Computational Methods to Study Diversification in Pathogens, and Invertebrate and Vertebrate Immune Systems

by

Supriya Munshaw

Department of Computational Biology and Bioinformatics
Duke University

Date: _____

Approved:

_____
Thomas B. Kepler, Advisor

_____
Lindsay G. Cowell

_____
Barton F. Haynes

_____
Garnett H. Kelsoe

_____
Katharina Koelle

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computational Biology and Bioinformatics
in the Graduate School of Duke University

2010

Abstract

(Computational Biology and Bioinformatics)

Computational Methods to Study Diversification in Pathogens, and Invertebrate and Vertebrate Immune Systems

by

Supriya Munshaw

Department of Computational Biology and Bioinformatics
Duke University

Date: _____

Approved:

_____
Thomas B. Kepler, Advisor

_____
Lindsay G. Cowell

_____
Barton F. Haynes

_____
Garnett H. Kelsoe

_____
Katharina Koelle

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computational Biology and Bioinformatics
in the Graduate School of Duke University

2010

Abstract

Pathogens and host immune systems use strikingly similar methods of diversification. Mechanisms such as point mutations and recombination help pathogens escape the host immune system and similar mechanisms help the host immune system attack rapidly evolving pathogens. Understanding the interplay between pathogen and immune system evolution is crucial to effective drug and vaccine development. In this thesis we employ various computational methods to study diversification in a pathogen, an invertebrate and a vertebrate immune system.

First, we develop a technique for phylogenetic inference in the presence of recombination based on the principle of minimum description length, which assigns a cost-the description length-to each network topology given the observed sequence data. We show that the method performs well on simulated data and demonstrate its application on HIV env gene sequence data from 8 human subjects.

Next, we demonstrate via phylogenetic analysis that the evolution of repeats in an immune-related gene family in *Strongylocentrotus purpuratus* is the result of recombination and duplication and/or deletion. These results support the evidence suggesting that invertebrate immune systems are highly complex and may employ similar mechanisms for diversification as higher vertebrates.

Third, we develop a probabilistic model of the immunoglobulin (Ig) rearrangement process and a Bayesian method for estimating posterior probabilities for the comparison of multiple plausible rearrangements. We validate the software using

various datasets and in all tests, SoDA2 performed better than other available software.

Finally, we characterize the somatic population genetics of the nucleotide sequences of > 1000 recombinant Ig pairs derived from the blood of 5 acute HIV-1 infected (AHI) subjects. We found that the Ig genes from the 20 day AHI PC showed extraordinary clonal relatedness among themselves; a single clone comprised of 52 members, with observed and inferred precursor antibodies specific for HIV-1 Env gp41. Antibodies from AHI patients show a decreased CDR3H length and an increased mutation frequency when compared to influenza vaccinated individuals. The high mutation frequency is coupled with a comparatively low synonymous to non-synonymous mutation ratio in the heavy chain. Our results may suggest presence of positive antigenic selection in previously triggered non-HIV-1 memory B cells in AHI.

Taken together, the studies presented in this thesis provide methods to study diversification in pathogens, and invertebrate and vertebrate immune systems.

Contents

Abstract **iv**

List of Tables **ix**

List of Figures **xi**

Acknowledgements **xiv**

1 Introduction **1**

1.1 Diversification in HIV-1 2

1.2 Diversification in an Invertebrate immune system 4

1.3 Diversification in a Vertebrate immune system 6

1.3.1 Antigent Independent Diversity 7

1.3.2 Antigen Dependent Diversity or Affinity Maturation 8

1.4 Characterizing the diverse B cell repertoire during an acute HIV-1 infection 9

2 An Information Theoretic Method for the Treatment of Plural Ancestry in Phylogenetics **11**

2.1 Method 15

2.1.1 Inference Procedure 15

2.1.2 Validation of Procedure 22

2.2 Results 26

2.2.1 HIV-1 26

2.3 Extension to Published Model 29

2.4 Conclusions & Future Directions 32

3 A rapidly diversifying host-defense gene cluster in the purple sea urchin *Strongylocentrotus purpuratus* 37

3.1 Repeat Structure of the genes 38

3.1.1 Family A 39

3.2 Validation of presence of recombination 40

3.3 Results of Validation 43

3.4 Discussion 43

4 SoDA2: A Hidden Markov Model Approach for Identification of Immunoglobulin Rearrangements 47

4.1 Methods 49

4.1.1 Determining the Type of Ig 49

4.1.2 V and J gene pre-alignment 50

4.1.3 HMM 52

4.1.4 Algorithm 52

4.2 Validation, Results & Discussion 54

4.2.1 Simulated Datasets 54

4.2.2 Clonally Related Datasets 55

4.2.3 Sequences from Genbank 57

4.3 Conclusion 58

5 Antibody Response to HIV-1 62

5.1 Background 62

5.2 Patients 63

5.3 Analysis pipeline 64

5.4 Results 68

5.4.1 Clonal Expansion 68

5.4.2 Polyclonal Activation 73
5.4.3 Polyreactivity 76
5.4.4 Ig Isotypes and Gene Segment Usage 77
5.4.5 Complementarity Determining Region 3 80
5.4.6 Somatic Mutations 82
5.5 Conclusions 84
6 Conclusions & Future Directions **89**
A SoDA2 **92**
A.1 Calculating the total probability of the alignment 92
A.1.1 The forward algorithm 93
A.1.2 The backward algorithm 96
A.1.3 Alignment with Highest Posterior Probability 99
B Antibodyome **100**
B.1 Ig Isotype & Gene Segment Usage 100
B.2 CDR3 length 103
B.3 Somatic Mutations 103
B.4 Clones 105
Bibliography **106**
Biography **117**

List of Tables

2.1 Parameter values for rates of mutation (k_μ), duplication (k_s) and recombination (k_r) for simulations. The values in the tables are mutations per split in the simulations. 23

2.2 Analysis of the simulated datasets with parameter values, description lengths (I_T) of the networks with (I_T2) and without (I_T1) plural ancestry (PA). [1]The Gene-wise type I error rate is defined as the number of genes falsely identified as mosaic divided by the total number of non-mosaic genes. [2]The Gene-wise type II error rate is defined as the number of genes falsely identified as non-mosaic divided by the total number of mosaic genes. [3]The Alignment-wise type II error is the proportion of datasets judged to have no recombination ($I_T1 > I_T2$) but do. 24

2.3 Phlyogenetic Analysis of the C2-V5 region of the *env* gene in HIV-1 in 8 patients with description lengths (I_T) of the networks with (I_T2) and without (I_T1) PA 28

2.4 Details of putative recombinants identified in patient 2. The breakpoint for each gene is the putative position of recombination. [1]Months after seroconversion 29

2.5 Simulation results of estimating the site-varying mutation rate ρ and the branch length τ 31

4.1 Number of correct rearrangements identified by each software out of 100 sequences tested at each mutation rate 55

4.2 Number of VH, DH and JH genes (with alleles) identified at each mutation rate for the simulated sequences 56

4.3 Results from 662 sequences from Genbank, showing the performance of the 5 programs. If 2 or more programs displayed the same rearrangement (including the alleles), it was believed to be the majority rearrangement. 58

5.1 Clinical Characteristics of Individuals in Antibodyome Project 63

5.2 Name and description of columns in SQL database. 70

5.3 Simpson's Index for all patients 72

5.4 Clones in 684-6, an AHI patient 20 days post transmission 72

5.5 Percentage of HIV-1, Influenza and other specific antibodies in the various groups of patients. 76

5.6 Percentage of antigens specific for the listed antigen in the listed patient group that also react with other antigens 77

5.7 Ig Isotype in antibodies of various specificities 78

5.8 Average mutation frequency and standard error in heavy chains of antibodies of various specificities 83

5.9 Ratio of Synonymous to Non-Synonymous (dS/dN) mutations in the entire sequence, CDRs and the Framework regions of heavy, kappa and lambda chains 84

B.1 Ig Isotype in antibodies from various groups 100

B.2 VH gene segment usage by specificity 100

B.3 DH gene segment usage by specificity 101

B.4 JH gene segment usage by specificity 102

B.5 Length of CDR3 in heavy, kappa and lambda chains of antibodies of various specificities 103

B.6 Mutation Frequencies in Heavy, Kappa and Lambda chains for all patients 103

B.7 Mutation Frequencies in Kappa and Lambda chains by specificity . . 103

B.8 Number of nucleotides in the CDRs and Frameworks in the heavy and light chains of different patient groups 104

B.9 Number of synonymous and non synonymous mutations in the CDRs and Frameworks of heavy and light chains 104

B.10 Clones in DFLU07001, a patient vaccinated with a Trivalent Inactivated Seasonal Vaccine 105

List of Figures

2.1 A) An example of a starting tree with 9 species. B) Branch swapping is applied and species 6 has moved. C) An example of node absorption where the parent of species 1 and 2 is combined with the root. D) An example of node splitting or fission where the parent of species 6, 7, 8 and 9 has been split. 8 and 9 have been assigned to the new node. E) An example of a phylogenetic network where species 6 is represented as a recombinant between the parent of species 3, 4, 5 and the parent of species 7, 8, and 9. 19

2.2 A) Network generated by simulation B) Network generated by our method 25

2.3 Network generated by applying our method to the C2-V5 region of the *env* gene of HIV-1 from patient 2. 9 putative recombinants were found. 35

2.4 Mosaic sequence AF137871 from patient 2 shown with its parents 1010 and 1346. The vertical line at position 202 shows denotes the putative point of recombination such that parent the first half of the gene is believed to have descended from 1010 and the second half from 1346. 36

3.1 Dot plot showing of one of the 185/333 genes, S185-14 against itself. The gene has multiple repeat elements. 39

3.2 Phylogenetic relationship between the 4 members, A1(pink), A2(blue), A3(orange) and A4(red) of repeat family A. S185-# is the identification number for a gene. 41

3.3 Potential evolutionary history of repeat family A. The numbers in the orange circles are the number of mutations along each branch. Red circles represent recombination events. Several duplication, recombination and mutation events are used to best explain the presence and absence of repeats in these genes. 42

3.4 Histograms of Log Likelihood Ratio values for a) Elements 1 and 27 b) TCR V and J genes c) Segments H3a and H3b from histone genes 44

4.1 The basic topology of the HMM for (a) heavy chains and (b) kappa and lambda chains. The HMM starts at the last base of the invariant cysteine of all high-likelihood V segments, runs through all DH segments and through all high-likelihood JH segments till the first base of the invariant tryptophan or phenylalanine. 51

4.2 Distribution of the empirical data for (a) VH gene recombination site choice (b) n nucleotides in the VD junction (c) 5‘ DH recombination site choice (d) 3‘ DH recombination site choice (e) n nucleotides in the DJ junction (f) 5‘ JH recombination site choice. All the data is fit to negative binomial distributions with varying parameters. 53

4.3 Shows a detailed topology of the HMM with all possible transitions. Each nucleotide in the observed sequence is treated as a separate state. The transition probabilities are derived from empirical data. The star denotes the start (3rd position of invariant cysteine) of the HMM and the + denotes the end (first position of invariant tryptophan/phenylalanine) 60

4.4 (a) Top rearrangement as chosen by SoDA2 with a higher mutation frequency than the alternative (b). This is calibrated with the mutation frequency in the VH region. 60

4.5 The alignment of CDR3H of sequence by 1154693 using IGHD1-21*01 by (a) SoDA2 (b) IMGT/V-QUEST, JOINSOLVER and iHHMune (c) SoDA. Rearrangements (b) and (c) were also provided by SoDA2 at a slightly lower probability. 61

5.1 An example that shows the quality score of a sequence, a plot of position vs quality score and the alignment matrix used to trim by quality score. The default value used is 10. 69

5.2 A phylogenetic relationship between a set of 52 clonally related antibodies from patient 684-6. The lengths of the branches represent number of nucleotide changes while the number on the branch represents the number of amino acid changes. The numbers in parenthesis indicate the number of identical pairs observed. The sequences at internal nodes are numbered in square boxes. Sequences marked in red showed binding to gp41. 74

5.3 Comparison of distribution of mutations in 2 clones derived from EBV infected individuals and the 52 member clone from 684-6 75

5.4 VH gene segment usage in all five groups compared to germline gene segment usage. (*) indicate statistical significance ($p < 0.01$). 79

5.5 (a) p values for a 2-paired T-test to see a difference in mean CDR3H length between the different groups. Values significant at $p < 0.05$ are marked in red (b) Cumulative distribution functions for the number of CDR3H length of the five groups. star indicates that the two distributions have a p value of $p < 0.05$ when compared to AHI in a Kolmogorov Smirnov Test 87

5.6 (a) Cumulative distribution functions for the number of mutations in the VH region of the five groups. (b) p values for a 2-paired T-test to see a difference in mean number of mutations between the different groups. Values significant at $p < 0.01$ are marked in red. 88

B.1 DH gene segment usage in all five groups compared to germline gene segment usage. 101

B.2 JH gene segment usage in all five groups. There are only six JH segments, making germline usage equal. 102

Acknowledgements

Very sincere thanks to

Dr. Thomas B. Kepler, my advisor, for his patience, perseverance, endless knowledge, and teaching me how to do research.

Dr. Lindsay Cowell, for being a wonderful committee member, always smiling, and helping me keep calm about my prelim and defense.

Dr. Barton Haynes, for always keeping an open door despite his busy schedule and for being a great mentor.

Dr. Garnett Kelsoe, for teaching me something new and interesting every time I see him.

Dr. Katia Koelle, for her energy and enthusiasm and providing valuable insight to my research.

Members of the Kepler and Cowell labs, past and present, for always being ready to help, no matter how silly my questions were.

Members of the Haynes and Liao labs, for their patience in explaining experimental procedures, for bearing with my data formatting demands and providing me with data.

The CBB crew, for an enjoyable time at Duke.

Sean, for never giving up on keeping me sane.

My family, for believing in me and supporting me in everything I do.

1

Introduction

[1] Pathogens persist in the environment by employing various survival strategies. Two such strategies include causing infection in naive individuals and evading the immune system in already exposed individuals by constantly altering their genetic content. While the first strategy is limited to the availability of naive hosts, the second gives the pathogen an ability to generate extraordinary diversity in its antigenic epitopes. Some pathogens, especially RNA viruses such as Human Immunodeficiency Virus - 1 (HIV-1) and influenza, have highly error prone replication and packaging machinery leading to point mutations and recombination, generating tremendous diversity. Others, such as bacteria undergo exchange of genetic material to produce new strains. This provides an ongoing challenge to the host that the pathogen infects. In response, the host immune system has evolved similar molecular mechanisms to recognize the large number of such rapidly evolving pathogens. Some of these mechanisms include site-specific rearrangement, homologous recombination and somatic hypermutation. Understanding the evolutionary interplay between the host immune system and pathogens is crucial to better drug and vaccine development. In this

[1] Parts of this introductory chapter have been peer-reviewed and published [1] [2] [3]

thesis we develop computational methods and explore diversification in a pathogen (HIV-1), an invertebrate immune system (*Strongylocentrotus purpuratus*) and a vertebrate immune system (human).

1.1 Diversification in HIV-1

Retroviruses like HIV, known for their high mutation rates, are present in the host organism as mixtures of genetically diverse populations known as quasispecies. In addition to point mutations, the HIV-1 genome also undergoes genetic recombination as an intrinsic part of its normal replication cycle [4]. Recombination can occur under the forced copy choice model when reverse transcriptase encounters a break in the RNA during the formation of the negative strand DNA molecule. In this case, it makes a switch to the co-packaged strand to avoid loss of genetic information [5]. It can also occur via the DNA strand displacement and assimilation model [6]. Here, the two copackaged RNA molecules are both copied and two molecules of minus-strand DNA are synthesized. Plus-strand DNA synthesis occurs in a discontinuous manner. One short stretch of plus-strand DNA molecule is displaced by the growing point of the 5‘ DNA. The displaced DNA fragment anneals to the minus-strand DNA of the other DNA molecule. After repair by the host machinery a recombinant is formed.

It has been shown that HIV-1 undergoes approximately two to three recombination events per genome per replication cycle [7]. In cases of superinfection (where the cell is infected by more than one virion), recombinants are generated using the genetic material of multiple virions. Splenic cells in three patients have been reported to contain 6% to 18% recombinants, which is significantly higher than the 0.5% expected with a PCR-mediated error [8]. Another study involving splenic cells of two patients reported an average of 3.2 proviruses per cell [9], which would inevitably lead to the production of recombinants. Hence, recombination is useful for

the repair of defective retroviral genomes, for generating viral diversity and assisting in the spread of beneficial mutations among viral quasispecies.

Constant changes in the HIV-1 genome make it essential to study patterns of its evolution. Generation of genetic diversity aids the virus escape the host immune system, develop drug resistance as well as makes it an extremely difficult target for vaccine development. The study of virus evolution may give us a clearer understanding of drug administration, disease progression and vaccine design. Although phylogenetic tree estimation is a prevalent method of such sequence analysis, the presence of recombination in the evolutionary history of HIV-1 has made it difficult to construct accurate phylogenetic relationships between its genes using current methods. An important assumption of most phylogeny reconstruction methods is that sequences are capable of being described by a single phylogeny. This assumption is violated by sequences that undergo recombination or are a result of plural ancestry. A mosaic sequence is a result of cross-over between two or more distinct ancestors and can thus be best explained not by a single tree but by a set of correlated trees over the sequence [10]. Analyzing mosaic sequences with methods that do not take recombination into account can lead to incorrect conclusions like overestimation of branch lengths [11]. Inclusion of recombination in evolutionary analysis is thus essential for getting a better understanding of evolutionary changes in HIV-1. So, we have developed a technique for phylogenetic inference in the presence of plural ancestry based on the principle of minimum description length, which assigns a cost-the description length-to each network topology given the observed sequence data. The description length combines the cost of poor data fit and model complexity in terms of information. This device allows us to search through network topologies to minimize the total description length. By comparing the best models obtained with and without plural ancestry, one can determine whether or not recombination has played an active role in the evolution of the genes under investigation, identify

those genes that appear to be mosaic, and infer the phylogenetic network that best represents the history of the alignment. In Chapter 2, we show that the method performs well on simulated data and demonstrate its application on HIV *env* gene sequence data from 8 human subjects.

1.2 Diversification in an Invertebrate immune system

Complex diversified immune responses have been traditionally believed to be limited to higher vertebrates. However, recent evidence has demonstrated that invertebrate immune responses may also be highly diversified and are often encoded within large gene families [12] [13]. The genome of the purple sea urchin *Strongylocentrotus purpuratus* contains a number of large immune-related gene families, many with considerably more members than their vertebrate homologues [14]. The 185/333 gene family is an example of a large diverse gene family that is putatively involved in the sea urchin immune response [15]. The genes are closely linked, are flanked by dinucleotide and trinucleotide repeats, and are highly expressed in response to immunological challenge with whole bacteria, lipopolysaccharide, β-1-3-glucan and double-stranded RNA [16] [17]. Sea urchins seem to be able to discriminate among these pathogen signatures through as yet unknown mechanisms and express unique suites of 185/333 genes in response to challenge [17]. The 185/333 transcripts constitute 6.45% of a cDNA library constructed from bacterially activated coelomocytes, as opposed to 0.086% in a nonactivated library, a 75-fold increase [16].

Although the function of the 185/333 proteins remains unknown, they localize to the cell surface of a subset of the coelomocytes (immune cells) of the sea urchin and may be involved in the formation of syncytia to immobilize invading pathogens [18]. Analysis of 185/333 protein expression by two-dimensional Western blot analysis suggests that distinct suites of proteins are expressed in response to lipopolysaccharide compared to peptidoglycan, and that individual sea urchins can express > 200

unique proteins [19]. These protein data therefore support the previous observations of transcript and gene diversity[14], [15] and [18] and emphasize the putative role of the 185/333 gene family in the *S. purpuratus* immune response.

In addition to the striking increase in expression following immune challenge, the 185/333 sequences are intriguing. Alignment of the 185/333 mRNAs requires the insertion of large gaps, creating blocks of similar sequences or repeats. These repeats are variably present or absent in different mRNAs, which have been used to define specific element patterns. Analysis of 185/333 genes indicates that the variation in transcript repeat patterns is likely the result of variations in patterns encoded by many genes, rather than the result of extensive alternative splicing among a few genes. The genes have two exons. The first encodes a hydrophobic leader, and the second encodes the remainder of the open reading frame, including the variable element patterns. A variety of imperfect repeats within the coding regions have been used to align the 185/333 sequences and to define elements based on the locations of gaps within the alignment, as well as the edges of the repeats.

The 185/333 sequence diversity is extremely high. From 16 *S. purpuratus* individuals, 872 185/333 sequences (183 genes and 689 transcripts) have been analyzed, of which 475 are unique, encoding 323 proteins with 37 different element patterns [15] [17]. Sequence diversity is the result of variation in element patterns, as well as point mutations and small indels. No identical gene sequences are shared among different animals, indicating that the nucleotide diversity occurs not only within the 185/333 gene family of individual sea urchins but also within the *S. purpuratus* population.

The results presented in Chapter 3 suggest that the highly diversified 185/333 gene family is subject to frequent recombination, gene duplication, and gene deletion. Because gaps introduced into sequence alignments to define the element patterns complicate phylogenetic analysis of full-length gene sequences, the evolution of this

gene family was analyzed from the perspective of the repeats within the genes. Phylogenetic analysis suggests that the repeats have arisen as a result of intragenic repeat duplication and/or deletion, recombination, and point mutations. Incongruent phylogenetic histories of a variety of elements and analysis of the distribution of element sequences across the genes suggest that the genes undergo frequent recombination, which is likely to be a mechanism for generating diversity within the gene family. Within this framework of gene diversity, however, there is a paradox of remarkably conserved element sequences, suggesting that the divergence from the last common ancestral gene for the extant 185/333 sequences occurred relatively recently. The 185/333 gene family therefore provides an intriguing addition to the growing body of evidence suggesting that invertebrate immune systems are far more complex than previously believed.

1.3 Diversification in a Vertebrate immune system

Vertebrate immune response to HIV-1 and most other pathogens consists of a cell mediated and a humoral response. For the purposes of this project, we focused on the humoral arm of the immune response. B cells are the major players of the humoral response. B cells express immunoglobulin (Ig) molecules on their outer surface and secrete them into the extracellular space. Secreted Ig is known as antibody. Antibodies serve as effector molecules that neutralize microbes by binding to exposed antigens and targeting them to other components of the immune system, such as phagocytic cells and complement, that effect clearance. In order to account for the presence of a large number of possible pathogens, some of them constantly mutating, Ig genes generate diversity in two stages, an antigen-independent stage and an antigen-dependent stage.

1.3.1 Antigent Independent Diversity

Antigen-independent diversity is generated in the bone marrow, where B cells originate, by combinatorial rearrangement of gene segments and junctional diversity.

Combinatorial rearrangement

Each antibody molecule comprises one heavy chain protein and one light chain protein. Both the light and heavy chain genes are encoded by gene segments that are genetically rearranged during a process known as V(D)J recombination [20] [21]. Heavy chains are made up of of three gene segments Variable (V), Diversity (D) and Joining (J) where as light chains only have a V and J segment. Recombination of these gene segments into a transcribable gene is mediated by the recombination activating genes, RAG1 and RAG2. For heavy chains, the D and J gene segments recombine first followed by the recombination of the V segment to the DJ gene. For light chains, the V segment directly recombines with the J segment. In humans, there are approximately 50 known functional V segments, 27 known functional D segments, and six known functional J segments [22] all located near the long-arm telomeric end of chromosome 14 available for assembly into heavy chain genes. This allows for approximately 8100 combinations in the heavy chain alone. Humans also have two light chain loci, κ [23] and λ [24]. Only one of these loci is expressed per cell so that each antibody either has a κ light chain or a λ light chain. Humans have 44 functional Vκ, 5 Jκ, 33 Vλ and 5 Jλ genes [22] resulting in 220 possible κ chains and 165 possible λ chains. Thus this combinatorial rearrangement alone allows for greater than 3 million antibodies.

Junctional Diversity

Additional diversity at the junctions of recombination is created during the rearrangement process. First the recombination site choice may be different for the same

gene in different recombination events. Second, non-templated (n) nucleotides are sometimes added at the junction by terminal deoxynucleotidyl transferase (TdT) between adjoining gene segments [25]. TdT is is the only known polymerase capable of adding nucleotides to a DNA strand without a template [26]. The presence of n nucleotides had previously not been seen in light chains as TdT is shown to be expressed only in pro-B cells, which is where the heavy chain rearrangement takes place [27] [28] [29]. But some later studies such as one by Bridges (1998) has confirmed the presence of n nucleotides in light chains [30]. The nucleotides added by TdT become part of complementarity determining region 3 (CDR3), which is the section of the gene that encodes one of the three antigen binding loops in the resulting protein. Additionally, presence of palindromic (p) nucleotides has also been seen in these junctions [31]. Both heavy and light chain genes encode a total of three loops through their three CDRs. Together, the six loop structures in the proteins from the expressed light and heavy chain genes form the antigen binding interface for the Ig molecule. Both CDR1 and CDR2 are within the rearrangeable germline V segments for the given locus. CDR3 begins at the 3‘ end of the V segment through to the 5‘ end of the J segment, encompassing the rearranged D segment and all n-nucleotides which makes it the most diverse region in the antibody sequence.

1.3.2 Antigen Dependent Diversity or Affinity Maturation

After the Ig gene has been rearranged the B cell leaves the bone marrow, it enters the periphery where it may or may not encounter antigen. Once it comes into contact with antigen and its affinity threshold is exceeded by binding to the antigen, the B cell becomes activated. An activated B cell does two things. Firstly, it secretes antibodies, which bind to pathogens and help neutralize them, or identify them to phagocytes and other innate system defenses, allowing to eliminate them. Secondly, they proliferate and express a B-cell specific factor called activation-induced cytidine

deaminase (AID), which causes mutations in the Ig genes at a rate of up to 10^6 times the normal background rate [32]. These point mutations are usually within the CDR and may help increase affinity for antigen since the CDRs form the antigen binding interface. The cells are subsequently selected for enhanced affinity for the eliciting antigen.

It is estimated that these processes of diversification can generate approximately 10^{12} different antibodies making it challenging to correctly identify the underlying germline gene segments and sub-sequently the sequences of the complementarity determining regions (CDRs). There are various tools that have been developed to solve this problem. An in-house tool called Somatic Diversification Analysis (SoDA) was developed using a novel three dimensional alignment algorithm [33]. We have developed SoDA2, which is based on a Hidden Markov Model and used to compute the posterior probabilities of candidate rearrangements and to find those with the highest values among them. Chapter 4 explains the algorithm in more detail.

1.4 Characterizing the diverse B cell repertoire during an acute HIV-1 infection

As of 2007, an estimated 33 million people are living with HIV. In the United States alone, the total number of persons with HIV was estimated to be between 1 million and 1.1 million (MMWR, 2008). Through its amazing ability to evolve rapidly, HIV has become one of the most successful pathogens in the history of the human race. The rapidly evolving nature of HIV-1 makes it difficult for the host immune system to develop a sustainable response. The humoral response of the host needs to elicit a broadly neutralizing response, one that is able to counteract irrespective of the changes in HIV-1. However, broadly neutralizing antibodies to HIV-1 are rare. Four such antibodies have been discovered; 4e12, 2G12, 2F5 and 1b12 [34] [35]; but never found at detectable levels in any patients since then. Two more potent

antibodies have also been recently discovered from an African donor [36] but the question of why these are only seen in a select few patients still remains unanswered. Although available therapies and vaccines have managed to prolong the time to Acquired Immunodeficiency Syndrome (AIDS), none of the treatments so far have been successful in eliminating the virus. The unexpected success of the Thai vaccine trial [37] and failures of other vaccine trials suggest that it may be advisable to step back and acquire a greater understanding of the fundamentals of the humoral immune response to HIV, and to examine the genetic interplay between the host immune system and virus in more detail than has been done before now. The work presented in this final chapter is a large-scale effort to understand and characterize the effect of transmitted HIV-I on the humoral arm of the immune response during an acute infection. The goal of the project is to gain a better understanding of the humoral response during an acute HIV-1 infection by profiling the plasma cell response in acutely infected individuals. We have developed statistical and computational tools to aid the study, which are explained in detail in Chapters 4 and 5.

2

An Information Theoretic Method for the Treatment of Plural Ancestry in Phylogenetics

[1] Several distinct classes of methods for the treatment of mosaic sequences have been developed. Each of these infers different types of information relevant to genetic mosaicism. One includes methods that identify recombination without phylogenetic inference. For instance, Hudson and Kaplan (1985) [38] developed a method that provides a lower bound on the number of recombination events that occurred in the history of a collection of sequences that does not require inference of a phylogeny. Sawyer (1989) developed methods for the detection of genetic mosaicism by looking at the distribution of segments of synonymous polymorphisms, again, without the use or inference of phylogenies. Methods that generate inferred phylogenies include that of Grassly and Holmes (1997) [39] which compares the likelihood of an evolutionary model constrained to use a single phylogenetic tree for all sites in the gene to those of models in which the tree topology is allowed to vary along the length of the gene. Although this method does not produce a single phylogenetic network, the

[1] Parts of this chapter have been peer-reviewed and published [1]

multiple inferred trees can in principle be superimposed to produce such a network.

