the edges of the analysed sequence. In the next chapter, when applying the model to real data, we control for this by enforcing a strict prior on the recombination rate, but this is clearly not optimal. As with many of these methods, sites are considered independently rather than considering all codon changes within the sequence together. Finally, in our regime we use both the forwards and backwards algorithms to determine the likelihood. However, similar speeds could be achieved using just the forwards algorithm, with a 2-fold reduction in the required memory.

In Chapter 7, we analyse a collection of real datasets. We first consider selection dependent on the host's drug regimen. We then focus on HLA associated selection. To do this we combine sequence datasets from the three major databases listed in 6.1 to create a reference dataset $\mathbf{D}_B$, and use all sequence data with associated host HLA data available to us to create $\mathbf{D}$.

7

HYPOTHESIS FREE ESTIMATION OF HLA ASSOCIATED SELECTION IN THE PRESENCE OF RECOMBINATION USING LARGE DATA SETS: REAL DATA

7.1 INTRODUCTION

In this chapter, we apply the model implemented and tested in Chapter 6 to detect drug associated selection and HLA associated selection along regions of HIV-1 *pol* which encode protease and reverse transcriptase. Using the wealth of paired drug regime/viral sequence data available on the Stanford drug resistance database [63], we are able to estimate drug associated selection along the region of interest by replacing 'HLA type' with 'patient drug regime' in the model. We use the large amount of literature detailing the location and impact upon drug sensitivity of drug resistance mutations to further validate the model. We determine whether we are able to detect known drug resistance mutations and examine our sensitivity in doing so. We then switch focus to HLA associated selection by merging data from a collection of databases and other sources. We perform tests to check that two potential sources of bias are not playing a role in our dataset. Results for both studies are discussed in detail. In our HLA associated selection analyses, we ask the following questions:

- Do we detect strong selection in well known HLA associated epitopes in the analysed regions?
- By considering collections of putative and known epitopes of increasing validity (using a computational approach, and two collections of validated epitopes, do we find that sites estimated by our method to be selected are within epitopes?

- Do HLA associated mutations exist which lead to drug resistance (or partial drug resistance)?
- Differences between HIV-1 subtype consensus sequences are thought to be mainly driven by founder effects. Are differences between HLA allele frequencies across populations also playing a role in driving these amino acid subtype differences?
- Are there any selected sites associated with potential epitopes currently not defined in the Los Alamos database [127]?

Finally, we summarise our findings and explore the potential to apply the model to other problems in the field of pathogen evolution.

7.2 DRUG RESISTANCE STUDY

We first apply our model to investigate drug associated selection in protease and reverse transcriptase. Doing so highlights that our model may be used for a variety of applications in which we seek to determine the change in selection pressure associated to some host variable (be it HLA type, drug regime or some other host attribute). It also serves as a validation of our approach because of the wealth of existing knowledge of drug resistance mutations in protease and reverse transcriptase. On average, we expect the selective advantage conferred by drug escape to be easier to detect than that for HLA associated CTL escape because of the constant and intense selection pressure imposed by drugs. The drug associated selection parameters we estimate represent the trade-off between detriment to replicative fitness imposed by drug associated mutations and the benefit gained through the associated reduction in drug sensitivity. We summarise the sensitivity of our model based on current knowledge, and determine if there are any sites of selection which are currently unidentified, which may then be tested *in vitro*.

Before we begin our analyses, we first explain why we chose protease and reverse transcriptase to examine selective pressures on HIV-1 in this chapter. There are known to be more CTL responses directed at *gag* and *nef*, and *pol* has the

added complication of drug resistance mutations as a source of noise and potential bias. However, it is precisely because protease and reverse transcriptase are used as drug targets that they are also the target of the most extensive sequencing of the HIV-1 genome, and huge sequence datasets are required if we wish to test our model to its full potential. Also, the importance of the enzymes protease and reverse transcriptase within the HIV-1 life cycle mean that they are relatively well conserved, and indels are rare. As such, alignment is relatively easy, even for very large sequence datasets. Finally, we may also test for drug associated selection, the subject of this section.

7.2.1 *Data*

To carry out our analysis, we require a reference dataset (consisting of a large number of sequences) and a query dataset (consisting of paired viral sequence and host drug regime data). The Stanford drug resistance database [63] is an online database that catalogues drug-related sequence variation in the enzymes reverse transcriptase, protease and integrase - targets of antiretroviral therapy, found in the *pol* gene of HIV-1. The database contains a large proportion of published reverse transcriptase, protease and integrase sequences, including submissions to GenBank [21], sequences published in journal articles and sequences of HIV-1 isolates from people participating in clinical trials. Most importantly for our purposes, sequences are linked to data about the source of the sequence, the antiretroviral drug treatment history of the person from whom the sequence was obtained and the results of *in vitro* drug susceptibility testing. Thus, instead of labelling hosts with their HLA profile, we may label them by their drug regime and determine drug associated selection upon the viral sequence. In this scheme, our output for the non-synonymous/synonymous substitution ratio in the absence of reversion pressure and HLA associated selection (ω) will encompass all selection not explained by drug associated selection or reversion. For example, selection due to CTL pressure, the antibody response, or the APOBEC3G response. We split the protease and reverse transcriptase sequence and drug regime data

into query and reference sets. we exclude integrase in our analysis, as there was substantially less data available to analyse: integrase sequence data was available from ∼8,000 people, compared to ∼78,000 for both protease and reverse transcriptase. To split the data, we defined the collection of sequences in hosts not receiving any therapy (directed at the region of interest) at time of sequencing as the reference set, and viral sequences in hosts receiving any treatment (directed at the region of interest) coupled with their drug regime data as the query dataset. We then randomly assign 2000 reference sequences to the query dataset. This was to ensure that any drug associated selection common across all drugs was not simply absorbed into estimates of ω, at that site.

7.2.2 *Data preparation and performing inference via MCMC*

Before we could begin our analyses, we first had to perform a series of data cleaning steps to both our reference sequences and query sequences in order to reduce sequence and drug ambiguity for the region under investigation. Full details of procedures are provided in Appendix B.1.1. Priors on parameters are as in Chapter 6, see Table 6.1. The consensus sequence is set as the consensus of the reference dataset. Between 30 and 40 independent MCMC runs were conducted for 5 days for both protease and reverse transcriptase. States were sampled every 200 steps in the chain, and runs were amalgamated by removing the first 100,000 states from each independent run and combining runs. Each independent run achieved convergence as measured by visual inspection. As an example, in Figure B.1 we show ten examples of independent runs for our HLA associated selection study. Convergence for the drug associated selection analysis was similar. This resulted in combined runs of length ∼40,000,000 and ∼12,000,000 for protease and reverse transcriptase respectively.

7.2.3 *Results*

We summarise parameter estimates in Figures 7.1-7.7, using the same method of visualising the data as in Section 6.5. In Figures 7.1-7.3 we display $\gamma_{i,h}$ for each of the seven protease inhibitors for which sufficient data was available for us to perform our analyses. We show ω_i (the non-synonymous/synonymous substitution ratio in the absence of drug associated selection and reversion pressure at site i) across protease in the last sub-figure of Figure 7.3. Consider for example Amprenavir (abbreviated APV, the first sub-figure in Figure 7.1). Recall that under our model $\gamma_{i,h}$ modifies ω_i through multiplication if the mutation is a non-synonymous base change away from the defined consensus codon at that site. Thus, a value of 1 would indicate that the drug is estimated to have no impact upon non-synonymous base changes from consensus, and a value significantly higher than 1 could be a strong candidate for a mutation which confers some reduction in sensitivity to the drug or compensates the fitness cost of such a mutation. A clear example for APV is the strong peak at site 50 for both the mean (black) and median (red) value of $\gamma_{50,APV}$. Looking now at the coloured vertical lines, notice that site 50 coincides with a red vertical line, which indicates a site of a known drug resistance mutation associated with the highest levels of phenotypic resistance and/or with the strongest clinical evidence for interfering with successful protease inhibitor therapy [63]. Orange and dashed orange lines mark sites of drug resistance mutations of decreasing significance: orange lines are at sites at which a mutation is known to reduce susceptibility or virological response to the drug, and dashed orange lines are sites of mutations which result in reduced susceptibility to the drug in the presence of other drug resistance mutations [63]. Finally, green vertical lines mark sites which distinguish the subtype B and subtype C consensus sequence at the amino acid level.

We are interested in sites at which we find a high average value of $\gamma_{i,h}$ which is well supported. Again, looking at the APV sub-figure; these two properties are summarised by the strip below the plotting region. Sites at which the median is greater than 1.5 and 2 are shown in orange and red respectively. If the 2.5% quantile is greater than 1, we outline it in black, and add the site to our

collection of candidate sites for APV associated selection. Site 50 is a clear candidate, but also note that site 84 (which coincides with another red vertical line) is also under significant APV associated selection. There are a number of 'false' negatives and 'false' positives. In the case of APV, an example of a false negative is site 82: a site where we see a red vertical line but no significant increase in APV associated selection. An example of a false positive is site 11. However, in the case of these two examples, if we investigate further we find that although a certain mutation at site 82 confers strong APV resistance, it is not selected for by APV, and that although site 11 is not the site of a major drug resistance mutation [63], it is known to be associated with minor reductions in APV susceptibility [289]. The full list of known drug resistance mutations for protease and reverse transcriptase can be found on the Stanford drug resistance database [63]. Notice that some drugs which we analyse do not have associated red, orange or dashed orange vertical lines. These are drugs for which the data was not available in the same format on the Stanford drug resistance database, either because the drug is no longer commonly used (DDC = zalcitabine, DLV = delavirdine), was an experimental drug that failed (αAPA = alpha-anilinophenylacetamide, ADV = adefovir), or is now used in small amounts with other drugs (RTV = ritonavir). Our sensitivity (under the definition that red and orange vertical lines in Figures 7.1-7.3 are indeed at sites under selection) in detecting the known major drug resistance mutations displayed in Figures 7.1 - 7.8 is shown in Table 7.1. Whilst initially the results may seem quite dissapointing, we note that our model will not attribute drug associated selection to mutations > 1 nucleotide from consensus, and cannot detect insertions. Also, some mutations may result in reduced drug sensitivity to a particular drug, but not be selected by it (as at site 82 for APV discussed above). Indeed, we point to the fact that our estimated drug associated scaling of selection combines both the fitness cost of non-synonymous change away from consensus and the benefit of the associated reduction in drug sensitivity. We therefore suggest that perhaps the plots that we generate are more meaningful in assessing the potential for escape from a particular drug in an *in vivo* setting. The 'gold standard' data used to generate the red and orange lines is a combination of *in vitro* drug resistance assays and clinical data. For this rea-

region	drug resistance site sensitivity	
	strongest major DRMs (red)	major DRMs (orange)
protease	17/27	11/14
reverse transcriptase	18/30	6/20

Table 7.1: Estimates of our sensitivity to major drug resistance sites. Strongest major DRMs (red) are sites at which drug resistance mutation(s) associated with the highest levels of phenotypic resistance and/or with the strongest clinical evidence for interfering with successful therapy [63] reside. Major DRMs (orange) are sites at which a mutation is known to reduce susceptibility or virological response to the drug.

son, we do not investigate specificity of our model. Resistance mutation scores on the Stanford drug resistance database (which determined the strongest major drug resistance mutations and major drug resistance mutations in Table 7.1 and Figures 7.1-7.8) are strongly influenced by studies linking mutations to decreased drug susceptibility *in vitro*: such mutations receive 'significant mutation penalty scores' [63]. Because of this, there is likely to be a lack of specificity in the database for mutations that are selected for by antiretrovirals *in vivo*.

7.2.3.1 *Novel drug associated sites*

In Tables 7.2 and 7.3, we display the collection of sites that we determine to be under significant drug associated selection away from the consensus codon of the reference sequences, but not present in the complete collection of drug resistance mutations on the Stanford drug resistance database [63]. In most cases, such sites are next to known drug resistance mutations, so we expect that they are combinations of compensatory mutations and false positives (which are caught due to the true positive at the neighbouring site). However, there is a smaller collection of sites which do not neighbour known drug resistance sites, the strongest of which is at site 95 in protease associated to IDV; shown in Figure 7.1 row 2, column 1. This site was also found to be under selection by IDV in [220], and we suggest that selection at this site should be investigated further.

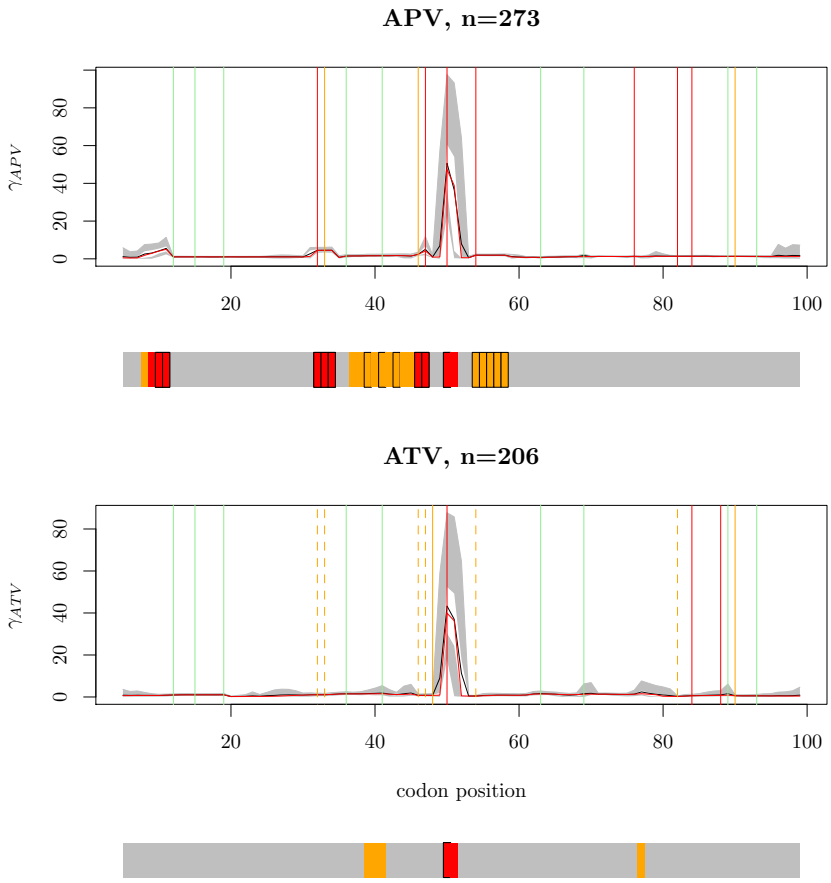


Figure 7.1: Results of protease inhibitor associated selection study. Drugs analysed are those for which > 10 sequences from distinct individuals taking the drug are available. The first seven subplots show drug dependent parameters which scale selection (ω) of non-consensus codons through non-synonymous mutations. The white and grey regions enclose the 50% and 95% credible intervals respectively. The median and mean scaling of ω associated to each drug (averaged over the MCMC run) are plotted in red and black respectively. Green vertical lines are drawn at sites which distinguish HIV-1 subtype B and subtype C at the amino acid level. The remaining coloured vertical lines are known major drug resistance mutations [63] associated to each drug. Red lines are at sites where at least one mutation is known to confer very strong reduction in drug efficacy, orange lines indicate sites where at least one mutation reduces drug susceptibility, and dashed orange lines indicate sites at which a mutation contributes to reduced susceptibility in combination with other mutations. Each subplot is summarised by the band beneath it: sites at which the median exceeds 1.5 and 2 are coloured orange and red respectively. If the 2.5% quantile exceeds 1, the site is outlined in black.
Key: APV = amprenavir; ATV = atazanavir; IDV = indinavir; LPV = lopinavir; NFV = nelfinavir; RTV = ritonavir; SQV = saquinavir.

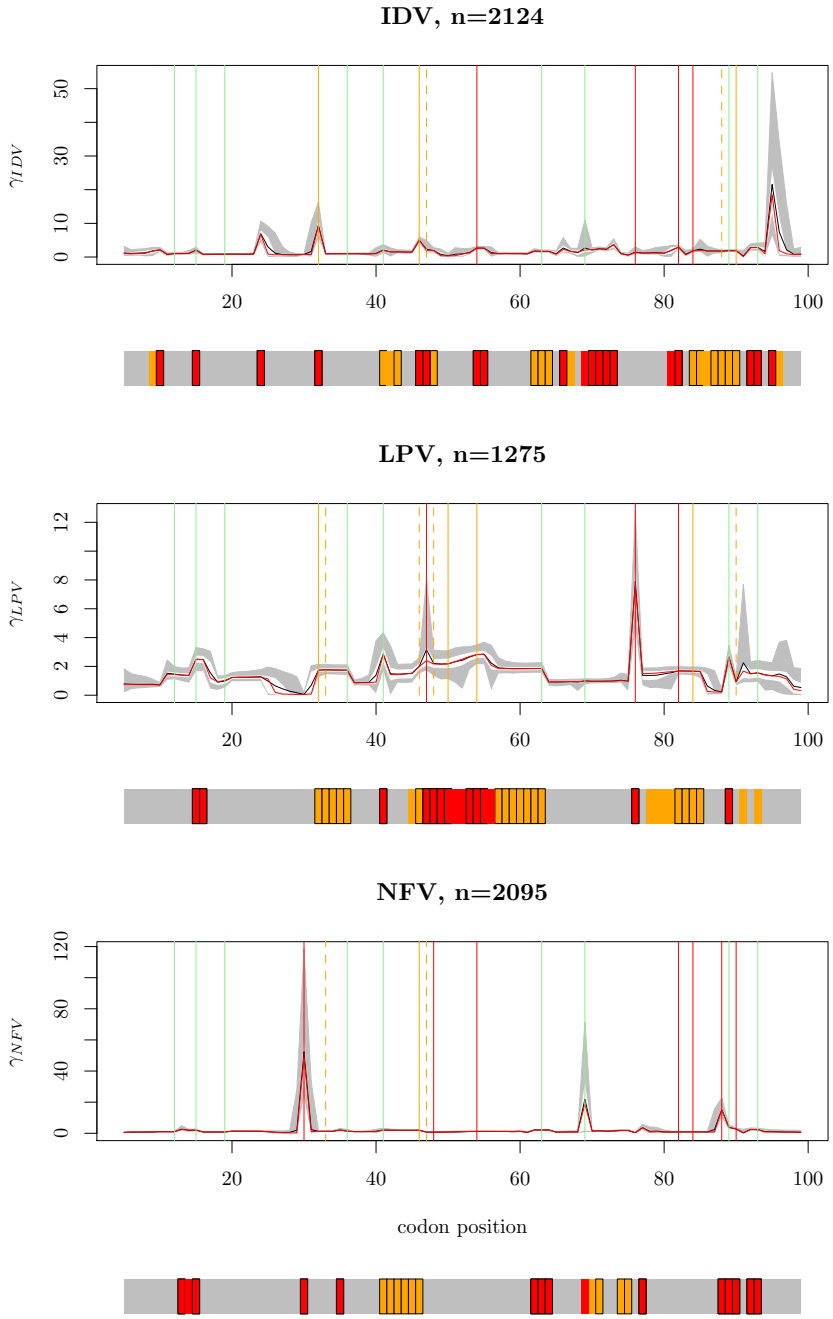


Figure 7.2: Results of protease inhibitor associated selection study (continued). The drugs analysed are those for which > 10 sequences from distinct individuals taking the drug are available. See Figure 7.1 caption for plotting details. Key: APV = amprenavir; ATV = atazanavir; IDV = indinavir; LPV = lopinavir; NFV = nelfinavir; RTV = ritonavir; SQV = saquinavir.

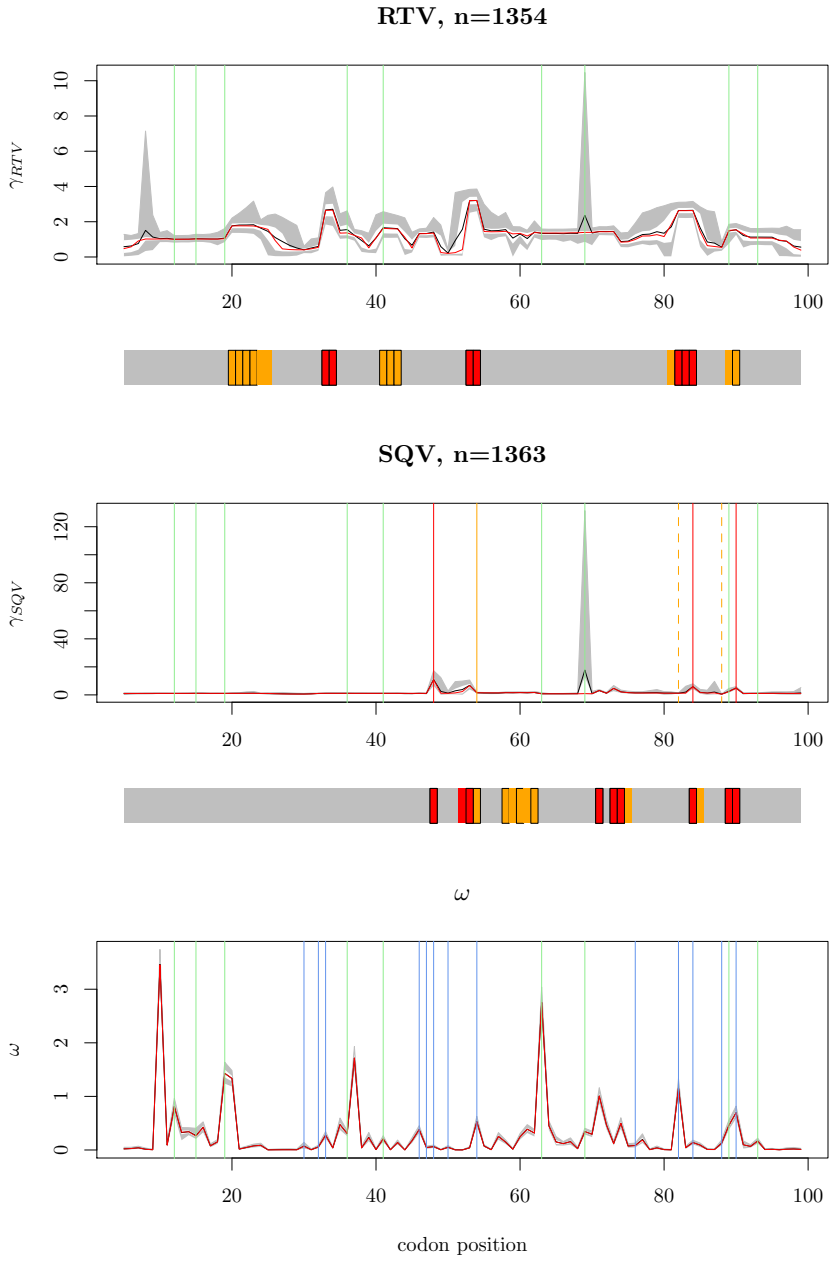


Figure 7.3: Results of protease inhibitor associated selection study (continued). The drugs analysed are those for which $>$ 10 sequences from distinct individuals taking the drug are available. See Figure 7.1 caption for plotting details of the first two plots.
The final subplot shows estimates of ω across protease. Blue vertical lines denote any major drug resistance mutations associated to a protease inhibitor.
Key: APV = amprenavir; ATV = atazanavir; IDV = indinavir; LPV = lopinavir; NFV = nelfinavir; RTV = ritonavir; SQV = saquinavir.

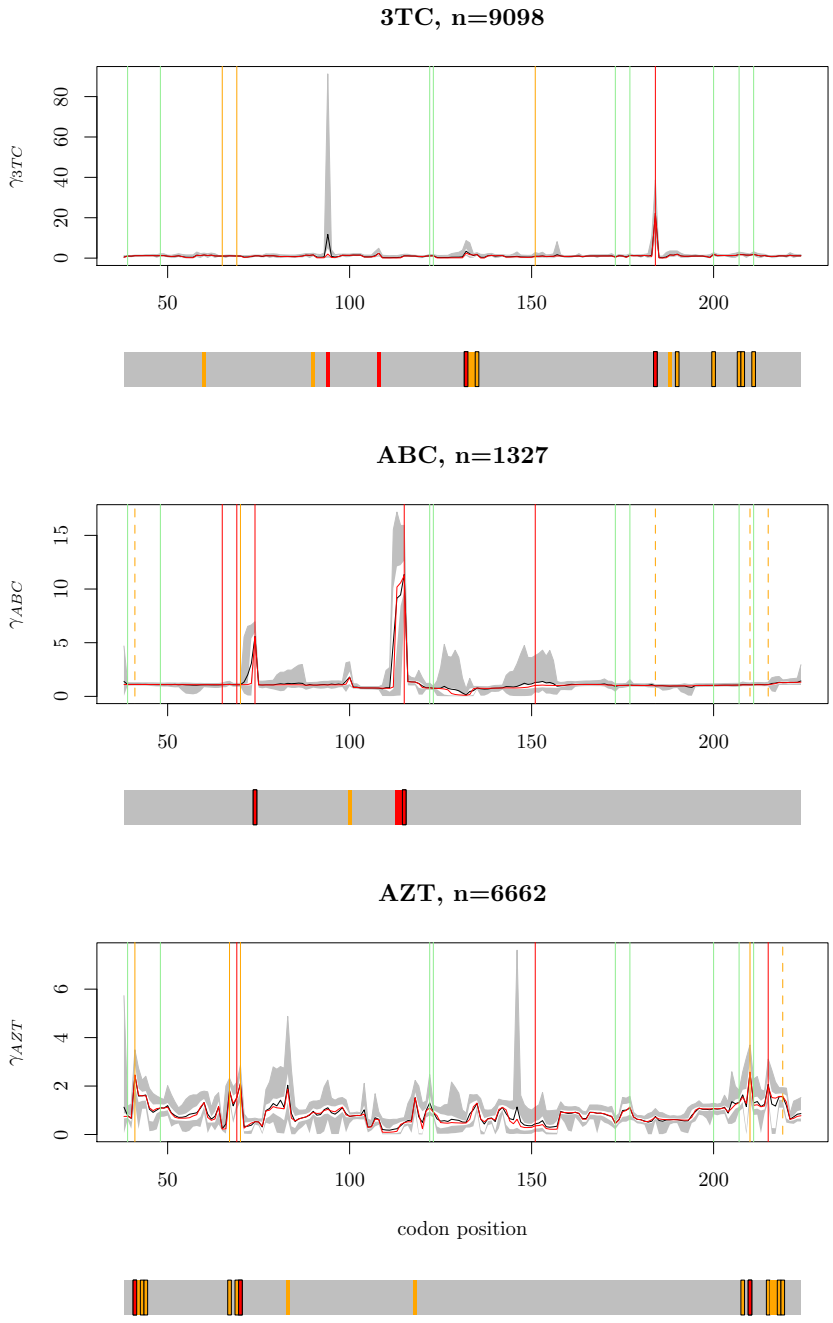


Figure 7.4: Results of reverse transcriptase inhibitor associated selection study: NRTIs. The drugs analysed are those for which > 10 sequences from distinct individuals taking the NRTI are available. See Figure 7.1 caption for plotting details.
Key: 3TC = lamivudine; ABC = abacavir; AZT = zidovudine; D4T = stavudine; DDC = zalcitabine; DDI = didanosine; FTC = emtricitabine; TDF = tenofovir.

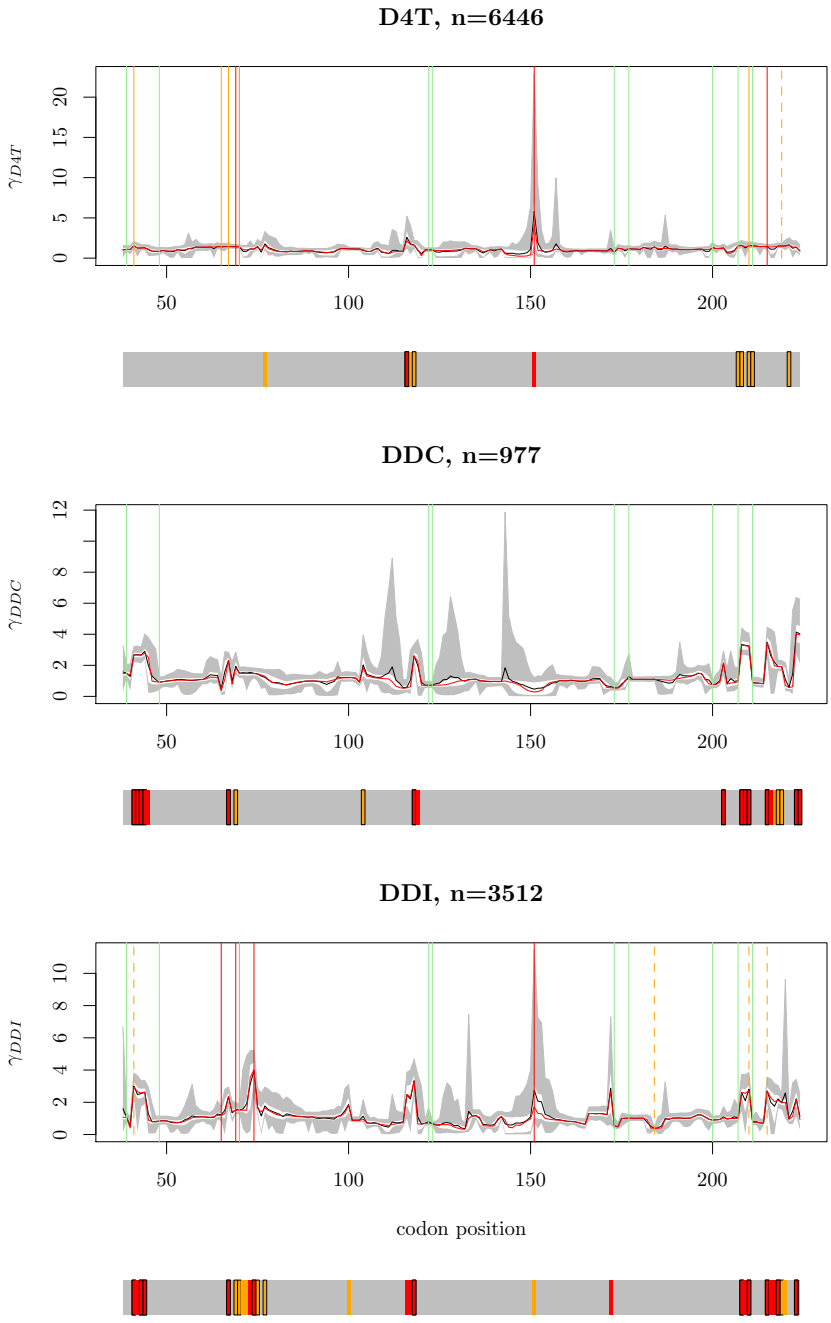


Figure 7.5: Results of reverse transcriptase inhibitor associated selection study: NRTIs (continued). The drugs analysed are those for which > 10 sequences from distinct individuals taking the NRTI are available. See Figure 7.1 caption for plotting details.
Key: 3TC = lamivudine; ABC = abacavir; AZT = zidovudine; D4T = stavudine; DDC = zalcitabine; DDI = didanosine; FTC = emtricitabine; TDF = tenofovir.

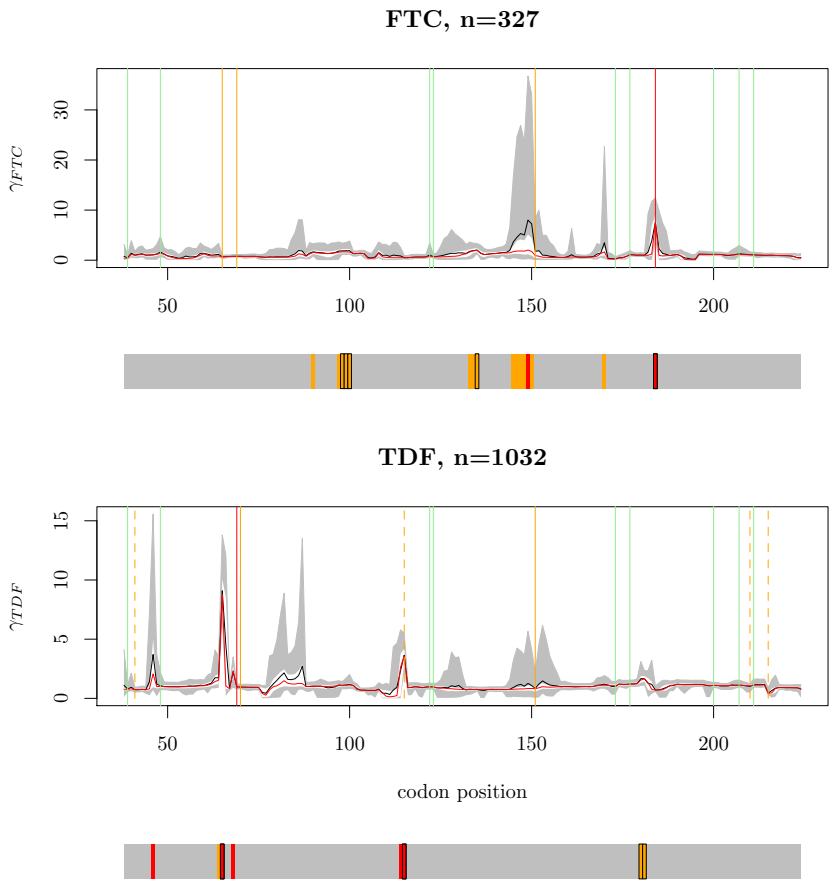


Figure 7.6: Results of reverse transcriptase inhibitor associated selection study: NRTIs (continued). The drugs analysed are those for which > 10 sequences from distinct individuals taking the NRTI are available. See Figure 7.1 caption for plotting details.
Key: 3TC = lamivudine; ABC = abacavir; AZT = zidovudine; D4T = stavudine; DDC = zalcitabine; DDI = didanosine; FTC = emtricitabine; TDF = tenofovir.

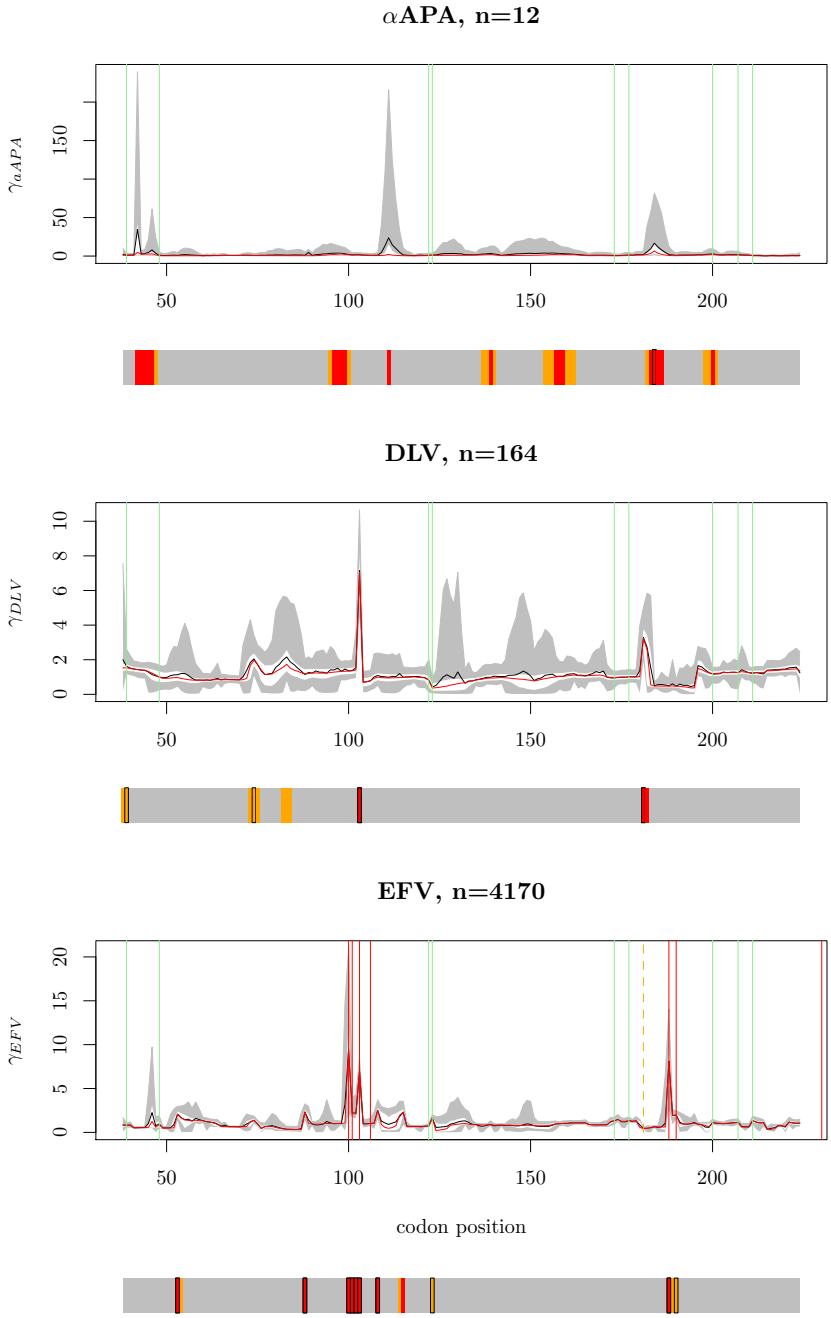


Figure 7.7: Results of reverse transcriptase inhibitor associated selection study: NNRTIs. The drugs analysed are those for which > 10 sequences from distinct individuals taking the NNRTI are available. See Figure 7.1 caption for plotting details.
Key: αAPA = alpha-anilinophenylacetamide; ADV = adefovir; DLV = delavirdine; EFV = efavirenz; NVP = nevirapine.

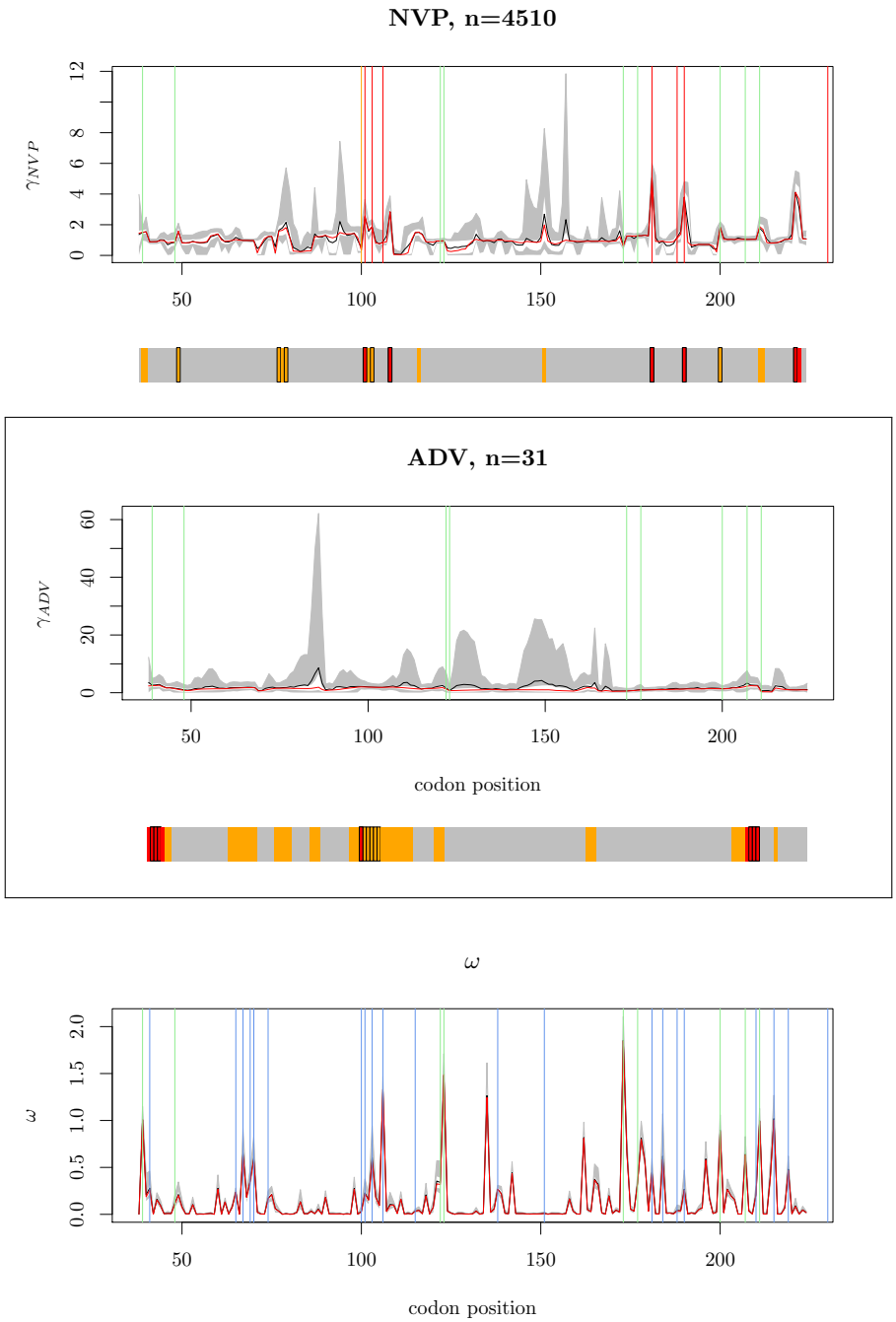


Figure 7.8: Results of reverse transcriptase inhibitor associated selection study: NNRTIs (continued) and NtRTIs. The drugs analysed are those for which > 10 sequences from distinct individuals taking the NNRTI are available. See Figure 7.1 caption for plotting details of the first two sublplots. ADV is boxed as it is an ntRTI.
The final subplot shows estimates of ω across protease. Blue vertical lines denote any major drug resistance mutations associated to a protease inhibitor.
Key: αAPA = alpha-anilinophenylacetamide; ADV = adefovir; DLV = delavirdine; EFV = efavirenz; NVP = nevirapine.

