MECHANISTIC AND GENETIC BIASES IN HUMAN IMMUNOGLOBULIN HEAVY CHAIN DEVELOPMENT

by

Joseph M. Volpe

Department of Computational Biology & Bioinformatics
Duke University

Date: _____

Approved:

_____
Thomas B. Kepler, Supervisor

_____
Lindsay G. Cowell

_____
Garnett H. Kelsoe

_____
Marcy K. Uyenoyama

_____
Jun Yang

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

2008

ABSTRACT

(Computational Biology & Bioinformatics)

MECHANISTIC AND GENETIC BIASES IN HUMAN IMMUNOGLOBULIN HEAVY CHAIN DEVELOPMENT

by

Joseph M. Volpe

Department of Computational Biology & Bioinformatics
Duke University

Date: _____

Approved:

_____
Thomas B. Kepler, Supervisor

_____
Lindsay G. Cowell

_____
Garnett H. Kelsoe

_____
Marcy K. Uyenoyama

_____
Jun Yang

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 & Bioinformatics
in the Graduate School of
Duke University

2008

Abstract

Broadly neutralizing antibodies against HIV are rare; most patients never develop them at detectable levels. The discovery of four such antibodies therefore warrants research into their origins and their presumed unique characteristics. Such studies, however, require baseline knowledge about commonalities and biases affecting human immunoglobulin development. Obtaining that knowledge requires large sets of gene sequence data and the appropriate statistical techniques and tools.

The Genbank repository provides a free and easily accessible source for such data. Several large datasets cumulatively comprising over 10,000 human Ig heavy chain genes were identified, downloaded, and carefully filtered. We then developed a special software tool called SoDA, which employs a unique dynamic programming algorithm to provide a statistical reconstruction of the events that led to a given antigen receptor gene. Once developed, tested, and peer-reviewed, we used SoDA to provide initial data about each downloaded gene with respect to gene segment usage, n-nucleotide addition, CDR3 length, and mutation frequency, thereby establishing the most precise estimates currently available for human Ig heavy chain gene segment usage frequencies.

We compared data from productive non-autoreactive Ig to non-productive Ig and found evidence for gene segment usage biases, D/J segment pairing preferences resulting from multiple sequential D-to-J recombination events, and biases in TdT action between the V-D and D-J. Further analysis of autoreactive Ig genes yielded evidence that n-nucleotide addition comes at a cost: the higher the ratio of n-nucleotides to

germline-encoded nucleotides for a given CDR3 length, the greater the probability of autoreactivity. These results suggest that the germline gene segments have been selected for lack of autoreactivity.

It has previously been shown that human Ig gene segments have evolved efficient evolvability under somatic hypermutation. We have now extended these results, showing that Ig gene sequences are tuned to preferentially produce consequential mutations in the antigen-binding domains, and synonymous mutations in the framework regions.

Together, these analyses provide new insights into the genetic and mechanistic biases shaping the human Ig repertoire.

Contents

List of Tables

List of Figures

Acknowledgements

I would first like to acknowledge, with the most sincere sentiments and gratitude, my advisor Dr. Thomas B Kepler. Through his encouragement, direction, and enthusiasm, he challenged me to constantly improve my skills and inspired me to produce the best research possible. From the day he interviewed me for acceptance into the program, to now as I prepare to graduate, he has believed in my ability to accomplish the difficult goal of earning a PhD. He consistently demonstrated patience in his pedagogy and pushed me to learn, question, and achieve. He has shown me, and truly exemplifies in his ethics, ingenuity, innovation, and perseverance, what being a scientist is all about.

I also would like to thank my committee members for having the patience and scientific curiosity to push me to better my scientific understanding, aptitude, and dialogue. In particular, I would like to thank Dr. Lindsay G Cowell who, with her compassion and generous nature, guided me through my first two successful years at Duke and provided me with support when I needed it most. She is always ready to lend a hand and an ear for all things scientific and personal. A thank you also to Anne Lieberman for her gramatical advisement and contagious happiness.

To the members of the DULCI lab, I appreciate your encouragement and support, in all its forms, including technical support with C++, Linux, and latex. I also thank the IGSP for its support of this PhD program.

To my fellow students in the Computational Biology & Bioinformatics program, I thank you for your camaraderie and encouragement. We stood together and cut new

paths for the success of this young PhD program. Thank you particularly to Supriya Munshaw for her support and friendship. Our daily dialogue often made it possible to get through those occasional dragging days. Thank you also to Ana Paula Sales and Haige Shen who patiently and valiantly tried to help me understand statistics.

A huge debt of gratitude is owed to Jeff Headd and Dave Orlando, who have been more like brothers than fellow students. Without their support, friendship, encouragement, and distraction, this old guy would have never made it through his first year, let alone five. They taught me to not take myself so seriously, and in doing so, helped me to achieve more than I thought possible.

Finally, a deep expression of gratitude to my parents and siblings. They believed in me even at times when I questioned myself, and my successes here at Duke were possible only through their prayers, guidance, and support. These words of acknowledgment could never sufficiently express my deep appreciation.

Chapter 1

Introduction

Broadly neutralizing antibodies against HIV are rare; most infected patients never develop them and instead require drug therapy to maintain healthy CD8+ T-cell counts and low viral levels. The discovery of four such antibodies, 2F5, 4E10, 2G12, and 1B12 [43, 44, 9], has been both exciting and frustrating. These antibodies maintain affinity for the virus and continue to effectively bind to it, even as the virus continues to evolve within the host, thereby blocking entry into target cells. That these antibodies were discovered presumes human capability to develop a broadly neutralizing humoral response to the virus, given the right conditions. The rarity of evidence for such effective antibodies arising naturally in patients, however, suggests that these antibodies are intrinsically difficult to develop somatically. Their rarity also suggests that there may potentially be something rare about the genes from which they are expressed. In order to better understand the assumed unique qualities of these genes, however, it is imperative to understand the typical characteristics of and biases present within the normal human antibody repertoire.

Antibodies are the soluble, secreted form of the immunoglobulin (Ig) molecule produced within all B-cells. This molecule, which serves as both receptor and effector, is a hallmark of adaptive immunity and its defining characteristic: somatic diversification of antigen receptor genes. Each Ig molecule is a homodimer of heterodimers, where each heterodimer comprises one heavy chain protein and one light chain protein. Both the light and heavy chain genes are encoded by ligated gene segments, genetically rearranged during a process known as V(D)J recombination [80, 93]. Light chain genes are made by recombination of one variable (V) and one joining (J) gene segment; heavy chain genes are made by recombination of V, J and an additional segment between them, the diversity (D) segment [20, 34] (Fig. 1.1). In humans, approximately 50 known functional V segments [46, 53, 11, 10], 27 known functional D segments [46, 34, 12], and six known functional J segments [46, 34, 72] are available within a single locus for assembly into heavy chain genes. This locus resides near the long-arm telomere of chromosome 14 and extends inward toward the centromere, with the V segments at the 5' end followed by the D segments and then J segments. Humans have two light chain loci, κ [49, 14] and λ [25], that can rearrange and contribute the required protein, though only one locus is expressed per cell. The recombination of these gene segments into a transcribable gene is mediated by the recombination activating genes, RAG1 and RAG2. In complex, RAG1/RAG2 bring together gene segments and play an important role in the splicing out of intervening sequence as the segments are joined together[67, 63].

During recombination, non-templated (n)-nucleotides may be added by terminal deoxynucleotidyl transferase (TdT) between adjoining gene segments [18]. TdT is

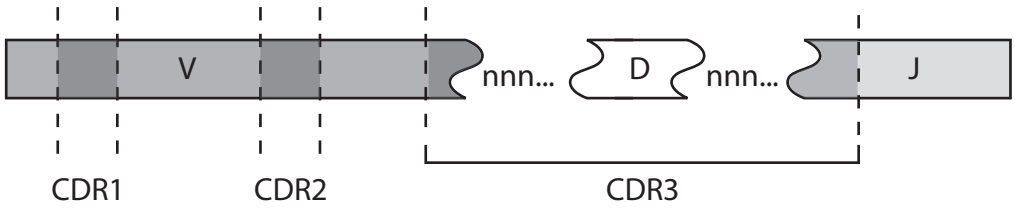

Figure 1.1: Immunoglobulin heavy chain genes are made from the ligation of three types of gene segments: variable (V), diversity (D), and joining (J). Non-templated (n)-nucleotides may be inserted in the junctions during recombination. Each heavy chain gene contains three complementarity determing regions (CDR) in which mutations are concentrated during somatic hypermutation. Two CDR are wholly contained withint the V segment, while CDR3 spans from the 3′ end of the V to the 3′ end of the J, entirely containing the D segment and all n-nucleotides.

structurally similar to DNA pol β [16] and belongs to a family of polymerases called pol X [36]. It is the only known polymerase capable of elongating DNA without a template [5]. The nucleotides added by TdT become part of complementarity determining region 3 (CDR3), which refers to a section of the gene that encodes one of the primary antigen binding loops in the resulting protein (Fig. 1.1). Both heavy and light chain genes encode a total of three loops through their three CDR. Together, the six loops 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 completely embedded within the rearrangeable germline V segments for the given locus. CDR3, however, spans the 3′ end of the V segment through to the 5′ end of the J segment, entirely encompassing the rearranged D segment and all n-nucleotides making it responsible for much of the Ig population diversity.

The complementarity determining regions are the primary targets for point mutations added to the gene during a process known as somatic hypermutation [54]. During an adaptive immune response, B-cells bearing surface Ig reactive with micro-

bial antigens are activated. Some will organize into germinal centers structures in the lymph nodes where they undergo affinity maturation ([37]; reviewed in [67] and [98]). During this process, the B-cells proliferate and express a B-cell specific factor called activation-induced cytidine deaminase (AID) [59, 58], which causes mutations in the Ig genes at rate of up to 10^6 times the normal background rate. These point mutations are concentrated within the CDR and may help to refine affinity for antigen since the CDR form the antigen binding interface. The cells are subsequently selected for enhanced affinity for the eliciting antigen.

In addition to combinatorial diversification, n-nucleotide addition, and somatic hypermutation, Ig genes are subject to randomness in the location of the recombination sites in all segments. This randomness potentially limits the usage of nucleotides at the ends of the gene segments [2]. Together, the processes and mechanisms involved in Ig gene formation enable the generation of over 10^{14} different protein specificities [81]. Thus, the adaptive immune system seems to employ a strategy wherein it randomly generates broad populations of antigen receptors to counter the random variability of antigenic microbes. With so much randomness in the formation of Ig genes, it is both likely and understandable that between 55% to 75% of B-cells developing in the bone marrow present self-reactive Ig on their surface [99]. Three primary mechanisms have been identified that prevent self-reactive Ig from harming their host, namely receptor editing, deletion, and anergy [29, 61, 92, 26, 30].

The genes for T-cell receptors are also created by the same aforementioned processes, though they are not subject to somatic hypermutation. Dynamic assembly and somatic diversification of these B- and T-cell antigen receptor genes are the hall-

marks of adaptive immunity. It is assumed that the processes that produce these antigen receptor genes are largely random, but any existing biases – deviations from strict randomness – potentially provide clues about the mechanisms by which the Ig gene formation processes operate. Much can be learned about V(D)J recombination, n-nucleotide addition, autoreactivity, and somatic hypermutation through statistical studies of these deviations from randomness, and several studies have been published reporting analyses of these biases. Using 71 productive Ig rearrangements from a single individual, Brezinschek et. al. [7] characterized V, D, and J segment usage by PCR analysis of genes from unstimulated B-cells, providing the first evidence for biased gene segment usage within an individual's immature B-cell repertoire. They showed, in particular, that the VH3 family is differentially over-represented among VH gene segments, and that JH6 is expressed more frequently than any of the other segments. In a follow-up study [6] the investigators used samples from two human subjects to study both productive and non-productive Ig rearrangements. By including non-productive sequences and comparing these unselected rearrangements to productive rearrangements subject to selection, they were able to attribute the differential usage to selection. Specifically, they showed that a certain VH4 family segments appeared to be selectively suppressed.

A 2001 study by Rosner et. al. [79] used cells from ten human subjects to study CDR3 length differences between mutated and non-mutated Ig genes. Their analysis led them to hypothesize that B-cells bearing Ig with shorter CDR3 are selected for antigen binding. In the course of this study, the authors established statistical baselines for typical n-nucleotide tract lengths in the V-D and D-J junctions of Ig genes

and provided some of the first statistics regarding D gene segment usage frequency and CDR3 length in the adult human Ig repertoire. More recently, Souto-Carneiro et. al. [90] gathered Ig sequences from several studies, including the aforementioned Brezinschek study, to statistically characterize CDR3 genetics using more sequences than had been previously available in a single study. They developed specialized software for the analysis of CDR3 amino acid composition, D segment usage, D segment reading frame, and provide one of the most complete statistical analyses of D gene segment usage, including evidence for the use of the controversial "irregular D segments" [34].

Inspired by the desire to deepen our understanding of the broadly neutralizing anti-HIV antibodies, we ventured to perform a comprehensive analysis of Ig molecules on a large scale to establish baseline statistics regarding Ig genetics. Thus, the research initiatives described herein differ from the aforementioned efforts in that they involve using a much larger set of Ig genes than was previously possible. Such large scale initiatives to study biases have only recently become tractable. Developments in laboratory methods and sequencing technologies have facilitated rapid production of large genetic datasets for Ig. The parallel rise of bioinformatics and systems biology has promoted methods for storage, analysis, and sharing of those data. Genbank, for example, currently holds over 20,000 human Ig heavy chain records. This profusion of freely available Ig sequence data presents an opportunity to study statistically the genetic and molecular details of Ig using a large dataset that only recently has become manageable.

The large number of human heavy chain records in Genbank have been submitted

as a result of hundreds of different studies with equally numerous different objectives. Still, much more can be learned from this freely available raw sequence data. Its reuse enabled broad and detailed studies of the genetic properties and molecular mechanisms that influence Ig formation. Simply due to the large scale, such analyses of productively rearragned Ig genes, including autoreactive Ig, in comparison to non-productively rearranged genes provided statistically-based foundational data regarding typical characteristics of and biases within the human Ig repertoire. A broad and large-scale study of the kind performed here requires two main components: sets of gene segments to analyze and compare, and a tool to facilitate the initial analysis of those gene segments. Presented in the following chapters is a detailed description of the construction of such a tool and the application of that tool on large datasets of Ig heavy chain sequences gathered from Genbank. This process not only provided the most precise estimates of Ig heavy chain gene segment usage currently available, but also enabled statistical analyses that provided unique insights into the genetic and mechanistic biases that shape the Ig heavy chain repertoire.

Chapter 2

Developing the Research Tool, SoDA

In order to study biases in the mechanisms and genetics of Ig, we need an efficient and simple tool to facilitate the understanding of the genetics heavy chain genes. For any given Ig gene, we want to know which processes occurred to produce a given Ig heavy chain DNA sequence, which gene segments were involved, and the extent that processes like TdT n=nucleotide addition and somatic hypermutation shaped the DNA sequence of the gene. The impact of complicating factors, like recombination site choice and somatic hypermutation of n-nucleotides, can never be known for sure, but can be probabilistically estimated. This makes for a challenging problem and research opportunity.

Stated more precisely, the problem we set out to solve is the statistical reconstruction of the recombination events that led to any given antigen receptor gene. The solution consists in identifying each of the V, D, and J gene segments used as well as the recombination sites, point mutations and n-nucleotides. Because of the uncertainty in the recombination sites, the aligned, untrimmed gene segments may

overlap. Therefore, alignments of the target gene to the individual gene segments cannot be treated as independent. This feature sets the V(D)J problem apart from the otherwise similar spliced alignment problem [27]. The substantial single-nucleotide diversification that occurs throughout all junctions due to somatic hypermutation and the possible addition of n-nucleotides adds further complication.

Several algorithms already existed for determining Ig and TCR gene segment composition. IMGT/V-QUEST is perhaps the most complete of these tools, having the ability to analyze both Ig and TCR sequences for human, mouse, sheep, and other organisms [28]. V-QUEST, however, is based on the BLAST algorithm; it is not as sensitive as dynamic programming methods for sequence alignment and does not guarantee finding the best alignment of two sequences [3]. Another approach, based on the identification of conserved motifs in the target gene, was taken by the authors of JOINSOLVER [90]. This program was developed for the specific task of analyzing the CDR3 region of rearranged Ig heavy chain sequences to characterize D segment usage, and thus does not analyze TCR sequences, though it does analyze Ig light chains. Neither program produces a final alignment that identifies the inferred origin of each nucleotide in the target sequence, nor does either program allow for gaps when performing alignments, although insertions and deletions are known to occur during somatic hypermutation [88].

Our solution, dubbed SoDA for Somatic Diversification Analysis, is based on dynamic programming, which has become the standard approach for pairwise sequence alignment because of its simplicity and guaranteed optimality. Traditional dynamic programming algorithms for sequence alignment [60],[89] rely on the equivalence of

pairwise alignments and paths through a two-dimensional lattice. Our algorithm is a generalization of these approaches based on an equivalence between V(D)J alignments and paths through a three-dimensional (3D) lattice as described in detail below.

2.1 Approach

The solution to the V(D)J problem consists of the identification of a set $\mathcal{G}^* = \{\mathcal{V}^*, \mathcal{D}^*, \mathcal{J}^*\}$ of gene segments and a set $\mathcal{M}^*$ of modifications – recombination sites, point mutations (including insertions and deletions) and n nucleotide insertions – that when applied to $\mathcal{G}$ produces the observed target sequence $\mathcal{T}$, that is $\mathcal{M}(\mathcal{G}) = \mathcal{T}$, and such that the posterior probability, $P(\mathcal{G}, \mathcal{M}|\mathcal{T})$, is maximized. The posterior probability quantifies our confidence in the inference.

Bayes' rule gives

$$P(\mathcal{G}, \mathcal{M}|\mathcal{T})P(\mathcal{T}) = I(\mathcal{M}(\mathcal{G}) = \mathcal{T})P(\mathcal{G})P(\mathcal{M}) \tag{2.1}$$

where I is the indicator function, equaling one when its argument is true and zero otherwise. The basic assumption here is that $\mathcal{M}$ and $\mathcal{G}$ are independent given the constraint that $\mathcal{M}(\mathcal{G}) = \mathcal{T}$. Practically speaking, the information $-\log P(\mathcal{M})$ is based on empirical frequencies and specifies the cost function for the alignment algorithm.

When we first developed SoDA, there had been limited research into preferential usage of specific Ig or TCR gene segments over others. Livak, *et.al.*, found possible preferences for certain D and J gene segments in TCR rearrangements and Marshall,

et.al., showed preferential usage of certain V gene segments in murine Ig rearrangements [48, 51]. Few studies, however, characterizing segment-by-segment usage for functional human rearrangements have been published, especially for Ig heavy chain rearrangements. One study by Brezinschek, *et.al.*, did find frequencies for V, D, and J gene segment usage in IgM, but the study was based on serum samples from only two male donors [6]. Thus, limited data on such frequencies in the population were not available at the time of development. So, in performing these analyses, SoDA assumes that all V, D, and J gene segments are equally likely, rendering the $P(\mathcal{G})$ term in equation 2.1 a uniform.

2.2 Algorithm

SoDA proceeds through two stages. In the first, the set of viable V, D and J segments is chosen by independent unconditional pairwise alignments between the target gene and each candidate gene segment. As stated above, the optimality of a solution based on independent unconditional alignment is not assured, but we can eliminate gene segments that score too low to participate in the optimal solution to simplify the computationally intensive second stage.