Hein (1990) [40] poses the problem within the context of dynamic programming and describes an algorithm to identify the phylogenies, varying across sites, that minimize the sum of mutation and recombination costs, allowing these costs to be arbitrarily fixed. Though this algorithm is computationally infeasible, Hein (1993) [41] also published a much simpler, fast algorithm based on a reasonable heuristic regarding the topologies that are likely to be important. These methods represent generalizations of the method of parsimony and are thus subject to the same criticisms [42]. Strimmer and Moulton (2000) [43] addressed these perceived shortcomings by applying maximum likelihood on directed acyclic graphs (DAGs) as an extension to Felsenstein‘s method for trees. Felsenstein‘s method is based on the conditional probability $P(y|x,t)$ of observing nucleotide y given parent x and intervening time t. Strimmer and Moulton adjust this definition to allow for plural ancestry by introducing priors $p1$, $p2$ for the probabilities that there are two parents, 1 or 2, respectively. The graph is then constructed using the conditionals $P(y|x_1,x_2,t) = p_1P(y|x_1,t) + p_2P(y|x_2,t)$.

In a more recent method, Jin et al (2006) [44] applied maximum likelihood methods to reticulate networks, which are obtained from trees by the addition of additional edges between edges in the original tree. Their extension is based on the decomposition of the phylogenetic network into overlapping sub-trees and uses a maximum likelihood criterion with a branch-and-bound heuristic to reconstruct the phylogenetic history of putative recombinants. The use of maximum likelihood methods in this context represents a significant step forward, but faces the challenge posed by the need to choose from among several models of differing sizes. As pointed out by the authors, adding a reticulation edge always increases the likelihood. Hypothesis testing can be applied, but the hierarchy of models thus formed is not necessarily transitive. For three models A, B, and C it is possible that A is rejected in favor of

B, B is rejected in favor of C, yet A is not rejected in favor of C. Indeed, the authors do not actually test for the presence of recombination but use a visual heuristic to choose the size of the final model.

We have developed an information-theoretic method for phylogenetic inference in the presence of plural ancestry. The method is an outgrowth of an approach we developed for the analysis of mosaicism in host defense genes [45] and is based on model selection by minimization of the description length, as we now explain.

The minimum description length (MDL) principle casts the model selection problem as that of finding the most efficient encoding of the data [46]. This approach arose from the theoretical work of Kolmogorov (1965) [47], Solomonoff (1964) [48] and Chaitin (1966) [49] on data complexity, a concept developed to provide a definition of randomness applicable to individual datastreams rather than exclusively to ensembles. The idea is to define the complexity of a datastream D as the length, in bits, of the shortest computer program that produces D as output. A random datastream is one that cannot be compressed at all-the shortest program essentially is PRINT D. The underlying intuition is that regularities allow data compression; randomness is the complete absence of regularity. Rissanen (1989) [50] extended these abstract ideas to the practical problem of statistical data analysis by placing restrictions on the kinds of computers and programs that can be considered.

The analogy Rissanen uses is that of communication: the datastream is a message to be sent after appropriate encoding. An efficient encoding takes advantage of the data‘s regularities, as Morse code encodes the frequently used letter "e" as "dot", and the rarer letter "q" as "dash dash dot dash". This strategy requires the transmission of the message in two parts. The first part contains the code key and the second contains the message itself, encoded using the key. In more familiar terms, the code key is analogous to the statistical model and the associated parameter estimates, and the "message itself" is analogous to the data residuals. The complexity of the

model and the lack of fit are thus both evaluated using information as the common currency.

Viewed from this perspective, a tree is primarily a very efficient means of encoding data, such as DNA sequences, that contain patterns of shared features. Data compression is achieved by encoding common characteristics at the root and adding more and more specific characteristics (shared by fewer individuals) as accumulating differences as the leaves are approached. It is worth pointing out in this regard a passage from Darwin (1859) [51] in which he describes classification by the "Natural System" as an artificial means for enunciating, as briefly as possible, general propositions, that is, by one sentence to give the characters common, for instance, to all mammals, by another those common to all carnivora, by another those common to the dog-genus, and then by adding a single sentence, a full description is given of each kind of dog. The ingenuity and utility of this system are indisputable. It is precisely the "data compression" properties of trees that he is referring to, and it is their extraordinary utility in this context that drove him to posit the propinquity of descent as the cause. There is an intimate relationship between code-length and probability captured by the Kraft inequalities [46], which provides a 1-1 map between coding schemes and likelihood functions. In particular, the asymptotic expansion in N, the size of the dataset, of the minimum description length has minus the log of the maximum likelihood as the leading coefficient (order N).

The method we have developed is realized in a stochastic minimization of the description length over the space of phylogenetic networks; each elementary transaction involves random modification of the network topology. We have implemented the method in software, validated its performance on simulated datasets, and demonstrated its application on a collection of genomic sequences of the HIV-1 envelope protein isolated longitudinally from each of eight subjects infected with HIV [52].

2.1 Method

2.1.1 Inference Procedure

The number of phylogenetic tree networks grows superexponentially with the size of the data [53], and the number of phylogenetic networks grows more rapidly still, making exhaustive consideration of all topologies infeasible. We have therefore designed our procedure as a stochastic optimization on the space of phylogenetic networks. The procedure uses simulated annealing [54] to perform the minimization of the minimum description length of the data plus the network as the objective function. The process is carried out in two stages. During stage 1, the tree-like topology is preserved by restricting the elementary operations that are allowed. This stage is intended to find a suitable starting point for the second stage, in which an enlarged set of elementary operations is applied, including some that do not preserve the tree topology. After each operation, the description length is evaluated and used to determine whether the result of the operation will be preserved or discarded. The procedure is explained below in greater detail.

The Scoring function - Minimum Description Length (MDL)

We have cast the determination of the phylogenetic history of a set of sequences explicitly as a problem of model selection in which we seek the model that minimizes the total information required to encode the data or in other words, the description length of the data given a model. A phylogenetic tree can be viewed as a hierarchical data structure that enables an efficient encoding of a set of similar DNA sequences through elimination of redundancies. Consider the encoding of two similar genes, each of length L. One could use a naive model and encode the two genes independently, at an information cost of $4L$ bits (1 nucleotide requires 2 bits to encode). Or one could encode the consensus sequence (costing $2L$ bits) and then

encode the changes required to recover each of the two genes from the consensus. Roughly speaking, each mutation requires $\log_2 3$ bits to specify the class of mutation and $\log_2 L$ bits to specify the nucleotide position at which it occurs. As long as the number of mutations, m, required for this encoding is small enough to satisfy $m(\log_2 3 + \log_2 L) < 2L$, the tree model is more efficient than the naive model. For larger gene sets, the process is similar, though the models are more complex. Each gene must specify its parent and those changes to the parent that produce the gene. The set of pointers from genes to their parents is equivalent to the topology of the tree and the encoding of the changes from parent to child is accomplished through the use of a model for mutations. We extend this basic coding scheme by allowing any node in the network to have two ancestors and thus be treated as a mosaic sequence. In this case, we need to specify both ancestors, the breakpoint where the mosaic switches from one to the other, and the mutations between this parental hybrid and the gene of interest. The point is to account accurately for the cost of allowing plural ancestry. For each sequence alignment we consider, we will use and compare two different models. One, designated M_1, prohibits plural ancestry, and the other, M_2, allows it. Although the method can be used with any mutation model, we use a relatively simple model in what follows here. This model allows for different mutation rates for each of the four classes of pairwise relationships between nucleotides: identity, transition, transversion 1 (G,A $\mapsto$ C,T), and transversion 2 (G,A $\mapsto$ T,C). We further assume that these rates are uniform across positions in the gene and over all branches. Finally, we assume equal branch-lengths throughout the tree. Under these conditions, the mutations are distributed according to a multinomial model. The probability density function is defined as

$$P(x|\vec{p}) = p_1^{x_1} p_2^{x_2} p_3^{x_3} p_4^{x_4} \tag{2.1}$$

where x is the unknown parameter vector of length four representing the relationships between nucleotides mentioned above (the sum of the components of this vector equals the total number of nucleotides in the network) and p contains the proportions of each relationship. Taking a uniform prior on the parameter vector, we integrate it out, and obtain the total information required to encode the mutations given this model:

$$I_\mu = \log_2 \int dp P(x|\vec{p}) P(\vec{p}) = \log_2 \binom{[(N-1)L+3]}{3} + \log_2 \binom{(N-1)L}{\vec{x}} \qquad (2.2)$$

where N is the total number of nodes in a given network (including the root) and L is the common length of the sequences. The arguments of the logarithms are the binomial and multinomial coefficients, respectively. Eq.(2) has a useful direct interpretation in terms of coding: The first term is the information required to specify the counts of the components of vector x given the total number, $(N-1)L$, of nucleotides in a network (the root is excluded since there are no mutations in the root). The second term accounts for the number of ways these mutations can be assigned to the NL nucleotides of the dataset given the counts in vector x. We ignore a term $2L$ that represents the amount of information required to encode the root. Since all networks in question will have a root with the same length, this term will not make any difference in the final description length. In addition to the information required to encode the mutations, we need to encode the ancestry as well. Under a singular ancestry model, each node must specify a parent. If N_I is the number of internal nodes and N_O is the number of observed or leaf nodes, the ancestor information is given by equation 2.3.

$$I_A = \log_2 (N_O - 1) + N_O \log_2 N_I + (N_I - 1) \log_2 (N_I - 1) \qquad (2.3)$$

where the first term represents the amount of information required to specify the number of internal nodes and hence all possible parents. The second term represents the amount of information required to specify a parent for an observed or leaf node and since an internal node cannot be its own parent, the third term represents the information to encode an ancestor for an internal node. Under the plural ancestry model, we must first specify the number of recombinants in a given network and the number of ways of distributing these recombinants. Each recombinant will then have a secondary parent from one of the N_I internal nodes (a recombinant cannot have the same parent twice, so we have $N_I - 1$ choices) and the location of the cross-over point, where the gene switches from similarity to parent 1 to similarity to parent 2. Since this switch cannot happen at the last nucleotide, specifying the recombination point takes $\log_2(L-1)$ bits. If there are N_R nodes with dual ancestry, the corresponding information cost is

$$I_R = \log_2(N-1) + \log_2 \binom{N-1}{N_R} + N_R \log_2(N_I - 1) + N_R \log_2(L-1) \qquad (2.4)$$

Equation 2.4 might be a slight overestimation (of order 1) in that internal nodes are allowed to be recombinants but they cannot be their own parents. Here we do not distinguish between observed node recombinants and internal node recombinants. The total information cost or the description length of the encoding under model M_2 is simply the sum of the three terms, $I_T = I_\mu + I_A + I_R$. The third term in the equation I_R is omitted when encoding under M_1. The optimal model is the one that minimizes the total information. The tradeoff of recombination is achieved by reducing the number of mutations, and thus decreasing I_μ by more than I_R.

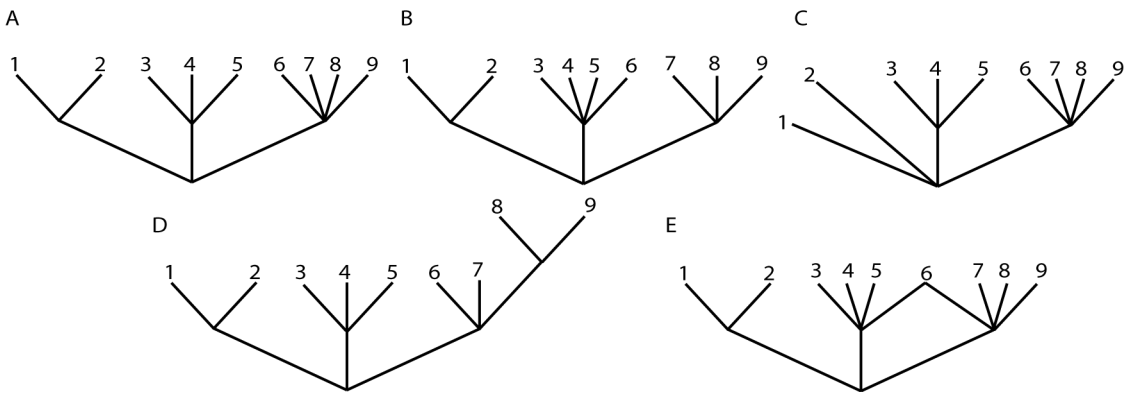


FIGURE 2.1: A) An example of a starting tree with 9 species. B) Branch swapping is applied and species 6 has moved. C) An example of node absorption where the parent of species 1 and 2 is combined with the root. D) An example of node splitting or fission where the parent of species 6, 7, 8 and 9 has been split. 8 and 9 have been assigned to the new node. E) An example of a phylogenetic network where species 6 is represented as a recombinant between the parent of species 3, 4, 5 and the parent of species 7, 8, and 9.

Simulated Annealing

As mentioned above, we start with a random initial topology. Figure 2.1A demonstrates an example of a topology with 9 species. The elementary operations we use allow any finite topology to be converted to any other finite topology in a finite number of steps. We then randomly assign our data to this topology and calculate the resulting description length. Note that each node in our network represents a DNA sequence. The leaf nodes represent the input sequences while the sequences at the interior nodes, i.e. nodes with children, are determined by the demand for minimizing the description length and hence contain the appropriate consensuses. We then allow the repeated application of the following elementary operations and their inverses: change parent, split a node, combine two nodes or add a second parent. There are two main steps in the program, each of which is described in detail below.

Stage 1

The operation to be applied is chosen from among the three possibilities with uniform probability: branch swapping, node absorption, and node fission.

- *B*ranch swapping - We choose two nodes at random: an interior node (which will become the new parent) and a child node. We break the link between the chosen child and its current parent and link the child to the new parent. An example of branch swapping is shown in 2.1B, where the branch between species 6 and its parent in Figure 2.1A has been swapped giving species 6 a new parent.

- *N*ode absorption - We randomly choose an interior node and merge it with its parent. In other words, the chosen node is deleted from the tree and all its children are made direct progeny of its parent. An example of node absorption is shown in Figure 2.1C, where the parent of species 1 and 2 in Figure 2.1A has been merged with the root.

- *N*ode fission - We randomly choose an interior node and create a new node that is assigned as a child of the chosen node. The children of the chosen node are then distributed between the new child and itself minimizing the local contribution to the description length using a simple 2-means algorithm (MacQueen, 1967). An example of node fission is shown in Figure 2.1D, where the parent of species 6, 7, 8 and 9 is split such that 8 and 9 have a separate parent.

The description length of the resulting tree is calculated at the end of each iteration within stage 1. The move is accepted if the description length of the resulting network is less than the existing one. If it is not, we use simulated annealing to

choose whether to accept or reject the move. This is necessary in order to avoid local minima. Step 1 runs till the completion of simulated annealing.

Stage 2

Stage 2 differs from stage 1 by replacing branch-swapping with adoption. Adoption proceeds by choosing a node at random from all non-root nodes, designating that the child and choosing new parent node from among all interior nodes.

- *I*f the chosen child node has one parent, add a branch to the new parent while maintaining the branch to the old parent. Determine the recombination point between the two parents that minimizes the local description length by minimizing the number of mutations between the parent hybrid and the sequence at the child node. An example of a move of this kind is shown in Figure 2.1E, where 6 now has two parents.

- *I*f the chosen child already has two parents, we consider five operations and choose the one that produces the smallest description length.

  - *R*epresent the child as a recombinant between the current primary parent and the new parent.
  - *R*epresent the child as a recombinant between the current secondary parent and the new parent.
  - *B*reak the link with both existing parents and represent the child with having only the new parent as a parent.
  - *B*reak the link with the secondary parent and represent the child by the current primary parent only. This move is a check and is done to allow reversibility.
  - *B*reak the link with the primary parent and represent the child by the current secondary parent only.

As in step 1, the description length of the tree resulting from the move is calculated and stage 2 is also iterated till the completion of simulated annealing. At the end of both steps we have a network that minimizes the description length of the data.

2.1.2 Validation of Procedure

We generated simulated datasets using a Markov Process model to grow a phylogenetic network with point mutations, recombinations and duplications and obtain a set of sequences related as if through this network. At any given time in the process, each node is considered independently and can acquire a single mutation, split into two sibling nodes, become recombination receptive, or do nothing. The rates of these processes are given by k_μ, k_s, and k_r, respectively. Mutations are random and incorporated at random locations along the gene. When a node becomes recombination receptive, a partner node is chosen from among all extant internal nodes and a recombination cross-over point is chosen at random along the common lengths of the parents. A new node with the resulting mosaic sequence is added as a shared child to the two parent node. For each of the combinations of $\frac{k_s}{k_r}$ and $\frac{k_\mu}{k_r}$ displayed in Table 2.1.2, we generated 12 replicate datasets and analyzed them using the methods described above (Table 2.1.2).

In order to determine whether the procedure not only detects the presence of recombination, but also accurately infers their evolutionary histories, we compared the simulated phylogenetic networks for each dataset to networks generated by our program. Figure 2.2 provides one such visual comparison of the simulated network and the network constructed using our procedure, showing that the two are very similar. Nine out of 12 mosaic sequences were identified as such with the appropriate ancestors and cross-over points. The mosaic sequence shown in the red dashed line could not be identified because the cross-over point was at position 297 of 300 and

Table 2.1: Parameter values for rates of mutation (k_μ), duplication (k_s) and recombination (k_r) for simulations. The values in the tables are mutations per split in the simulations.

$\frac{k_\mu}{k_r}$ \ $\frac{k_s}{k_r}$	4	8	16
16	4	2	1
32	8	4	2
64	16	8	4
128	32	16	8
256	64	32	16
512	128	64	32
1024	256	128	64
2048	512	256	126
4096	1024	512	256

there was too little information transferred from the secondary parent for reliable identification. The mosaic sequence shown in the blue dashed line was not identified because one of its ancestors lost all its non-recombinant descendants from the simulation. Hence, its mosaic descendants could not be identified as such. Note that the inferred network is not expected to be identical to the true network since the method generates multifurcating networks when there is insufficient information to resolve the pattern of bifurcations within any multifurcating node, a condition that is the norm for datasets of this size and complexity. For example, given 3 sequences AGA, AAC and TAC, any combination of these sequences would make an equally feasible bifurcating tree. In this case, our program simply chooses a multifurcating tree since there is no one best resolution. To estimate the alignment-wise type I error rate (for a whole alignment), we generated 15 simulated datasets with no recombination for each of the following mutation to duplication rate ratios: 1, 2, 4, 8, 16, 32, 64, 128, 256, 512 and 1024, resulting in a total of 165 total datasets. There was an average of 46.9 genes per dataset. Alignment-wise false positives as defined above were counted in 11 of the 165 datasets, 9 of which had identified a single gene as mosaic

Table 2.2: Analysis of the simulated datasets with parameter values, description lengths (I_T) of the networks with (I_T2) and without (I_T1) plural ancestry (PA). [1]The Gene-wise type I error rate is defined as the number of genes falsely identified as mosaic divided by the total number of non-mosaic genes. [2]The Gene-wise type II error rate is defined as the number of genes falsely identified as non-mosaic divided by the total number of mosaic genes. [3]The Alignment-wise type II error is the proportion of datasets judged to have no recombination ($I_T1 > I_T2$) but do.

$\frac{k_s}{k_r}$	$\frac{k_\mu}{k_r}$	$\frac{mutations}{split}$	I_T1 (without PA)	I_T2 (with PA)	Gene-Wise Type I error rate[1]	Gene-Wise Type II error rate[2]	Alignment-Wise Type II error rate[3]
16	16	1	1480	1488	0.0	1	0.92
8	16	2	2258	2223	0.0	0.76	0.58
16	32	2	1964	1955	0.0	0.79	0.56
4	16	4	3933	3412	0.003	0.36	0.0
8	32	4	3306	3073	0.002	0.36	0.33
16	64	4	3312	3105	0	0.30	0
4	32	8	5823	5685	0.0	0.28	0.16
8	64	8	5173	5085	0.005	0.20	0.0
16	128	8	4546	4506	0.004	0.18	0.0
4	64	16	9330	9122	0.009	0.18	0.0
8	128	16	9154	8754	0.009	0.49	0.25
16	256	16	7363	7324	0	0.2	0.2
4	128	32	13276	12933	0.002	0.33	0.0
8	256	32	13340	13105	0.003	0.14	0.0
16	512	32	12225	12139	0	0.14	0.0
4	256	64	21061	20903	0.006	0.40	0.0
8	512	64	19952	19602	0.009	0.31	0.08
16	1024	64	19673	19628	0.002	0.32	0.2
4	512	128	25946	25873	0.002	0.60	0.17
8	1024	128	26664	26644	0	0.72	0.50
16	2048	128	26067	26064	0.0	0.67	0.57
4	1024	256	29786	29791	0.0	1	1
8	2048	256	29830	29836	0.0	1	1
16	4096	256	29755	29760	0.0	1	1
4	2048	512	29992	29997	0.0	1	1
8	4096	512	30168	30173	0.0	1	1
4	4096	1024	30232	30238	0.0	1	1

while the other 2 had 2 and 3 mosaic genes respectively. These mosaic genes were generated in datasets with mutation to duplication rate ratio of 4 to 64. The higher mutation rates led to extremely diverse datasets resulting in poor resolution. The overall alignment-wise type I error rate is thus 6.67%. The average gene-wise error rate within each of the family of datasets where alignment-wise false positives were found is less than 0.4%. Note that where I_T2 is smaller than I_T1 and there are no putatively mosaic genes, we can say with certainty that the annealing in stage 1 was incomplete. This shortcoming can be reduced by improving the cooling schedule.

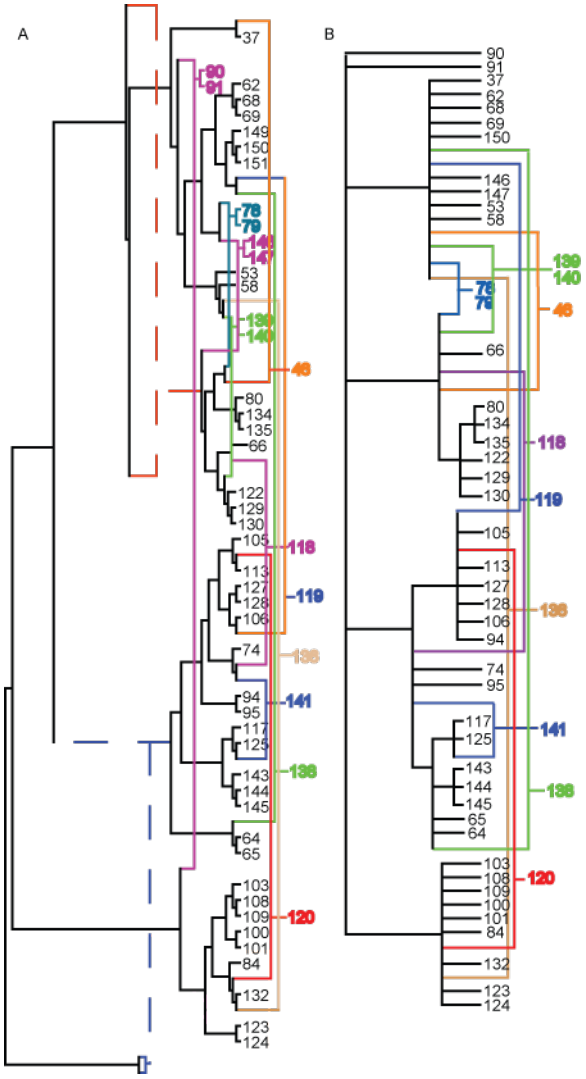


FIGURE 2.2: A) Network generated by simulation B) Network generated by our method

Observations

We made the following observations: 1) The gene-wise type I error rate, which is defined as the number of genes falsely identified as mosaic divided by the total number of non-mosaic genes, is fairly low at all mutation rates. This number remains less than or equal to 10% for all parameter values used in testing. 2) The gene-wise type II error rate, defined as the number of genes falsely identified as non-mosaic

divided by the total number of mosaic genes, is reasonable for a range of intermediate relative mutation rates (4-64 mutations per split) but becomes intolerably large for more extreme values at very low and very high rates. Failure of this type is inevitable: mutations both provide and destroy the information necessary to identify recombination. Recombination between identical or nearly identical genes cannot be detected by analysis of the resulting sequence, yet very high rates corrupt the signal. 3) Similarly, the alignment-wise type II error, the proportion of datasets judged to have no recombination ($I_T1 > I_T2$) when they in fact do, is high when the mutations per branch are extreme. 4) Finally, the overall alignment-wise type I error rate, the proportion of datasets that are falsely identified to have recombination is 0.007.

2.2 Results

2.2.1 HIV-1

Shankarappa et al (1999) [52] studied the divergence and diversity of the C2-V5 region of the HIV-1 *env* gene in nine homosexual men with moderate or slow disease progression. They obtained gene sequences at an average of 12 time points per person, covering 6 to 12 years of infection. These datasets were used because they provided the opportunity to study the patterns of HIV-1 evolution within single individuals and 8 of these datasets were available on NCBI (accession numbers AF137629 to AF137765 (patient 1), AF137767 to AF137897 (patient 2), AF137898 to AF138003 (patient 3), AF138004 to AF138163 (patient 5), AF138166 to AF138263 (patient 6), AF138305 to AF138411 (patient 7), AF138531 to AF138643 (patient 9) and AF138652 to AF138703 (patient 11)). We aligned the sequences using the CLUSTALW application in BioEdit. Because our program is not yet meant to handle extensive insertions and deletions within alignments, we discarded the ends of alignments where such indels were present. Sequences from one patient had indels throughout the length of the gene, and was therefore excluded from the study. The

sequence lengths given in the results represent only the part of the alignment used for the analysis. We ran each dataset 5 times with different random initial conditions and accepted as mosaic genes those that were identified as mosaic at least 3 out of 5 times. Table 2.2.1 shows the total number of sequences and mosaic genes found in each of the datasets as well as the average description lengths for each of the datasets. Comparison of the description lengths at stages 1 and 2 (with and without plural ancestry) suggests the presence of recombination in all cases but patient 11. Patient 11 had no putative recombinants at all. Each of the other patients had multiple mosaic genes and substantially smaller description length in stage 2. The mutations per split for the datasets ranged from 3.9-14.3. This range was covered in our simulations; the results of the simulation can be taken as relevant to the analysis of the HIV datasets. 2.3 shows the phylogenetic network of the sequences extracted from patient 2. This particular set of sequences was shown to have approximately 5.6 mutations per split and 9 putative mosaic genes. According to our simulations, the type II error for 5.6 mutations per split could be between 0.18 and 0.36 (between 4 and 8 mutations per split). Hence, we can be confident that we have identified approximately at least 64% of the recombinants. The type I error rate in this case is between 0.0 and 0.005, which is fairly low. Figure 2.4 shows an example of a mosaic sequence and its inferred parents. This sequence (AF137871) was isolated 126 months after seroconversion from patient 2. Table 2.2.1 shows that putative recombinants in patient 2 were amplified starting 30 months after seroconversion.

Further evidence that these genes really are mosaic comes from an analysis of the times at which they were sampled. Shankarappa and colleagues sequenced virus samples at each of several times from each individual-typically from 3 month post-seroconveresion to seven or more years (the latest sampling time was more than 11 years). We expect to find genes arising from viral recombination to appear later in the infection, since detectable recombination requires superinfection of target cells

Table 2.3: Phlyogenetic Analysis of the C2-V5 region of the *env* gene in HIV-1 in 8 patients with description lengths (I_T) of the networks with (I_T2) and without (I_T1) PA

Patient	Total Sequences	Length	Mutations/ split	Putative recombinants	p	I_T1 (without PA)	I_T2 (with PA)
1	133	405	5.2	8	0.12	13701	6523
2	129	381	5.6	9	0.046	7383	6448
3	106	585	12.8	0	NA	8006	8756
5	159	578	6.6	6	1	8458	7990
6	98	627	3.9	8	0.06	15229	8282
7	107	609	14.3	24	0.07	19366	10177
9	113	296	5.2	13	0.019	11918	7055
11	52	387	5.5	0	NA	3269	3274

with genetically distinguishable viruses, the chances of which clearly increase over time. We therefore analyzed the sampling times for the sets of genes identified as mosaic by determining, for each patient, the earliest time giving rise to a sample containing a mosaic gene. For the i^{th} patient, denote that sampling time t_i. Take the null hypothesis to be that the genes identified as mosaic are chosen at random from among all genes sampled. Under this null, the probability π that the earliest sampling time in the ith patient sample is t_i or later is

$$p_i = \frac{(N_i - k_i)! n_i(t_i)!}{N_i!(n_i(t_i) - k_i)!} \tag{2.5}$$

where N_i is the total number of genes sampled, k_i is the size of the subsample of putatively mosaic genes, and $n_i(t_i)$ is the number of genes in the complete sample with sampling time t_i or later. The probabilities computed for the data sets are shown in Table 2.2.1. Mosaic genes are first detected in patients at an average of 37 months after seroconversion. On analyzing the sampling times, we found that the mosaic sequences for patients 2, 3, 7 and 9 appeared significantly late. We also found that patient 11, the slowest progressor in the study, showed no presence of recombination.

