THE DEVELOPMENT AND GENETIC ORIGIN OF BROADLY NEUTRALIZING HIV ANTIBODIES

By

Bryan S. Briney

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Microbiology and Immunology

August, 2012

Nashville, Tennessee

Approved:

Professor James E. Crowe, Jr.

Professor James W. Thomas

Professor Christopher R. Aiken

Professor Spyros A. Kalams

Professor Billy G. Hudson

For my parents, always supportive

and

For my wife, Mandie. Most. Done.

ACKNOWLEDGEMENTS

This work would not have been possible without the extraordinary leadership and guidance of my mentor, Dr. James Crowe. His unrivaled passion for big science and core commitment to push the limits of technology in pursuit of discovery were a driving force behind my research and are an ongoing inspiration. I share his love of the search for beauty in data, and his mentorship has enabled me to become a more complete scientist.

I would also like to express my sincere gratitude to my dissertation committee, Dr. Tom Thomas, Dr. Chris Aiken, Dr. Spyros Kalams, and Dr. Billy Hudson for their exemplary scientific guidance and direction. I would like to especially thank my committee chairman, Dr. Tom Thomas, for his leadership and support.

This work was supported by all members of the Crowe Lab, but special thanks go to those who directly contributed to my dissertation work, Dr. Mark Hicar and Jordan Willis. The Vanderbilt Flow Cytometry Core and Genome Sciences Resource both provided much needed assistance and experimental advice. Of course, this work would not have been possible without funding sources, which included the NIH, the NIAID, and the Vanderbilt Cell and Molecular Biology Training Grant.

Finally, I would like to thank my family. My parents have been the biggest fans of my educational career since before kindergarten. Their constant encouragement has been a joy and an inspiration. My wife has been, without a doubt, my biggest supporter. Her willingness to move halfway across the country, away from friends and family, to allow me to pursue my dream is worth more than she will ever know. Her support has been endless, and her uncanny ability to express interest and excitement while I drone on about intensely boring topics has helped preserve my sanity.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Organization of the Human Antibody Germline Locus

Antibodies are antigen receptors expressed by B cells and are critical to the eradication of pathogenic infection and establishment of immunological memory. The antibody protein consists of two identical heavy chains associated with two identical light chains. The antibody genes that encode heavy and light chains are located in three primary locations in the human genome: heavy chain genes (IGH) are located on chromosome 14, light chain kappa genes (IGK) are located on chromosome 2, and light chain lambda genes (IGL) are located on chromosome 22 (Brochet et al., 2008). Each of these loci consists of multiple variable (V) and joining (J) gene segments, and the IGH locus also contains several diversity (D) gene segments. Sequencing of the human IGH locus revealed 44 functional V

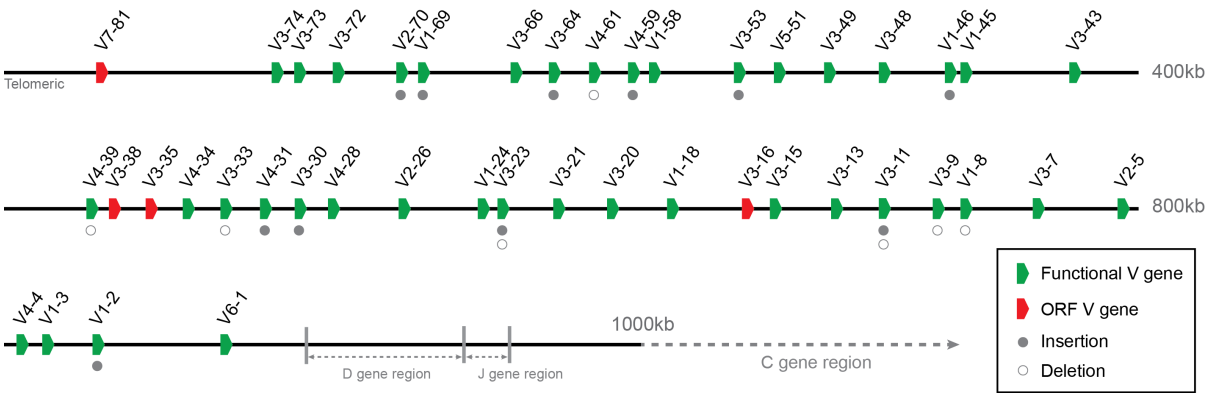


Figure 1. Organization of the human heavy chain locus. All functional heavy chain variable genes are shown (green) as well as all ORF variable genes (red). Variable genes that are suspected to reside in regions of insertion (filled circles) or deletion (open circles) polymorphisms are indicated. Image adapted from IMGT.org.

genes (and 85 pseudogenes), 27 D genes (23 functional), and nine J genes (6 functional) (Matsuda et al., 1998). Further sequencing has resulted in the identification of 11 additional functional V gene segments that were not present in the original reference genome (Lefranc et al., 2009). A simplified diagram of the IGH locus is shown in Figure 1.

The human variable genes (and, at the IGH locus, the diversity genes) can be phylogenetically grouped into families based on sequence similarity. Heavy chain variable genes are organized into seven families and homology within gene families is typically above 80%. Sequence homology between heavy chain variable genes belonging to different families is less than 70%. The 23 functional human diversity genes are also organized into seven families. In the germline locus, the diversity genes are uniquely positioned in four tandemly oriented groups (Siebenlist et al., 1981; Corbett, 1997), and are located approximately 8kb upstream of Cμ, the nearest constant region gene. The unique positioning of diversity genes in regularly spaced 9kb clusters suggests that the locus may have been created by a series of duplications (Siebenlist et al., 1981; Ichihara et al., 1988). The standard IMGT nomenclature for human V and D genes follows the following pattern: the chain and gene description (IGHV for variable genes, IGHD for germline genes), the family, the gene number (determined by position in the germline locus), and the allele. The gene number is separated from the family with a hyphen and the allele is separated from the gene number with an asterisk. Thus, an example variable gene nomenclature appears as: IGHV1-2*03.

V(D)J Recombination

Since the discovery that RAG-mediated recombination of variable, diversity and joining genes generates virtually unlimited sequence diversity in the antibody repertoire (Brack et al., 1978; Alt and Baltimore, 1982; Tonegawa, 1983; Schatz et al., 1989; Oettinger

et al., 1990), much progress has been made in determining the genetic and mechanistic elements that participate in the antibody recombination process. It is generally understood that recombination signal sequences (RSS), which flank V, D and J genes and are composed of conserved AT-rich heptamer and nonamer sequences separated by spacers of either 12 or 23 nucleotides, are recognized and bound by recombination activating gene (RAG1 and RAG2) proteins at the initiation of the recombination process (Hesse et al., 1989; Alt et al., 1992). RAG binding is highly dependent on the heptamer and nonamer sequences, and alterations to either sequence results in decreased RAG binding (Cuomo et al., 1996; Difilippantonio et al., 1996; Nadel et al., 1998). The length of the spacer sequence is critical to recombination, and there is evidence of sequence conservation within the spacer region (Ramsden et al., 1994; Lee et al., 2003; Montalbano et al., 2003).

Recombination typically occurs only between RSS elements of different spacer lengths, in a model commonly referred to as the 12/23 rule of recombination (Ramsden et al., 1996; Steen et al., 1996; van Gent et al., 1996; Schatz, 2004). After binding to one 12-bp RSS and one 23-bp RSS, the RAG complex induces single-strand DNA nicks between the

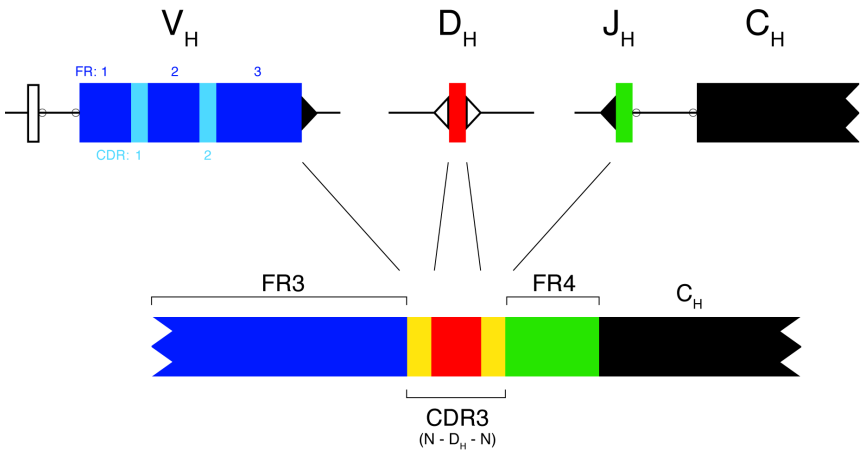


Figure 2. V(D)J recombination at the human heavy chain locus. Recombination between V_H (blue), D_H (red) and J_H (green) genes is mediated by 12bp RSSs (open arrows) and 23bp RSSs (filled arrows). The final recombinant includes N-addition at the V-D and D-J junctions (yellow).

coding sequence and the heptamer of each RSS, resulting in hairpin formation on each of the coding ends and a blunt double-stranded break on each signal end (Roth et al., 1992; Schlissel et al., 1993; McBlane et al., 1995; Sadofsky, 2001). The hairpins are opened, nucleotides may be added to or removed from the coding ends, and the double strand breaks at the coding ends are joined into a single coding strand (Lewis, 1994; Mahajan et al., 1999; Shockett and Schatz, 1999; Walker et al., 2001; Mansilla-Soto and Cortes, 2003; Roth, 2003).

In antibody heavy chain genes, diversity genes are flanked by 12-bp RSSs on either side, while variable and joining genes are flanked by 23-bp RSSs (Early et al., 1980; Kurosawa and Tonegawa, 1982). Recombination of the light chain is the result of a single V_L-J_L recombination; heavy chain recombination (Figure 2) occurs in step-wise fashion, with D-J_H recombination preceding V_H-D recombination (Alt et al., 1987; Schatz et al., 1992).

Somatic Hypermutation

Establishment of a diverse primary (or naïve) antibody repertoire is a result of RAG-mediated recombination. Diversification of the secondary antibody repertoire, which arises in response to antigenic stimulus, is accomplished primarily through somatic hypermutation (SHM) (Brenner and Milstein, 1966; Kelsoe, 1994). Naïve, antigen inexperienced B cells undergo the SHM process upon recognition of an infectious agent. It is through the SHM process, which occurs primarily in secondary lymphoid tissue, that they mutate the variable region of their antibody genes (MacLennan et al., 1992; Li et al., 2004). Many of these mutations have no effect on antigen recognition and many have deleterious effects on either antigen recognition or proper folding of the antibody protein, however, some mutations produce antibodies with improved affinity for the target pathogenic epitope (Casali et al.,

2006). Thus, the SHM process provides a basis for the positive selection of high affinity antibodies that are characteristic of a mature immune response (MacLennan, 1994).

Many components of the SHM machinery are known, but the complete process and the mechanisms by which it is targeted specifically to the immunoglobulin loci are still poorly understood. SHM introduces point mutations at a frequency of approximately 10^{-3} mutations per base pair, which is 10^{6}-fold higher than the rate of spontaneous mutation in other genes (Rajewsky et al., 1987). Mutations begin approximately 150-bp downstream of the transcription start site and the mutation frequency decreases exponentially with increasing distance from the transcription start site (Rada and Milstein, 2001). Activation-induced cytidine deaminase (AID) is required for SHM and initiates the SHM process by the deamination of cytosine nucleotides, which results in the conversion of cytosine to uracil (Muramatsu et al., 1999; 2000). Deamination thus produces a uracil-guanine mismatch, and several possible processes result in the error-prone repair of the mismatch. The precise mechanism(s) responsible for error-prone repair during SHM are not known, although several DNA repair mechanisms have been shown to be critical to the SHM process, including base excision repair and mismatch repair (Phung et al., 1998; Rada et al., 1998; Wiesendanger et al., 2000; Di Noia and Neuberger, 2002; Zheng et al., 2005).

Antibody complementarity determining regions (CDRs, also referred to as hypervariable regions) are the primary region of antigen recognition and are preferentially targeted for affinity maturation by the SHM machinery, making them the most variable regions of the antibody gene (Capra and Kehoe, 1975; Kabat et al., 1992). There are several structural and genetic reasons for the preferential targeting of CDRs by SHM. Genetically, SHM is known to preferentially target the WRCY hotspot motif (or its reverse complement, RGYW) (Dörner et al., 1998), and the frequency of these hotspots is increased in CDRs (Wagner et al., 1995; Shapiro and Wysocki, 2002; Pham et al., 2003). Further, codon usage is biased in CDRs toward codons that are easily mutable, enhancing the

likelihood that a nucleotide substitution induced by SHM results in an amino acid change (Motoyama et al., 1991; Wagner et al., 1995; Kepler, 1997). Structurally, the CDRs are largely loop-based, which make them sufficiently flexible to incorporate the substitutions and short indels introduced by SHM without compromising structural integrity. Framework regions (FRs), by contrast, are highly structured and less able to accommodate somatic mutations (Celada and Seiden, 1996).

Secondary Mechanisms of Affinity Maturation

V(DD)J and direct V_H-J_H recombination

Direct V_H-J_H joining and V(DD)J recombination (also referred to as D-D fusion) are in direct violation of the 12/23 rule, but such recombination events have been demonstrated in both *in vitro* and *in vivo* systems (Sanz, 1991; Kiyoi et al., 1992; Raaphorst et al., 1997; Koralov et al., 2005; 2006; Watson et al., 2006). Even in model systems designed to induce such recombination events, however, non-12/23 recombinations are much less efficient than recombinations that adhere to the 12/23 rule (Akira et al., 1987; Hesse et al., 1989; Akamatsu et al., 1994).

V(DD)J recombinants are the result of an aberrant recombination process by which two or more diversity genes are joined into a single recombinant. The joining of two diversity genes, which are flanked on both sides by 12-bp RSSs, can only be accomplished in clear violation of the 12/23 rule, but such recombination events have been seen both *in vivo* and *in vitro* (Sanz, 1991; Kiyoi et al., 1992; Raaphorst et al., 1997; Watson et al., 2006). V(DD)J recombinations have been estimated to occur in as many as 5-11% of all recombinations (Sanz, 1991; Kiyoi et al., 1992; Raaphorst et al., 1997), but the true frequency of V(DD)J recombinations is difficult to determine. Identification of V(DD)J recombinants relies on the accurate detection of two diversity genes within a single recombinant, but N-addition

mimicry of diversity gene segments, which is genetically indistinguishable from true V(DD)J recombination, likely inflates current estimates of V(DD)J recombination (Watson et al., 2006).

V_H replacement and receptor revision

V_H replacement is a process by which a secondary V_H-V(D)J recombination can occur, resulting in replacement of the variable gene while preserving the original D-J_H recombination (Nemazee, 2006). V_H replacement differs from receptor editing, which is the process of secondary rearrangement in light chains. Receptor editing results in an entirely new V_L-J_L recombination through the recombination of a V_L gene segment upstream of the original recombination with a J_L gene segment downstream of the original recombination (Papavasiliou et al., 1997; Retter and Nemazee, 1998). Thus, receptor editing proceeds without violating the 12/23 rule. V_H replacement utilizes a cryptic RSS (cRSS) found near the 3′ end of most human variable genes (Radic and Zouali, 1996), and this cRSS is used to recombine with the normal RSS at the 3′ end of the invading variable gene. The cRSS contains a heptamer sequence, but lacks an identifiable nonamer or spacer sequence, and recombination with the cRSS is inefficient, much like other forms of non-12/23 recombination (Koralov et al., 2006; Lutz et al., 2006).

V_H replacement was observed first in transformed murine pre-B cells (Kleinfield et al., 1986; Reth et al., 1986), with subsequent studies identifying V_H replacement *in vivo* (Taki et al., 1993; Chen et al., 1995). In the most informative work done on V_H replacement in the human repertoire, a genetic fingerprint of V_H replacement was identified in the human peripheral blood repertoire (Zhang et al., 2003). Identification of V_H replacement events in the peripheral repertoire relies on detection of short pentameric sequences that are located between the cRSS and the 3′ end of variable genes. These pentamers remain even after V_H

replacement, providing an identifiable remnant of the replaced variable gene. Short pentameric sequences are easily mimicked through random N-addition, making reliable detection of V_H replacement difficult. As such, estimates of V_H recombination frequency in the peripheral blood repertoire have varied widely, from 5% to 22% of the total repertoire (Zhang et al., 2003; Koralov et al., 2006; Watson et al., 2006).

Somatic hypermutation-associated insertions and deletions

Although the somatic hypermutation process typically results in single nucleotide substitutions, deletion of germline nucleic acids or insertion of non-germline nucleic acids does occur in association with somatic hypermutation (Goossens et al., 1998; Wilson et al., 1998a; Bemark and Neuberger, 2003). These insertions and deletions are rare, with somatic hypermutation-associated (SHA) insertions or deletions estimated to be present in 1.3 to 6.5% of circulating B cells (Goossens et al., 1998; Wilson et al., 1998a; Bemark and Neuberger, 2003). Short SHA indels are much more common than long SHA indels, with most insertions and deletions being 1-2 codons in length (Goossens et al., 1998; Wilson et al., 1998a; Bemark and Neuberger, 2003). Although infrequent, SHA insertion and deletion events add substantially to the diversity of the human antibody repertoire (Wilson et al., 1998b; de Wildt et al., 1999; Reason and Zhou, 2006).

SHA insertions and deletions also have been shown to play a critical role in the antibody response against viral and bacterial pathogens, including HIV, influenza, and *Streptococcus pneumoniae* (Zhou et al., 2004; Walker et al., 2009; Wu et al., 2010a; Krause et al., 2011; Pejchal et al., 2011; Walker et al., 2011). Of particular interest, structural analysis of an SHA insertion in the anti-influenza antibody 2D1 identified a substantial structural alteration induced by the insertion (Krause et al., 2011). This insertion, although located in a framework region, caused a large conformational change in a CDR and allowed antibody-antigen interactions that were sterically hindered without the insertion-induced

conformational change. In addition to 2D1, the extremely broad and potently neutralizing HIV antibody VRC01 contained a six nucleotide deletion in the CDR1 of the light chain (CDR-L1) (Wu et al., 2010a). This SHA deletion shortened the CDR-L1 loop, thereby removing potential clashes with loop D of the HIV envelope protein and allowing direct interaction between the HIV antigen and the CDR-L2 loop of VRC01 (Zhou et al., 2010).

Affinity maturation and antigen contact by antibody framework regions

While much affinity maturation effort is expended on the CDRs, there are other regions that are important to antigen recognition. T cell receptors (TCRs) contain a fourth hypervariable region (HV4, sometimes referred to as CDR4), which is highly variable, surface exposed, and is involved in superantigen and accessory molecule recognition (Choi et al., 1990; Garcia et al., 1996; Li et al., 1998).

Emerging evidence suggests that an HV4-like region may exist in antibodies as well as TCRs. Recent crystallographic work on the anti-influenza antibody CR6261 has shown that the HV4-like region of FR3 was somatically mutated (Throsby et al., 2008) and directly contributed to antigen binding (Ekiert et al., 2009). The anti-influenza antibody 2D1 contains a three codon insertion in a HV4-like region of FR3 which, while not directly involved in antigen recognition, causes a critical conformational shift in nearby CDRs that is required for antigen recognition (Krause et al., 2011). A unique example of HV4-like contribution to antigen recognition is the anti-HIV antibody 21c (Diskin et al., 2010). 21c binds to the HIV co-receptor binding pocket, which is only exposed following binding of CD4, the primary host receptor. Interestingly, while the majority of the binding surface of 21c is in contact with the HIV envelope protein, the HV4-like region of 21c binds to CD4, forming a cross-protein epitope. In addition to 21c, the broadly neutralizing anti-HIV antibody VRC03 contains a surprisingly long seven codon insertion in the HV4-like region of FR3 (Wu et al., 2010a). Finally, the HV4-like FR3 region of antibody heavy chains of the VH3 family has been shown

to interact with Staphylococcal protein A, a known superantigen (Potter et al., 1996), mimicking the superantigen-binding activity of the HV4 region in TCRs. While the HV4-like regions that have been identified to date are not somatically mutated to the same extent as antibody CDRs, the ability of this HV4-like region to tolerate a substantial number of somatic mutations and genetic insertions suggests the existence of a somewhat flexible region that has an under-appreciated ability to accommodate affinity maturation modifications.

Human Immunodeficiency Virus

Structure, function and diversification of HIV Env

Human Immunodeficiency Virus Type 1 (HIV-1), a member of the *Retroviridae* family,

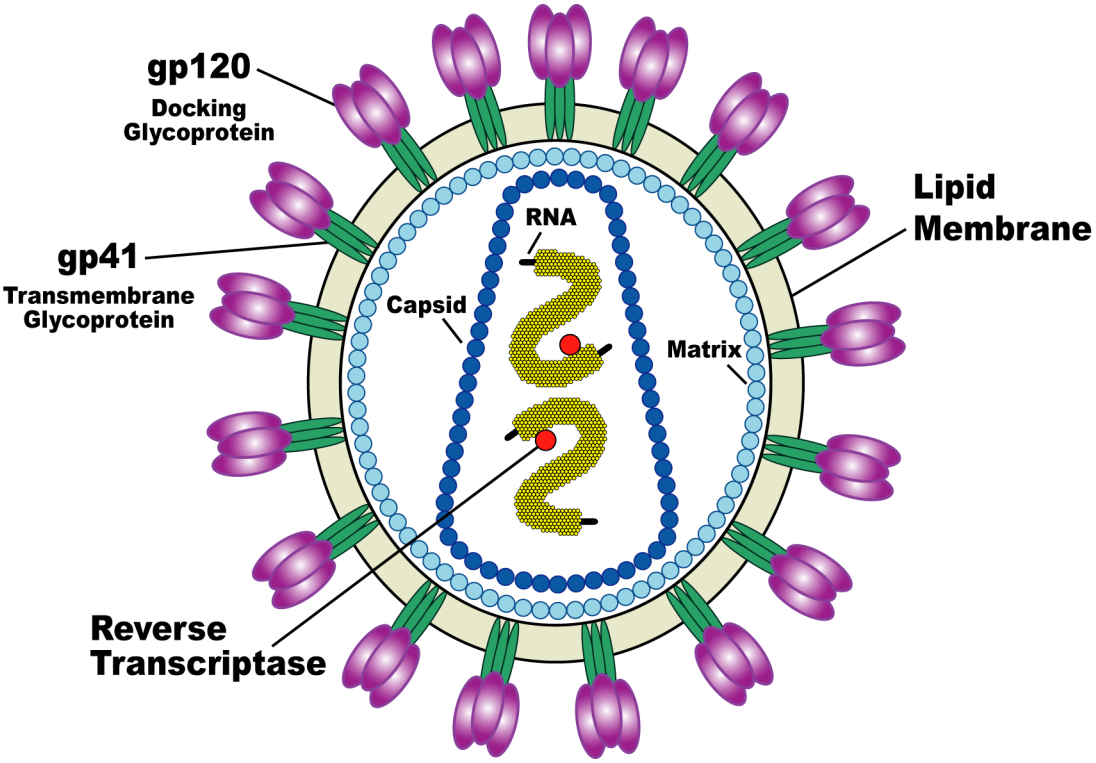


Figure 3. Structure of the HIV virion particle. Image adapted from AIDSreagents.org.

is an enveloped virus containing a positive-strand RNA genome (Figure 3). During replication, the RNA genome is converted to double-stranded DNA by the highly error prone reverse transcriptase enzyme, which generates point mutations and intergenomic recombination and creates massive diversity in the viral envelope (Env) protein as well as other viral proteins. Based on genome diversity, HIV isolates can be classified into three groups: M (Major), N (Non-M, Non-O) and (Outlier). Group M, which is the most prevalent, is further classified into subtypes or clades (A-D, F-H, J and K). Clade B is the most frequently found circulating clade in the United States, while clade C is the most common subtype in India, China and Sub-Saharan Africa. In addition to the major clades, circulating recombinant forms of HIV exist and display genetic characteristics of several clades, presumably formed through intergenomic recombination.