	Protease inhibitor						
codon	ATV	APV (FPV)	IDV	LPV	NFV	RTV	SQV
15			✓	✓	✓		
16							
21							
22							
34		✓				✓	
39							
41		✓	✓	✓	✓	✓	
42							
44					✓		
45					✓		
49							
55		✓	✓	✓			
56		✓					
57		✓		✓			
59							
60				✓			✓
61							
62			✓	✓	✓		✓
64			✓		✓		
66			✓				
70							
72			✓				
75					✓		
87							
92			✓		✓		
95			✓				

Table 7.2: Summary of significant unknown drug associated mutations in protease. Drug columns highlighted in grey do not have resistance mutation information available in the Stanford drug resistance database [63].

	NRTIs								NNRTI				ntRTI
codon	3TC	FTC	ABC	DDI	TDF	D4T	AZT	DDC	NVP	EFV	αAPA	DLV	ADV
39												✓	
42								✓					
43				✓			✓	✓					
49									✓				
53										✓			
79									✓				
88										✓			
99		✓											✓
102										✓			✓
104								✓					
108									✓	✓			
123										✓			
132	✓												
135	✓	✓											
180									✓				
200	✓								✓				
203								✓					
207	✓					✓							
208	✓			✓		✓	✓	✓					✓
209								✓					✓
210													✓
211						✓							
218				✓			✓	✓					
223				✓				✓					
224								✓					

Table 7.3: Summary of significant unknown drug associated mutations in reverse transcriptase. Drug columns highlighted in grey do not have resistance mutation information available in the Stanford drug resistance database [63].

7.3 HLA ASSOCIATED SELECTION

In this portion of our analysis we look for increased selection associated to HLA types. In our model this is encoded in our $\gamma_{i,h}$ parameters. Recall that these parameters scale the non-synonymous/synonymous substitution ratio away from consensus in the presence of HLA type h. Thus, $\gamma_{i,h} > 1$ means that HLA type h is associated with increased selection away from the consensus codon at site i. Again we focus our attention on protease and reverse transcriptase for the same reasons as outlined in Section 7.2. For our HLA association study, we combine all available sequence data from the publicly accessible HIV-1 sequence databases to generate our reference sequences $\mathbf{D}_B$ and query datasets. For our query datasets we require paired data consisting of host HLA types $\mathcal{H}$, and corresponding within host consensus viral sequence data $\mathbf{D}$. A large number of sequences in $\mathbf{D})B$ yields short coalescence times between any member of $\mathbf{D}$ and $\mathbf{D}_B$, increasing our power to estimate parameters.

7.3.1 *Sources used for our reference datasets*

- Los Alamos HIV sequence database: Collection of HIV sequences, fed by bi-weekly downloads of new HIV sequence data submitted to GenBank, and as such contains much of the published HIV sequence data for which the sequence data was made publicly available. Meta-data is extracted from GenBank submissions (e.g. submission date, subtype, country of isolation, HLA type of host etc), and manually annotated with additional information taken from the corresponding publications and through direct interaction with the authors [170].
- Stanford Drug resistance database: Collection of protease and reverse transcriptase sequences with associated patient drug prescription data. Further details are provided in Section 7.2.1.

- UCLA HIV positive selection mutation database: Protease and reverse transcriptase sequences taken from HIV-1 patient plasma samples by Specialty Laboratories from 1999 to 2002 [47]. The database was created in order to identify regions of drug selection using estimates of non-synonymous/synonymous base changes along the viral sequence.

7.3.2 *Summaries of paired host HLA type and viral sequence data used for query datasets*

- **Swiss Spainish intermittent treatment trial (SSITT) dataset** [73, 82]: Swiss portion of this dataset. Region: various cities across Switzerland, $n = 79$, sampling date: 2000, treatment: HAART.
- **Durban dataset** [146, 145]: Region: Durban, South Africa, $n = 1268$, sampling date: 1999-2006, treatment: ART naive.
- **Mma Bana (mother and child) study** [257]: Adult portion of this study. Region: Gaborone, Botswana, $n = 428$, sampling date: 2007, treatment: ART naive.
- **Bloemfontein dataset** [123, 122]: Region: Bloemfontein, South Africa, $n = 261$, sampling date: February-September 2006, treatment: ART naive.
- **The short pulse antiretroviral therapy at seroconversion (SPARTAC) dataset** [4]: UK portion of this dataset. Region: UK, $n = 331$, sampling date: 2003-2007, study requirements: primary infection: though definition of this is complex, see the SPARTAC trial investigators [4] for details.
- All remaining sequences with associated host HLA data available in the Los Alamos HIV sequence database [256] which were not present in the above studies. Protease $n = 404$, reverse transcriptase $n = 327$.

7.3.3 *Data preparation*

As in Section 7.2, we first had to perform a series of data cleaning steps to reduce sequence ambiguity across the region under investigation, and to reduce repetition of sequences as much as possible when extracting data from multiple sources. We also used a series of scripts to amalgamate the various HLA notations, and truncate HLA data to serological resolution. Detailed descriptions of our sequence and HLA dataset preparation are provided in Appendix B.1.2. On completion of data preparation, for each of the HLA class I genes for both protease and reverse transcriptase, we have a query and reference dataset. In each, overlaps and repeated patient identifiers were removed such that each sequence in $\mathbf{D}_B \cup \mathbf{D}$ has a unique patient identifier (assuming unique patient identifiers $\Rightarrow$ unique patients). Initially, we set the consensus sequence $\mathscr{C}$ (using the notation of Chapter 6) as the consensus of the subtype B query sequences (as defined using the program RIP [229]). We chose to analyse each of the HLA class I genes separately. Our reasoning behind this decision was that many of the individuals for which some HLA data was available were not typed at all three class I loci, and we sought to retain power to infer parameters. However, we note that taking this approach can (and does) lead to spurious signals of association through linkage disequilibrium between class I alleles across the three genes. We address this *post hoc*. Priors on parameters are as in Chapter 6, see Table 6.1. 15 independent MCMC runs were performed for 5 days.

As part of our analyses, we wish to discern whether there is evidence of selection conditional on the HLA allele frequencies present in different populations. In order to investigate this, we run a further collection of MCMC runs, setting the consensus $\mathscr{C}$ as the consensus of the subtype C sequences in our query dataset (again, as defined by RIP [229]). This will allow us to examine whether there is a push away from a given subtype consensus (B or C) conditional on a particular HLA allele distribution, see Section 7.3.9. Note that we only consider $\mathscr{C} =$ subtype B and C, as these are by far the most common HIV-1 subtypes present in our query sequences $\mathbf{D}$. To check that any effects that we observe aren't the

result of grouping together viral subtypes in our query dataset, we also run a further four collections of MCMC runs for each region:

1. $\mathbf{D}$ = subtype C query sequences,
   $\mathscr{C}$ = consensus of subtype B query sequences.
2. $\mathbf{D}$ = subtype C query sequences,
   $\mathscr{C}$ = consensus of subtype C query sequences.
3. $\mathbf{D}$ = subtype B query sequences,
   $\mathscr{C}$ = consensus of subtype B query sequences.
4. $\mathbf{D}$ = subtype B query sequences,
   $\mathscr{C}$ = consensus of subtype C query sequences.

For each collection, we conduct 10 independent MCMC runs for 5 days. In each case, we sample every 500 states and remove a burn-in of 100,000 states, before combining the runs. All independent runs achieved convergence as measured by visual inspection (see Figure B.1 for 10 examples of γ_{B^*07} run summaries), and the resultant amalgamated runs were $> 25,000,000$ states long.

7.3.4 *Checking for potential sources of bias*

Whilst carrying out this work we were acutely aware of history in the field of mistaking geographical HLA structure in the data for a signal of association between HLA type and viral sequence evolution [189]. We highlight and investigate two potential sources of bias and check if such effects are present in our filtered query and reference datasets:

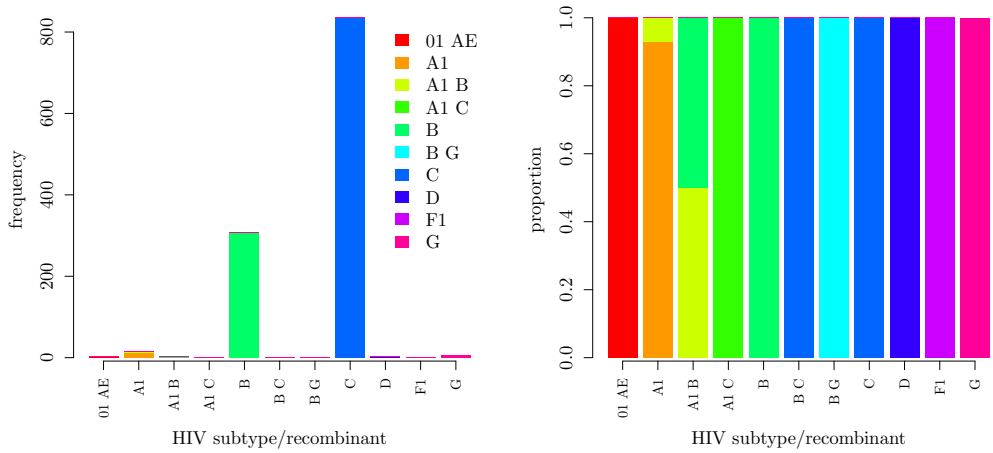
- **Highly connected sequences**. To increase computational efficiency we use a subset of the reference sequence set to estimate the likelihood for each query sequence, see Section 6.4. Given this approach, a potential confounder would be a single or small collection of reference sequences which are close to a large number of query sequences by Hamming distance (the metric we use to define sequences of importance on which to determine the likelihood). Such sequences would have a strong impact upon the likelihood

and therefore have the potential to dramatically alter parameter estimates. To check, we determine the closest reference sequence to each member of our query sequences. We find that in the collection of reference sequences closest to at least one query sequence, ∼80% in both protease and reverse transcriptase were closest to a single sequence in the query dataset. >95% were closest to ≤3 query sequences, and the maximum such number for any reference sequence was 14 and 15 in protease and reverse transcriptase respectively (a single sequence in each case). We therefore conclude that this possible source of bias is not an issue in our datasets.

- **Query sequences which are closest to reference sequences of a different subtype**. If this is the case, then we would see false signals of selection associated to sites which distinguish between HIV-1 subtypes. To check this, we identify the HIV-1 subtype for each query sequence [229] and then determine HIV-1 subtype of the closest reference sequence by Hamming distance, and compare the two. The resulting plot for HLA-B reverse transcriptase queries and their corresponding closest sequence is shown in Figure 7.9 (results for the other class I genes in reverse transcriptase and protease are similar, and omitted). In each case >99% of closest sequences are of the same subtype (unless it is a circulating recombinant form, in which case the closest subtype is one of the subtype recombinants). Thus, we conclude that this potential source of bias is not present in our datasets.

7.3.5 *Results*

We address the questions posed in Section 7.1. First we look to see if we detect well known epitopes. We then examine lesser known and putative epitopes: we quantify overlap between collections of candidate sites for HLA associated selection (as determined by our model) and epitope sets either described in the literature or determined by known binding motifs. We look for sites at which drug associated selection and HLA associated selection coincide, and then move on to check for evidence of dissimilarity of HLA allele frequencies between continents



(a) (b)

Figure 7.9: Subtype of closest sequences by Hamming distance of each subtype represented in the reverse transcriptase query sequences for which HLA-B alleles were determined in the host. Figure 7.9a shows the absolute number of subtypes in the query sequence dataset, with bars coloured by the subtype of the corresponding closest sequence in the reference dataset by Hamming distance. 7.9b shows the corresponding distribution of subtypes of the closest sequences in the reference dataset as measured by Hamming distance. Note that here, subtype A and F are further split and the sub-subtypes A1 and F1 are represented. Also, a small number of the CRF 01 AE (which was previously known as subtype E [241]) and other recombinants.

playing a role in driving the amino acid differences we see between HIV-1 subtypes. Finally, we look for potential sites of HLA associated selection currently not present in the Los Alamos Immunology database [127].

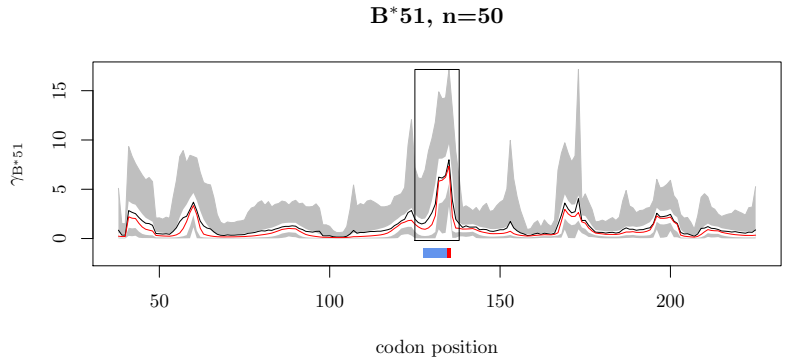
7.3.6 *Famous epitopes: Do we detect strong selection in famous HLA associated epitopes in the analysed portions of the viral genome?*

The density of known epitopes across *pol* is low relative to *gag* and *nef*. However, there are several epitopes in *pol* which can be used to test our approach. Described epitopes are categorised into lists, known as the 'A-list' and the 'B-list'. The B-list refers to the entire collection of epitopes reported in the literature [205]. The A-list is a collection of the best-defined HIV-1 CTL epitopes, according to Brander *et al.*, which is updated yearly [173]. Amongst the A-list epitopes, a subset has documented variants which result in escape (or partial escape) from the CTL immune response. In this chapter we define an 'escape' site as either:

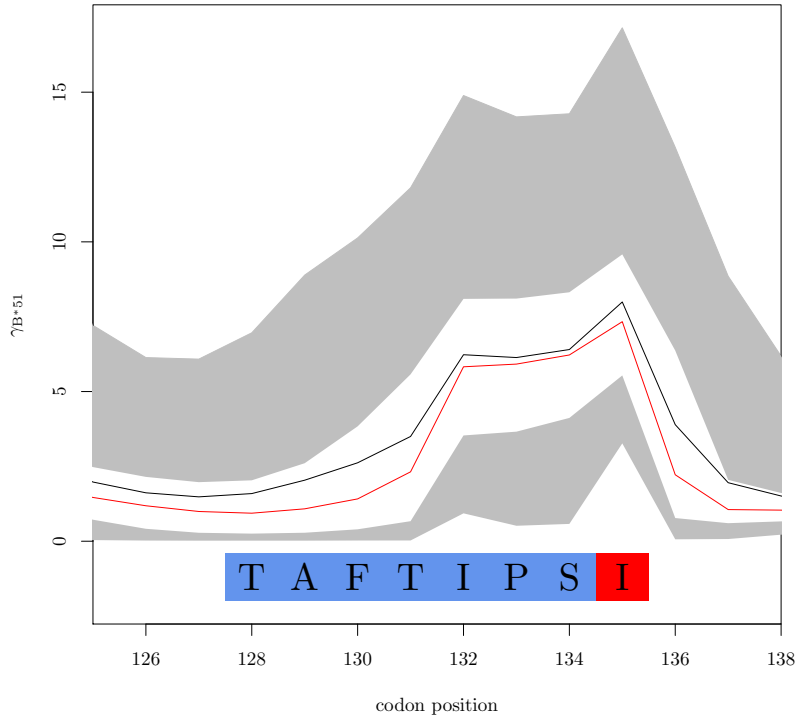
- A site reported as an escape.
- A site at which a mutation reduces CTL binding.
- A site at which a mutation results in a non-susceptible form (no CTL response when patient cells are challenged with the variant peptide).

We summarise this subset for the regions under analysis in Table 7.4, and name this list 'famous epitopes'.

Most notable among our famous epitopes is the B*51 restricted epitope `TAFTIPSI`, at sites 128-135 of reverse transcriptase. Figure 7.10a shows B*51 associated scaling of selection across reverse transcriptase, with `TAFTIPSI` outlined. If we zoom in on the epitope in Figure 7.10b, we find that despite the relatively modest number of B*51 positive individuals ($n = 50$), there is a signal of selection within the epitope and the escape site at position 135 (which we highlight in red) experiences the strongest scaling of selection. Whilst a variant at site 129 is known to abrogate CTL recognition, it is known to be highly conserved *in vivo* [185].



(a)



(b)

Figure 7.10: B*51 associated selection within TAFTIPSI using results from the MCMC run using all query data, and setting consensus as subtype C. Figure 7.10a shows the scaling of ω associated to B*51 within the available portion of reverse transcriptase after data filtering, γ_{B^*51}. TAFTIPSI is boxed. Figure 7.10b zooms in to TAFTIPSI and we align the epitope with the plot of γ_{B^*51}. The site of the escape mutation is under the strongest HLA associated selection pressure. The red line tracks the median value of γ_{B^*51} over the portion of reverse transcriptase, and the black line tracks the mean. The 50% credible interval is defined by the central white band, and the grey region coupled with the white is the 95% credible interval. $n = 50$ is the number of individuals in the query dataset with B*51. The TAFTIPSI epitope is lined up in both, with the escape site highlighted in red.

region	epitope	sites	restricting HLA	escape site(s)
protease	LVGPTPVNI	76-84	A*0201	7 [134]
	EEMNLPGRW	34-42	B44	2 [192, 77]
reverse transcriptase	ALVEICTEM	33-41	A*0201	3 [134]
	YTAFTIPSV	127-135	A2	9 [62]
	AIFQSSMTK	158-166	A*0301	1&3 [303], 3&9 [185], 4 [303], 5 [35], 5&9 [185], 8 [10], 9 [156, 185]
	AIFQSSMTK	158-166	A*1101	9 [189], 5&9 [185], 3&9 [185]
	TAFTIPSI	128-135	B*5101	2 [185], 9 [82, 94, 172, 185, 263, 189, 92, 138]
	NETPGIRYQY	137-146	B18	2 [146, 31] , 6 [31]
	SPAIFQSSM	156-164	B7	7 [29, 31]
	HPDIVIYQY	175-183	B*3501	3 [189, 137], 5 [251, 185]
	IEELRQHLL	202-210	B*4001	1&6 [62], 6 [62]

Table 7.4: Famous epitopes. These are the collection of A-list epitopes for which a described escape variant exists in the literature.

In Section 7.3.6.2 we step through the remaining famous epitopes, but first we describe the information contained in Figures 7.11 - 7.12 and B.2 - B.4.

7.3.6.1 *Summary figures for selection parameters γ_h and ω*

In order to display all of our parameter estimates of HLA associated selection in a single plot for each region in Figures 7.11 - 7.12, 7.12 - B.5, we summarise plots such as that shown in Figure 7.10a. These are plotted as grey strips. There are a maximum of six strips for each HLA type. Each of the strips summarises a collection of MCMC runs, with the following ordering:

1. $\mathbf{D}$ = all query sequences,
   $\mathscr{C}$ = consensus of subtype B query sequences.
2. $\mathbf{D}$ = all query sequences,
   $\mathscr{C}$ = consensus of subtype C query sequences.
3. $\mathbf{D}$ = subtype C query sequences,
   $\mathscr{C}$ = consensus of subtype B query sequences.

4. **D** = subtype C query sequences,
   $\mathscr{C}$ = consensus of subtype C query sequences.
5. **D** = subtype B query sequences,
   $\mathscr{C}$ = consensus of subtype B query sequences.
6. **D** = subtype B query sequences,
   $\mathscr{C}$ = consensus of subtype C query sequences.

The grey colouring of the strips alternate between light and dark as we switch between HLA types to help guide the eye. If a band is excluded, then there were insufficient query sequences ($n < 10$) within hosts with that HLA type. Numbers of query sequences residing in hosts with each HLA type are shown as a barplot on the right hand side of the figures. On each band, putative selected sites are overlaid in colour. If the median $\gamma_{i,h} > 2$, we colour the site red, if it is > 1.5 we colour the site orange (as in Section 7.2). The lowest six strips in the plot summarise the median value of ω (the non-synonymous/synonymous substitution ratio in the absence of HLA dependent selection and reversion pressure) for each of the six collections of runs. They are coloured according to the key at the foot of the figures. Finally, blue and green vertical lines annotate major drug resistance mutations and amino acid differences between HIV-1 subtype B and C consensus.

In this section we will determine whether we find increased selection away from consensus in known or putative epitope collections, and in famous epitopes. In order to test this, we require a method to define candidate sites arising from these analyses.

We define top-tier candidate sites associated to an HLA as those where both the median value of $\gamma_{i,h}$ taken from the MCMC run is ≥ 2 and the lower 2.5% quantile lies above 1. These sites are outlined in black. We define second-tier candidate sites associated to an HLA type as those at which the lower 10% of $\gamma_{i,h}$ lies above 1. We outline these sites in white. Finally, if we determine HLA associated selection at a site in two HLA types found to be in linkage disequilibrium ($p < 0.05$ (after Bonferroni correction) and $r^2 > 0.1$) we mark that site with a black square below the row. Consider selection associated to B*51 as an example.

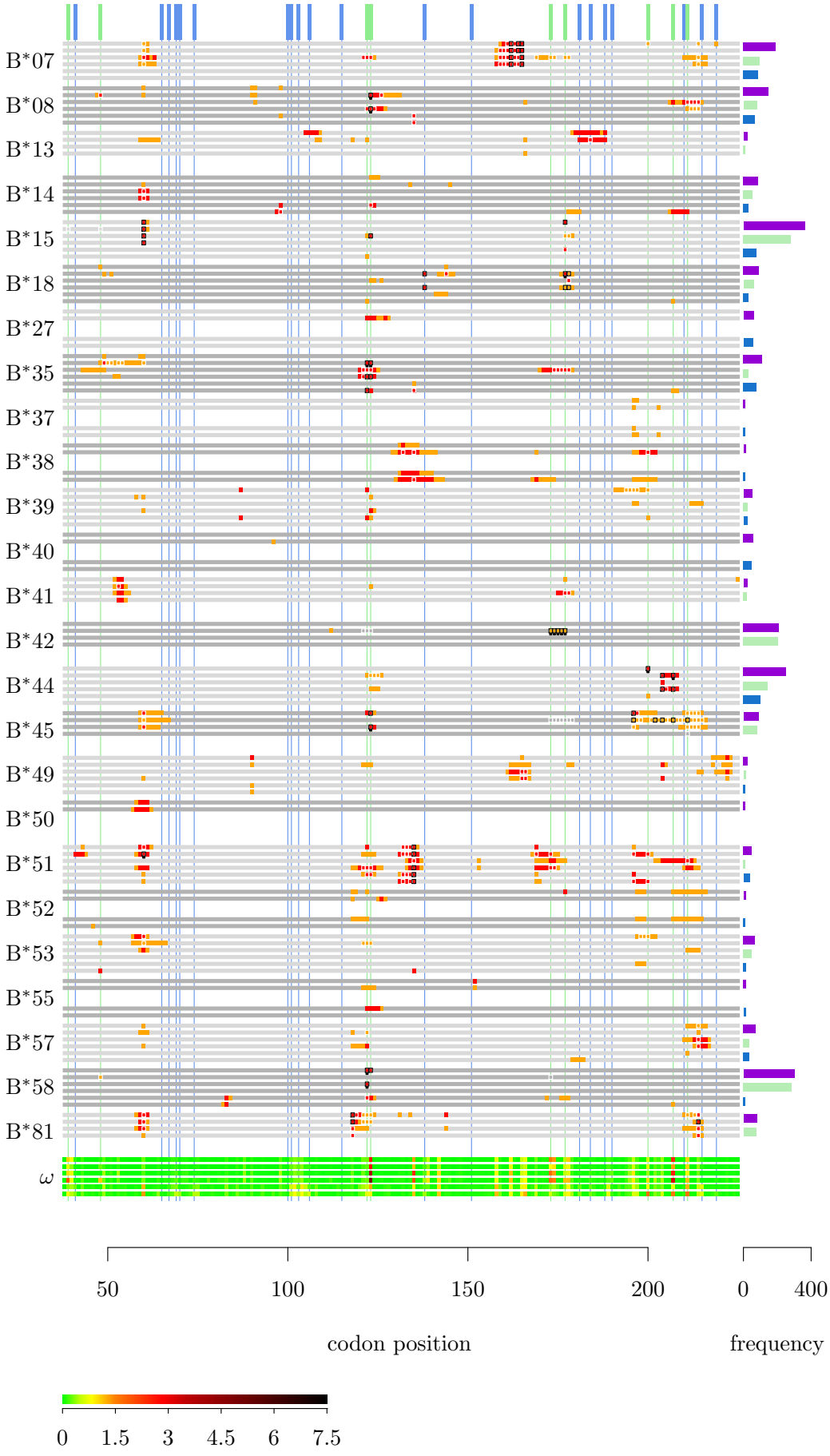


Figure 7.11: Summary of HLA-B associated selection in reverse transcriptase. See Section 7.3.6.1 for full details of the plot.

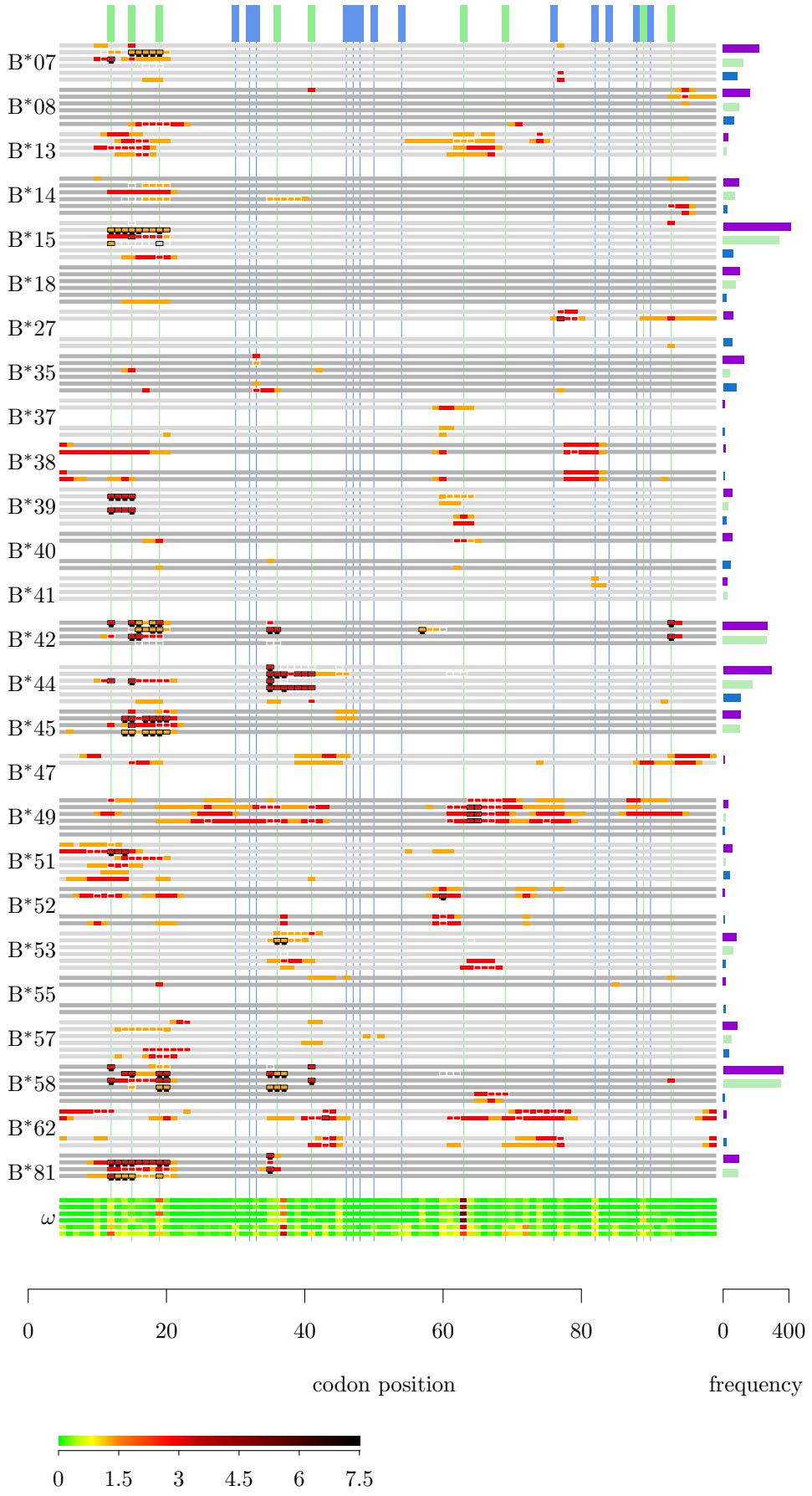


Figure 7.12: Summary of HLA-B associated selection in protease. See Section 7.3.6.1 for full details of the plot.

The resulting band in Figure 7.11 summarising 7.10a is the second of the B*51 strips. In the Sections 7.3.6.2 - 7.3.8 we refer to results of our analysis using all query sequences and $\mathscr{C}$ = subtype C consensus unless otherwise stated. This is because the majority of our query sequences are subtype C, and so selection away from this nucleotide consensus will lead to the strongest signal of HLA dependent selection.

7.3.6.2 *Famous epitopes other than* TAFTIPSI

We now examine the remaining epitopes in our famous list, starting with HLA-B epitopes in reverse transcriptase. See Figures 7.11 and 7.13 for a summary of our results. In Figure 7.13, putative CTL escape sites (according to the definition in Section 7.4) within HLA-B associated A-list epitopes are highlighted in red. Notice that many A-list epitopes are not variable, and so do not survive into our 'famous' list.

*B*07 epitope* SPAIFQSSM *(sites 156-164 in reverse transcriptase).*
We detect strong B*07 dependent selection at the known escape site (position 162) [29, 31]. We also find two more top-tier sites and several second-tier sites within the epitope. On further investigation of the literature we find that many variants exist for this epitope which result in non-susceptible forms in some individuals and susceptible forms in others [121].

*B*18 epitope* NETPGIRYQY *(sites 137-146 in reverse transcriptase).*
We detect strong B*18 dependent selection at the first anchor residue of NETPGIRYQY (site 138) [146, 31].

*B*35 epitope* HPDIVIYQY *(or* NPDIVIYQY*, sites 175-183 in reverse transcriptase).*
We do not detect B*35 dependent selection in this epitope when considering all of our query sequences. However, look at the third B*35 band in Figure 7.11. We find second-tier sites within HPDIVIYQY when considering only subtype C sequences in our query dataset and setting $\mathscr{C}$ as the subtype B consensus sequence.

*B*40 epitope* `IEELRQHLL` *(sites 202-210 in reverse transcriptase).*

We do not detect B*40 dependent selection in `IEELRQHLL`. B*40 is a relatively rare HLA type in our query datasets which is particularly rare in hosts with subtype B and C viral sequences (<10 sequence/host HLA pairs have B*40 for each viral subtype collection), thus any signal would likely come from the remaining subtypes represented in the query dataset, which may be abrogated due to subtype consensus differences.

*B*44 epitope* `EEMNLPGRW` *(sites 34-42 in protease).*

Moving on to protease in Figures 7.12 and 7.14, we detect a strong signal of selection within the B*44 epitope `EEMNLPGRW`. This signal is particularly strong at the first anchor residue (site 35) where escape is known to occur [192, 77].

We now look at the famous HLA-A epitopes, again starting with reverse transcriptase. See Figures B.4 and B.6 (note there are no HLA-C epitopes in our famous list):

*A*02 epitope* `YTAFTIPSV` *(sites 127-135 in reverse transcriptase).*

We do not detect A*02 associated selection in `YTAFTIPSV`. The subtype B and C consensus at the escape site (position 135) is the defined escape [62], explaining why we do not detect a signal. However, we notice that a signal at the site is picked up when only considering subtype B viruses, see the fifth and sixth A*02 strips of Figure B.4. This perhaps suggests that other mutations at the same residue can also result in a diminished CTL response.

*A*03 epitope* `AIFQSSMTK` *(sites 158-166 in reverse transcriptase).*

Escape is known to occur at the second anchor residue (site 166) [156, 185]. This site is in our collection of second tier candidates when setting $\mathscr{C}$ = subtype C consensus, see Figure B.4 and B.8. However, we do not detect any A*03 associated selection elsewhere in the epitope.

*A*11 epitope* `AIFQSSMTK` *(sites 158-166 in reverse transcriptase).*

We do not detect A*11 dependent selection in this epitope. A*11 is a relatively rare HLA type in our query dataset which is virtually absent in hosts with sub-

type B or C viruses (<10 individuals for each), and so for the same reasons as for B*40, we expect any potential signal to be weak.

*A*02 epitope* `LVGPTPVNI` *(sites 76-84 in protease).*
Finally, looking at Figures B.2 and B.7 we see that A*02 dependent selection is not detected in `LVGPTPVNI`. Interestingly, some literature suggests that the wild-type sequence is not recognised, and a drug induced mutation at the escape site (position 35) results in epitope recognition [273]. These conflicting signals would go some way to explaining why we do not detect A*02 associated selection at the known escape site.

Together with our drug analysis and simulated data study in Section 6.5, these findings further demonstrate the ability of our method to accurately recapitulate some sites known to be under HLA associated selection.

7.3.7 *Putative epitopes: Do we find that our method estimates sites of selection to be within epitopes?*

Beyond these well known epitopes, we wish to determine whether we are capturing selection at sites that lie within known or likely epitopes. To investigate this, we perform a series of tests of overlap between our top-tier and second-tier sites with collections of regions thought to be epitopes with increasing confidence. We use two methods to determine p-values and compare our top-tier and second-tier candidate sites to these putative epitope collections. The three epitope collections are obtained in the following ways:

1. **Motif scan**: Using known binding motifs [178] of anchor residues where both residues bind strongly, we scan along the region to detect candidates for epitopes. We allow epitopes to range from eight to twelve amino acids in length.

2. **B-list**: The collection of B-list epitopes [205] for which a restricting HLA type is known.

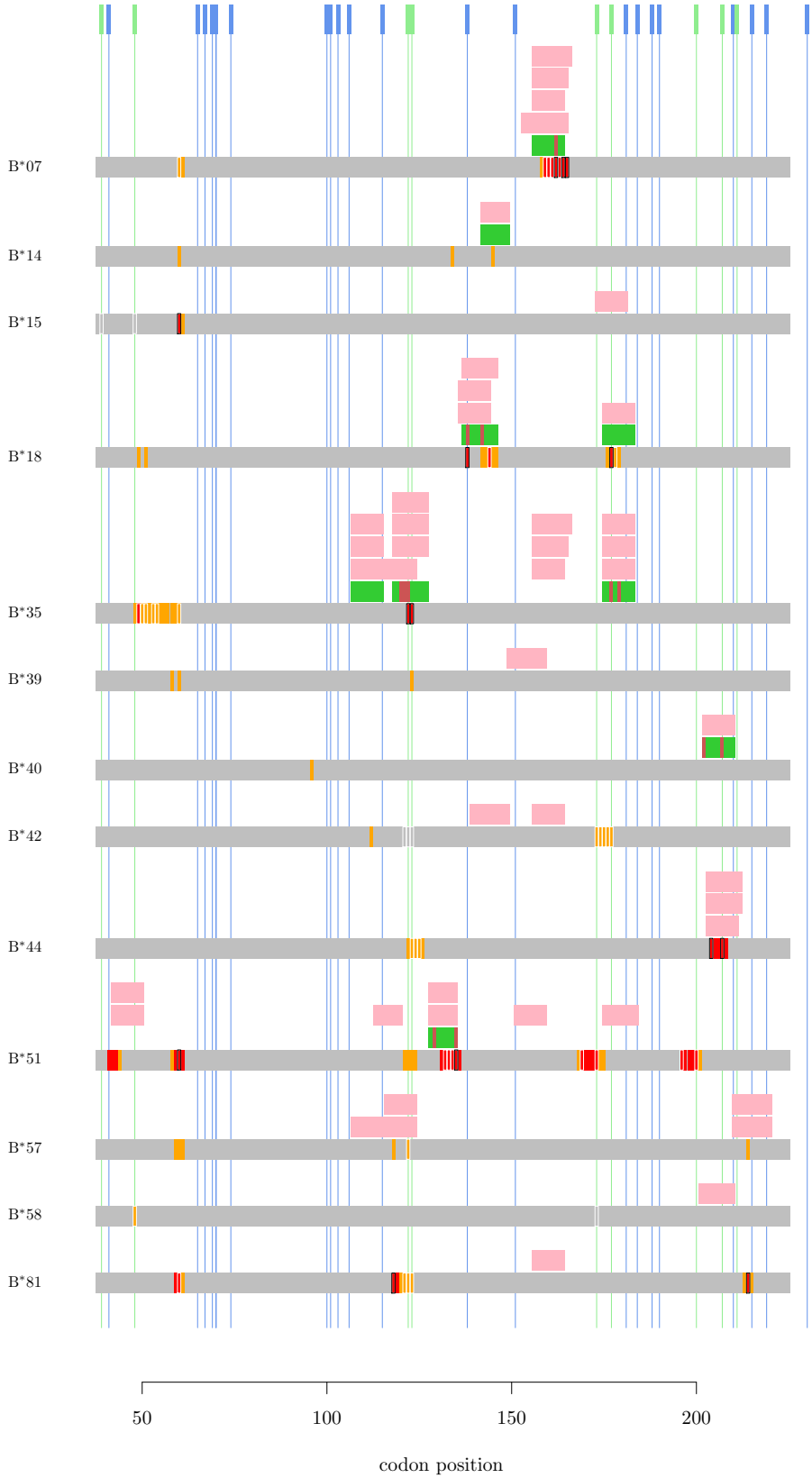


Figure 7.13: A-list and B-list epitopes across reverse transcriptase: HLA-B alleles. Epitope plot showing epitopes associated to HLA-B alleles as defined by the A-list [174] and B-list [205]. Grey bars denote the region analysed, and are overlaid with colours based on our MCMC results as in Figure B.2. Blue vertical lines denote drug resistance sites. Green vertical lines denote sites at which there is an amino acid difference between the subtype B and subtype C consensus viral sequence. Colouring of epitopes refers to the A-list and B-list which are coloured green and pink respectively. In addition, we colour sites in the A-list epitopes where known escape variants occur in red. The grey strips correspond to line two of Figure 7.11 ($\mathscr{C} = C$, all query sequences)

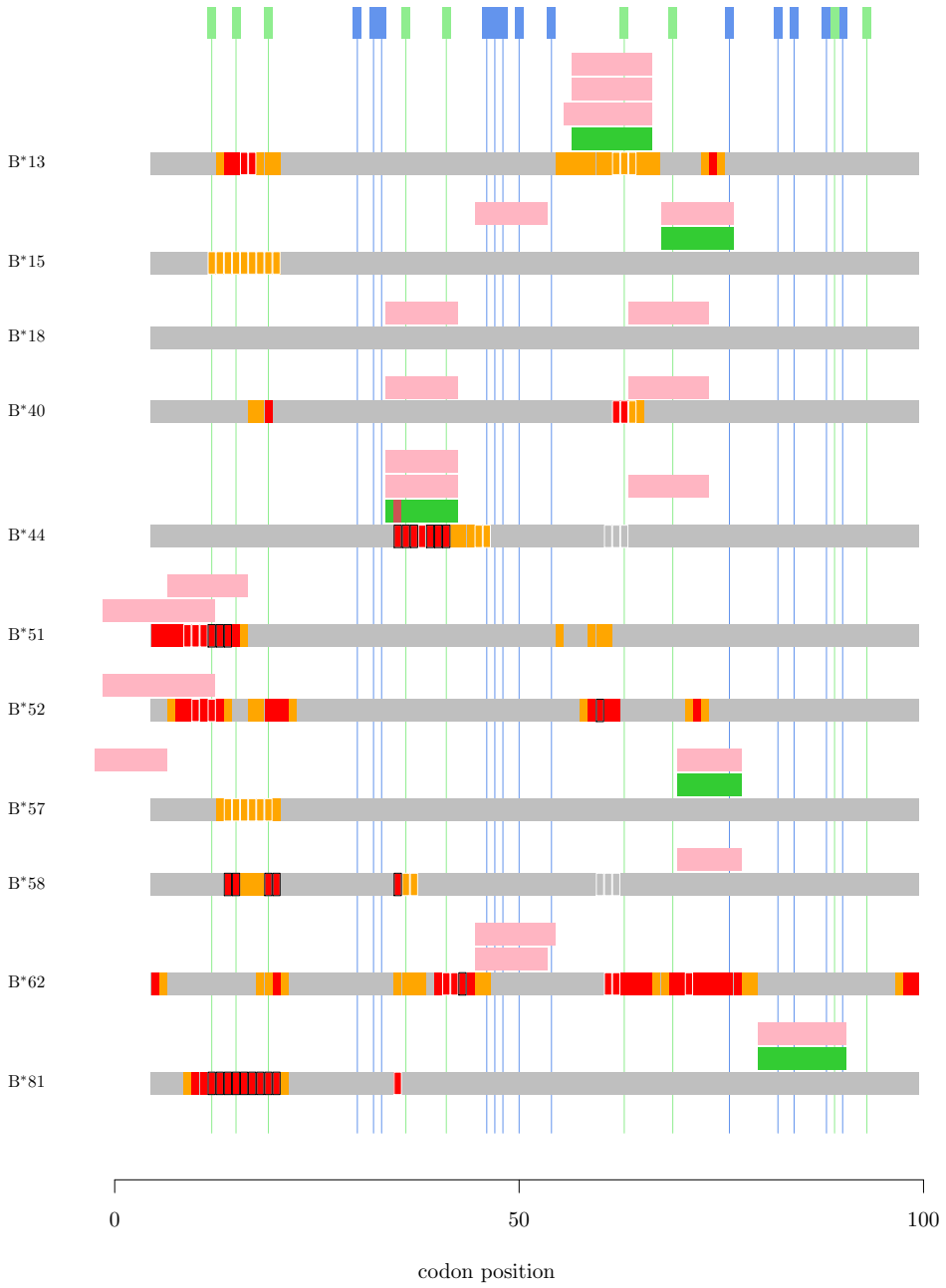


Figure 7.14: A-list and B-list epitopes across protease: HLA-B alleles. Epitope plot showing epitopes associated to HLA-B alleles as defined by the A-list [174] and B-list [205]. Plotting is as in Figure 7.13.

3. **A-list**: The collection of A-list epitopes [174] as defined by Brander *et al.* [173].

We also considered creating further collections of epitopes using the program netMHCpan [198] which uses a machine learning approach to estimate binding affinities at potential anchor residues. However, the resulting collections of putative epitopes obtained using netMHCpan formed a very large proportion of the analysed sequence, so we disregarded this method.