Table 2.4: Details of putative recombinants identified in patient 2. The breakpoint for each gene is the putative position of recombination. [1]Months after seroconversion

Recombinant	Sampling time[1]	Breakpoint	Mutations saved by PA
AF137802	30	265	3
AF137820	51	271	5
AF137823	51	271	3
AF137824	61	213	5
AF137836	73	124	3
AF137864	103	259	3
AF137871	126	202	6
AF137892	68	158	4
AF137897	68	180	5

2.3 Extension to Published Model

MDL methods, like likelihood-based methods, provide flexibility, allowing the use of more complex evolutionary models. We have here used a simple model for evolution, because doing so gives us access to a simple closed form for the description length for any network and any data and thus greatly facilitates numerical minimization, but this simplification does not come without costs of its own. For example, Crandall et al (1999) examined sequence data from eight patients, and found several positions at which parallel mutations related to drug resistance became prevalent. Had they been isolated from a single individual, such homoplasies would erroneously be taken as contributing evidence to the hypothesis of plural ancestry. A model that accounts for mutation rate heterogeneity among positions and/or selection would be required to distinguish these cases. Similarly, a model that allows for varying branch lengths will also be useful. We started constructing such a model only to find that for the case of HIV-1, varying branch lengths is not as important as site heterogeneity. The model is explained below.

Let the length of the i^{th} branch in given topology be denoted τ_i, the mutation rate at the j^{th} site of a given sequence be ρ_j and the overall mutation rate for the

tree is μ. Then the probability of a mutation at the j^{th} site along the i^{th} branch under a simple Jukes-Cantor model is

$$p_{ij} = \frac{3}{4}(1 - exp(-\frac{4}{3}\tau_i\rho_j\mu)) \tag{2.6}$$

Since μ is independent of branch and position, lets assume that $\rho_j = \rho_j\mu$. The likelihood of a mutation at the i^{th} branch and j^{th} position is given by the expression below:

$$\log L = \sum_{ij}[X_{ij}\log p_{ij} + (1 - X_{ij})\log 1 - p_{ij}] \tag{2.7}$$

where X_{ij} equals 0 if there is no mutation at the i^{th} branch and j^{th} position and 1 if there is a mutation. In order to solve for τ_i, the partial derivative of the log likelihood gives us

$$\frac{\partial}{\partial\tau_i}\log L = \sum_j[\frac{X_{ij}}{p_{ij}} - \frac{1 - X_{ij}}{1 - p_{ij}}](1 - \frac{4}{3}p_{ij})\rho_j = \sum_j[\frac{X_{ij} - p_{ij}}{p_{ij}(1 - p_{ij})}](1 - \frac{4}{3}p_{ij}) \tag{2.8}$$

A series expansion yields

$$X_{i.} - \tau_i(p_{.} - \frac{1}{3}\sum_j \rho_j X_{ij}) - \frac{1}{3}\tau_i^2\sum_j \rho_j^2(1 + \frac{5}{9}X_{ij}) = 0 \tag{2.9}$$

This quadratic equation can be solved and the unique positive root taken. A similar expression can be derived for ρ_j, and the two can be solved iteratively to find the overall solution.

In order to test that our model results in accurate branch lengths and site-varying mutation rates, we created various simulations. For each of the simulations, we first generated a topology with values for branch lengths and mutation rates for each site. This topology remained constant within the simulation. The branch lengths and mutation rates were either constant or generated from an exponential and gamma

Table 2.5: Simulation results of estimating the site-varying mutation rate ρ and the branch length τ

Simulation	μ, number of sequences (n)	ρ	τ	% of expected ρ lies in 95% percentile of observed	% of expected τ lies in 95% percentile of observed	Average percentile rank (ρ, τ)
1	0.1, n=100	$\rho_{1\ldots L} = 1$	$\tau_{1\ldots N} = 1$	100	100	0.55, 0.49
2	0.1, n=100	$\rho_{1\ldots \frac{L}{2}}$=0.5, $\rho_{\frac{L}{2}\ldots L} = 1$	$\tau_{1\ldots N} = 1$	100	100	0.51, 0.5
3	0.1, n=100	$\rho_{1\ldots \frac{L}{3}} = 0.25$, $\rho_{\frac{L}{3}\ldots \frac{L*2}{3}} = 0.5$, $\rho_{\frac{L*2}{3}\ldots L} = 1$	$\tau_{1\ldots N} = 1$	100	100	0.54, 0.51
4	0.1, n=100	$\rho_{1\ldots \frac{L}{2}}$=0.5, $\rho_{\frac{L}{2}\ldots L} = 1$ ($\rho_j = 0.05$, $\rho_{j+1} = 0.1$)	$\tau_{1\ldots N} = 1$	100	100	0.51, 0.5
5	0.1, n=100	$\rho_{1\ldots L} \approx$ Uniform(0,1)	$\tau_{1\ldots N} = 1$	99	100	0.53, 0.48
6	0.1, n=100	$\rho_{1\ldots L} \approx$ Uniform(0,1)	$\tau_{1\ldots \frac{N}{2}} = 0.5$, $\tau_{\frac{N}{2}\ldots N} = 1.5$	99	92	0.59,0.52
7	0.1, n=100	$\rho_{1\ldots L} \approx \Gamma(1,1)$	$\tau_{1\ldots N} \approx$ Exp(1)	98	72	0.52, 0.44
8	0.05, n=100	$\rho_{1\ldots L} \approx \Gamma(1,1)$	$\tau_{1\ldots N} = 1$	100	100	0.57, 0.51
9	0.05, n=100	$\rho_{1\ldots L} \approx \Gamma(1,1)$	$\tau_{1\ldots \frac{N}{2}} = 0.5$, $\tau_{\frac{N}{2}\ldots N} = 1.5$	98	92	0.6, 0.53
10	0.05, n=100	$\rho_{1\ldots L} \approx \Gamma(1,1)$	$\tau_{1\ldots N} \approx$ Exp(1)	99	92	0.68, 0.47
11	0.05, n=10,000	$\rho_{1\ldots L} \approx \Gamma(1,1)$	$\tau_{1\ldots N} \approx$ Exp(1)	97	34	0.62, 0.5
13	0.1, n=100	$\rho_{1\ldots L} \approx \Gamma(1,1)$	$\tau_{1\ldots N} \approx$ Exp(1)	99	46	0.68, 0.55
14	0.05, n=10,000	$\rho_{1\ldots L} \approx \Gamma(1,1)$	$\tau_{1\ldots N} \approx$ Exp(1)	99	29	0.66,0.55
15	0.1, n=10000	$\rho_{1\ldots L} \approx \Gamma(1,1)$	$\tau_{1\ldots N} \approx$ Exp(1)	92	32	0.54,0.49

distribution respectively. For a given topology, we generated n datasets or different sets of sequences. In our case, n is either 100 or 10000. We then used the external sequences in these datasets and the topology as an input to the iterative estimator. The results are shown in table 2.5. The average percentile rank represents the average number of times the observed values were smaller than the expected values.

Our procedure estimates the site-varying mutation rates 98.6% of the time i.e. the estimated ρ_j falls in the 95 percentile of the observed values 98.6% of the time. However, branch length estimation only gives good results when branch lengths are assumed equal. This is due to the following reason. Given two siblings, one with a longer branch length and another with a short branch length, it becomes impossible to determine the exact branch length without knowledge of the ancestor. In determining phylogenies, ancestral sequences are not seen and can only be estimated by making consensus. In order to best estimate the ancestor of two sequences, the consensus will randomly choose the ancestor at each position without any information about the branch length. We found that in our large sample sizes, we have 3 equal subsets in our estimated values of τ_i - one where τ_i is overestimated, another where

it is underestimated and a third where the expected value lies within 95 percentile of the observed. The τ_i in each one of these subsets was not significantly consistent i.e. our procedure did not under or over-estimate a particular branch length every time.

For the purposes of our research of studying recombination in HIV-1, it is more important to include site-varying mutation rates than estimating branch lengths. The assumption of equal branch length will not worsen our results because we assume that the entire pool of viruses at the beginning of the infection has had the same amount of time to mutate. This assumption is violated when you have a subsequent infection event for the same patient. The next step is to simplify this model, assume equal branch lengths and continue to use our estimator ρ for site-varying mutation rates.

2.4 Conclusions & Future Directions

Our method depends entirely on the use of a single well-behaved cost function on phylogenetic networks given the observed sequence data so that the power of numerical minimization can be utilized. Methods based on pairwise hypothesis testing Frequentist methods such as the likelihood ratio test do not provide a cost function of this kind. Our choice of cost function, the MDL, is intended to balance the cost of lack of model fit to the data against the cost of model complexity in a principled and natural way, by supplying a common currency-information-for both. Consistency is essential in this context since the complexity of the underlying network changes with the number of putatively mosaic genes. Other methods, including those of Hein (1990), allow the cost of recombinations relative to mutations to be fixed arbitrarily but do not determine the appropriate value for this ratio. MDL naturally sets this ratio adaptively: the total cost of mutations is not simply proportional to the number of mutations; the information cost of a single mutation decreases as the total number

of mutations increases. Similarly, the cost of a recombination event decreases as the total number of such events increases and the cost of either one depends on the total number of nucleotides in the dataset.

We ran several of our simulated datasets using Hein's web-based recombination tool Recpars (http://www.daimi.au.dk/~compbio/recpars/recpars.html). Under its default values, we ran 10 of our simulated datasets and observed a gene-wise type I error of 2% and type II error of 43%. The same datasets with our program had a 0.2% gene-wise type I error and 28% type II error. Out of 75 negative controls (datasets with no recombinants), we found that Recpars erroneously identified 32 datasets (42%) as having recombination events, compared to the 9% observed using our method. Moreover, the web version of Recpars failed to give us any results for datasets with mutation to duplication ratios of 32 and higher.

Furthermore, the stochastic search methods we use are not particularly sophisticated. Substantially improved performance could likely be achieved with greater attention paid to the numerical optimization process. The computation time depends almost entirely on the stochastic search method. Using our current search parameters, we calculated the computation time for 8 datasets ranging from 20 sequences to 100 sequences of length 300 with an average of 8 mutations/split (the mean mutations per branch for the HIV dataset). The largest dataset of 100 sequences ran for 115 minutes (real user time on a 64 bit machine, 2.19GHz processor, 4GB RAM) while the smallest dataset ran for 8 minutes. The average time was 48 minutes. A better optimization scheme with the use of Markov Chain Monte Carlo methods could lower the computation time significantly. Models allowing differing branch lengths and heterogeneous mutation rates across alignment positions will require substantially greater computation; good approximation schemes will certainly prove helpful in this regard.

In spite of the oversimplifications and relatively large computational effort re-

quired, we have shown that the method as described works well, with tolerable error rates on simulated data and biologically plausible results on clinical data from HIV infected patients.

In the next few chapters, we move on to studying recombination as a diversification mechanism in the host immune response. We begin by looking at an invertebrate, *Strongylocentrotus purpuratus*, an organism that lacks an adaptive immune system.

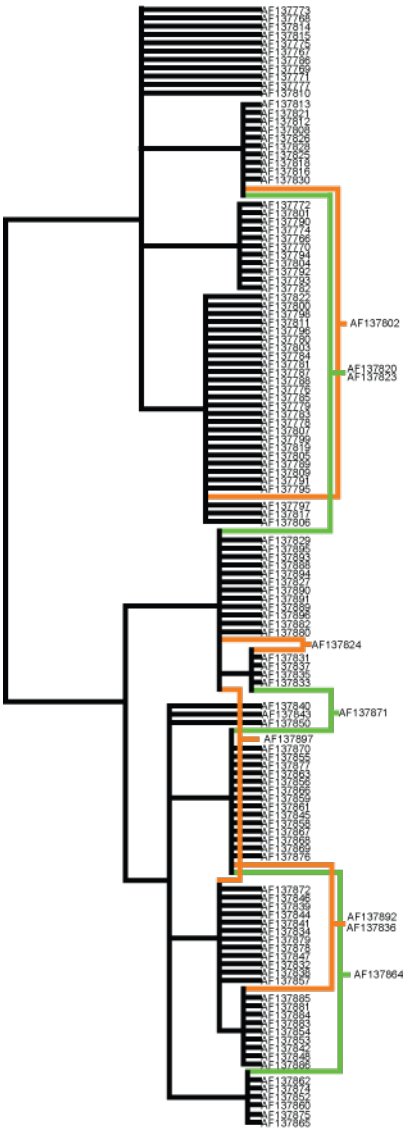


FIGURE 2.3: Network generated by applying our method to the C2-V5 region of the *env* gene of HIV-1 from patient 2. 9 putative recombinants were found.

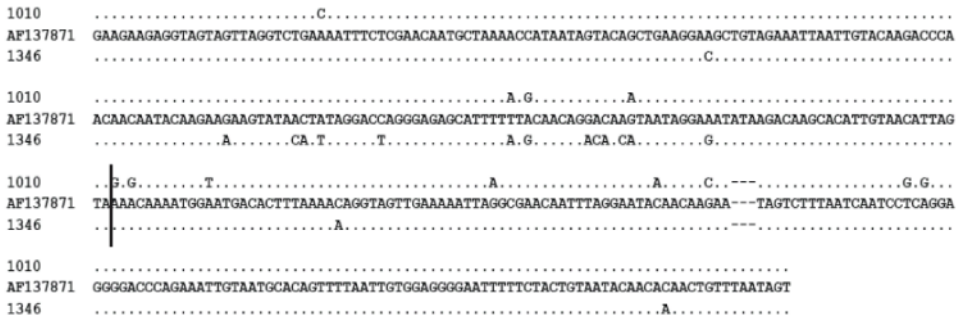


FIGURE 2.4: Mosaic sequence AF137871 from patient 2 shown with its parents 1010 and 1346. The vertical line at position 202 shows denotes the putative point of recombination such that parent the first half of the gene is believed to have descended from 1010 and the second half from 1346.

3

A rapidly diversifying host-defense gene cluster in the purple sea urchin *Strongylocentrotus purpuratus*

[1]In the past few years, many researchers have found that despite lacking an adaptive immune system, immune related genes in invertebrates tend to undergo recombination [45] [12] [2]. The genome of the purple sea urchin contains large diverse gene families, one of which, *185/333*, was found to be highly expressed in response to whole bacteria, lipopolysaccharide, β-1-3-glucan and double-stranded RNA [16] [17]. These genes show a 75 fold increase in activated coelomocytes, immune cells of the sea urchin [16]. Although the function of the 185/333 proteins remains unknown, they localize to the cell surface of a subset of the coelomocytes (immune cells) of the sea urchin and may be involved in the formation of syncytia to immobilize invading pathogens [18]. Analysis of 185/333 protein expression by two dimensional Western blot analysis suggests that distinct suites of proteins are expressed in response to lipopolysaccharide compared to peptidoglycan, and that individual sea urchins can

[1] Parts of this chapter are peer reviewed and published. Only the analysis done by S. Munshaw have been included. [2]

express > 200 unique proteins [19]. These protein data therefore support the previous observations of transcript and gene diversity [15] [55] [17] and emphasize the putative role of the 185/333 gene family in the *S. purpuratus* immune response. The DNA sequences of these genes show extraordinary diversity. Only 475 unique sequences were found from the 875 analyzed sequences obtained from 16 urchins. This diversity is the result of point mutations and various insertions and deletions [15] [55]. The alignments of the genes showed the presence or absence of 27 large blocks which were defined as elements [15]. In addition to this, we showed that the *185/333* genes are also subject to frequent recombination, gene duplication and gene deletion.

3.1 Repeat Structure of the genes

The 185/333 gene set showed an extraordinary pattern of repeats. The elements in the genes were defined by the presence and absence of these repeat elements. Figure 3.1 shows a dot plot of one of the sequences from the gene set against itself. The sequences to be compared are arranged along the margins of the matrix in the dot plot. The dot plot in Figure 3.1 is constructed using a window of 10 nucleotides with up to 2 mismatches. At every point in the matrix where the two sequences are identical for a stretch of 10 nucleotides a dot is placed (i.e. at the intersection of every row and column that have the same 10 letters in both sequences). A diagonal stretch of dots will indicate regions where the two sequences are similar. The solid line on the main diagonal is a reflection of the trivial fact that every base of the sequence is identical to itself. This solid line means that there is a stretch of nucleotides that is identical at two different positions in the genes. The repeats in the 5‘ end of the gene are about 75 nucleotides long. The 3‘ end of this gene has a large repeat along with some shorter repeats embedded in the large one. All unique genes have this repetitive structure and hence, they were aligned based on these repeats. We found 5 main families of repeats and these are present approximately 3-4 times in the genes.

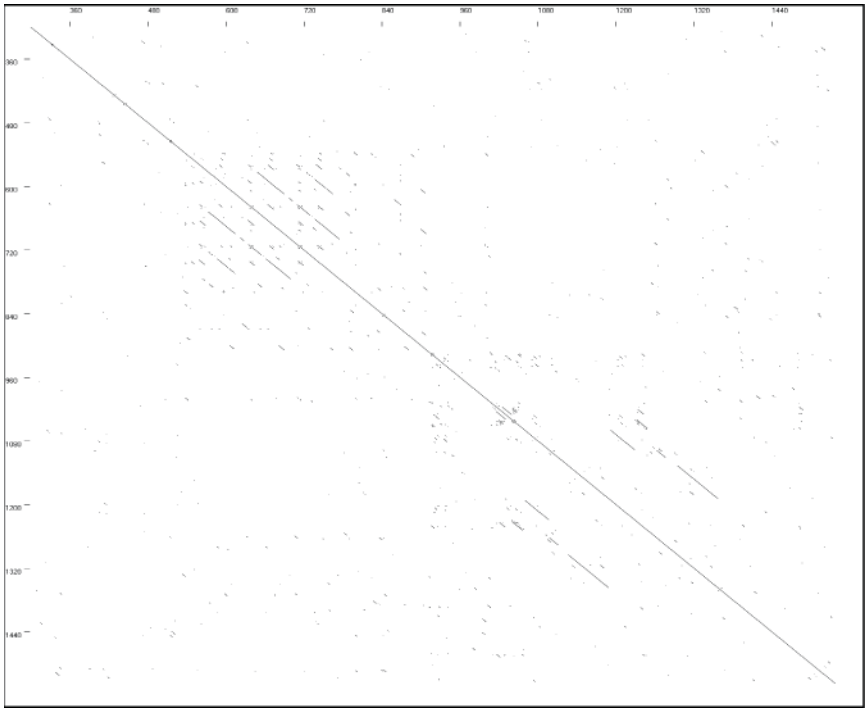


FIGURE 3.1: Dot plot showing of one of the 185/333 genes, S185-14 against itself. The gene has multiple repeat elements.

The 5‘ end had a long 75 nucleotide repeat whereas the 3‘ end of most genes contains shorter repeats that were embedded in this long repeat. Next, we looked more closely at the long 75 nucleotide repeat named Family A.

3.1.1 Family A

Family A of repeats had 4 members, each 75 nucleotides in length as seen in te 5‘ end of the gene in Figure 3.1. Each of the 185/333 genes either contained or lacked one of these members. The designation A1-A4 was given by sequence similarity. All but 3 genes in the set contained repeat A1 and repeat A4. Fewer contain repeat A2 and only 2 genes contain repeat A3. Between genes, each of these repeats differed by point mutations. Figure 3.2, constructed using dnaml in Phylip [56], shows the relationship between all members of repeat family A. The tree shows that there is a clear divergence between repeat A1 (pink) and repeat A4 (red). The tree also

shows that although repeat A1 (pink) and A2 (blue) are distinct on the tree, the distance between them is less than that between repeat A1 and repeat A4. This would suggest that if this were a duplication event, the A1-A2 divergence happened more recently than A1-A4. Knowing this, we also constructed evolutionary history for family A by point mutations. Figure 3.3 shows the potential evolutionary history of repeat family A. Since repeat A4 is present in all the genes, we assume that A4 is the ancestral sequence. We then suggest a duplication event followed by point mutations leading to distinct sequences for repeat A1 and repeat A4. The numbers in the orange circles are the number of mutations along each branch. Red circles represent recombination events. Several duplication, recombination and mutation events are used to best explain the presence and absence of repeats in these genes. Next, we wanted to validate the presence of recombination in these genes.

3.2 Validation of presence of recombination

Our aim for this analysis was to determine whether or not the elements, as defined, appear to have been recombined. The approach is to estimate evolutionary histories for the two elements and determine whether they appear to share a common genealogy. Here, we compare the two-tree vs. one-tree model. The two-tree model implies that given a pair of elements from a gene, they have evolved independently from each other, resulting in two different trees showing a possibility of recombination between the elements. On the other hand, the one-tree model is one where the elements have the same evolutionary histories and hence, can be represented with the same tree. This idea comes from the concept that a sequence that is a result of cross-over between two or more distinct ancestors can be best explained not by a single tree but by a set of correlated trees over the sequence [10]. We used dnaml from the PHYLIP 3.66 package [56]to estimate maximum likelihood (ML) trees from elements 1, 6 and 27 since they were present in all genes. We also estimated ML trees for all pairs

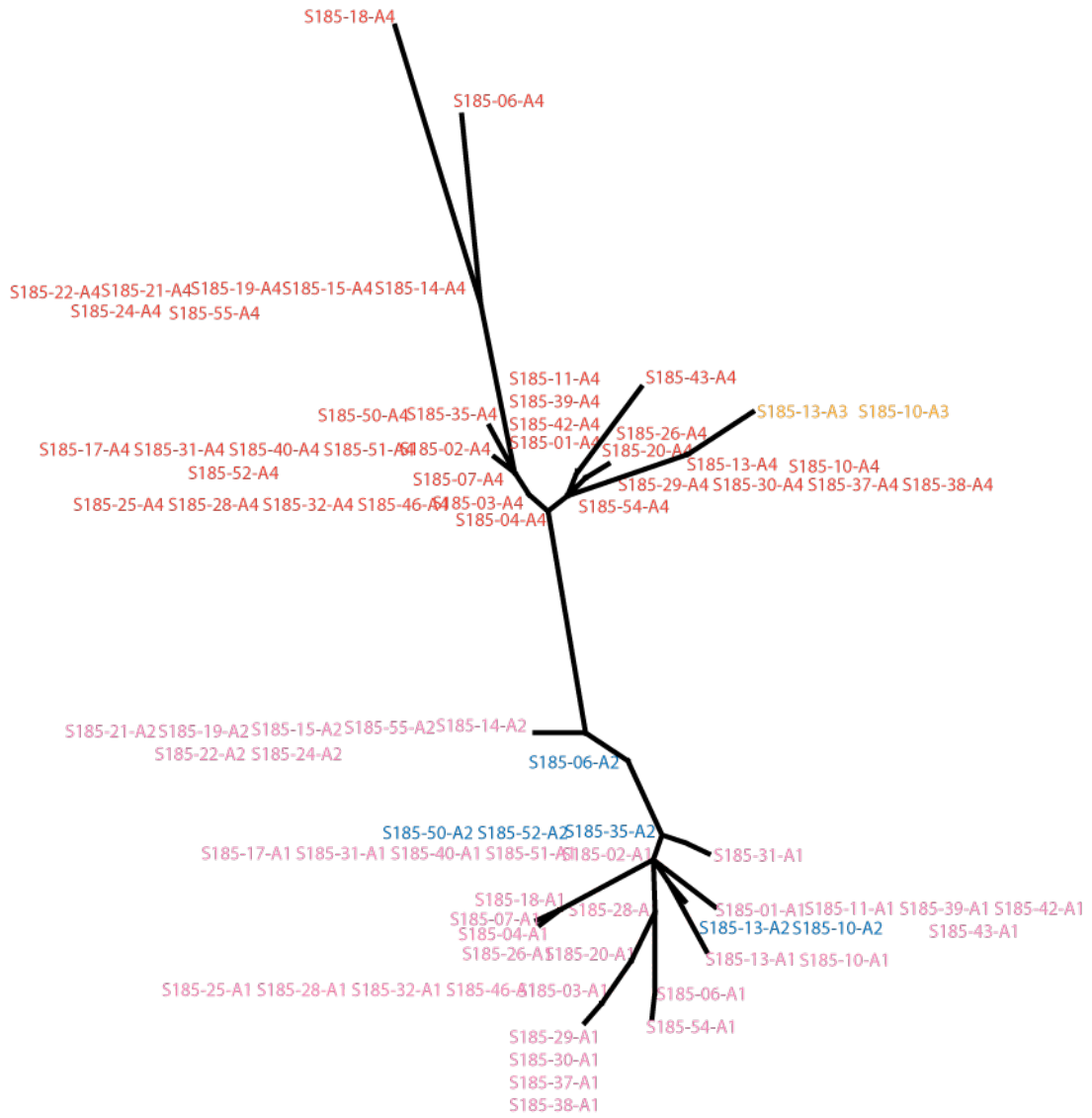


FIGURE 3.2: Phylogenetic relationship between the 4 members, A1(pink), A2(blue), A3(orange) and A4(red) of repeat family A. S185-# is the identification number for a gene.

of elements. A measure of comparing two nested models such as the one-tree and two-tree models is the log of the likelihood ratio: $\Delta = \log L_A + \log L_B - \log L_{A+B}$. The value of Δ will be larger as the true difference between the trees on elements A and B increases. It is important to determine whether the apparent difference in the element trees might have arisen by chance alone acting in the context of the

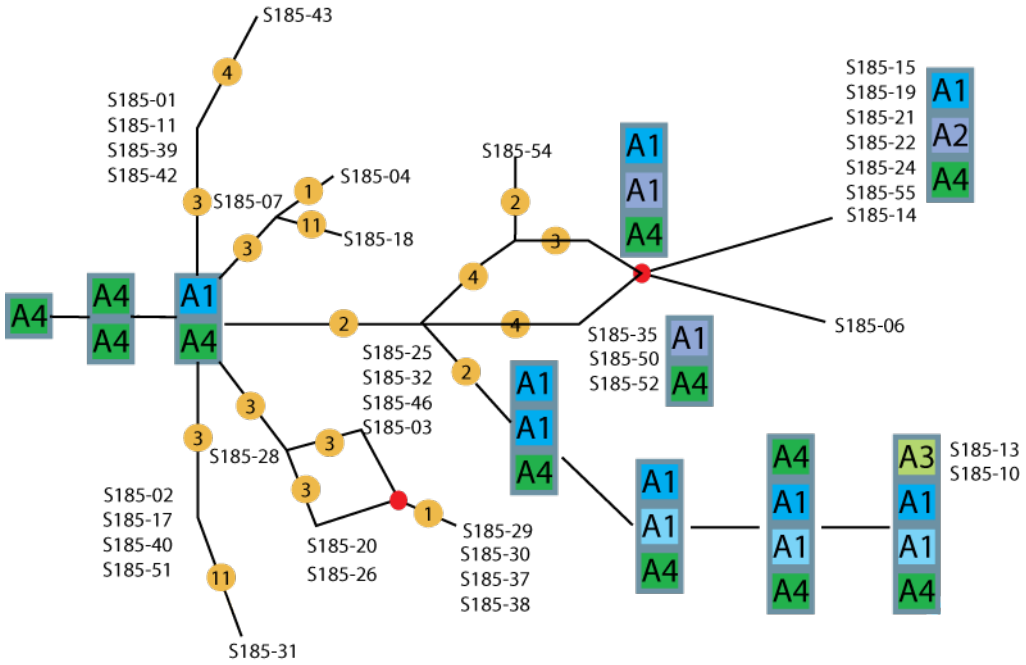


FIGURE 3.3: Potential evolutionary history of repeat family A. The numbers in the orange circles are the number of mutations along each branch. Red circles represent recombination events. Several duplication, recombination and mutation events are used to best explain the presence and absence of repeats in these genes.

single-tree process. Ideally, we would know what the expected distribution of Δ values is under the null hypothesis that there is just one tree shared by the two elements so we could judge just how rare the observed value of Δ is. Unfortunately, the models are too complicated for this distribution to be derived. We therefore turn to resampling methods. To estimate the distribution of Δ under the null hypothesis, we generate 100 permuted alignments. That is, we preserve individual columns in the alignment and reorder them randomly in such a way that each column is used exactly once. Here, we use the example of elements 1 and 27. For each of these permuted alignments we take the first 92 nucleotides, which is the length of element 1, to be our resampled element 1 and the remaining 135 nucleotides, the length of element 27, to be our resampled element 27. We estimate ML trees on them. As a check, we estimated the ML tree on the resampled complete alignment which should

be similar to the original complete alignment since all columns are used. For each of the 100 resampled alignments we have a value for Δ. Under the null hypothesis i.e. the elements co-evolved and there is no evidence of recombination, the observed value should be drawn from the same distribution as the resampled values. We used the human T cell receptor (TCR) genes as a positive control. It is widely known that antigen receptors in humans use recombination to generate a large repertoire of specificities against pathogens [57]. 25 unique sea urchin early histone H3 genes served as negative controls [58]. We also used element 6 as a negative control. We divided element 6 into a 5‘ half and 3‘ half. Since this is a small element, we would not expect any recombination to occur within element 6. Hence we would expect that both halves can be represented by the same tree.