The functional spike on the surface of the virion is the Env glycoprotein. The Env glycoprotein complex is originally produced as a single-chain glycoprotein precursor, gp160, which is cleaved by a cellular protease. Cleavage of gp160 results in the cell-surface attachment protein gp120 and the membrane-spanning protein gp41. The gp160 cleavage products are noncovalently linked and assemble into a trimer of gp120-gp41 heterodimers that is expressed on the virion surface (Kowalski et al., 1987; Lu et al., 1995). Gp120 is heavily glycosylated, with nearly half the total mass being the result of N-linked glycans (Poignard et al., 2001). It is composed of five variable regions (V1-V5) interspersed with five constant regions (C1-C5) (Starcich et al., 1986).

The principle function of the glycoprotein spike is to facilitate cell entry by binding to the primary cell-surface receptor, CD4, and one of the two co-receptors, CCR5 and CXCR4. Binding to the receptor and co-receptor is accomplished by gp120, and fusion of the viral and cell membranes is mediated by gp41 (Zwick et al., 2004b). Gp41 is relatively well conserved, but most of its surface appears to be shielded from antibody recognition before attachment and fusion (Weiss, 2003).

Evasion of the humoral immune response

The enormous variability of HIV is an effective mechanism for evading neutralizing antibody. The sequence variation in one isolate from a single HIV-infected individual sampled a few years after infection is greater than the global variation of an influenza epidemic strain during a flu season (Korber et al., 2001). In HIV Env, sequence variability is concentrated in the variable loops (V1–V5), which appear to be a major target for neutralizing antibody responses. Escape from these responses is readily achieved by mutations in the loops that have only minor consequences for viral fitness. Longitudinal studies in humans show that a neutralizing antibody response to the dominant virus does develop but, once a threshold is reached, an escape variant is selected (Richman et al., 2003; Wei et al., 2003). In turn, an antibody response to this variant develops over time that results in the selection of a new escape variant and further repetition of the process. Numerous monoclonal antibodies (mAbs) to Env have been isolated from humans, but only a few of these can effectively neutralize most strains of the virus (Burton et al., 1994; Binley et al., 2004; Zolla-Pazner et al., 2004; Burton et al., 2005a; Cardoso et al., 2005; Haynes and Montefiori, 2006; Walker et al., 2009; Wu et al., 2010a; Walker et al., 2011). This highlights one of the central features of the humoral immune response against HIV-1: Most of the anti-Env antibodies generated during natural infection are directed to regions of gp120 or gp41 that are not exposed on the mature functional virus spike. This may arise because gp120 readily dissociates from gp41 and because certain epitopes on monomeric forms of gp120 and exposed regions of gp41 are highly immunogenic (Haynes and Montefiori, 2006; Pantophlet and Burton, 2006). This results in a large proportion of antibodies that can be detected by gp120 or gp140 binding assays, but are unable to bind the native virus spike and neutralize the virus. However, sera from some HIV-1-infected subjects are able to potently neutralize diverse isolates of HIV-1 (Dhillon et al., 2007; Li et al., 2007; Binley et al.,

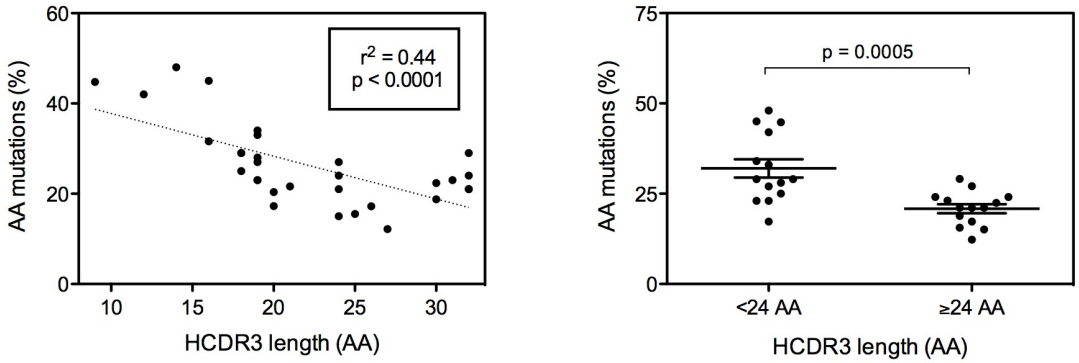


Figure 4. HCDR3 length and somatic mutation frequency in HIV bnAbs. HCDR3 length and mutation frequency are negatively correlated (left panel). HIV bnAbs with long HCDR3s (24AA, or two standard deviations above the mean HCDR3 length for the total repertoire) are significantly less mutated than bnAbs with HCDR3s shorter than 24AA in length (right panel).

2008; Sather et al., 2009; Simek et al., 2009), which demonstrates that there are vulnerable regions on the functional Env trimer. Antibodies that can recognize many different variants of HIV, so-called broadly neutralizing antibodies (bnAbs), are thought to evolve slowly and only in some individuals. It is precisely this type of antibody that one would like to elicit by vaccination, and so they have attracted special interest.

Broadly Neutralizing HIV Antibodies

A small but growing number of broadly neutralizing monoclonal antibodies, antibodies capable of neutralizing HIV isolates from different clades, have been identified (Burton et al., 2004). These mAbs are important in vaccine design but, unexpectedly, elucidation of their 3D structures has provided remarkable insight into the adaptability of antibodies when challenged by a virus that has developed an extensive repertoire of antibody evasion techniques. For nearly a decade, only a small panel of broadly neutralizing antibodies had been identified. These antibodies targeted three distinct epitopes, and it was long presumed that all neutralizing epitopes had been identified. However, recently

described bnAbs revealed a novel neutralizing epitope (Walker et al., 2011), raising the tantalizing possibility that there are further undiscovered neutralizing epitopes on the HIV Env protein.

While most HIV-infected individuals mount a neutralizing antibody response to autologous virus isolates, only rarely do individuals mount a broadly neutralizing antibody response. It is not clear why bnAbs are so rarely generated, but it is possible that antibodies able to neutralize diverse HIV isolates require uncommonly found genetic elements to facilitate broad neutralization. In fact, the growing panel of broadly neutralizing HIV antibodies has revealed the use several unique genetic structural and genetic features to exploit the few weaknesses in the HIV glycoprotein. Two of the most commonly found unique features are the use of extremely long HCDR3 loops and extensive somatic mutation. These two features are negatively correlated (Figure 4), indicating that either a long HCDR3 or extensive somatic mutation is required, but typically not both.

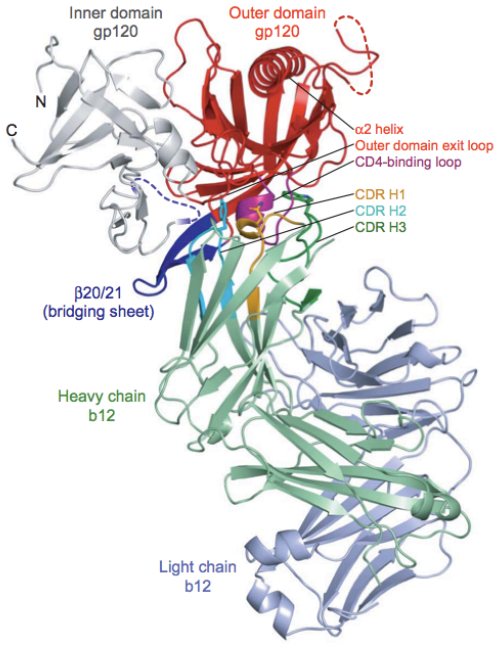


Figure 5. Crystal structure of b12 in complex with HIV gp120. b12 binds to the CD4bs using only heavy chain residues, and three heavy chain residues (Asn31, Tyr53 and Trp100) make up over 40% of the contact surface. From Zhou et al, 2007.

CD4 binding site antibodies

The antibody b12, which binds to an epitope that overlaps the CD4 binding site (CD4bs) on gp120 (Burton et al., 1991; Barbas et al., 1993; Burton et al., 1994), has been studied extensively with regard to its neutralization activity *in vitro* and *in vivo*.

At the time of discovery, b12 was the most broad and potently neutralizing anti-gp120 antibody. In *in vitro* neutralization assays involving a panel of 90 viruses, the antibody was able to reduce viral infection by 50% for approximately half of the viruses within the test panel at an inhibitory concentration (IC50) of less than 20 μg/mL (Binley et al., 2004). *In vivo* studies using macaque models of HIV infection have shown that b12 can also protect animals against viral challenge (Parren et al., 2001; Veazey et al., 2003). The most prominent features of the b12 combining site are a long (18 amino acid) heavy chain complementarity determining region 3 (HCDR3) that extends directly out from the surface of the antibody like a finger (Figure 5; from Zhou, 2007) and the exclusive use of heavy chain residues in the binding interaction (Burton et al., 2005b; Zhou et al., 2007). Computer docking and mutagenesis studies on gp120 and b12 have been used to argue that the finger-like HCDR3 is able to reach into the recessed CD4 binding site of gp120 (Saphire et al., 2001; Pantophlet and Ollmann Saphire, 2003; Zwick et al., 2003).

After identification of a panel of HIV-infected individuals whose serum was broadly neutralized and focused on the CD4bs, a structure-guided technique was used to engineer gp120 by resurfacing the molecule to alter antigenic surfaces while preserving the CD4bs (Wu et al., 2010a). This resurfaced, stabilized core protein (referred to as RSC3), in combination with a null variant which was identical to RSC3 with the exception of a single point mutation that eliminated antibody binding to the CD4bs, was able to effectively obscure many highly immunogenic non-neutralizing epitopes and allow focusing on the CD4bs. By sorting single B cells using RSC3, three bNmAbs were identified (VCR01-03) that neutralized over 90% of circulating HIV-1 isolates with an average IC50 of less than 1 μg/mL. All three VRC antibodies display extensive somatic mutation, with as many as 40% of the variable gene amino acids altered from the germline encoded sequence. In addition, VRC01 and VRC03, which are clonally related, contain a six nucleotide deletion in the light chain CDR2 (Wu et al., 2010a). A high-resolution crystal structure of VRC01 in complex with

a core gp120 protein suggests that VRC01 optimally approaches the CD4bs and requires minimal conformational changes in order to bind critical contact residues (Zhou et al., 2010). It is thought that the breadth of VRC01 is due to the precise targeting of the antibody to the CD4 binding site with little interaction with the variable surrounding region, however the mechanism of neutralization by VRC01 has been shown to be surprisingly distinct from a CD4-like interaction (Li et al., 2011).

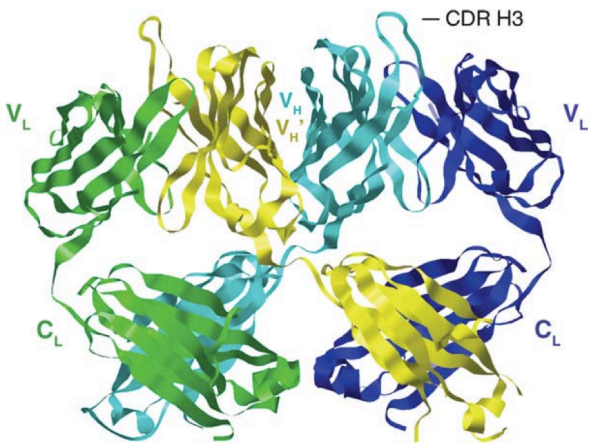


Figure 6. The unique domain-exchanged structure of bnAb 2G12. The heavy chain variable region of one antigen-binding fragment (V_H, cyan) pairs with the light chain variable region of the other antigen-binding fragment (V_L, dark blue). From Pantophlet et al., 2006.

Glycan-reactive antibodies

A dense array of N-linked glycans, collectively referred to as a glycan shield, covers much of the surface of gp120 in envelope spikes. The close spacing between carbohydrates on gp120, which is unusual for mammalian glycoproteins, is thought to impose conformational constraints on these glycans. It has been postulated that this dense and relatively rigid presentation of carbohydrates on gp120, stabilized by a network of hydrogen bonds, provides the basis for immunological discrimination of these complex glycan networks from self glycans by glycan-binding antibodies (Scanlan et al., 2007).

The bnAb 2G12 recognizes a cluster of α1 → 2-linked mannose residues on the distal ends of oligomannose sugars located on the carbohydrate-covered silent face of the gp120 outer domain (Trkola et al., 1996; Sanders et al., 2002; Scanlan et al., 2002). The crystal structure of 2G12 revealed that the antibody is able to achieve nanomolar binding

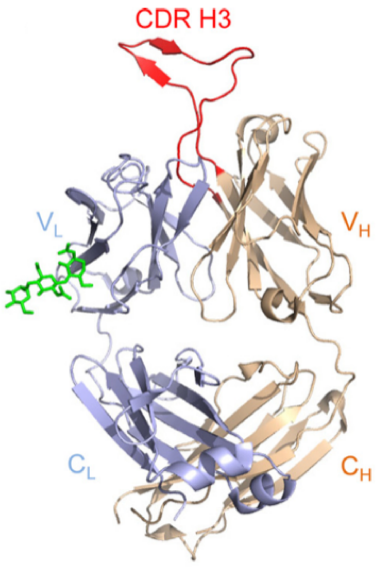


Figure 7. Crystal structure of PG16. The ax-head shaped HCDR3 (red) has unique secondary structure. From Pancera et al., 2010.

affinity to a glycan array due to the unusual configuration of the antigen binding fragments (Figure 6; from Pantophlet et al., 2006), which form a domain-swapped structure – the variable heavy chains (V_H) from each Fab arm have exchanged positions to interact with the light chain (V_L) of the neighboring antigen binding fragment (Calarese et al., 2003). The result is a multivalent platform for the binding interaction with carbohydrate, which allows the antibody to be uniquely specific for the carbohydrate pattern on the surface of gp120 (Doores et al., 2010). Co-crystallization of 2G12 with Man9 GlcNAc2 indicates that this multivalent platform allows 2G12 to interact with up to four oligomannose sugars simultaneously. Two of these glycans are located, as expected, within the conventional antigen binding sites formed by V_L and V_H, whereas the other two are bound within the newly formed secondary binding site formed by the V_H/V_H interface (Calarese et al., 2003).

In the past several years, there has been an avalanche of potently neutralizing, glycan-specific HIV antibodies. Using large-scale microneutralization screening assays, the monoclonal antibodies PG9 and PG16 were isolated from a single African donor infected with a clade A isolate (Walker et al., 2009). The neutralization profiles of both antibodies were measured to be about an order of magnitude greater than those of previously identified bNAbs, and the epitope characterization of PG9 and PG16 revealed unique and complex specificities (Walker et al., 2009; Pancera et al., 2010). They were found to interact with an epitope formed from the conserved region of the V1/V2 and V3 variable loops, to preferentially bind functional Env structure targeting a quaternary epitope, and binding was

found to be heavily dependent on glycans (Doores and Burton, 2010; Pejchal et al., 2010). In addition, PG9 and PG16, which are presumed to be clonally related, encode what are among the longest human antigen-specific HCDR3s described to date (Figure 7; from Pancera, 2010).

Following the discovery of PG9 and PG16 and using the same high-throughput microneutralization techniques, seventeen new bnAbs were identified from a group of four HIV-infected donors that exhibited exceptionally broad and potent neutralization (Pejchal et al., 2011). These antibodies, collectively named the PGT antibodies, range from 10-100 times more potent than any previously described HIV antibody and targeted novel glycan epitopes on HIV gp120. It has been speculated that the enhanced potency of the PGT family of antibodies may be partially due to their potential ability to cross-link Env trimers on the virion surface (Pejchal et al., 2011; Walker et al., 2011). The PGT antibodies can be further segregated by genetic and epitope similarities. Four antibodies, PGT141-145, are purported to target the same quaternary epitope as PG9 and PG16, are dependent on carbohydrate interactions, and encode HCDR3s that are even longer than the exceptionally long HCDR3s encoded by PG9 and PG16. Several other members of the PGT family, PGT125-128 and PGT130-131, specifically target $Man_{8/9}$ glycans on gp120 and target an epitope that is similar to that of 2G12 (Walker et al., 2011). Of notable importance, several of the PGT antibodies are of complementary neutralization breadth, such that a cocktail of PGT antibodies provides high potency and extremely broad coverage of HIV isolates (Walker et al., 2011).

MPER-specific antibodies

Although most of the surface of gp41 appears to be occluded from antibody binding on Env spikes, a region close to the viral membrane, the membrane-proximal external region (MPER), has some accessibility to the neutralizing human antibodies 2F5 and 4E10

(Muster et al., 1993; 1994; Purtscher et al., 1994; Stiegler et al., 2001; Zwick et al., 2001). Some evidence has emerged that the epitopes of these two antibodies are still accessible, and possibly more so, following CD4 binding to gp120 (Binley et al., 2003). Both antibodies are thought to recognize linear gp41 epitopes because they bind with relatively high affinity to short peptides corresponding to cognate gp41 sequences, and their neutralizing activity can be effectively inhibited by such peptides. To date, all attempts to generate neutralizing activity by immunization with sequences from these epitopes, either as peptides or incorporated into proteins, have been unsuccessful.

The structure of 2F5, which was described in complex with a 17-mer peptide derived from the core linear epitope, revealed the bound peptide epitope in a relatively extended conformation (Ofek et al., 2004). Notably, the peptide makes contact with the base of the 22 amino acid long HCDR3 region but not with much else of this loop. However, mutagenesis studies have shown that changes in the apex of the loop can reduce antibody binding to peptide and have an even more pronounced adverse effect on neutralization that alterations near the base of the loop (Zwick et al., 2004a). These results suggest not only that the long loop is required for the creation of the peptide binding site, but also that the tip of the loop may be involved in further interactions. A favored current hypothesis is that the loop may interact with the viral membrane, and some support for this view is provided by observations of enhanced 2F5 binding to MPER peptides when attached to a membrane (Ofek et al., 2004).

In contrast to 2F5, a 13-residue peptide containing the core epitope for 4E10 adopted a helical conformation in complex with antibody (Cardoso et al., 2005). The key contact residues mapped to one face of the helix and bound in a highly hydrophobic pocket in the 4E10 antibody combining site. As with 2F5, the 4E10 CDRH3 is long and apparently not directly involved in peptide binding. The epitope for 4E10 is even closer to the viral membrane than that for 2F5, with only a few residues separating the putative C terminus of

the epitope and the transmembrane domain. Therefore, the possible interaction of tryptophan residues in the CDRH3 with the viral membrane is an attractive hypothesis. This notion was supported by enhanced antibody binding to the MPER in the context of a membrane (Ofek et al., 2004). A somewhat controversial study highlighted cross-reactivity of 2F5 and 4E10 with cardiolipin (Haynes et al., 2005), although this cross-reactivity may reflect the hydrophobic nature of the combining sites of these antibodies rather than any autoimmune origin.

Although not currently published, oral presentations have revealed the discovery of 10e8, another highly potent neutralizing antibody directed against the MPER epitope. It has been reported that 10e8 is the broadest antibody known, with the ability to neutralize over 98% of tested isolates. Further, it is reported to lack the autoreactive qualities of other MPER-specific antibodies.

Other neutralizing epitopes

The third variable loop (V3) of gp120, as its name implies, contains substantial sequence divergence between different isolates of HIV. However, the crown of the loop contains a relatively conserved sequence motif GPGR or GPGQ that may be important in binding to the co-receptor (Hartley et al., 2005). The antibody 447–52D, the most broadly neutralizing antibody known to target the V3 loop region, neutralizes a range of isolates bearing the GPGR motif, although the neutralization is less broad and potent than other bnAbs (Gorny et al., 1992; Binley et al., 2004; Zolla-Pazner et al., 2004). The crystal structure of 447–52D in complex with a V3-loop peptide reveals how an antibody can evolve to recognize a motif with a conserved core but with a good deal of flanking variation (Stanfield et al., 2004). The antibody interacts specifically with the GPGR crown of the V3 loop, but the flanking sequence is bound by interaction with main-chain atoms. Perhaps in response to selection pressure by antibodies akin to 447–52D, many primary viruses appear

to have reduced accessibility of the V3 loop to the point where they are no longer recognized by such antibodies (Binley et al., 2004).

