Evolutionary Models of Antibody Lineages



KENNETH B. HOEHN

Oriel College
University of Oxford
Michaelmas Term 2017

A THESIS SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY IN GENOMIC MEDICINE AND STATISTICS

Abstract

Evolutionary Models of Antibody Lineages

Kenneth B. Hoehn, Oriel College

Doctor of Philosophy in Genomic Medicine and Statistics

Antibodies are proteins released into the blood and mucosa to identify and neutralize invading pathogens by binding to structures on their surface (antigens). Before being exported as antibodies, these vital components of the adaptive immune system are expressed, and refined, as membrane-bound B cell receptors (BCRs). BCRs are initially formed through somatic recombination of germline DNA, creating a large repertoire of unique sequences. After encountering antigen for the first time, BCRs undergo an evolutionary process of somatic mutation and clonal selection leading to improved antigen binding. Recently, next-generation sequencing has provided an unprecedented ability to characterize the genetic diversity of BCRs within individuals. Chapter 1 of this thesis overviews the work done elsewhere in the field until now. Chapter 2 uses summary statistics applied to high-throughput sequence data from a clinical trial to explore the genetic changes that occur in the repertoire during HIV infection. The results of these analyses motivated a more rigorous, model-based approach to understanding BCR diversity. Chapter 3 introduces a new phylogenetic substitution model that relaxes common model assumptions that are violated by somatic hypermutation. Chapter 4 expands this model to incorporate previously defined empirical models of somatic hypermutation, providing a more complete model of B cell maturation. Chapter 5 uses the models developed in Chapters 4 and 5 to explore dynamics of clonal selection during the maturation of three HIV broadly neutralizing antibody lineages. Finally, Chapter 6 shows how this framework may be scaled up to characterize data from entire BCR repertoires from a phylogenetic perspective.

Statement of Contribution

Chapter 1 (Introduction)

Portions of this chapter (Sections 1.1-1.6 and 1.8), were included in a modified form in review paper which I co-authored with Dr. Anna Fowler and my supervisors Profs. Gerton Lunter and Oliver Pybus:

Hoehn K. B., A. Fowler, G. Lunter, O. G. Pybus, 2016 The Diversity and Molecular Evolution of B-Cell Receptors during Infection. Mol. Biol. Evol. 30: 11471157.

Chapter 2

Patients were sampled by investigators of the SPARTAC clinical trial. All lab work, including RT-PCR amplification of BCRs (Section 2.2.2) and sequencing (Section 2.2.3) was performed by Dr. Rachael Bashford-Rogers and Dr. Astrid Gall at the Sanger Institute, Cambridge University. Initital sequence processing and reference mapping (Section 2.2.3), as well as generation of Figure 2.1 were performed by Dr. Rachael Bashford-Rogers at the Sanger Centre. I performed all other analyses. The content of this chapter was published as:

Hoehn K. B., A. Gall, R. Bashford-Rogers, S. J. Fidler, S. Kaye et al., 2015 Dynamics of immunoglobulin sequence diversity in HIV-1 infected individuals. Philos. Trans. R. Soc. B Biol. Sci. 370: 20140241.

Chapter 3

This chapter is wholly my own work, with input from my supervisors Profs. Gerton Lunter and Oliver Pybus. The contents of this chapter were published as:

Hoehn K. B., G. Lunter, O. G. Pybus, 2017 A Phylogenetic Codon Substitution Model for Antibody Lineages. Genetics 206: 417427.

Chapters 4 and 5

These chapters are wholly my own work, with input from my supervisors Profs. Gerton Lunter and Oliver Pybus.

Chapter 6

All of the PCR and sequencing procedures in this section was performed by Dr. Astrid Gall at the University of Cambridge, as well as the sequencing core facility at the Wellcome Trust Sanger Institute. The computational analyses presented are wholly my own work.

For my parents, Mike and Fran Hoehn.

Acknowledgments

I would like to thank my supervisor, Prof. Oliver Pybus, for his persistent support and guidance, and for challenging my work to be better. I would also like to thank my co-supervisor Prof. Gerton Lunter, who has provided invaluable input to the work I have done during my degree.

I would like to thank the Marshall Aid Commemoration Commission for its generous financial support throughout my work on this thesis. I would also like to thank Oriel College, the European Research Council, and the Oxford University Medical Sciences department for their financial support of my degree as well.

I could not have written this thesis without the love and support of my parents, Mike and Fran Hoehn.

I would also like to thank all of the friends and colleagues I've had the pleasure of spending time with while in the UK. I would particularly like to thank Shangnon Fei and Arjelle Agupitan, who have been invaluable companions in my time at Oxford.

Contents

Abstract i

Statement of Contribution ii

Acknowledgements iii

List of Figures x

1 Introduction 1

1.1 Overview 1

1.2 B cell development 4

1.3 Sequencing the BCR repertoire 8

1.4 Measuring BCR diversity 9

1.5 Clonal lineage assignment and clustering 13

1.6 Untangling mutation and selection 15

1.7 BCR dynamics during HIV infection 17

1.8 BCR phylogenetics 21

1.9 Phylogenetic substitution models and the maximum likelihood framework 24

1.10 Overview of thesis 30

2 Dynamics of Immunoglobulin sequence diversity in HIV-1 infected individuals 31

2.1 Introduction 31

2.2 Methods 34

2.2.1 HIV patients 34

2.2.2 RT-PCR . . . 35
2.2.3 Sequencing and reference-based V-D-J assignment . . . 35
2.2.4 Alignment . . . 36
2.2.5 Clonal assignment . . . 37
2.2.6 Measurement of BCR diversity . . . 38
2.2.7 Subsampling . . . 39
2.2.8 Analysis of BCR diversity . . . 40
2.2.9 Comparison with healthy controls . . . 41
2.2.10 Statement of contribution . . . 41
2.3 Results . . . 42
2.3.1 Untreated patients . . . 43
2.3.2 Treated patients . . . 43
2.3.3 Relationship between clinical variables and B cell diversity . 44
2.3.4 Analysis of BCR diversity and subsampling . . . 50
2.3.5 Comparison with healthy controls . . . 50
2.4 Discussion . . . 51

3 A phylogenetic codon substitution model for antibody lineages 57
3.1 Introduction . . . 57
3.2 Materials and Methods . . . 60
3.2.1 A molecular evolutionary model for antibody lineages . . . 60
3.2.2 Likelihood calculation and ancestral state reconstruction . . 64
3.2.3 Implementation . . . 67
3.2.4 bnAb lineage sequences and multiple sequence alignment . 67
3.2.5 Hotspot model selection . . . 69
3.2.6 Effectiveness of the mean field approximation . . . 70
3.2.7 Effects on ancestral state reconstruction . . . 73
3.3 Results . . . 74
3.3.1 A codon substitution model for phylogenies undergoing somatic hypermutation . . . 74

3.3.2 Hotspot model selection 80
3.3.3 Ancestral state reconstruction 81
3.4 Discussion 81

4 Semi-empirical substitution models **87**
4.1 Introduction 87
4.2 Materials and Methods 92
4.2.1 Discrete and semi-empirical substitution models 92
4.2.2 Multiple sequence alignment 97
4.2.3 Verifying effectiveness of WAG model 101
4.2.4 Model fitting and tree estimation 101
4.2.5 Efficient topology searching under HLP17 102
4.2.6 Reconstructing germline CDR3 sequences 102
4.3 Results 103
4.3.1 Characterizing parameter space of the semi-empirical WAG model 103
4.3.2 The best-fit model combination varies by lineage 106
4.3.3 Differences in topology estimated under different model combinations and published topologies 107
4.3.4 Ancestral sequence reconstructions 110
4.4 Discussion 114

5 Variation in clonal selection during maturation of HIV broadly neutralizing antibody lineages **120**
5.1 Introduction 120
5.2 Materials and Methods 124
5.2.1 bnAb sequences and phylogenetic tree construction 124
5.2.2 Monoclonal antibody neutralization profiles 125
5.2.3 Characterizing selection pressure in bnAb lineages 129
5.2.4 Detecting variation in selection pressure within a lineage . . 130

5.2.5 Multiple hypothesis test correction and method validation . 133
5.3 Results . . . 136
5.3.1 CH103 and VRC26 show significant differences in ω between CDR and FWR . . . 136
5.3.2 CH103 and VRC26 show significant ω_{CDR} heterogeneity among clades . . . 136
5.3.3 Low ω_{CDR} is generally associated with later timepoints . . . 144
5.3.4 Broad neutralization is not associated with reduced ω_{CDR} among clades . . . 145
5.3.5 Method validation . . . 147
5.4 Discussion . . . 150

6 Phylogenetic analysis of B cell repertoires 155
6.1 Introduction . . . 155
6.2 Methods . . . 158
6.2.1 Hierarchical substitution models . . . 158
6.2.2 Study subjects and repertoire sequencing . . . 159
6.2.3 Sequence processing and clonal clustering . . . 161
6.2.4 Lineage analysis . . . 162
6.2.5 Statement of contribution . . . 164
6.3 Results . . . 164
6.3.1 Clonal clusters are dominated by low diversity clones at single timepoints . . . 164
6.3.2 Some SHM parameter estimates are consistent among patients 167
6.3.3 Variation in selective effects within and between repertoires 170
6.4 Discussion . . . 177

7 Discussion 179
7.1 HLP17 and phylogenetic substitution models . . . 180
7.2 The past, present, and near future of antibody phylogenetics . . . 182

7.3 A broader look at the future . . . 185
7.4 Conclusion . . . 187

Appendices **189**

A Supplemental information for Chapter 2 **190**
A.1 Human B-cell receptor PCR primers . . . 190
A.2 Illustration of the BCR sequencing pipeline . . . 191
A.3 Illustration of the sequence alignment pipeline . . . 192
A.4 Relationship between Gini index and read depth between patients . 193
A.5 Gini index and read depth . . . 194
A.6 CD4+ cell counts and seroconversion information . . . 195
A.7 Effect of sub-sampling on BCR statistics . . . 196

B Supplemental information for Chapter 3 **197**
B.1 Likelihood computation and ancestral state reconstruction . . . 197
B.1.1 Calculating likelihood at each node . . . 198
B.1.2 Marginal ancestral state reconstruction . . . 200
B.2 Summary statistics of bNAb sequence data . . . 202
B.3 Parameter estimates from different motif models . . . 203
B.4 Analysis with random motifs . . . 204
B.5 Parameter results from fully context dependent simulation . . . 205
B.6 Analysis of interaction between h and ω . . . 207
B.7 Simulation analysis using empirical codon frequencies . . . 211
B.8 Ancestral reconstruction results in bnAb lineages . . . 212
B.9 Ancestral reconstruction results from simulations . . . 213

Bibliography **214**

List of Figures

1.1 Somatic rearrangement and functional division of the immunoglobulin heavy (IGH) chain locus during B cell development. The IGH locus begins with multiple V, D, and J segments. Of these, one of each type is spliced together to form the rearranged variable region. After transcription, the joined VDJ segments are spliced together with a neighbouring C region [Owen et al., 2013]. The re-arranged variable region is divided into non-consecutive framework (FWR) and complementary determining (CDR) regions. After translation, the heavy chain protein folds such that FWRs bind together, exposing the CDRs for use in antigen binding. This protein is paired with a light chain, which undergoes a similar somatic recombination process (VJ recombination) and also contains three exposed CDRs [Owen et al., 2013]. The full B cell receptor is displayed on the cell surface, with its contant regions associated with intracellular signalling proteins Igα and Igβ. See [Murphy et al., 2012] Figures 4.6, 5.1, and [Sompayrac, 2016], p. 29. 6

1.2 Chord diagrams showing the pairing of V and J segments within (a) productive and (b) non-productive IgM sequences from a single healthy individual. Chord widths represent the proportion of sequences with a given V (coloured) and J (grey) segment pairing. The five most common V segments in productive rearrangements (and all J segments) are labelled. Note that IGHV3-23/IGHJ4 was significantly more common in productive vs. non-productive rearrangements, which may indicate functional bias of that pairing. The figure was generated from data in [Elhanati et al., 2015], which was aligned to the IMGT reference ([Lefranc et al., 2009] using IgBLAST [Ye et al., 2013]. Productive rearrangements were subsampled to the same read depth as non-productive rearrangements ($2x10^5$ reads); the values displayed in (a) are means of 100 subsampling repetitions. 7

1.3 Network diagrams that visualise the diversity and clonal structure of BCR sequences. BCR sequences were obtained from (a) two healthy people, (b) two individuals infected with HIV-1, sampled during early infection, and (c) two patients with chronic lymphocytic leukaemia (CLL), a B cell cancer. Each point/circle represents a unique BCR sequence, the size of which is proportional to how common that sequence is. Edges are drawn between pairs of sequences that differ by exactly one nucleotide change. Note that samples do differ by read depth (approximately $3.4x10^4$ and $3.6x10^4$ for part (a), $9.2x10^4$ and $3.6x10^5$ for part (b), $5.1x10^4$ and $2.6x10^4$ for part (c). Parts (a) and (c) are reproduced from Bashford-Rogers et al. [Bashford-Rogers et al., 2013] and parts (b) from Chapter 2. 12

1.4 Observed mutation (sequence difference from germline) frequency among productive heavy chain immunoglobulin sequences across the V-gene sequence (horizontal axis, IMGT unique numbering). The distribution of mutations across the region is strongly non-uniform, with mutations more likely to occur at certain positions. The CDR2 region (middle shaded box) has a high rate of observed mutations and is thought to be more important in antigen binding than the surrounding framework regions (FWR2 and FWR3). This figure was generated from the same data set as Figure 1.2. 17

1.5 Representation of the *Env* protein on the HIV capsid. a) shows the Env protein, made of trimeric gp120 and gp41 subunits, without its glycan sheild. The CD4 binding site (epitope of CH103 and VRC01 antibodies) and V1V2 loop (epitope of VRC26 antibody) are labeled. b) shows *Env* with its glycan shield intact. c) shows a gp41 'stump' that is frequently left on the viral surface after degradation of gp120, and is often the target of non-neutralizing antibodies. See more detailed representation in [McCoy and Weiss, 2013, Burton and Hangartner, 2016] 21

1.6 Maximum likelihood phylogenetic tree of the VRC01 bnAb lineage sampled at ten time points over 15 years of diversification within a single individual infected with HIV-1. Each tip represents a BCR heavy chain sequence; terminal branches are coloured by time point of sampling (see key). The red circle at the root represents the germline sequence (IGHV1-2*02 and IGHJ1*01, D region left unassigned). Note the general, but not complete, trend of increasing genetic divergence from the root with sampling time. Late-sampled sequences near the root indicate very high rate heterogeneity among lineages; these sequences might represent inactive memory B cells. BnAb sequences were obtained through cell sorting [Wu et al., 2010] followed by high-throughput sequencing data to identify related BCRs [Zhou et al., 2013]. See Wu et al. [Wu et al., 2015] for full experimental details. Sequences for this tree were obtained from GenBank [Wu et al., 2015] and aligned using MUSCLE [Edgar, 2004]. A maximum likelihood phylogeny was estimated using the GTRGAMMA substitution model in RAxML [Stamatakis, 2014], and re-rooted to position the germline sequence at the root with a divergence of zero. Scale bar represents genetic distance (expected substitutions per nucleotide site). 23

1.7 Example phylogenetic tree used in Eq 1.4 and 1.5. It is rooted at node R. A, B, and C are external nodes representing sequences. . . . 26

2.1 Network visualisation of the diversity of BCR sequences obtained from untreated patient 3 at week 4 (part a) and week 108 (part b), and from treated patient 4 at week 4 (part c) and week 108 (part d). Each vertex represents a unique sequence, and the size of each vertex is proportional to the number of reads identical to that sequence. Edges are drawn between vertices that are one nucleotide substitution apart. See Methods section 2.2.6 for further details. . . 47

2.2 BCR diversity statistics and viral load values for patients 1 - 4 (parts a to d) over the study period of 108 weeks. Patients 1 - 3 were untreated, whilst patient 4 received ART until week 48. Four plots are provided in each panel. The top plot shows viral load, with each point representing a clinical sample. The second plot shows vertex Gini index values of the BCR sequences obtained at each time point. The third plot shows the proportion of reads at each time point that belong to 'large' clones, i.e. those that occupy >0.1% of reads at any time point. The bottom plot in each panel shows the proportion of all reads at each time point that are occupied by the 20 largest clones observed across all time points. Each of the 20 largest clones is represented by a bar of a different colour, and lines connect bars at adjacent time points that represent the same clone. 48

2.3 BCR diversity statistics and viral load values for patients 5 - 8 (parts a to d) over the study period of 108 weeks. All patients received ART until week 48. See Figure 2.2 legend for details. 49

2.4 Comparison of three BCR summary statistics among healthy uninfected individuals (blue) and HIV-infected patients. Values from HIV+ patients are calculated separately for weeks 4-24 (HIV+ early; green) and weeks 52-108 (HIV+ late; black). For each period statistics were averaged using linear interpolation (see Methods section 2.2.9 for details). Each box and solid line represents the first, third, and second quartiles, respectively, whilst whiskers indicate range. Data points are superimposed on the box and whisker plots. 51

3.1 Proposed evolutionary model for antibody lineages. All sequences descend from given germline node G, which has sequence g. Arrows indicate the direction of evolutionary change. Note that this known ancestor G is not necessarily the most recent common ancestor of the lineage, which is node R and sequence x. See similarities to tree in Boussau and Gouy [Boussau and Gouy, 2006]. 65

3.2 Proportional error in parameter estimation compared to true values for the VRC01 B cell lineage fully context dependent simulations. Values of ω, κ, tree length, and ratio of internal to external branch lengths are shown in panels a), b), c), and d), respectively. Estimates obtained under the NY98 are in orange (h=0) and estimates obtained under the HLP17 model are in blue (h estimated using maximum likelihood). The edges and centres of boxplots show the 1st, 2nd, and 3rd quartiles, while the whiskers show range. Similar results for B cell lineages CH103 and VRC26 are shown in Appendix Figure B.4-B.5. 75

4.1 Graphical representation of empirical antibody motif and amino acid models. a) S5F mutability estimates for the central nucleotide in all possible pentamers. Pentamers are grouped by the type of canonical motif (if any) that they fall under. The canonical motifs listed mutate at the nucleotide in boldface. Circles are also coloured by mutability. b) WAG amino acid exchangeability matrix, showing relative instantaneous substitution rates for each amino acid (from blue to red, lowest to highest). c) Similarly, the AB amino acid exchangeability matrix (see Section 4.1). 92

4.2 Multiple sequence alignment process for bnAb lineages. V and J segments are aligned by combining their pairwise alignments to their reference sequence (here, gaps inserted at position indicated by purple **X**). CDR3s are aligned using MUSCLE. 99

4.3 Validation of the WAG exchangeability matrix. The y axis shows the maximum likelihood (ℓ) obtained from a particular WAG:S5F model, while the x axis shows the ML estimation of m_o or y_o. For comparison, the ℓ of the discrete NY98:S5F model is shown in the horizontal dotted line, and the ML estimate of ω in the NY98:S5F model is shown by the vertical dotted line. 104

4.4 ML estimates of m_o and y_o using WAG:S5F models with permuted WAG exchangeability matrices. For comparison, the ML estimate of ω for NY98:S5F is shown by the vertical dotted line. 106

4.5 AIC of model combinations for each lineage. Each subplot shows all model combinations for each lineage, with amino acid models on the x axis and mutability models on the y axis. Cells display the AIC for each model combination, and are shaded by AIC value as well from red to blue, with blue being the lowest AIC and best model fit. See Table 4.1 for details on each model. 108

4.6 ML topologies of the CH103 lineage estimated under models NY98:H0, and WAG:S5F, respectively. External branches are coloured by their sample date (from 53 weeks to 144 weeks post-infection, green to purple, respectively). Monoclonal antibody sequences are labelled, as is the germline sequence (V4-59:J4). Branch lengths are in units of substitutions per site. 111

4.7 ML topologies of the VRC26 lineage estimated under models NY98:H0, and WAG:FCH, respectively. External branches are coloured by their sample date (from 38 weeks to 206 weeks post-infection, green to purple, respectively). Monoclonal antibody sequences are labelled, as is the germline sequence (V3-30:J3). Branch lengths are in units of substitutions per site. 112

4.8 ML topologies of the VRC01 lineage estimated under models NY98:H0, and WAG:FCH, respectively. External branches are coloured by their sample date (from 1995 to 2009, green to purple, respectively). Monoclonal antibody sequences are labelled, as is the germline sequence (V1-2:J1). Branch lengths are in units of substitutions per site. 113

5.1 Process for creating a composite tree summarizing results of the clade search analysis. All clades with at least two descendant sequences are enumerated (left side). For each, I estimate a separate value of ω_{CDR} for blue vs black branches. To the right of each tree is the ML estimate of ω_{CDR} for the blue branches and the p value for the test. Because in this example I only fit one additional ω_{CDR} on a tree with four clades with >1 taxa, I determine significance using a Bonferroni corrected p value cut-off of $0.05/4 = 0.0125$. The right side shows how the composite tree is initialized with the ω_{CDR} from the $2:\omega$ model, and then 'overwritten' by clades that show a significantly better fit under the $3:\omega$ model than the $2:\omega$ model. . 135

5.2 Profile likelihood curves for the CH103 and VRC26. One curve in each lineage shows the estimates of ω under the $2:\omega$ model, with the LR computed by comparing the ℓ values of the $2:\omega$ model to those under the $1:\omega$ model. Other curves show clades from the $3:\omega$ model, with their LRs computed by comparing the ℓ values from the $3:\omega$ model to those under the $2:\omega$ model. 138

5.3 Composite trees showing significant differences in ω_{CDR} among clades in the a) CH103, b) VRC26, and c) VRC01 lineages. Branches are coloured by ω_{CDR} value, from cyan (ω_{CDR}=0) to red (ω_{CDR}=1.1). Monoclonal antibody sequences are labelled (Table 5.1). Branch lengths are in units of time specified on the scale bars below each tree. See Methods and Figure 5.1. 140

5.4 Composite trees showing significant differences in ω_{FWR} among clades in the a) CH103, b) VRC26, and c) VRC01 lineages. Branches are coloured by ω_{FWR} value, from cyan (ω_{FWR}=0) to red (ω_{FWR}=1.1). Monoclonal antibody sequences are labelled (Table 5.1). Branch lengths are in units of time specified on the scale bars below each tree. See Methods and Figure 5.1. 141

5.5 Composite trees showing significant differences in ω_{CDR} among clades in the a) CH103, b) VRC26, and c) VRC01 lineages. Branches are coloured by ω_{CDR} value, from cyan (ω_{CDR}=0) to red (ω_{CDR}=1.1). Monoclonal antibody sequences are labelled (Table 5.1). Branch lengths are in units of genetic distance (substitutions per site) specified on the scale bars below each tree. See Methods and Figure 5.1. 142

5.6 Composite trees showing significant differences in ω_{FWR} among clades in the a) CH103, b) VRC26, and c) VRC01 lineages. Branches are coloured by ω_{FWR} value, from cyan (ω_{FWR}=0) to red (ω_{FWR}=1.1). Monoclonal antibody sequences are labelled (Table 5.1). Branch lengths are in units of genetic distance (substitutions per site) specified on the scale bars below each tree. See Methods and Figure 5.1. 143

5.7 Distribution LRS statistics for 20 simulated alignments of 50 sequences under the null hypothesis of no variation in ω_{CDR} among branches ($2 : \omega$). Density estimate of LRS values is shown in grey, while the chi-squared probability density function with one degree of freedom is shown in red. 148

5.8 Sensitivity and specificity of different p value correction methods explored using simulations. Two groups of 20 alignments were simulated. The first used tree a), in which all black branches used a ω_{CDR} = 0.78, while all blue branches used ω_{CDR} = 0.1; the second used tree c), with all branches using ω_{CDR} = 0.78. In panels b) and d), each clade within a tree is coloured based on the number of times it is identified as having a significantly different ω_{CDR} from the rest of the tree (scaled from black to blue, 0-20, respectively). Trees in b) and d) show how different p value corrections affect the results (see 4.3.5 above). Sensitivity (true positive rate) is the proportion of constrained internal branches correctly identified (100 in a/b, 0 in c/d), while specificity (true negative rate) is the proportion of all non-constrained branches correctly identified across all 20 alignments for each simulation type (a/b vs c/d). 149

5.9 Proposed fitness landscape showing positive relationship between fitness, neutralization breadth, and purifying selection. Fitness valleys (red) are associated with low breadth, high substitution rate, and diversifying selection, while peaks (blue) are associated with wider breadth, lower substitution rate, and purifying selection. Dotted lines are clades from an antibody lineage originating from the 'V(D)J' spot in the top left. Arrows show each clade's location at the time of sampling. This example is consistent with the CH103 lineage at week 92 in Figure 5.3a: one clade has ascended a local peak and shifted to low ω_{CDR}, while the others, including the putative bnAb clade, have not yet ascended any peaks and show no signal of lowered ω_{CDR}. Map outline freely available from blog.spoongraphics.co.uk.153

6.1 Genetic distance (proportion of junctional nucleotide differences) of each sequence to its nearest neighbour (most similar) sequence. The lower and upper modes of this distribution are generally interpreted as within and between clone comparisons, respectively. In these figures all sequence comparisons to the left of each vertical red line are grouped together as a clone. 163

6.2 Proportion of reads in clonal lineages, binned by mean pairwise difference (MPD) of the sequences in the lineage on the y axis, and by the number of timepoints in which the lineage was detected on the x axis. Note how the majority of reads in all three patients are found in lineages found only in one timepoint, and at low genetic diversity (MPD < 0.1) 166

6.3 ML Estimates of base frequencies under $r : 1 : \omega$ and $r : n : \omega$ models are consistent between patients. These 12 parameters (four bases at 3 codon positions) are used to calculate $\boldsymbol{\pi}$ values used in $\mathbf{Q}$ matrix. Note that these estimates are consistent between patients, but different from their empirically calculated base frequencies ('Emp. FC3X4'). These models were applied to the 10NUS datasets of each patient. 168

6.4 ML Estimates of κ and S5F scaling parameters (y_h and m_h) under $r : 1 : \omega$ and $r : n : \omega$ models. κ and y_h are fairly consistent between patients, but not m_h. These models were applied to the 10NUS datasets of each patient. 169

6.5 Histograms of read counts contained within lineages with different values of ω estimated under the $r : n : \omega$ model. Each read is represented by the ω value estimated for the lineage containing it under the $r : n : \omega$ model. The count of reads for each ω are shown in the above histograms. The ML estimate of ω under the $r : 1 : \omega$ model for each patient is represented by a vertical dotted line. Note that PS133, PS517, and PS021 had 3, 3, and 1 lineages with an estimate of $\omega > 2$ that are not shown in this plot. 172

6.6 Estimate of ω for individual lineages under the $r : n : \omega$ model for each patient. Each circle represents an individual lineage. The size of each circle corresponds to the lineage's size, while the colour corresponds to its genetic diversity (dark blue = low MPD, light blue = high MPD). The ML estimate of ω under the $r : 1 : \omega$ model for each patient is represented by a vertical dotted line. The horizontal line corresponds to an LR of 1.9, which corresponds to an LRS of 3.8, which corresponds to a p value < 0.05 for a likelihood ratio test (Eq. 1.8). However, the LR values computed for each lineage of the repertoire here cannot be compared in this way, so this is only shown as a reference. Note that PS133, PS517, and PS021 had 3, 3, and 1 lineages with an estimate of $\omega > 2$ that are not shown in this plot. 173

6.7 Eight largest lineages of PS133, coloured by ω_{CDR} estimate (blue=0, red=1.1) 174

6.8 Eight largest lineages of PS517, coloured by ω_{CDR} estimate (blue=0, red=1.1) 175

6.9 Eight largest lineages of PS021, coloured by ω_{CDR} estimate (blue=0, red=1.1) 176

A.1 Human B-cell receptor PCR primers used in Chapter 2 190

A.2 (summarized from main text) MiSeq reads were filtered for base quality (median >32) using QUASR (http://sourceforge.net/projects/quasr) [Watson et al., 2013]. Sequences were concatenated and a gap inserted between the forward and reverse reads (average gap length ~35 nucleotides. Non-immunoglobulin sequences were removed; only those reads with significant similarity to reference IgHV and IgHJ genes from the ImMunoGeneTics (IMGT) database [Lefranc et al., 2009] were retained, as determined using BLAST [Altschul et al., 1990] with E-value thresholds of $1x10^{-10}$ and $1x10^{-3}$, respectively. Primer sequences were then trimmed from the reads, and sequences retained for analysis only if both primer sequences were identified and if both forward and reverse sequence lengths were >100 bp. Lastly, a second BLAST analysis was used to identify the IgHV genes of retained reads. Specifically, BLAST [Altschul et al., 1990] was used to align reads against known BCR sequences from the IMGT database [Lefranc et al., 2009] (E-value thresholds for IgHV and IgHJ-genes were 10^{-70} and 10^{-20} respectively, due to their different gene lengths). The combined per-base error rate for the RT-PCR and sequencing process of the MiSeq platform was $2.06x10^{-4}$ [Bashford-Rogers et al., 2013, Bashford-Rogers et al., 2014]. 191

A.3 (summarized from main text) The first read of each pair was aligned with the closest matching sequence in the V references through gapless pairwise alignment. To further optimize the alignment each read was matched using only the 5′ half of the V segments. Next, the second read of each pair was pairwise aligned with the closest sequence in the J references, again using gapless alignment. Once the first and second reads had been aligned to the V and J references, respectively, gaps were added 3′ of the first read, and 5′ of the second read, such that each were 300 nucleotides in length. The two reads were then concatenated, and excess gaps were trimmed between each read pair. Because we expect sequences from the same clone to match the same, or highly similar, V and J segments, this process generates an approximation of the optimal gapless alignment for sequences from the same clone. 192

A.4 Each point represents a sampling time point and is coloured by patient (see key). The fitted linear regression has a slope of $5.2x10^{-3}$ (p = $5.69x10^{-5}$) and a correlation coefficient of 0.55. This relationship remained significant when patient means were used to correct for temporal autocorrelation ($p < 0.03$). 193

A.5 I model the effect of read depth on Gini index using a simple population of B cells consisting of a single dominant clone occupying one third of all B cells and a large pool of unique B cells, which comprise the rest of the system. As sequencing depth increases, one third of each sample joins the dominant clone, which grows larger. However, all other clones remain unique, leading to a larger population of singleton clones. This means the ratio of reads in the dominant clone to each of the singletons grows with sampling depth, leading to a more skewed size distribution and higher Gini index. Specific values are provided in the table above. 194

A.6 CD4+ cell counts (cell/μL) for each patient and available timepoint. 195

A.7 BCR statistics of HIV+ patients calculated from full data versus data sub-sampled data. Read depth of full data had a median of 567,936 and a range of 74,861 to 1,040,936 reads per patient per time point. Sub-sampled data were randomly sub- sampled to a depth of 70,000 reads per patient per time point. Dots are coloured by patient. . . . 196

B.1 Evolutionary model reproduced from 3.1 197

B.2 Distribution of pairwise codon differences from each tip sequences to the germline sequence of each lineage. 202

B.3 Distribution of h values for all possible motifs fulfilling the following criteria: a) trimers which contained two IUPAC letters standing for two possible nucleotides (R, Y, S, W, K, and M), which were followed by an unambiguous nucleotide (A, C, G, or T), and b) dimer which contained one IUPAC letters standing for two possible nucleotides, which was followed by an unambiguous nucleotide. We then fit symmetric HLP17 models using each of these motifs and their reverse complements, with all other h parameters set to zero. 204

B.4 Proportional error in parameter estimation compared to true values for the CH103 B cell lineage fully context dependent simulations. Values of ω, κ, tree length, and ratio of internal to external branch lengths are shown in panels a), b), c), and d), respectively. Estimates obtained under the NY98 are in orange (h=0) and estimates obtained under the HLP17 model are in blue (h estimated using maximum likelihood). The edges and centres of boxplots show the 1st, 2nd, and 3rd quartiles, while the whiskers show range. 205

B.5 Proportional error in parameter estimation compared to true values for the VRC26 B cell lineage fully context dependent simulations. Values of ω, κ, tree length, and ratio of internal to external branch lengths are shown in panels a), b), c), and d), respectively. Estimates obtained under the NY98 are in orange (h=0) and estimates obtained under the HLP17 model are in blue (h estimated using maximum likelihood). The edges and centres of boxplots show the 1st, 2nd, and 3rd quartiles, while the whiskers show range. 206

B.6 Proportion of fully context-dependent simulated data sets that incorrectly reject the true symmetric WRC/GYW motif model. Note that this generally increases as both h and ω increase. 207

B.7 Bias in h from data sets simulated using different values of h and ω under a fully context-dependent symmetric WRC/GYW model. Note that in the asymmetric model h^{WRC} and h^{GYW} are biased in opposite directions as h and ω increase, while in the symmetric model h^{WRC} shows no such consistent bias. Note that these values are absolute bias, not proportional error. This would be obtained by dividing each value in each cell by the true h used in simulation (left axis). a) Bias in the value of h^{WRC} estimated from fitting an asymmetric WRC/GYW model. b) Bias in the value of h^{GYW} estimated from fitting an asymmetric WRC/GYW model. c) Bias in the value of $h^{WRC/GYW}$ estimated from fitting a symmetric WRC/GYW model. 207

B.8 Bias in ω from data sets simulated using different values of h and ω under a fully context-dependent symmetric WRC/GYW model. Note that, while a trend exists in the asymmetric model, the bias is ultimately very small. a) Bias in the value of ω estimated from fitting an asymmetric WRC/GYW model. b) Bias in the value of ω estimated from fitting a symmetric WRC/GYW model. 208

B.9 Proportion of fully context-dependent simulated data sets simulated under the symmetric WRC/GYW model that incorrectly reject the true values of h and ω. 208

B.10 Cumulative distribution of b^a_{ij} parameters among synonymous vs non-synonymous substitutions in the **Q** matrix formed from different motifs individually using the ML codon frequency results from CH103. Results are nearly identical in the other two lineages. . . . 209

B.11 Cumulative distribution of b^a_{ij} parameters among synonymous vs non-synonymous substitutions in the **Q** matrix formed from different motifs together and symmetrically using the ML codon frequency results from CH103. Results are nearly identical in the other two lineages. 210

B.12 Differences between predicted ancestors for the best fit HLP17 model (Supplemental Figure 3) and the NY98 model for three bnAb lineages by cladistics degree (nodes from root sequence). Solid line shows the mean difference for each node, dashed lines show the minimum and maximum values for each degree. 212

Chapter 1

Introduction

1.1 Overview

The adaptive immune system ensures the survival of humans and other vertebrates in the face of rapidly evolving and genetically diverse infectious diseases. B lymphocytes are an essential component of this system and express receptors on their cell surface (B cell receptors; BCRs) capable of specifically binding foreign antigens. BCRs are membrane-bound immunoglobulins composed of two large heavy chain molecules and two smaller light chain molecules, encoded in humans by the genes IGH and IGL (or IGK) respectively. The diversity of BCRs expressed by an individual's B cells is vast, and comprises both naive receptors that are randomly generated from the germline during development, as well as receptors that are retained after successfully binding antigen during previous infections. Populations of BCRs can rapidly improve antigen binding during infection through an evolutionary process of mutation and selection known as affinity maturation [Eisen and Siskind, 1964]. Because BCR sequences are diverse and diverge rapidly, concepts from molecular evolution should be beneficial in understanding the dynamics of the adaptive immune system within individuals.

Until recently, there was a paucity of data on within-individual BCR sequence diversity for researchers to explore. That situation has now changed with the application of next generation sequencing to BCRs [Boyd et al., 2009, Six et al., 2013]. With this technique, researchers can directly observe the somatic genetic changes that generate the diversity of the BCR repertoire, providing an unprecedented picture of the adaptive immune system as an evolving population of cells. Analyses of these data from an evolutionary perspective have led to insights into the aging of the B-cell repertoire [Wang et al., 2014b] and into the process of affinity maturation [Elhanati et al., 2015, Yaari et al., 2015], and have many applications across a broad range of diseases (see Table 1.1).

Despite these advances, there are important challenges in applying models and methods from molecular evolution to BCR sequences, which stem both from the complex biology of B cells and the nature of available data. Unlike most natural populations, naive BCR sequences do not descend from a single common ancestor via a process of point mutation, but are instead generated from a diverse set of germline gene segments via a process of somatic recombination (see below). In addition, the mutation process during affinity maturation is strongly dependent on the sequence context of flanking nucleotides [Yaari et al., 2013] and selection on the resulting amino acid sequence is complex and site-specific, driven by the need to avoid dangerous self-reactivity while concurrently enhancing pathogen binding. Finally, BCRs are a complex of heavy and light chain immunoglobulin molecules, and information from both is necessary for a complete understanding of BCR evolution and function.

Application	Diseases/ infections	Significance	Challenges
Identification of broadly neutralizing antibodies (bnAbs)	Infection with rapidly evolving pathogens such as HIV, Hepatitis C virus, influenza viruses	– Potential for use as vaccine targets (Haynes et al. 2012). – Provide model system for understanding affinity maturation, co-evolution and immune development (e.g. Wu et al. 2015).	– Lineages are often large and diverse, often with high levels of hypermutation, long CDR3s, and auto/poly reactivity (Haynes et al. 2012). – Lineages may have complex dynamics of mutation and selection (Sheng et al. 2016; Chapters 4 and 5).
Study of vaccine responses	Any disease for which vaccines are used or being developed, e.g. Influenza, typhoid, or Ebola	– A model system for immune response with a known stimulus (Galson et al. 2014). – Identifying correlates of immune protection following vaccination.	– Distinguishing vaccine-specific changes to the BCR repertoire from healthy repertoire diversity. – Variable immune responses among individuals to the same stimulus (Ademokun et al. 2011; Jiang et al. 2013). – Vaccine responses may be primed by previous exposure to other antigens (Williams et al. 2015).
Tracking B cell migration and development within the body	Autoimmune diseases such as multiple sclerosis; cancers.	– Identifies migration of B cells between tissue compartments – Can identify the sites at which B cells mature (Stern et al. 2014)	– Modeling potentially complex migration patterns. – Differential sampling between tissues. – Difficult to have serial sampling in multiple tissues from humans.
Disease diagnosis	Autoimmune diseases such as multiple sclerosis and rheumatoid arthritis; cancers, in particular B cell lymphoma.	– A direct and potentially cheap diagnosis tool. – Improved understanding of disease. – Provide clinical markers of disease progression (Robinson et al. 2013).	– Complex and multiple disease epitopes induce complex responses. – Low level presence of B cells associated with disease (e.g. Chapter 1).

Table 1.1: Applications of BCR repertoire sequencing. Note that accurate assignment of B cell clonal lineages is an important challenge for all applications listed.

1.2 B cell development

The initial diversity of the BCR repertoire is the result of a somatic recombination process called V(D)J recombination. This process brings together one each of the variable (V), diversity (D) and joining (J) segments of the IGH locus on chromosome 14 to form an exon in the heavy chain immunoglobulin gene, and one each of the V and J segments of the IGL (or IGK) locus to form the light chain. Not all gene segments are utilized. Of the 123-129 IGHV gene segments, 44 contain ORFs; further, 25 of the 27 D segments and 6 of the 9 J segments have been shown to be used for somatic recombination in the heavy chain [Lefranc and Lefranc, 2001, Li, 2004]. During this process, additional sequence diversity is generated by random deletion or insertion of nucleotides at segment junctions. This process combines highly variable sequence regions that determine antigen binding (the complementarity determining regions; CDRs) with more conserved framework regions (FWRs) that provide structural support (illustrated in Figure 1.1). Thus each naive B cell has its own BCR sequence, and the number of possible BCR sequences is huge, with models predicting at least 10^{18} sequences [Elhanati et al., 2015], far greater than the number of B cells in the body. The process may generate non-productive (e.g. out-of-frame) coding sequences. When this happens, the B cell may recombine its second copy of the IGH gene. If this too fails to produce a viable recombinant sequence then the cell undergoes apoptosis, which further modulates the background genetic diversity of receptors (Figure 1.2). Immature B cells within the bone marrow are tested for binding to self-antigen, and those that are autoreactive are eliminated through clonal deletion [Owen et al., 2013]. These cells then undergo two more rounds of selection in the spleen. More specifically, they are once again negatively selected against binding for self-antigen reactivity, and then move to the splenic follicles where they are positively selected for functional receptor expression [Owen et al., 2013]. Cells which pass these checkpoints are released and circulate through the peripheral blood and lymphoid organs.

Once a naive B cell is activated by binding a foreign antigen, it undergoes cell division (clonal expansion) and initiates processes that somatically alter the BCR sequence and diversify the clonal population. In parallel, a mechanism called class switching alters the constant region of the heavy chain, changing the type and function of the BCR and its interaction with other molecules. While this does not affect its antigen-binding properties, it does leave a molecular mark that can be used to separate naive BCRs from those that have undergone affinity maturation. Affinity maturation modifies antigen binding via a process of random and rapid sequence change, termed somatic hypermutation (SHM), and by clonal selection. SHM involves greatly increased mutation rates of ~10^{-3} changes per nucleotide per cell division, corresponding to ~1 mutation per cell division in the relevant locus [Teng and Papavasiliou, 2007, Victora and Nussenzweig, 2012]. Mechanistically, these mutations are induced by the enzyme activation-induced cytodine deanimase (AID), which deaminates cytosine to uracil during transcription [Muramatsu et al., 2000, Teng and Papavasiliou, 2007, Peled et al., 2008]. Importantly, for evolutionary analysis, SHM is a stochastic and strongly non-uniform process, and clearly distinct from the processes of germline mutation and evolution. In particular, SHM is context-dependent such that the probability of mutation at a site is strongly influenced by neighbouring nucleotides [Shapiro et al., 2003, Yaari et al., 2013]. The resulting mutations are further shaped by a round of selection, in which B cells compete for survival and replication signals by competitively binding to antigens [Peled et al., 2008]. The combination of these processes shapes both the type and rate of observed mutations across the IGH and IGL loci.

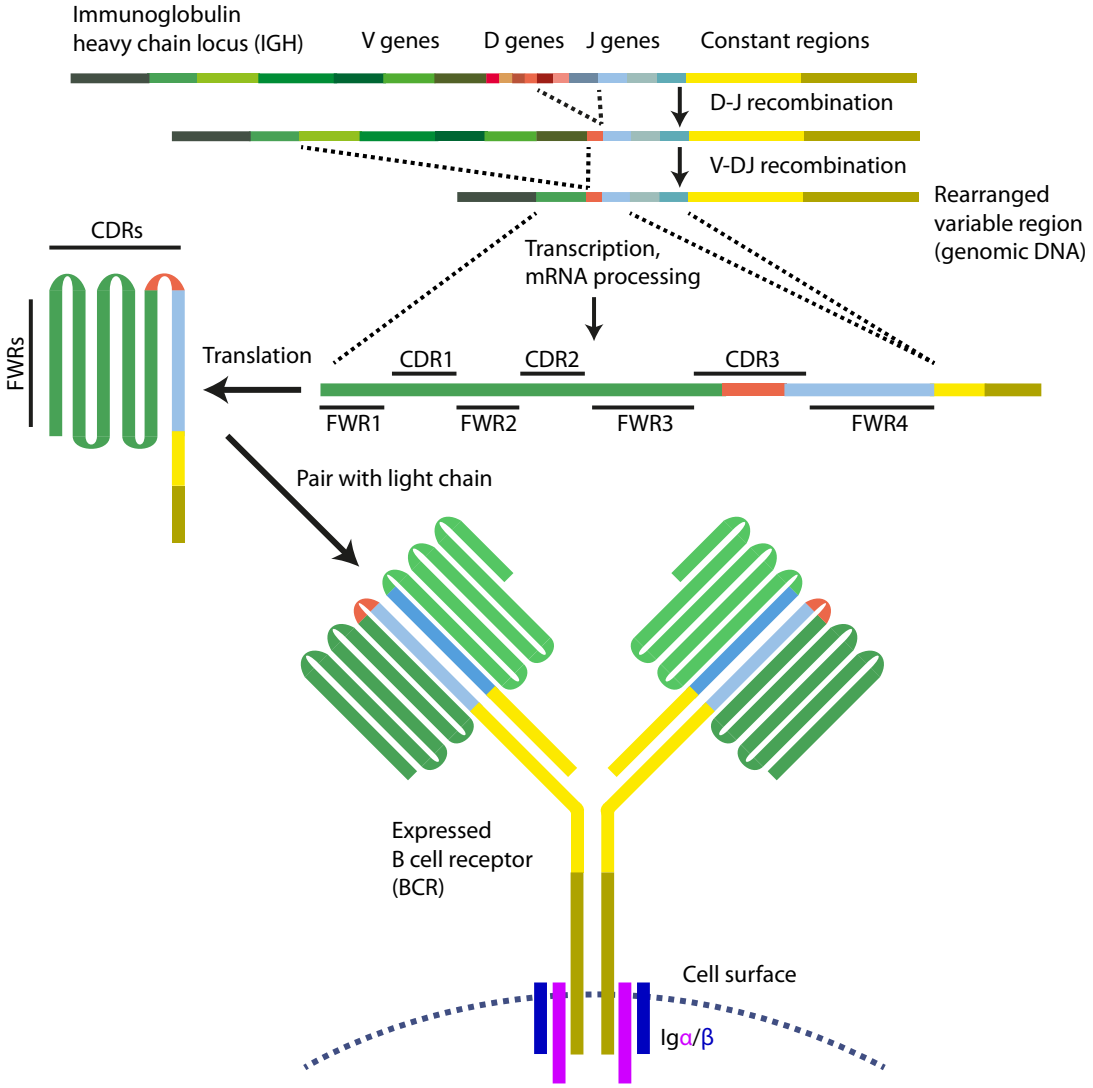


Figure 1.1: Somatic rearrangement and functional division of the immunoglobulin heavy (IGH) chain locus during B cell development. The IGH locus begins with multiple V, D, and J segments. Of these, one of each type is spliced together to form the rearranged variable region. After transcription, the joined VDJ segments are spliced together with a neighbouring C region [Owen et al., 2013]. The rearranged variable region is divided into non-consecutive framework (FWR) and complementary determining (CDR) regions. After translation, the heavy chain protein folds such that FWRs bind together, exposing the CDRs for use in antigen binding. This protein is paired with a light chain, which undergoes a similar somatic recombination process (VJ recombination) and also contains three exposed CDRs [Owen et al., 2013]. The full B cell receptor is displayed on the cell surface, with its contant regions associated with intracellular signalling proteins Igα and Igβ. See [Murphy et al., 2012] Figures 4.6, 5.1, and [Sompayrac, 2016], p. 29.

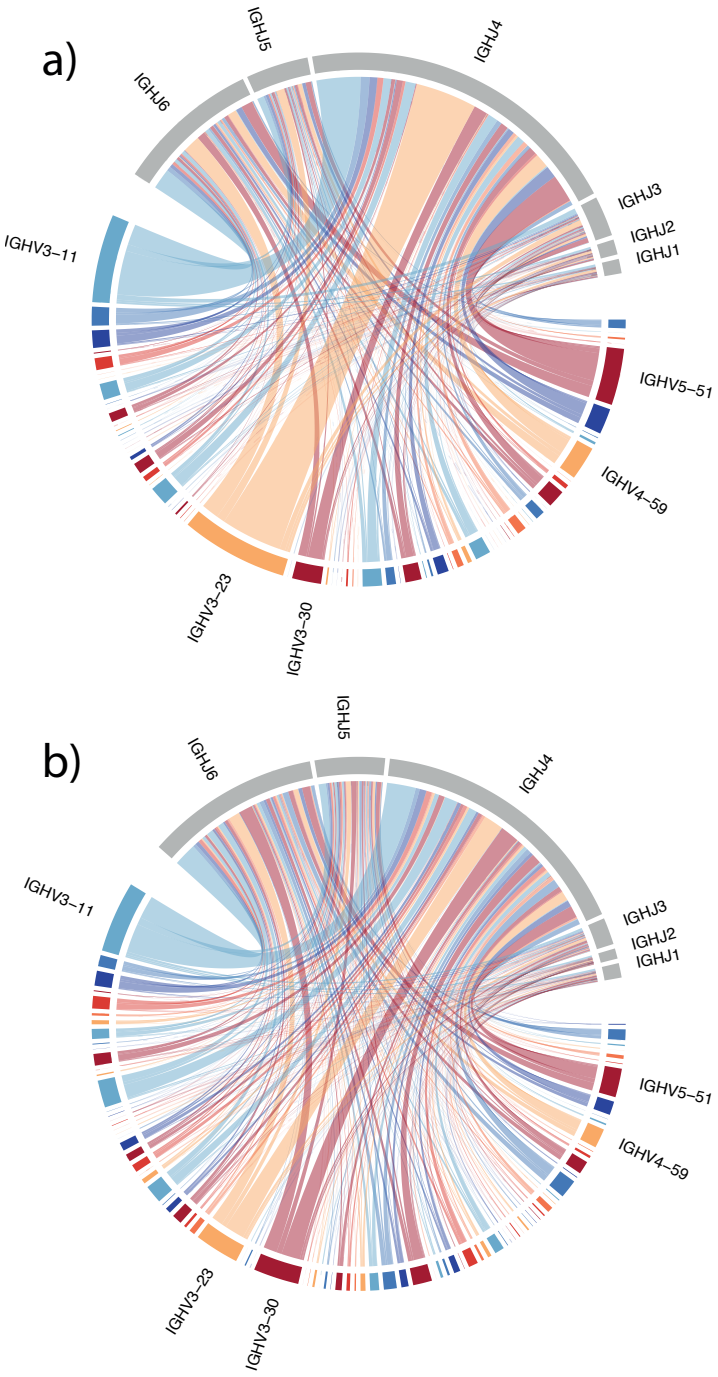


Figure 1.2: Chord diagrams showing the pairing of V and J segments within (a) productive and (b) non-productive IgM sequences from a single healthy individual. Chord widths represent the proportion of sequences with a given V (coloured) and J (grey) segment pairing. The five most common V segments in productive rearrangements (and all J segments) are labelled. Note that IGHV3-23/IGHJ4 was significantly more common in productive vs. non-productive rearrangements, which may indicate functional bias of that pairing. The figure was generated from data in [Elhanati et al., 2015], which was aligned to the IMGT reference ([Lefranc et al., 2009] using IgBLAST [Ye et al., 2013]. Productive rearrangements were subsampled to the same read depth as non-productive rearrangements ($2x10^5$ reads); the values displayed in (a) are means of 100 subsampling repetitions.

1.3 Sequencing the BCR repertoire

Many of the biological processes that drive the B cell response in the adaptive immune system - V(D)J recombination, somatic hypermutation, and clonal selection - leave clear marks in the genomes of B cells. Because of this, it is possible, at least in principle, to study these processes using DNA sequencing of the BCR variable region. Though more recent approaches use high-throughput sequencing technologies to generate massive datasets containing millions of reads (e.g. Chapter 2), BCR sequencing data has been obtained even from low-throughput sequencing technologies [Shlomchik et al., 1987]. Though proportion of the repertoire sampled was lower in these studies, BCR sequencing data was used to study diseases such as B cell cancers [Spencer and Dunn-Walters, 1999, van Dongen et al., 2003], multiple sclerosis [Colombo et al., 2000, Colombo et al., 2003], and autoimmunity in model organisms [Shlomchik et al., 1987]. Low-throughput BCR genetics led to the use of phylogenetic trees to represent the development of B cell clones [Shlomchik et al., 1987] and even the development of B-cell specific phylogenetic methods [Dunn-Walters et al., 2002, Barak et al., 2008].

High- and low-throughput targeted sequencing technologies face similar challenges in dealing with the extraordinary variability of BCR. Rearranged VDJ segments are flanked by introns, so targeting germline DNA requires a cocktail of PCR primers [van Dongen et al., 2003, Larimore et al., 2012]. A challenge for this approach is to control for PCR bias, which could skew the frequency of sequenced variants and obscure the signal of clonal expansion. An alternative approach that can significantly reduce the problem of PCR bias is to target expressed mRNA, in which case the constant regions flanking the VDJ segments in mature mRNA (see Figure 1.1) can be used for PCR priming [Colombo et al., 2000, Galson et al., 2015]. The challenges for mRNA sequencing are to (i) disentangle variation in sequence frequency that is due to differential expression, which can be extensive, from that

due to clonal expansion, and (ii) to ensure that sequencing error and subsequent bioinformatic processing do not introduce systematic biases into subsequent evolutionary analyses. A more detailed discussion of of BCR repertoire sequencing is given in [Benichou et al., 2012].

Sequencing of the somatically altered heavy chain has the potential to reveal the clonal structure and dynamics of the B cell population through time, and this thesis focuses on the analysis of sequence data from this region. However, it is important to note that although the majority of variation in BCR sequences is concentrated in the heavy chain, and in particular the CDRs [Xu and Davis, 2000, Georgiou et al., 2014], the light chain also contains mutations that may affect antigen binding. If one's goal is to characterise entire antibodies, or to understand the binding properties of a given heavy chain sequence, then knowledge of paired heavy and light chain sequences is required. Computational approaches have previously sought to infer how heavy and light chain sequences are paired from independently sequenced sets of sequences by using relative frequencies [Reddy et al., 2010], or the shapes of phylogenetic trees [Zhu et al., 2013] of heavy and light chain sequences. Recently, single cell technologies have enabled natively-paired heavy and light chains to be sequenced by attaching unique barcodes to cDNA from individual cells [Busse et al., 2014]. Alternatively, oligo-dT beads that link heavy and light chains from a single cell have been used [DeKosky et al., 2013].

1.4 Measuring BCR diversity

Once BCR sequences are generated, statistical and computational approaches are necessary to explore and summarise their diversity, in order to reveal associations with immune responses or disease status, or to identify BCR sequences of specific interest. The exceptional diversity of the BCR repertoire, and its dynamic nature, makes comparative study within and among individuals challenging.

Several different measures have been proposed, and can be distinguished into those that characterise raw sequence variability versus those that depend on the frequency of BCR lineages, clones or clusters (i.e. groups of identical or similar sequences; see next section: 1.5). In the context of viral infection, both the number of somatic mutations [Chen et al., 2012, Wang et al., 2014b, Galson et al., 2015] and V, J gene usage [Zhou et al., 2013] have proved useful. CDR3 sequence length also has been used to distinguish repertoires after pneumococcal vaccination [Ademokun et al., 2011, Chen et al., 2012, Galson et al., 2015]. Diversity statistics such as the Gini index or mean clone size are also used to investigate BCR diversity (Chapter 2; [Bashford-Rogers et al., 2013, Galson et al., 2015]). Figure 1.3 provides a graphical representation of BCR diversity under different conditions. These figures were produced using network diagrams that visualise the diversity and clonal structure of BCR sequences - each point represents a unique BCR sequence, and edges are drawn between pairs of sequences that differ by exactly one nucleotide change.

Other approaches seek to characterize BCR diversity using statistical models. Mora et al. [Mora et al., 2010] introduced a maximum entropy model that characterises the repertoire as a statistical distribution, whilst Elhanati et al. [Elhanati et al., 2015] used probabilistic inference to quantify the process of V(D)J recombination and somatic hypermutation. Greiff et al. [Greiff et al., 2015] proposed employing entropy measures developed in ecology research, which unify a range of diversity measures into a single profile.

A common assumption of these approaches is that clonal expansions observed in infected individuals correspond to B-cell responses against the pathogen under study. Given that rapid cell division of B cells is a key part of the adaptive immune response during infection [Victora and Nussenzweig, 2012], this seems like a reasonable assumption. Further, some studies, e.g. [Wang et al., 2013], have shown a

correlation between clonal expansion and infection (in that case with Epstein-Barr virus). However, it is not directly known how much this process affects differences in clone size at the level of the whole repertoire, especially in instances of co-infection with multiple pathogens. Further, some infections may manipulate host immune responses through so-called superantigens (e.g. staphylococcal protein A), which trigger large, non-specific clonal expansions that disrupt antigen-specific affinity maturation [Thammavongsa et al., 2015]. Such phenomena do not in general prevent clonal expansions from being useful indicators of immune dynamics, but require them to be carefully interpreted in the context of the particular host-pathogen system.

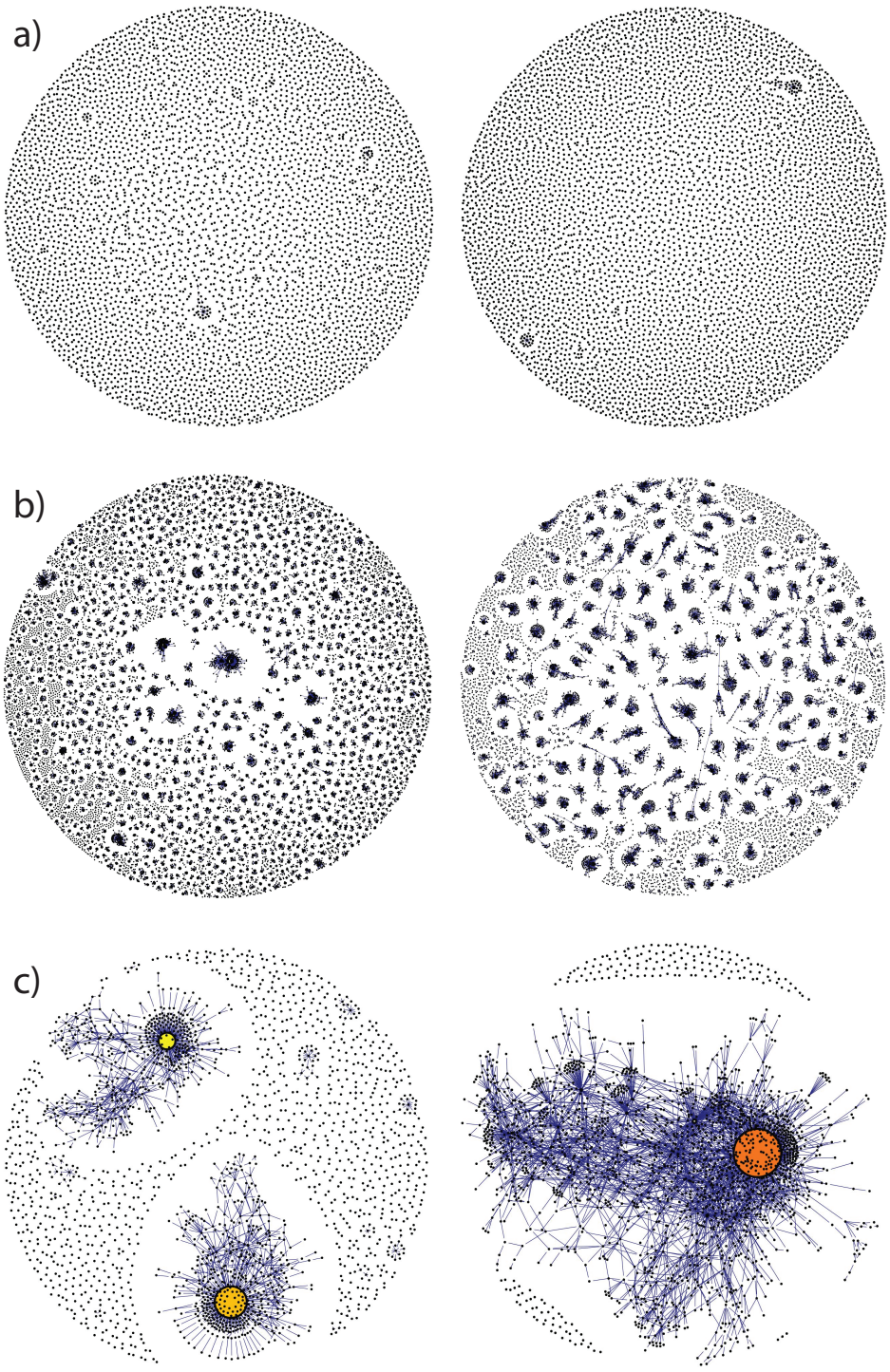


Figure 1.3: Network diagrams that visualise the diversity and clonal structure of BCR sequences. BCR sequences were obtained from (a) two healthy people, (b) two individuals infected with HIV-1, sampled during early infection, and (c) two patients with chronic lymphocytic leukaemia (CLL), a B cell cancer. Each point/circle represents a unique BCR sequence, the size of which is proportional to how common that sequence is. Edges are drawn between pairs of sequences that differ by exactly one nucleotide change. Note that samples do differ by read depth (approximately $3.4x10^4$ and $3.6x10^4$ for part (a), $9.2x10^4$ and $3.6x10^5$ for part (b), $5.1x10^4$ and $2.6x10^4$ for part (c). Parts (a) and (c) are reproduced from Bashford-Rogers et al. [Bashford-Rogers et al., 2013] and parts (b) from Chapter 2.

1.5 Clonal lineage assignment and clustering

Molecular phylogenetics is an undeniably powerful tool for analysis of sequence diversity. However, its application to BCR repertoires is impeded by the V(D)J recombination process, the existence of which means that not all BCR sequence differences are due to point mutation through descent from a common ancestor. Consequently, sequences must be grouped by lineage, each representing sequences that descend from a single ancestral B cell, before they can be analysed phylogenetically [Hershberg and Luning Prak, 2015].

A key step in this process is the alignment of BCR sequences to reference datasets of V, D, and J gene segments, in order to determine their germline origin. Several such alignment methods are available: IMGT/V-Quest [Brochet et al., 2008] is popular and provides a well curated reference dataset; IgBLAST [Ye et al., 2013] can be run with a user-specified reference dataset, and IgSCUEAL uses phylogenetic relationships between germline genes to increase the accuracy of assignment [Frost et al., 2015]. Other tools include iHMM-Align [Gaeta et al., 2007], which implements a hidden Markov model. Other techniques, using methods adopted from phylogenetic ancestral state reconstruction, assign V(D)J segments while also quantifying uncertainty in assignment [Kepler, 2013]. However, germline V, D and J segments vary considerably among individuals and new alleles are still being discovered, so the reference dataset may be inaccurate [Gadala-Maria et al., 2015]. Segment similarity, junctional diversity, somatic hypermutation, and sequencing errors all further increase the difficulty of unambiguously assigning BCR sequences to specific germline segments. This is particularly true for D segments, due to their short length (11 to 37 nucleotides; [Lefranc and Lefranc, 2001] and the frequent occurrence of deletions during V(D)J recombination, which may remove part or all of a D segment [Elhanati et al., 2015]. Chapter 2 of this thesis is a high-throughput repertoire analysis that avoids the issue of D gene assignment

by clustering sequences into clones if they differ by at most one nucleotide across their entire length.

Several studies sidestep the problem of germline D gene assignment, and instead use clustering approaches to group similar BCR sequences by using either the entire V(D)J region sequence (Figure 1.3; [Bashford-Rogers et al., 2013]) or the CDR3 region [Jiang et al., 2013, Sok et al., 2013, Laserson et al., 2014]. A threshold number of differences (edit-distance) is often used to determine if sequences belong to the same or different clusters. One difficulty with this approach is the choice of threshold. Some studies address this by exploring multiple thresholds; edit-distances of 3 to 5 differences have been chosen by looking at how cluster numbers and sizes change as the threshold is increased [Yaari et al., 2013, Laserson et al., 2014]. However, a more principled approach is clearly needed to test how closely these clustering techniques reconstruct the true clonal structure of the B cell population. At present they appear well suited for the detailed analysis of recently diverged clones [Laserson et al., 2014], and for quantifying the diversity of the BCR repertoire in general [Bashford-Rogers et al., 2013, Trck et al., 2015]. However, it is unlikely that clustering approaches based on edit-distances will be effective in accurately identifying large, diverse lineages. For example, broadly neutralizing HIV lineages often show high levels of genetic diversity (Figure 1.6; Chapter 3; [Wu et al., 2015]), and the intermediate (i.e. ancestral) sequences necessary for accurate clustering may not be available in many cases, because affinity maturation occurs in the germinal centres rather than in peripheral blood [Parham, 2009]. Studies that have successfully isolated large and diverse B-cell lineages from HIV-infected patients have generally done so by combining sequence analysis with detailed experimental work [Zhou et al., 2013].

1.6 Untangling mutation and selection

The enzyme-driven nature of somatic hypermutation (SHM) poses a challenge for studying the molecular evolution of BCRs. Standard nucleotide substitution models typically assume that sites (either nucleotides or codons) evolve independently [Felsenstein, 1981]. However, SHM is strongly context dependent, to the extent that observed mutation rates vary more than tenfold across sites (See Figure 1.4) [Elhanati et al., 2015]. Consequently, traditional methods for identifying positive and negative selection that rely on uniform-rate independent-site models can generate false positives when applied to BCR sequences, e.g. within non-productive (out-of-frame) sequences that are not subject to selection [Dunn-Walters and Spencer, 1998].

Models of SHM based on empirical data have been developed and include di-, tri-, penta-, and hepta-nucleotide models [Smith et al., 1996, Shapiro et al., 2002, Shapiro et al., 2003, Yaari et al., 2013, Elhanati et al., 2015], and have been used to investigate selection on the naive B cell repertoire [Elhanati et al., 2015]. Selection has been explored using the 'focused' binomial test, which determines whether the observed number of replacement mutations is significantly different from that expected under a null model of biased mutation but no selection [Hershberg et al., 2008]. This framework has subsequently been extended using Bayesian inference (BASELINe; [Yaari et al., 2012]). In common with other components of the acquired immune system (e.g. class I and II MHC glycoproteins; [Yang and Swanson, 2002, Furlong and Yang, 2008]), analyses indicate that BCR sequences are a mosaic of regions under a mixture of positive and purifying selection (i.e. CDRs) and structural regions whose evolution is highly constrained by purifying selection (i.e. FWRs) [Yaari et al., 2012, McCoy et al., 2015, Yaari and Kleinstein, 2015].

It should also be possible to detect the action of antigen-driven selection from

the shape of BCR lineage phylogenies, which represent the common ancestry of a sample of sequences from a lineage of clonally related B cells [Dunn-Walters et al., 2002]. Computer simulation of lineage trees generated by affinity maturation under a variety of scenarios found seven measures of tree shape that correlated strongly with immunological parameters [Shahaf et al., 2008]. However, recent analyses using these measures concluded that they are affected by experimental factors that are difficult to control, such as the number of sequences sampled from a lineage and the number of cell divisions since initial VDJ rearrangement [Uduman et al., 2014]. Utilizing lineage information, such as excluding terminal branch mutations, has been shown to increase the sensitivity of methods based on the expected number of replacement mutations [Uduman et al., 2014]. It is interesting to note that very similar approaches were developed independently in viral phylogenetics, specifically in studies of HIV-1 and influenza populations under strong positive selection (e.g. [Bush et al., 1999, Lemey et al., 2007]).