Before we begin, define an overlap between a 'selected site' (top-tier or second-tier site as detected by our model) and a putative epitope as 'the selected site falls within the epitope'.

We use the following two tests to measure the strength of overlap of the selected sites as estimated using our model and the putative epitopes as determined using the three methods above:

Test 1. **Random placement of selected sites**: For each HLA type, we have some number of selected sites; n say. We randomly place n simulated selected sites across the region and repeat for each HLA in one of the three class I genes. We then count the number of overlaps with the epitope set for each HLA type and then sum these overlaps. After repeating this procedure $1,000,000$ times we obtain an estimate of the distribution of overlaps under this null model of placement of selected sites. We then compare this distribution to the number of epitope overlaps with putative selected sites as estimated by our model, and obtain an estimate of the p-value.

Test 2. **Shuffling of HLA type labellings**: We permute the labellings of HLA types for a class I gene $1,000,000$ times, and count the number of overlaps with the putative epitope sets for each permutation. This leads to an estimate of an alternative null distribution (in this case we account for the fact that some portions of the region may be epitope rich across HLA types). We then compare this distribution of counts of overlapping sites to the observed number of epitope overlaps with our putative selected sites to obtain p-values.

An example of the putative epitope collections generated using motif scan is shown in Figure 7.15 (we take the subtype C consensus sequence in the analysed region of reverse transcriptase to scan and look for known HLA-B motifs). Epitopes are colour coded according to the number of strongly binding anchor residues and publication. Corresponding plots for HLA-A and C, and the class I motifs across protease are similar and are omitted. Figures 7.13, 7.14, B.6 - B.9 display all the Los Alamos A-list [174] and B-list [205] collections associated to HLA types present in our query datasets across protease and reverse transcriptase. Results taking subtype B as the consensus were similar and are omitted. In A-list epitopes, known escape sites (as described in Section 7.3.6.1) are highlighted in red. Note that many epitopes do not have any well defined escape sites. Also, it is very likely that many of HIV-1's epitopes remain to be described. This is because:

- Many HLA types are rare in the populations which are most studied.
- Subtype B viruses are the most well studied: there may be epitopes in other subtypes which do not exist in the majority of subtype B viral sequences.
- Some HLA types are rare across populations, so epitopes restricted by these HLAs will be inherently difficult to detect, even with increased sampling from less well studied populations.

Despite this, for the A-list and B-list epitope collections we see strong overlap between known epitopes and sites which we determine to be under HLA-B associated selection (see Figure 7.13). Looking at Figure 7.15 notice that using motif scan does not appear to be particularly helpful, as a large proportion of sites are in potential epitopes across the region. We now assess these overlaps quantitatively. For HLA-B associated selection over reverse transcriptase (the longest of the two regions examined with the greatest number of known HLA associations), we see the sort of trend which we would expect using the two tests of strength of overlap. Throughout, the most conservative test is Test 2. In Table 7.5, as our confidence in the epitope collection increases, p-values for Test 2 tend to reduce when considering either our top-tier or second-tier sites. Putative epitopes in motif scan were generated using the subtype C consensus (results using the subtype

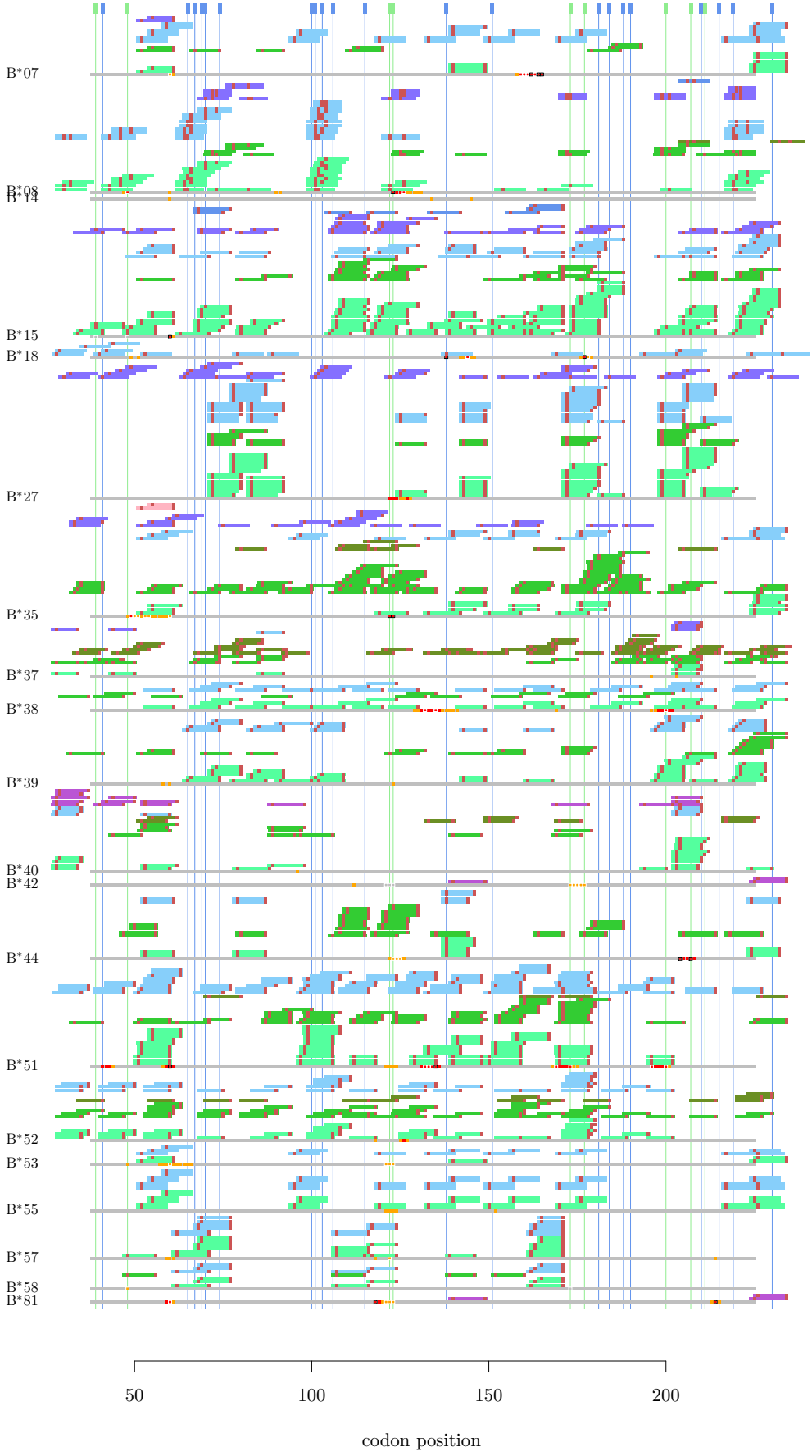


Figure 7.15: *See next page...*

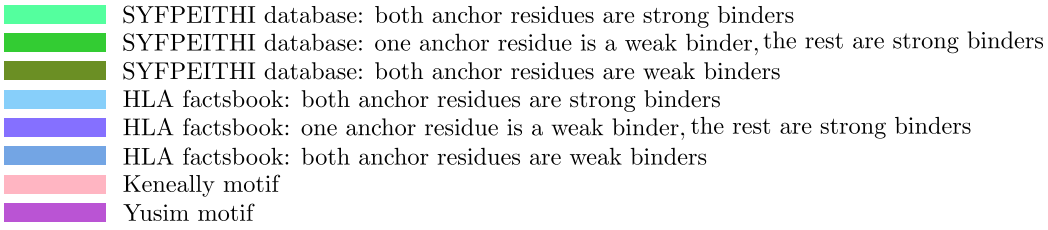


Figure 7.15: Motif scan across reverse transcriptase subtype C consensus: HLA-B alleles. Using validated binding motifs in the literature, we scan reverse transcriptase for HLA-B associated epitopes. Grey bars denote the region analysed, and are overlaid with colours based on our MCMC results as in Figure B.2. Blue vertical lines denote drug resistance sites. Green vertical lines denote sites at which there is an amino acid difference between the subtype B and subtype C consensus viral sequence. Potential epitopes are plotted above them, with the anchor residues marked in red. Putative epitopes are coloured according to the key which details each reference used: the SYFPEITHI MHC and peptide database [235], the HLA factsbook [178], Yusim [314] and Keneally [143].

region	HLA	our sites	epitope collection		
			motif scan	B-list	A-list
protease	A	top-tier	0.522 (0.818)	0.400 (9.44)	1 (0)
		second-tier	0.613 (2.48)	0.250 (2.05)	0.400 (0.942)
	B	top-tier	1 (0.292)	0.0643 (2.99)	0.0497 (6.22)
		second-tier	0.988 (0.395)	0.0206 (2.24)	0.0496 (2.87)
reverse transcriptase	A	top-tier	1 (2.00)	0.184 (2.79)	0.500 (1.29)
		second-tier	0.040 (1.61)	0.0521 (4.66)	0.250 (6.60)
	B	top-tier	0.166 (1.14)	0.000295 (24.7)	0.00277 (47.0)
		second-tier	0.310 (1.37)	0.00487 (19.0)	0.00289 (19.0)

Table 7.5: p-values for overlap using Test 2: HLA-A and HLA-B associated selection in protease and reverse transcriptase. p-values for strength of overlap with B-list and A-list epitopes using Test 2 are displayed. Odds ratios are shown in parentheses.

B consensus were similar). This trend was also present when considering overlap between anchor residues in putative epitopes and our top-tier and second-tier sites (results not shown). Such trends were less apparent for the other HLA loci, due in part to the paucity of A and B-list epitopes in HLA C, coupled with the likely inclusion of spurious selected sites due to linkage disequilibrium between the HLA class I genotypes. To summarise, taking runs with subtype C as the consensus, we find significant overlap of our top-tier sites with A-list epitopes in HLA-B associated epitopes in reverse transcriptase and protease. In protease, we find that as the number of defined HLA-B epitopes is increased to include B-list epitopes, we obtain an increase in the significance of overlap with our collection

of second-tier candidate sites. This is likely due to the low number of protease A-list epitopes in which escape variants are observed *in vivo*. Significant overlap is detected between our second-tier sites and the HLA-A associated motif scan epitopes in reverse transcriptase. No significant overlap is detected for HLA-A associated epitopes in protease. In each case p-values for Test 1 were lower and followed the same trends. Overall, the results suggest that we are capturing sites under selection in known HLA-B epitopes.

7.3.8 **Drug resistance mutations**: *Do escape mutations exist which lead to drug resistance (or partial drug resistance) or vice versa?*

We examine overlap between drug associated selection and HLA associated selection. If we find that there is an overlap, the combined selection pressure could lead to both drug resistance and immune escape. In Figures B.2 - B.5 we overlay the major drug resistance mutations (as defined in the Stanford drug resistance database [63]) in light blue vertical lines. We find a number of candidates for overlaps of drug and HLA associated selection in both protease and reverse transcriptase, the strongest of which is at position 138 in reverse transcriptase, associated to B*18 populations infected with subtype C. Whilst not mentioned in the Los Alamos HIV immunology database's table of escape variants [128] for A-list epitope `NETPGIRYQY`, we find that mutations in the overlapping B-list epitope `NETRGVRY` are associated with diminished or loss of ELISPOT response, and these same two sites are described as under selection in [31]. The same mutation which leads to partial immune escape is associated with decreased sensitivity to both Etravirine (ETR) and Rilpivirine (RPV) - two second generation NNRTIs (designed to be effective against viruses with resistance mutations to first generation NNRTIs). We are not the first to notice this potential issue in patients with the B*18 allele. Gatanaga *et al.* [93] test for selection at site 138 and examine the impact on RPV resistance, finding that it is selected for and can confer up to 7.3 fold resistance to RPV. We share the view of the authors that drug resistance testing for RPV (and ETR) should be carried out in B*18 patients before treatment.

region	HLA type	site
protease	A*25	46
	B*35	33
	A*25	47
	A*25	48
	A*66	76
	B*35	181
	B*35	188
reverse transcriptase	A*43	140
	B*13	184
	B*51	41
	B*57	214

Table 7.6: Summary of significant unknown HLA-B associated mutations in protease and reverse transcriptase. Second-tier and top-tier candidate sites are shown, with top-tier candidates outlined.

The NETRGVRY epitope mutation was the only site in our top-tier collection which overlapped with a major drug resistance mutation [63]. Within our second-tier sites, strong candidates are defined as those sites with a median > 1.5 which are not part of a region under selection overlapping with a selected (top-tier or second-tier) subtype B/C amino acid difference. We exclude these types of sites to reduce the potential for false positives caught by selection acting at subtype B/C differences. We also note that there are a number of candidates at which the median is greater than 2 but does not reach significance. These two collections of mutations are summarised in Table 7.6. Also of note is site 213 in reverse transcriptase, which is next to the drug resistance site 214 and under HLA associated selection by multiple HLA types. Many of these HLA types which are potentially associated to amino acid change at drug resistance sites are rare in both subtype populations. Therefore, further paired host HLA/viral sequence data or longitudinal viral sequence data in HLA typed individuals or binding assays are required to determine if these putative associations are real.

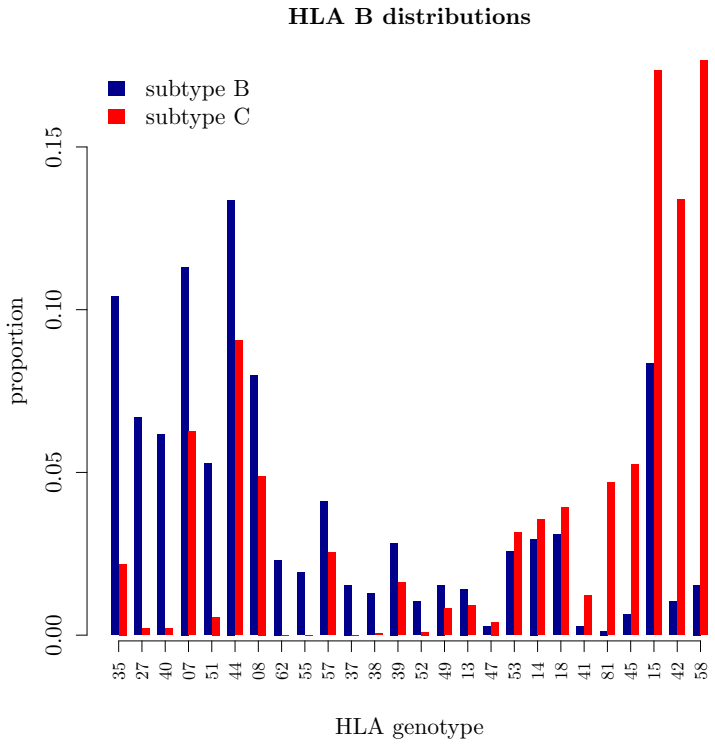


Figure 7.16: Distributions of HLA-B types subtype B infected hosts and subtype C infected hosts. We estimate viral subtypes of query sequences using RIP [229]. The subtypes most frequently defined in our query set were subtype B and subtype C. Consider the set of subtype B sequences. Using the corresponding HLA information we determine the HLA distribution of hosts infected with these viruses. This is an approximation of the HLA distribution in populations there the predominant subtype is subtype B. We do the same for subtype C. Bars are ordered by the absolute difference between HLA proportions in the subtype B and subtype C query sequence sets.

7.3.9 **Differences between HIV-1 subtypes**: *Are differences between HLA allele frequencies across populations playing a role in driving HIV-1 subtype differences at the amino acid level?*

Before we start, we make the following observations:

- The set of HLA frequencies of hosts infected with subtype B viruses differs from that of hosts infected with subtype C viruses. Figure 7.16 shows the differences for the HLA-B frequencies.
- Looking at Figures 7.11-7.12, B.2-B.5 we can observe that sites at which subtype B and C amino acid consensus sequences differ are under strong selection (both γ_h and ω are often high at these sites).

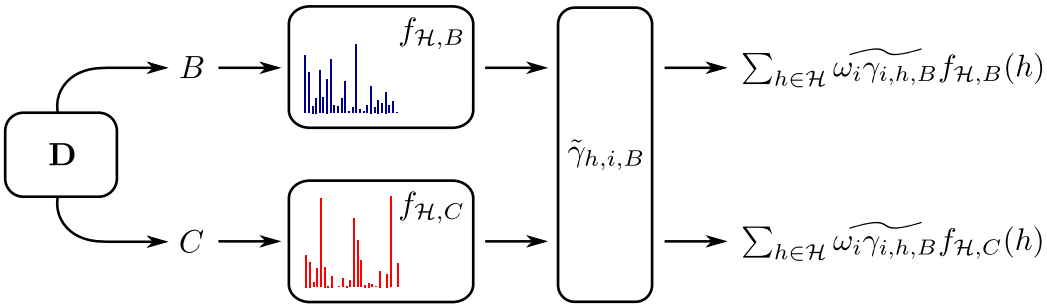


Figure 7.17: Flow diagram: Is HIV-1 diversification at the level of subtype partially driven by host HLA? We split **D** into subtypes, and consider subtype B and subtype C sequences. Using HLA information, we determine the HLA frequencies in hosts infected with subtype B and subtype C viruses. We then scale HLA associated selection by these frequencies to determine an approximation of the effect of HLA allele frequencies on selection away from some defined consensus sequence.

These two points suggest that CTL immune evasion plays an important role at sites at which subtype B and C differ. In this section, we investigate these sites further. We ask, do different population HLA frequencies act to push and pull away from HIV-1 subtype consensus sequences? To address this question, we obtain summary values which describe the average selection (away from some predefined consensus) at each site conditional on a collection of HLA frequencies. By substituting HLA frequencies from different populations, we may look for differences in selection between them. Follow the flow diagram shown in Figure 7.17. We first stratify query sequences **D** by HIV-1 subtype using RIP [229]. We then determine the HLA allele frequencies of the collection of hosts in which these viral sequences reside. As subtypes B and C are by far the most common HIV-1 subtypes in our dataset (see Figure 7.9), we concentrate our analysis on these two subtypes. We let $f_{\mathcal{H},B}(h)$ denote the frequency of HLA type $h \in \mathcal{H}$ in hosts infected with subtype B viruses ($f_{\mathcal{H},C}(h)$ is defined similarly for subtype C infected hosts). $f_{\mathcal{H},B}(h)$ therefore represents an approximation of the HLA frequencies present in a region in which subtype B is the majority HIV-1 subtype.

Let $\widetilde{\omega_i\gamma_{i,h}}$ denote the median HLA h associated selection away from consensus $\mathscr{C}$ at site i. Then $\sum_{h\in\mathcal{H}} f_{\mathcal{H},B}(h)\widetilde{\omega_i\gamma_{i,h}}$ provides an approximation of selection imposed by $f_{\mathcal{H},B}$ away from consensus $\mathscr{C}$ at site i.

Introduce $\gamma_{i,h,B}$ and $\gamma_{i,h,C}$ to denote the HLA h associated selection parameter at site i when setting $\mathscr{C}$ = subtype B or C consensus respectively. We will consider $\mathcal{H}$ = HLA-B types and consider selection away from $\mathscr{C}$ = subtype

B consensus first. This amounts to evaluating the sums $\sum_{h\in\mathcal{H}} \widetilde{\omega_i\gamma_{i,h,B}} f_{\mathcal{H},B}$ and $\sum_{h\in\mathcal{H}} \widetilde{\omega_i\gamma_{i,h,B}} f_{\mathcal{H},C}$ at each site. By plotting these two values against each other, we have an indication of how $f_{\mathcal{H},B}$ and $f_{\mathcal{H},C}$ differ in their influence on selection away from subtype B consensus at each site. We can perform the exact same calculations setting $\mathscr{C}$ = subtype C consensus. The resulting scatterplots showing selection away from subtype B and C consensus for sites in protease are shown in Figures 7.19 and B.12b respectively (corresponding plots for HLA-A and HLA-C are displayed in Figures B.11 and B.12). We mark any subtype B/C consensus amino acid sequence differences in red and annotate the codon position. In Figure 7.18 we show the codon distribution at these sites which distinguish subtype B and subtype C amino acid consensus sequences.

7.3.9.1 *Selection away from HIV-1 subtype consensus sequences.*

Looking at Figure 7.19 we see that the majority of subtype B/C differences in protease are under strong selection in both populations. Some of the sites which distinguish subtype B and C are highly polymorphic and are typically seen to have high HLA dependent selection in hosts infected with subtype B or subtype C viruses. The most prominent example in protease is site 63, see Figure 7.18. Some of the subtype B/C differences are under stronger selection away from subtype B consensus by $f_{\mathcal{H},C}$ than $f_{\mathcal{H},B}$ (e.g. sites 12, 15, 19, 93). To test if the difference in selection away from subtype B consensus between $f_{\mathcal{H},B}$ and $f_{\mathcal{H},C}$ is significant, we consider $A := \sum_{i\in S}\sum_{h\in\mathcal{H}} \widetilde{\gamma}_{i,h,B}\,(f_{h,B} - f_{h,C})$. Here S is the set of subtype B/C differences, and $\widetilde{\gamma}_{i,h,B}$ is the median estimate of $\gamma_{i,h,B}$ determined from the MCMC run where $\mathscr{C}$ = subtype B consensus and all query sequences are used. Note that we avoid $\widetilde{\omega_i\gamma_{i,h,B}}$ as the subtype B/C differences are enriched for sites with high ω. To obtain a p-value we permute the labellings of subtype B/C differences 1,000,000 times and re-evaluate A. This gives us a distribution in which to place the true A obtained from the true sites of subtype B/C differences. Our hypothesis in each case is one-tailed. We expect greater selection away from subtype B consensus by $f_{\mathcal{H},C}$ and greater selection away from subtype C consensus by $f_{\mathcal{H},B}$. The resulting p-values for differences in selection between $f_{\mathcal{H},B}$ and $f_{\mathcal{H},C}$ for each of the HLA class I alleles are shown in Table 7.7.

Figure 7.18: Codon distributions at sites which distinguish subtype B and subtype C consensus in protease at the amino acid level. We plot the distribution of codons alphabetically from left to right at site which distinguish subtype B and C. Subtype B distributions are shown on the top row, and subtype C distributions are plotted on the bottom row. We plot subtype C distributions upside down to guide the eye to difference between the two subtypes more clearly. Sites at which non-synonymous codon differences between subtype consensus at the site is the result of >1 base change within the codon are coloured in red.

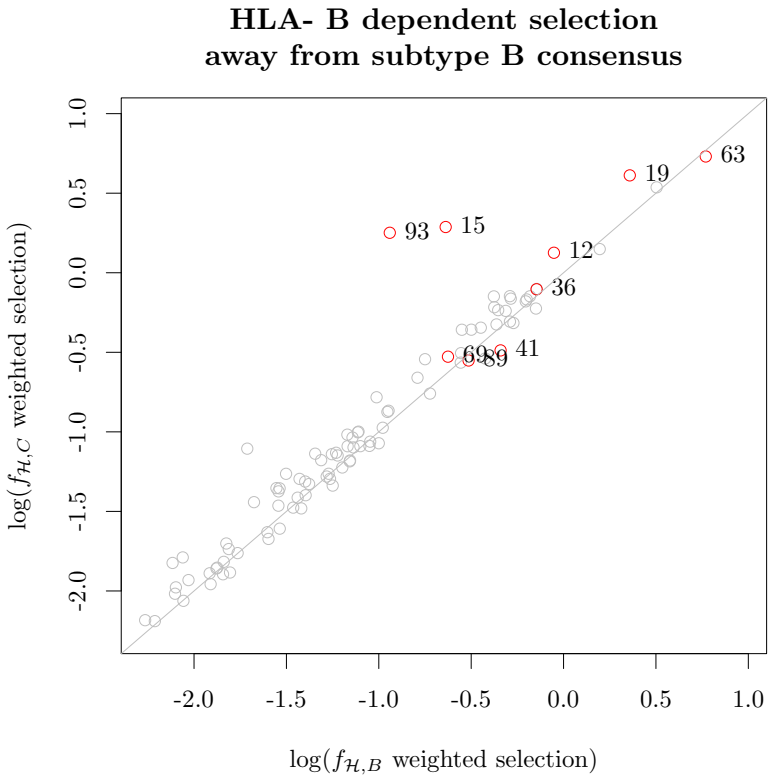


Figure 7.19: Scatter-plot showing selection away from subtype B consensus conditional on HLA distribution: Protease, HLA B.
Query sequences were split into subtypes using RIP [229]. HLA distributions associated to subtype B and subtype C HIV-1 sequences were determined. where $\widetilde{\omega_i \gamma_{h,i,B}}$ is the median in the MCMC run in which all query sequences were used. We plot $\sum_h \widetilde{\omega_i \gamma_{h,i,B}} f_{h,B}$ against $\sum_h \widetilde{\omega_i \gamma_{h,i,B}} f_{h,C}$, where $\widetilde{\omega_i \gamma_{h,i,B}}$ is the median value of $\omega_i \gamma_{i,h,B}$ in the MCMC run where all query sequences were used. Differences between subtype B and subtype C amino acid consensus are coloured red and labelled by codon position.

region	HLA	subtype of $\mathscr{C}$	p-value
protease	A	B	0.00011
	B		0.00058
	C		$< 10^{-6}$
	A	C	0.87
	B		0.95
	C		0.027
reverse transcriptase	A	B	2.9×10^{-5}
	B		$< 10^{-6}$
	C		$< 10^{-6}$
	A	C	0.047
	B		0.048
	C		0.048

Table 7.7: p-values for observing a value of A at least as great as that seen in the data.

Overall, we find strong evidence that HLA alleles more commonly found in hosts infected with subtype C viruses select away from $\mathscr{C}$ = subtype B consensus. This is consistent with the notion that local host CTL pressures shape viral evolution at the population level. However, most striking in Table 7.7 is that the reverse is not true. We do not see evidence that HLA alleles more commonly found in hosts infected with subtype B viruses select away from $\mathscr{C}$ = subtype C consensus. In Figures B.12-B.14 all subtype B/C differences appear more or less on the line $y = x$. The lack of a difference in the signal of selection away from the subtype C consensus amino acid sequence between $f_{\mathcal{H},B}$ and $f_{\mathcal{H},C}$ may be due in part to the larger geographical regions over which subtype B viral sequences in our query datasets were taken. Subtype B is prevalent in Europe and North America, so combining sequences across these regions may mask any underlying signal associated to HLA frequencies in a single population. By contrast, the majority of our query subtype C viral sequences are sampled from either Botswana or South Africa, far closer geographically by comparison. Overall, we find that our data supports the hypothesis that subtype diversification occurred as a result of both founder effects and differences between HLA frequencies between populations. However, we have no way of determining the directonality of this effect from our analysis. It may be due a boost by $f_{\mathcal{H},C}$ or the lack of a boost by $f_{\mathcal{H},B}$ at the subtype B/C differences.

7.3.9.2 **Novel selected sites**: *Are there any HLA associated sites currently not defined in the Los Alamos database?*

We now consider potential HLA associated sites which do not play a primary role in the subtype B/C amino acid differences described in Section 7.3.9. We define such candidate sites as HLA associated sites which do not form a run of selected sites overlapping with a subtype B/C amino acid difference, we also remove any sites which are contained within a known (A-list or B-list) epitope associated to that HLA. We notice in reverse transcriptase that there is selection at site 60 due to differences between subtype B and C at the nucleotide level, so we treat this site as a subtype B/C amino acid difference as above. Using our MCMC run in which we consider all query sequences and set $\mathscr{C}$ =subtype C consensus, we show top-tier (outlined and higlighted in grey) and second-tier (for this section, we are more strict for second-tier candidate sites, only selecting the subset with a median $\gamma_{i,h}$>1.5) candidate sites for HLA-B associated selection in Table 7.8 (for potential sites associated to HLA-A and C, see Tables B.3 and B.4).

We show HLA-B associated sites in the main text, as the HLA-B locus is known to have the strongest association with CTL response [145]. However, care must be taken before forming strong conclusions about these sites due to linkage disequilibrium between the HLA class I genes, shown in Tables B.1 and B.2 (e.g. B*81 and Cw*18, among others).

7.4 DISCUSSION

Applying the model introduced in Chapter 6 to two classes of query data, we find that we are able to detect many known drug resistance mutations and well defined escape mutations within CTL epitopes. As mentioned in Chapter 6, the modelling procedure that we apply represents a large step way from traditionally used techniques in the field. It allows us to make use of the vast amounts of sequence data available to us (that will only increase at a greater and greater rate), and accounts for recombination in our analyses. We detect a number of

region	HLA-B allele	site(s)
protease	B*27	77
	B*42	57
	B*52	60
	B*08	95
	B*13	16, 17
	B*27	78, 79
	B*38	79
	B*42	58
	B*42	59
	B*44	45, 46
	B*62	61
	B*62	61
	B*62	62
	B*62	71
	B*81	35
reverse transcriptase	B*45	196
	B*81	118 ,214
	B*08	126
	B*13	184
	B*35	53, 54
	B*38	132, 135
	B*41	53
	B*45	197, 198
	B*51	169, 196

Table 7.8: Top-tier and second-tier candidate sites for HLA associated selection that occur outside known epitopes and are not at sites that distinguish subtype B and subtype C amino acid consensus.

unknown candidate drug resistance mutations, displayed in Tables 7.2 and 7.3. Although our drug selection sensitivity is discouraging at first glance, we note that our method measures the trade-off between replicative capacity and drug sensitivity *in vivo* rather than simply the impact upon drug sensitivity used to define drug resistance mutations in many papers and databases.

Before commencing our HLA associated selection analysis we first check for two potential sources of bias, and find that they are not present in our datasets. We recapitulate many well defined HLA/virus associations, particularly association to HLA-B alleles in reverse transcriptase. Using a simple method to define what we call top-tier and second-tier candidate sites, we find that sites that we estimate to be under selection align with well defined epitopes in the literature (and known escape sites within these epitopes) with good specificity. Furthermore, we are able to explain apparent false negatives. In Section 7.3.8 we detect one site subject to strong drug and HLA associated selection. After further investigation, we find that the site has recently been tested for HLA and drug escape with these concerns in mind [93]. We also identify a collection of less well supported candidates for combined drug and HLA associated selection, often associated to rarer HLA types.

By stratifying our query sequences by HIV-1 subtype and looking at the associated HLA distributions after this stratification, we find evidence for selection away from particular viral subtypes conditional on HLA distributions associated to other HIV-1 subtypes. This is particularly apparent in the case of selection by the HLA distribution associated to subtype C viruses (primarily found in southern Africa in our query datasets) away from the subtype B viral consensus amino acid sequence. This supports the notion that both founder effects *and* HLA distribution differences between populations helped shape the viral subtypes that we see today. However, further analysis is required to confirm this assertion. Of concern, a strong driver of this signal at site 93 is B*42, but no B*42 motifs exist which span the site (see Figure 7.15).

Even ignoring the sites of subtype B/C differences where we detect an enrichment of selection across HLA types, we find quite a large number of candidates for HLA associated selection which we list in Section 7.3.9.2 and Tables 7.8, B.3

and B.4. Care should be taken before making any strong statements about these sites due to linkage disequilibrium between the HLA class I alleles. This is a drawback of splitting the class I alleles as we did, though considering all HLA types in our analysis together to help distinguish the true signal would lead to a substantially reduced query dataset (due to a lack of data across all three class I alleles). Carlson *et al* [36] allow HLA genotypes from only one class I allele to select for escape at a given site in an attempt to account for linkage disequilibrium, but we are reluctant to do this for fear of losing some real signal.

Some HLA genotypes are common and could be further split down to 4 digits where available (e.g. A*02), indeed some HLA alleles are known to restrict different epitopes at the 4 digit level. Such a stratification is the ideal approach in the presence of huge query datasets, but the lack or such vast datasets make this difficult. A method to make use of any correlation structure between HLA types would be useful in this case.

We have made the assumption that selection associated to HLA types acts multiplicatively at each site, but there is evidence to suggest that they act additively [164]. In further analyses we can alter the likelihood to account for this (and other selection regimes) and compare the results.

Potential extensions to the model include the incorporation of further parameters to measure covariance between HLA types. Doing so would allow us to gain more power from the data and generate insights about any existing correlation structure between HLA types, and ask a variety of questions: Does any correlation structure present at a locus relate to host HLA sequence similarity, HLA serotype or HLA supertype? Are any correlations preserved across different regions of the virus? This would also be useful in drug associated selection analysis. Do we see correlation between selection drug dependent selection across drugs of the same class? Should particular drug combinations not be administered?

Currently we assume that the host associated variable (drug regime or HLA types in our analyses) only select for amino acids that are one nucleotide away from consensus at a given site. In the case of drug resistance mutations this is sometimes not the case; some selected amino acids are two or more nucleotides away from consensus and thus would not be detected by our model. This could

explain some of our lack of sensitivity at drug escape sites. Changes to the model could be applied to account for this issue.

We have created, tested, and implemented a model which allows us to make use of huge publicly available sequence data. The method works, and could be extended to allow for the incorporation of longitudinal sequence data and transmission chains. We envisage applying the method to other regions of the HIV-1 genome in the first instance, but also for searching for antibiotic driven resistance mutations in bacteria, and blood group influences upon the malaria genome.

8

DISCUSSION AND FUTURE WORK

8.1 DISCUSSION

Over the course of the last five chapters we sought to answer the collection questions laid out in Chapter 1. We now carefully go through these questions one by one, before describing a collection of extensions and further applications of the modelling framework outlined in Chapter 6. In Chapter 3 we sought to create a model of CTL escape and reversion which took the shared ancestry of sampled viral sequences into consideration. We asked:

- Can CTL escape and reversion rates be estimated using a model that correctly accounts for the shared ancestry of viral sequences?

Yes. Using existing population genetics software [65] in combination with the model of Fryer *et al.* in exponential growth, we create a modelling framework which accounts for the dependency structure present in viral sequence data.

- Can such a model reliably recapitulate known escape and reversion rates from simulated data?

Yes. We presented a collection of simulation studies in Chapter 6. Our integrated method performs better than the ODE method of Fryer *et al.* particularly for slow escape and reversion rates.

- How does accounting for this dependency structure change estimates of escape and reversion rates when using real data?

We find that escape rate estimates obtained using our integrated method are highly correlated with those estimated by the ODE model of Fryer *et al.*, see Figures 5.4a-5.4c. In the case of reversion, we find that our integrated method

does a much better job of estimating lower reversion rates in simulation studies, and this is perhaps seen in Figures 5.4d-5.4f: under the ODE model, often reversion rate estimates are essentially zero, but we detect a signal of reversion using our integrated approach. However, we find that incorporating the dependency structure present in the viral genealogy can sometimes lead to very different rate estimates, see Figure 3.14.

Also, although we find a strong correlation between estimates using relatively small datasets in Chapter 5, this could change if we had the ability to analyse very large sequence datasets with associated host HLA information. When the sampling proportion is low in a population under exponential growth, the resulting genealogy has long terminal branches. Thus, in an exponential coalescent setting with small samples our integrated method is well approximated by the ODE model (which unknowingly assumed a star-like genealogy). This approximation becomes poorer as more and more sequences are added. Unfortunately though, we are unable to make use of large sequences datasets due to the computational constraints imposed by performing MCMC sampling on the space of genealogies. This drawback motivated the model and inference regime proposed in Chapter 6.

- Are rates of escape and reversion broadly similar in different populations?

In the case of escape, Yes. In Chapter 5 we find a strong correlation between escape rate estimates in two geographical populations in which subtype B is the most common HIV-1 subtype. The relatively small sample size of a South African dataset coupled with the difference between HLA distributions across populations (often common HLAs with known epitope associations are rare in South Africa) meant that rate estimation was difficult. Using the ODE approach of Fryer *et al* (and the understanding that rate estimates under our model were correlated with those of the ODE model, at least for small sample sizes) on a much larger African dataset, we find evidence of a correlation between CTL escape rates across widely separated geographical areas. These results were very interesting as they suggest that epistasis does not play a major role in the escape process: if it did, then the large sequence divergence between subtype B and C would likely result in complex interactions with escape sites, leading to differ-

ences in escape rates.

In contrast, we find no significant evidence for a correlation between reversion rates across populations. This may reflect a greater presence of epistasis in the process of reversion, or merely the inherent difficulty of estimating reversion rates.

- Is the rate of reversion of a CTL escape mutation positively correlated to its replicative fitness cost?

We do not observe any significant correlation between replicative fitness as measured by *in vitro* fitness assays and reversion rate estimates. One possible explanation is that the path through the space of viral sequences required to restore replicative capacity is not necessarily as simple as a flip back to the 'wildtype' amino acid at escape sites.

- Is the rate of escape of a CTL escape mutation negatively correlated to its replicative fitness cost?

Yes, we find evidence to suggest that there is a non-linear negative correlation between escape rate as measured by our integrated method and the *in vitro* replicative capacity of a subtype B lab-strain after site directed mutagenesis is applied to generate specific CTL escape mutants.

We find this correlation remarkable. Despite the myriad of simplifying assumptions that both models make to estimate replicative fitness *in vivo*, we observe a correlation. Again, this observation suggests a general lack of epistasis in the process of CTL escape (replicative capacity estimates are for single CTL mutants applied to a lab strain background, whereas our estimates reflect the rates of escape on the viral backgrounds seen in real world within host populations).

Between Chapters 5 and 6 we reflect on the lessons learned in creating the method and results of Chapters 3 - 5. We lay out a series of desirable features of a modelling and inference framework with which to detect HLA dependent selection. We decided that such a model should:

1. **Take the process of recombination into consideration**: The majority of analyses in the field ignore recombination, or focus on datasets where it is less apparent.

2. **Allow for incorporation of HIV-1 sequence data for which no associated host HLA information is available**: A wealth of sequence data is publicly available and has the potential to provide increased power to detect our parameters of interest.

3. **Use a codon model**: A more realistic model of sequence evolution which allows us to determine non-synonymous sequence alterations conditional on HLA, potentially indicative of CTL escape or compensatory mutations.

4. **Not require prior knowledge to inform rate estimation**: This will allow us to potentially detect unseen biological correlations and validate documented HLA/epitope associations.

After creating and successfully implementing a model with these properties, we asked:

- Do we successfully recapitulate HLA associated selection parameters used to create simulated datasets under our model, even in the presence of high recombination rates?

Yes, despite difficulty in estimating parameters to describe reversion, we find that our method is able to accurately distinguish regions of increased HLA dependent selection for different selection regimes, even in the presence of high recombination rates (we considered up to $r = 0.05$, for which we expect $\sim$5 recombination events per query sequence in protease). We also find that we have the potential to do even better given a more successful method of restricting the reference set of sequences ($\mathbf{D}_B$) for each query sequence, see Figure 6.18.

- Can our model detect known drug resistance mutations?

Yes. We make use of patient drug regime information available on the Stanford drug resistance database [63] to estimate drug associated selection pressure in protease and reverse transcriptase sequences. We find strong selection for non-synonymous mutations away from consensus associated to each of the drugs we analyse. Although our sensitivity to drug resistance mutations was not as good as we had initially hoped for, we note that assays used to determine the strength of drug resistance do not measure selection pressure within the host.

Our method provides estimates of parameters which reconcile the trade-off between reduction in drug sensitivity and impact upon replicative viral fitness. Therefore, our estimated drug associated selection parameters are perhaps more meaningful for predicting drug escape *in vivo*. We then went on to use our model on data where host HLA type is the host variable of interest, and ask:

- Are well known HLA genotype/epitope pairings detected using our model?

Yes. We find strong HLA dependent selection in all but two of the HLA-B A-list epitopes in which variants are known to occur. For the remaining two, we find that the apparent lack of signal of HLA dependent selection could be explained. In one case it could be attributed to the rarity of the allele. In the other, escape at a site within the epitope was the consensus amino acid in subtype B and C viral sequences. As a result HLA associated selection under our model could not be detected.

- How sensitive is the method?

We consider collections of peptides defined as epitopes with increasing confidence. Selected sites as estimated by our model overlap most significantly with well defined A-list and B-list epitopes taken from the Los Alamos database [127, 205]. As expected, this was most apparent for the HLA B associated epitopes.

- Are unknown HLA/epitope pairings detected?

Yes. We find a collection of HLA associated selection at sites not currently defined as being in epitopes. Many of these putative selected sites are associated to HLA alleles which are rare in populations infected with subtype B viruses or rare in general. This makes sense, as the majority of HIV-1 research is carried out in Western populations on subtype B viruses.

- Is there evidence of HLA genotype distributions shaping the differences we observe in consensus viral sequences between populations?

Yes. We stratify query sequences by the two most commonly sampled HIV-1 subtypes (B and C) and use the associated HLA data to define HLA frequency distributions of hosts infected with these viral subtypes. To estimate selection

associated to an entire HLA distribution, we re-weight HLA associated selection by HLA frequency and sum over HLA types. We find that the sites under the strongest selection were more likely to be sites that distinguish the two subtypes at the amino acid level. Moreover, the subtype C HLA distribution selected away from the subtype B consensus. This effect was not apparent in the opposite direction, perhaps due to the wide ranging geographies from which HIV-1 subtype B samples were taken. We emphasise that further work is required to validate this finding, but the results are intriguing.

8.2 FUTURE WORK

Given our findings, the future directions which we explore will be based on the model introduced and implemented in Chapters 6 and 7. Despite improvements over our model described in Chapter 3 and other models in the literature there are still a variety of modifications and improvements that can be made:

1. We make the coalescent approximation with fixed population size in Chapter 6. The possibility of relaxing the fixed population size assumption and/or removing the coalescent approximation should be explored.

2. If a virus (ν say) in the smaller query sequence dataset ($\mathbf{D}$) is more closely related to another members of $\mathbf{D}$ than any members of the reference sequence dataset ($\mathbf{D}_B$) then the model will not perform as well as it could. Although not statistically correct, redefining $\mathbf{D}_B$ as $(\mathbf{D}_B \cup \mathbf{D}) \setminus \nu$ when considering each virus ν may work well in practice.

3. $\mathbf{D}_B$ is currently heavily biased towards subtype B sequences, but a larger number of $\mathbf{D}$ are subtype C. The homogeneous mixing assumption is also clearly incorrect , particularly at the between subtype level. We could modify the coalescence rate dependent on HIV-1 subtype or sampling population to partially account for these to problems.