Antibodies which target the HIV co-receptor binding site are typically only able to bind in the presence of soluble CD4, indicating that the co-receptor binding pocket is at least partially occluded until ligation of the primary receptor (Salzwedel et al., 2000; Xiang et al., 2002). Antibodies to this CD4-induced (CD4i) site are typically weakly neutralizing and display breadth only within clades. Although CD4i antibodies are not potently neutralizing or especially broad, they do contain unique features that enable better understanding of HIV-host interactions. For example, the CD4i antibody 47e contains a tyrosine sulfation post-translational modification in the HCDR3 loop which allows accurate mimicry of the HIV co-receptor CCR5, which also contains a sulfated tyrosine (Choe et al., 2003; Huang et al., 2004).

CHAPTER II

RATIONALE AND EXPERIMENTAL DESIGN

History of HIV Vaccination

Since Edward Jenner's success with smallpox immunization in 1796, there have been dramatic immunization-related reductions in disease incidence for a number of viral diseases including smallpox, polio, measles, mumps, rubella, hepatitis B, and influenza. For each of these, protection has been achieved by mimicking infection with the pathogen and thereby establishing immunologic memory that can rapidly respond should an actual infection occur. This has been perhaps best achieved with the use of live attenuated virus vaccines, such as the mumps and measles vaccines, which infect the host but do not cause disease and induce strong and long-lasting immune responses. A second successful approach involves the use of killed virus vaccines, an example of which is the Salk polio vaccine, used in the first widespread polio vaccination campaigns in the mid-1900's. A third approach, referred to as a subunit vaccine, employs exposure of the immune system to recombinant viral proteins alone, as in the highly successful hepatitis B vaccine.

The goal of each of these vaccination strategies is to have an immunologic barrier in place that will prevent infection or, failing that, minimize symptoms of disease caused by virus infection. Successful vaccines have typically been generated against pathogens for which the immune response prevents serious disease in a substantial fraction of those infected. Remarkably, knowledge of how vaccine barriers function to protect against infection remains limited. What is known is largely built upon observations from animal models but, although these are widely accepted as relevant to human vaccines, direct mechanistic evidence for how any vaccine works in humans is sparse.

The main gatekeeper in most vaccination strategies is thought to be neutralizing antibody. Preexisting serum or mucosal antibody induced by an earlier infection or through vaccination can bind to free virus particles, prevent viral entry to host cells, and therefore prevent the establishment of infection. Even if some host cells are infected, antibody can bind to such cells, trigger their elimination via host effector systems, and perhaps contribute to aborting infection. Immunity generated in these types of scenarios is termed “sterilizing”, and neutralizing antibodies administered intravenously or through mucosal routes can alone provide sterilizing immunity against several viruses in experimental animal models. Realistically, for most human vaccines, antibodies are unlikely to provide sterilizing immunity. Rather, they limit the initial burst of virus replication such that it can then be contained by ongoing immune responses without substantial disease symptoms, which is the most likely mechanism for most successful vaccines.

The tremendous global success with other viral vaccines raises the question as to why HIV vaccine development has been so difficult. Many of the difficulties lie in distinct properties of this virus compared with other viruses. Foremost among these is HIV's enormous sequence diversity. Because of an error-prone reverse transcriptase, a high propensity for recombination, and an extremely rapid turnover in vivo, HIV's capacity for mutation and adaptation is enormous. Viruses even within the same HIV clade may differ by up to 20%, and in places such as Africa where there are multiple subtypes, circulating viruses can differ within the highly variable envelope protein by up to 38%. Indeed, the amount of HIV diversity within a single infected individual can exceed the variability generated over the course of a global influenza epidemic, the latter of which results in the need for a new vaccine each year. With more than 33 million people currently infected with HIV, and the need for a vaccine that simultaneously protects against all potential exposures, HIV sequence diversity alone represents a staggering challenge.

Early vaccine trials, including the much-publicized Vaxgen trial, used HIV Env-based immunogens and were designed to elicit sterilizing immunity through the induction of neutralizing antibodies. Passive transfer of neutralizing antibodies had been shown protective, so the induction of neutralizing antibodies with similar potency was seen as an attractive route to protection. These early vaccines were able to generate antibody responses against the Env protein, but these responses were non-neutralizing and thus unable to protect against infection. This was not wholly unsurprising, since the majority of anti-HIV antibodies in the serum of HIV-infected individuals are directed against non-neutralizing epitopes or epitopes that are not present on the mature Env trimer.

In part because of the apparent difficulties in eliciting neutralizing antibodies through vaccination, the field switched its focus away from vaccines that induce sterilizing immunity toward those that control viral load after infection and thus reduce secondary transmission. This shift was prompted by data showing that T cell–mediated immunity was critical for resistance to immunodeficiency viruses. For example, depletion of $CD8^+$ cells in simian immunodeficiency virus (SIV)-infected macaques led to a resurgence of viral load, and in HIV-infected individuals known as “elite controllers,” the control of viral load was associated with potent and broad cellular immune responses. Unfortunately, the STEP HIV vaccine trial, which was designed to induce a strong CD8+ response through the use of an adenovirus (Ad5) vector, failed to protect Ad5-seronegative individuals from HIV infection and may have enhanced the likelihood of infection in individuals with previous adenovirus exposure.

Rationale

The failure of the STEP trial, which tested the viability of inducing $CD8^+$ responses to HIV, has renewed interest in the development of vaccines and immunogens designed to

elicit broadly neutralizing antibody responses. The failure of the early HIV vaccines, however, has taught us valuable lessons about the importance of careful immunogen design in order to focus the immune response on epitopes of interest and to minimize non-neutralizing antibody responses. Over the past several years, intense effort has been expended in the search for HIV bnAbs, with the goal of more fully understanding neutralizing epitopes and the mechanisms of neutralization. At the beginning of my thesis research, after studying the moderate successes and nearly complete failures of various HIV vaccine trials, I thought that a completely new approach to vaccine development might be necessary. This new approach, instead of relying on mimicry of viral proteins or viral epitopes, would require immunogens designed to rationally guide a naïve antibody, specifically the unmutated progenitor of an HIV bnAb, toward the affinity matured, broadly neutralizing antibody. There are several immediate objections to this sort of vaccination strategy, each of which, if correct, would render such a vaccination scheme close to impossible. First, detailed knowledge of the HIV-uninfected antibody repertoire would be necessary, and a massive database of antibody sequences from several individuals would be required to accumulate a sufficiently detailed understanding of the HIV-uninfected antibody repertoire. A strategy designed to rationally mutate a germline antibody sequence toward a broadly neutralizing antibody sequence must, by definition, understand the genetic and structural detail of both the germline and the broadly neutralizing antibody. Therefore, a deep and comprehensive understanding of the HIV-uninfected repertoire is paramount. Second, HIV bnAbs often contain genetic or structural features that are extremely uncommon. These features, including extensive somatic hypermutation, SHA indels and long HCDR3 loops, may be generated only after years or decades of constant antigen exposure, which accompanies chronic HIV infection. If these genetic features, which are critical to the neutralizing potency of most bnAbs, require antigen exposure over a long time period, it is unlikely that a brief vaccination regimen will be able to effectively elicit the required bnAbs. Finally, since very

few bnAbs have been discovered and since broadly neutralizing serum activity is rare even in HIV-infected individuals, it has been postulated that most people are incapable of generating the unique genetic and structural features that are commonly found in bnAbs. Under this assumption, the rarity of HIV bnAbs is explained primarily by the rarity of individuals capable of generating them and a vaccine designed to induce such antibodies would only be effective in a small subset of the population.

The design of an immunogen the effectively elicits bnAbs is a monumental task, and one that would extend far past the time constraints of a PhD thesis. My thesis work focused on the fist step of such a process: the determination of whether or not the three objections discussed above are valid.

Experimental Design

High throughput antibody repertoire sequencing

To effectively evaluate the validity of the three objections detailed previously, an extremely detailed analysis of the human antibody repertoire is required. During my studies, advances in high throughput sequencing technology made such analyses possible. Specifically, the 454 sequencing platform, which enables the simultaneous sequencing of up to 1 million DNA fragments, released the equipment and reagents necessary for read lengths of up to 500 bases per fragment. Although this read length is not comparable to Sanger sequencing, for which 1200 base reads are often possible, it is sufficient for complete coverage of an antibody heavy chain variable region, which is approximately 400bp in length. Although sequencing advances made antibody repertoire sequencing possible, two major hurdles remain: quantitative amplification of a diverse repertoire of antibody sequences, and bioinformatics analysis of the sequencing results.

Quantitative primer design

In order to perform a valid quantitative analysis of the human antibody repertoire, it is imperative that the initial PCR amplification of the antibody variable regions be accomplished with as little primer-induced bias as possible. This is no small feat: there are 55 functional variable genes and 6 joining genes, which makes multiplexed primer design extremely complex. Further, somatic hypermutation within the primer-annealing region would skew amplification away from a subset of somatically mutated sequences and provide a misleading picture of the overall antibody repertoire. In order to identify primers that produce as little bias as possible, I sampled several dozen multiplexed sets of primers for evidence of primer bias. Through this exhaustive sampling process, I identified a single set of primers for which the identified germline gene repertoire of the amplification product consistently matched that of individually amplified, single sorted B cells, the gold standard of repertoire analysis. Further, the variable gene primer-annealing region is in framework region 1 (FR1), which is highly conserved among variable gene families and is infrequently the target of somatic hypermutation.

Development of a sequence analysis pipeline

After resolution of the primer bias issue, another technological hurdle still remains. Although high throughput sequencing is not new and although many tools exist to analyze and process the resulting mountain of data, there are no antibody-specific analysis tools capable of handling the enormous amount of data generated by even a single high throughput sequencing run. A flexible database structure and schema must also be developed, preferably using industry-standard database software, which will enable rapid identification of sequences that fit desired search criteria. Several software packages exist which allow identification of putative germline genes from an antibody sequence, as well as a varying amount of somatic mutation analysis. One of the gold-standard tools of the

antibody field, the IMGT-V/QUEST webserver, allows batch upload of a maximum of 50 sequences. This is obviously not feasible when the sequence pool to be analyzed contains hundreds of thousands of sequences. I was afforded the opportunity, however, to be among a group of researchers selected by IMGT to beta test a new version of the webserver, HighV-QUEST, which was specifically designed to accommodate high throughput antibody sequencing and allows batch upload of up to 150,000 sequences at a time. In total, the HighV-QUEST webserver has enough computational power to analyze approximately 8 million antibody sequences in 24 hours. I have also created a custom relational database using MySQL database software. MySQL is an open-source implementation of the industry-standard Structured Query Language (SQL), a programming language designed specifically for managing data in relational database systems. The database is designed so that the output from the HighV-QUEST webserver can be directly imported into the database and rapidly queried. Although the learning curve for programming in MySQL is steep, it offers extraordinary speed and flexibility when handling very large relational data sets.

Selection of bnAbs for analysis

In order to evaluate the presence and frequency of germline precursors of bnAbs in the HIV-uninfected repertoire, one or more bnAbs must first be selected for analysis. Selecting the proper bnAb for which to search for germline precursors is critical, because some of the bnAbs that have been discovered may be very difficult or impossible to induce in a large segment of the population. Others may be difficult to generate through a vaccination regimen of feasible duration. I have previously shown that bnAbs typically have either extensive somatic hypermutation or encode a long HCDR3 loop. Antibody sequences with extensive somatic hypermutation are, by definition, quite distinct from any possible germline precursor. The feasibility of generating such extensive somatic hypermutation through vaccination has not been shown; it is likely to be very difficult and will potentially

involve more than simply immunogen design. It has been speculated that, through the use of yet-to-be-determined adjuvants, the rate of somatic hypermutation may be altered such that a large number of somatic mutations can be rapidly induced with a small number of inoculations. This is yet to be shown, however, and even a temporary alteration in the rate of somatic hypermutation may have substantial unintended effects.

On the other hand, antibodies with long HCDR3s may already be present in the naïve, or unmutated, B cell population. Selection of pre-existing antibodies encoding long HCDR3 loops from the germline repertoire would, in theory, only require design of the proper immunogen for selection of bnAb precursors. For this reason, we selected two somatically-related bnAbs, PG9 and PG16, which use exceptionally long HCDR3s as the primary mechanism of binding and neutralization. These antibodies are not extensively somatically mutated and most residues that are critical to binding and neutralization are encoded in the germline sequence, raising the intriguing possibility that germline precursors to these bnAbs, if they are found to be present in HIV-uninfected individuals, may have some level of binding and/or neutralization capacity even without somatic hypermutation. Thus, elicitation of PG-like neutralizing antibodies may require only minimal somatic hypermutation.

CHAPTER III

GENETIC ANALYSIS OF THE HUMAN ANTIBODY REPERTOIRE IN PERIPHERAL BLOOD AND MUCOSAL AND LYMPHOID TISSUES

Introduction

A diverse human antibody repertoire is a key element of the acquired immune response and is critical to the effective prevention and clearance of microbial infections (Crotty and Ahmed, 2004). Vast diversity in the antibody repertoire is generated initially through a process of combinatorial rearrangement in which variable (V), diversity (D), and joining (J) gene segments are assembled into a complete immunoglobulin sequence (Tonegawa, 1983; Schatz et al., 1989). This initial diversity is increased through the use of antigen-driven somatic hypermutation and class-switch recombination (Neuberger and Milstein, 1995; Wilson et al., 1998a; Neuberger, 2008; Schroeder and Cavacini, 2010). These affinity maturation processes result in the creation of distinct memory populations that contain only antigen-experienced B cells (Jackson et al., 2008).

Previous analysis of the antibody gene repertoire in peripheral blood B cell subsets did not detect significant differences in germline V, D or J gene segment usage between the naïve and memory populations (Tian et al., 2007). This finding was somewhat surprising, since it was expected that the memory subset might contain an altered germline gene repertoire that was biased by antigen selection. Although more narrowly focused work was previously able to identify differential J_H gene use in memory subsets (Rosner et al., 2001), this study only analyzed the fraction of sequences that contain the V_H6 gene segment and not the total repertoire. A study using a larger sequence pool was able to identify differences in both J_H gene and V_H gene family use in naïve and memory subsets, with the memory

population displaying an increase in J_H4 gene use, a corresponding decrease in J_H6 gene use, and differential use of several V_H gene families (Wu et al., 2010b). During my studies, there were several studies that leveraged high throughput amplicon sequencing to perform in-depth analysis of human antibody repertoires (Boyd et al., 2009; 2010; Wu et al., 2010b; Arnaout et al., 2011; Prabakaran et al., 2011). These previous studies have been limited in scope, however, due to analysis of the use of V, D or J genes in isolation. In the peripheral blood antibody repertoire, individual V, D and J genes are not expressed in isolation, but are linked by recombination. Thus, it is imperative to study gene segment usage not only by individual gene segment use, but also in the context of complete V(D)J pairings to gain a more complete understanding of the antibody repertoire. Here, I present a thorough examination of V(D)J recombinants in the human peripheral blood repertoire. The studies reveal stark repertoire differences between circulating B cell subsets and provide genetic evidence for global control of repertoire diversity in both naïve and memory subsets.

Although repertoire differences between the peripheral blood B cell subsets have been identified, the overall peripheral blood repertoires were more similar than expected. Since each individual has experienced a unique progression of pathogenic encounters and since these pathogenic encounters have been shown to elicit antibody responses with skewed repertoires (Tian et al., 2007; 2008), the resulting memory B cell subsets would be expected to be equally unique in each individual. This raises the possibility that many antigen-specific cells reside in locations other than the peripheral blood. From an immunological perspective, this makes sense; having the appropriate memory responses located at the most common sites of pathogen encounter would aid in the rapid generation of a robust antibody response.

The antibody repertoires of mucosal tissues have been previously shown to consist predominantly of class-switched B cells and that these B cells are antigen-specific (Benckert, 2011). In general, IgA is the most abundant antibody isotype at mucosal

surfaces, although IgG and IgD make up a substantial portion of mucosal secretions (10, 18, 19). In addition, IgE is measurable in several mucosal secretions when allergy is present (37). Although the isotype composition of mucosal tissues is fairly well studied, little is known about the genetic repertoire composition of the tissue antibody repertoires.

Genetic Analysis of the Human Healthy Donor Antibody Repertoire

With assistance from the Vanderbilt Flow Cytometry core, we separately isolated naïve, IgM memory and IgG memory B cells from four healthy individuals using flow cytometric sorting, extracted mRNA and performed RT-PCR to amplify antibody genes from those cells, and subjected the resulting amplicons to high throughput DNA sequencing. The primers were selected for their ability to produce accurate, reproducible amplification of both naïve and mutated antibody repertoires (Boyd et al., 2009; 2010), and the variable gene use of our repertoire closely matched repertoire analysis in which amplification was performed on single B cells (Tian et al., 2007; 2008). After selecting only high-quality antibody sequences, we obtained a total of 294,232 naïve cell sequences, 161,313 IgM memory cell sequences and 94,841 IgG memory cell sequences.

I analyzed the V(D)J recombinant repertoire in each of the three B cell subsets, and created Circos plots showing the relative prominence of each V(D)J recombination within the repertoire of each cell subset (Figure 8A-C). These plots revealed a large number of trends that were apparent only when analyzing the repertoire in the context of complete V(D)J recombinations. In this chapter, I will focus on three of the most prominent trends.

First, virtually all of the major V_H-J_H pairings (identified by colored ribbons in Figure 8) follow a similar pattern: increased use of V_H-J_H pairings that contained heavy chain joining gene 4 (J_H4) and decreased use of pairings that contained J_H5 or J_H6 in both memory subsets, as compared to naïve cells. Use of J_H4 has been shown previously to be increased

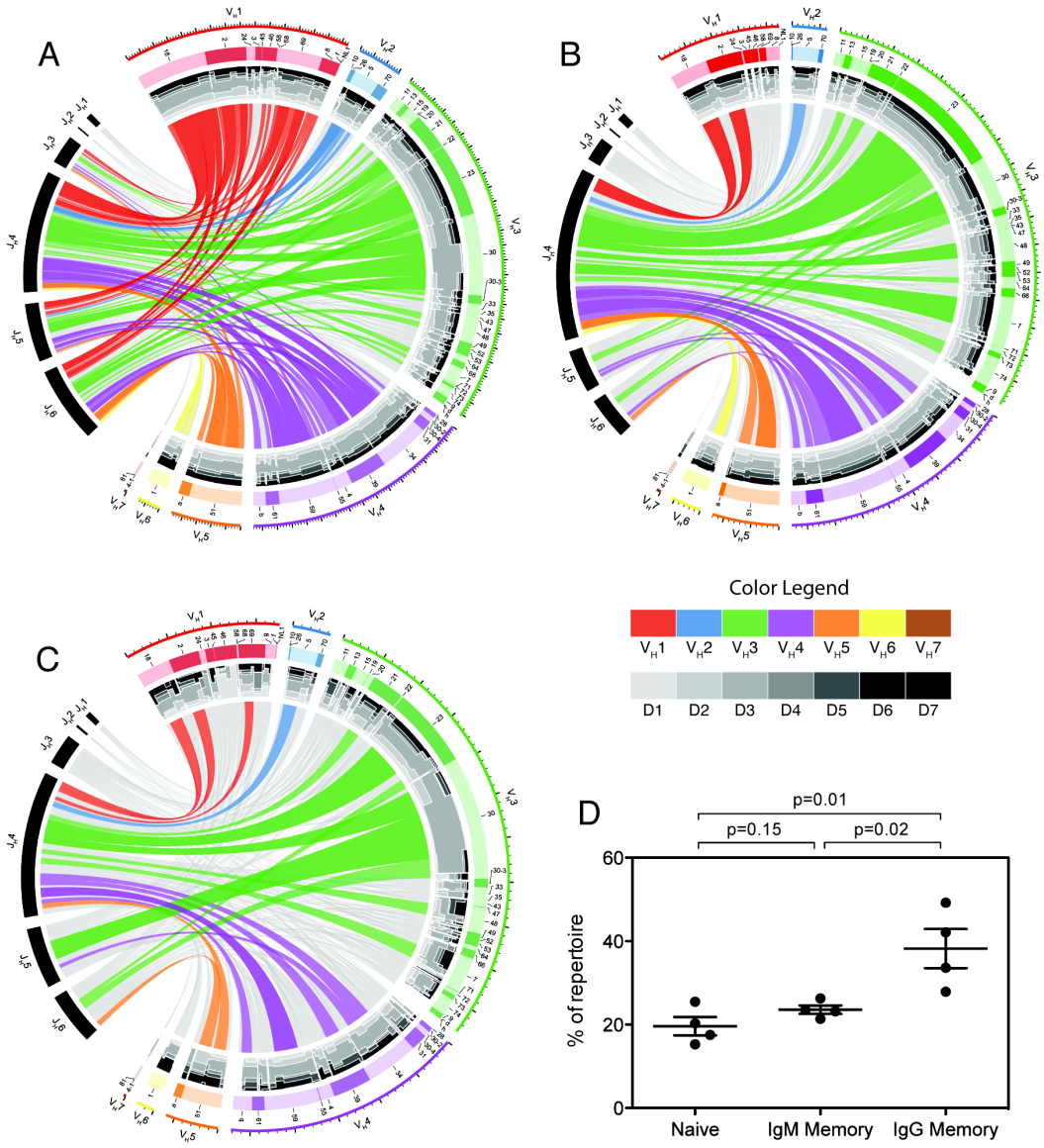


Figure 8. Circos diagrams display linked V_H, D, and J_H gene use in naïve, IgM memory or IgG memory B cell subsets. Circos plots were generated for (A) naïve, (B) IgM memory, or (C) IgG memory repertoires. On the right half of each plot (colored segments), each heavy chain variable gene family is shown in a separate color. The arc length of each segment corresponds the relative frequency of each heavy chain variable (VH) gene family in the identified B cell subset. The smaller ticks on the outer ring represent 1,000 sequences, with major ticks marking 5,000 sequences. The left half of each plot (black segments) shows each heavy chain joining (JH) gene segment. The arc length of each segment corresponds to the frequency of JH gene use in combination with the given VH gene within the identified subset. The scale of arc lengths on the JH gene side of the diagram is 1/4 that of the arcs on the VH gene side. Just inside the tick ring on the VH gene side of the diagram, segmented arcs identify the individual VH genes within each VH gene family. Within each plot, pairing of JH genes to VH gene segments is represented by colored ribbons. The color of the ribbon indicates VH gene use; increased ribbon width and color intensity corresponds to increased frequency of the represented VH-JH pairing. Just outside the VH-JH links on the VH gene side of the diagram, a stacked histogram indicates D gene use for each VH-JH pairing. Diversity gene families D1, D2, D3, D4, D5, D6 and D7 are plotted in increasingly darker shades of grey (D1 is closest to the center in lightest grey; D7 is furthest from the center in darkest grey). (D) The contribution of the fifty most common V(D)J recombinations for each subset is plotted. Bars indicate mean ± SEM. The p values for pairwise comparisons were determined using a two-tailed Student's T test.

in memory subsets (Wu et al., 2010b), but the consistency with which the broad spectrum of V_H-J_H pairings exhibited increased J_H4 use is surprising. Second, use of diversity gene family 3 (D3) was increased dramatically in recombinations that used heavy chain variable gene 3-30 (V_H3-30) and either J_H4 or J_H5 in both naïve and IgG memory subsets. In the IgM memory cell subset, however, diversity gene use in recombinations that used V_H3-30 and J_H4 or J_H5 was much lower than in the naïve or IgG memory subsets and was comparable to that of other genes in the V_H3 family. These data support emerging evidence that the IgM memory repertoire is genetically distinct from the IgG memory repertoire and that this difference is likely the result of different stimuli (Tian et al., 2007; Wu et al., 2010b). Finally, the three Circos plots reveal the increased oligoclonality of both memory subsets compared to naïve. The colored ribbons in each plot represent V_H-J_H pairings that comprise at least 1% of the total subset repertoire. In the naïve subset, there are 66 different V_H-J_H pairings that each account for at least 1% of the total naïve repertoire. In the IgM memory subset, only 27 different V_H-J_H pairings exceed 1% of the total subset repertoire, and only 19 V_H-J_H pairings exceed 1% in the IgG memory subset. These data indicate that the memory subsets become increasingly oligoclonal, with a small selection of V_H-J_H pairings comprising a larger fraction of the total subset repertoire.