Using a standard local alignment approach, SoDA first finds an alignment score for each V gene segment in the library. Once all V gene segments are aligned and scored, they are sorted; only those above a viability cutoff are retained. The process then repeats for J gene segments. For Ig heavy chains and TCR β chains, candidate D segments are evaluated by alignment against that part of target sequence that lies between the invariant cysteine encoded by V and the invariant

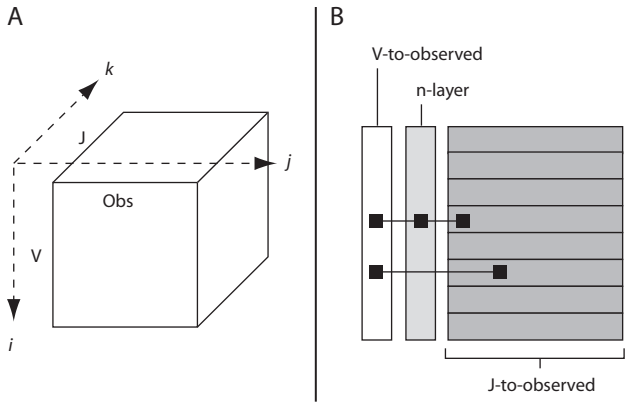

Figure 2.1: **A.** For aligning two contiguous sequences to a target sequence, the j axis represents the target sequence, while the other sequences are represented by the i and k axes. **B.** Aligning the J sequence to the target sequence horizontally allows a traceback path to traverse from any point in V to any point in J, while still allowing for movement in the n-layer if necessary.

tryptophan/phenylalanine encoded by J, the positions of which are discovered by the initial alignments to V and J.

The second stage of SoDA is the simultaneous alignment of all segments in the remaining viable sets. This stage uses a novel 3-dimensional dynamic programming alignment algorithm to make its final sequence inference, which we now describe in detail.

2.2.1 Two Contiguous Sequences (Light Chains)

We started by designing an algorithm to align a target sequence to two contiguous sequences which can overlap and/or have a random number of additional nucleotides in their junction, as in the case of light chain rearrangements. Using a three dimensional matrix, we conceived the front vertical plane to represent the alignment of the V gene segment to the target sequence, where the vertical axis i represents the V gene segment and the horizontal axis j represents the observed input sequence

(Fig. 2.1A). The third axis k would then represent the J gene segment, and thus each horizontal plane in the matrix represents an alignment of the J segment with the target sequence. The front V plane is further separated from the matrix by an n layer, which allows for random n-nucleotide additions (Fig. 2.2A). A path, then, for the alignment through the matrix would first follow an increase in the i and j axes while k remained constant, representing an alignment of the V gene segment with the target sequence. Then, at some point, the path would pass into the n layer and traverse through it, meaning that j continues to increase while movement in the i and k dimensions remain constant, thereby representing n-nucleotides. Or, the path would jump directly from some point in the V plane to some point in one of the horizontal J planes to account for situations where there had been no n-nucleotide additions and instead an overlap of the V and J segments (Fig. 2.1B). From that point on, the path would be horizontal only, with i remaining constant and j and k increasing, representing an alignment of the target sequence with the J gene segment. Accounting for exonuclease activity in rearrangements requires performing this alignment in three dimensions, to allow for a path from any point in the V segment to any point in the J segment while still being able to move through the n-layer.

The algorithm was written such that the dynamic programming matrix h was computed by individual vertical layers. Let $s(X)$ be the length of a gene segment X, where X can be the V, D, J, or target sequence (T). Then, we define one vertical layer of the dynamic programming matrix h as $h[1 \dots s(V)][1 \dots s(T)][k]$ for all integers k, such that $1 \leq k \leq s(J)$. Note that one vertical layer constitutes one unit along the k axis. Computing the matrix by layers is necessary since the criteria evaluated for

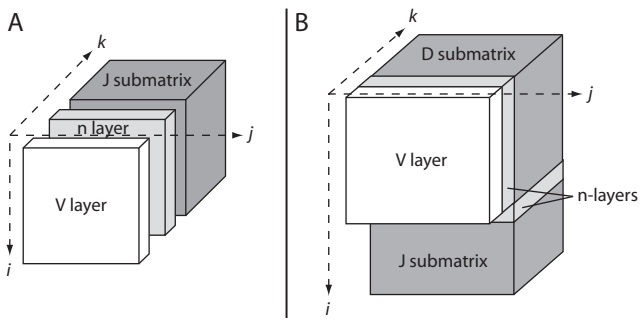

Figure 2.2: **A.** Allowing for a random number of n-nucleotides necessitates the addition of a special layer in the dynamic programming matrix. The layer is inserted between the V-to-target sequence alignment plane and the J-to-target alignment submatrix. **B.** The matrix for aligning three sequences requires that the second n-layer and submatrix be appended to the bottom, and offset by 2 units in the k direction, to account for the front vertical V and n-layers.

each matrix position differs per layer for the V, n-, and initial J submatrix layers. First the V layer is filled in completely, calculating the value at each position using the typical dynamic programming alignment algorithm rules:

$\forall$ i, j, and k, such that $1 \leq i \leq s(V)$, $1 \leq j \leq s(T)$, and for k=1:

$$h[i][j][1] = \max \begin{cases} h[i-1][j][k] - gap \\ h[i-1][j-1][k] + m \\ h[i][j-1][k] - gap \end{cases}$$

where m is the score of a match or mismatch of the nucleotides being compared.

Once complete, the algorithm moves to the next layer back, where k=2. This is the n-layer, which is a buffer layer between the V and J layers to accommodate the addition of n-nucleotides, so the rules for calculating $h[i][j][k]$ here are different. Generally, vertical movement within a layer would imply gaps in the target sequence. In the context of the problem, however,an insertion or deletion that occurs within the set of n-nucleotides is simply interpreted as the existence or non-existence of an n-nucleotide. Thus, in the n-layer, vertical movement along the i axis not defined:

$\forall$ i, j, and k, such that $1 \leq i \leq s(V)$, $1 \leq j \leq s(T)$, and for k=2,

$$h[i][j][2] = \max \begin{cases} h[i][j-1][k] \\ h[i][j-1][k-1] \end{cases}$$

The algorithm then proceeds to fill in the submatrix for the J layers, where $k \geq 3$. Here, each calculation must evaluate not only adjacent positions, as in typical alignments, but must also consider positions in both the n- and V layers due to the constraint of having a path that can move from any position in the J sequence to any position in the V sequence, and thus there are five maximum value options:

$\forall$ i, j, and k, such that $1 \leq i \leq s(V)$, $1 \leq j \leq s(T)$, and $3 \leq k \leq s(J)$,

$$h[i][j][k] = \max \begin{cases} {}^{1}h[i][j-1][k] - gap \\ {}^{2}h[i][j-1][k-1] + m \\ {}^{3}h[i][j-1][2] + m \\ {}^{4}h[i][j-1][1] + m \\ {}^{5}h[i][j][k-1] - gap \end{cases}$$

Options 3 and 4 above are specifically for considering the n- and V layers, respectively. Note that when k=3, options 2 and 3 are redundant. Thus, for that layer only, there are four options.

As with standard alignment algorithms, a traceback path is created and saved as the matrix is computed. When the matrix is complete, traceback begins at the position in the J submatrix where the value is the greatest.

2.2.2 Three Contiguous Sequences (Heavy Chains)

Building upon the concepts for light chain rearrangements, we developed the algorithm necessary to accommodate heavy chain rearrangements, where there is a third

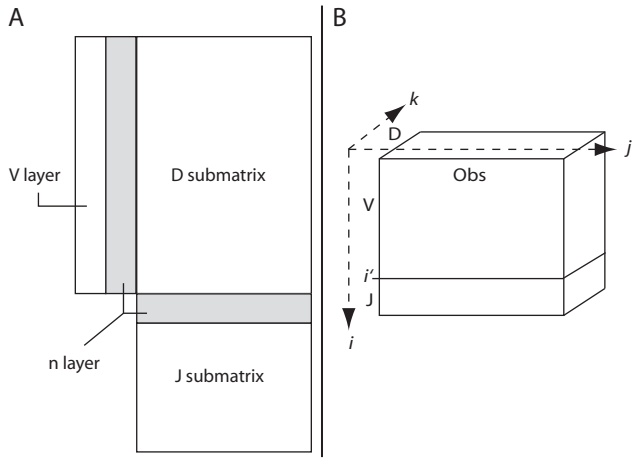

Figure 2.3: **A.** To account for the addition of a third sequence, a second n-layer and second submatrix are added beneath the original alignment matrix. **B.** The third sequence addition requires associating two of the sequences to one of the axes. For Ig heavy chains, the J sequence is represented along the i axis with the V sequence.

contiguous sequence to be aligned to the target sequence. The key to the algorithm is to replicate the idea of an n-layer and a second 3D submatrix for the additional sequence, but to append both beneath the existing layers required for the alignment of the first two sequences (Fig. 2.3A). This is accomplished by designating the vertical axis i to represent the V sequence, an n-layer, and the J sequence, so that $i = s(V) + s(J) + 1$. The D sequence, then, is represented along the k axis, and the target sequence is still represented by the j axis (Figs.2.2B,2.3B). Arranging the sequences in this way allows for n-nucleotide additions and a path from any point in J to any point in D, while maintaining those same properties of movement for the D and V sequences. The matrix, then, is initially filled out using the same calculations as the two sequence alignment procedure, but with the additional steps of filling in the n-layer for the D-J junction and the submatrix for the J gene segment alignment to the target sequence. Thus, the algorithm switches from computing vertical layers,

to computing horizontal layers. Computing this second n-layer is similar to computing the first, except that movement along the k axis is undefined instead of the i axis. Thus, the second n-layer is filled out using the following rules:

For $i = i' + 1$ and $\forall$ j and k, such that $1 \leq j \leq s(T)$ and for $k \geq 3$,

$$h[i][j][k] = \max \begin{cases} h[i][j-1][k] \\ h[i-1][j-1][k] \end{cases}$$

where i' is the position along the i axis at which the best alignment of D to the target sequence, relative to V, occurs. At each i, the horizontal alignment of D to the target sequence is similar, differing only by values that come from the vertical V layer. Thus, at some horizontal layer at height i, the alignment of D to the target sequence will be maximized relative to a position in V, which we define as the vertical position i' where the traceback path will pass through into the V layer (Fig.2.3B). In terms of alignment to the target sequence, the V and J sequences are independent of each other. So, the computations for the second n-layer and the subsequent J submatrix begin at $i'+1$ and k begins at 3, since k=1 and k=2 represent the vertical V and n-layers, respectively.

The J submatrix is filled out in a similar manner to the D submatrix, considering values in the n-layer and in the D submatrix horizontal layer at height i', in addition to adjacent positions within the matrix:

$\forall$ i, j, and k, such that $i' + 2 \leq i \leq s(J)$, $1 \leq j \leq s(T)$, and $3 \leq k \leq s(D)$,

$$h[i][j][k] = \max \begin{cases} {}^1h[i][j-1][k] - gap \\ {}^2h[i-1][j-1][k] + m \\ {}^3h[i'+1][j-1][k] + m \\ {}^4h[i'][j-1][k] + m \\ {}^5h[i-1][j][k] - gap \end{cases}$$

Note that when $i = i' + 2$, options 2 and 3 are redundant, and thus for the topmost horizontal layer in the J matrix, the algorithm chooses from only four options.

2.2.3 Traceback

Once the matrix is entirely computed, the traceback procedure begins at the cell in the J submatrix containing the maximum value. In general, traceback proceeds by moving concurrently along the i and j axes, while k remains constant, finding the alignment of the J sequence to the target sequence, respectively. Then, the traceback path moves into either the n-layer and then the D submatrix, or directly into the D submatrix. Movement within the n-layer is solely along the j axis. At $i = i'$, traceback proceeds through the D submatrix by moving concurrently along the j and k axes, finding the alignment between the D gene segment and the target sequence. Then, the path moves into the vertical n-layer for the V-D junction, or jumps directly into the V layer. Again, movement in the n-layer is solely along the j axis. Once in the V layer, movement proceeds concurrently along the i and j axes until either i=1 or j=1, finding the alignment of the V gene segment to the target sequence. All movement within the n-layers is designated as the addition of n nucleotides in the junction between the respective gene segments. Thus, once the traceback is complete, the resulting alignment is the best possible alignment of the gene segments, with n-nucleotides and gaps, to the target sequence.

2.2.4 Expanding the 3D Algorithm Beyond Three Sequences

Though the 3D algorithm described here only aligns three subsequences to a target sequence, we believe that the fundamental concepts established here allow for expanding this alignment algorithm beyond three subsequences. In such a case, each additional sequence would be represented either by the k or the i axis, alternating by each addition. Also, a new 3D submatrix for aligning this new sequence would need to be appended in the same direction as the axis that represents the new sequence. In general, odd numbered sequences would be represented by the i axis, while even numbered sequences would be represented by the k axis. For example, in the case of our SoDA implementation, the third sequence, the J gene segment, is represented along the i axis with the V gene segment and the submatrix for its alignment to the target sequence is appended below the matrix for the D gene segment, the second sequence. So, adding a fourth sequence would require that the k axis represents the new sequence and that an additional submatrix for aligning this sequence to the target would be added in the direction of the k axis, adjacent to the lower submatrix for the third subsequence alignment. Adding further sequences would thus create a set of submatrices arranged like a staircase, moving in the direction of the i and k axes.

2.3 Results and Discussion

To test SoDA, we developed a simulation program to generate artificial Ig sequences. The program simulates V(D)J recombination by randomly selecting a V, D, and J gene segment, and selecting the effective recombination site for each segment from a

	2.5% mutation			5% mutation			10% mutation			15% mutation		
Correct	SoDA	JS[a]	VQ[b]	SoDA	JS	VQ	SoDA	JS	VQ	SoDA	JS	VQ
V	30	30	30	30	26	29	30	25	29	30	23	28
D	28	26	16	28	28	10	27	25	18	22	19	16
J	30	29	29	30	30	28	29	28	29	28	30	30
VDJ	28	26	16	28	25	8	27	20	15	22	15	15

Table 2.1: [a] JOINSOLVER [90], [b] IMGT/V-QUEST [28]. At each mutation rate, 2.5 %, 5%, 10%, and 15%, we generated 30 sequences and tested them using the above mentioned programs. The results show the number of correct inferences out of 30 that each of the three programs made for each gene segment and for the rearrangement as a whole. A correct inference for the rearrangement means that all three gene segments were correctly identified, while allowing for mismatched alleles.

uniform distribution extending five nucleotides to either side of the nominal recombination site. N nucleotide addition is simulated by randomly choosing the number of n-nucleotides to add to each junction from the uniform distribution up to 12. Somatic hypermutation is simulated by introducing point mutations independently at each position with transition/transversion ratio fixed by empirical frequencies [88]. The probability of mutation per nucleotide was varied to produce expected mutation frequencies of 2.5%, 5%, 10%, and 15%; we generated 30 artificial sequences at each mutation rate.

We used these sequence sets to compare the inferences generated by SoDA, JOINSOLVER and V-QUEST. All three programs use the same V,D, and J gene segment libraries from IMGT, except that JOINSOLVER adds five "irregular D" segments to the D segment library. We chose four sets of 30 sequences as a reasonable test, given that only SoDA among the three programs can process sequences in batch. The results of our tests are shown in Table 2.1. As expected, as the mutation rate increased, SoDA's accuracy declined, though at each mutation rate, SoDA identified more rearrangements correctly than did either JOINSOLVER or V-QUEST. SoDA

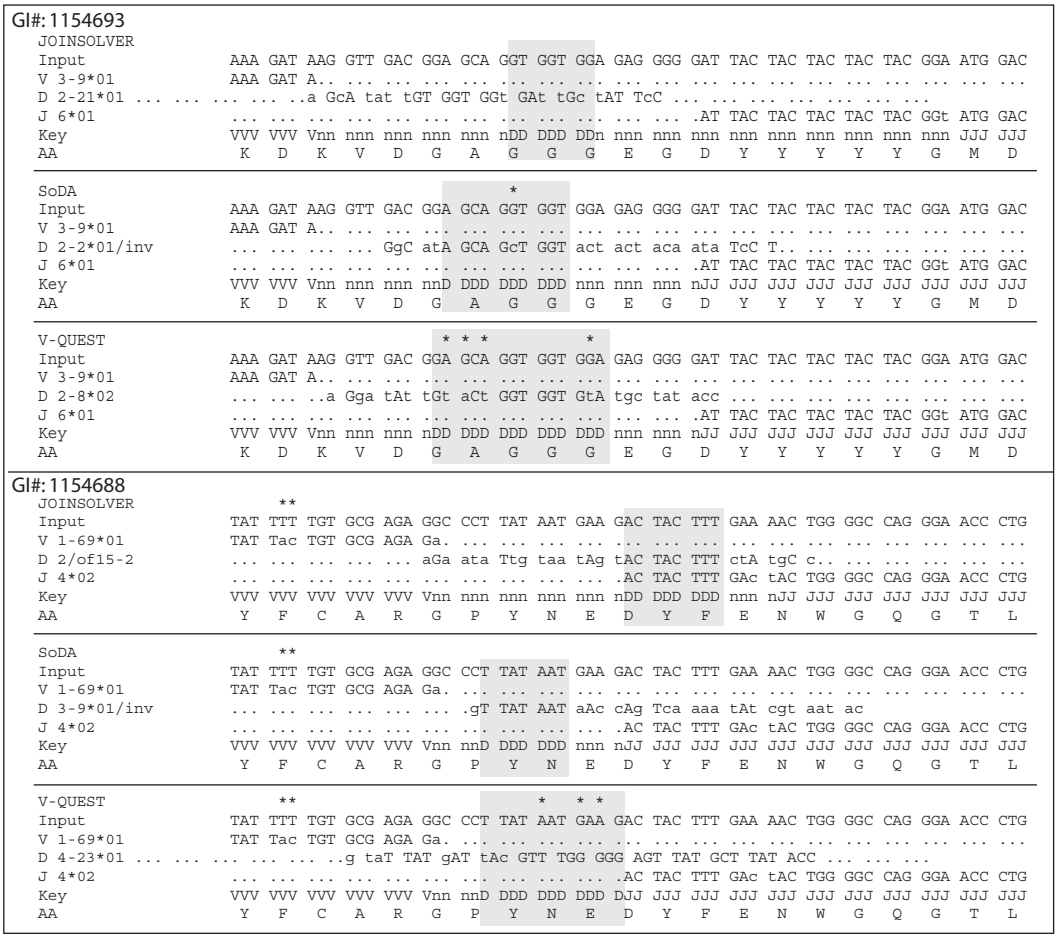

Figure 2.4: For both GI numbers 1154693 and 1154688, each of the three programs, SoDA, JOINSOLVER, and V-QUEST, returned the same V and J gene segment inference, but a different D segment. Above is a summary of the alignments, per program, of the D gene segment selected and its junctions with the V and J segments. The italicized n's in the key for JOINSOLVER show where we believe JOINSOVLER's inference can be improved, by changing them to Js. For GI1154693, SoDA's D gene segment identification seems to produce the most favorable alignment, having only one mutation compared to the four mutations of V-QUEST's alignment, and having a significantly lower number of n-nucleotide additions than JOINSVOLER's alignment. For GI1154688, JOINSOLVER's D gene segment identification produces the longest D segment alignment, but SoDA's choice produces the best overall alignment between the V, D, and J gene segments.

identified the correct VDJ combination in 28, 28, 27, and 23 cases out of 30 at 2.5%, 5%, 10%, and 15% mutation rates, respectively. Also, for each test, SoDA correctly identified all 30 of the V segments correctly. JOINSOLVER did not perform as well

identifying V gene segments (though this is not unexpected since JOINSOLVER was developed specifically for analyzing CDR3 regions, and not V gene segments). V-QUEST identified fewer D segments correctly. All of the programs were able to identify the J gene segments correctly, and at the 15% mutation rate, JOINSOLVER and V-QUEST both performed slightly better than SoDA at this task.