Recently, two further approaches to analysing B cell selection have been developed. Kepler et al. [Kepler et al., 2014b] used a statistical model of selection and an empirical model of sequence mutability to study their interplay along the BCR sequences of an antibody lineage. Alternatively, one can adjust and control for the motif-targeted nature of SHM by studying and comparing productive and non-productive BCR rearrangements within a given dataset (see section 1.2 above; [Larimore et al., 2012, Elhanati et al., 2015, McCoy et al., 2015]. McCoy et al. [McCoy et al., 2015] combined this information with a statistical model of trait evolution [Lemey et al., 2012] in order to derive a per-residue map of natural selection along the BCR.

Although antigen-driven positive selection is of great interest, of equal importance to the evolution of antigen-specific BCR sequences is the influence of purifying selection, which results from the removal of self-reactive and non-productive recep-

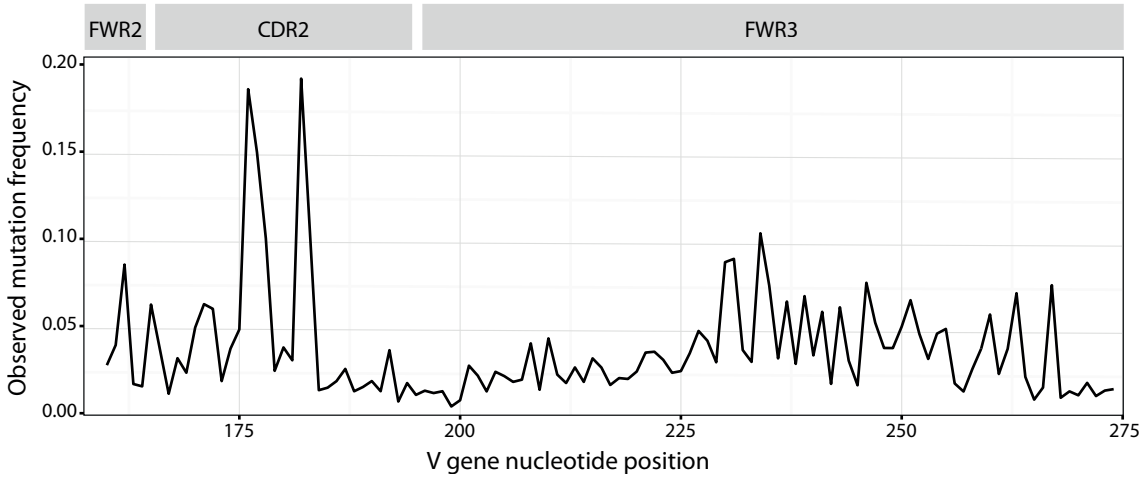


Figure 1.4: Observed mutation (sequence difference from germline) frequency among productive heavy chain immunoglobulin sequences across the V-gene sequence (horizontal axis, IMGT unique numbering). The distribution of mutations across the region is strongly non-uniform, with mutations more likely to occur at certain positions. The CDR2 region (middle shaded box) has a high rate of observed mutations and is thought to be more important in antigen binding than the surrounding framework regions (FWR2 and FWR3). This figure was generated from the same data set as Figure 1.2.

tors and partly precedes the affinity maturation stage (see section 1.2). This initial selection can be studied by comparing the mutation profiles of non-productive and productive BCR sequences; the latter often have shorter CDR3 sequences post-selection, and exhibit complex and position-dependent selection for and against particular amino-acids [Elhanati et al., 2015].

1.7 BCR dynamics during HIV infection

The human immunodeficiency virus (HIV) is a lentivirus that infects CD4+ T cells and, if untreated, can lead to the development of acquired immunodeficiency syndrome (AIDS). HIV is highly genetically variable, both globally and within individuals during chronic infection. The virus originated from related simian immunodeficiency viruses (SIVs) in primates and has crossed from primate species to humans multiple times. HIV-2 was transmitted to humans from sooty mangabey monkeys, while HIV-1 was transmitted in at least four different

events: three from chimpanzees (forming groups M, N, and O; [Keele et al., 2006]) and one from gorillas (group P; [Plantier et al., 2009, Maartens et al., 2014]).The global pandemic strain, HIV-1 group M, consists of nine subtypes - subtype C predominates in South Africa and India while subtype B predominates in Western Europe and the Americas [Hemelaar et al., 2011].

HIV infection within hosts is complex, and consists of both acute and chronic phases. The virus is spread primarily through sexual contact, with 80% of mucosally transmitted HIV-1 B and C infections established by a single founder virus [Keele et al., 2008, Abrahams et al., 2009]. Transmission is followed by the eclipse phase of ˜10 days before the virus is detectable in blood plasma. Once the virus reaches the lymph nodes, it replicates rapidly and spreads throughout the lymphatic tissues, causing substantial loss of active germinal centres. Viral load typically peaks between 21-28 days after infection and then decreases to a stable viral load over 12-20 weeks (termed the 'set-point viral load'), which is maintained by an equilibrium between viral replication and evolution and host immune responses during over chronic infection [McMichael et al., 2010]. Over many years of untreated chronic infection, CD4+ T cell counts slowly decrease until germinal centre reactions cannot be sustained, leading to a rebound in HIV viral counts and a general collapse of the adaptive immune response (AIDS). Patients who develop AIDS often succumb to opportunistic infections which would otherwise be innocuous in healthy patients.

Evasion of a neutralizing antibody responses is an important part of HIV pathogenesis. The most important functional and antigenic region of HIV is the viral surface envelope (*Env*) glycoprotein, which binds to the CD4 receptor of CD4 positive T cells, which is the mechanism of viral entry into T cells [Wilen et al., 2012]. *Env* is composed of two trimeric subunits: gp41, which attaches to the viral capsid, and gp120, which contains the CD4 binding site (CD4bs; [Burton

and Mascola, 2015]). As detailed by Burton and Mascola [Burton and Mascola, 2015], there are three main types of antibody responses to *Env*. The first are non-neutralizing antibodies stimulated by the virus, which are antibodies that bind to antigenic regions of *Env* that do not actually impede viral entry. Over time many HIV functional spikes degrade to leave gp41 'stumps' on their surface, which act as non-neutralizing epitopes. The second category is strain-specific neutralizing antibodies that bind to antigenically variable regions of the functional *Env*. These typically arise early in infection and because they bind to epitopes that are easily varied are usually escaped by new viral variants that evolve during chronic infection. The third category, and the primary interest of many vaccine development efforts, is broadly neutralizing antibodies (bnAbs), which neutralize a wide spectrum of HIV variants observed worldwide. The first of the modern broad, potent HIV neutralizing antibodies, PG9 and PG16, were first discovered in 2009 [Walker et al., 2009]. These antibodies neutralized up to 80% of the $>$ 150 HIV strains tested [Walker et al., 2009]. Since then, multiple bnAb lineages targeting a variety of epitopes have been characterized, some of which neutralize up to 98% of HIV strains tested (see review in [McCoy and Weiss, 2013]). Locations of these different epitopes are shown in Figure 1.5.

Anti-HIV bnAbs typically only appear in ˜20% of patients after years of chronic infection [Haynes et al., 2012, Burton and Mascola, 2015, McCoy and McKnight, 2017]. These antibodies have a number of unusual features such as high levels of SHM, long CDR3s, and frequent use of insertions and deletions [Haynes et al., 2012, Burton and Mascola, 2015]. Conserved epitopes of *Env* are often shielded by glycans that prevent neutralizing antibodies from reaching these epitopes [Sattentau, 2011]. A number of bnAb features, especially long CDR3 length, appear to be related to the ability to 'pierce' the glycan shield to bind to the underlying conserved *Env* epitopes [Doria-Rose et al., 2014, Burton and Mascola, 2015]. Two frequent targets of bnAbs are the CD4bs (Figure 1.5;[Wu et al., 2010, Liao et al.,

2013]), and conserved motifs of the V1 and V2 variable loops of gp120 (Figure 1.5;[Doria-Rose et al., 2014]).

A large effort is currently underway to use stimulation of anti-HIV bnAb lineages to develop an HIV vaccine. Typical approaches to viral vaccination involve empirical testing with killed or live attenuated viruses [Haynes et al., 2012], or, more recently, recombinant protein and viral vectored vaccines (see [McCoy and Weiss, 2013]), though these strategies have consistently failed to elicit protective antibody responses to HIV infection. This is due in large part to the fact that most neutralizing antibody responses elicited by vaccination and early infection are strain-specific and do not provide protection from most HIV strains [Haynes et al., 2012]. Several studies have shown that passive transfusion of bnAb antibodies can be protective against SHIV (simian/human immunodeficiency virus) infection in macaques [Parren et al., 2001, Mascola et al., 2000]. However, these broadly neutralizing antibody lineages have proven difficult to elicit in humans due in part to their strange sequence characteristics, and the fact that they are often auto-reactive [Haynes et al., 2012].

The fact that HIV bnAb lineages tend to have complex maturation histories has motivated the in-depth study of antibody phylogenetics, with the hope that understanding the dynamics of natural bnAb development in HIV patients may provide clues on how to stimulate bnAb lineages in uninfected patients through vaccination [Haynes et al., 2012]. Three chapters of this thesis use three well-characterized bnAb lineages: i) CH103, a CD4bs binding antibody isolated from a patient tracked for three years starting from acute infection [Liao et al., 2013], ii) VRC26, a V1V2 binding antibody with a long CDR3 which penetrates the glycan shield, tracked from a patient over four years starting within the first year of infection [Doria-Rose et al., 2014]; and iii) VRC01, a CD4bs binding antibody isolated from a patient followed for over 15 years of chronic infection [Wu et al., 2015].

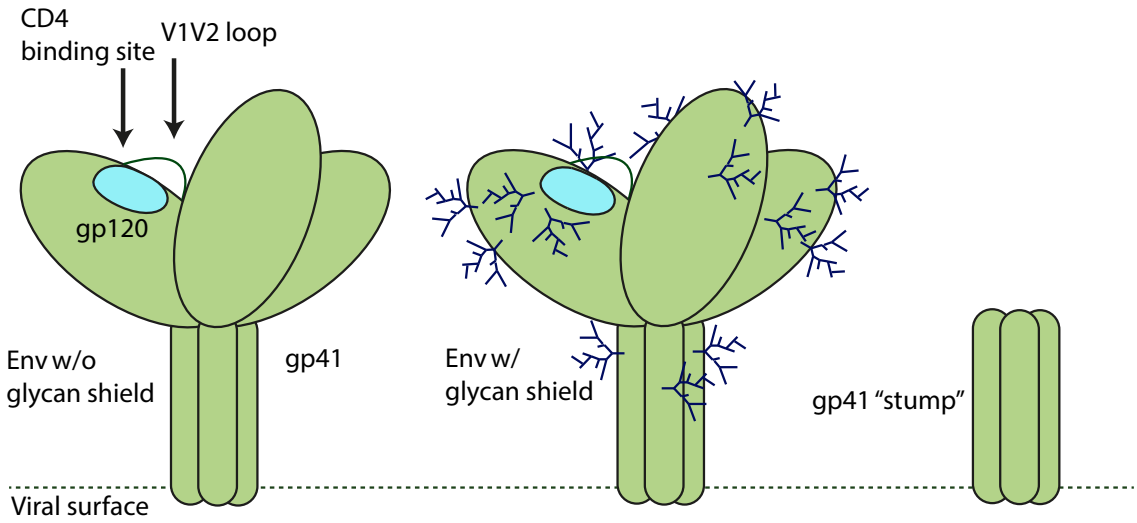


Figure 1.5: Representation of the *Env* protein on the HIV capsid. a) shows the Env protein, made of trimeric gp120 and gp41 subunits, without its glycan sheild. The CD4 binding site (epitope of CH103 and VRC01 antibodies) and V1V2 loop (epitope of VRC26 antibody) are labeled. b) shows *Env* with its glycan shield intact. c) shows a gp41 'stump' that is frequently left on the viral surface after degradation of gp120, and is often the target of non-neutralizing antibodies. See more detailed representation in [McCoy and Weiss, 2013, Burton and Hangartner, 2016]

These lineages were chosen for a number of reasons: i) they represent a diverse range of maturation times (3 to 15 years), ii) their neutralization profiles and antibody structures have been well-characterized previously, and iii) the dynamics of their maturation have been the subject of multiple major studies on the subject [Wu et al., 2015, Sheng et al., 2016].

1.8 BCR phylogenetics

The process of affinity maturation generates rapid sequence evolution, so it is unsurprising that phylogenetic approaches are now routinely used to visualise how B cell lineages undergo diversification and divergence in response to an antigen. Phylogenies have been used to address important problems, such as reconstructing ancestral BCR sequences within a lineage [Kepler, 2013, Sok et al., 2013], detecting and measuring selection on B cell populations [Uduman et al., 2014], and

studying how broadly neutralizing antibodies sometimes evolve in response to HIV infection [Wu et al., 2015]. Ancestral state reconstruction is particularly important in HIV vaccination because many efforts in HIV vaccination research focus on understanding the mutations that led to a particular antibody sequence [Haynes et al., 2012]. Further integration of phylogenetic concepts, including those from fields such as viral phylodynamics [Grenfell et al., 2004, Volz et al., 2013], may improve our understanding of affinity maturation dynamics during infection. For example, the rate of SHM evolution in a lineage over time (and its variability among lineages) could, in theory, be revealed by using molecular clock models to analyse BCR sequences sampled at different times. Further, asymmetric tree shapes might help to identify the action of strong positive selection on serially-sampled antibody lineages (Figure 1.6), analogous to phylogenetic footprint left by recurrent selection on some influenza virus lineages [Grenfell et al., 2004]. However, as noted in section 1.6, BCR lineage tree shapes may be subject to biases that are not yet fully understood [Uduman et al., 2014], so for the time being they should be interpreted with caution. Further, current phylogenetic models may not represent adequately the particular processes of growth and mutation that generate BCR lineages (see Section 1.2) and therefore they should be applied with caution. It is likely that current evolutionary models will need to be substantially modified or carefully selected before we can be confident in evolutionary inferences from BCR sequence data. Addressing this is a major task of this thesis, which is detailed in Chapters 3, 4, and 5.

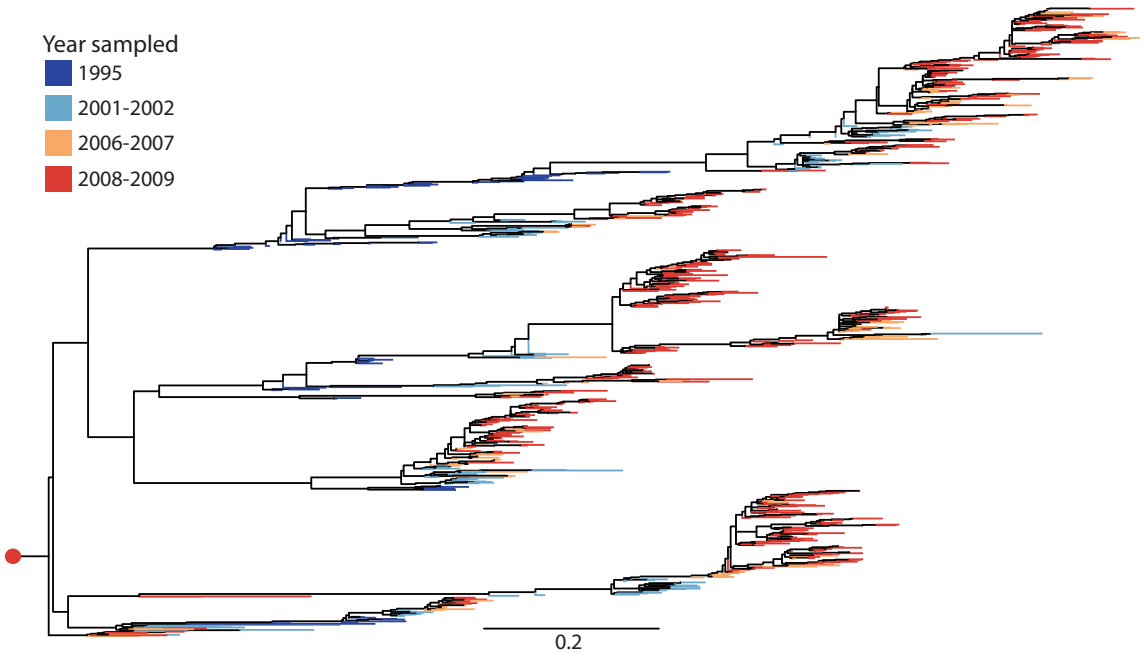


Figure 1.6: Maximum likelihood phylogenetic tree of the VRC01 bnAb lineage sampled at ten time points over 15 years of diversification within a single individual infected with HIV-1. Each tip represents a BCR heavy chain sequence; terminal branches are coloured by time point of sampling (see key). The red circle at the root represents the germline sequence (IGHV1-2*02 and IGHJ1*01, D region left unassigned). Note the general, but not complete, trend of increasing genetic divergence from the root with sampling time. Late-sampled sequences near the root indicate very high rate heterogeneity among lineages; these sequences might represent inactive memory B cells. BnAb sequences were obtained through cell sorting [Wu et al., 2010] followed by high-throughput sequencing data to identify related BCRs [Zhou et al., 2013]. See Wu et al. [Wu et al., 2015] for full experimental details. Sequences for this tree were obtained from GenBank [Wu et al., 2015] and aligned using MUSCLE [Edgar, 2004]. A maximum likelihood phylogeny was estimated using the GTRGAMMA substitution model in RAxML [Stamatakis, 2014], and re-rooted to position the germline sequence at the root with a divergence of zero. Scale bar represents genetic distance (expected substitutions per nucleotide site).

1.9 Phylogenetic substitution models and the maximum likelihood framework

The process of nucleotide change along a given phylogenetic tree is modelled as a continuous time Markov chain in which each state depends on the immediately preceding state and the amount of time that has elapsed between states [Felsenstein, 1981]. This process is characterized by the **Q** matrix, which contains the instantaneous rates of change from each state to each other state, typically from each nucleotide to each other nucleotide. For single nucleotide models, the **Q** matrix is parameterized as:

$$\boldsymbol{Q} = \begin{bmatrix} - & r_{TC}\pi_C & r_{TA}\pi_A & r_{TG}\pi_G \\ r_{CT}\pi_T & - & r_{CA}\pi_A & r_{CG}\pi_G \\ r_{AT}\pi_T & r_{AC}\pi_C & - & r_{AG}\pi_G \\ r_{GT}\pi_T & r_{GC}\pi_C & r_{GA}\pi_A & - \end{bmatrix} \tag{1.1}$$

Where r_{xy} is the relative substitution rate parameter from nucleotide x to y. If the **Q** matrix is symmetric (i.e. $r_{xy} = r_{yx}$ for all x and y), π_y is the equilibrium frequency of nucleotide y, which is the probability of observing nucleotide y if this Markov chain were to run over an infinite time. Larger r_{xy} values indicate a greater relative instantaneous rate for a particular nucleotide substitution. Combining these parameters gives the instantaneous rate of substitution from each nucleotide to each other nucleotide. Diagonal elements are computed such that all rows sum to zero. For instance, $r_{TT} = -1 * (r_{TC}\pi_C + r_{TA}\pi_A + r_{TG}\pi_G)$.

Different substitution models place different constraints on the values of their **Q** matrix. Almost all substitution models in common use are time-reversible, meaning their relative substitution rates are equal in either direction (e.g. $r_{TC} = r_{CT}$). This halves the number of substitution parameters to be considered, and also

has several useful properties that can be used to speed up likelihood calculation [Felsenstein, 1981]. Among time-reversible substitution models (reviewed in [Lio and Goldman, 1998]), the simplest is the Jukes-Cantor (JC69) in which all r_{xy} parameters are equal, and all $\pi_y = 1/4$. Other models estimate transition/transversion bias by estimating κ, an additional rate parameter applied to transitions (e.g. the elements of the **Q** matrix $q_{TC} = \kappa\pi_C$ and $q_{AC} = \pi_C$; HKY85), while the general time reversible model (GTR; [Yang, 1994]) uses a separate r_{xy} for all six possible combinations of two different nucleotides.

The values within the **Q** matrix are instantaneous transition rates and not probabilities of substitution between nucleotides. The probability of observing a particular nucleotide at time $s + t$ depends on the nucleotide at time s, but also on the amount of time t that has passed between the two points. In instances where t is very small, we are likely to observe the same nucleotide at each state. If t is very large, we expect the probability of observing a nucleotide at state s to be equal to the equilibrium frequency of that nucleotide. The transition matrix P^t contains the probability of observing any nucleotide at time $s + t$ given all possible nucleotides at state s, and t. This matrix is obtained by multiplying the **Q** matrix by the branch length t and computing the matrix exponential [Lio and Goldman, 1998]:

$$P^t = e^{Qt} \tag{1.2}$$

The values within the P^t matrix give the probability of transition from each nucleotide to each other nucleotide over a given branch length t. In contrast to the **Q** matrix, the rows of the P^t matrix sum to 1.

It is important to note that 'time' in this context (t, $s + t$) is not time as measured in days and weeks, but in units of genetic distances (i.e. substitutions per site). Over a given interval t, the expected number of substitutions per site depends not

only on the value of t, but also on the values within **Q**. To make branch lengths interpretable, the elements of **Q** are scaled such that their mean rate = 1. Where q_{xx} is the diagonal element for the row/column corresponding to nucleotide x in the **Q** matrix:

$$\mu = \sum_{x\epsilon(A,C,G,T)} -\pi_x q_{xx} \tag{1.3}$$

Each value in the **Q** matrix is then divided by μ, making the expected number of substitutions = 1 when t =1 using Eq. 1.2. This means branch lengths of the phylogenetic tree ($\boldsymbol{t}$) are interpretable as the expected number of substitutions per site between two nodes. Under a strict molecular clock model, in which mutation is assumed to occur at a uniform rate over time these lengths directly relate to time in units of days or weeks. Other models exist to allow variation in substitution rate, which more flexibly converts these branch lengths from units of substitution to units of time [Drummond and Rambaut, 2007]. However, in this thesis branch lengths represent genetic distance and will be in units of substitutions per site unless otherwise noted.

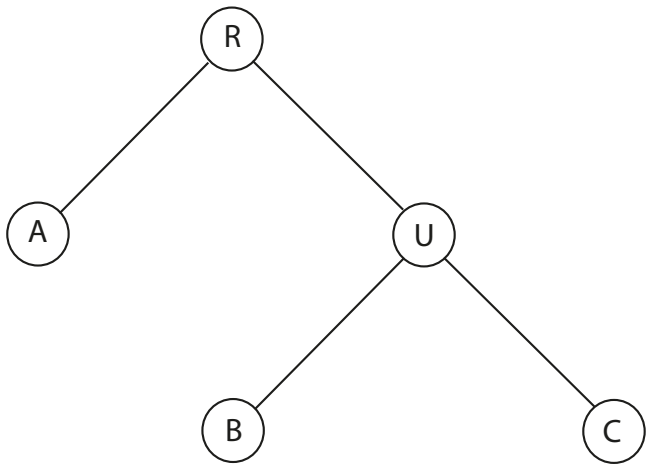


Figure 1.7: Example phylogenetic tree used in Eq 1.4 and 1.5. It is rooted at node R. A, B, and C are external nodes representing sequences.

It is possible to use Eq. 1.2 to calculate the likelihood of a set of sequences given a **Q** matrix, branch lengths, and phylogenetic tree topology. For the tree given in Figure 1.7, the likelihood of the data (X_i; nodes A, B, and C in Figure 1.7) at site i given a set of parameters ($\boldsymbol{v}$) used within the **Q** matrix, tree topology (**T**, shown in Figure 1.7), and set of branch lengths ($\boldsymbol{t}$) is the probability of the observed data averaged over all possible nucleotide histories of the internal nodes R and U. Specifically:

$$P(X_i|\boldsymbol{v},\boldsymbol{T},\boldsymbol{t}) = \sum_{r\epsilon(A,C,G,T)} \sum_{a\epsilon(A,C,G,T)} \sum_{u\epsilon(A,C,G,T)} \sum_{b\epsilon(A,C,G,T)} \sum_{c\epsilon(A,C,G,T)} \pi_r P_{ra} P_{A=a} P_{ru} P_{ub} P_{B=b} P_{uc} P_{C=c} \tag{1.4}$$

Where P_{ra} is the probability of nucleotide transition from nucleotide r to a over time, represented by edge RA in Figure 1.7. $P_{A=a}$ is an indicator function equal to one if the sequence $A = a$ at site i. Summation operators may be moved according to the tree structure in Figure 1.7:

$$P(X_i|\boldsymbol{v},\boldsymbol{T},\boldsymbol{t}) = \sum_{r\epsilon(A,C,G,T)} \pi_r \sum_{a\epsilon(A,C,G,T)} P_{ra} P_{A=a} \sum_{u\epsilon(A,C,G,T)} P_{ru} \sum_{b\epsilon(A,C,G,T)} P_{ub} P_{B=b} \sum_{c\epsilon(A,C,G,T)} P_{uc} P_{C=c} \tag{1.5}$$

Equation 1.5 shows how the likelihood of a single site in an alignment may be caluclated given model parameters, tree topology, and branch lengths. Expanding this equation to multiple sites, however, rapidly becomes computationally impractical because the likelihood must consider all possible nucleotide states of the internal nodes. For instance, a two-site example would involve each of the summation operators in Equation 1.5 operating over all dimers of nucleotides $r\epsilon(AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT)$. A three-

site example would involve performing summations over all trimers of nucleotides, and so on. To make likelihood calculations computationally practical for a larger number of sites, each site is assumed to mutate independently of all other sites. This means the likelihood may be calculated for each site individually using Equation 1.5. More specifically, given that sites are independent, the full likelihood is the product of the likelihood at each of n sites in the sequence alignment:

$$P(X|\boldsymbol{v}, \boldsymbol{T}, \boldsymbol{t}) = \prod_{i=1}^{n} P(X_i|\boldsymbol{v}, \boldsymbol{T}, \boldsymbol{t}) \tag{1.6}$$

Equations 1.5 and 1.6 are the key components of Felsenstein's pruning algorithm which, under the assumption of independent mutation among sites, dramatically improves efficiency in likelihood computation from its full definition in Eq. 1.4 [Felsenstein, 1981]. This algorithm is used for likelihood computation in virtually all likelihood-based phylogenetic estimation.

The goal of maximum likelihood phylogenetics is to estimate tree topology (**T**), branch lengths ($\boldsymbol{t}$), and parameters within the **Q** matrix ($\boldsymbol{v}$) that gives the highest likelihood value for the data. More specifically, the goal is to solve the function:

$$ML = argmax_{\boldsymbol{v},\boldsymbol{T},\boldsymbol{t}} P(X|\boldsymbol{v}, \boldsymbol{T}, \boldsymbol{t}) \tag{1.7}$$

Where **ML** is the set of parameters under which $P(X|\boldsymbol{v}, \boldsymbol{T}, \boldsymbol{t})$ is maximized. Substitution parameters and branch lengths may be optimized individually using hill-climbing optimization (e.g. Brent [Brent, 1971] or Newton-Raphson optimization, reviewed in [Ypma, 1995]), or together using multivariate optimization such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [Broyden, 1970, Fletcher, 1970, Goldfarb, 1970, Shanno, 1970]. Because the space of possible tree topologies is far too large to exhaustively explore in most datasets, ML estimates of tree topology are typically found heuristically by iteratively rearranging tree topology until no

improvement in likelihood is found [Gil et al., 2013].

Eq. 1.7 provides a natural framework for hypothesis testing, which is particularly useful if the set of substitution parameters v have clear biological interpretations. This typically compares models that are nested; i.e. when one is a special case of the other. For example, for a given dataset we may compare an HKY model in which κ is estimated by Eq. 1.7 (H_1) to a HKY model in which κ=2 (H_0). The maximum likelihood values (ℓ) obtained by these two models when solving Eq. 1.7 may be compared using a likelihood ratio test (LRT):

$$LRS = -2 * (log(\ell(H_0)) - log(\ell(H_1))) \tag{1.8}$$

Because H_0 is a special case of H_1, $\ell(H_0) \leq \ell(H_1)$ and therefore $LRS \geq 0$. As detailed in Huelsenbeck and Rannala [Huelsenbeck and Rannala, 1997], the distribution of LRS statistics under H_0 is approximately chi-squared distributed with degrees of freedom equal to the number of additional freely estimated parameters in H_1. In the previous paragraph's example, H_1 has one free parameter (κ) that is fixed in H_0. We may then compute a p value for the LRT comparing these two tests using the chi-squared cumulative distribution function (CDF) with one degree of freedom.

The phylogenetic models developed in this thesis are codon substitution models, which operate under the same framework outlined above. However, the **Q** matrix of codon substitution models are 61x61 matrices describing instantaneous rates of substitution of all sense codons to all others [Goldman and Yang, 1994]. This requires each site to be defined as a codon instead of a single nucleotide, and all summations in Eqs. 1.4 and 1.5 are performed across all sense codons instead of over all nucleotides. Codon models are especially useful for hypothesis testing because some of their substitution parameters have clear biological interpretations

relating to the degree and type of natural selection operating on the sequences [Yang and Swanson, 2002].

1.10 Overview of thesis

This thesis begins with an analysis of B cell repertoire diversity in eight HIV-1 infected patients during a clinical trial (Chapter 2). This analysis was non-phylogenetic in nature and attempted to relate the wholesale genetic diversity of the repertoire to clinical measurements such as viral load. This study revealed the challenges of applying descriptive summary statistics to such a complex biological problem. It therefore motivated a more in-depth, model based approach of analysing individual lineages. However, as detailed in sections 1.6 and 1.8, certain features of SHM violate important assumptions in standard phylogenetic substitution models. Chapter 3 introduces a phylogenetic codon substitution model that partially incorporates these unique features, and I apply it to three HIV bnAb lineages. Chapter 4 expands this substitution model by incorporating previously defined empirical models of somatic hypermutation and amino acid exchangeability. Chapter 5 further expands this framework to test hypotheses relating to variation in clonal selection among CDR and FWR regions, and among different clades of bnAb lineages. Chapter 6 brings this framework full circle, scaling these models up to be applied to entire B cell repertoires.

Chapter 2

Dynamics of Immunoglobulin sequence diversity in HIV-1 infected individuals

2.1 Introduction

The recent application of high-throughput sequencing technology to immunology, and to the characterisation of B cell and T cell receptor diversity in particular, has the potential to reveal immune system dynamics in unprecedented detail [Boyd et al., 2009, Six et al., 2013, Robins, 2013]. This work has led to advances in our understanding of antibody dynamics after vaccination [Laserson et al., 2014, Jiang et al., 2013], the effects of aging and infection on the B cell repertoire [Wang et al., 2013], the development of B cell cancers [Niklas et al., 2014, Bashford-Rogers et al., 2013], and the onset of autoimmune disorders such as multiple sclerosis [Stern et al., 2014]. Many of these studies have taken an 'antibodyome' perspective to viral infection, in which the immune response to infection is investigated by the bulk sequencing from patient samples of the antigen-binding regions of B cell immunoglobulin genes. These sequences are then analysed to explore how the anti-

body repertoire changes in response to perturbation arising from viral evolution and vaccination. This approach has been applied, for example, to influenza virus [Jiang et al., 2013, Vollmers et al., 2013], varicella-zoster virus [Wang et al., 2014a], and dengue virus [Parameswaran et al., 2013].

Some of the most notable findings in this field have come from detailed investigation of the interactions between the humoral immune system and HIV-1 infection. Much of this work has focused on observing and understanding the development of broadly neutralizing antibodies (bnAbs) in order to aid the development of preventative HIV-1 vaccines [Haynes et al., 2012]. Interestingly, bnAb precursors may be preferred in vaccine design over observed bnAbs because they may have a wider binding profile, and the sequences of such precursors can be inferred using methods of phylogenetic ancestral sequence reconstruction [Sok et al., 2013]. This requires an understanding of the diversity and dynamics of B cell clones during infection, and significant strides have been made, both in tracking the co-evolution of B cell and viral lineages, [Fera et al., 2014, Liao et al., 2013, Gao et al., 2014] and in finding potential bnAb precursors using phylogenetic methods [Sok et al., 2013].

Despite these advances, more general trends in B cell diversity and clonal dynamics during HIV-1 infection are still poorly understood, particularly during early infection. While some studies have contrasted the plasmablast, naive, and memory B cell content between early and chronic infections [Moir et al., 2010], the diversity of patient 'antibodyomes' has not yet been characterized, either in comparison to healthy controls or through time in individuals that have recently seroconverted. Although early anti-retroviral therapy (ART) has been shown to have a significant effect on viral divergence (e.g.[Gall et al., 2013]), the corresponding effect on B cell clonal diversity under ART is unknown. While it may be expected that B cell populations in HIV+ individuals are more clonal than in uninfected individuals, and

that the degree of B cell clonal relatedness is related to the size of the concurrent viral population, these associations have not been explicitly tested.

Most previous studies that investigated the repertoire of B cell receptor (BCR) diversity in peripheral blood (as opposed to those that focused on specific clonal lineages, e.g.[Fera et al., 2014]) have typically used non-phylogenetic approaches to summarize B cell diversity. One successful approach has been to use single-linkage clustering to group sequences into clusters (or clones), and then to infer clonal expansions by measuring properties of the size distribution of clones using entropy scores, such as the Gini index [Bashford-Rogers et al., 2013]. Others studies have used alternative statistics, including mean clone size, the number of unique IgHV-D-J allele combinations, and genetic distances between IgV segment sequences and their respective germline homologs (e.g. [Niklas et al., 2014]). Both sets of approaches have shown promise in studying B cell cancers, which result from significant clonal expansions of usually one BCR lineage [Chiorazzi et al., 2005], but it is unknown if a similar approach will be informative when studying immune responses to HIV infection.

To better understand B cell repertoire dynamics during early HIV infection and the degree to which it is modulated by ART, I used deep-sequencing data to capture the diversity of immunoglobulin (Ig) heavy-chain sequences from eight HIV patients enrolled in the Short Pulse Anti-Retroviral Therapy At SeroConversion (SPARTAC) trial [The SPARTAC Trial Investigators, 2013]. This is, to my knowledge, the first time high-throughput BCR sequencing has been applied to a study of ART and early HIV infection. Patients were enrolled an estimated 12 to 95 days after seroconversion, and were sampled at up to eight time points over approximately two years. At each time point B cell repertoire sequencing was performed and both viral load and CD4+ T cell counts were measured. Three patients were untreated, whilst five received ART for the first 48 weeks of the study only. This

study design not only allowed for tracking of individual B cell clones during early infection, but also testing for associations between the dynamics of B cell sequence diversity and clinical variables, including treatment status.

For each patient and each time point, the high-throughput Illumina MiSeq platform was used to obtain paired-end reads from Ig heavy chain sequences that represent the mixture of antibody classes in peripheral blood. Within each patient, I extended a previous single-linkage clustering approach [Bashford-Rogers et al., 2013] to classify sequences into clones and track their relative frequencies through time. I also explored a number of statistics in order to quantify BCR sequence diversity from HIV-1 infected patients, and to compare this diversity to that observed in a cohort of HIV-negative controls. While some general patterns were observed, overall I found a high degree of heterogeneity in B cell clonal dynamics both among patients and through time.

2.2 Methods

2.2.1 HIV patients

Peripheral blood mononuclear cells (PBMCs) were isolated from eight patients with primary HIV-1 infection recruited from the Short Pulse Anti-Retroviral Therapy At SeroConversion (SPARTAC) study [The SPARTAC Trial Investigators, 2013]. All had recently seroconverted before enrolment in the trial (an estimated 12-93 days before enrolment; median = 56 days). Patients 1-3 were untreated during the study period. Patients 4-8 received an ART regimen from week 0 to week 48, after which treatment was suspended. Patients were sampled between six and eight times over 108 weeks of the study (all time points are defined as weeks after start of the study, defined as week 0). All patients were sampled at weeks 4, 16, 24, 52, 60, and 108, whereas four patients were also sampled at week 0, and six at week

12. To ensure consistency among patients, and to ensure an equal number of time points during and after ART, only the former time points were used in analysis.

2.2.2 RT-PCR

Sequence data were obtained from mRNA of PBMCs amplified by RT-PCR. RT-PCR reagents were purchased from Invitrogen and primers (supplied by Sigma Aldrich) are described by Van Dongen et al.[van Dongen et al., 2003] and in Appendix Table A.1. Reverse transcription (RT) was performed using 500ng of total PBMC RNA mixed with 1μl JH reverse primer (10μM), 1μl dNTPs (0.25mM), and RNase free water added to make a total volume of 11μl. This was incubated for 5 minutes at 65oC, and 4μl First strand buffer, 1μl DTT (0.1M), 1μl RNase-OUT Recombinant Ribonuclease Inhibitor and 1μl SuperScript III reverse transcriptase (200units/μl) was added. RT was performed at 50oC for 60 minutes before heat-inactivation at 70oC for 15 minutes. PCR amplification of cDNA (5μl of the RT product) was performed with the JH reverse primer and the FR1 forward primer set pool (0.25 μM each), using 0.5μl Phusion High-Fidelity DNA Polymerase (Finnzymes), 1μl dNTPs (0.25mM), 1μl DTT (0.25mM), per 50μl reaction. The following PCR program was used: 3 minutes at 94oC, 35 cycles of 30 seconds at 94oC, 30 seconds at 60oC and 1 minute at 72oC, with a final extension cycle of 7 minutes at 72oC on an MJ Thermocycler.

2.2.3 Sequencing and reference-based V-D-J assignment

MiSeq libraries were prepared using Illumina protocols and sequenced by 150bp paired-ended MiSeq (Illumina). MiSeq reads were filtered for base quality (median >32) using QUASR (http://sourceforge.net/projects/quasr) [Watson et al., 2013]. Sequences were concatenated and a gap inserted between the forward and reverse reads (average gap length ˜35 nucleotides; Appendix Figure A.2).

Non-immunoglobulin sequences were removed; only those reads with significant similarity to reference IgHV and IgHJ genes from the ImMunoGeneTics (IMGT) database [Lefranc et al., 2009] were retained, as determined using BLAST [Altschul et al., 1990] with E-value thresholds of $1x10^{-10}$ and $1x10^{-3}$, respectively. Primer sequences were then trimmed from the reads, and sequences retained for analysis only if both primer sequences were identified and if both forward and reverse sequence lengths were >100 bp. Lastly, a second BLAST analysis was used to identify the IgHV genes of retained reads. Specifically, BLAST [Altschul et al., 1990] was used to align reads against known BCR sequences from the IMGT database [Lefranc et al., 2009] (E-value thresholds for IgHV and IgHJ-genes were 10^{-70} and 10^{-20} respectively, due to their different gene lengths). The combined per-base error rate for the RT-PCR and sequencing process of the MiSeq platform was $2.06x10^{4}$ [Bashford-Rogers et al., 2013, Bashford-Rogers et al., 2014].

2.2.4 Alignment

The data comprised non-overlapping paired-end reads separated by a gap of variable and unknown length. This necessitated certain heuristic measures during data processing and the two reads of each pair were aligned separately to ensure positional homology. As above, reference IGHV and IGHJ segments were obtained from IMGT [Lefranc et al., 2009]. Gaps were removed from the reference sequences, then IGHV sequences were clustered using CD-HIT [Fu et al., 2012] with a 95% similarity threshold. Only one sequence was retained from each cluster, resulting in a smaller set of V segments. The IHGJ reference sequences were not down-sampled, however. The first read of each pair was aligned with the closest matching sequence in the V references through gapless pairwise alignment. To further optimize the alignment each read was matched using only the 5′ half of the V segments. Next, the second read of each pair was pairwise aligned with the closest sequence in the J references, again using gapless alignment. Once the first

and second reads had been aligned to the V and J references, respectively, gaps were added 3′ of the first read, and 5′ of the second read, such that each were 300 nucleotides in length. The two reads were then concatenated, and excess gaps were trimmed between each read pair. A graphical representation of this process is provided in Appendix Figure A.3. Because I expect sequences from the same clone to match the same, or highly similar, V and J segments, this process generates an approximation of the optimal gapless alignment for sequences from the same clone. Further, the process does not require sequences to match the same V and J segments in order to be compared, a requirement that may be problematic when dealing with highly mutated sequences. However, the process may produce sub-optimal alignments when indel diversity is present within a clone. A small percentage of reads showed high similarity between forward and reverse reads (longest common sub-sequence of 20bp or greater) and were excluded during this step.

2.2.5 Clonal assignment

Sequences from all time points within each patient were pooled and assigned to clones using single-linkage agglomerative clustering using the Hamming distance between two sequences [Bashford-Rogers et al., 2013], while ignoring gap variation. This latter condition was necessary because each aligned sequence had an arbitrary gap between its two reads that resulted from paired end sequencing, and not from a natural process of immunological diversification. To make my results comparable with previous work [Bashford-Rogers et al., 2013], I also chose a maximum Hamming distance of one for clustering two sequences together into the same clone.

2.2.6 Measurement of BCR diversity

The diversity of BCR sequences was qualitatively visualised using the network approach developed by Bashford-Rogers et al [Bashford-Rogers et al., 2013]. Briefly, each vertex represents a unique BCR sequence, whose relative size is proportional to the number of sequence reads identical to the vertex sequence. Edges are then drawn between vertices whose sequences differ by at most one nucleotide change. Networks were computed using igraph, as implemented in R (http://igraph.sourceforge.net; see [Bashford-Rogers et al., 2013] for details). Because each clone is shown in proportion to its relative frequency in the BCR sequence population, these networks provide an intuitive visualisation of the clone size distribution. Examples of these plots from representatives of the treated and untreated patient groups are shown in Figure 2.1.

Three sample summary statistics were applied to the BCR sequences obtained at each time point to summarise their clonal diversity: (i) mean clone size, (ii) vertex Gini index, and (iii) the proportion of reads in 'large clones.' Each of these three statistics is described below. Given an empirical cumulative distribution (D) of clone sizes, and a theoretical cumulative distribution (E) obtained from the same number of clones of equal size, the Gini index is the total area between distributions D and E, divided by the area under E. In other words, it is the proportional difference between an empirical cumulative distribution and the expectation of that distribution under perfect equality. See review of the index's history and use in [Ceriani and Verme, 2012]

For each time point and each clone, clone size was defined as the number of reads in the clone at that time point divided by the total number of reads at that time point. Mean clone size was calculated as the arithmetic mean of this distribution of clone sizes. This is equivalent to $\frac{1}{\text{Total number of clones}}$. We used two different Gini index measures to characterize how unequally reads were distributed among

clones. Next, the vertex Gini index was used to summarise inequality in BCR sequence frequency; the index equals 1 if all reads have the same BCR sequence (maximum inequality), and equals 0 if each BCR sequence is equally common (maximum equality). As defined above, vertex size equals the number of identical reads with a unique BCR sequence, so the vertex Gini index was computed directly from the vertex size distribution as formed using a clustering cut-off of zero. Bashford-Rogers et al [Bashford-Rogers et al., 2013] explored both the vertex and cluster Gini indices, and found that the former correlated better with clinical parameters in Chronic Lymphocytic Leukaemia (CLL) patients; hence only the vertex Gini index is used here.

Clones were classified as 'large clones' if they comprised at least 0.1% of the reads sequenced at any of the time points in which they were found. I quantified these by calculating the proportion of reads at each time point that belong to 'large clones'. Additionally, I wished to describe changes in the very upper tail of the clonal size distribution. To do so, I plotted the sizes of the 20 largest clones as a proportion of the total number of reads at each time point. These values are for illustrative purposes only and are not used as sample statistics.

2.2.7 Subsampling

Preliminary statistical analysis using ANOVA showed that Gini index values were significantly associated both with patient identity (p=0.02) and read depth (p=0.001). A plot of Gini index values against read depth for each time point highlights the variation in read depth among patients (Appendix Figure A.4). To ensure that comparisons among patients are reliable and not an artefact of read depth variation, all further analyses were undertaken on subsamples of the original data. Specifically, 70,000 sequences were randomly sampled from each time point in each HIV+ patient (the lowest number of reads available for a HIV+ patient sam-

ple was 74861). When comparing HIV+ patients with healthy controls (see below), then 5000 sequences were randomly sampled from each control and HIV+ data set (the lowest number of reads available for a control patient was 5087). In each case random subsampling of reads was repeated ten times. Since variance in BCR statistics among these ten repetitions was very low ($< 1x10^{-4}$) I only report the mean values of the ten subsampling repetitions.

2.2.8 Analysis of BCR diversity

To examine the relationship between BCR sequence diversity and clinical variables, I compared each of the above BCR statistics to log(viral load) and CD4+ counts using linear regression. P-values were adjusted for multiple hypothesis testing using the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995] (as implemented in the R function p.adjust). To account for possible temporal autocorrelation within patients, all data points within each patient were summarised using two means. Because ART will have a significant effect on many variables, such as viral load, I calculated separate means for values from early time points (weeks 4 - 24; i.e. during ART for treated patients) and from late time points (weeks 52 - 108; when no patients were receiving ART). These means were calculated by linearly interpolating between adjacent time points, and then calculating the mean of the resulting function between the first and last points (weeks 4 and 24 for early only, 52 and 108 for late only). ANOVAs were used to explore how the Gini index and other BCR summary statistic values vary among patients, time points, treatment status, and read depth. To make linear regression coefficients of each regression more comparable, all variables besides viral load and CD4+ count were scaled to a mean of zero and variance of one before linear regression and ANOVA analysis.

2.2.9 Comparison with healthy controls

To test whether BCR sequence diversity differs between HIV-infected and uninfected individuals, I analysed previously published data from a cohort of six healthy patients [Bashford-Rogers et al., 2014]. For each, non-gapped Illumina MiSeq reads were obtained from peripheral blood mononuclear cells. See [Bashford-Rogers et al., 2014] for full details of RNA capture, PCR, and sequencing for the healthy controls. To ensure that these non-gapped sequences are directly comparable to the gapped sequences obtained from HIV patients, I only used the first and last 110bp of each read, placing a gap of unknown length between. Because these patients were sequenced at a lower read depth (5087 - 8475 reads) than HIV+ patients, all data sets from healthy controls were randomly subsampled (as described above) to remove any potential bias arising from variable read depth. Further, when comparing with healthy controls, only within-time point edges were allowed when identifying clones in HIV patients (because sequences from only one time point were available from the control individuals). Differences in BCR diversity statistics between control and HIV-infected individuals were tested using Wilcoxon tests.

2.2.10 Statement of contribution

Patients were sampled by investigators of the SPARTAC clinical trial. All lab work, including RT-PCR amplification of BCRs (Section 2.2.2) and sequencing (Section 2.2.3) was performed by Dr. Rachael Bashford-Rogers and Dr. Astrid Gall at the Sanger Institute, Cambridge University. Initital sequence processing and reference mapping (Section 2.2.3), as well as generation of Figure 2.1 were performed by Dr. Rachael Bashford-Rogers at the Sanger Centre. I performed all other analyses.

2.3 Results

Figure 2.1 provides a qualitative visualisation of BCR sequence diversity, and its change between the beginning and end of the study period (weeks 4 and 108, respectively). The data shown is from Patients 3 and 4 and the plots were computed using the network approach developed in [Bashford-Rogers et al., 2013]. For both patients the sequences at week 4 are clearly more clonally clustered compared with those at week 108, although the BCR diversity for Patient 4 (Figure 2.1c) appears more polyclonal than for Patient 3 (Figure 2.1a) at this time. As expected, at neither time point is the BCR population dominated by a single large clone, as was often observed in samples from B cell lymphoma patients [Bashford-Rogers et al., 2013]. The plots shown in Figure 2.1 are largely typical of those observed for other patients.

Full results from all patients are summarized in Figures 2.2 and 2.3. All patients had seroconverted recently (median = 56 days) before the start of the study (week zero). Patients 1-3 were untreated through the course of the study period (weeks 0 to 108). Patients 4-8 received an ART regimen between weeks 0 and 48, after which they were untreated. BCR sequencing was performed on PMBC samples from each time point for each patient (yielding $7.4x10^4$ to $1.0x10^6$ filtered BCR reads per sample). BCR network analysis was applied to these sequencing datasets to quantify the clonal architecture of these samples according to [Bashford-Rogers et al., 2013]. Four plots are shown for each patient, which show the change through time in (i) viral load, (ii) the vertex Gini index of BCR sequences, (iii) the proportion of BCR sequences in 'large clones', and (iv) the size distributions of the 20 largest clones as a proportion of the total number of reads at each time point.

2.3.1 Untreated patients

All three untreated patients (1-3) showed peak viral loads at or within four weeks of enrolment, and reduced, but persistent viraemia after 52 weeks corresponding to set point viral load. Patient 1 had an additional peak in viral load at 24 weeks, suggesting a superinfection event, and showed a higher viral load for a longer duration than Patients 2 and 3. In all untreated patients Gini indexes and the proportion of reads in large clones were typically higher during the first half of the study, when viral loads were higher. Hence at these times a higher proportion of the BCR sequence population was composed of common sequences/clones. When 'large clones' dominated the BCR population (e.g. week 24 in patient 2), that dominance was not the result of a single dominant clone, but rather resulted from the concurrent increase in frequency of a large number of different clones (see bottom plot of each panel in Figure 2.2). Further, the relative frequencies of individual clones that were found at high frequencies (colours in bottom plot of each panel in Figure 2.2) fluctuated through time with no discernable pattern; in almost all cases clones were common at only one time point.

2.3.2 Treated patients

All five treated patients (4-8) showed a rapid decrease in viral load soon after the start of the study, with counts in weeks 16-24 below 100 copies/ml (often $<$50, or undetectable), coinciding with the period of ART. As expected, viral load values across all time points were significantly lower in the treated group than the untreated group (one sided t-test; $p < 0.0001$). Patients 5-8 showed a rebound in viral load at week 52, after the cessation of ART, whilst the rebound in Patient 4 was not as strong and viral loads remained relatively low ($<$21,000 copies/ml) throughout the study period. In Patients 5-7 viral load decreased in week 60 before stabilizing at 108 weeks at a similar level to that in untreated patients.

The dynamics of B cell diversity and clonality in the treated patients were highly variable. Patient 7 showed a striking pattern after cessation of ART: at week 60, there was a polyclonal expansion (such that most BCR sequences fell into 'large' clones) which remained in week 108. Patient 8 also showed polyclonal responses and an increase in the proportion of sequences in large clones at weeks 52 and 108.

BCR diversity dynamics in Patient 6 in many ways resembled those of untreated Patients 2 and 3, with a peak in the prevalence of 'large clones' in week 24, and low clonality in week 108. Patient 4 exhibited high viral load and clonality at week 4, after which both remained low for the rest of the study. Patient 5 showed a comparatively stable level of clonality through time, with the proportion of sequences in large clones ranging from 0.13 to 0.45. Notably, Patient 5 is the only patient to exhibit a 'large clone' that persisted throughout the study period (red bars in the bottom panel of Figure 2.3a). Across time points this persistent clone occupied between 2.4% and 5.1% of all BCR sequences.

2.3.3 Relationship between clinical variables and B cell diversity

I sought to test whether statistics of BCR diversity were associated with clinical variables, such as viral load, time since seroconversion, and CD4+ cell count (Appendix Figure A.6). To enable direct comparisons among patients, all data sets were sub-sampled to exactly 70,000 reads (see Methods section 2.2.7). For each patient, mean values during the early and late periods were used to account for autocorrelation (see Methods). As shown in Table 2.1, there was no significant association between log viral load, CD4+ cell count, or estimated days since seroconversion with any of the three BCR diversity statistics. In all cases, correlations were non-significant both before and after correction for multiple hypothesis testing. The results also remained non-significant when repeated without adjusting for temporal autocorrelation (i.e. each time point in each patient counted as an

independent data point).

Clinical Variable	BCR diversity statistic	Coefficient	Adjusted *p* value
log(viral load), Early period	Mean clone size	-0.005	0.99
	Gini index	-0.13	0.92
	Proportion of reads in large clones	-0.19	0.92
log(viral load), Late period	Mean clone size	0.3	0.92
	Gini index	0.37	0.92
	Proportion of reads in large clones	0.28	0.92
CD4 count, Early period	Mean clone size	-0.3	0.92
	Gini index	-0.43	0.92
	Proportion of reads in large clones	0.13	0.92
CD4 count, Late period	Mean clone size	-0.34	0.92
	Gini index	-0.47	0.92
	Proportion of reads in large clones	-0.34	0.92
Days from seroconversion at week 0, Early period	Mean clone size	-0.17	0.92
	Gini index	-0.19	0.92
	Proportion of reads in large clones	0.25	0.92
Days from seroconversion at week 0, Late period	Mean clone size	-0.29	0.92
	Gini index	-0.36	0.92
	Proportion of reads in large clones	-0.21	0.92

Table 2.1: Regression analysis of BCR diversity statistics and clinical variables. P-values were adjusted using the Benjamini-Hochberg procedure

Response	Factor	Mean Square	*p* value
Mean clone size	Patient	0.99	0.63
	Week	0.49	0.63
	Treatment	0.6	0.63
Gini index	Patient	1.7	0.57
	Week	2.3	0.57
	Treatment	0.2	0.74
Proportion of reads in large clones	Patient	0.48	0.87
	Week	0.7	0.63
	Treatment	1.1	0.63

Table 2.2: ANOVA analysis of BCR diversity statistics. P-values were adjusted using the Benjamini-Hochberg procedure

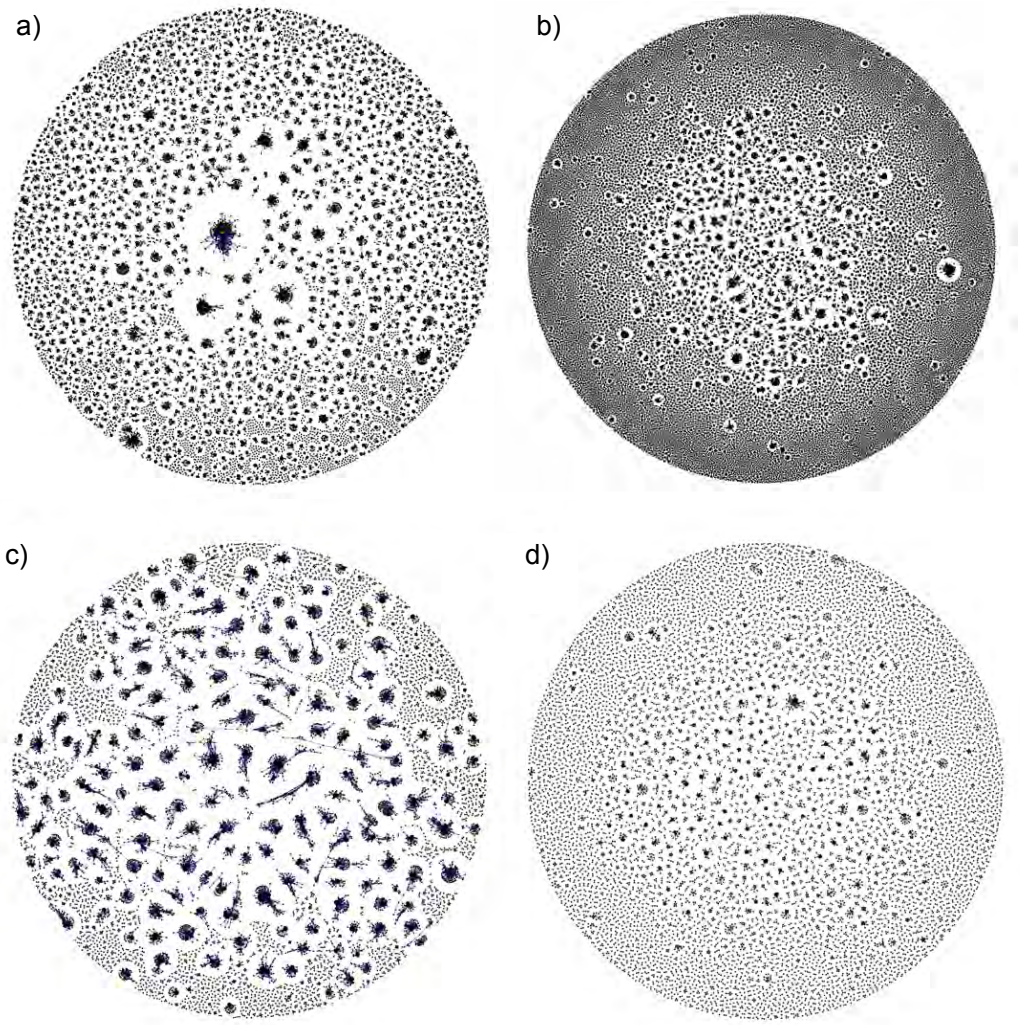


Figure 2.1: Network visualisation of the diversity of BCR sequences obtained from untreated patient 3 at week 4 (part a) and week 108 (part b), and from treated patient 4 at week 4 (part c) and week 108 (part d). Each vertex represents a unique sequence, and the size of each vertex is proportional to the number of reads identical to that sequence. Edges are drawn between vertices that are one nucleotide substitution apart. See Methods section 2.2.6 for further details.

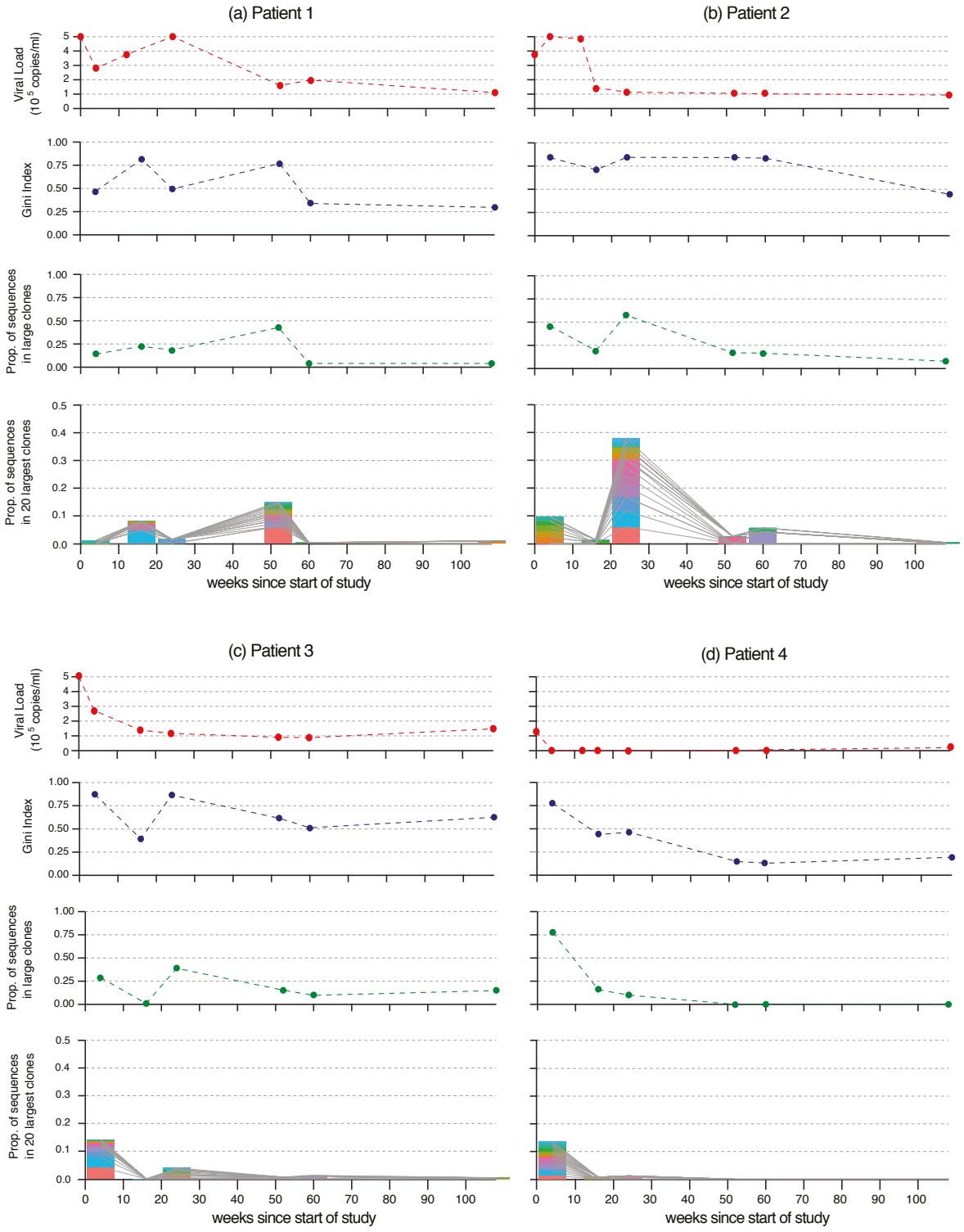


Figure 2.2: BCR diversity statistics and viral load values for patients 1 - 4 (parts a to d) over the study period of 108 weeks. Patients 1 - 3 were untreated, whilst patient 4 received ART until week 48. Four plots are provided in each panel. The top plot shows viral load, with each point representing a clinical sample. The second plot shows vertex Gini index values of the BCR sequences obtained at each time point. The third plot shows the proportion of reads at each time point that belong to 'large' clones, i.e. those that occupy >0.1% of reads at any time point. The bottom plot in each panel shows the proportion of all reads at each time point that are occupied by the 20 largest clones observed across all time points. Each of the 20 largest clones is represented by a bar of a different colour, and lines connect bars at adjacent time points that represent the same clone.

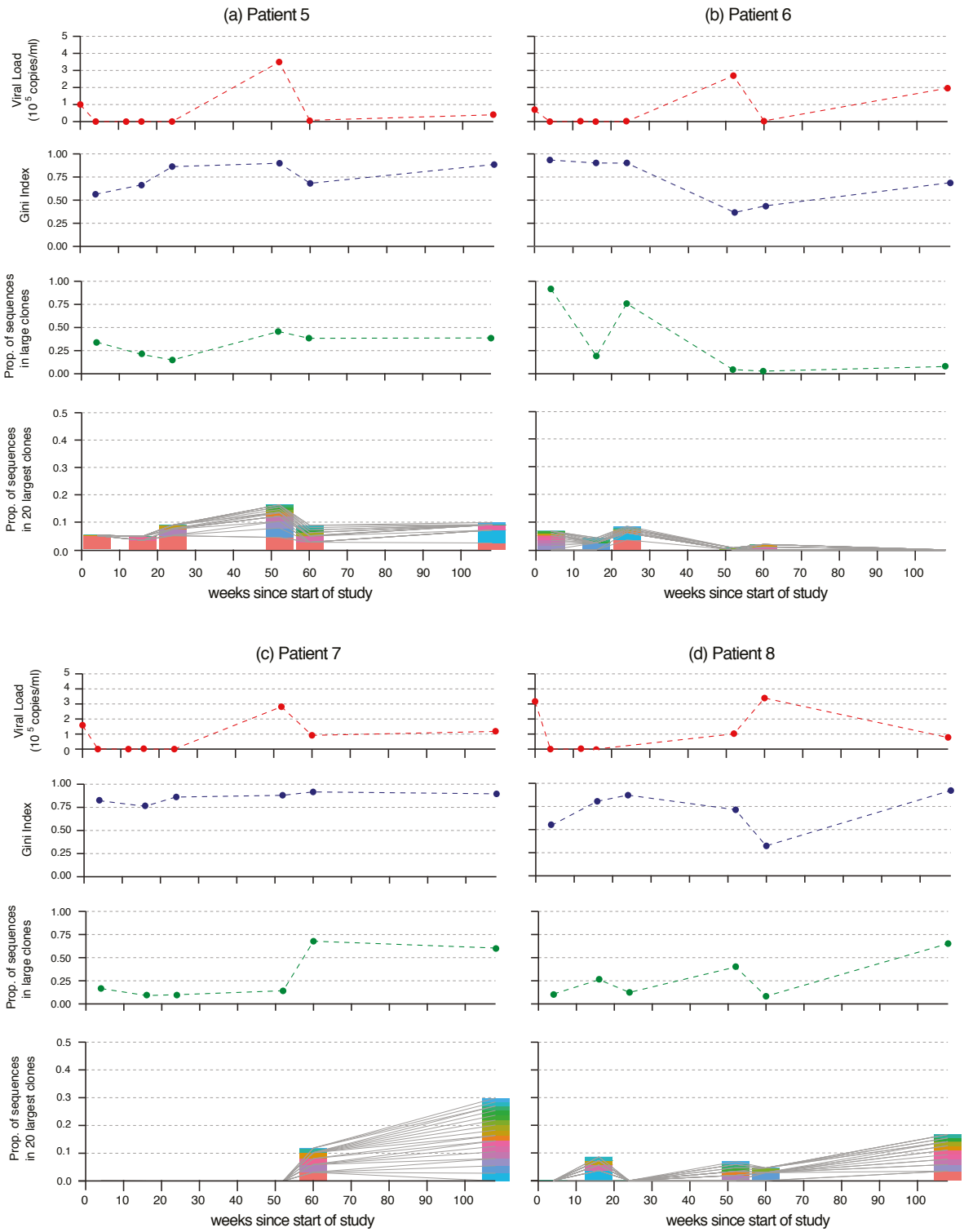


Figure 2.3: BCR diversity statistics and viral load values for patients 5 - 8 (parts a to d) over the study period of 108 weeks. All patients received ART until week 48. See Figure 2.2 legend for details.

2.3.4 Analysis of BCR diversity and subsampling

To explore possible sources of variation in the three BCR statistics, I used ANOVAs to test whether each statistic was associated with patient identity, week of sampling, and treatment group at each time point (Table 2.2). All data sets were subsampled to 70,000 reads. None of these factors explained a significant proportion of variation in any of the BCR diversity statistics. Thus the association between Gini index and patient identity observed prior to subsampling (see Supplementary Figure A.4) was most likely an artefact due to variation in read depth among patients.

I also explored the effects of sub-sampling on each of the three BCR diversity statistics (Appendix Figure A.7). As might be expected, the 'proportion of reads in large clones' statistic was almost unaffected by sub-sampling. Although vertex Gini indices were slightly higher before subsampling, this inflation was relatively constant within patients. Mean clone size was substantially altered by changes in read depth, justifying my use of sub-sampling when comparing data from different patients (Tables 2.1 and 2.2; Figure 2.4).