Further analysis of the variable gene use in each of the subsets revealed contrasts with recently published work on the antibody repertoire in human cord blood (Prabakaran et al., 2011). In the cord blood repertoire, V_H1-2 was found to be the most commonly used germline gene. In the naïve, IgM memory and IgG memory subsets, not only was V_H1-2 not the most commonly used variable gene, it was not even the most commonly used V_H1 family gene. Both V_H1-18 and V_H1-69 were more frequently used in the naïve and IgG memory populations and V_H1-18 was used more frequently in the IgM memory population. In all three peripheral blood subsets, either V_H3-23 (naïve and IgM

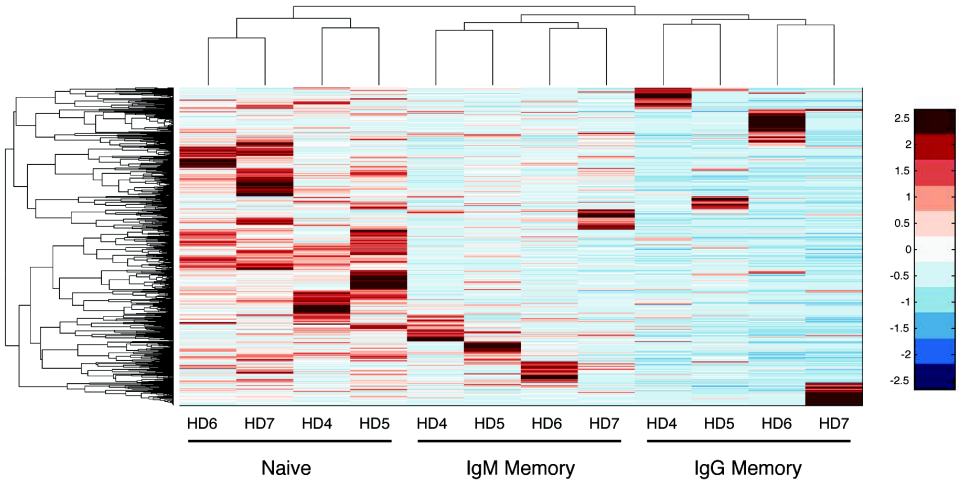


Figure 9. Clustergram of V(D)J recombinants reveals global control of expressed antibody repertoires in B cell subsets. The frequency of each V(D)J recombination was determined for the naïve, IgM memory or IgG memory subsets for four healthy donors, and a clustergram was created. V(D)J recombinants were clustered by relative frequency in each donor and subset, and the resulting phylogenetic tree is shown on the left. Results for three B cell subsets from each donor were clustered by the V(D)J repertoire, and the resulting phylogenetic tree is shown at the top. The frequency variation for each V(D)J recombination across all subsets was determined, and standardized to a range of -3 to 3. The colorimetric scale for the heat map is shown.

memory) or V_H3-30 (IgG memory) was the most commonly used variable gene, which is consistent with previous data (Boyd et al., 2009; 2010; Wu et al., 2010b; Arnaout et al., 2011).

The oligoclonality of each subset was further analyzed by determining the contribution of the 50 most commonly used V(D)J recombinations to the total repertoire of each subset (Figure 8D). In the IgG memory subset, which was most oligoclonal, the top 50 V(D)J recombinations accounted for 38.2% of the subset repertoire. Conversely, in the naïve and IgM memory subsets, which were significantly less oligoclonal than the IgG memory subset, the top 50 V(D)J recombinations accounted for only 19.6% (naïve, p=0.01) or 23.6% (IgM memory, p=0.02) of the total subset repertoire. Additionally, there was a trend toward reduced oligoclonality in the naïve subset compared to IgM memory that fell short of statistical significance (p=0.15).

With assistance from Brett McKinney, we next performed a clustering analysis on the V(D)J recombinant repertoire for each donor and subset (Figure 9). We found patterns that were robust to different clustering metrics and linkage types. Specifically, Brett performed hierarchical bi-clustering with the Euclidean and Pearson correlation metric in combination with single linkage, complete linkage and average linkage. In all scenarios, the samples clustered in the categories shown in Figure 9. Prior to clustering a variance filter was used to remove genes with very low variation across subjects. Interestingly, repertoires of the same subset from different donors (inter-donor subsets) were more closely related than different subset repertoires from the same donor (intra-donor subsets). In fact, phylogenetic clustering of the subset repertoires of all four donors revealed clustering exclusively among inter-donor subsets, with no observed intra-donor subset clustering. This finding was unexpected, since each donor has experienced a unique set and sequence of pathogen encounters, and each donor might be expected to generate unique memory repertoires appropriately skewed by prior histories of infection. Also of note, donor pairs were consistently clustered together, regardless of subset. Donors HD4 and HD5 clustered most closely in each of the three subsets, as did donors HD6 and HD7. In addition to the inter-clonal subset clustering, the tight groupings of highly over-represented V(D)J recombinations within each individual memory repertoire provide further evidence of the oligoclonality of the memory subsets. The unique V(D)J recombinations that comprise the tight groupings are not shared between like subsets of different donors and, surprisingly, similar groupings are not present in the naïve subset from the same donors. This finding indicates that while the frequency of germline gene family use may appear similar between naïve and memory populations when the V_H , D or J_H families are analyzed in isolation, deeper analysis of the V(D)J recombinant repertoires of these subsets uncovers stark repertoire differences.

The substantial differences in each subset repertoire at the individual V(D)J recombinant level, coupled with the overarching similarities of the repertoires at the germline gene family level, present something of a paradox. These data seem to suggest the presence of a broad, global mechanism for repertoire regulation at the germline gene family level. While there is no direct evidence of such a mechanism, several recent studies indicate the presence of repertoire-based regulatory mechanisms in circulating B cell subsets. Despite the tendency of pathogen-specific antibody responses to exhibit biased germline gene repertoires (Weitkamp et al., 2005; Tian et al., 2008; Gorny et al., 2009), the frequency of gene family use in naïve and memory subsets is remarkably consistent across individuals (Tian et al., 2007; Wu et al., 2010b). Further, alteration of this gene family homeostasis in the circulating B cell repertoire is associated with disease states (David et al., 1995; 1996; Zouali, 1996; Scamurra et al., 2000). In recent work, long-lived plasma cells were shown to contain significantly fewer autoreactive B cells than the circulating IgG memory subset, and this difference was attributed to differential repertoire regulation within the two subsets (Scheid et al., 2011). Thus, while no mechanism for regulation of circulating human antibody repertoires has been identified, mounting indirect evidence, including the data presented in this report, suggests the presence of such regulation.

Differences Between Peripheral Blood and Mucosal and Lymphoid Tissue Antibody Repertoires

Antibody variable gene use

I obtained total RNA from a variety of tissue samples derived from a pool of donors: peripheral blood leukocytes, bone marrow, small intestine, lung, stomach, lymph node, tonsil, spleen and thymus. The expressed antibody variable genes were amplified using RT-PCR, and the resulting amplicons were subjected to high-throughput DNA sequencing. After

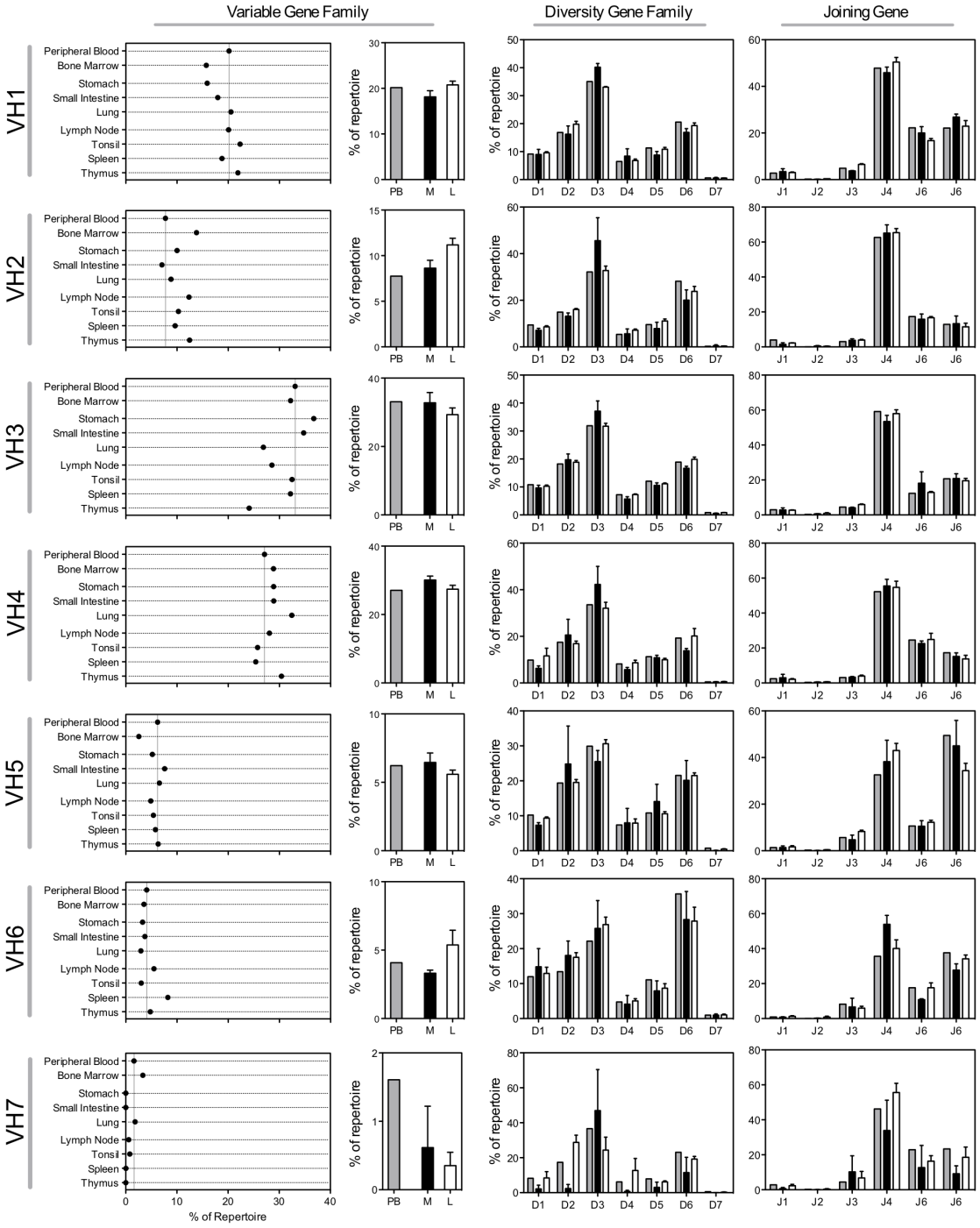


Figure 10. Antibody variable gene family usage in peripheral blood or tissues. Antibody variable gene family usage is shown for each sample as a percentage of the total sample antibody repertoire. VH3 was the most common variable gene family used in six of the eight samples. VH7 was the least common variable gene family used in seven of the eight samples. To facilitate comparison between tissues and peripheral blood, a vertical grey line indicates the peripheral blood frequency for each variable gene family.

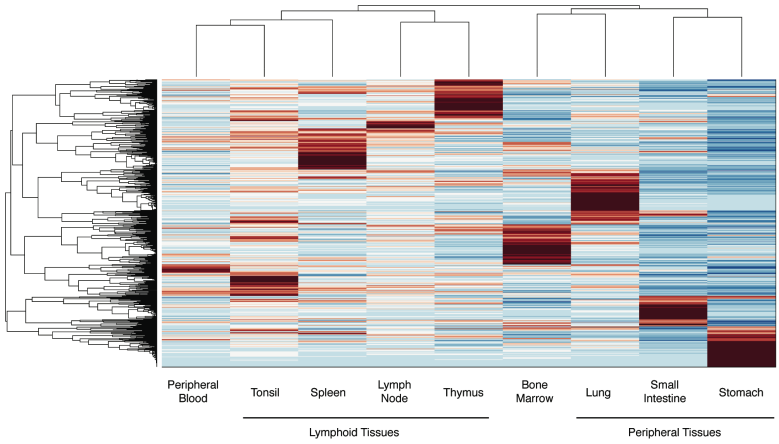


Figure 11. Clustergram of antibody repertoires. The frequency of each V(D)J recombination was determined for each of nine tissues, and a clustergram was created. V(D)J recombinants were clustered by relative frequency in each tissue-specific repertoire, and the resulting phylogenetic tree is shown on the left. Tissue-specific repertoires were clustered by the overall V(D)J usage of each repertoire, and the resulting phylogenetic tree is shown at the top. The frequency variation for each V(D)J recombination across all tissue-specific repertoires was determined, and standardized to a range of -3 to 3. The colorimetric scale for the heat map is shown.

selecting only high-quality, non-redundant antibody sequences, I obtained a total of 1,412,943 sequences.

I first determined the frequency of use of each antibody variable gene family in each tissue (Figure 10), and discovered substantial differences in the gene family use in tissues compared to that in peripheral blood. A large number of differences were noted; I focus here on three of the most prominent trends. First, heavy chain variable gene family 2 (V_H2) use was increased in every tissue except for small intestine. V_H2 was found in 7.7% of peripheral blood sequences, and the largest increases were found in bone marrow (13.8%), thymus (12.5%) and lymph node (12.3%). Second, while V_H3 was the most common variable gene family in most of the samples, the lung and thymus samples used the V_H4 family more frequently than V_H3. Finally, repertoires in mucosal sites (stomach, lung, small intestine) differed significantly from that in lymphoid tissues (lymph node, spleen, tonsil, thymus). Each

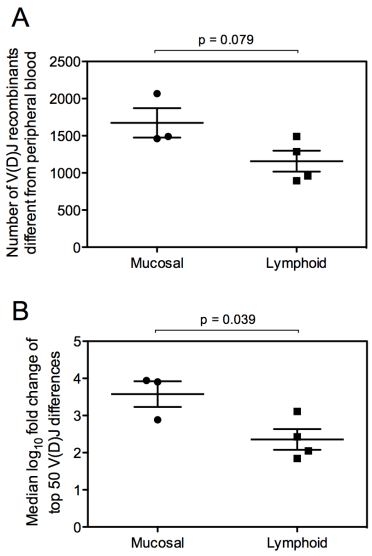


Figure 12. Comparison of V(D)J use in lymphoid and mucosal tissues to peripheral blood. (A) The frequency of each V(D)J recombination was calculated for each tissue and compared to peripheral blood. (B) The frequency of each V(D)J recombination was determined for each tissue and compared to peripheral blood. The $\log_{10}$ fold change of the 50 most different V(D)J recombinations is shown for each tissue.

of the four lymphoid tissue samples showed reduced use of the V_H3 family compared to peripheral blood, while two of the three mucosal tissues showed increased use of V_H3 family genes. Similarly, each of the three mucosal samples showed reduced V_H6 gene family use compared to peripheral blood, but three of the four lymphoid tissue samples used V_H6 more frequently than peripheral blood. Analysis of the diversity (D) gene family and heavy chain joining (J_H) gene use within each variable gene family produced similar trends for both the mucosal and lymphoid groups.

Differences in the recombined $V_H(D)J_H$ repertoire

I next performed a clustering analysis on the recombined $V_H(D)J_H$ repertoire for each tissue sample (Figure 11). Interestingly, the repertoire in lymphoid tissues (tonsil, spleen, lymph node and thymus) clustered with each other and with peripheral blood. Repertoires from mucosal tissues (lung, small intestine, stomach) also clustered together, along with bone marrow. This finding indicates that the recombined $V_H(D)J_H$ repertoires of mucosal tissues differ substantially from both peripheral blood and lymphoid tissues.

To more closely investigate these repertoire differences, I first determined the germline gene use in each recombined $V_H(D)J_H$ sequence in each tissue by calculating the frequency of each $V_H(D)J_H$ combination. I then compared these frequencies to that in the

pooled peripheral blood sample (Figure 12A). I found a trend toward more differences between mucosal tissues and peripheral blood than were present between lymphoid tissues and blood ($p=0.079$). I also analyzed the magnitude of the top 50 differences from peripheral blood for each of the mucosal and lymphoid samples (Figure 12B), and found that differences in $V_H(D)J_H$ frequency between mucosal tissue samples and peripheral blood were significantly larger than differences between lymphoid tissue samples and peripheral blood ($p=0.039$).

Mutation frequency analysis

Sequences from each tissue subset were grouped by mutation frequency, and a mutation histogram was created for each tissue sample (Figure 13A). The peripheral blood sample contained a large number of sequences with few or no mutations, which is consistent with the previous reports that the peripheral blood B cell compartment contains a large proportion of naïve cells (Wu et al., 2010b; Briney et al., 2012b). In contrast, the bone marrow sample contained very few sequences with few or no mutations, which is somewhat surprising, since bone marrow contains many progenitor and precursor B cells, which are presumably unmutated. The absence of unmutated sequences is likely due to the use of mRNA as template for the antibody gene amplification. Since bone marrow resident long-lived plasma cells transcribe the antibody gene at a much higher rate than immature B cells (Smith et al., 2009), it is likely that oversampling of plasma cell expressed sequences in the amplified mRNA skewed the bone marrow sequence repertoire toward highly mutated sequences.

All three mucosal tissue samples (small intestine, stomach and lung) showed a dramatic paucity of sequences with few or no mutations, indicating that naïve B cells are less frequent in tissue. Interestingly, each of the lymphoid tissues (lymph node, tonsil, spleen and thymus) contained a higher frequency of antibody genes with few or no

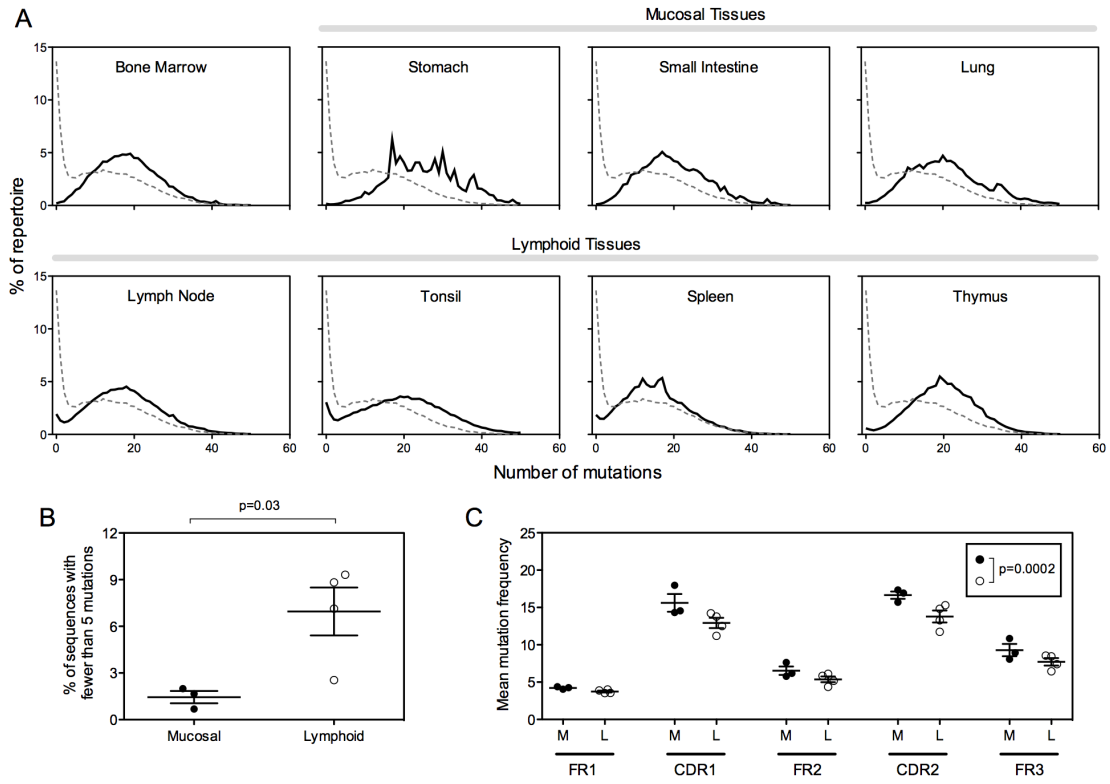


Figure 13. Mutation frequency for peripheral blood, bone marrow and mucosal and lymphoid tissues. Mutation histograms are shown for each sample. Sequences for each tissue sample were grouped by the number of somatic mutations, and the frequency of each mutation count was determined. To ease comparisons between tissue samples and peripheral blood, the mutation histogram for peripheral blood is shown as a dashed line on each plot.

mutations (7.0%) than mucosal tissues (1.4%; p=0.03), but a lower frequency than peripheral blood (30.6%; Figure 13B). A more detailed breakdown of mutation frequency by antibody gene region (Figure 13C) showed a reduction in mutation frequency in lymphoid tissue repertoires across all framework regions (FRs) and complementarity determining regions (CDRs) when compared to mucosal tissue repertoires (p=0.0002).