We then tested each of these programs on a set of 30 sequences selected randomly from a set of 650 rearranged Ig heavy chain sequences. Of these 650 sequences, 547 came from the set of sequences used by [90] to test JOINSOLVER (Genbank accession numbers Z68345-487 and Z80363-770), and the remaining 103 sequences were added to this set after searching Genbank for "human immunoglobulin heavy chain variable region" (accession numbers AM050894-1008). After testing the three programs using all 30 sequences, we found that the programs agreed in their V, D, and J identifications on 18 of these genes. For the remaining 12 sequences, SoDA never made a V or J identification that was not supported by one of the other two programs. In two of these 12 cases, the only difference was in JOINSOLVER's V segment prediction. In one case JOINSOLVER's V choice differed from that of SoDA and V-QUEST, and the D identified by V-quest differed from that of SoDA and JOINSOLVER. In three other cases, V-QUEST simply did not identify both the D and J gene segments, though JOINSOLVER and SODA agreed on all three gene segments. There was one instance where JOINSOLVER's D choice differed from that of the other two programs, and a similar instance where V-QUEST's selection of a D segment was the only difference. For the remaining four instances, the three programs chose different D gene segments. The alignments for two of these cases are

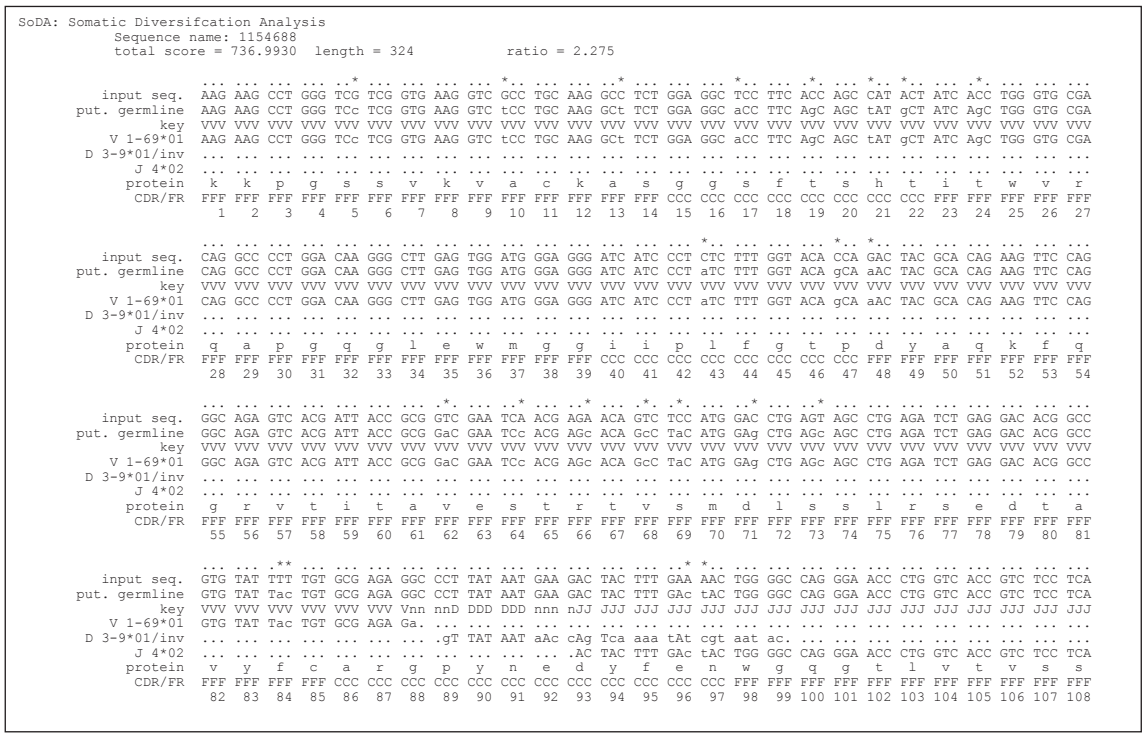

Figure 2.5: The complete output of SoDA's inference for sequence 1154693. SoDA provides a text file in this format for each sequence it analyzes.

shown in Fig.2.4.

For sequence 1154693 (GI number), SoDA's D gene choice appears arguably to provide the best fit (Fig. 2.5). JOINSOLVER's inference can be improved as shown where the key is italicized. Nevertheless, we believe SoDA's alignment to be the best due to its smaller relative mutation frequency and the smaller number of n-nucleotides required. For sequence 1154688 (GI number), JOINSOLVER found the longest D gene segment alignment without mutations, but uses an unusual reading frame in gene segment not found in the IMGT IGHD library (2/OF15-2) and uses 15 n-nucleotides in the V-D junction. V-QUEST aligns 11 bases of its D prediction (4-23*01) with 3 mutations, while SoDA's uses a common D segment (3-9*01), though

it is inverted, with no mutations and only four n-nucleotides in the V-D junction. Though both SoDA and JOINSOLVER use four n-nucleotides in the D-J junction, JOINSOLVER's D alignment forces the J alignment further downstream, and therefore misses aligning the first 12 bases of the J segment to the target sequence. This example illustrates the dependence among segment alignments that motivated the development of SoDA. Although JOINSOLVER has found a D segment alignment that score higher than that of the SoDA solution, when the respective complete alignments are considered, SoDA's use of a lower-scoring D segment is more than offset by the J segment alignment that this choice allows (Fig 2.5).

2.4 Additional Features

In addition to humans, SoDA can analyze both Ig and TCR sequences from mouse and opossum, and Ig sequences from rhesus macaque. SoDA also allows for modification of the mutation frequency input parameter, which directly affects the scoring matrix calculations, enabling for stricter or more liberal predictions. Such a setting is necessary for analyzing heavily mutated Ig sequences, like anti-HIV antibodies or Ig from chronic B-cell lymphoma patients. SoDA also provides graphical analyses of hydrophobicity and charge comparisons between the input and the statistically predicted germline sequences. Most recently, the ability for SoDA to report lists of top gene segments along with its prediction was added. This enables users to select and view alignments using segments that SoDA did not choose in its optimal alignment. Currently, SoDA lacks the ability to infer multiple tandem D segments within TCR and Ig genes: a deficit shared by V-QUEST but not by JOINSOLVER.

SoDA, however, is currently the only tool that can analyze up to 8000 sequences in batch, reporting to the user completion of the analysis via email. Additionally, improvements to the software enable a user to substitute other gene segments for the ones chosen in the optimal alignment to in order to test other possible alignments.

2.5 Summary

The antigen receptors of adaptive immunity are unique in their capacity for somatic diversification and consequently present unique challenges for bioinformatics. We have developed a software package built on a novel 3D generalization of pairwise alignment algorithms to find the most likely recombination scenario, including the gene segments and their somatic modifications, to account for any given TCR or Ig sequence. We tested our system, SoDA, against a set of artificial Ig sequences produced by simulating the rearrangement process, and a set of real human Ig genes chosen from Genbank. SoDA performed well, and compared favorably to existing programs JOINSOLVER and V-QUEST. The cost of this performance enhancement is a substantially increased computational effort. Although validated here on human Ig heavy chains only, the program performs comparably on all Ig and TCR loci in humans. Therefore, the information SoDA provides allows for a detailed analysis of the population somatic genetics of the immune response.

Chapter 3

Data Collection and Filtering

Research objectives frequently require the use of specialized tools such as particular reagents, machines, statistical techniques, or software. SoDA is such a tool. Its development was in itself a research project, but now that it has been tested and peer reviewed[96], its true value comes in its usefulness as a research tool. SoDA made the type of large-scale analysis we wanted to perform possible and efficient. Such an analysis, however, also required a large set of Ig heavy chain sequence data.

3.1 DNA Sequences

We set out to compile sets of human immunoglobulin heavy chain gene sequences that are representative of natural immunoglobulin diversity, excluding clonally related genes and genes of perinatal origin. To do so, we submitted the search "human[orgn] heavy[titl] immunoglobulin[titl]" to the Genbank nucleotide database which returned 16,870 results. We label this large collection of Genbank sequences as the P dataset, for "productive Ig". We then identified 1,167 autoreactive Ig rearrangements in Gen-

bank using keyword searches with terms such as "autoreactive", "immunoglobulin", "autoantibody", and "heavy". We refer to this set of genes as the autoreactive dataset (A). We further identified two additional gene sets that would facilitate particular analyses. We identified a set of 608 gene sequences from a study conducted of the synovial B-cells of rheumatoid arthritis patients [57], which we refer to as the RA dataset. The fourth dataset is a collection of 6,329 genes used in a study by Ohm-Laursen, et.al., that were gathered from the serum of 28 healthy human adult volunteers using a primer for a specific VH gene segment, VH3-23 (17). We refer to these genes as the OL dataset. We downloaded each set of DNA sequences, preprocessed and filtered them as described below, and analyzed them, using SoDA, for gene segment usage, point mutations, n-nucleotide addition, and recombination junctional diversity.

3.2 Classification by Productivity

Each dataset was then divided into three groups on the basis of the inferred original, pre-somatic mutation productivity. We classified those sequences that appeared to have been originally rearranged out of frame by virtue of the V segment being mutually out of frame with the J segment, excluding indels, as non-productive. These non-productive sequences were grouped together into a fifth dataset (NP). We classified those sequences that had no stop codons and both invariant V cysteines and the invariant J tryptophan in-frame and intact as productive. All others were classified as indeterminate and omitted from further consideration.

Many non-productive sequences come from B-cells that have rearranged gene

segments on both alleles. In this situation, the first rearrangement did not yield a transcribable gene due to the presence of stop codons in the coding sequence or out-of-frame gene segments which code for a protein incapable of producing the necessary tertiary structure. These B-cells get a second chance to rearrange to produce a functional receptor, and thereby avoid deletion, by rearranging the second allele. Successful rearrangement of this allele yields a B-cell that presents a functional receptor, yet also carries a non-functional rearrangement. Thus, these non-productive genes provide a set of sequences that represent raw products of recombination that have been uninfluenced by selection pressures for antigen, therefore enabling comparisons that can show the effects of selection on the Ig heavy chain repertoire.

3.3 Filtering

We filtered the SoDA results in each set to remove clonal duplicates, which we defined to be those sequences that were inferred to use the same V, D, and J gene segments, had the same inferred CDR3 length, and have similar Genbank accession numbers. Sequences containing these matches and differing only by point mutations were still considered clonal duplicates. Where groups of clonally related genes were identified, a single representative was chosen at random and the others were omitted. We then grouped the sequences in each set by study of origin and removed any large sets of sequences that came from the same study.

The large P dataset was further filtered to remove those sequences that, by their own Genbank annotations, indicated origin from neonates or cordblood. These sequences were also filtered to remove any sequences that may be from autoreactive

patients due to the inclusion of at least one of the following words in the Genbank record: "auto*", "anti*", "self-reactive", "anti-self", "lupus", "rheumatoid", "sjogren", "diabetes", "sclerosis", "wegener", "crohn", "addison", "scleroderma", "grave", "psoriasis", "celiac", "vasculitis", "colitis", and "thyroiditis".

Each sequence in the autoreactive dataset was then manually screened to ensure that its Genbank record indicated autoreactivity. The A dataset was also filtered to remove those genes that were specifically anti-DNA, since these genes would bias the CDR charge measurements. The final set of productive, non-autoreactive genes (P) contained 6490 sequences; the final set of non-productive genes (NP) contained 325 sequences; the final set of autoreactive genes (A) contained 264 sequences; the final set of rheumatoid arthritis genes (RA) contained 608 sequences; and the final set of OL sequences contained 5403 productive sequences (OL-P) and 434 non-productive sequences. These character designations for the filtered datasets are used throughout the remainder of this document.

Chapter 4

Analysis of Human Ig Heavy Chain Genes

Having assembled and filtered several large datasets of human heavy chain genes, and having built a tool to enable the initial analyses, we set out to study the genetic properties of these genes with a focus on studying biases that may exist in the mechanisms that form these genes. We initially focused on analyzing the large set of P sequences, comparing and contrasting their genetic and molecular characteristics to those of the NP genes. Since the NP genes have not been subject to selection, this comparison should help to elucidate selection-induced biases in the repertoire. What follows are the results of a comprehensive characterization of the nearly 6,500 human Ig P genes in terms of V, D, and J gene segment usage, n-nucleotide addition, and CDR3 length, and an analysis of the molecular mechanisms involved in their gene creation. We include a detailed characterization and comparison of those sequences to the 325 non-functional NP genes. One of our more striking findings is the existence of strong pairing preferences among D and J gene segments. We hypothesize that these results may be due to repeated sequential rearrangement of D and J seg-

ments and present a statistical model that illustrates the efficacy of this mechanism for producing the observations. In addition, we have found that the n-nucleotide tract lengths in both he V-D and D-J junctions are well-fit by a negative binomial distribution. Differences in tract length distributions between these two junctions are characterized by specific differences in the parameters of the distribution, which can be interpreted in terms of mechanisms of n-nucleotide polymerization.

4.1 Results

4.1.1 Preferred Pairing Among Gene Segments

We performed contingency table analyses to investigate whether there is preferred gene segment pairing between D and J segments in the P genes. We tabulated the frequency of occurrence of each D-J pair and used a contingency table to compare these frequencies with those expected under the null hypothesis of independent selection. The extremely low p-value ($p < 10^{-50}$) for the chi-square analysis indicates that the D and J segments are not independent. To measure the degree of departure from independence of each pair, we calculated adjusted residuals, which are approximately independent and distributed as standard normals [21]. So, values greater than 1.96 or less than -1.96 for particular D-J pairs represent a significant departure from the expected value at a 95% confidence level and are therefore evidence for a correlation between that particular D and J segment. We analyzed the P sequences only since the number of NP sequences was insufficient for this analysis.

Our data show that certain pairs of D-J segments have frequencies significantly different from what is expected under the null hypothesis (Table 1, Fig. 4.1). For

	P sequences		NP sequences	
	Obs.	Rel. freq.	Obs.	Rel. freq.
D1-1	133	0.020	2	0.006
D2-2	811	0.125	76	0.234
D3-3	498	0.077	26	0.080
D4-4	78	0.012	1	0.003
D5-5	192	0.030	8	0.025
D6-6	141	0.022	5	0.015
D1-7	99	0.015	3	0.009
D2-8	129	0.020	4	0.012
D3-9	246	0.038	6	0.018
D3-10	547	0.084	17	0.052
D5-12	144	0.022	3	0.009
D6-13	295	0.045	22	0.068
D1-14	56	0.009	3	0.009
D2-15	268	0.041	22	0.068
D3-16	313	0.048	17	0.052
D4-17	263	0.041	10	0.031
D6-19	425	0.065	14	0.043
D1-20	10	0.002	0	0.000
D2-21	184	0.028	8	0.025
D3-22	527	0.081	28	0.086
D4-23	94	0.014	4	0.012
D5-24	153	0.024	3	0.009
D6-25	26	0.004	0	0.000
D1-26	364	0.056	13	0.040
D7-27	101	0.016	7	0.022
D0-IR	72	0.011	5	0.015
D1-IR1	94	0.014	6	0.018
D1-OR15	11	0.002	1	0.003
D2-IR2	33	0.005	2	0.006
D2-OF15	88	0.014	6	0.018
D3-OR15	47	0.007	2	0.006
D4-OR15	30	0.005	0	0.000
D5-OR15	18	0.003	1	0.003
	6490		325	

Table 4.1: Observed counts and relative frequencies of individual D segment usage in the P and NP datasets.

	J1	J2	J3	J4	J5	J6
D1-1	0.262	-0643	0.160	0.972	1.438	-2.153
D2-2	-3.052	5.502	-5.819	-10.729	-1.142	16.705
D3-3	-2.665	-0.817	0.264	-2.898	2.346	2.891
D4-4	-0.010	0.309	-1.419	1.349	-0.075	-0.459
D5-5	-0.672	-0.914	-0.674	2.137	-0.834	-0.611
D6-6	0.977	0.214	-1.120	1.048	-0.964	-0.138
D1-7	0.499	0.459	-0.884	0.600	1.039	-1.182
D2-8	-0.111	-0.085	1.620	-1.408	1.314	-0.546
D3-9	-1.979	-1.085	0.356	2.825	-0.816	-1.541
D3-10	1.078	-1.438	0.332	-2.126	1.722	0.977
D5-12	-2.501	-1.265	1.295	1.843	-0.001	-1.448
D6-13	0.302	0.500	-0.985	1.888	2.619	-3.665
D1-14	4.568	-0.614	1.465	-2.389	4.321	-3.452
D2-15	2.245	0.120	-1.303	-0.997	1.810	-0.268
D3-16	0.079	-1.350	2.424	-2.115	-1.612	2.233
D4-17	0.096	1.605	-1.754	3.728	-2.160	-1.972
D6-19	-0.328	-0.777	-1.426	5.127	-0.653	-3.773
D1-20	-0.652	-0.578	-0.288	0.480	0.803	-0.387
D2-21	1.332	-0.400	3.541	-0.635	-0.401	-2.168
D3-22	2.207	-1.038	2.178	2.217	-2.151	-3.191
D4-23	0.619	-0.022	-0.705	1.483	0.612	-1.866
D5-24	-0.093	0.026	0.243	3.806	-2.301	-2.776
D6-25	1.933	-0.934	1.517	-0.417	0.566	-1.620
D1-26	0.874	-0.844	2.953	2.112	-1.672	-3.503
D7-27	-0.055	0.418	3.214	1.234	0.644	-4.517

Figure 4.1: A heat map showing adjusted residual values for D-J segment pairings based on contingency table analysis of the P sequence data. Adjusted residuals are approximately independent and distributed as standard normals. Values greater than 1.96 (white) or less than -1.96 (dark gray) represent a significant departure from the expected value at a 95% confidence level.

example, based on the marginal frequencies of D2-2 and J6 (13.3% and 25.3%, respectively), we expected a frequency of 3.3% for the D2-2/J6 pair. The observed frequency, however, was 6.5%, an increase of 94% over what was expected. Our segment pair observations highlight an interesting pattern of D-J correlations within the data. Several 5′ D segments showed increased frequency of pairing with the most 3′ J segments (J5 and J6) and decreased frequency of pairing with closer (chromoso-

mal distance) J segments (J1-J4). Some 3′ D segments, however, showed increased frequency of pairing with the closest J segments (J1-J4), but a decreased frequency for the furthest J segments, J5 and J6. These findings led us to the hypothesize that multiple successive D-J recombinations may occur prior to adjoining a V segment to the D-J pair. This hypothesis has been put forth before, but little evidence has been offered for this occurrence in humans [74, 90, 91].