3.3 Results of Validation

We ran the likelihood ratio analysis on all pairs of elements 1, 6 and 27 as well as on TCR sequences and histone gene sequences. We could reject the null hypothesis ($p < 0.01$) for the *185/333* genes as well as the TCR sequences. The histone genes ($p = 0.1$) and 5ànd 3ȟalves of element 6 ($p = 0.08$) show no evidence for recombination. Figure 3.4 shows the histograms for the log likelihood values for elements 1 and 27, histone genes and TCR sequences. From this analysis, we can conclude that there is evidence for scrambling or recombination in the *185/333* genes.

3.4 Discussion

The size and diversity of the 185/333 gene family provide an interesting system for studying the complexity of the innate immune system in *S. purpuratus*. The data presented in this chapter suggest that the diversity of the 185/333 gene family may, in part, be the explained by recombination events, in addition to point mutations. In particular the repeat families, located at the 5‘ end of the genes, appear to have

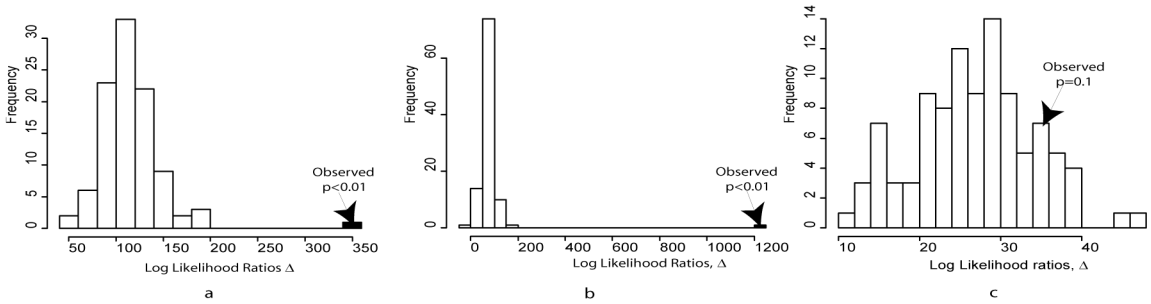


FIGURE 3.4: Histograms of Log Likelihood Ratio values for a) Elements 1 and 27 b) TCR V and J genes c) Segments H3a and H3b from histone genes

originated as a result of recombination, duplication, and deletion. Incongruent phylogenetic histories and analysis of the distribution of specific element sequences across genes also suggest that the genes undergo frequent recombination. Presence of different copies of repeats suggests that duplication may be an important mechanism for the diversification of this gene family.

Over the last few years, gene recombination has become increasingly evident in immune systems of invertebrates. In addition to the multistep assembly process of diversity in antigen receptors of jawless fish [12] and FREP genes expressed in freshwater snails in response to trematode parasites [59] are also believed to diversify through point mutation and somatic recombination using a limited set of source genes [13]. Different sets of FREP genes are present in parent snails compared to offspring, yet both parents and offspring have been shown to have identical source genes, implying somatic recombination of germline DNA in all tissues. The FREP genes therefore represent examples of germline-encoded immune diversity in medium to large families that function in organisms surviving on innate immunity.

This study provides evidence for yet another set of immune-related genes from a different invertebrate system. The 185/333 is a large gene family that employs recombination as a mechanism for diversification. Recombination is evident both between and within repeats, suggesting that these events are not limited to the

repeat borders, but instead may occur throughout the entire gene sequence. We did not identify any specific recombination hot spots, which is likely due to the presence of multiple repeats throughout the sequence. Frequent gene recombination is strongly implied from the sequences of the repeats of family A, as well as from the intact genes. In addition to the repeats within the coding regions, there are repeats that surround each gene in the intergenic region. We have not investigated these repeats in detail, but speculate that they may promote duplication and/or deletion of intact genes or blocks of tandem genes. The intergenic repeats plus the 185/333 genes themselves, which may act as large repeats, may facilitate meiotic mispairing, resulting in variations in the numbers of 185/333 genes in individuals. Based on the multiple types of repeats that we have identified in this gene cluster, the level of genomic instability and the rate of recombination may be more extreme than that observed for the FREP families. This would predict that the 185/333 gene family is the product of numerous ongoing and recent recombination events and that the extant members of the gene family are relatively young. This is in agreement with molecular clock analysis indicating that the 185/333 elements are not $>$ 10.8 million years old, about the same age as the *S. purpuratus* species.

The 185/333 gene family presents an intriguing addition to what is currently known about the complexity of invertebrate immune systems. The diversity is based on both variations in nucleotide sequence and mosaic combinations of repeats into distinct patterns, generating a diverse repertoire of transcripts and proteins [19] in response to immune challenge. This diversity may reflect diversification pressure placed on *S. purpuratus* by the microbes present within their marine environment. Marine microbial rRNA sequences isolated from Eastern Pacific seawater suggest that there are $2 * 10^6$ bacteria/ml and $5 * 10^5$ archaea/ml [60]. Given this level of constant pathogen exposure, it is only reasonable to expect that any organism living in this environment would survive based on a complex immune system that incor-

porates mechanisms to keep pace with the swift evolutionary variations in microbial pathogens.

In the next chapter we continue to study host diversification mechanisms by moving on to developing a method to study recombination in the vertebrate immune system, specifically in B cells.

4

SoDA2: A Hidden Markov Model Approach for Identification of Immunoglobulin Rearrangements

[1]As described in Chapter 1, antibody genes are a result of rearrangement of 2 or more independent gene segments, n nucleotide addition by TDT and somatic mutation. The inference of the recombination and mutation events that produced a given Ig gene is of great importance in the study of humoral immunity and has been tackled in many different ways. The goal of such inference is to identify each of the component gene segments used as well as the recombination sites, point mutations and n nucleotides. The aligned gene segments usually overlap, which is why alignments of the target gene to the individual gene segments cannot be treated as independent. Somatic mutations, n nucleotide addition and recombination site choice make this task more challenging. The short length of the DH gene segment makes it especially difficult to identify the CDR3 region of the heavy chain, which is the most diverse region in the antibody sequence. This leads to many possible gene segment combinations that can result in a given antibody gene. Hence, it is necessary to report all

[1] This chapter is currently under peer review [3]

such rearrangements and assign a probability to each of the combinations, making it easy to compare all possible rearrangements.

Several algorithms have been developed for inferring Ig gene segment composition. IMGT/V-QUEST is one of the first and most complete of these tools and has the ability to analyze both Ig and TCR sequences for a variety of organisms including human and mouse [61]. V-QUEST, however, is based on the BLAST algorithm; it does not guarantee finding the best alignment of two sequences [62]. Additionally, the implementation of the algorithm only allows for running a maximum of 50 sequences at a time. Another tool, JOINSOLVER, is based on the identification of conserved motifs in the target gene [63]. Both JOINSOLVER and V-QUEST provide multiple gene segment possibilities but the implementation only provides junction analysis for the topmost choice. Somatic Diversification Analysis (SoDA) [33] uses a 3D alignment algorithm that allows for insertions and deletions. The algorithm uses dynamic programming and is an extension of the Smith-Waterman local alignment Algorithm [64]. The 3D alignment allows for a continuous alignment through all the states of the recombination. SoDA infers only a single highest-scoring alignment, and ignores other solutions that may have equal or nearly equal scores. SoDA's guarantee of optimality in the inferred rearrangement is obtained at the cost of computational effort; SoDA takes more CPU time than either JOINSOLVER or V-QUEST. A major shortcoming for all the programs above is that they do not provide a meaningful comparison of the different possible rearrangements. iHMMune-align [65] partially solves the problem and provides a probabilistic model using an HMM to infer the rearrangement. iHMMune-align uses the Viterbi algorithm [66], to find the most probable path through the alignment matrix, but does not sum over paths or provide results on sub-optimal alignments. This choice becomes an issue when selecting an appropriate DH gene segment for Ig heavy chains. The DH gene is the shortest of all gene segments, and is typically the most difficult to align. We have found Ig

genes that present an equally good alignment with different DH genes (see Results and Figure 4.5). iHMMune-align or SoDA gives only the solution with the highest score even if the highest score is not significantly better than the second highest score and so on (iHMMune-align does provide the option of viewing the top 10 VH gene alignments, but not D).

Among these four methods, only SoDA allows for gaps when performing alignments, although insertions and deletions are known to occur at non-negligible frequencies during somatic hypermutation [67],and alignment without gaps when gaps are present leads to dramatically erroneous inferences.

The method we are introducing here is an update of SoDA-we call it SoDA2. It employs a probability mass function-based alignment for determining gene segments and a probabilistic HMM for the inference of CDR3. The system calculates the posterior probability over all paths using a particular set of gene segments by the forward and backward algorithms. It then provides the alignment path with the highest posterior probability. If the sequence does not hold enough information to unambiguously select a gene segment, SoDA2 reports all alignments that do not differ significantly. We tested this method using a simulated dataset constructed from the statistics of observed rearrangements and compared these results with those obtained using existing methods. We also used two natural datasets, a set of clonally related Ig genes and a random set of sequences from NCBI. Each test indicates that SoDA2 provides the most thorough and accurate results among all programs in addition to providing the most statistically complete results.

4.1 Methods

4.1.1 Determining the Type of Ig

The first step consists of aligning the target sequence with a consensus-like sequence of the VH, Vκ and Vλ families to determine if the input sequence is a heavy, kappa or

lambda chain. These consensus sequences are pre-created by separate alignments of the VH, Vκ and Vλ segments. We use the AHO numbering scheme [68]which is based on the spatial alignment of known three-dimensional structures of immunoglobulin domains. The gaps are placed to minimize the average deviation from the averaged structure of the aligned domain so that the position of the CDRs remains consistent. The consensus is represented by a probability mass function (pmf), a $L * 5$ matrix where L is the length of the V genes in this case [69]. For each nucleotide position in the gene, we determine the frequency of use for each nucleotide state (including "gap") at that position in the family. For the target antibody, we create a similar pmf using the quality scores of the input sequence. The quality score is proportional to the log probability of the estimated sequencing error and is provided by the user's base-calling software [70]. If quality scores are not provided, we treat the input sequence as well-determined and all mismatches as due to somatic mutation. The pmf at each position of the target sequence depends on the quality score, which varies at each position, and a mutation frequency μ which is assumed to be constant over positions. For each position, we then have the probability of observing the 5 bases (including a gap) at that position given the quality score and the mutation frequency. We use the pmf of the target antibody gene and the pmf of the VH, Vκ and Vλ sequences as scores to create a local alignment [64] [69].

4.1.2 V and J gene pre-alignment

Assume, for example, that our target sequence has been determined to be a heavy chain. We use the traceback path generated by aligning the pmf matrix of VH to our target and obtain the pmf for each member of the VH family. The mutation frequency μ is recalculated after observing mismatches in the highest scoring alignment. All VH segments with sufficiently high likelihood alignments are then submitted to the HMM. Sufficiency thresholds for the likelihood were established using a simu-

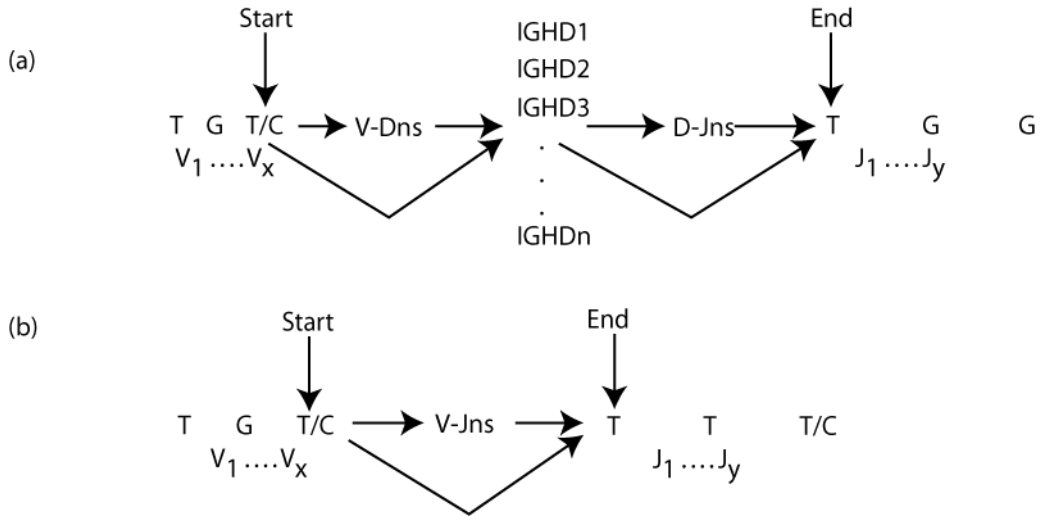


FIGURE 4.1: The basic topology of the HMM for (a) heavy chains and (b) kappa and lambda chains. The HMM starts at the last base of the invariant cysteine of all high-likelihood V segments, runs through all DH segments and through all high-likelihood JH segments till the first base of the invariant tryptophan or phenylalanine.

lated dataset (see Results and Discussion). The position of the invariant cysteine is determined. The target sequence is then aligned past the invariant cysteine with all the appropriate JH segments, using the pmf based alignment mentioned above. The Js with the highest likelihood are selected for submission to the HMM. The target sequence is further trimmed at the invariant tryptophan/phenylalanine, and only the remaining region, CDR3, is used as our target sequence for the HMM. The 3‘ ends (post-invariant cysteine) of all significant VH gene segments and the 5‘ ends (before invariant tryptophan/phenylalanine) of all JH segments from the pre-alignment are also chosen for the HMM. Since DH segments are most difficult to identify, we submit all DH segments to the HMM. The mutation frequency of the final trimmed target sequence to be considered for the HMM is set at 1.5x since the CDR3 region is subject to higher mutation than the VH region [65] [71]. Figure 4.1 shows the basic set-up of the HMM for heavy chains (4.1a) and light chains (4.1b) with an overview of the states and allowed transitions.

4.1.3 HMM

We implemented a pair HMM with 10 non-silent states - Match/Mistmatch state in V gene (Mv), Insertion in V gene (Iv), Deletion in V gene (Dv), V-D junction n nucleotides (N1), Match/Mismatch in DH gene (Md), Insertion in the DH gene (Id), Deletion in the DH gene (Dd), D-J junction n nucleotides (N2), Match/Mismatch in the J gene (Mj), Insertion in the J gene (Ij) and Deletion in the J gene (Dj). Our HMM must begin in the Match/Mismatch state of the V gene since the invariant cysteine is encoded by the V. The end state must be Match/Mismatch in the J gene at the beginning of the invariant tryptophan/phenylalanine.

The emission probabilities in every state are determined by the likelihood vector calculated using the quality scores and the mutation rate μ. For target sequences with unknown quality scores, high quality scores are assumed, making the probability of the observed base depend only on μ. Emission probabilities for the N nucleotide states are determined based on empirical data [72]. Transition probabilities between states are determined by fitting a negative binomial distribution (see Figure 4.2) to the recombination site choice for VH, DH and JH and number of n nucleotides in the junctions as determined in a set of 293 unmutated rearranged sequences [73]. Figure 4.3 shows a detailed implementation of the HMM with transition probabilities.

4.1.4 Algorithm

Once we have the appropriately trimmed target and germline sequences, we calculate the log of the total probability of a proposed rearrangement using the forward and backward algorithms [74] [75]. We select the gene segments that lead to the highest posterior probabilities, and perform a Posterior Viterbi algorithm with traceback [76] to select the path with the highest posterior probability for each possible rearrangement. We report the probability of the most probable path for each of the equally probable gene segment sets. For a heavy chain, a DH gene alignment of less than 3

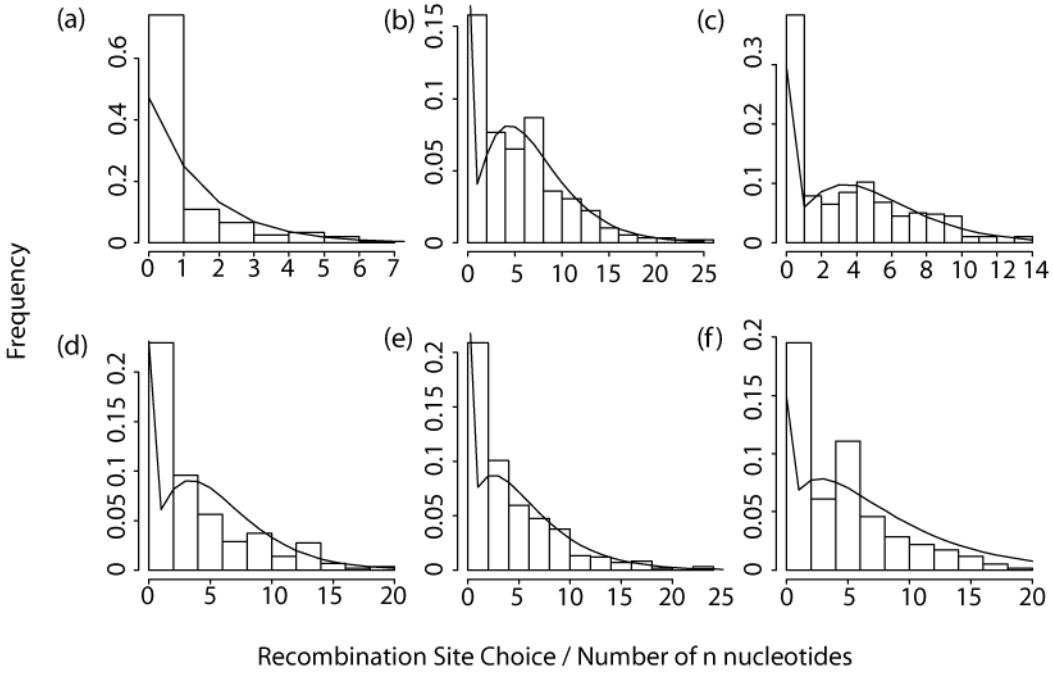


FIGURE 4.2: Distribution of the empirical data for (a) VH gene recombination site choice (b) n nucleotides in the VD junction (c) 5‘ DH recombination site choice (d) 3‘ DH recombination site choice (e) n nucleotides in the DJ junction (f) 5‘ JH recombination site choice. All the data is fit to negative binomial distributions with varying parameters.

nucleotides is flagged as "Unreliable D Alignment". The functionality of an antibody gene is determined as follows and reported with the results. A functional Ig chain must have no stop codons and the invariant cysteine at the start of CDR3 must be in-frame and intact. For heavy chains, the invariant tryptophan at the end of CDR3 must be in-frame and intact; for light chains, CDR3 must end with an in-frame and intact phenylalanine. We provide color coded output in HTML, text and excel formats to allow the user to use the information in ways most convenient to his or her needs. (See Appendix A for algorithm details)

4.2 Validation, Results & Discussion

4.2.1 Simulated Datasets

We created simulated datasets of 100 sequences each with mutation frequencies of 2.5%, 5%, 10% and 20%. Recombination site choice and number of n nucleotides for these simulations were drawn from a negative binomial distribution. To avoid any bias towards our HMM, the parameters for these simulations were estimated using a set of 662 sequences obtained from Genbank. Furthermore, rearrangements for these sequences were determined using IMGT/VQuest [61] rather than SoDA or SoDA2. IMGT Junction Analysis was used to determine empirical data for deriving the distributions [77]. Mutations were introduced such that the average mutation frequency across the gene was 2.5%, 5%, 10% and 20%, and the mutation frequency in the CDRs was 2x than that in the framework. Each of these datasets was used to test SoDA2, SoDAv1.0, IMGT/VQuest, JOINSOLVER and iHMMune-align. Inverted DH segments were omitted from the simulations because IMGT/VQuest and iHMMune-align do not allow for alignments against them. Table 4.2.1 shows the results of running our simulated datasets using the various available software. The table shows the number of rearrangements (all VH, DH and JH with alleles) identified correctly at each mutation rate by each program out of the 100 sequences tested in each group. For all our tests, we only compare the highest scoring rearrangement provided by SoDA2 with the highest scoring ones provided by other programs. For our simulated data, we see that SoDA2 performs better in identifying the complete rearrangement (including correct alleles) than other programs under all mutation rates (Table 4.2.1). In particular, SoDA2 outperforms all other programs in DH segment identification (Table 4.2.1). SoDA2 falls slightly behind JOINSOLVER in VH and JH gene identification due to the trade-off between accuracy and efficiency. We employ a computationally efficient alignment algorithm that aligns the target gene

Table 4.1: Number of correct rearrangements identified by each software out of 100 sequences tested at each mutation rate

	0.025	0.05	0.10	0.2
SoDAv2.0	73	65	47	28
IMGT/V-QUEST	52	47	42	16
JOINSOLVER	59	47	34	11
iHmmune-align	41	31	22	12
SoDA	46	30	31	6

to consensus sequences of alleles, which can lead to the identification of the incorrect allele in a very few cases. Aligning the target gene to every allele would decrease this error but increase computation time significantly. Such errors are seen rarely and do not change the overall superior performance of SoDA2 shown in Table 4.2.1. If the score for multiple rearrangements is equal for any of the programs, all rearrangements are considered. Although SoDA2s performance falls at the 20% mutation rate, it still performs better than other software. We only report all alignments that are equally probable and leave it up to the user to select and view any number of V, J or complete alignments he or she wants. For sequences where SoDA2 failed to identify the correct rearrangement as the most probable one, we found a median difference of 0.67 in the natural log of the probability between the highest scoring rearrangement and the correct one at the 5% mutation rate. Thus, if allowed to include rearrangements with low differences (<1) in the natural log of the probability from the top scoring alignment, SoDA2 would have identified correct rearrangements for 22 additional sequences at the 5% mutation rate, yielding a possible 87% success rate.

4.2.2 Clonally Related Datasets

In order to test real biological data, we used two clonally related datasets that were used to test iHMMune-align (Gaeta et. al, 2007) derived from tonsilar IgD

Table 4.2: Number of VH, DH and JH genes (with alleles) identified at each mutation rate for the simulated sequences

	0.025			0.05			0.10			0.2		
	V	D	J	V	D	J	V	D	J	V	D	J
SoDA2	97	76	98	94	73	94	87	58	88	85	42	78
IMGT/V-QUEST	90	65	98	83	61	92	81	42	85	72	20	76
JOINSOLVER	99	52	94	97	49	93	93	45	89	88	23	82
iHmmune-align	79	65	92	77	68	87	76	42	69	69	20	45
SoDA	87	48	90	86	42	87	78	32	86	69	21	61

class-switched B cells (Zheng et al., 2004). Because they are clonally related, sequences within a given set should have identical rearrangements and differ only by somatic mutation. We analyzed this dataset using VQuest, JOINSOLVER, SoDA and iHMMune-align to determine the number of times each of the programs resulted in the same rearrangement as was done by Gaeta et al. (2007). We ran the sequences through all the programs and found that iHMMune-align selected 47/57 identical rear-rangements for the first group of sequences, while SoDA2 selected 34/57 identical rearrangements. IMGT/VQuest, JOINSOLVER and SoDA identified 37, 25 and 18 identical rearrangements re-spectively. SoDA2 returned a minority DH gene segment in 17 ca-ses, a minority JH allele in 5 cases, and a minority VH allele in 4 cases. In cases where SoDA2 failed to select the majority VH or JH gene segment, all the other programs, including iHMMune-align also failed to select the majority gene segment. It can be seen in these cases that mutation had obliterated the information necessary to make the correct inference. For the 17 cases where SoDA2 did not return the majority DH segment, the DH segment that was returned was typically judged more probable than the majority segment due to the balancing of n-nucleotide use and mutations. An example of this phenomenon is the inference for AF262199 (Figure 4). In this case, In this case, the mutation frequency in the VH gene segment is approximately 7%. SoDA2 selects IGHD1-26*01 requiring 3 muta-

tions (8.5% mutation frequency in CDR3) and 7 n-nucleotides, while IGHD7-27*01 requires 2 mutations (5.5% mutation frequency) and 10 n-nucleotides. For the second dataset of 99 sequences, both iHMMune-align and SoDA2 identified 68 out of 99 identical rearrange-ments while IMGT/VQuest, JOINSOLVER and SoDA identified 56, 41 and 37 identical rearrangements, respectively.

4.2.3 Sequences from Genbank

We tested a set of 662 sequences collected from Genbank and previously used for testing iHMMune-align and SoDA (Genbank accession nos Z68345-487 and Z80363-770). 113 out of 662 sequences produced inferences on which all five programs agreed. There was no agreement from any of the programs on 140 sequences. This means that they either could not infer a rearrangement at all or they all differed in their inference. From the rest, SoDA2 agreed with the majority of the programs on 300 rearrangements (see Table 4.2.3). These did not include those where SoDA and SoDA2 were the only two in agreement and the chosen rearrangement was the majority. SoDA2 performs considerably better than other programs in this test. We closely examined sequences for which SoDA2 failed to agree with 2 or more programs. We found a median difference of 1.05 between the top scoring rearrangement and the majority rearrangement. We also found that in all cases, SoDA2 selected an alternative rearrangement equally likely as the majority one. Figure 4.5 shows an example of one such sequence where SoDA2 selected IGHD2-21*01 to be the best fitting DH alignment with a score of -785.07 (4.5a). On allowing alignments with a slightly higher probability, we found both the rearrangement chosen by the majority of the programs (VQuest, JOINSOLVER and iHMMune-align, Figure 4.5(b) and also the rearrangement selected by SoDA (4.5c). The difference in the natural log of the probability is 0.63 in the first case and 0.93 in the second. This shows that allowing rearrangements within a reasonable range of probabilities in SoDA2

Table 4.3: Results from 662 sequences from Genbank, showing the performance of the 5 programs. If 2 or more programs displayed the same rearrangement (including the alleles), it was believed to be the majority rearrangement.

	Number of Rearrangements
All Programs Agree	113
All Programs disagree	140
SoDA2 agrees with 2 or more programs	300
VQuest agrees with 2 or more programs	255
iHMMune-align agrees with 2 or more programs	137
JOINSOLVER agrees with 2 or more programs	272
SoDA agrees with 2 or more programs	244
SoDA2 agrees only with SoDA (no other programs agree)	11

would give an accurate and thorough picture of the various rearrangements possible for a given immunoglobulin sequence. It is important to note that SoDA2 considers factors such as recombination site choices for each gene segment and numbers of n nucleotides at both junctions derived from empirical data in inferring rearrangements while alignment algorithms used by SoDA, VQuest and JOINSOLVER base their results on sequence similarity matrices which may not accurately represent the process of V(D)J recombination.

4.3 Conclusion

The problem of inferring the correct rearrangement for antigen receptors is difficult due to the stochastic nature of the process, but the task is important for an increased understanding of the population somatic genetics of the immune response. In this paper we present a method based on an HMM that provides a statistical basis for identifying rearrangements of Ig genes. In addition to providing the posterior probability of the top rearrangement candidate, SoDA2 also provides all rearrangements with sufficiently high posterior probabilities, thus giving the user a statistically complete picture of the observed sequence's origins.

We tested SoDA2 against simulated datasets that were created using empirically observed recombination site choices for each of the gene segments and numbers of n nucleotides in the junctions. We also tested it on two clonally-related datasets as well as a set of Ig heavy chains chosen randomly from Genbank. Our software performed as well as or better than available software on two out of three validation tests. The one test where SoDA2 did not outperform all of the others involved a single rearrangement. On the identical test with a different rearrangement, SoDA2 did as well as its nearest competitor. It is important to realize that the key feature of this paper is to provide a tool based entirely on a probability model, and that therefore returns results interpretable as posterior probabilities rather than arbitrary scores. As with other inferential procedures, it is important to not only identify the optimal solution, but to identify near-optimal solutions and have a method for the absolute comparison among these alternatives. This performance and thorough result reporting leads to a substantially longer computation time. SoDA2 takes approximately 15s of real user time per set of VH and JH segment for a given heavy chain target sequence on a 64 bit machine with a 2.19GHz processor and 4GB RAM, but the investment of computational effort seems worthwhile.