Mucosal tissue repertoires encode longer HCDR3s than lymphoid tissue repertoires

Sequences from each tissue repertoire were grouped by HCDR3 length and the frequency of each HCDR3 length group was determined (Figure 14A). To ease comparisons, the HCDR3 length histogram for the peripheral blood repertoire is displayed

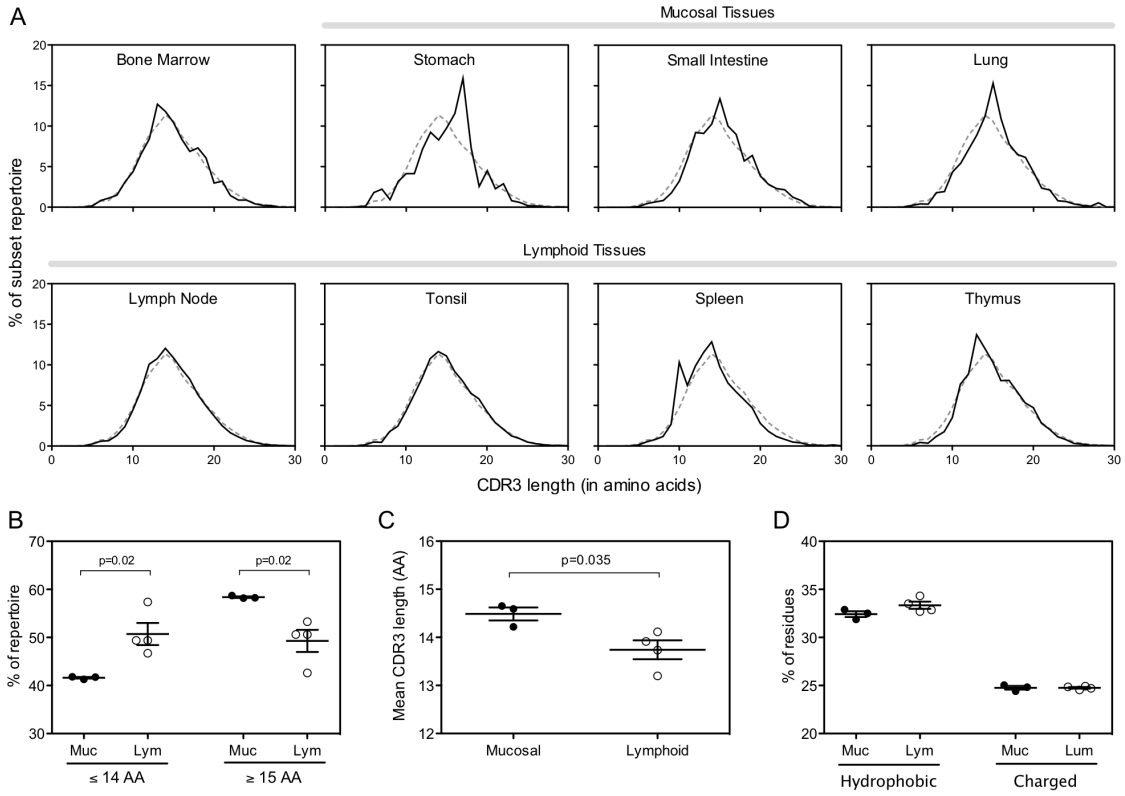


Figure 14. Mucosal tissue repertoires encode longer CDR3s and are more mutated than lymphoid tissue repertoires. (A) The mean CDR3 length was determined for each repertoire. Bars indicate mean ± SEM. (B) The mean mutation frequency was determine for each repertoire. Bars indicate mean ± SEM.

alongside each tissue HCDR3 histogram. Tissue repertoires then were divided into two groups based on HCDR3 length: short HCDR3s (≤14 amino acids) and long HCDR3s (≥15 amino acids) (Figure 14B). The repertoires of lymphoid tissues were split approximately evenly between short (50.7%) and long (49.3%) HCDR3 lengths. In contrast, repertoires from mucosal tissues contained a significantly higher frequency of long HCDR3s (58.4%) and a significantly lower frequency of short HCDR3s (41.6%) than lymphoid repertoires (p=0.02). Further, the overall mean HCDR3 length of the mucosal tissue repertoires was significantly longer than the overall mean HCDR3 length of lymphoid tissue repertoires (Figure 14C; p=0.035). This finding was surprising since mucosal repertoires contain a

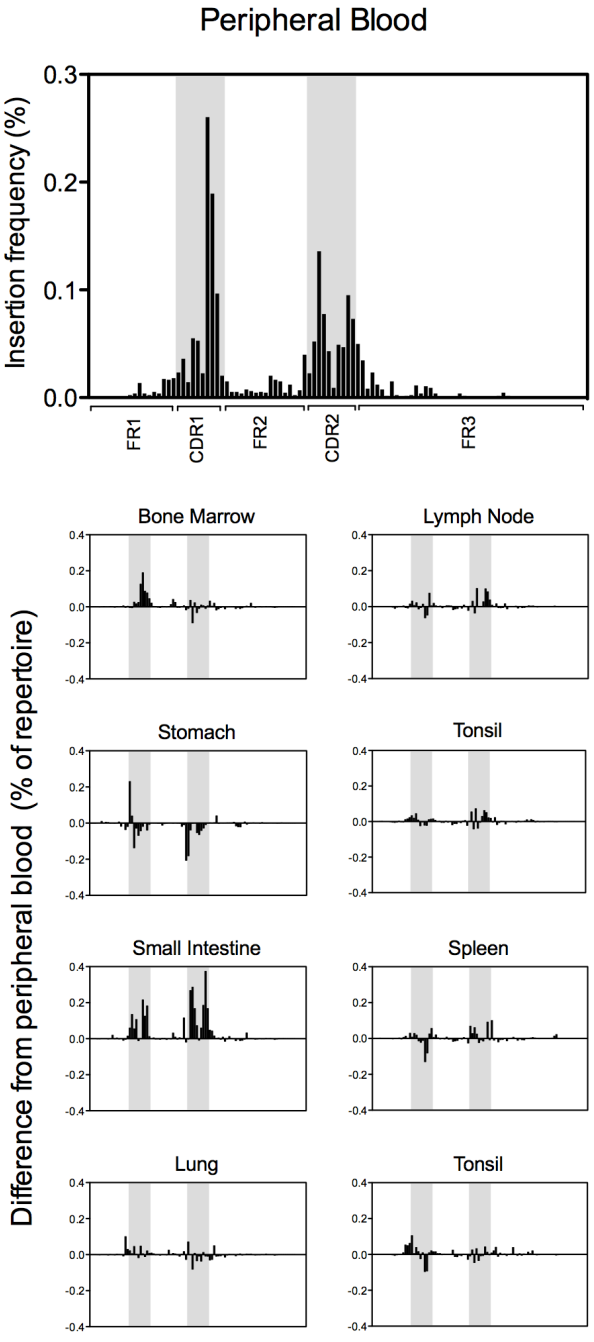


Figure 15. Frequency and position of DNA fragments encoding non-frameshift insertions. (A) The presence and frequency of non-frameshift insertions is shown in peripheral blood. The location of CDR1 and CDR2 are highlighted in grey. (B) The difference in insertion frequency when compared to peripheral blood is shown for each tissue. As before, CDR1 and CDR2 are highlighted in grey.

higher fraction of highly mutated sequences than lymphoid repertoires (Figure 13), but there are reports that highly mutated memory B cell subsets may encode shorter HCDR3s than the unmutated naïve subset (Wu et al., 2010b). Since long HCDR3s tend to have a lower frequency of hydrophobic and charged residues (Wu et al., 2010b), we determined the frequency of both hydrophobic and charged HCDR3 residues in mucosal and lymphoid repertoires (Figure 14D). While there was a trend toward reduced frequency of hydrophobic HCDR3 residues in the mucosal tissue repertoires compared to lymphoid repertoires (p=0.13), there was no difference in the frequency of charged HCDR3 residues in mucosal repertoires (23.4%) compared to lymphoid repertoires (23.2%; p=0.61).

Somatic hypermutation-associated insertions and deletions

Short nucleotide insertions or deletions (abbreviated: indels) are associated with the somatic hypermutation process (Wilson et al., 1998b), and antibodies encoding these somatic hypermutation-associated indels (SHA indels) have been shown critical to the immune response against pathogens that initiate infection at mucosal surfaces (Wu et al., 2010a; Krause et al., 2011; Walker et al., 2011). The heavy chain sequences for antibodies from all tissue repertoires were analyzed for the presence of codon-length nucleotide indels in the antibody variable gene region. Sequences from the peripheral blood repertoire containing SHA indels were analyzed further to determine the location of each indel, and the frequency of insertions (Figure 15) or deletions (Figure 7) at each codon position of the variable gene

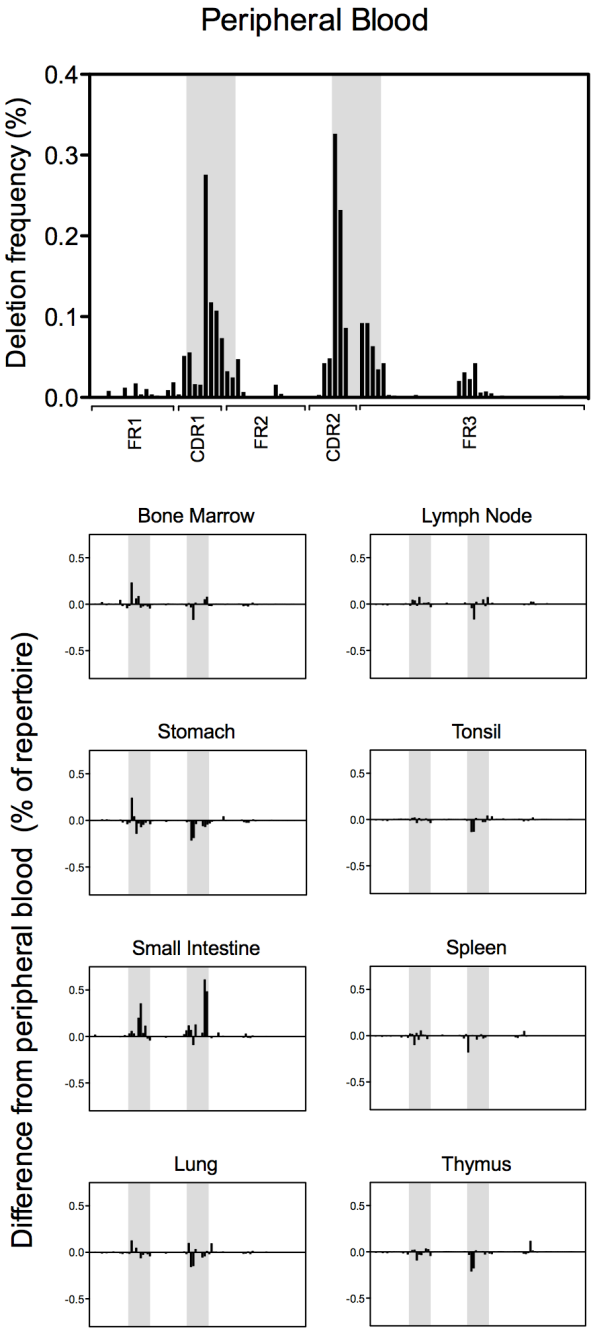


Figure 16. Frequency and position of DNA fragments encoding non-frameshift deletions. (A) The presence and frequency of non-frameshift deletions is shown in peripheral blood. The frequency is plotted as the percent of sequences in the repertoire displaying deletions at each codon position. The location of CDR1 and CDR2 are highlighted in grey. (B) The difference in deletion frequency when compared to peripheral blood is shown for each tissue. As before, CDR1 and CDR2 are highlighted in grey.

was calculated. For each additional tissue repertoire, the difference in frequency, compared to peripheral blood, was calculated at each codon position. Insertions and deletions were both located predominantly in CDRs as opposed to frameworks, and were roughly equally distributed in location between heavy chain CDR1 (HCDR1) and HCDR2.

A large increase in frequency of both insertions and deletions in the small intestine antibody repertoire was noted. This finding was even more surprising since the frequency of SHA indels correlates with the frequency of somatic hypermutation events (Wilson et al., 1998b), and a corresponding increase in somatic mutations was not observed in the small intestine antibody repertoire (Figure 13A). The presence of a large increase in SHA indel frequency without a corresponding increase in mutation frequency suggests that increased frequency of SHA indels in the small intestine antibody repertoire likely was not driven simply by an overall increase in affinity maturation associated events. Instead, these data suggest that SHA indels are enriched specifically in the small intestine repertoire.

Non-12/23 Recombinations in the Human Peripheral Blood Repertoire

Presence and frequency of putative V(DD)J recombinants in peripheral blood B cells

To limit the number of falsely identified V(DD)J recombinations (that is, recombinations with N-addition regions that contain stretches of similarity to diversity genes), I developed a stringent filtering procedure. All sequences were analyzed with the IMGT High/V-QUEST webserver, with the number of accepted diversity genes set to 2. The antibody region identified by IMGT as a putative diversity gene is designated here as the “match region” (Figure 17). Our filtering process required the match regions to contain a maximum of one nucleotide difference from the germline diversity gene segment. The length of the match region, minus any mismatches between the match region and the germline diversity gene, is designated the “match score”. The match score, which represents the

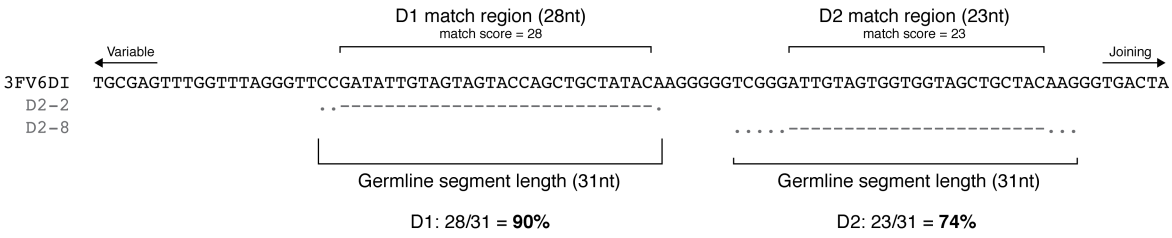


Figure 17. Stringent filtering of V(DD)J recombinants. A sample V(DD)J recombinant (3FV6DI) is shown, along with the germline diversity gene assignments for both the 5′ D (D2-2) and 3′ D (D2-8) positions. Dashes indicate conservation between germline and 3FV6DI and dots indicate mismatches. The match region is indicated above the 3FV6DI sequence, along with the match score (the match region length minus any mismatches within the match region). The germline segment length is shown below the sequence alignment. Below the germline segment length, the scoring calculation is shown. Sequences that contained scores of >60% for both diversity gene positions were considered V(DD)J rearrangements.

number of identically matched nucleotides between the match region and the germline diversity gene segment, was required to be at least 60% of the overall length of the germline diversity gene segment, except in the case of the short IGHD7-27 gene segment for which we required a match score of 72% of the germline diversity gene length (8 of 11 nucleotides).

The sequences obtained from each of the three cell subsets were examined for the presence of junctions containing multiple diversity gene segments using a high stringency filtering procedure. The frequency of V(DD)J recombination in each of the three sorted B cell subsets is shown in Figure 18A. The mean V(DD)J recombinant frequency in the naïve population (0.12%) was more than 10-fold higher than in the IgM memory population (0.01%, p=0.0095). Interestingly, the IgG memory population did not contain a single predicted V(DD)J recombination event that passed our filtering procedure. It was possible that our filtering procedure, which allowed only a single mutation in the match region, preferentially rejected prediction of V(DD)J recombinants from the somatically mutated memory populations while retaining V(DD)J recombinants in the mutation-free naïve population. A less stringent analysis, which allowed mutations, revealed increased V(DD)J

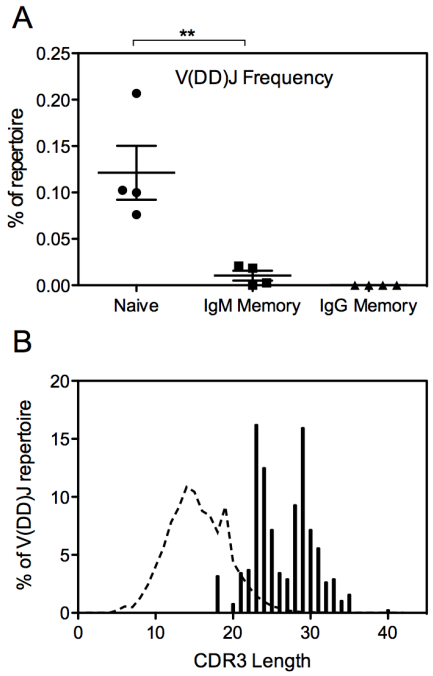


Figure 18. Frequency and HCDR3 length of putative V(DD)J recombinants. (A) V(DD)J recombinant frequency (as a percent of the subset repertoire) of naïve, IgM memory or IgG memory B cell subsets isolated from the peripheral blood of four healthy individuals. The V(DD)J frequency for each donor ± SEM for each subset is shown. Pairwise comparisons of V(DD)J recombinant frequency between different subsets were determined using a one-way ANOVA with Bonferroni's correction. (B) Histogram of CDR3 length distribution of V(DD)J recombinations (filled bars). The length distribution for the entire repertoire (dashed line) is also shown for comparison. ** = p<0.01.

recombinant frequency in all subsets including IgG memory, but still showed a significant reduction in V(DD)J recombination frequency in the IgM memory (0.08%, p=0.0077) and IgG memory (0.04%, p=0.0048) subsets when compared to the naïve subset (0.23%). Memory subsets have a reduced frequency of long HCDR3 loops (Tian et al., 2007) and the V(DD)J population is dominated by long HCDR3-containing antibodies (Figure 18B), with the average HCDR3 length of 26.5 amino acids.

Germline diversity gene usage in putative V(DD)J recombinants

The D gene usage in V(DD)J recombinants was compared to D gene usage in the total naïve cell repertoire to identify if there was a preferential use of particular D genes in V(DD)J recombinants (Figure 19). Interestingly, D gene use at the 3′ D position in V(DD)J recombinants was very similar to D gene usage in the total naïve repertoire, showing only an increase in D7 gene family usage (difference between means of 6.93 ± 2.58, p=0.036) and an approximately equivalent decrease in D2 gene family usage (difference between means of -10.02 ± 2.40, p=0.0058). D gene use at the 5′ D position in V(DD)J recombinants showed a significant increase in the D2 gene family (p=0.0022) and a

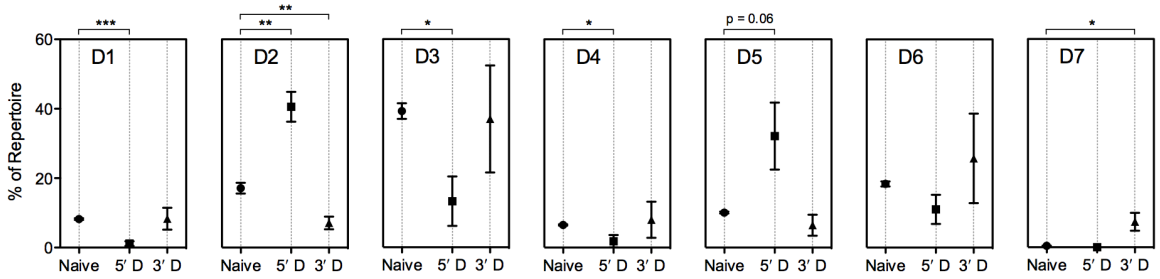


Figure 19. Diversity gene use in putative V(DD)J recombinants differs from that in the total naïve repertoire. Diversity gene family use for the total naïve repertoire (Naïve) or for the 5′D and 3′D positions in V(DD)J recombinants. Mean ± SEM for each donor is shown. Pairwise comparisons were determined using a two-way ANOVA with Bonferroni's correction. * = p<0.05; ** = p<0.01; *** = p<0.001.

significant decrease in the D1, D3 and D4 gene families (p<0.0001, p=0.0132 and p=0.0414, respectively). The 5′ D position also showed a strong trend toward increased D5 gene family use (p=0.06) and a notable absence of D7 family members.

N-addition length and GC content of V(DD)J recombination sites indicate true D-D fusion

While it is highly unlikely that random insertion of nucleotides at the VD or DJ junction would result in long sequence stretches that identically match germline diversity gene segments, I examined the N-addition lengths of all recombination sites in the total peripheral blood repertoire to determine if the N-addition regions flanking both diversity genes matched that of the normal V(D)J repertoire. I determined the mean N-addition length for all VD and DJ recombination sites in the total repertoire and the VD (also referred to as N1), DD (N2), and DJ (N3) recombination sites in the putative V(DD)J repertoire (Figure 20A). I found that there was no statistically distinguishable difference between the mean length of the recombination sites in the total repertoire and in the putative V(DD)J repertoire.

I next compared the GC content of all N-addition (N) and D gene regions in the total naïve repertoire to the GC content of both assigned diversity genes (D1 and D2) and the N1, N2 and N3 N-addition sites in the V(DD)J repertoire (which correspond to the VD, DD and

DJ recombination sites, respectively) (Figure 20B). There was a highly significant decrease in the GC content of both D1 ($p<0.0001$) and D2 ($p=0.0051$) regions when compared to the N-addition regions in the total naïve repertoire. In contrast, neither of the assigned D gene segments in the V(DD)J repertoire differed from the GC content of assigned D genes in the total naïve repertoire, Thus, the D1 and D2 regions better resembled the GC content of D gene segments than N-addition regions.