To test our hypothesis, we developed a statistical model and estimated its parameters using a Markov Chain Monte Carlo (MCMC) method. We fit our model to the observed data by estimating probability vectors for D and J segment usage and a multiple recombination rate (MRR) parameter, ρ. Each component of the probability vectors gives the relative probability that the corresponding segment will be chosen during the recombination process at any stage. The MRR is the probability of a subsequent recombination occurring given that one just occurred and that segments of the same type remain to produce another recombination. The D and J parameter vectors are initialized to the marginal frequencies calculated from the observed D-J pair frequencies, and ρ is initialized to 0.10. The algorithm begins by first running a set of 600,000 recombination trials using the initialized D and J vectors. When the trials are complete, the D-J pair frequencies are compared with the observed frequencies and a chi-square value is established. One of the parameters in the D or J vectors, or ρ, is then selected at random and altered slightly and a new set of 600,000 primary recombination trials begins. For each primary recombination, a D and J segment are initially selected. All intervening segments between those selected are designated as unavailable and the probabilities of the remaining segments

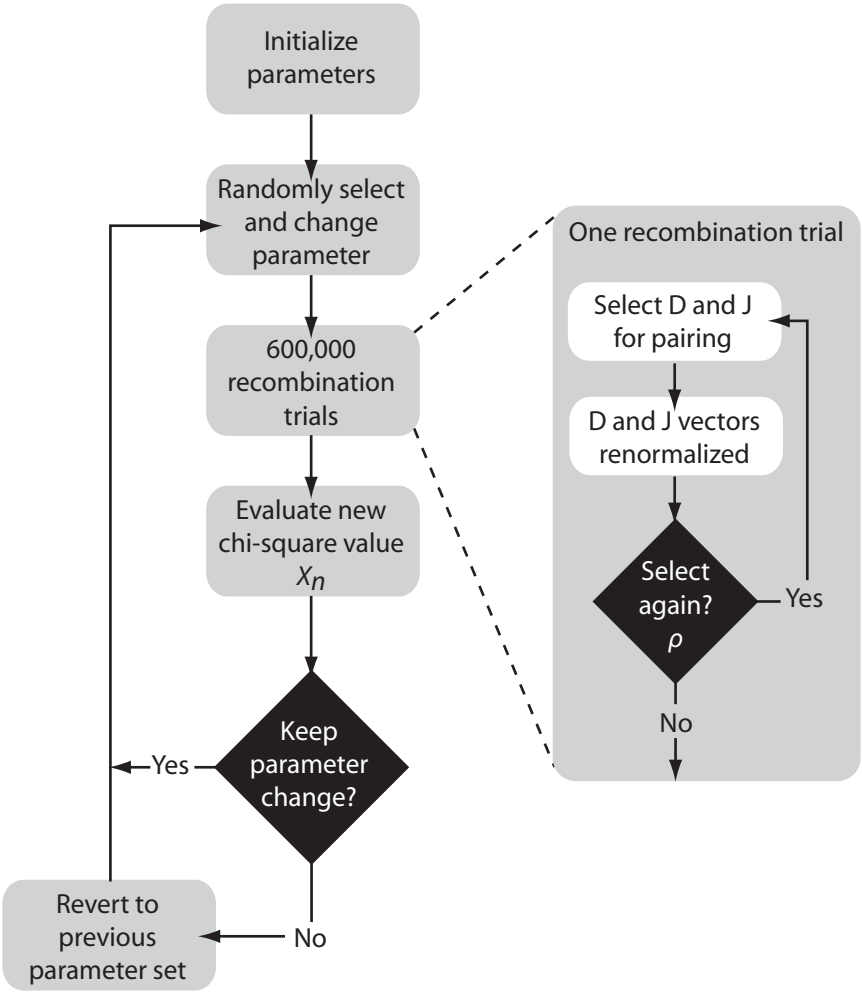

Figure 4.2: Flow diagram depicting the steps for estimating the multiple recombination parameters in our statistical model. Rho (ρ) is the parameter for multiple recombination; it represents the probability of a subsequent recombination occurring given that one just occurred and that segments are available for another recombination. Changes to the parameters are accepted stochastically according to the Metropolis-Hastings criterion: with probability 1 if the new chi-square (χ^2_{new}) value is lower than the previously computed value (χ^2_{old}), or with probability $\exp(0.5(\chi^2_{old} - \chi^2_{new}))$ [31].

are recalculated, normalizing them to represent the new restricted set of available segments. Then, with probability ρ, a subsequent recombination may occur. If this secondary recombination does occur, the probabilities of the remaining segments are again normalized. Subsequent recombinations may continue to occur in this manner

provided that there are segments available to recombine. If at any stage, the algorithm does not choose to make a subsequent rearrangement, the process terminates. It also terminates when no more segments can be recombined. At the completion of all 600,000 trials, the D-J pair frequencies of the trials are compared with the observed values, and a chi-square value is computed. The new parameter values are accepted stochastically according to the Metropolis-Hastings criterion: with probability 1 if the new chi-square (χ^2_{new}) value is lower than the old value (χ^2_{old}), or with probability $\exp(0.5(\chi^2_{old} - \chi^2_{new}))$ [31]. Otherwise, the algorithm reverts back to the previous set of parameters. This enables the algorithm to occasionally accept non-improving moves and thereby avoid being trapped in local minima. The algorithm then repeats, altering another parameter and performing a new set of trials (Fig. 4.2). The output of the algorithm represents a sample from the Bayesian posterior density on the parameters.

We ran the algorithm for 300,000 iterations in which each iteration included 600,000 recombination trials were performed. We found the best fit of our data to the model at $\rho = 0.198$. At this multiple recombination rate, the model produced a chi-square value of 503. This is very statistically different from 635 ($p < 10^{-30}$; chi-square test with 1 degree of freedom), the chi-square value observed when $\rho = 0$.

4.1.2 CDR3 Statistics

Our data show statistically significant differences in the length of CDR3 between the P and NP sequences ($p < 10^{-10}$). The P sequences have a shorter mean CDR3 length of 15.49 amino acids while the NP sequences have a mean length of 18.00

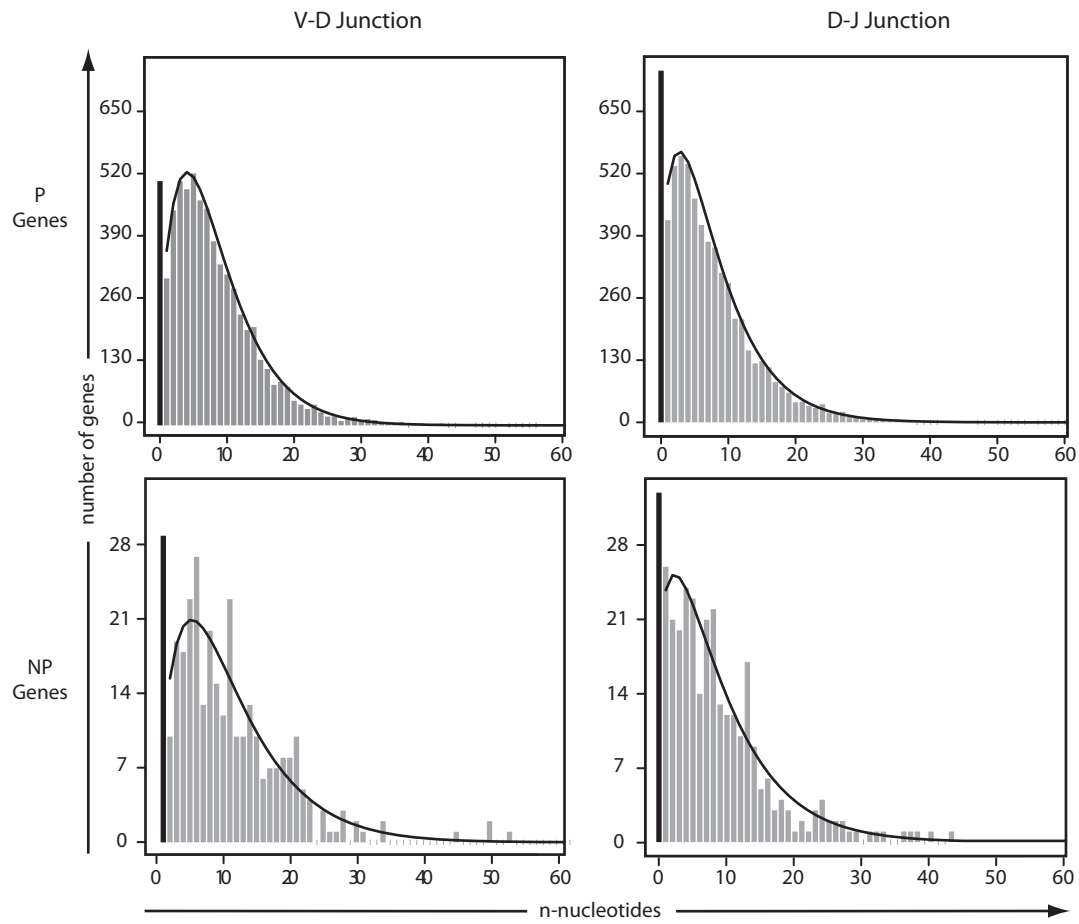

Figure 4.3: Plots of the observed n-nucleotide data for both the P and NP genes in both the VD and DJ junctions fit to a zero-inflated negative binomial distribution.

amino acids. In the V-D junction, an average of 7.86 and 9.78 n-nucleotides were added to the P and NP sequences, respectively: a statistically significant difference ($p < 0.001$). For the D-J junction, the data show statistically different averages of 7.04 and 8.26 n-nucleotides for the P and NP sequences, respectively ($p < 0.01$).

Plots of the observed n-nucleotide frequencies resembled plots of a zero-inflated negative binomial distribution. The negative binomial distribution is a discrete probability distribution for the number of independent Bernoulli trials required to achieve a fixed number, r, of successes. For both P and NP data of n-nucleotide additions in

	Junction	r	p	Mean n addition
P	V-D	2.24	0.21	7.86
	D-J	1.76	0.19	7.04
NP	V-D	1.85	0.15	9.78
	D-J	1.48	0.15	8.26

Table 4.2: Negative binomial parameters r and p produced from fits of the observed n-nucleotide data from the V-D and D-J junctions for both the P and NP gene sets. In our model, we interpret r to mean the number of detachments TdT experiences from the DNA.

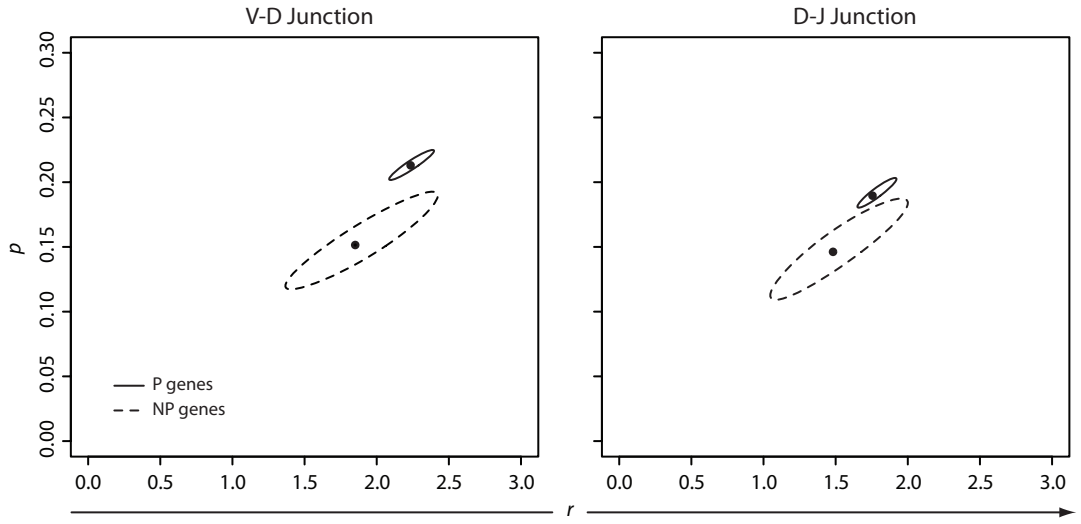

Figure 4.4: We fit our observed n-nucleotide addition data to the negative binomial distribution, and calculated both the maximum likelihood estimators plotted at (r,p) and the corresponding confidence regions.

both the V-D and D-J junctions, we fit our data to the negative binomial distribution and calculated the maximum likelihood estimator (MLE) for the parameters r and p, where p is the probability of getting a success in any given trial. (Table 2, Fig. 4.3). We then calculated 95% confidence regions (Fig. 4.4). We found that for the P sequences, $r < 2$ for the D-J junction but $r > 2$ for the V-D junction.

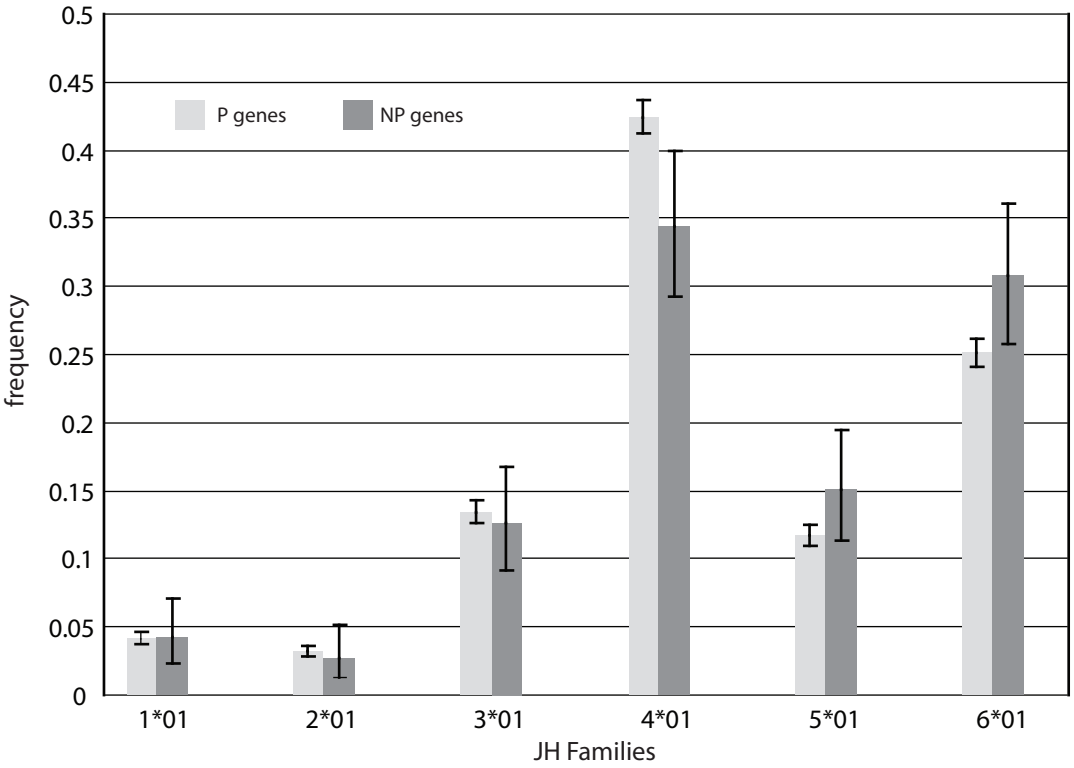

Figure 4.5: Observed relative frequencies of JH gene segment usage in the P and NP gene sets. Error bars represent 95% confidence intervals.

4.1.3 Gene Segment Usage Frequencies

The ability to detect biases statistically is made easier when the number of total categories is small. The J locus has fewer gene segments than either of the other heavy chain loci, and thus provides the best opportunity for the discovery of such bias in gene segment usage. Indeed, we find very strong departure from uniform segment usage in both P and NP sets ($p < 10^{-12}$) which both show a strong preference for J4 and J6 and substantially reduced frequency of J1 and J2 (Fig. 4.5). There are also differences in relative frequencies of J segment usage between P and NP genes ($p = 0.03$), with J4 under-represented by 18% among NP genes relative to P, and J5 and J6, over-represented by 27% and 21% in NP compared to P, respectively.

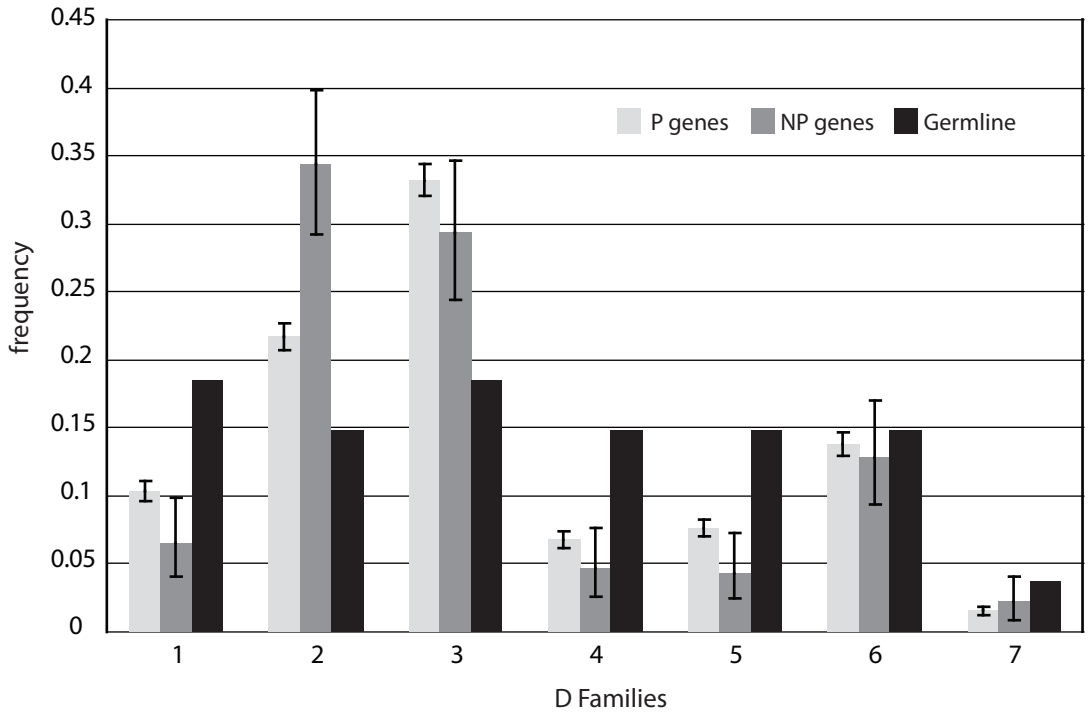

Figure 4.6: Relative observed frequencies of DH gene segment usage by family in the P and NP gene sets, and comparison to germline complexity of each gene segment family. The germline complexity refers to the number of segments within the locus assigned to each family. Error bars represent 95% confidence intervals.

D segments, which outnumber J segments by more than a factor of four, are organized into seven families based on sequence homology. At a family level, we compared usage of segments of both the P and NP sets to the genomic complexity of each family, which is the number of segments assigned to each family within the locus, and found a significant departure from these proportions as well ($p < 10^{-12}$) (Fig. 4.6). Again, we observe statistically significant differences in relative frequencies of usage between the P and NP sequences at both the family and individual gene segment levels ($p < 10^{-5}$). Family D2 is over represented among NP genes by 55% relative to the P sequences, but families D4 and D5 are under represented by 29% and 41%, respectively, relative to the P sequences (Fig. 4.6). Individually, we again

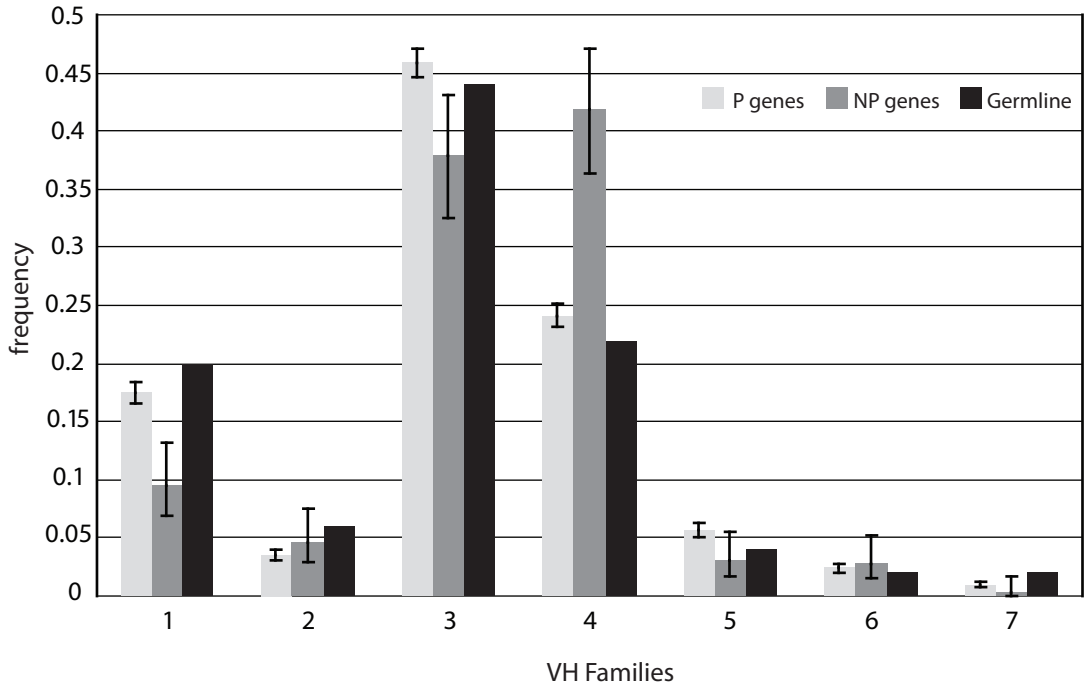

Figure 4.7: Observed relative frequencies of VH gene segment usage by family in P and NP sequences, and comparison to germline complexity of each gene segment family. Error bars represent 95% confidence intervals.

find a strong departure from uniform segment usage in both the P and NP sets ($p < 10^{-12}$). The most notable disparity is with segment D2-2, which is significantly over represented in the NP genes by 82% relative to the P genes. There was not a statistically significant difference in the number of inverted D segments observed between the P and NP genes.