2.3.5 Comparison with healthy controls

To enable direct comparisons between HIV+ patients and healthy controls, all data sets were sub-sampled to 5000 reads (see Methods 2.2.9). As before, values from HIV patients were grouped into two periods, early (weeks 4-24) and late (weeks 52-108) to avoid temporal autocorrelation. Comparisons of BCR sequence diversity are shown in Figure 2.4. All three statistics (mean clone size, Gini index, the proportion of reads in large clones) were significantly higher in early-period HIV+ patients than in healthy controls ($p = 0.01$, 0.02, and 0.03, respectively). However, all comparisons between healthy controls and late-period HIV+ patients were non significant.

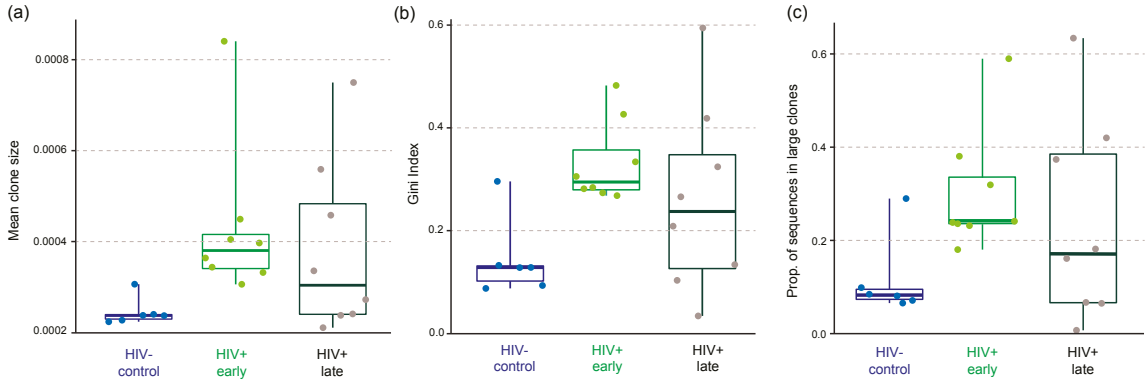


Figure 2.4: Comparison of three BCR summary statistics among healthy uninfected individuals (blue) and HIV-infected patients. Values from HIV+ patients are calculated separately for weeks 4-24 (HIV+ early; green) and weeks 52-108 (HIV+ late; black). For each period statistics were averaged using linear interpolation (see Methods section 2.2.9 for details). Each box and solid line represents the first, third, and second quartiles, respectively, whilst whiskers indicate range. Data points are superimposed on the box and whisker plots.

2.4 Discussion

The results presented here demonstrate that BCR sequence diversity during HIV infection is highly dynamic and heterogeneous, both within and among patients. Individual clones fluctuate substantially in relative frequency, and at no time point did a single clone dominate the BCR sequence population in any patient. This suggests an actively evolving antibody response against HIV, likely due to the high turnover of HIV genetic diversity within each patient generating novel antigenic targets (e.g.[Wei et al., 2003]).

I observed significantly greater clonal structure in HIV+ patients than in uninfected controls during the first half of the study period, but not during the second half. This is consistent with a pronounced humoral response to viraemia during early infection. BCR diversity statistics during the late period were highly variable, suggesting that strong clonal structure was maintained in some patients but

lost or significantly reduced in others (Figure 2.4). However, when comparing HIV+ patients I found no significant association between BCR diversity statistics and clinical variables, such as viral load or CD4+ count, suggesting a highly complex relationship between the adaptive immune system and viral dynamics. Previous studies have shown highly heterogeneous B cell clonal responses to immune perturbations such as vaccines over a timescale of weeks [Laserson et al., 2014], so the variation observed here over a period of approximately two years fits with established findings. Further, once differences in read depth were accounted for, BCR diversity statistics did not vary significantly with treatment group, time of sampling, or patient identity.

There are several possible explanations for the heterogeneity I observed among HIV+ patients and through time. The first and simplest explanation is a lack of statistical power and sampling resolution. With only 8 patients, the power of this study is almost certainly constrained by small sample size. However, a previous study of B-cell lymphoma patients that had a similar sample size and which used similar analytical techniques was able to establish significant associations between BCR clonality and clinical outcome in CLL patients, likely due to the exceptionally large B cell clones generated by this cancer [Bashford-Rogers et al., 2013]. Another potentially important factor in this study is the coarseness of temporal sampling relative to the dynamics of the BCR sequence population. Each patient was sampled six to eight times over two years, with all time points at least four weeks apart. Further, the earliest sample taken after ART ended in the treated group was obtained in week 52. This is potentially significant because recent studies of vaccine responses that sampled more intensively have shown that B cell clones originating at vaccination tend to expand within the first week after inoculation and afterwards rapidly contract, with little persistent expansion after four weeks [Laserson et al., 2014]. Though I might expect clonal responses to chronic viral infection to be more persistent than those to vaccination, the sampling scheme of

this study (which was not initially designed with B cell dynamics in mind) may not reveal consistent trends in BCR clonality during HIV infection. Fluctuations that might be associated with acute infection may be difficult to discern because (i) some patients were given early treatment, (ii) the time between seroconversion and the start of the study varied substantially (12-95 days), and (iii) in one patient there was evidence of HIV-1 superinfection during the study period. Further, BCR diversity changes associated with viral rebound after cessation of ART may not be apparent because patients were sampled several weeks later. It is also important to consider the effects of viral escape variants on sequence diversity of B cells. For instance, because it involves constantly changing antigens, persistent viral escape may lead to expansion of many small B cell clones, rather than a single large, easily detectable clone.

There are also biological explanations for absence of significant relationships between clinical variables and BCR diversity scores. Whilst the BCR metrics used here were informative when applied to B cell lymphoma patients, those patients often exhibited extreme clonality, such that one or a few B cell clones comprised the majority of BCR reads [Bashford-Rogers et al., 2013]. However the response of the BCR sequence population to viral infection is likely to be far more subtle and diverse. Because clonal expansion during infection is related to binding affinity to viral epitopes, BCR diversity will be related not only to the amount of virus present at the time of sampling, but also to the antigenic diversity of the virus population, B-cell longevity, and complex ecological dynamics arising from antigenic cross-reactivity [Gao et al., 2014]. Further, the dynamical interaction between viraemia, viral sequence variation, helper T-cell counts, and T-cell activation of HIV-specific B cells may further obscure relationships between BCR diversity and other measurements. The network visualisations (Figure 2.1) and BCR statistics (Figures 2.2 and 2.3) reported here indicate that the BCR clone size distribution during early HIV infection is much less skewed than that observed for B

cell cancers [Bashford-Rogers et al., 2013], meaning that HIV-specific clonal expansions may be difficult to detect from bulk antibody samples without some form of functional filtering or screening (e.g. antigen-specific cell sorting or isolation of plasmablasts [Palanichamy et al., 2014]). If so, then further progress will require very close collaboration between experimental researchers and theoreticians interested in developing new analytical approaches. Another process of relevance to studies of HIV infection is B cell hypermutation, which has been shown to generate substantial intra-clonal sequence diversity and divergence through time [Sok et al., 2013, Fera et al., 2014]. If high rates of hypermutation are common then the conservative clustering algorithm used here (which only links sequences that differ by one mutation) may split large and diverse lineages into distinct groups, thereby reducing its ability to capture BCR dynamics. Further work is needed to discriminate between inter- and intra-clonal diversity in a statistically rigorous manner.

The sequence data used in this study was somewhat limited by the fact that the sequence reads did not overlap sufficiently to be assembled into the full variable sequence. Fortunately, the junction region was captured in the reverse read, leaving a gap of indeterminate length in the V gene region. Because most genetic diversity is concentrated in the CDR3 junctional region [Xu and Davis, 2000], the gapped nature of our sequence data likely did not have a major impact on our results.

Our results are more successful in highlighting the potential usability and performance of different statistics of BCR diversity. Although the vertex Gini index proved informative in prior work [Bashford-Rogers et al., 2013], I found that it is positively correlated with sequencing read depth (Appendix Figure A.4). Application of the Gini index to simple hypothetical data sets demonstrates why this correlation may arise (see Appendix Table A.5). In short, as read depth increases,

large clones grow in size yet more unique singleton reads are also added, thereby increasing the skew of the clone size distribution, resulting in a higher Gini index. Bashford-Rogers et al. [Bashford-Rogers et al., 2013] also explored the dependence of the Gini index on read depth, via random sub-sampling of empirical data (Supplemental Figure 9 in [Bashford-Rogers et al., 2013]), although in their study the effect was small compared to the difference between case and control patients. In analyses of HIV infection meaningful differences in Gini indices are likely more subtle and thus this statistic may be less effective. One BCR statistic used here, the 'proportion of reads in large clones', showed no significant relationship with read depth and was unaffected by subsampling (Appendix Figure A.7) and is thus worth considering further. Importantly, while the Gini index attempts to characterize the entire clone size distribution, and is thus affected by changes in the relative frequency of very rare variants, statistics such as the proportion of reads in 'large clones' focus on the dynamics of large clones. Substantial theoretical work is needed to understand the sampling properties of BCR diversity statistics and to develop more powerful alternatives to those used to date. It is likely that multiple informative statistics will be needed to adequately summarise BCR repertoires obtained through bulk sequencing, and may need to consider properties of the data other than the distribution of clone sizes.

Further methodological challenges face the analysis of BCR repertoire data obtained through bulk sequencing. A formal, phylodynamic analysis of BCR sequence data is appealing because it would allow for model-based inference of parameters such as hypermutation rates and the relative growth rates of clonal populations. However, because B cell receptor lineages are created through somatic V(D)J recombination, bulk BCR sequences do not coalesce to a single common ancestor and must be separated into distinct lineages before analysis can proceed. Further, probabilistic models of sequence evolution will be computationally very demanding when applied to BCR data sets, some of which - including those used

here - contain millions of unique sequences per patient.

Chapter 3

A phylogenetic codon substitution model for antibody lineages

3.1 Introduction

Recent advances in sequencing technology are giving an unprecedented view into the genetic diversity of the immune system during infection, especially in the context of chronic infections caused by viruses. Broadly neutralizing antibody (bnAb) lineages, which produce B cell receptors (BCRs) capable of binding a wide range of viral epitopes, are of particular interest [Haynes et al., 2012]. Within such lineages, all B cells descend from a shared common ancestor and are capable of rapid sequence evolution through the processes of somatic hypermutation (SHM) and clonal selection. For chronically infecting viruses such as HIV-1, this co-evolutionary process may continue for years [Wu et al., 2015]. Because immunoglobulin gene sequences from bnAb lineages undergo rapid molecular evolution, selection and diversification, they would appear to be suitable for evolutionary and phylogenetic analysis, and these methods have already been applied to various immunological questions such as inferring the intermediate sequences of bnAb lineages [Sok et al., 2013]. Intermediate ancestors of B cell lineages are

of particular interest because they may act as targets for stimulation by vaccines [Haynes et al., 2012].

However, the biology of mutation and selection during somatic hypermutation is different from that which occurs in the germline, and therefore it is unlikely that standard phylogenetic techniques will be directly applicable to studying bnAb lineages without suffering some bias and error. One of the most important assumptions of likelihood-based phylogenetics is that evolutionary changes at different nucleotide or codon sites are statistically independent. Without this assumption, likelihood calculations become computationally impractical as the length and number of sequences increases [Felsenstein, 1981]. Unfortunately, in contrast to germline mutations, somatic hypermutation of BCR sequences is driven by a collection of enzymes that cause some sequence motifs (between two and seven base pairs) to mutate at a higher rate than others [Smith et al., 1996, Teng and Papavasiliou, 2007, Elhanati et al., 2015]. Context sensitivity of mutation during affinity maturation is driven largely by the targeting preferences of AID (Activation-Induced Cytidine deaminase), an enzyme that induces G:U mismatches in DNA and largely drives increased mutation rate during affinity maturation [Muramatsu et al., 1999]. For example, AA<u>C</u> and <u>G</u>TA act as hotspots that mutate with higher probability at their underlined base while CC<u>C</u> and <u>G</u>GG are coldspots that are less likely to mutate at their underlined base [Smith et al., 1996, Yaari et al., 2013]. This context sensitivity clearly violates the assumption of independent evolution among sites. Furthermore, because hotspot motifs are, by definition, more mutable than non-hotspot motifs, their frequency within a B-cell lineage may decrease over time as they are replaced with more stable motifs [Sheng et al., 2016]. These changes will not be passed on to subsequent organismal generations through the germline because the mutational process is somatic; thus, B cell lineages cannot be assumed to have converged to their stationary nucleotide frequency distribution as is usually done in phylogenetic analysis. This effect may have a number

of consequences for molecular evolutionary inference, for example it may render inappropriate the common practice of estimating equilibrium frequencies from the sequences themselves. At present it is unknown how the violation of these assumptions will affect phylogenetic inference of BCR sequences in practice, and the problem of ameliorating such effects remains an open issue.

Some approaches have been developed to study the substitution process in BCR data in the context of biased mutation. Some of these are non-phylogenetic in nature [Hershberg et al., 2008, Yaari et al., 2012] and focus on the expected number of germline to tip replacement mutations in comparison to a null model. Kepler et al [Kepler et al., 2014b] developed a non-linear regression model approach that, combined with an empirical model of mutation rate at each site, allowed the authors to test for the effects of selection and mutation on BCR genetic diversity. The substitution model detailed in McCoy et al [McCoy et al., 2015] accounts for biased mutation by comparing values of nonsynonymous/synonymous mutation rate ratios inferred from a given data set to those inferred from out-of-frame rearrangements, and focuses on analysing mostly non-phylogenetically related sequences from an entire BCR repertoire.

However, no existing approaches explicitly parameterize the effect of biased mutation of BCR hotspot motifs within a phylogenetic substitution model, and this is a crucial step in uniting the well established field of model-based molecular phylogenetics with BCR genetics and immunology. The aim of this chapter is to develop such a model. Specifically I introduce a model that can partially account for the effect of context dependent mutability of hot- and coldspot motifs, and explicitly model descent from a known germline sequence. The model I introduce is a modification of the NY98 [Nielsen and Yang, 1998] codon substitution model. This model is a widely used simplification of the model introduced in [Goldman and Yang, 1994], and is, confusingly, often referred to as the 'GY94'

model. Although an approximation, this new model can detect and quantify the effect of somatic hypermutation on BCR sequences whist preserving the assumption of independence among codon sites in order to maintain computational feasibility. This model shows a significantly better fit than the standard NY98 model to three previously published and long-lived bnAb lineages in HIV-1 infected patients. Through simulations, I validate the effectiveness of the model, and show its ability to reduce bias in the estimation of other evolutionary parameters such as tree length. Further, I use this model as a framework for testing hypotheses of hotspot motif symmetry and hierarchy of mutability, and I explore its potential applications such as improved ancestral state reconstruction.

3.2 Materials and Methods

3.2.1 A molecular evolutionary model for antibody lineages

The model I propose differs from most substitution models in some important ways. First, codon substitutions are weighted by the probability that they occurred within a particular hotspot motif, which makes the substitution process non-reversible and has important implications for parameter inference and likelihood calculation. Second, BCR sequences are the direct descendants of a germline ancestor, the sequence of which is assumed to be known or inferred a priori. Both these conditions are illustrated graphically in Figure 3.1.

In order to represent the molecular evolution of long-lived B cell lineages more accurately, I develop here a new substitution model that models the effects of motif-specific mutation across BCR sequences. This model, named the HLP17 model, is a modification of the NY98 substitution model (more specifically, it is a modification of the NY98 model, because ω is kept constant among sites and lineages; [Nielsen and Yang, 1998]). Specifically, I add to the NY98 model an additional parameter,

h^a, which represents the change in relative substitution rate of a hotspot/coldspot mutation in motif a. Explicitly modelling the full context dependence of hotspot motifs would make likelihood calculations computationally infeasible. Instead, I weight h^a by b^a_{ij}, which is the probability that the mutation from codon i to codon j was a hotspot mutation in motif a, averaged across all possible combinations of codons on the 5′ and 3′ flanks of the target codon. This is a mean field approximation (i.e. the expected effect is averaged across all possible scenarios) and is similar to the singlet-doublet-triplet model of [Whelan and Goldman, 2004]. A 'hotspot mutation' is defined as a mutation occurring within the underlined base of the specified motif (e.g. the trimer motif and its reverse complement WR<u>C</u>/<u>G</u>YW; nucleotides represented using the IUPAC coding scheme with W = A or T, Y = T or C, and R = A or G).

In the HLP17 model, each entry q_{ij} in the transition rate matrix **Q** is parameterised by:

π_j = Baseline frequency of codon j.

κ = Transition/transversion mutation relative rate ratio.

ω = Nonsynonymous to synonymous mutation relative rate ratio.

a = Motif in which mutation rate is modified at underlined base. Here, $a \in$ (WR<u>C</u>, <u>G</u>YW, W<u>A</u>, <u>T</u>W, SY<u>C</u>, <u>G</u>RS), but in principle any other motif $\leq$ 4nt long could be used.

h^a = Change in mutability due to mutation in motif a; $h^a \geq$ -1.

and the transition matrix **Q** itself is defined by

$$q_{ij} = \begin{cases} 0 & \text{i}\Rightarrow\text{j} > 1 \text{ nucleotide change} \\ \pi_j(1 + \sum_a b_{ij}^a h^a) & \text{i}\Rightarrow\text{j synonymous transversion} \\ \kappa\pi_j(1 + \sum_a b_{ij}^a h^a) & \text{i}\Rightarrow\text{j synonymous transition} \\ \omega\pi_j(1 + \sum_a b_{ij}^a h^a) & \text{i}\Rightarrow\text{j nonsynonymous transversion} \\ \omega\kappa\pi_j(1 + \sum_a b_{ij}^a h^a) & \text{i}\Rightarrow\text{j nonsynonymous transversion} \end{cases} \tag{3.1}$$

where b_{ij}^a is the probability that a mutation from i to j involves the underlined base in motif a. The values of b_{ij}^a are calculated by marginalizing over all possible 5′ and 3′ flanking sense codons as follows:

$$b_{ij}^a = \sum_{k=1}^{61} \sum_{m=1}^{61} \pi_k \pi_m \text{I}(i,j,k,m,a) \tag{3.2}$$

where I is the indicator function:

$$I(i,j,k,m,a) = \begin{cases} 1 & kim\Rightarrow kjm \text{ is a mutation from motif } a \\ 0 & otherwise \end{cases} \tag{3.3}$$

This model, though an approximation, has several useful properties. Most importantly, because codon changes are modelled as occurring independently of each other, the phylogenetic likelihood can still be calculated using Felsensteins pruning algorithm, which greatly reduces computational time [Felsenstein, 1981]. This substitution model also has the intuitive property that, if no hotspot motif is specified, then all $h^a = 0$ and the model simplifies to the NY98 model. Thus the NY98 submodel of the NY98 model, when using the same given germline sequence, is a special case of the HLP17 model.

While in standard NY98-type models the vector $\boldsymbol{\pi}$ represents the equilibrium frequencies of codons, this is not the case for the HLP17 model. This can be checked

by direct calculation of the total flux in and out of a codon j; in general $\sum_{i \neq j} \pi_i q_{ij} \neq \sum_{i \neq j} \pi_j q_{ji}$ for HLP17 because the matrix b^a_{ij} is generally not symmetric in i and j. Reversible substitution models such as REV/GTR [Yang, 1994] fulfill the condition of reversibility ($\sum_{i \neq j} \pi_i q_{ij} = \sum_{i \neq j} \pi_j q_{ji}$), meaning the total rate of change into a particular character is equal to the total rate of change away from that character. This implies that the frequencies of each character are stable through time, and that $\boldsymbol{\pi}$ are the equilibrium frequencies of each character [Yang, 1994, Yang, 2014]. This condition is not necessarily fulfilled if the substitution matrix is not symmetric, so the $\boldsymbol{\pi}$ values in the HLP17 model cannot be interpreted as equilibrium frequencies. Although equilibrium frequencies do exist (and can be calculated numerically), we are in fact interested in the models non-equilibrium behaviour, since the ancestral sequence is likely to be far from equilibrium, and by the time the codons are observed the sequence is unlikely to have reached equilibrium. As a result, the best-fit values of $\boldsymbol{\pi}$ may even change according to the time at which a B-cell lineage is sampled. Thus the values of $\boldsymbol{\pi}$ in this model are more appropriately interpreted as best-fit constant codon frequencies given the data and other model parameters, and should not be directly interpreted as equilibrium frequencies. More specifically, I use the CF3X4 model as implemented in codonPhyML [Kosakovsky Pond et al., 2010, Gil et al., 2013] to find the best-fitting codon frequencies. In this model, the frequencies of A, C, G, and T at each of the three codon positions are estimated through ML as twelve additional parameters.

Because the NY98 model is a special case of the HLP17 model (i.e. when all h parameters = 0) the two models are nested and can be compared using a likelihood ratio test when fit to the same sequence data and rooted using the same germline sequence. Let $\ell(HLP17)$ and $\ell(NY98)$ be the maximum likelihoods obtained under the HLP17 and NY98 models, respectively. The likelihood ratio statistic (Equation 1.8) is then approximately chi-squared distributed with degrees of freedom equal to the number of additional h parameters [Huelsenbeck and Rannala, 1997].

For each bnAb dataset, I calculate $\ell(HLP17)$ by co-optimising h and other model parameters, whereas $\ell(NY98)$ is calculated by constraining all $h^a = 0$ whilst optimising the other model parameters. Importantly, because HLP17 uses a given germline sequence rather than codon frequencies during likelihood calculations at the root (see 3.2.2), it is only nested with NY98 if NY98 also uses the same sequence as its direct ancestor. Using codon frequencies in absence of a known root, as is usually done during likelihood calculations [Felsenstein, 1981], would make the likelihood values of these models incomparable through a likelihood ratio test.

Within this framework, a hierarchical network of hotspot models can be specified by fixing certain values of h^a to zero and by setting some values of h^a to be equal. For instance, a symmetric WR$\underline{\text{C}}$/$\underline{\text{G}}$YW model is specified by setting $h^{WR\underline{C}} = h^{\underline{G}YW}$ and by setting all other values of h^a to zero, leaving just one parameter ($h^{WR\underline{C}}$) to be estimated using maximum likelihood. Pairs of models that are nested (e.g. strand symmetric vs. asymmetric motifs) can be formally compared using likelihood ratio tests; non-nested models may be compared using the Akaike information criterion (AIC; [Akaike, 1974]).

3.2.2 Likelihood calculation and ancestral state reconstruction

Our implementation of the HLP17 model requires a non-standard approach to likelihood calculations. In contrast to most substitution models, the rate parameters in the **Q** matrix of the HLP17 model are not necessarily time-reversible, i.e. the model does not necessarily satisfy the detailed balance condition $\pi_i q_{ij} = \pi_j q_{ji}$. Time reversibility is useful because it means that likelihood calculations can be undertaken on an unrooted tree, which subsequently can be rooted on any branch. This property, known as the 'pulley principle', only holds for reversible models and helps to speed up search algorithms for maximum likelihood tree inference [Felsenstein, 1981] and maximum likelihood marginal ancestral sequence recon-

structions [Koshi and Goldstein, 1996, Bollback et al., 2007].

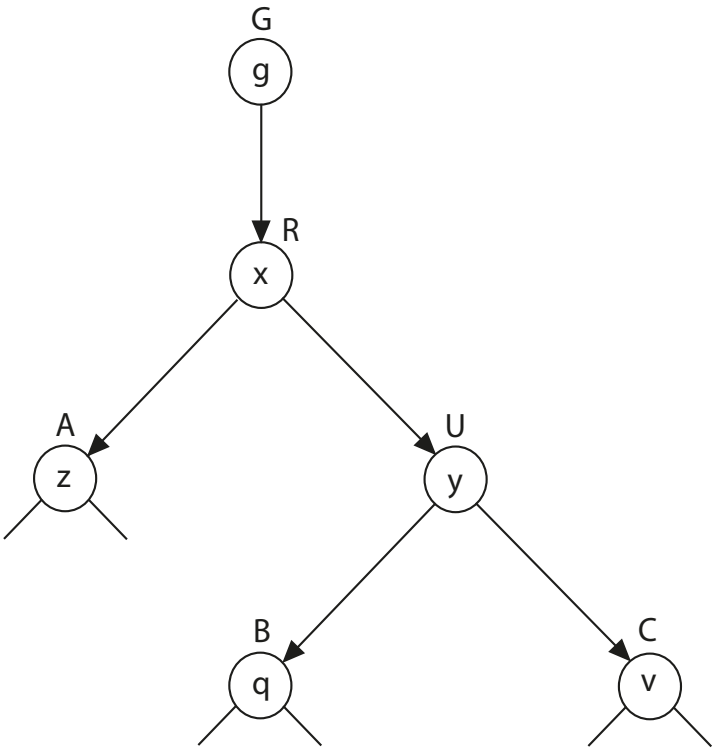


Figure 3.1: Proposed evolutionary model for antibody lineages. All sequences descend from given germline node G, which has sequence g. Arrows indicate the direction of evolutionary change. Note that this known ancestor G is not necessarily the most recent common ancestor of the lineage, which is node R and sequence x. See similarities to tree in Boussau and Gouy [Boussau and Gouy, 2006].

However, in the case of B cell lineage evolution it is necessary to root the phylogeny at the germline sequence during parameter estimation. Fortunately Boussau and Gouy [Boussau and Gouy, 2006] have introduced an algorithm that uses partial likelihood values to calculate the likelihood at each node in the tree for non-reversible models. This provides a similar increase in computational efficiency during likelihood evaluation to that provided by the pulley principle in reversible models. The algorithm I use for likelihood calculations is the same as that in [Boussau and Gouy, 2006], with the exception that I require tip sequences to descend from a known germline sequence (node G; Figure 3.1).

Using effectively the same formulas as in [Boussau and Gouy, 2006, Boussau et al., 2008], I use two partial likelihood functions - L_{low} and L_{upp} - which describe the partial likelihoods of the branches above and below a given edge. Given the tree in 3.1, for site s in the sequence alignment:

$$L_{low}(UC, v) = \begin{cases} 1 & \text{v = character at tip} \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

$$L_{low}(RU, y) = \sum_{q} P_{yq} L_{low}(UB, q) \sum_{v} P_{yv} L_{low}(UC, v) \tag{3.5}$$

$$L_{upp}(RU, x) = P_{gx} \sum_{z} P_{xz} L_{low}(RA, z) \tag{3.6}$$

$$L_{upp}(UB, y) = \sum_{x} P_{xy} L_{upp}(RU, x) \sum_{v} P_{yv} L_{low}(UC, v) \tag{3.7}$$

where P_{gx} is the probability of a transition from codon g to codon x along edge GR (Figure 3.1. Importantly, because the character at the root node G is given as g, equilibrium frequencies are not used at the root node during likelihood calculations. The likelihood for site s (L_s) at any edge, for instance UB, is then:

$$L_{s,UB} = P(X|G = g) = \sum_{y} L_{upp}(UB, y) \sum_{q} P_{yq} L_{low}(UB, q) \tag{3.8}$$

where $\boldsymbol{X}_s$ is the sequence data at site s. Using these formulas, it is possible to update the values of L_{upp} and L_{low} once for each edge in the tree, and then perform branch length optimization using only the values in L_{upp} and L_{low}, and the updated branch length. This substantially improves the efficiency of branch length optimization, compared to performing a pruning algorithm across the entire tree. As shown similarly in [Boussau et al., 2008] equation 3.8 can be modified slightly

to allow efficient relative likelihood calculations for a given codon q at node B:

$$P(X|G = g, B = q)_{UB} = \sum_{y} L_{upp}(UB, y) P_{yq} L_{low}(UB, q) \tag{3.9}$$

Marginal ancestral state reconstruction can then be performed by calculating this value for each possible value of q, and selecting the codon with the highest relative probability. Alternatively, the most likely amino acid can be found by summing the marginal probabilities for all codons which code for that amino acid, and selecting the amino acid with the highest probability. For analyses in this thesis I compare marginal amino acid reconstructions. For reference, proofs of equations 3.8 and 3.9, which are essentially the same as that in [Boussau and Gouy, 2006], are given in Appendix Figure B.1.

3.2.3 Implementation

I implement this substitution model and likelihood calculation framework in IgPhyML, a program modified from the source code of codonPhyML [Gil et al., 2013]. IgPhyML implements the rate matrix in equation 3.1 estimates the parameters h^q using maximum likelihood, together with the other model parameters. ML optimization is performed using the BFGS optimization algorithm that is implemented within codonPhyML [Broyden, 1970, Fletcher, 1970, Goldfarb, 1970, Shanno, 1970, Gil et al., 2013]. IgPhyML is available for download through: https://github.com/kbhoehn/IgPhyML.

3.2.4 bnAb lineage sequences and multiple sequence alignment

I applied the HLP17 model to the heavy chain sequences of the three bnAb lineages analysed in [Wu et al., 2015]. The lineage of greatest duration was VRC01, which was sampled over 15 years [Wu et al., 2015], followed by CAP256-VRC26

(hereafter VRC26), which was sampled over four years [Doria-Rose et al., 2014], and CH103, which was sampled over three years [Liao et al., 2013]. Sequences from each bnAb lineage were downloaded from GenBank and translated into amino acids, aligned to their putative germline V gene segment using IgBlast [Ye et al., 2013], and then re-translated back into codons. Putative germline segment assignments (V4-59*01 for CH103, V3-30*18 for VRC26, and V1-2*01 for VRC01) were obtained from bnAber [Eroshkin et al., 2013] and sequences were obtained from the IMGT V-Quest human reference set [Lefranc and Lefranc, 2001]. Because of considerable uncertainty in D and J germline assignments for each lineage, only the V segment was used. Insertions relative to the germline sequence were removed, so that all sequences within each lineage were aligned to the same germline sequence. Removing these insertions brought together two nucleotides that are not actually adjacent, creating false motifs. To prevent this, the 3' nucleotide of the region joined together from the removal of the insertion was converted to an 'N'. To keep results consistent among lineages, only nucleotide positions from the beginning of the first framework region (FWR1) to the end of FWR3 were used (see Figure 1.1. Sampling dates of each sequence were extracted from the sequence ID tags provided on GenBank. Eleven sequences were excluded from CH103 because this information was not available.

The three lineages showed substantial differences both in divergence and level of indels. While CH103 sequences showed no insertions relative to its germline sequence and VRC26 showed only two, VRC01 showed 394, many of which were the same length and occurred in the same position, and were thus likely inherited (i.e. synapomorphic rather than homoplasic). It is worth noting that indel polymorphisms often have important functional roles in bnAb lineages [Kepler et al., 2014a]. However, in this study I am primarily interested in the role of hot- and coldspot motifs in bnAb evolution, and there does not seem to be any compelling reason to believe removing indel regions would have a large overall effect

on the results of these analyses. In the final alignments for CH103, VRC26, and VRC01, 0.03%, 0.07%, and 0.5% of nucleotides were 'N' characters respectively; this state included gaps, germline-relative insertions, and ambiguous characters. Two sequences from VRC26 and two from VRC01 were removed due to difficulty in aligning them to the germline sequence, either because of high divergence or large indels. All three lineages showed high levels of root-tip divergence, with a mean of 27.6%, 21.0%, and 62.7% of codons differing between the root and tips for CH103, VRC26, and VRC01, respectively (Appendix Figure B.2).

3.2.5 Hotspot model selection

Because I want to compare ancestral sequence reconstruction for specific nodes (see Section 3.2.7), and because previous studies have reported difficulties in finding optimal tree topologies with non-reversible models [Boussau and Gouy, 2006], I chose to apply the HLP17 model to a fixed tree topology, and the problem of co-estimating topology will be addressed in Chapter 4. For each data set, the tree topology used was that inferred using the standard NY98 version of the NY98 model in codonPhyML, which was subsequently re-rooted in order to place the germline sequence as the common ancestor of all tips.

Because WR$\underline{C}$/$\underline{G}$YW hotspots are generally cited as among the strongest hotspot motifs in somatic hypermutation, and are generally thought to act in a strand symmetric manner [Smith et al., 1996, Teng and Papavasiliou, 2007, Yaari et al., 2013], the results of the symmetric WR$\underline{C}$/$\underline{G}$YW hotspot model fit are shown in detail in Table 3.1. By placing different constraints on the h^a parameters, I tested ten different hotspot models on the three bnAb lineages CH103, VRC26, and VRC01. The specific constraints used to define each hotspot model, and the results of model testing are shown in Table 3.3. Full results from each model fit are provided in Appendix Table B.1). Importantly, all models that were compared using likeli-

hood ratio tests used the same, pre-specified root sequence. The NY98 model was specified as the HLP17 model with all h^a fixed to zero.

Further, to ensure that the effects I observe are particular to the hotspot and coldspot motifs under investigation, I compared estimated h values for specific hotspot motifs to those obtained from all other possible trimer motifs with similar characteristics. Specifically, I generated all possible motifs and their reverse complements that (i) were 3nt in length, (ii) contained two IUPAC letters standing for two possible nucleotides (R, Y, S, W, K, and M), and (iii) subsequently contained an unambiguous nucleotide (i.e. A, C, G, or T). I then fit the HLP17 model using these each of these 144 motifs individually and compared how estimated h values for these motifs compared to the values for WR$\underline{C}$/$\underline{G}$YW and SY$\underline{C}$/$\underline{G}$RS. I repeated this process for dimer motifs but with the constraints that motifs (i) were 2nt in length, (ii) contained one IUPAC letter standing for two possible nucleotides and (iii) subsequently contained an unambiguous nucleotide. I fitted the HLP17 model to the same data using these 24 dimer motifs and compared them to the results from W$\underline{A}$/$\underline{T}$W motifs. Results from this analysis are shown in Appendix Figure B.3.

3.2.6 Effectiveness of the mean field approximation

I evaluated and validated the effectiveness of the HLP17 model by simulating data sets under different values of h and testing how accurately model parameters were inferred. For brevity, I considered only symmetric WR$\underline{C}$/$\underline{G}$YW hotspot motifs in this analysis ($h^{WR\underline{C}} = h^{\underline{G}YW}$; hereafter in this section hereafter referred to as h). Because the HLP17 model is a mean field approximation it will not fully account for the context dependency of somatic hypermutation. To measure the degree of this effect, I generated simulated datasets using a modified version of HLP17 that does fully account for the context dependence of adjacent codon sites. Specifically:

1. I estimated a maximum likelihood phylogeny for each bnAb lineage data set using the standard NY98 model. During estimation I optimised ω, κ, $\boldsymbol{\pi}$ vector, branch lengths and the tree topology. The resulting ML tree was re-rooted at the germline sequence with a branch length of zero.

2. For each value of h investigated (0, 1, 2, and 4), I simulated 20 alignments along each of these trees using a version of the model that fully accounts for context dependency. Specifically, in forward simulation, the 3′ and 5′ flanking codons of each site are known. This allowed us to create a **B** matrix for each site in each sequence with b_{ij} equal to either 1 or 0 depending on whether or not the substitution was a hotspot mutation in a WR$\underline{\text{C}}$/$\underline{\text{G}}$YW motif. The process begins at the given root sequence, and gives rise to a separate **B** and **Q** matrix at each site in the sequence. This leads to different substitution rates (**Q** matrices) for each site at each internal node. For each descendant branch, the matrix is multiplied by the branch length and then exponentiated to give the transition probability matrix; the next codon is selected randomly with respect to these probabilities. This process continues at each site to create two descendant sequences, then repeats for descendant nodes down the tree until all tips are filled. Simulations were undertaken using the estimated values of ω, κ, and $\boldsymbol{\pi}$ obtained in step (1) for the corresponding bnAb lineage data set.

3. For each of the replicates defined in step (2), I performed three different ML calculations: (i) h was optimised using ML (with $\hat{h}$ as the MLE estimate of h), (ii) h was fixed to zero and (iii) h was fixed to the true value used in simulation. These three scenarios enable us to test type 1 and type 2 error rates, by determining whether $\hat{h}$ was significantly different to h or to zero, respectively. Statistical significance was determined using the chi-squared approximation to the likelihood ratio statistic, as described above. In all calculations, the tree topology was fixed to that inferred in step (1).

4. For each data set and for each set of simulations under a particular value of h, I estimated $\hat{h}$ and then calculated the properties of this estimator as follows:
   (a) Bias in estimation: (Mean($\hat{h}$) - h)
   (b) Variance in estimation: Variance($\hat{h}$)
   (c) Type 1 error rate: The proportion of simulated data sets in which h was outside of the 95% confidence interval for $\hat{h}$.
   (d) Type 2 error rate: The proportion of simulated data sets in which $h > 0$, but failed to reject the null hypothesis ($h = 0$).

Biased mutation during somatic hypermutation has been shown to give false signatures of natural selection using approaches that compare the expected number of replacement and silent mutations [Dunn-Walters and Spencer, 1998]. I hypothesised that the HLP17 model might partially reduce this bias. To test this, and to explore whether the HLP17 model improved estimation of other evolutionary parameters, I compared the percentage error under the HLP17 and NY98 models of estimates of (i) ω, (ii) κ, (iii) tree length (sum of all branch lengths) and (iv) the ratio of internal to external branch lengths. These results are provided in Figure 3.2 and Appendix Figure B.4-B.5.

Because ω and κ are estimated along with h^a, it is possible that estimation of h^a may interfere with the other parameters, or vice-versa. To test this, I performed the fully context dependent simulation procedure again for the CH103 lineage, but with different values of $h^{WR\underline{C}/\underline{G}YW}$ (0, 1, 2 and 4) and ω (0.1, 0.5, 1, 2). I then fit two models to each data set - the null model using the true values of $h^{WR\underline{C}/\underline{G}YW}$ and ω, and the alternative model estimating both via maximum likelihood - and compared them using the likelihood ratio test. In these cases the alternative model introduced two free parameters (h and ω) compared to the null model, so the likelihood ratio test used a chi square test with two degrees of freedom. The results of this simulation analysis are given in Appendix Figure B.6-B.10. To see whether

this had any effect on hotspot model selection, I also fit an asymmetric WRC/GYW model for each simulated data set, to see if different values of h and ω led to more frequent rejection of the symmetric model. Further I compared the values of b^a_{ij}, ω, and κ within the **Q** matrix of each best-fit HLP17 model, to see if any significant relationship existed between the parameters (Appendix Figure B.6-B.10).

The fact that bnAb lineages cannot be expected to be in equilibrium when they are sampled has interesting implications for the use of Markov substitution models. Typically, it is assumed that nucleotide or codon frequencies are at equilibrium at the time of sampling, and empirical codon frequencies are often used as estimates of equilibrium frequencies. In the case of long-lived B cell lineages, however, sampled sequences are almost certainly not in equilibrium, making empirical codon frequencies inaccurate approximations for equilibrium frequencies. Because changes from SHM are not inherited through the germline, each BCR lineage is expected to begin out of sequence equilibrium, potentially converging to its equilibrium distribution as it evolves. To test how the use of empirical codon frequencies in such circumstances might affect estimation of h, I repeated the simulation procedure above using empirical equilibrium frequencies from each data set. These results are included in Appendix Figure B.2.

3.2.7 Effects on ancestral state reconstruction

One of the key applications of molecular phylogenetics to BCR sequence data is the reconstruction of ancestral sequences within a B-cell lineage [Kepler, 2013]. For each bnAb lineage, I found the amino acid with the highest marginal probability at each site of each node using equation 3.9. I performed this process for both the NY98 and best-fit HLP17 trees and parameters (i.e. those obtained in Section 3.2.5). This allowed us to compare whether the HLP17 models showed any tangible difference in maximum likelihood ancestral sequence reconstruction

over NY98. However, because the ancestral sequences for the bnAb lineages are not known, I repeated this analysis on the lineages simulated under full context-dependence, to see which model gave more accurate ancestral state reconstructions. I thus performed the same maximum likelihood ancestral sequence reconstruction procedure for each simulation replicate, and for two of the models (h=ml, h=0) described in step (3) in Section 3.2.6 above. These ancestral sequences were then used to compare the accuracy of reconstructions under the HLP17 model with those obtained using the NY98 model (h=0). In each simulation replicate, accuracy of ancestral sequence reconstruction was measured by calculating the mean number of amino acid differences between the predicted and true sequences at each node. These are shown in Appendix Figure B.12 and Appendix Table B.3, respectively.

3.3 Results

3.3.1 A codon substitution model for phylogenies undergoing somatic hypermutation

All three bnAb lineages showed a significant improvement in likelihood under the symmetric WR$\underline{C}$/$\underline{G}$YW HLP17 model compared to the NY98 model. The maximum likelihood values of h for the three data sets were $\hat{h}^{WR\underline{C}} = \hat{h}^{\underline{G}YW}$ = 1.86, 1.81, and 2.03, for CH103, VRC26, and VRC01, respectively. In each case the simpler NY98 model (all h=0) could be rejected using the likelihood ratio test ($p < 0.0001$ for all three lineages). These results are summarized in Table 3.1. These $\hat{h}$ values represent up to a three-fold increase in the relative rate of change at hotspot locations (depending on the values of $b_{ij}^{WR\underline{C}/\underline{G}YW}$).

The mean field approximation used in this model did not dramatically affect parameter estimation when applied to data sets simulated under a fully context de-

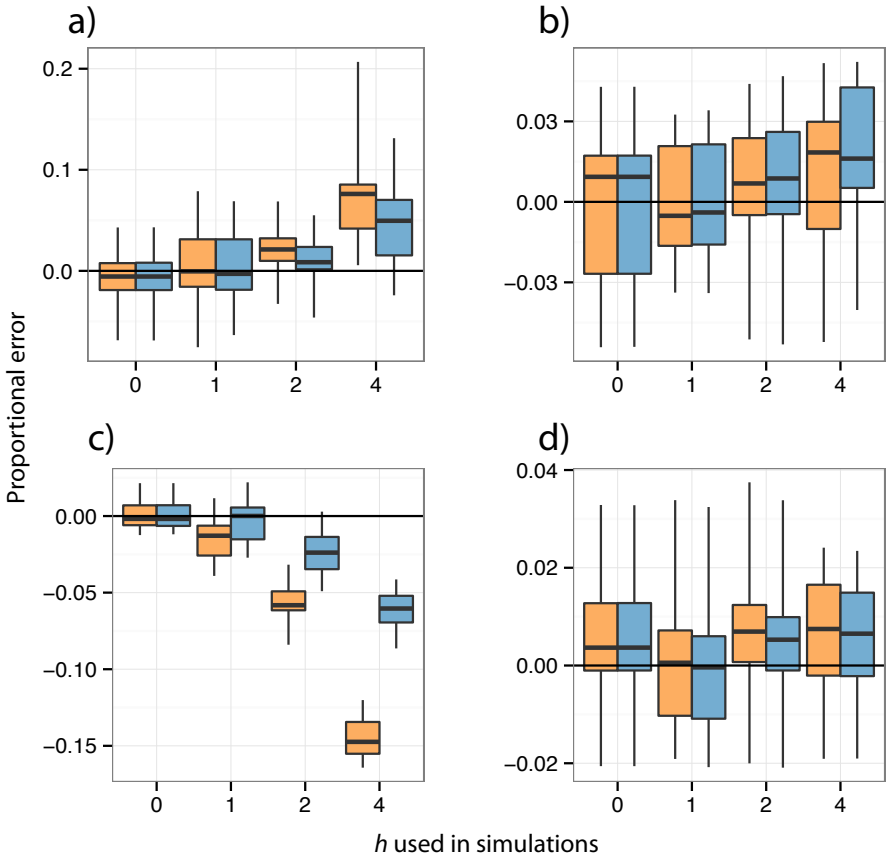


Figure 3.2: Proportional error in parameter estimation compared to true values for the VRC01 B cell lineage fully context dependent simulations. Values of ω, κ, tree length, and ratio of internal to external branch lengths are shown in panels a), b), c), and d), respectively. Estimates obtained under the NY98 are in orange (h=0) and estimates obtained under the HLP17 model are in blue (h estimated using maximum likelihood). The edges and centres of boxplots show the 1st, 2nd, and 3rd quartiles, while the whiskers show range. Similar results for B cell lineages CH103 and VRC26 are shown in Appendix Figure B.4-B.5.

pendent model, at least for the parameter space of the three empirical bnAb lineages (Table 3.2). Mean $\hat{h}$ values from simulations in which $0 \leq h^{WR\underline{C}/\underline{G}YW} \leq 2$ were close to their true h values and exhibited low absolute bias and variability (maximum -0.204 and 0.028, respectively, when h =2). Of these simulated data sets, 12.2% incorrectly rejected the correct parameter value (i.e. 12.2% of datasets simulated with $0 \leq h^{WR\underline{C}/\underline{G}YW} \leq 2$ had an ML estimate of h significantly different from the true value of h used in the simulations). The theoretical expectation under $\alpha = 0.05$ is that only 5% of such datasets would incorrectly reject that true value. It is unclear how to obtain principled performance boundaries, given that

Lineage	$\hat{h}^{WRC/GYW}$	Log-likelihood		2*LR	p value
		$h^{WRC/GYW}$=mle	$h^{WRC/GYW}$=0		
CH103	1.86 (1.4, 2.4)	-14600.4	-14702	203.2	$< 1x10^{-15}$
VRC26	1.81 (1.5, 2.1)	-37238.5	-37516.4	555.8	$< 1x10^{-15}$
VRC01	2.03 (1.7, 2.4)	-43647.1	-43945.3	596.4	$< 1x10^{-15}$

Table 3.1: Significance was determined using the likelihood ratio test under a chi-squared distribution with one degree of freedom. 90% confidence intervals for $\hat{h}$ are shown in parentheses in the second column.

the HLP17 model is a biased approximation, but 12.2% appears to be reasonably good performance. Further, none of the datasets simulated with $h > 0$ failed to reject the null hypothesis that $h = 0$, demonstrating good statistical power. Bias generally increased if h was raised beyond that observed in the empirical bnAb linages. Performance was worse when $h = 4$, which resulted in a mean type 2 error of 0.62 and a mean bias of -0.35. This behaviour is as expected because, as h increases, the mean field approximation will become less accurate. I found that using empirical codon frequencies decreased the performance of h estimation; using empirical frequencies resulted in higher bias and type 2 error rates than using ML frequencies (Appendix Figure B.2). Discussion of why empirical codon frequencies are unlikely to be suitable for long-lived B-cell lineage phylogenies is provided in the Methods section 3.2.1.

Within the parameter space of the empirical data sets ($0 \leq h^{WRC/GYW} \leq 2$), there was no substantial difference in estimation of other model parameters compared to the standard NY98 model, except for the tree length parameter in some simulations (Figure 3.2, Appendix Figure B.4-B.5). However, when this h is large (4, in these simulations), the NY98 model substantially underestimates tree length in each of the simulated lineages. In contrast, the HLP17 model, while not completely eliminating this effect, substantially reduced it. In simulations based on the long-lived VRC01 and VRC26 lineages in which this $h = 4$, the NY98 model

Set	h	Mean $\hat{h}$	Bias	Variability	Type 1 error	Type 2 error
CH103	0.00	-0.012	-0.012	0.008	-	0.10
	1.00	1.091	0.091	0.018	0.00	0.05
	2.00	2.095	0.095	0.028	0.00	0.05
	4.00	4.146	0.146	0.093	0.00	0.05
VRC26	0.00	0.004	0.004	0.003	-	0.05
	1.00	0.965	-0.035	0.005	0.00	0.00
	2.00	1.841	-0.159	0.013	0.00	0.25
	4.00	3.501	-0.499	0.026	0.00	0.85
VRC01	0.00	-0.002	-0.002	0.002	-	0.05
	1.00	0.965	-0.035	0.008	0.00	0.05
	2.00	1.796	-0.204	0.022	0.00	0.50
	4.00	3.291	-0.709	0.049	0.00	0.95

Table 3.2: HLP17 performance under fully-context dependent simulations for symmetric WRC/GYW hotspots. Maximum likelihood estimates of h and likelihood ratio tests for symmetric WRC/GYW model. Type 1 error rate shows the proportion of data sets that incorrectly failed to reject the null hypothesis of $h = 0$. Type 2 error rate shows the proportion of data sets that rejected the true value of h shown in the first column. Both hypothesis tests for type 1 and 2 errors used an alpha value of 0.05. Importantly, data in these analyses were not simulations under HLP17, but a fully context dependent variation of it.

overestimated the ω parameter; this bias was not obvious in simulations based on the CH103 lineage, which was sampled for a shorter duration. The HLP17 model performed slightly better in ω estimation under these circumstances. Interestingly for data sets CH103, HLP17 showed a higher error in κ for $h = 4$.

The distribution of b_{ij} weights for individual motifs, and to a lesser extent combined symmetric motifs, were noticeably different for synonymous and non-synonymous substitutions (Appendix Figures B.10 and B.11). To see how this relationship may affect my results, I simulated data sets under the symmetric WRC/GYW model with different values of h and ω, and fit both symmetric and asymmetric WRC/GYW models to each data set. Interestingly, under the symmetric model I found no consistent pattern of rejecting the true values of h and ω as these parameters increased. The asymmetric model did, however, did show an increase in incorrectly rejecting the symmetric model as h and ω increased. This appears to

have been driven by bias in $h^{\underline{G}YW}$, which is overestimated, and to a lesser extent $h^{WR\underline{C}}$, which was underestimated, as h and ω increased (Appendix Figure B.7). Interestingly, there was no such consistent bias in $h^{WR\underline{C}/\underline{G}YW}$ under the symmetric WR$\underline{\text{C}}$/$\underline{\text{G}}$YW motif, and neither symmetric nor asymmetric models showed a robust bias in ω (Appendix Figure B.8).

	Constraint/optimization of each h^a						p values from LR tests		
Model name	$h^{\underline{WRC}}$	$h^{\underline{GYW}}$	$h^{\underline{WA}}$	$h^{\underline{TW}}$	$h^{\underline{SYC}}$	$h^{\underline{GRS}}$	CH103	VRC26	VRC01
Symmetric WRC/GYW*	ML	$h^{\underline{WRC}}$	0	0	0	0	$4.6\text{x}10^{-15}$	$6.3\text{x}10^{-5}$	$5.2\text{x}10^{-3}$
Asymmetric WRC/GYW	ML	ML	0	0	0	0			
Symmetric WA/TW*	0	0	ML	$h^{\underline{WA}}$	0	0	$<1\text{x}10^{-15}$	$<1\text{x}10^{-15}$	$<1\text{x}10^{-15}$
Asymmetric WA/TW	0	0	ML	ML	0	0			
Symmetric SYC/GRS*	0	0	0	0	ML	$h^{\underline{SYC}}$	0.53	$2.0\text{x}10^{-13}$	$4.2\text{x}10^{-3}$
Asymmetric SYC/GRS	0	0	0	0	ML	ML			
Uniform hotspots*	ML	$h^{\underline{WRC}}$	$h^{\underline{WRC}}$	$h^{\underline{WRC}}$	0	0	$4.2\text{x}10^{-15}$	$<1\text{x}10^{-15}$	$<1\text{x}10^{-15}$
Hierarchical hotspots	ML	$h^{\underline{WRC}}$	ML	$h^{\underline{WA}}$	0	0			
SCAH*	ML	ML	ML	ML	ML	$h^{\underline{SYC}}$	0.65	$1.2\text{x}10^{-6}$	$9.1\text{x}10^{-4}$
FCH	ML	ML	ML	ML	ML	ML			

Table 3.3: Hotspot model selection. Models of hotspot hierarchy (degree of mutability) and symmetry, specified by placing constraints on how different values of h are optimized. Columns 2-7 show how the parameter h^a is obtained for a particular model. A value of '0' indicates that h is fixed at zero, 'ML' indicates that a parameter is optimised by maximum likelihood, and h^a indicates that h parameter is equal to another value of h. For instance, in 'Symmetric WR$\underline{C}$/$\underline{G}$YW', $h^{\underline{GYW}}$ is equal to its reverse complement $h^{\underline{WRC}}$, which is ML optimised. However, in 'Asymmetric WR$\underline{C}$/$\underline{G}$YW', both are ML optimised. Note that each model marked with an asterisk * is nested with the model immediately below it by one free parameter, allowing hypothesis testing using a likelihood ratio test. Rows 8-10 show p values obtained from likelihood ratio tests of each of these nested hotspot models for the bnAb lineage specified in each column. SCAH = symmetric coldspots, asymmetric hotspots; FCH = free coldspots and hotspots. Parameters, log likelihood, and AIC of each fit are shown in Appendix Table B.1.

3.3.2 Hotspot model selection

All hotspot motif models tested gave a significantly higher likelihood than the standard NY98 model when applied to the CH103, VRC26, and VRC01 B-cell lineages. Likelihoods were considerably higher for asymmetric models. Using a LRT, the asymmetric WR$\underline{C}$/$\underline{G}$YW model showed a significantly better fit than the corresponding nested symmetric model (p = $4.6x10^{-15}$, $6.3x10^{-5}$, and $5.2x10^{-3}$, for lineages CH103, VRC26, and VRC01, respectively). Similarly the asymmetric W$\underline{A}$/$\underline{T}$W model fitted the data better than its symmetric counterpart (p $< 1x10^{-15}$ for all three lineages). Allowing different hotspot motifs to have different h values also resulted in significantly higher likelihoods than using a uniform value of h for all hotspots (p $< 1x10^{-14}$ for all three lineages). Interestingly, VRC26 and VRC01 showed a significantly higher likelihood under asymmetric SY$\underline{C}$/$\underline{G}$RS coldspot motifs (p = $2.0x10^{-13}$ and $4.2x10^{-3}$), but CH103 did not (p=0.53). This difference was also reflected in the best-fit (lowest AIC) model for each lineage. For VRC26 and VRC01the best-fit model was the 'Free coldspots and hotspots' model, in which all motifs and their reverse complements are given separate h values. However, for CH103 the best-fit model was the 'Symmetric coldspots, asymmetric hotspots' model, in which each hotspot and its reverse complement are given separate h values, but coldspots remain symmetric.

The randomization analysis confirmed that the hotspots I selected do indeed evolve differently to similarly constrained random motifs. Known hotspots were hotter, and known coldspots were colder, than nearly all such random motifs. Further, random motifs that showed better fit than known hotspots were similar to known hotspots, differing only in the site furthest from the mutable position. WR$\underline{C}$/$\underline{G}$YW motifs exhibited a larger value of h, and a higher likelihood, than any other trimer motif analysed. Further, SY$\underline{C}$/$\underline{G}$RS motifs resulted in h values that were lower than 142 of the 143 other trimer motifs tested (only KYC/GRM showed a lower h, and was the only other trimer motif that showed a higher likelihood and negative

h). W$\underline{A}$/$\underline{T}$W motifs showed a higher *h* value than all other dimer motifs analysed, and a higher likelihood than 21 of the 23 other dimer motifs (only RC/GY and YC/GR showed higher likelihoods). These results are shown in Appendix Figure B.3.

3.3.3 Ancestral state reconstruction

In fully context dependent simulations, I also found that the HLP17 model provided an accuracy of ancestral state reconstructions that was similar to the NY98 model where $h < 4$, and that HLP17 noticeably improved accuracy at $h = 4$ (Appendix Table B.3). While true ancestral sequences are not available for the three empirical bnAb lineages, I did observe differences between ancestral sequences reconstructed using the HLP17 and NY98 models. For each lineage, I compared the two models by calculating the mean number of amino acid differences between the predicted ancestral sequences at all internal nodes of each tree. Performing this ancestral state reconstruction on each of the three bnAb lineages showed a mean of 0.04, 0.1, and 0.09 amino acid sequence difference across all internal nodes, with a maximum difference of 3, 2, and 4 amino acid differences in a single node for CH103, VRC26, and VRC01, respectively. Differences were somewhat more concentrated in the basal third of the phylogeny. Typically, I would expect the uncertainty in ancestral state reconstruction to increase as I move from the tree tips towards the root; however, B-cell lineages are unusual in that the root sequence is also known as it corresponds to the germline sequence.

3.4 Discussion

Molecular phylogenetics has already been used in a variety of applications in the study BCR genetic diversity and the molecular evolution of B cell lineages [Kepler, 2013, Sok et al., 2013, Kepler et al., 2014b]. However, the process of somatic

hypermutation is known to occur in ways that violate fundamental assumptions of most phylogenetic substitution models. Here, I demonstrate that failing to account for this has tangible effects on phylogenetic inference from sets of sequences from long-lived bnAb lineages. I develop and implement a new codon substitution model (HLP17) that, whilst only an approximation, is capable of mitigating these effects.

Perhaps the most salient difference between standard substitutions models and the biology of somatic hypermutation is the context dependency of mutation in BCRs. This biased mutation process at hotspot motifs, for which a variety of empirical models have been developed to characterise the process at di, tri, penta, and heptamer levels [Smith et al., 1996, Yaari et al., 2013, Elhanati et al., 2015], has long been known to give false signature of selection in BCRs [Dunn-Walters and Spencer, 1998]. This effect was observed in some of my simulations (Figure 3.2, Appendix Figure B.4-B.5), as a failure to account for the increased rate of substitution at hotspot motifs led to overestimation of the ω (dn/ds) parameter. However, these simulations used an h value of 4, which was outside of the range of what I observed for empirical bnAb lineages.

Other approaches have been taken to study the effect of context dependent mutation in phylogenetic substitution models. Many have focused on modelling the substantially increased mutability of CpG motifs [Lunter and Hein, 2004]). These approaches are attempts to account for the full context dependency of CpG hypermutation, and require significantly more complex models. In the case of somatic hypermutation in BCRs, the increased mutability of BCR hotspot mutations (˜3 fold) is not as great as CpG motifs (˜18 fold; [Lunter and Hein, 2004], so a simpler, approximate approach is still effective (Table 3.2). The mean-field approximation has also been used previously, but in a reversible codon model, to take into account di- and trinucleotide substitutions [Whelan and Goldman, 2004].

The HLP17 codon substitution model detailed here is a relatively straightforward modification of the widely used NY98 submodel of the NY98 model. Although it requires special techniques for efficient likelihood computation and ancestral state reconstruction, I expect it to be not only useable, but statistically preferable to the NY98 model when applied to any BCR data set whose diversity may have been shaped by somatic hypermutation. Further, the HLP17 model does not rely on an empirical model to incorporate the effect of biased mutation, but instead attempts to explicitly model the context-dependent mutational process by estimating the parameter(s) h^a directly from the data. I note, however, that the HLP17 model is a mean-field approximation and does not capture the full context of motif driven evolution. Therefore I do not expect it to fully disentangle interactions between selection and biased mutation, and estimated values of ω should be interpreted carefully. Importantly, empirical analyses on bnAb lineages performed here were using tree topologies that were optimal under NY98, rather than HLP17. This is expected to make the estimation of each h conservative in these analyses, but it is not clear how the optimal topology of the HLP17 model will differ from that under NY98.

Our model selection results suggest that different hotspot motifs have highly variable effects on sequence evolution in B-cell lineages. It is generally thought that increased mutation in WRC/GYW motifs (or the tetramer motifs WRCY/RGYW) reflect the action of AID targeting, while in WA/TW motifs it is the result of error-prone polymerase repair [Teng and Papavasiliou, 2007]. Consistent with these separate mechanisms, WRC/GYW motifs have generally been found to be strand symmetric, but WA/TW motifs are strand-biased, with WA mutating at a higher rate than TW [Bransteitter et al., 2004, Spencer and Dunn-Walters, 2005, Teng and Papavasiliou, 2007]. It is interesting, then, that in all three lineages tested here show a significantly better fit for asymmetric vs. symmetric WRC/GYW (Table

3.3; Appendix Table B.1). However, I believe these results do not necessarily conflict with previous findings on the targeted nature of SHM, and can be explained by other factors. Our simulation analyses using the CH103 lineage show that increasing values of h and ω can lead to an artificially high $h^{\underline{G}YW}$, an artificially low $h^{WR\underline{C}}$, and an increased rate of incorrectly rejecting the symmetric WR$\underline{C}$/$\underline{G}$YW model in favour of the asymmetric version (Appendix Figure B.6 and B.7). This may be due to how, in the **Q** matrix, ω interacts differently between $b^{WR\underline{C}}$ and $b^{\underline{G}YW}$ (Appendix Figure B.10). Importantly, this bias appears to be limited to the h parameter in the asymmetric model, as the same simulations showed no consistent bias in ω for the symmetric model, and only a weak bias for asymmetric model (Appendix Figure B.8). Further, increasing ω in these simulations and analysing them using the symmetric WR$\underline{C}$/$\underline{G}$YW model did not reveal bias in $h^{WR\underline{C}/\underline{G}YW}$ (Appendix Figure B.7), or an increased rejection of the true values of ω and $h^{WR\underline{C}/\underline{G}YW}$ (Appendix Figure B.9). Thus, while the symmetric WR$\underline{C}$/$\underline{G}$YW model likely does fit better due to the increased mutability of WR$\underline{C}$/$\underline{G}$YW motifs (see also Appendix Figure B.3), the biological significance of its rejection in favour of the asymmetric WR$\underline{C}$/$\underline{G}$YW model should be interpreted with caution.

Another common assumption in phylogenetic analysis is that the codons or nucleotides sampled for analysis are at their equilibrium frequencies throughout the phylogeny. Because codon frequencies in the B cell's germline ancestor may be different from the equilibrium frequencies of the somatic substitution process, the somatic evolutionary process may begin out of equilibrium. As a result, mean codon frequencies may change through time within a B-cell lineage, until convergence. This phenomenon is not related to the non-time-reversibility of this model; it will occur even under a time-reversible model. I dealt with this problem in two ways. The first is in only using codon frequencies in the **Q** matrix, rather than also using them at the root during likelihood calculations. This effectively separates the codon frequencies from the germline substitution process (the root sequence)

from the somatic substitution process (the **Q** matrix). For simplicity I estimated the codon frequencies of the somatic substitution process by maximum likelihood within the model. This provided an improvement, both in maximum likelihood and in parameter estimation, over using empirical codon frequencies (Appendix Figure B.2). However, it is not yet clear if this is the most efficient or the most effective way of dealing with the problem of sequences that have not converged to their equilibrium distribution. While ML optimization finds the best fitting single set of codon frequency values (under the CF3X4 model), in reality codon frequencies may change over the course of the phylogeny, and a model that accounts for that would likely be more appropriate. While it should be possible to calculate the expected change in codon frequencies over a defined branch length given a non-reversible substitution model, I leave that problem for future analyses.

The decay of hotspot motifs in bnAb lineages may have important implications for our understanding of host-virus coevolution. More specifically, the loss of hotspot motifs may lead to a decrease in sequence mutability, and therefore a decline in overall rate of evolution over time for a given lineage [Sheng et al., 2016]. This hypothesis has several interesting implications. If the slowdown in mutation rate over time, arising from hotspot decay, is an intrinsic property of activated B cell lineages, then BCR sequence divergence from a germline ancestor (and thus affinity maturation) may be intrinsically constrained. Consequently, while BCR lineages may be able to rapidly evolve binding affinity and co-evolve with pathogens for an initial period after activation, over longer periods of time the ratio of the rate of BCR sequence change to pathogen sequence change may decline. I hypothesise that in extreme cases the rates of BCR evolution within a lineage may eventually fail to keep up with the rapid evolution of chronically infecting viruses, such as HIV-1, due to the exhaustion of available BCR hotspot motifs. The notion that biased mutation will lead to decreased mutability and evolutionary rate was explored recently by Sheng et al. [Sheng et al., 2016]. They concluded that the

observed mutation rate decreases in bnAb lineages was most likely due to a shift from positive to purifying selection, although the loss of hotspot motifs may also play a role and the issue is not yet fully resolved.

I have implemented this model in the software package IgPhyML, a modified version of codonPhyML [Gil et al., 2013]. This program can perform all of the substitution model analyses performed here. Source code is available at: https://github.com/kbhoehn/IgPhyML.

Chapter 4

Semi-empirical substitution models

4.1 Introduction

As detailed in Chapter 3, both the mutational and selective processes of B cell evolution are different in important ways from the evolution of germline genes in organismal populations. Most saliently, the mutational process that drives affinity maturation - somatic hypermutation (SHM) - is sequence context sensitive and does not occur in a reversible manner. Further, antibody sequences are strongly shaped by different degrees of selection, both among sites in their sequences and through time [Sheng et al., 2016]. Most phylogenetic substitution models are unable to take these differences into account, particularly context sensitivity, due to the assumptions built into them, such as time reversibility and independent evolution at each nucleotide. To address these issues, I introduced in Chapter 3 a non-reversible codon substitution model, HLP17, that allows for non-reversible mutation of AID hot- and coldspot motifs.

The HLP17 model incorporated some context-dependent mutation biases that are characteristic of SHM by estimating the additional parameter h^a, which is the change in mutability for motif a. Motif a can, in theory, correspond to any arbi-

trary motif less than or equal to four nucleotides in length. However, in Chapter 3 I only considered previously characterized hotspots (e.g. a = WR$\underline{\text{C}}$), for which $h^a > 0$, and coldspots (e.g. SY$\underline{\text{C}}$), for which $-1 \leq h^a < 0$ and (W=A/T, R=A/G, S=G/C, Y=C/T; only the underlined C in both example motifs experiences altered substitution rate). If a codon substitution occurs within motif a (e.g. AAC to AAG is a WR$\underline{\text{C}}$ substitution) its instantaneous rate of substitution is modified by h^a. However, motifs greater than 1 nucleotide in length may span multiple codons. This presents a problem because sites within a multiple sequence alignment must be treated as evolving independently [Felsenstein, 1981], so it is not possible for the model to 'know' the adjacent codons of a particular site when performing these calculations. In Chapter 3, I detailed how this problem may be addressed using a mean field approximation, in which, for each codon substitution i to j, all possible combinations of flanking codons are enumerated and weighted by their predicted frequencies, and h^a is multiplied by the proportion of these combinations that occur within motif a. Thus, the instantaneous rate from codon substitution i to j is modified by the change in mutability of motif a (h^a), which is multiplied by the probability i to j occurred within motif a.