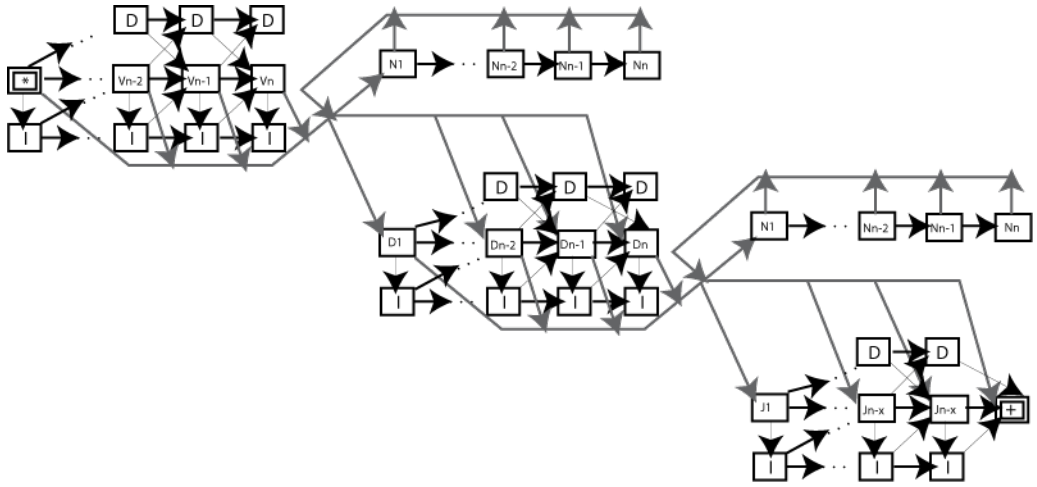


FIGURE 4.3: Shows a detailed topology of the HMM with all possible transitions. Each nucleotide in the observed sequence is treated as a separate state. The transition probabilities are derived from empirical data. The star denotes the start (3rd position of invariant cysteine) of the HMM and the + denotes the end (first position of invariant tryptophan/phenylalanine)

```
(a)     AF262199          TGT GTG AGG AAT ACT GGG AAT CGG GGT GCT TTT GAT ATC TGG
        IGDH1-26*01        TGT GCG AGG TAT AGT GGG AAT CGG GGT GCT TTT GAT ATC TGG
        Key               VVV VVV VVD DDD DDD DDD Dnn nnn nnJ JJJ JJJ JJJ JJJ JJJ
        InputAA             C   V   R   N   T   G   N   R   G   A   F   D   I   W
        GermAA              C   A   R   Y   S   G   N   R   G   A   F   D   I   W

        AF262199          TGT GTG AGG AAT ACT GGG AAT CGG GGT GCT TTT GAT ATC TGG
        IGDH7-27*01       TGT GCG AGG AAT ACT GGG GAT CGG GGT GCT TTT GAT ATC TGG
        Key               VVV VVV VVn nnn DDD DDD DDn nnn nnJ JJJ JJJ JJJ JJJ JJJ
(b)     InputAA             C   V   R   N   T   G   N   R   G   A   F   D   I   W
        GermAA              C   A   R   Y   T   G   D   R   G   A   F   D   I   W
```

FIGURE 4.4: (a) Top rearrangement as chosen by SoDA2 with a higher mutation frequency than the alternative (b). This is calibrated with the mutation frequency in the VH region.

```
    SoDA2
    SoDA2 Score          -785.07
(a) 1154693     TGT GCA AAA GAT AAG GTT GAC GGA GCA GGT GGT GGA GAG GGG GAT TAC TAC TAC TAC TAC GGA ATG GAC GTC TGG
    IGHD2~21*01 TGT GCA AAA GAT AAG GTT GAC GGA GCA GGT GGT GGT GAG GGG GAT TAC TAC TAC TAC TAC GGT ATG GAC GTC TGG
    Key         VVV VVV VVV VVV Vnn nnn nnn nnn nnn nDD DDD DDD Dnn nnn nJJ JJJ JJJ JJJ JJJ JJJ JJJ JJJ JJJ JJJ JJJ
    InputAA      C   A   K   D   K   V   D   G   A   G   G   G   E   G   D   Y   Y   Y   Y   Y   G   M   D   V   W
    Germ AA      C   A   K   D   K   V   D   G   A   G   G   G   E   G   D   Y   Y   Y   Y   Y   G   M   D   V   W


(b) VQuest, JOINSOLVER, iHMMUNE
    SoDA2 Score          -785.7
    1154693     TGT GCA AAA GAT AAG GTT GAC GGA GCA GGT GGT GGA GAG GGG GAT TAC TAC TAC TAC TAC GGA ATG GAC GTC TGG
    IGHD6~13*01 TGT GCA AAA GAT AAG GTT GAC GGA GCA GCT GGT GGA GAG GGG GAT TAC TAC TAC TAC TAC GGT ATG GAC GTC TGG
    Key         VVV VVV VVV VVV Vnn nnn nnn nnD DDD DDD DDD nnn nnn nnn nJJ JJJ JJJ JJJ JJJ JJJ JJJ JJJ JJJ JJJ JJJ
    InputAA      C   A   K   D   K   V   D   G   A   G   G   G   E   G   D   Y   Y   Y   Y   Y   G   M   D   V   W
    Germ AA      C   A   K   D   K   V   D   G   A   A   G   G   E   G   D   Y   Y   Y   Y   Y   G   M   D   V   W


(c) SoDA1
    SoDA2 Score          -786
    1154693     TGT GCA AAA GAT AAG GTT GAC GGA GCA GGT GGT GGA GAG GGG GAT TAC TAC TAC TAC TAC GGA ATG GAC GTC TGG
    IGHD2~2*01R TGT GCA AAA GAT AAG GTT GAC GGA GCA GCT GGT GGA GAG GGG GAT TAC TAC TAC TAC TAC GGT ATG GAC GTC TGG
    Key         VVV VVV VVV VVV Vnn nnn nnn nnD DDD DDD DDD nnn nnn nnn nJJ JJJ JJJ JJJ JJJ JJJ JJJ JJJ JJJ JJJ JJJ
    InputAA      C   A   K   D   K   V   D   G   A   G   G   G   E   G   D   Y   Y   Y   Y   Y   G   M   D   V   W
    Germ AA      C   A   K   D   K   V   D   G   A   A   G   G   E   G   D   Y   Y   Y   Y   Y   G   M   D   V   W
```

FIGURE 4.5: The alignment of CDR3H of sequence by 1154693 using IGHD1-21*01 by (a) SoDA2 (b) IMGT/V-QUEST, JOINSOLVER and iHHMune (c) SoDA. Rearrangements (b) and (c) were also provided by SoDA2 at a slightly lower probability.

5

Antibody Response to HIV-1

5.1 Background

The primary goal of this project was twofold: the first is to characterize the plasma cell response in an acute HIV-1 infection and second, to compare this response to an influenza vaccination and influenza infection. We chose to study plasma cells in each of the cases because it is well known that the first B cells that respond to infection or vaccination are present in blood as plasma cells [28]. While HIV-1 infection is difficult to diagonize, chronic HIV-1 infection is well studied. Various studies have shown that hypergammaglobulinemia and increased B cell activation are characteristic of the chronic infection [78]. Their work suggested that both virus-specific IgG and polyclonal IgM and IgA responses are present in HIV-infected individuals. A decrease in B cell memory response has also been shown [79] where influenza-specific memory B cell responses were significantly lower in HIV-infected than in HIV-negative individuals. Very few studies have looked at early HIV-1 infection. One such study showed a decrease in peripheral B cell counts but no changes in memory B cell numbers in individuals with a primary infection of less than 60 days [80]. This study focuses

Table 5.1: Clinical Characteristics of Individuals in Antibodyome Project

Patient ID	Days post infection	Viral Load (copies/ml)	CD4 count (cells/mm^3)	Treatment Details
001-4	17	4,750,000	164	No treatment given
065-0	17	1,860,000	269	No treatment given
684-6	20	1,500,000	225	No treatment given
681-7	20	9,000,000	230	No treatment given
068-9	30	1280	611	On ART for 1 week
DFLU07001	7	NA	NA	Vaccinated with Fluzone 2007
DFLU07021	7	NA	NA	Vaccinated with Fluzone 2008
DFLU07004	7	NA	NA	Vaccinated with Fluzone 2007
DFLU07024	7	NA	NA	Vaccinated with Fluzone 2008
KFLU08002	7	NA	NA	Influenza Infected, Asymptomatic
KFLU08003	7	NA	NA	Influenza Infected, Asymptomatic
KFLU08005	7	NA	NA	Influenza Infected, Symptomatic
KFLU08007	7	NA	NA	Influenza Infected, Symptomatic
KFLU08012	7	NA	NA	Influenza Infected, Symptomatic
KFLU08013	7	NA	NA	Influenza Infected, Asymptomatic
0223	AHI Uninfected	NA	NA	NA
0239	AHI Uninfected	NA	NA	NA

on the immediate effect of the virus on the B cell arm of the immune system. The work presented in this thesis is a part of this project and focuses specifically on the somatic population genetics of the antibody repertoire in an acute HIV-1 infection.

5.2 Patients

We have antibody pairs from single plasma cells from the blood leukapheresis and plasma from five acutely infected (AHI) homosexual men. All infections are believed to be sexually transmitted. The exact day of transmission in all these patients was recorded because of excellent patient-doctor relationships. Our disease controls are peripheral blood from four influenza vaccinated individual (FV), samples drawn 7 days post vaccination with trivalent inactivated seasonal vaccine, three individuals with an asymptomatic influenza infection (FIA) and three individuals with a symptomatic influenza infection (FIS), blood samples 7 days post infection. The volunteers were infected with H3N2 A/Wisconsin/67/2005. FIA patients were those that had contracted the disease as measured by seroconversion, but never developed any symptoms. Influenza infection and vaccination were selected to serve as controls where a viral infection is successfully cleared and a successful memory response is

developed. Although our patients have never seen the particular vaccine or strain they have been vaccinated with or infected with influenza in our study, the possibility of previous exposure remains a concern. The presence of cross-reactive antibodies should be kept in mind while making comparisons. Our negative control is two AHI uninfected individuals. Table 5.1 shows the clinical information of the individuals.

5.3 Analysis pipeline

In order to facilitate analysis of the large scale antibody sequencing of the antibodyome project, we have developed an efficient computational pipeline that ensures quality control.

- Single plasmablasts and plasma cells were sorted were sorted from the patients [81] and the antibody genes were sequenced from each cell in the forward and reverse direction using multiple primers, one for each VH, VK and VL family.

- Chromatograms generated by the sequencing facility were used as input to an industrial strength base calling software called Phred [82] [70]. This software reads the chromatograms in ab1 files, calls bases and assigns a quality score to each base. The quality score of a given position is defined by equation 5.1 where P_e is the probability of error of the base call at a position. It is the log transformed probability of observing an error at that position. To study the somatic genetics of antibodies, it is necessary to distinguish between a bad quality base and one that has changed due to somatic mutation. Hence, quality scores are extremely important for this project.

$$Q = -10 \log_{10} P_e \tag{5.1}$$

  For each chromatogram there are two output files - a FASTA file that contains the base calls and another FASTA file with the quality score at each position.

- We assembled the forward and reverse strand to consolidate the information from both. For this we used a likelihood based alignment algorithm that aligns the sequences on the basis of its quality scores. This was done via a standard Smith-Waterman local alignment [64] algorithm with a difference in scoring scheme. Instead of using a scoring matrix or the actual base at each position, we use the quality score at each position. The quality score Q_i at i is transformed into a probability mass function (pmf) vector L_i of size 5 where positions 0,1,2,3,4 correspond to the likelihood of bases A,C,G,T,-respectively at position i. Values for L_i were calculated in the following way.

$$L_{ij} = 1 - 10^{\frac{-Q_i}{10}} \tag{5.2}$$

if j = base call at i. The second term in this case refers to the probability of an error at i

$$L_{ij} = (1 - \delta) * \frac{10^{\frac{-Q_i}{10}}}{3} \tag{5.3}$$

if $j \neq$ base call at i and $j \neq$ -. δ in this case is the probability of a deletion and the probability of error is equally divided by the other 3 bases.

$$L_{i4} = \frac{\phi}{4 * 10^{\frac{-Q_i}{10}}} \tag{5.4}$$

in case of a gap. ϕ is the probability of an insertion. We set $\phi = 4\delta$ so that the scoring matrix is reversible.

The match/mismatch score in an alignment of sequence x at position i and sequence y at position j is then calculated as the dot product of the likelihood vectors of x and y at positions i and j, so that the score would be higher if it were a match and lower if otherwise. The gap score is calculated by first calculating the average quality score of the two positions between which the

gap is to be inserted. Lets call that *laqs*. The likelihood vector of the position i with the gap is then calculated by

$$L_{i0\ldots3} = \delta * 10^{\frac{-laqs}{10}} \tag{5.5}$$

for all bases.

$$L_{i4} = (1 - \phi * 10^{\frac{-laqs}{10}}) \tag{5.6}$$

for the gap. For each position in the sequence, we now have the the probability of observing each base at that position given the quality score. Finally, after the sequences have been aligned, we make a consensus sequence from the forward and reverse strand that can be used for further analysis. This was done by selecting the base at each position which has the maximum (max) likelihood calculated by multiplying the likelihoods at each position for the two vectors. The quality score for that position is calculated by

$$Q_i = -10 \log_{10} (1 - max) \tag{5.7}$$

This gives us a consensus sequence and its quality. This method has been adapted from a recently published study of assembling interferon genes from the bat [69]

- Next, we trimmed the sequence based on the forward and reverse primers to ensure that we don't have any primer dependent sequence. For this, we ran a Smith-Waterman local alignment [64] once again and trimmed at the end of the alignment.

- Next, we trimmed the sequence based on the quality score to ensure that only good quality sequence is used for further analysis. Base calls at the beginning and end of sequences are usually of a lower quality. Our aim was to use the best quality sequence. Figure 5.1 shows an example of the quality score for a

sequence and a plot of position vs. score. The arrows show positions to trim when using the default value as 10. The algorithm is as follows:

- Set a default value for quality score (in the example figure 5.1, it is set to 10).
- Initialize an n x 3 matrix M where n is the length of the sequence. Set $M_{0j} = Q_0$. Initialize every other position to 0.
- Fill out the matrix by using the following conditions:

  For $i = 1..n$

$$M_{i0} = M_{(i-1)0} + default \tag{5.8}$$

$$M_{i1} = M_{(i-1)1} + Q_i \tag{5.9}$$

$$M_{i2} = max(M_{(i-1)2} + default, M_{(i-1)2} + Q_i) \tag{5.10}$$

  In tracing back, we find the two points that correspond to the two arrows in figure 5.1 by looking for the maximum value for each column. The first point is when the maximum value switches from being in row 1 to row 2 and the second point is when the maximum value switches from row 2 to row 3. These two points are used to trim the sequence. Figure 5.1 also shows an example of the matrix with arrows for trimming positions.

- Now we have a good quality antibody sequence that we can analyze. We ran the sequence through Somatic Diversification Analysis (SoDA, [33]) and determine the germline rearrangement. The information from SoDA is then stored into a SQL database. The database has 21 columns, names and decriptions of which are shown in Table 5.3. This database made it easier to access specific information about the antibody sequences.

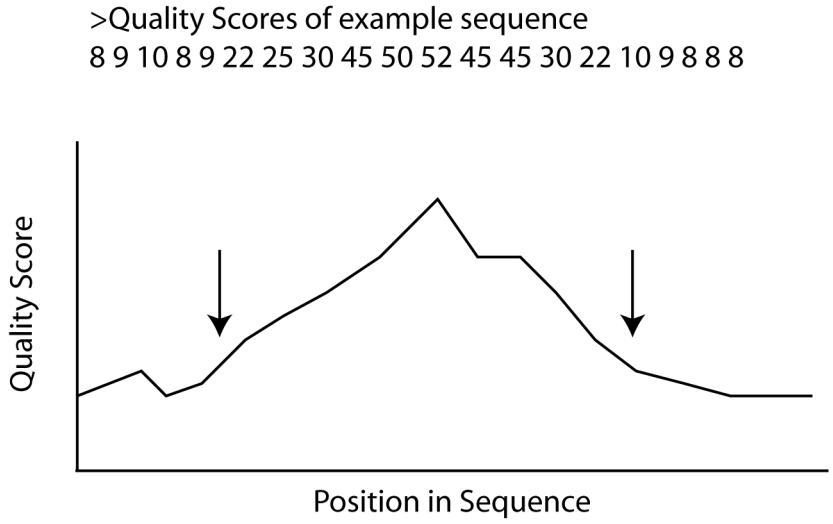
- Next, we determined which of these sequences resulted into valid antibody pairs. We searched the database for sequences from the same patient, plate and well that had a functional heavy and a functional light chain. Wells with more than 1 heavy chain and with more that 2 light chains were discarded. Sequences from wells with a pair were then given ID numbers (H000# for heavy chains and K000# or L000# for kappa or lambda respectively). The pair information was stored in a separate database with the IDs of each pair. Reactivity for the functional antibodies were tested using ELISA and Luminex assays.

- We performed statistical analysis of the antibody pairs by patient and by specificity. For specificity, our categories were gp41 specific, other HIV-1 specific, autoreactive, Hemagglutinin (HA)-specific, Non-HA Flu specific, gut flora specific, other specificities and unknown specificities. gp41 specific antibodies are those that bind to consensus few autologous envelope proteins gp41, gp140 or gp160. HIV-1 specific antibodies are ones that bind to other HIV-1 proteins such as the consensus sequences of p31, reverse transcriptase, Tat, p55, Gag or AT-inactivated virions. Cardiolipin and Anti-nuclear antibody (Hep-2 cells) were used to test for autoreactivity. Fluzone 2007 killed influenza vaccine and HA from various strains were used for influenza specificity, gut flora and lipid A are in the gut flora category while tetanus toxoid, killed cryptococcus and killed candida fall in the category of other specificities.

5.4 Results

5.4.1 Clonal Expansion

As mentioned in Chapter 1, B cells undergo a process known as affinity maturation when they encounter antigen. This process allows the B cell to undergo somatic



8	18	28	38	48	58	68	78	88	98	108	118	128	138	148	158	168	178	188	198
8	17	27	35	44	66	91	121	166	216	268	313	358	388	410	420	429	437	445	453
8	18	27	37	45	66	91	121	166	216	268	313	358	388	410	420	430	439	447	455

FIGURE 5.1: An example that shows the quality score of a sequence, a plot of position vs quality score and the alignment matrix used to trim by quality score. The default value used is 10.

mutation which in turn increases its affinity to the antigen that activated it. It was recently shown that the antibody response to the influenza vaccine upto 7 days post vaccination was dominated by only a few B cell clones that had the same germline rearrangement but extensive intraclonal diversity from somatic mutations [83]. 10 out of 12 multiple sclerosis patients showed expansion of one single B cell clone in their cerebrospinal fluid [84]. Clonally related B cells were also seen in the germinal centers of two Epstein Barr Virus infected tonisllar sections approximately 7 and 14 days after infection [85]. A study done for HIV-1 observed that only patients with hypergammaglobulinemia showed B cell clonal expansion [86]. However, all of the HIV-1 patients enrolled in this study had been infected for over a year. Another study showed 134 clones in a study that isolated 502 B cells in chronically infected elite

Table 5.2: Name and description of columns in SQL database.

Name of Column	Description
SeqID	Identification number of the sample
TypeOfIg	Heavy or Kappa or Lambda
TotalScore	Score of Alignment
Length	Length of Alignment
Functionality	Whether the antibody has a functional rearrangement
V	V segment used
Vmuts	Number of mutations in the V segment
V-Dns	Number of n nucleotides in the V-D junction. Would be NA in light chains
D	D segment used
Dmuts	Number of mutations in the D segment
D-Jns	Number of n nucleotides in the D-J junction. Would be NA in light chains
V-Jns	Number of n nucleotides in the V-D junction. Would be NA in heavy chains
J	J segment used
Jmuts	Number of mutations in the J segment
CDR3aaseq	Amino acid sequence of CDR3
CDR3dnaseq	DNA sequence of CDR3
Inputseq	The input DNA sequence
PutativeGermline	The germline DNA rearrangement as determined by SoDA
AAseq	The translated input sequence
Quality	Quality Score of the input sequence
IgIsotype	For heavy chains, the constant region is identified
ID	This identifier is unique and can be null. It only given to a sequence if it is a part of a functional pair

controllers [87]. No studies have been done with acute patients due to the difficulty of diagnosing early stage HIV-1 infection. To determine the extent of clonal relatedness in each of the patients, we calculated a variation of the Simpson's index, a measure of diversity for antibodies from our patients [88]. This index was calculated for each patient i as

$$S_i = \sum_{j=1}^{k_i} [\frac{n_{ij}}{N_i}]^2 \tag{5.11}$$

where k_i is the total number of clones in patient i, n_{ij} is the number of antibody pairs in clone j of patient i and N_i is the total number of pairs from patient i. The higher the Simpson's index, the less diverse the antibody repertoire. We observed extraordinary clonal relatedness in the FV individual and 684-6, one of the AHI patients. The blood sample from this patient was drawn approximately 20 days post infection. Table 5.4.1 shows the extent of diversity in each of the patients. The

Simpson's index is artificially high when N is very small. Thus, an additional measure of diversity would be to look at the proportion of non-clonally related sequences in the patient. The higher this proportion, the more diverse the repertoire. For both tests, we see that the FV individuals and one of the day 20 acute HIV infection patients, 684-6 exhibited extraordinary clonal relatedness. Such clones are expected to expand in presence of an antigen. So, we wanted to look at what proportion of these clones were antigen specific. Table 5.4.1 shows the clones from our 20 day AHI patient. A similar table for DFLU07001 is available in Appendix B, Table B.4. We found that at least 1 antibody in all of the clones from the FV individual was antigen-specific with 8 out of 17 clones being 100% antigen specific. For the AHI patient, only 1 out of 6 clones was 100% gp41 specific. An additional 2 clones had a few gp41 specific antibodies. An interesting observation in the FV individual was that approximately 16% of the clones used VH4-59. VH4-59 usage has been seen in a previous study showing that usage increases as a function of age [89]. Another striking feature was that half of these 16% VH4-59 clones used VK1-39 for their light chain. About 50% of the clones in our 684-6 patient also utilize VK1-39. This may suggest a larger role of the light chain in antibody binding than previously thought. Next, we looked more closely at the 52 member clone from 684-6.

684-6 Clone

This set of clonally related sequences used VH3-7, JH5 and were all IgG3s. The light chains were VK1-39 and JK4. Figure 5.2 shows the phylogenetic relationship between these clonally related antibodies. The lengths of the branches represent number of nucleotide changes while the number on the branch represents the number of amino acid changes. The antibodies in this clone, especially the kappa chain is heavily mutated. There are two pieces of evidence that may suggest that this large clone is a pre-existing one and due to a cross reactive antigen in the setting of acute HIV

Table 5.3: Simpson's Index for all patients

Patient ID	Simpson's Index	N	Unique Antibodies	Proportion of unique pairs
0014	0.004	263	250	0.95
0223	0.005	195	192	0.98
0239	0.013	83	79	0.95
0650	0.01	96	89	0.92
0689	0.001	467	458	0.98
6817	0.008	126	124	0.98
684-6	0.06	215	141	0.66
DFLU07001	0.02	272	152	0.56
DFLU07021	0.07	41	27	0.66
DFLU07004	0.1	23	16	0.69
DFLU07024	0.016	116	98	0.84
KFLU08002	0.03	34	34	1.00
KFLU08003	0.006	139	138	0.99
KFLU08005	0.007	154	143	0.92
KFLU08007	0.03	40	40	1
KFLU08012	0.016	62	61	0.98
KFLU08013	0.013	92	86	0.93

Table 5.4: Clones in 684-6, an AHI patient 20 days post transmission

Patient	Number of pairs in Clone	VH	JH	Light Chain	VL	JL	% antigen specific
684-6	52	3-7	5	kappa	1-39	4	7.7
684-6	2	1-8	5	kappa	1-39	1	0
684-6	8	3-74	5	kappa	1-39	4	100
684-6	3	3-48	6	kappa	3-20	4	0
684-6	2	3-23	1	lambda	6-57	2	0
684-6	5	3-23	5	kappa	1-39	4	60
684-6	3	1-3	6	kappa	1-39	4	0
684-6	2	3-30	4	lambda	6-57	2	100
684-6	2	3-30	4	kappa	1-39	4	100
684-6	3	4-39	6	kappa	1-33	4	0
684-6	2	3-30-3	4	lambda	3-19	2	0

infection is being driven to affinity mature to react with gp41. First, if this was a HIV specific clone, we would expect to find unmutated antibodies or antibodies with very low mutations. We compared the distribution of mutations in the heavy chains of antibodies in this clone to 2 clones isolated from EBV infected cases [85]. Figure 5.3 shows that the for both EBV clones, we see a proportion of antibodies without any mutations as well as at low mutation rates. In the case of 684-6, the lowest mutated antibody gene was 2.5%, which is approximately 15 nucleotide mutations in the entire antibody gene compared to the inferred germline. The heavy chain had 5 (˜2%) mutations while the kappa had 15 mutations (˜3%). Second, all of the antibodies in this clone are class-switched to IgG3. One would expect presence of class-switched

antibodies in the FV individuals due to the presence of cross-reactive antibodies to previously encountered influenza strains. In an acute HIV-1 infection we would expect some non class-switched antigen specific antibodies from HIV-1. However, we do not see any such IgM or IgD clones suggesting a possible activation of memory cells specific for a different antigen. To test this hypothesis, we estimated the precursor sequences at the internal nodes of the tree using a maximum likelihood estimation using Phylip3.83 [56]. Antibodies were made from these intermediate sequences and their specificity was tested. We also observed that the germline antibody did not bind to gp41 and binding affinity for gp41 increased as more mutations developed. We saw a loss of binding in certain antibody pairs. If the clone was developed in response to gp41, we would expect that the germline antibody would bind weakly to the molecule. Figure 5.2 shows the phylogenetic relationship between the observed clonally related sequences. Intermediate sequences are marked and colored in red if they bind to gp41. Similarly observed sequences marked in red also displayed gp41 binding in an ELISA or Luminex assay. In a different analysis, the affinity of these antibodies was measured to gut flora and we found that some of the intermediate and observed sequences displayed cross-reactivity with gut flora (unpublished).

5.4.2 Polyclonal Activation

On studying the various specificities of the isolated antibodies, we found that an average of about 62% of the antibodies from the four FV individuals were influenza specific while only approximately 12% of those from the AHI patient antibodies were specific for any HIV antigens (Table 5.4.2). The rest were specific for a wide range of other antigens such as Gut Flora, autoantigens and untested antigens. Of all acute HIV-1 patients, 684-6 displayed most antigen specific antibodies at 15%. Although never been shown early (17 - 30 days) in infection, this polyclonal activation was supportive evidence to a different study that showed polyclonal B cell activation in

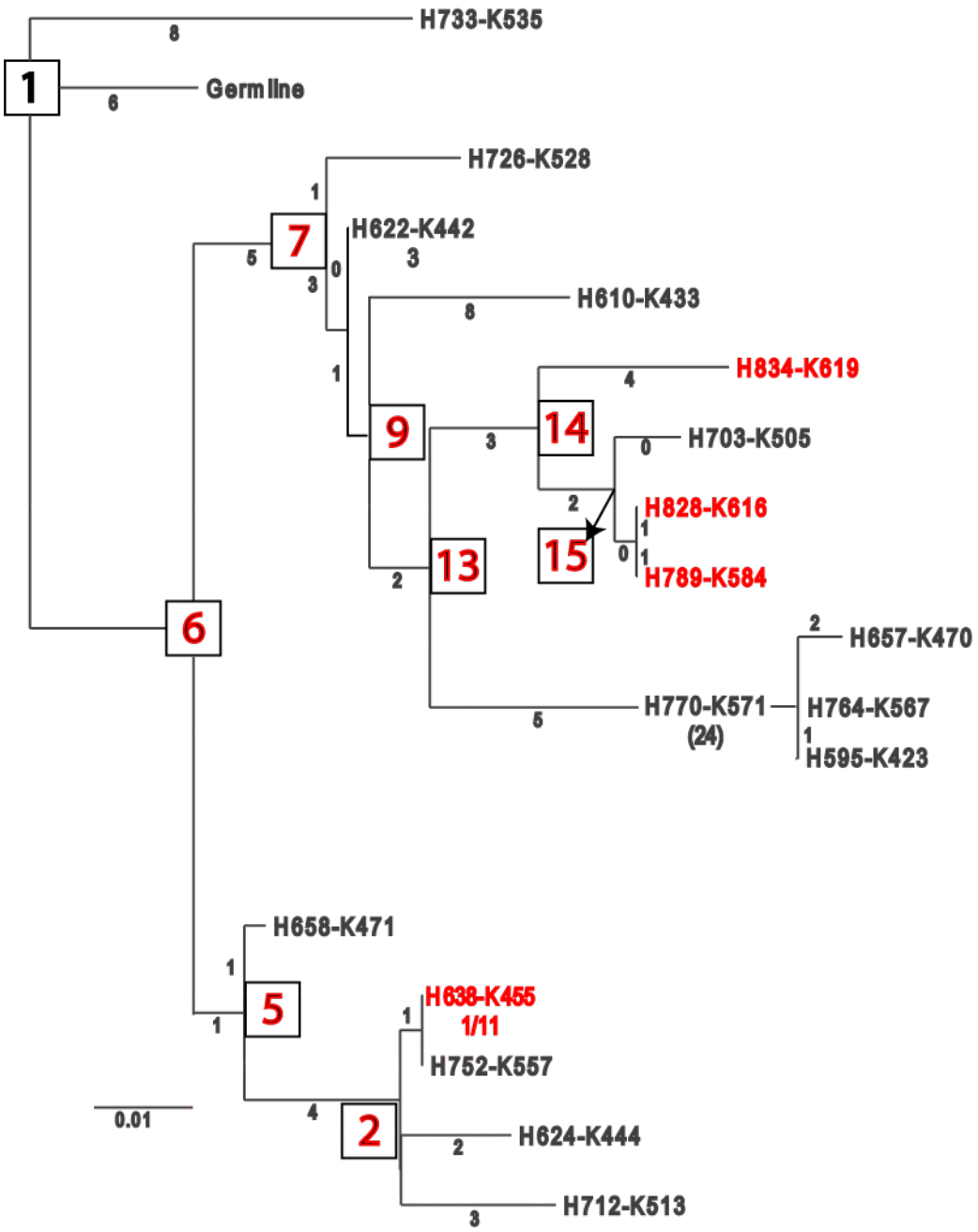


FIGURE 5.2: A phylogenetic relationship between a set of 52 clonally related antibodies from patient 684-6. The lengths of the branches represent number of nucleotide changes while the number on the branch represents the number of amino acid changes. The numbers in parenthesis indicate the number of identical pairs observed. The sequences at internal nodes are numbered in square boxes. Sequences marked in red showed binding to gp41.