Like the D segments, V segments are also classified into seven families based on sequence homology. The frequencies of V gene segment usage by family in the P and NP gene sets, compared to the genomic complexity of the V locus, are shown in figure 4.7. Though the observed frequencies produce a very significant rejection of the hypothesis that the usage of V segments in the P and NP gene sets are exactly proportional to the number of genes in each family ($p < 10^{-10}$), the P genes more closely

resemble the genomic complexity of the V locus than do the NP genes. Relative to the P genes, the NP genes differ significantly ($p < 10^{-10}$), with usage frequencies 44% and 17% below what is expected for families V1 and V3, respectively, but 67% greater than what is expected for family V4. Concerning individual segments, both P and NP genes used segments V3-23 and V4-34 most frequently, though V3-23 was the top segment in the P genes, but second to V4-34 in the NP genes.

4.2 Discussion

We provide here the most precise estimates of gene segment usage frequency currently available. The quantity of data that we assembled and analyzed has enabled us to estimate V, D, and J segment usage frequencies with tight confidence intervals. These data potentially give insight into the structural basis for differential segment usage in terms of either raw expression or somatic selection, though such elucidations are left for further research.

In addition, comparison of our results with previously published usage frequencies provides validation of our data collection methods and confidence that our P sequence dataset is representative of natural diversity as intended. In particular, an extensive study of Ig CDR3 diversity based on de novo sampling of Ig using a primer for a single VH gene shows D segment usage results remarkably similar to our own, based on a Spearman rank correlation score of 0.93 [64]. This in spite of the fact that D segments are notoriously challenging to identify within Ig genes due to recombination site choice, flanking 5' and 3' n-nucleotide addition [18], and somatic mutation [54, 98, 67]. With J segments, furthermore, our data are consistent with published findings that

indicate that segment J4 is used most frequently, followed in descending order by segments J6, J5, J3, J1, and J2 [7, 103, 100] (Fig. 4.5).

For V segments, our data again provide statistical evidence in support of published findings. With individual segments, our data support previous results showing that segment V3-23 is the most frequently used [7] in productive rearrangements, and that gene V4-34, which we found to be used second most frequently, has high usage within adult peripheral lymphocytes [100]. Like the J segments, individual segment usage can vary, but in spite of that, segment usage at the family level approximates expected usage based on literature. Our data support findings that show that segments in family V3 are used most frequently, followed in descending order by V4, V1, V5, V2, V6, and lastly V7 [7, 6]. We have also shown consistency with findings that, with some variation, the distribution of V gene usage by family shows similarity to germline complexity of the known segments [7] (Fig. 4.7). Our data showed this to be especially true for families V1, V3, and V4.

The NP sequences showed an enhancement of segment usage from family V4 at the expense of segments from family V1, due primarily to a 67% increase in usage of segment V4-34 from what was expected. Segment V4-34 has been reported to be over-represented in the adult human repertoire [41], and has also been implicated in generating autoreactive B-cells in SLE patients and against cold agglutinins [68, 95, 4]. Since the NP sequences are not subject to selection, those sequences coding for autoreactive receptors would not be deleted from the repertoire. Also, V4-34 has been shown previously to be limited by selection in the expressed human Ig repertoire due to lowered usage of this segment between IgM and IgG populations

[42]. Thus, V4-34 is likely not enhanced in autoimmune disorders, but instead is selectively limited in the P sequences.

Having validated our data collection methods, we focused on analyzing the genetic mechanisms involved in V(D)J recombination. One such mechanism is n-nucleotide addition by TdT. The zero-inflated negative binomial model fits these data well enough for us to seek an interpretation of its three parameters. We develop this interpretation in terms of two states: TdT attached to one of the unjoined DNA ends, or unattached. The probability that TdT never attaches is the first parameter, the zero-inflation factor. When attached, TdT either adds another nucleotide or becomes detached, with probabilities p and $1 - p$, respectively. In this context, the final parameter, r, has a natural interpretation as the number of times TdT detaches before the joint is closed.

We found that for the P sequences, $r < 2$ for the D-J junction but $r > 2$ for the V-D junction (Table 2). This pattern is consistent with a greater TdT concentration during the V-D joining process relative to that during the D-J process.

Studies of TdT expression during B-cell ontogeny show high levels TdT mRNA during the pro-B and late pro-B stages of development – the stages in which the D-to-J and the V-to-DJ rearrangements occur, respectively [47, 101]. Specifically, it has been shown that TdT expression is upregulated as the B-cell moves from the pre-pro-B stage, undergoing D-to-J recombination, and that expression peaks as the V-to-DJ rearrangement occurs in the late pro-B stage [101]. TdT expression then quickly declines as the cell progresses into the pre-B stage. This observation is consistent with our result, that there are more detachments (and hence more

attachments) before end-joining in the V-D junction relative to the D-J junction (Table 2).

We also investigated the mechanisms involved in gene segment recombination. Our findings regarding D-J segment correlations raise an interesting hypothesis that multiple successive D-J rearrangements may occur prior to recombination with a V segment. Previously, Reth et.al. tested the possibility of this hypothesis in murine 300-19 cells cultured in vitro by assaying for the presence of a designated D-J insert and found that such multiple successive recombinations can and do occur [74]. Other studies analyzing nonproductive human Ig rearrangements have hypothesized, based on their observations, that multiple successive D-J rearrangements at the human heavy locus are likely [90, 91]. We here provide evidence for this hypothesis for human Ig. This rearrangement mechanism differs from that observed in receptor editing in the heavy chain via V_H replacement [39] or at the light chain loci by secondary de novo rearrangements [23, 71]. Our analysis suggests that multiple D-J rearrangements may occur up to 15% of the time prior to the V-to-DJ rearrangement, with each successive D-J recombination replacing the previous one via excision.

The processes involved in D-J recombination are complex and likely require more parameters to better model the system. Still, the results of our modeling, with such extreme differences in p-value and chi-square values, are sufficient to support our hypothesis for the observed patterns in our P sequences. These data provide the first statistically supported observations of multiple successive recombinations in productive human Ig sequences. Considering V-D pairings, we did not perform a similar contingency table analysis since the greater number of possible pairs dramat-

icallyreduces the statistical power. For the NP sequences, the relatively low number of sequences in this set did not allow for this analysis.

These analyses prompted us to speculate about the observed J segment frequencies. Our multiple recombination model can help explain the lower usage frequency of segments J1 and J2, but prompts one to question why V5 is not used as frequently as V6, yet instead has a similar frequency to V3. Of the remaining four segments, J4 and J6 are used most frequently, followed by J5 and J3. It is possible that there are structural reasons for these observations concerning DNA access and histone acetylation. We propose, however, that the observed trends may instead be due to selection for tyrosine residues. Analyses of the $5'$ portion of the functional J segments, up to the invariant tryptophan residue, show that both J3 and J5 lack tyrosine residues, while J4 has two and J6 has five. Tyrosine has biochemical and structural properties that make it beneficial in protein binding interfaces, such as CDR [56]. Also, studies of amino acid profiles in human Ig have shown that tyrosine is one of the most abundant residues found in CDR, and specifically within CDR3, it locates most often at the C-terminus end of the CDR3 loop [56, 105]. Any residues contributed to CDR3 by J segments would be found at the C-terminus end of CDR3. The desirability of tyrosine residues and their frequent location at the $3'$ end of CDR3 suggests biased selection toward proteins comprised of J segments that contribute such residues, namely J4 and J6.

With regard to CDR3 length, we found that the P sequences had a statistically shorter mean compared with the NP sequences. The higher mean CDR3 observed in the NP sequences may be due to a lack of selection. It has been previously shown

that negative selection occurs in the bone marrow against B-cells presenting Ig with long CDR3 [84]. This may be because Ig with long CDR3 have been correlated with polyreactive specificity, including specificity for self peptide [15]. Since the NP sequences are not subject to selection in the bone marrow, these data provide evidence that negative selection restricts CDR3 length in the human Ig repertoire.

4.3 Conclusions

We applied a statistical approach to the study of the mechanisms involved in Ig gene formation by utilizing the wealth of publicly available data. Amassing sequence data from Genbank may be precarious. Yet, our observations of gene segment frequencies aligned well with previous reports, validating our approach and allowing us to provide novel statistical evidence for interesting mechanisms that shape the human Ig heavy chain repertoire.

We provide here the most precise estimates of human heavy chain gene usage frequency currently available. Additionally, we provide here the first statistical evidence in humans for sequential D to J recombination at the human heavy chain locus.

Chapter 5

Analysis of Autoreactive Ig Heavy Chains

Beliefs that Ig with long CDR3 are autoreactive have become common in immunological circles even though such assertions are the product of unsupported conclusions drawn from reasoning about experimental results. Experiments using oligonucleotide site-directed mutagenesis and CDR3 replacement show that heavy chain CDR3 provides the essential structural correlates necessary for polyreactive Ig [35, 52]. Further studies showed a correlation between Ig with long CDR3 and polyreactive specificity, including specificity for self peptide [15]. Additionally, Shiokawa, *et.al.* showed evidence for selection against Ig with longer CDR3, as the cells displaying such molecules are frequently deleted from the immature bone marrow B-cell population [84].

Since Ig molecules with long CDR3 have been associated with polyreactivity, and cells containing Ig with long CDR3 are negatively selected against in the bone marrow, then Ig molecules with long CDR3 must be autoreactive and are negatively selected against. The aforementioned studies tempt one to arrive at this conclusion, but such a conclusion is not directly supported by the literature. Some literature, in

fact, indicates that autoantibodies are commonly present in healthy human serum [45] and that in murine models, positive selection for some autoreactive antibodies has been observed [32]. Still, the belief that Ig with long CDR3 are autoreactive has become popular in immunology partly since a structural argument can be made. When an Ig protein folds, CDR3 becomes a loop, which is conformationally diverse. Long CDR3 make for larger loops that may be more likely to accept different conformations, thereby increasing the likelihood for polyreactivity and autoreactivity. It is possible, however, and not refuted by the literature, that Ig with long CDR3 are not autoreactive, or that autoreactive Ig have a short or average CDR3 length.

To further examine the role of CDR3 in autoreactive specificity and to explore possible genetic biases between autoreactive and non-autoreactive Ig genes, we performed a comprehensive and comparative study of over 7,300 Ig. We assembled four sets of human heavy chain genes for comparison: a set of productive, non-autoreactive genes; a heterogeneous set of autoreactive genes; a set of autoreactive genes from a specific autoimmune disease, rheumatoid arthritis; and a set of non-productively rearranged genes. We performed a detailed analysis of each gene set in terms of gene segment composition, CDR3 length, n-nucleotide addition, and mutation frequency, and employed statistical methods to detect biases that may exist between these sets. We find differential biases in gene segment usage and n-nucleotide tract length, but not in CDR3 length between autoreactive and non-autoreactive productive genes. We did, however, find an increase in the proportion of n-nucleotides in CDR3 from autoreactive genes, suggesting that germline D and J segments are selected for lack of autoreactivity. Our data further suggest that diversification via n-nucleotide addi-

tion comes at a cost: as the n-nucleotide proportion increases, so does the probability of autoreactivity.

5.1 Results

5.1.1 Complementarity Determining Region 3

We observed significant differences for the mean n-nucleotide tract lengths in both the V-D and D-J junctions between the four gene sets (Fig. 5.1). In particular, the A, RA, and NP sequences differed significantly from the P sequences, with relative increases of 14%, 9%, and 24%, respectively, in the V-D junction and relative increases of 26%, 21%, and 17%, respectively, in the D-J junction. In neither junction did we observe differences between the A, RA, and NP sequences sets.

For CDR3 length, we observed means of 15.49aa, 15.58aa, and 15.35aa for the P, A, and RA genes, respectively (Fig. 5.1). The slight differences among these means are not statistically significant. We did, however, observe a statistically significant enhancement of CDR3 length in the NP genes compared to each of the productive gene sets. Plotting the cumulative distribution functions for the CDR3 data for each gene set shows that the CDR3 lengths of the P, A, and RA genes are distributed similarly, while the distribution of lengths for NP gene CDR3 lengths differs (Fig. 5.2).

We further computed the ratio of n-nucleotides to germline-encoded (D segment) nucleotides in each of the three productive gene sets. Figure 5.3 provides the mean ratios and shows that there is a statistically significant difference in the mean ratio between the A and RA genes relative to the P genes, but no such difference between

V-D n-nucleotides

	P	A	RA	NP	
mean	7.86	8.99	8.57	9.78	
Std. err. (±)	0.01	0.22	0.06	0.25	
p-values		0.01	0.01	$< 10^{-7}$	P
			0.40	0.26	A
				0.06	RA

D-J n-nucleotides

	P	A	RA	NP	
mean	7.04	8.87	7.90	8.26	
Std. err. (±)	0.01	0.21	0.08	0.19	
p-values		$< 10^{-5}$	$< 10^{-3}$	$< 10^{-3}$	P
			0.40	0.26	A
				0.06	RA

CDR3 length

	P	A	RA	NP	
mean (aa)	15.49	15.58	15.35	18.00	
Std. err. (±)	0.02	0.64	0.17	0.79	
p-values		0.72	0.18	$< 10^{-11}$	P
			0.24	$< 10^{-9}$	A
				$< 10^{-11}$	RA

V gene segment usage

	A	RA	NP	
p-values	0.003	$< 10^{-20}$	$< 10^{-20}$	P
		$< 10^{-20}$	$< 10^{-20}$	A
			$< 10^{-20}$	RA

D gene segment usage

	A	RA	NP	
p-values	0.91	0.08	$< 10^{-4}$	P
		0.55	0.03	A
			$< 10^{-4}$	RA

J gene segment usage

	A	RA	NP	
p-values	0.01	0.27	0.04	P
		0.02	0.01	A
			0.03	RA

Figure 5.1: The left side shows mean n-nucleotide tract lengths for the V-D and D-J junctions and mean CDR3 lengths observed in the four sets of genes, along with p-values determing the statistical significance for the differences between each combination of genes. The right side shows tables of p-values for each combination of gene sets indicating the level of statistically significant difference between each pair for V, D, and J gene segment usage.

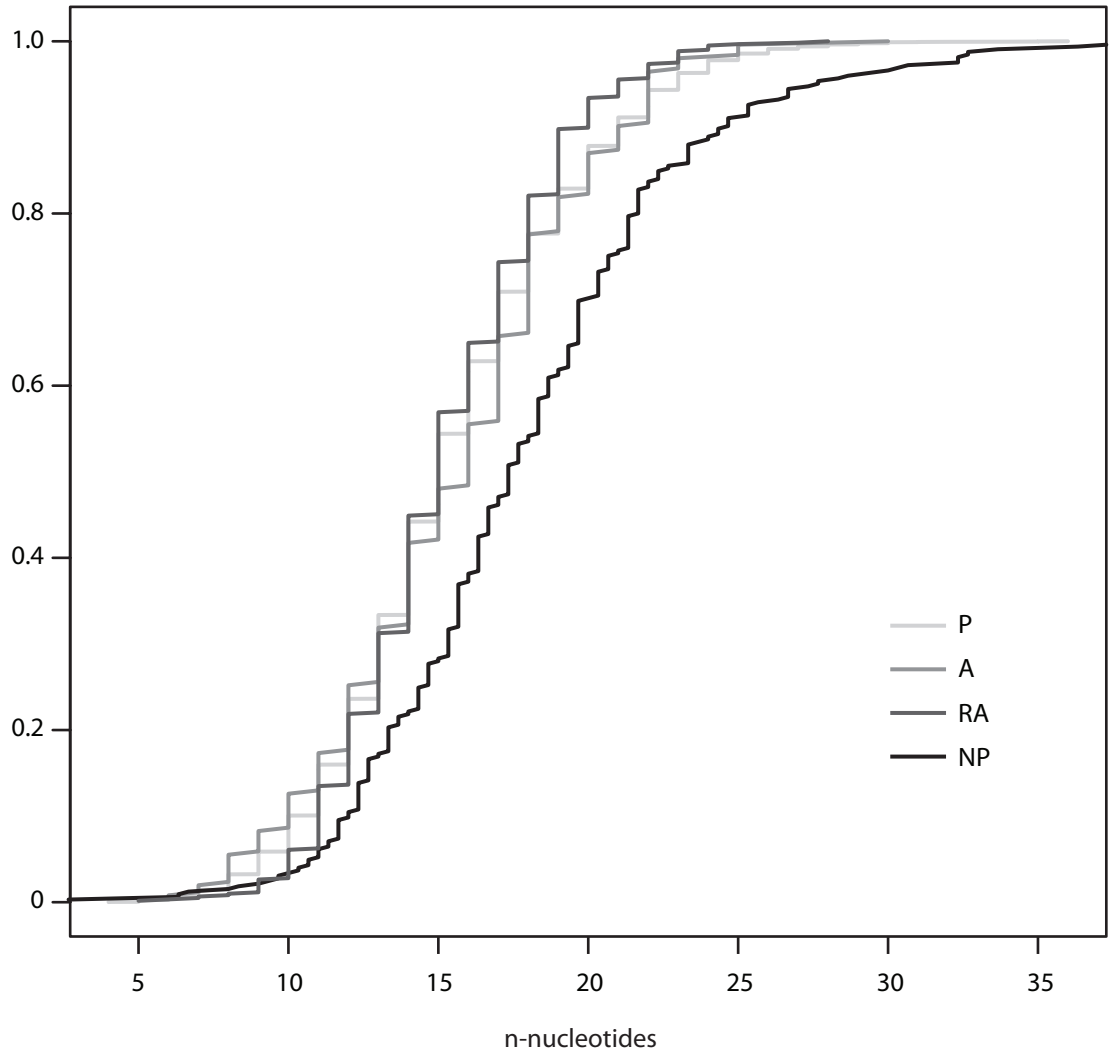

Figure 5.2: Cumulative distribution functions for the number of n-nucleotides observed in CDR3 of the four gene sets.

the A and RA genes themselves.

5.1.2 Somatic Mutations

We observed significantly higher mutation frequencies of 4.7% and 4.9% for the A and RA genes, respectively, compared to the P genes ($p - values < 10^{-10}$). These values represent a 46% and 47% relative increase over the observed mutation frequency

Mutation frequency

	P	A	RA	
Mut. freq	0.032	0.047	0.049	
p-values		$< 10^{-20}$	$< 10^{-20}$	P
			0.04	A

Synonymous vs nonsynonymous

	P	A	RA	
S/NS	0.47	0.50	0.47	
p-values		0.22	0.75	P
			0.35	A

n/D nucleotide ratio

	P	A	RA	
mean	0.54	0.66	0.63	
Std. err. (±)	$<10^{-4}$	0.006	$<10^{-4}$	
p-values		$<10^{-4}$	$<10^{-4}$	P
			0.60	A

Figure 5.3: Tables showing the p-values indicating the statistally significant differences between comparisons of each pair of functional gene sets in terms of overall mutations, ratio of synonymous to non-synonymous mutations, and n-nucleotide to D-segment nucleotide ratio.

for the P genes of 3.2%. Despite higher mutation frequencies, we did not observe a significant difference in the number non-synonymous mutations observed in the three gene sets (Fig. 5.3).

5.1.3 Gene Segment Usage

For the J segments, both the A and NP gene sets showed significant differences compared to every other gene set, as did the RA and P gene sets, except when compared to each other (Figs. 5.1, 5.4). Out of all possible combinations between the four gene sets, only the RA genes compared with the P genes showed no statistically significant difference in J segment usage.