4. For our purposes the Hamming distance metric provided a simple method to restrict $\mathbf{D}_B$ to collections of sequences which are most likely to be closely

related to a given member of $\mathbf{D}$. However, in the presence of high levels of recombination or when computational overheads mean that $\mathbf{D}_B$ must be further restricted for each member of $\mathbf{D}$ (e.g. if the analysed region is significantly longer) we should use more sophisticated approaches. One obvious example is to use a Viterbi algorithm. We may assume some constant prior on recombination rate across the analysed region and determine a collection of likely paths through $\mathbf{D}_B$. These paths could then be used to restrict $\mathbf{D}_B$ for each member of $\mathbf{D}$.

5. We may allow for longitudinal samples or transmission chains in our query dataset, $\mathbf{D}$. For example, a transmission chain $a \rightarrow b \rightarrow c$ could be dealt with evaluating $P(a|D_B, \Theta)P(b|a, \Theta)P(c|b, \Theta)$ as part of the likelihood. Similar modifications could be applied to exploit the extra power to estimate parameters present in longitudinal datasets.

6. Linkage disequilibrium is a problem. We considered the three HLA class I loci independently. Although it would reduce the number of query sequences, we should consider HLA types together. We may then determine if doing so removes some of this problem by comparing the resulting distributions of HLA associated selection parameters for HLA types in strong linkage disequilibrium.

7. We could examine the use of alternative models of HLA associated selection. Throughout Chapter 6 and 7 we assumed that distinct HLAs associated selection parameters combine multiplicatively to select away from the consensus amino acid at a given site. Other schemes should be considered and the resulting fit and parameter estimates compared. For example, rates could combine additively or immunodominance could be approximated.

8. There may be correlation structure between $\gamma_{h,i}$ across the analysed region. We can currently analyse this correlation *post hoc* to determine if particular HLAs act similarly across the analysed region, or if particular drug treatments have similar selection profiles. If this *post hoc* analysis suggests that this is the case, a potential extension to the model would be to co-estimate this correlation structure in the MCMC scheme.

9. Currently a lot of time is spent performing quality control on and alignment of the sequences to be analysed. This is cumbersome and means that it can take a lot of work before we even begin our analysis of a new dataset. For any given member of **D**, we don't actually care about the majority of the multiple sequence alignment, only the alignment of sequences locally (by which I mean sequences close-by according to some measure). Graphical methods of constructing a more general class of alignment could be useful here. Such approaches are less computationally intensive than multiple sequence alignments and have desirable properties for our purposes. They would also open up our method for use in regions that are far more variable in HIV-1 (*env*) and other pathogens.

10. We should check to determine if recombination is really playing a role in our parameter estimation by enforcing a strict zero prior on the recombination rates across the analysed region and comparing the resulting parameter estimates to those estimated when allowing for recombination.

11. Simulated datasets were generated under the model. We should generate more realistic simulated datasets and determine our ability to recapitulate parameters.

12. Throughout our simulation studies our ability to estimate reversion parameters was poor, and we underestimate the non synonymous/synonymous substitution ratio (ω) as the recombination rate is increased, see Figure 6.17. We have a collection of arguments for why this is the case, but need to examine this thoroughly to determine the exact source(s) of the problem.

There are straightforward applications of the model to other datasets. We can apply the framework to other regions of the HIV-1 genome, e.g. *gag* and *nef*.
Moving on to applications to other pathogens, the most obvious candidates are Hepatitis-C and HIV-2 for which sequences and host data could be analysed in exactly the same way as the HIV-1 data used in Chapter 6; estimating both HLA and drug dependent selection.
We are not restricted to viruses. With some modifications we could apply the method to look for mutations selected for by antibiotics in bacterial populations,

or look for host associated selection in other pathogens such as malaria.

We wish to make the program publicly available and easy to use, with extensive documentation and plotting functionality. BEAST [65] has demonstrated that doing so can lead to far greater use and could potentially result in extensions and modifications being made by other researchers.

During the course of this thesis there have been a number of setbacks to overcome: Chapter 3 solved one problem whilst simultaneously creating another. In a world of rapidly expanding sequence datasets genealogy based inference has been fantastically useful. However, performing inference on the full data likelihood (coalescent based or birth-death based) becomes intractable as the number of sequence samples gets large. In Chapter 6 we took a step back and created a modelling and inference framework which utilises an approximation to the full likelihood that can be used for a range of applications and extended in numerous ways. We are excited to apply our procedure to ask further interesting biological questions.

A

HYPOTHESIS FREE ESTIMATION OF HLA ASSOCIATED SELECTION IN THE PRESENCE OF RECOMBINATION USING LARGE DATASETS: THE GREAT ESCAPE

A.1 DERIVATION OF EXPECTED DIVERGENCE TIME TO CLOSEST RELATIVE IN A COALESCENT TREE WITH k LEAVES

Consider adding a new lineage to an existing tree with k lineages. We are interested in the the distribution of this time, t. To derive the mean, let T_k be the total tree length of a coalescent tree with k lineages. Then,

$$E(T_k) = \sum_{i=2}^{k} iE\left(Exp\left(\binom{i}{2}\right)\right) \tag{199}$$

$$= \sum_{i=2}^{k} \frac{i}{\frac{i(i-1)}{2}} = 2\sum_{i=2}^{k} \frac{1}{i-1}. \tag{200}$$

Thus, since $t = T_{k+1} - T_k$,

$$E(t) = E(T_{k+1}) - E(T_k) \tag{201}$$

$$= \frac{2}{k}. \tag{202}$$

Thus, an approximation which has the correct mean is an exponential with parameter $\frac{k}{2}$. It can be shown that the moment generating function (mgf) of the distribution is

$$g_k(\phi) = \sum_{i=2}^{k} \frac{i-1}{\binom{k}{2}} \prod_{j=i}^{k} \left(1 - \binom{i}{2}^{-1} \phi\right)^{-1} \tag{203}$$

A.2 SIMULATIONS USING THE GOLD STANDARD RESTRICTION OF $\mathbf{D}_B$

In the following figures, the restriction of D_B to D_{B_i} for each sequence contains the true sequences copied from. The underlying recombination probability is 0.05 between each site along the codon sequence. Figures A.1 - A.3 display estimates of HLA associated selection. Figure A.4 shows estimates and coverage of the reversion parameters and recombination probabilities.

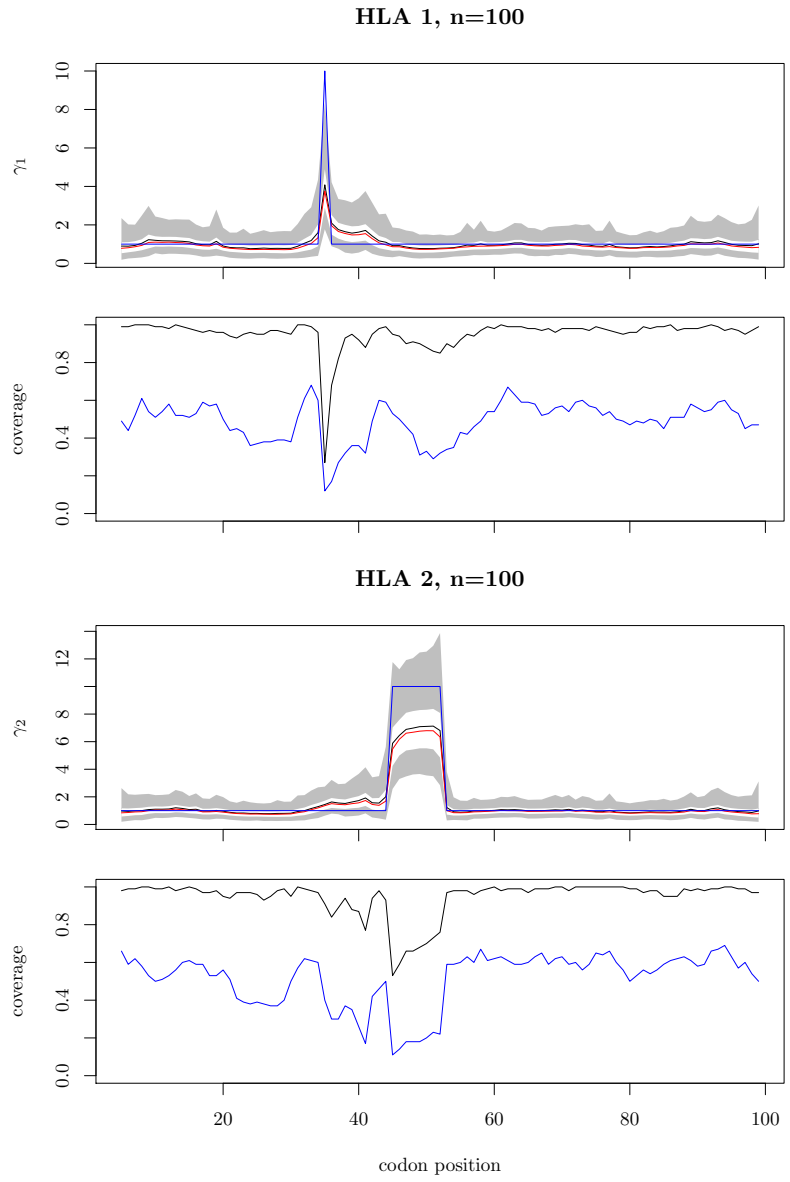


Figure A.1: Estimating HLA associated selection across a region using seqences copied from; $\boldsymbol{r = 0.05}$. D_{B_i} used to generate estimates for each run contained the actual sequences copied from to generate sequence i. Odd rows show average estimates and credible intervals of HLA associated selection (γ_h) across the region for the first two HLA types of the simulated sequences. Averages are taken over 100 independent MCMC runs on independent simulated data with the same underlying parameters. The mean estimate is plotted in black, the median is plotted in red, and the true underlying parameter value is plotted in blue. The white band denotes the 50% credible interval within the 95% credible interval denoted by the grey band. Even rows display the coverage, the black line shows the proportion of the 100 estimates that lie within the 95% credible interval for each run, and the blue line shows the corresponding proportion lying within the 50% credible interval.

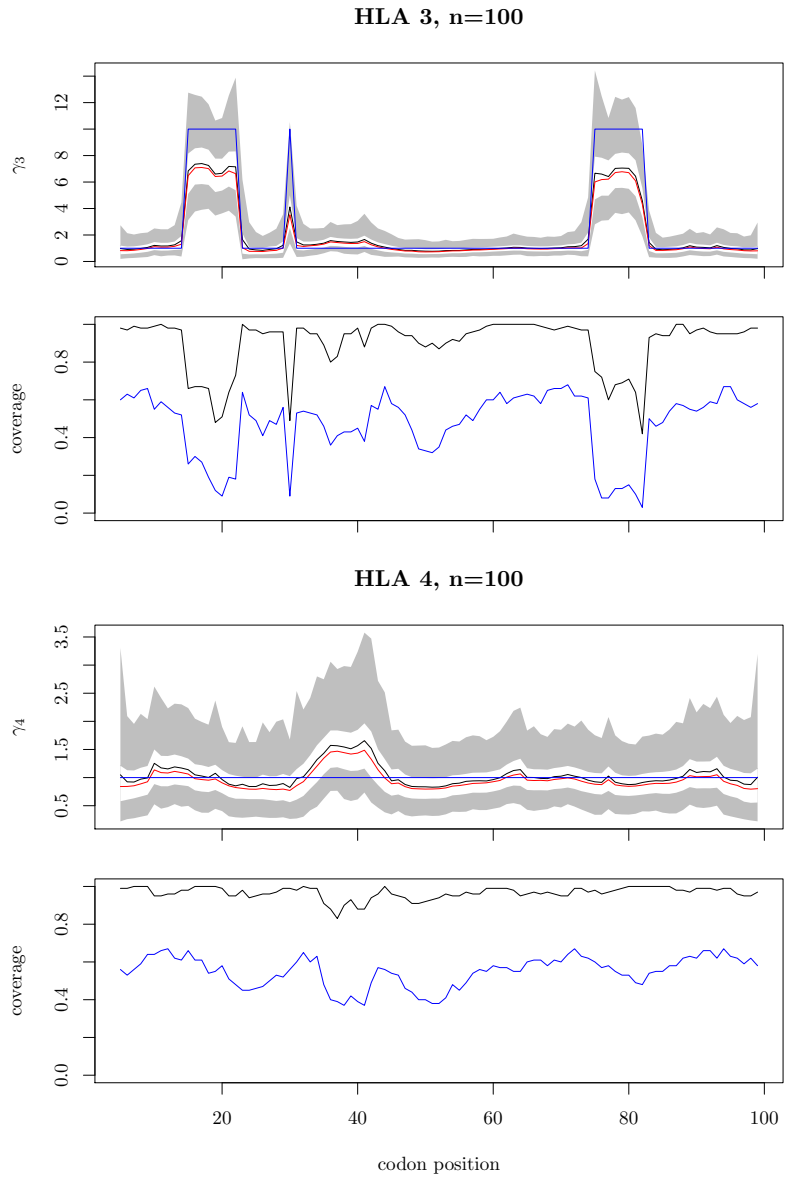


Figure A.2: Estimating HLA associated selection across a region using seqences copied from; $\boldsymbol{r = 0.05}$. D_{B_i} used to generate estimates for each run contained the actual sequences copied from to generate sequence i. Plotting is as in Figure A.1

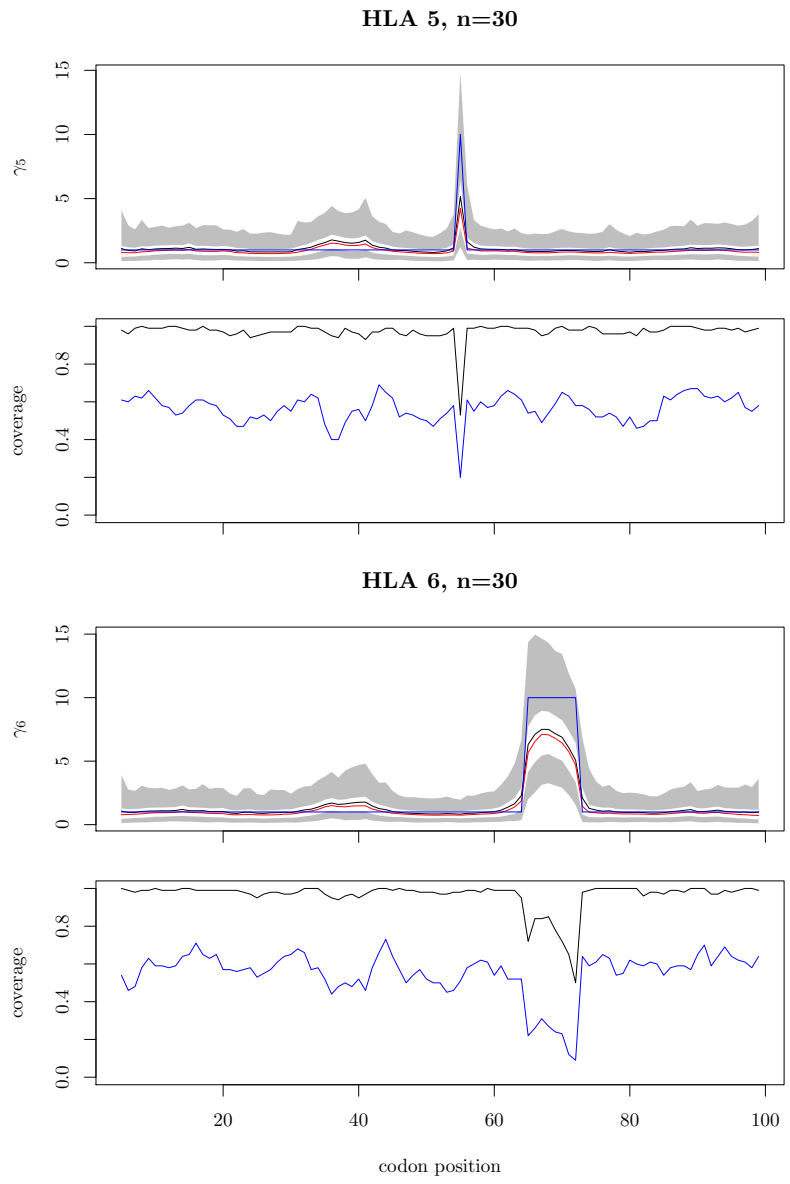


Figure A.3: Estimating HLA associated selection across a region using seqences copied from; $\boldsymbol{r = 0.05}$. D_{B_i} used to generate estimates for each run contained the actual sequences copied from to generate sequence i. Plotting is as in Figure A.1.

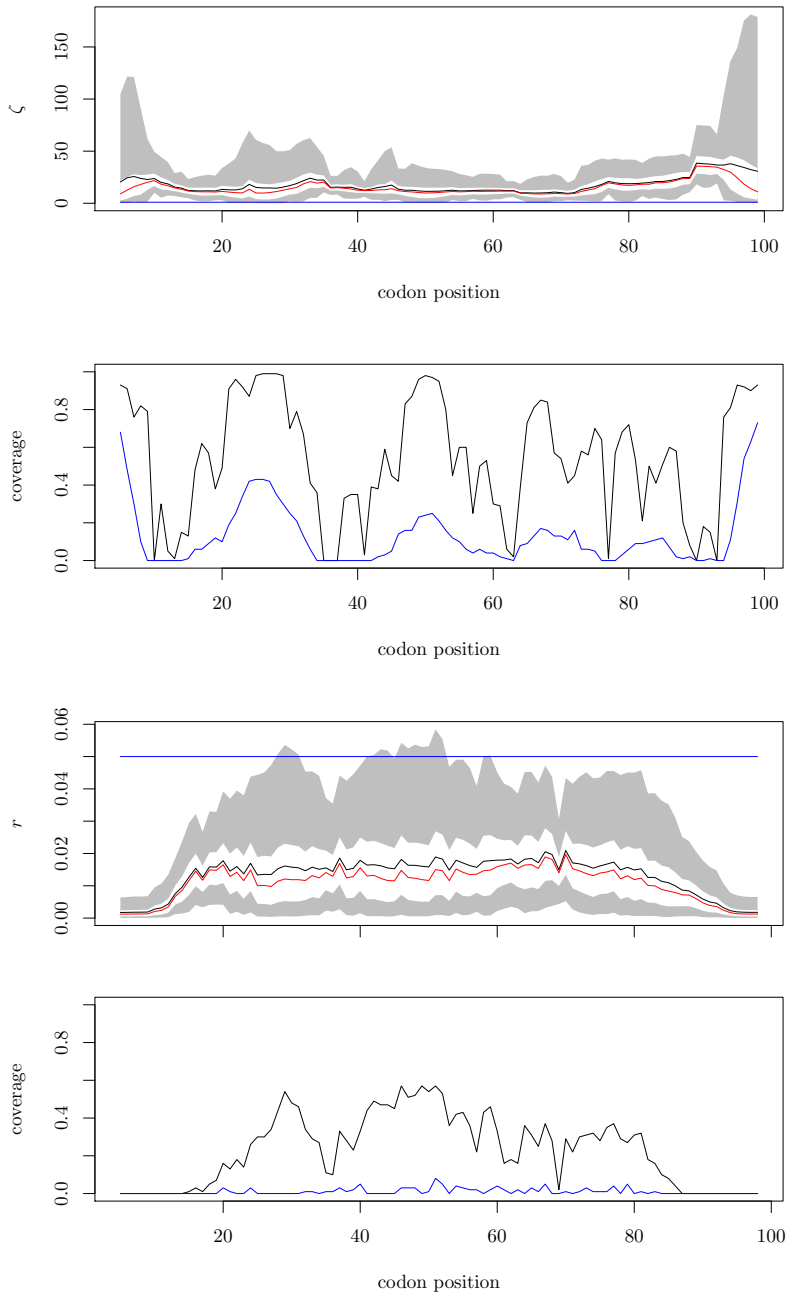


Figure A.4: Estimating scaling of selection due to reversion and recombination probability across a region; $\boldsymbol{r = 0.05}$. D_{B_i} used to generate estimates for each run contained the actual sequences copied from to generate sequence i. Odd rows show average estimates and credible intervals of parameters describing reversion (ζ) and recombination probability between adjacent sites (r). Averages are taken over 100 independent MCMC runs on independent simulated data with the same underlying parameters. The mean estimate is plotted in black, the median is plotted in red, and the true underlying parameter value is plotted in blue. The white band denotes the 50% credible interval within the 95% credible interval denoted by the grey band. Even rows display the coverage, the black line shows the proportion of the 100 estimates that lie within the 95% credible interval for each run, and the blue line shows the corresponding proportion lying within the 50% credible interval.

B

HYPOTHESIS FREE ESTIMATION OF HLA ASSOCIATED SELECTION IN THE PRESENCE OF RECOMBINATION USING LARGE DATA SETS: REAL DATA

B.1 DETAILED DESCRIPTION OF QUERY AND REFERENCE DATASET PREPARATION

B.1.1 *Drug associated selection analyis*

All sequences and their associated host drug regime data were downloaded from the Stanford drug resistance database [63]. We assumed that the alignment was correct.

Sites at which there was $\geq$5% ambiguity were removed, followed by sequences for which there was $\geq$5% nucleotide ambiguity across the region of interest. We did it this way round as in reverse transcriptase, a $\sim$550 nucleotide region is sequenced much more frequently than the remainder of reverse transcriptase, and we wish to retain power to estimate parameters.

One sequence per patient was included in the combined alignment of reference and query dataset by removing repeat entries of any given patient identifier uniformly at random (assuming unique patient identifiers $\Rightarrow$ unique patients).

The dataset is then split in two, sequences for which the host was not receiving any drugs $\mathbf{D}_B$, and sequences for which the host was receiving drugs $\mathbf{D}$. 2000 sequences were then randomly chosen from $\mathbf{D}_B$ and placed in $\mathbf{D}$. $\mathbf{D}_B$ defines our reference dataset. We perform some further data cleaning steps to create the query dataset from $\mathbf{D}$.

Sequences were removed from $\mathbf{D}$ if 'RTI', 'NNRTI', 'NRTI' or 'PI' is used as a

descriptor of one of the drugs in a patients drug regime, as this was not specific enough for our purposes. Finally, any sequences which were associated to a drug present in **D** fewer than 10 times were removed. **D** now defines the collection of query sequences. Drug information associated to the query sequences was reformatted into a binary .csv file so that it could be read by our C program.
The resulting collections of reference and query datasets consisted of 49,535 reference sequences and 6,080 query sequences with associated drug regime data for the protease dataset, and 45,254 and 13,111 query sequences with associated drug regime data for the reverse transcriptase dataset.

B.1.2 *HLA associated selection analysis*

Obtaining reference and query datasets for an investigation into HLA associated selection was more cumbersome. We wished to incorporate as much data as was publicly available into our analyses. In order to do this we investigate the most highly sequenced portions of the viral genome: protease and reverse transcriptase. To create the reference datasets we concentrated our efforts on three major sequence resources:

1. Los Alamos HIV sequence database.
2. Stanford drug resistance database.
3. UCLA HIV positive selection mutation database.

Before filtering, the combination of these three databases represented a total of $131,708 + 64,418 + 45,161 = 241,287$ and $160,225 + 67,515 + 45,161 = 272,901$ sequences in protease and reverse transcriptase respectively. For the Los Alamos HIV sequence database portion of the reverse transcriptase reference dataset, we allow sequence chunks of length >500 nucleotides.
To create our query datasets, we combine viral sequence data with associated host HLA information from across the datasets outlined in Section 7.3.1. In order to arrive at the final collection of reference sequences and query sequences combined with HLA information we apply a number of filtering steps.

B.1.2.1 *Creating the reference sequence dataset.*

Filtering steps

1. **Remove overlap between the Stanford drug resistance database and Los Alamos HIV sequence database**: For any patient identifiers present in both databases, we remove all sequences corresponding to that patient from the Los Alamos database sequences.

2. **Remove sequences with associated host HLA information from both the Stanford drug resistance database and Los Alamos HIV sequence database sequences**: This was acheived by comparing patient identifiers. The appropriate sequences were removed from the Stanford drug resistance database and Los Alamos HIV sequence database sequences.

3. **Remove repeat patient entries**: Again, this was acheived using patient identifiers. For any repeated patient identifier, one sequence was kept (chosen uniformly at random from among the sequences with the same patient identifier).

Sequence alignment

The next stage of the reference dataset preparation was to align the collection of sequences. This is a non-trivial task given the vast amount of sequence data we have assembled. Luckily, the alignment of protease in the Stanford drug resistance database is well maintained, so we assume that this portion of the alignment is accurate. Note that this is why we remove the Los Alamos database copy where there is overlap between the Stanford drug resistance database and Los Alamos sequence databases. In the case of the sequences taken from the Los Alamos HIV sequence database, we find that attempting to generate an alignment when downloading $\sim$70,000 sequences yields nonsense results. However, there is a lack of indels in the protease region. We therefore simply considered sequences of exactly 297 nucleotides for the alignment (after removal of gaps from the alignment proposed using the Los Alamos database default alignment software). Performing this step lead to a loss of $<$500 sequences out of a total of $\sim$70,000. We then checked the resulting distribution of nucleotides across the

region. Finally, the UCLA positive selection database was assumed to be aligned correctly.

Such a simple alignment procedure was not available to us for the reverse transcriptase sequences due to the increased prevalence of indels within reverse transcriptase, coupled with the longer sequence length. We reduced the size of the collection of sequences to be aligned by assuming sequences taken from the Stanford drug resistance database were correctly aligned. Aligning all the remaining sequences at once using MUSCLE [71, 70] was prohibitive computationally, as was using the align option in MUSCLE (which can be used to combine alignments: this procedure also tended to produce artefacts in the overall alignment). We therefore considered the approach of aligning subsets of 5000 sequences and comparing these 'subalignments'. We found that although more common than within protease, indels are rare. Removing them resulted in a collection of subalignments that aligned with both each other and the consensus sequences for both the Stanford drug resistance database and UCLA positive selection database alignments. After this alignment step $50,246 + 55,179 + 45,161 = 150,586$ and $67,847 + 57,576 + 45,161 = 170,584$ protease and reverse transcriptase remained in our reference datasets.

Quality control

We removed ambiguous codons and sequences from both the protease and reverse transcriptase alignments by first deleting codons with $> 5\%$ ambiguity and then sequences with $> 5\%$ sequence ambiguity as in Section B.1.1. Removal of sequence ambiguity resulted in a total of 146,523 and 165,629 reference sequences for protease and reverse transcriptase respectively.

B.1.2.2 *Creating the query dataset*

Filtering steps

As in the reference sequence filtering step, we remove repeat patient entries using patient identifiers: For any repeated patient identifier, one sequence was kept (chosen uniformly at random from among the sequences with the same patient identifier.

Sequence alignment

We aligned the query sequences using MUSCLE [71, 70]. We found that no gaps were introduced (aside from at the beginning and end of the sequences). Both the protease and reverse transcriptase query sequences align nicely with their respective reference sequence datasets.

Quality control

We checked ambiguity of codons and sequences within the query datasets using the same criteria as in Section B.1.2.1, removing codon positions and sequences as required.

Throughout, HLA typing was restricted to two digits. We consider HLA-A, B and C separately in our analyses, and so create three query datasets (consisting of HLA data and viral sequence data). For each, we remove any sequences for which associated host HLA data is not available for the class I molecule under investigation. In the cases where individuals were partially HLA typed for A, B or C, we assume that the individual only has the HLA type which has been determined. After all of this data filtering, 1523, 1514 and 1444 protease query datasets remained for HLA-A, B and C analysis respectively. 1320, 1298 and 1172 reverse transcriptase query sequences remained for HLA-A, B and C analysis respectively.

In both the query and reference datasets after filtering we obtain a relatively constant sequence ambiguity across the newly defined regions of protease and reverse transcriptase, with slight increases at the start and end of the region as we would expect. Consensus nucleotide proportion across both newly defined regions averaged across all sites is 0.941 and 0.943 for protease and reverse transcriptase respectively, with any large drop below this coinciding with a subtype B/C consensus sequence difference at the nucelotide level.

The resulting cleaned datasets that we pass to our program to perform MCMC are:

- A sequence reference dataset $\mathbf{D}_B$ consisting of sequences taken from Los Alamos sequence database, the Stanford Drug resistance database and the UCLA positive selection database for which no host HLA information was available.

- A query dataset which consists of:
  1. Sequence data **D** taken the datasets outlined in 7.3.1.
  2. Host HLA data $\mathcal{H}$ taken to two digits associated to the viral sequences in **D**.

B.2 SUPPLEMENTARY FIGURES AND TABLES

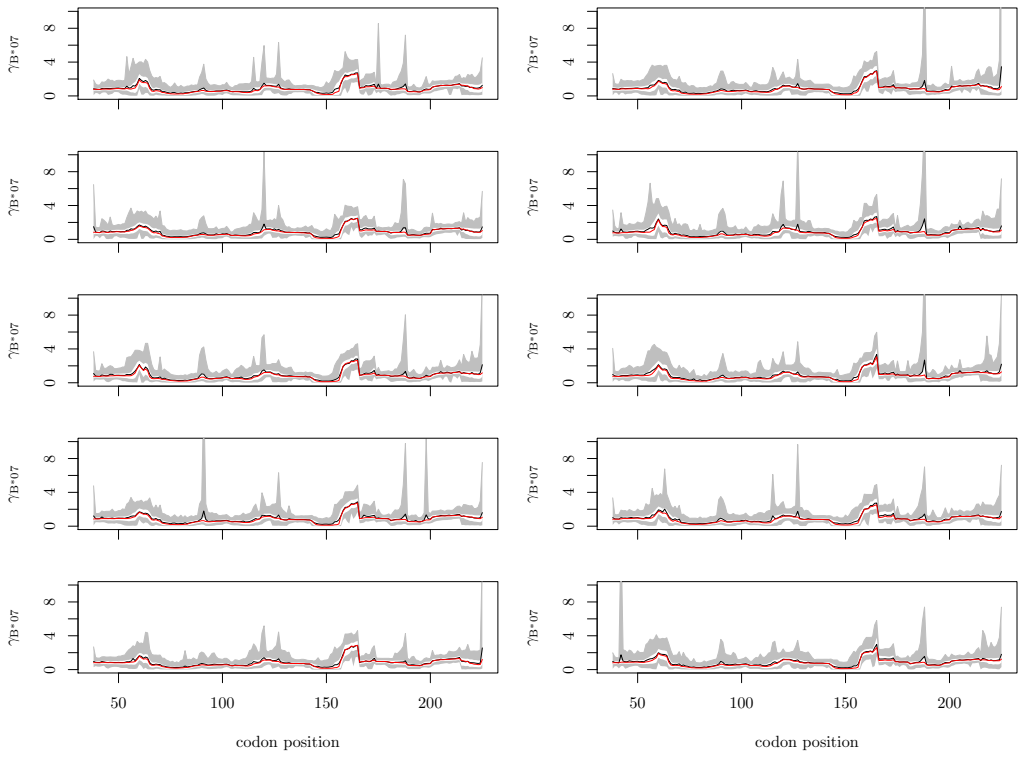


Figure B.1: Results are consistent across independent MCMC runs. Results of an 10 independent MCMC runs applied to all the reference sequence data and query HLA-B dataset in Section 7.3. We show summaries of the posterior distribution of γ_{B^*07}. The white and grey regions enclose the 50% and 95% credible intervals respectively. The median and mean scaling of ω associated to B*07 (averaged over the independent MCMC run) are plotted in red and black respectively.

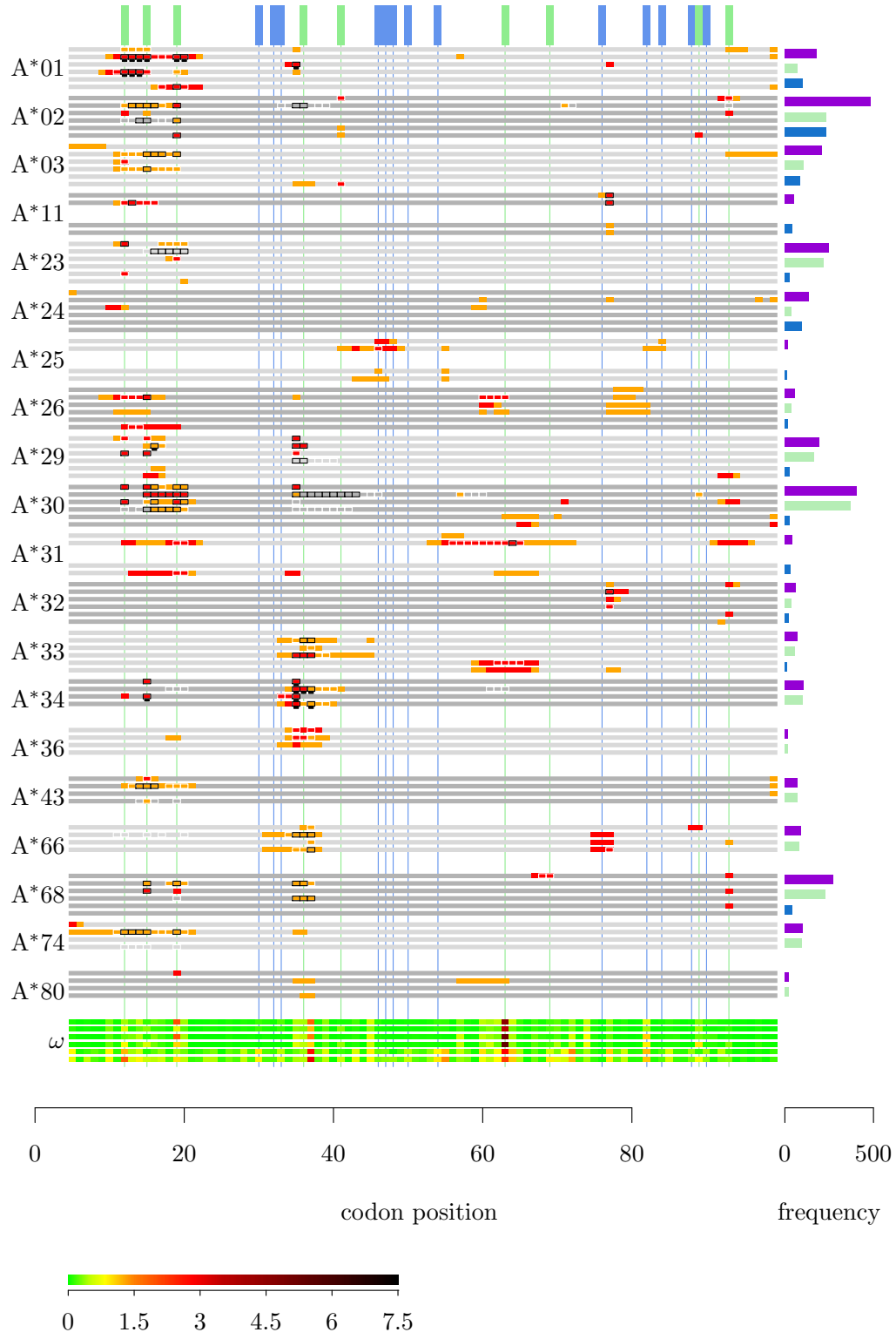


Figure B.2: Summary of HLA-A associated selection in protease. For each HLA subtype we plot six summaries of selection across the region. See Section 7.3.6.1 for full details of the plot.

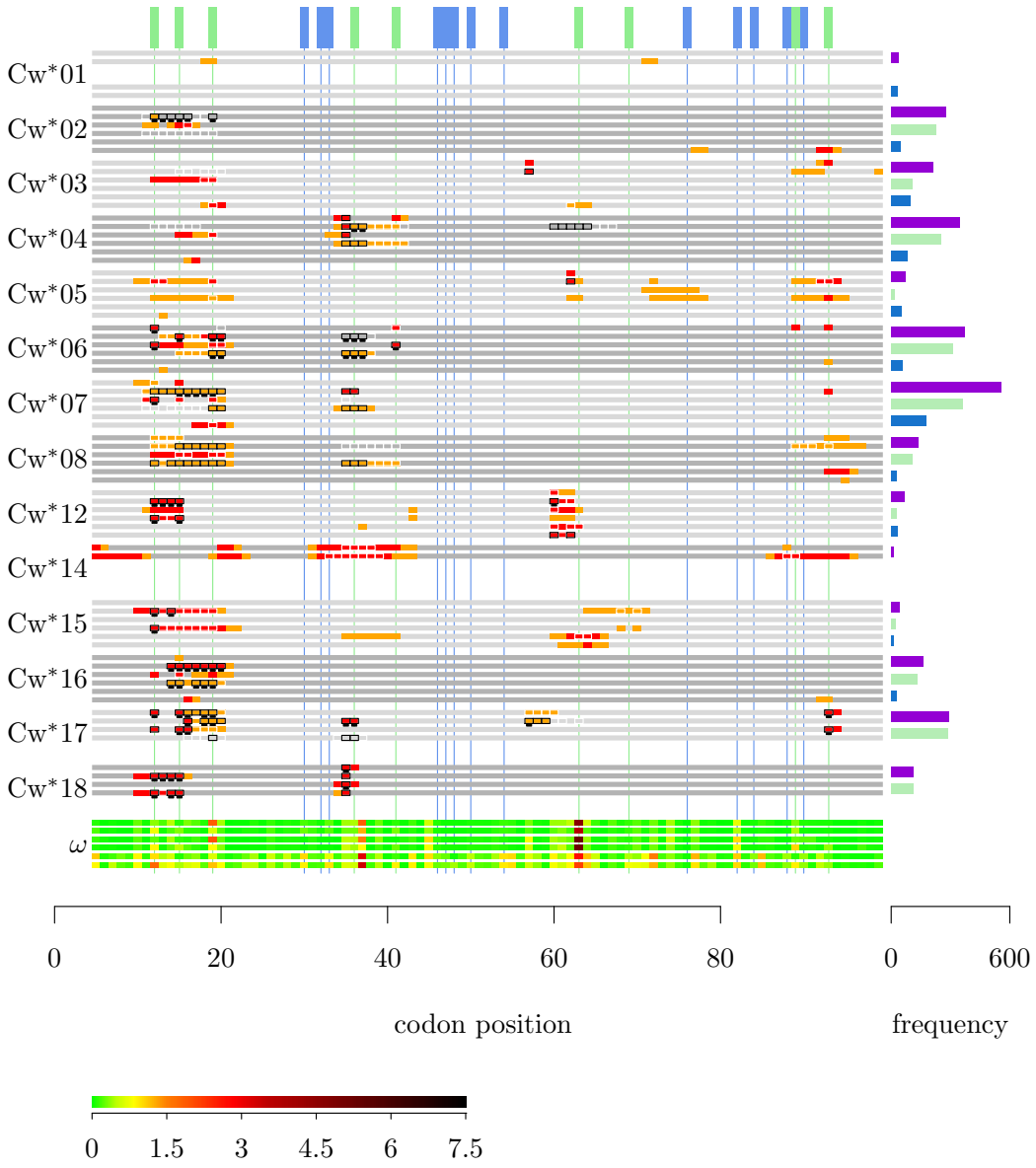


Figure B.3: Summary of HLA-C associated selection in protease. See Section 7.3.6.1 for full details of the plot.

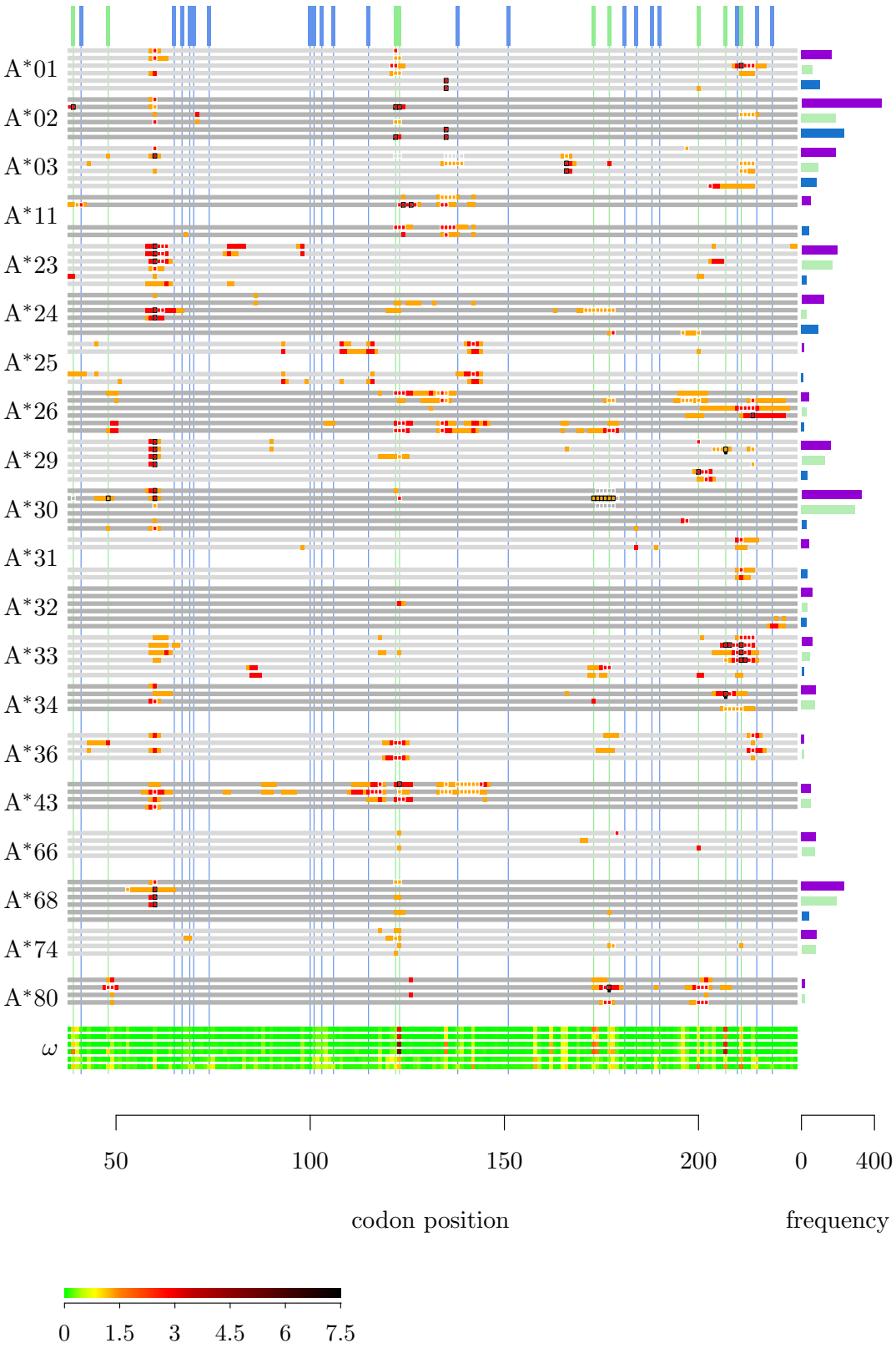


Figure B.4: Summary of HLA-A associated selection in reverse transcriptase. See Section 7.3.6.1 for full details of the plot.