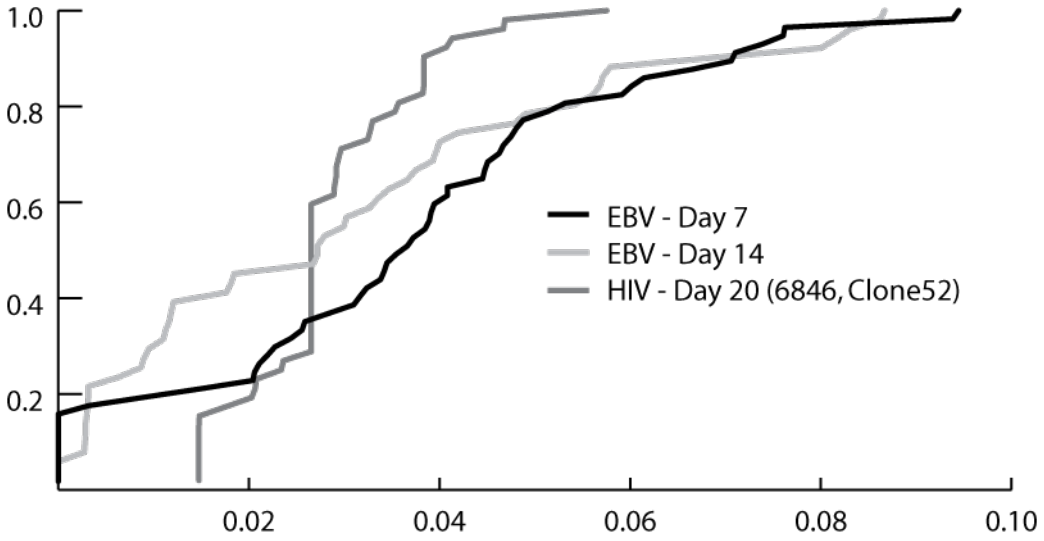


FIGURE 5.3: Comparison of distribution of mutations in 2 clones derived from EBV infected individuals and the 52 member clone from 684-6

subjects with an acute infection of less than 40 days [78]. A majority of the HIV response in AHI patients was directed towards gp41; none of the antibodies neutralized. This was expected as it has been shown that the first anti-HIV-1 antibody response was to gp41 and appeared 13 days after the appearance of plasma virus [90]. In contrast, envelope gp120-specific antibodies were delayed an additional 14 days. 3.7% of the antibodies in uninfected individuals were also HIV specific. We also see a higher percentage of antibodies specific for Gut (1.5%) and autoantigens (3.7%) in AHI than in other infections. This further suggests that a larger proportion of the initial antibody response in an influenza vaccination is targeted to the antigen while that in the acute HIV-1 infection, we see a polyclonal activation as early as day 20. In contrast to an influenza vaccination, we observe that only ˜14% of the antibody response in FIA individuals was influenza specific and 2.4% in FIS individuals. Thus we see an evidence of polyclonal activation not only in an HIV-1 infection but also in an influenza infection.

Table 5.5: Percentage of HIV-1, Influenza and other specific antibodies in the various groups of patients.

Patient Type	gp41	Other HIV	HA	Fluzone (Non-HA)	Gut	Auto
AHI	11.45%	0.45%	0%	1.53%	1.53%	3.70%
FIS	0%	0%	1.99%	0.40%	0%	0%
FIA	0%	0%	10%	3.46%	0%	0%
FV	0.45%	0%	50.33%	11.14%	0%	1.34%
Uninfected	3.70%	0%	0.37%	0.74%	0.37%	1.48%

5.4.3 Polyreactivity

A polyreactive antibody is one which binds to a wide variety of antigens usually at a low affinity. We studied the reactivities of each of our antibodies to see how many of these antibodies were polyreactive. We found the highest polyreactivity in our AHI patients where 2.34% of all antibodies were reactive for more than 1 antigen. This was significantly ($p < 0.05$) higher than the 0.4% we see in uninfected individuals and 0% in FIS and FIA. FV had also had a lower proportion of 1.1% but this was not statistically significant from AHI. Polyreactive IgG antibodies may be important in the early phase of infection since previous studies have shown that patients lacking these are more susceptible to bacterial infection [91].

We also looked at the reactivity of the HIV specific antibodies and found that over 7% of the gp41 specific antibodies from our AHI patients were polyreactive, over 20% of those specific for another HIV antigen were also polyreactive, usually with a part of the fluzone vaccine excluding HA. This polyreactivity in gp41 antibodies is significantly higher than what is seen for influenza specific (HA and Non-HA) antibodies in the FV, FIA and FIS patients. Table 5.4.3 summarizes these results.

Table 5.6: Percentage of antigens specific for the listed antigen in the listed patient group that also react with other antigens

Patient Group	HIV-gp41	HIV-other	HA	Fluzone (Non-HA)
AHI	7.90%	20%	NA	47%
FV	50%	NA	2.20%	10%
FIA	NA	0%	0%	0%
FIS	NA	0%	0%	0%
Uninfected	0%	NA	0%	0%

5.4.4 Ig Isotypes and Gene Segment Usage

Naive mature B cells produce both IgM and IgD, which are the first two isotype segments in the immunoglobulin locus. Hence, IgMs or IgDs are the first antibodies to appear in response to initial exposure to antigen. After activation by antigen, they undergo antibody class switching to produce IgG, IgA or IgE antibodies. During class switching, the constant region of the immunoglobulin heavy chain changes but the variable regions, and therefore antigen specificity, stay the same. The antibody thus retains affinity for the same antigens, but can interact with different effector molecules. IgG antibodies are predominately involved in the secondary immune response. Presence of specific IgG generally corresponds to maturation of the antibody response. IgA plays an important role in mucosal immunity while IgE primarily participates in responses to allergens. Our AHI and influenza infected individuals (FIA and FIS) both had a significantly higher proportion of IgM than the other two groups showing characteristics of a primary infection ($p < 0.01$, Table B.1). IgGs were the primary isotype observed in our FV and uninfected individuals at 70% and 48% respectively (Table B.1). Although only statistically significant when compared to FV and FIS individuals, we found a higher proportion of IgAs in our AHI than any other group (46%, Table B.1). gp41 specific antibodies from AHI are primarily IgA and IgG. The proportion of IgAs in HIV specific antibodies, both gp41 and other is significantly higher than seen in influenza specific antibodies seen in FV, FIA and

Table 5.7: Ig Isotype in antibodies of various specificities

IgIsotype	FIS		FV		FIA		AHI	
	HA	Non-HA	HA	Non-HA	HA	Non-HA	gp41	HIV-Other
IgA	0%	0%	14.60%	12%	23.08%	22.22%	41.73%	40%
IgG	60%	100%	84.07%	74%	61.54%	66.67%	42.52%	60%
IgM	40%	0%	0.88%	14%	15.38%	11.11%	14.96%	0%

FIS (see Table 5.4.4). Polyclonal activation in the gut as a result of HIV-1 has been previously observed in HIV-1 infected individuals and this study only reconfirms the finding [92].

VH gene segments are classified into seven families based on sequence homology. Each of these families consists of sub-families, which increases the genomic complexity of the locus. The germline usage for each of the VH families is calculated by the number of members in each family. Figure 5.4 shows the germline or expected usage of the gene segments compared to the usage in our AHI patients, FV patients, FIS and FIA patients and uninfected individuals. A majority of the VH usage agrees with germline usage. We see no difference between the groups for usage of VH1 and VH2. All groups show an over-usage of VH3 over germline. There was however a significant over-usage of VH4 in the FV patient (31%), specifically VH4-59 ($p < 0.01$). This is also true of the influenza specific antibodies from FV. A previous study has shown an over-usage of VH4 in nonproductive rearrangements [93]. Conversely, we found a significant under-usage of VH-4 in our AHI patients (16%) compared to uninfected individuals (23%). The expected germline usage of VH4 is ˜19%. We also found an over-usage of VH6 in the acute HIV-1 patients. The germline usage of VH6 is under 2% while that in acute HIV-1 patients (cumulative) is 4%. More specifically, patient 065-0 and 068-9 had a significant excess of VH6 with 6.4% and 9.7% respectively ($p < 0.01$). VH6 was also significantly overused in the AHI gp41 specific antibodies and autoreactive antibodies from all patients when compared to HA and Non-HA specific antibodies in FV. FIS and FIA have very few influenza specific antibodies

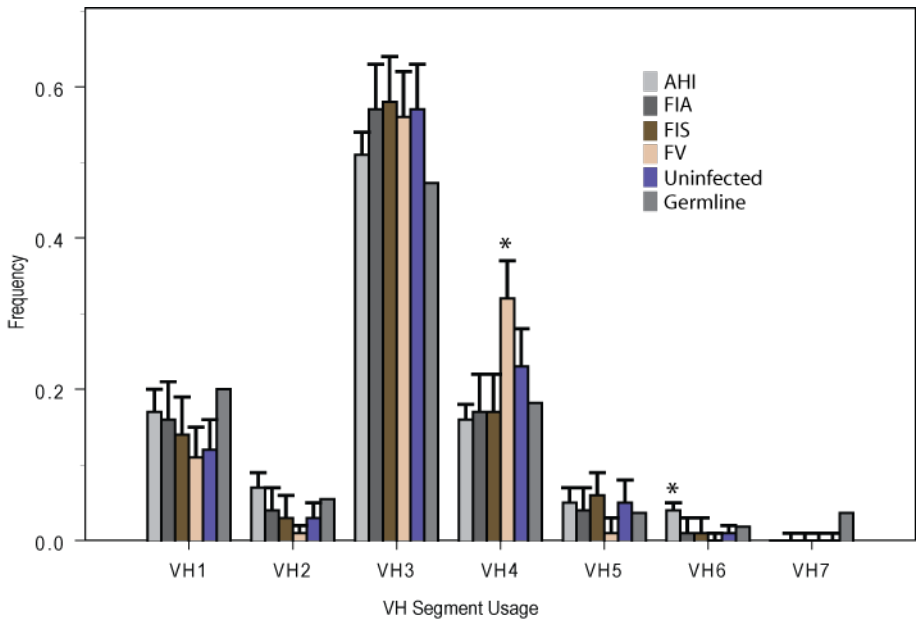


FIGURE 5.4: VH gene segment usage in all five groups compared to germline gene segment usage. (*) indicate statistical significance ($p < 0.01$).

and hence a valid comparison cannot be made. However, both groups have 0% VH6 usage. Additionally, all 5 of the non-gp41 HIV specific antibodies utilize VH3. Interestingly, VH1 and especially VH1-69 are significantly overused by Non-HA specific antibodies from FIA (Table B.1). VH1-69 is also overused in gp41 specific antibodies from AHI patients. A majority of the recently discovered broadly neutralizing antibodies to HA in influenza [94], well known HIV-1 antibody 4e10 [35] as well as antibodies to E2 on Hepatitis C [95]have been shown to use VH1-69. VH1-69 is the only VH gene that consistently encodes two hydrophobic residues at the tip of its CDR-H2 loop. It is the only germline gene to encode a phenylalanine at this position, which may play an important role in interacting with the HIV-1 envelope [96].

DH gene segments, like VH are classified into 8 families, DH0 - DH7. These gene segments are short and high somatic mutation and presence of n nucleotides makes it difficult to identify them with certainty. We compared the DH segment usage in the heavy chains of our patients. We found a significant over-usage of DH2 in all of

our infection groups and a particular over-usage of DH2-2 in one our FV individual, DLFU07001. ($p < 0.01$). Once again, this gene segment was found in excess in non-productive rearrangements in a previous study [93]. We also found that AHI patients used an excess of DH1 when compared to FIA individuals. There were no interesting biases in DH gene segment us by specificity (Figure B.1).

The JH locus is the simplest of all three gene segments in that it contains only 6 genes. JH gene segments are 16 - 20 amino acids long and encode the invariant tryptophan essential for the folding of the protein. Preferential usage of JH5 and JH6 in non-productive immunoglobulin rearrangements while that of JH4 in productive immunoglobulin rearrangements (Figure B.1) [93]. Once again, we found a significant under-usage of JH4 and over-usage of JH6 in FV patients (30% and 40% respectively) ($p < 0.01$) as well as in HA specific antibodies in FV patients. Due to this extensive over-usage of JH6, usage of all other JH segments in FV individuals was significantly lower than other groups We also found that AHI patients used significantly lower JH2 gene segments than FV and uninfected individuals (2% vs 5.3% and 4.5% respectively). HIV specific antibodies significantly overused JH5 gene segments (Table B.1).

5.4.5 Complementarity Determining Region 3

The complementarity determining region (CDR) is the part of the antibody that determines their specificity to a particular antigen. CDRs are subject to high somatic mutation and help antibodies recognize a vast repertoire of antigens. There are three CDRs - CDR1, CDR2 and CDR3 in the variable region of the heavy chain and an additional three in the light chain. Of these six, the CDR3s of both chains show the greatest variability by junctional diversity and somatic mutation. In most cases, the CDR3 loop is believed to be the one that makes antigenic contact. Therefore, it is important to look at differences in CDR3 length in our patients. We found

a mean CDR3H length of 14.84 amino acids (aa) in our AHI. This mean was not statistically different from the FIS individuals (mean of 14.78aa) and uninfected individuals (mean of 14.88aa). The gp41 specific and other HIV-1 specific antibodies in AHI had longer CDR3s than average at 14.9 and 16 amino acids respectively. FV individuals and FIA individuals had significantly higher means of 16.98aa and 15.6aa respectively (see Figure 5.5(a)). Antibodies specific for HA in influenza primarily from FV individuals but also from FIS and FIA had a significantly long CDR3 region (17.7aa) in the heavy chain. Kappa and lambda chains of HA specific antibodies from FIV are also significantly longer that all HIV specific antibodies from AHI (see Table B.2). There was also a significant difference in the distributions of the CDR3 lengths between these groups as measured by a Kolmogorov Smirnov Test (see Figure 5.5(b)).

Hydrophobic CDR3s are believed to be deleted in B cell tolerance mechanisms [97]. For example, broadly neutralizing mAbs 2F5 and 4E10 both have long hydrophobic CDR3 regions and have been observed of to bind to lipid autoantigens [98]. We first looked at the hydrophobicity of the full length CDR3H using the Kyte and DooLittle Hydropathy Index [99]. According to this scale, the CDR3H of 2F5 and 4e10 have a hydropathy index of -0.004 and -0.27 respectively. If we assume CDR3H is a loop where only the middle portion interacts with the antigen and calculate the hydrophobicity in the middle 3rd of the CDR3 loop, 2F5 and 4e10 have hydropathy indices of 1.3 and 0.014 respectively. From our patients, we found that CDR3H of the antibodies from FV individuals had the highest average hydropathy index at -0.34. This average was significantly higher than all other groups in a 2-paired T-test ($p < 10^{-7}$). Uninfected, FIA and FIS patients had the second highest and relatively equal hydropathy index at -0.57, -0.59 and -0.59 respectively. Even though we did not find any evidence of highly hyrdophobic CDR3H when compared with 2F5 and 4e10 in any of our patient groups, hydrophobicity in the middle 3rd of the CDR3 loop in FV and AHI was significantly higher than other groups at -0.02 and -0.13

respectively. HA specific antibodies had the most hydrophobic CDR3 regions, but this is expected since over 65% of antibodies in FV are HA specific. There were no other significant differences in CDR3 hydrophobicity based on specificity.

5.4.6 Somatic Mutations

Somatic mutations are accumulated in immunoglobulins during affinity maturation. It is an important process for increasing the diversity of the antibody repertoire. As described in Chapter 1 when B cells are activated by an antigen, they form germinal centers and proliferate. At this time AID is activated, which induces point mutations in the antibody genes at a rate of about one nucleotide substitution per division. This enhances the affinity of the antibody to the antigen. We found that our uninfected individuals had the highest mutation frequency of 6.6% in the heavy chains, 4.75% in lambda, and 4.52% in kappa chains (Table B.3). A high mutation rate in plasma cells of an individual with no disease symptoms or an acute infection would be expected since these would be from pre-existing activated memory cells and these are typical of the values observed [100]. We also observed a higher mutation rate in the heavy chains in our AHI and FIS individuals than in the FV and FIA individuals. Figure 5.6(a) shows the cumulative distribution of the mutations in the VH regions of the groups while Figure 5.6(b) shows the p values for the difference in average mutation frequency in the heavy chains.

Table 5.4.6 shows the average mutation frequencies and the standard errors for antibodies specific for the various antigens. Heavy chains for gp41 specific antibodies were heavily mutated at 5%. Kappa and lambda chains involved in gp41 binding were also substantially mutated at 4.9% and 4.6% respectively (see Table B.3). All influenza specific antibodies also showed substantially high mutation in their heavy and light chains. The significantly high mutation frequency in antigen specific antibodies supports the previously mentioned hypothesis of activation of previously

Table 5.8: Average mutation frequency and standard error in heavy chains of antibodies of various specificities

Patient Group	Specificity	Mean Mutation Frequency	Standard Error
AHI	HIV-other	7.3%	0.7%
AHI	gp41	5%	0.3%
FIA	HA	6.7%	0.4%
FIA	Non-HA	5.6%	1.2%
FIS	HA	5.7%	1.3%
FIS	Non-HA	7.6%	NA
FV	HA	5.2%	0.2%
FV	Non-HA	5.4%	0.4%

existing B cells in both HIV and influenza infections.

Mutations can be either synonymous or non-synonymous. Synonymous mutations are those that do not change the amino acid while non synonymous mutations cause an amino acid change. Usually, most non-synonymous changes would be expected to be eliminated by purifying selection, but under certain conditions positive selection may lead to their retention. Investigating the number of synonymous and non-synonymous substitutions may therefore provide information about the degree of selection operating on these antibodies. Therefore, we look at the frequency of these mutations in the heavy chains in our patients. We further classify the heavy chains into CDR and framework regions. Table 5.4.6 shows the synonymous to non-synonymous mutation frequency ratio. The FIS patients has a significantly smaller ratio than all other groups suggesting that antibodies from these patients are under higher positive selection. Alternatively, uninfected and FV patients show the highest ratio indicating a lower selection force than FIS. This could be explained by activation of memory B cells with cross-reactivity. Lastly, the AHI patients shows a dS/dN ratio significantly higher than FIS but lower than the other three groups in the heavy, kappa and lambda chains. A previously reported high mutation rate (5.6%) with this dS/dN ratio could suggest an activation of previously triggered non-HIV-1

Table 5.9: Ratio of Synonymous to Non-Synonymous (dS/dN) mutations in the entire sequence, CDRs and the Framework regions of heavy, kappa and lambda chains

	Heavy			Kappa			Lambda		
Patient	Total	CDR	FR	Total	CDR	FR	Total	CDR	FR
AHI	0.43	0.28	0.53	0.47	0.26	0.61	0.44	0.25	0.58
FIA	0.46	0.27	0.59	0.51	0.27	0.69	0.45	0.27	0.59
FIS	0.40	0.25	0.50	0.45	0.24	0.61	0.36	0.18	0.52
FV	0.45	0.25	0.59	0.46	0.19	0.67	0.43	0.30	0.56
Uninfected	0.49	0.29	0.61	0.58	0.29	0.79	0.54	0.26	0.76

memory B cells, mildly cross-reactive with HIV-1 and undergoing positive selection in the presence of the antigen. AHI and FIS have a significantly ($p < 0.01$) lower dS/dN ratio in their framework regions of heavy, kappa and lambda chains suggesting stronger positive selection in those regions (see 5.4.6). Framework regions play an important role in the folding and structure of the antibodies. Larger proportions of non synonymous mutations in these regions is usually unexpected. However, selective forces acting on the B cells in the presence of an active replicating antigen are seen in both regions in the case of AHI and FIS. No such difference was seen in the CDR regions. Tables B.3 and B.3 in Appendix B show the absolute number of mutations in each category.

5.5 Conclusions

In this chapter, we characterize the initial plasma cell response during an acute HIV infection and make a comparison of this response to one in an influenza vaccination and influenza infection. In doing so, we made the following three observations.

- First, acute HIV-1 infection and acute influenza infection show evidence of polyclonal activation and the production of non-pathogen directed antibodies. This conclusion is supported by the fact that only 12%, 14% and 2.4% of the response in AHI, FIA and FIS patients was antigen specific. In contrast over 60% of the response in FV patients was antigen specific. We not only see a very small proportion of antigen specific antibodies in the infected patients but also

observe antibodies to gut flora, autoantigens and other antigens, especially in AHI. We also observed that the number of unmutated heavy chains in AHI, FIS and FIA were 4.63%, 2.64% and 0.78% respectively. While only 4% of the heavy chains in an EBV infected patient at day 3 were unmutated, the proportion rose to 15% in the day 14 patient [85]. This supports evidence of an antigen-directed response in an EBV infection at day 14 versus a polyclonal response in the AHI, FIA and FIS. Additionally we observe significantly shorter CDR3s in the heavy chains from AHI and FIS compared to FIA and FV. Both FV individuals and FIA individuals are able to control the viral infection effectively. Previously and recently discovered broadly neutralizing antibodies to HIV-1 have been shown to have long CDR3s in the heavy chain [35] [36]. 4 out of 6 broadly neutralizing antibodies to influenza also have a CDR3H length of greater than 15aa [94]. This may suggest that long CDR3s may be essential for effective clearance of the virus. The short CDR3 in the AHI and FIS individuals may suggest that prolonged exposure of the virus may be necessary to see a plasma cell repertoire with long CDR3s in heavy chains.

- Second, we see evidence that an acute HIV-1 infection may hamper B cell tolerance mechanisms. We observe that the center of the CDR3 loop in the heavy chains of the AHI patients is significantly more hydrophobic than those in FIA, FIS or uninfected individuals. A biased usage of VH6 seen in AHI patients, HIV specific antibodies and autoreactive antibodies has been previously observed in B cell acute lymphoblastic leukemia samples [101], Autoimmune Idiopathic Thrombocytopenic Purpura and in human insulin-dependent diabetes mellitus [102]. Lastly, 3.7% of the antibodies from AHI patients are autoreactive, higher than any other group.

- Finally, the possibility exists that a component of the intial antibody response

to gp41 is a cross-reactive response involving pre-existing memory B cells to an unrelated antigen. Antibodies from AHI patients have significantly higher mutation frequency than FIA and FV individuals and although not significant higher than FIS individuals. In particular, antibodies to gp41 from AHI patients had an average mutation frequency of 5% and those to other HIV-1 antigens had a 7.3% mutation frequency in the heavy chain. We also did not find any unmutated antibodies in our clonally related sequences unlike the clones found from the EBV patients. Additionally, some members of the 684-6 clone displayed cross-reactivity to gut flora (unpublished).

Taken together, these data suggest that HIV induces massive polyclonal activation, a possible breakdown of tolerance mechanisms, an activation of pre-existing humoral responses and demonstrates the profound perturbation of the B cell arm of the immune system soon after HIV-1 transmission.

(a)

	Uninfected	FIS	FIA	FV
AHI	0.92	0.73	0.01	0.00
FV	0.01	0.00	0.01	NA
FIA	0.04	0.02	NA	
FIS	0.69	NA		

(b)

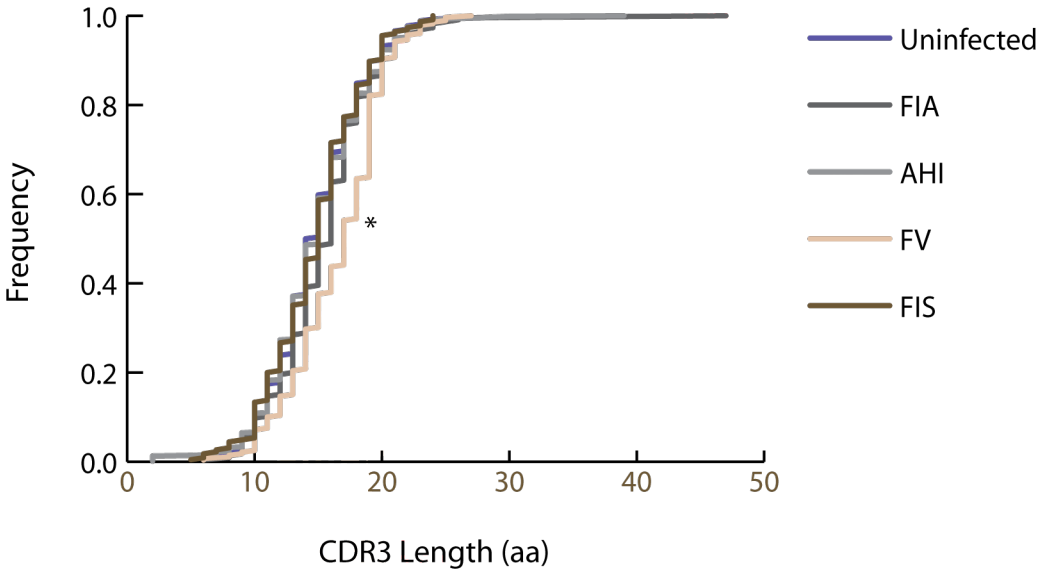
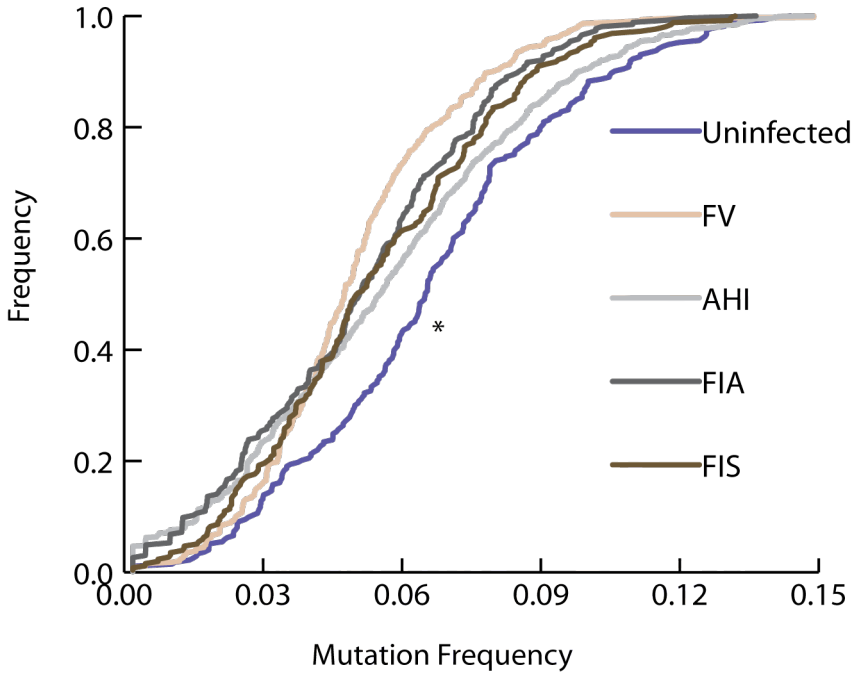


FIGURE 5.5: (a) p values for a 2-paired T-test to see a difference in mean CDR3H length between the different groups. Values significant at $p < 0.05$ are marked in red (b) Cumulative distribution functions for the number of CDR3H length of the five groups. star indicates that the two distributions have a p value of $p < 0.05$ when compared to AHI in a Kolmogorov Smirnov Test

(a)



(b)

	Uninfected	FIS	FIA	FV
AHI	**0.000**	0.189	**0.007**	**0.000**
FV	**0.000**	**0.073**	0.756	na
FIA	**0.000**	0.215	na	
FIS	**0.000**	na		

FIGURE 5.6: (a) Cumulative distribution functions for the number of mutations in the VH region of the five groups. (b) p values for a 2-paired T-test to see a difference in mean number of mutations between the different groups. Values significant at $p < 0.01$ are marked in red.