An implicit assumption in the HLP17 model is that the effects of SHM are discrete; namely, the effects of context dependent mutation bias (h^a) are modelled as occurring only and uniformly within predefined motifs. For example, the substitutions AGC to AGG, and AAC to AAG, are both WR$\underline{\text{C}}$ mutations and therefore experience the same modification in substitution rate from $h^{WR\underline{C}}$, while the substitution ATC to ATG is not a WR$\underline{\text{C}}$ motif and experiences no modification at all. This discretizing assumption is useful in that it allows for the estimation of context-dependent substitution bias of previously defined motifs with a modest number of additional parameters (at most one per motif), as was performed in Chapter 3 for the set six canonical motifs (WR$\underline{\text{C}}$, $\underline{\text{G}}$YW, W$\underline{\text{A}}$, $\underline{\text{T}}$W, SY$\underline{\text{C}}$, and $\underline{\text{G}}$RS). However, this assumption is only an approximation, as the context dependent ef-

fects of SHM have been shown to vary substantially both within and outside of the canonical motifs [Yaari et al., 2013].

Relaxing the assumption of SHM operating on a small set of discrete motifs has the potential to create a more biologically realistic substitution model. One way of doing this would be to extend the HLP17 model defined in Chapter 3 to estimate a separate value of h^a for all possible motifs of a given size. This would add an enormous number of parameters, making the model impractical for most data sets. For instance, if we model all trimer motifs starting or ending with mutable nucleotides (e.g. WRC and GYW) that would add 64*2 = 128 additional h parameters.

Fortunately, modelling the effects of SHM by estimating a separate parameter for each motif is unnecessarily naive. A number of previous studies have characterized the targeting preferences of SHM in order to develop empirical models of SHM targeting. The best-characterized of these models is known as the S5F (Synonymous, 5-mer, Functional; [Yaari et al., 2013]). This is an empirical model consisting of all 1024 possible nucleotide pentamers and the relative mutability of their central nucleotide. This model was produced by assembling a large dataset of BCR repertoire sequences from seven humans, separating the sequences into clones, aligning them to their predicted germline sequences, and using this data to calculate the relative probability that the central nucleotide in each pentamer was synonymously mutated (this mitigates bias from nonsynonymous vs synonymous substitutions; [Yaari et al., 2013]). The mutability of a pentamer motif is defined as the 'probability of the central base in the motif being targeted for SHM relative to all other motifs' [Yaari et al., 2013]. These mutability values are shown graphically in Figure 4.1a. While the S5F model is not a phylogenetic model, in this chapter I will show how its mutability information can be incorporated into the HLP17 model to describe a much wider range of context dependence in SHM

with a minimal number of additional free parameters.

The HLP17 model makes a similar discretizing assumption when dealing with amino acid substitutions. As with the classic NY98 model [Nielsen and Yang, 1998] on which it was based, the HLP17 model in Chapter 3 considers only whether a codon is substituted with the same amino acid (synonymous substitution) or a different amino acid (nonsynonymous substitution). This is useful because the ratio of nonsynonymous to synonymous substitution rates (ω) is often used to make inferences relating to the effects of natural selection [Yang and Swanson, 2002]. However, assuming a single ω for all codons may not be biologically realistic because amino acids have varying degrees of functional similarity and are thus not expected to be equally exchangeable [Goldman and Yang, 1994]. A naive means of incorporating amino acid exchangeability would be to estimate a separate exchangeability parameter for each amino acid, which would add an additional 179 free parameters to the model (assuming exchange rates are time reversible). Similar to the problem of context dependence in SHM, however, it should be possible to incorporate outside empirical information to solve this problem with only a few additional free parameters.

The approach explored in this chapter uses previously defined empirical models of amino acid substitution to describe amino acid exchangeability. These are substitution models, similar to HLP17, but describe the rates at which different amino acids substitute each other. The $\mathbf{Q}$ matrix of amino acid substitution models can be separated into a 20x20 exchangeability matrix (usually symmetric with 180 substitution rate parameters) that describes the relative instantaneous rates at which each amino acid substitution occurs, and a vector of the 20 amino acid frequencies. The exchangeability of two amino acids is thought to reflect their degree of functional similarity. Because these models have so many parameters, their parameters are usually estimated from large datasets of sequences from mul-

tiple lineages, and then fixed as empirical estimates in practice. For instance, the parameters of WAG amino acid model [Whelan and Goldman, 2001] were estimated from a large dataset of multiple protein families in order to describe the exchangeability of amino acids very generally. These parameters were then fixed, and the resulting model can be used as a substitution model to estimate phylogenetic trees from amino acid sequences without re-estimation of any of its rate parameters. Other amino acid models have been developed for more specific contexts. For instance, the AB model has a similar structure to the WAG model but its parameters were estimated using a large dataset of B cell receptor sequences, and is thus intended specifically for use in reconstructing antibody lineage trees [Mirsky et al., 2015]. An illustration of the WAG and AB exchangeability matrices are shown in Figure 4.1b and c. Other empirical amino acid exchangeability models include those that have been developed to characterize substitution in specific domains of life, such as retroviruses [Dimmic et al., 2002], or to incorporate different features of substitution models, such as rate variation among sites [Le and Gascuel, 2008]. Others have developed empirical codon substitution models using similar techniques as empirical amino acid models, though these have far more parameters (61x61 matrices) and require large datasets for their initial parameter estimation [Kosiol et al., 2007]

In this chapter, I will show how the HLP17 framework introduced in Chapter 3 may be augmented with empirical information at both the nucleotide and amino acid level. This leads to the generation of two classes of HLP17 models - 'discrete' (Chapter 3), and 'semi-empirical' (this chapter). I explore the different strengths of each model class: parameter estimates of discrete models are more readily interpretable, while semi-empirical models are more biologically informed and may give higher maximum likelihood estimates. Further, I explore here the effects of these models on estimation of phylogenetic topology. This was not addressed in

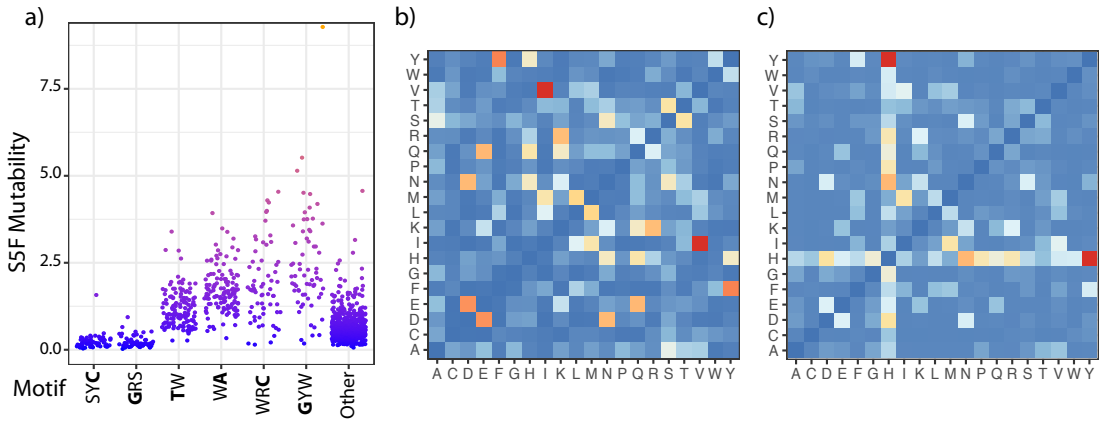


Figure 4.1: Graphical representation of empirical antibody motif and amino acid models. a) S5F mutability estimates for the central nucleotide in all possible pentamers. Pentamers are grouped by the type of canonical motif (if any) that they fall under. The canonical motifs listed mutate at the nucleotide in boldface. Circles are also coloured by mutability. b) WAG amino acid exchangeability matrix, showing relative instantaneous substitution rates for each amino acid (from blue to red, lowest to highest). c) Similarly, the AB amino acid exchangeability matrix (see Section 4.1).

Chapter 3, due to the issue of inefficient topology searching under a non-reversible model. Here, I resolve that problem by implementing algorithms for efficient topology searching under a non-reversible model as described in Boussau and Gouy [Boussau and Gouy, 2006], and explore the effects of different HLP17 models on the topology of these lineages.

4.2 Materials and Methods

4.2.1 Discrete and semi-empirical substitution models

A phylogenetic codon substitution model is defined by its instantaneous rate matrix, **Q**, which is a 61 x 61 matrix that describes the instantaneous rates at which each sense codon changes to each other sense codon. The **Q** matrix of the HLP17 model detailed in Chapter 3 can be generalized by using H_{ij} to describe the effects of sequence mutability, and O_{ij} to describe rates of change at the amino acid level. Specifically, the instantaneous rate matrix **Q** can be defined as:

$$q_{ij} = \begin{cases} 0 & \text{i}\Rightarrow\text{j} > 1 \text{ nucleotide change} \\ \pi_j H_{ij} & \text{i}\Rightarrow\text{j synonymous transversion} \\ \kappa\pi_j H_{ij} & \text{i}\Rightarrow\text{j synonymous transition} \\ \pi_j H_{ij} O_{ij} & \text{i}\Rightarrow\text{j nonsynonymous transversion} \\ \kappa\pi_j H_{ij} O_{ij} & \text{i}\Rightarrow\text{j nonsynonymous transversion} \end{cases} \tag{4.1}$$

In the discrete HLP17 (Chapter 3), O_{ij} and H_{ij} are equal to:

$$O_{ij} = \omega \tag{4.2}$$

$$H_{ij} = 1 + \sum_a h^a \sum_{k=1}^{61} \sum_{m=1}^{61} \pi_k \pi_m \begin{cases} 1 & kim\Rightarrow kjm \text{ is a mutation from motif } a \\ 0 & otherwise \end{cases} \tag{4.3}$$

Where ω is the nonsynonymous to synonymous mutation relative rate ratio, and a is the motif in which mutation rate is modified (at the underlined base). Usually $a \in$ (WR$\underline{\text{C}}$, $\underline{\text{G}}$YW, W$\underline{\text{A}}$, $\underline{\text{T}}$W, SY$\underline{\text{C}}$, $\underline{\text{G}}$RS), but in principle any other motif $\leq$ 4nt long could be used. h^a is the change in mutability due to mutation in motif a; $h^a \geq -1$. Note that if all $h^a = 0$, then $H_{ij} = 1$ for all i and j, and, if Eq. 4.2 is used for O_{ij}, Eq. 4.1 collapses to the NY98 model.

The original form of the HLP17 model (Equation 4.3) uses a discrete value of hotspot mutability (h^a); namely, changes in mutability due to sequence context are modelled in a binary manner, based on presence or absence of a particular motif (e.g. WR$\underline{\text{C}}$). To create a model that describes a wider variety of motifs, I incorporate the S5F (section 4.1) empirical mutability information by using it in place of h^a parameters in the discrete model:

$$H_{ij} = 1 + \sum_{k=1}^{61} \sum_{m=1}^{61} \pi_k \pi_m (y_h + m_h M_{kim,kjm}) \tag{4.4}$$

Under Eq. 4.1, all 9-mers *kim* and *kjm* considered in Eq. 4.4 differ by one and only one position (site d). $M_{kim,kjm}$ is the relative mutability of the nucleotide at site d in the 9-mer *kim*, given its sequence context. Here, these values are the mutability estimates from the S5F model using the pentamer from positions d-2 and d+2 in 9-mer *kim* ; however, they could in principle be taken from other similar mutability models. I define y_h and m_h as scaling coefficients (similar to the slope and intercept of a linear model) that are estimated by maximum likelihood and transform the S5F mutability values (which range from 0 to ˜10) to that of h^a (range -1 to ∞). All H_{ij} values < 0 are set to 0. This model has an advantage over the discrete HLP17 model in that it describes a much broader spectrum of SHM mutability while only adding two additional parameters to the NY98 model [Nielsen and Yang, 1998] upon which both are based. As with the discrete HLP17, this uses a mean-field approximation to adjust for mutability across codon boundaries, and remains nested with the NY98 model. However, it has a disadvantage compared to the discrete case: it is ultimately only as good as the empirical model (S5F) it is based on. In instances where that model is not realistic, such as in non-human antibody lineages, the S5F-incorportated HLP17 model may be outperformed by the discrete HLP17.

The original formulation of the HLP17 model (Eq. 3.1) also discretizes the effect of amino acid exchangeability by using ω - a single rate scalar describing the relative rates of nonsynonymous and synonymous substitutions. This makes the biologically unrealistic assumption that all amino acids have equal exchangeability rates. Empirical models already exist to describe the relative exchangeability of amino acids, both for proteins in general (e.g. the WAG model; [Whelan and Goldman, 2001]) and antibody lineages specifically (the AB model; [Mirsky et al., 2015]). I

incorporate this information by using, in place of ω:

$$O_{ij} = y_o e^{m_o log(E_{ij})} \tag{4.5}$$

Where E_{ij} is the value in the exchangeability matrix corresponding to the amino acid change from codon i to j, obtained from either the WAG or AB models. Like the semi-empirical hotspot model (Eq. 4.4), y_o and m_o are scaling coefficients estimated by ML. This form has several useful and biologically intuitive properties. While the relationship between ω and Eq. 4.5 is complex, y_o is interpretable as a 'baseline' ratio of nonsynonymous/synonymous change, while m_o is the effect of heterogeneity in amino acid exchangeability. When $m_o = 0$, Eq. 4.5 becomes equivalent to Eq. 4.2 with $y_o = \omega$. When $m_o > 0$, larger values of E_{ij} lead to higher instantaneous substitution rates, while an E_{ij} value of zero leads to a substitution rate of zero, regardless of y_o. In the unrealistic situation in which all E_{ij}'s are equal, Eq. 4.5 is equivalent to Eq. 4.2, with the ML estimates of $\hat{\omega} = \hat{y_o} e^{\hat{m_o} log(c)}$, and y_o and m_o being non-identifiable. Further, exchangeability matrices that differ in scale by a constant value will return the same maximum likelihood value and differ only in their estimate of y_o. That is, if all E_{ij}'s are multiplied by a constant value c, Eq. 4.5 becomes $y_o e^{log(cE_{ij})} = y_o e^{log(c)} e^{log(E_{ij})}$, and thus the new ML estimate of y_o ($\hat{y_o}'$) will be related to the previous estimate by $\hat{y_o} = \hat{y_o}' e^{\hat{m_o} log(c)}$. It should be noted that this form is similar to the original formulation of the GY94 model [Goldman and Yang, 1994], which incorporated amino acid exchangeability in place of ω.

The discrete amino acid exchangeability model (Eq. 4.2) has the advantage that ω is readily interpretable as the ratio of nonsynonymous/synonymous mutations, an important parameter that can reflect the type and degree of natural selection. While y_o and m_o depend heavily on the amino acid model chosen and are consequently less directly interpretable, I can still recover the ratio T_N/T_S, which I refer to as the 'effective ω', or ω_e. Specifically:

$$T_N = \sum_{i=1}^{61}\sum_{j=1}^{61} q_{ij} \text{ i}\Rightarrow\text{j is nonsynon.} \tag{4.6}$$

$$T_S = \sum_{i=1}^{61}\sum_{j=1}^{61} q_{ij} \text{ i}\Rightarrow\text{j is synon. and i} \neq \text{j} \tag{4.7}$$

The ratio T_N/T_S describes the ratio of the sums of instantaneous nonsynonymous to synonymous rates in the **Q** matrix. I then calculate T_N and T_S under a **Q** matrix in which all other parameters are the same, except $\omega = 1$ (or $y_o = 1$ and $m_o = 0$). Comparing these two ratios:

$$\omega_e = \frac{\frac{T_N}{T_S}}{\frac{T_{N,\omega=1}}{T_{S,\omega=1}}} \tag{4.8}$$

This gives the sum of instantaneous nonsynonymous/synonymous rates under a specified **Q** matrix compared to the same model if $\omega = 1$. Under the discrete model (Eq. 4.2) ω_e is identical to ω, and under a semi-empirical model (Eq. 4.5) ω_e provides a similar interpretation, as the relative rate of nonsynonymous to synonymous substitutions.

Because H_{ij} and O_{ij} do not share parameters, any combination of these discrete and semi-empirical models may be used interchangeably, allowing the user to decide which is the most appropriate given the data and hypotheses being tested. Hereafter the models will be referred to by their amino acid exchangeability model (WAG, AB, or NY98) and motif model (e.g. H0, WR<u>C</u>/<u>G</u>RS, S5F). For instance, NY98:H0 indicates the NY98 (discrete) amino acid model using Eq. 4.2, paired with the H0 (all h=0) model, and is equivalent to the NY98 model with a known germline root sequence. The WAG:FCH model uses Eq. 4.5 to incorporate the WAG amino acid exchangeability matrix, and uses Eq. 4.3 to estimate a separate h^a for all six canonical SHM motifs (FCH = Free Coldspots and Hotspots; Chapter

3). The AB:S5F model uses Eq. 4.5 to incorporate the AB amino acid exchangeability matrix, and Eq. 4.4 to incorporate the S5F mutability model. Referring simply to the NY98 model without specifying a motif model indicates the reversible, unrooted codon substitution model of [Nielsen and Yang, 1998]. See Table 4.1 for details on each model tested.

Model Type	Model Name	Equation	Additional Parameters	Description	Reference
Amino acid models	NY98	4.2	ω	Single parameter for dN/dS	(Nielsen and Yang 1998)
	WAG	4.5	m_o, y_o	WAG amino acid exchangeabilities	(Whelan and Goldman 2001)
	AB	4.5	m_o, y_o	AB amino acid exchangeabilities	(Mirsky *et al.* 2015)
Motif models	H0	4.3	None	All $h = 0$	(Yang *et al.* 2000)
	WRC/GYW	4.3	h^{WRC}, h^{GYW}	Strand-asymmetric WRC/GYW	Chapter 3
	WA/TW	4.3	h^{WA}, h^{TW}	Strand-asymmetric WA/TW	Chapter 3
	SYC/GRS	4.3	h^{SYC}, h^{GRS}	Strand-asymmetric SYC/GRS	Chapter 3
	FCH	4.3	h^{WRC}, h^{GYW}, h^{WA}, h^{TW}, h^{SYC}, h^{GRS}	Separate h value for all six canonical motifs	Chapter 3
	S5F	4.4	m_h, y_h	S5F mutability model	(Yaari *et al.* 2013)

Table 4.1: Details of amino acid and motif models. A complete model combination consists of one amino acid and one motif model.

4.2.2 Multiple sequence alignment

I applied these models to the three bnAb lineages analysed in Wu et al. [Wu et al., 2015]. Namely, the CH103 lineage, sampled over three years [Liao et al., 2013], the VRC26 lineage, sampled over four years [Doria-Rose et al., 2014], and the VRC01 lineage, sampled over 15 years [Wu et al., 2015]. Heavy chain sequences from each bnAb lineage were downloaded from GenBank. Putative germline segment

assignments (V4-59*01 and J4 for CH103, V3-30*18 and J3 for VRC26, and V1-2*01 and J1 for VRC01) were obtained from bnAber [Eroshkin et al., 2013] and sequences were obtained from the IMGT V-Quest human reference set [Lefranc and Lefranc, 2001]. In contrast to Chapter 3, I included the entire sequence - not simply the V gene segment - in the multiple sequence alignment (MSA). This is challenging for two reasons. First, the germline CDR3 segment cannot be reliably inferred due to the short length of D gene segments and the random nature of junction nucleotides. Second, because BCR sequences are short and highly polymorphic, it is likely that standard multiple sequence alignment methods, based on matching nucleotides within columns, will introduce large numbers of artificial gaps and place nucleotides in alignment that are not actually homologous.

To address these issues, I instead used a hybrid approach that aligned each sequence to its predicted germline V and J segment, and aligned the CDR3 regions of each MSA using MUSCLE [Edgar, 2004]. First, I used *partis* [Ralph and Matsen, 2016] to align each sequence to its predicted V and J gene segment germline allele (see paragraph above). This placed each sequence in alignment with its predicted germline sequence. I then added gaps to each sequence and its aligned germline sequence concurrently, so that the V and J segments of all sequences' predicted germlines were aligned with each other. Because these gap additions were replicated in the B-cell Ig sequences, this effectively aligned all the V and J segments of all sequences in the MSA to a common germline, placing them in alignment with each other. Because the germline CDR3 regions cannot be inferred as reliably as the V and J segments, the CDR3 regions were aligned separately as amino acids using MUSCLE [Edgar, 2004], and re-inserted into the alignment. This process is shown graphically in Figure 4.2.

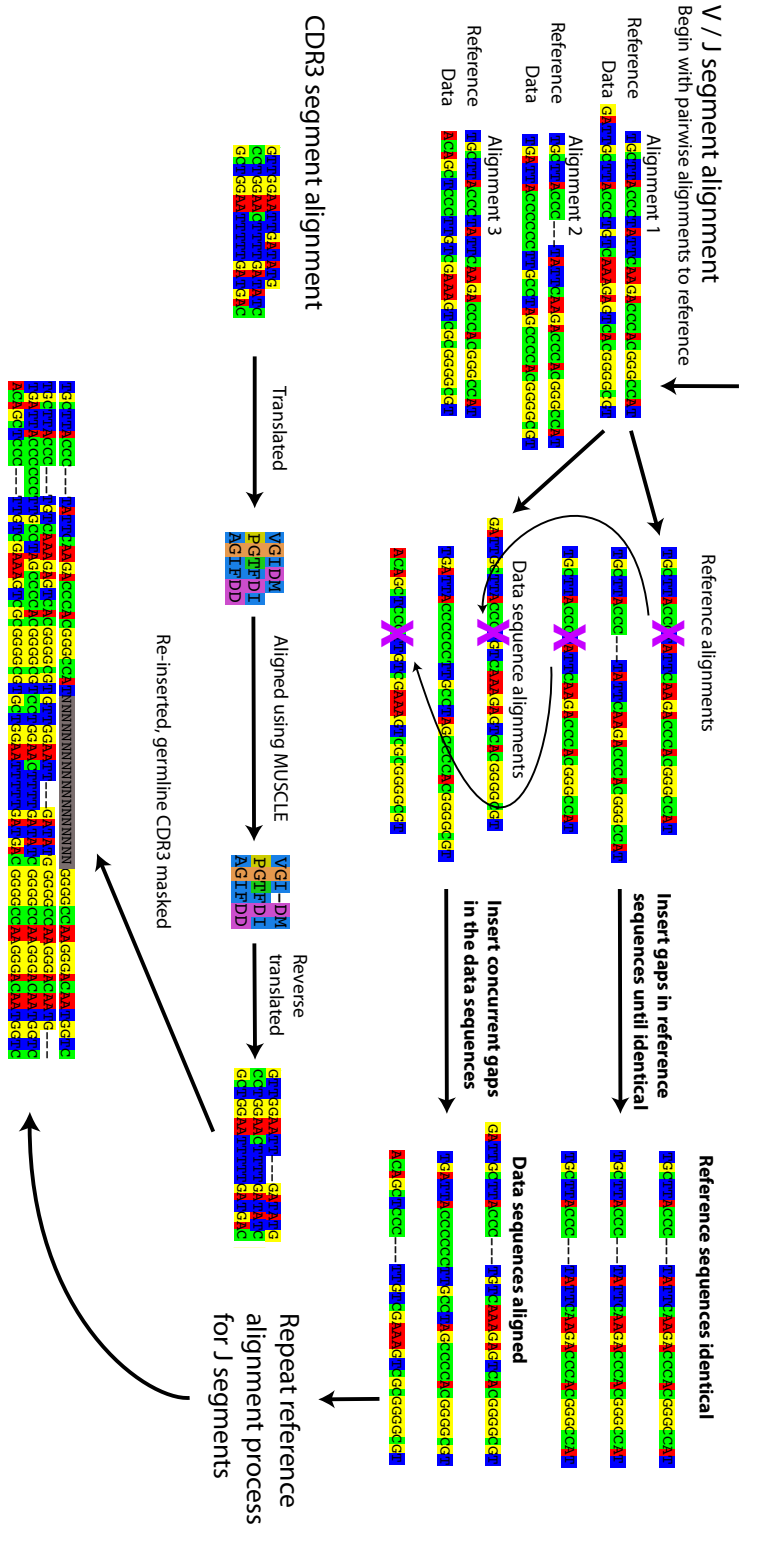


Figure 4.2: Multiple sequence alignment process for bnAb lineages. V and J segments are aligned by combining their pairwise alignments to their reference sequence (here, gaps inserted at position indicated by purple **X**). CDR3s are aligned using MUSCLE.

Gaps inserted by partis that disrupted codons were augmented with additional nucleotide insertions such that all codons were left intact. One seven-codon region in VRC01 was highly polymorphic for germline-relative insertions, and was ultimately removed from the final MSA due to difficulties in alignment. This region contained a seven-codon insertion which was identified in [Wu et al., 2015] as providing increased binding surface on gp120. However, it seems unlikely that removing this region will have a large impact on the overall phylogenetic results presented in this chapter. After processing, each sequence was translated into an amino acid sequence, and all subsequent alignment was performed at the amino acid level. Final amino acid alignments were un-translated into nucleotide sequences, giving a codon alignment. Germline V and J segments for each lineage were manually aligned to these MSAs, leaving the CDR3 region as a large gap. Some sequences were discarded by partis, leaving 441, 675, and 1005 sequences in the CH103, VRC26, and VRC01 alignments respectively, including germline sequences. This process is shown graphically in Figure 4.2.

Due to the computational complexity of phylogenetic topology estimation under complex codon models with a large number of taxa, each lineage was subsampled to 200 sequences, including its predicted germline sequence. I then added previously identified monoclonal antibody sequences to each alignment, which serve as experimentally verified markers for phenotypic properties such as neutralization breadth. Ultimately, 11 monoclonal antibodies were added to CH103 [Liao et al., 2013], 12 to VRC26 [Doria-Rose et al., 2014] and 20 to VRC01 [Wu et al., 2015]. See Section 5.2.2 for a detailed explanation of how these antibodies were obtained and how neutralization assays were performed.

4.2.3 Verifying effectiveness of WAG model

To verify the effectiveness of semi-empirical amino acid exchangeability models, I used the CH103 lineage to estimate the maximum likelihood parameters of 100 WAG:S5F models in which the exchangeability values within the WAG matrix were permuted. I performed these permutation analyses to verify that the WAG exchangeability matrix contains useful biological information about the exchangeability of amino acids in bnAb lineages. If permuted WAG matrices show log likelihood values similar to the true WAG exchangeability matrix, this would indicate the true WAG exchangeability matrix does not accurately describe amino acid substitutions in the CH103 lineage. During these estimations, the tree topology was fixed to the maximum likelihood topology estimated under the NY98 model. I also estimated parameters of the WAG:S5F model using ML under an amino acid exchangeability matrix in which all values were equal to one. I compared the log-likelihood values, and the ML estimates of y_o and m_o from each of these models, to those obtained under the true WAG matrix.

4.2.4 Model fitting and tree estimation

To undertake model selection for each lineage, I first estimated the ML phylogenetic topology under the standard NY98 model for each lineage using SPR moves. I then used this topology as a starting tree for estimating ML phylogenetic topology, branch lengths, and substitution parameters under all 18 combinations of motif and one amino acid models given in Table 4.1. For each comparison, I estimated the ML phylogenetic topology using NNI moves. Because not all model combinations are nested with each other (though all are nested with the NY98:H0 combination) I compared model fit using AIC [Akaike, 1974].

4.2.5 Efficient topology searching under HLP17

As explained in section 1.9, most phylogenetic analysis programs use the pulley principle [Felsenstein, 1981] to dramatically speed up branch length and topology estimation. Boussau and Gouy [Boussau and Gouy, 2006] introduced a framework for using a similar algorithm to perform efficient likelihood calculations in non-reversible models applied to rooted phylogenetic trees, which I implemented for branch length estimation in Chapter 3. The same approach of storing partial likelihoods can also be used to dramatically speed up topology searching via nearest neighbour interchanges (NNI's; [Boussau and Gouy, 2006]), and subtree pruning and regraft moves (SPR). While NNI and SPR search routines were previously implemented in codonPhyML [Gil et al., 2013], this implementation was not compatible with non-reversible models. Hence, I implemented these algorithms in a revised version of IgPhyML, which is now capable of efficient topology searching with the non-reversible HLP17 model.

I selected the best-fit model combination from each lineage (outlined in section 4.2.4), and re-estimated ML parameters and phylogenetic topology using SPR moves, which more thoroughly search the space of possible topologies than NNI moves. As a control, I performed the same parameter and topology estimation using the NY98:H0 model. Both of these models used the ML estimate of phylogenetic topology under the NY98 model as the starting tree. Branch supports were calculated for each branch of each tree using aLRT [Anisimova et al., 2006]. I report ML estimates of key substitution parameters, as well as statistics such as tree length and the ratio of internal/external branch lengths.

4.2.6 Reconstructing germline CDR3 sequences

Given the ML topologies estimated in section 4.2.5, and using the same ancestral sequence reconstruction (ASR) algorithms used in Chapter 3 [Boussau et al., 2008],

I reconstructed the most likely marginal codon and amino acid for each site within the germline CDR3 of each lineage. Because the HLP17 model allows for non-reversible hotspot mutations, the predicted germline sequence from these models are expected to have a greater 'hotness' when allowing for hotspot mutability, compared to the standard NY98:H0 model. I used mean S5F mutability [Yaari et al., 2013] as a proxy for this. It is unfortunately not possible to directly test the accuracy of each approach in reconstructing the germline sequence, because the true germline CDR3 sequence is not known for these lineages. To test which sub-model more accurately reconstructed the germline sequence, I repeated the ASR analysis but with the CDR1 and CDR2 regions of the germline masked by replacing their sequence information with 'N's, in addition to the unknown CDR3 region. More specifically, I repeated topology estimation as described in section 4.2.5 and ASR as described previously in this section, but with the germline CDR1 and CDR2 of each lineage masked (CDR3s were already masked in section 4.2.2). I then compared the reconstructed CDR1 and CDR2 to the sequences in the actual V reference gene.

4.3 Results

4.3.1 Characterizing parameter space of the semi-empirical WAG model

The values within the WAG exchangeability matrix are estimates of the rates at which each amino acid is substituted. These can be substantially different between pairs of amino acids (Figure 4.1b) and are thought to reflect functional similarity between amino acids. Because of this, I expect that a WAG model using the true WAG matrix will show a significantly better maximum likelihood than a model using a WAG matrix with permuted exchangeability values when applied to antibody lineage sequences (or any protein coding gene). Further, the semi-empirical

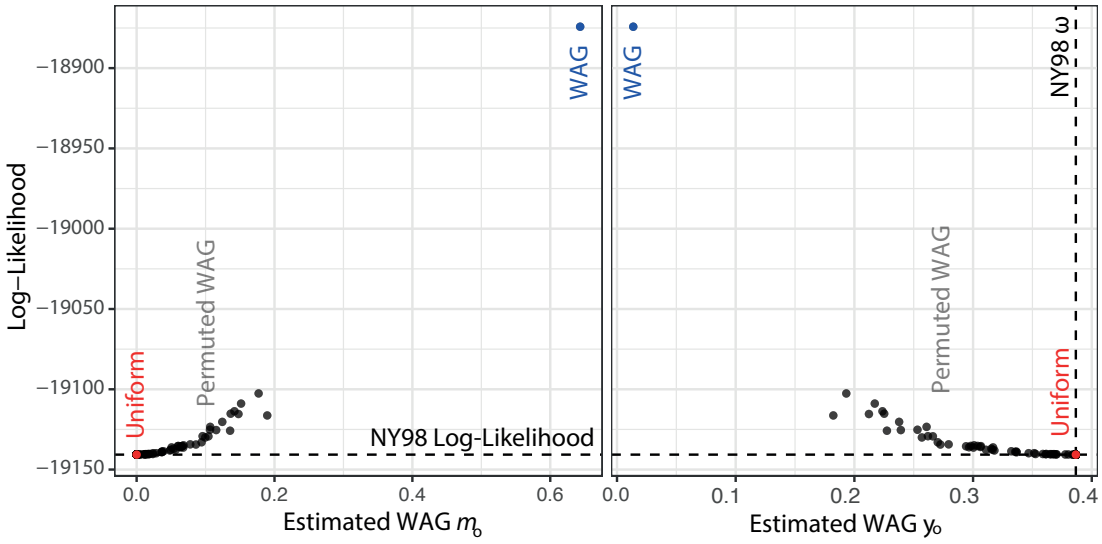


Figure 4.3: Validation of the WAG exchangeability matrix. The y axis shows the maximum likelihood (ℓ) obtained from a particular WAG:S5F model, while the x axis shows the ML estimation of m_o or y_o. For comparison, the ℓ of the discrete NY98:S5F model is shown in the horizontal dotted line, and the ML estimate of ω in the NY98:S5F model is shown by the vertical dotted line.

amino acid models tested in this chapter use Eq. 4.5, making them nested with the NY98 model. This is because the NY98:H0 model is a special case of Eq. 4.5 in which $m_o = 0$. Thus, all semi-empirical models using Eq. 4.5 are expected to show a maximum likelihood value that is at least as high as NY98:H0, regardless of their exchangeability matrix. To verify that the WAG semi-empirical amino acid exchangeability model does behave in this way, I permuted the values in the WAG exchangeability matrix and used them to perform parameter estimation on the CH103 lineage under the WAG:S5F model (Table 4.1).

As expected, the maximum likelihood obtained from models using permuted WAG exchangeability values was at least as high as the maximum likelihood obtained under the NY98 model, with the lowest likelihood permutations having the lowest m_o values (lower bound of zero) and y_o (bounded by the ML estimate of ω under NY98) values. Further, using WAG:S5F models with all exchangeability values equal to one yielded a maximum likelihood identical to the NY98:S5F model,

with $m_o = 0$ and $y_o = 0.4$. All of the models using permuted WAG matrices had much lower likelihood values compared to the model using the actual WAG matrix. These results are all shown in Figure 4.3.

ML estimates of m_o and y_o showed a strong negative relationship to each other (Figure 4.4), with the model using the true WAG matrix showing the highest m_o and lowest y_o. Models using a permuted WAG exchangeability values show lower estimated m_o and higher y_o. This pattern is easy to interpret if we consider the semi-empirical amino acid models (Eq. 4.5) as a combination of two models, with the scaling parameters changing the relative contributions of each. The parameter y_o represents a model in which codon changes are considered only as nonsynonymous or synonymous, while m_o represents a model in which rates of codon changes are derived from the empirical amino acid exchangeability matrix. In cases where the amino acid exchangeability matrix is a poor estimate for the rates of amino acid exchangeability in the data being analysed (e.g. most permuted WAG matrices), codon changes are primarily modelled as nonsynonymous/synonymous changes (low m_o, high y_o). In cases where the values in the exchangeability matrix are accurate estimates of amino acid exchangeability in the data (e.g. the true WAG matrix), codon changes are more represented by the exchangeability matrix (high m_o, low y_o). As discussed in 4.2.1, y_o also reflects the scaling of the exchangeability matrix, so the ML estimates of m_o and y_o are not directly comparable between all amino acid models (for instance, multiplying all values in the WAG matrix by a constant value would lead to the same maximum likelihood value but a lower ML estimate of y_o).

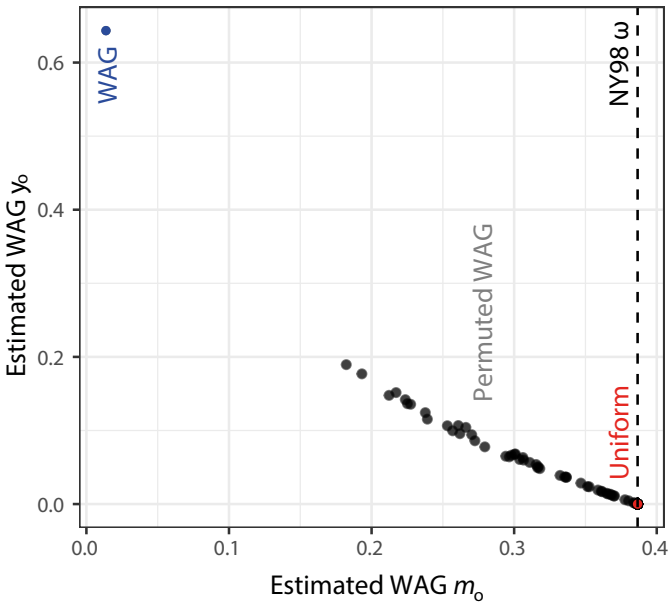


Figure 4.4: ML estimates of m_o and y_o using WAG:S5F models with permuted WAG exchangeability matrices. For comparison, the ML estimate of ω for NY98:S5F is shown by the vertical dotted line.

4.3.2 The best-fit model combination varies by lineage

The most straightforward means of comparing how well a particular substitution model fits a set of data is to use the Akaike information criterion (AIC; [Akaike, 1974]), which compares the maximum likelihood attained from the model fit to the number of free parameters in the model. More specifically, the AIC is defined as:

$$AIC = 2 * f - 2 * log(\ell(H_i)) \tag{4.9}$$

Where $\ell(H_i)$ is the maximum likelihood value obtained from the model i, and f is the number of free parameters. A lower AIC value indicates a better fit. The score is dependent on the maximum likelihood (ℓ) of the model on the data, but is penalized by adding extra parameters. Compared to the NY98:H0 model (Table 4.1), nested in all HLP17 models, WAG and AB amino acid models both add a single additional free parameter (Eq. 4.5). Asymmetric discrete models describing a single motif and its reverse complement (WRC/GYW, WA/TW, and SYC/GRS)

add two free parameters each (Eq. 3.1), as does the S5F semi-empirical model (Eq. 4.4) . The free coldspots and hotspots (FCH) discrete model, which fits a different h for all three canonical motifs and their reverse complements, adds six additional free parameters. Here, I fit different combinations of discrete and semi-empirical models to the three bnAb lineages (Figure 4.5).

For all datasets and all submodels tested, the WAG model exhibited a better/lower AIC score than both the NY98 and even the antibody-specific AB model. The S5F semi-empirical model showed a much lower AIC than any other two parameter motif model tested, likely because it was able to capture a more full range of mutability than single asymmetric motifs. Results comparing the S5F model to the more parameter-rich FCH discrete model were mixed. The S5F model outperformed FCH in the CH103 lineage, but not in VRC26 and VRC01 where FCH showed the best fit.

4.3.3 Differences in topology estimated under different model combinations and published topologies

Interestingly, the CH103 lineage showed a similar estimate of topology under the NY98:H0 and WAG:S5F models (Figure 4.6). Tree topologies estimated under the NY98:H0 and WAG:FCH models for the VRC01 lineage were similar as well 4.8). For the CH103 tree, the monoclonal antibody sequences had similar placements in the NY98:H0 and WAG:S5F estimated topologies. However, some taxa from late time points were closely related to the CH103-6 antibody sequences in the NY98:H0 estimated topology, but were more closely related to 1AH92U in the WAG:S5F estimated topology. The placement of monoclonal antibody sequences was also very similar to the published topology in Liao et al. [Liao et al., 2013], with 1A102RI6 and 1AH92U being less related to the CH103-6 sequences but within the same clade. For the VRC01 lineage, the clade groupings are consistent

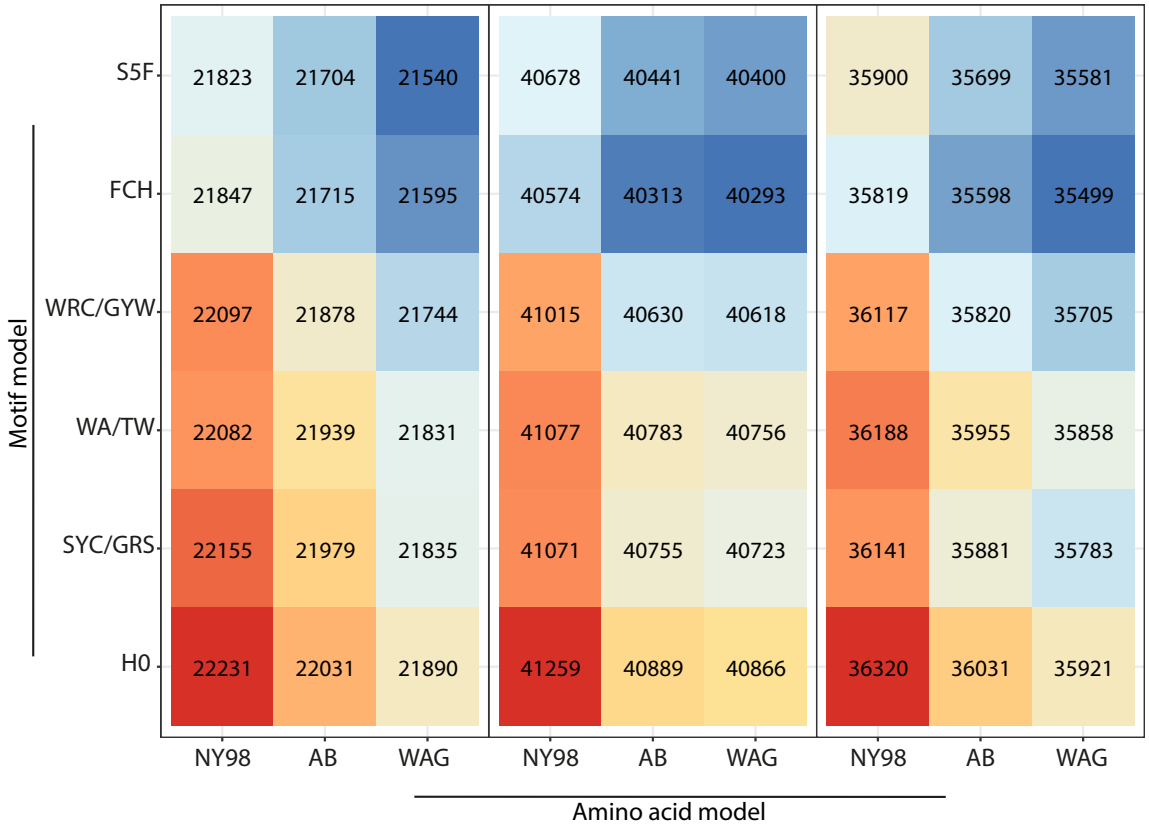


Figure 4.5: AIC of model combinations for each lineage. Each subplot shows all model combinations for each lineage, with amino acid models on the x axis and mutability models on the y axis. Cells display the AIC for each model combination, and are shaded by AIC value as well from red to blue, with blue being the lowest AIC and best model fit. See Table 4.1 for details on each model.

with the published topology [Wu et al., 2015] which, given the large divergence between clade in this tree, is not entirely surprising. However, the topologies estimated under both the NY98:H0 and WAG:FCH models have the monoclonal antibody sequence H5 as being in the most basal clade, followed by the rest of the monoclonal antibody sequences, while the published topology lists H5 as grouping with H4, followed by H03+06 [Wu et al., 2015].

The VRC26 lineage, on the other hand, did show some noticeable differences between topologies estimated under the NY98:H0 and NY98:FCH models (Figure 4.7). While most monoclonal antibody sequences tended to cluster together within the same groups in both trees, their placement within the trees changed. In NY98:H0, MC3, 4, and 7 are most closely related to MC8 and 9, but in NY98:FCH, MC3, 4, and 7 are more closely related to MC12. Similarly, under the NY98:H0 model, MC5 is most closely related to MC10, but under the WAG:FCH model, it is most closely related to MC2 and 6, with MC10 belonging to a separate clade (Figure 4.7). Compared to the published topology [Doria-Rose et al., 2014], the WAG:FCH tree shows the same grouping of MC8 and 9 being most closely related to MC12, and then MC3, 4, and 7. The published topology also showed MC6, 5, and 2 as fairly equally unrelated; however, topologies estimated under both the NY98:H0 and WAG:FCH models placed MC2 and 6 much closer together, with MC5 as a more distant. NY98:H0 actually placed MC10 within the same clade as these three sequences, while in WAG:FCH and the published topology MC10 and MC11 form a separate clade.

As also shown in Chapter 3, topologies estimated under the best-fit HLP17 models showed an increase in total tree length (number of mutations) compared to the NY98:H0 models (Table 4.1). This may be due to increased allowance of homoplasy due to increased mutability of hotspots. It is interesting that all three lineages showed a fairly consistent increase in branch length (~10%) for each lin-

eage. Further, no lineage showed a large difference in internal/external branch length ratio between the models tested (Table 4.1). In the VRC26 and VRC01 lineages, germline CDR3 reconstructions were predictably more mutable under HLP17 models than under the NY98:H0 model, though this was not true in the CH103 lineage (Table 4.1).

4.3.4 Ancestral sequence reconstructions

Reconstructing the rearranged, un-mutated germline CDR3 sequence for an antibody lineage is an important application of phylogenetic BCR sequence analysis, particularly for applications such as HIV vaccine design that focus on stimulating the germline precursors of bnAb lineages [Haynes et al., 2012]. I tested how CDR3 ASRs might differ between each lineage's best-fit HLP17 and NY98:H0 models, by estimating the ML tree topology (using an SPR heuristic search and the ML estimated topology using the NY98 model as a starting tree) for each model and reconstructing its germline CDR3 sequence using the formulas detailed in Chapter 3 and Boussau et al. [Boussau et al., 2008]. All three lineages showed at least one nucleotide difference between models along their CDR3s, with the VRC26 and VRC01 lineages showing 5 and 2 nucleotide differences, respectively (Table 4.1). Interestingly, CH103 actually showed a slightly lower mutability for the WAG:S5F model than NY98:H0. VRC26 and VRC01, however, both showed higher mutability under WAG:FCH, as expected (Table 4.2).

To test which sub-model may more accurately reconstruct the un-mutated germline sequence, I repeated the above analysis but with the CDR1 and CDR2 regions of the germline masked by replacing their sequence information with Ns, in addition to the unknown CDR3 region. This allowed me to compare the phylogenetically

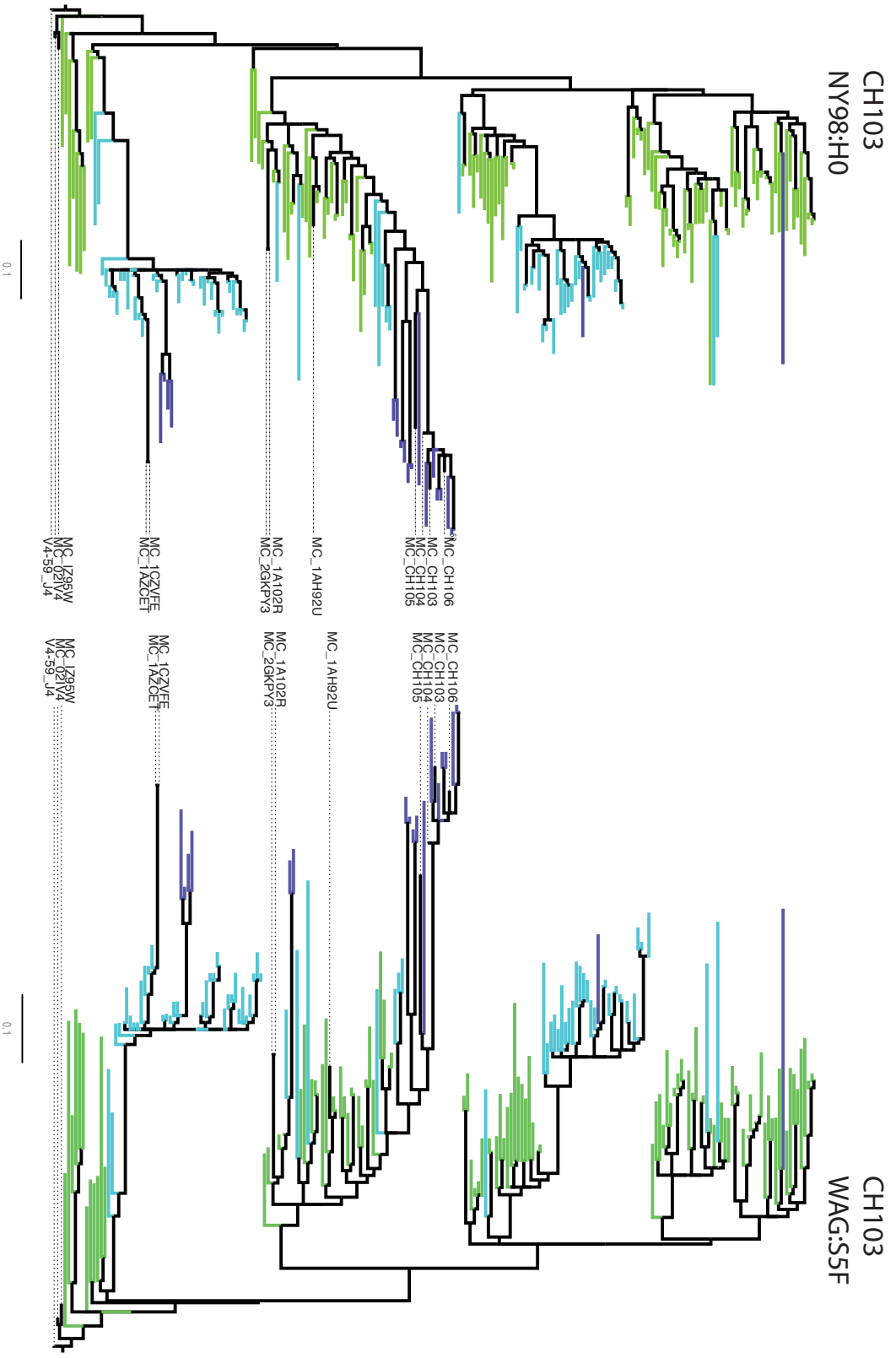


Figure 4.6: ML topologies of the CH103 lineage estimated under models NY98:H0, and WAG:S5F, respectively. External branches are coloured by their sample date (from 53 weeks to 144 weeks post-infection, green to purple, respectively). Monoclonal antibody sequences are labelled, as is the germline sequence (V4-59:J4). Branch lengths are in units of substitutions per site.

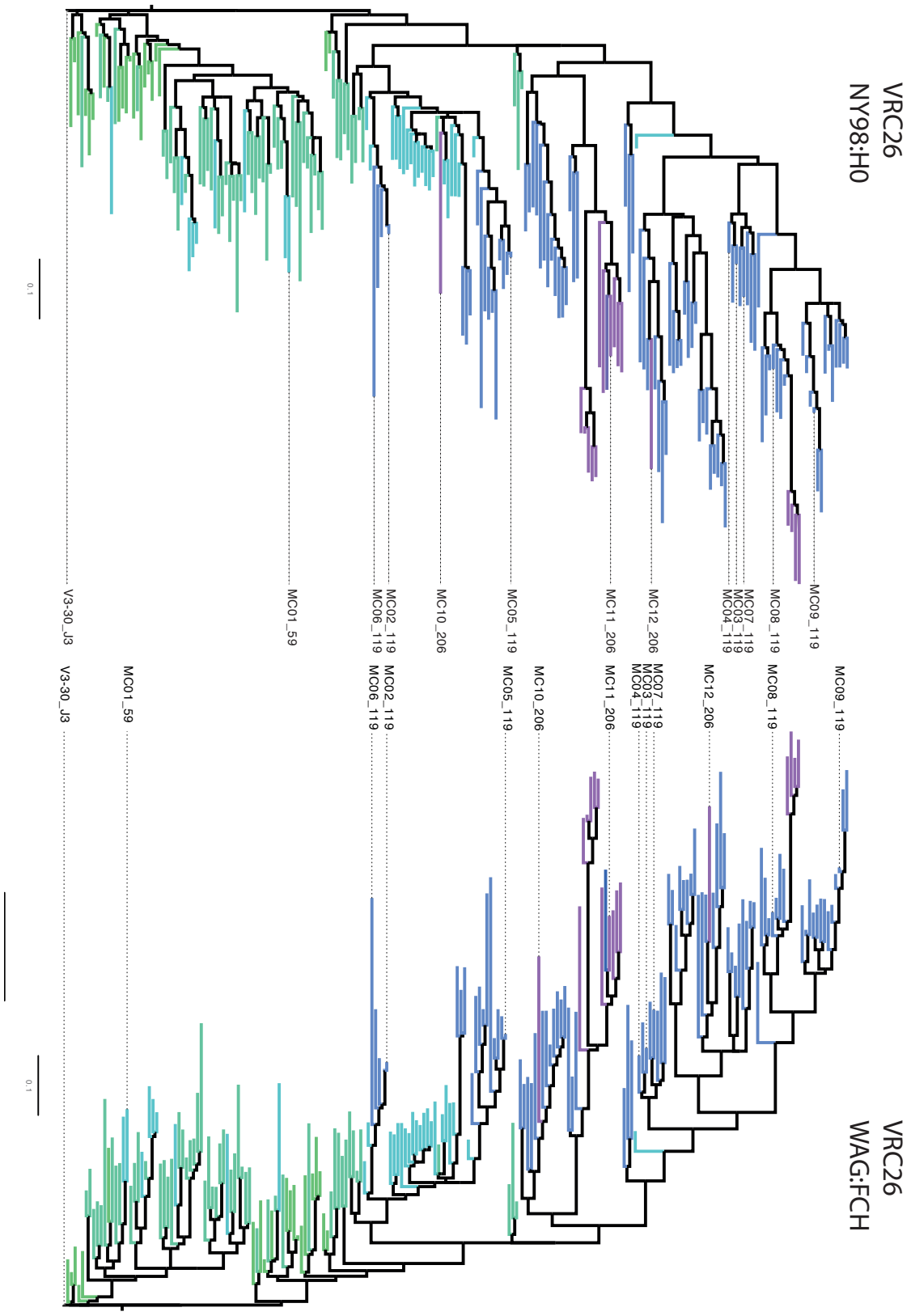


Figure 4.7: ML topologies of the VRC26 lineage estimated under models NY98:H0, and WAG:FCH, respectively. External branches are coloured by their sample date (from 38 weeks to 206 weeks post-infection, green to purple, respectively). Monoclonal antibody sequences are labelled, as is the germline sequence (V3-30:J3). Branch lengths are in units of substitutions per site.

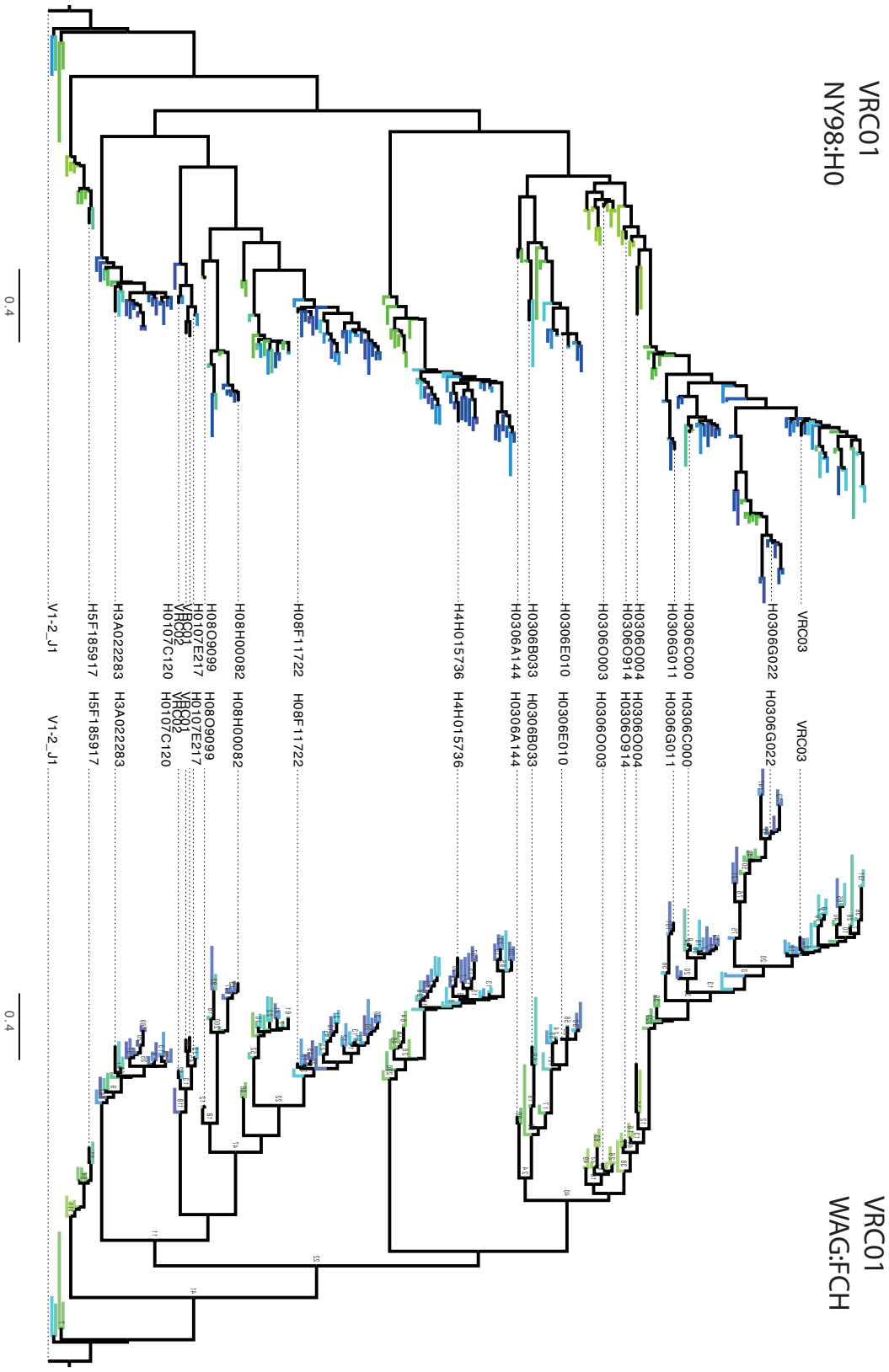


Figure 4.8: ML topologies of the VRC01 lineage estimated under models NY98:H0, and WAG:FCH, respectively. External branches are coloured by their sample date (from 1995 to 2009, green to purple, respectively). Monoclonal antibody sequences are labelled, as is the germline sequence (V1-2:J1). Branch lengths are in units of substitutions per site.

reconstructed germline sequence to the germline sequence predicted through the traditional approach of aligning sequences to known databases of V and J segments through tools such as IMGT-VQuest [Brochet et al., 2008] or IgBlast [Ye et al., 2013]. Interestingly, the best-fit HLP17 and NY98:H0 models reconstructed the same CDR1 and CDR2 germline sequences for each of the three lineages.

	CH103		**VRC26**		**VRC01**	
Model	NY98:H0	WAG:S5F	NY98:H0	WAG:FCH	NY98:H0	WAG:FCH
Tree length	20.32	22.25	28.09	30.25	32.85	36.49
Internal/External branch ratio	0.35	0.35	0.29	0.28	0.51	0.52
Mean S5F mutability score of germline CDR3 ASRs	1.014	1.002	0.852	0.932	0.858	0.954
Germline CDR3 ASR differences	1 nuc, 0 aa		7 nuc, 5 aa		2 nuc, 0 aa	
RF dist. from NY98	42	142	14	112	10	18
ω_e	0.46	0.41	0.57	0.55	0.34	0.34
Difference from true germline, masked CDR reconstructions	0 nuc	0 nuc	0 nuc	0 nuc	2 nuc	2 nuc

Table 4.2: Summary statistics of ML trees and germline sequence reconstructions for NY98:H0 and the best-fit HLP17 model for each lineage.

4.4 Discussion

Perhaps the most surprising result of these analyses is that the AB amino acid exchangeability model, which was designed specifically to better describe antibody evolution [Mirsky et al., 2015], was handily outperformed by the older and much more general WAG model [Whelan and Goldman, 2001]. The parameters of both of these models were developed using similar methods - using a large number of sequence alignments and maximum likelihood optimization in order to estimate a symmetric amino acid exchangeability matrix. Further, because both the AB and WAG semi-empirical models use empirical exchangeability matrices with the same number of fixed parameters, this result is robust to any potential problems

with applying AIC to a comparison of empirical or parametric substitution models.

Model comparisons between semi-empirical and discrete amino acid models using AIC could, however, be questioned. AIC is parameterized by the number of free parameters estimated from the data. The NY98:H0 model, for instance, has 14 free substitution parameters, while WAG:H0 has 15. However, the WAG exchangeability matrix used by the WAG:H0 model also contains 180 rate parameters, which are not included in the AIC calculation because they are fixed during parameter estimation. However, the fact that the AB exchangeability matrix was also tested here suggests that, in the context of my study, these multiple rate parameters were not fixed during analyses. One can imagine, for instance, producing thousands of different amino acid exchangeability matrices, then selecting the matrix with the highest ℓ value when fit to the data, and comparing it to the ℓ of the NY98:H0 model. In this extreme example, one should obviously count these amino acid rate parameters as estimated parameters in the AIC calculation. It is not immediately clear where the model-fitting analyses performed in this chapter lie in the spectrum of treating amino acid rate parameters as fixed parameters or as free parameters for the purposes of AIC calculation. However, considering that only two amino acid exchangeability matrices were tested in this chapter, and that these matrices were previously characterized and biologically informed, treating exchangeability rate parameters as fixed seems justifiable.

Importantly, the WAG model used alignments from a large database of different protein families, in an effort to make an exchangeability matrix that was as general as possible [Whelan and Goldman, 2001]. The AB model, on the other hand, was estimated from a large database of antibody sequence alignments [Mirsky et al., 2015]. One possible explanation for WAG's superiority is that the substitution processes operating in these bnAb lineages is significantly different from

the processes in the alignments used to fit the AB model; thus, a better fit is obtained from a model capturing the general patterns of amino acid change. This also suggests that the amino acid substitution process is highly variable among bnAb lineages. It is important to note that the semi-empirical models introduced in this chapter used only the exchangeability matrices, not amino acid frequencies, of these amino acid substitution models. For this reason, these analyses are not perfect comparisons of the full WAG and AB models.

Another surprising result is that, while the S5F mutability model [Yaari et al., 2013] outperformed all others for the CH103 lineage, it was outperformed by FCH for the VRC26 and VRC01 lineages. One explanation is simply that empirical models such as the S5F rely on the appropriateness of the data used to fit them, and it is possible that the datasets used to produce the S5F model were generated by an evolutionary process more similar to the process that generated the CH103 lineage than the VRC26 and VRC01 lineages. It is also possible that VRC26 and VRC01 experienced some other un-modelled substitution process and that heterogeneity can be accommodated by the many additional h parameters of FCH. It is notable that there did not appear to be a substantial interaction between mutability models and amino acid models - the best fit mutability model outperformed the others regardless of the amino acid model it was paired with, and vice versa for amino acid models.

I had expected a larger difference in ML topology estimated under the NY98:H0 and best-fit HLP17 model for each lineage. The CH103 and VRC01 lineages, at least with regards to their monoclonal antibody sequences, showed very similar topologies. This may be a limitation brought about by the fact that tree estimation under both models used the ML tree estimated under the NY98 model as the starting topology, which may have kept the heuristic tree estimation process in a local optimum near the starting tree. Alternative starting trees may give differ-

ent answers, though the issue of starting tree effects is a feature of ML topology analysis generally, and one that should have been mitigated by using SPR moves instead of NNIs. It is also possible that only a few sites in the CH103 and VRC01 lineages are significantly better represented by models that account for hotspot mutability and amino acid exchangeability, while the majority of sites follow the 'standard' substitution process described in the NY98 model. This would explain why the more complex models introduced in this chapter do not seem to have a large effect on ML estimated topology for these lineages. The VRC26 lineage, however, did show some differences in topology between the models and between the published topology [Doria-Rose et al., 2014]. However, because I currently do not have a model to predict antigen binding of ancestral sequences, it is unclear how these changes in topology change the biological interpretation of the tree, at least regarding the development of the VRC26 lineage outlined in Doria-Rose et al. [Doria-Rose et al., 2014].

Other tree-based statistics behaved approximately as expected. Germline CDR3 reconstructions were predictably more mutable ('hotter') in the best fit HLP17 models than the NY98:H0 models, at least in the VRC26 and VRC01 lineages. This is likely due to the HLP17 model better accounting for a decrease in hotspot motifs during affinity maturation, which, in the reverse process of ancestral state reconstruction, leads to an increase in hotspot motifs in the reconstructed germline. It is not clear why the CH103 lineage does not show this pattern as well. Further, and consistent with the results from Chapter 3, best-fit HLP17 models lead to an ~10% increase in tree size (Table 4.1) indicating more substitutions were predicted in these trees. This may be due to an increased frequency of homoplasy (the same state change happening independently on separate branches of the tree), resulting from accounting for hotspot mutations. This is similar to the issue of long branch attraction in which homoplasies are erroneously placed on internal branches of the tree. By allowing hotspot motifs to be more mutable, the HLP17 models may

be more permissive of these mutations occurring on separate branches, leading to an overall increase in tree length. Interestingly, however, trees estimated in these analyses showed similar ratios of internal/external branch lengths regardless of the model used (Table 4.2). Because failing to account for homoplasy is expected to place more mutations near the root of the tree instead of near the tips (i.e. make homoplasies shared derived characters instead of identical, independent mutations), we expected the internal/external branch ratio to increase as these homoplasies were corrected. This is related to the parsimony score of a tree [Fitch, 1971], and the total length of the ML tree, because both would necessarily be increased by 'moving' mutations from internal to external branches. However, these results suggest that either the ratio of internal/external branch lengths is a poor measure of homoplasy, or that the differences in tree length are not due to homoplasy, but to more substitutions across the tree generally.

It is also interesting that the NY98:H0 and best-fit HLP17 models showed the same reconstructions for CDR1 and CDR2 when these regions were masked in each lineage, and leaves open the question of which approach is actually more accurate. This result is likely tied to the issue of topology estimation, especially considering that the CH103 and VRC01 lineages both show similar ASRs and similar ML estimates of phylogenetic topology under the NY98:H0 model and their best fit HLP17 model. VRC26, on the other hand, shows more substantial difference in both ASR and ML topology estimated under the NY98:H0 and WAG:FCH models (Figures 4.6, 4.7, and 4.8, Table 4.2).

By incorporating available empirical information of both the nucleotide and amino acid substitution biases, I have created a more complex model of substitution processes during affinity maturation than has been previously available. These models show a significantly better fit to these data than previous models, and allow the estimation of interesting biological parameters, such as the rate of nonsyn-

onymous to synonymous substitutions (ω/y_o) and altered mutability of hot- and coldspot motifs. These models are obviously useful in instances where knowing the value of those parameters is important. Unfortunately, it is not obvious that incorporating these parameters significantly improves aspects of phylogenetic tree estimation and ancestral sequence reconstruction on the three lineages tested in this chapter, at least under my current implementation within the ML framework of IgPhyML. Future work may incorporate this model into programs that allow topology searching and ancestral state reconstruction in a Bayesian paradigm, e.g. BEAST [Drummond and Rambaut, 2007]. This would much more readily account for the inherent uncertainty in the topology.