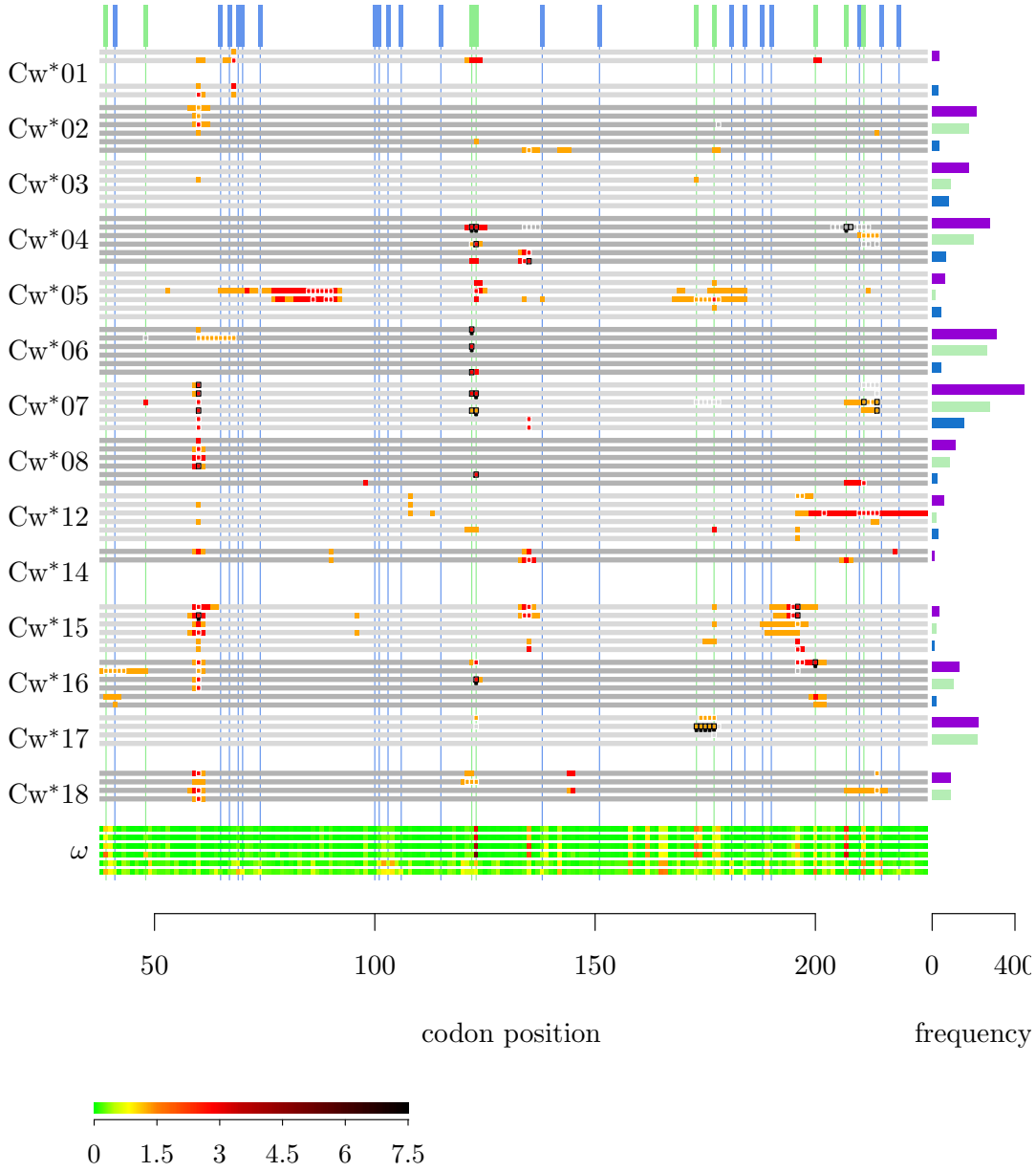


Figure B.5: Summary of HLA-C associated selection in reverse transcriptase. See Section 7.3.6.1 for full details of the plot.

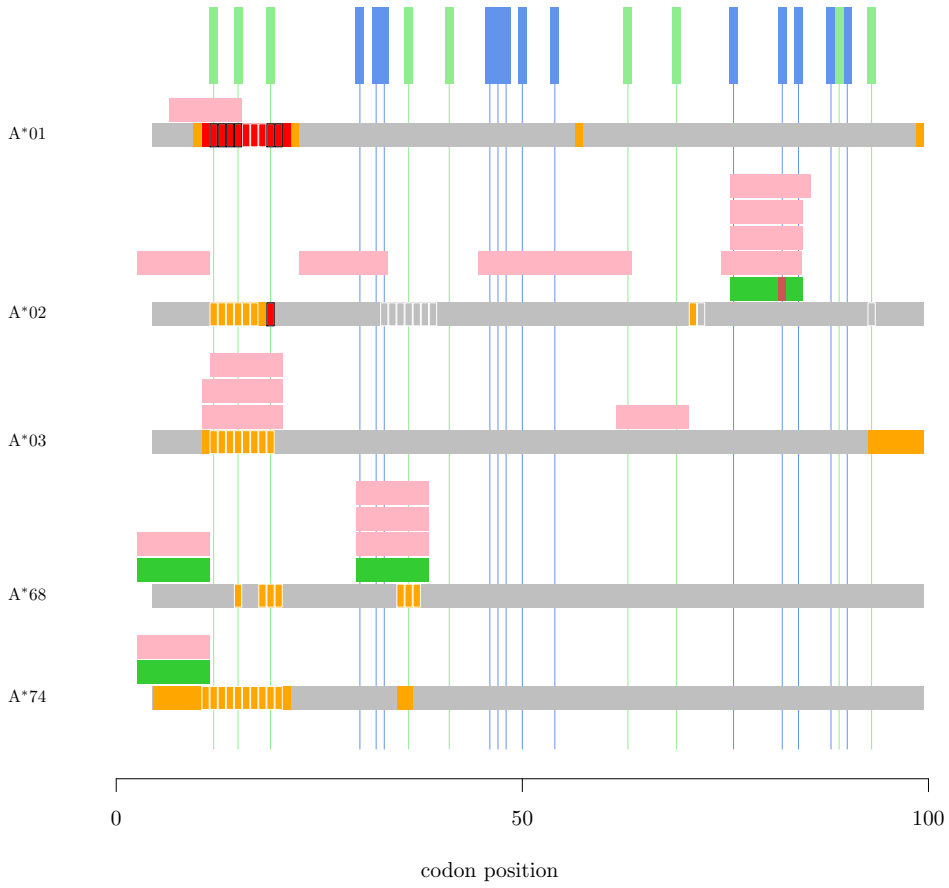


Figure B.6: A-list and B-list epitopes across protease: HLA-A alleles. Epitope plot showing epitopes associated to HLA-A alleles as defined by the A-list [174] and B-list [205]. Plotting is as in Figure 7.13

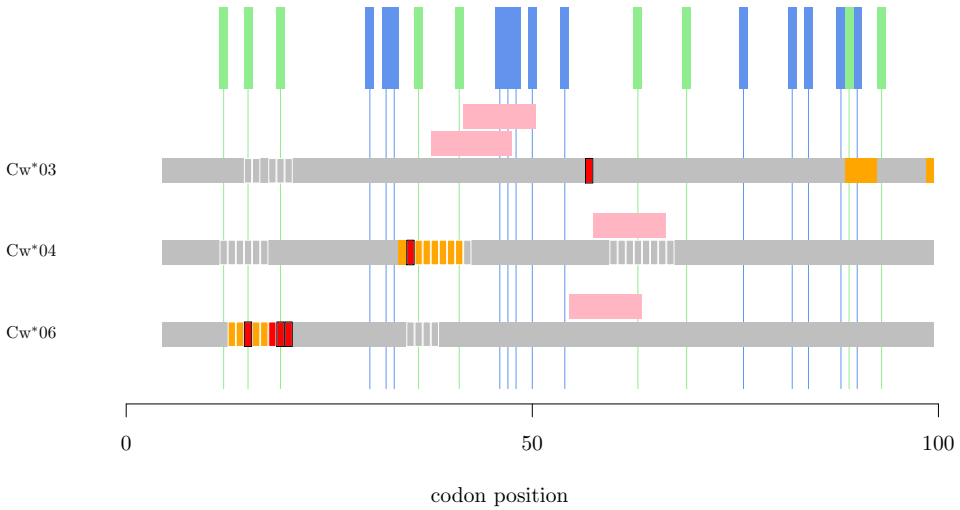


Figure B.7: A-list and B-list epitopes across protease: HLA-C alleles. Epitope plot showing epitopes associated to HLA-C alleles as defined by the A-list [174] and B-list [205]. Plotting is as in Figure 7.13

HLA 1	HLA 2	r^2	p-value
B*14	Cw*08	0.803	8.89×10^{-31}
B*35	Cw*04	0.422	2.02×10^{-30}
B*44	Cw*05	0.352	7.17×10^{-26}
B*08	Cw*07	0.201	1.81×10^{-21}
B*07	Cw*07	0.217	1.37×10^{-20}
B*40	Cw*03	0.242	2.40×10^{-18}
A*01	B*08	0.225	1.12×10^{-16}
B*57	Cw*06	0.258	5.71×10^{-15}
B*27	Cw*02	0.218	3.14×10^{-13}
B*51	Cw*15	0.275	5.03×10^{-12}
B*15	Cw*03	0.144	1.45×10^{-11}
A*29	Cw*16	0.206	9.76×10^{-10}
B*44	Cw*16	0.125	1.05×10^{-9}
B*52	Cw*12	0.218	2.68×10^{-9}
B*37	Cw*06	0.156	4.22×10^{-9}
B*38	Cw*12	0.189	2.25×10^{-8}
B*27	Cw*01	0.121	8.42×10^{-8}
A*25	B*18	0.130	9.71×10^{-6}

Table B.1: Significant correlations between HLA type occurrences in HIV subtype B populations. p-value/(number of tests)< 0.05, and $r^2 > 0.1$.

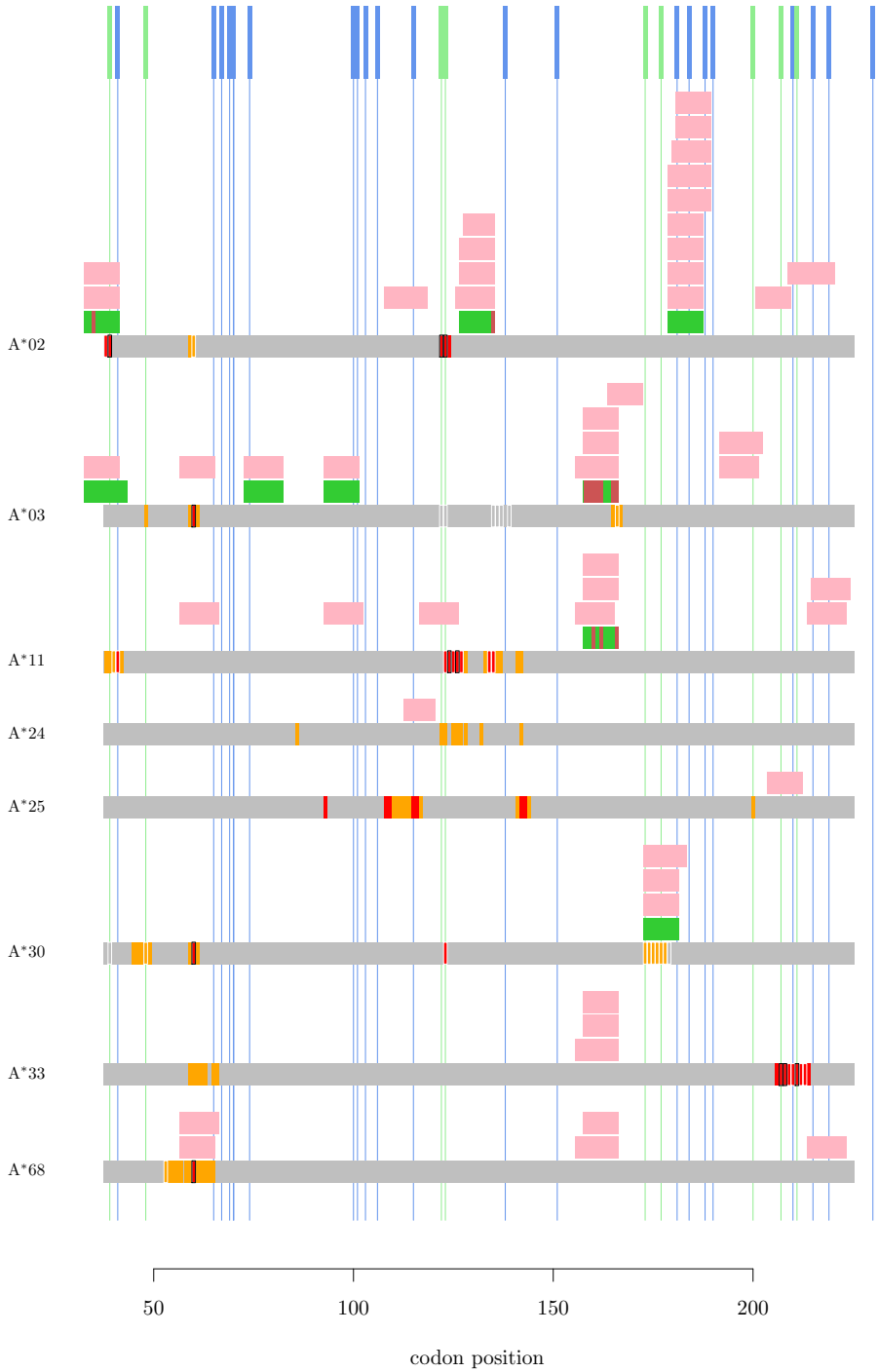


Figure B.8: A-list and B-list epitopes across reverse transcriptase: HLA-A alleles. Epitope plot showing epitopes associated to HLA-A alleles as defined by the A-list [174] and B-list [205]. Plotting is as in Figure 7.13

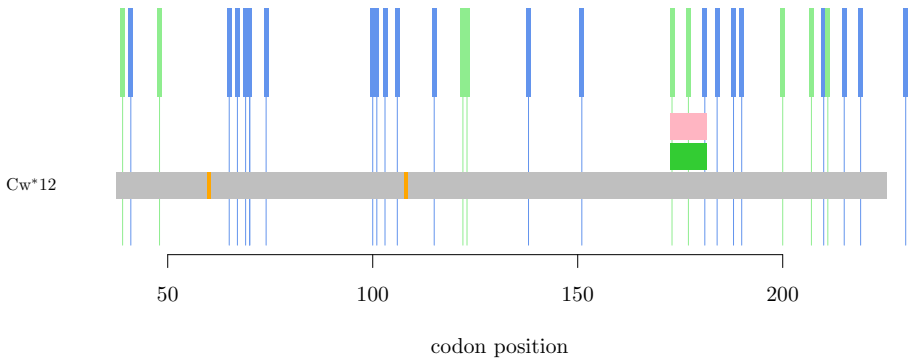


Figure B.9: A-list and B-list epitopes across reverse transcriptase: HLA-C alleles. Epitope plot showing epitopes associated to HLA-C alleles as defined by the A-list [174] and B-list [205]. Plotting is as in Figure 7.13

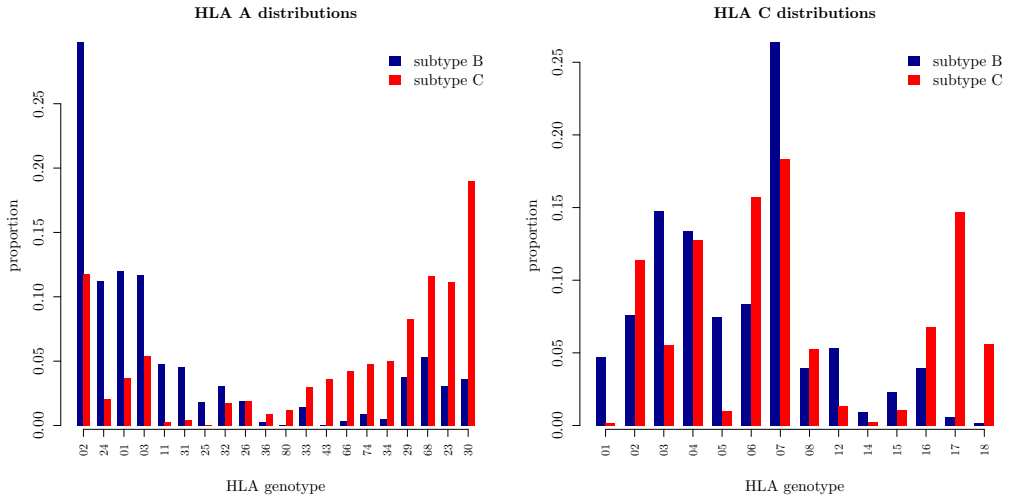


Figure B.10: Distribution of HLA-A and C alleles across subtype B and subtype C populations. We estimate viral subtypes of query sequences using RIP [229]. The subtypes most frequently defined in our query set were subtype B and subtype C. Consider the set of subtype B sequences. Using the corresponding HLA information we determine the HLA distribution in hosts associated to these viruses. This is an approximation of the HLA distribution in populations mostly infected with HIV subtype B. We do the same for subtype C. Bars are ordered by the absolute difference between HLA proportions in the subtype B and subtype C query sequence sets.

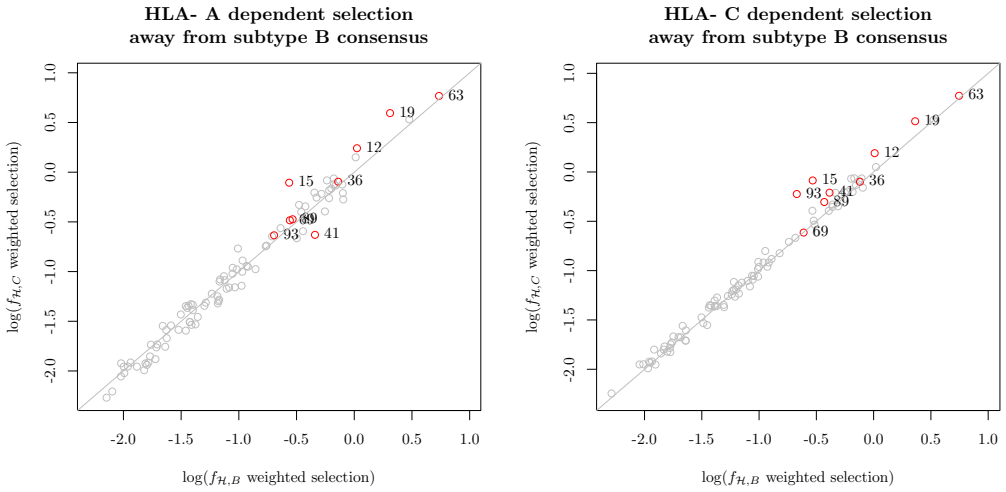
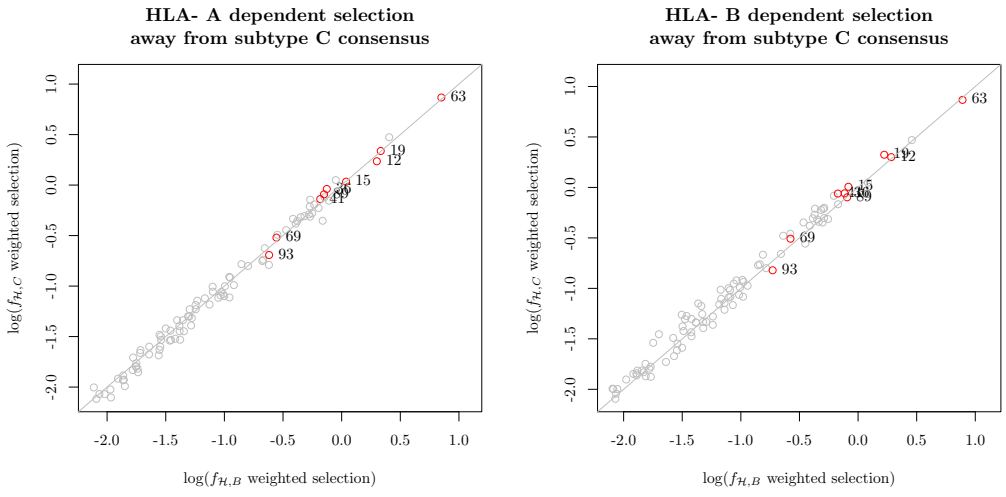


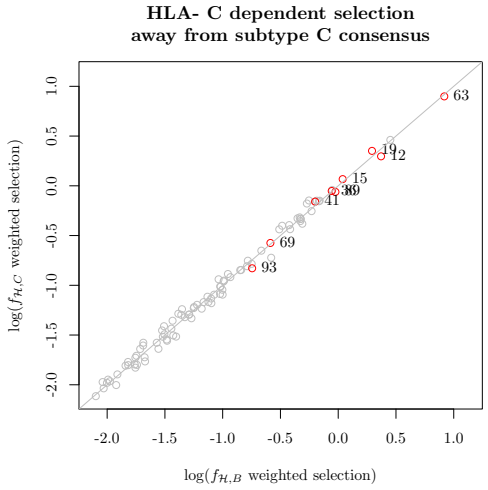
Figure B.11: Scatterplots showing selection away from a subtype consensus conditional on HLA distribution: protease, HLA-A and HLA-C. Plot annotation is as in Figure 7.19.

HLA 1	HLA 2	r^2	p-value
B*42	Cw*17	0.846	7.81×10^{-204}
B*58	Cw*06	0.520	1.26×10^{-122}
B*14	Cw*08	0.551	2.64×10^{-69}
B*45	Cw*16	0.451	5.02×10^{-67}
A*01	Cw*18	0.304	4.22×10^{-41}
A*01	B*81	0.318	3.56×10^{-40}
B*81	Cw*18	0.279	5.21×10^{-40}
B*15	Cw*02	0.152	2.27×10^{-34}
B*07	Cw*07	0.135	4.51×10^{-31}
B*53	Cw*04	0.146	4.71×10^{-29}
B*08	Cw*07	0.122	1.89×10^{-28}
B*39	Cw*12	0.439	3.75×10^{-27}
A*34	B*44	0.134	2.45×10^{-24}
A*80	B*18	0.209	2.84×10^{-21}
A*74	B*35	0.137	3.75×10^{-18}
A*36	B*53	0.133	1.01×10^{-12}
B*39	Cw*15	0.124	1.43×10^{-10}

Table B.2: Significant correlations between HLA type occurrences in HIV subtype C populations. p-value/(number of tests)< 0.05, and $r^2 > 0.1$.



(a) (b)



(c)

Figure B.12: Scatterplots showing selection away from a subtype C consensus conditional on HLA distribution: protease. Plot annotation is as in Figure 7.19. Plots are for the HLA-A, B and C frequencies respectively.

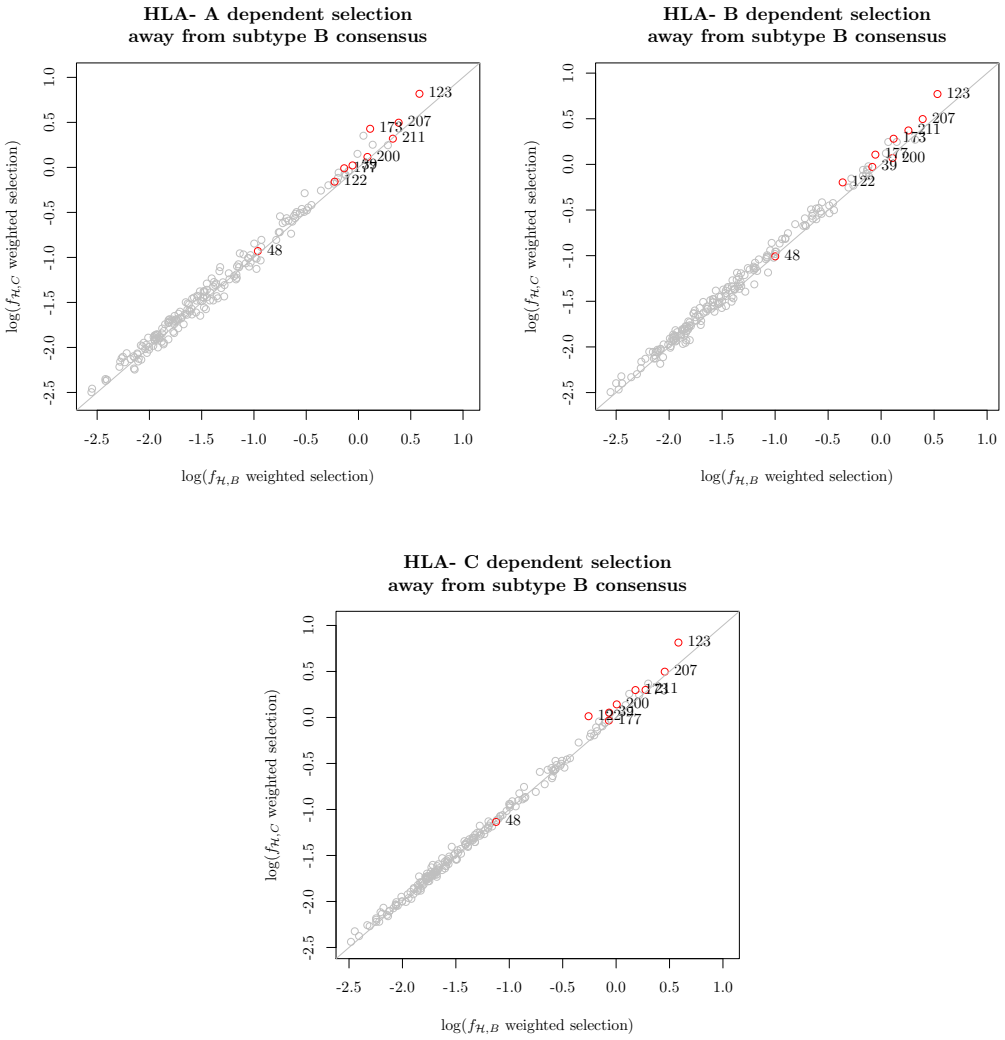
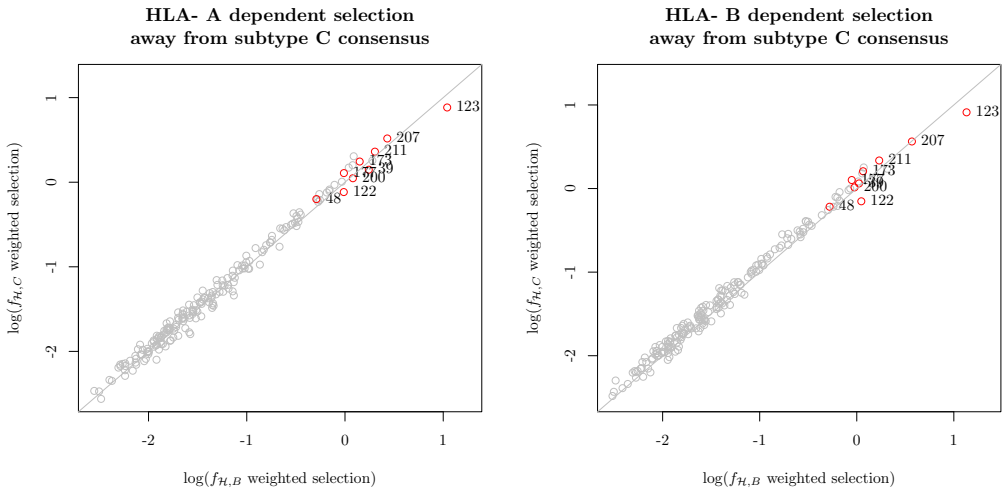


Figure B.13: Scatterplots showing selection away from subtype B consensus conditional on HLA frequencies: reverse transcriptase. Plots are as described in Figure 7.19. Plots are for HLA-A, B and C frequencies respectively.

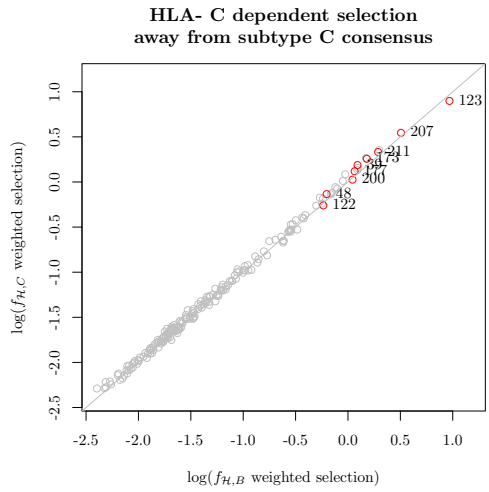
region	HLA-A allele	site(s)
protease	A*11	77
	A*32	77
	A*02	71
	A*25	46
reverse transcriptase	A*11	127, 134, 135
	A*26	134, 135, 196, 197, 198, 214
	A*29	214
	A*43	139, 140, 141, 147, 148, 149, 152, 153, 154, 155, 156
	A*68	53

Table B.3: Top-tier and second-tier candidate sites for HLA-A associated selection that occur outside known epitopes are not at sites that segregrate subtype B and subtype C. As in Section 7.3.9.2, top-tier candidate sites are outlined and highlighted in grey.



(a)

(b)



(c)

Figure B.14: Scatterplots showing selection away from a subtype C consensus conditional on HLA distribution: reverse transcriptase. Plot annotation is as in Figure 7.19. Plots are for the HLA-A, B and C frequencies respectively.

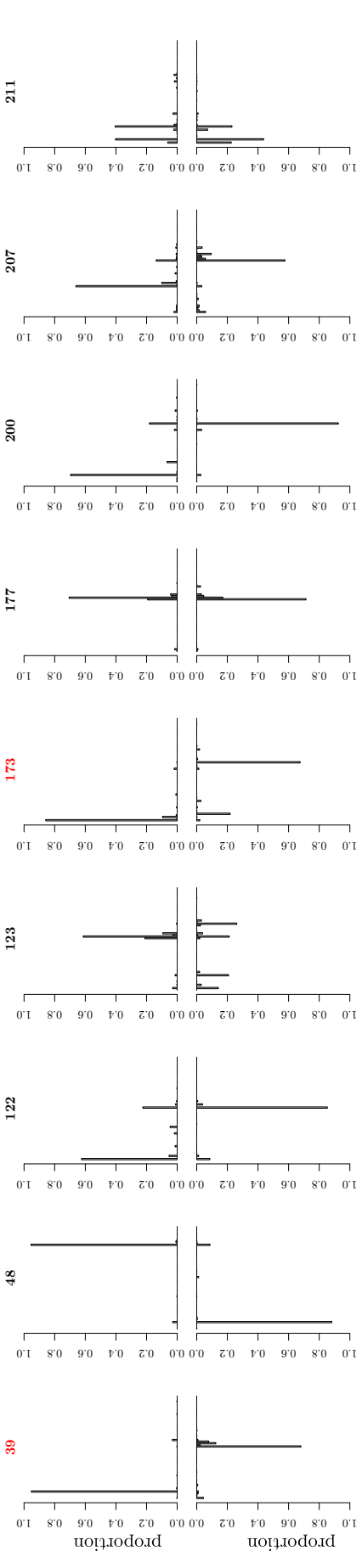


Figure B.15: Codon distribution at sites segregating subtype B and subtype C sequences in reverse transcriptase at the amino acid level. We plot the distribution of codons alphabetically from left to right at site which distinguish subtype B and C. Subtype B distributions are shown on the top row, and subtype C distributions are plotted on the bottom row. We plot subtype C distributions upside down to guide the eye to difference between the two subtypes more clearly. Sites at which non-synonymous codon differences between subtype consensus at the site is the result of >1 base change within the codon are coloured in red.

region	HLA-C allele	site(s)
protease	Cw*03	57
	Cw*05	62
	Cw*12	60
	Cw*18	35
	Cw*12	61, 62
	Cw*15	62, 64
	Cw*17	57, 58, 59
reverse transcriptase	Cw*15	196
	Cw*01	68
	Cw*12	202
	Cw*14	148
	Cw*15	147, 148

Table B.4: Top-tier and second-tier candidate sites for HLA-C associated selection that occur outside known epitopes are not at sites that segregrate subtype B and subtype C. As in Section 7.3.9.2, top-tier candidate sites are outlined and highlighted in grey.

BIBLIOGRAPHY

[1] Notices. *Cladistics*, 5(2):163–166, June 1989. doi: 10.1111/j.1096-0031.1989.tb00562.x. URL http://dx.doi.org/10.1111/j.1096-0031.1989.tb00562.x.

[2] *The Evolution of HIV*. Johns Hopkins University Press, 1 edition, March 1999. ISBN 0801861519. URL http://www.worldcat.org/isbn/0801861519.

[3] Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, June 2007. ISSN 0028-0836. doi: 10.1038/nature05911. URL http://dx.doi.org/10.1038/nature05911.

[4] Short-Course Antiretroviral Therapy in Primary HIV Infection. *N Engl J Med*, 368(3):207–217, January 2013. doi: 10.1056/nejmoa1110039. URL http://dx.doi.org/10.1056/nejmoa1110039.

[5] M-R R. Abrahams, J. A. Anderson, E. E. Giorgi, C. Seoighe, K. Mlisana, L-H H. Ping, G. S. Athreya, F. K. Treurnicht, B. F. Keele, N. Wood, J. F. Salazar-Gonzalez, T. Bhattacharya, H. Chu, I. Hoffman, S. Galvin, C. Mapanje, P. Kazembe, R. Thebus, S. Fiscus, W. Hide, M. S. Cohen, S. Abdool Karim, B. F. Haynes, G. M. Shaw, B. H. Hahn, B. T. Korber, R. Swanstrom, C. Williamson, CAPRISA Acute Infection Study Team, and 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. *Journal of virology*, 83(8): 3556–3567, April 2009. ISSN 1098-5514. doi: 10.1128/jvi.02132-08. URL http://dx.doi.org/10.1128/jvi.02132-08.

[6] Karsten Ahnert and Mario Mulansky. Odeint – Solving Ordinary Differential Equations in C + +. *AIP Conference Proceedings*, 1389(1):1586–1589, 2011. doi: 10.1063/1.3637934. URL http://dx.doi.org/10.1063/1.3637934.

[7] S. Munir Alam, Richard M. Scearce, Robert J. Parks, Kelly Plonk, Steven G. Plonk, Laura L. Sutherland, Miroslaw K. Gorny, Susan Zolla-Pazner, Stacie Vanleeuwen, M. Anthony Moody, Shi-Mao M. Xia, David C. Montefiori, Georgia D. Tomaras, Kent J. Weinhold, Salim Abdool A. Karim, Charles B. Hicks, Hua-Xin X. Liao, James Robinson, George M. Shaw, and Barton F. Haynes. Human immunodeficiency virus type 1 gp41 antibodies that mask membrane proximal region epitopes: antibody binding kinetics, induction, and potential for regulation in acute infection. *Journal of virology*, 82(1): 115–125, January 2008. ISSN 1098-5514. doi: 10.1128/jvi.00927-07. URL http://dx.doi.org/10.1128/jvi.00927-07.

[8] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell: Reference Edition*. Garland Science, 5 edition, November 2007. ISBN 0815341113. URL http://www.worldcat.org/isbn/0815341113.

[9] Todd M. Allen, Marcus Altfeld, Xu G. Yu, Kristin M. O'Sullivan, Mathias Lichterfeld, Sylvie Le Gall, Mina John, Bianca R. Mothe, Paul K. Lee, Elizabeth T. Kalife, Daniel E. Cohen, Kenneth A. Freedberg, Daryld A. Strick, Mary N. Johnston, Alessandro Sette, Eric S. Rosenberg, Simon A. Mallal, Philip J. Goulder, Christian Brander, and Bruce D. Walker. Selection, transmission, and reversion of an antigen-processing cytotoxic T-lymphocyte escape mutation in human immunodeficiency virus type 1 infection. *Journal of virology*, 78(13):7069–7078, July 2004. ISSN 0022-538X. doi: 10.1128/jvi.78.13.7069-7078.2004. URL http://dx.doi.org/10.1128/jvi.78.13.7069-7078.2004.

[10] Marcus Altfeld, Todd M. Allen, Xu G. Yu, Mary N. Johnston, Deepak Agrawal, Bette T. Korber, David C. Montefiori, David H. O'Connor, Ben T. Davis, Paul K. Lee, Erica L. Maier, Jason Harlow, Philip J. Goulder, Christian Brander, Eric S. Rosenberg, and Bruce D. Walker. HIV-1 superinfection

despite broad CD8+ T-cell responses containing replication of the primary virus. *Nature*, 420(6914):434–439, November 2002. ISSN 0028-0836. URL http://view.ncbi.nlm.nih.gov/pubmed/12459786.

[11] Marcus Altfeld, Marylyn M. Addo, Eric S. Rosenberg, Frederick M. Hecht, Paul K. Lee, Martin Vogel, Xu G. Yu, Rika Draenert, Mary N. Johnston, Daryld Strick, Todd M. Allen, Margaret E. Feeney, James O. Kahn, Rafick P. Sekaly, Jay A. Levy, Jürgen K. Rockstroh, Philip J. Goulder, and Bruce D. Walker. Influence of HLA-B57 on clinical presentation and viral control during acute HIV-1 infection. *AIDS (London, England)*, 17(18):2581–2591, December 2003. ISSN 0269-9370. URL http://view.ncbi.nlm.nih.gov/pubmed/14685052.

[12] Maria Anisimova, Rasmus Nielsen, and Ziheng Yang. Effect of Recombination on the Accuracy of the Likelihood Method for Detecting Positive Selection at Amino Acid Sites. *Genetics*, 164(3):1229–1236, July 2003. ISSN 1943-2631. URL http://www.genetics.org/content/164/3/1229.abstract.

[13] Richard Apps, Ying Qi, Jonathan M. Carlson, Haoyan Chen, Xiaojiang Gao, Rasmi Thomas, Yuko Yuki, Greg Q. Del Prete, Philip Goulder, Zabrina L. Brumme, Chanson J. Brumme, Mina John, Simon Mallal, George Nelson, Ronald Bosch, David Heckerman, Judy L. Stein, Kelly A. Soderberg, M. Anthony Moody, Thomas N. Denny, Xue Zeng, Jingyuan Fang, Ashley Moffett, Jeffrey D. Lifson, James J. Goedert, Susan Buchbinder, Gregory D. Kirk, Jacques Fellay, Paul McLaren, Steven G. Deeks, Florencia Pereyra, Bruce Walker, Nelson L. Michael, Amy Weintrob, Steven Wolinsky, Wilson Liao, and Mary Carrington. Influence of HLA-C Expression Level on HIV Control. *Science*, 340(6128):87–91, April 2013. doi: 10.1126/science.1232685. URL http://dx.doi.org/10.1126/science.1232685.

[14] Becca Asquith and Angela R. McLean. In vivo CD8+ T cell control of immunodeficiency virus infection in humans and macaques. *Proceedings of the National Academy of Sciences*, 104(15):6365–6370, April 2007. doi: 10.1073/pnas.0700666104. URL http://dx.doi.org/10.1073/pnas.0700666104.

[15] Becca Asquith, Charles T. T. Edwards, Marc Lipsitch, and Angela R. McLean. Inefficient Cytotoxic T Lymphocyte-Mediated Killing of HIV-1-Infected Cells In Vivo. *PLoS Biol*, 4(4):e90+, March 2006. doi: 10.1371/journal.pbio.0040090. URL http://dx.doi.org/10.1371/journal.pbio.0040090.

[16] Ahidjo Ayouba, Sandrine Souquières, Bernadette Njinku, Paul M. V. Martin, Michaela C. Müller-Trutwin, Pierre Roques, Françoise Barré-Sinoussi, Philippe Mauclère, François Simon, and Eric Nerrienet. HIV-1 group N among HIV-1-seropositive individuals in Cameroon. *AIDS*, 14(16):2623–2625, November 2000. ISSN 0269-9370. doi: 10.1097/00002030-200011100-00033. URL http://dx.doi.org/10.1097/00002030-200011100-00033.

[17] Katharine J. Bar, Hui Li, Annie Chamberland, Cecile Tremblay, Jean P. Routy, Truman Grayson, Chuanxi Sun, Shuyi Wang, Gerald H. Learn, Charity J. Morgan, Joseph E. Schumacher, Barton F. Haynes, Brandon F. Keele, Beatrice H. Hahn, and George M. Shaw. Wide Variation in the Multiplicity of HIV-1 Infection among Injection Drug Users. *J. Virol.*, 84(12):6241–6247, June 2010. doi: 10.1128/jvi.00077-10. URL http://dx.doi.org/10.1128/jvi.00077-10.

[18] F. Barre-Sinoussi, J. C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnier. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, 220(4599):868–871, May 1983. ISSN 1095-9203. doi: 10.1126/science.6189183. URL http://dx.doi.org/10.1126/science.6189183.

[19] István Bartha, Jonathan M. Carlson, Chanson J. Brumme, Paul J. McLaren, Zabrina L. Brumme, Mina John, David W. Haas, Javier Martinez-Picado, Judith Dalmau, Cecilio López-Galíndez, Concepción Casado, Andri Rauch, Huldrych F. Günthard, Enos Bernasconi, Pietro Vernazza, Thomas Klimkait, Sabine Yerly, Stephen J. O'Brien, Jennifer Listgarten, Nico Pfeifer, Christoph Lippert, Nicolo Fusi, Zoltán Kutalik, Todd M. Allen,

Viktor Müller, P. Richard Harrigan, David Heckerman, Amalio Telenti, Jacques Fellay, and Gil McVean. A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *eLife*, 2, October 2013. ISSN 2050-084X. doi: 10.7554/elife.01123. URL http://dx.doi.org/10.7554/elife.01123.