6

Conclusions & Future Directions

This thesis presents three distinct yet closely related topics: diversification in HIV-1, *Strongylocentrotus purpuratus* and in human B cells. All three employ various strategies including recombination and point mutation to be successful in their function.

In Chapter 2 we found the presence of HIV-1 recombinants in chronically infected homosexual men who were classified as long term non-progressors [52]. The emergence of recombination was correlated with the time of the infection such that we found more recombinants towards the later stage of infection than in the beginning. Intra-patient recombination has been associated with the emergence of co-receptor switched viruses [103] and development of anti-viral drug resistance [104]. One study found recombinants in a patient with rapid progression to AIDS [105]. However, this study was done on a single patient, hence a cause and effect conclusions cannot be made. Another follow-up study suggested an association between dual HIV-1 infection and rapid disease progression [106]. Whether this progression is due to emergence of recombinants or higher pressure on the immune system to develop responses to two rapidly evolving viruses is unknown. Using our method for a large scale longitudinal study would be helpful in determining the association of recombinants to

various clinical and behavioral factors. Since a root is necessary for phylogenetic analysis with recombination, one would need patients where the transmitted virus is known. For example, one can assume that the sample of viruses found in our AHI patients in Chapter 5 have a known transmitted virus. If unavailable, we can also use the virus/es at the time of first sampling as the root. Periodic sequencing of the virus in addition to tracking clinical factors such as evidence of drug resistance, CD4 counts and viral load and behavioral factors such as drug use and sexual history would enable us to design association studies between these factors.

In the next chapter, we presented a study on an interesting gene set from an invertebrate system. We found evidence of recombination and point mutations in *S. purpuratus*, an organism that lacks an adaptive immune response. A follow up for this study would be to identify proteins coded by the genes as well as to identify the genomic origin for the mRNA segments in the now sequenced *S. purpuratus* genome. To identify the function of the protein, one could start by implementing computational prediction methods [107]. A putative function can be assigned this way. Designing knock-out sea-urchin models or expressing the protein in a simpler system such as bacteria is another possibility. For genomic origin of the transcripts, there are two possible hypothesis. The first is that each of the ESTs found in the study is accounted for in the genome and that over time the genes have arisen through duplication and recombination. The second, more revolutionary idea would be the presence of repeats as germline gene segments in the genome. This would suggest a rearrangement mechanism as seen in vertebrates [21]. In order to determine which one of these is correct, we could use tools such as BLAST and other sequence alignment and phylogenetic methods from the already sequenced genome. An alternative mechanism would be to sequence these genes from the genome by designing primers from the various repeat elements.

Finally in Chapters 4 and 5, we developed methods and analyzed the plasma cell

repertoire in patients with an acute HIV-1 infection. Although a very important study in itself, the use of these methods could be applied to answer many different questions. One area of interest is to study the the interplay between the evolution of the virus with the evolution of the antibody repertoire. One could use methods developed in Chapter 2 combined with the ideas and methods of the Antibodyome project for such a study. This would enable us to study the changing plasma cell repertoire to the changing epitopes in HIV-1. Although it may take several years, a study of this kind would allow us to tackle issues such as the differences between elite controllers and rapid progressors. Sorting memory B cells from AHI over time would allow us to determine which of the gp41 specific plasma cells are pre-existing or are saved as memory cells at a later time point. Since early samples are already collected for the study in Chapter 5, the study proposed above would require sampling virus and sorting memory and plasma cells at different time points throughout the infection. Such a study would provide valuable information to fill gaps in our basic understanding of HIV-1 infection. It will also aid in better drug and vaccine design. In addition to designing new studies, one cannot forget the mediocre yet positive results from the recently conducted vaccine trial of ALVAC and AIDSVAX in Thailand. Despite the success of this phase IIB trial, the immune response to this combination vaccine is unknown. With the methods and analysis pipeline developed in Chapter 4 and 5, a high throughput study of the antibody response to this vaccine would be in order. In addition, the ongoing search for broadly neutralizing antibodies can also be facilitated with such methods by a study on elite controllers.

Appendix A

SoDA2

A.1 Calculating the total probability of the alignment

The main aim of modeling this system as an HMM is to obtain the full probability of observing a sequence of recombination events i.e. $P(x, y)$ where x is the antibody sequence and y is the putative germline. The full probability is simply the sum of all possible alignment paths so that $P(x, y) = \sum_{\pi} P(x, y, \pi)$. This is calculated by summing over all possible paths in the forward and backward direction. Instead of maximizing at each step, we calculate the total probability of the alignment. The forward and backward algorithms are explained in this section. The abbreviations used in the explanation are given below:

- m - length of the antibody gene
- vn - length of the V segment we're aligning against
- dn - length of the D segment. In case of a light chain $dn = 0$.
- jn - length of the J segment
- $P_e(x_i, y_j|k)$ - emission probability at x_i and y_j given state k.

- $P_t(l|k, j-1)$ - transition probability of moving into state l at position j when position j - 1 is in state k.

Our states are as follows:

- Match/Mistmatch state in V gene (Mv)
- Insertion in V gene (Iv)
- Deletion in V gene (Dv)
- V-D junction n nucleotides (N1)
- Match/Mismatch in D gene (Md)
- Insertion in the D gene (Id)
- Deletion in the D gene (Dd)
- D-J junction n nucleotides (N2)
- Match/Mismatch in the J gene (Mj)
- Insertion in the J gene (Ij) and Deletion in the J gene (Dj).

A.1.1 The forward algorithm

We define forward matrices F for each of the states. We set $F_{Mv}[1,1] = 1$ and the rest are set to 0. The alignment must begin in the V match/mismatch state.

FOR i in $x_{1..m}$

FOR j in $y_{1..vn}$ (except $x_1 y_1$) #We first start in the V region

FOR k = Mv, Iv & Dv, do the following:

$$F_{Mv}[i,j] * = P_e(x_{i-1}, y_{j-1}|Mv) F_k[i-1, j-1] P_t(Mv|k, j-1)$$

$$F_{Iv}[i,j]* = P_e(x_{i-1}, -|Iv)F_k[i,j-1]P_t(Iv|k,j-1)$$

$$F_{Dv}[i,j]* = P_e(-, y_{j-1}|Dv)F_k[i-1,j]P_t(Dv|k,j)$$

END FOR

Move into N1 from Mv

$$F_{N1}[i]* = P_e(x_{i-1}|N1)[F_{Mv}[i-1,j-1]P_t(N1|Mv,j-1)$$

END FOR

Move into the Md (heavy) or Mj (light) from the Mv

FOR z in $y_{0..dn}$ or $y_{0..jn}$

$$F_{Md}[i,z]* = P_e(x_{i-1}, y_{z-1}|Md)F_{Mv}[i-1,j-1]P_t(Md|Mv,j-1)$$

$$F_{Mj}[i,z]* = P_e(x_{i-1}, y_{z-1}|Mj)F_{Mv}[i-1,j-1]P_t(Mj|Mv,j-1)$$

END FOR

Next, we account for moving within the N1 state where you can only make a linear movement from $N1_{i-1}$ to $N1_i$.

FOR j in $x_{1..i}$

$$F_{N1}[i]* = P_e(x_{i-1}, n|N1)F_{N1}[i-1]P_t(N1|N1,j-1)$$

END FOR

Next, we account for moving from $N1$ to Md.

Move into the Md (heavy) or Mj (light) from N1

FOR z in $y_{0..dn}$

$$F_{Md}[i,z]* = P_e(x_{i-1}, y_{z-1}|Md)F_{N1}[i-1]P_t(Md|N1,i-1)$$

END FOR

#We move on to the D region

FOR j in $y_{1..dn}$ (except x_1y_1)

FOR k = Md, Id & Dd, do the following:

$F_{Md}[i,j]* = P_e(x_{i-1}, y_{j-1}|Md)F_k[i-1,j-1]P_t(Md|k,j-1)$

$F_{Id}[i,j]* = P_e(x_{i-1}, -|Id)F_k[i,j-1]P_t(Id|k,j-1)$

$F_{Dd}[i,j]* = P_e(-, y_{j-1}|Dd)F_k[i-1,j]P_t(Dd|k,j)$

END FOR

Move into N2 from Md

$F_{N2}[i]* = P_e(x_{i-1}|N2)[F_{Md}[i-1,j-1]P_t(N2|Md,j-1)$

Move into the Mj from Md

FOR z in $y_{0..jn}$

$F_{Md}[i,z]* = P_e(x_{i-1}, y_{z-1}|Mj)F_{Md}[i-1,j-1]P_t(Mj|Md,j-1)$

END FOR

END FOR

Next, we account for moving within the N2 state where you can only make a linear movement from $N2_{i-1}$ to $N2_i$.

FOR j in $x_{1..i}$

$F_{N2}* = P_e(x_{i-1}, n|N2)F_{N2}[i-1]P_t(N2|N2,j-1)$

END FOR

Next, we account for moving from $N2$ to Mj.

Move into Mj from the N2

FOR z in $y_{0..jn}$

$F_{Mj}[i,z]* = P_e(x_{i-1}, y_{z-1}|Mj)F_{N2}[i-1]P_t(Mj|N2,i-1)$

END FOR

We move on to the J region. Once here, we stay here

FOR j in $y_{1..jn}$ (except x_1y_1)

FOR k = Mj, Ij & Dj, do the following:

$F_{Mj}[i,j]* = P_e(x_{i-1}, y_{j-1}|Mj)F_k[i-1, j-1]P_t(Mj|k, j-1)$

$F_{Ij}[i,j]* = P_e(x_{i-1}, -|Ij)F_k[i, j-1]P_t(Ij|k, j-1)$

$F_{Dj}[i,j]* = P_e(-, y_{j-1}|Dj)F_k[i-1, j]P_t(Dj|k, j)$

END FOR

END FOR

END FOR

The total probability in the forward direction ($P(x, y)$) is stored in $F_{Mj}[m, jn]$, the only position the HMM is allowed to end.

A.1.2 The backward algorithm

We define backward matrices B for each of the states. We set $B_{Mv}[m, jn] = 1$ and the rest are set to 0.

FOR i in $x_{m-1..0}$

#Starting in the J region

FOR j in $y_{jn-1..0}$

#To be in Mj at i, j, you can come from Mj, Ij or Dj

$B_{Mj}[i,j]* = P_e(x_{i+1}, y_{j+1}|Mj)B_{Mj}[i+1, j+1]P_t(Mj|Mj, j+1)$

$B_{Mj}[i,j]* = P_e(x_{i+1}, -|Ij)B_{Ij}[i, j+1]P_t(Mj|Ij, j+1)$

$B_{Mj}[i,j]* = P_e(-, y_{j+1}|Dj)B_{Dj}[i+1, j]P_t(Mj|Dj, j)$

#To be in Ij at i, j, you can come from Mj or Ij

$B_{Ij}[i,j]* = P_e(x_{i+1}, y_{j+1}|Mj)B_{Mj}[i+1, j+1]P_t(Ij|Mj, j+1)$

$B_{Ij}[i,j]* = P_e(x_{i+1}, -|Ij)B_{Ij}[i, j+1]P_t(Ij|Ij, j+1)$

#To be in Dj at i, j, you can come from Mj or Dj

$B_{Dj}[i,j]* = P_e(x_{i+1}, y_{j+1}|Mj)B_{Mj}[i+1,j+1]P_t(Dj|Mj,j)$

$B_{Dj}[i,j]* = P_e(-, y_{j+1}|Dj)B_{Dj}[i+1,j]P_t(Mj|Dj,j)$

Moving to Mj from N2 is also allowed

$B_{N2}[i]* = P_e(x_{i+1}, n|N2)B_{Mj}[i+1,j]P_t(Mj|N2,j+1)$

END FOR

You can also move from the N2 region to the N2 region

FOR j in $x_{m-1..0}$

$B_{N2}[i]* = P_e(x_{i+1}, n|N2)B_{N2}[i+1]P_t(N2|N2,j+1)$

END FOR

#Moving on to the D region

FOR j in $y_{dn-1..0}$

#To be in Md at i, j, you can come from Md, Id, Dd or any position in Mj

$B_{Md}[i,j]* = P_e(x_{i+1}, y_{j+1}|Md)B_{Md}[i+1,j+1]P_t(Md|Md,j+1)$

$B_{Md}[i,j]* = P_e(x_{i+1}, -|Id)B_{Id}[i,j+1]P_t(Md|Id,j+1)$

$B_{Md}[i,j]* = P_e(-, y_{j+1}|Dd)B_{Dd}[i+1,j]P_t(Md|Dd,j)$

FOR z in $y_{jn-1...0}$

$B_{Md}[i,j]* = P_e(x_{i+1}, y_{j+1}|Md)B_{Mj}[i+1,j+1]P_t(Md|Mj,j+1)$

END FOR

#To be in Id at i, j, you can come from Md or Id

$B_{Id}[i,j]* = P_e(x_{i+1}, y_{j+1}|Mj)B_{Mj}[i+1,j+1]P_t(Ij|Mj,j+1)$

$B_{Id}[i,j]* = P_e(x_{i+1}, -|Ij)B_{Ij}[i,j+1]P_t(Ij|Ij,j+1)$

#To be in Dj at i, j, you can come from Mj or Dj

$B_{Dd}[i,j]* = P_e(x_{i+1}, y_{j+1}|Md)B_{Md}[i+1,j+1]P_t(Dd|Md,j)$

$B_{Dd}[i,j]* = P_e(-, y_{j+1}|Dd)B_{Dd}[i+1,j]P_t(Md|Dd,j)$

Moving to Md from N1 is also allowed

$B_{N1}[i]* = P_e(x_{i+1}, n|N1)B_{Md}[i+1,j]P_t(Md|N1,j+1)$

END FOR

You can also move from the N1 region to the N1 region

FOR j in $x_{m-1..0}$

$B_{N1}[i]* = P_e(x_{i+1}, n|N1)B_{N1}[i+1]P_t(N1|N1,j+1)$

END FOR

FOR j in $y_{dn-1..0}$ #Moving on to the V region

#To be in Md at i, j, you can come from Mv, Iv, Dv or any position in Md

$B_{Mv}[i,j]* = P_e(x_{i+1}, y_{j+1}|Mv)B_{Md}[i+1,j+1]P_t(Mv|Mv,j+1)$

$B_{Mv}[i,j]* = P_e(x_{i+1}, -|Iv)B_{Iv}[i,j+1]P_t(Mv|Iv,j+1)$

$B_{Mv}[i,j]* = P_e(-, y_{j+1}|Dv)B_{Dv}[i+1,j]P_t(Mv|Dv,j)$

FOR z in $y_{dn-1...0}$

$B_{Mv}[i,j]* = P_e(x_{i+1}, y_{j+1}|Mv)B_{Mv}[i+1,j+1]P_t(Mv|Md,j+1)$

END FOR

#If it is a light chain, you can go from Mj to Mv

FOR z in $y_{jn-1...0}$

$B_{Mv}[i,j]* = P_e(x_{i+1}, y_{j+1}|Mv)B_{Mv}[i+1,j+1]P_t(Mv|Mj,j+1)$

END FOR

#To be in Id at i, j, you can come from Md or Id

$B_{Iv}[i,j]* = P_e(x_{i+1}, y_{j+1}|Mj)B_{Mj}[i+1,j+1]P_t(Ij|Mj,j+1)$

$B_{Iv}[i,j]* = P_e(x_{i+1}, -|Ij)B_{Ij}[i,j+1]P_t(Ij|Ij,j+1)$

#To be in Dj at i, j, you can come from Mj or Dj

$B_{Dv}[i,j]* = P_e(x_{i+1}, y_{j+1}|Mv)B_{Mv}[i+1, j+1]P_t(Dv|Mv, j)$

$B_{Dv}[i,j]* = P_e(-, y_{j+1}|Dv)B_{Dv}[i+1, j]P_t(Mv|Dv, j)$

END FOR

END FOR

The total probability in the backward direction ($(P(x,y))$) is stored in $B_{Mv}[0,0]$, the only position the HMM is allowed to start. The total probability in the forward and backward directions should be equal.

A.1.3 Alignment with Highest Posterior Probability

The forward and backward algorithms gave us the total probability of the alignment. The posterior probability of a given alignment path is given by: $P(\pi_{ij} = k|x,y) = \frac{f^k ib^k(i)}{P(x,y)}$. We define posterior probability matrices for each state using the above equation. In order to find the path with the highest posterior probability, we use the Posterior Viterbi Algorithm [76]. Starting at the first position of the target and a match/mismatch in v, the algorithm goes forward, selecting the step with the highest posterior probability. The path π selected at the end is the one that has the overall highest score.

Appendix B

Antibodyome

B.1 Ig Isotype & Gene Segment Usage

Table B.1: Ig Isotype in antibodies from various groups

	AHI	FIA	FV	FIS	Uninfected
IgA	46.22%	39.56%	19.35%	33.33%	43.18%
IgD	0.71%	0.45%	0.44%	0.36%	0.38%
IgG	36.63%	35.11%	70.25%	48.44%	48.11%
IgM	16.43%	24.89%	9.68%	18.22%	8.33%

Table B.2: VH gene segment usage by specificity

	FIS	FIS	FIA	FIA	FV	FV	AHI	AHI
	HA	Non-HA	HA	Non-HA	HA	Non-HA	HIV-other	gp41
VH1	80%	0%	22.22%	11.11%	3.98%	2%	0%	10.24%
VH1-69	0%	0%	11.11%	0%	1.33%	0%	0%	6.30%
VH2	20%	0%	44.44%	0%	3.54%	0%	0%	4.72%
VH3	0%	0%	22.22%	55.56%	42.04%	54%	100%	56.69%
VH4	0%	100%	0%	11.11%	49.12%	44%	0%	17.32%
VH5	0%	0%	0%	22.22%	0%	0%	0%	0%
VH6	0%	0%	0%	0%	0%	0%	0%	4.72%

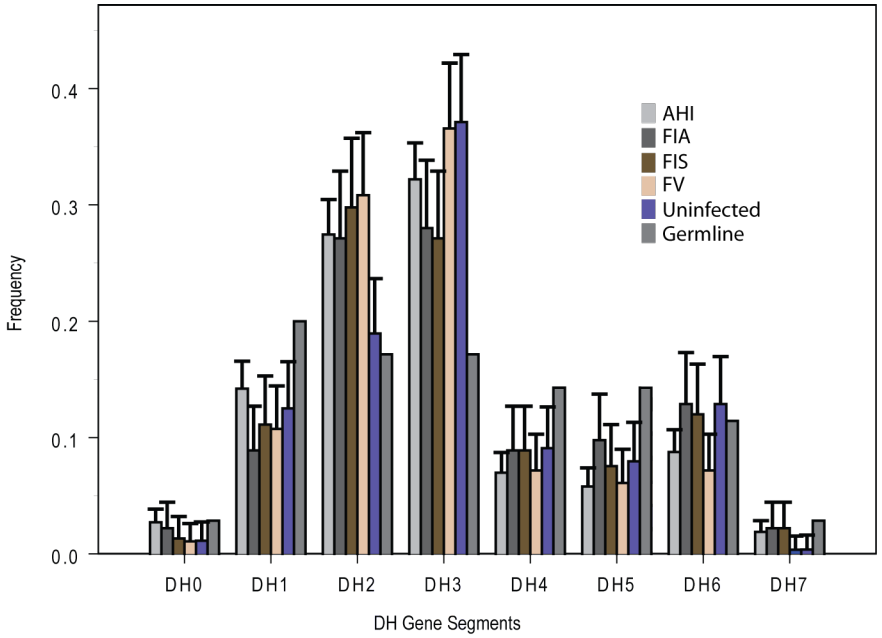


FIGURE B.1: DH gene segment usage in all five groups compared to germline gene segment usage.

Table B.3: DH gene segment usage by specificity

	FIS		FIA		FV		AHI	
	HA	Non-HA	HA	Non-HA	HA	Non-HA	gp41	HIV-other
DH0	0%	0%	3.85%	0%	0%	0%	0%	0.79%
DH1	20%	100%	3.85%	0%	11.06%	6%	0%	11.81%
DH2	20%	0%	23.08%	11.11%	34.96%	42%	60%	24.41%
DH3	20%	0%	46.15%	33.33%	33.19%	38%	40%	41.73%
DH4	40%	0%	7.69%	11.11%	10.62%	6%	0%	5.51%
DH5	0%	0%	7.69%	44.44%	6.19%	0%	0%	3.94%
DH6	0%	0%	7.69%	0%	3.98%	8%	0%	4.72%
DH7	0%	0%	0%	0%	0%	0%	0%	7.09%

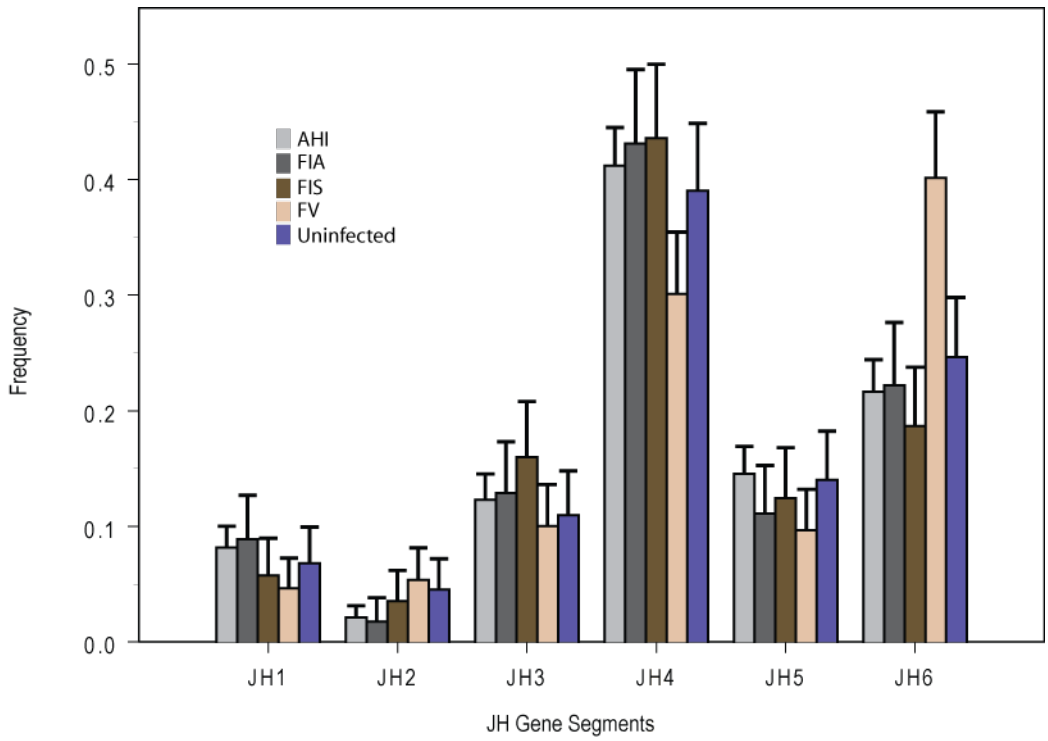


FIGURE B.2: JH gene segment usage in all five groups. There are only six JH segments, making germline usage equal.

Table B.4: JH gene segment usage by specificity

	FIS		FIA		FV		AHI	
	HA	Non-HA	HA	Non-HA	HA	Non-HA	gp41	HIV-other
JH1	0%	0%	3.85%	22.22%	3.98%	0%	20%	9.45%
JH2	0%	0%	0%	0%	2.65%	2%	0%	0.79%
JH3	20%	0%	23.08%	22.22%	10.62%	14%	20%	11.81%
JH4	60%	100%	34.62%	22.22%	18.58%	30%	20%	32.28%
JH5	20%	0%	11.54%	0%	5.75%	2%	40%	25.20%
JH6	0%	0%	26.92%	33.33%	58.41%	52%	0%	20.47%

B.2 CDR3 length

Table B.5: Length of CDR3 in heavy, kappa and lambda chains of antibodies of various specificities

Patient Group	Antigen Specificity	Heavy	Kappa	Lambda
AHI	gp41	14.9 ± 0.4	9	10.4 ± 0.3
	HIV-other	16 ± 0.8	9.1 ± 0.1	10.5 ± 0.6
FV	HA	17.7 ± 0.2	9.4 ± 0.1	11.1 ± 0.1
	Non-HA	16.9 ± 0.5	9.6 ± 0.5	11.2 ± 0.1
FIA	HA	15.4 ± 0.7	9.1 ± 0.1	10.9 ± 0.1
	Non-HA	15.6 ± 1.5	9.1 ± 0.1	10
FIS	HA	14.8 ± 0.4	8.8 ± 0.5	NA
	Non-HA	13	NA	12

B.3 Somatic Mutations

Table B.6: Mutation Frequencies in Heavy, Kappa and Lambda chains for all patients

	Heavy	Kappa	Lambda
FIS	$5.29\% \pm 0.17\%$	$3.32\% \pm 0.2\%$	$3.27\% \pm 0.37\%$
FIA	$4.98\% \pm 0.18\%$	$2.82\% \pm 0.18\%$	$2.8\% \pm 0.22\%$
AHI	$5.59\% \pm 0.1\%$	$4.22\% \pm 0.12\%$	$3.68\% \pm 0.13\%$
FV	$4.92\% \pm 0.12\%$	$3.59\% \pm 0.11\%$	$3.21\% \pm 0.11\%$
Uninfected	$6.6\% \pm 0.2\%$	$4.75\% \pm 0.21\%$	$4.52\% \pm 0.27\%$

Table B.7: Mutation Frequencies in Kappa and Lambda chains by specificity

Patient Group	Antigen Specificity	Mean Mutation Frequency (kappa)	Std. Err. (kappa)	Mean Mutation Frequency (lambda)	Std. Err. (lambda)
AHI	gp41	3.1%	0.3%	3.1%	0.4%
	HIV-other	4.7%	NA%	5.1%	1.3%
FV	HA	3.7%	0.1%	3.5%	0.1%
	Non-HA	4.3%	0.4%	3.1%	0.3%
FIA	HA	3.7%	0.5%	4.9%	0.7%
	Non-HA	2.8%	0.8%	5.4%	0.6%
FIS	HA	2.9%	1.6%	NA	NA
	Non-HA	NA	NA	3.5%	NA

Table B.8: Number of nucleotides in the CDRs and Frameworks in the heavy and light chains of different patient groups

Patient	CDR3H	FRH	CDR3K	FRK	CDR3L	FRL
AHI	109989	288356	44038	185071	26362	98825
FIA	25389	65351	10890	45087	5259	19504
FIS	23898	63713	10463	43871	4974	18443
FV	44847	111846	15819	67374	11418	42857
Uninfected	26004	68651	11998	49706	4830	18212

Table B.9: Number of synonymous and non synonymous mutations in the CDRs and Frameworks of heavy and light chains

	HCDR		HFR		KCDR		KFR		KCDR		KFR	
Patient	S	N	S	N	S	N	S	N	S	N	S	N
AHI	1585	5661	4701	8916	646	2523	2298	3789	336	1349	1084	1870
FIA	329	1223	1002	1692	130	474	422	608	62	230	164	276
FIS	324	1275	917	1841	132	553	461	753	48	264	157	303
FV	526	2082	1750	2986	167	897	789	1173	186	624	354	631
Uninfected	442	1504	1472	2416	219	745	830	1052	74	290	287	376

B.4 Clones

Table B.10: Clones in DFLU07001, a patient vaccinated with a Trivalent Inactivated Seasonal Vaccine

Patient	Number of pairs in Clone	VH	JH	Light Chain	VL	JL	% antigen specific
DFLU07001	2	4-59	6	kappa	1-39	4	100
DFLU07001	5	4-59	6	kappa	1-39	1	100
DFLU07001	11	4-4	6	lambda	3-19	2	100
DFLU07001	18	4-59	6	kappa	1-39	3	94.4
DFLU07001	5	4-59	6	kappa	1-17	1	100
DFLU07001	21	4-59	6	kappa	1-39	2	100
DFLU07001	5	4-59	6	kappa	3-11	4	100
DFLU07001	8	3-30-3	6	lambda	1-44	3	100
DFLU07001	15	3-49	4	lambda	1-51	2	100
DFLU07001	3	3-30	6	kappa	1-33	4	100
DFLU07001	2	4-59	6	kappa	1-39	4	100
DFLU07001	4	3-30	6	kappa	3-20	2	100
DFLU07001	2	4-31	2	lambda	3-9	2	0
DFLU07001	2	4-39	6	lambda	6-57	2	100
DFLU07001	3	1-2	6	lambda	3-21	2	100
DFLU07001	4	3-9	6	kappa	1-33	4	100
DFLU07001	2	4-59	5	lambda	2-14	1	100
DFLU07001	2	3-15	3	lambda	1-40	2	100
DFLU07001	2	4-59	6	lambda	1-44	3	100
DFLU07001	3	4-31	6	kappa	3-20	3	66.67
DFLU07001	2	4-59	6	kappa	1-39	2	100
DFLU07001	4	4-b	6	lambda	7-43	3	75
DFLU07001	4	4-59	6	kappa	3-20	1	100
DFLU07001	2	4-59	6	lambda	1-44	2	100
DFLU07001	2	4-59	6	kappa	1-39	2	100
DFLU07001	2	3-49	4	kappa	1-33	4	0
DFLU07001	3	4-59	6	kappa	1-39	4	100
DFLU07001	3	4-30-4	6	kappa	3-20	2	66.67
DFLU07001	2	4-30-4	6	kappa	3-20	2	50
DFLU07001	2	3-48	5	lambda	3-21	1	100
DFLU07001	2	4-59	6	kappa	3-20	4	100
DFLU07001	2	3-23	3	kappa	3-15	4	100
DFLU07001	2	3-30-3	3	lambda	1-40	3	0
DFLU07001	2	4-59	6	lambda	3-21	2	100
DFLU07001	2	3-7	5	lambda	3-1	2	100
DFLU07021	2	1-2	4	lambda	3-21	2	100
DFLU07021	8	2-5	3	kappa	1-17	2	100
DFLU07021	2	1-2	5	kappa	1-39	5	100
DFLU07021	3	4-39	1	kappa	2-28	5	100
DFLU07021	3	4-39	3	kappa	1-5	1	100
DFLU07021	2	3-30-3	2	kappa	1-39	1	100

Bibliography

[1] Munshaw S and Kepler TB. An information theoretic method for the treatment of plural ancestry in phylogenetics. *Mol. Biol. Evol.*, 6:1199–1208, 2008.