Chapter 5

Variation in clonal selection during maturation of HIV broadly neutralizing antibody lineages

5.1 Introduction

Affinity maturation is the process whereby a population of naive B cells expressing weakly binding B cell receptors (BCRs) are replaced by descent with modification by a population of B cells expressing receptors with much stronger binding affinities. Somatic hypermutation (SHM) provides the fuel for this process by generating a large amount of mutational diversity within the genes that encode the BCR. However, the changes that occur during affinity maturation tend towards increasing binding affinity of the BCR. This directionality comes from clonal selection, the process whereby germinal centre B cells survive and reproduce based on how effectively they bind to antigen on the surface of follicular dendritic cells. This closely parallels the process of Darwinian selection observed in organismal populations. Because clonal selection is such an important factor in affinity maturation, quantifying and characterizing it is important to understanding the process of

adaptive immunity. Previous studies have provided evidence that selection pressure varies among regions of the BCR sequence [Yaari et al., 2012], and over time during affinity maturation [Sheng et al., 2016]. However, these studies have been limited by the methods typically used in antibody genetic studies to characterize selective pressure. In this chapter, I use the HLP17 phylogenetic codon model to explore variation in natural selection operating within and between three broadly neutralizing antibody lineages (bnAbs).

While methods for inferring natural selection by estimating the rates ratio of nonsynonymous to synonymous (dN/dS) substitution have existed in molecular evolution for over a decade [Nielsen and Yang, 1998, Yang and Swanson, 2002], the field of antibody genetics has largely eschewed these phylogenetic models in favour of counting methods, collectively referred to here as 'R/S methods' [Hershberg et al., 2008, Uduman et al., 2011, Yaari et al., 2012]. These methods operate by calculating the number of amino acid replacement (R; nonsynonymous) and silent (S; synonymous) substitutions between a single data sequence and its inferred germline sequence and then comparing R to the expected number of replacement substitutions (E(R)) under an empirically derived null model of neutral SHM and the predicted germline sequence [Hershberg et al., 2008]. For example, as detailed in Uduman et al. [Uduman et al., 2011], the expected number of replacement substitutions (E(R)) is calculated by:

$$E(R) = \sum_{i=1}^{n} \sum_{d \epsilon (A,C,G,T)} f(g_i) M(g_i, d) I \begin{cases} 1 & g_i \Rightarrow d = \text{AA replacement} \\ 0 & g_i \Rightarrow d \neq \text{AA replacement} \end{cases} \tag{5.1}$$

Where n is the number of sites in the alignment, g_i is the nucleotide at position i in the germline, $f(g_i)$ is the scalar change in mutation rate from the sequence context of g_i obtained from an empirical model (e.g. the S5F model [Yaari et al., 2013]), $M(g_i, d)$ is the relative probability of substitution of g_i by d compared to the two

other possible nucleotides. These values are multiplied by an indicator function equal to one if the substitution from g_i to d is a replacement mutation and zero otherwise. In this framework $R < E(R)$ is indicative of purifying selection, while $R > E(R)$ is indicative of positive selection. Significance is determined by comparing R to a binomial cumulative density function parameterized using $n = (R + S)$ and $p = E(R)/(E(R) + E(S))$ [Hershberg et al., 2008, Uduman et al., 2011]. There are many variations of this test, including those that use R and S from the framework (FWR) and complementary determining (CDR) regions in different ways [Hershberg et al., 2008, Uduman et al., 2011], a Bayesian framework for quantifying selection (BASELINe; [Yaari et al., 2012], and tests that consider substitutions occurring on different branches of parsimony trees [Uduman et al., 2014, Yaari et al., 2015].

These R/S methods have been used to infer variation in selective pressure operating on different sites in antibody sequences and at different time points during affinity maturation. Considering different regions of the Ig heavy chain sequence, framework regions (FWRs) have been consistently shown to be under purifying selection in a variety of contexts while CDRs typically show more neutral evolution [Yaari et al., 2012]. Sheng et al [Sheng et al., 2016] recently used an R/S method to investigate variation in selection pressure over long periods of affinity maturation in three HIV bnAb lineages. Briefly, this was done by (i) estimating a phylogenetic for each lineage, (ii) reconstructing the sequence of each internal node, and (iii) using these reconstructed sequences in place of germline reference sequences to estimate R and S occurring over particular time periods of each lineage. It was argued that these analyses provided evidence that clonal selection on the CDRs of HIV bnAb lineages shifts from diversifying to purifying selection as broad neutralizing affinity is attained in bnAb lineages. It was generally hypothesized that, as neutralizing antibody lineages developed affinity for more conserved Env epitopes, their overall antibody structure would be more conserved

than in early development when affinity is likely much weaker and more amino acid changes were required to increase neutralization. This means the amino acid composition of a bnAb BCR is expected to be more conserved later in development, and thus is hypothesized to experience increased purifying selection compared to diversifying selection early during the development of the same lineage. Further, this shift in selective pressure was associated with an overall decrease in substitution rate in these lineages [Sheng et al., 2016]. This is thought to occur because an increase in purifying selection largely restricts nucleotide change to those at degenerate 3rd codon positions, leading to an overall decrease in substitution rate. This relationship between neutralization breadth, purifying selection, and decreased substitution rate is known as the affinity maturation selection (AMS) model [Sheng et al., 2016].

Despite popularity in application, R/S methods suffer from a number of drawbacks. In particular, they do not consider intermediate stages between data sequences (at the tips) and germline reference sequences. They generally consider sequences (and therefore shared derived mutations) as independent observations, and use summary statistics (R and S) rather than explicitly modelling the substitution process. These issues have been addressed in the broader field of molecular evolutionary biology using phylogenetic codon substitution models [Nielsen and Yang, 1998, Yang and Swanson, 2002]. These models parameterize the rate of nonsynonymous to synonymous (dN/dS) substitution using the parameter ω, which has a similar interpretation to the R/S ratio in previously described counting methods. Yang and Swanson [Yang and Swanson, 2002] demonstrated how such models could be used in a variety of forms to estimate different values of ω for different sites in a multiple sequence alignment ('partitioned ω model'), and for different branches along a phylogenetic tree ('branch ω model').

In this chapter, I extend the HLP17 codon substitution model within IgPhyML

to allow variation in ω among sites within an alignment and among branches within a phylogenetic tree. I then use these models and characterize the variation in selective pressure between and within three bnAb lineages to provide a finer resolution of the selective effects during affinity maturation than previously available through R/S methods. I apply this framework to analyse the same three bnAb lineages as those investigated in Chapter 3 and 4 as well as Sheng et al. [Sheng et al., 2016], which were sampled over a wide range of time scales (between 3 and 15 years). My results are broadly consistent with the hypothesis that dN/dS decreases over time in bnAb lineages. However, I find that in the younger lineages, CH103 and VRC26, lower dN/dS among clades is not actually associated with broad neutralizing breadth, seemingly at odds with Sheng et al's [Sheng et al., 2016] affinity maturation selection (AMS) hypothesis. Taken together, these results suggest that within a particular lineage, bnAb clades early in development are actually characterized by more persistent co-evolution than other clades, and only shift towards increased purifying selection in their CDRs late in development.

5.2 Materials and Methods

5.2.1 bnAb sequences and phylogenetic tree construction

Immunoglobulin heavy (IgH) chain multiple sequence alignments of the CH103 [Liao et al., 2013], VRC26 [Doria-Rose et al., 2014], and VRC01 [Wu et al., 2015] bnAb lineages used here are the same as those used in Chapter 4 (section 4.2.2). These alignments were a combination of sequences obtained from high-throughput repertoire sequencing, experimentally characterized monoclonal antibody sequences, and the putative germline sequence of each lineage (with its CDR3 masked with 'N' characters due to uncertainty in D gene assignment). Briefly, these sequences were downloaded from GenBank (www.ncbi.nlm.nih.gov/genbank) using acces-

sion numbers obtained from Liao et al. [Liao et al., 2013], Doria-Rose et al. [Doria-Rose et al., 2014], and Wu et al. [Wu et al., 2015] for the CH103, VRC26, and VRC01 lineages, respectively. I also downloaded between 10 and 20 experimentally validated monoclonal antibody (mAb) sequences for each lineage from Genbank using accession numbers provided by these three publications (see section 4.2.2 for details). As detailed more thoroughly in section 4.2.2, these sequences were aligned to their predicted V and J germline reference sequences, while their CDR3s were aligned using MUSCLE [Edgar, 2004]. I subsampled each of these alignments to include 199 sequences obtained through next-generation sequencing and combined this with each lineage's predicted germline V and J sequences, yielding 200 sequences total. I then manually aligned each lineage's respective monoclonal antibody sequences to these multiple sequence alignments. For simplicity, all phylogenetic analyses performed in this chapter use the NY98:H0 model, which is equivalent to tree estimation under the NY98 model using the lineage's germline sequence as a known root (see section 4.2.1).

5.2.2 Monoclonal antibody neutralization profiles

HIV broadly neutralizing antibody lineages are usually first characterized by the isolation of broadly neutralizing antibodies, which are then cultured as monoclonal antibodies and studied experimentally. In the CH103 and VRC01 lineages, peripheral blood mononuclear cells (PBMCs) were sorted from each patient based on their binding affinity to the resurfaced stabilized core 3 antigen (RSC3), an engineered HIV Env protein probe designed to expose the CD4 binding site (a broadly neutralizing epitope) while eliminating other antigenic regions [Wu et al., 2010]. Antibodies for the VRC26 lineage were identified by a different approach: by isolating IgG+ PBMCs from an HIV+ donor, culturing them, and then screening for neutralization against different HIV strains [Doria-Rose et al., 2014]. In all cases, isolated PMBCs were cultured as mAbs, sequenced, and tested for breadth and

potency of neutralization.

Neutralization breadth and potency for mAbs isolated for the CH103, VRC26, and VRC01 lineages [Liao et al., 2013, Doria-Rose et al., 2014] were measured using the TZM-bl assay [Montefiori, 2005, Montefiori et al., 2012]. This assay uses TZM-bl, a HeLa cell line containing a bioluminescent luciferase reporter gene which is activated after HIV infection by the viral Tat protein [Todd et al., 2012]. The assay is performed by mixing cell-free HIV of a particular strain with different dilutions of antibodies, and then adding TZM-bl cells. Luciferase expression in these cells is then quantified by luminescence and is proportional to the number of virus that enter the cells, which is assumed to be proportional to the number of infectious viruses in solution [Montefiori, 2005]. Neutralization potency is usually measured as the half-maximum inhibitory concentration, IC50 (μg/ml), which in this assay is the antibody concentration that results in a 50% reduction in luminescence compared to a negative control. These titrations generally begin at antibody concentrations of 50 μg/ml, so IC50 values above this concentration are usually labelled as $>$50 μg/ml and the antibody is classified as non-neutralizing for that strain.

TZM-bl assay is used on multiple HIV strains for each mAb to estimate the breadth and potency of the antibody's neutralization. Breadth is the percentage of HIV strains neutralized by the antibody (in Table 5.1 'neutralization' is IC50 $<$ 50 μg/ml, though this varies). Potency is usually summarized using the geometric mean IC50 value, either across all strains tested or only in neutralized strains. Table 5.1 summarizes these three statistics for the mAbs used in this chapter. These values were either obtained directly from, or re-calculated from data provided in the studies of each respective lineage [Wu et al., 2010, Wu et al., 2015, Liao et al., 2013, Doria-Rose et al., 2014]. Note that, while the canonical antibodies (CH103, VRC01, etc) of these lineages were tested against a wide array of HIV

strains representing global HIV diversity, the additional clonal variant antibodies of these different lineages were assayed against different numbers of HIV strains. This means the breadth and potency statistics reported here are not directly comparable between lineages. For example, the VRC26 mAbs used in this chapter were tested against 47 HIV strains, while CH103 mAbs were tested against only 6. VRC01 clonal variant mAbs come from two different publications - most were tested against 10 HIV strains, though the VRC01-3 were from an earlier publication and only tested against 7 of the same strains [Wu et al., 2010, Wu et al., 2015].

Lineage	MCA ID	Clade	Time (months)	Breadth (%)	Potency, neutralized strains	Potency, all strains	HIV strains	Source
CH103	1AZCET15	-	32.3	0	50	50.00	6	Liao et al. (2013). All and neutralizing strain potency calculated from data in Supp. Table 14.
	1A102R16	D	15.2	0	50	50.00		
	1AH92U	D	15.2	17	22	43.61		
	CH103	C, D	31.4	100	1.40	1.40		
	CH104	C, D	31.4	100	1.17	1.17		
	CH105	C, D	31.4	100	3.58	3.58		
	CH106	C, D	31.4	100	0.75	0.75		
VRC26	MC1	F	13.6	19	1.88	26.68	47	Doria-Rose et al. (2014). Neutralized strain potency from Figure 1b, all strain potency calculated from data in Extended Data Figure 2a.
	MC2	C	27.5	17	0.40	21.95		
	MC3	B, C	27.5	36	0.08	4.89		
	MC4	B, C	27.5	30	0.32	11.13		
	MC5	C	27.5	21	0.10	13.32		
	MC6	C	27.5	17	0.38	21.81		
	MC7	B, C	27.5	13	1.51	31.98		
	MC8	A, B, C	27.5	47	0.11	2.90		
	MC9	A, B, C	27.5	47	0.07	2.38		
	MC10	C	47.5	23	0.60	17.78		
	MC11	C	47.5	26	0.94	18.13		
	MC12	B, C	47.5	6	0.49	37.24		
VRC01	H3A022283	-	76	70	0.12	0.72	10	Wu et al. (2015). All strain potency given in Figure 3a, neutralizing strain potency calculated from these values.
	H0306C000	-	136	90	0.02	0.03		
	H08F11722	-	154	100	0.06	0.06		
	H4H015736	-	171	40	3.71	17.66		
	H5F185917	-	154	100	0.24	0.24		
	H0306E010	-	148	40	1.44	12.08		
	H0306G022	-	161	80	0.04	0.15		
	H08H00082	-	171	100	0.04	0.04		
	H0306B033	-	85	90	0.11	0.17		
	H0306O914	-	0	80	0.05	0.21		
	H0306O003	-	0	60	0.80	4.15		
	H0107E217	-	148	100	0.03	0.03		
	H0306G011	-	161	100	0.03	0.03		
	H0107C120	-	136	100	0.03	0.03		
	H0306O004	-	0	80	0.06	0.23		
	H0306A144	-	76	30	0.12	7.99		
	H08O9099	-	0	100	0.05	0.06		
	VRC03	-	161	40	0.13	1.64	7	Wu et al. (2010). Both potencies calculated from Supp. Table 2b-d
	VRC02	-	161	100	0.06	0.06		
	VRC01	-	161	100	0.07	0.07		

Table 5.1: Monoclonal antibody sequences used in this chapter. Breadth is the percentage of HIV strains tested with IC50 < 50 μg/ml, potency is measured as the geometric mean IC50 (μg/ml) of either all strains tested, or against strains neutralized (IC50 < 50 μg/ml). The clade in Figures 5.3 and 5.4 containing the antibody is also labelled. Concentrations listed as IC50 > 50 μg/ml in their original publication were recorded as 50 μg/ml.

5.2.3 Characterizing selection pressure in bnAb lineages

The IgH variable region sequence is divided into non-consecutive regions (see Figure 1.1) generally involved in antigen binding (complementary determining regions; CDR) and structural support (framework region; FWR). These regions are thought to undergo distinct selective pressures, with purifying selection operating on FWRs and neutral or positive selection acting on CDRs [Yaari et al., 2012]. To test these evolutionary hypotheses for the CH103, VRC26, and VRC01 lineages, I used a codon partitioned dataset and model to estimate separate ω parameters for the FWR and CDR regions (hereafter ω_{FWR} and ω_{CDR}) of each lineage. This model is referred to as the $2:\omega$ model (Table 5.2). The locations of these IMGT defined [Lefranc and Lefranc, 2001] regions were determined using IgBlast [Ye et al., 2013]. The $2:\omega$ is nested with the single ω model ($1:\omega$, Table 5.2), differing by a single free parameter. Because of this, the significance of the difference in fit between these two models can be compared by a likelihood ratio test, in which the likelihood ratio (LR) statistic (LRS; Equation 1.8) is approximated by a chi-squared distribution with one degree of freedom [Huelsenbeck and Rannala, 1997]. 95% confidence intervals for ω_{CDR} were estimated using profile likelihood curves. Briefly, ω_{CDR} values were fixed to specific values, and all other parameters were estimated using ML. This was repeated for a range of ω_{CDR} values. The lower confidence limit was determined as the highest ω_{CDR} value lower than the ML ω_{CDR} estimate with an LR at least 1.9 below the ML $\omega_{\hat{CDR}}$ LRS (where p=0.05 under the chi-squared approximation). The upper interval is the lowest ω_{CDR} value above the ML ω_{CDR} estimate with an LR at least 1.9 below the ML ω_{CDR} LRS. Profile likelihood curves are shown in Figure 5.2.

Model name	Type of variation	Description
1:ω	Partition and branch homogenous	Estimates a single ω for all alignment partitions and branches.
2:ω	Partition heterogeneous, branch homogenous	Estimates a separate ω for CDR and FWR partitions (ω_{FWR} and ω_{CDR}), but both are equal across all branches.
3:ω	Partition heterogeneous, branch heterogeneous	Estimates a separate ω for CDR and FWR partitions, but one partition uses a single ω for all branches and the other uses two ω's. Unless specified, the 3:ω analyses in this chapter use a single ω_{FWR} for all branches, and two ω_{CDR} parameters that are used on either side of a specified internal branch.

Table 5.2: The HLP17 models used in this chapter, with different types of variation in ω.

5.2.4 Detecting variation in selection pressure within a lineage

As detailed in section 5.1, previous work has suggested that selective pressure on the CDRs changed significantly over the course of development of the CH103, VRC26, and VRC01 bnAb lineages [Sheng et al., 2016]. This was shown by reconstructing ML sequences at each internal node of a bnAb lineage phylogenetic tree, and using these sequences to estimate the relative change in replacement vs silent mutations in the CDRs at different time points along the tree. In this section, I introduce a codon substitution model framework to more clearly test the hypothesis that selective pressure varies over the branches of a phylogenetic tree.

To estimate variation in selective pressure among branches in a phylogenetic tree, I use the partition and branch heterogeneous $3:\omega$ model (Table 5.2). Like the $2:\omega$ model, the $3:\omega$ model estimates separate ω values for FWR and CDR data partitions. However, the $3:\omega$ model estimates two separate ω values for two different sets of branches of a tree for one of these partitions (either FWR or CDR), while keeping the other homogenous across all branches in the tree. In principle, either

ω_{FWR} or ω_{CDR} could vary among branches in a tree; however, because [Sheng et al., 2016] only documented significant change in selective pressure within the CDRs of the three bnAb lineages investigated here, I use the $3:\omega$ model to estimate two ω_{CDR} values and one ω_{FWR} for each lineage, unless otherwise specified.

Estimating parameters under the $3:\omega$ and comparing the resulting likelihood to that obtained under the $2:\omega$ model allows one to test whether there is significant variation in ω_{CDR}, given two sets of branches within a tree. However, it leaves open the larger problem of determining which, if any, sets of branches show significant variation in ω_{CDR}. Though the number of possible sets of branches in a phylogenetic tree is far too vast to test exhaustively (especially using the computationally demanding models used here), it is not necessary to consider all possible combinations of branches. Within bnAb lineages, variation in ω_{CDR} is hypothesized to result from a shift to purifying selection after broad neutralizing affinity had been attained [Sheng et al., 2016]. Under the hypothesis that changes in antibody structure become less extreme as neutralizing (including broadly neutralizing) activity develops, we expect dN/dS to decrease over the course of bnAb evolution. If this is true, ω_{CDR} is expected to vary specifically among monophyletic clades within a bnAb lineage. This is a similar assumption used in correlated molecular clock models [Drummond and Suchard, 2010]. Thus, the method described and used in this chapter tests the $3:\omega$ model on all monophyletic clades of at least two sequences within a tree. More specifically, the 'clade search' method is applied to each bnAb lineage as follows:

1. The ML (rooted) topology, branch lengths, and κ, $\boldsymbol{\pi}$, and ω parameters are estimated under the $1:\omega$ model. This topology is fixed for the rest of the analysis.

2. Estimate ML branch lengths, κ, $\boldsymbol{\pi}$, ω_{FWR} and ω_{CDR} under the $2:\omega$ model. The maximum likelihood value obtained here, $\ell(2:\omega)$, is the log-likelihood under the hypothesis that ω varies among FWR/CDR partitions but is equal

among all branches in the tree.

3. Select an internal branch b. Using the $3:\omega$ model, estimate by ML one value of ω_{CDR} for this branch and all its descendant branches (i.e. the clade defined by b), and another ω_{CDR} for all other branches in the tree. At the same time, co-estimate branch lengths, κ, $\boldsymbol{\pi}$, and ω_{FWR} using ML. The maximum log-likelihood obtained from this model is $\ell(3:\omega)$.

4. Given the division of tree branches defined by internal branch b (step 3), calculate the likelihood ratio statistic (LRS) to determine improvement in likelihood for the $3:\omega$ model compared to the $2:\omega$ model. Here, $LRS = 2 * (\ell(3:\omega) - \ell(2:\omega))$.

5. Under the null hypothesis $(2:\omega)$, the distribution of LRS is expected to approximate a chi-squared distribution with one degree of freedom. I therefore calculate a p value using $1 - CDF(chi - squared(LRS, df = 1)))$.

6. Repeat steps 3-5 for all *n-1* internal branches of the tree, where n is the number of taxa.

To graphically summarize significant variation in ω_{CDR} among clades in the tree, I use the results of the hypothesis tests on each clade (step 4) to create a composite tree. First, I discard all hypothesis tests with a p value below a specified threshold of significance (see Section 5.2.4). Each branch within a phylogenetic tree belongs to multiple clades. To ensure that all clades that show a significantly higher LR for the $3:\omega$ model, compared to the $2:\omega$, model are at least partially represented in the composite tree, I group each branch with the smallest clade showing significant ω_{CDR} variation in which it is a member. I begin this process by labelling all branches with the ω_{CDR} estimated from the $2:\omega$ model. Starting from the root, I then move down the descendant branches of each node. When a clade defined by an internal branch shows a significantly higher LRS under the $3:\omega$ model than the $2:\omega$ model, all branches within that clade are grouped together. This 'overwrites' the grouping of the branches in that clade. I repeat this process

across the entire tree. Once grouped, the branches of these trees can be labelled by their ω_{CDR} or ω_{FWR} to show variation in these parameters across the tree. This process is shown graphically in Figure 5.1. The resulting tree will display branch lengths in units of genetic distance (substitutions per site; see section 1.9). However, some hypotheses I would like to explore using these analyses (e.g. the AMS model of [Sheng et al., 2016]), involve changes in selection over units of time. To make it easier to interpret these trees within that context, I use the least squares dating method [To et al., 2016] to convert the genetic distance trees obtained from IgPhyML into time-scaled trees whose branches are in units of weeks/months. These clock trees are shown in Figures 5.3 and 5.4, while trees with branch lengths in units of genetic distance (substitutions per site) are shown in Figures 5.5 and 5.6.

5.2.5 Multiple hypothesis test correction and method validation

Because the null model ($2 : \omega$) is a special case of the $3 : \omega$ model, in which all branches use the same value of ω_{CDR}, the two models are nested and can be compared on a single split using a likelihood ratio test, in which the LRS (Section 5.2.4) is approximately chi-squared distributed with one degree of freedom [Huelsenbeck and Rannala, 1997]; Figure 5.7). However, because I test *n-1* clades for each tree, the probability of obtaining a significant difference under the null hypothesis (i.e. type I error) for a particular tree increases with n. A common way of adjusting for this is to use a Bonferroni adjustment, in which p values are adjusted by multiplying them by the number of hypotheses tested (*n-1* in this case). The Bonferroni adjustment is expected to be conservative, however, because it assumes all hypotheses tested are independent. This is clearly not the case in phylogenetic split analysis, because the p values of nested clades are expected to be correlated. In such a situation, other techniques for controlling the false dis-

covery rate (FDR) are recommended, such as the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995] which controls the FDR under some conditions of positive dependence.

To explore the behaviour of the LRS statistic, as well as the sensitivity and specificity of different multiple test correction techniques, I performed a series of parametric bootstrap analyses using the CH103 lineage. For these analyses, I first subsampled the CH103 alignment to 50 sequences, including the germline V4-59/J4 sequence. I replicated the full procedure above (i.e. that in section 5.2.4) by estimating the ML tree topology, substitution parameters, and branch lengths for this alignment under the $1 : \omega$ model. Using the ML topology under the $1 : \omega$ model, I then re-estimated all substitution parameters and branch lengths under the $2 : \omega$ model (with ω_{CDR} and ω_{FWR} estimated as 0.78 and 0.49, respectively). Using the results of this partition model, I simulated 40 new sequence alignments comprising 50 sequences each. The first set of 20 alignments were simulated using the ML parameters that were estimated from the $2 : \omega$ model on the real empirical data, with ω_{CDR} and ω_{FWR} constant over the tree. These alignments constituted simulation under the null hypothesis of no variation of ω_{CDR} over the tree. I then simulated 20 more datasets using the same topology and substitution parameters, but constrained one clade of 6 sequences to have a $\omega_{CDR} = 0.1$, while the rest of the tree had a $\omega_{CDR} = 0.78$. These simulations represent the alternate hypothesis that ω_{CDR} varies among some clades of the tree (i.e. the $3 : \omega$ model). I then performed the clade search analysis (Section 5.2.4) on all 40 data sets, and characterized the sensitivity and specificity of the method under different FDR controlling strategies.

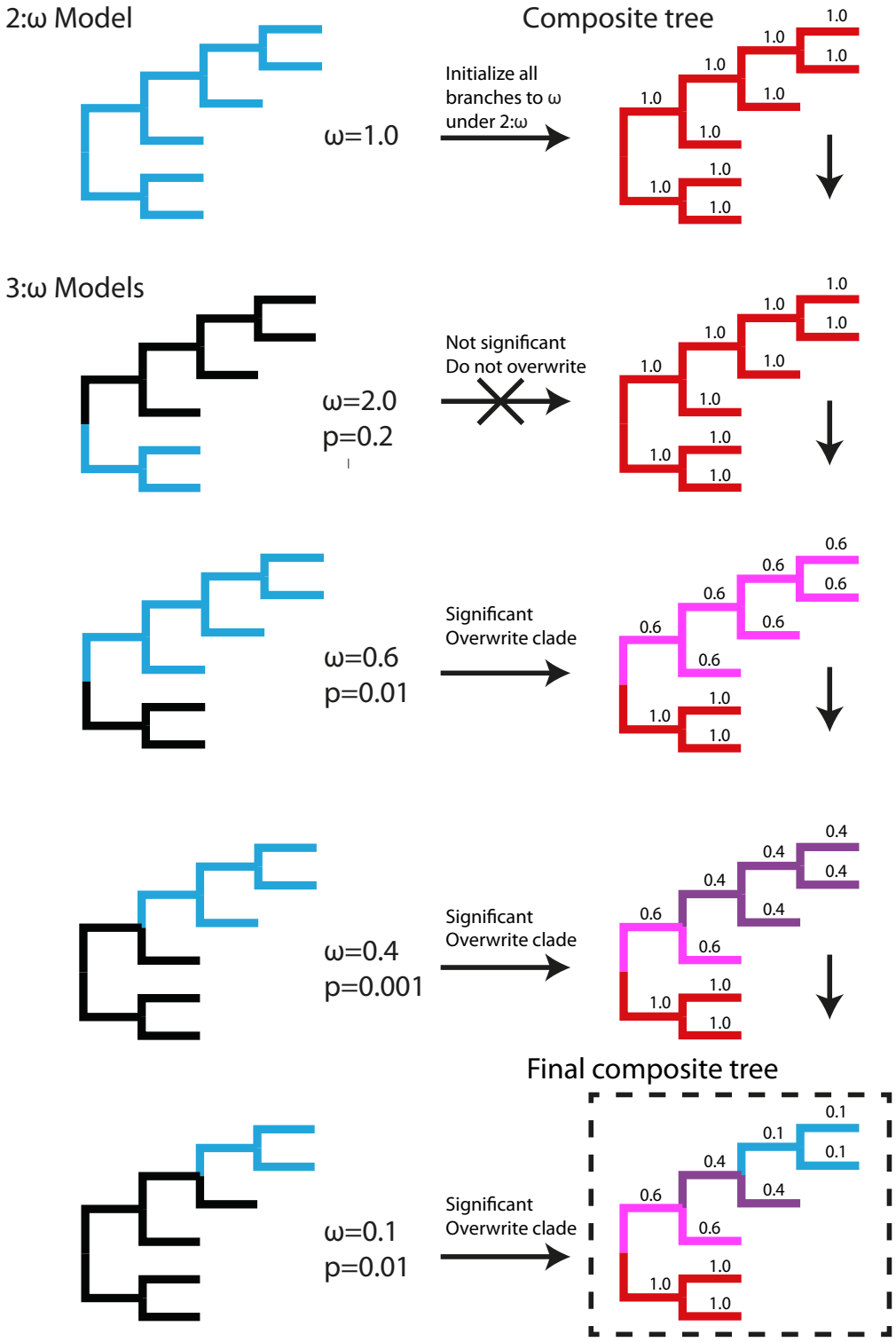


Figure 5.1: Process for creating a composite tree summarizing results of the clade search analysis. All clades with at least two descendant sequences are enumerated (left side). For each, I estimate a separate value of ω_{CDR} for blue vs black branches. To the right of each tree is the ML estimate of ω_{CDR} for the blue branches and the p value for the test. Because in this example I only fit one additional ω_{CDR} on a tree with four clades with >1 taxa, I determine significance using a Bonferroni corrected p value cut-off of $0.05/4 = 0.0125$. The right side shows how the composite tree is initialized with the ω_{CDR} from the $2:\omega$ model, and then 'overwritten' by clades that show a significantly better fit under the $3:\omega$ model than the $2:\omega$ model.

5.3 Results

5.3.1 CH103 and VRC26 show significant differences in ω between CDR and FWR

CH103 and VRC26 both showed significantly lower ML values of ω in their FWRs than CDRs (Table 5.3). Interestingly, VRC01 showed both a lower ω_{CDR} and ω_{FWR} than either CH103 or VRC26, and did not show a significantly better fit for the $2:\omega$ model over the $1:\omega$ model. These results for CH103 and VRC26 are consistent with previous analyses showing stronger purifying selection in FWRs and more neutral selection in CDRs [Yaari et al., 2012].

	CH103		VRC26		VRC01	
	ω	*L*	ω	*L*	ω	*L*
1:ω	0.46	-11097	0.57	-20614.9	0.34	-18146.1
2:ω FWR	0.42	-11093.9	0.49	-20606	0.34	-18146.1
2:ω CDR	0.54 (0.45, 0.64)		0.66 (0.58, 0.74)		0.34 (0.28, 0.4)	

Table 5.3: Maximum likelihood values of ω for each lineage, with accompanying maximum log likelihood ln(L). These are compared to the ω and L values for a partitioned model estimating a separate ω for CDR and FWRs. The value of L for the FWR/CDR models shown here is used as $\ell(2:\omega)$ in the 'clade search' analysis. 95% confidence intervals are shown in parentheses for ω_{CDR} estimates.

5.3.2 CH103 and VRC26 show significant ω_{CDR} heterogeneity among clades

For each of the three lineages, I performed the clade search procedure in section 5.2.4, allowing ω_{CDR} to vary on either side of a given branch, but constrained ω_{FWR} to be constant across the tree. Given the results of the simulation analysis (section 5.3.5), I used a Bonferroni adjusted p value to determine significance for the clade search procedure. CH103 and VRC26 both had significant differences in

ω_{CDR} among different clades, while VRC01 had no significant differences among clades. These results are shown graphically in Figure 5.3, and numerically in Table 5.4. Clades of interest are labelled in Figure 5.3, and are referred to by their lineage and letter (e.g. CH103:A refers to clade A in Figure 5.3a).

Analyses of the CH103 lineage indicated that a single clade exhibited a significantly lower ω_{CDR} than the rest of the tree (Clade CH103:B; Figure 5.3a). The ML estimate of ω_{CDR} = 0.54 (95% CI: 0.45, 0.64) under the $2:\omega$ model, which estimates a single ω_{CDR} for all branches in the CH103 lineage (Table 5.3). However, under the $3:\omega$ model, clade CH103:B had an ML estimate of ω_{CDR} = 0.23 (95% CI: 0.13, 0.37), and its subclade CH103:A had an even lower ML estimate of ω_{CDR} = 0.1 (95% CI: 0.05, 0.21). These results are summarized in Table 5.4.

The VRC26 lineage contained multiple clades with significant differences in ω_{CDR}, shown graphically in Figure 5.3b. Two large clades (VRC26:F and C) diverge early in the development of the lineage. VRC26:F contains the MC1 mAb sequence and is only sampled until week 59. VRC26:C, on the other hand, contains the mAb sequences MC2-12 and was also sampled at later time point. Some of its subclades eventually developed broad neutralizing affinity over three years (e.g. MC8 and 9; Table 5.1). Under the $2:\omega$ model, the whole VRC26 lineage had an ML estimate of ω_{CDR} = 0.66 (95% CI: 0.58, 0.74; Table 5.3). Under the $3:\omega$ model, the VRC26:F clade had a significantly higher ML estimate of ω_{CDR} = 1.01 (95% CI: 0.81, 1.25), while clade VRC26:C had a significantly lower estimate of ω_{CDR} = 0.54 (95% CI: 0.48, 0.6; Table 5.4). The subclades of VRC26:C also exhibited significant heterogeneity in ω_{CDR}, with the most extreme shift in ω_{CDR} being present in VRC26:D, which had an ML estimate of ω_{CDR} = 0.011 (95% CI: 0.05, 0.26; Table 5.4).

I repeated the clade search procedure using the $3:\omega$ model, but estimated a separate ω_{FWR} among branches of each tree while constraining ω_{CDR} to be constant

among branches. None of the three lineages tested had a significantly better fit under the $3:\omega$ model compared to the $2:\omega$ model (Figure 5.4a-c). Further, under the $2:\omega$ model ML estimates of ω_{FWR} were generally more consistent between the three lineages than ω_{CDR}, ranging from 0.34 in VRC01 to 0.49 in VRC26 (Table 5.3). Both of these results are consistent with the hypothesis that FWRs have a persistent role in supporting antibody structure and therefore experience stronger and more consistent purifying selection than the CDR region over during affinity maturation.

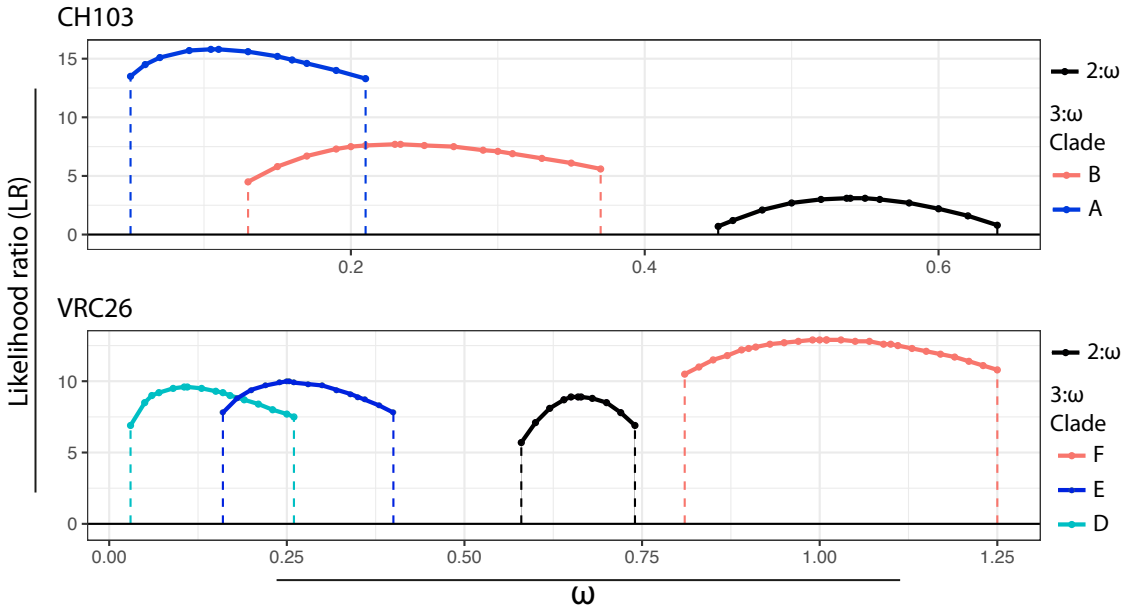


Figure 5.2: Profile likelihood curves for the CH103 and VRC26. One curve in each lineage shows the estimates of ω under the $2:\omega$ model, with the LR computed by comparing the ℓ values of the $2:\omega$ model to those under the $1:\omega$ model. Other curves show clades from the $3:\omega$ model, with their LRs computed by comparing the ℓ values from the $3:\omega$ model to those under the $2:\omega$ model.

Lineage	Model	Clade	ω_{CDR}	95% Confidence Interval		LRS
CH103	2:ω	-	0.54	0.45	0.64	-
	3:ω	A	0.1	0.05	0.21	15.4
		B	0.23	0.13	0.37	31.6
		C	0.57	0.34	0.98	0
		D	0.69	0.5	0.96	3.8
VRC26	2:ω	-	0.66	0.58	0.74	-
	3:ω	A	0.46	0.32	0.66	4.8
		B	0.52	0.42	0.65	8.6
		C	0.54	0.48	0.6	32.2
		D	0.11	0.03	0.26	19.2
		E	0.25	0.16	0.4	20
		F	1.01	0.81	1.25	25.8

Table 5.4: Estimates of ω_{CDR} under the $2 : \omega$ for entire lineage trees, and $3 : \omega$ models for labelled clades within the CH103 and VRC26 lineages. LRS = 2*($\ell(3 : \omega)$ - $\ell(2 : \omega)$) where L($2 : \omega$) is the maximum log-likelihood under the $2 : \omega$ model. See clade positions in Figure 5.3. See profile likelihood curves in Figure 5.2.

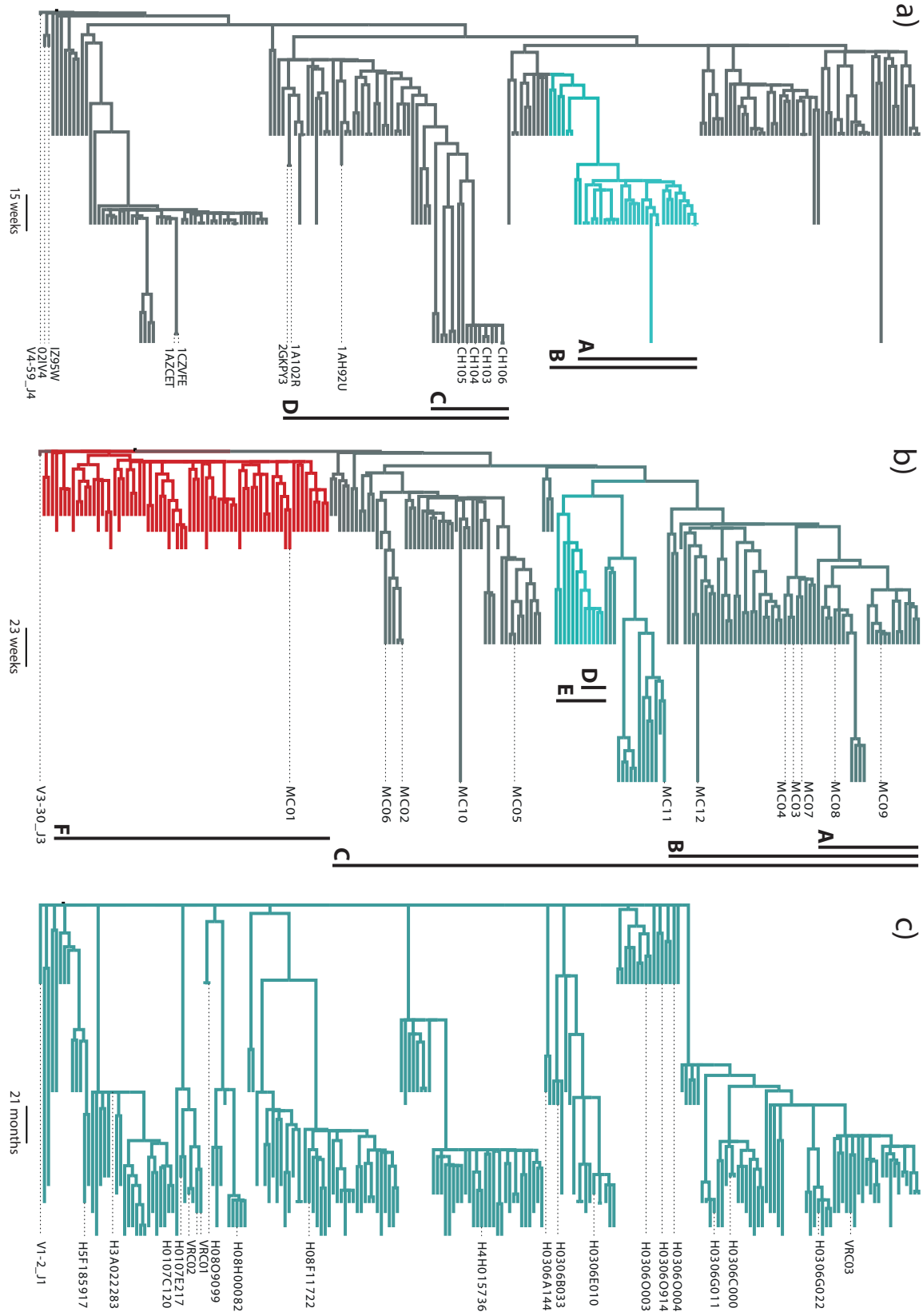


Figure 5.3: Composite trees showing significant differences in ω_{CDR} among clades in the a) CH103, b) VRC26, and c) VRC01 lineages. Branches are coloured by ω_{CDR} value, from cyan (ω_{CDR}=0) to red (ω_{CDR}=1.1). Monoclonal antibody sequences are labelled (Table 5.1). Branch lengths are in units of time specified on the scale bars below each tree. See Methods and Figure 5.1.

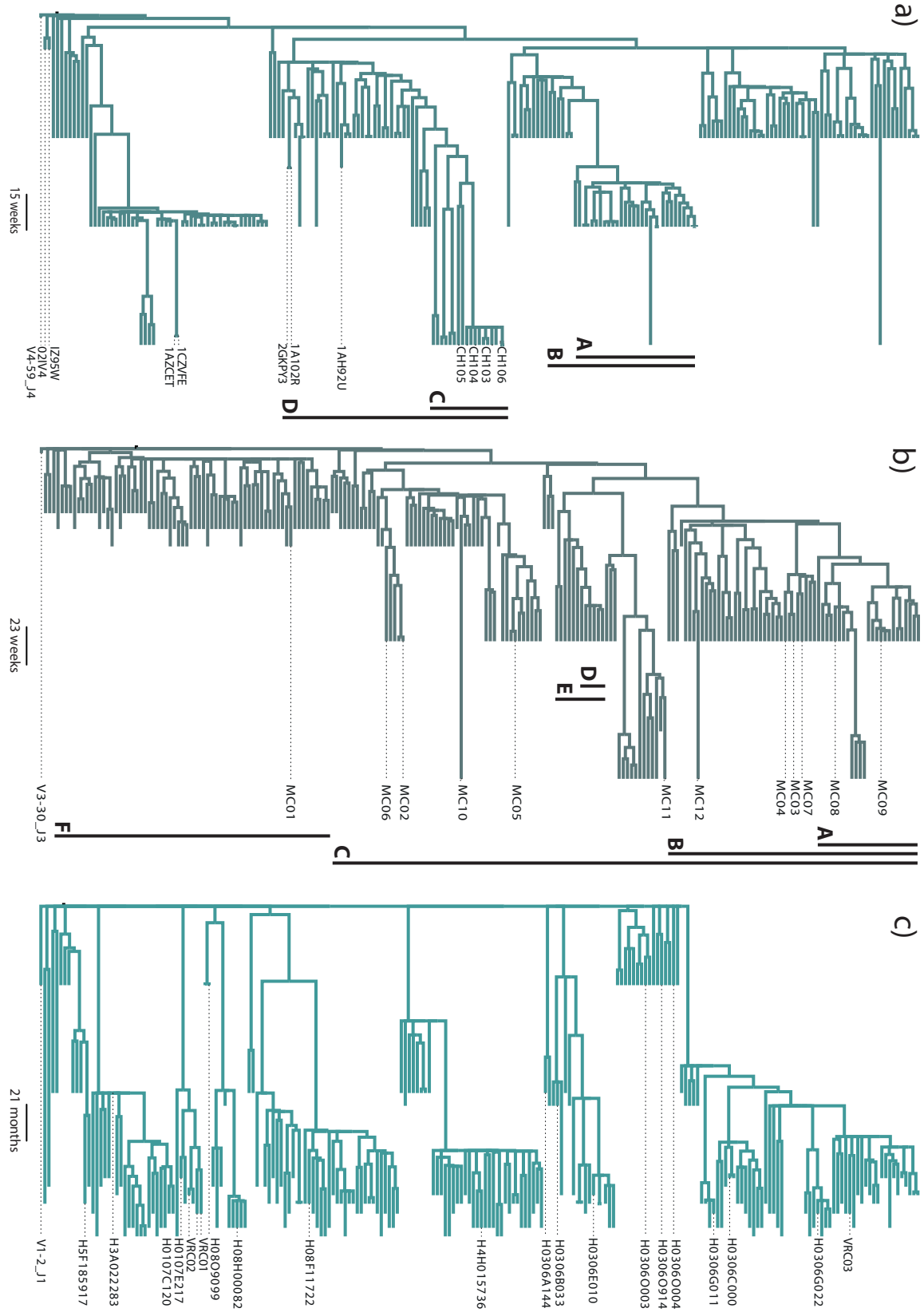


Figure 5.4: Composite trees showing significant differences in ω_{FWR} among clades in the a) CH103, b) VRC26, and c) VRC01 lineages. Branches are coloured by ω_{FWR} value, from cyan (ω_{FWR}=0) to red (ω_{FWR}=1.1). Monoclonal antibody sequences are labelled (Table 5.1). Branch lengths are in units of time specified on the scale bars below each tree. See Methods and Figure 5.1.

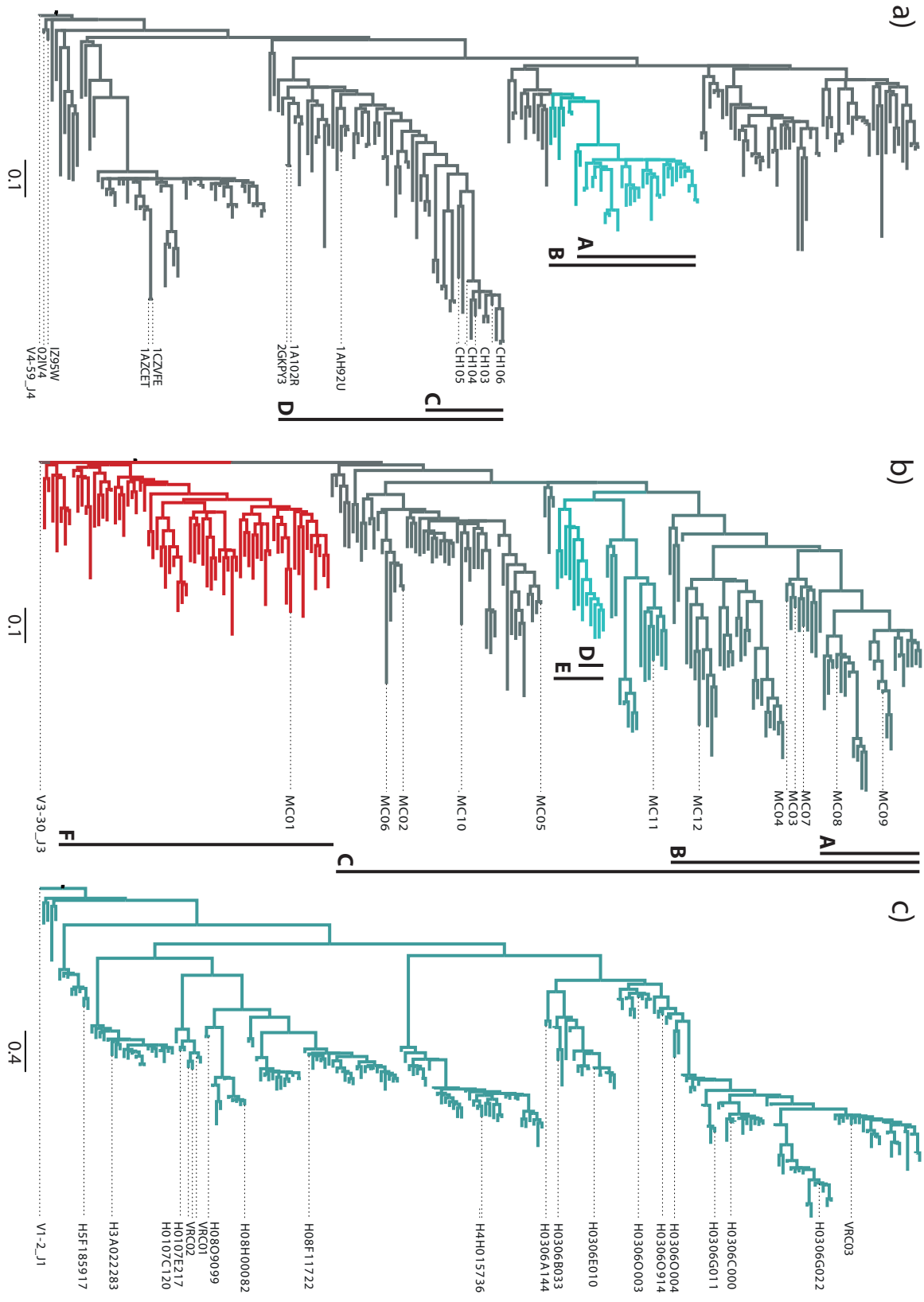


Figure 5.5: Composite trees showing significant differences in ω_{CDR} among clades in the a) CH103, b) VRC26, and c) VRC01 lineages. Branches are coloured by ω_{CDR} value, from cyan (ω_{CDR}=0) to red (ω_{CDR}=1.1). Monoclonal antibody sequences are labelled (Table 5.1). Branch lengths are in units of genetic distance (substitutions per site) specified on the scale bars below each tree. See Methods and Figure 5.1.

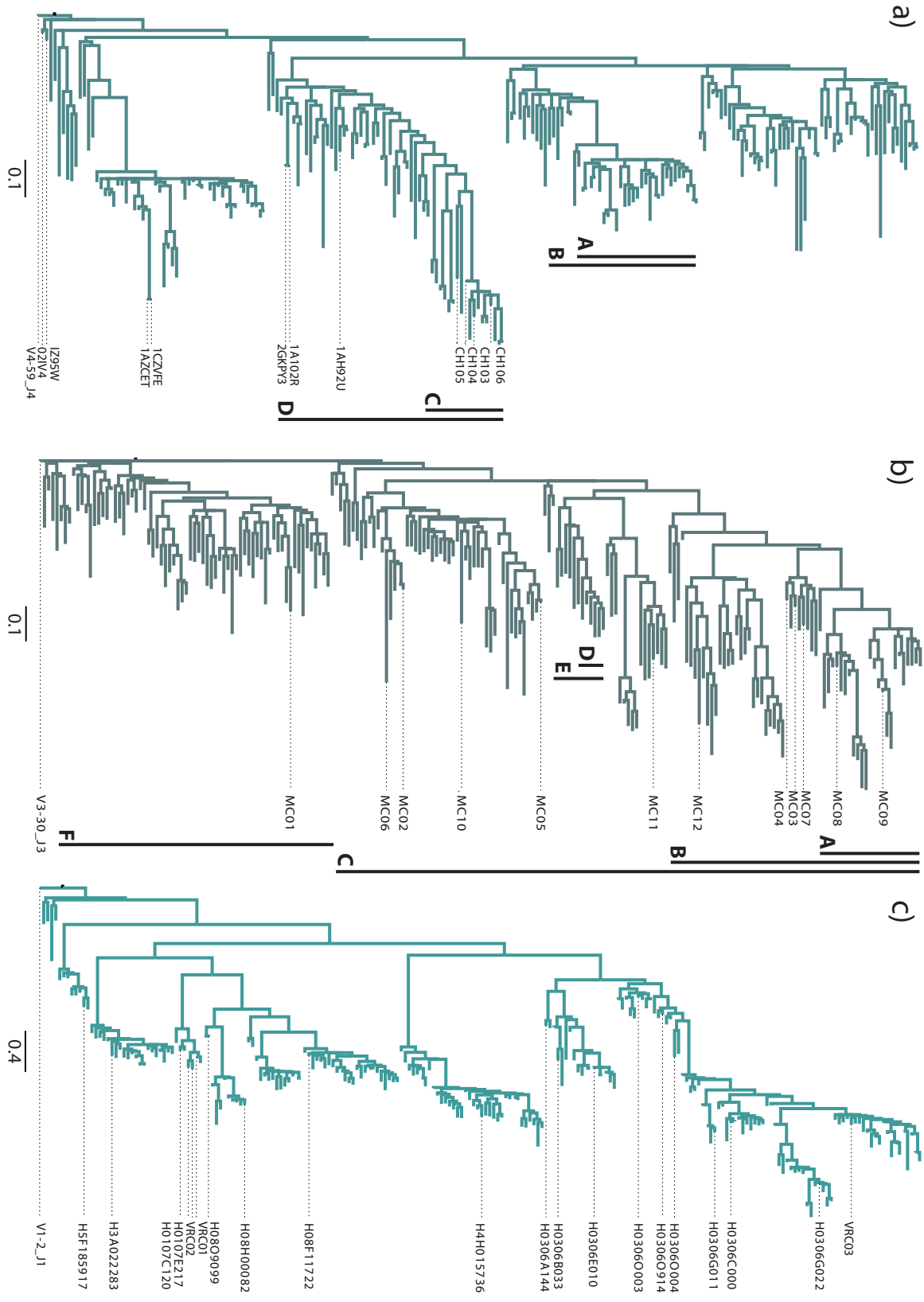


Figure 5.6: Composite trees showing significant differences in ω_{FWR} among clades in the a) CH103, b) VRC26, and c) VRC01 lineages. Branches are coloured by ω_{FWR} value, from cyan (ω_{FWR}=0) to red (ω_{FWR}=1.1). Monoclonal antibody sequences are labelled (Table 5.1). Branch lengths are in units of genetic distance (substitutions per site) specified on the scale bars below each tree. See Methods and Figure 5.1.

5.3.3 Low ω_{CDR} is generally associated with later timepoints

As shown in Figure 5.3a and Table 5.4, CH103:A, the clade within the CH103 lineage that shows the lowest likelihood ratio-test p value and the lowest ML estimate of ω_{CDR}, contains primarily sequences from week 92, a single sequence from week 144, and no sequences from week 53. Within the VRC26 lineage, the VRC26:F clade had a significantly higher ω_{CDR} estimate (1.01) than any other clade, and was not sampled at all after week 59. The VRC26:C clade, which was sampled from the earliest (week 38) until the latest timepoint (week 206), showed a lower ω_{CDR} estimate (0.54) than VRC26:F, with one of its subclades (VRC26:D) showing an even lower ω_{CDR} estimate (0.11) at week 144 (Figure 5.3b and Table 5.4).

In the VRC01 lineage, there was no evidence for significant difference in ω_{CDR} among clades in the tree under the $3:\omega$ model (Figure 5.3c). However, under the $2:\omega$ model, the ML estimate of ω_{CDR} (0.34; 95% CI: 0.28, 0.4; Table 5.3) across the VRC01 tree was lower than the ML ω_{CDR} estimates of the $2:\omega$ model and of most clades tested under the $3:\omega$ model in the CH103 and VRC26 lineages (Table 5.3, Figure 5.3). Crucially, while sampling of the CH103 and VRC26 lineages began soon after primary HIV infection, VRC01 is thought to have been sampled years after primary infection. Thus, the entire lineage represents much later timepoints in the course of HIV infection than those represented by the CH103 and VRC26 lineages.

Taken together, all of these results - both within and between these three bnAb lineages - are consistent with the hypothesis that the CDRs of bnAb lineages undergo increased purifying selection over time during chronic infection. I do not explicitly model a decrease in ω_{CDR} over time. However, all clades that show significantly lower ω_{CDR} estimates compared to ω_{CDR} under the $2:\omega$ model (e.g. CH103:A) are sampled at later timepoints in CH103 and VRC26, the two 'younger' lineages.

The clades of the VRC01 lineage do not show significant variation in ω_{CDR} under the $3 : \omega$ model, but the ω_{CDR} estimate across all branches of the 'older' VRC01 tree is lower than most in the VRC26 and CH103 lineages.

5.3.4 Broad neutralization is not associated with reduced ω_{CDR} among clades

It was hypothesized by Sheng et al. [Sheng et al., 2016] that the development of broad neutralizing affinity results in a decrease in ω_{CDR}. As detailed above in section 5.3.3, this is broadly consistent with my results, if B cell lineages as a whole are considered. However, a more complex picture emerges when considering differences among clades in estimated binding affinity. Strikingly, clades CH103:C and CH103:D, which contain the only broadly neutralizing monoclonal antibody sequences isolated from the CH103 lineage (CH103-6), did not show a significant difference in ω_{CDR} from the $2 : \omega$ model (Table 5.1, Figure 5.3a, Table 5.4). The nested clades that did show significant variation in ω_{CDR}, CH103:A and B, do not have any available monoclonal antibodies developed to characterize them, so it is not currently possible to directly measure and test the breadth of their binding affinity. It may be that CH103:A/B is actually a more potent bnAb than the previously characterized CH103-6 antibodies, which would be consistent with the AMS model [Sheng et al., 2016].

The VRC26 lineage showed similar results. VRC26:A, the clade that contains the most broad and potent monoclonal antibody sequences, MC8 and 9, had a ML estimate of ω_{CDR} = 0.46 (95% CI: 0.32, 0.66; Table 5.1, Table 5.4) and LRS=4.8, which under the Bonferroni correction is not significant enough to reject the $2 : \omega$ model. VRC26:B, which contains the monoclonal antibody sequences MC8 and 9, but also MC3, 4, 7, and 12, did show a significantly better fit under the $3 : \omega$ model, but had an ML estimate of ω_{CDR} = 0.52 (95% CI: 0.42, 0.65; Table 5.1, Table 5.4). The much

larger VRC26:C, which contains all monoclonal antibody sequences except MC1, also had a significantly better fit under the $3:\omega$ model, but similar ω_{CDR} estimate to VRC26:B, with ω_{CDR} = 0.54 (95% CI: 0.48, 0.6; Table 5.2, Table 5.4). Thus, all of the clades tested that contain MC8 and 9 and are able to reject the $2:\omega$ model show estimates of ω_{CDR} comparable to the $2:\omega$ model of the CH103 lineage (Table 5.3).

Clades VRC26:B and C both rejected the $2:\omega$ model, which had an ω_{CDR} = 0.66 (Table 5.3). However, because VRC26:C and VRC26:F define almost all sequences in the VRC26 lineage, it is likely that their rejection of the $2:\omega$ model was driven by high ω_{CDR} values of the VRC26:F clade. To test this, I repeated the analysis excluding the sequences from clades VRC26:F and VRC26:E, which resulted in an ML estimate of ω_{CDR} = 0.57 (95% CI: 0.48, 0.66) under the $2:\omega$ model, with no clade showing a significantly better fit under the $3:\omega$ model. This is consistent with the idea that the rejection of the $2:\omega$ model in clades containing MC8 and 9 (VRC26:A/B/C) was actually driven by strong variation in ω_{CDR} in VRC26:F and VRC26:E, rather than significant variation in ω_{CDR} in VRC26:A/B/C themselves.

Within the VRC26 lineage, only the nested VRC26:D/E clades showed both a signficantly higher LR under the $3:\omega$ model and had a much lower estimate of ω_{CDR} (0.11 and 0.25, respectively). However, the monoclonal antibody sequence most closely related to VRC26:D/E (MC11) shows much lower breadth and potency compared to MC3, 4, 8, and 9, which are contained in VRC26:B (Table 5.1).

Thus the developmental pathways leading to the development of the most potent bnAb sequences in both CH103 and VRC26 do not show significant variation in ω_{CDR} from the rest of their trees, and both have moderate ω_{CDR} values between 0.52 and 0.57 (Table 5.3, Table 5.4). Further, clades in the CH103 lineage containing the CH103-6 broadly neutralizing monoclonal antibody sequences (CH103:C/D)

show an asymmetric branching pattern, often taken to indicate persistent positive selection in viruses, such as influenza [Grenfell et al., 2004]. Both of these suggest that the selective dynamics of clades that develop into bnAbs are characterized more by persistent co-evolution than shifts to purifying selection, at least early in bnAb lineage development.

5.3.5 Method validation

To validate the clade search method, I simulated 20 alignments of 50 sequences each under similar parameters as those estimated from a subsampled CH103 dataset and the $2 : \omega$ model. I then performed the clade search method on each of these alignments. As expected, the distribution of all 980 LRS values observed followed a chi-squared distribution with one degree of freedom (Figure 5.7), supporting the argument that the standard chi-squared approximation for the LRS test [Huelsenbeck and Rannala, 1997] is valid for this method.

I next simulated a further 20 datasets using the same parameters, but constrained a clade of six taxa (hereafter the 'constrained' clade; coloured blue in Figure 5.8a) to have a $\omega_{CDR} = 0.1$ (i.e. the $3 : \omega$ model). I then performed the clade search method on each simulated data set. Simply using uncorrected p values correctly rejected the $2 : \omega$ model on all branches in the constrained clade in 19/20 alignments (Figure 5.8b). However, this technique also erroneously rejected the $2 : \omega$ model on other branches outside the constrained clade. Using uncorrected p values also spuriously rejected the $2 : \omega$ model for some clades when analysing alignments simulated under the null ($2 : \omega$) model as well (Figure 5.8d). In contrast, the Bonferroni correction yielded high sensitivity and specificity under the parameters tested. Interestingly, the Bonferroni correction performed at least as well as both the BH [Benjamini and Hochberg, 1995] and BY (Benjamini and Yekutieli

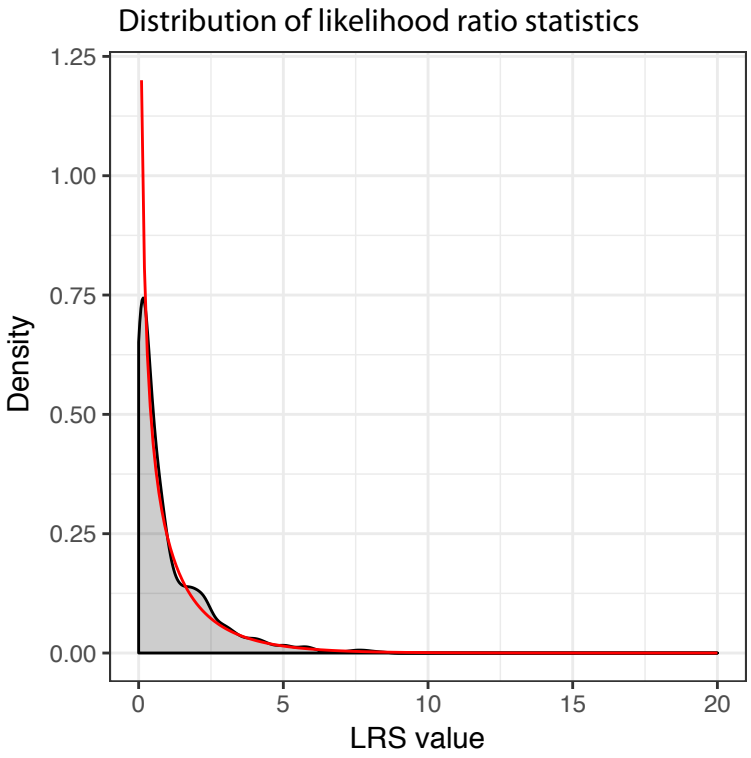


Figure 5.7: Distribution LRS statistics for 20 simulated alignments of 50 sequences under the null hypothesis of no variation in ω_{CDR} among branches ($2 : \omega$). Density estimate of LRS values is shown in grey, while the chi-squared probability density function with one degree of freedom is shown in red.

2001) FDR strategies in sensitivity and specificity (Figure 5.8b). When applied to datasets simulated under the $2 : \omega$ model, all three p value correction strategies correctly failed to reject the $2 : \omega$ model on 980 clades tested (Figure 5.8c and 5.8d).