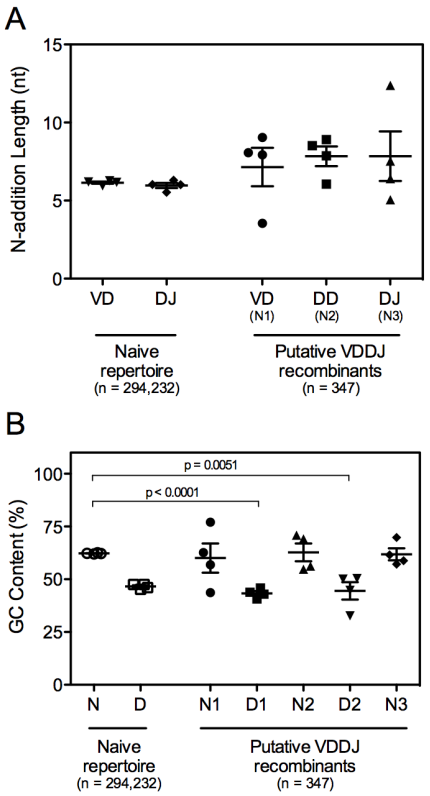


Figure 20. Putative V(DD)J recombinants contain normal N-addition lengths and diversity genes, with low GC content. (A) N-addition length for each recombination site in the total naïve repertoire or in the V(DD)J repertoire. The mean N-addition length for each of four healthy individuals ± SEM is shown for each recombination site. (B) GC content (as a percent of the region sequence) for each N-addition region or diversity gene segment in the V(DD)J repertoire or the total naïve repertoire. Combined N-addition at the VD and DJ junctions (N) or diversity gene region (D) are shown for the total repertoire. Diversity genes at the 5′D position (D1) and 3′D position (D2) and N-addition sites at the VD junction (N1), DD junction (N2) and DJ junction (N3) are shown for putative V(DD)J recombinants.

D gene order in V(DD)J recombinants matches the order of those D genes in the genome.

I analyzed 5′ D and 3′ D pairings in the V(DD)J repertoire and discovered that every V(DD)J recombinant contained D-D pairings in an orientation that matched the orientation of the genomic locus (Figure 21A). I also found that V(DD)J recombination occurred across the spectrum of D genes, using every D gene with the exceptions of D4-11, D4-14 and D6-25. D4-11 (<0.001%), D4-14 (0.895%) and D6-25 (0.60%) were the least frequently observed D genes in the total repertoire (Figure 5C), so the lack of these D genes in the V(DD)J repertoire was likely due to their rarity. As expected, the 5′ D position contained a high proportion of D gene segments located at the 5′ end of the

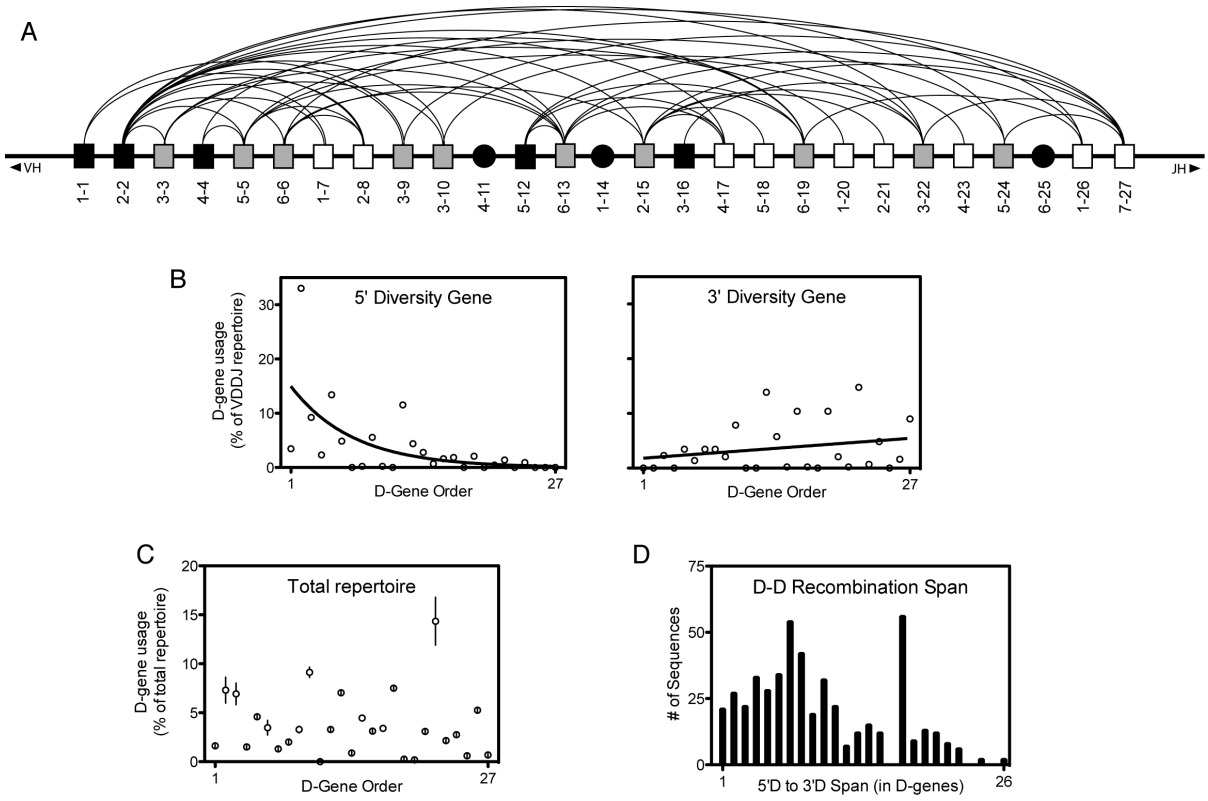


Figure 21. Genomic orientation of diversity genes matches the orientation in putative V(DD)J recombinants and explains diversity gene use bias at 5′D and 3′D positions. (A) All functional diversity genes are represented in the genomic orientation, with arcs connecting diversity genes found paired in a V(DD)J recombinant. Black boxes indicate diversity genes found only in the 5′D, white boxes indicate diversity genes found only in the 3′D position, and grey boxes indicate diversity genes found in both positions. D4-11, D1-14 and D6-25, the only diversity genes not found in any V(DD)J recombinants, are represented by black circles. The orientation of diversity genes in V(DD)J recombinants matches the genomic orientation in every instance, so the leftmost member of any linked pair of diversity genes shown in this diagram was always in the 5′D position of the V(DD)J recombination. (B) The frequency of each diversity gene in either the 5′D or 3′D positions is shown with the diversity genes ordered and labeled by position from the 5′ end of the genomic locus. (C) The frequency of each diversity gene in the total repertoire (including all B cell subsets). Shown are the mean frequency ± SEM for each donor. (D) The frequency of each recombination span, defined as the distance between the 5′D gene and the paired 3′D gene in V(DD)J recombinants (measured in diversity gene segments and including non-functional genes). Recombination between adjacent diversity genes results in a recombination span of 1.

genomic locus (Figure 21B), with the frequency of D gene presence in the 5′ D position decreasing exponentially with distance from the 5′ end of the genomic locus. A similar preference for upstream D genes was not seen in when analyzing the frequency of D gene use in the entire repertoire (Figure 21C). Only three of the ten furthest downstream (3′) D genes were ever found in the 5′ D position of a V(DD)J recombinant. There also was a weak

trend toward increased usage of downstream D genes in the 3′ D position of V(DD)J recombinants (p=0.2089).

I next determined whether or not there was a preference for D-D recombination events between D genes located close to each other in the genomic locus. For each V(DD)J recombination, we determined a recombination span, calculated by subtracting the position number of the 5′ D gene from the position of the 3′ D gene. Recombination between adjacent D genes resulted in a recombination span of 1, while recombination between the first and last (27th) diversity genes resulted in a recombination span of 26. We observed a strong trend toward decreased use of the most distant pairings (Figure 21D; p=0.0568). Notably, although there was a global trend toward decreased pairings of distant diversity genes, the most frequently observed recombination span was 17, of which there were several D-D combinations accounting for over 10% of all V(DD)J recombinants.

Skewed germline gene usage in 5′ D and 3′ D positions was likely the result of diversity gene orientation in the genomic locus.

Understanding that the frequency of diversity gene use in the 5′ D position of V(DD)J recombinants depended on location in the genomic locus (Figure 21B), we investigated whether or not orientation of the genomic locus was likely to be the cause of the skewed diversity gene usage seen in 5′ D and 3′ D positions of V(DD)J recombinants, or whether other genetic or mechanistic factors were the dominant force behind the skewed diversity gene repertoire.

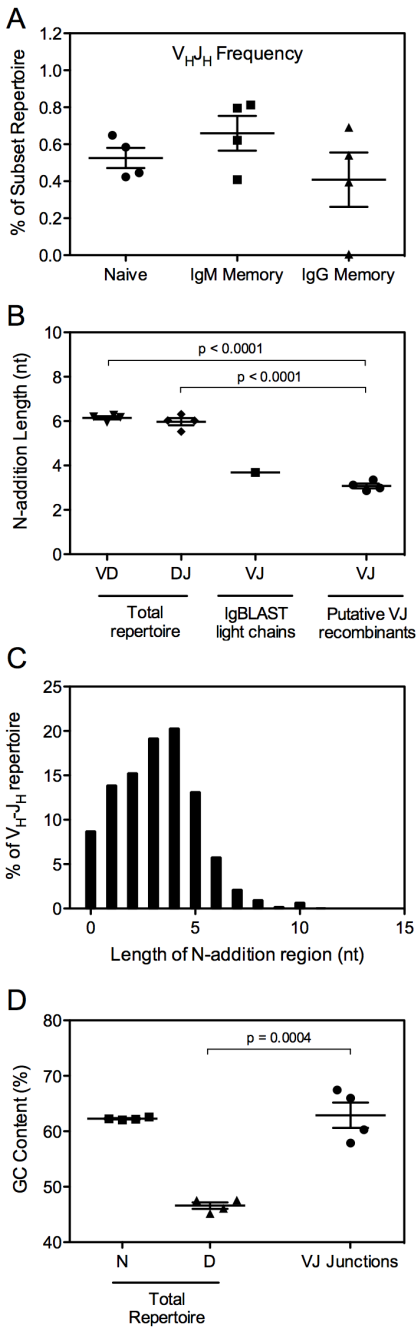
At the 5′ D position, the complete lack of D7 family use was readily explained by the location of the only D7 family member, D7-27, at the 3′ end of the genomic locus: in-order V(DD)J rearrangements with D7-27 in the 5′ D position are not possible. The increase in D2 family use at the 5′ D position also was likely attributable to genomic orientation. The most commonly used D2 gene member in the naive repertoire, D2-2, accounted for over half of

D2 family use in the naïve repertoire and is located one position from the 5′ end of the genomic locus. In addition, three of the four D2 family members, accounting for over 80% of D2 family use in the naive repertoire, were found in the 15 (of 27) most upstream positions of the genomic locus. Much like the increase in D2 family usage, the decrease in D1 family likely was due to the extreme downstream position of the most commonly used D1 family member, D1-26, which accounted for over 50% of D1 family usage in the naive repertoire. Finally, the trend toward increased D5 family frequency was possibly due to the positioning of the two most common D5 gene members, used in over 75% of D5 family use in the naïve repertoire, in the 5′ half of the genomic locus.

The increase in D7 family usage in 3′ D positions likely was attributable to the fact that D7-27, the only D7 family member, is positioned at the furthest 3′ position of the genomic locus, allowing for in-order V(DD)J recombination with every other diversity gene. Alternatively, the decreased use of D2 family use in 3′ D positions was possibly explained by the fact that the most commonly used D2 family member, D2-2, is positioned such that only one possible V(DD)J recombination exists with D2-2 in the 3′ D position.

Frequency of direct V_H-J_H recombinants is similar in naïve and memory subsets

Since the occurrence of recombinants with multiple D gene suggested that non-classical events are tolerated at some level during V(D)J recombination, I considered whether an alternate mechanism occurs in which a D gene is not incorporated. Naïve, IgM memory and IgG memory subsets were examined for the presence of recombinants that did not contain a D gene. These putative V_H-J_H recombinants were defined as sequences that contained an identifiable V_H gene, an identifiable J_H gene, but no identifiable D gene. The frequency of putative V_H-J_H recombinants did not differ statistically in naïve, IgM memory or IgG memory subsets (Figure 22A). This finding was in contrast to V(DD)J recombinants, where the frequency of V(DD)J recombinants was reduced in both memory cell subsets.



The variable gene family and joining gene use in putative V_H-J_H recombinants did not differ significantly from germline usage in the total repertoire (data not shown).

N-addition length in putative V_H-J_H recombinants was lower than in conventional V_HD_H and D_HJ_H regions but is similar to V_L-J_L N-addition length

The nucleotide length of all junctions in putative V_H-J_H recombinants was examined and compared to the N-addition length for the V_HD_H and D_HJ_H recombination sites in the total naïve repertoire (Figure 22B). The N-addition region at the VJ junction of potential V_H-J_H recombinants was significantly shorter than at the V_HD_H ($p<0.0001$) or the D_HJ_H ($p<0.0001$) recombination sites in the total naïve repertoire. Interestingly however, the mean length of N-addition regions in V_H-J_H junctions (3.01 nt) was very similar to the corresponding junction in light chains (3.69 nt),

Figure 22. The junctional regions of putative direct V_H-J_H recombinants were GC-rich and were similar to V_L-J_L recombinants in length. (A) Frequency of putative direct V_H-J_H recombinants in peripheral blood B cell subsets. (B) N-addition length for the VD and DJ junction in the total naïve repertoire and of VJ junctions in putative VJ recombinants. The mean N-addition length for each of four healthy individuals ± SEM is shown for each recombination site. The mean N-addition length at the VJ recombination site for all light chain antibody sequences in the IgBLAST database is also shown. (C) Histogram showing the frequency of N-addition length (in nucleotides) as a percentage of the putative direct V_H-J_H recombinant repertoire. (D) GC content for the N-addition (N) or diversity gene (D) regions of the total naïve repertoire or of the VJ junction in putative VJ recombinants. The mean GC content for four healthy individuals ± SEM is shown for each recombination site.

which are encoded only by V and J segments. A histogram showing the length of the putative N-addition regions in all V_H-J_H recombinants (Figure 22C) revealed that very short N-addition lengths are highly preferred in V_H-J_H junctions, with nearly 9% of all putative V_H-J_H recombinants containing no N-addition and only a single sequence (0.03%) containing an N-addition region longer than 10 nucleotides.

GC content of putative V_H-J_H junctions resembled that in N-addition regions but differed from that in D gene regions

The GC content of the VJ junction region in putative V_H-J_H recombinants was determined, as well as the GC content for N-addition and diversity gene regions in the entire conventional naïve repertoire (Figure 22D). The VJ junction regions in putative V_H-J_H recombinants contained GC content that was consistent with that in N-addition regions but, in contrast, was significantly higher than that in D gene regions ($p=0.0004$). Producing putative V_H-J_H recombinants by trimming of D genes would require degrading the ends of D gene segments in a non-random manner, consistently leaving behind D gene remnants that contain GC content significantly different from the D gene repertoire as a whole. Thus, it is likely that these recombinants are generated by direct V_H-J_H recombination.

Identical matching of pentamers from the 3′ end of variable gene segments revealed V_H replacement frequency

Since coincident N-addition matches to the relatively short pentamer sequences in the panel would be expected to artificially inflate the V_H gene replacement estimate, I examined both the N1 and N2 regions for identical matches to the pentamer panel. N2 corresponds to the N-addition region between the diversity and joining genes and where V_H replacement does not occur, and serves as a control to define the expected frequency of random N-addition matches to the pentamer panel. To eliminate the possibility of somatic

mutations altering pentamer sequences and producing either false positive or false negative results, I initially examined only sequences from the naïve B cell subset. I found that a sequence in the pentamer panel matched to a sequence in an N2 region (presumably randomly) in 5.1% of sequences, which is much higher than in previous studies (Zhang et al., 2003). The frequency of identical N1 region matches to a sequence in the pentamer panel was 6.5% (Figure 23A). The increased occurrence of pentamer matches in the N1 region was highly significant (p=0.0048), and the difference of means suggests the frequency of V_H replacement events to be ≥1.43% (95% CI of 0.8-2.0%). This frequency is somewhat lower than previous estimates, which were not limited to analysis of sequences from naïve B cell subsets.

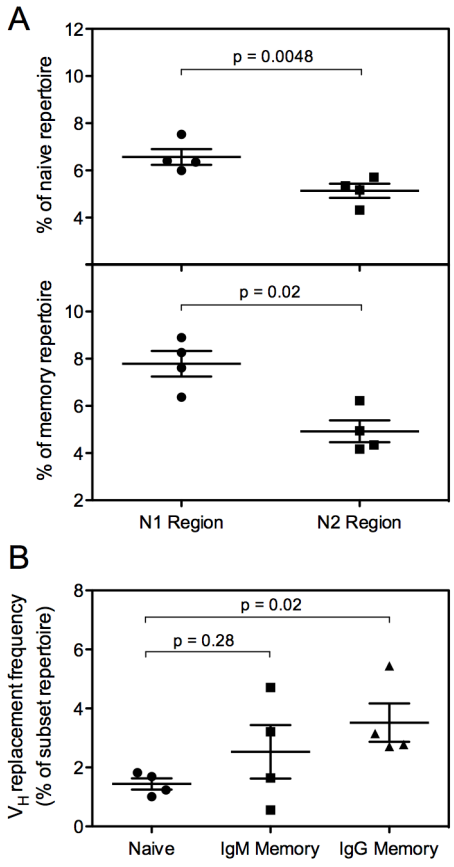


Figure 23. Frequency of V_H replacement in the peripheral blood repertoire. Frequency of V_H replacement events based on identical matches in V-D recombination region (N1) to a panel of pentameric nucleotide sequences representing the 3′ end of all functional variable genes. The frequency of the same panel of pentameric sequences in the D-J recombination region (N2) where V_H replacement does not occur was used to define the expected frequency of random matches, as a negative control. The frequency of pentameric matches in four healthy individuals ± SEM is shown for each region. (B) V_H replacement frequency was determined for each B cell subset by subtracting the mean N2 pentamer match frequency from the mean N1 pentamer match frequency. Plotted is the difference in means for each donor ± SEM.

To determine whether circulating memory B cells express sequences with an increased frequency of V_H replacement, we pooled the IgM and IgG memory B cell subset sequences and performed the same analysis on the entire memory subset (Figure 23A). This analysis revealed a frequency of matches in the N1 region (4.9%) that was similar to the frequency observed in the naïve subset, indicating that somatic hypermutation does not

produce a significant number of falsely positive matches to the pentamer panel. Interestingly, analysis of the V_H recombination frequency (defined here as the mean frequency of identical pentamer matches in the N1 region minus the number in the N2 region) showed a significant increase in V_H recombination frequency in the class-switched IgG memory subset compared to the naïve subset ($p=0.02$, Figure 23C).

Discussion

The substantial differences in each subset (naïve, IgM memory and IgG memory) of the human peripheral blood repertoire at the individual V(D)J recombinant level, coupled with the overarching similarities of the repertoires at the germline gene family level, present something of a paradox. These data seem to suggest the presence of a broad, global mechanism for repertoire regulation at the germline gene family level. While there is no direct evidence of such a mechanism, several recent studies indicate the presence of repertoire-based regulatory mechanisms in circulating B cell subsets. Despite the tendency of pathogen-specific antibody responses to exhibit biased germline gene repertoires (Weitkamp et al., 2005; Tian et al., 2008; Gorny et al., 2009), the frequency of gene family use in naïve and memory subsets is remarkably consistent across individuals (Tian et al., 2007; Wu et al., 2010b). Further, alteration of this gene family homeostasis in the circulating B cell repertoire is associated with disease states (David et al., 1995; 1996; Zouali, 1996; Scamurra et al., 2000). In recent work, long-lived plasma cells were shown to contain significantly fewer autoreactive B cells than the circulating IgG memory subset, and this difference was attributed to differential repertoire regulation within the two subsets (Scheid et al., 2011). Thus, while no mechanism for regulation of circulating human antibody repertoires has been identified, mounting indirect evidence, including the data presented in this report, suggests the presence of such regulation.

Three different types of recombination events were identified in the human peripheral blood repertoire that appear to violate the conventional 12/23 rule of recombination: $V_H(D_HD_H)J_H$ recombination, direct V_H-J_H recombination without any D_H segment, and V_H replacement. While the occurrence of such events has been identified previously in both *in vitro* and *in vivo* systems, experiments in many of these systems were designed to induce these events in a non-physiologic manner (Akira et al., 1987; Hesse et al., 1989; Akamatsu et al., 1994). Further, two of these events, V(DD)J recombination and VH replacement, may result in antibodies that encode exceptionally long HCDR3 loops and may point to the genetic origin of such antibodies.

I identified putative V(DD)J recombinants in the naïve B cell subset at a frequency of approximately one in 800 circulating cells. Previous estimates of V(DD)J frequency have varied widely, partly due to differing criteria for identification of putative V(DD)J recombinants. Several methods used previously involved low-stringency filters that identified putative D genes using as few as four homologous nucleotides. When using such a short homology region to identify D-D fusion events, it is likely that many of the identified D-D fusions were false attributions. The presence of false positives in the identified V(DD)J repertoire could dilute or counteract genetic trends in the true V(DD)J population, so limiting falsely positive results was of critical importance. The stringent approach used in this study, while limiting false positives, also likely underestimated to some degree the actual frequency of cells that express antibodies encoded by DNAs derived from V(DD)J recombination. Thus, the V(DD)J frequency estimate produced in this study describes the minimum expected frequency of such events in the human peripheral blood repertoire. This estimate, while likely an underestimate of V(DD)J recombinant frequency, is approximately 20-fold lower than the frequency of long HCDR3s in the human peripheral blood antibody repertoire (Briney et al., 2012a), making it unlikely that V(DD)J recombination is the primary mechanism responsible for the generation of long HCDR3s.

To further show that the sequences in putative V(DD)J recombinants identified in this study were generated by true D-D fusion and not alignment of coincidental N-addition mediated sequences, I closely analyzed the nucleic acid characteristics of sequence regions within putative V(DD)J recombinants and determined that it is highly unlikely that N-addition mimicry was the primary cause of these putative V(DD)J recombinants. Analysis of N-addition length, GC-content and orientation of the 5′ and 3′ D gene segments in V(DD)J recombinants all suggest that the N-addition mimicry mechanism is unlikely to account for these sequences. Random addition of nucleotides that coincidentally match germline D genes while satisfying all of the conditions identified above would require an extraordinarily unlikely set of events.