[2] Buckley KM, Munshaw S, Kepler TB, and Smith LC. The 185/333 gene family is a rapidly diversifying host-defense gene cluster in the purple sea urchin. *J. Mol. Biol.*, 379(4):912–928, 2008.

[3] Munshaw S and Kepler TB. SoDA2: A Hidden Markov Model approach for identification of immunoglobulin rearrangements. *Bioinformatics*, page In press, 2010.

[4] Clavel F, Hoggan MD, Willey RL, Strebel K, Martin MA, and Repaske R. Genetic recombination of Human Immunodeficiency Virus. *J. Virol.*, 63:1455–1459, 1989.

[5] Coffin JM. Structure, replication and, recombination of retrovirus genomes: Some unifying hypotheses. *J. Gen. Virol.*, 42(1):1–26, 1979.

[6] Junghans RP, Boone LR, and Skalka AM. Retroviral DNA H structures: Displacement-assimilation model of recombination. *Cell*, 30:53–62, 1982.

[7] Jetzt AE, Yu H, Klarmann GJ, Ron Y, Preston B, and Dougherty JP. High rate of recombination throughout the Human Immunodeficiency Virus Type I genome. *J. Virol*, 74(3):1234–1240, 2000.

[8] Gratton S, Cheynier R, Dumaurier M, Oksenhendler E, and Wain-Hobson S. Highly restricted spread of HIV-1 and multiply infected cells within splenic germinal centers. *Proc. Natl. Acad. Sci.*, 97(26):14566–14571, 2002.

[9] Jung A, Maier R, Vartanian J, Bocharov G, Jung V, Fischer U, Meese E, Wain-Hobson S, and Meyerhans A. Multiply infected spleen cells in HIV patients. *Nature*, 418:144, 2002.

[10] Hudson RR. Properties of a neutral allele model with intragenic recombination. *Theor. Pop. Biol*, 23:183–201, 1983.

[11] Schierup MH and Hein J. Consequences of recombination on traditional phylogenetic analysis. *Genetics*, 156:879–891, 2000.

[12] Pancer Z, Amemiya CT, Ehrhardt RA, Ceitlin J, Gartland GL, and Cooper MD. Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. *Nature*, 430:174–180, 2004.

[13] Zhang S, Adema CM, Kepler TB, and Loker ES. Diversification of Ig superfamily genes in an invertebrate. *Science*, 305(5681):251–254, 2004.

[14] Rast JP, Smith LC, Loza-Coll M, Hibino T, and Litman GW. Genomic insights into the immune system of the sea urchin. *Science*, 314:952–956, 2006.

[15] Buckley KM and Smith LC. Extraordinary diversity among members of the large gene family 185/333, from the purple sea urchin, *Strongylocentrotus purpuratus*. *BMC Mol. Biol.*, 8:68, 2007.

[16] Nair SV, Del Valle H, Gross PS, Terwilliger DP, and Smith LC. Microarray analysis of coelomocyte gene expression in response to LPS in the sea urchin. Identification of unxpected immune diversity in an invertebrate. *Physiol. Genomics*, 22:33–47, 2005.

[17] Terwilliger DP, Buckley KM, Brockton V, Ritter NJ, and Smith LC. Distincitve expression patterns of 185/333 genes in the purple sea urchin *Strongylocentrotus purpuratus*: an unexpectedly diverse family of transcripts in response to LPS, beta-1-3-glucan, and dsRNA. *BMC Mol. Biol.*, 8:16, 2007.

[18] Brockton V, Henson JH, Raftos DA, Majeske AJ, Kim YO, and Smith LC. Localization and diversity of 185/333 proteins from the purple sea urchin - unexpected protein-size range and protein expression in a new coelomocyte type. *J. Cell Sci.*, 121:339–348, 2008.

[19] Dheilly NM, Nair SV; Smith LC, and Raftos DA. Highly variable immune-response proteins (185/333) from the sea urchin, *Strongylocentrotus purpuratus*: proteomic analysis identifies diversity within and between individuals. *J. Immunol.*, 182(4):2203–2212, 2009.

[20] Sakano H, Maki R, Kurosawa Y, Roeder W, and Tonegawa S. Two types of somatic recombination are necessary for the generation of complete immunoglobulin heavy-chain genes. *Nature*, 286(5774):676–683, 1980.

[21] Tonegawa S. Somatic generation of antibody diversity. *Nature*, 302(5909):575–581, 1983.

[22] LeFranc MP. IMGT, the international ImMunoGeneTics database. *Nuc. Acids Res.*, 29(1):207–209, 2001.

[23] Lorenz W, Straubinger B, and Zachau HG. Physical map of the immunoglobulin K locus and its implications for the mechanisms of VK-JK rearrangement. *Nuc. Acids Res.*, 15(23):9667–9676, 2001.

[24] Frippiat JP, Williams SC, Tomlinson IM, Cook GP, Cherif D, LePaslier L, Collins JE, Dunham I, Winter G, and LeFranc MP. Organization of the human immunoglobulin lambda light-chain locus on chromosome 22q11.2. *Hum. Mol. Genet.*, 4(6):983–991, 1995.

[25] Desiderio SV, Yancopoulos GD, Paskind M, Thomas E, Boss MA, Landau N, Alt FW, and Baltimore D. Insertion of N regions into heavy-chain genes is correlated with the expression of terminal deoxytransferase in B-cells. *Nature*, 311(5988):752–757, 1984.

[26] Bollum FJ. *Terminal deoxynucleotidyl transferase*. New York, New York, 10 edition, 1974.

[27] Sakano H, Hppi K, Heinrich G, and Tonegawa S. Sequences at the somatic recombination sites of immunoglobulin light-chain genes. *Nature*, 280:288–294, 1979.

[28] Goldsby RA, Kindt TJ, Osborne BA, and Kuby J. *Immunology*. WH Freeman and Company, 5 edition, 1859.

[29] Goldsby RA, Kindt TJ, Osborne BA, and Kuby J. *Fundamentals of Immunology*. WH Freeman and Company, 5 edition, 1859.

[30] Bridges SL. Frequent N addition and clonal relatedness among immunoglobulin lambda light chains expressed in rheumatoid arthritis synovia and PBL, and the influence of V lambda gene segment utilization on CDR3 length. *Mol Med.*, 4(8):525–553, 1998.

[31] Lafaille JJ, DeCloux A, Marc Bonneville M, Takagaki Y, and Tonegawa S. Junctional sequence of t cell receptor gamma delta genes: implications for gamma delta t cell linages and for a novel intermediate of v-(d)-j joining. *Cell*, 59(5):859–870, 1989.

[32] Muramatsu M, Kinoshita S, Fagarasan S, Yamada S, Shinkai Y, and Honjo T. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell*, 102(5):553–563, 2000.

[33] Volpe JM, Cowell LG, and Kepler TB. SoDA: implementation of a 3D alignment algorithm for inference of antigen receptor recombinations. *Bioinformatics*, 22(4):438–444, 2006.

[34] Kunert R, Ruker F, and Katinger H. Molecular characterization of five neutralizing anti-HIV type i antibodies: identification of nonconventional D segments in the human monoclonal antibdoies 2g12 and 2f5. *AIDS Res Hum Retroviruses*, 14(13):1115–1128, 1998.

[35] Kunert R, Wolbank S, Stiegler G, Weik R, and Katinger H. Characterization of molecular features, antigen-binding and in vitro properties of IgG and IgM of 4e10, an anti-HIV type I neutralizing monoclonal antibody. *AIDS Res Hum Retroviruses*, 14(13):1115–1128, 1998.

[36] Walker LM, Phogat SM, Chan-Hui P, Wagner D, Phung P, Goss JL, Wrin T, Simek MD, Fling S, Mitcham JL, Lehrman JK, Priddy FH, Olsen OA, Frey SM, Hammond PW, Kaminsky S, Zamb T, Moyle M, Koff WC, Poignard P, and Burton DR. Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. *Science*, 326(5950):285–289, 2009.

[37] Rerks-Ngarm S et al. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in thailand. *N Engl J Med*, NEJMoa0908492:1–12, 2009.

[38] Hudson RR and Kaplan NL. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111(1):147–164, 1985.

[39] Grassly NC and Holmes EC. A likelihood method for the detection of selection and recombination using nucleotide sequences. *Mol. Biol. Evol.*, 14:239–247, 1997.

[40] Hein J. Reconstructing evolution of sequences subject to recombination using parsimony. *Math Biosci.*, 98(2):185–200, 1990.

[41] Hein J. A heuristic method to reconstruct the history of sequences subject to recombination. *J. Mol. Evol.*, 36(4):396–405, 1993.

[42] Felsenstein J. Phylogenies from molecular sequence: inference and reliability. *Ann. Rev. Genet.*, 22:521–565, 1988.

[43] Strimmer K and Moulton V. Likelihood analysis of phylogenetic networks using directed graphical models. *Mol. Biol. Evol.*, 17(6):875–881, 2000.

[44] Jin G, Nakhleh L, Snir S, and Tuller T. Maximum likelihood of phylogenetic networks. *Bioinf*, 22(21):2604–2611, 2006.

[45] Zhang SM, Adema CM, Kepler TB, and Loker ES. Diversification of Ig superfamily genes in an invertebrate. *Science*, 305:251–254, 2004.

[46] Grunwald PD. *The Minimum Description Length Principle*. Cambrdige:MIT Press, 2007.

[47] Kolmogorov A. Three approaches to the quantitative definition of information. *Inform. Trans.*, 1:1–7, 1965.

[48] Solmonoff R. A formal theory of inductive inteference. *Inform. Control*, 7:1–22, 1964.

[49] Chaitin G. On the length of programs for computing finite binary sequences. *J. ACM*, 13:547–569, 1966.

[50] Rissanen J. Stochastic complexity. *J. Royal Statist. Soc. B*, 49:223–239, 1989.

[51] Darwin CR. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray, 1 edition, 1859.

[52] Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, Gupta P, Rinaldo CR, Learn GH, He X, Huang XL, and Mullins JI. Consistent viral evolutionary changes associated with the progression of Human Immunodeficiency Virus Type I infection. *J. Virol.*, 73(12):10489–10502, 1999.

[53] Felsenstein J. The number of evolutionary trees. *Sys. Zoology*, 27(1):27–33, 1978.

[54] Kirkpatrick S, Gelatt CD, and Vecchi MP. Optimization by simulated annealing. *Science*, 220:2604–2611, 1983.

[55] Terwilliger DP, Buckley KM, Mehta D, Moorjani PG, and Smith LC. Unexpected diversity displayed in cdnas expressed by the immune cells of the purple sea urchin, *Strongylocentrotus purpuratus*. *Physiol. Genomics*, 26:134–144, 2006.

[56] Felsenstein J. PHYLIP (phylogeny inference package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle*, 2005.

[57] Siu G, Clark SP, Yoshikai Y, Malissen M, Yanagi Y, Strauss E, Mak TW, and Hood L. The human T cell antigen receptor is encoded by variable, diversity and joining gene segments that rearrange to generate a complete V gene. *Cell*, 37:393–401, 1984.

[58] Marzluff WF, Sakallah S, and Kelkar H. The sea urchin histone gene complement. *Dev. Biol.*, 300:308–320, 2006.

[59] Adema CM, Lynn A Hertel LA, Miller RD, and Loker ES. A family of fibrinogen-related proteins that precipitates parasite-derived molecules is produced by an invertebrate after infection. *Proc. Natl. Acad. Sci.*, 94(16):8691–8696, 1997.

[60] Massana R, Guillou L, Diez B, and Pedros-Alio C. Unveiling the organisms behind novel eukaryotic ribosomal DNA sequences from the ocean. *Appl. Environ. Microbiol.*, 68:4554–4558, 2002.

[61] Giudicelli1 V, Chaume D, and Lefranc MP. IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor VJ and VDJ rearrangement analysis. *Nuc. Acids Res.*, 32:435–440, 2004.

[62] Altschul SF, Warren G, Miller W, Myers EW, and Lipman DJ. Basic local alignment tool. *J. Mol. Biol.*, 215:403–410, 1990.

[63] Souto-Carneiro MM, Longo NS, Russ DE, Sun H, and Lipsky PE. Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER. *J. Immunol.*, 172:6790–6802, 1990.

[64] Smith TF and Waterman MS. Identification of common molecular subsequences. *J. Mol. Biol.*, 147(1):195–197, 1981.

[65] Gata BA, Malming HR, Jackson KJL, Bain ME, Wilson P, and Collins AM. iHMMune-align: hidden markov model-based alignment and identification of

germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics*, 23(13):1580–1587, 2007.

[66] Rabiner LR. A tutorial on hidden markov-models and selected applications in speech recognition. *Proc. IEEE*, 77:257–286, 1989.

[67] Smith DS, Creadon G, Jena PK, Portanova JP, Kotzin BL, and Wysocki LJ. Di- and trinucleotide target preferences of somatic mutagenesis in normal and autoreactive B cells. *J. Immunol.*, 156:2642–2652, 1996.

[68] Honegger A and Plckthun A. Yet another numbering scheme for immunoglobulin variable domains: An automatic modeling and analysis tool. *J. Mol. Biol*, 309:657–670, 2001.

[69] Kepler TB, Sample C, Haines A, Roach J, Walsh A, and Rambsurg EA. Characterization of chiropteran type-i interferon genes inferred from genome sequencing traces by a novel gene-family assembler. *J. Mol. Biol*, 309:657–670, 2009.

[70] Ewing B and Green P. Basecalling of automated sequences traces using phred. II. error probabilities. *Genome Res.*, 8:186–194, 1998.

[71] Cowell LG, Kim H, Humaljoki T, Berek C, and Kepler TB. Enhanced evolvability in immunoglobulin v genes under somatic hypermutation. *J. Mol. Evol.*, 49:23–26, 1999.

[72] Basu M, Hegde MV, and Modak MJ. Synthesis of compositionally unique DNA by terminal deoxynucleotidyl transferase. *Biochem. Biophys. Res. Comm*, 111:1105–1112, 1983.

[73] Jackson KJ, Gaeta B, Sewell W, and Collins AM. Exonuclease activity and P nucleotide addition in the generation of the expressed immunoglobulin repertoire. *BMC Immunol.*, 2:5–19, 2004.

[74] Durbin R, Eddy SR, Krogh A, and Mitchison G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambrdige University Press, 1998.

[75] Majoros WH. *Methods for Computational Gene Prediction*. Cambridge University Press, 2007.

[76] Fariselli P, Martelli PL, and Casadio R. A new decoding algorithm for hidden markov models improves the prediction of the topology of all-beta membrane proteins. *BMC Bioinformatics*, 6(4):S12, 2005.

[77] Monod MY, Giudicelli1 V, Chaume D, and Lefranc MP. IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J junctions. *Bioinformatics*, 20:i379–i385, 2004.

[78] Shirai A, Cosentino M, Leitman-Klinman SF, and Klinman DM. Human Immunodeficiency Virus infection induces both polyclonal and virus-specific B cell activation. *J. Clin. Investigation*, 89:561–566, 1992.

[79] Malaspina A, Moir S, Orsega SM, Vasquez J, Miller NJ, Donoghue ET, Kottilil S, Gezmu M, Follmann D, Vodeiko GM, Levandowski RA, Mican JM, and Fauci AS. Compromised B cell responses to Influenza vaccination in HIV-infected individuals. *J. Infec. Diseases*, 191:1442–1450, 2005.

[80] Titanjia K, Chiodia F, Belloccob R, Schepisa D, Osorioa L, Tassandinc C, Tambussic G, Grutzmeiera S, Lopalcoc L, and Militoa A. Primary HIV-1 infection sets the stage for important B lymphocyte dysfunctions. *AIDS*, 19:1947–1955, 2005.

[81] Liao H, Levesque MC, Negal A, Dixon A, Zhang R, Walter E, Parks R, Whitesides J, Masrshall DJ, Hwang K, Yang Y, Gao F, Munshaw S, Kepler TB, Denny T, Moody MA, and Haynes BF. High-throughput isolation of immunoglobulin genes from single human B cells and expression as monoclonal antibodies. *J. Virol. Methods*, 158(1-2):171–179, 2009.

[82] Ewing B and Green P. Basecalling of automated sequences traces using phred. I. accuracy assessment. *Genome Res.*, 8:175–185, 1998.

[83] Wrammert J, Smith K, Miller J, Langley W, Kokko K, Larsen C, Zheng N, Mays I, Garman L, Helms C, James J, Air G, Capra JD, Ahmed R, and Wilson PC. Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature*, 453(5):667–672, 2008.

[84] Qin Y, Duquette P, Zhang Y, Talbot P, Poole R, and Antel J. Clonal expansion and somatic hypermutation of V(H) genes of B cells from cerebrospinal fluid in Multiple Sclerosis. *J. Clin. Invest.*, 102(5):1045–1050, 1998.

[85] Kurth J, Spieker T, Wustrow J, Strickler JG, Hansmann M, Rajewsky K, and Kuppers R. EBV-infected B cells in infectious mononucleosis: Viral strategies

for spreading in the B cell compartment and establishing latency. *Immunity*, 13:485–495, 2000.

[86] Chong Y, Ikematsu H, Ariyama I, Chijiwa K, Li W, Yamaji K, Kashiwagi S, and Hayashi J. Evidence of B cell clonal expansion in HIV type 1-infected patients. *AIDS Res. Hum. Retroviruses*, 17(16):1507–1515, 2001.

[87] Scheid JF, Mouquet H, Feldhahn N, Seaman MS, Velinzon1 K, Pietzsch J, Ott RG, Anthony RM, Zebroski H, Hurley A, Phogat A, Chakrabarti B, Li Y, Connors C, Pereyra F, Walker BD, Wardemann H, Ho D, Wyatt RT, Mascola JR, Ravetch JV, and Nussenzweig MC. Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals. *Nature*, 458:636–640, 2009.

[88] Simpson EH. Measurement of diversity. *Nature*, 163:688, 1949.

[89] Weksler ME and Szabo P. The effect of age on the B cell repertoire. *J Clin Immunol*, 20:240–249, 2000.

[90] Tomaras GD, Yates NL, Liu P, Li Qin L, Fouda GG, Chavez LL, Decamp AC, Parks RJ, Ashley VC, Lucas JT, Cohen M, Eron J, Hicks CB, Liao HX, Self SG, Landucci G, Forthal DN, Weinhold KJ, Keele BF, Hahn BH, Greenberg ML, Morris L, Karim SS, Blattner WA, Montefiori DC, Shaw GM, Perelson AS, and Haynes BF. Initial B-cell responses to transmitted Human Immunodeficiency Virus Type I: virion-binding immunoglobulin M (IgM) and IgG antibodies followed by plasma anti-gp41 antibodies with ineffective control of initial viremia. *J. Virol.*, 82(24):12449–12463, 2008.

[91] Alachkar H, Taubenheim N, Haeney MR, Durandy A, and Arkwright PD. Memory switched B cell percentage and not serum immunoglobulin concentration is associated with clinical complications in children and adults with specific antibody deficiency and common variable immunodeficiency. *Clin Immunol.*, 120:310–318, 2006.

[92] Levesque MC, Moody MA, Hwang K, Marshall DJ, Whitesides JF, Amos JD, Gurley TC, Allgood S, Haynes BB, Vandergrift NA, Plonk S, Parker DC, Cohen MS, Tomaras GD, Goepfert PA, Shaw GM, Schmitz JE, Eron JJ, Shaheen NJ, Hicks CB, Liao HX, Markowitz M, Kelsoe GH, Margolis DM, and Haynes BF. Polyclonal B cell differentiation and loss of gastrointestinal tract germinal centers in the earliest stages of HIV-1 infection. *PLOS Med.*, 6(7):e1000107, 2009.

[93] Volpe JM and Kepler TB. Large-scale analysis of human heavy chain V(D)J recombination patterns. *Immunome Res.*, 4(1), 2008.

[94] Sui J, Hwang WC, Perez S, Wei G, Aird D, Chen L, Santelli E, Stec B, Cadwell G, Ali M, Wan H, Murakami A, Yammanuru A, Han T, Cox NJ, Bankston LA, Donis RO, Liddington RC, and Marasco WA. Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. *Nat. Struct. Mol. Biol.*, 16:265–273, 2009.

[95] Chan CH, Hadlock KG, Foung SKH, and Levy S. VH1-69 gene is preferentially used by Hepatitis C virus-associated B cell lymphomas and by normal B cells responding to the E2 viral antigen. *Blood*, 97:1023–1026, 2001.

[96] Huang C, Venturi M, Majeed S, Moore MJ, Phogat S, Zhang M, Dimitrov DS, Hendrickson WA, Robinson J, Sodroski J, Wyatt R, Choe H, Farzan M, and Kwong PD. Structural basis of tyrosine sulfation and VH-gene usage in antibodies that recognize the HIV type 1 coreceptor-binding site on gp120. *Proc. Natl. Acad. Sciences*, 101:2706–2711, 2004.

[97] Raaphorst FM, Raman CS, Nall BT, and Teale JM. Molecular mechanisms governing reading frame choice of immunoglobulin diversity genes. *Immunology Today*, 18(1):37–43, 1997.

[98] Haynes BF, Fleming J, Clair WS, Katinger H, Stiegler G, Kunert R, Robinson J, Scearce RM, Plonk K, Staats HF, Ortel TL, Liao HX, and Munir Alam SM. Cardiolipin polyspecific autoreactivity in two broadly neutralizing HIV-1 antibodies. *Science*, 308(5730):1906–1908, 2005.

[99] Kyte J and Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157(1):105–132, 1982.

[100] Klein U, Rajewsky K, and Kuppers R. Human immunoglobulin (Ig)M+IgD+ peripheral blood B cells expressing the CD27 cell surface antigen carry somatically mutated variable region genes: CD27 as a general marker for somatically mutated (memory) B cells. *J. Exp. Med*, 188:1679–1689, 1998.

[101] Berman JE, Nickerson KG, Pollock RR, Barth JE, Schuurman RKB, Knowles DM, Chess L, and Alt FW. VH gene usage in humans: biased usage of the VH6 gene in immature B lymphoid cells. *European J. Immunol*, 21(5):1311–1314, 2005.

[102] Soderstrom I, Dijk-Hard I, Feld S, Hillorn V, Holmberg D, and Lundkvist I. Altered VH6-D-JH repertoire in human insulin-dependent diabetes mellitus

and autoimmune idiopathic thrombocytopenic purpura. *European J. Immunol*, 29(9):2853–2862, 1999.

[103] Mild M, Esbjrnsson J, Fenyo EM, and Medstrand P. Frequent intrapatient recombination between Human Immunodeficiency Virus Type I R5 and X4 envelopes: Implications for coreceptor switch. *J. Virol.*, 81(7):3369–3376, 2007.

[104] Kellam P and Larder BA. Retroviral recombination can lead to linkage of reverse transcriptase mutations that confer increased Zidovudine resistance. *J. Virol.*, 69(2):669–674, 1995.

[105] Liu SL, Mittler JE, Nickle DC, Mulvania TM, Shriner D, Rodrigo AG, Kosloff B, He X, Corey L, and Mullins JI. Selection for Human Immunodeficiency Virus Type I recombinants in a patient with rapid progression to AIDS. *J. Virol.*, 76(21):10674–10684, 1995.

[106] Gottlieb GS, Nickle DC, Jensen MA, Wong KG, Grobler J, Li F, Liu S, Rademeyer C, Learn GH, Abdool Karim SS, Williamson C, Corey L, Margolick JB, and Mullins JI. Dual HIV-1 infection associated with rapid disease progression. *The Lancet*, 363(9409):619–622, 2004.

[107] Song L, Kalyanaraman C, Fedorov AA, Fedorov EV, Glasner ME, Brown S, Imker HJ, Babbitt PC, Almo SC, Jacobson MP, and Gerlt JA. Prediction and assignment of function for a divergent N-succinyl amino acid racemase. *Nat. Chem. Biol.*, 3:486–491, 2007.

[108] Sawyer S. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.*, 6:528–538, 1989.

[109] Buckley KM, Terwilliger DP, and Smith LC. Sequence variations in 185/333 messages from the purple sea urchin suggest posttranscriptional modifications to increase immune diversity. *J. Immunol*, 181:8585–8594, 2008.

[110] Simpson EH. Measurement of diversity. *Nature*, 163:688, 1949.

Biography

Supriya Munshaw was born and brought up in Ahmedabad, India. She moved to the United States in 2000 to attend Bard College, Annandale on Hudson NY. She was one of the first three students to study Bioinformatics at Bard where graduated with a Bachelor of Arts in Natural Sciences and Mathematics in May 2004. Supriya joined the second class of the PhD program in Computational Biology in August 2004 immediately after graduation. Her first rotation with Dr. Barton Haynes' group at the Duke Human Vaccine Institute motivated her to work with HIV-1. She joined Dr. Thomas Kepler's group in 2005 where she worked on various projects, most of which have been published in peer reviewed journals. These include

- Munshaw S and Kepler TB. 2009. SoDA2: A Hidden Markov Model Approach for Inference of Immunoglobulin Rearrangements. In Press.
- Liao H, Levesque MC, Negal A, Dixon A, Zhang R, Walter E, Parks R, Whitesides J, Masrshall DJ, Hwang K, Yang Y, Gao F, Munshaw S, Kepler TB, Denny T, Moody MA, and Haynes BF. 2009. High-throughput isolation of immunoglobulin genes from single human B cells and expression as monoclonal antibodies. J. Virol. Methods.158(1-2):171-9
- Munshaw S and Kepler TB. 2008. An Information-Theoretic Method for the Treatment of Plural Ancestry in Phylogenetics. Mol. Biol. Evol. 25(6):1199-1208.
- Buckley KM, Munshaw S, Kepler TB, Smith LC. 2008. The 185/333 gene

family is a rapidly diversifying host-defense gene cluster in the purple sea urchin, Strongylocentrotus purpuratus. J. Mol Biol. 379(4):912-28.

She has also presented the above projects at various international conferences. She has received travel awards to attend

- The International Society for Developmental and Comparative Immumology Conference (Charleston SC in July 2006) to present her work on the purple sea urchin and,
- Multi-Scale Modeling of Host/Pathogen Interactions (Pittsburgh PA in June 2009),
- AIDS Vaccine 2009 (Paris, France in October 2009) and
- Keystone Symposia on HIV Vaccines (X5) (Banff, Canada in March 2010) to present her work with humoral response to HIV-1

Outside of work, Supriya likes to travel the world, exercise, and continue her endless quest to learn French.