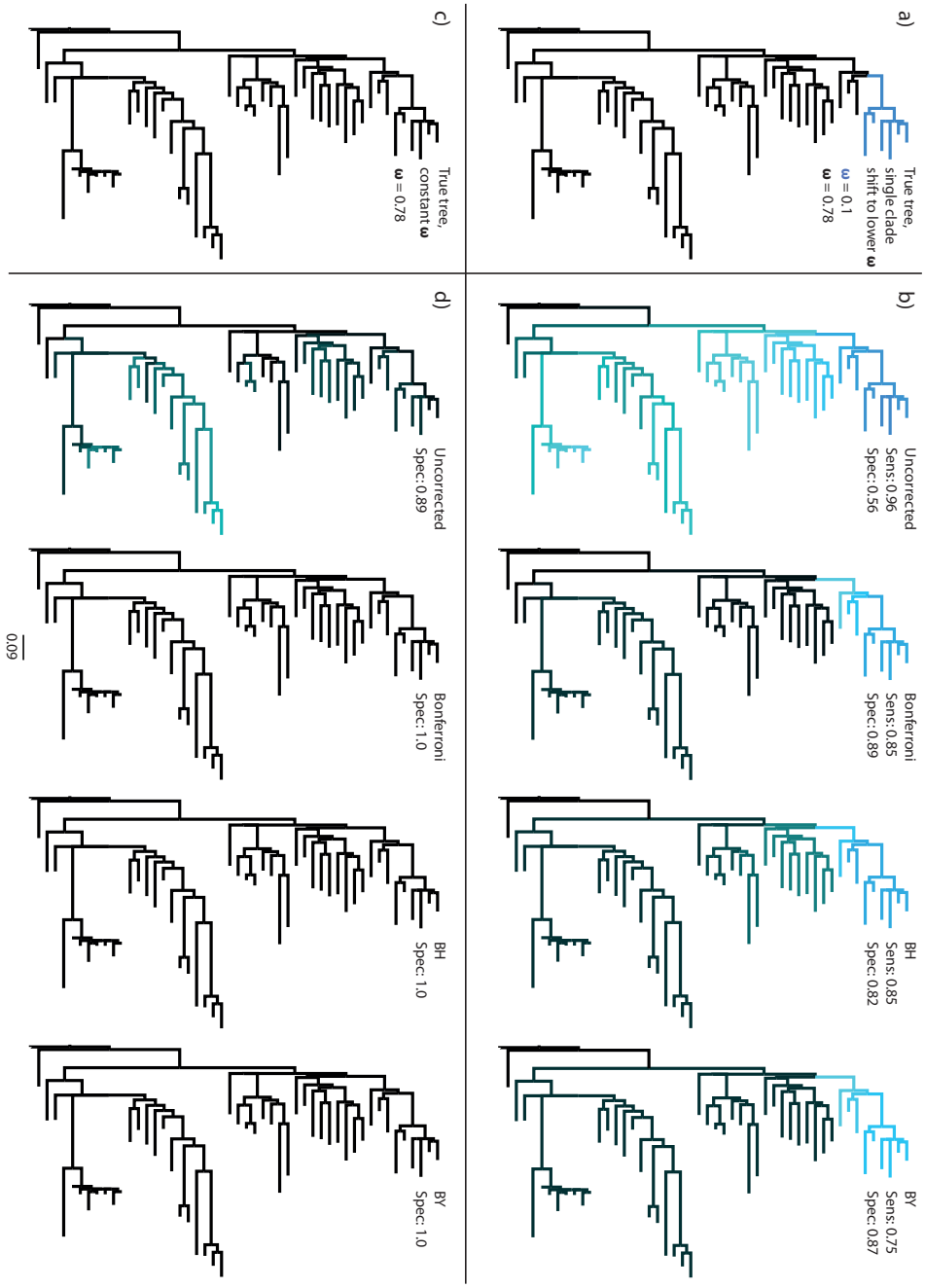


Figure 5.8: Sensitivity and specificity of different p value correction methods explored using simulations. Two groups of 20 alignments were simulated. The first used tree a), in which all black branches used a ω_{CDR} = 0.78, while all blue branches used ω_{CDR} = 0.1; the second used tree c), with all branches using ω_{CDR} = 0.78. In panels b) and d), each clade within a tree is coloured based on the number of times it is identified as having a significantly different ω_{CDR} from the rest of the tree (scaled from black to blue, 0-20, respectively). Trees in b) and d) show how different p value corrections affect the results (see 4.3.5 above). Sensitivity (true positive rate) is the proportion of constrained internal branches correctly identified (100 in a/b, 0 in c/d), while specificity (true negative rate) is the proportion of all non-constrained branches correctly identified across all 20 alignments for each simulation type (a/b vs c/d).

5.4 Discussion

The hypothesis that dN/dS in the CDR regions of the immunoglobulin heavy chain (estimated as ω_{CDR} in this chapter) decreases as broad neutralizing affinity is attained was first proposed by Sheng et al. [Sheng et al., 2016] as means of explaining an apparent decrease in the substitution rate of the CDR over time in bnAb lineages [Wu et al., 2015]. It is hypothesized that, as neutralizing activity increases within a lineage, fewer amino acid changes are likely to lead to increased affinity. Thus, bnAb lineages are expected to experience increased purifying selection compared to early on in the affinity maturation. This is known as the affinity maturation selection (AMS) model. Though Sheng et al. [Sheng et al., 2016] did not explicitly estimate dN/dS in their phylogenetic trees, they showed evidence (described in section 5.1) that this shift in dN/dS was associated with decreasing rates of substitution in the same three lineages tested here. They also showed that a competing hypothesis - that the change in substitution rate was caused by a decrease in total sequence mutability - was unlikely to explain the decrease in dN/dS as differences in sequence mutability correlated much more weakly with the change in substitution rate.

Though the analyses in this chapter explicitly model ω_{CDR} differences among clades within a B cell lineage tree they do not explicitly model a change in ω_{CDR} over time, and, as such, do not directly test the hypothesis that ω_{CDR} decreases over time in the CDRs of bnAb lineages. The 'clade search' method implemented here is further limited in that it can only model differences among separate clades. If all clades have shifted to purifying selection concurrently, ancestral branches that experienced higher ω_{CDR} will not be detected. Given these constraints, however, my results are consistent with the hypothesis that decreased ω_{CDR} occurs in later timepoints during bnAb lineage development. The only significant shift towards diversifying selection was detected in a clade consisting of early timepoints

(Figure 5.3b: VRC26:F clade), while shifts to purifying selection occur in clades sampled primarily in later timepoints (Figure 5.3: CH103:A/B, VRC26:D/E, VRC01).

However, my results also suggest that the AMS model is currently incomplete. Monoclonal antibodies from different clades of the CH103 and VRC26 lineages show substantial variation in binding breadth and potency (Table 5.1;[Liao et al., 2013, Doria-Rose et al., 2014]), and clades within these lineages show substantial variation in ω_{CDR} (Figure 5.3). The AMS model predicts that broadly neutralizing clades would be associated with lower ω_{CDR} values. However, this is not the case. The clades that do show a shift towards purifying selection are clearly not associated with the bnAb sequences of the CH103 and VRC26 lineages. As detailed in section (5.3.4), there is no evidence of a significant shift towards lower ω_{CDR} in the clades leading to the characterized bnAb sequences in the CH103 or VRC26 lineages. However, without experimental verification through monoclonal antibodies it is not possible to know neutralizing profiles of the clades that did show lower estimates of ω_{CDR} values (CH103:A/B, VRC26:D/E).

I believe these results may be explained through a simple generalization of the AMS model. The AMS model recognizes that neutralization breadth, purifying selection, and fitness (expected number of offspring cells) should be closely correlated in antibody lineages. Wide neutralization breadth leads to better antigen binding and thus higher fitness, and is also expected to lead to selection for conservation of BCR structure (i.e. purifying selection). However, this does not apply only to broadly neutralizing antibodies - any neutralizing antibody should show this relationship.

One can visualize affinity maturation as occurring on a fitness landscape over the space of possible BCR sequences (Figure 5.9), with fitness valleys containing antibodies of narrow neutralizing breadth, higher substitution rate, and diversifying

selection (high ω_{CDR}) interspersed with fitness peaks of wider breadth, lower substitution rate, and purifying selection (low ω_{CDR}). Most fitness peaks are relatively low- to mid-breadth neutralizing antibodies. Lineages that climb these peaks increase their fitness but are then effectively stuck at local maxima (so long as the landscape remains constant). Lineages that eventually develop broad neutralizing breadth do not get stuck climbing these local peaks, and thus spend most of their long development in the valleys with higher ω_{CDR} values. This model suggests that in the CH103 and VRC26 lineages, the shifts observed in ω_{CDR} are in clades that ascended local fitness peaks, while putative bnAb clades were still in the process of finding the broad neutralizing peak and thus do not yet show significant signs of lowered ω_{CDR} (Figure 5.9). All major sampled clades of the VRC01, by contrast, have had time to ascend broad neutralizing peaks, and thus only show low ω_{CDR} values.

Of course, this fitness landscape is not static, especially in chronic HIV infection. Fitness peaks are constantly shifting as HIV evolves and adapts in response to selective pressure from the host immune response, with some peaks disappearing and others emerging. Non-broadly neutralizing antibody peaks correspond to epitopes that may be easily lost due to HIV escape variants, and thus shift more readily. It is possible that the neutralizing antibody lineages that ascend them may be more susceptible to elimination when unable to respond effectively to HIV escape variants outside of their neutralization breadth. This would explain why clades CH103:A and VRC26:E, which show the lowest estimates of ω_{CDR} are found at low frequency at their lineages' latest timepoint. Peaks representing broad neutralizing affinity, on the other hand, are expected to be rare but far more static than narrowly neutralizing antibody peaks. No peak within this landscape is expected to be fully static - even in HIV patients who do develop broadly neutralizing antibodies, viral populations inevitably escape neutralizing antibody responses once established. This fitness landscape concept may provide a useful model for de-

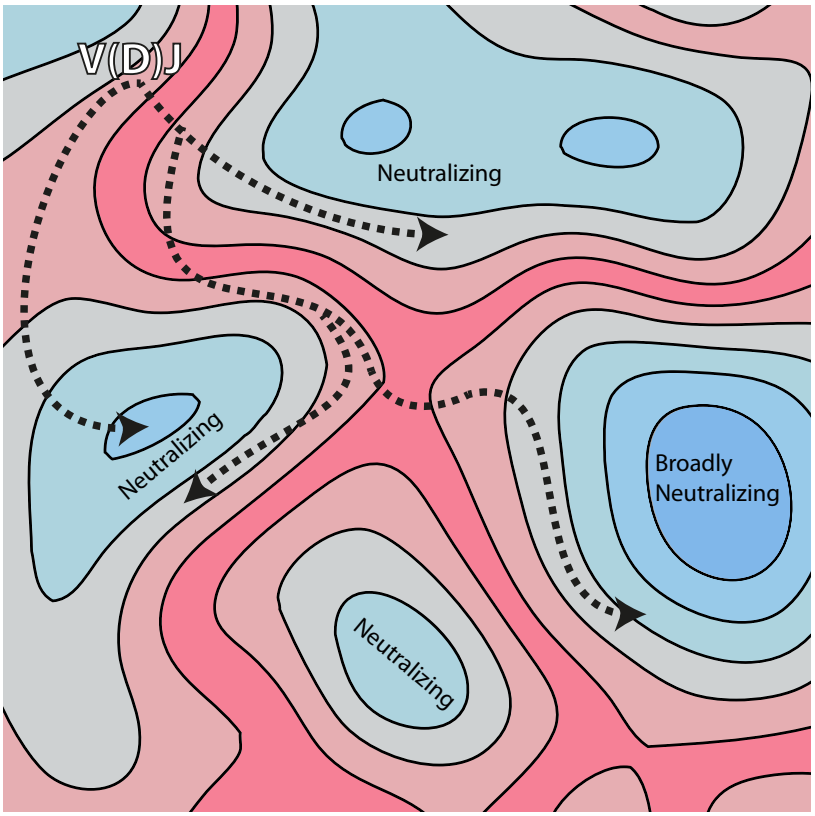


Figure 5.9: Proposed fitness landscape showing positive relationship between fitness, neutralization breadth, and purifying selection. Fitness valleys (red) are associated with low breadth, high substitution rate, and diversifying selection, while peaks (blue) are associated with wider breadth, lower substitution rate, and purifying selection. Dotted lines are clades from an antibody lineage originating from the 'V(D)J' spot in the top left. Arrows show each clade's location at the time of sampling. This example is consistent with the CH103 lineage at week 92 in Figure 5.3a: one clade has ascended a local peak and shifted to low ω_{CDR}, while the others, including the putative bnAb clade, have not yet ascended any peaks and show no signal of lowered ω_{CDR}. Map outline freely available from blog.spoongraphics.co.uk.

scribing cooperation among B cell lineages - with one driving evolution in HIV to be more sensitive to the other (i.e. one lineage 'pushing down' a local fitness peak, which 'pushes up' a broad fitness peak that is ascended by a different lineage; for example in [Gao et al., 2014].

The dynamics of clonal selection early in HIV bnAb development is important for HIV vaccine design, whose ultimate goal is to elicit broad neutralizing responses by directing antibody lineage evolution towards broad neutralizing affinity [Haynes et al., 2012]. The results presented here suggest that during early

antibody lineage development, putative bnAb clades are characterized by persistent selection, and only shift towards purifying selection in their CDRs late in development once broad affinity has been attained. These results suggest complex co-evolution between HIV and bnAb lineages in the early years of chronic infection. Given this, it seems likely that only by considering BCR and HIV evolution together will further progress will be made in understanding how the dynamics of clonal selection lead to the development of broad neutralizing affinity.

Chapter 6

Phylogenetic analysis of B cell repertoires

6.1 Introduction

Model-based phylogenetic methods have shown promise in elucidating important information about the adaptive immune response in humans. Chapters 3, 4, and 5 focused on phylogenetic analysis of three broadly neutralizing antibody (bnAb) lineages. These lineages were identified by screening peripheral blood mononuclear cells (PBMCs) from HIV infected individuals for broad neutralizing activity, culturing these cells which then secrete monoclonal antibodies (mAbs), sequencing their immunoglobulin transcripts, and matching these resulting Ig sequences to bulk high-throughput repertoire sequencing data from the same patient [Wu et al., 2010, Liao et al., 2013, Doria-Rose et al., 2014]. Previously, I developed a series of phylogenetic substitution models (HLP17) designed to characterize the evolution of these lineages by estimating additional model parameters that represent important features of affinity maturation, such as somatic hypermutation (SHM) hotspot motifs (Chapters 3 and 4), and variation in clonal selection pressure (Chapter 5).

However, phylogenetic analysis of individual, experimentally-verified B cell lineages is only a small part of the larger field of antibody genetics. Many B cell repertoire analyses use large collections of lineages that have been identified entirely by clustering sequences into clonal clusters from high-throughput repertoire sequencing (RepSeq) data, typically based on their predicted V and J germline ancestor, and CDR3 sequence similarity (e.g. [Laserson et al., 2014, Stern et al., 2014, Heiden et al., 2017]). Many lineage analyses stop after the clustering step and use the size distribution of these clonal clusters to characterize the behaviour of the repertoire as a whole. This may be sufficient to detect large-scale perturbations of the B cell repertoire, such as those caused by lymphoma [Bashford-Rogers et al., 2013], and acute viral infection (Chapter 2). Other analyses use the tree topology of specific lineages to detect more subtle processes, such as B cell migration between tissues. For instance, Stern et al. [Stern et al., 2014] used maximum parsimony lineage trees to determine that B cells migrate from the cervical lymph nodes to the brain in patients with multiple sclerosis, and Tabibian-Keissar et al. [Tabibian-Keissar et al., 2008] used lineage tree topology to demonstrate migration between different gut locations.

One problem with RepSeq lineage analysis is that the lineages predicted by clonal clustering techniques typically contain fewer unique sequences, and are less genetically diverse than bnAb lineages. This, coupled with the large number of individual lineages in the dataset may be one reason why non-parametric tree building approaches such as maximum parsimony have remained standard practice for lineage analysis of RepSeq datasets [Barak et al., 2008, Gupta et al., 2015], whilst parametric phylogenetic approaches are more often used in bnAb lineage analysis (Chapters 3, 4, and 5; [Liao et al., 2013, Doria-Rose et al., 2014, Wu et al., 2015, Sheng et al., 2016].

The lack of genetic diversity in many of these lineages is especially problematic for a complex, highly parametric substitution model such as HLP17. The number of parameters in the HLP17 model is unlikely to have been a problem for the analyses of bnAb lineages in the last three chapters, because each of the three lineages analysed contained several hundred sequences and high levels of genetic diversity, making estimates of additional parameters more reliable (see simulation analyses in Chapter 3). Estimating HLP17 parameters from smaller and more homogenous lineages would likely lead to model over-fitting, reulting in unreliable ML parameter estimates and little improvement in likelihood.

This is not a unique situation in phylogenetic analysis. Those studying within-host evolution of viruses in cohorts of multiple individuals often face a similar situation of multiple lineages; for example, when each patient is represented by only a small amount of genetic diversity [Rodrigo et al., 2003, Edo-Matas et al., 2011]. One means of solving this issue is to use 'hierarchical' phylogenetic models. These are phylogenetic models in which contain multiple 'levels' of parameters, some of which are shared between lineages, while others are estimated for each lineage individually [Suchard et al., 2003]. This allows estimation of some parameters using the whole data set. For example, Rodrigo et al. [Rodrigo et al., 2003] applied such a hierarchical approach to a set of HIV-infected patients in order to jointly estimate the rate of substitution rate and the proportion of individuals that did not respond to antiretroviral therapy (i.e. experienced on-going viral evolution).

This chapter shows how a hierarchical framework may be used to characterize entire B cell repertoires. While individual lineages identified within RepSeq data may lack the high genetic diversity exhibited by long-lived bnAb lineages, the genetic diversity of the entire collection of lineages that comprise the repertoire is large. Further, some biological processes represented by parameters of the HLP17 model, such as targeting preferences of SHM, should be consistent among lineages

during affinity maturation. This would allow some parameters to be shared between lineages, allowing their simultaneous estimation from all lineages within a repertoire dataset. In this chapter, I provide a proof-of-principle demonstration that the HLP17 model can be expanded into a hierarchical model that takes this information into account. I apply these hierarchical HLP17 models to three RepSeq datasets isolated from non-progressing HIV+ children.

6.2 Methods

6.2.1 Hierarchical substitution models

The goal of these repertoire analyses is to estimate parameters shared between all lineages in the repertoire, which requires likelihood calculations at the repertoire level. As an extension of equation 1.6, the likelihood of the repertoire of n lineage trees (L_R) is the product of likelihoods of its individual lineage trees, which are assumed to be conditionally independent given their germline sequences (G_i):

$$L_R = \prod_{i=1}^{n} P(X_i|\boldsymbol{Q}_i, \boldsymbol{t}_i, T_i, G_i) \tag{6.1}$$

where X_i is the multiple sequence alignment, and $\boldsymbol{Q}_i$, $\boldsymbol{t}_i$, T_i, and G_i are, respectively, the HLP17 $\mathbf{Q}$ matrix (Eq. 4.1), tree branch lengths, tree topology, and germline sequence for clonal lineage i. In the analyses presented here, T_i is estimated for each lineage individually using the NY98 model [Nielsen and Yang, 1998] and fixed during the repertoire analysis. Branch lengths in the vector $\boldsymbol{t_i}$ are independent among lineages and are estimated by ML. Likelihoods of each tree are computed using Eq. 1.6.

The parameters within each $\boldsymbol{Q}_i$ matrix can vary depending on the type of hierarchical model chosen. The motif targeting preferences of SHM are expected to be

the same for all lineages, so the h^q parameters for discrete models (Eq. 4.3) and the scaling coefficients for semi-empirical models (Eq. 4.4) are here shared between lineages (i.e. they will the same in each $\boldsymbol{Q}_i$). Further, my previous analyses have shown that κ is consistently ˜2 in bnAb lineages (Figure 3.2, Appendix Figures B.4 and B.5), which is consistent with previous estimates [Cowell and Kepler, 2000], so κ is also shared among lineages. The 12 base frequency parameters used to calculate codon frequencies ($\boldsymbol{\pi}$) are also shared between lineages.

Lineages potentially vary in the type and degree of clonal selection experienced, thus allowing linages to have separate ω parameters, which may be useful for detecting lineages undergoing interesting selection processes. The ω parameter could potentially vary in a variety of ways. In the simplest case, ω is constant between all partitions and lineages, and all $\boldsymbol{Q}_i$ matrices are identical, thereby estimating a single ω for the entire repertoire ($r : 1 : \omega$, equivalent to the $1 : \omega$ model, Table 5.2). Though not explored here, this model could be extended to estimate two ω values for the repertoire - a separate ω for CDR and FWR partitions, but each being constant across all lineages ($r : 2 : \omega$ model, similar to $2 : \omega$ in Chapter 5, Table 5.2). It is also possible to allow ω to vary among lineages, thus estimating a separate ω for each lineage (referred to as the $r : n : \omega$ model) or even to allow separate estimation of a ω_{CDR} and ω_{FWR} for each lineage (referred to as the $r : 2n : \omega$ model).

6.2.2 Study subjects and repertoire sequencing

Data used in this chapter were taken from a subset of 11 perinatally HIV-infected South African children followed longitudinally from birth, as part of the larger Paediatric Early HAART and Strategic Treatment Interruption Study (PEHSS) [Mphatswe et al., 2007]. These children were termed paediatric slow progressors (PSPs) because, prior to universal ART recommendation for paediatric HIV infection, they

had not met ART initiation criteria by the time they reached 5 years of age (i.e. $CD4 > 350$ cells/mm3 or WHO stage III or IV) [Leitman et al., 2017]. The three children studied in this chapter were chosen because they were the three best characterized in [Leitman et al., 2017]. Leitman et al.[Leitman et al., 2017] divided their patients into two groups - those with higher viral load (viraemic controllers; VC) and those with lower viral load (viraemic non-controllers; VNC). In general, viral loads were 0.87 log_{10} copies/ml lower in the VNC group than the VC group [Leitman et al., 2017]. PS133 was a viraemic controller while PS021 and PS517 were viraemic non-controllers, indicating they had a low or high viral count, respectively.

Children were sampled ten times each over ten years, from birth. Sample dates were an average of 366, 413, and 385 days apart for PS133, PS527, and PS021, respectively. The first 5 samples of PS133 contained no PBMCs, and neither did the fourth sample of PS517. At each of these dates, PBMCs were taken, sorted into memory or naive cells, and a 3' multiplex PCR with barcoded primers was then used to amplify the IGHV region, which was then sequenced using high-throughput paired-end Illumina MiSeq sequencing. More specifically, RNA samples were first reverse-transcribed into cDNA. This was performed in a 20 μl reaction containing 10μl total RNA ($<$ 500ng), 2 μl nuclease free H_2O, 1 μl of a single JH-R primer (Table 6.1) (10μM) and 1 μl dNTP mix (10μM). This was heated to 65C for 5 minutes and incubated on ice for at least 1 minute to bind primers. 4μl 5X First-Strand buffer, 1μl DTT (0.1M) and 1μl SuperScript RT were added to this mixture for a total volume of 20μl. This was incubated at 50C for 60 minutes and 70C for 15 minutes to complete reverse transcription. This cDNA was purified using a 1.8x Agencourt AMPure XP bead cleanup. A PCR was conducted to amplify the variable region, and used 25μl 2x KAPA buffer, 1μl UNI3-JH-C reverse primer (10μM), 1μl forward V gene multiplex primer mix (10μM each primer) (Table 6.1) 11μl H_2O and 12μl purified cDNA sample in a total reaction volume of 50μl. The

thermocycler conditions of the PCR are given in Table 6.2.

Primer sequence	Name (Reaction)
TGTCCAGCACGCTTCAGGCTNNNNTNNNNTNNNNCT TACCTGAGGAGACGGTGACC	JH-R (RT-PCR)
TGTCCAGCACGCTTCAGGC	UNI3-JH-C-R (RACE PCR)
GGCCTCAGTGAAGGTCTCCTGCAAG	VH1-FR1 (RACE PCR)
GTCTGGTCCTACGCTGGTGAAACCC	VH2-FR1 RACE PCR)
CTGGGGGGTCCCTGAGACTCTCCTG	VH3-FR1 RACE PCR)
CTTCGGAGACCCTGTCCCTCACCTG	VH4-FR1 RACE PCR)
CGGGGAGTCTCTGAACATCTCCTGT	VH5-FR1 RACE PCR)
TCGCAGACCCTCTCACTCACCTGTG	VH6-FR1 RACE PCR)

Table 6.1: PCR primers used in these analyses.

Number of cycles	Temperature	Time
1 cycle	95C	5min
5 cycles	98C	5sec
	72C	2min
5 cycles	98C	5sec
	65C	10sec
	72C	2min
30 cycles	98C	20sec
	62C	1min
	72C	2min
1 cycle	72C	7min

Table 6.2: PCR cycle procedure used in these analyses.

6.2.3 Sequence processing and clonal clustering

Raw paired end sequence data from memory B cells were assembled and filtered of low quality reads (Phred score $>$ 20) and non-functional re-arrangements using pRESTO [Heiden et al., 2014]. I then aggregated all reads from each patient and aligned them to IMGT reference V and J alleles using IgBlast [Ye et al., 2013]. Many sequences had germline relative insertions, which are removed by Change-

O [Gupta et al., 2015]. This brought together flanking nucleotides that are not actually adjacent in the sequence, potentially creating false codons (even premature stop codons in many cases). Because these sequences were going to be analysed as codon alignments, I masked the left flanking nucleotide of these insertions (i.e. replaced it with an 'N') to prevent the creation of false codons. I then used the SHazaM R package [Gupta et al., 2015] to determine an optimal clustering distance cut-off. In this clustering method, sequences are first separated into groups based on their predicted V and J germline segments and their CDR3 lengths, all of which must be identical within a clone. From there, the nucleotide distance (i.e. proportion of nucleotide differences in the CDR3) between each sequence and its nearest neighbour is calculated. These measurements of 'nearest neighbour distances' are usually bimodal (Figure 6.1). The lower and upper modes of this distribution are thought to represent within and among clone distance comparisons, and a cut-off for grouping two sequences together within a clone is typically chosen between them, shown in Figure 6.1 (see [Gupta et al., 2017]. I then clustered these sequences using the chosen cut-offs to generate clonal clusters, with matched germline V and J reference sequences and germline CDR3s masked with 'N' characters using Change-O [Gupta et al., 2015]. I then exported the IMGT aligned sequences for each clonal group and trimmed nucleotides from the beginning of the sequences in each alignment so that all sequences were in the correct reading frame. I found that many sequences were identical across their IMGT-aligned regions but variable outside of these regions. These sequences were collapsed with other identical sequences at the same sample point.

6.2.4 Lineage analysis

For each lineage identified by the clonal clustering process in section 6.2.3, I calculated (i) the mean pairwise difference (MPD), which is the mean number of nucleotide differences between all pairwise comparisons of sequences within the lin-

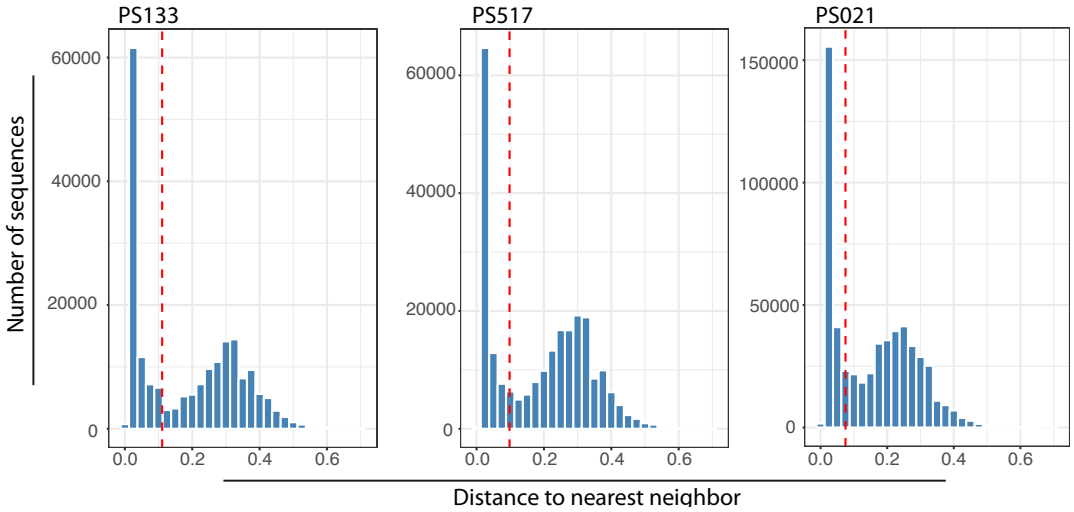


Figure 6.1: Genetic distance (proportion of junctional nucleotide differences) of each sequence to its nearest neighbour (most similar) sequence. The lower and upper modes of this distribution are generally interpreted as within and between clone comparisons, respectively. In these figures all sequence comparisons to the left of each vertical red line are grouped together as a clone.

eage, excluding the germline sequence, and (ii) the number of unique sequences (NUS) in the lineage excluding the germline sequence - sequences from different timepoints were counted as unique even if they had the same sequence. I then selected all lineages with at least 10 NUS from each patient (hereafter the 10NUS dataset). This resulted in 173, 317, and 673 lineages for PS133, PS517, and PS021, respectively. For computational efficiency I estimated the ML topology for each of these lineages individually using the standard, unrooted NY98 model. This topology was fixed (for each lineage) throughout the rest of the analysis. I then estimated ML substitution parameters and branch lengths for the $r : 1 : \omega$ and $r : n : \omega$ models for each patient using the hierarchical approach described in section 6.2.1. These models both estimated κ, y_h, m_h, and $\boldsymbol{\pi}$ across lineages within a patient. The $r : 1 : \omega$ model estimates a single ω across all lineages, while $r : n : \omega$ model estimates a separate ω for each lineage.

I also explored how ω_{CDR} varied among clades within each lineage using the clade search method under the $2 : \omega$ and $3 : \omega$ models introduced in Chapter 5 (Table 5.2). Because the clade search method is computationally intensive, I selected a

subset of lineages from each patient for these analyses. Because the power to detect heterogeneity in ω among clades likely depends on the number of unique sequences and genetic diversity of the lineage, I multiplied the MPD and NUS of each lineage, and selected the 8 lineages from each patient with the highest resulting value of MPD * NUS. For each of these lineages, I fixed the κ, y_h, m_h, and π parameter values to the ML estimates obtained under the $r : n : \omega$ model as applied to the 10NUS dataset for each patient. I then repeated the clade search method (Section 5.2.4) by first estimating a separate ω_{CDR} and ω_{FWR} for each lineage, and then allowing ω_{CDR} to vary among clades. To increase computational efficiency, I only considered clades with ≥ 3 sequences for the $3 : \omega$ model, compared to ≥ 2 sequences in Section 5.2.4. I created composite trees for each of these 8 lineages per patient (see Section 5.2.4).

6.2.5 Statement of contribution

All of the PCR and sequencing procedures in this section was performed by Dr. Astrid Gall at the University of Cambridge, as well as the sequencing core facility at the Wellcome Trust Sanger Institute. The computational analyses presented are wholly my own work.

6.3 Results

6.3.1 Clonal clusters are dominated by low diversity clones at single timepoints

The majority of sequences in all three patients are found in low-diversity clones ($MPD < 0.01$) that sampled at a single timepoint only (62%, 73%, and 74% in

PS021, 517, and 133, respectively; Figure 6.2). While many of these clones contain only a few sequences, others are massive - the largest is in patient PS517 and contains 25,167 sequences, only 126 of which are unique. It is not immediately clear what these large, homogenous lineages represent.

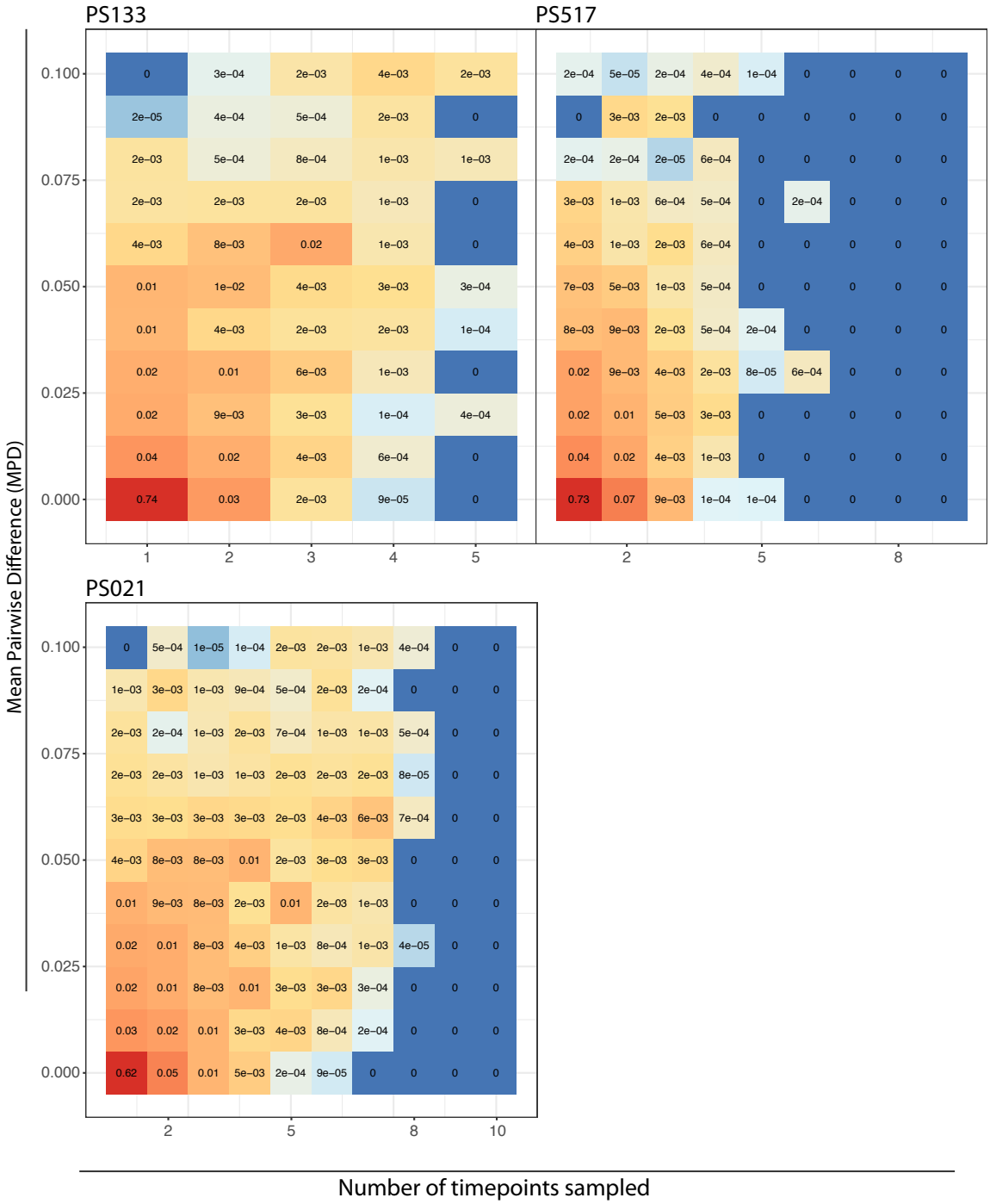


Figure 6.2: Proportion of reads in clonal lineages, binned by mean pairwise difference (MPD) of the sequences in the lineage on the y axis, and by the number of timepoints in which the lineage was detected on the x axis. Note how the majority of reads in all three patients are found in lineages found only in one timepoint, and at low genetic diversity (MPD < 0.1)

6.3.2 Some SHM parameter estimates are consistent among patients

Base frequency parameters are shared between lineages because they are expected to be consistent within a single patient. Interestingly, however, estimates of these parameters are also consistent among all three patients, regardless of whether the $r:1:\omega$ or $r:n:\omega$ models were used (Figure 6.3). Further, as predicted in Chapter 3, these ML estimated base frequencies are distinct from their empirical estimates obtained from the same datasets.

Estimates of κ were also consistently ~2 between all three patients (Figure 6.4). Interestingly, estimated S5F scaling coefficient parameters (see Eq. 4.4) were noticeably different between the $r:1:\omega$ and $r:n:\omega$ models for patients PS021 and PS133; specifically, m_h was ~2 under the $r:n:\omega$ model, but ~1 under the $r:1:\omega$ model. Generally, parameter estimates between PS021 and PS133 were quite similar, while PS517 was more divergent. This may be due to the fact that lineages in patients PS021 and PS133 tended to be more genetically diverse, leading to more stable parameter estimates for these patients. It is also interesting that patient PS517 showed very similar parameter estimates for the $r:1:\omega$ and $r:n:\omega$ models. This may be because, compared to the PS021 and PS133 repertoires, fewer individual PS517 lineages show increases in LRs under the $r:n:\omega$ model compared to the $r:1:\omega$ model (see Section 6.3.3).

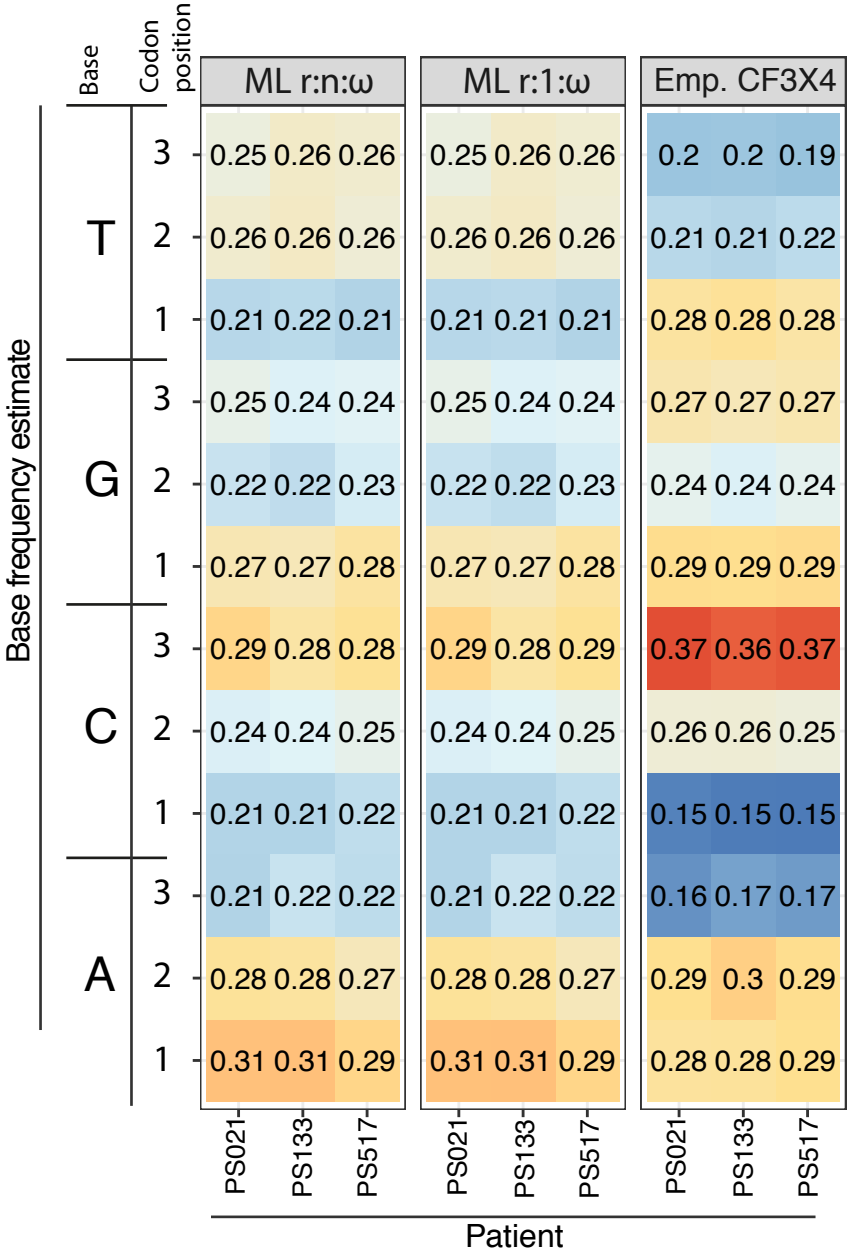


Figure 6.3: ML Estimates of base frequencies under $r:1:\omega$ and $r:n:\omega$ models are consistent between patients. These 12 parameters (four bases at 3 codon positions) are used to calculate $\boldsymbol{\pi}$ values used in $\mathbf{Q}$ matrix. Note that these estimates are consistent between patients, but different from their empirically calculated base frequencies ('Emp. FC3X4'). These models were applied to the 10NUS datasets of each patient.

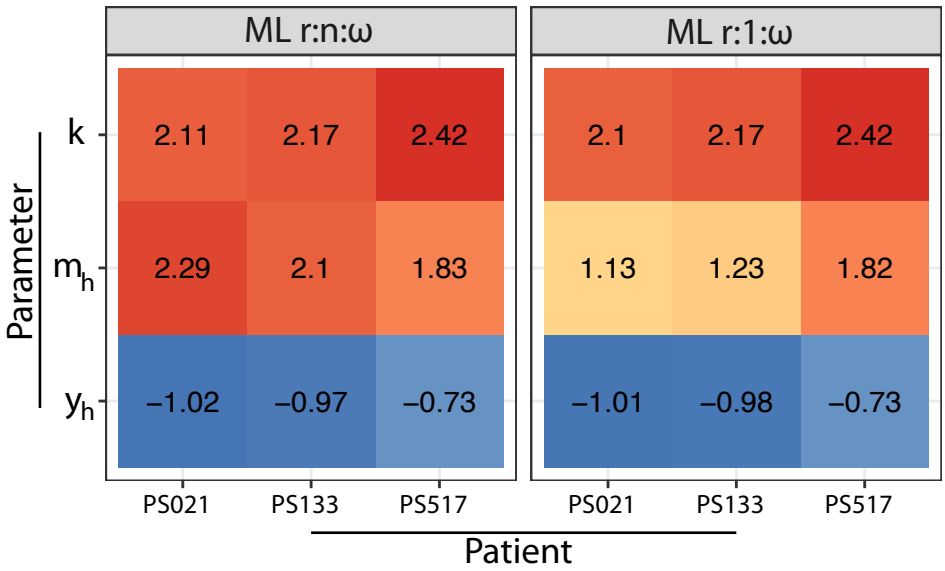


Figure 6.4: ML Estimates of κ and S5F scaling parameters (y_h and m_h) under $r:1:\omega$ and $r:n:\omega$ models. κ and y_h are fairly consistent between patients, but not m_h. These models were applied to the 10NUS datasets of each patient.

6.3.3 Variation in selective effects within and between repertoires

Individual lineages within repertoires show variation in ω under the $r : n : \omega$ model. Results make it clear that the models are only nested at the level of the repertoire, but not for each subtree: ~5% of lineages in PS133 and PS021, as well as ~0.3% (one lineage) in PS517, actually show an $LR < 0$ under the $r : n : \omega$ model. This is possible in these analyses because other substitution parameters besides ω are estimated at the repertoire level under the $r : n : \omega$ model. In other words, some of the estimated parameters of the $r : 1 : \omega$ model are different to those under the $r : n : \omega$ model, which can lead to lower likelihoods for some lineages (Figure 6.6).

Figure 6.5 shows the distribution of ω estimates under the $r : n : \omega$ model within the repertoire. These results show that most reads are contained in lineages with ML estimates of ω similar to the ω estimate obtained under the $r : 1 : \omega$ model. This is as expected, because the estimate under the $r : 1 : \omega$ model is effectively the average ω across all lineages in the repertoire. However, there is also substantial variation in ω estimates among lineages and, as shown in Figure 6.6, many of the lineages with divergent estimates of ω show a noticeable improvement in LR.

Figure 6.6 shows how the ω esimate for each lineage under the $r : n : \omega$ model relates to the likelihood ratio (LR) for that lineage under the $r : n : \omega$ vs. $r : 1 : \omega$ models. These LR values followed a roughly parabolic relationship to ω, centred at the ω estimate from the $r : 1 : \omega$ model. This makes sense, considering that lineages with an ML value of ω close to the estimate under the $r : 1 : \omega$ model are expected to show no improvement in likelihood by estimating their own value of ω. Those lineages with ML ω values further away from the $r : 1 : \omega$ estimate will show a larger improvement in likelihood under the $r : n : \omega$ model.

I also repeated the 'clade search' method detailed in Chapter 5 (Section 5.2.4) for

the largest eight lineages in each of the three patients. These are shown in Figures 6.7, 6.8, and 6.9. These lineages showed a variety of tree shapes ranging from asymmetric (e.g. PS133 tree 8, Figure 6.7) to balanced (e.g. PS133 tree 1, Figure 6.7). Among these lineages, PS021 tree 8 (Figure 6.9) showed a significantly better fit under the $3:\omega$ model, with one clade (A) having a higher ω_{CDR} estimate ($\omega_{CDR} = 2.56$), and another (B) with a lower estimate ($\omega_{CDR} = 0.22$), compared to its estimate under the $2:\omega$ model ($\omega_{CDR} = 0.51$). PS133 tree 7 also showed showed a significantly better fit under the $3:\omega$ model for one clade (A), which had a lower ω_{CDR} estimate ($\omega_{CDR} = 0.26$) compared to its estimate under the $2:\omega$ model ($\omega_{CDR} = 0.36$; Figure 6.7). It is interesting to note that the largest lineages of PS517 have far fewer sequences compared to the other patients. This may be one reason why so few PS517 lineages showed large increase in LR under the $r:n:\omega$ model compared to the $r:1:\omega$ (Figure 6.6).

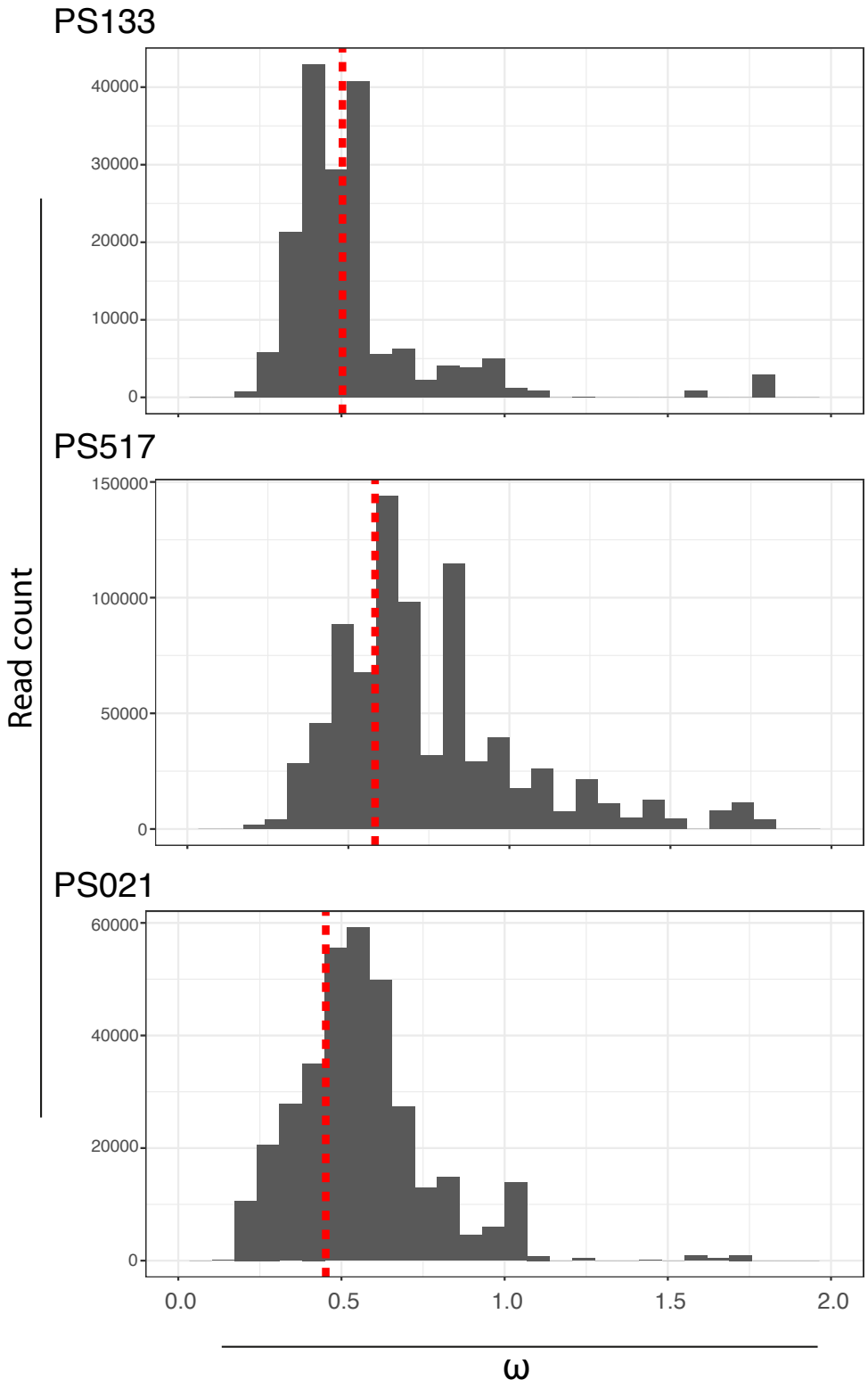


Figure 6.5: Histograms of read counts contained within lineages with different values of ω estimated under the $r : n : \omega$ model. Each read is represented by the ω value estimated for the lineage containing it under the $r : n : \omega$ model. The count of reads for each ω are shown in the above histograms. The ML estimate of ω under the $r : 1 : \omega$ model for each patient is represented by a vertical dotted line. Note that PS133, PS517, and PS021 had 3, 3, and 1 lineages with an estimate of $\omega > 2$ that are not shown in this plot.

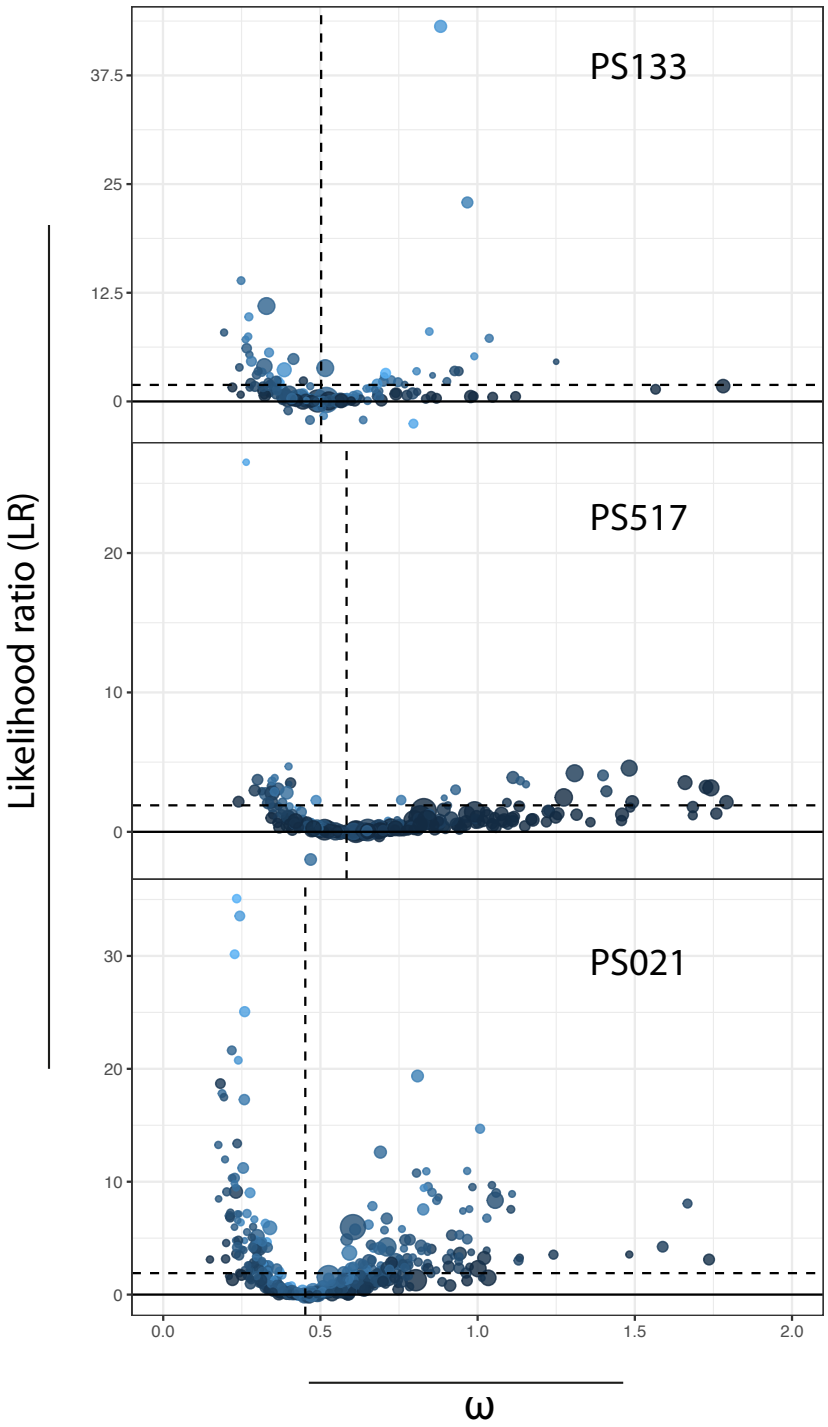


Figure 6.6: Estimate of ω for individual lineages under the $r : n : \omega$ model for each patient. Each circle represents an individual lineage. The size of each circle corresponds to the lineage's size, while the colour corresponds to its genetic diversity (dark blue = low MPD, light blue = high MPD). The ML estimate of ω under the $r : 1 : \omega$ model for each patient is represented by a vertical dotted line. The horizontal line corresponds to an LR of 1.9, which corresponds to an LRS of 3.8, which corresponds to a p value < 0.05 for a likelihood ratio test (Eq. 1.8). However, the LR values computed for each lineage of the repertoire here cannot be compared in this way, so this is only shown as a reference. Note that PS133, PS517, and PS021 had 3, 3, and 1 lineages with an estimate of $\omega > 2$ that are not shown in this plot.

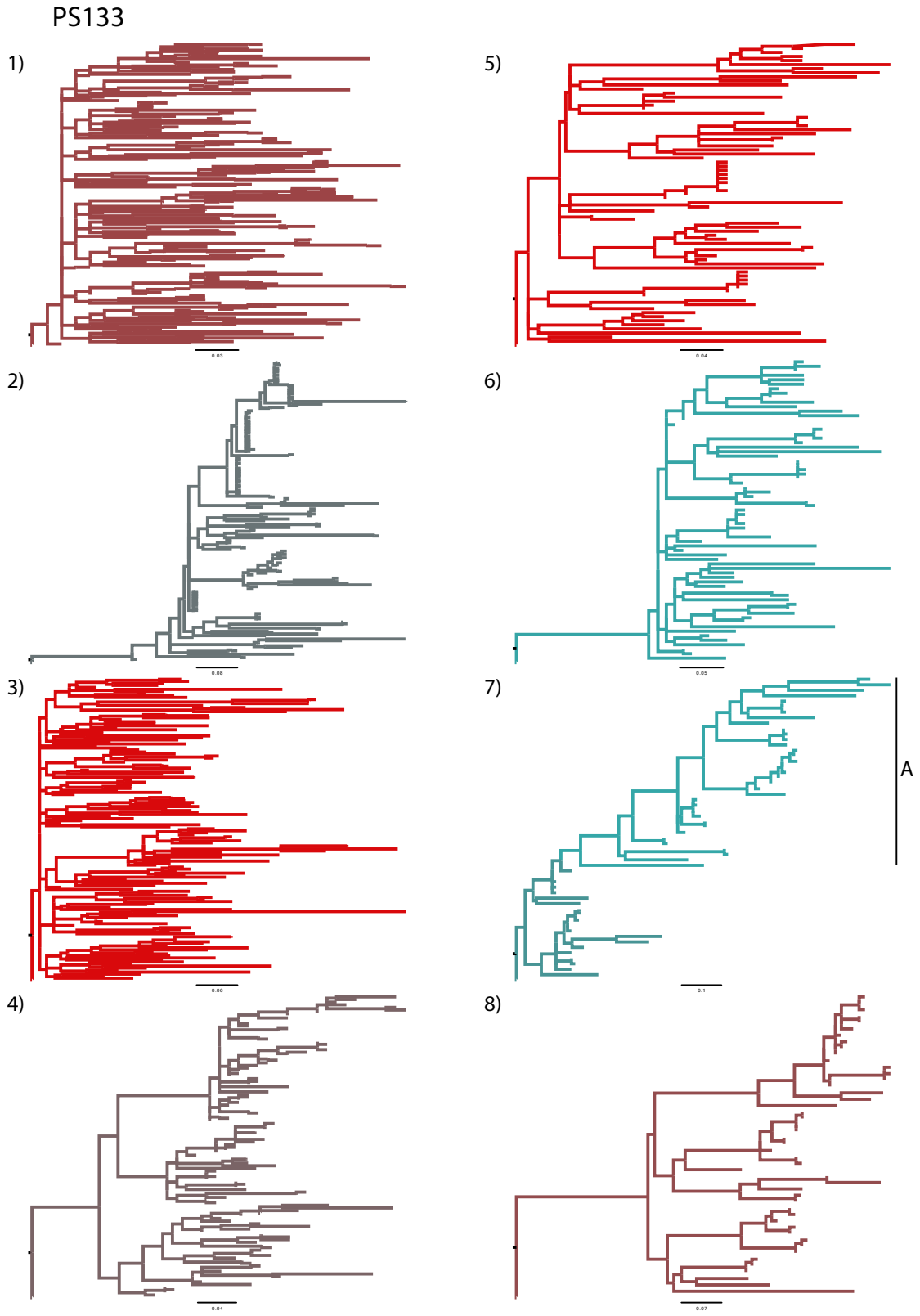


Figure 6.7: Eight largest lineages of PS133, coloured by ω_{CDR} estimate (blue=0, red=1.1)

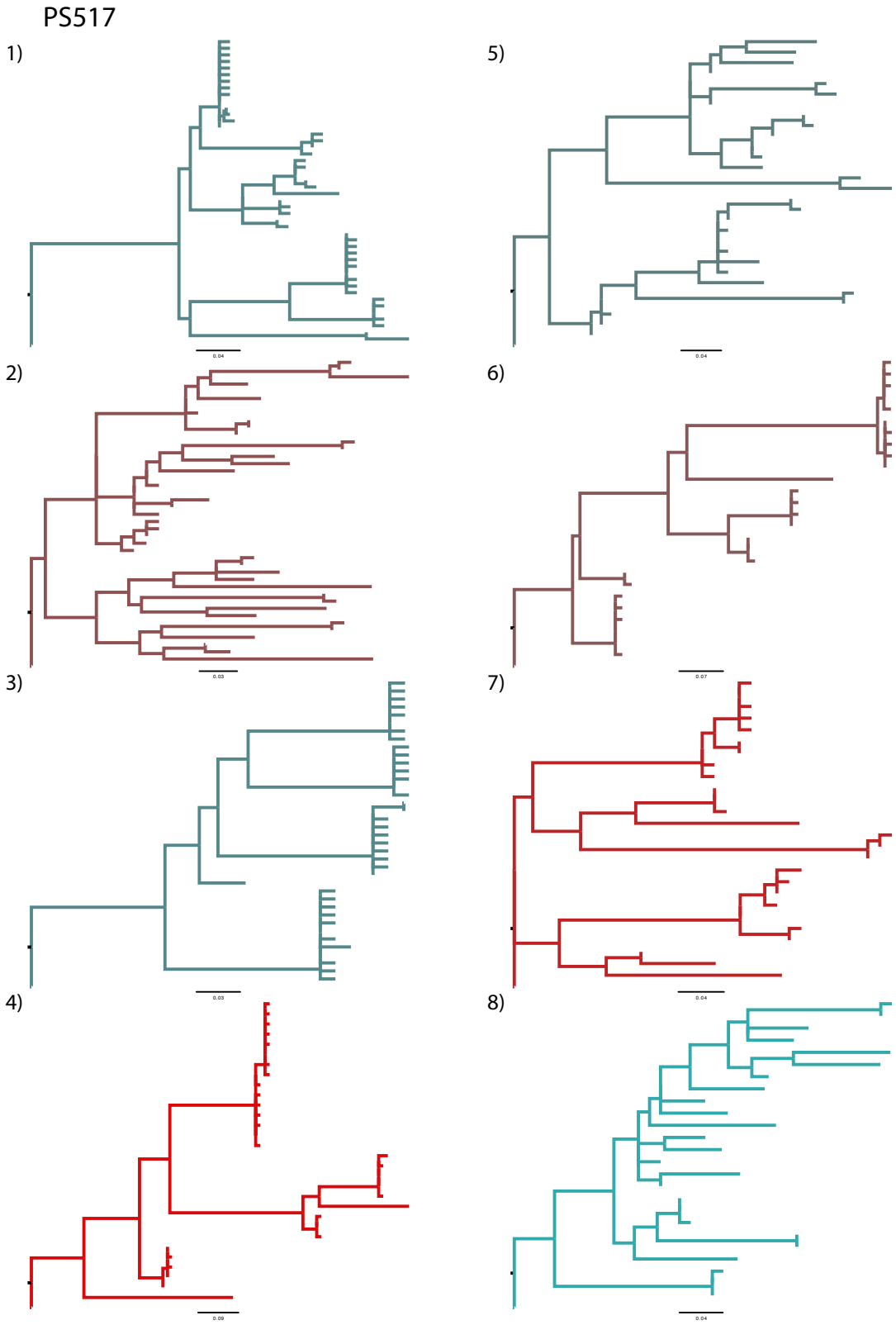


Figure 6.8: Eight largest lineages of PS517, coloured by ω_{CDR} estimate (blue=0, red=1.1)

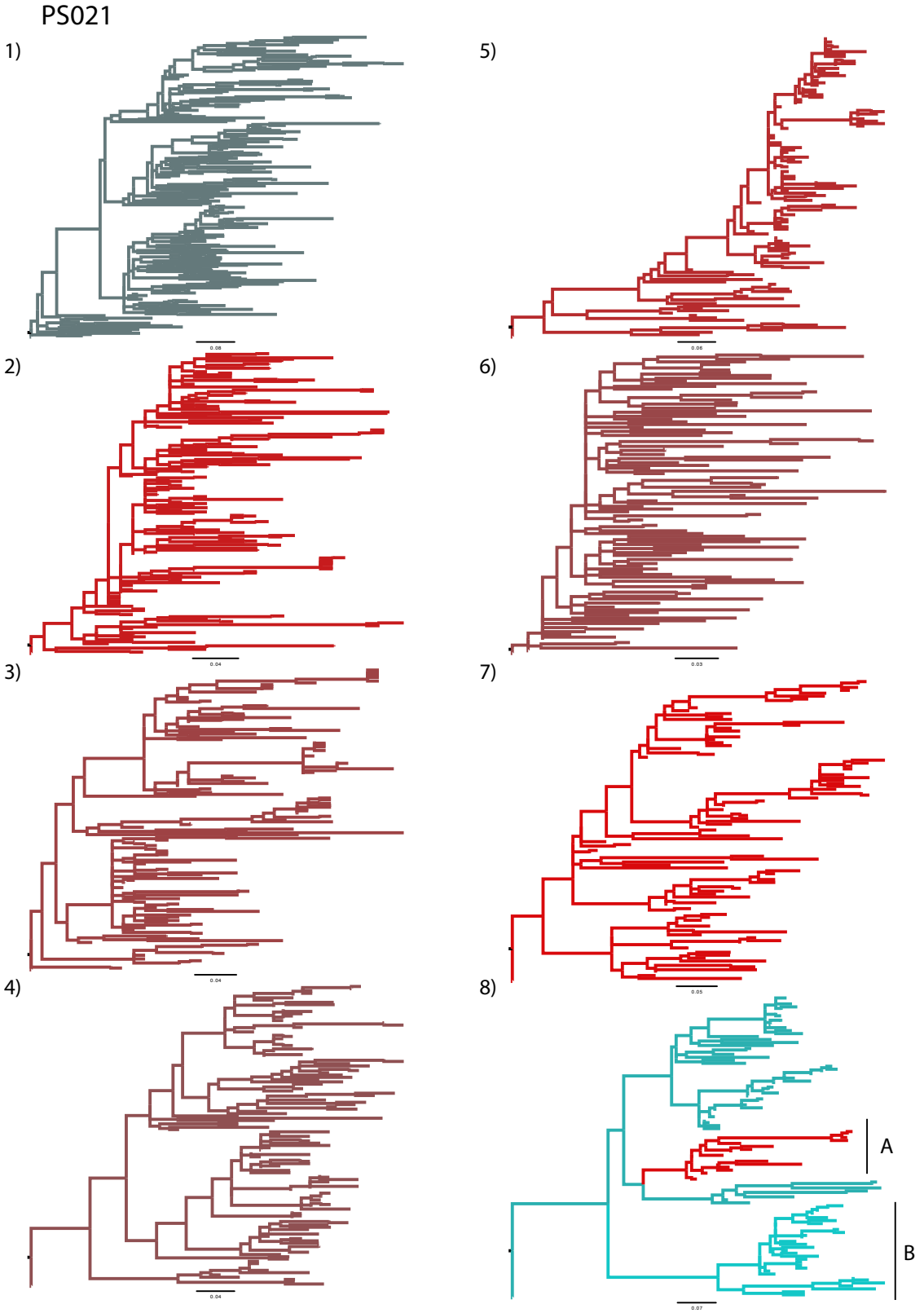


Figure 6.9: Eight largest lineages of PS021, coloured by ω_{CDR} estimate (blue=0, red=1.1)

6.4 Discussion

While the complexity of B cell repertoires as a whole makes biologically informed inferences difficult to make (e.g. Chapter 2), the repertoire is fundamentally a collection of antibody lineages, and breaking it into its constituent components should help to provide more biological insight into the processes that generate it. One approach would be to use the HLP17 model to undertake parameter estimation process for each lineage individually, as was done in the previous three chapters. One problem with this approach is that the HLP17 model is highly parametric compared to many other phylogenetic models. This does not present a significant issue for the HIV bnAb lineages analysed in Chapters 3, 4, and 5, because they are often genetically diverse enough to estimate its large number of parameters. However, as shown here, antibody lineages identified *de novo* from bulk repertoire sequencing data are generally less genetically diverse. This makes analyses more prone to over-fitting.

Here, I demonstrate how the full genetic diversity of the B cell repertoire may be exploited to estimate HLP17 parameters that correspond to antibody evolutionary processes that are expected to be consistent among lineages. This allows parameter estimation in lineages with smaller amounts of genetic diversity, because fewer (if any) parameters are estimated for each lineage individually. Interestingly, many parameters explored here (e.g. π) are consistent between patients, raising the possibility that in future analyses it may be unnecessary to estimate these parameters at all. Instead, parameters may be empirically estimated a under large dataset and fixed for future analyses, similar to many amino acid substitution models (e.g. the WAG model; [Whelan and Goldman, 2001]. This was not the case with all parameters estimated at the repertoire level, however, with estimates of m_h differing substantially between patients and models (Figure 6.4). It is not clear why these estimates would differ between patients, since targeting prefer-

ences of SHM are generally thought to be shared among humans. However, there is some evidence here that these differences may be due to interaction with ω. Namely, PS133 and PS021, but not PS517, showed substantial difference between m_h estimated under the $r:n:\omega$ and $r:1:\omega$ models. PS133 and PS021 also had a number of lineages showing a large improvement in likelihood under the $r:n:\omega$ model, while PS517 had comparatively few (Figure 6.6).