The frequency and genetic composition of direct V_H-J_H recombination events in the human peripheral blood repertoire was also determined. While the possibility that some of these sequences are the product of extensive trimming of D genes cannot definitively be excluded, analysis of the nucleic acid characteristics of the junctional region in putative V_H-J_H recombinants revealed that it is highly unlikely that D gene trimming was the primary source of these recombinants. In contrast to V(DD)J recombinants, the frequency of putative V_H-J_H recombinants, which encode much shorter CDR3 loops than the putative V(DD)J recombinants, was similar in naïve and memory cell subsets. This finding raises an interesting question: why do clones with the very short HCDR3 loops generated by putative direct V_H-J_H recombination transition from naïve to memory at a much higher frequency than clones that possess the longer CDR3 loops that are encoded in V(DD)J recombinants? B cells of the IgM memory and IgG memory subsets have significantly shorter mean CDR3 lengths than the total naïve subset (Tian et al., 2007). It is possible that flat or convex epitopes, which would likely not pair well with antibodies containing long, protruding HCDR3s, are targeted more frequently by naïve B cells, resulting in an increased fraction of short HCDR3s in the memory repertoire.

V_H replacement, a process of secondary recombination between variable genes and previously completed V(D)J recombined genes, has been shown in humans and animal model systems to facilitate the rescue of cells with non-functional or autoreactive primary recombinations (Zhang et al., 2004; Lutz et al., 2006). Published frequency of V_H replacement in the peripheral blood antibody repertoire has varied widely (Zhang et al., 2003; Koralov et al., 2006; Watson et al., 2006). I found matches to the pentamer panel in the N1 region at similar frequencies to those reported by Zhang, *et al*. There was far more coincidental matching of the pentamer panel in the N2 region than was noted in that study, however, indicating that the true V_H replacement frequency is substantially lower than previously thought. Although the frequency of V_H replacement is similar to the frequency of long HCDR3s, V_H replacement typically only adds 5-6 nucleotides to the overall HCDR3 length. Thus, V_H replacement alone is likely not sufficient to convert an HCDR3 of normal length (approximately 15 amino acids) into an HCDR3 of extreme length like that of PG9 or PG16 (30 amino acids).

While the frequency of V_H replacement reported here is lower than that of previous work, the data presented here constitutes the first report of an increase in V_H replacement in memory cell subsets when compared to cells in the naïve subset. This surprising discovery suggests either that B cells containing antibodies generated by V_H replacement transition to the memory subset more frequently that B cells that have not undergone V_H replacement, or that V_H replacement can occur in mature B cell populations, raising the possibility of antigen-driven V_H recombination in humans. Other studies have shown that stimulation of B cells with IL-4 or IL-7 and CD40 can induce RAG expression, that RAG proteins are expressed in germinal center B cells (Han et al., 1997; Papavasiliou et al., 1997) and that secondary recombinations occur in mouse germinal centers (Wang and Diamond, 2008). Thus, it is feasible that there is a window for V_H replacement in mature human B cells at or near the time of stimulation.

CHAPTER IV

GENETIC FEATURES OF BROADLY NEUTRALIZING HIV ANTIBODIES ARE PRESENT IN THE PERIPHERAL BLOOD REPERTOIRE OF UNINFECTED INDIVIDUALS

Introduction

Antibodies with long HCDR3s

Antibodies containing long heavy chain complementarity determining region 3 (HCDR3) loops have been shown to efficiently neutralize a wide variety of pathogens, including HIV, malaria, and African trypanosomes (Stijlemans et al., 2004; Burton et al., 2005b; Henderson et al., 2007). In some cases, the unique feature of long HCDR3 antibodies is that the extended loop structure facilitates interaction with epitopes that are otherwise occult because of extensive glycosylation or location in recessed structures on the pathogen surface.

For HIV, several of the most broad and potently neutralizing antibodies have extremely long HCDR3 loops. Two exceptionally broad and potent anti-HIV antibodies, PG9 and PG16, encode among the longest human antigen-specific antibodies described to date and form secondary structure through the use of a complex hydrogen bonding network in the HCDR3 (Walker et al., 2009; Pancera et al., 2010). These antibodies target a currently undefined quaternary epitope and preferentially bind cell surface expressed trimeric envelope protein (Doores and Burton, 2010; Pejchal et al., 2010). In addition to long HCDR3s, both antibodies contain sulfated tyrosine residues within the HCDR3, a post-translational modification that has only recently been observed in antibody combining sites (Walker et al., 2009). Two additional HIV antibodies, designated 2.5b and 2909, target a

similar quaternary epitope and contain long HCDR3s, but are able to neutralize only a very limited panel of virus isolates (Changela et al., 2011; Spurrier et al., 2011). A panel of recently described antibodies, PGT141-PGT145, are purported to target the same quaternary epitope as PG9 and PG16, have a similar strong preference for membrane-bound, trimeric envelope, and encode HCDR3s that are even longer than the exceptionally long HCDR3s seen in PG9 and PG16 (Walker et al., 2011). The broadly neutralizing HIV antibody b12 contains a long HCDR3 and is able to neutralize by targeting the conserved CD4 binding site (Burton et al., 1991; Barbas et al., 1993; Burton et al., 1994). The b12 antibody uses only heavy chain interactions at the antigen binding interface, and passive administration of b12 has been shown to be protective against low-dose repeated challenge in macaques (Saphire et al., 2001; Hessell et al., 2009). Two other broadly neutralizing antibodies with long HCDR3s, 4E10 and 2F5, target a conserved membrane-proximal region and have been shown to protect against mucosal SHIV challenge alone and in combination with the anti-HIV antibody 2G12 (Mascola et al., 1999; 2000; Stiegler et al., 2001; Hessell et al., 2010), and the long HCDR3 of 2F5 is critical to the neutralizing ability of 2F5 (Zwick et al., 2004a). Antibody 447-52D contains a long HCDR3 loop and is able to neutralize a broad range of clade B HIV-1 isolates by targeting a conserved epitope on the V3 loop of gp120 (Stanfield et al., 2004; Jiang et al., 2010). Finally, the neutralizing antibody 17b targets the HIV co-receptor binding site and facilitates neutralization by preventing co-receptor binding and reducing affinity for the primary receptor, cluster of differentiation 4 (CD4) (Kwong et al., 1998). Thus, antibodies containing long HCDR3s comprise a sizeable fraction of the neutralizing HIV antibodies described to date, including many of the most broad and potently neutralizing antibodies.

Although induction of such long HCDR3 antibodies is likely to be critical to the design of an effective HIV vaccine strategy, it is still unclear how to induce such antibodies. Previous work has speculated as to a potential mechanism for inducing such antibodies by

vaccination (Pejchal et al., 2009; Pancera et al., 2010). It is known that the affinity maturation process is associated with codon-length insertion events that are likely caused by the somatic hypermutation machinery (Wilson et al., 1998a; 1998b; Reason and Zhou, 2006). It is thought, then, that repeated rounds of affinity maturation, resulting in multiple short insertion events within the HCDR3, could gradually lengthen HCDR3 loops in the affinity matured antibodies. This observation fits well with the known kinetics of broadly neutralizing antibody generation during HIV infection: potently neutralizing HIV antibodies are produced later than is common in other viral infections (Richman et al., 2003; Wei et al., 2003), suggesting that many rounds of affinity maturation may be necessary to develop broad and potent neutralizing capacity.

It is also possible that long HCDR3 loops are not generated primarily through the affinity maturation process, however, and are instead primarily created during the recombination process through the introduction of extensive numbers of N- and P-insertions and the selective use of optimal germline gene segments. If the primary source of long HCDR3 antibodies in the peripheral blood is not affinity maturation, the process of inducing an antibody response containing antibodies with long HCDR3s would consist of exhaustive sampling of the repertoire to select those B cells encoding what are, presumably, rare antibodies (Pancera et al., 2010; Pejchal et al., 2010). While it has previously been shown that antibodies containing long HCDR3s are present in immature B cell populations in both man (Ivanov et al., 2005) and mouse (Schelonka et al., 2008; Vale et al., 2010) as well as in human perinatal liver (Schroeder and Wang, 1990), it is unclear how frequently these antibodies are able to successfully navigate the autoreactivity screening process and enter the mature B cell population. Extensive work has been done in characterizing short and long HCDR3s in mice (Ippolito et al., 2006; Schelonka et al., 2007; Schroeder et al., 2010), but much of the work was done model systems under non-physiologic conditions. An examination of hundreds of thousands of circulating human antibody sequences has

identified an upper limit to the number of unique HCDR3s in a single individual (Arnaout et al., 2011). The upper bound, 3 to 9 million unique HCDR3s per individual, is much lower than previously estimated, but this study did not describe the length distribution of these HCDR3s. It has been shown that many B cells encoding HCDR3s of extreme length are eliminated before reaching the periphery (Ivanov et al., 2005), likely because antibodies with long HCDR3s tend to have autoreactive properties (Crouzier et al., 1995; Aguilera et al., 2001; Wardemann et al., 2003; Haynes et al., 2005). Detailed study and genetic characterization of the long HCDR3 antibody population in human peripheral blood has been limited by the rarity of such sequences. In this study, I examined expressed antibody sequences from four healthy donors and determined that antibodies containing long HCDR3s are more common in the naïve subset than in memory, indicating that affinity maturation is not the primary source of such antibodies. Further, extensive genetic characterization identified several conserved sequence elements in the long HCDR3 peripheral blood antibody population. Thus, human peripheral blood antibodies containing long HCDR3s are not generated primarily through repetitive rounds of affinity maturation, but are typically formed at the time of the original recombination.

Antibodies with somatic hypermutation-associated insertions and deletions

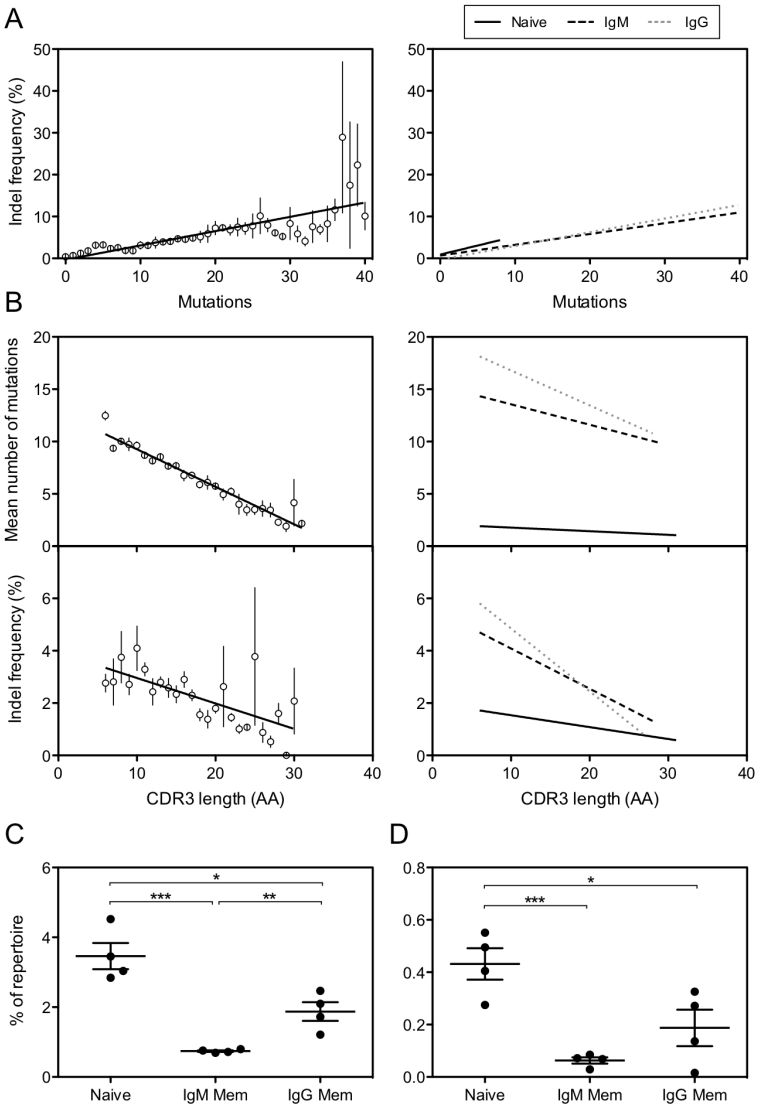
The somatic hypermutation (SHM) process typically results in single point mutations, but occasionally produces insertions or deletions of varying length (citations). Several antibodies have recently been described for which somatic hypermutation-associated insertions and deletions (SHA indels) are critical to proper antigen binding. For HIV, there are several bnAbs that require SHA indels for potent neutralization. In the case of VRC01, a bnAb that targets the CD4bs, a six nucleotide deletion in the light chain CDR1 (LCDR1) reduces the size of the LCDR1 and removes steric clashes with the target antigen (Wu et al., 2010a). When the deletion is restored, binding and neutralization are severely reduced.

As with long HCDR3-encoding antibodies, much effort is being directed toward the design of immunogens that efficiently elicit bnAbs. In the case of bnAbs that require SHA indels for appropriate function, it is critical to know whether antibodies encoding these SHA indels must be induced by the vaccine or whether they are already present in the HIV-uninfected antibody repertoire. The latter prospect is especially appealing, since selection of antibodies containing pre-existing SHA indels is thought to be much simpler that designing immunogens that specifically elicit such SHA indels. Thus, examination of the HIV-uninfected peripheral blood repertoire and characterizing the frequency and distribution of SHA indels is an important first step toward the design of immunogens that successfully induce a broadly neutralizing HIV response.

The Genetic Origin of Antibodies Encoding Long HCDR3s

Increased HCDR3 length was not associated with an increased number of somatic mutations or insertions.

Three major subsets of B cells in the peripheral blood were considered: naïve B cells, which are antigen inexperienced and lack somatic mutations or class-switching; IgM memory B cells, which express the surface memory marker CD27 and show evidence of somatic hypermutation but not class-switching; and IgG memory cells, which express CD27 and have undergone both somatic hypermutation and class-switching. Naïve, IgM memory and IgG memory B cells were separately isolated from four healthy individuals (hereafter, designated Group 1) using flow cytometric sorting and subjected the transcribed antibody heavy chain genes from those cells to high throughput sequencing. After selecting only non-redundant, high-quality antibody sequences, a total of 294,232 naïve cell sequences, 161,313 IgM memory cell sequences and 94,841 IgG memory cell sequences were obtained from Group 1 donors. I also subjected the transcribed peripheral blood heavy chain antibody



genes of three additional healthy donors (hereafter, designated Group 2) and four HIV-infected donors to high throughput sequencing. The B cells from these additional donors were not sorted by B cell subset prior to sequencing, but instead represent a sampling of the total peripheral blood B cell repertoire from each donor.

Peripheral blood antibody sequences

Figure 24. Increased HCDR3 length does not correlate with affinity maturation events. (A) Peripheral blood antibody sequences were grouped by mutation frequency and the percent of sequences in each group that contained codon-length (non-frameshift) insertions was calculated for each donor. All values for healthy donors from Group 1 (n=4) and Group 2 (n=3) are shown in the left panel, with the mean percentage ± SEM shown for each mutation value. In the right panel, sequences from Group 1 healthy donors were segregated by B cell subset, and the best-fit linear regression for each subset is shown. (B) Peripheral blood antibody sequences were grouped by HCDR3 length (in amino acids) and the mean number of mutations for each HCDR3 length group was calculated for each donor. As in Figure 1A, the left panel shows the mean ± SEM for all donors in Group 1 and Group 2. The right panel shows the best-fit linear regression of each B cell subset for Group 1 donors. The percent of sequences within each HCDR3 length group containing non-frameshift insertions also was calculated. In the left panel, the mean percentage ± SEM for all donors in Group 1 and Group 2 is shown. In the right panel, the best-fit linear regression of each B cell subset is shown for Group 1 donors. (C) Peripheral blood antibody sequences from Group 1 healthy donors were grouped by donor into naïve and memory subsets and the percent of sequences containing long HCDR3s (≥24 amino acids in length, or two standard deviations above the mean HCDR3 length). The percentages for each donor are shown, with the mean ± SEM. (D) The donor groups from Figure 1C were analyzed for the frequency of very long HCDR3s (≥28 amino acids in length, *i.e.*, three standard deviations above the mean HCDR3 length). The percentages for each donor are shown, with the mean ± SEM. The p values were determined using a one-way ANOVA. All statistically significant differences are indicated. * = $p<0.05$, ** = $p<0.01$, *** = $p<0.001$

from Group 1 and Group 2 healthy donors were grouped by mutation frequency, and the fraction of sequences containing insertions was determined for each group (Figure 24A, left panel). In agreement with previously published data (Wilson et al., 1998a), I observed a strong positive correlation between number of mutations and insertion frequency (r^2=0.77, p<0.0001). Insertions were present only in a minority of the most highly mutated sequences, however, suggesting either that the somatic hypermutation process is inefficient at introducing insertions in a functional reading frame or that antibodies seldom are able to tolerate genetic insertions while retaining functionality. When separately analyzing each of the B cell subsets from Group 1 healthy donors, I observed that the correlation between mutations and insertions also was found for each of the memory B cell subsets (Figure 24A, right panel).

I next grouped the sequences by HCDR3 length and determined the mean number of mutations and insertion frequency for each HCDR3 length group (Figure 24B, left panels). Interestingly, HCDR3 length was negatively correlated with both mutation frequency (r^2=0.64, p<0.0001) and insertion frequency (r^2=0.13, p<0.0001), suggesting that genetic processes that accomplish somatic hypermutation typically do not alter HCDR3 length. A similar trend was seen in each of the Group 1 healthy donor B cell subsets (Figure 24B, right panels). It has been shown previously that the mean HCDR3 length in circulating memory B cell subsets is shorter than in the naïve B cell subset by approximately a single amino acid (Wu et al., 2010b). Analyzing the mean HCDR3 length, however, does not allow determination of whether there is a broadly-distributed, overall shortening of the entire HCDR3 repertoire (perhaps caused by somatic hypermuation-induced deletions in the HCDR3 that reduce the length of both long and short HCDR3s) or whether the lower mean HCDR3 length in memory is predominantly due to a strong preference against long HCDR3s in the memory subset. I examined the antibody sequences from naïve, IgM memory, and IgG memory B cell subsets from Group 1 healthy donors for presence of long HCDR3s

(defined here as HCDR3s ≥ 24 amino acids, which corresponds to 2 SD above the mean HCDR3 length) and for very long HCDR3s (defined here as HCDR3s ≥ 28 amino acids, which corresponds to 3 SD above the mean HCDR3 length). The naïve population contained a significantly higher fraction of long HCDR3s (3.5%, Figure 24C) than both the IgM memory subset (0.74%, p=0.0003) and the IgG memory subset (1.9%, p=0.014). The naïve subset also contained a significantly higher fraction of very long HCDR3s (0.43%, Figure 24D) than either the IgM memory subset (0.06%, p=0.001) or the IgG memory subset (0.19%, p=0.038). Interestingly, the IgM memory subset showed a significantly reduced frequency of long HCDR3s when compared with the IgG memory subset (p=0.0056), which supports emerging data that the IgM memory subset does not function primarily as a transition state in progression from the naïve to IgG memory repertoire (Wu et al., 2010b). In summary, introduction of point mutations and insertions during somatic hypermutation had little effect on HCDR3 length, and I noted the somewhat surprising presence of a large population of long HCDR3s in the naïve repertoire. These data suggested that long HCDR3s are not built primarily through repeated rounds of affinity maturation using genetic insertions, but are present in the naïve repertoire before B cells begin the affinity maturation process.

Antibody sequences encoding long or very long HCDR3s display skewed germline gene usage.