The V and D segments outnumber J segments by factors of eight and four, respectively and are divided into seven families each based on sequence homology. Looking at the frequency of V segment usage by family, we observe statistically significant

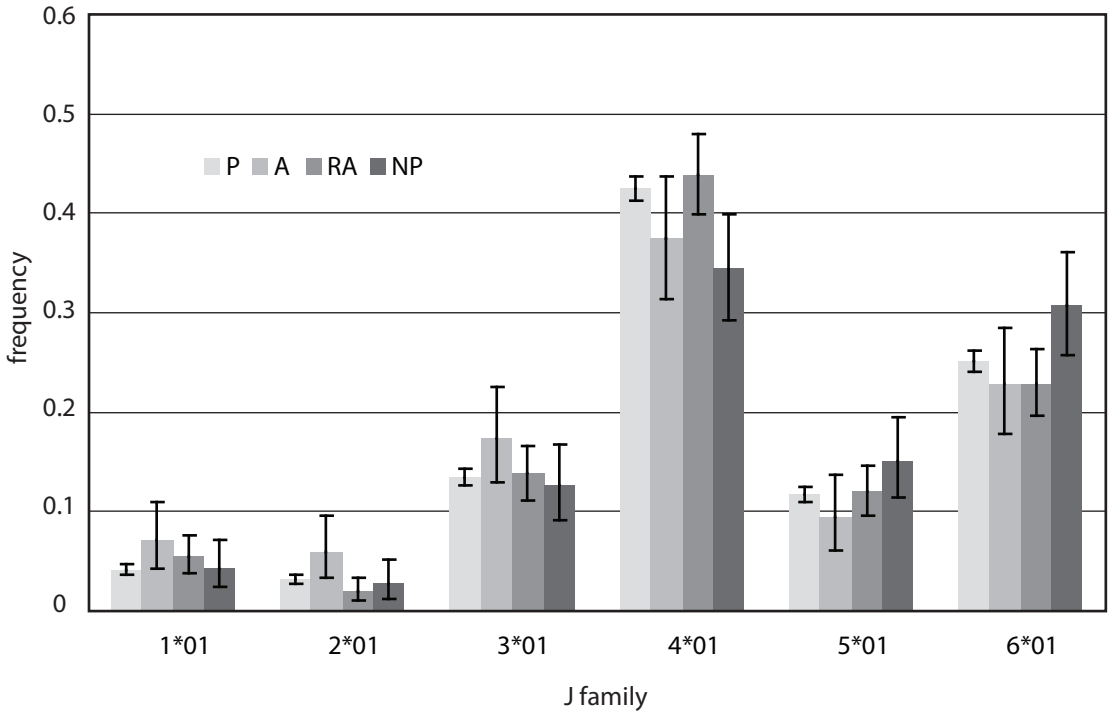

Figure 5.4: Relative frequencies of J segment usage for the P, A, AR, and NP genes.

differences for each gene set compared to every other gene set (Fig. 5.1). For both the A and the RA genes, the most notable observation was the relative increase of 50% and 130%, respectively, of segments from family VH1 relative to the P genes, but a 26% and 67% decrease, respectively, in the relative use of segments from family VH4 (Fig. 5.5). The RA genes also showed dramatic enhancements in usage of segments from family VH5 by 102% relative to the P genes (Fig. 5.5). For the NP genes, we observed the opposite trend in family VH4, with 67%, 115%, and 34% relative increases of segment usage relative to the P, A, and RA genes, respectively.

Among the productive gene sets, we did not observe statistically significant differences in D segment usage by family (Fig. 5.1). The NP genes, however, did differ significantly from each of the three productive gene sets. We also observed significant

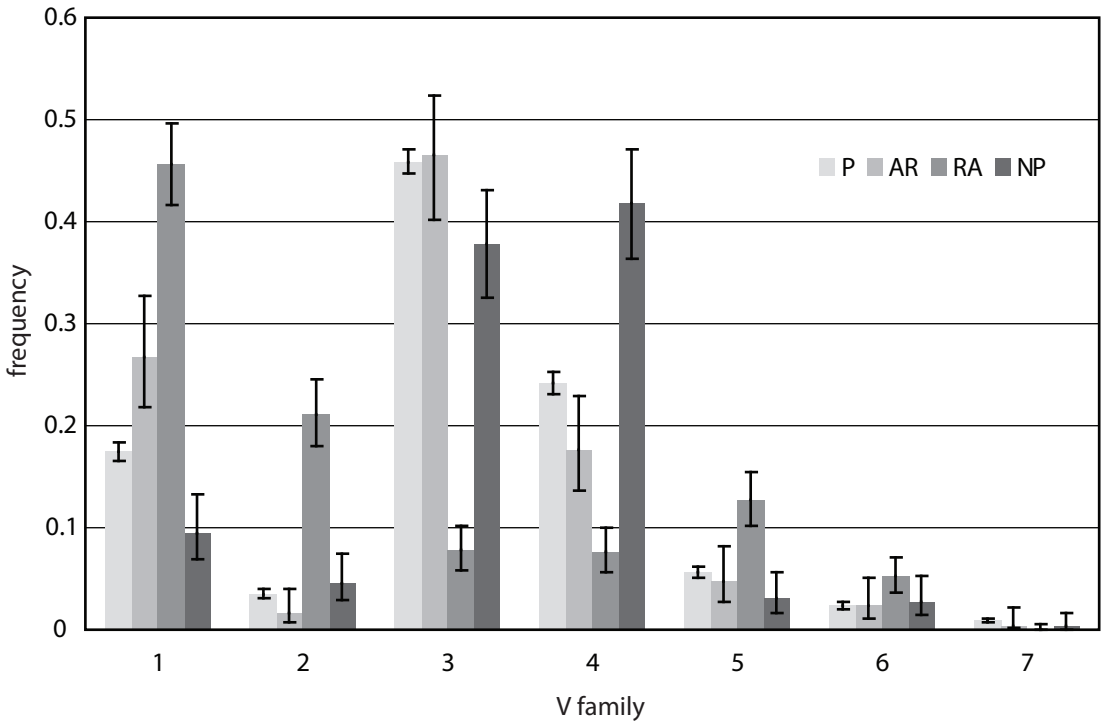

Figure 5.5: Relative frequencies of V segment usage for the P, A, AR, and NP genes.

enhancements of 33% and 136% in the number of inverted D-segments used in the A and RA genes, respectively, relative to the P genes ($p < 0.02$).

5.2 Discussion

We set out to investigate subtle genetic biases that exist between non-autoreactive and autoreactive genes by looking at gene segment usage, CDR3 n-nucleotide composition, and CDR3 length. Our data show that for CDR3 length, no such bias exists and therefore, notions regarding a long CDR3 as evidence for autoreactivity of Ig genes [15, 40, 1] are not substantiated. Our data are consistent with previous studies showing negative selection against Ig with longer CDR3, since the mean observed CDR3 length of the NP genes is significantly longer than that of the productive

genes. The means of the P, A, and RA genes, however, are nearly identical and have nearly overlapping cumulative distribution functions. Thus, CDR3 length alone is not a good predictor of autoreactivity.

We also did not observe statistically significant differences in D segment usage between the P, A, and RA genes. Evolution, via selection, presumably has shaped the human heavy D and J segment repertoire to eliminate segments that may confer or promote autoreactivity. Our data suggest that the D and J segment repertoire has indeed been selected for enhanced CDR3 diversity and against autoreactivity.

We did observe, however, significant differences in n-nucleotide tract lengths in both segment junctions of the A and RA genes, relative to the P genes, despite the lack of statistically significant differences in length. These data indicate that a relationship between the number of germline-encoded nucleotides and n-nucleotides within CDR3 plays a role in autoreactivity. Given a particular CDR3 length, the possibility of autoreactivity increases with an increase in the ratio of n-nucleotides to germline-encoded nucleotides. Due to recombination site choice, exonucleolytic activity during segment ligation, or even D-segment inversion, the A and RA genes contain fewer D and J segment nucleotides, and therefore contain more randomness compared to the P genes. This shift in nucleotide origin comes at a cost and may promote the development of autoreactivity. Studies with TdT deficient mice crossbred with autoimmune-prone mice provide evidence for our hypothesis. These crossbred mice, which make B- and T-cells with nearly no n-nucleotides, showed lower incidence of autoimmune disease and longer life spans compared to non-TdT deficient, autoimmune-prone controls [24, 77]. These studies underscore the importance of a

lower n-nucleotide to germline-encoded nucleotide ratio in autoimmunity.

These data also support published hypotheses that development of autoreactivity may occur in the periphery or secondary lymphoid organs. Studies have shown that autoreactive Ig can arise during somatic hypermutation [86, 85, 102, 19, 8, 73]. During this process, B-cells presenting Ig, which as a result of the mutations have lost affinity for the presented antigen, are deleted, while those that acquire greater affinity are positively selected. Occasionally, some of the surviving cells with affinity for self antigen are not selected against, due to molecular mimicry of the antigenic epitope to self-antigen (reviewed in [76]). Our data show that even though the proportion of non-synonymous to synonymous mutations is statistically equivalent, the mutation frequencies in the A and RA genes are significantly higher. These higher mutation frequencies within genes containing less germline-encoded and more randomly added nucleotides may simultaneously generate greater antigen specificity as well as autoreactivity.

Regarding the V gene segments, we observed an enhancement of VH1 family segment usage in the A genes that appears to be at the expense of segments from family VH4, since these segments showed attenuation relative to the same segments in the P genes. The RA genes reflect these same segment families biases relative to the P genes, but to an even greater degree. This is striking given that family VH4 is typically second highest in usage frequency and has been well reported to provide some of the most frequently used segments in adult human lymphocytes, particularly V4-34 [7, 6, 41]. The data here imply specific selection within the A genes against segments in family VH4, but for segments in family VH1.

Reviewing the methods and primer sequences involved in the creation of the RA dataset shows no bias in the way that samples were collected [57, 22]. Thus, the dataset seems representative of Ig genes from the synovial fluid of RA patients. We can therefore assume that the effects we observe are not artifactual.

The enhancement of segments from family VH1 in the A and RA genes may be due to the framework region of these gene segments. Studies of autoimmune anti-erythrocyte human protein antibodies and non-autoimmune anti-staphylococcal antibodies show that framework region 1 may play a critical role in effective binding to antigen [47, 70]. The VH1 and VH5 family genes are unique in their incorporation of a KK protein motif in framework 1, which would make for a positively charged hotspot in the folded protein. Such pairings of positively charged residues do not exist in the other V segments. We did also observe an enhancement of VH5 segment usage in the RA genes. Thus, it may be possible that part of the framework region 1 of the VH1 family genes promotes binding to self-antigens.

5.3 Conclusions

We have shown subtle biases between autoreactive and non-autoreactive Ig genes at the genetic level. By compiling a broad panel of autoreactive genes, and then a more specific set of genes from a particular autoimmune disease, we have been able to detect such biases through comparison to a large sample of diverse non-autoreactive genes. We provide substantial statistic evidence that in autoreactive genes, VH1 family gene segments are enhanced while VH4 family gene segments are attenuated. We have also shown that CDR3 length in itself is not indicative of

autoreactivity. Instead, the composition of CDR3, viewed as a ratio of randomly inserted to germline-encoded nucleotides, affects the probability of autoreactivity.

Chapter 6

Mutational Biases in Ig Heavy Chains

Somatic hypermutation (SHM), described in chapter 1, is an important process for increasing the diversity of the Ig repertoire. B-cells bearing surface Ig reactive with microbial antigens are activated; some enter the germinal centers, where their Ig genes experience point mutations at a rate of about one nucleotide substitution per division. They are subsequently selected for enhanced affinity for the eliciting antigen.

Biological processes like SHM require at least two main components: a molecular mechanism and a substrate. Presumably, the efficiency of such a process relies on how co-evolved the mechanism is with its substrate. It has been shown that the mechanism centers on Activation-Induced Cytidine Deaminase (AID; [59, 58, 75, 69]), but involves many other components of the DNA synthesis and repair pathways [62, 17]. The substrate is the Ig gene which is directly but controllably mutated due in part to inherent DNA binding motif preferences of AID. Even before the molecular mechanism became known, the DNA motif for AID was determined statistically to

be RGYW [78].

Since the purpose of SHM is to increase the receptor affinity of the responding cells for the eliciting antigen, the strategy of adaptive immunity seems to be to match the random antigenic variability of microbes with the random generation of receptor specificities. There is a crucial difference between the two strategies, however. Whereas the microbes diversify to escape, the immune system diversifies to pursue. Pathogens benefit from any mutation that disrupt host recognition, but the host benefits only from the substantially smaller class of mutations that specifically improve antigen receptor affinity. Biases that influence mutation process are desirable and advantageous for the host in that they provide efficient evolution of Ig and produce less waste. In theory then, the affinity maturation process benefits when genomic evolution shapes the Ig locus to promote AID action, through inclusion of its mutator hotspots, in ways that enhance the evolvability of the gene at the cellular level. Thus, we are interested in evolution in two contexts: pre-selection genomic evolution as it shapes the human Ig locus to contain pre-selection biases, and the rapid microevolution under selection of the Ig gene at the cellular level during the SHM process.

We and others have shown that codons predicted to be susceptible to SHM were observed more frequently in the complementarity determining regions (CDR) than in the framework regions (FR) in both humans and mice [97, 38, 65]. Subsequent studies of non-productively rearranged human genes confirmed this prediction directly, finding an excess of mutations in CDR compared to FR [13]. Such observations suggest a pre-selection mutation-rate bias inherent within the germline gene segments. They

also support our hypothesis of efficient evolvability during affinity maturation since nonsynonymous mutations within the FR regions are not only likely to be unhelpful, but also deleterious. It is possible then, due to the nature of the SHM mechanism and its sensitivity to the local DNA sequence, that this process is finely tuned to promote mutation rate biases through differential enhancement of synonymous and nonsynonymous mutations between FR and CDR. Such enhancements were considered by Shapiro et al [83] who predicted that mutations would favor the third codon positions in FR and second codon positions in CDR of human Vκ genes. Still, some mutations introduced during SHM will inevitably render the target gene's expressed protein structurally unstable. Such molecules are removed by selection, but their production imposes a kind of opportunity cost on the system: it would be best to not produce them at all. Thus, SHM efficiency improves when the set of likely mutations is limited to those that render the post-selection repertoire as similar to the pre-selection repertoire while improving antigenic affinity, thereby reducing the probability of introducing structurally deleterious mutations.

Our purpose here is to explore this hypothesis by looking at how finely adapted the germline is for efficient evolvability under SHM through analyses of synonymous and nonsynonymous mutations in the CDR and FR of Ig. To do so, we established a model using the large set of Ohm-Laursen genes, which contain cumulatively over 5,000 productive (OL-P) and non-productive (OL-NP) Ig genes that all use the same VH gene (VH3-23) (see section 3.1). We then applied our model to the productive (GB-P) and non-productive (GB-P) genes Ig gathered from Genbank and used for the analyses discussed in chapters four and five. We find and present here observations

showing that the germline is exquisitely tuned to promote efficient evolution of Ig under SHM such that the post-selection repertoire closely resembles the pre-selection repertoire with respect to mutational biases. We also find no evidence for what has been referred to as positive selection on the basis of relative NS mutation rates in the CDR.

6.1 Mutation Analysis

The Genbank-derived sequences used for this analysis, though gathered in the same way as described in chapter three, were slightly less filtered than the sets used for the analyses in chapters four and five. Specifically, the need to check Genbank records and remove sequences containing a keyword that may indicate autoreactivity was not necessary for this analysis. Consequently, the GB-P and GB-NP datasets used here are slightly larger, containing 7176 and 339 genes, respectively.

All sequences were analyzed to identify mutations from the inferred original rearrangement using our own software. Only positions in VH $5'$ of and including the $3'$ invariant cysteine were examined. Each mutation was classified as either synonymous (S) or nonsynonymous (NS). We computed the pre-selection proportion of synonymous to nonsynonymous mutations expected in the absence of mutator sequence specificity for the CDR and FR from the codon usage in the gene segments assuming that one-half of all mutations are transitions [50]. Our results are not, however, sensitive to this parameter value.

To further our analysis and test our hypothesis, we defined a conditional excess mutation rate, $\varepsilon_{ij}(X)$, as the amount of observed mutation over the expected value,

such that:

$$\varepsilon_{ij}(X) = \log_2(\mu_{ij}(X)) \tag{6.1}$$

and

$$\mu_{ij}(X) = \frac{\left(\frac{m_{ij}}{n_{ij}}\right)}{\left(\frac{m_{\bullet j}}{n_{\bullet j}}\right)}$$

where i is the region (FR/CDR), j is the type of mutation (S/NS), m is the observed number of mutations, n is the computed number of potential mutations, and X is the dataset, either P or NP. This gives four possible conditional excess mutation rate computations per data set: FR/S, FR/NS, CDR/S, and CDR/NS.

6.2 Results

6.2.1 Establishing the Model

We started with the gene sequences from the Ohm-Laursen (OL) data set, as these sequences all use the same VH gene segment (see 3.1). Analyzing a set of genes that all use the same VH segment will help to establish a model that may apply to the entire VH gene repertoire, represented in the Genbank dataset. Considering both the P and NP sequences, CDR positions account for about 16.7% of the total nucleotides analyzed, with an average of 16.0 codons assigned to CDR and 79.8 codons assigned to FR (Table 6.1). The overall mutation frequency observed in the OL-P genes of 3.26% is approximately 28% greater than the observed mutation frequency of 2.54% in the OL-NP genes. The relative mutation rates in the FR and CDR of the OL-P genes were 2.45% and 7.30%, respectively.

		Sequences	FR nucs	CDR nucs
OL	P	5403	1294954	259699
	NP	434	102313	20381
	Total	5837	1397267	280080
GB	P	7176	1523848	320283
	NP	339	69867	17622
	Total	7515	1593715	337905

Table 6.1: Number of sequences per dataset and the absolute number of nucleotides analyzed in the FR and CDR of those sequence sets.

In the absence of sequence specificity, we computed the expected proportion of S mutations in the FR and CDR to be 26% and 25%, respectively. However, we observed statistically significant increases in the proportion of S mutations over these expectations in the FR of 48% and 8% for the OL-P and OL-NP, respectively ($p - values < 0.02$). For the CDR, we observed decreases of 3.50% and 7.67% from the expected proportion of S mutations in the OL-P and OL-NP sequences, respectively, though only the decrease in the OL-P sequences was statistically significant ($p < 0.01$). We then computed the conditional excess mutation rates for both the OL-P and OL-NP sets, and plotted them together as shown in figure 6.1A. The plot shows a plausibly linear relationship between the OL-P and OL-NP sequences regarding observed conditional excess S and NS mutations rates.