The analyses presented in this chapter serve as a proof-of-principle demonstration of how the HLP17 framework developed over the previous three chapters may be expanded to study not just individual lineages, but entire BCR repertoires. This is in contrast to Chapter 2, where I explored bulk BCR repertoire data using summary statistics of genetic diversity. Here, repertoires are characterized by the parameters of an evolutionary model. These parameters have biological interpretations, and exist within a flexible framework that allows naturally for hypothesis testing. Not only do these hierarchical models allow for more reliable estimation of substitution parameters in individual lineages within the repertoire, they provide a model-based characterization of the substitution process that generated the repertoire itself. While I primarily explore ω is explored here, this hierarchical framework could potentially be expanded to characterize repertoire-wide evolutionary processes such as class-switching and migration as well.

Chapter 7

Discussion

B cell receptors, released as antibodies, are the cornerstone of adaptive immunity, and next-generation sequencing has recently given researchers unprecedented potential to take apart the complexities of this system. However, BCR repertoire sequencing on its own is not informative - it must be analysed in a coherent manner, which is difficult given the complex, dynamic, and interacting processes acting on BCR sequence evolution. This thesis lays a model-based foundation for such a framework. Chapter 2 is an explorative, albeit cautionary tale to this effect. While I was able to show some relationship between the total genetic diversity of the B cell repertoire and aspects of early HIV infection, I did not find any significant relationship with other clinical variables. The next three chapters eschewed bulk repertoire data altogether, and instead focused exclusively on three bnAb lineages from HIV-infected patients. Chapter 3 introduced the HLP17 substitution model, which provided the backbone for the rest of the research in this thesis. This model, for the first time, incorporated some of the unique features of SHM into a substitution model. In addition, Chapter 3 introduced IgPhyML, a practical implementation of the HLP17 model, which builds upon the software package codonPhyML [Gil et al., 2013]. Chapter 4 expanded this framework by incorporating different types of mutability information, and other information about amino acid exchangeability. Chapter 5 further elaborated the HLP17 framework, enabling it to study the

dynamics of selection in early bnAb lineage development. Chapter 6 brings my thesis full circle by showing, in a short, proof-of-principle study, how the phylogenetic framework developed in Chapters 3, 4, and 5 can be scaled up to study entire B cell repertoires within a phylogenetic context.

7.1 HLP17 and phylogenetic substitution models

The HLP17 substitution model and its associated framework, developed in this thesis, synthesizes models and algorithms from different areas within phylogenetics. As with almost all likelihood-based phylogenetic models, this work is based on the work of Joe Felsenstein [Felsenstein, 1981], which introduced the pruning algorithm used for likelihood calculation, thereby making model-based phylogenetics computationally practical. Two important assumptions in Felsenstein [Felsenstein, 1981] - that evolution occurs independently among sites and that the substitution process is time reversibile - have remained in most phylogenetic models widely used today, and the fact that these assumptions are violated by features of the SHM process is what motivated the development of the HLP17 model.

Previous work in phylogenetics has relaxed the assumptions of independence among sites and time reversibility separately. The most direct and perhaps appropriate means of accounting for the effects of neighbouring sites is to create a more expansive substitution model, which describes rates of change between groups of nucleotides, for instance between all nucleotide dimers. Codon substitution models, such as the GY94 [Goldman and Yang, 1994] are a natural step up in terms of modelling dependence among certain nucleotides. However, creating larger and larger substitution matrices rapidly becomes computationally impractical - a single nucleotide matrix contains 16 elements, a codon model contains 3721, and a pentamer substitution matrix would contain 390,625 elements. For

this reason, codon models such as the NY98 model of [Nielsen and Yang, 1998] still assume independence among codon sites. The HLP17 model also assumes independence among codon sites, but expands on the NY98 model by using a mean-field approximation to take into account possible neighbouring sites of a target site. This approach was previously used by Whelan and Goldman [Whelan and Goldman, 2004] to account for the effects of multiple concurrent nucleotide substitution events.

Within maximum likelihood phylogenetics, time-reversibility is primarily used as a means of speeding up likelihood calculations during branch length estimation and topology searching. Under what Felsenstein [Felsenstein, 1981] dubbed the 'pulley principle,' it is possible to caluclate the same likelihood value using any point on any branch on the tree as the 'root' - the most ancestral point of the tree, which therefore defines the direction of evolution. This means that, rather than re-calculating the likelihood for the entire tree (Eq. 1.5) when optimizing a single branch length, it is possible to 're-root' the tree at that branch, and, using the partial likelihoods on either side, rapidly re-calculate likelihood under different branch lengths. However, the pulley principle relies on the relationship that, for all characters i and j, $\pi_i Q_{ij} = \pi_j Q_{ji}$; namely, it requires a reversible substitution model and that equilibrium frequencies exist at the root sequence. HLP17 breaks the pulley principle in two ways - it is non-reversible and uses a known root sequence. Hence, the pulley principle does not hold even under the reversible NY98:H0 model (Table 4.1). Implementing the HLP17 model thus required special likelihood memoization algorithms [Boussau and Gouy, 2006] in order to render branch length and topology estimation computationally practical. Further, the non-reversible nature of the HLP17 **Q** matrix necessitated use of a different means of matrix exponentiation for calculating the **P** matrix. There are a myriad of different means of calculating a matrix exponential, each with different limitations [Moler and Van Loan, 1978]. By far the most common method used in phyloge-

netics is eigenvalue decomposition, which has advantages in computational efficiency [Schranz et al., 2008], and is especially useful for reversible substitution models because their **Q** matrices have real eigenvalues, thereby enabling stable and efficient algorithms for eigenvalue calculation [Keilson, 1979, Yang, 2014]. This is not true for all matrices, and Schranz et al [Schranz et al., 2008] found that matrix exponentiation using eigenvalue decomposition is unstable when applied to certain non-reversible rate matrices they empirically derived from microbial DNA sequences. This study showed the Pade approximation with scaling and squaring was a far more robust means of exponentiating the nonreversable matrices encountered. This method was implemented within codonPhyML [Serrano Zanetti, 2013, Gil et al., 2013] and was not changed during my development of IgPhyML.

Thus, while many of the key pieces of HLP17 and IgPhyML existed before - codon substitution models [Goldman and Yang, 1994, Nielsen and Yang, 1998], mean-field approximation of flanking site effects [Whelan and Goldman, 2004], non-reversible optimization algorithms [Boussau and Gouy, 2006], and even most of the source code [Gil et al., 2013], these pieces had not been brought together into a single framework, nor, more importantly, had they been applied to more realistically model the evolution of antibody lineages.

7.2 The past, present, and near future of antibody phylogenetics

Large-scale computational analysis of antibody genetics is a fairly new field. Though the theory of clonal selection driving diversity of immune repertoire has existed for decades [Burnet, 1976], the molecular mechanisms behind its action have been only recently elucidated. AID, the protein that drives both V(D)J recombination

and SHM, was discovered only in 1999 [Muramatsu et al., 1999, Muramatsu et al., 2000]. It is important to note, however, that B cell genetics generally is far from a new field. Though the amount of sequence data available was lower, researchers have for decades been able to use BCR sequences to study diseases such as B cell cancers [Spencer and Dunn-Walters, 1999, van Dongen et al., 2003] and autoimmune conditions [Shlomchik et al., 1987, Colombo et al., 2000, Colombo et al., 2003]. However, studying large scale repertoire dynamics was only made possible in 2009 with the development and application of next-generation sequencing to BCR repertoires [Weinstein et al., 2009, Boyd et al., 2009].

In an interesting example of (academic) ontogeny recapitulating phylogeny, the field of antibody phylogenetics, including this thesis, has largely repeated the developmental history of phylogenetic methods itself. Early analyses of B cell lineages focused on creating 'lineage tree' using maximum parsimony [Shlomchik et al., 1987, Dunn-Walters et al., 2002, Barak et al., 2008]. These trees usually collapse internal nodes such that internal nodes may be multifurcating, that sampled sequences are allowed to be internal nodes, and that branches are in units of single nucleotide mutations [Barak et al., 2008]. These provided an easily interpretable visual representation of the development of an antibody lineage. Dynamics of affinity maturation, such as clonal selection, were approximated by measuring different characteristics of the tree shape, such as the number of outgoing branches from a single node [Dunn-Walters et al., 2002, Dunn-Walters et al., 2004, Shahaf et al., 2008]. The other predominant means of characterizing selection in B cell lineages was counting of the absolute number of observed amino acid changes (R/S ratio) between germline and data sequences [Chang and Casali, 1995]. These techniques were shown to be biased towards detecting positive selection [Dunn-Walters and Spencer, 1998], but later efforts corrected these using empirical models of SHM targeting [Hershberg et al., 2008]. This paradigm of counting replacement vs. silent mutations went through several iterations of increasing sophisti-

cation, eventually leading to the development of a Bayesian approach to quantify selection [Uduman et al., 2011], 2014; [Yaari et al., 2012].

Maximum parsimony techniques were also the preferred method of phylogenetic tree-building before the development of model-based phylogenetics [Felsenstein, 1978, Felsenstein, 1981]. A key motivation for using likelihood-based phylogenetics in tree estimation was identifying instances in which maximum parsimony gives misleading estimates of topology even under increasing amounts of data [Felsenstein, 1978]. Methods to characterize selection by counting nonsynonymous to synonymous codon changes (similar to R/S methods) also predominated in phylogenetics until the development of codon model-based approaches [Nielsen and Yang, 1998]. These methods generally operated by counting the number of synonymous and nonsynonymous sites between two sequences, counting the number of synonymous and nonsynonymous differences between the two sequences, and correcting for multiple substitutions [Nei and Gojobori, 1986, Nielsen and Yang, 1998]. These methods had similar disadvantages to the R/S methods currently in use in BCR genetics, in that they are non-phylogenetic in nature; namely they fundamentally operate on pairs of sequences rather than on modelling shared ancestry between multiple sequences, and must deal with codons with multiple substitutions in a fairly ad-hoc manner. Parameterizing dN/dS within a codon substitution model provides a clear, statistically rigorous means of modelling and testing this ratio [Yang and Swanson, 2002], and automatically accounts for correlation among sequences due to shared ancestry.

Both maximum likelihood and Bayesian model based phylogenetic approaches are fairly new to antibody phylogenetics, and seem to be largely connected to the study of HIV bnAb lineages [Liao et al., 2013, Doria-Rose et al., 2014]. One of the first new ML phylogenetic methods designed specifically for antibody lineages was that developed by Kepler [Kepler, 2013], which combined phylogenetic an-

cestral state reconstruction algorithms with known germline reference datasets to reconstruct the germline ancestor of a B cell lineage. Since then, more complex models have been used, including relaxed clock models, though these methods have not been adapted specifically for B cell evolution [Wu et al., 2015, Sheng et al., 2016].

The work of this thesis also fits within this narrative of repeating the history of phylogenetic methods. Chapters 3 and 4 aimed to bring antibody phylogenetics to approximately the same place that Nielsen and Yang [Nielsen and Yang, 1998] brought molecular evolution; namely, these chapters introduced functional codon and semi-empirical amino acid models capable of tree-estimation and able to test some biologically interesting hypotheses. The obvious next step, then, is to use the HLP17 framework as the basis for a rigorous and flexible suite of tests for clonal selection to replace the counting methods currently in wide use (i.e. the antibody genetics equivalent of [Yang and Swanson, 2002]). Chapter 5 moved significantly in that direction by estimating ω values for CDR/FWR partitions and for a set of tree branches. However, these models could be expanded even further by implementing, for instance, 'random sites' models that can identify specific sites in the BCR under diversifying selection [Yang and Swanson, 2002]. One challenge to future development is the fact that HLP17 is already highly parametric compared to other codon substitution models. However, Chapter 6 demonstrates that many of the substitution parameters in HLP17 are consistent among lineages from different individuals, suggesting that a semi-empirical HLP17 model, in which most parameters are fixed, may be able to solve this issue.

7.3 A broader look at the future

Analysis of bulk BCR repertoire data is an odd task - not uncommon in the era of high-throughput genomics - in which the existence of massive amounts of data

has preceded a coherent and structured analytical framework for understanding it. The biological processes shaping the genetic diversity of B cell repertoires are numerous, and, as shown by the analyses in Chapter 2, it is difficult to make firm conclusions from statistical analysis without additional biological information about particular lineages. Because of this, I believe there will be two primary drivers of advancement in this field: (i) finding new ways to associate biological information with individual sequences or lineages, and (ii), developing more rigorous models to use the combined information to test biological hypotheses. One reason that bnAb lineages have spurred so much development in the field (e.g. Chapters 3-5) is because they are confirmed to have useful neutralization properties, and their associated monoclonal antibodies may be used for further study. Without associated mAb sequences, it would be difficult to draw any firm conclusions about the development of neutralization within an antibody lineage, and the lineage itself would be simply be one of thousands present in the repertoire, without clear biological utility. Developing mAbs for all lineages in bulk RepSeq data would be impractical; however other types of data, such as Ig class and tissue location, are easily associated with individual sequences [Stern et al., 2014, Heiden et al., 2017]. Incorporating this information will require clearly defined models of discrete trait evolution and migration, both of which have been developed for phylogenetic methods in other applications [Lemey et al., 2009].

From a broader view, the HLP17 model, combined with the efficient non-reversible optimization algorithms I have implemented in IgPhyML is a highly generalizable tool, far less constrained than classical (i.e. time-reversible) approaches with respect to the range of substitution phenomena it is capable of modelling. Because of this, it may find broad utility in a diverse range of other applications that experience non-reversible evolution and/or motif-targeted hypermutation, such as characterizing APOBEC-mediated hypermutation operating on HIV sequences [Holtz et al., 2013] and even lineages of cancer cells [Alexandrov et al., 2013]. HIV

proviruses within T cells are often rendered nonfunctional by the cytodine deaminase protein APOBEC3G, which is evolutionarily related to AID and helps to control retroviral insertion by hypermutating integrated proviruses [McMichael et al., 2010]. APOBEC3G has its own hotspot motifs (e.g. GGG; [Monajemi et al., 2014]) and the HLP17 model may be used to study hypermutation in a phylogenetic context. In addition to branch-specific ω values, IgPhyML can also estimate branch-specific h^a (Eq 3.1) values, potentially allowing the detection of clades of HIV that have been exposed to APOBEC3G among other lineages which have not. Further, many types of cancer, particularly those resulting from high levels of environmentally-triggered mutation (such as skin and lung cancer) experience context-mediated hypermutation as well [Pfeifer et al., 2005, Alexandrov et al., 2013]. These biases, either in the form of specific motifs or empirical mutability models similar to the S5F [Yaari et al., 2013] could be easily incorporated into this framework using Eqs 4.3 and 4.4.

7.4 Conclusion

From an evolutionary biologist's viewpoint, the recent availability of high-throughput B cell repertoire data presents a rare opportunity to develop models and methods for a new kind of genetics. Much of the groundwork for this new field was laid decades ago with the development of model-based phylogenetics, and it is tempting to simply use these techniques as they currently stand. However, the evolutionary process of affinity maturation violates fundamental model assumptions that date back to the inception of molecular phylogenetics [Felsenstein, 1981], and are built into the foundations of most of modern phylogenetic software packages. Creating a more complete framework that relaxes these assumptions means returning to these methodological foundations and working back up. Fortunately, many problems encountered along the way have been addressed before in different contexts (e.g. [Boussau and Gouy, 2006]). From here, the future will likely be

in developing means of using this new framework to test biological hypotheses, which will rely both on new types of data becoming available for B cell sequences and lineages, and finding ways to incorporate that data into these models.

Appendices

Appendix A

Supplemental information for Chapter 2

A.1 Human B-cell receptor PCR primers

Primer	Sequence	
JH reverse	CTTACCTGAGGAGACGGTGACC	
VH1-FR1 forward	GGCCTCAGTGAAGGTCTCCTGCAA G	
VH2-FR1 forward	GTCTGGTCCTACGCTGGTGAAACCC	
VH3-FR1 forward	CTGGGGGGTCCCTGAGACTCTCCTG	FR1 primer set*
VH4-FR1 forward	CTTCGGAGACCCTGTCCCTCACCTG	
VH5-FR1 forward	CGGGGAGTCTCTGAACATCTCCTGT	
VH6-FR1 forward	TCGCAGACCCTCTCACTCACCTGTG	

Figure A.1: Human B-cell receptor PCR primers used in Chapter 2

A.2 Illustration of the BCR sequencing pipeline

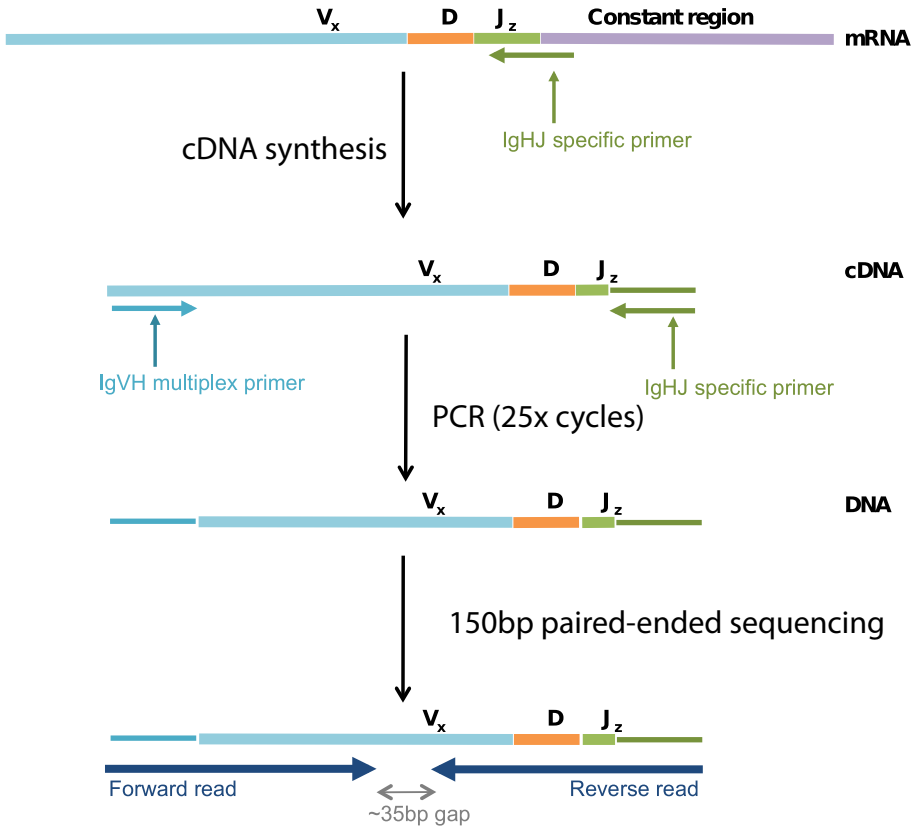


Figure A.2: (summarized from main text) MiSeq reads were filtered for base quality (median >32) using QUASR (http://sourceforge.net/projects/quasr) [Watson et al., 2013]. Sequences were concatenated and a gap inserted between the forward and reverse reads (average gap length ˜35 nucleotides. Non-immunoglobulin sequences were removed; only those reads with significant similarity to reference IgHV and IgHJ genes from the ImMunoGeneTics (IMGT) database [Lefranc et al., 2009] were retained, as determined using BLAST [Altschul et al., 1990] with E-value thresholds of $1x10^{-10}$ and $1x10^{-3}$, respectively. Primer sequences were then trimmed from the reads, and sequences retained for analysis only if both primer sequences were identified and if both forward and reverse sequence lengths were >100 bp. Lastly, a second BLAST analysis was used to identify the IgHV genes of retained reads. Specifically, BLAST [Altschul et al., 1990] was used to align reads against known BCR sequences from the IMGT database [Lefranc et al., 2009] (E-value thresholds for IgHV and IgHJ-genes were 10^{-70} and 10^{-20} respectively, due to their different gene lengths). The combined per-base error rate for the RT-PCR and sequencing process of the MiSeq platform was $2.06x10^{-4}$ [Bashford-Rogers et al., 2013, Bashford-Rogers et al., 2014].

A.3 Illustration of the sequence alignment pipeline

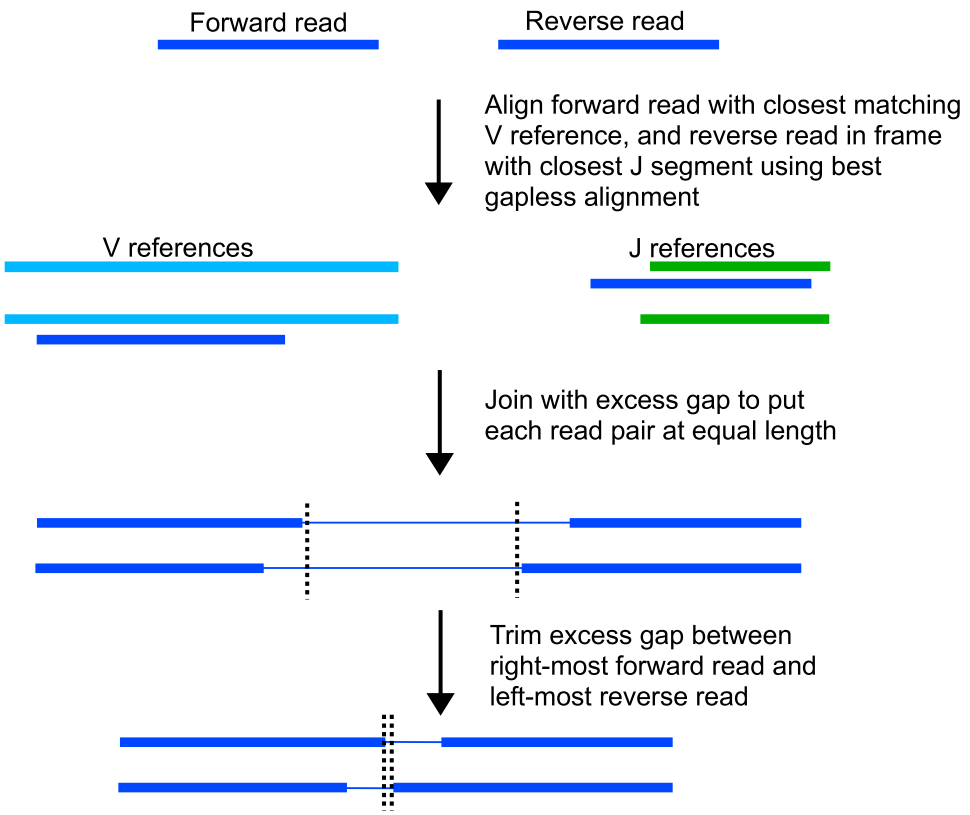


Figure A.3: (summarized from main text) The first read of each pair was aligned with the closest matching sequence in the V references through gapless pairwise alignment. To further optimize the alignment each read was matched using only the 5′ half of the V segments. Next, the second read of each pair was pairwise aligned with the closest sequence in the J references, again using gapless alignment. Once the first and second reads had been aligned to the V and J references, respectively, gaps were added 3′ of the first read, and 5′ of the second read, such that each were 300 nucleotides in length. The two reads were then concatenated, and excess gaps were trimmed between each read pair. Because we expect sequences from the same clone to match the same, or highly similar, V and J segments, this process generates an approximation of the optimal gapless alignment for sequences from the same clone.

A.4 Relationship between Gini index and read depth between patients

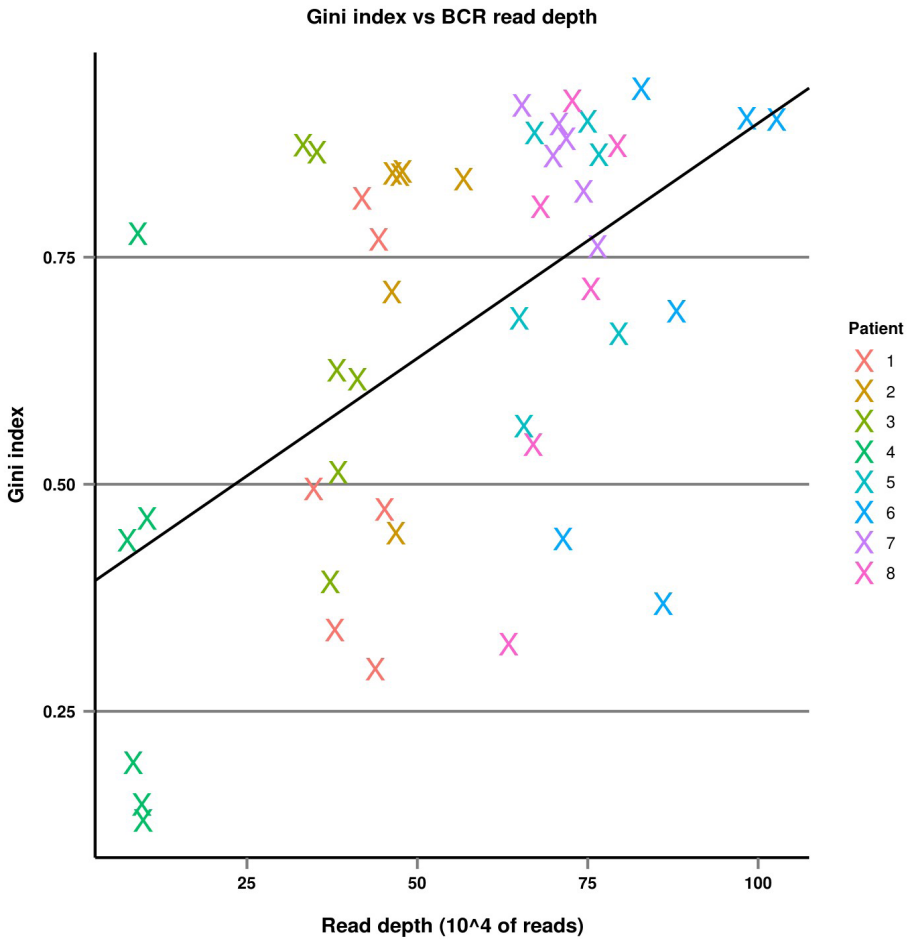


Figure A.4: Each point represents a sampling time point and is coloured by patient (see key). The fitted linear regression has a slope of $5.2x10^{-3}$ (p = $5.69x10^{-5}$) and a correlation coefficient of 0.55. This relationship remained significant when patient means were used to correct for temporal autocorrelation ($p < 0.03$).

A.5 Gini index and read depth

Depth (total reads)	Largest clone (reads)	No. of Singletons	Gini index
15	5	10	0.24
30	10	20	0.29
45	15	30	0.3
60	20	40	0.31
75	25	50	0.31

Figure A.5: I model the effect of read depth on Gini index using a simple population of B cells consisting of a single dominant clone occupying one third of all B cells and a large pool of unique B cells, which comprise the rest of the system. As sequencing depth increases, one third of each sample joins the dominant clone, which grows larger. However, all other clones remain unique, leading to a larger population of singleton clones. This means the ratio of reads in the dominant clone to each of the singletons grows with sampling depth, leading to a more skewed size distribution and higher Gini index. Specific values are provided in the table above.

A.6 CD4+ cell counts and seroconversion information

Patient	Estimated days since seroconversion at week 0	Week 4	Week 16	Week 24	Week 52	Week 60	Week 108
1	95	470	NA	440	480	50	470
2	34	NA	430	460	470	410	NA
3	12	550	690	630	530	440	NA
4	76	1610	1230	1180	1200	870	1460
5	93	830	620	850	500	650	510
6	68	490	550	480	600	740	500
7	35	590	570	660	470	360	NA
8	44	770	390	NA	660	320	NA

Figure A.6: CD4+ cell counts (cell/μL) for each patient and available timepoint.

A.7 Effect of sub-sampling on BCR statistics

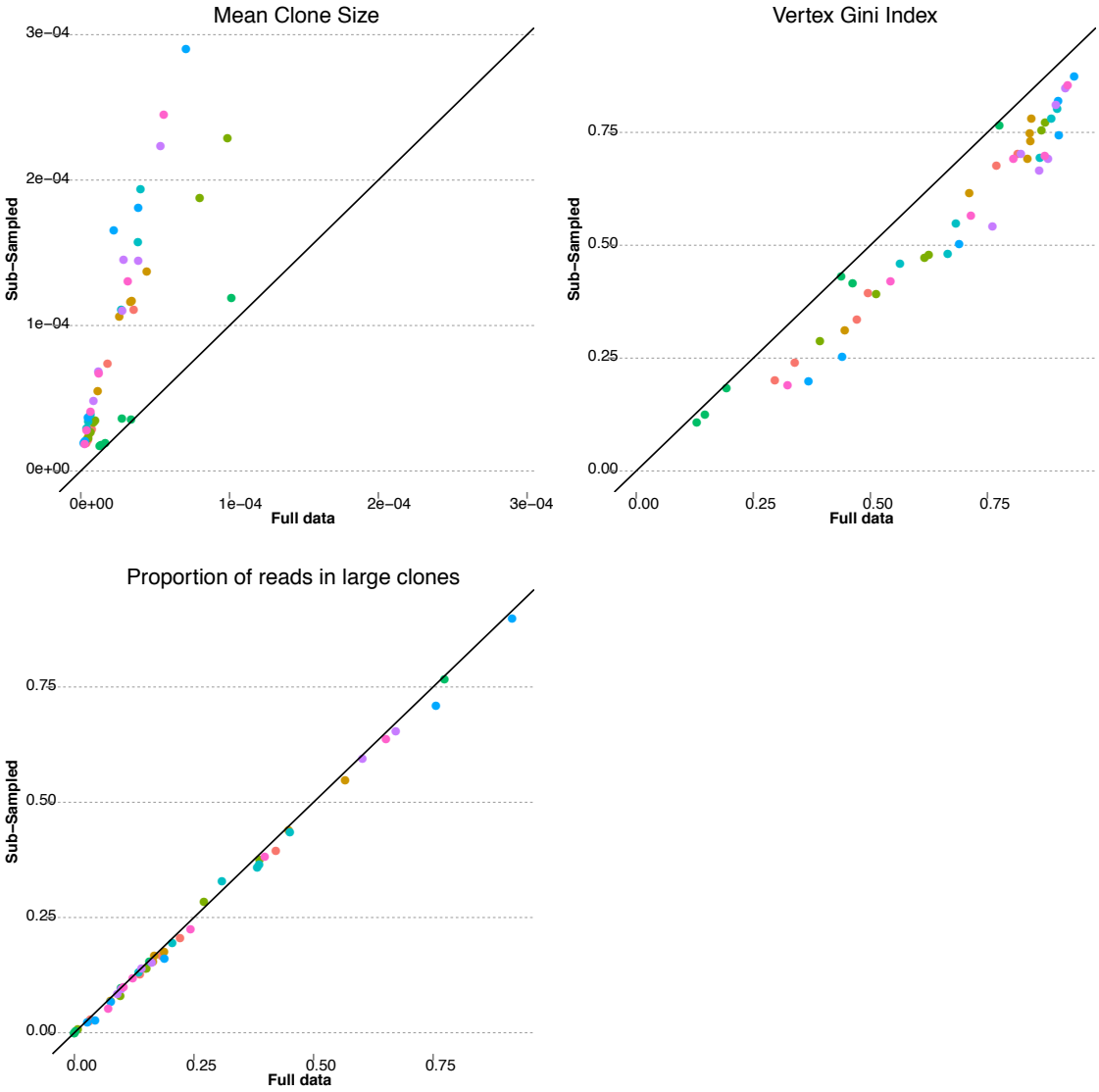


Figure A.7: BCR statistics of HIV+ patients calculated from full data versus data sub-sampled data. Read depth of full data had a median of 567,936 and a range of 74,861 to 1,040,936 reads per patient per time point. Sub-sampled data were randomly sub- sampled to a depth of 70,000 reads per patient per time point. Dots are coloured by patient.

Appendix B

Supplemental information for Chapter 3

B.1 Likelihood computation and ancestral state reconstruction

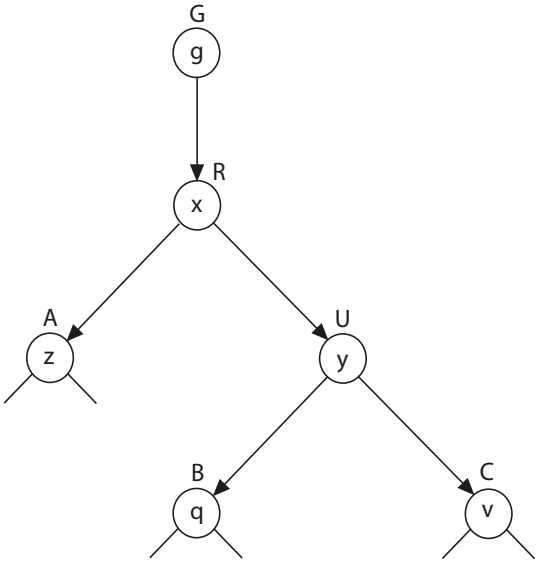


Figure B.1: Evolutionary model reproduced from 3.1

This section details an algorithm for efficient likelihood computation and marginal ancestral sequence reconstruction with a non-reversible model on sequences with

a known common ancestor. See Figure B.1. All sequences descend from given germline node G, which has sequence g. Note that this known ancestor G is not necessarily the most recent common ancestor of the lineage, which is node R and sequence x. See similarities to tree and terminology in [Boussau and Gouy, 2006].

P_{gx} = Transition probability from g to x along branch GR.

$$L_{low}(UC, v) = \begin{cases} 1 & \text{v = character at tip} \\ 0 & \text{otherwise} \end{cases} \tag{B.1}$$

$$L_{low}(RU, y) = \sum_{q} P_{yq} L_{low}(UB, q) \sum_{v} P_{yv} L_{low}(UC, v) \tag{B.2}$$

$$L_{upp}(RU, x) = P_{gx} \sum_{z} P_{xz} L_{low}(RA, z) \tag{B.3}$$

$$L_{upp}(UB, y) = \sum_{x} P_{xy} L_{upp}(RU, x) \sum_{v} P_{yv} L_{low}(UC, v) \tag{B.4}$$

B.1.1 Calculating likelihood at each node

This is effectively the same proof from (Boussau and Gouy 2006) but with P_{gx} in place of equilibrium frequencies.

Full likelihood P(data, given germline character g):

$$L_s = P(X|G = g) = \sum_x P_{gx} \sum_z P_{xz} L_{low}(RA, z) \sum_y P_{xy} \sum_q P_{yq} L_{low}(UB, q) \sum_v P_{yv} L_{low}(UC, v) \tag{B.5}$$

Likelihood from a germline edge:

$$\begin{aligned} L_{s,GR} = P(X|G = g) &= \sum_x P_{gx} L_{low}(GR, x) \\ &= \sum_x P_{gx} \sum_z P_{xz} L_{low}(RA, z) \sum_y P_{xy} L_{low}(RU, v) \\ &= \sum_x P_{gx} \sum_z P_{xz} L_{low}(RA, z) \sum_y P_{xy} \sum_q P_{yq} L_{low}(UB, q) \sum_v P_{yv} L_{low}(UC, v) \\ &= L_s \end{aligned} \tag{B.6}$$

Likelihood from a root edge:

$$\begin{aligned} L_{s,RU} = P(X|G = g) &= \sum_x L_{upp}(RU, x) \sum_y P_{xy} L_{low}(RU, y) \\ &= \sum_x P_{gx} \sum_z P_{xz} L_{low}(RA, z) \sum_y P_{xy} \sum_q P_{yq} L_{low}(UB, q) \sum_v P_{yv} L_{low}(UC, v) \\ &= L_s \end{aligned} \tag{B.7}$$

Likelihood from a non-root edge:

$$
\begin{aligned}
L_{s,UB} = P(X|G=g) &= \sum_y L_{upp}(UB,y) \sum_q P_{yq} L_{low}(UB,q) \\
&= \sum_y \sum_x P_{xy} L_{upp}(RU,x) \sum_v P_{yv} L_{low}(UC,v) \sum_q P_{yq} L_{low}(UB,q) \\
&= \sum_x L_{upp}(RU,x) \sum_y P_{xy} \sum_v P_{yv} L_{low}(UC,v) \sum_q P_{yq} L_{low}(UB,q) \\
&= \sum_x L_{upp}(RU,x) \sum_y P_{xy} L_{low}(RU,y) \\
&= L_{s,RU} = L_s
\end{aligned}
\tag{B.8}
$$

B.1.2 Marginal ancestral state reconstruction

The essential operation of marginal ancestral state reconstruction, is, for a given node N, find the character n which maximizes $P(X|G=g, N=n)$.

Reconstruction at node X:

$$
P(X|G=g) = \sum_x P_{gx} \sum_z P_{xz} L_{low}(RA,z) \sum_y P_{xy} \sum_q P_{yq} L_{low}(UB,q) \sum_v P_{yv} L_{low}(UC,v)
\tag{B.9}
$$

$$
P(X|G=g, X=x) = P_{gx} \sum_z P_{xz} L_{low}(RA,z) \sum_y P_{xy} \sum_q P_{yq} L_{low}(UB,q) \sum_v P_{yv} L_{low}(UC,v)
\tag{B.10}
$$

Then beginning at branch GR:

$$\begin{aligned} P(X|G=g,X=x)_{GR} &= P_{gx}L_{low}(GR,x) \\ &= P_{gx}\sum_{z}P_{xz}L_{low}(RA,z)\sum_{y}P_{xy}L_{low}(RU,v) \\ &= P_{gx}\sum_{z}P_{xz}L_{low}(RA,z)\sum_{y}P_{xy}\sum_{q}P_{yq}L_{low}(UB,q)\sum_{v}P_{yv}L_{low}(UC,v) \\ &= P(X|G=g,X=x) \end{aligned} \tag{B.11}$$

Reconstruction at node B:

$$P(X|G=g,B=q) = \sum_{x}P_{gx}\sum_{z}P_{xz}L_{low}(RA,z)\sum_{y}P_{xy}P_{yq}L_{low}(UB,q)\sum_{v}P_{yv}L_{low}(UC,v) \tag{B.12}$$

Then beginning at branch UB:

$$\begin{aligned} P(X|G=g,B=q)_{UB} &= \sum_{y}L_{upp}(UB,y)P_{yq}L_{low}(UB,q) \\ &= \sum_{y}\sum_{x}P_{xy}L_{upp}(RU,x)\sum_{v}P_{yv}L_{low}(UC,v)P_{yq}L_{low}(UB,q) \\ &= \sum_{y}\sum_{x}P_{xy}P_{gx}\sum_{z}P_{xz}L_{low}(RA,z)\sum_{v}P_{yv}L_{low}(UC,v)P_{yq}L_{low}(UB,q) \\ &= \sum_{x}P_{gx}\sum_{z}P_{xz}L_{low}(RA,z)\sum_{y}P_{xy}P_{yq}L_{low}(UB,q)\sum_{v}P_{yv}L_{low}(UC,v) \\ &= P(X|G=g,B=q) \end{aligned} \tag{B.13}$$

B.2 Summary statistics of bNAb sequence data

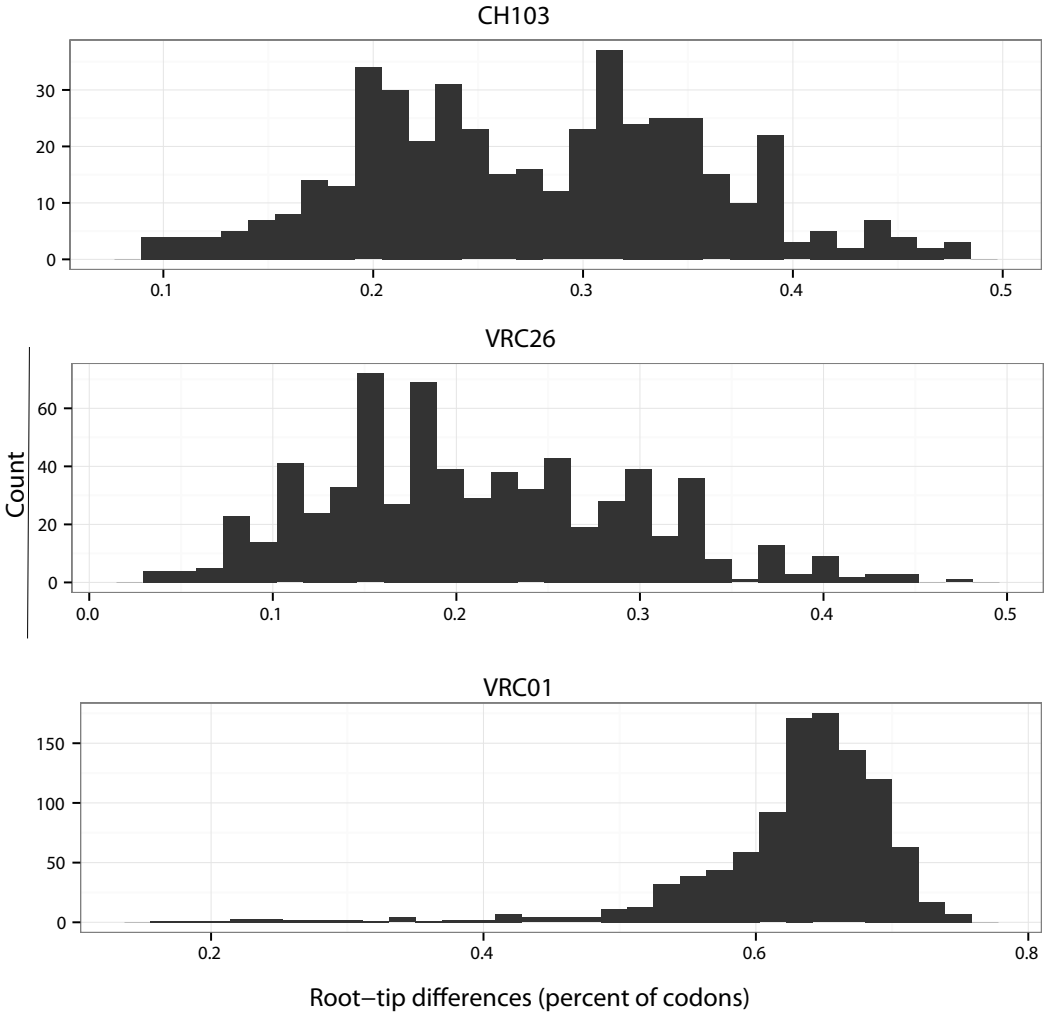


Figure B.2: Distribution of pairwise codon differences from each tip sequences to the germline sequence of each lineage.

B.3 Parameter estimates from different motif models

Set	Hotspot model	κ	ω	h^{WRC}	h^{GYW}	h^{WA}	h^{TW}	h^{SYC}	h^{GRS}	Log L.hood	AIC
CH103	Symm. WRC/GYW*	1.91	0.43	1.86	1.86	0	0	0	0	-14600.4	29230.8
	Asymm. WRC/GYW	1.98	0.42	1.19	3.50	0	0	0	0	-14569.7	29171.4
	Symm. WA/TW*	1.89	0.41	0	0	0.64	0.64	0	0	-14667.7	29365.4
	Asymm. WA/TW	1.84	0.37	0	0	1.54	-0.18	0	0	-14605	29242
	Symm. SYC/GRS*	1.88	0.41	0	0	0	0	-0.57	-0.57	-14658.3	29346.6
	Asymm. SYC/GRS	1.88	0.41	0	0	0	0	-0.59	-0.54	-14658.1	29348.2
	Uniform hotspots*	1.87	0.39	2.31	2.31	2.31	2.31	0	0	-14513	29056
	Hierarchical hotspots	1.88	0.40	3.87	3.87	1.96	1.96	0	0	-14482.2	28996.4
	SCAH*	1.90	0.34	2.31	5.84	3.23	0.59	-0.21	-0.21	-14388.8	28815.6
	FCH	1.90	0.35	2.30	5.86	3.22	0.59	-0.19	-0.24	-14388.7	28817.4
VRC26	Symm. WRC/GYW*	1.87	0.56	1.81	1.81	0	0	0	0	-37238.5	74507
	Asymm. WRC/GYW	1.87	0.56	2.16	1.48	0	0	0	0	-37230.5	74493
	Symm. WA/TW*	1.89	0.54	0	0	0.74	0.74	0	0	-37386.3	74802.6
	Asymm. WA/TW	1.88	0.52	0	0	1.13	0.29	0	0	-37344.4	74720.8
	Symm. SYC/GRS*	1.88	0.56	0	0	0	0	-0.62	-0.62	-37385.4	74800.8
	Asymm. SYC/GRS	1.89	0.57	0	0	0	0	-0.38	-0.78	-37358.4	74748.8
	Uniform hotspots*	1.85	0.50	2.55	2.55	2.55	2.55	0	0	-36915.8	73861.6
	Hierarchical hotspots	1.85	0.51	4.09	4.09	2.18	2.18	0	0	-36837.4	73706.8
	SCAH*	1.83	0.49	4.37	3.34	2.62	1.33	-0.14	-0.14	-36794	73626
	FCH	1.83	0.50	4.19	3.39	2.56	1.29	0.15	-0.39	-36782.2	73604.4
VRC01	Symm. WRC/GYW*	2.22	0.33	2.03	2.03	0	0	0	0	-43647.1	87324.2
	Asymm. WRC/GYW	2.23	0.33	1.80	2.31	0	0	0	0	-43643.2	87318.4
	Symm. WA/TW*	2.22	0.33	0	0	0.38	0.38	0	0	-43899.2	87828.4
	Asymm. WA/TW	2.21	0.32	0	0	0.80	-0.09	0	0	-43823.8	87679.6
	Symm. SYC/GRS*	2.19	0.33	0	0	0	0	-0.69	-0.69	-43742.5	87515
	Asymm. SYC/GRS	2.19	0.33	0	0	0	0	-0.74	-0.61	-43738.4	87508.8
	Uniform hotspots*	2.21	0.33	1.49	1.49	1.49	1.49	0	0	-43596.8	87223.6
	Hierarchical hotspots	2.21	0.33	3.51	3.51	1.19	1.19	0	0	-43436.9	86905.8
	SCAH*	2.19	0.32	2.42	3.01	1.32	0.30	-0.48	-0.48	-43327.1	86692.2
	FCH	2.18	0.31	2.50	2.96	1.35	0.32	-0.57	-0.34	-43321.6	86683.2

Table B.1: ML parameters, log-likelihood, and AIC values for each model fit on each bnAb lineage. See Table 3.3 for model descriptions. SCAH = Symmetric coldspots, asymmetric hotspots. FCH = Free coldspots and hotspots.

B.4 Analysis with random motifs

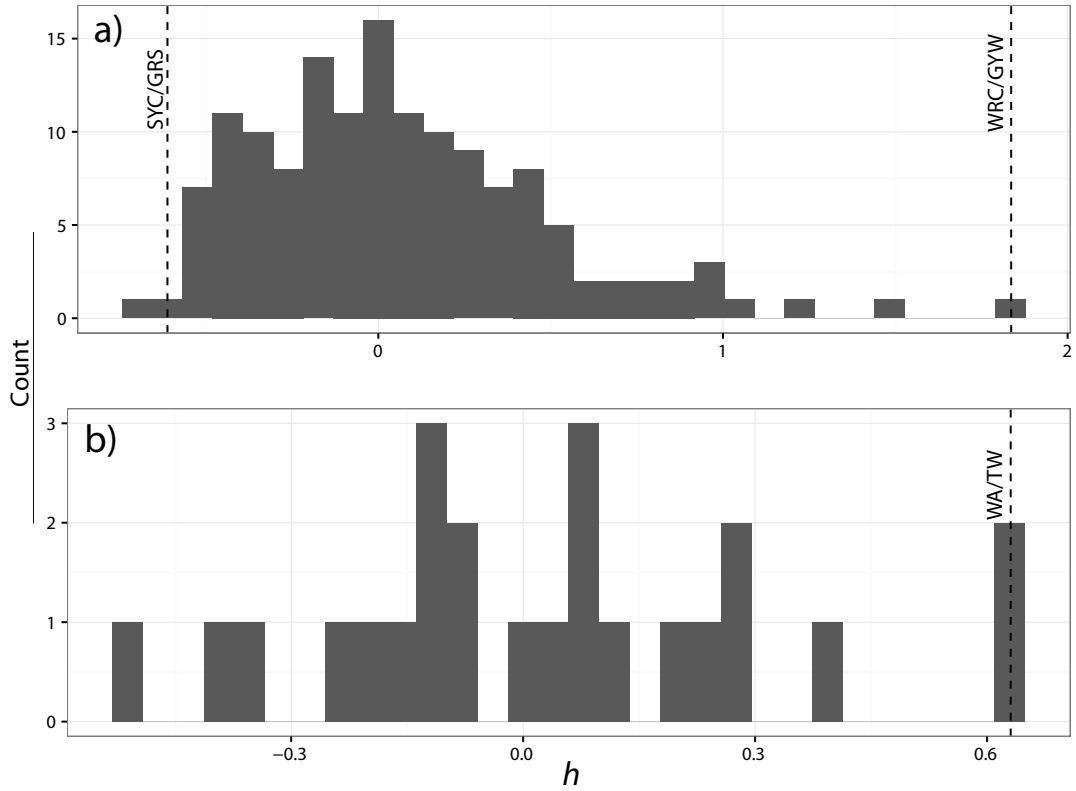


Figure B.3: Distribution of h values for all possible motifs fulfilling the following criteria: a) trimers which contained two IUPAC letters standing for two possible nucleotides (R, Y, S, W, K, and M), which were followed by an unambiguous nucleotide (A, C, G, or T), and b) dimer which contained one IUPAC letters standing for two possible nucleotides, which was followed by an unambiguous nucleotide. We then fit symmetric HLP17 models using each of these motifs and their reverse complements, with all other h parameters set to zero.

B.5 Parameter results from fully context dependent simulation

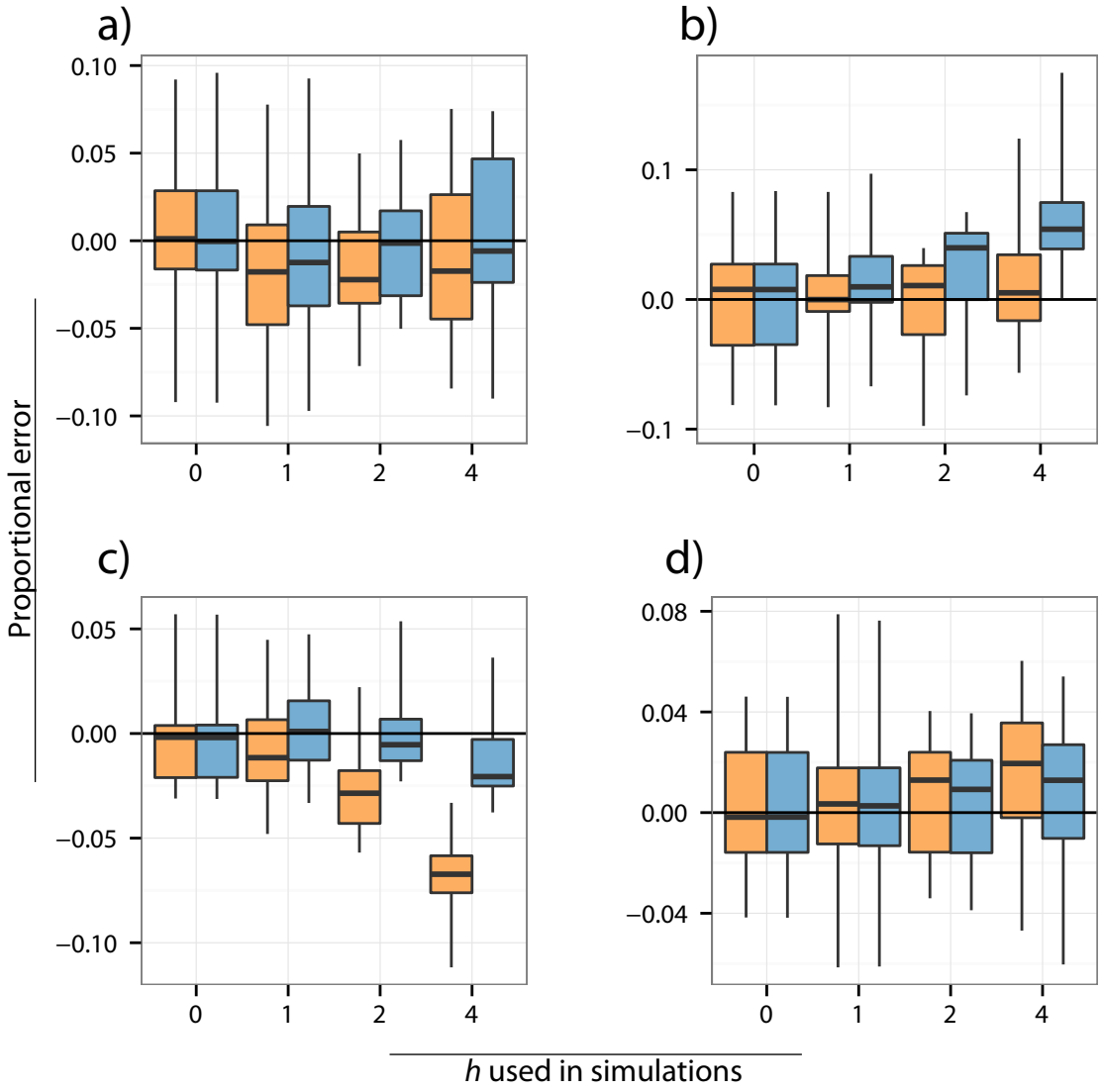


Figure B.4: Proportional error in parameter estimation compared to true values for the CH103 B cell lineage fully context dependent simulations. Values of ω, κ, tree length, and ratio of internal to external branch lengths are shown in panels a), b), c), and d), respectively. Estimates obtained under the NY98 are in orange (h=0) and estimates obtained under the HLP17 model are in blue (h estimated using maximum likelihood). The edges and centres of boxplots show the 1st, 2nd, and 3rd quartiles, while the whiskers show range.

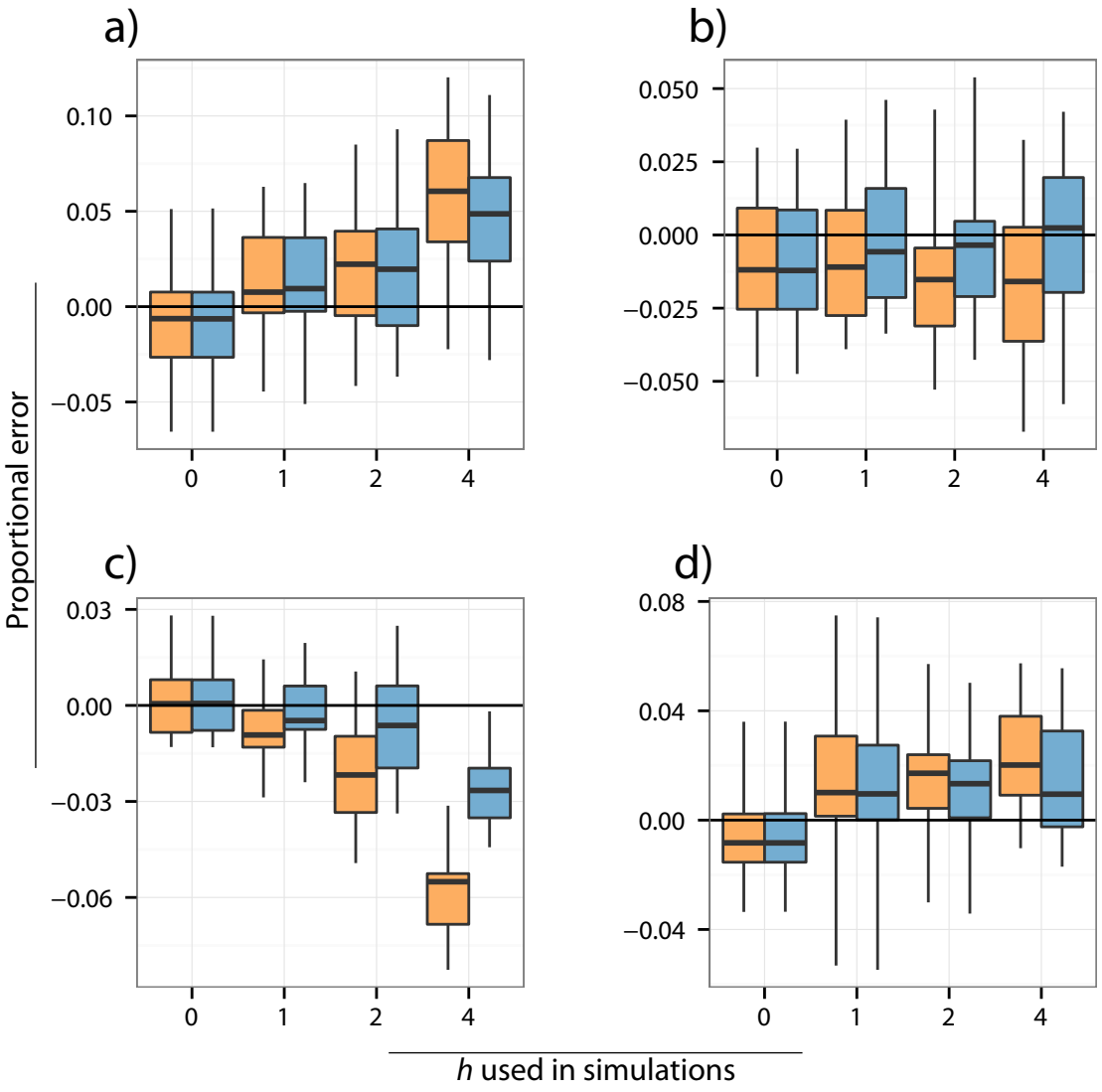


Figure B.5: Proportional error in parameter estimation compared to true values for the VRC26 B cell lineage fully context dependent simulations. Values of ω, κ, tree length, and ratio of internal to external branch lengths are shown in panels a), b), c), and d), respectively. Estimates obtained under the NY98 are in orange (h=0) and estimates obtained under the HLP17 model are in blue (h estimated using maximum likelihood). The edges and centres of boxplots show the 1st, 2nd, and 3rd quartiles, while the whiskers show range.

B.6 Analysis of interaction between h and ω

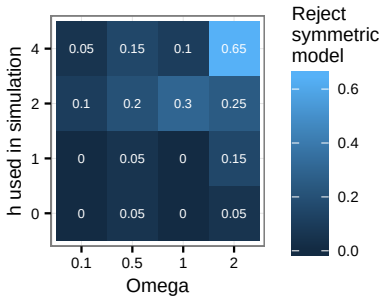


Figure B.6: Proportion of fully context-dependent simulated data sets that incorrectly reject the true symmetric WRC/GYW motif model. Note that this generally increases as both h and ω increase.

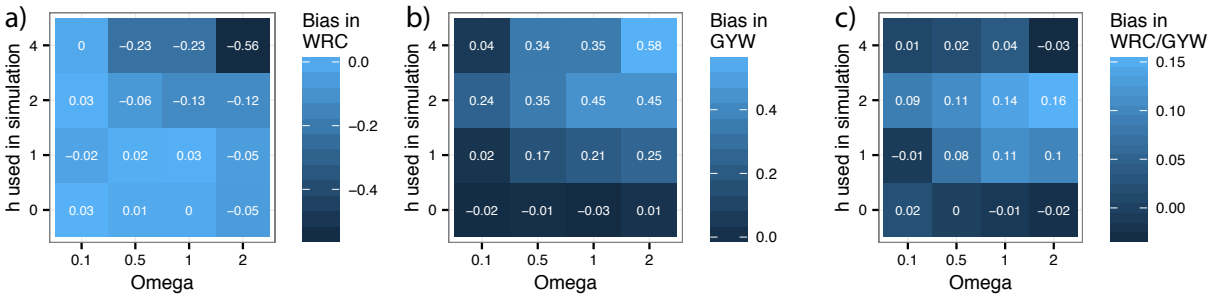


Figure B.7: Bias in h from data sets simulated using different values of h and ω under a fully context-dependent symmetric WRC/GYW model. Note that in the asymmetric model h^{WRC} and h^{GYW} are biased in opposite directions as h and ω increase, while in the symmetric model h^{WRC} shows no such consistent bias. Note that these values are absolute bias, not proportional error. This would be obtained by dividing each value in each cell by the true h used in simulation (left axis).
a) Bias in the value of h^{WRC} estimated from fitting an asymmetric WRC/GYW model.
b) Bias in the value of h^{GYW} estimated from fitting an asymmetric WRC/GYW model.
c) Bias in the value of $h^{WRC/GYW}$ estimated from fitting a symmetric WRC/GYW model.

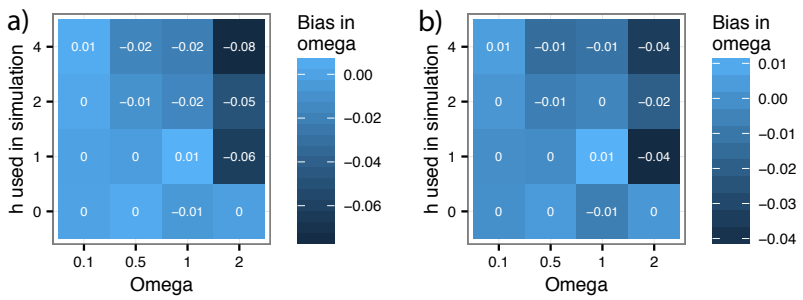


Figure B.8: Bias in ω from data sets simulated using different values of h and ω under a fully context-dependent symmetric WRC/GYW model. Note that, while a trend exists in the asymmetric model, the bias is ultimately very small.
a) Bias in the value of ω estimated from fitting an asymmetric WRC/GYW model.
b) Bias in the value of ω estimated from fitting a symmetric WRC/GYW model.

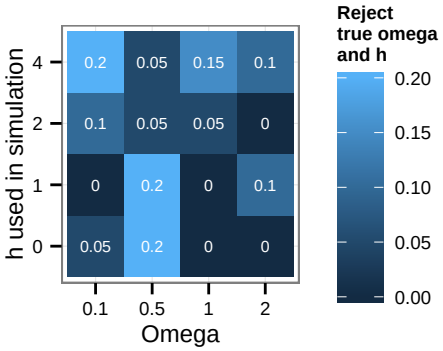


Figure B.9: Proportion of fully context-dependent simulated data sets simulated under the symmetric WRC/GYW model that incorrectly reject the true values of h and ω.

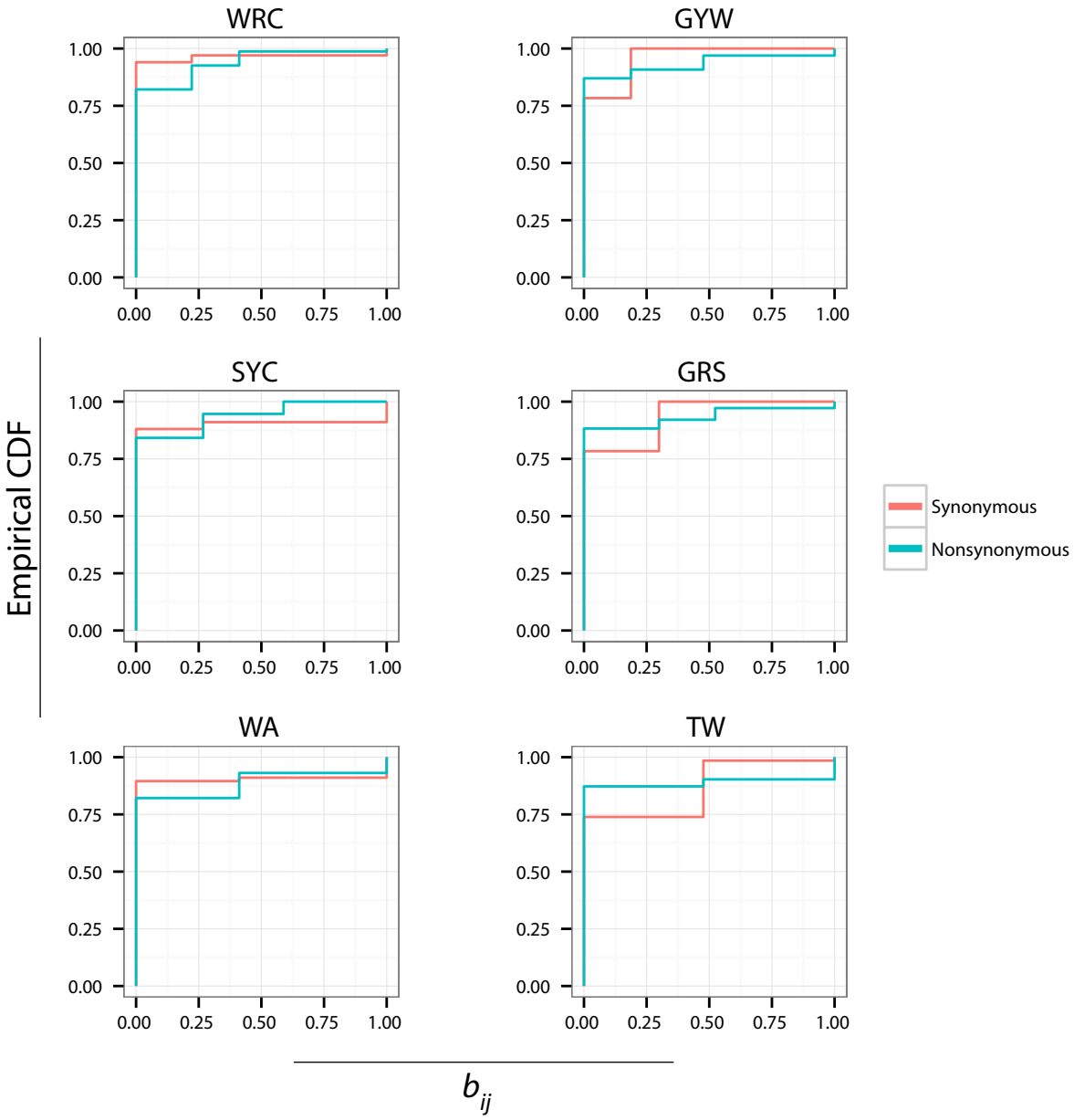


Figure B.10: Cumulative distribution of b^a_{ij} parameters among synonymous vs non-synonymous substitutions in the **Q** matrix formed from different motifs individually using the ML codon frequency results from CH103. Results are nearly identical in the other two lineages.