[20] W. J. Bellini and P. A. Rota. Genetic diversity of wild-type measles viruses: implications for global measles elimination programs. *Emerging infectious diseases*, 4(1):29–35, 1998. ISSN 1080-6040. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2627654/.

[21] Dennis A. Benson, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. GenBank. *Nucleic Acids Research*, 42(D1): D32–D37, January 2014. ISSN 1362-4962. doi: 10.1093/nar/gkt1030. URL http://dx.doi.org/10.1093/nar/gkt1030.

[22] Tanmoy Bhattacharya, Marcus Daniels, David Heckerman, Brian Foley, Nicole Frahm, Carl Kadie, Jonathan Carlson, Karina Yusim, Ben McMahon, Brian Gaschen, Simon Mallal, James I. Mullins, David C. Nickle, Joshua Herbeck, Christine Rousseau, Gerald H. Learn, Toshiyuki Miura, Christian Brander, Bruce Walker, and Bette Korber. Founder Effects in the Assessment of HIV Polymorphisms and HLA Allele Associations. *Science*, 315(5818):1583–1586, March 2007. doi: 10.1126/science.1131528. URL http://dx.doi.org/10.1126/science.1131528.

[23] Roman Biek, J. Caroline Henderson, Lance A. Waller, Charles E. Rupprecht, and Leslie A. Real. A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. *Proceedings of the National Academy of Sciences*, 104(19):7993–7998, May 2007. doi: 10.1073/pnas.0700741104. URL http://dx.doi.org/10.1073/pnas.0700741104.

[24] Monogram Biosciences. http://http://www.monogrambio.com.

[25] Joseph Bockhorst, Fangli Lu, Joel H. Janes, Jon Keebler, Benoit Gamain, Philip Awadalla, Xin-zhuan Su, Ram Samudrala, Nebojsa Jojic, and Joseph D. Smith. Structural polymorphism and diversifying selection

on the pregnancy malaria vaccine candidate VAR2CSA. *Molecular and Biochemical Parasitology*, 155(2):103–112, October 2007. ISSN 01666851. doi: 10.1016/j.molbiopara.2007.06.007. URL http://dx.doi.org/10.1016/j.molbiopara.2007.06.007.

[26] P. Borrow, H. Lewicki, B. H. Hahn, G. M. Shaw, and M. B. Oldstone. Virus-specific CD8+ cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. *J. Virol.*, 68(9):6103–6110, September 1994. URL http://jvi.asm.org/cgi/content/abstract/68/9/6103.

[27] Persephone Borrow, Hanna Lewicki, Xiping Wei, Marc S. Horwitz, Nancy Peffer, Heather Meyers, Jay A. Nelson, Jean E. Gairin, Beatrice H. Hahn, Michael B. A. Oldstone, and George M. Shaw. Antiviral pressure exerted by HIV-l-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. *Nature Medicine*, 3 (2):205–211, February 1997. doi: 10.1038/nm0297-205. URL http://dx.doi.org/10.1038/nm0297-205.

[28] P. Borst. Control of VSG gene expression sites in Trypanosoma brucei. *Molecular and Biochemical Parasitology*, 91(1):67–76, March 1998. ISSN 01666851. doi: 10.1016/s0166-6851(97)00184-9. URL http://dx.doi.org/10.1016/s0166-6851(97)00184-9.

[29] Christian L. Boutwell and M. Essex. Identification of HLA class I-associated amino acid polymorphisms in the HIV-1C proteome. *AIDS research and human retroviruses*, 23(1):165–174, January 2007. ISSN 0889-2229. URL http://view.ncbi.nlm.nih.gov/pubmed/17263647.

[30] Christian L. Boutwell, Jonathan M. Carlson, Tien-Ho Lin, Aaron Seese, Karen A. Power, Jian Peng, Yanhua Tang, Zabrina L. Brumme, David Heckerman, Arne Schneidewind, and Todd M. Allen. Frequent and Variable Cytotoxic-T-Lymphocyte Escape-associated Fitness Costs in the Human Immunodeficiency Virus Type 1 Subtype B Gag Proteins. *Journal of Virology*, 87(7):3952–3965, January 2013. ISSN 1098-5514. doi: 10.1128/jvi.03233-12. URL http://dx.doi.org/10.1128/jvi.03233-12.

[31] Zabrina L. Brumme, Chanson J. Brumme, Jonathan Carlson, Hendrik Streeck, Mina John, Quentin Eichbaum, Brian L. Block, Brett Baker, Carl Kadie, Martin Markowitz, Heiko Jessen, Anthony D. Kelleher, Eric Rosenberg, John Kaldor, Yuko Yuki, Mary Carrington, Todd M. Allen, Simon Mallal, Marcus Altfeld, David Heckerman, and Bruce D. Walker. 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. *Journal of virology*, 82(18):9216–9227, September 2008. ISSN 1098-5514. doi: 10.1128/jvi.01041-08. URL http://dx.doi.org/10.1128/jvi.01041-08.

[32] Zabrina L. Brumme, Mina John, Jonathan M. Carlson, Chanson J. Brumme, Dennison Chan, Mark A. Brockman, Luke C. Swenson, Iris Tao, Sharon Szeto, Pamela Rosato, Jennifer Sela, Carl M. Kadie, Nicole Frahm, Christian Brander, David W. Haas, Sharon A. Riddler, Richard Haubrich, Bruce D. Walker, P. Richard Harrigan, David Heckerman, and Simon Mallal. HLA-Associated Immune Escape Pathways in HIV-1 Subtype B Gag, Pol and Nef Proteins. *PLoS ONE*, 4(8):e6687+, August 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0006687. URL http://dx.doi.org/10.1371/journal.pone.0006687.

[33] Zabrina L. Brumme, Denis R. Chopera, and Mark A. Brockman. Modulation of HIV reservoirs by host HLA: bridging the gap between vaccine and cure. *Current Opinion in Virology*, 2(5):599–605, October 2012. ISSN 18796257. doi: 10.1016/j.coviro.2012.08.003. URL http://dx.doi.org/10.1016/j.coviro.2012.08.003.

[34] Dennis R. Burton, Ronald C. Desrosiers, Robert W. Doms, Wayne C. Koff, Peter D. Kwong, John P. Moore, Gary J. Nabel, Joseph Sodroski, Ian A. Wilson, and Richard T. Wyatt. HIV vaccine design and the neutralizing antibody problem. *Nature immunology*, 5(3):233–236, March 2004. ISSN 1529-2908. doi: 10.1038/ni0304-233. URL http://dx.doi.org/10.1038/ni0304-233.

[35] H. Cao, P. Kanki, J. L. Sankalé, A. Dieng-Sarr, G. P. Mazzara, S. A. Kalams, B. Korber, S. Mboup, and B. D. Walker. Cytotoxic T-lymphocyte cross-reactivity among different human immunodeficiency virus type 1 clades: implications for vaccine development. *Journal of virology*, 71(11):8615–8623, November 1997. ISSN 0022-538X. URL http://view.ncbi.nlm.nih.gov/pubmed/9343219.

[36] Jonathan M. Carlson, Zabrina L. Brumme, Christine M. Rousseau, Chanson J. Brumme, Philippa Matthews, Carl Kadie, James I. Mullins, Bruce D. Walker, P. Richard Harrigan, Philip J. R. Goulder, and David Heckerman. Phylogenetic Dependency Networks: Inferring Patterns of CTL Escape and Codon Covariation in HIV-1 Gag. *PLoS Comput Biol*, 4(11):e1000225+, November 2008. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000225. URL http://dx.doi.org/10.1371/journal.pcbi.1000225.

[37] Jonathan M. Carlson, Malinda Schaefer, Daniela C. Monaco, Rebecca Batorsky, Daniel T. Claiborne, Jessica Prince, Martin J. Deymier, Zachary S. Ende, Nichole R. Klatt, Charles E. DeZiel, Tien-Ho Lin, Jian Peng, Aaron M. Seese, Roger Shapiro, John Frater, Thumbi Ndung'u, Jianming Tang, Paul Goepfert, Jill Gilmour, Matt A. Price, William Kilembe, David Heckerman, Philip J. R. Goulder, Todd M. Allen, Susan Allen, and Eric Hunter. Selection bias at the heterosexual HIV-1 transmission bottleneck. *Science*, 345 (6193):1254031+, July 2014. ISSN 1095-9203. doi: 10.1126/science.1254031. URL http://dx.doi.org/10.1126/science.1254031.

[38] A. Carmichael, X. Jin, P. Sissons, and L. Borysiewicz. Quantitative analysis of the human immunodeficiency virus type 1 (HIV-1)-specific cytotoxic T lymphocyte (CTL) response at different stages of HIV-1 infection: differential CTL responses to HIV-1 and Epstein-Barr virus in late disease. *The Journal of Experimental Medicine*, 177(2):249–256, February 1993. doi: 10.1084/jem.177.2.249. URL http://dx.doi.org/10.1084/jem.177.2.249.

[39] Mary Carrington and Stephen J. O'Brien. The Influence of HLA Genotype on AIDS*. *Annual Review of Medicine*, 54(1):535–551, 2003. doi: 10.

1146/annurev.med.54.101601.152346. URL http://dx.doi.org/10.1146/annurev.med.54.101601.152346.

[40] Mary Carrington, George W. Nelson, Maureen P. Martin, Teri Kissner, David Vlahov, James J. Goedert, Richard Kaslow, Susan Buchbinder, Keith Hoots, and Stephen J. O'Brien. HLA and HIV-1: Heterozygote Advantage and B*35-Cw*04 Disadvantage. *Science*, 283(5408):1748–1752, March 1999. doi: 10.1126/science.283.5408.1748. URL http://dx.doi.org/10.1126/science.283.5408.1748.

[41] Gabriel Catano, Hemant Kulkarni, Weijing He, Vincent C. Marconi, Brian K. Agan, Michael Landrum, Stephanie Anderson, Judith Delmar, Vanessa Telles, Li Song, John Castiblanco, Robert A. Clark, Matthew J. Dolan, and Sunil K. Ahuja. HIV-1 disease-influencing effects associated with ZNRD1, HCP5 and HLA-C alleles are attributable mainly to either HLA-A10 or HLA-B*57 alleles. *PloS one*, 3(11), 2008. ISSN 1932-6203. doi: 10.1371/journal.pone.0003636. URL http://dx.doi.org/10.1371/journal.pone.0003636.

[42] Centers for Disease Control (CDC). Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men–New York City and California. *MMWR. Morbidity and mortality weekly report*, 30(25):305–308, July 1981. ISSN 0149-2195. URL http://view.ncbi.nlm.nih.gov/pubmed/6789108.

[43] Centers for Disease Control (CDC). Pneumocystis pneumonia–Los Angeles. *MMWR. Morbidity and mortality weekly report*, 30(21):250–252, June 1981. ISSN 0149-2195. URL http://view.ncbi.nlm.nih.gov/pubmed/6265753.

[44] Bikram Chakraborty, Luisa Valer, Carmen De Mendoza, Vicente Soriano, and Miguel E. Quiñones Mateu. Failure to detect human immunodeficiency virus type 1 superinfection in 28 HIV-seroconcordant individuals with high risk of reexposure to the virus. *AIDS research and human retroviruses*, 20(9):1026–1031, September 2004. ISSN 0889-2229. URL http://view.ncbi.nlm.nih.gov/pubmed/15585092.

[45] David C. Chan and Peter S. Kim. HIV Entry and Its Inhibition. *Cell*, 93 (5):681–684, May 1998. ISSN 00928674. doi: 10.1016/s0092-8674(00)81430-0. URL http://dx.doi.org/10.1016/s0092-8674(00)81430-0.

[46] Lamei Chen and Christopher Lee. Distinguishing HIV-1 drug resistance, accessory, and viral fitness mutations using conditional selection pressure analysis of treated versus untreated patient samples. *Biology direct*, 1(1): 14+, May 2006. ISSN 1745-6150. doi: 10.1186/1745-6150-1-14. URL http://dx.doi.org/10.1186/1745-6150-1-14.

[47] Lamei Chen, Alla Perlina, and Christopher J. Lee. Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. *Journal of virology*, 78(7):3722–3732, April 2004. ISSN 0022-538X. doi: 10.1128/jvi.78.7.3722-3732.2004. URL http://dx.doi.org/10.1128/jvi.78.7.3722-3732.2004.

[48] Ya-Lin Chiu, Vanessa B. Soros, Jason F. Kreisberg, Kim Stopak, Wes Yonemoto, and Warner C. Greene. Cellular APOBEC3G restricts HIV-1 infection in resting CD4+ T cells. *Nature*, 435(7038):108–114, April 2005. ISSN 0028-0836. doi: 10.1038/nature03493. URL http://dx.doi.org/10.1038/nature03493.

[49] Denis R. Chopera, Zenda Woodman, Koleka Mlisana, Mandla Mlotshwa, Darren P. Martin, Cathal Seoighe, Florette Treurnicht, Debra A. de Rosa, Winston Hide, Salim A. Karim, Clive M. Gray, Carolyn Williamson, , and the CAPRISA 002 Study Team. Transmission of HIV-1 CTL Escape Variants Provides HLA-Mismatched Recipients with a Survival Advantage. *PLoS Pathog*, 4(3):e1000033+, March 2008. ISSN 1553-7374. doi: 10.1371/journal.ppat.1000033. URL http://dx.doi.org/10.1371/journal.ppat.1000033.

[50] J. M. Coffin. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science*, 267(5197):483–489, January 1995. ISSN 0036-8075. doi: 10.1126/science.7824947. URL http://dx.doi.org/10.1126/science.7824947.

[51] The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, October 2005. ISSN 0028-0836. doi: 10.1038/nature04226. URL http://dx.doi.org/10.1038/nature04226.

[52] Dana C. Crawford, Tushar Bhangale, Na Li, Garrett Hellenthal, Mark J. Rieder, Deborah A. Nickerson, and Matthew Stephens. Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature Genetics*, 36(7):700–706, June 2004. ISSN 1061-4036. doi: 10.1038/ng1376. URL http://dx.doi.org/10.1038/ng1376.

[53] Deborah Cromer, Steven M. Wolinsky, and Angela R. McLean. How fast could HIV change gene frequencies in the human population? *Proceedings of the Royal Society B: Biological Sciences*, 277(1690):1981–1989, July 2010. ISSN 1471-2954. doi: 10.1098/rspb.2009.2073. URL http://dx.doi.org/10.1098/rspb.2009.2073.

[54] Cyril Dalmasso, Wassila Carpentier, Laurence Meyer, Christine Rouzioux, Cécile Goujard, Marie-Laure L. Chaix, Olivier Lambotte, Véronique Avettand-Fenoel, Sigrid Le Clerc, Laure Denis D. de Senneville, Christiane Deveau, Faroudy Boufassa, Patrice Debré, Jean-François F. Delfraissy, Philippe Broet, Ioannis Theodorou, and ANRS Genome Wide Association 01. Distinct genetic loci control plasma HIV-RNA and cellular HIV-DNA levels in HIV-1 infection: the ANRS Genome Wide Association 01 study. *PloS one*, 3(12), 2008. ISSN 1932-6203. URL http://view.ncbi.nlm.nih.gov/pubmed/19107206.

[55] UK HIV Drug Resistance Database. http://www.ctu.mrc.ac.uk/hivrdb.

[56] Tulio de Oliveira, Oliver G. Pybus, Andrew Rambaut, Marco Salemi, Sharon Cassol, Massimo Ciccozzi, Giovanni Rezza, Guido C. Gattinara, Roberta D'Arrigo, Massimo Amicosante, Luc Perrin, Vittorio Colizzi, Carlo F. Perno, and Benghazi Study Group. Molecular Epidemiology: HIV-1 and HCV sequences from Libyan outbreak. *Nature*, 444(7121): 836–837, December 2006. ISSN 0028-0836. doi: 10.1038/444836a. URL http://dx.doi.org/10.1038/444836a.

[57] Lot de Witte, Alexey Nabatov, Marjorie Pion, Donna Fluitsma, Marein A. W. P. de Jong, Tanja de Gruijl, Vincent Piguet, Yvette van Kooyk, and Teunis B. H. Geijtenbeek. Langerin is a natural barrier to HIV-1 transmission by Langerhans cells. *Nature Medicine*, 13(3):367–371, March 2007. ISSN 1078-8956. doi: 10.1038/nm1541. URL http://dx.doi.org/10.1038/nm1541.

[58] Cynthia A. Derdeyn, Julie M. Decker, Frederic Bibollet-Ruche, John L. Mokili, Mark Muldoon, Scott A. Denham, Marintha L. Heil, Francis Kasolo, Rosemary Musonda, Beatrice H. Hahn, George M. Shaw, Bette T. Korber, Susan Allen, and Eric Hunter. Envelope-constrained neutralization-sensitive HIV-1 after heterosexual transmission. *Science*, 303(5666):2019–2022, March 2004. ISSN 1095-9203. doi: 10.1126/science.1093137. URL http://dx.doi.org/10.1126/science.1093137.

[59] Alexander Dilthey, Stephen Leslie, Loukas Moutsianas, Judong Shen, Charles Cox, Matthew R. Nelson, and Gil McVean. Multi-Population Classical HLA Type Imputation. *PLoS Comput Biol*, 9(2):e1002877+, February 2013. doi: 10.1371/journal.pcbi.1002877. URL http://dx.doi.org/10.1371/journal.pcbi.1002877.

[60] J. Donelson. Antigenic variation and the African trypanosome genome. *Acta Tropica*, 85(3):391–404, March 2003. ISSN 0001706X. doi: 10.1016/s0001-706x(02)00237-1. URL http://dx.doi.org/10.1016/s0001-706x(02)00237-1.

[61] Rika Draenert, Sylvie Le Gall, Katja J. Pfafferott, Alasdair J. Leslie, Polan Chetty, Christian Brander, Edward C. Holmes, Shih-Chung C. Chang, Margaret E. Feeney, Marylyn M. Addo, Lidia Ruiz, Danni Ramduth, Prakash Jeena, Marcus Altfeld, Stephanie Thomas, Yanhua Tang, Cori L. Verrill, Catherine Dixon, Julia G. Prado, Photini Kiepiela, Javier Martinez-Picado, Bruce D. Walker, and Philip J. Goulder. Immune selection for altered antigen processing leads to cytotoxic T lymphocyte escape in chronic HIV-1 infection. *The Journal of experimental medicine*, 199(7):905–915, April 2004. ISSN 0022-1007. doi: 10.1084/jem.20031982. URL http://dx.doi.org/10.1084/jem.20031982.

[62] Rika Draenert, Todd M. Allen, Yang Liu, Terri Wrin, Colombe Chappey, Cori L. Verrill, Guillem Sirera, Robert L. Eldridge, Matthew P. Lahaie, Lidia Ruiz, Bonaventura Clotet, Christos J. Petropoulos, Bruce D. Walker, and Javier Martinez-Picado. Constraints on HIV-1 evolution and immunodominance revealed in monozygotic adult twins infected with the same virus. *The Journal of experimental medicine*, 203(3):529–539, March 2006. ISSN 0022-1007. doi: 10.1084/jem.20052116. URL http://dx.doi.org/10.1084/jem.20052116.

[63] Stanford drug resistance database. http://hivdb.stanford.edu/.

[64] A. J. Drummond, A. Rambaut, B. Shapiro, and O. G. Pybus. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Molecular Biology and Evolution*, 22(5):1185–1192, May 2005. ISSN 1537-1719. doi: 10.1093/molbev/msi103. URL http://dx.doi.org/10.1093/molbev/msi103.

[65] Alexei Drummond and Andrew Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(1):214+, November 2007. ISSN 1471-2148. doi: 10.1186/1471-2148-7-214. URL http://dx.doi.org/10.1186/1471-2148-7-214.

[66] Alexei J. Drummond, Geoff K. Nicholls, Allen G. Rodrigo, and Wiremu Solomon. Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data. *Genetics*, 161(3):1307–1320, July 2002. ISSN 0016-6731. URL http://www.genetics.org/content/161/3/1307.abstract.

[67] Alexei J. Drummond, Simon Y. W. Ho, Matthew J. Phillips, and Andrew Rambaut. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biol*, 4(5):e88+, March 2006. doi: 10.1371/journal.pbio.0040088. URL http://dx.doi.org/10.1371/journal.pbio.0040088.

[68] Alexei J. Drummond, Marc A. Suchard, Dong Xie, and Andrew Rambaut. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology*

and evolution, 29(8):1969–1973, August 2012. ISSN 1537-1719. doi: 10.1093/molbev/mss075. URL http://dx.doi.org/10.1093/molbev/mss075.

[69] Maya Dymova, Oleksander Liashenko, Petro Poteiko, Valeriy Krutko, Eugeny Khrapov, and Maxim Filipenko. Genetic variation of Mycobacterium tuberculosis circulating in Kharkiv Oblast, Ukraine. *BMC Infectious Diseases*, 11(1):77+, March 2011. ISSN 1471-2334. doi: 10.1186/1471-2334-11-77. URL http://dx.doi.org/10.1186/1471-2334-11-77.

[70] Robert C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, March 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh340. URL http://dx.doi.org/10.1093/nar/gkh340.

[71] Robert C. Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, 5(1):113+, August 2004. ISSN 1471-2105. doi: 10.1186/1471-2105-5-113. URL http://dx.doi.org/10.1186/1471-2105-5-113.

[72] David T. Evans, H. O'ConnorDavid, Peicheng Jing, John L. Dzuris, John Sidney, Jack da Silva, Todd M. Allen, Helen Horton, John E. Venham, Richard A. Rudersdorf, Thorsten Vogel, C. David Pauza, Ronald E. Bontrop, Robert DeMars, Alessandro Sette, Austin L. Hughes, and David I. Watkins. Virus-specific cytotoxic T-lymphocyte responses select for amino-acid variation in simian immunodeficiency virus Env and Nef. *Nature Medicine*, 5(11):1270–1276, November 1999. ISSN 1078-8956. doi: 10.1038/15224. URL http://dx.doi.org/10.1038/15224.

[73] Catherine Fagard, Annette Oxenius, Huldrych Günthard, Felipe Garcia, Michelle Le Braz, Gabriel Mestre, Manuel Battegay, Hansjakob Furrer, Pietro Vernazza, Enos Bernasconi, Amalio Telenti, Rainer Weber, Dominique Leduc, Sabine Yerly, David Price, Sara J. Dawson, Thomas Klimkait, Thomas V. Perneger, Angela McLean, Bonaventura Clotet, Jose M. Gatell, Luc Perrin, Montserrat Plana, Rodney Phillips, Bernard Hirschel, and Swiss HIV Cohort Study. A prospective trial of structured treatment

interruptions in human immunodeficiency virus infection. *Archives of internal medicine*, 163(10):1220–1226, May 2003. ISSN 0003-9926. doi: 10.1001/archinte.163.10.1220. URL http://dx.doi.org/10.1001/archinte.163.10.1220.

[74] Nuno R. Faria, Andrew Rambaut, Marc A. Suchard, Guy Baele, Trevor Bedford, Melissa J. Ward, Andrew J. Tatem, João D. Sousa, Nimalan Arinaminpathy, Jacques Pépin, David Posada, Martine Peeters, Oliver G. Pybus, and Philippe Lemey. The early spread and epidemic ignition of HIV-1 in human populations. *Science*, 346(6205):56–61, October 2014. ISSN 1095-9203. doi: 10.1126/science.1256739. URL http://dx.doi.org/10.1126/science.1256739.

[75] Jacques Fellay, Dongliang Ge, Kevin V. Shianna, Sara Colombo, Bruno Ledergerber, Elizabeth T. Cirulli, Thomas J. Urban, Kunlin Zhang, Curtis E. Gumbs, Jason P. Smith, Antonella Castagna, Alessandro Cozzi-Lepri, Andrea De Luca, Philippa Easterbrook, Huldrych F. Günthard, Simon Mallal, Cristina Mussini, Judith Dalmau, Javier Martinez-Picado, José M. Miro, Niels Obel, Steven M. Wolinsky, Jeremy J. Martinson, Roger Detels, Joseph B. Margolick, Lisa P. Jacobson, Patrick Descombes, Stylianos E. Antonarakis, Jacques S. Beckmann, Stephen J. O'Brien, Norman L. Letvin, Andrew J. McMichael, Barton F. Haynes, Mary Carrington, Sheng Feng, Amalio Telenti, David B. Goldstein, and NIAID Center for HIV/AIDS Vaccine Immunology (CHAVI). Common genetic variation and the control of HIV-1 in humans. *PLoS genetics*, 5(12), December 2009. ISSN 1553-7404. URL http://view.ncbi.nlm.nih.gov/pubmed/20041166.

[76] Joseph Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, November 1981. ISSN 0022-2844. doi: 10.1007/bf01734359. URL http://dx.doi.org/10.1007/bf01734359.

[77] Guido Ferrari, Bette Korber, Nilu Goonetilleke, Michael K. Liu, Emma L. Turnbull, Jesus F. Salazar-Gonzalez, Natalie Hawkins, Steve Self, Sydeaka Watson, Michael R. Betts, Cynthia Gay, Kara McGhee, Pierre Pellegrino,

Ian Williams, Georgia D. Tomaras, Barton F. Haynes, Clive M. Gray, Persephone Borrow, Mario Roederer, Andrew J. McMichael, and Kent J. Weinhold. Relationship between functional profile of HIV-1 specific CD8 T cells and epitope variability with the selection of escape mutants in acute HIV-1 infection. *PLoS pathogens*, 7(2), 2011. ISSN 1553-7374. URL http://view.ncbi.nlm.nih.gov/pubmed/21347345.

[78] Diana Finzi, Joel Blankson, Janet D. Siliciano, Joseph B. Margolick, Karen Chadwick, Theodore Pierson, Kendall Smith, Julianna Lisziewicz, Franco Lori, Charles Flexner, Thomas C. Quinn, Richard E. Chaisson, Eric Rosenberg, Bruce Walker, Stephen Gange, Joel Gallant, and Robert F. Siliciano. Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nature Medicine*, 5(5):512–517, May 1999. doi: 10.1038/8394. URL http://dx.doi.org/10.1038/8394.

[79] Alan D. Frankel and John A. T. Young. HIV-1: Fifteen Proteins and an RNA. *Annual Review of Biochemistry*, 67(1):1–25, 1998. ISSN 0066-4154. doi: 10.1146/annurev.biochem.67.1.1. URL http://dx.doi.org/10.1146/annurev.biochem.67.1.1.

[80] Christophe Fraser, Christl A. Donnelly, Simon Cauchemez, William P. Hanage, Maria D. Van Kerkhove, T. Déirdre Hollingsworth, Jamie Griffin, Rebecca F. Baggaley, Helen E. Jenkins, Emily J. Lyons, Thibaut Jombart, Wes R. Hinsley, Nicholas C. Grassly, Francois Balloux, Azra C. Ghani, Neil M. Ferguson, Andrew Rambaut, Oliver G. Pybus, Hugo Lopez-Gatell, Celia M. Alpuche-Aranda, Ietza Bojorquez B. Chapela, Ethel Palacios P. Zavala, Dulce Ma Espejo M. Guevara, Francesco Checchi, Erika Garcia, Stephane Hugonnet, Cathy Roth, and WHO Rapid Pandemic Assessment Collaboration. Pandemic potential of a strain of influenza A (H1N1): early findings. *Science (New York, N.Y.)*, 324(5934):1557–1561, June 2009. ISSN 1095-9203. doi: 10.1126/science.1176062. URL http://dx.doi.org/10.1126/science.1176062.

[81] A. J. Frater, C. T. T. Edwards, N. McCarthy, J. Fox, H. Brown, A. Milicic, N. Mackie, T. Pillay, J. W. Drijfhout, S. Dustan, J. R. Clarke, E. C. Holmes, H. T. Zhang, K. Pfafferott, P. J. Goulder, M. O. McClure, J. Weber, R. E. Phillips, and S. Fidler. Passive Sexual Transmission of Human Immunodeficiency Virus Type 1 Variants and Adaptation in New Hosts. *J. Virol.*, 80(14):7226–7234, July 2006. doi: 10.1128/jvi.02014-05. URL http://dx.doi.org/10.1128/jvi.02014-05.

[82] A. J. Frater, H. Brown, A. Oxenius, H. F. Günthard, B. Hirschel, N. Robinson, A. J. Leslie, R. Payne, H. Crawford, A. Prendergast, C. Brander, P. Kiepiela, B. D. Walker, P. J. R. Goulder, A. McLean, R. E. Phillips, and the Swiss HIV-Cohort Study. Effective T-Cell Responses Select Human Immunodeficiency Virus Mutants and Slow Disease Progression. *Journal of Virology*, 81(12):6742–6751, June 2007. ISSN 1098-5514. doi: 10.1128/jvi.00022-07. URL http://dx.doi.org/10.1128/jvi.00022-07.

[83] Thomas C. Friedrich, Elizabeth J. Dodds, Levi J. Yant, Lara Vojnov, Richard Rudersdorf, Candice Cullen, David T. Evans, Ronald C. Desrosiers, Bianca R. Mothe, John Sidney, Alessandro Sette, Kevin Kunstman, Steven Wolinsky, Michael Piatak, Jeffrey Lifson, Austin L. Hughes, Nancy Wilson, David H. O'Connor, and David I. Watkins. Reversion of CTL escape-variant immunodeficiency viruses in vivo. *Nature Medicine*, 10(3):275–281, February 2004. ISSN 1078-8956. doi: 10.1038/nm998. URL http://dx.doi.org/10.1038/nm998.

[84] Simon D. W. Frost and Erik M. Volz. Viral phylodynamics and the search for an 'effective number of infections'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1548):1879–1890, June 2010. ISSN 1471-2970. doi: 10.1098/rstb.2010.0060. URL http://dx.doi.org/10.1098/rstb.2010.0060.

[85] Helen R. Fryer, John Frater, Anna Duda, Mick G. Roberts, Rodney E. Phillips, Angela R. McLean, and The SPARTAC Trial Investigators. Modelling the Evolution and Spread of HIV Immune Escape Mutants. *PLoS Pathog*, 6(11):e1001196+, November 2010. ISSN 1553-7374. doi: 10.1371/

journal.ppat.1001196. URL http://dx.doi.org/10.1371/journal.ppat.1001196.

[86] R. C. Gallo, P. S. Sarin, E. P. Gelmann, M. Robert-Guroff, E. Richardson, V. S. Kalyanaraman, D. Mann, G. D. Sidhu, R. E. Stahl, S. Zolla-Pazner, J. Leibowitch, and M. Popovic. Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). *Science*, 220(4599):865–867, May 1983. ISSN 1095-9203. doi: 10.1126/science.6601823. URL http://dx.doi.org/10.1126/science.6601823.

[87] Vitaly V. Ganusov, Nilu Goonetilleke, Michael K. P. Liu, Guido Ferrari, George M. Shaw, Andrew J. McMichael, Persephone Borrow, Bette T. Korber, and Alan S. Perelson. Fitness Costs and Diversity of the Cytotoxic T Lymphocyte (CTL) Response Determine the Rate of CTL Escape during Acute and Chronic Phases of HIV Infection. *Journal of Virology*, 85 (20):10518–10528, October 2011. ISSN 1098-5514. doi: 10.1128/jvi.00655-11. URL http://dx.doi.org/10.1128/jvi.00655-11.

[88] F. Gao, E. Bailes, D. L. Robertson, Y. Chen, C. M. Rodenburg, S. F. Michael, L. B. Cummins, L. O. Arthur, M. Peeters, G. M. Shaw, P. M. Sharp, and B. H. Hahn. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature*, 397(6718):436–441, February 1999. ISSN 0028-0836. URL http://view.ncbi.nlm.nih.gov/pubmed/9989410.

[89] Feng Gao, Elizabeth Bailes, David L. Robertson, Yalu Chen, Cynthia M. Rodenburg, Scott F. Michael, Larry B. Cummins, Larry O. Arthur, Martine Peeters, George M. Shaw, Paul M. Sharp, and Beatrice H. Hahn. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature*, 397(6718): 436–441, February 1999. ISSN 0028-0836. doi: 10.1038/17130. URL http://dx.doi.org/10.1038/17130.

[90] Xiaojiang Gao, George W. Nelson, Peter Karacki, Maureen P. Martin, John Phair, Richard Kaslow, James J. Goedert, Susan Buchbinder, Keith Hoots, David Vlahov, Stephen J. O'Brien, and Mary Carrington. Effect of a Single Amino Acid Change in MHC Class I Molecules on the Rate of Progression to AIDS. *New England Journal of Medicine*, 344(22):1668–1675, May

2001. doi: 10.1056/nejm200105313442203. URL http://dx.doi.org/10.1056/nejm200105313442203.

[91] B. Gaschen, C. Kuiken, B. Korber, and B. Foley. Retrieval and on-the-fly alignment of sequence fragments from the HIV database. *Bioinformatics (Oxford, England)*, 17(5):415–418, May 2001. ISSN 1367-4803. URL http://view.ncbi.nlm.nih.gov/pubmed/11331235.

[92] Hiroyuki Gatanaga, Hirotaka Ode, Atsuko Hachiya, Tsunefusa Hayashida, Hironori Sato, Masafumi Takiguchi, and Shinichi Oka. Impact of human leukocyte antigen-B*51-restricted cytotoxic T-lymphocyte pressure on mutation patterns of nonnucleoside reverse transcriptase inhibitor resistance. *AIDS (London, England)*, 24(5), March 2010. ISSN 1473-5571. URL http://view.ncbi.nlm.nih.gov/pubmed/20160632.

[93] Hiroyuki Gatanaga, Hayato Murakoshi, Atsuko Hachiya, Tsunefusa Hayashida, Takayuki Chikata, Hirotaka Ode, Kiyoto Tsuchiya, Wataru Sugiura, Masafumi Takiguchi, and Shinichi Oka. Naturally selected rilpivirine-resistant HIV-1 variants by host cellular immunity. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 57(7):1051–1055, October 2013. ISSN 1537-6591. URL http://view.ncbi.nlm.nih.gov/pubmed/23797286.

[94] Mark J. Geels, Marion Cornelissen, Hanneke Schuitemaker, Kiersten Anderson, David Kwa, Jolanda Maas, John T. Dekker, Elly Baan, Fokla Zorgdrager, Remco van den Burg, Martijn van Beelen, Vladimir V. Lukashov, Tong-Ming M. Fu, William A. Paxton, Lia van der Hoek, Sheri A. Dubey, John W. Shiver, and Jaap Goudsmit. Identification of sequential viral escape mutants associated with altered T-cell responses in a human immunodeficiency virus type 1-infected individual. *Journal of virology*, 77(23):12430–12440, December 2003. ISSN 0022-538X. doi: 10.1128/jvi.77.23.12430-12440.2003. URL http://dx.doi.org/10.1128/jvi.77.23.12430-12440.2003.

[95] Tanja Gernhard. The conditioned reconstructed process. *Journal of Theoretical Biology*, 253(4):769–778, August 2008. ISSN 00225193. doi: 10.1016/j.jtbi.2008.04.005. URL http://dx.doi.org/10.1016/j.jtbi.2008.04.005.

[96] Robert J. Gifford, Tulio de Oliveira, Andrew Rambaut, Oliver G. Pybus, David Dunn, Anne-Mieke M. Vandamme, Paul Kellam, Deenan Pillay, and UK Collaborative Group on HIV Drug Resistance. Phylogenetic surveillance of viral genetic diversity and the evolving molecular epidemiology of human immunodeficiency virus type 1. *Journal of virology*, 81(23):13050–13056, December 2007. ISSN 1098-5514. doi: 10.1128/jvi.00889-07. URL http://dx.doi.org/10.1128/jvi.00889-07.

[97] Matthew J. Gonzales, Eric Delwart, Soo-Yon Y. Rhee, Rose Tsui, Andrew R. Zolopa, Jonathan Taylor, and Robert W. Shafer. Lack of detectable human immunodeficiency virus type 1 superinfection during 1072 person-years of observation. *The Journal of infectious diseases*, 188(3):397–405, August 2003. ISSN 0022-1899. URL http://view.ncbi.nlm.nih.gov/pubmed/12870121.

[98] Nilu Goonetilleke, Michael K. P. Liu, Jesus F. Salazar-Gonzalez, Guido Ferrari, Elena Giorgi, Vitaly V. Ganusov, Brandon F. Keele, Gerald H. Learn, Emma L. Turnbull, Maria G. Salazar, Kent J. Weinhold, Stephen Moore, B. CHAVI Clinical Core, Norman Letvin, Barton F. Haynes, Myron S. Cohen, Peter Hraber, Tanmoy Bhattacharya, Persephone Borrow, Alan S. Perelson, Beatrice H. Hahn, George M. Shaw, Bette T. Korber, and Andrew J. McMichael. The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection. *The Journal of Experimental Medicine*, 206(6):1253–1272, June 2009. ISSN 1540-9538. doi: 10.1084/jem.20090365. URL http://dx.doi.org/10.1084/jem.20090365.

[99] Heinrich G. Göttlinger. HIV-1 Gag: a Molecular Machine Driving Viral Particle Assembly and Release. *HIV Sequence Compendium 2001*, 2001. URL http://74.125.127.132/scholar?q=cache:5qNWxSQ6tRwJ:scholar.google.com/+HIV-1+Gag:+a+Molecular+Machine+Driving+ViralParticle+Assembly+and+Release&hl=en&as_sdt=0,5.

[100] P. J. Goulder, R. E. Phillips, R. A. Colbert, S. McAdam, G. Ogg, M. A. Nowak, P. Giangrande, G. Luzzi, B. Morgan, A. Edwards, A. J. McMichael, and S. Rowland-Jones. Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS. *Nature medicine*, 3(2):212–217, February 1997. ISSN 1078-8956. URL http://view.ncbi.nlm.nih.gov/pubmed/9018241.

[101] Philip J. R. Goulder and Bruce D. Walker. HIV and HLA Class I: An Evolving Relationship. *Immunity*, 37(3):426–440, September 2012. ISSN 10747613. doi: 10.1016/j.immuni.2012.09.005. URL http://dx.doi.org/10.1016/j.immuni.2012.09.005.

[102] Philip J. R. Goulder and David I. Watkins. HIV and SIV CTL escape: implications for vaccine design. *Nat Rev Immunol*, 4(8):630–640, August 2004. doi: 10.1038/nri1417. URL http://dx.doi.org/10.1038/nri1417.

[103] Philip J. R. Goulder, Rodney E. Phillips, Robert A. Colbert, Stephen McAdam, Graham Ogg, Martin A. Nowak, Paul Giangrande, Graz Luzzi, Barbara Morgana, Anne Edwards, Andrew J. McMichael, and Sarah Rowland-Jones. Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS. *Nature Medicine*, 3(2):212–217, February 1997. doi: 10.1038/nm0297-212. URL http://dx.doi.org/10.1038/nm0297-212.

[104] Philip J. R. Goulder, Christian Brander, Yanhua Tang, Cecile Tremblay, Robert A. Colbert, Marylyn M. Addo, Eric S. Rosenberg, Thi Nguyen, Rachel Allen, Alicja Trocha, Marcus Altfeld, Suqin He, Michael Bunce, Robert Funkhouser, Stephen I. Pelton, Sandra K. Burchett, Kenneth McIntosh, Bette T. M. Korber, and Bruce D. Walker. Evolution and transmission of stable CTL escape mutations in HIV infection. *Nature*, 412 (6844):334–338, July 2001. ISSN 0028-0836. doi: 10.1038/35085576. URL http://dx.doi.org/10.1038/35085576.

[105] Elin S. Gray, Maphuti C. Madiga, Penny L. Moore, Koleka Mlisana, Salim S. Abdool Karim, James M. Binley, George M. Shaw, John R. Mascola, and Lynn Morris. Broad neutralization of human immunodeficiency virus type

1 mediated by plasma antibodies against the gp41 membrane proximal external region. *Journal of virology*, 83(21):11265–11274, November 2009. ISSN 1098-5514. doi: 10.1128/jvi.01359-09. URL http://dx.doi.org/10.1128/jvi.01359-09.

[106] Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, December 1995. ISSN 1464-3510. doi: 10.1093/biomet/82.4.711. URL http://dx.doi.org/10.1093/biomet/82.4.711.

[107] Bryan T. Grenfell, Oliver G. Pybus, Julia R. Gog, James L. N. Wood, Janet M. Daly, Jenny A. Mumford, and Edward C. Holmes. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science*, 303(5656):327–332, January 2004. ISSN 1095-9203. doi: 10.1126/science.1090727. URL http://dx.doi.org/10.1126/science.1090727.

[108] R. C. Griffiths and S. Tavaré. Sampling theory for neutral alleles in a varying environment. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 344(1310):403–410, June 1994. ISSN 0962-8436. doi: 10.1098/rstb.1994.0079. URL http://dx.doi.org/10.1098/rstb.1994.0079.

[109] R. C. Griffiths and Simon Tavare. Ancestral Inference in Population Genetics. *Statistical Science*, 9(3):307–319, August 1994. ISSN 0883-4237. doi: 10.1214/ss/1177010378. URL http://dx.doi.org/10.1214/ss/1177010378.

[110] Stéphane Guindon, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3):307–321, March 2010. ISSN 1076-836X. doi: 10.1093/sysbio/syq010. URL http://dx.doi.org/10.1093/sysbio/syq010.

[111] L. G. Gürtler, P. H. Hauser, J. Eberle, A. von Brunn, S. Knapp, L. Zekeng, J. M. Tsague, and L. Kaptue. A new subtype of human immunodeficiency virus type 1 (MVP-5180) from Cameroon. *Journal of virology*, 68(3):1581–

1585, March 1994. ISSN 0022-538X. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC236615/.

[112] Joris Hemelaar, Eleanor Gouws, Peter D. Ghys, Saladin Osmanov, and WHO-UNAIDS Network for HIV Isolation and Characterisation. Global trends in molecular epidemiology of HIV-1 during 2000-2007. *AIDS (London, England)*, 25(5):679–689, March 2011. ISSN 1473-5571. doi: 10.1097/qad.0b013e328342ff93. URL http://dx.doi.org/10.1097/qad.0b013e328342ff93.