6.2.2 Applying the Model

Having established a model by which we can test our hypothesis, we applied our model to the over 7,000 sequences collected from Genbank (GB). Again, we analyzed both productively (GB-P) and nonproductively (GB-NP) rearranged sequences separately. Considering both the GB-P and GB-NP sequences, CDR positions account

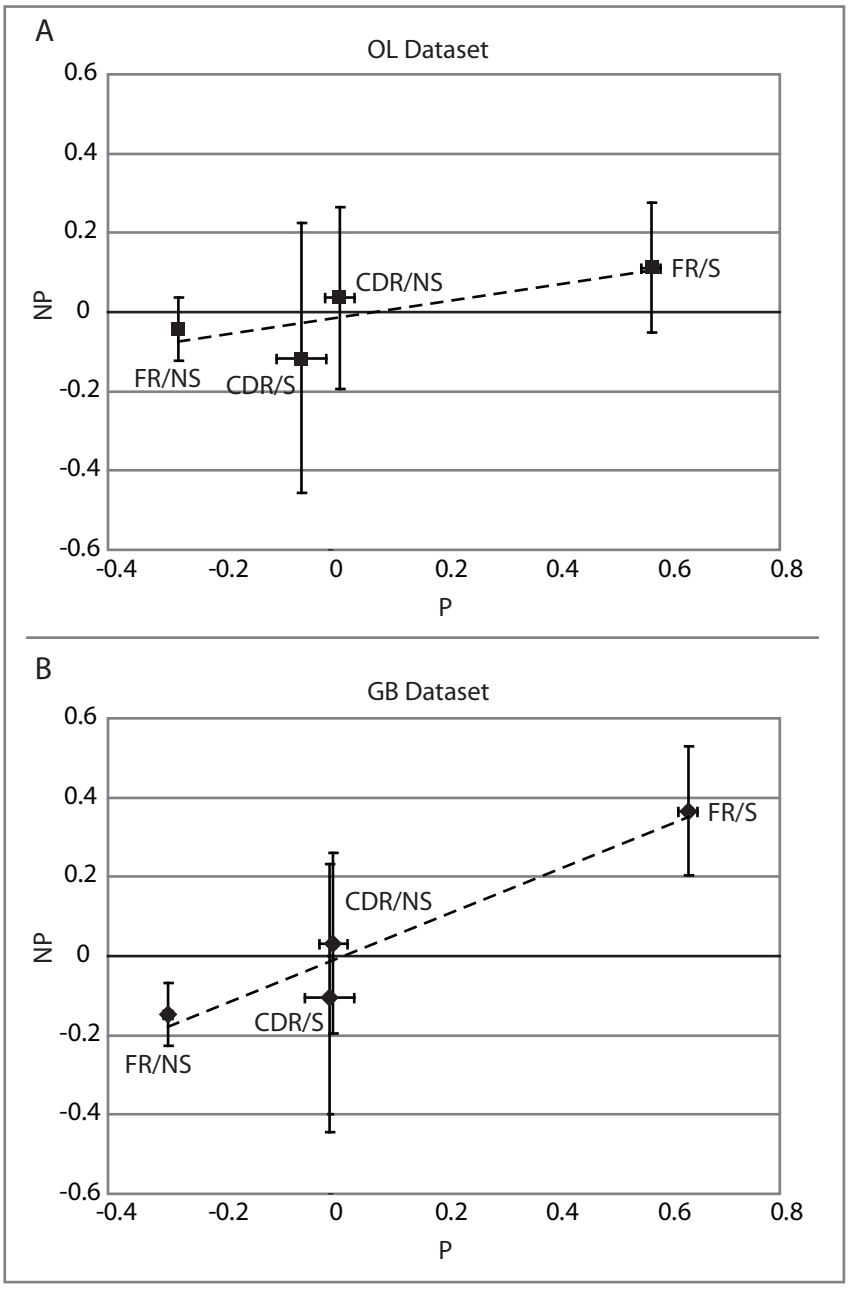

Figure 6.1: Plots of the conditional excess mutation rates for the OL (A) and GB (B) datasets. The conditional excess mutation rates gives a measure of how much over or under the observed relative mutation rate is comared to what is expected.

for about 17.5% of the total nucleotides analyzed, with an average of 15.0 codons assigned to CDR and 70.7 assigned to Furs (Table 6.1). The overall mutation fre-

		FR	CDR
OL-P	S	12258	4575
	NS	19505	14390
OL-NP	S	606	224
	NS	1542	749
GB-P	S	19888	5755
	NS	31397	18116
GB-NP	S	535	148
	NS	1120	507

Table 6.2: Number of sequences per dataset and the absolute number of nucleotides analyzed in the FR and CDR of those sequence sets.

quency in the GB-P genes of 4.08% is approximately 54.2% greater than the observed mutation frequency of 1.58% in the GB-NP genes. The relative mutation rates for the Furs and CDR in the GB-P sequences were 3.36% and 7.45%, respectively.

We computed the expected proportion of S mutations in the FR and CDR to be 25% and 24%, respectively, in the absence of sequence specificity. We observed, however, very statistically significant increases in the proportion of S mutations within the FR of 55.3% and 28.8% for the GB-P and GB-NP sequences, respectively ($p - values < 10^{-10}$). Observed decreases of S mutations in the CDR of both the GB-P and GB-NP sequences were not statistically significant. We again computed the conditional excess mutation rates for both gene sets, plotted them together as shown in figure 6.1B, and observed a linear relationship between the GB-P and GB-NP sequences.

6.2.3 Mutator Target Motifs Are Positionally Biased

We tested the human VH germline repertoire for inclusion and positioning of the known AID mutator motif, RGYW. AID mutates the second nucleotide, G, in the

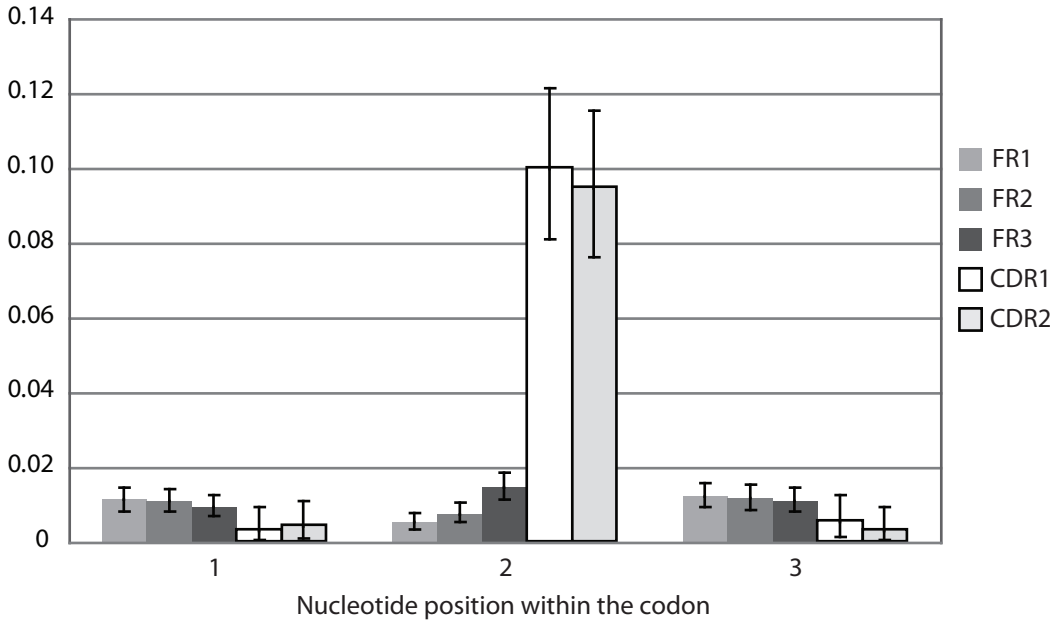

Figure 6.2: This chart shows the relative frequency of the G nucleotide within the AID hotspot motif (RGYW) occurring at the first, second, and third positions of codons within the FR and CDR of the human VH gene segments.

motifs it finds in DNA. Thus, we scanned each gene segment, recording counts of the nucleotide position within a codon where the G nucleotide is in the identified motifs (Fig. 6.2). Of the motifs occurring in the FR, there is no statistically significant difference in placement of the G nucleotide within the FR codons. Within the CDR, however, there is a statistically significant enhancement of the G nucleotide within identified motifs occurring at the second nucleotide position in the CDR codons ($p < 10^{-10}$). Relative to the FR regions, the observed frequency of the motifs presence in the CDR represents a statistically significant enhancement of 84% ($p < 10^{-10}$).

6.3 Discussion

We set out to test the hypothesis that the SHM process has evolved to encourage such efficient mutation that the post-selection repertoire resembles the pre-selection repertoire with respect to mutational biases. We analyzed the genetic composition of both pre- and post-selection Ig genes, and made comparisons between S and NS regional mutation rates within these genes. The use of the OL sequence sets, in which all genes used the same VH gene, was essential in establishing our model. The observations made in the OL genes with VH gene should be applicable to all VH gene segments if our hypothesis holds. Thus, we then applied our model to the more diverse set of GB genes and found that the patterns observed in the OL dataset were even stronger, providing evidence in support of our hypothesis.

Based on just the RGYW motif alone, we show that mutations are much more likely in the CDR, which supports previous studies showing greater mutation rates in the CDR relative to the FR [13, 82]. Under our hypothesis, we expect to observe such trends not only in the P genes, but particularly in the NP genes as well. Our observations of differential mutations rates in the NP genes are evidence for pre-selection bias in the germline. Full support for our hypothesis, however, would be observations of a linear relationship between conditional excess S and NS mutation rates of the P and NP sequences, such that $\varepsilon_{ij}(P) = c\varepsilon_{ij}(NP)$. Such a relationship implies that the pre-selection biases are exploited in the post-selection repertoire such that mutations in the post-selection repertoire are from the set established by the pre-selection biases. Indeed, such a linear relationship, within the bounds of the confidence intervals, was observed for both the GB and OL datasets (Fig. 6.1).

Under the null hypothesis, we expect that the relative rates of S mutations in the absence of sequence selection would be about 25-26% in both the CDR and FR of both the OL and GB sequences. Much work, however, has shown that sequence specificity for AID does exist and is important for SHM [38, 83, 82]. Thus it was not surprising that our data support these studies and our hypothesis. What was striking, however, was the extent of the positioning of the RGYW hotspot motifs in the germline VH gene segments which provide much of the pre-selection NS mutation rate bias in the NP sequences. Previous authors have predicted the second nucleotide position of CDR codons in human Vκ to be highly mutable and we provide here direct evidence for this prediction in human VH gene segments through the observations of NS mutations in the CDR and observed enhancements of S mutations in the FR over what is expected.

We have also observed a lack of evidence for what has been referred to as "positive selection", indicated by an excess of NS mutations in the CDR. In fact, we did not find a statistically significant difference in the observed proportions of S and NS mutations in the CDR of the GB genes. The diagnosis of antigenic selection in Ig has been an area of great interest in the past; several statistical methods have been developed for the purpose [50]. These techniques are based on comparisons of mutation frequencies among FR and CDR, and/or S and NS positions and make assumptions about the equality of mutation rates among these sites under the null hypothesis. One of the implications of our findings here is that these methods are inherently unreliable. The null hypotheses do not hold even in the absence of selection.

Finally, we address the possibility that the results obtained here result from the

contamination of the NP sequence set. There is concern regarding the possibility of contamination of the NP sequences with sequences that were originally rearranged in frame, experienced mutation with selection, and suffered insertion or deletion mutations in CDR3 that rendered them out-of-frame afterward, thus giving them the appearance of NP sequences, but having experienced selection nevertheless. We regard this scenario as unlikely, for the following reasons. First, insertions and deletions are relatively infrequent events, occurring at a rate of about 1 per 100 point mutations [88]. Second, we have filtered out all sequences that are inferred to have had insertions or deletions in VH, DH, or JH in CDR3. We cannot rule out the possibility that any given sequence has suffered a frame-shift mutation in an n-nucleotide tract, since that event would be undetectable. We do, however, argue that these events should be sufficiently rare that they are unlikely to be the cause of the effects we observe. Furthermore, the NP genes have a lower average mutation frequency than the P genes in both datasets.

6.4 Conclusion

Taken together, our observations support our hypothesis that the germline has evolved to promote efficient evolvability under SHM such that the post-selection repertoire resembles the pre-selection repertoire. We provide evidence to suggest that evolution has shaped the heavy chain VH gene segment repertoire such that mutations incurred during SHM are of the smaller set of possible mutations that promote efficient evolvability and thus their general impact is positive and not deleterious. We also have shown a lack of evidence for positive selection, which is characterized

by an excess of NS in the CDR. Our data show that the rates of S and NS mutations in the CDR are statistically equivalent, even in the NP genes, and therefore are not indicative of selection during affinity maturation.

Chapter 7

Conclusions

In conducting the research reported here, we used a large collection of genes harvested from Genbank. One could argue the precariousnous of gathering data in this way for analysis. We, however, would counter that the Genbank repository exists, in part, so that DNA sequence data deposited there can be reused in research initiatives. We performed very careful, exacting, and thorough filtering of the Genbank genes that ended up in our large but useful datasets. Compiling and filtering such a large set of genes enabled us to perform novel statistical analyses that would otherwise have been impossible.

Our analysis of non-autoreactive, productively rearranged genes provides the most precise estimates currently available for heavy chain gene segment usage frequency in non-neonatal humans. This baseline information may be informative for researchers looking at specialized Ig molecules, such as for HIV vaccines. More importantly, however, our analyses led us to uncover and report on influential biases in the mechanisms that are responsible for human Ig repertoire development. For

example, we are the first to show statistical evidence for sequential D-to-J recombinations at the heavy chain locus during B-cell ontogeny. We have also shown that TdT action is biased between the V-D and D-J junctions in the heavy chain such that more n-nucleotides are likely in the V-D junction than in the D-J junction.

We have also been able to show that genetic biases exist in the germline that influence repertoire development. Our analysis of P genes from Genbank leads us to speculate that certain JH segments, particularly JH3 and JH5, may be less preferable in Ig gene rearrangements due to their inherent lack of code for Tyrosine residues. We also provided strong statistical evidence to suggest that the ratio of non-templated to germline-encoded nucleotides within the CDR3 of a given Ig gene has an impact on the molecule's potential for autoreactivity, which implies that the germline gene segments, particularly the D gene segments, are biased against autoreactive potential. We also showed through mutation analysis that the germline is biased towards encouraging certain types of mutations in certain regions of the gene so as to provide the most efficient evolutionary path for a genes under affinity maturation.

The processes and mechanisms that go into Ig formation are often talked about as being random, though it is understood that they are not truly random. However, the extent of bias for some of these processes and mechanisms had never been studied using dataset large enough to reveal statistically significant evidence. The contribution presented here uncovers and quantifies some of the biases that exist in the genetics and mechanisms that are necessary for Ig heavy chain repertoire development. Our analyses also provide data that can now be used to update SoDA and other such programs with gene segment frequencies so that they yield better, more

informed results. These data also provide a foundation and guidance on which analyses of broadly neutralizing anti-HIV antibodies should be based. Analysis of these molecules was complicated before by the lack of statistics regarding the genetics of Ig heavy chain genes. The work here now provides the necessary background data to make informed analyses and hypotheses about the origins and unique characteristics of anti-HIV antibodies.

7.1 Future Directions

We have provided statistical support for biases in the mechanisms and genetics of immunoglobulin development. Showing experimental support for some of our observations would be further validation of this work. Our analysis showing statistical evidence for multiple sequential rearrangements of D to J gene segments could be further supported by lab studies. Such studies might involve isolating excision circles from developing B-cells, and using specially designed primers to specifically locate and amplify D-J pairs in those circles. For our analysis of autoreactive Ig, experiments using normal and autoreactive mice could be performed. Populations of Ig could be collected from each set of mice, and testing for binding affinity against a panel of antigens could be performed. Then, sequencing of those Ig to determine CDR3 length and n/D nucleotide ratios would help validate our results. Our analysis of mutational biases is a purely statistical study based on observations of synonymous and nonsynonymous mutations in the genes and does not lend itself to further laboratory exploration.

Bibliography

[1] I. Aguilera, J. Melero, A. Nu nez Roldan, and B. Sanchez. Molecular structure of eight human autoreactive monoclonal antibodies. *Immunology*, 102(3):273–280, 2001.

[2] F.W. Alt and D. Baltimore. Joining of immunoglobulin heavy chain gene segments: implications from a chromosome with evidence of three D-J_H fusions. *Proc Natl Acad Sci*, 79(13):4118–4122, 1982.

[3] S.F. Altschul, W. Gish, W. Miller, E.W Meyers, and D.J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, 1990.

[4] N.M. Bhat, L.M. Lee, R.F. van Vollenhoven, N.N Teng, and M.M Bieber. VH4-34 encoded antibody in systemic lupus erythematosus: effect of isotype. *J Rheumatol*, 29(10):2114–2121, 2002.

[5] F.J. Bollum. *Terminal deoxynucleotidyl transferase. In Boyer, P. (ed.), The Enzymes*. New York, New York, 10 edition, 1974.

[6] H. Brezinschek, S.J. Foster, R.I. Brezinschek, T. Dörner, R. Domiati-Saad, and P.E. Lipsky. Differential effects of selection and somatic hypermutation on human peripheral $CD5^+IgM^+$ and $CD5^-IgM^+$ B cells. *J Clin Invest*, 99(10):2488–2501, 1997.

[7] H.P. Brezinschek, R.I. Brezinschek, and P.E. Lipsky. Analysis of the heavy chain repertoire of human peripheral B cells using single-cell polymerase chain reaction. *J Immunol*, 155(1):190–202, 1995.

[8] L.P. Casson and T. Manser. Random mutagenesis of two complementarity determining region amino acids yields an unexpectedly high frequency of antibodies with increased affinity for both cognate antigen and autoantigen. *J Exp Med*, 182(3):743–750, 1995.

[9] 3rd C.F. Barbas, E. Björling, F. Chiodi, N. Dunlop, D. Cababa, T.M. Jones, S.L. Zebedee, M.A. Persson, P.L Nara, and E. Norrby *et.al.* Recombinant human Fab fragments neutralize human type 1 immunodeficiency virus in vitro. *Proc Natl Acad Sci*, 89(19):9339–9343, 1992.

[10] G.P. Cook and I.M. Tomlinson. The human immunoglobulin V repertoire. *Immunol Today*, 16(5):237–242, 1995.

[11] G.P. Cook, I.M. Tomlinson, G. Walter, H. Riethman, N.P. Carter, L. Buluwela, G. Winter, and T.H Rabbitts. A map of the human immunoglobulin V locus completed by analysis of the telomeric region of chromosome 14q. *Nat Genet*, 7(2):162–168, 1994.

[12] S.J. Corbett, I.M. Tomlinson, E.L. Sonnhammer, D. Buck, and G. Winter. Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, "minor" D segments or D-D recombination. *J Mol Biol*, 270(4):587–597, 1997.

[13] L.G. Cowell, H.J. Kim, T. Humaljoki, C. Berek, and T.B. Kepler. Enhanced evolvability in immunoglobulin V genes under somatic hypermutation. *J Mol Evol*, 49(1):23–26, 1999.

[14] J.P. Cox, I.M. Tomlinson, and G. Winter. A directory of human germ-line V kappa segments reveals a strong bias in their usage. *Eur J Immunol*, 24(4):827–836, 1994.

[15] R. Crouzier, T. Martin, and J.L. Pasquali. Heavy chain variable region, light chain variable region, and heavy chain CDR3 influences on the mono- and polyreactivity and on the affinity of human monoclonal rheumatoid factors. *J Immunol*, 154(9):4526–4535, 1995.

[16] M. Delarue, J.B. Boulé, J. Lescar, N. Expert-Bezançon, N. Jourdan, N. Sukumar, F. Rougeon, and C. Papanicolaou. Crystal structures of a template-independent DNA polymerase: murine terminal deoxynucleotidyl transferase. *EMBO J*, 21(3):427–439, 2002.

[17] F. Delbos, S. Aoufouchi, A. Faili, J.C. Weill, and C.A Reynaud. DNA polymerase eta is the sole contributor of A/T modifications during immunoglobulin gene hypermutation in the mouse. *J Exp Med*, 204(1):17–23, 2007.

[18] S.V. Desiderio, G.D. Yancopoulos, M. Paskind, E. Thomas, M.A. Boss, N. Landau, F.W. Alt, and D. Baltimore. Insertion of N regions into heavy-chain genes is correlated with expression of terminal deoxytransferase in B-cells. *Nature*, 311(5988):752–755, 1984.

[19] B. Diamond, J.B. Katz, E. Paul, C. Aranow, D. Lustgarten, and M.D. Scharff. The role of somatic mutation in the pathogenic anti-DNA response. *Eur J Immunol*, 10:731–757, 1992.

[20] P. Early, H. Huang, M. Davis, K. Calame, and L. Hood. An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: V_H, D and J_H. *Cell*, 19(4):981–992, 1980.

[21] B. S. Everitt. *The Analysis of Contingency Tables*. Chapman and Hall/CRC, Boca Raton, 1977.

[22] F. Fais, F. Ghiotto, S. Hashimoto, B. Sellars, A. Valetto, S.L. Allen, P. Schulman, V.P. Vinciguerra, K. Rai, L.Z. Rassenti, T.J. Kipps, G. Dighiero, H.W. Schroeder Jr, M. Ferrarini, and N. Chiorazzi. Chronic lymphocytic leukemia B-cells express restricted sets of mutated and unmutated antigen receptors. *J Clin Invest*, 102(8):1515–1525, 1998.

[23] R.M. Feddersen and B.G. Van Ness. Double recombination of a single immunoglobulin kappa-chain allele: implications for the mechanism of rearrangement. *Proc Natl Acad Sci*, 82(14):4793–4797, 1985.

[24] A.J. Feeney, B.R. Lawson, D.H Kono, and A.N. Theofilopoulos. Terminal deoxynucleotidyl transferase deficiency decreases autoimmune disease in MRL-Fas(lpr) mice. *J Immunol*, 167(6):3486–3493, 2001.