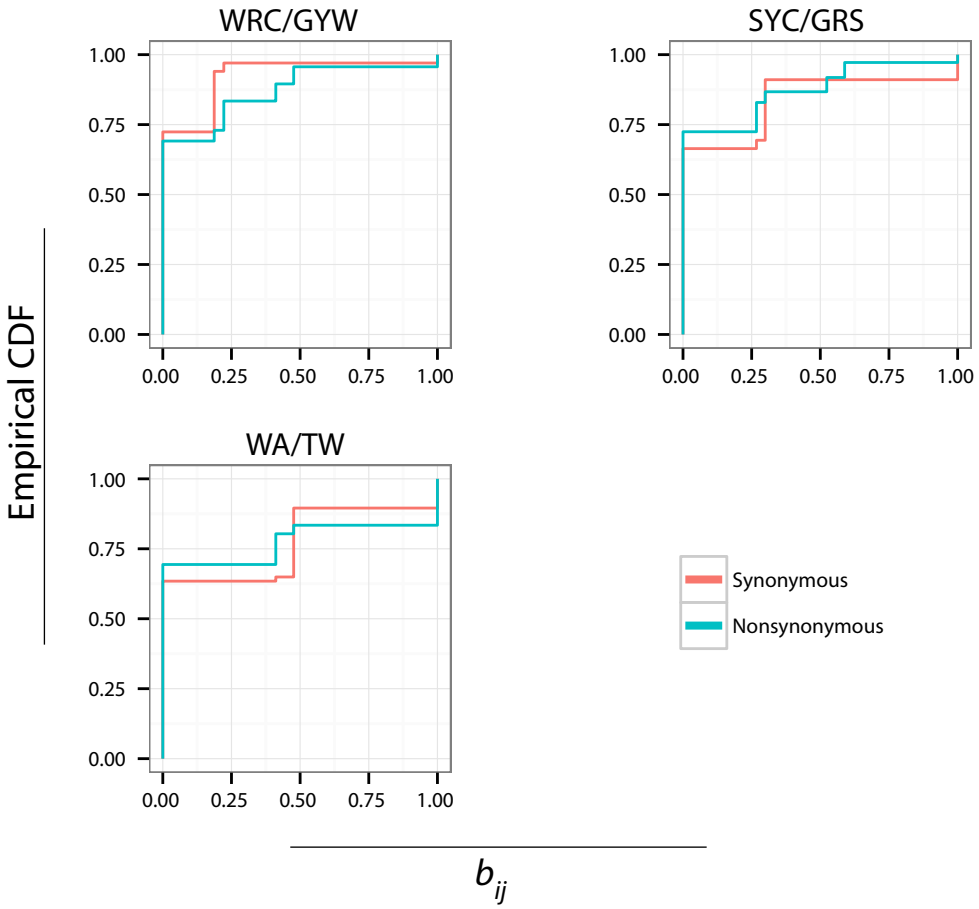


Figure B.11: Cumulative distribution of b^a_{ij} parameters among synonymous vs non-synonymous substitutions in the **Q** matrix formed from different motifs together and symmetrically using the ML codon frequency results from CH103. Results are nearly identical in the other two lineages.

B.7 Simulation analysis using empirical codon frequencies

Set	h	MLE h	Bias	Variability	Type 1 error	Type 2 error
CH103	0	-0.042	-0.042	0.009	-	0.25
	1	0.829	-0.171	0.021	0	0.35
	2	1.546	-0.454	0.020	0	0.85
	4	2.935	-1.065	0.062	0	1
VRC26	0	-0.011	-0.011	0.003	-	0
	1	0.894	-0.106	0.006	0	0.35
	2	1.719	-0.281	0.008	0	0.85
	4	3.126	-0.874	0.042	0	1
VRC01	0	0.032	0.032	0.006	-	0.25
	1	0.758	-0.242	0.012	0	0.8
	2	1.270	-0.730	0.031	0	1
	4	2.167	-1.833	0.028	0	1

Table B.2: Effects of empirical equilibrium frequencies on h estimation. Type 1 error rate shows the proportion of data sets that incorrectly failed to reject the null hypothesis of $h = 0$. Type 2 error rate shows the proportion of data sets that rejected the true value of h shown in the first column. Both of these hypothesis tests used an alpha value of 0.05. See Methods for explanation of how summary statistics were calculated.

B.8 Ancestral reconstruction results in bnAb lineages

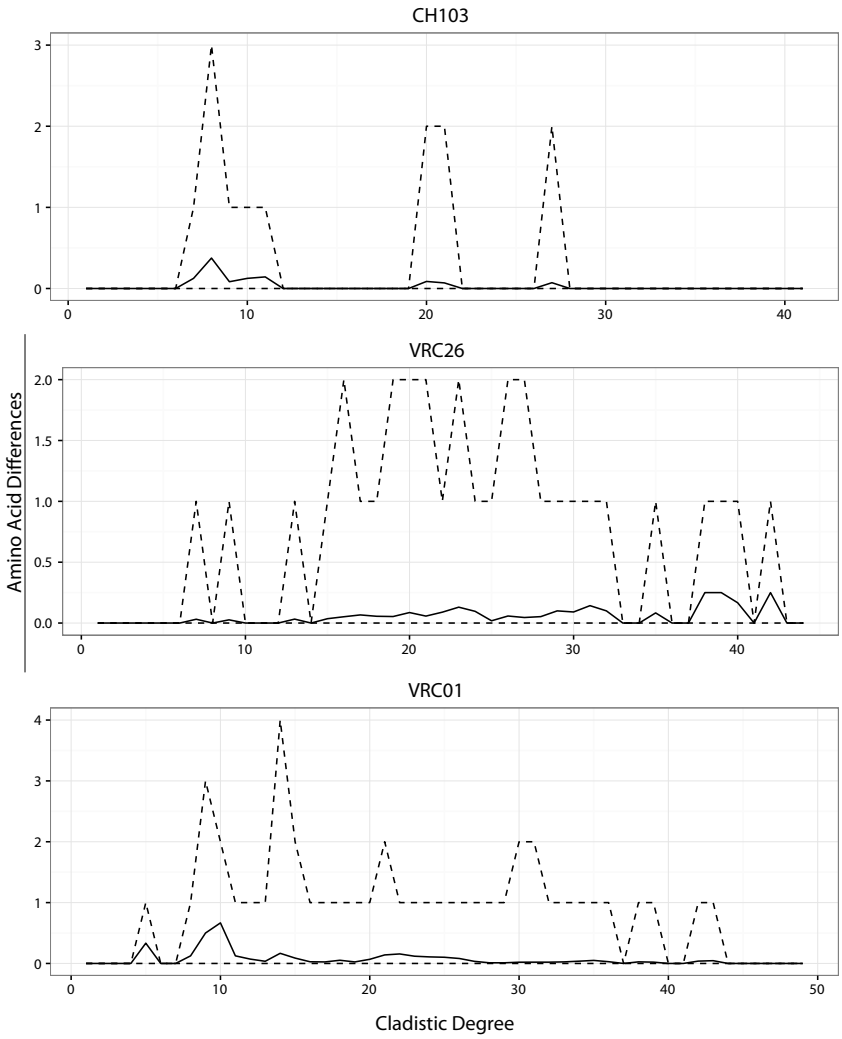


Figure B.12: Differences between predicted ancestors for the best fit HLP17 model (Supplemental Figure 3) and the NY98 model for three bnAb lineages by cladistics degree (nodes from root sequence). Solid line shows the mean difference for each node, dashed lines show the minimum and maximum values for each degree.

B.9 Ancestral reconstruction results from simulations

Set	h	Reconstructed sites per dataset	Mean incorrect amino acids per dataset		Improvement for h=MLE
			h=MLE	h=0	
CH103	0	36207	25.05	25.2	0.15
	1		25.65	25.4	-0.25
	2		25.1	25.85	0.75
	4		24.95	28.6	3.65
VRC26	0	64992	62.55	62.7	0.15
	1		64.2	64.05	-0.15
	2		67.05	68	0.95
	4		67.35	72.9	5.55
VRC01	0	98880	88.9	88.75	-0.15
	1		85.25	85.45	0.2
	2		87.25	91.45	4.2
	4		89.85	99.3	9.45

Table B.3: Accuracy of ancestral state reconstruction in simulation analyses. The mean number of incorrectly reconstructed amino acids per data set are shown for the h=MLE and h = 0 models, for each set of h parameters used in simulations.

Bibliography

[Abrahams et al., 2009] Abrahams, M.-R., Anderson, J. A., Giorgi, E. E., Seoighe, C., Mlisana, K., Ping, L.-H., Athreya, G. S., Treurnicht, F. K., Keele, B. F., Wood, N., Salazar-Gonzalez, J. F., Bhattacharya, T., Chu, H., Hoffman, I., Galvin, S., Mapanje, C., Kazembe, P., Thebus, R., Fiscus, S., Hide, W., Cohen, M. S., Karim, S. A., Haynes, B. F., Shaw, G. M., Hahn, B. H., Korber, B. T., Swanstrom, R., Williamson, C., and Consortium, f. t. C. A. I. S. T. a. t. C. f. H.-A. V. I. (2009). 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.

[Ademokun et al., 2011] Ademokun, A., Wu, Y.-C., Martin, V., Mitra, R., Sack, U., Baxendale, H., Kipling, D., and Dunn-Walters, D. K. (2011). Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages. *Aging Cell*, 10(6):922–930.

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

[Alexandrov et al., 2013] Alexandrov, L., Nik-Zainal, S., Wedge, D., Campbell, P., and Stratton, M. (2013). Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports*, 3(1):246–259.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410. 53092.

[Anisimova et al., 2006] Anisimova, M., Gascuel, O., and Sullivan, J. (2006). Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. *Systematic Biology*, 55(4):539–552.

[Barak et al., 2008] Barak, M., Zuckerman, N. S., Edelman, H., Unger, R., and Mehr, R. (2008). IgTree: Creating Immunoglobulin variable region gene lineage trees. *Journal of Immunological Methods*, 338(12):67–74.

[Bashford-Rogers et al., 2014] Bashford-Rogers, R. J., Palser, A. L., Idris, S. F., Carter, L., Epstein, M., Callard, R. E., Douek, D. C., Vassiliou, G. S., Follows, G. A., Hubank, M., and Kellam, P. (2014). Capturing needles in haystacks: a comparison of B-cell receptor sequencing methods. *BMC Immunology*, 15(1):29.

[Bashford-Rogers et al., 2013] Bashford-Rogers, R. J. M., Palser, A. L., Huntly, B. J., Rance, R., Vassiliou, G. S., Follows, G. A., and Kellam, P. (2013). Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Research*, 23(11):1874–1884.

[Benichou et al., 2012] Benichou, J., Ben-Hamo, R., Louzoun, Y., and Efroni, S. (2012). Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology*, 135(3):183–191.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

[Bollback et al., 2007] Bollback, J. P., Gardner, P. P., and Nielsen, R. (2007). Estimating the history of mutations on a phylogeny. In *Ancestral Sequence Reconstruction*, pages 69–79. Oxford University Press, Oxford, United Kingdon.

[Boussau et al., 2008] Boussau, B., Blanquart, S., Necsulea, A., Lartillot, N., and Gouy, M. (2008). Parallel adaptations to high temperatures in the Archaean eon. *Nature*, 456(7224):942–945.

[Boussau and Gouy, 2006] Boussau, B. and Gouy, M. (2006). Efficient Likelihood Computations with Nonreversible Models of Evolution. *Systematic Biology*, 55(5):756–768.

[Boyd et al., 2009] Boyd, S. D., Marshall, E. L., Merker, J. D., Maniar, J. M., Zhang, L. N., Sahaf, B., Jones, C. D., Simen, B. B., Hanczaruk, B., Nguyen, K. D., Nadeau, K. C., Egholm, M., Miklos, D. B., Zehnder, J. L., and Fire, A. Z. (2009). Measurement and Clinical Monitoring of Human Lymphocyte Clonality by Massively Parallel V-D-J Pyrosequencing. *Science Translational Medicine*, 1(12):12ra23–12ra23. 00147 PMID: 20161664.

[Bransteitter et al., 2004] Bransteitter, R., Pham, P., Calabrese, P., and Goodman, M. F. (2004). Biochemical Analysis of Hypermutational Targeting by Wild Type and Mutant Activation-induced Cytidine Deaminase. *Journal of Biological Chemistry*, 279(49):51612–51621.

[Brent, 1971] Brent, R. P. (1971). An algorithm with guaranteed convergence for finding a zero of a function. *The Computer Journal*, 14(4):422–425.

[Brochet et al., 2008] Brochet, X., Lefranc, M.-P., and Giudicelli, V. (2008). IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Research*, 36(Web Server issue):W503–508. 00303 PMID: 18503082.

[Broyden, 1970] Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms: 2. The new algorithm. *IMA journal of applied mathematics*, 6(3):222–231.

[Burnet, 1976] Burnet, F. M. (1976). A modification of jerne's theory of antibody production using the concept of clonal selection. *CA: A Cancer Journal for Clinicians*, 26(2):119–121.

[Burton and Hangartner, 2016] Burton, D. R. and Hangartner, L. (2016). Broadly Neutralizing Antibodies to HIV and Their Role in Vaccine Design. *Annual Re-*

view of Immunology, 34(1):635–659.

[Burton and Mascola, 2015] Burton, D. R. and Mascola, J. R. (2015). Antibody responses to envelope glycoproteins in HIV-1 infection. *Nature Immunology*, 16(6):571–576.

[Bush et al., 1999] Bush, R. M., Bender, C. A., Subbaro, K., Cox, N. J., and Fitch, W. M. (1999). Predicting the Evolution of Human Influenza A. *Science*, 286(5446):1921–1925.

[Busse et al., 2014] Busse, C. E., Czogiel, I., Braun, P., Arndt, P. F., and Wardemann, H. (2014). Single-cell based high-throughput sequencing of full-length immunoglobulin heavy and light chain genes. *European Journal of Immunology*, 44(2):597–603.

[Ceriani and Verme, 2012] Ceriani, L. and Verme, P. (2012). The origins of the Gini index: extracts from Variabilit e Mutabilit (1912) by Corrado Gini. *The Journal of Economic Inequality*, 10(3):421–443.

[Chang and Casali, 1995] Chang, B. and Casali, P. (1995). A Sequence Analysis of Human Germline Ig VH and VL Genes. *Annals of the New York Academy of Sciences*, 764(1):170–179.

[Chen et al., 2012] Chen, W., Prabakaran, P., Zhu, Z., Feng, Y., Streaker, E. D., and Dimitrov, D. S. (2012). Characterization of human IgG repertoires in an acute HIV-1 infection. *Experimental and Molecular Pathology*, 93(3):399–407.

[Chiorazzi et al., 2005] Chiorazzi, N., Rai, K. R., and Ferrarini, M. (2005). Chronic Lymphocytic Leukemia. *New England Journal of Medicine*, 352(8):804–815.

[Colombo et al., 2003] Colombo, M., Dono, M., Gazzola, P., Chiorazzi, N., Mancardi, G., and Ferrarini, M. (2003). Maintenance of B lymphocyte-related clones in the cerebrospinal fluid of multiple sclerosis patients. *European Journal of Immunology*, 33(12):3433–3438.

[Colombo et al., 2000] Colombo, M., Dono, M., Gazzola, P., Roncella, S., Valetto, A., Chiorazzi, N., Mancardi, G. L., and Ferrarini, M. (2000). Accumulation of Clonally Related B Lymphocytes in the Cerebrospinal Fluid of Multiple Sclerosis Patients. *The Journal of Immunology*, 164(5):2782–2789.

[Cowell and Kepler, 2000] Cowell, L. G. and Kepler, T. B. (2000). The Nucleotide-Replacement Spectrum Under Somatic Hypermutation Exhibits Microsequence Dependence That Is Strand-Symmetric and Distinct from That Under Germline Mutation. *The Journal of Immunology*, 164(4):1971–1976.

[DeKosky et al., 2013] DeKosky, B. J., Ippolito, G. C., Deschner, R. P., Lavinder, J. J., Wine, Y., Rawlings, B. M., Varadarajan, N., Giesecke, C., Drner, T., Andrews, S. F., Wilson, P. C., Hunicke-Smith, S. P., Willson, C. G., Ellington, A. D., and Georgiou, G. (2013). High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nature Biotechnology*, 31(2):166–169. 00052.

[Dimmic et al., 2002] Dimmic, M. W., Rest, J. S., Mindell, D. P., and Goldstein, R. A. (2002). rtREV: An Amino Acid Substitution Matrix for Inference of Retrovirus and Reverse Transcriptase Phylogeny. *Journal of Molecular Evolution*, 55(1):65–73.

[Doria-Rose et al., 2014] Doria-Rose, N. A., Schramm, C. A., Gorman, J., Moore, P. L., Bhiman, J. N., DeKosky, B. J., Ernandes, M. J., Georgiev, I. S., Kim, H. J., Pancera, M., Staupe, R. P., Altae-Tran, H. R., Bailer, R. T., Crooks, E. T., Cupo, A., Druz, A., Garrett, N. J., Hoi, K. H., Kong, R., Louder, M. K., Longo, N. S., McKee, K., Nonyane, M., ODell, S., Roark, R. S., Rudicell, R. S., Schmidt, S. D., Sheward, D. J., Soto, C., Wibmer, C. K., Yang, Y., Zhang, Z., NISC Comparative Sequencing, Mullikin, J. C., Binley, J. M., Sanders, R. W., Wilson, I. A., Moore, J. P., Ward, A. B., Georgiou, G., Williamson, C., Abdool Karim, S. S., Morris, L., Kwong, P. D., Shapiro, L., and Mascola, J. R. (2014). Developmental pathway for potent V1v2-directed HIV-neutralizing antibodies. *Nature*, 509(7498):55–62.

[Drummond and Rambaut, 2007] Drummond, A. J. and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(1):214. 05112 PMID: 17996036.

[Drummond and Suchard, 2010] Drummond, A. J. and Suchard, M. A. (2010). Bayesian random local clocks, or one rate to rule them all. *BMC Biology*, 8:114.

[Dunn-Walters et al., 2002] Dunn-Walters, D. K., Belelovsky, A., Edelman, H., Banerjee, M., and Mehr, R. (2002). The Dynamics of Germinal Centre Selection as Measured by Graph-Theoretical Analysis of Mutational Lineage Trees. *Developmental Immunology*, 9(4):233–243.

[Dunn-Walters et al., 2004] Dunn-Walters, D. K., Edelman, H., and Mehr, R. (2004). Immune system learning and memory quantified by graphical analysis of B-lymphocyte phylogenetic trees. *Biosystems*, 76(13):141–155.

[Dunn-Walters and Spencer, 1998] Dunn-Walters, D. K. and Spencer, J. (1998). Strong intrinsic biases towards mutation and conservation of bases in human IgVH genes during somatic hypermutation prevent statistical analysis of antigen selection. *Immunology*, 95(3):339–345.

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

[Edo-Matas et al., 2011] Edo-Matas, D., Lemey, P., Tom, J. A., Serna-Bolea, C., Blink, V. D., E, A., Wout, V. T., B, A., Schuitemaker, H., and Suchard, M. A. (2011). Impact of CCR5delta32 Host Genetic Background and Disease Progression on HIV-1 Intrahost Evolutionary Processes: Efficient Hypothesis Testing through Hierarchical Phylogenetic Models. *Molecular Biology and Evolution*, 28(5):1605–1616.

[Eisen and Siskind, 1964] Eisen, H. N. and Siskind, G. W. (1964). Variations in affinities of antibodies during the immune response. *Biochemistry*, 3(7):996–1008.

[Elhanati et al., 2015] Elhanati, Y., Sethna, Z., Marcou, Q., Callan, C. G., Mora, T., and Walczak, A. M. (2015). Inferring processes underlying B-cell repertoire diversity. *Phil. Trans. R. Soc. B*, 370(1676):20140243.

[Eroshkin et al., 2013] Eroshkin, A. M., LeBlanc, A., Weekes, D., Post, K., Li, Z., Rajput, A., Butera, S. T., Burton, D. R., and Godzik, A. (2013). bNAber: database of broadly neutralizing HIV antibodies. *Nucleic Acids Research*, 42(D1):D1133–D1139.

[Felsenstein, 1978] Felsenstein, J. (1978). Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. *Systematic Zoology*, 27(4):401–410.

[Felsenstein, 1981] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376.

[Fera et al., 2014] Fera, D., Schmidt, A. G., Haynes, B. F., Gao, F., Liao, H.-X., Kepler, T. B., and Harrison, S. C. (2014). Affinity maturation in an HIV broadly neutralizing B-cell lineage through reorientation of variable domains. *Proceedings of the National Academy of Sciences*, 111(28):10275–10280. 00001.

[Fitch, 1971] Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specified tree topology. *Systematic Zoology*, 20(4):406–416.

[Fletcher, 1970] Fletcher, R. (1970). A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317–322.

[Frost et al., 2015] Frost, S. D. W., Murrell, B., Hossain, A. S. M. M., Silverman, G. J., and Pond, S. L. K. (2015). Assigning and visualizing germline genes in antibody repertoires. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1676):20140240.

[Fu et al., 2012] Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152. 00057 PMID: 23060610.

[Furlong and Yang, 2008] Furlong, R. F. and Yang, Z. (2008). Diversifying and Purifying Selection in the Peptide Binding Region of DRB in Mammals. *Journal of Molecular Evolution*, 66(4):384–394.

[Gadala-Maria et al., 2015] Gadala-Maria, D., Yaari, G., Uduman, M., and Kleinstein, S. H. (2015). Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proceedings of the National Academy of Sciences*, 112(8):E862–E870.

[Gaeta et al., 2007] Gaeta, B. A., Malming, H. R., Jackson, K. J., Bain, M. E., Wilson, P., and Collins, A. M. (2007). iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics*, 23(13):1580–1587. 00044.

[Gall et al., 2013] Gall, A., Kaye, S., Hu, S., Bonsall, D., Rance, R., Baillie, G. J., Fidler, S. J., Weber, J. N., McClure, M. O., Kellam, P., and others (2013). Restriction of V3 region sequence divergence in the HIV-1 envelope gene during antiretroviral treatment in a cohort of recent seroconverters. *Retrovirology*, 10(1):8–1. 00003.

[Galson et al., 2015] Galson, J. D., Clutterbuck, E. A., Trck, J., Ramasamy, M. N., Mnz, M., Fowler, A., Cerundolo, V., Pollard, A. J., Lunter, G., and Kelly, D. F. (2015). BCR repertoire sequencing: different patterns of B-cell activation after two Meningococcal vaccines. *Immunology and Cell Biology*.

[Gao et al., 2014] Gao, F., Bonsignori, M., Liao, H.-X., Kumar, A., Xia, S.-M., Lu, X., Cai, F., Hwang, K.-K., Song, H., Zhou, T., Lynch, R., Alam, S. M., Moody, M. A., Ferrari, G., Berrong, M., Kelsoe, G., Shaw, G., Hahn, B., Montefiori, D., Kamanga, G., Cohen, M., Hraber, P., Kwong, P., Korber, B., Mascola, J., Kepler, T., and Haynes, B. (2014). Cooperation of B Cell Lineages in Induction of HIV-1-Broadly Neutralizing Antibodies. *Cell*, 158(3):481–491. 00005.

[Georgiou et al., 2014] Georgiou, G., Ippolito, G. C., Beausang, J., Busse, C. E.,

Wardemann, H., and Quake, S. R. (2014). The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature Biotechnology*, 32(2):158–168.

[Gil et al., 2013] Gil, M., Zanetti, M. S., Zoller, S., and Anisimova, M. (2013). CodonPhyML: Fast Maximum Likelihood Phylogeny Estimation under Codon Substitution Models. *Molecular Biology and Evolution*, 30(6):1270–1280.

[Goldfarb, 1970] Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26.

[Goldman and Yang, 1994] Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, 11(5):725–736.

[Greiff et al., 2015] Greiff, V., Bhat, P., Cook, S. C., Menzel, U., Kang, W., and Reddy, S. T. (2015). A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Medicine*, 7(1):49.

[Grenfell et al., 2004] Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A., and Holmes, E. C. (2004). Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science*, 303(5656):327–332.

[Gupta et al., 2017] Gupta, N. T., Adams, K. D., Briggs, A. W., Timberlake, S. C., Vigneault, F., and Kleinstein, S. H. (2017). Hierarchical Clustering Can Identify B Cell Clones with High Confidence in Ig Repertoire Sequencing Data. *The Journal of Immunology*, 198(6):2489–2499.

[Gupta et al., 2015] Gupta, N. T., Vander Heiden, J. A., Uduman, M., Gadala-Maria, D., Yaari, G., and Kleinstein, S. H. (2015). Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data: Table 1. *Bioinformatics*, page btv359.

[Haynes et al., 2012] Haynes, B. F., Kelsoe, G., Harrison, S. C., and Kepler, T. B. (2012). B-cell-lineage immunogen design in vaccine development with HIV-1

as a case study. *Nature Biotechnology*, 30(5):423–433. 00112.

[Heiden et al., 2017] Heiden, J. A. V., Stathopoulos, P., Zhou, J. Q., Chen, L., Gilbert, T. J., Bolen, C. R., Barohn, R. J., Dimachkie, M. M., Ciafaloni, E., Broering, T. J., Vigneault, F., Nowak, R. J., Kleinstein, S. H., and OConnor, K. C. (2017). Dysregulation of B Cell Repertoire Formation in Myasthenia Gravis Patients Revealed through Deep Sequencing. *The Journal of Immunology*, page 1601415.

[Heiden et al., 2014] Heiden, J. A. V., Yaari, G., Uduman, M., Stern, J. N. H., OConnor, K. C., Hafler, D. A., Vigneault, F., and Kleinstein, S. H. (2014). pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics*, 30(13):1930–1932. 00000 PMID: 24618469.

[Hemelaar et al., 2011] Hemelaar, J., Gouws, E., Ghys, P. D., and Osmanov, S. (2011). Global trends in molecular epidemiology of HIV-1 during 20002007. *AIDS (London, England)*, 25(5):679–689.

[Hershberg and Luning Prak, 2015] Hershberg, U. and Luning Prak, E. T. (2015). The analysis of clonal expansions in normal and autoimmune B cell repertoires. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1676):20140239.

[Hershberg et al., 2008] Hershberg, U., Uduman, M., Shlomchik, M. J., and Kleinstein, S. H. (2008). Improved methods for detecting selection by mutation analysis of Ig V region sequences. *International Immunology*, 20(5):683–694.

[Holtz et al., 2013] Holtz, C. M., Sadler, H. A., and Mansky, L. M. (2013). APOBEC3g cytosine deamination hotspots are defined by both sequence context and single-stranded DNA secondary structure. *Nucleic Acids Research*, 41(12):6139–6148.

[Huelsenbeck and Rannala, 1997] Huelsenbeck, J. P. and Rannala, B. (1997). Phylogenetic Methods Come of Age: Testing Hypotheses in an Evolutionary Context. *Science*, 276(5310):227–232.

[Jiang et al., 2013] Jiang, N., He, J., Weinstein, J. A., Penland, L., Sasaki, S., He, X.-S., Dekker, C. L., Zheng, N.-Y., Huang, M., Sullivan, M., Wilson, P. C., Greenberg, H. B., Davis, M. M., Fisher, D. S., and Quake, S. R. (2013). Lineage Structure of the Human Antibody Repertoire in Response to Influenza Vaccination. *Science Translational Medicine*, 5(171):171ra19–171ra19. 00040.

[Keele et al., 2008] Keele, B. F., Giorgi, E. E., Salazar-Gonzalez, J. F., Decker, J. M., Pham, K. T., Salazar, M. G., Sun, C., Grayson, T., Wang, S., Li, H., Wei, X., Jiang, C., Kirchherr, J. L., Gao, F., Anderson, J. A., Ping, L.-H., Swanstrom, R., Tomaras, G. D., Blattner, W. A., Goepfert, P. A., Kilby, J. M., Saag, M. S., Delwart, E. L., Busch, M. P., Cohen, M. S., Montefiori, D. C., Haynes, B. F., Gaschen, B., Athreya, G. S., Lee, H. Y., Wood, N., Seoighe, C., Perelson, A. S., Bhattacharya, T., Korber, B. T., Hahn, B. H., and Shaw, G. M. (2008). Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proceedings of the National Academy of Sciences*, 105(21):7552–7557.

[Keele et al., 2006] Keele, B. F., Heuverswyn, F. V., Li, Y., Bailes, E., Takehisa, J., Santiago, M. L., Bibollet-Ruche, F., Chen, Y., Wain, L. V., Liegeois, F., Loul, S., Ngole, E. M., Bienvenue, Y., Delaporte, E., Brookfield, J. F. Y., Sharp, P. M., Shaw, G. M., Peeters, M., and Hahn, B. H. (2006). Chimpanzee Reservoirs of Pandemic and Nonpandemic HIV-1. *Science*, 313(5786):523–526.

[Keilson, 1979] Keilson, J. (1979). *Markov Chain Models Rarity and Exponentiality | J. Keilson | Springer*. Springer, New York.

[Kepler et al., 2014a] Kepler, T., Liao, H.-X., Alam, S. M., Bhaskarabhatla, R., Zhang, R., Yandava, C., Stewart, S., Anasti, K., Kelsoe, G., Parks, R., Lloyd, K., Stolarchuk, C., Pritchett, J., Solomon, E., Friberg, E., Morris, L., Karim, S. A., Cohen, M., Walter, E., Moody, M. A., Wu, X., Altae-Tran, H., Georgiev, I., Kwong, P., Boyd, S., Fire, A., Mascola, J., and Haynes, B. (2014a). Immunoglobulin Gene Insertions and Deletions in the Affinity Maturation of HIV-1 Broadly Reactive Neutralizing Antibodies. *Cell Host & Microbe*, 16(3):304–313.

[Kepler, 2013] Kepler, T. B. (2013). Reconstructing a B-cell clonal lineage. I. Statistical inference of unobserved ancestors. *F1000Research*, 2(103).

[Kepler et al., 2014b] Kepler, T. B., Munshaw, S., Wiehe, K., Zhang, R., Yu, J.-S., Woods, C. W., Denny, T. N., Tomaras, G. D., Alam, S. M., Moody, M. A., Kelsoe, G., Liao, H.-X., and Haynes, B. F. (2014b). Reconstructing a B-cell clonal lineage. II. Mutation, selection, and affinity maturation. *B Cell Biology*, 5:170.

[Kosakovsky Pond et al., 2010] Kosakovsky Pond, S., Delport, W., Muse, S. V., and Scheffler, K. (2010). Correcting the Bias of Empirical Frequency Parameter Estimators in Codon Models. *PLoS ONE*, 5(7):e11230.

[Koshi and Goldstein, 1996] Koshi, J. M. and Goldstein, R. A. (1996). Probabilistic reconstruction of ancestral protein sequences. *Journal of molecular evolution*, 42(2):313–320.

[Kosiol et al., 2007] Kosiol, C., Holmes, I., and Goldman, N. (2007). An Empirical Codon Model for Protein Sequence Evolution. *Molecular Biology and Evolution*, 24(7):1464–1479.

[Larimore et al., 2012] Larimore, K., McCormick, M. W., Robins, H. S., and Greenberg, P. D. (2012). Shaping of Human Germline IgH Repertoires Revealed by Deep Sequencing. *The Journal of Immunology*, 189(6):3221–3230.

[Laserson et al., 2014] Laserson, U., Vigneault, F., Gadala-Maria, D., Yaari, G., Uduman, M., Heiden, J. A. V., Kelton, W., Jung, S. T., Liu, Y., Laserson, J., Chari, R., Lee, J.-H., Bachelet, I., Hickey, B., Lieberman-Aiden, E., Hanczaruk, B., Simen, B. B., Egholm, M., Koller, D., Georgiou, G., Kleinstein, S. H., and Church, G. M. (2014). High-resolution antibody dynamics of vaccine-induced immune responses. *Proceedings of the National Academy of Sciences*, 111(13):4928–4933.

[Le and Gascuel, 2008] Le, S. Q. and Gascuel, O. (2008). An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution*, 25(7):1307–

1320.

[Lefranc et al., 2009] Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., Regnier, L., Ehrenmann, F., Lefranc, G., and Duroux, P. (2009). IMGT(R), the international ImMunoGeneTics information system(R). *Nucleic Acids Research*, 37(Database):D1006–D1012.

[Lefranc and Lefranc, 2001] Lefranc, M.-P. and Lefranc, G. (2001). *The Immunoglobulin FactsBook*. Academic Press.

[Leitman et al., 2017] Leitman, E. M., Thobakgale, C. F., Adland, E., Ansari, M. A., Raghwani, J., Prendergast, A. J., Tudor-Williams, G., Kiepiela, P., Hemelaar, J., Brener, J., Tsai, M.-H., Mori, M., Riddell, L., Luzzi, G., Jooste, P., Ndungu, T., Walker, B. D., Pybus, O. G., Kellam, P., Naranbhai, V., Matthews, P. C., Gall, A., and Goulder, P. J. (2017). Role of HIV-specific CD8+ T cells in pediatric HIV cure strategies after widespread early viral escape. *The Journal of Experimental Medicine*, page jem.20162123.

[Lemey et al., 2007] Lemey, P., Kosakovsky Pond, S. L., Drummond, A. J., Pybus, O. G., Shapiro, B., Barroso, H., Taveira, N., and Rambaut, A. (2007). Synonymous Substitution Rates Predict HIV Disease Progression as a Result of Underlying Replication Dynamics. *PLoS Computational Biology*, 3(2):e29.

[Lemey et al., 2012] Lemey, P., Minin, V. N., Bielejec, F., Pond, S. L. K., and Suchard, M. A. (2012). A counting renaissance: combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. *Bioinformatics*, 28(24):3248–3256.

[Lemey et al., 2009] Lemey, P., Rambaut, A., Drummond, A. J., and Suchard, M. A. (2009). Bayesian Phylogeography Finds Its Roots. *PLOS Computational Biology*, 5(9):e1000520.

[Li, 2004] Li, A. (2004). Utilization of Ig heavy chain variable, diversity, and join-

ing gene segments in children with B-lineage acute lymphoblastic leukemia: implications for the mechanisms of VDJ recombination and for pathogenesis. *Blood*, 103(12):4602–4609.

[Liao et al., 2013] Liao, H.-X., Lynch, R., Zhou, T., Gao, F., Alam, S. M., Boyd, S. D., Fire, A. Z., Roskin, K. M., Schramm, C. A., Zhang, Z., Zhu, J., Shapiro, L., NISC Comparative Sequencing Program, Mullikin, J. C., Gnanakaran, S., Hraber, P., Wiehe, K., Kelsoe, G., Yang, G., Xia, S.-M., Montefiori, D. C., Parks, R., Lloyd, K. E., Scearce, R. M., Soderberg, K. A., Cohen, M., Kamanga, G., Louder, M. K., Tran, L. M., Chen, Y., Cai, F., Chen, S., Moquin, S., Du, X., Joyce, M. G., Srivatsan, S., Zhang, B., Zheng, A., Shaw, G. M., Hahn, B. H., Kepler, T. B., Korber, B. T. M., Kwong, P. D., Mascola, J. R., and Haynes, B. F. (2013). Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature*, 496(7446):469–476. 00132.

[Lio and Goldman, 1998] Lio, P. and Goldman, N. (1998). Models of molecular evolution and phylogeny. *Genome research*, 8(12):1233–1244.

[Lunter and Hein, 2004] Lunter, G. and Hein, J. (2004). A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics*, 20(suppl 1):i216–i223.

[Maartens et al., 2014] Maartens, G., Celum, C., and Lewin, S. R. (2014). HIV infection: epidemiology, pathogenesis, treatment, and prevention. *The Lancet*, 384(9939):258–271.

[Mascola et al., 2000] Mascola, J. R., Stiegler, G., VanCott, T. C., Katinger, H., Carpenter, C. B., Hanson, C. E., Beary, H., Hayes, D., Frankel, S. S., Birx, D. L., and Lewis, M. G. (2000). Protection of macaques against vaginal transmission of a pathogenic HIV-1/SIV chimeric virus by passive infusion of neutralizing antibodies. *Nature Medicine*, 6(2):207.

[McCoy et al., 2015] McCoy, C. O., Bedford, T., Minin, V. N., Bradley, P., Robins,

H., and Matsen, F. A. (2015). Quantifying evolutionary constraints on B-cell affinity maturation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1676):20140244.

[McCoy and McKnight, 2017] McCoy, L. E. and McKnight, . (2017). Lessons learned from humoral responses of HIV patients:. *Current Opinion in HIV and AIDS*, 12(3):195–202.

[McCoy and Weiss, 2013] McCoy, L. E. and Weiss, R. A. (2013). Neutralizing antibodies to HIV-1 induced by immunization. *The Journal of Experimental Medicine*, 210(2):209–223.

[McMichael et al., 2010] McMichael, A. J., Borrow, P., Tomaras, G. D., Goonetilleke, N., and Haynes, B. F. (2010). The immune response during acute HIV-1 infection: clues for vaccine development. *Nature reviews. Immunology*, 10(1):11–23.

[Mirsky et al., 2015] Mirsky, A., Kazandjian, L., and Anisimova, M. (2015). Antibody-Specific Model of Amino Acid Substitution for Immunological Inferences from Alignments of Antibody Sequences. *Molecular Biology and Evolution*, 32(3):806–819.

[Moir et al., 2010] Moir, S., Buckner, C. M., Ho, J., Wang, W., Chen, J., Waldner, A. J., Posada, J. G., Kardava, L., O'Shea, M. A., Kottilil, S., Chun, T.-W., Proschan, M. A., and Fauci, A. S. (2010). B cells in early and chronic HIV infection: evidence for preservation of immune function associated with early initiation of antiretroviral therapy. *Blood*, 116(25):5571–5579. 00089 Characterization of lymphocytes including B cells during early versus chronic HIV infection is important for understanding the impact of chronic viremia on immune cell function. In this setting, we investigated B cells before and after reduction of HIV plasma viremia by antiretroviral therapy (ART). At baseline, peripheral blood B-cell counts were significantly lower in both early and chronic HIV-infected individuals compared with uninfected controls. Similar to CD4+ but

not CD8+ T cells, B-cell numbers in both groups increased significantly after ART. At baseline, B cells of early HIV-infected individuals were composed of a higher percentage of plasmablasts and resting memory B cells compared with chronic HIV-infected individuals whose B cells were composed of a higher percentage of immature/transitional and exhausted B cells compared with their early infection counterparts. At 1 year after ART, the percentage of resting memory B cells remained higher in early compared with chronic HIV-infected individuals. This difference translated into a better functional profile in that memory B-cell responses to HIV and non-HIV antigens were superior in early- compared with chronic-treated HIV infected individuals. These findings provide new insights on B cells in HIV infection and how early initiation of ART may prevent irreversible immune system damage.

[Moler and Van Loan, 1978] Moler, C. and Van Loan, C. (1978). Nineteen dubious ways to compute the exponential of a matrix. *SIAM review*, 20(4):801–836.

[Monajemi et al., 2014] Monajemi, M., Woodworth, C. F., Zipperlen, K., Gallant, M., Grant, M. D., and Larijani, M. (2014). Positioning of APOBEC3g/F Mutational Hotspots in the Human Immunodeficiency Virus Genome Favors Reduced Recognition by CD8+ T Cells. *PLoS ONE*, 9(4):e93428.

[Montefiori, 2005] Montefiori, D. C. (2005). Evaluating neutralizing antibodies against HIV, SIV, and SHIV in luciferase reporter gene assays. *Current protocols in immunology*, pages 12–11.

[Montefiori et al., 2012] Montefiori, D. C., Karnasuta, C., Huang, Y., Ahmed, H., Gilbert, P., Souza, D., S, M., McLinden, R., Tovanabutra, S., Laurence-Chenine, A., Sanders-Buell, E., Moody, M. A., Bonsignori, M., Ochsenbauer, C., Kappes, J., Tang, H., Greene, K., Gao, H., LaBranche, C. C., Andrews, C., Polonis, V. R., Rerks-Ngarm, S., Pitisuttithum, P., Nitayaphan, S., Kaewkungwal, J., Self, S. G., Berman, P. W., Francis, D., Sinangil, F., Lee, C., Tartaglia, J., Robb, M. L., Haynes, B. F., Michael, N. L., and Kim, J. H. (2012). Magnitude and Breadth of the Neu-

tralizing Antibody Response in the RV144 and Vax003 HIV-1 Vaccine Efficacy Trials. *The Journal of Infectious Diseases*, 206(3):431–441.

[Mora et al., 2010] Mora, T., Walczak, A. M., Bialek, W., and Callan, C. G. (2010). Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences*, 107(12):5405–5410.

[Mphatswe et al., 2007] Mphatswe, W., Blanckenberg, N., Tudor-Williams, G., Prendergast, A., Thobakgale, C., Mkhwanazi, N., McCarthy, N., Walker, B. D., Kiepiela, P., and Goulder, P. (2007). High frequency of rapid immunological progression in African infants infected in the era of perinatal HIV prophylaxis. *AIDS*, 21(10):1253–1261.

[Muramatsu et al., 2000] Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y., and Honjo, T. (2000). Class Switch Recombination and Hypermutation Require Activation-Induced Cytidine Deaminase (AID), a Potential RNA Editing Enzyme. *Cell*, 102(5):553–563.

[Muramatsu et al., 1999] Muramatsu, M., Sankaranand, V. S., Anant, S., Sugai, M., Kinoshita, K., Davidson, N. O., and Honjo, T. (1999). Specific Expression of Activation-induced Cytidine Deaminase (AID), a Novel Member of the RNA-editing Deaminase Family in Germinal Center B Cells. *Journal of Biological Chemistry*, 274(26):18470–18476.

[Murphy et al., 2012] Murphy, K., Travers, P., Walport, M., and Janeway, C. (2012). *Janeway's immunobiology*. Garland Science, New York. OCLC: 733935898.

[Nei and Gojobori, 1986] Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, 3(5):418–426.

[Nielsen and Yang, 1998] Nielsen, R. and Yang, Z. (1998). Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene. *Genetics*, 148(3):929–936.

[Niklas et al., 2014] Niklas, N., Prll, J., Weinberger, J., Zopf, A., Wiesinger, K., Krismer, K., Bettelheim, P., and Gabriel, C. (2014). Qualifying high-throughput immune repertoire sequencing. *Cellular Immunology*, 288(12):31–38.

[Owen et al., 2013] Owen, J. A., Punt, J., Stranford, S. A., and Jones, P. P. (2013). *Kuby Immunology: International Edition*. W.H. Freeman, New York. OCLC: 820117219.

[Palanichamy et al., 2014] Palanichamy, A., Apeltsin, L., Kuo, T. C., Sirota, M., Wang, S., Pitts, S. J., Sundar, P. D., Telman, D., Zhao, L. Z., Derstine, M., Abounasr, A., Hauser, S. L., and Bdingen, H.-C. v. (2014). Immunoglobulin class-switched B cells form an active immune axis between CNS and periphery in multiple sclerosis. *Science Translational Medicine*, 6(248):248ra106–248ra106. 00001 PMID: 25100740.

[Parameswaran et al., 2013] Parameswaran, P., Liu, Y., Roskin, K., Jackson, K. L., Dixit, V., Lee, J.-Y., Artiles, K. L., Zompi, S., Vargas, M., Simen, B., Hanczaruk, B., McGowan, K., Tariq, M., Pourmand, N., Koller, D., Balmaseda, A., Boyd, S., Harris, E., and Fire, A. (2013). Convergent Antibody Signatures in Human Dengue. *Cell Host & Microbe*, 13(6):691–700.

[Parham, 2009] Parham, P. (2009). *The immune System*. Garland Science, London [u.a.

[Parren et al., 2001] Parren, P. W. H. I., Marx, P. A., Hessell, A. J., Luckay, A., Harouse, J., Cheng-Mayer, C., Moore, J. P., and Burton, D. R. (2001). Antibody Protects Macaques against Vaginal Challenge with a Pathogenic R5 Simian/Human Immunodeficiency Virus at Serum Levels Giving Complete Neutralization In Vitro. *Journal of Virology*, 75(17):8340–8347.

[Peled et al., 2008] Peled, J. U., Kuang, F. L., Iglesias-Ussel, M. D., Roa, S., Kalis, S. L., Goodman, M. F., and Scharff, M. D. (2008). The Biochemistry of Somatic Hypermutation. *Annual Review of Immunology*, 26(1):481–511.

[Pfeifer et al., 2005] Pfeifer, G. P., You, Y.-H., and Besaratinia, A. (2005). Mutations induced by ultraviolet light. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 571(12):19–31.

[Plantier et al., 2009] Plantier, J.-C., Leoz, M., Dickerson, J. E., Oliveira, F. D., Cordonnier, F., Leme, V., Damond, F., Robertson, D. L., and Simon, F. (2009). A new human immunodeficiency virus derived from gorillas. *Nature Medicine*, 15(8):871.

[Ralph and Matsen, 2016] Ralph, D. K. and Matsen, IV, F. A. (2016). Likelihood-Based Inference of B Cell Clonal Families. *PLOS Computational Biology*, 12(10):e1005086.

[Reddy et al., 2010] Reddy, S. T., Ge, X., Miklos, A. E., Hughes, R. A., Kang, S. H., Hoi, K. H., Chrysostomou, C., Hunicke-Smith, S. P., Iverson, B. L., Tucker, P. W., Ellington, A. D., and Georgiou, G. (2010). Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nature Biotechnology*, 28(9):965–969.

[Robins, 2013] Robins, H. (2013). Immunosequencing: applications of immune repertoire deep sequencing. *Current Opinion in Immunology*, 25(5):646–652.

[Rodrigo et al., 2003] Rodrigo, A. G., Goode, M., Forsberg, R., Ross, H. A., and Drummond, A. (2003). Inferring Evolutionary Rates Using Serially Sampled Sequences from Several Populations. *Molecular Biology and Evolution*, 20(12):2010–2018.

[Sattentau, 2011] Sattentau, Q. J. (2011). Vaccinology: A sweet cleft in HIV's armour. *Nature*, 480(7377):324.

[Schranz et al., 2008] Schranz, H. W., Yap, V., Easteal, S., Knight, R., and Huttley, G. A. (2008). Pathological rate matrices: from primates to pathogens. *BMC Bioinformatics*, 9(1):550.

[Serrano Zanetti, 2013] Serrano Zanetti, M. (2013). Markovian codon models of evolution for phylogeny inference.

[Shahaf et al., 2008] Shahaf, G., Barak, M., Zuckerman, N. S., Swerdlin, N., Gorfine, M., and Mehr, R. (2008). Antigen-driven selection in germinal centers as reflected by the shape characteristics of immunoglobulin gene lineage trees: A large-scale simulation study. *Journal of Theoretical Biology*, 255(2):210–222.

[Shanno, 1970] Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656.

[Shapiro et al., 2002] Shapiro, G. S., Aviszus, K., Murphy, J., and Wysocki, L. J. (2002). Evolution of Ig DNA Sequence to Target Specific Base Positions Within Codons for Somatic Hypermutation. *The Journal of Immunology*, 168(5):2302–2306.

[Shapiro et al., 2003] Shapiro, G. S., Ellison, M. C., and Wysocki, L. J. (2003). Sequence-specific targeting of two bases on both DNA strands by the somatic hypermutation mechanism. *Molecular Immunology*, 40(5):287–295.

[Sheng et al., 2016] Sheng, Z., Schramm, C. A., Connors, M., Morris, L., Mascola, J. R., Kwong, P. D., and Shapiro, L. (2016). Effects of Darwinian Selection and Mutability on Rate of Broadly Neutralizing Antibody Evolution during HIV-1 Infection. *PLOS Comput Biol*, 12(5):e1004940.

[Shlomchik et al., 1987] Shlomchik, M. J., Marshak-Rothstein, A., Wolfowicz, C. B., Rothstein, T. L., and Weigert, M. G. (1987). The role of clonal selection and somatic mutation in autoimmunity. *Nature*, 328(6133):805–811.

[Six et al., 2013] Six, A., Mariotti-Ferrandiz, E., Chaara, W., Magadan, S., Pham, H.-P., Lefranc, M.-P., Mora, T., Thomas-Vaslin, V., Walczak, A. M., and Boudinot, P. (2013). The past, present, and future of immune repertoire biology the rise of next-generation repertoire analysis. *T Cell Biology*, 4:413.

[Smith et al., 1996] Smith, D. S., Creadon, G., Jena, P. K., Portanova, J. P., Kotzin, B. L., and Wysocki, L. J. (1996). Di- and trinucleotide target preferences of somatic mutagenesis in normal and autoreactive B cells. *The Journal of Immunology*, 156(7):2642–2652.

[Sok et al., 2013] Sok, D., Laserson, U., Laserson, J., Liu, Y., Vigneault, F., Julien, J.-P., Briney, B., Ramos, A., Saye, K. F., Le, K., Mahan, A., Wang, S., Kardar, M., Yaari, G., Walker, L. M., Simen, B. B., St. John, E. P., Chan-Hui, P.-Y., Swiderek, K., Kleinstein, S. H., Alter, G., Seaman, M. S., Chakraborty, A. K., Koller, D., Wilson, I. A., Church, G. M., Burton, D. R., and Poignard, P. (2013). The Effects of Somatic Hypermutation on Neutralization and Binding in the PGT121 Family of Broadly Neutralizing HIV Antibodies. *PLoS Pathogens*, 9(11). 00003 PMID: 24278016.

[Sompayrac, 2016] Sompayrac, L. (2016). *How the Immune System Works*. Wiley Blackwell, Chichester, United Kingdom.

[Spencer and Dunn-Walters, 1999] Spencer, J. and Dunn-Walters, D. K. (1999). Somatic hypermutation and B-cell malignancies. *The Journal of Pathology*, 187(2):158–163.

[Spencer and Dunn-Walters, 2005] Spencer, J. and Dunn-Walters, D. K. (2005). Hypermutation at A-T Base Pairs: The A Nucleotide Replacement Spectrum Is Affected by Adjacent Nucleotides and There Is No Reverse Complementarity of Sequences Flanking Mutated A and T Nucleotides. *The Journal of Immunology*, 175(8):5170–5177.

[Stamatakis, 2014] Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

[Stern et al., 2014] Stern, J. N. H., Yaari, G., Heiden, J. A. V., Church, G., Donahue, W. F., Hintzen, R. Q., Huttner, A. J., Laman, J. D., Nagra, R. M., Nylander, A.,

Pitt, D., Ramanan, S., Siddiqui, B. A., Vigneault, F., Kleinstein, S. H., Hafler, D. A., and OConnor, K. C. (2014). B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Science Translational Medicine*, 6(248):248ra107–248ra107.

[Suchard et al., 2003] Suchard, M. A., Kitchen, C. M. R., Sinsheimer, J. S., Weiss, R. E., and Yang, Z. (2003). Hierarchical Phylogenetic Models for Analyzing Multipartite Sequence Data. *Systematic Biology*, 52(5):649–664.

[Tabibian-Keissar et al., 2008] Tabibian-Keissar, H., Zuckerman, N. S., Barak, M., Dunn-Walters, D. K., Steiman-Shimony, A., Chowers, Y., Ofek, E., Rosenblatt, K., Schiby, G., Mehr, R., and Barshack, I. (2008). B-cell clonal diversification and gut-lymph node trafficking in ulcerative colitis revealed using lineage tree analysis. *European Journal of Immunology*, 38(9):2600–2609.

[Teng and Papavasiliou, 2007] Teng, G. and Papavasiliou, F. N. (2007). Immunoglobulin Somatic Hypermutation. *Annual Review of Genetics*, 41(1):107–120.

[Thammavongsa et al., 2015] Thammavongsa, V., Kim, H. K., Missiakas, D., and Schneewind, O. (2015). Staphylococcal manipulation of host immune responses. *Nature Reviews Microbiology*, 13(9):529–543.

[The SPARTAC Trial Investigators, 2013] The SPARTAC Trial Investigators (2013). Short-Course Antiretroviral Therapy in Primary HIV Infection. *New England Journal of Medicine*, 368(3):207–217. 00000.

[To et al., 2016] To, T.-H., Jung, M., Lycett, S., and Gascuel, O. (2016). Fast Dating Using Least-Squares Criteria and Algorithms. *Systematic Biology*, 65(1):82–97.

[Todd et al., 2012] Todd, C. A., Greene, K. M., Yu, X., Ozaki, D. A., Gao, H., Huang, Y., Wang, M., Li, G., Brown, R., Wood, B., D'Souza, M. P., Gilbert, P., Montefiori, D. C., and Sarzotti-Kelsoe, M. (2012). Development and implementation of an international proficiency testing program for a neutralizing

antibody assay for HIV-1 in TZM-bl cells. *Journal of Immunological Methods*, 375(1):57–67.

[Trck et al., 2015] Trck, J., Ramasamy, M. N., Galson, J. D., Rance, R., Parkhill, J., Lunter, G., Pollard, A. J., and Kelly, D. F. (2015). Identification of Antigen-Specific B Cell Receptor Sequences Using Public Repertoire Analysis. *The Journal of Immunology*, 194(1):252–261.

[Uduman et al., 2014] Uduman, M., Shlomchik, M. J., Vigneault, F., Church, G. M., and Kleinstein, S. H. (2014). Integrating B Cell Lineage Information into Statistical Tests for Detecting Selection in Ig Sequences. *The Journal of Immunology*, 192(3):867–874. 00001 PMID: 24376267.

[Uduman et al., 2011] Uduman, M., Yaari, G., Hershberg, U., Stern, J. A., Shlomchik, M. J., and Kleinstein, S. H. (2011). Detecting selection in immunoglobulin sequences. *Nucleic Acids Research*, page gkr413.

[van Dongen et al., 2003] van Dongen, J. J. M., Langerak, A. W., Brggemann, M., Evans, P. a. S., Hummel, M., Lavender, F. L., Delabesse, E., Davi, F., Schuuring, E., Garca-Sanz, R., van Krieken, J. H. J. M., Droese, J., Gonzlez, D., Bastard, C., White, H. E., Spaargaren, M., Gonzlez, M., Parreira, A., Smith, J. L., Morgan, G. J., Kneba, M., and Macintyre, E. A. (2003). Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: Report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia*, 17(12):2257–2317.

[Victora and Nussenzweig, 2012] Victora, G. D. and Nussenzweig, M. C. (2012). Germinal Centers. *Annual Review of Immunology*, 30(1):429–457.

[Vollmers et al., 2013] Vollmers, C., Sit, R. V., Weinstein, J. A., Dekker, C. L., and Quake, S. R. (2013). Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proceedings of the National Academy of Sciences*, 110(33):13463–13468.

[Volz et al., 2013] Volz, E. M., Koelle, K., and Bedford, T. (2013). Viral Phylodynamics. *PLoS Comput Biol*, 9(3):e1002947.

[von Bdingen et al., 2012] von Bdingen, H.-C., Kuo, T. C., Sirota, M., van Belle, C. J., Apeltsin, L., Glanville, J., Cree, B. A., Gourraud, P.-A., Schwartzburg, A., Huerta, G., Telman, D., Sundar, P. D., Casey, T., Cox, D. R., and Hauser, S. L. (2012). B cell exchange across the blood-brain barrier in multiple sclerosis. *Journal of Clinical Investigation*, 122(12):4533–4543.

[Walker et al., 2009] Walker, L. M., Phogat, S. K., Chan-Hui, P.-Y., Wagner, D., Phung, P., Goss, J. L., Wrin, T., Simek, M. D., Fling, S., Mitcham, J. L., Lehrman, J. K., Priddy, F. H., Olsen, O. A., Frey, S. M., Hammond, P. W., Investigators, P. G. P., Kaminsky, S., Zamb, T., Moyle, M., Koff, W. C., Poignard, P., and Burton, D. R. (2009). Broad and Potent Neutralizing Antibodies from an African Donor Reveal a New HIV-1 Vaccine Target. *Science*, 326(5950):285–289.

[Wang et al., 2014a] Wang, C., Liu, Y., Cavanagh, M. M., Saux, S. L., Qi, Q., Roskin, K. M., Looney, T. J., Lee, J.-Y., Dixit, V., Dekker, C. L., Swan, G. E., Goronzy, J. J., and Boyd, S. D. (2014a). B-cell repertoire responses to varicella-zoster vaccination in human identical twins. *Proceedings of the National Academy of Sciences*, page 201415875.

[Wang et al., 2013] Wang, C., Liu, Y., Xu, L. T., Jackson, K. J. L., Roskin, K. M., Pham, T. D., Laserson, J., Marshall, E. L., Seo, K., Lee, J.-Y., Furman, D., Koller, D., Dekker, C. L., Davis, M. M., Fire, A. Z., and Boyd, S. D. (2013). Effects of Aging, Cytomegalovirus Infection, and EBV Infection on Human B Cell Repertoires. *The Journal of Immunology*, page 1301384. 00000 PMID: 24337376.

[Wang et al., 2014b] Wang, C., Liu, Y., Xu, L. T., Jackson, K. J. L., Roskin, K. M., Pham, T. D., Laserson, J., Marshall, E. L., Seo, K., Lee, J.-Y., Furman, D., Koller, D., Dekker, C. L., Davis, M. M., Fire, A. Z., and Boyd, S. D. (2014b). Effects of Aging, Cytomegalovirus Infection, and EBV Infection on Human B Cell Repertoires. *The Journal of Immunology*, 192(2):603–611.

[Watson et al., 2013] Watson, S. J., Welkers, M. R. A., Depledge, D. P., Coulter, E., Breuer, J. M., Jong, M. D. d., and Kellam, P. (2013). Viral population analysis and minority-variant detection using short read next-generation sequencing. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 368(1614):20120205. RNA viruses within infected individuals exist as a population of evolutionary-related variants. Owing to evolutionary change affecting the constitution of this population, the frequency and/or occurrence of individual viral variants can show marked or subtle fluctuations. Since the development of massively parallel sequencing platforms, such viral populations can now be investigated to unprecedented resolution. A critical problem with such analyses is the presence of sequencing-related errors that obscure the identification of true biological variants present at low frequency. Here, we report the development and assessment of the Quality Assessment of Short Read (QUASR) Pipeline (<http://sourceforge.net/projects/quasr>) specific for virus genome short read analysis that minimizes sequencing errors from multiple deep-sequencing platforms, and enables post-mapping analysis of the minority variants within the viral population. QUASR significantly reduces the error-related noise in deep-sequencing datasets, resulting in increased mapping accuracy and reduction of erroneous mutations. Using QUASR, we have determined influenza virus genome dynamics in sequential samples from an in vitro evolution of 2009 pandemic H1N1 (A/H1N1/09) influenza from samples sequenced on both the Roche 454 GSFLX and Illumina GAIIx platforms. Importantly, concordance between the 454 and Illumina sequencing allowed unambiguous minority-variant detection and accurate determination of virus population turnover in vitro .

[Wei et al., 2003] Wei, X., Decker, J. M., Wang, S., Hui, H., Kappes, J. C., Wu, X., Salazar-Gonzalez, J. F., Salazar, M. G., Kilby, J. M., Saag, M. S., Komarova, N. L., Nowak, M. A., Hahn, B. H., Kwong, P. D., and Shaw, G. M. (2003). Antibody neutralization and escape by HIV-1. *Nature*, 422(6929):307–312.

[Weinstein et al., 2009] Weinstein, J. A., Jiang, N., White, R. A., Fisher, D. S., and Quake, S. R. (2009). High-Throughput Sequencing of the Zebrafish Antibody Repertoire. *Science*, 324(5928):807–810. 00143 PMID: 19423829.

[Whelan and Goldman, 2001] Whelan, S. and Goldman, N. (2001). A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Molecular Biology and Evolution*, 18(5):691–699.

[Whelan and Goldman, 2004] Whelan, S. and Goldman, N. (2004). Estimating the Frequency of Events That Cause Multiple-Nucleotide Changes. *Genetics*, 167(4):2027–2043.

[Wilen et al., 2012] Wilen, C. B., Tilton, J. C., and Doms, R. W. (2012). HIV: Cell Binding and Entry. *Cold Spring Harbor Perspectives in Medicine*, 2(8).

[Williams et al., 2015] Williams, W. B., Liao, H.-X., Moody, M. A., Kepler, T. B., Alam, S. M., Gao, F., Wiehe, K., Trama, A. M., Jones, K., Zhang, R., Song, H., Marshall, D. J., Whitesides, J. F., Sawatzki, K., Hua, A., Liu, P., Tay, M. Z., Seaton, K. E., Shen, X., Foulger, A., Lloyd, K. E., Parks, R., Pollara, J., Ferrari, G., Yu, J.-S., Vandergrift, N., Montefiori, D. C., Sobieszczyk, M. E., Hammer, S., Karuna, S., Gilbert, P., Grove, D., Grunenberg, N., McElrath, M. J., Mascola, J. R., Koup, R. A., Corey, L., Nabel, G. J., Morgan, C., Churchyard, G., Maenza, J., Keefer, M., Graham, B. S., Baden, L. R., Tomaras, G. D., and Haynes, B. F. (2015). Diversion of HIV-1 vaccine-induced immunity by gp41-microbiota cross-reactive antibodies. *Science*, 349(6249):aab1253–aab1253.

[Wu et al., 2010] Wu, X., Yang, Z.-Y., Li, Y., Hogerkorp, C.-M., Schief, W. R., Seaman, M. S., Zhou, T., Schmidt, S. D., Wu, L., Xu, L., Longo, N. S., McKee, K., ODell, S., Louder, M. K., Wycuff, D. L., Feng, Y., Nason, M., Doria-Rose, N., Connors, M., Kwong, P. D., Roederer, M., Wyatt, R. T., Nabel, G. J., and Mascola, J. R. (2010). Rational Design of Envelope Identifies Broadly Neutralizing Human Monoclonal Antibodies to HIV-1. *Science*, 329(5993):856–861.

[Wu et al., 2015] Wu, X., Zhang, Z., Schramm, C., Joyce, M. G., DoKwon, Y., Zhou, T., Sheng, Z., Zhang, B., ODell, S., McKee, K., Georgiev, I., Chuang, G.-Y., Longo, N., Lynch, R., Saunders, K., Soto, C., Srivatsan, S., Yang, Y., Bailer, R., Louder, M., Mullikin, J. C., Connors, M., Kwong, P. D., Mascola, J. R., and Shapiro, L. (2015). Maturation and Diversity of the VRC01-Antibody Lineage over 15 Years of Chronic HIV-1 Infection. *Cell*, 161(3):470–485.

[Xu and Davis, 2000] Xu, J. L. and Davis, M. M. (2000). Diversity in the CDR3 Region of VH Is Sufficient for Most Antibody Specificities. *Immunity*, 13(1):37–45.

[Yaari et al., 2015] Yaari, G., Benichou, J. I. C., Vander Heiden, J. A., Kleinstein, S. H., and Louzoun, Y. (2015). The mutation patterns in B-cell immunoglobulin receptors reflect the influence of selection acting at multiple time-scales. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1676):20140242.

[Yaari and Kleinstein, 2015] Yaari, G. and Kleinstein, S. H. (2015). Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Medicine*, 7(1).

[Yaari et al., 2012] Yaari, G., Uduman, M., and Kleinstein, S. H. (2012). Quantifying selection in high-throughput Immunoglobulin sequencing data sets. *Nucleic Acids Research*, page gks457.

[Yaari et al., 2013] Yaari, G., Vander Heiden, J. A., Uduman, M., Gadala-Maria, D., Gupta, N., Stern, J. N. H., O'Connor, K. C., Hafler, D. A., Laserson, U., Vigneault, F., and Kleinstein, S. H. (2013). Models of Somatic Hypermutation Targeting and Substitution Based on Synonymous Mutations from High-Throughput Immunoglobulin Sequencing Data. *Frontiers in Immunology*, 4. 00003 PMID: 24298272.

[Yang, 1994] Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution*, 39(1):105–111.

[Yang, 2014] Yang, Z. (2014). *Molecular Evolution: A Statistical Approach*. Oxford University Press. Google-Books-ID: yILnCwAAQBAJ.

[Yang and Swanson, 2002] Yang, Z. and Swanson, W. J. (2002). Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Molecular biology and evolution*, 19(1):49–57.

[Ye et al., 2013] Ye, J., Ma, N., Madden, T. L., and Ostell, J. M. (2013). IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research*, 41(W1):W34–W40.

[Ypma, 1995] Ypma, T. (1995). Historical Development of the NewtonRaphson Method. *SIAM Review*, 37(4):531–551.

[Zhou et al., 2013] Zhou, T., Zhu, J., Wu, X., Moquin, S., Zhang, B., Acharya, P., Georgiev, I., Altae-Tran, H., Chuang, G.-Y., Joyce, M., DoKwon, Y., Longo, N., Louder, M., Luongo, T., McKee, K., Schramm, C., Skinner, J., Yang, Y., Yang, Z., Zhang, Z., Zheng, A., Bonsignori, M., Haynes, B., Scheid, J., Nussenzweig, M., Simek, M., Burton, D., Koff, W., Mullikin, J., Connors, M., Shapiro, L., Nabel, G., Mascola, J., and Kwong, P. (2013). Multidonor Analysis Reveals Structural Elements, Genetic Determinants, and Maturation Pathway for HIV-1 Neutralization by VRC01-Class Antibodies. *Immunity*, 39(2):245–258. 00030.

[Zhu et al., 2013] Zhu, J., Ofek, G., Yang, Y., Zhang, B., Louder, M. K., Lu, G., McKee, K., Pancera, M., Skinner, J., Zhang, Z., Parks, R., Eudailey, J., Lloyd, K. E., Blinn, J., Alam, S. M., Haynes, B. F., Simek, M., Burton, D. R., Koff, W. C., NISC Comparative Sequencing Program, Mullikin, J. C., Mascola, J. R., Shapiro, L., Kwong, P. D., Becker, J., Benjamin, B., Blakesley, R., Bouffard, G., Brooks, S., Coleman, H., Dekhtyar, M., Gregory, M., Guan, X., Gupta, J., Han, J., Hargrove, A., Ho, S.-l., Johnson, T., Legaspi, R., Lovett, S., Maduro, Q., Masiello, C., Maskeri, B., McDowell, J., Montemayor, C., Mullikin, J., Park, M., Riebow, N., Schandler, K., Schmidt, B., Sison, C., Stantripop, M., Thomas, J., Thomas,

P., Vemulapalli, M., and Young, A. (2013). Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proceedings of the National Academy of Sciences*, 110(16):6470–6475. 00027.