[113] Joshua T. Herbeck, Geoffrey S. Gottlieb, Xiuhong Li, Zheng Hu, Roger Detels, John Phair, Charles Rinaldo, Lisa P. Jacobson, Joseph B. Margolick, and James I. Mullins. Lack of Evidence for Changing Virulence of HIV-1 in North America. *PLoS ONE*, 3(2):e1525+, February 2008. doi: 10.1371/journal.pone.0001525. URL http://dx.doi.org/10.1371/journal.pone.0001525.

[114] Trevor Hinkley, Joao Martins, Colombe Chappey, Mojgan Haddad, Eric Stawiski, Jeannette M. Whitcomb, Christos J. Petropoulos, and Sebastian Bonhoeffer. A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nat Genet*, 43(5):487–489, May 2011. ISSN 1061-4036. doi: 10.1038/ng.795. URL http://dx.doi.org/10.1038/ng.795.

[115] Martin S. Hirsch, Huldrych F. Günthard, Jonathan M. Schapiro, Françoise B. Vézinet, Bonaventura Clotet, Scott M. Hammer, Victoria A. Johnson, Daniel R. Kuritzkes, John W. Mellors, Deenan Pillay, Patrick G. Yeni, Donna M. Jacobsen, and Douglas D. Richman. Antiretroviral Drug Resistance Testing in Adult HIV-1 Infection: 2008 Recommendations of an International AIDS Society-USA Panel. *Clinical Infectious Diseases*, 47 (2):266–285, July 2008. doi: 10.1086/589297. URL http://dx.doi.org/10.1086/589297.

[116] J. Hiscott, H. Kwon, and P. Génin. Hostile takeovers: viral appropriation of the NF-kappaB pathway. *The Journal of clinical investigation*, 107(2):143–151, January 2001. ISSN 0021-9738. doi: 10.1172/jci11918. URL http://dx.doi.org/10.1172/jci11918.

[117] David D. Ho, Avidan U. Neumann, Alan S. Perelson, Wen Chen, John M. Leonard, and Martin Markowitz. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature*, 373(6510):123–126, January 1995. ISSN 0028-0836. doi: 10.1038/373123a0. URL http://dx.doi.org/10.1038/373123a0.

[118] Eddie C. Holmes, Sean Nee, Andrew Rambaut, Geoff P. Garnett, and Paul H. Harvey. Revealing the History of Infectious Disease Epidemics through Phylogenetic Trees. *Philosophical Transactions: Biological Sciences*, 349(1327), 1995. ISSN 09628436. doi: 10.2307/56121. URL http://dx.doi.org/10.2307/56121.

[119] Edward C. Holmes. Evolutionary History and Phylogeography of Human Viruses. *Annual Review of Microbiology*, 62(1):307–328, 2008. ISSN 0066-4227. doi: 10.1146/annurev.micro.62.081307.162912. URL http://dx.doi.org/10.1146/annurev.micro.62.081307.162912.

[120] Edward C. Holmes. *The Evolution and Emergence of RNA Viruses (Oxford Series in Ecology and Evolution)*. Oxford University Press, USA, 1 edition, August 2009. ISBN 0199211132. URL http://www.worldcat.org/isbn/0199211132.

[121] Ilka Hoof, Carina L. Pérez, Marcus Buggert, Rasmus K. Gustafsson, Morten Nielsen, Ole Lund, and Annika C. Karlsson. Interdisciplinary analysis of HIV-specific CD8+ T cell responses against variant epitopes reveals restricted TCR promiscuity. *Journal of immunology (Baltimore, Md. : 1950)*, 184 (9):5383–5391, May 2010. ISSN 1550-6606. URL http://view.ncbi.nlm.nih.gov/pubmed/20363973.

[122] Kuan-Hsiang G. Huang, Dominique Goedhals, Jonathan M. Carlson, Mark A. Brockman, Swati Mishra, Zabrina L. Brumme, Stephen Hickling, Christopher S. W. Tang, Toshiyuki Miura, Chris Seebregts, David Heckerman, Thumbi Ndung'u, Bruce Walker, Paul Klenerman, Dewald Steyn, Philip Goulder, Rodney Phillips, Cloete van Vuuren, John Frater, and Bloemfontein-Oxford Collaborative Group. Progression to AIDS in South Africa Is Associated with both Reverting and Compensatory Viral

Mutations. *PLoS ONE*, 6(4):e19018+, April 2011. doi: 10.1371/journal.pone.0019018. URL http://dx.doi.org/10.1371/journal.pone.0019018.

[123] Kuan-Hsiang Gary H. Huang, Dominique Goedhals, Helen Fryer, Cloete van Vuuren, Aris Katzourakis, Tulio De Oliveira, Helen Brown, Sharon Cassol, Chris Seebregts, Angela McLean, Paul Klenerman, Rodney Phillips, John Frater, and Bloemfontein-Oxford Collaborative Group. Prevalence of HIV type-1 drug-associated mutations in pre-therapy patients in the Free State, South Africa. *Antiviral therapy*, 14(7):975–984, 2009. ISSN 1359-6535. doi: 10.3851/imp1416. URL http://dx.doi.org/10.3851/imp1416.

[124] R. R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical population biology*, 23(2):183–201, April 1983. ISSN 0040-5809. URL http://view.ncbi.nlm.nih.gov/pubmed/6612631.

[125] Gareth J. Hughes, Esther Fearnhill, David Dunn, Samantha J. Lycett, Andrew Rambaut, Andrew J. Leigh Brown, and on behalf of the UK HIV Drug Resistance Collaboration. Molecular Phylodynamics of the Heterosexual HIV Epidemic in the United Kingdom. *PLoS Pathog*, 5(9):e1000590+, September 2009. doi: 10.1371/journal.ppat.1000590. URL http://dx.doi.org/10.1371/journal.ppat.1000590.

[126] Janis F. Hutchinson. The Biology and Evolution of HIV. *Annual Review of Anthropology*, 30, 2001. ISSN 00846570. doi: 10.2307/3069210. URL http://dx.doi.org/10.2307/3069210.

[127] Los Alamos HIV immunology database. http://www.hiv.lanl.gov/content/immunology.

[128] Los Alamos HIV immunology database: CTL associated variants. http://www.hiv.lanl.gov/content/immunology/variants/ctl_variant.html.

[129] Astrid K. Iversen, Guillaume Stewart-Jones, Gerald H. Learn, Natasha Christie, Christina Sylvester-Hviid, Andrew E. Armitage, Rupert Kaul, Tara Beattie, Jean K. Lee, Yanping Li, Pojchong Chotiyarnwong, Tao Dong, Xiaoning Xu, Mark A. Luscher, Kelly MacDonald, Henrik Ullum, Bente Klarlund-Pedersen, Peter Skinhøj, Lars Fugger, Søren Buus,

James I. Mullins, E. Yvonne Jones, P. Anton van der Merwe, and Andrew J. McMichael. Conflicting selective forces affect T cell receptor contacts in an immunodominant human immunodeficiency virus epitope. *Nature immunology*, 7(2):179–189, February 2006. ISSN 1529-2908. URL http://view.ncbi.nlm.nih.gov/pubmed/16388312.

[130] Gareth M. Jenkins, Andrew Rambaut, Oliver G. Pybus, and Edward C. Holmes. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *Journal of molecular evolution*, 54(2):156–165, February 2002. ISSN 1432-1432. doi: 10.1007/s00239-001-0064-3. URL http://dx.doi.org/10.1007/s00239-001-0064-3.

[131] Nuria Jiménez-Hernández, Manuela Torres-Puente, Maria A. Bracho, Inmaculada García-Robles, Enrique Ortega, Juan del Olmo, Fernando Carnicer, Fernando González-Candelas, and Andrés Moya. Epidemic dynamics of two coexisting hepatitis C virus subtypes. *Journal of General Virology*, 88(1):123–133, January 2007. doi: 10.1099/vir.0.82277-0. URL http://dx.doi.org/10.1099/vir.0.82277-0.

[132] Welkin E. Johnson and Ronald C. Desrosiers. VIRAL PERSISTENCE: HIV's Strategies of Immune System Evasion. *Annual Review of Medicine*, 53(1): 499–518, 2002. doi: 10.1146/annurev.med.53.082901.104053. URL http://dx.doi.org/10.1146/annurev.med.53.082901.104053.

[133] T. Jombart, R. M. Eggo, P. J. Dodd, and F. Balloux. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*, 106(2):383–390, February 2011. ISSN 1365-2540. doi: 10.1038/hdy.2010.78. URL http://dx.doi.org/10.1038/hdy.2010.78.

[134] Annika C. Karlsson, Steven G. Deeks, Jason D. Barbour, Brandon D. Heiken, Sophie R. Younger, Rebecca Hoh, Meghan Lane, Matti Sällberg, Gabriel M. Ortiz, James F. Demarest, Teri Liegler, Robert M. Grant, Jeffrey N. Martin, and Douglas F. Nixon. Dual pressure from antiretroviral therapy and cell-mediated immune response on the human immunodeficiency virus type 1 protease gene. *Journal of virology*, 77(12):6743–6752,

June 2003. ISSN 0022-538X. URL http://view.ncbi.nlm.nih.gov/pubmed/12767994.

[135] R. A. Kaslow, M. Carrington, R. Apple, L. Park, A. Munoz, A. J. Saah, J. J. Goedert, C. Winkler, S. J. O'Brien, C. Rinaldo, R. Detels, W. Blattner, J. Phair, H. Erlich, and D. L. Mann. Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection. *Nature Medicine*, 2(4):405–411, April 1996. doi: 10.1038/nm0496-405. URL http://dx.doi.org/10.1038/nm0496-405.

[136] Richard A. Kaslow, Tevfik Dorak, and James J. Tang. Influence of Host Genetic Variation on Susceptibility to HIV Type 1 Infection. *Journal of Infectious Diseases*, 191(Supplement 1):S68–S77, February 2005. doi: 10.1086/425269. URL http://dx.doi.org/10.1086/425269.

[137] A. Kawana, H. Tomiyama, M. Takiguchi, T. Shioda, T. Nakamura, and A. Iwamoto. Accumulation of specific amino acid substitutions in HLA-B35-restricted human immunodeficiency virus type 1 cytotoxic T lymphocyte epitopes. *AIDS research and human retroviruses*, 15(12):1099–1107, August 1999. ISSN 0889-2229. URL http://view.ncbi.nlm.nih.gov/pubmed/10461830.

[138] Yuka Kawashima, Katja Pfafferott, John Frater, Philippa Matthews, Rebecca Payne, Marylyn Addo, Hiroyuki Gatanaga, Mamoru Fujiwara, Atsuko Hachiya, Hirokazu Koizumi, Nozomi Kuse, Shinichi Oka, Anna Duda, Andrew Prendergast, Hayley Crawford, Alasdair Leslie, Zabrina Brumme, Chanson Brumme, Todd Allen, Christian Brander, Richard Kaslow, James Tang, Eric Hunter, Susan Allen, Joseph Mulenga, Songee Branch, Tim Roach, Mina John, Simon Mallal, Anthony Ogwu, Roger Shapiro, Julia G. Prado, Sarah Fidler, Jonathan Weber, Oliver G. Pybus, Paul Klenerman, Thumbi Ndung'u, Rodney Phillips, David Heckerman, P. Richard Harrigan, Bruce D. Walker, Masafumi Takiguchi, and Philip Goulder. Adaptation of HIV-1 to human leukocyte antigen class I. *Nature*, 458(7238):641–645, April 2009. ISSN 1476-4687. doi: 10.1038/nature07746. URL http://dx.doi.org/10.1038/nature07746.

[139] Brandon F. Keele, Fran Van Heuverswyn, Yingying Li, Elizabeth Bailes, Jun Takehisa, Mario L. Santiago, Frederic Bibollet-Ruche, Yalu Chen, Louise V. Wain, Florian Liegeois, Severin Loul, Eitel M. Ngole, Yanga Bienvenue, Eric Delaporte, John F. Y. Brookfield, Paul M. Sharp, George M. Shaw, Martine Peeters, and Beatrice H. Hahn. Chimpanzee Reservoirs of Pandemic and Nonpandemic HIV-1. *Science*, 313(5786):523–526, July 2006. doi: 10.1126/science.1126531. URL http://dx.doi.org/10.1126/science.1126531.

[140] Brandon F. Keele, Elena E. Giorgi, Jesus F. Salazar-Gonzalez, Julie M. Decker, Kimmy T. Pham, Maria G. Salazar, Chuanxi Sun, Truman Grayson, Shuyi Wang, Hui Li, Xiping Wei, Chunlai Jiang, Jennifer L. Kirchherr, Feng Gao, Jeffery A. Anderson, Li-Hua H. Ping, Ronald Swanstrom, Georgia D. Tomaras, William A. Blattner, Paul A. Goepfert, J. Michael Kilby, Michael S. Saag, Eric L. Delwart, Michael P. Busch, Myron S. Cohen, David C. Montefiori, Barton F. Haynes, Brian Gaschen, Gayathri S. Athreya, Ha Y. Lee, Natasha Wood, Cathal Seoighe, Alan S. Perelson, Tanmoy Bhattacharya, Bette T. Korber, Beatrice H. Hahn, and George M. Shaw. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proceedings of the National Academy of Sciences of the United States of America*, 105(21):7552–7557, May 2008. ISSN 1091-6490. doi: 10.1073/pnas.0802203105. URL http://dx.doi.org/10.1073/pnas.0802203105.

[141] A. D. Kelleher, C. Long, E. C. Holmes, R. L. Allen, J. Wilson, C. Conlon, C. Workman, S. Shaunak, K. Olson, P. Goulder, C. Brander, G. Ogg, J. S. Sullivan, W. Dyer, I. Jones, A. J. McMichael, S. Rowland-Jones, and R. E. Phillips. Clustered mutations in HIV-1 gag are consistently required for escape from HLA-B27-restricted cytotoxic T lymphocyte responses. *The Journal of experimental medicine*, 193(3):375–386, February 2001. ISSN 0022-1007. doi: 10.1084/jem.193.3.375. URL http://dx.doi.org/10.1084/jem.193.3.375.

[142] David G. Kendall. On the Generalized "Birth-and-Death" Process. *The Annals of Mathematical Statistics*, 19(1):1–15, March

1948. URL http://www.jstor.org/action/exportSingleCitation?singleCitation=true&suffix=10.2307%2F2236051.

[143] Amanda Kenneally, Bin Liang, and Linda D. Barber. The peptide-binding motif of HLA-B*3505. 51(10):866–868, 2000. doi: 10.1007/s002510000220. URL http://dx.doi.org/10.1007/s002510000220.

[144] Stephen J. Kent, Caroline S. Fernandez, C. Jane Dale, and Miles P. Davenport. Reversion of immune escape HIV variants upon transmission: insights into effective viral immunity. *Trends in Microbiology*, 13(6):243–246, June 2005. ISSN 0966842X. doi: 10.1016/j.tim.2005.03.011. URL http://dx.doi.org/10.1016/j.tim.2005.03.011.

[145] Photini Kiepiela, Alasdair J. Leslie, Isobella Honeyborne, Danni Ramduth, Christina Thobakgale, Senica Chetty, Prinisha Rathnavalu, Corey Moore, Katja J. Pfafferott, Louise Hilton, Peter Zimbwa, Sarah Moore, Todd Allen, Christian Brander, Marylyn M. Addo, Marcus Altfeld, Ian James, Simon Mallal, Michael Bunce, Linda D. Barber, James Szinger, Cheryl Day, Paul Klenerman, James Mullins, Bette Korber, Hoosen M. Coovadia, Bruce D. Walker, and Philip J. Goulder. Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature*, 432(7018):769–775, December 2004. ISSN 1476-4687. doi: 10.1038/nature03113. URL http://dx.doi.org/10.1038/nature03113.

[146] Photini Kiepiela, Kholiswa Ngumbela, Christina Thobakgale, Dhanwanthie Ramduth, Isobella Honeyborne, Eshia Moodley, Shabashini Reddy, Chantal de Pierres, Zenele Mncube, Nompumelelo Mkhwanazi, Karen Bishop, Mary van der Stok, Kriebashnie Nair, Nasreen Khan, Hayley Crawford, Rebecca Payne, Alasdair Leslie, Julia Prado, Andrew Prendergast, John Frater, Noel McCarthy, Christian Brander, Gerald H. Learn, David Nickle, Christine Rousseau, Hoosen Coovadia, James I. Mullins, David Heckerman, Bruce D. Walker, and Philip Goulder. CD8+ T-cell responses to different HIV proteins have discordant associations with viral load. *Nature medicine*, 13(1):46–53, January 2007. ISSN 1078-8956. doi: 10.1038/nm1520. URL http://dx.doi.org/10.1038/nm1520.

[147] Gustavo H. Kijak, Viviana Simon, Peter Balfe, Jeroen Vanderhoeven, Sandra E. Pampuro, Carlos Zala, Claudia Ochoa, Pedro Cahn, Martin Markowitz, and Horacio Salomon. Origin of Human Immunodeficiency Virus Type 1 Quasispecies Emerging after Antiretroviral Treatment Interruption in Patients with Therapeutic Failure. *J. Virol.*, 76(14):7000–7009, July 2002. doi: 10.1128/jvi.76.14.7000-7009.2002. URL http://dx.doi.org/10.1128/jvi.76.14.7000-7009.2002.

[148] J. F. C. Kingman. Exchangeability and the Evolution of Large Populations. In G. Koch and F. Spizzichino, editors, *Exchangeability in Probability and Statistics*, pages 97–112. North-Holland, Amsterdam, 1982.

[149] J. F. C. Kingman. On the Genealogy of Large Populations. *Journal of Applied Probability*, 19:27–43, 1982. ISSN 00219002. doi: 10.2307/3213548. URL http://dx.doi.org/10.2307/3213548.

[150] Ireen E. Kiwelu, Vladimir Novitsky, Lauren Margolin, Jeannie Baca, Rachel Manongi, Noel Sam, John Shao, Mary F. McLane, Saidi H. Kapiga, and M. Essex. Frequent intra-subtype recombination among HIV-1 circulating in Tanzania. *PloS one*, 8(8), 2013. ISSN 1932-6203. URL http://view.ncbi.nlm.nih.gov/pubmed/23940702.

[151] J. Kjems, M. Brown, D. D. Chang, and P. A. Sharp. Structural analysis of the interaction between the human immunodeficiency virus Rev protein and the Rev response element. *Proceedings of the National Academy of Sciences of the United States of America*, 88(3):683–687, February 1991. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC50877/.

[152] Florian Klein, Hugo Mouquet, Pia Dosenovic, Johannes F. Scheid, Louise Scharf, and Michel C. Nussenzweig. Antibodies in HIV-1 Vaccine Development and Therapy. *Science*, 341(6151):1199–1204, September 2013. ISSN 1095-9203. doi: 10.1126/science.1241144. URL http://dx.doi.org/10.1126/science.1241144.

[153] P. Klenerman, S. Rowland-Jones, S. McAdam, J. Edwards, S. Daenke, D. Lalloo, B. Köppe, W. Rosenberg, D. Boyd, and A. Edwards. Cytotoxic

T-cell activity antagonized by naturally occurring HIV-1 Gag variants. *Nature*, 369(6479):403–407, June 1994. ISSN 0028-0836. doi: 10.1038/369403a0. URL http://dx.doi.org/10.1038/369403a0.

[154] P. Klenerman, U. C. Meier, R. E. Phillips, and A. J. McMichael. The effects of natural altered peptide ligands on the whole blood cytotoxic T lymphocyte response to human immunodeficiency virus. *European journal of immunology*, 25(7):1927–1931, July 1995. ISSN 0014-2980. doi: 10.1002/eji.1830250720. URL http://dx.doi.org/10.1002/eji.1830250720.

[155] Katia Koelle and David A. Rasmussen. Rates of coalescence for common epidemiological models at equilibrium. *Journal of The Royal Society Interface*, 9(70):997–1007, September 2011. ISSN 1742-5662. doi: 10.1098/rsif.2011.0495. URL http://dx.doi.org/10.1098/rsif.2011.0495.

[156] Tomohiko Koibuchi, Todd M. Allen, Mathias Lichterfeld, Stanley K. Mui, Kristin M. O'Sullivan, Alicja Trocha, Spyros A. Kalams, R. Paul Johnson, and Bruce D. Walker. Limited sequence evolution within persistently targeted CD8 epitopes in chronic human immunodeficiency virus type 1 infection. *Journal of virology*, 79(13):8171–8181, July 2005. ISSN 0022-538X. URL http://view.ncbi.nlm.nih.gov/pubmed/15956562.

[157] Aleksandr Lazaryan, Elena Lobashevsky, Joseph Mulenga, Etienne Karita, Susan Allen, Jianming Tang, and Richard A. Kaslow. Human leukocyte antigen B58 supertype and human immunodeficiency virus type 1 infection in native Africans. *Journal of virology*, 80(12):6056–6060, June 2006. ISSN 0022-538X. doi: 10.1128/jvi.02119-05. URL http://dx.doi.org/10.1128/jvi.02119-05.

[158] Andrew J. Leigh Brown, Samantha J. Lycett, Lucy Weinert, Gareth J. Hughes, Esther Fearnhill, and David T. Dunn. Transmission Network Parameters Estimated From HIV Sequences for a Nationwide Epidemic. *Journal of Infectious Diseases*, 204(9):jir550–1469, September 2011. ISSN 1537-6613. doi: 10.1093/infdis/jir550. URL http://dx.doi.org/10.1093/infdis/jir550.

[159] T. Leitner and J. Albert. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proceedings of the National Academy of Sciences of the United States of America*, 96(19):10752–10757, September 1999. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC17955/.

[160] Philippe Lemey, Oliver G. Pybus, Andrew Rambaut, Alexei J. Drummond, David L. Robertson, Pierre Roques, Michael Worobey, and Anne-Mieke Vandamme. The Molecular Population Genetics of HIV-1 Group O. *Genetics*, 167(3):1059–1068, July 2004. ISSN 0016-6731. doi: 10.1534/genetics.104.026666. URL http://dx.doi.org/10.1534/genetics.104.026666.

[161] Philippe Lemey, Andrew Rambaut, and Oliver G. Pybus. HIV evolutionary dynamics within and among hosts. *AIDS reviews*, 8(3):125–140, 2006. ISSN 1139-6121. URL http://view.ncbi.nlm.nih.gov/pubmed/17078483.

[162] Philippe Lemey, Andrew Rambaut, Alexei J. Drummond, and Marc A. Suchard. Bayesian Phylogeography Finds Its Roots. *PLoS Comput Biol*, 5 (9):e1000520+, September 2009. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000520. URL http://dx.doi.org/10.1371/journal.pcbi.1000520.

[163] A. J. Leslie, K. J. Pfafferott, P. Chetty, R. Draenert, M. M. Addo, M. Feeney, Y. Tang, E. C. Holmes, T. Allen, J. G. Prado, M. Altfeld, C. Brander, C. Dixon, D. Ramduth, P. Jeena, S. A. Thomas, A. St John, T. A. Roach, B. Kupfer, G. Luzzi, A. Edwards, G. Taylor, H. Lyall, G. Tudor-Williams, V. Novelli, J. Martinez-Picado, P. Kiepiela, B. D. Walker, and P. J. R. Goulder. HIV evolution: CTL escape mutation and reversion after transmission. *Nature Medicine*, 10(3):282–289, February 2004. ISSN 1078-8956. doi: 10.1038/nm992. URL http://dx.doi.org/10.1038/nm992.

[164] Alasdair Leslie, Philippa C. Matthews, Jennifer Listgarten, Jonathan M. Carlson, Carl Kadie, Thumbi Ndung'u, Christian Brander, Hoosen Coovadia, Bruce D. Walker, David Heckerman, and Philip J. R. Goulder. Additive Contribution of HLA Class I Alleles in the Immune Control of HIV-1 Infection. *Journal of Virology*, 84(19):9879–9888, October 2010. ISSN 1098-

5514. doi: 10.1128/jvi.00320-10. URL http://dx.doi.org/10.1128/jvi.00320-10.

[165] Fraser Lewis, Gareth J. Hughes, Andrew Rambaut, Anton Pozniak, and Andrew J. Leigh Brown. Episodic Sexual Transmission of HIV Revealed by Molecular Phylodynamics. *PLoS Med*, 5(3):e50+, March 2008. ISSN 1549-1277. doi: 10.1371/journal.pmed.0050050. URL http://dx.doi.org/10.1371/journal.pmed.0050050.

[166] Bin Li, Adrianne D. Gladden, Marcus Altfeld, John M. Kaldor, David A. Cooper, Anthony D. Kelleher, and Todd M. Allen. Rapid Reversion of Sequence Polymorphisms Dominates Early Human Immunodeficiency Virus Type 1 Evolution. *J. Virol.*, 81(1):193–201, January 2007. doi: 10.1128/jvi.01231-06. URL http://dx.doi.org/10.1128/jvi.01231-06.

[167] Na Li and Matthew Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, December 2003. ISSN 0016-6731. URL http://www.genetics.org/content/165/4/2213.abstract.

[168] Sophie Limou, Sigrid Le Clerc, Cédric Coulonges, Wassila Carpentier, Christian Dina, Olivier Delaneau, Taoufik Labib, Lieng Taing, Rob Sladek, Christiane Deveau, Rojo Ratsimandresy, Matthieu Montes, Jean-Louis L. Spadoni, Jean-Daniel D. Lelièvre, Yves Lévy, Amu Therwath, François Schächter, Fumihiko Matsuda, Ivo Gut, Philippe Froguel, Jean-François F. Delfraissy, Serge Hercberg, Jean-François F. Zagury, and ANRS Genomic Group. Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). *The Journal of infectious diseases*, 199(3):419–426, February 2009. ISSN 0022-1899. doi: 10.1086/596067. URL http://dx.doi.org/10.1086/596067.

[169] Michael K. Liu, Natalie Hawkins, Adam J. Ritchie, Vitaly V. Ganusov, Victoria Whale, Simon Brackenridge, Hui Li, Jeffrey W. Pavlicek, Fangping Cai, Melissa Rose-Abrahams, Florette Treurnicht, Peter Hraber, Catherine Riou,

Clive Gray, Guido Ferrari, Rachel Tanner, Li-Hua H. Ping, Jeffrey A. Anderson, Ronald Swanstrom, CHAVI Core B, Myron Cohen, Salim Abdool S. Karim, Barton Haynes, Persephone Borrow, Alan S. Perelson, George M. Shaw, Beatrice H. Hahn, Carolyn Williamson, Bette T. Korber, Feng Gao, Steve Self, Andrew McMichael, and Nilu Goonetilleke. Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. *The Journal of clinical investigation*, 123(1):380–393, January 2013. ISSN 1558-8238. doi: 10.1172/jci65330. URL http://dx.doi.org/10.1172/jci65330.

[170] Tommy F. Liu and Robert W. Shafer. Web Resources for HIV Type 1 Genotypic-Resistance Test Interpretation. *Clinical Infectious Diseases*, 42 (11):1608–1618, June 2006. ISSN 1537-6591. doi: 10.1086/503914. URL http://dx.doi.org/10.1086/503914.

[171] Yi Liu, David C. Nickle, Daniel Shriner, Mark A. Jensen, Gerald H. Learn, John E. Mittler, and James I. Mullins. Molecular clock-like evolution of human immunodeficiency virus type 1. *Virology*, 329(1):101–108, November 2004. ISSN 00426822. doi: 10.1016/j.virol.2004.08.014. URL http://dx.doi.org/10.1016/j.virol.2004.08.014.

[172] Yi Liu, John McNevin, Jianhong Cao, Hong Zhao, Indira Genowati, Kim Wong, Sherry McLaughlin, Matthew D. McSweyn, Kurt Diem, Claire E. Stevens, Janine Maenza, Hongxia He, David C. Nickle, Daniel Shriner, Sarah E. Holte, Ann C. Collier, Lawrence Corey, M. Juliana McElrath, and James I. Mullins. Selection on the human immunodeficiency virus type 1 proteome following primary infection. *Journal of virology*, 80(19): 9519–9529, October 2006. ISSN 0022-538X. doi: 10.1128/jvi.00575. URL http://dx.doi.org/10.1128/jvi.00575.

[173] Anuska Llano, Ashley Williams, Alex Olvera, Sandra Silva-Arrieta, and Christian Brander. Best-characterized hiv-1 ctl epitopes: The 2013 update. http://www.hiv.lanl.gov/content/immunology/pdf/2013/optimal_ctl_article.pdf.

[174] Anuska Llano, Ashley Williams, Alex Olvera, Sandra Silva-Arrieta, and Christian Brander. Best-Characterized HIV-1 CTL Epitopes: The 2013 Up-

date. Technical report, 2013. URL http://www.hiv.lanl.gov/content/immunology/pdf/2013/optimal_ctl_article.pdf.

[175] Franco Lori, Mark G. Lewis, Jianqing Xu, Georg Varga, Donald E. Zinn, Carrol Crabbs, Wendeline Wagner, Jack Greenhouse, Peter Silvera, Jake Yalley-Ogunro, Carmine Tinelli, and Julianna Lisziewicz. Control of SIV Rebound Through Structured Treatment Interruptions During Early Infection. *Science*, 290(5496):1591–1593, November 2000. doi: 10.1126/science.290.5496.1591. URL http://dx.doi.org/10.1126/science.290.5496.1591.

[176] P. C. Mahalanobis. On the generalised distance in statistics. In *Proceedings National Institute of Science, India*, volume 2, pages 49–55, April 1936. URL http://ir.isical.ac.in/dspace/handle/1/1268.

[177] Mariana Marin, Kristine M. Rose, Susan L. Kozak, and David Kabat. HIV-1 Vif protein binds the editing enzyme APOBEC3G and induces its degradation. *Nature Medicine*, 9(11):1398–1403, October 2003. doi: 10.1038/nm946. URL http://dx.doi.org/10.1038/nm946.

[178] Steven G. E. Marsh, Peter Parham, and Linda D. Barber. *The HLA FactsBook*. Academic Press, 1 edition, January 2000. ISBN 0125450257. URL http://www.worldcat.org/isbn/0125450257.

[179] Alexey E. Masharsky, Elena N. Dukhovlinova, Sergei V. Verevochkin, Olga V. Toussova, Roman V. Skochilov, Jeffrey A. Anderson, Irving Hoffman, Myron S. Cohen, Ronald Swanstrom, and Andrei P. Kozlov. A Substantial Transmission Bottleneck among Newly and Recently HIV-1-Infected Injection Drug Users in St Petersburg, Russia. *Journal of Infectious Diseases*, 201(11):1697–1702, June 2010. ISSN 1537-6613. doi: 10.1086/652702. URL http://dx.doi.org/10.1086/652702.

[180] Philippa C. Matthews, Andrew Prendergast, Alasdair Leslie, Hayley Crawford, Rebecca Payne, Christine Rousseau, Morgane Rolland, Isobella Honeyborne, Jonathan Carlson, Carl Kadie, Christian Brander, Karen Bishop, Nonkululeko Mlotshwa, James I. Mullins, Hoosen Coovadia, Thumbi

Ndung'u, Bruce D. Walker, David Heckerman, and Philip J. R. Goulder. Central Role of Reverting Mutations in HLA Associations with Human Immunodeficiency Virus Set Point. *Journal of Virology*, 82(17):8548–8559, September 2008. ISSN 1098-5514. doi: 10.1128/jvi.00580-08. URL http://dx.doi.org/10.1128/jvi.00580-08.

[181] Philippa C. Matthews, Emily Adland, Jennifer Listgarten, Alasdair Leslie, Nompumelelo Mkhwanazi, Jonathan M. Carlson, Mikkel Harndahl, Anette Stryhn, Rebecca P. Payne, Anthony Ogwu, Kuan-Hsiang Gary H. Huang, John Frater, Paolo Paioni, Henrik Kloverpris, Pieter Jooste, Dominique Goedhals, Cloete van Vuuren, Dewald Steyn, Lynn Riddell, Fabian Chen, Graz Luzzi, Thambiah Balachandran, Thumbi Ndung'u, Søren Buus, Mary Carrington, Roger Shapiro, David Heckerman, and Philip J. Goulder. HLA-A*7401-mediated control of HIV viremia is independent of its linkage disequilibrium with HLA-B*5703. *Journal of immunology (Baltimore, Md. : 1950)*, 186(10):5675–5686, May 2011. ISSN 1550-6606. doi: 10.4049/jimmunol.1003711. URL http://dx.doi.org/10.4049/jimmunol.1003711.

[182] Andrew J. McMichael, Persephone Borrow, Georgia D. Tomaras, Nilu Goonetilleke, and Barton F. Haynes. The immune response during acute HIV-1 infection: clues for vaccine development. *Nature reviews. Immunology*, 10(1):11–23, January 2010. ISSN 1474-1741. doi: 10.1038/nri2674. URL http://dx.doi.org/10.1038/nri2674.

[183] Ute-Christiane Meier, Paul Klenerman, Philip Griffin, William James, Britta Köppe, Brendan Larder, Andrew McMichael, and Rodney Phillips. Cytotoxic T Lymphocyte Lysis Inhibited by Viable HIV Mutants. *Science*, 270 (5240):1360–1362, November 1995. doi: 10.1126/science.270.5240.1360. URL http://dx.doi.org/10.1126/science.270.5240.1360.

[184] J. W. Mellors, A. Muñoz, J. V. Giorgi, J. B. Margolick, C. J. Tassoni, P. Gupta, L. A. Kingsley, J. A. Todd, A. J. Saah, R. Detels, J. P. Phair, and C. R. Rinaldo. Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. *Annals of internal medicine*, 126(12):946–954, June 1997. ISSN 0003-4819. URL http://view.ncbi.nlm.nih.gov/pubmed/9182471.

[185] L. Menéndez-Arias, A. Mas, and E. Domingo. Cytotoxic T-lymphocyte responses to HIV-1 reverse transcriptase (review). *Viral immunology*, 11 (4):167–181, 1998. ISSN 0882-8245. URL http://view.ncbi.nlm.nih.gov/pubmed/10189185.

[186] Stephen A. Migueles, M. Shirin Sabbaghian, W. Lesley Shupert, Maria P. Bettinotti, Francesco M. Marincola, Lisa Martino, Clair W. Hallahan, Sara M. Selig, David Schwartz, John Sullivan, and Mark Connors. HLA B*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors. *Proceedings of the National Academy of Sciences*, 97(6):2709–2714, March 2000. ISSN 1091-6490. doi: 10.1073/pnas.050567397. URL http://dx.doi.org/10.1073/pnas.050567397.

[187] Iliyana Mikell, D. Noah Sather, Spyros A. Kalams, Marcus Altfeld, Galit Alter, and Leonidas Stamatatos. Characteristics of the earliest cross-neutralizing antibody response to HIV-1. *PLoS pathogens*, 7(1), 2011. ISSN 1553-7374. URL http://view.ncbi.nlm.nih.gov/pubmed/21249232.

[188] Vladimir N. Minin, Erik W. Bloomquist, and Marc A. Suchard. Smooth Skyride through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics. *Molecular Biology and Evolution*, 25(7):1459–1471, July 2008. ISSN 1537-1719. doi: 10.1093/molbev/msn090. URL http://dx.doi.org/10.1093/molbev/msn090.

[189] Corey B. Moore, Mina John, Ian R. James, Frank T. Christiansen, Campbell S. Witt, and Simon A. Mallal. Evidence of HIV-1 Adaptation to HLA-Restricted Immune Responses at a Population Level. *Science*, 296 (5572):1439–1443, May 2002. doi: 10.1126/science.1069660. URL http://dx.doi.org/10.1126/science.1069660.

[190] Penny L. Moore, Elin S. Gray, Isaac A. Choge, Nthabeleng Ranchobe, Koleka Mlisana, Salim S. Abdool Karim, Carolyn Williamson, Lynn Morris, and CAPRISA 002 Study Team. The c3-v4 region is a major target of autologous neutralizing antibodies in human immunodeficiency virus type 1 subtype C infection. *Journal of virology*, 82(4):1860–1869, February 2008.

ISSN 1098-5514. doi: 10.1128/jvi.02187-07. URL http://dx.doi.org/10.1128/jvi.02187-07.

[191] Dilys Morgan, Cedric Mahe, Billy Mayanja, J. Martin Okongo, Rosemary Lubega, and James A. Whitworth. HIV-1 infection in rural Africa: is there a difference in median time to AIDS and survival compared with that in industrialized countries? *AIDS (London, England)*, 16(4):597–603, March 2002. ISSN 0269-9370. URL http://view.ncbi.nlm.nih.gov/pubmed/11873003.

[192] Sandra M. Mueller, Birgit Schaetz, Kathrin Eismann, Silke Bergmann, Michael Bauerle, Matthias Schmitt-Haendle, Hauke Walter, Barbara Schmidt, Klaus Korn, Heinrich Sticht, Bernd Spriewald, Ellen G. Harrer, and Thomas Harrer. Dual selection pressure by drugs and HLA class I-restricted immune responses on human immunodeficiency virus type 1 protease. *Journal of virology*, 81(6):2887–2898, March 2007. ISSN 0022-538X. URL http://view.ncbi.nlm.nih.gov/pubmed/17202219.

[193] Dennis J. Murphy and James R. Brown. Computational biology in anti-tuberculosis drug discovery. *Infectious disorders drug targets*, 9(3):319–326, June 2009. ISSN 1871-5265. URL http://view.ncbi.nlm.nih.gov/pubmed/19519485.

[194] Gary J. Nabel. Mapping the future of HIV vaccines. *Nat Rev Micro*, 5 (7):482–484, July 2007. ISSN 1740-1526. doi: 10.1038/nrmicro1713. URL http://dx.doi.org/10.1038/nrmicro1713.

[195] John C. Nash. *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation*. CRC Press, 2 edition, January 1990. ISBN 085274319X. URL http://www.worldcat.org/isbn/085274319X.

[196] Sean Nee, Robert M. May, and Paul H. Harvey. The Reconstructed Evolutionary Process. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1309):305–311, May 1994. ISSN 1471-2970. doi: 10.1098/rstb.1994.0068. URL http://dx.doi.org/10.1098/rstb.1994.0068.

[197] Claudia Neuhauser and Stephen M. Krone. The Genealogy of Samples in Models With Selection. *Genetics*, 145(2):519–534, February 1997. ISSN 1943-2631. URL http://www.genetics.org/content/145/2/519.abstract.

[198] Morten Nielsen, Claus Lundegaard, Thomas Blicher, Kasper Lamberth, Mikkel Harndahl, Sune Justesen, Gustav Røder, Bjoern Peters, Alessandro Sette, Ole Lund, and Søren Buus. NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PloS one*, 2(8), 2007. ISSN 1932-6203. URL http://view.ncbi.nlm.nih.gov/pubmed/17726526.

[199] Rasmus Nielsen and Ziheng Yang. Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene. *Genetics*, 148(3):929–936, March 1998. ISSN 1943-2631. URL http://www.genetics.org/content/148/3/929.abstract.

[200] W. Nietfield, M. Bauer, M. Fevrier, R. Maier, B. Holzwarth, R. Frank, B. Maier, Y. Riviere, and A. Meyerhans. Sequence constraints and recognition by CTL of an HLA-B27-restricted HIV-1 gag epitope. *The Journal of Immunology*, 154(5):2189–2197, March 1995. URL http://www.jimmunol.org/content/154/5/2189.abstract.

[201] Tamara Nora, Charlotte Charpentier, Olivier Tenaillon, Claire Hoede, François Clavel, and Allan J. Hance. Contribution of recombination to the evolution of human immunodeficiency viruses expressing resistance to antiretroviral treatment. *Journal of virology*, 81(14):7620–7628, July 2007. ISSN 0022-538X. doi: 10.1128/jvi.00083-07. URL http://dx.doi.org/10.1128/jvi.00083-07.

[202] M. Notohara. The coalescent and the genealogical process in geographically structured population. 29(1):59–75, 1990. doi: 10.1007/bf00173909. URL http://dx.doi.org/10.1007/bf00173909.

[203] V. Novitsky, P. Gilbert, T. Peter, M. F. McLane, S. Gaolekwe, N. Rybak, I. Thior, T. Ndung'u, R. Marlink, T. H. Lee, and M. Essex. Association between Virus-Specific T-Cell Responses and Plasma Viral Load in Human

Immunodeficiency Virus Type 1 Subtype C Infection. *Journal of Virology*, 77(2):882–890, January 2003. ISSN 1098-5514. doi: 10.1128/jvi.77.2.882-890.2003. URL http://dx.doi.org/10.1128/jvi.77.2.882-890.2003.

[204] M. A. Nowak, R. M. May, R. E. Phillips, S. Rowland-Jones, D. G. Lalloo, S. McAdam, P. Klenerman, B. Köppe, K. Sigmund, and C. R. Bangham. Antigenic oscillations and shifting immunodominance in HIV-1 infections. *Nature*, 375(6532):606–611, June 1995. ISSN 0028-0836. doi: 10.1038/375606a0. URL http://dx.doi.org/10.1038/375606a0.

[205] Collection of B-list epitopes. http://www.hiv.lanl.gov/content/immunology/tables/ctl_summary.html.

[206] Maria I. Oliveira, Paul A. Rota, Suely P. Curti, Cristina A. Figueiredo, Ana M. Afonso, Marcia Theobaldo, Luiza T. Souza, Stephanie L. Liffick, William J. Bellini, Jose C. Moraes, Klaus E. Stevien, and Edison L. Durigon. Genetic homogeneity of measles viruses associated with a measles outbreak, São Paulo, Brazil, 1997. *Emerging infectious diseases*, 8(8):808–813, August 2002. ISSN 1080-6040. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2732519/.

[207] Adewunmi Onafuwa-Nuga and Alice Telesnitsky. The remarkable frequency of human immunodeficiency virus type 1 genetic recombination. *Microbiology and molecular biology reviews : MMBR*, 73(3), September 2009. ISSN 1098-5557. doi: 10.1128/mmbr.00012-09. URL http://dx.doi.org/10.1128/mmbr.00012-09.

[208] World Health Organisation. The top 10 causes of death 2012. http://www.who.int/mediacentre/factsheets/fs310/en/.

[209] J. Overbaugh and L. M. Rudensey. Alterations in potential sites for glycosylation predominate during evolution of the simian immunodeficiency virus envelope gene in macaques. *Journal of virology*, 66(10):5937–5948, October 1992. ISSN 0022-538X. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC241471/.