[25] J.P. Frippiat, S.C. Williams, I.M. Tomlinson, G.P. Cook, D. Cherif, D. Le Paslier, J.E. Collins, I. Dunham, G. Winter, and M.P Lefranc. Organization of the human immunoglobulin lambda light-chain locus on chromosome 22q11.2. *Hum Mol Genet*, 4(6):983–991, 1995.

[26] D. Gay, T. Saunders, S. Camper, and M. Weigert. Receptor editing: an approach by autoreactive B cells to escape tolerance. *J Exp Med*, 177(4):999–1008, 1993.

[27] M.S. Gelfand, A.A. Mironov, and P.A. Pevzner. Gene recognition via spliced sequence alignment. *Proc Natl Acad Sci*, 93(17):9061–9066, 1996.

[28] V. Giudicelli, D. Chaume, and M. Lefranc. IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. *Nucleic Acids Res*, 32(Web server issue):W435–W440, 2004.

[29] C. Goodnow, J. Crosbie, S. Adelstein, T.B. Lavoie, S.J. Smith-Gill, R.A. Brink, H. Pritchard-Briscoe, J.S. Wotherspoon, R.H. Loblay, K. Raphael, R.J. Trent, and A. Basten. Altered immunoglobulin expression and functional silencing

of self-reactive B lymphocytes in transgenic mice. *Nature*, 334(6184):676–682, 1988.

[30] C. Goodnow, J. Sprent, B. Fazekas de St. Groth, and C.G. Vinuesa. Cellular and genetic mechanisms of self tolerance and autoimmunity. *Nature*, 435(7042):590–597, 2005.

[31] W.K. Hastings. Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[32] K. Hayakawa, M. Asano, S.A. Shinton, M. Gui, D. Allman, C.L. Stewart, J. Silver, and R.R. Hardy. Positive selection of natural autoreactive B-cells. *Science*, 285(5424):113–116, 1999.

[33] Jr H.W. Schroeder, F. Mortari S. Shiokawa, P.M. Kirkham, R.A. Elgavish, and 3rd F.E. Bertrand. Developmental regulation of the human antibody repertoire. *Ann N Y Acad Sci*, 764:242–260, 1995.

[34] Y. Ichihara, H. Matsuoka, and Y. Kurosawa. Organization of human immunoglobulin heavy chain diversity gene loci. *EMBO J*, 7(13):4141–4150, 1988.

[35] Y. Ichiyoshi and P. Casali. Analysis of the structural correlates for antibody polyreactivity by multiple reassortments of chimeric human immunoglobulin heavy and light chain V segments. *J Exp Med*, 180(3):885–895, 1994.

[36] J. Ito and D.K. Braithwaite. Compilation and alignment of DNA polymerase sequences. *Nucleic Acids Res*, 19(15):4045–4057, 1991.

[37] J. Jacob, R. Kassir, and G. Kelsoe. In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. I. The architecture and dynamics of responding cell populations. *J Exp Med*, 173(5):1165–1175, 1991.

[38] T.B. Kepler. Codon bias and plasticity in immunoglobulins. *Mol Biol Evol*, 14(6):637–643, 1997.

[39] R. Kleinfield, R.R. Hardy, D. Tarlinton, J. Dangl, L.A. Herzenberg, and M. Weigert. Recombination between an expressed immunoglobulin heavy-chain gene and a germline variable gene segment in a Ly 1+ b-cell lymphoma. *Nature*, 322(6082):843–846, 1986.

[40] K.D. Klonowski, L.L. Primiano, and M. Monestier. Atypical VH-D-JH re-

arrangements in newborn autoimmune MRL mice. *J Immunol*, 162(3):1566–1572, 1999.

[41] P. Kraj, D.F. Friedman, F. Stevenson, and L.E. Silberstein. Evidence for the overexpression of the VH4-34 (VH4.21) Ig gene segment in the normal adult human peripheral blood B cell repertoire. *J Immunol*, 154(12):6406–6420, 1995.

[42] P. Kraj, S.P. Rao, A.M Glas, R.R. Hardy, E.C. Milner, and L.E. Silberstein. The human heavy chain Ig V region gene repertoire is biased at all stages of B-cell ontogeny, including early pre-B cells. *J Immunol*, 158(12):5824–5832, 1997.

[43] R. Kunert, F. Rüker, and H. Katinger. Molecular characterization of five neutralizing anti-HIV type 1 antibodies: identification of nonconventional D segments in the human monoclonal antibodies 2G12 and 2F5. *AIDS Res Hum Retroviruses*, 14(13):1115–1128, 1998.

[44] R. Kunert, S. Wolbank, G. Stiegler, R. Weik, and H. Katinger. Characterization of molecular features, antigen-binding, and in vitro properties of IgG and IgM variants of 4E10, an anti-HIV type 1 neutralizing monoclonal antibody. *AIDS Res Hum Retroviruses*, 20(7):755–762, 2004.

[45] S. Lacroix-Desmazes, S.V. Kaveri, L. Mouthon, A. Ayouba, E. Malanchère, A. Coutinho, and M.D. Kazatchkine. Self-reactive antibodies (natural autoantibodies) in healthy individuals. *J Immunol Methods*, 216(1-2):117–137, 1998.

[46] M.P. Lefranc. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res*, 29(1):207–209, 2001.

[47] Y.S. Li, K. Hayakawa, and R.R. Hardy. The regulated expression of B lineage associated genes during B-cell differentiation in bone marrow and fetal liver. *J Exp Med*, 178(3):951–960, 1993.

[48] F. Livak, D.B. Burtrum, L. Rowen, D.G. Schatz, and H.T. Petrie. Genetic modulation of T cell receptor gene segment usage during somatic recombination. *J Exp Med*, 192(8):1191–1196, 2000.

[49] W. Lorenz, B. Straubinger, and H.G. Zachau. Physical map of the human immunoglobulin K locus and its implications for the mechanisms of VK-JK rearrangement. *Nucleic Acids Res*, 15(23):96679676, 1987.

[50] I.S. Lossos, R. Tibshirani, B. Narasimhan, and R. Levy. The inference of antigen selection on Ig genes. *J Immunol*, 165(9):5122–5126, 2000.

[51] A.J. Marshall, G.E. Wu, and C.J. Paige. Frequency of V(H)81x usage during B-cell development: Initial decline in usage is independent of Ig heavy chain cell surface expression. *J Immunol*, 156(6):2077–2084, 1996.

[52] T. Martin, R. Crouzier, J.C. Weber, T.J. Kipps, and J.L. Pasquali. Structure-function studies on a polyreactive (natural) autoantibody. Polyreactivity is dependent on somatically generated sequences in the third complementarity-determining region of the antibody heavy chain. *J Immunol*, 152(12):5988–5996, 1994.

[53] F. Matsuda, E.K. Shin, H. Nagaoka, R. Matsumara, M. Haino, Y. Fukita, S. Taka-ishi, T. Imai, J.H. Riley, R. Anand, E. Soeda, and T. Honjo. Structure and physical map of 64 variable segments in the 3'0.8-megabase region of the human immunoglobulin heavy-chain locus. *Nat Genet*, 3(1):88–94, 1993.

[54] D. McKean, K. Hüppi, M. Bell, L. Staudt, W. Gerhard, and M. Weigert. Generation of antibody diversity in the immune response of BALBc mice to influenza virus hemagglutinin. *Proc Natl Acad Sci*, 81(10):3180–3184, 1984.

[55] C.J. McMahan and P.J. Fink. Receptor revision in peripheral T-cells creates a diverse V beta repertoire. *J Immunol*, 165(12):6902–6907, 2000.

[56] I.S. Mian, A.R. Bradwell, and A.J. Olson. Structure, function and properties of antibody binding sites. *J Mol Biol*, 217(1):133–151, 1991.

[57] Y. Miura, C.C. Chu, D.M. Dines, S.E. Asnis, R.A. Furie, and N. Chiorazzi. Diversification of the Ig variable region gene repertoire of synovial B lymphocytes by nucleotide insertion and deletion. *Mol Med*, 9(5-8):166–174, 2003.

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

[59] M. Muramatsu, V.S. Sankaranand, S. Anant, M. Sugai, K. Kinoshita, N.O. Davidson, and T. Honjo. Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal center B-cells. *J Biol Chem*, 274(26):18470–18476, 1999.

[60] S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–453, 1970.

[61] D.A. Nemazee and K. Bürki. Clonal deletion of B lymphocytes in a transgenic mouse bearing anti-MHC class I antibody genes. *Nature*, 337(6207):562–566, 1989.

[62] M.S. Neuberger and C. Rada. Somatic hypermutation: activation-induced deaminase for C/G followed by polymerase eta for A/T. *J Exp Med*, 204(1):7–10, 2007.

[63] M.A. Oettinger, D.G. Schatz, C. Gorka, and D. Baltimore. RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. *Science*, 248(4962):1517–1523, 1990.

[64] L. Ohm-Laursen, M. Nielsen, S.R. Larsen, and T. Barington. No evidence for the use of DIR, D-D fusions, chromosome 15 open reading frames or VH replacement in the peripheral repertoire was found on application of an improved algorithm, JointML, to 6329 human immunoglobulin H rearrangements. *Immunology*, 119(2):265–277, 2006.

[65] M. Oprea and T.B. Kepler. Genetic plasticity of V genes under somatic hypermutation: statistical analyses using a new resampling-based methodology. *Genome Res*, 9(12):1294–1304, 1999.

[66] T. Ota and M. Nei. Divergent evolution and evolution by the birth-and-death process in the immunoglobulin V_H gene family. *Mol Biol Evol*, 11(3):469–482, 1994.

[67] F.N. Papavasiliou and D.G. Schatz. Somatic hypermutation of immunoglobulin genes: merging mechanisms for genetic diversity. *Cell*, 109(Suppl):S35–S44, 2002.

[68] V. Pascual, K. Victor, D. Lelsz, M.B. Spellerberg, T.J. Hamblin, K.M. Thompson, I. Randen, J. Natvig, J.D. Capra, and F.K. Stevenson. Nucleotide sequence analysis of the V regions of two IgM cold agglutinins. evidence that the VH4-21 gene segment is responsible for the major cross-reactive idiotype. *J Immunol*, 146(12):4385–4391, 1991.

[69] S.K Petersen-Mahrt, R.S. Harris, and M.S. Neuberger. AID mutates E. coli

suggesting a DNA deamination mechanism for antibody diversification. *Nature*, 418(6893):99–103, 2002.

[70] K.N. Potter, Y. Li, and J.D. Capra. Staphylococcal protein A simultaneously interacts with framework region 1, complementarity-determining region 2, and framework region 3 on human VH3-encoded Igs. *J Immunol*, 157(7):2982–2988, 1996.

[71] M.Z. Radic and M. Zouali. Receptor editing, immune diversification, and self-tolerance. *Immunity*, 5(6):505–511, 1996.

[72] J.V. Ravetch, U. Siebenlist, S. Korsmeyer, T. Waldmann, and P. Leder. Structure of the human immunoglobulin mu locus: characterization of embryonic and rearranged J and D genes. *Cell*, 27(3 Pt 2):583–591, 1981.

[73] S.K. Ray, C. Putterman, and B. Diamond. Pathogenic autoantibodies are routinely generated during the response to foreign antigen: a paradigm for autoimmune disease. *Proc Natl Acad Sci*, 93(5):2019–2024, 1996.

[74] M.G. Reth, S. Jackson, and F.W. Alt. VHDJH formation and DJH replacement during pre-B differentiation: non-random usage of gene segments. *EMBO J*, 5(9):21312138, 1986.

[75] P. Revy, T. Muto, Y. Levy, F. Geissmann, A. Plebani, O. Sanal, N. Catalan, M. Forveille, R. Dufourcq-Labelouse, A. Gennery, I. Tezcan, F. Ersoy, H. Kayserili, A.G. Ugazio, N. Brousse, M. Muramatsu, L.D. Notarangelo, K. Kinoshita, T. Honjo, A. Fischer, and A. Durandy. Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). *Cell*, 102(5):565–575, 2000.

[76] J.S. Rice, C. Kowal, B.T. Volpe, L.A. DeGiorgio, and B. Diamond. Molecular mimicry: anti-DNA antibodies bind microbial and nonnucleic acid self-antigens. *Curr Top Microbiol Immunol*, 296:137–151, 2005.

[77] I.F. Robey, M. Peterson, M.S. Horwitz, F.H. Kono, T. Stratmann, A.N. Theofilopoulos, N. Sarvetnick, L. Teyton, and A.J. Feeney. Terminal deoxynucleotidyltransferase deficiency decreases autoimmune disease in diabetes-prone nonobese diabetic mice and lupus-prone MRL-Fas(lpr) mice. *J Immunol*, 172(7):4624–4629, 2004.

[78] I.B. Rogozin and N.A. Kolchanov. Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. *Biochim*

Biophys Acta, 1171(1):11–18, 1992.

[79] K. Rosner, D.B. Winter, R.E. Tarone, G.L Skovgaard, A. Bohr, and P.J. Gearhart. Third complementarity-determining region of mutated V_H immunoglobulin genes contains shorter V, D, J, P, and N components than non-mutated genes. *Immunology*, 103(2):179–187, 2001.

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

[81] I. Sanz. Multiple mechanisms participate in the generation of diversity of human H chain CDR3 regions. *J. Immunol*, 147(5):1720–1729, 1991.

[82] G.S. Shapiro, K. Aviszus, D. Ikle, and L.J. Wysocki. Predicting regional mutability in antibody V genes based solely on di- and trinucleotide sequence composition. *J Immunol*, 163(1):259–268, 1999.

[83] G.S. Shapiro, K. Aviszus, J. Murphy, and L.J. Wysocki. Evolution of Ig DNA sequence to target specific base positions within codons for somatic hypermutation. *J Immunol*, 168(5):2302–2306, 2002.

[84] S. Shiokawa, F. Mortari, J.O Lima, C. Nu nez, 3rd F.E. Bertrand, P.M. Kirkham, S. Zhu, A.P. Dasanayake, and Jr H.W. Schroeder. IgM heavy chain complementarity-determining region 3 diversity is constrained by genetic and somatic mechanisms until two months after birth. *J Immunol*, 162(10):6060–6070, 1999.

[85] M. Shlomchik, M. Mascelli, H. Shan, M.Z. Radic, D. Disetsky, A. Marshak-Rothstein, and M. Weigert. Anti-DNA antibodies from autoimmune mice arise by clonal expansion and somatic mutation. *J Exp Med*, 171(1):265–292, 1990.

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

[87] C.A. Siegrist. Neonatal and early life vaccinology. *Vaccine*, 19(25-26):3331–3346, 2001.

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

[89] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147(1):195–197, 1981.

[90] M.M. Souto-Carneiro, N.S. Longo, R.E. Daniel, H. Sun, and P.E. Lipsky. Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOVLER. *J Immunol*, 172(11):6790–6802, 2004.

[91] M.M. Souto-Carneiro, G.P. Sims, H. Girschik, J. Lee, and P.E. Lipsky. Developmental changes in the human heavy chain CDR3. *J Immunol*, 175(11):7425–7436, 2005.

[92] S.L. Tiegs, D.M. Russell, and D. Nemazee. Receptor editing in self-reactive bone marrow B cells. *J Exp Med*, 177(4):1009–1020, 1993.

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

[94] A. Trkola, A.B. Pomales, H. Yuan, B. Korber, P.J. Maddon, G.P. Allaway, H. Katinger, C.F. Barbas 3rd, D.R. Burton, and D.D. Ho *et.al.* Cross-clade neutralization of primary isolates of human immunodeficiency virus type 1 by human monoclonal antibodies and tetrameric CD4-IgG. *J Virol*, 69(11):6609–6617, 1995.

[95] R.F. van Vollenhoven, M.M Bieber, M.J Powell, P.K. Gupta, N.M Bhat, K.L. Richards, S.A. Albano, and N.N. Teng. VH4-34 encoded antibodies in systemic lupus erythematosus: a specific diagnostic marker that correlates with clinical disease characteristics. *J Rheumatol*, 26(8):1727–1733, 1999.

[96] J.M Volpe, L.G. Cowell LG, and T.B. Kepler. SoDA: implementation of a 3D alignment algorithm for inference of antigen receptor recombinations. *Bioinformatics*, 22(4):438–444, 2006.

[97] S.D. Wagner, C. Milstein, and M.S. Neuberger. Somatic hypermuation of immunoglobulin genes. *Nature*, 376(6543):732, 1995.

[98] S.D. Wagner and M.S. Neuberger. Codon bias targets mutation. *Annu. Rev. Immunol.*, 14:441–457, 1996.

[99] H. Wardemann, S. Yurasov, A. Schaefer, J.W. Young, E. Meffre, and M.C. Nussenzweig. Predominant autoantibody production by early human B cell precursors. *Science*, 301(5638):1374–1377, 2003.

[100] R. Wasserman, Y. Ito, N. Galili, M. Yamada, B.A. Reichard, S. Shane, B. Lange, and G. Rovera. The pattern of joining (JH) gene usage in the human IgH chain is established predominantly at the B precursor cell stage. *J Immunol*, 149(2):511–516, 1992.

[101] R. Wasserman, Y.S. Li, and R.R. Hardy. Down-regulation of terminal deoxynucleotidyl transferase by Ig heavy chain in B lineage cells. *J Immunol*, 158(3):1133–1138, 1997.

[102] T.H. Winkler, H. Fehr, and J.R. Kalden. Analysis of immunoglobulin variable region genes from human IgG anti-DNA hybridomas. *Eur J Immunol*, 22(7):1719–1728, 1992.

[103] M. Yamada, R. Wasserman, B.A. Reichard, S. Shane, A.J Caton, and G. Rovera. Preferential utilization of specific immunoglobulin heavy chain diversity and joining segments in adult human peripheral blood B lymphocytes. *J Exp Med*, 173(2):395–407, 1991.

[104] M. Zemlin, K. Bauer, M. Hummel, S. Pfeiffer, S. Devers, C. Zemlin, H. Stein, and H.T. Versmold. The diversity of rearranged immunoglobulin heavy chain variable region genes in peripheral blood B cells of preterm infants is restricted by short third complementarity-determining regions but not by limited gene segment usage. *Blood*, 97(5):1511–1513, 2001.

[105] M. Zemlin, M. Klinger, J. Link, C. Zemlin, K. Bauer, J.A. Engler, H.W. Schroeder Jr, and P.M. Kirkham. Expressed murine and human CDRH3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. *J Mol Biol*, 334(4):733–749, 2003.

Biography

Joseph Michael Volpe (Joe) was one of the first five graduate students to matriculate in the Computational Biology & Bioinformatics program at Duke University in 2003. He attended Wake Forest University for his undergraduate education, where he was named to the Dean's list every semester during his tenure and earned his bachelors degree in computer science, graduating Magna Cum Laude in 1999. Joe came to Duke after spending four years working in business for Forrester Research and IBM Global Services. He joined the Duke University Laboratory for Computational Immunology (DULCI) lab in 2005, where, under the direction of Dr. Tom Kepler, he learned to apply computational and statistical techniques to immunological problems. His work has focused on understanding biases that affect development of the human immunoglobulin heavy chain repertoire. During his tenure at Duke, Joe has given two symposia at the American Association of Immunologists conferences, one in Miami Beach in 2007 and the second in San Diego in 2005. In 2007, he was awarded the American Association of Immunologists-Huang Foundation Young Investigator award for his work on human immunoglobulin mutational patterns and broadly neutralizing anti-HIV antibodies. He also gave an oral presentation at the 2004 Duke University Graduate Student Symposium for his work on the human Ig heavy chain gene segment repertoire.

Joe has published several first author manuscripts, including one in *Bioinformatics* and one in *Immunome Research*, and has collaborated on publications with other groups outside of Duke. His first publication in *Bioinformatics* described research

involved in creating a new software tool called SoDA that provides a statistical reconstruction of the recombination and diversity generating events that led to a given antigen receptor gene. The software is now being used in a major HIV vaccine initiative to study the structure of anti-HIV antibodies, as well as by researchers tracking T-cell repertoire development in thymus transplant recipients.

Outside of school, Joe enjoys cooking, especially pastries and Italian food. He is also a composer and pianist, and has recorded an album containing nine of his original piano compositions. He is the third of four children in his family, and also the third so far to earn the title of "Doctor".