[210] Duncan Palmer, John Frater, Rodney Philips, Angela McLean, and Gil McVean. Integrating genealogical and dynamical modelling to infer escape and reversion rates in HIV epitopes, February 2013. URL http://arxiv.org/abs/1302.1098.

[211] Duncan Palmer, John Frater, Rodney Phillips, Angela R. McLean, and Gil McVean. Integrating genealogical and dynamical modelling to infer escape and reversion rates in HIV epitopes. *Proceedings of the Royal Society B: Biological Sciences*, 280(1762), July 2013. ISSN 1471-2954. doi: 10.1098/rspb.2013.0696. URL http://dx.doi.org/10.1098/rspb.2013.0696.

[212] Calvin Pan, Joseph Kim, Lamei Chen, Qi Wang, and Christopher Lee. The HIV positive selection mutation database. *Nucleic acids research*, 35 (Database issue):D371–D375, January 2007. ISSN 1362-4962. doi: 10.1093/nar/gkl855. URL http://dx.doi.org/10.1093/nar/gkl855.

[213] Paul W. H. I. Parren, Dennis R. Burton, and Quentin J. Sattentau. HIV-1 antibody - debris or virion? *Nature Medicine*, 3(4):366–367, April 1997. doi: 10.1038/nm0497-366d. URL http://dx.doi.org/10.1038/nm0497-366d.

[214] Martine Peeters, Aissatou Gueye, Souleymane Mboup, Frederic Bibollet-Ruche, Euloge Ekaza, Claire Mulanga, Rasmata Ouedrago, Regine Gandji, Pierre Mpele, Germaine Dibanga, Brahima Koumare, Mamadou Saidou, Eka Esu-Williams, Jean-Piere Lombart, Wanta Badombena, Nkandu Luo, Marleen V. Haesevelde, and Eric Delaporte. Geographical distribution of HIV-1 group O viruses in Africa. *AIDS*, 11(4):493–498, March 1997. ISSN 0269-9370. doi: 10.1097/00002030-199704000-00013. URL http://dx.doi.org/10.1097/00002030-199704000-00013.

[215] A. S. Perelson, A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science (New York, N.Y.)*, 271(5255):1582–1586, March 1996. ISSN 0036-8075. doi: 10.1126/science.271.5255.1582. URL http://dx.doi.org/10.1126/science.271.5255.1582.

[216] Alan S. Perelson, Paulina Essunger, Yunzhen Cao, Mika Vesanen, Arlene Hurley, Kalle Saksela, Martin Markowitz, and David D. Ho. Decay characteristics of HIV-1-infected compartments during combination therapy. *Nature*, 387(6629):188–191, May 1997. doi: 10.1038/387188a0. URL http://dx.doi.org/10.1038/387188a0.

[217] Luc Perrin, Laurent Kaiser, and Sabine Yerly. Travel and the spread of HIV-1 genetic variants. *The Lancet Infectious Diseases*, 3(1):22–27, January 2003. ISSN 14733099. doi: 10.1016/s1473-3099(03)00484-5. URL http://dx.doi.org/10.1016/s1473-3099(03)00484-5.

[218] Rodney E. Phillips, Sarah Rowland-Jones, Douglas F. Nixon, Frances M. Gotch, Jon P. Edwards, Afolabi O. Ogunlesi, John G. Elvin, Jonathan A. Rothbard, Charles R. M. Bangham, Charles R. Rizza, and Andrew J. McMichael. Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. *Nature*, 354(6353):453–459, December 1991. doi: 10.1038/354453a0. URL http://dx.doi.org/10.1038/354453a0.

[219] Jean-Christophe Plantier, Marie Leoz, Jonathan E. Dickerson, Fabienne De Oliveira, Francois Cordonnier, Veronique Lemee, Florence Damond, David L. Robertson, and Francois Simon. A new human immunodeficiency virus derived from gorillas. *Nature Medicine*, 15(8):871–872, August 2009. ISSN 1078-8956. doi: 10.1038/nm.2016. URL http://dx.doi.org/10.1038/nm.2016.

[220] Art F. Y. Poon, Sergei L. Kosakovsky Pond, Douglas D. Richman, and Simon D. W. Frost. Mapping Protease Inhibitor Resistance to Human Immunodeficiency Virus Type 1 Sequence Polymorphisms within Patients. *Journal of Virology*, 81(24):13598–13607, December 2007. ISSN 1098-5514. doi: 10.1128/jvi.01570-07. URL http://dx.doi.org/10.1128/jvi.01570-07.

[221] Kate Poropatich and David J. Sullivan. Human immunodeficiency virus type 1 long-term non-progressors: the viral, genetic and immunological basis for disease non-progression. *The Journal of general virology*, 92(Pt 2): 247–268, February 2011. ISSN 1465-2099. doi: 10.1099/vir.0.027102-0. URL http://dx.doi.org/10.1099/vir.0.027102-0.

[222] D. Posada and K. A. Crandall. Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). *Molecular biology and evolution*, 18(6):897–906, June 2001. ISSN 0737-4038. URL http://mbe.oxfordjournals.org/content/18/6/897.abstract.

[223] David Posada. jModelTest: phylogenetic model averaging. *Molecular biology and evolution*, 25(7):1253–1256, July 2008. ISSN 1537-1719. doi: 10.1093/molbev/msn083. URL http://dx.doi.org/10.1093/molbev/msn083.

[224] B. D. Preston, B. J. Poiesz, and L. A. Loeb. Fidelity of HIV-1 reverse transcriptase. *Science (New York, N.Y.)*, 242(4882):1168–1171, November 1988. ISSN 0036-8075. URL http://view.ncbi.nlm.nih.gov/pubmed/2460924.

[225] D. A. Price, P. J. Goulder, P. Klenerman, A. K. Sewell, P. J. Easterbrook, M. Troop, C. R. Bangham, and R. E. Phillips. Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. *Proceedings of the National Academy of Sciences of the United States of America*, 94(5): 1890–1895, March 1997. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC20013/.

[226] David A. Price, Philip J. R. Goulder, Paul Klenerman, Andrew K. Sewell, Philippa J. Easterbrook, Maxine Troop, Charles R. M. Bangham, and Rodney E. Phillips. Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. *Proceedings of the National Academy of Sciences*, 94(5):1890–1895, March 1997. URL http://www.pnas.org/content/94/5/1890.abstract.

[227] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*, 26(7):1641–1650, July 2009. ISSN 1537-1719. doi: 10.1093/molbev/msp077. URL http://dx.doi.org/10.1093/molbev/msp077.

[228] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. FastTree 2 - Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS*

ONE, 5(3):e9490+, March 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0009490. URL http://dx.doi.org/10.1371/journal.pone.0009490.

[229] Recombinant Identification Program. http://www.hiv.lanl.gov/content/sequence/RIP/RIP.html.

[230] O. G. Pybus, A. Rambaut, and P. H. Harvey. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, 155(3):1429–1437, July 2000. ISSN 0016-6731. URL http://www.genetics.org/content/155/3/1429.abstract.

[231] Oliver G. Pybus and Andrew Rambaut. Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet*, 10(8):540–550, August 2009. ISSN 1471-0056. doi: 10.1038/nrg2583. URL http://dx.doi.org/10.1038/nrg2583.

[232] Oliver G. Pybus, Michael A. Charleston, Sunetra Gupta, Andrew Rambaut, Edward C. Holmes, and Paul H. Harvey. The Epidemic Behavior of the Hepatitis C Virus. *Science*, 292(5525):2323–2325, June 2001. doi: 10.1126/science.1058321. URL http://dx.doi.org/10.1126/science.1058321.

[233] Andrew Rambaut, David Posada, Keith A. Crandall, and Edward C. Holmes. The causes and consequences of HIV evolution. *Nat Rev Genet*, 5(1):52–61, January 2004. ISSN 1471-0056. doi: 10.1038/nrg1246. URL http://dx.doi.org/10.1038/nrg1246.

[234] Bertha C. Ramirez, Etienne Simon-Loriere, Roman Galetto, and Matteo Negroni. Implications of recombination for HIV diversity. *Virus Research*, 134 (1-2):64–73, June 2008. ISSN 01681702. doi: 10.1016/j.virusres.2008.01.007. URL http://dx.doi.org/10.1016/j.virusres.2008.01.007.

[235] H. Rammensee, J. Bachmann, N. P. Emmerich, O. A. Bachor, and S. Stevanović. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*, 50(3-4):213–219, November 1999. ISSN 0093-7711. URL http://view.ncbi.nlm.nih.gov/pubmed/10602881.

[236] Bharat Ramratnam, Sebastian Bonhoeffer, James Binley, Arlene Hurley, Linqi Zhang, John E. Mittler, M. Markowitz, John P. Moore, Alan S.

Perelson, and David D. Ho. Rapid production and clearance of HIV-1 and hepatitis C virus assessed by large volume plasma apheresis. *The Lancet*, 354(9192):1782–1785, November 1999. ISSN 01406736. doi: 10.1016/s0140-6736(99)02035-8. URL http://dx.doi.org/10.1016/s0140-6736(99)02035-8.

[237] David A. Rasmussen, Oliver Ratmann, and Katia Koelle. Inference for Nonlinear Epidemiological Models Using Genealogies and Time Series. *PLoS Comput Biol*, 7(8):e1002136+, August 2011. doi: 10.1371/journal.pcbi.1002136. URL http://dx.doi.org/10.1371/journal.pcbi.1002136.

[238] Douglas D. Richman, Terri Wrin, Susan J. Little, and Christos J. Petropoulos. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. *Proceedings of the National Academy of Sciences of the United States of America*, 100(7):4144–4149, April 2003. ISSN 0027-8424. doi: 10.1073/pnas.0630530100. URL http://dx.doi.org/10.1073/pnas.0630530100.

[239] C. Rinaldo, X. L. Huang, Z. F. Fan, M. Ding, L. Beltz, A. Logar, D. Panicali, G. Mazzara, J. Liebmann, and M. Cottrill. High levels of anti-human immunodeficiency virus type 1 (HIV-1) memory cytotoxic T-lymphocyte activity and low viral load are associated with lack of disease in HIV-1-infected long-term nonprogressors. *Journal of virology*, 69(9):5838–5842, September 1995. ISSN 0022-538X. URL http://jvi.asm.org/content/69/9/5838.abstract.

[240] J. D. Roberts, K. Bebenek, and T. A. Kunkel. The accuracy of reverse transcriptase from HIV-1. *Science*, 242(4882):1171–1173, November 1988. ISSN 0036-8075. doi: 10.1126/science.2460925. URL http://dx.doi.org/10.1126/science.2460925.

[241] D. L. Robertson, J. P. Anderson, J. A. Bradac, J. K. Carr, B. Foley, R. K. Funkhouser, F. Gao, B. H. Hahn, M. L. Kalish, C. Kuiken, G. H. Learn, T. Leitner, F. McCutchan, S. Osmanov, M. Peeters, D. Pieniazek, M. Salminen, P. M. Sharp, S. Wolinsky, and B. Korber. HIV-1 nomenclature proposal. *Science (New York, N.Y.)*, 288(5463):55–56, April 2000. ISSN 0036-8075. URL http://view.ncbi.nlm.nih.gov/pubmed/10766634.

[242] James Robinson, Jason A. Halliwell, Hamish McWilliam, Rodrigo Lopez, Peter Parham, and Steven G. Marsh. The IMGT/HLA database. *Nucleic acids research*, 41(Database issue):D1222–D1227, January 2013. ISSN 1362-4962. doi: 10.1093/nar/gks949. URL http://dx.doi.org/10.1093/nar/gks949.

[243] A. G. Rodrigo, E. G. Shpaer, E. L. Delwart, A. K. Iversen, M. V. Gallo, J. Brojatsch, M. S. Hirsch, B. D. Walker, and J. I. Mullins. Coalescent estimates of HIV-1 generation time in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 96(5):2187–2191, March 1999. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC26758/.

[244] Fredrik Ronquist, Maxim Teslenko, Paul van der Mark, Daniel L. Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A. Suchard, and John P. Huelsenbeck. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology*, 61(3):539–542, May 2012. ISSN 1076-836X. doi: 10.1093/sysbio/sys029. URL http://dx.doi.org/10.1093/sysbio/sys029.

[245] Pierre Roques, David L. Robertson, Sandrine Souquière, Cristian Apetrei, Eric Nerrienet, Françoise Barré-Sinoussi, Michaela Müller-Trutwin, and François Simon. Phylogenetic characteristics of three new HIV-1 N strains and implications for the origin of group N. *AIDS (London, England)*, 18(10): 1371–1381, July 2004. ISSN 0269-9370. URL http://view.ncbi.nlm.nih.gov/pubmed/15199313.

[246] Noah A. Rosenberg and Magnus Nordborg. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat Rev Genet*, 3(5): 380–390, May 2002. ISSN 1471-0056. doi: 10.1038/nrg795. URL http://dx.doi.org/10.1038/nrg795.

[247] J. S. Rota, J. L. Heath, P. A. Rota, G. E. King, M. L. Celma, J. Carabaña, R. Fernandez-Muñoz, D. Brown, L. Jin, and W. J. Bellini. Molecular epidemiology of measles virus: identification of pathways of transmission and implications for measles elimination. *The Journal of infectious diseases*, 173

(1):32–37, January 1996. ISSN 0022-1899. URL http://view.ncbi.nlm.nih.gov/pubmed/8537679.

[248] Paul A. Rota and William J. Bellini. Update on the Global Distribution of Genotypes of Wild Type Measles Viruses. *Journal of Infectious Diseases*, 187(Supplement 1):S270–S276, May 2003. doi: 10.1086/368042. URL http://dx.doi.org/10.1086/368042.

[249] Christine M. Rousseau, Gerald H. Learn, Tanmoy Bhattacharya, David C. Nickle, David Heckerman, Senica Chetty, Christian Brander, Philip J. Goulder, Bruce D. Walker, Photini Kiepiela, Bette T. Korber, and James I. Mullins. Extensive intrasubtype recombination in South African human immunodeficiency virus type 1 subtype C infections. *Journal of virology*, 81(9): 4492–4500, May 2007. ISSN 0022-538X. doi: 10.1128/jvi.02050-06. URL http://dx.doi.org/10.1128/jvi.02050-06.

[250] Christine M. Rousseau, Marcus G. Daniels, Jonathan M. Carlson, Carl Kadie, Hayley Crawford, Andrew Prendergast, Philippa Matthews, Rebecca Payne, Morgane Rolland, Dana N. Raugi, Brandon S. Maust, Gerald H. Learn, David C. Nickle, Hoosen Coovadia, Thumbi Ndung'u, Nicole Frahm, Christian Brander, Bruce D. Walker, Philip J. Goulder, Tanmoy Bhattacharya, David E. Heckerman, Bette T. Korber, and James I. Mullins. HLA class I-driven evolution of human immunodeficiency virus type 1 subtype c proteome: immune escape and viral load. *Journal of virology*, 82 (13):6434–6446, July 2008. ISSN 1098-5514. doi: 10.1128/jvi.02455-07. URL http://dx.doi.org/10.1128/jvi.02455-07.

[251] S. Rowland-Jones, T. Dong, P. Krausa, J. Sutton, H. Newell, K. Ariyoshi, F. Gotch, S. Sabally, T. Corrah, J. Kimani, K. MacDonald, F. Plummer, J. Ndinya-Achola, H. Whittle, and A. McMichael. The role of cytotoxic T-cells in HIV infection. *Developments in biological standardization*, 92:209–214, 1998. ISSN 0301-5149. URL http://view.ncbi.nlm.nih.gov/pubmed/9554277.

[252] Andres C. Ruiz, Juan Morales, and Maria Zambrano. Analysis of the genetic variation in Mycobacterium tuberculosis strains by multiple

genome alignments. *BMC Research Notes*, 1(1):110+, 2008. ISSN 1756-0500. doi: 10.1186/1756-0500-1-110. URL http://dx.doi.org/10.1186/1756-0500-1-110.

[253] Victor Sanchez-Merino, Siwei Nie, and Katherine Luzuriaga. HIV-1-Specific CD8+ T Cell Responses and Viral Evolution in Women and Infants. *The Journal of Immunology*, 175(10):6976–6986, November 2005. URL http://www.jimmunol.org/content/175/10/6976.abstract.

[254] J. E. Schmitz, M. J. Kuroda, S. Santra, V. G. Sasseville, M. A. Simon, M. A. Lifton, P. Racz, K. Tenner-Racz, M. Dalesandro, B. J. Scallon, J. Ghrayeb, M. A. Forman, D. C. Montefiori, E. P. Rieber, N. L. Letvin, and K. A. Reimann. Control of viremia in simian immunodeficiency virus infection by CD8+ lymphocytes. *Science (New York, N.Y.)*, 283(5403):857–860, February 1999. ISSN 0036-8075. URL http://view.ncbi.nlm.nih.gov/pubmed/9933172.

[255] Stephanie J. Schrag, Paul A. Rota, and William J. Bellini. Spontaneous Mutation Rate of Measles Virus: Direct Estimation Based on Mutations Conferring Monoclonal Antibody Resistance. *J. Virol.*, 73(1):51–54, January 1999. URL http://jvi.asm.org/cgi/content/abstract/73/1/51.

[256] Los Alamos HIV sequence database. http://www.hiv.lanl.gov/.

[257] R. L. Shapiro, M. D. Hughes, A. Ogwu, D. Kitch, S. Lockman, C. Moffat, J. Makhema, S. Moyo, I. Thior, K. McIntosh, E. van Widenfelt, J. Leidner, K. Powis, A. Asmelash, E. Tumbare, S. Zwerski, U. Sharma, E. Handelsman, K. Mburu, O. Jayeoba, E. Moko, S. Souda, E. Lubega, M. Akhtar, C. Wester, R. Tuomola, W. Snowden, M. Martinez-Tristani, L. Mazhani, and M. Essex. Antiretroviral Regimens in Pregnancy and Breast-Feeding in Botswana. *N Engl J Med*, 362(24):2282–2294, June 2010. doi: 10.1056/nejmoa0907736. URL http://dx.doi.org/10.1056/nejmoa0907736.

[258] Paul M. Sharp and Beatrice H. Hahn. Origins of HIV and the AIDS pandemic. *Cold Spring Harbor perspectives in medicine*, 1(1):a006841+, Septem-

ber 2011. ISSN 2157-1422. doi: 10.1101/cshperspect.a006841. URL http://dx.doi.org/10.1101/cshperspect.a006841.

[259] Ann M. Sheehy, Nathan C. Gaddis, Jonathan D. Choi, and Michael H. Malim. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature*, 418(6898):646–650, July 2002. ISSN 0028-0836. doi: 10.1038/nature00939. URL http://dx.doi.org/10.1038/nature00939.

[260] Daniel Shriner, David C. Nickle, Mark A. Jensen, and James I. Mullins. Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genetics Research*, 81:115–121, March 2003. ISSN 1469-5073. doi: 10.1017/s0016672303006128. URL http://dx.doi.org/10.1017/s0016672303006128.

[261] Peter Simmonds. Genetic diversity and evolution of hepatitis C virus - 15 years on. *Journal of General Virology*, 85(11):3173–3188, November 2004. ISSN 0022-1317. doi: 10.1099/vir.0.80401-0. URL http://dx.doi.org/10.1099/vir.0.80401-0.

[262] F. Simon, P. Mauclère, P. Roques, I. Loussert-Ajaka, M. C. Müller-Trutwin, S. Saragosti, M. C. Georges-Courbot, F. Barré-Sinoussi, and F. Brun-Vézinet. Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. *Nature medicine*, 4(9):1032–1037, September 1998. ISSN 1078-8956. doi: 10.1038/2017. URL http://dx.doi.org/10.1038/2017.

[263] N. V. Sipsas, S. A. Kalams, A. Trocha, S. He, W. A. Blattner, B. D. Walker, and R. P. Johnson. Identification of type-specific cytotoxic T lymphocyte responses to homologous viral proteins in laboratory workers accidentally infected with HIV-1. *The Journal of clinical investigation*, 99(4):752–762, February 1997. ISSN 0021-9738. URL http://view.ncbi.nlm.nih.gov/pubmed/9045880.

[264] M. Slatkin and R. R. Hudson. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, 129

(2):555–562, October 1991. ISSN 0016-6731. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1204643/.

[265] Davey M. Smith, Joseph K. Wong, George K. Hightower, Caroline C. Ignacio, Kersten K. Koelsch, Eric S. Daar, Douglas D. Richman, and Susan J. Little. Incidence of HIV superinfection following primary infection. *JAMA*, 292(10):1177–1178, September 2004. ISSN 1538-3598. URL http://view.ncbi.nlm.nih.gov/pubmed/15353529.

[266] Gavin J. D. Smith, Dhanasekaran Vijaykrishna, Justin Bahl, Samantha J. Lycett, Michael Worobey, Oliver G. Pybus, Siu K. Ma, Chung L. Cheung, Jayna Raghwani, Samir Bhatt, J. S. Malik Peiris, Yi Guan, and Andrew Rambaut. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*, 459(7250):1122–1125, June 2009. ISSN 0028-0836. doi: 10.1038/nature08182. URL http://dx.doi.org/10.1038/nature08182.

[267] Tanja Stadler. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *Journal of theoretical biology*, 261 (1):58–66, November 2009. ISSN 1095-8541. doi: 10.1016/j.jtbi.2009.07.018. URL http://dx.doi.org/10.1016/j.jtbi.2009.07.018.

[268] Tanja Stadler. Sampling-through-time in birth-death trees. *Journal of Theoretical Biology*, 267(3):396–404, December 2010. ISSN 00225193. doi: 10.1016/j.jtbi.2010.09.010. URL http://dx.doi.org/10.1016/j.jtbi.2010.09.010.

[269] Tanja Stadler, Roger Kouyos, Viktor von Wyl, Sabine Yerly, Jürg Böni, Philippe Bürgisser, Thomas Klimkait, Beda Joos, Philip Rieder, Dong Xie, Huldrych F. Günthard, Alexei J. Drummond, Sebastian Bonhoeffer, and the Swiss HIV Cohort Study. Estimating the Basic Reproductive Number from Viral Sequence Data. *Molecular Biology and Evolution*, 29(1):347–357, January 2012. ISSN 1537-1719. doi: 10.1093/molbev/msr217. URL http://dx.doi.org/10.1093/molbev/msr217.

[270] Tanja Stadler, Denise Kühnert, Sebastian Bonhoeffer, and Alexei J. Drummond. Birth-death skyline plot reveals temporal changes of epidemic

spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences*, 110(1):201207965–233, December 2012. ISSN 1091-6490. doi: 10.1073/pnas.1207965110. URL http://dx.doi.org/10.1073/pnas.1207965110.

[271] A. Stamatakis, T. Ludwig, and H. Meier. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, 21(4):456–463, February 2005. ISSN 1460-2059. doi: 10.1093/bioinformatics/bti191. URL http://dx.doi.org/10.1093/bioinformatics/bti191.

[272] Leonidas Stamatatos, Lynn Morris, Dennis R. Burton, and John R. Mascola. Neutralizing antibodies generated during natural HIV-1 infection: good news for an HIV-1 vaccine? *Nature Medicine*, 15(8):866–870, June 2009. ISSN 1078-8956. doi: 10.1038/nm.1949. URL http://dx.doi.org/10.1038/nm.1949.

[273] Ivan Stratov, C. Jane Dale, Socheata Chea, James McCluskey, and Stephen J. Kent. Induction of T-cell immunity to antiretroviral drug-resistant human immunodeficiency virus type 1. *Journal of virology*, 79(12):7728–7737, June 2005. ISSN 0022-538X. URL http://view.ncbi.nlm.nih.gov/pubmed/15919925.

[274] Korbinian Strimmer and Oliver G. Pybus. Exploring the Demographic History of DNA Sequences Using the Generalized Skyline Plot. *Molecular Biology and Evolution*, 18(12):2298–2305, December 2001. ISSN 1537-1719. URL http://mbe.oxfordjournals.org/content/18/12/2298.abstract.

[275] The International HIV Controllers Study. The Major Genetic Determinants of HIV-1 Control Affect HLA Class I Peptide Presentation. *Science*, 330 (6010):1551–1557, December 2010. ISSN 1095-9203. doi: 10.1126/science.1195271. URL http://dx.doi.org/10.1126/science.1195271.

[276] Yoshiyuki Suzuki and Masatoshi Nei. Origin and Evolution of Influenza Virus Hemagglutinin Genes. *Molecular Biology and Evolution*, 19(4):501–

509, April 2002. URL http://mbe.oxfordjournals.org/content/19/4/501.abstract.

[277] J. Taylor and B. Korber. HIV-1 intra-subtype superinfection rates: estimates using a structured coalescent with recombination. *Infection, Genetics and Evolution*, 5(1):85–95, January 2005. ISSN 15671348. doi: 10.1016/j.meegid.2004.07.001. URL http://dx.doi.org/10.1016/j.meegid.2004.07.001.

[278] J. Taylor and G. Rudenko. Switching trypanosome coats: what's in the wardrobe? *Trends in Genetics*, 22(11):614–620, November 2006. ISSN 01689525. doi: 10.1016/j.tig.2006.08.003. URL http://dx.doi.org/10.1016/j.tig.2006.08.003.

[279] E. A. Thompson. *Human Evolutionary Trees*. Cambridge University Press, October 1975. ISBN 0521099455. URL http://www.worldcat.org/isbn/0521099455.

[280] Michael M. Thomson, Lucía Pérez-Álvarez, and Rafael Nájera. Molecular epidemiology of HIV-1 genetic forms and its significance for vaccine development and therapy. *The Lancet Infectious Diseases*, 2(8):461–471, August 2002. ISSN 14733099. doi: 10.1016/s1473-3099(02)00343-2. URL http://dx.doi.org/10.1016/s1473-3099(02)00343-2.

[281] L. S. Tiley, M. H. Malim, H. K. Tewary, P. G. Stockley, and B. R. Cullen. Identification of a high-affinity RNA-binding site for the human immunodeficiency virus type 1 Rev protein. *Proceedings of the National Academy of Sciences of the United States of America*, 89(2):758–762, January 1992. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC48318/.

[282] Pénélope Troude, Marie-Laure Chaix, Laurent Tran, Christiane Deveau, Rémonie Seng, Jean-François Delfraissy, Christine Rouzioux, Cécile Goujard, and Laurence Meyer. No evidence of a change in HIV-1 virulence since 1996 in France. *AIDS*, 23(10):1261–1267, June 2009. ISSN 0269-9370. doi: 10.1097/qad.0b013e32832b51ef. URL http://dx.doi.org/10.1097/qad.0b013e32832b51ef.

[283] Kun-Nan Tsai and Guang-Wu Chen. Influenza genome diversity and evolution. *Microbes and Infection*, 13(5):479–488, May 2011. ISSN 12864579. doi: 10.1016/j.micinf.2011.01.013. URL http://dx.doi.org/10.1016/j.micinf.2011.01.013.

[284] C. Turner. The rate of antigenic variation in fly-transmitted and syringe-passaged infections of Trypanosoma brucei. *FEMS Microbiology Letters*, 153(1):227–231, August 1997. ISSN 03781097. doi: 10.1016/s0378-1097(97)00266-8. URL http://dx.doi.org/10.1016/s0378-1097(97)00266-8.

[285] C. M. R. Turner and J. D. Barry. High frequency of antigenic variation in Trypanosoma brucei rhodesiense infections. *Parasitology*, 99(01):67–75, 1989. doi: 10.1017/s0031182000061035. URL http://dx.doi.org/10.1017/s0031182000061035.

[286] UNAIDS. Global reports - unaids report on the global aids epidemic 2013. http://www.unaids.org/en/media/unaids/contentassets/documents/epidemiology/2013/gr2013/UNAIDS_Global_Report_2013_en.pdf.

[287] Ana Vallari, Vera Holzmayer, Barbara Harris, Julie Yamaguchi, Charlotte Ngansop, Florence Makamche, Dora Mbanya, Lazare Kaptué, Nicaise Ndembi, Lutz Gürtler, Sushil Devare, and Catherine A. Brennan. Confirmation of Putative HIV-1 Group P in Cameroon. *Journal of Virology*, 85(3):1403–1407, February 2011. ISSN 1098-5514. doi: 10.1128/jvi.02005-10. URL http://dx.doi.org/10.1128/jvi.02005-10.

[288] Fran Van Heuverswyn, Yingying Li, Cecile Neel, Elizabeth Bailes, Brandon F. Keele, Weimin Liu, Severin Loul, Christelle Butel, Florian Liegeois, Yanga Bienvenue, Eitel M. Ngolle, Paul M. Sharp, George M. Shaw, Eric Delaporte, Beatrice H. Hahn, and Martine Peeters. Human immunodeficiency viruses: SIV infection in wild gorillas. *Nature*, 444(7116):164, November 2006. ISSN 0028-0836. doi: 10.1038/444164a. URL http://dx.doi.org/10.1038/444164a.

[289] Gerard J. van Westen, Alwin Hendriks, Jörg K. Wegner, Adriaan P. Ijzerman, Herman W. van Vlijmen, and Andreas Bender. Significantly im-

proved HIV inhibitor efficacy prediction employing proteochemometric models generated from antivirogram data. *PLoS computational biology*, 9 (2), 2013. ISSN 1553-7358. URL http://view.ncbi.nlm.nih.gov/pubmed/23436985.

[290] Nicole Vidal, Martine Peeters, Claire Mulanga-Kabeya, Nzila Nzilambi, David Robertson, Wantabala Ilunga, Hurogo Sema, Kazadi Tshimanga, Beni Bongo, and Eric Delaporte. Unprecedented Degree of Human Immunodeficiency Virus Type 1 (HIV-1) Group M Genetic Diversity in the Democratic Republic of Congo Suggests that the HIV-1 Pandemic Originated in Central Africa. *Journal of Virology*, 74(22):10498–10507, November 2000. ISSN 1098-5514. doi: 10.1128/jvi.74.22.10498-10507.2000. URL http://dx.doi.org/10.1128/jvi.74.22.10498-10507.2000.

[291] Benjamin F. Voight, Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K. Pritchard. A map of recent positive selection in the human genome. *PLoS biology*, 4(3):e72+, March 2006. ISSN 1545-7885. doi: 10.1371/journal.pbio.0040072. URL http://dx.doi.org/10.1371/journal.pbio.0040072.

[292] Erik M. Volz, Sergei L. Kosakovsky Pond, Melissa J. Ward, Andrew J. Leigh Brown, and Simon D. W. Frost. Phylodynamics of Infectious Disease Epidemics. *Genetics*, 183(4):1421–1430, December 2009. ISSN 1943-2631. doi: 10.1534/genetics.109.106021. URL http://dx.doi.org/10.1534/genetics.109.106021.

[293] Erik M. Volz, James S. Koopman, Melissa J. Ward, Andrew L. Brown, and Simon D. W. Frost. Simple Epidemiological Dynamics Explain Phylogenetic Clustering of HIV from Patients with Recent Infection. *PLoS Comput Biol*, 8(6):e1002552+, June 2012. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002552. URL http://dx.doi.org/10.1371/journal.pcbi.1002552.

[294] John Wakeley. *Coalescent Theory: An Introduction*. Roberts & Company Publishers, 1 edition, June 2008. ISBN 0974707759. URL http://www.worldcat.org/isbn/0974707759.

[295] Laura M. Walker, Sanjay K. Phogat, Po-Ying Chan-Hui, Denise Wagner, Pham Phung, Julie L. Goss, Terri Wrin, Melissa D. Simek, Steven Fling, Jennifer L. Mitcham, Jennifer K. Lehrman, Frances H. Priddy, Ole A. Olsen, Steven M. Frey, Phillip W. Hammond, Protocol G. Principal Investigators, Stephen Kaminsky, Timothy Zamb, Matthew Moyle, Wayne C. Koff, Pascal Poignard, and Dennis R. Burton. Broad and Potent Neutralizing Antibodies from an African Donor Reveal a New HIV-1 Vaccine Target. *Science*, 326(5950):285–289, October 2009. ISSN 1095-9203. doi: 10.1126/science.1178746. URL http://dx.doi.org/10.1126/science.1178746.

[296] Polly R. Walker, Oliver G. Pybus, Andrew Rambaut, and Edward C. Holmes. Comparative population dynamics of HIV-1 subtypes B and C: subtype-specific differences in patterns of epidemic growth. *Infection, Genetics and Evolution*, 5(3):199–208, April 2005. ISSN 15671348. doi: 10.1016/j.meegid.2004.06.011. URL http://dx.doi.org/10.1016/j.meegid.2004.06.011.

[297] M. P. Wand. Fast Computation of Multivariate Kernel Estimators. *Journal of Computational and Graphical Statistics*, 3(4):433–445, December 1994. doi: 10.1080/10618600.1994.10474656. URL http://dx.doi.org/10.1080/10618600.1994.10474656.

[298] M. P. Wand and M. C. Jones. *Kernel Smoothing (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 1 edition, December 1994. ISBN 0412552701. URL http://www.worldcat.org/isbn/0412552701.

[299] Yaoyu E. Wang, Bin Li, Jonathan M. Carlson, Hendrik Streeck, Adrianne D. Gladden, Robert Goodman, Arne Schneidewind, Karen A. Power, Ildiko Toth, Nicole Frahm, Galit Alter, Christian Brander, Mary Carrington, Bruce D. Walker, Marcus Altfeld, David Heckerman, and Todd M. Allen. Protective HLA class I alleles that restrict acute-phase CD8+ T-cell responses are associated with viral escape mutations located in highly conserved regions of human immunodeficiency virus type 1. *Journal of virology*,

83(4):1845–1855, February 2009. ISSN 1098-5514. doi: 10.1128/jvi.01061-08. URL http://dx.doi.org/10.1128/jvi.01061-08.

[300] Swiss Confederation website. http://www.bag.admin.ch/hiv_aids/.

[301] Xiping Wei, Sajal K. Ghosh, Maria E. Taylor, Victoria A. Johnson, Emilio A. Emini, Paul Deutsch, Jeffrey D. Lifson, Sebastian Bonhoeffer, Martin A. Nowak, Beatrice H. Hahn, Michael S. Saag, and George M. Shaw. Viral dynamics in human immunodeficiency virus type 1 infection. *Nature*, 373 (6510):117–122, January 1995. ISSN 0028-0836. doi: 10.1038/373117a0. URL http://dx.doi.org/10.1038/373117a0.

[302] Xiping Wei, Julie M. Decker, Shuyi Wang, Huxiong Hui, John C. Kappes, Xiaoyun Wu, Jesus F. Salazar-Gonzalez, Maria G. Salazar, J. Michael Kilby, Michael S. Saag, Natalia L. Komarova, Martin A. Nowak, Beatrice H. Hahn, Peter D. Kwong, and George M. Shaw. Antibody neutralization and escape by HIV-1. *Nature*, 422(6929):307–312, March 2003. ISSN 0028-0836. doi: 10.1038/nature01470. URL http://dx.doi.org/10.1038/nature01470.

[303] C. C. Wilson, R. C. Brown, B. T. Korber, B. M. Wilkes, D. J. Ruhl, D. Sakamoto, K. Kunstman, K. Luzuriaga, I. C. Hanson, S. M. Widmayer, A. Wiznia, S. Clapp, A. J. Ammann, R. A. Koup, S. M. Wolinsky, and B. D. Walker. Frequent detection of escape from cytotoxic T-lymphocyte recognition in perinatal human immunodeficiency virus (HIV) type 1 transmission: the ariel project for the prevention of transmission of HIV from mother to infant. *Journal of virology*, 73(5):3975–3985, May 1999. ISSN 0022-538X. URL http://view.ncbi.nlm.nih.gov/pubmed/10196293.

[304] Daniel J. Wilson and Gilean McVean. Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics*, 172(3): 1411–1425, March 2006. ISSN 0016-6731. doi: 10.1534/genetics.105.044917. URL http://dx.doi.org/10.1534/genetics.105.044917.

[305] J. D. Wilson, G. S. Ogg, R. L. Allen, C. Davis, S. Shaunak, J. Downie, W. Dyer, C. Workman, S. Sullivan, A. J. McMichael, and S. L. Rowland-Jones. Direct visualization of HIV-1-specific cytotoxic T lymphocytes dur-

ing primary infection. *AIDS (London, England)*, 14(3):225–233, February 2000. ISSN 0269-9370. URL http://view.ncbi.nlm.nih.gov/pubmed/10716497.

[306] Dominik Wodarz and Martin A. Nowak. Specific therapy regimes could lead to long-term immunological control of HIV. *Proceedings of the National Academy of Sciences*, 96(25):14464–14469, December 1999. doi: 10.1073/pnas.96.25.14464. URL http://dx.doi.org/10.1073/pnas.96.25.14464.

[307] Dominik Wodarz and Martin A. Nowak. Mathematical models of HIV pathogenesis and treatment. *Bioessays*, 24(12):1178–1187, 2002. doi: 10.1002/bies.10196. URL http://dx.doi.org/10.1002/bies.10196.

[308] Evan Wood, Robert S. Hogg, Viviane Dias D. Lima, Thomas Kerr, Benita Yip, Brandon D. Marshall, and Julio S. Montaner. Highly active antiretroviral therapy and survival in HIV-infected injection drug users. *JAMA : the journal of the American Medical Association*, 300(5):550–554, August 2008. ISSN 1538-3598. doi: 10.1001/jama.300.5.550. URL http://dx.doi.org/10.1001/jama.300.5.550.

[309] Michael Worobey, Marlea Gemmel, Dirk E. Teuwen, Tamara Haselkorn, Kevin Kunstman, Michael Bunce, Jean-Jacques Muyembe, Jean-Marie M. Kabongo, Raphael M. Kalengayi, Eric Van Marck, Gilbert, and Steven M. Wolinsky. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature*, 455(7213):661–664, October 2008. ISSN 0028-0836. doi: 10.1038/nature07390. URL http://dx.doi.org/10.1038/nature07390.

[310] Z. Yang and B. Rannala. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Molecular Biology and Evolution*, 14(7):717–724, July 1997. ISSN 1537-1719. URL http://mbe.oxfordjournals.org/content/14/7/717.abstract.

[311] Melvyn W. Yap, Sébastien Nisole, Clare Lynch, and Jonathan P. Stoye. Trim5α protein restricts both HIV-1 and murine leukemia virus. *Proceedings of the National Academy of Sciences of the United States of Amer-*

ica, 101(29):10786–10791, July 2004. doi: 10.1073/pnas.0402876101. URL http://dx.doi.org/10.1073/pnas.0402876101.

[312] Yoshiyuki Yokomaku, Hideka Miura, Hiroko Tomiyama, Ai Kawana-Tachikawa, Masafumi Takiguchi, Asato Kojima, Yoshiyuki Nagai, Aikichi Iwamoto, Zene Matsuda, and Koya Ariyoshi. Impaired processing and presentation of cytotoxic-T-lymphocyte (CTL) epitopes are major escape mechanisms from CTL immune pressure in human immunodeficiency virus type 1 infection. *Journal of virology*, 78(3):1324–1332, February 2004. ISSN 0022-538X. doi: 10.1128/jvi.78.3.1324-1332.2004. URL http://dx.doi.org/10.1128/jvi.78.3.1324-1332.2004.

[313] Xianghui Yu, Yunkai Yu, Bindong Liu, Kun Luo, Wei Kong, Panyong Mao, and Xiao-Fang Yu. Induction of APOBEC3G Ubiquitination and Degradation by an HIV-1 Vif-Cul5-SCF Complex. *Science*, 302(5647):1056–1060, November 2003. ISSN 1095-9203. doi: 10.1126/science.1089591. URL http://dx.doi.org/10.1126/science.1089591.

[314] K. Yusim, J. J. Szinger, I. Honeyborne, C. Calef, P. J. Goulder, and B. T. Korber. Enhanced Motif Scan: A Tool to Scan for HLA Anchor Residues in Proteins. Technical report, 2004.

[315] Shao Chong C. Zhang, Eric Martin, Mariko Shimada, Sasha Blue B. Godfrey, Jennifer Fricke, Shirley Locastro, Nicole Y. Lai, Paul Liebesny, Jonathan M. Carlson, Chanson J. Brumme, Oluwatobi Awele A. Ogbechie, Huabiao Chen, Bruce D. Walker, Zabrina L. Brumme, Daniel G. Kavanagh, and Sylvie Le Gall. Aminopeptidase substrate preference affects HIV epitope presentation and predicts immune escape patterns in HIV-infected individuals. *Journal of immunology (Baltimore, Md. : 1950)*, 188(12):5924–5934, June 2012. ISSN 1550-6606. URL http://view.ncbi.nlm.nih.gov/pubmed/22586036.

[316] Ping Zhu, Jun Liu, Julian Bess, Elena Chertova, Jeffrey D. Lifson, Henry Grisé, Gilad A. Ofek, Kenneth A. Taylor, and Kenneth H. Roux. Distribution and three-dimensional structure of AIDS virus envelope spikes.

Nature, 441(7095):847–852, May 2006. ISSN 0028-0836. doi: 10.1038/nature04817. URL http://dx.doi.org/10.1038/nature04817.

[317] Ping Zhu, Hanspeter Winkler, Elena Chertova, Kenneth A. Taylor, and Kenneth H. Roux. Cryoelectron Tomography of HIV-1 Envelope Spikes: Further Evidence for Tripod-Like Legs. *PLoS Pathog*, 4(11):e1000203+, November 2008. doi: 10.1371/journal.ppat.1000203. URL http://dx.doi.org/10.1371/journal.ppat.1000203.

[318] Rosario Zuñiga, Aldo Lucchetti, Patricia Galvan, Shyla Sanchez, Carmen Sanchez, Ana Hernandez, Hugo Sanchez, Nicole Frahm, Caitlyn H. Linde, Hannah S. Hewitt, William Hildebrand, Marcus Altfeld, Todd M. Allen, Bruce D. Walker, Bette T. Korber, Thomas Leitner, Jorge Sanchez, and Christian Brander. Relative Dominance of Gag p24-Specific Cytotoxic T Lymphocytes Is Associated with Human Immunodeficiency Virus Control. *Journal of Virology*, 80(6):3122–3125, March 2006. ISSN 1098-5514. doi: 10.1128/jvi.80.6.3122-3125.2006. URL http://dx.doi.org/10.1128/jvi.80.6.3122-3125.2006.