With the understanding that long HCDR3s are not produced primarily by the affinity maturation process, I examined the antibody repertoire for evidence of recombination events that correlate with HCDR3 length. First, I grouped the peripheral blood antibody sequences from each Group 1 and Group 2 healthy donor by three criteria: (1) all HCDR3s, which included all antibodies containing any HCDR3 length; (2) long HCDR3s, which

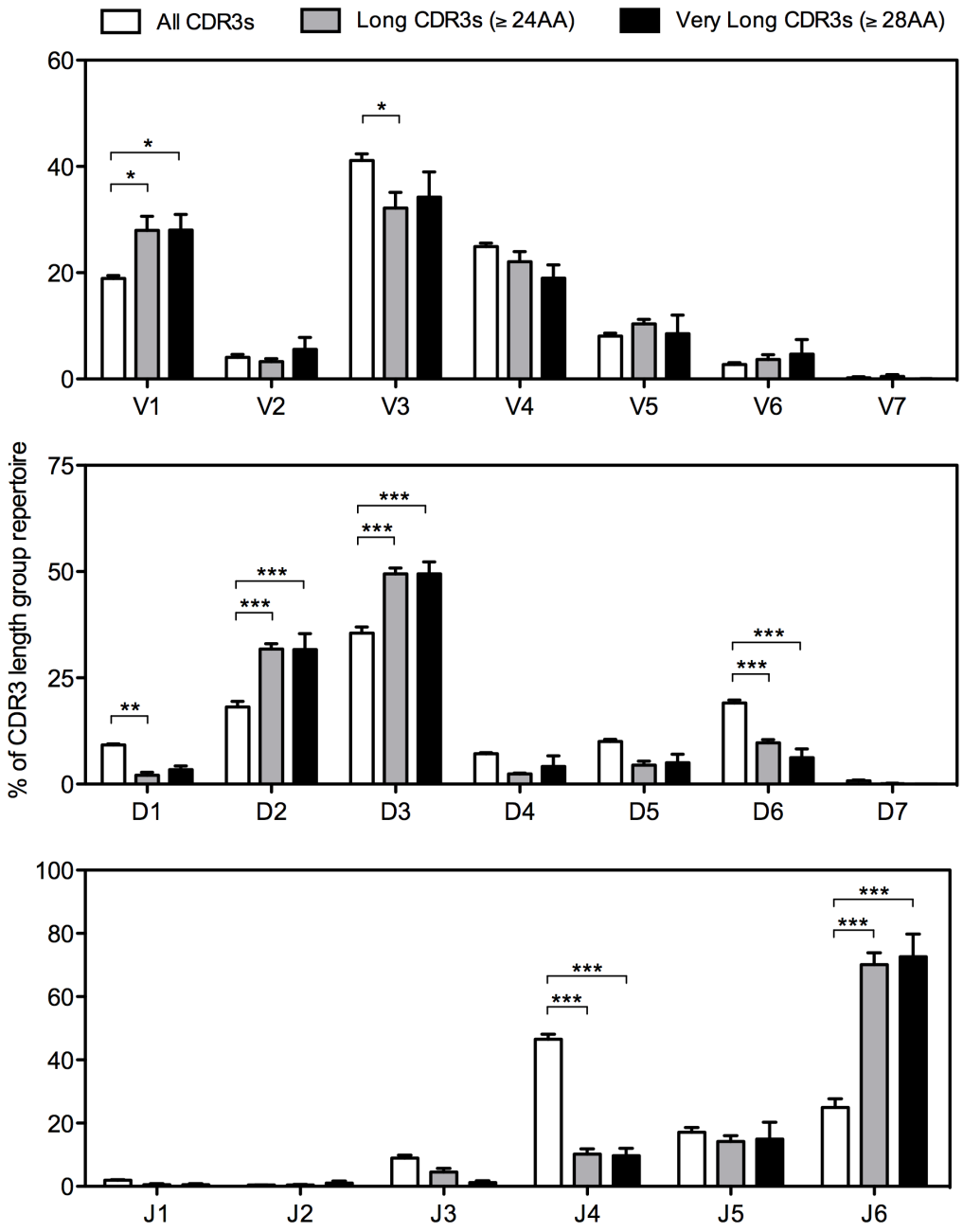


Figure 25. Skewed germline gene usage in antibodies containing long or very long HCDR3s. Peripheral blood antibody sequences from Group 1 (n=4) and Group 2 (n=3) healthy donors were assembled into the following three groups by HCDR3 length: (1) all HCDR3s, which contains all sequences of any HCDR3 length; (2) long HCDR3s, which contains only sequences with a HCDR3 length ≥ 24 amino acids; and (3) very long HCDR3s, which contains only sequences with a HCDR3 length ≥ 28 amino acids. The frequency of each germline variable gene family, diversity gene family, and joining gene was determined for each HCDR3 length group. The mean frequency ± SEM is shown. All HCDR3 lengths were calculated using the IMGT numbering system. The p values were determined using a two-way ANOVA with Bonferroni post-tests. All statistically significant differences are indicated. * = $p<0.05$, ** = $p<0.01$, *** = $p<0.001$

included only sequences with HCDR3 lengths ≥ 24 amino acids long; and (3) very long HCDR3s, which included only sequences with HCDR3 lengths ≥ 28 amino acids long.

I analyzed the germline V, D and J gene segment use in each of the three sequence groups (Figure 25). The most remarkable finding noted was the strong association of two particular D gene families and one J gene segment with longer HCDR3s. Use of D gene families D2 and D3 was increased in long HCDR3s (both D2 and D3: $p<0.001$) and very long HCDR3s (both D2 and D3: $p<0.001$). A significant decrease in use of D gene family D6 was seen in long and very long HCDR3s ($p<0.01$), and a decrease in D1 gene family use was seen in long HCDR3s ($p<0.01$). Use of J gene J_H6 was increased markedly in both long ($p<0.001$) and very long HCDR3s ($p<0.001$), while joining gene J_H4 use was decreased in long ($p<0.001$) and very long HCDR3s ($p<0.001$). This was not surprising, considering that J_H6 is the longest J_H gene segment and J_H4 is the shortest. Interestingly, however, use of J_H1 and J_H2, which are two amino acids longer than J_H4, was not increased significantly in long or very long HCDR3 groups. We also noted some variation in V_H gene usage. Use of variable gene family V_H1 was increased in the long HCDR3 ($p<0.05$) and very long HCDR3 ($p<0.05$) groups compared to the group with all HCDR3 lengths ($p<0.05$). V_H3 family use was decreased in long HCDR3s ($p<0.05$) but not in the very long HCDR3 group. Separate analysis of the individual B cell subsets in Group 1 donors indicated that the trends observed for the total repertoire were largely mirrored in each subset repertoire.

Deeper analysis of individual D gene use (as opposed to analysis of the D gene families shown in Figure 2) revealed that the increase in D2 and D3 gene families is driven almost completely by increased use of just three of the nine D2 and D3 gene family members: D2-2, D2-15 and D3-3 (Figure 27A). Separate analysis of the frequency of the three D genes that were increased in the total repertoire in each B cell subset (Figure 27B-D) showed similar trends, although the reduced sample size of each subset, as well as the

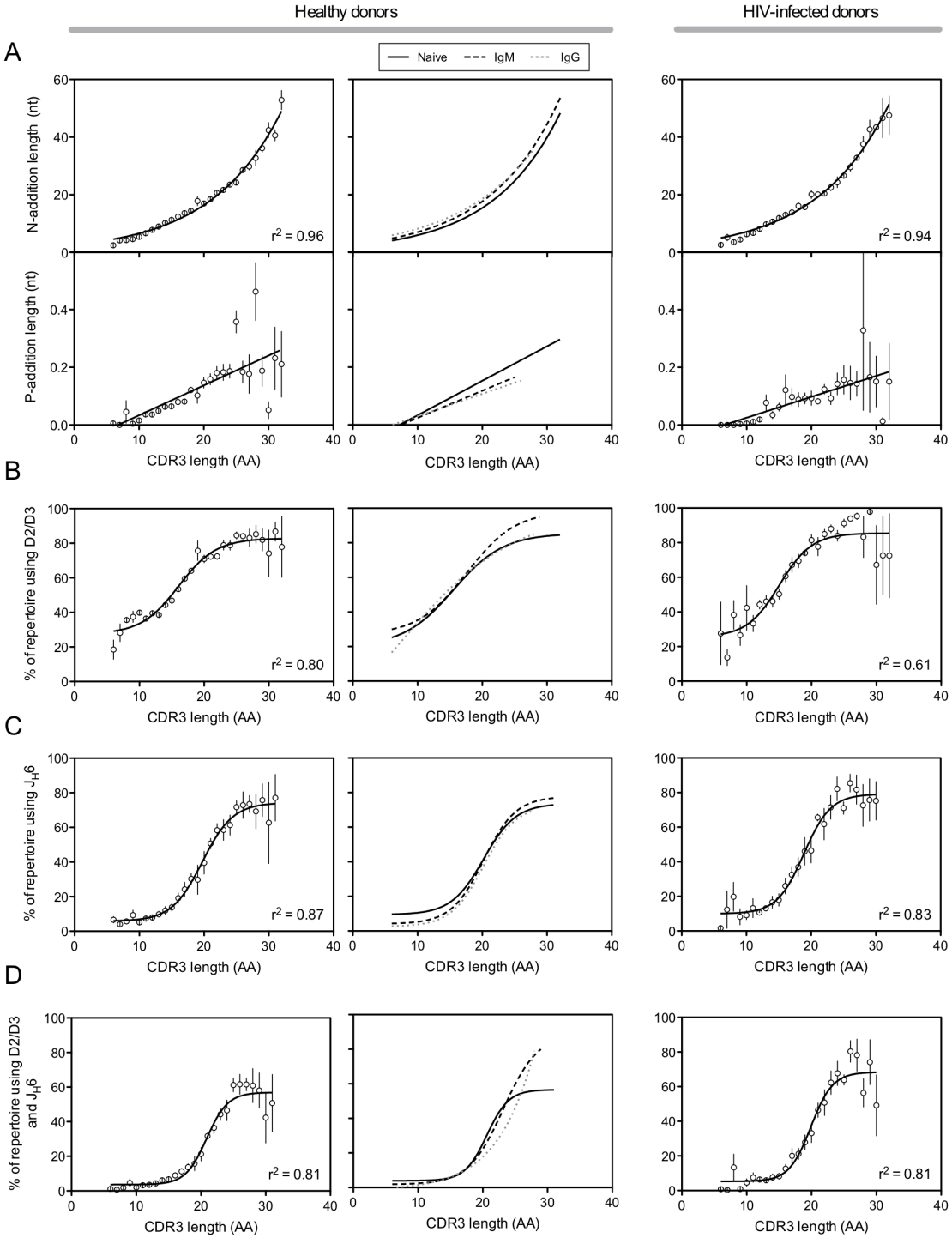

Healthy donors
HIV-infected donors
A
Naive
IgM
IgG
N-addition length (nt)
P-addition length (nt)
r2 = 0.96
r2 = 0.94
CDR3 length (AA)
B
% of repertoire using D2/D3
r2 = 0.80
r2 = 0.61
C
% of repertoire using JH6
r2 = 0.87
r2 = 0.83
D
% of repertoire using D2/D3 and JH6
r2 = 0.81
r2 = 0.81

Figure 26. (previous page) Long HCDR3s correlate with N-addition, P-addition and germline gene usage. For all figure sections, the leftmost panel corresponds to peripheral blood antibody sequences from Group 1 (n=4) and Group 2 (n=3) healthy donors. The middle panel corresponds to antibody sequences from Group 1 donors, segregated by B cell subset. The rightmost panel corresponds to peripheral blood antibody sequences from HIV-infected donors (n=4). (A) Peripheral blood antibody sequences were grouped by HCDR3 length (in amino acids) and the average N-addition length and P-addition length (both in nucleotides) was calculated for each HCDR3 length group. The mean length ± SEM is shown. Regression analysis of N-addition length produced a non-linear, exponential curve of best fit. Regression analysis of P-addition length produced a linear best fit. (B) Peripheral blood antibody sequences were grouped by HCDR3 length and the frequency of sequences encoding either diversity gene family 2 (D2) or diversity gene family 3 (D3) was calculated for each HCDR3 length group. The mean frequency of D2/D3 gene family use ± SEM for each HCDR3 length group is shown. Regression analysis produced a non-linear, sigmoidal curve of best fit. (C) Peripheral blood antibody sequences were grouped by HCDR3 length and the frequency of sequences encoding joining gene 6 (J_H6) was calculated for each HCDR3 length group. The mean frequency of D2/D3 gene family use ± SEM for each HCDR3 length group is shown. Regression analysis produced a non-linear, sigmoidal curve of best fit. (D) The frequency of sequences encoding both the J_H6 germline gene and D2/D3 germline gene family members was determined for each HCDR3 length group. The mean frequency of J_H6/D2/D3 gene family use ± SEM for each HCDR3 length group is shown. Non-linear regression analysis produced a sigmoidal curve of best fit.

infrequency of long HCDR3s in both memory subsets, resulted in trends that were less robust. The pattern of diversity gene use in long HCDR3s was somewhat surprising, since D3-16, which is not increased in long or very long HCDR3s, is two amino acids longer than any of the three preferred D genes. Further, four additional D genes, D2-8, D3-9, D3-10 and D3-22 are the same length as the three preferred genes, but are not significantly more common in long or very long HCDR3s than in the total repertoire. Thus, while the D2 and D3 gene families encode the longest D genes found in the repertoire, the increased frequency of only a select few of the diversity genes in these families indicates that length is not the only factor driving the increased frequency of these D gene segments in long and very long HCDR3 repertoires.

Increased HCDR3 length correlated with genetic features associated with recombination.

I next grouped all of the peripheral blood antibody sequences from Group 1 and Group 2 healthy donors by HCDR3 length and determined the average N-addition length and P-addition length (Figure 26A, left panels) for each HCDR3 length group. Both features showed positive correlations, with N-addition increasing exponentially (r^2=0.96) and

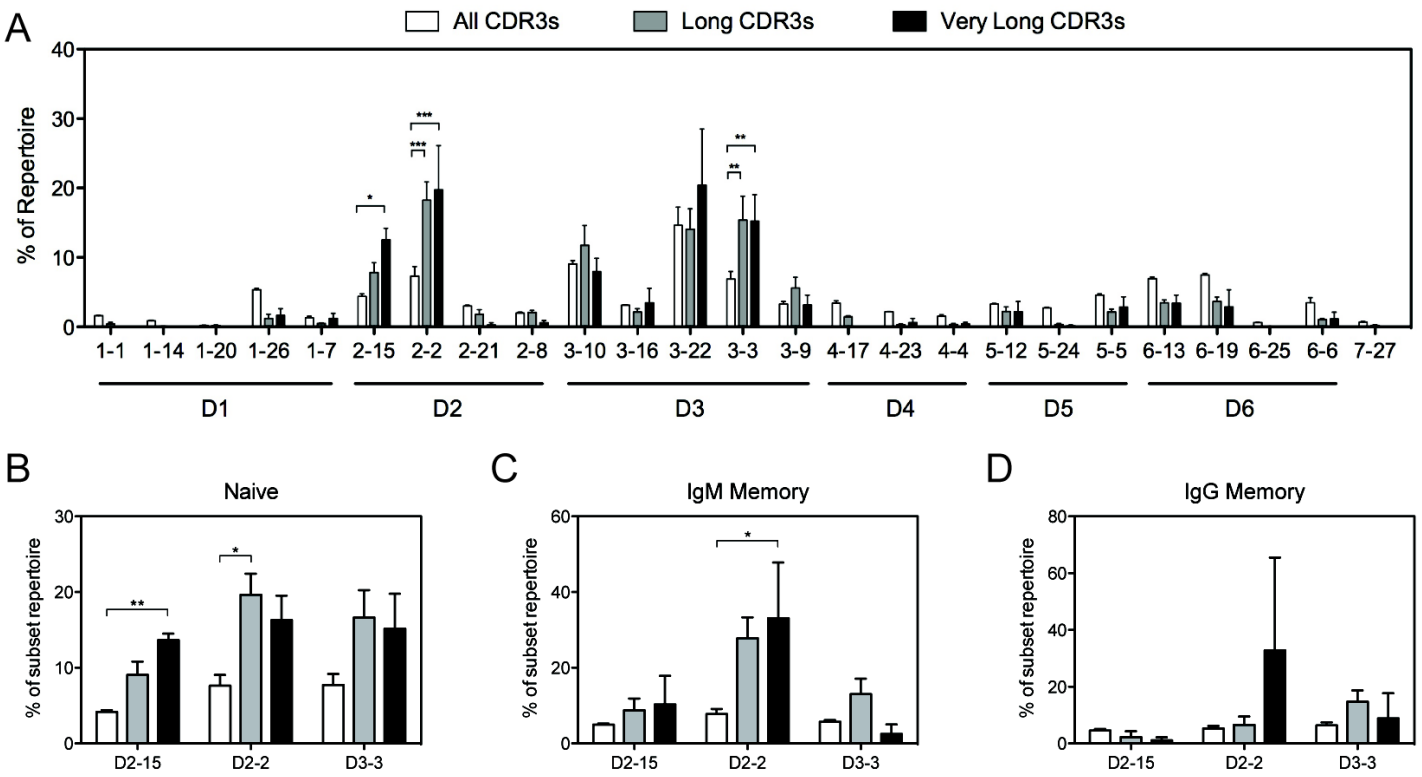


Figure 27. Frequency of D gene use in long and very long HCDR3s. Antibody sequences from Group 1 healthy donors (n=4) were assembled into the following three groups by HCDR3 length: (1) all HCDR3s, which contains all sequences of any HCDR3 length; (2) long HCDR3s, which contains only sequences with a HCDR3 length ≥ 24 amino acids (at least two standard deviations above the mean HCDR3 length for the entire repertoire); or (3) very long HCDR3s, which contains only sequences with a HCDR3 length ≥ 28 amino acids (at least three standard deviations above the mean). The frequency of diversity gene use in each of the three groups was determined for Group 1 healthy donor sequences (A). For the diversity genes that showed significant increases in (A), frequency was calculated for the naive (B), IgM memory (C), and IgG memory (D) subsets from the same Group 1 donors. The mean frequency ± SEM is shown. All HCDR3 lengths were calculated using the IMGT numbering system. The p values were determined using a two-way ANOVA with Bonferroni post tests. All statistically significant differences are indicated. * = $p<0.05$, ** = $p<0.01$, ***=$p<0.001$

P-addition increasing linearly (r^2=0.35, p<0.0001) with increasing HCDR3 length. Similar correlations were seen when analyzing individual B cell subsets from Group 1 healthy donors (Figure 26A, middle panels) and peripheral blood antibody sequences from HIV-infected donors (Figure 26A, right panels).

Following the observation that germline gene usage was skewed in long HCDR3s (Figure 25), I examined the use of D2/D3 (*i.e.*, D2 or D3) family genes as a combined group more closely. We first examined peripheral blood antibody sequences from Group 1 and Group 2 healthy donors (Figure 26B, left panel), and non-linear regression analysis revealed a sigmoidal correlation between D2/D3 gene use (r^2=0.80) and HCDR3 length, with combined D2/D3 gene use exceeding 80% in the longest HCDR3s. The increased use of D2/D3 genes was not surprising, as the D2 and D3 nucleotide sequences are the longest in diversity gene families. This correlation also was seen when analyzing the individual subsets of Group 1 healthy donors (Figure 26B, middle panel) and peripheral blood antibody sequences from HIV-infected donors (Figure 26B, right panel; r^2=0.61). Analysis of J_H6 gene use in the peripheral blood antibody repertoire of Group 1 and Group 2 healthy donors (Figure 26B, left panel) revealed a sigmoidal correlation between J_H6 gene use and increasing HCDR3 length (r^2=0.87). J_H6 was used in less than 10% of the shortest HCDR3s, but was present in over 75% of the longest HCDR3s. I observed a similar correlation when analyzing the individual subsets of Group 1 healthy donors (Figure 26C, middle panel) and peripheral blood antibody sequences from HIV-infected donors (Figure 26C, right panel; r^2=0.83). The observed preference for J_H6 was expected, as J_H6 is by far the longest J gene, adding as many as five more codons to the HCDR3 than the shortest J genes. I also examined the peripheral blood repertoire of Group 1 and Group 2 healthy donors to determine the frequency of sequences that use both J_H6 and D2/D3 genes (Figure 26D, left panel) and found a strong sigmoidal correlation between combined J_H6 and D2/D3 use (r^2=0.81). While combined J_H6 and D2/D3 use was unusual in the shortest HCDR3s (3.0%),

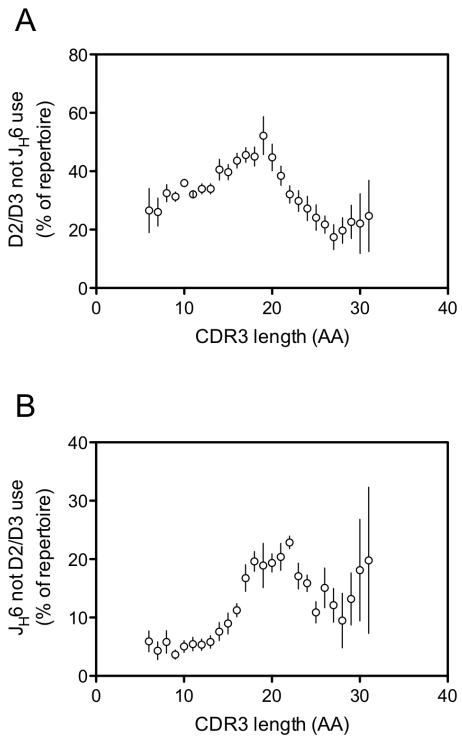


Figure 28. Limited preference for isolated use of D2/D3 or J_H6 in the longest HCDR3s. Peripheral blood antibody sequences from Group 1 (n=4) and Group 2 (n=3) healthy donors were grouped by HCDR3 length (in amino acids) and the frequency of (A) sequences encoding germline genes from the D2 or D3 families but not J_H6 or (B) sequences encoding J_H6 but not either of the D2 or D3 families. The mean frequency ± SEM is shown. All HCDR3 lengths were calculated using the IMGT numbering system.

D2-J_H6 and D3-J_H6 encoded antibodies comprised the majority of the repertoire of the longest of HCDR3s (58.8%). The trend toward increased use of J_H6 and D2/D3 also was seen in individual subsets of Group 1 healthy donors (Figure 26D, middle panel) and in HIV-infected donors (Figure 26D, right panel; r^2=0.81). Further analysis of Group 1 and Group 2 healthy donors revealed a correlation between HCDR3 length and use of J_H6 in sequences that do not incorporate D2/D3 (Figure 28A) and between HCDR3 length and in use of D2/D3 in sequences that do not incorporate J_H6 (Figure 28A), but the positive correlation only extended to HCDR3 lengths of approximately 20 amino acids. This finding suggested that while use of one germline gene (either J_H6 or D2/D3) is sufficient to allow generation of relatively long HCDR3s, both germline genes are required to generate the longest HCDR3s.

Diversity gene reading frame 2 is used preferentially in long HCDR3s.

I next analyzed the reading frame preferences in long HCDR3s. We used the ImMunoGeneTics (IMGT) method for calculating the D gene reading frame, which determines the reading frame based on the first codon of the D gene nucleotide sequence. The functional reading frame equivalent to IMGT reading frame 2 (RF2) has been shown to

be the most common reading frame in the overall repertoire (Ivanov et al., 2005; Ippolito et al., 2006; Schelonka et al., 2008; Zemlin et al., 2008; Schroeder et al., 2010), however, there was a significantly increased preference for RF2 in long HCDR3s in the peripheral blood antibody repertoire of Group 1 and Group 2 healthy donors (Figure 29A, left panel; $p<0.001$) and in HIV-infected donors (Figure 29A, right panel; $p<0.05$). A significant increase in RF2 use was also seen in very long HCDR3s in both healthy and HIV-infected donors (Figure 29A; $p<0.001$ and $p<0.01$, respectively). Although this reading frame is identified by IMGT as RF2, alternate methods of determining the reading frame, which are based on analysis of the amino acid sequence instead of the nucleotide sequence, produce a different reading frame nomenclature. The reading frame identified as RF2 by IMGT would be identified as RF1 using the alternate "functional" reading frame determination system. To keep confusion to a minimum, the IMGT nomenclature will be used for the remainder of this report. The increased RF2 use in long HCDR3s was unexpected because alternating the reading frame of the D gene should not affect the overall sequence length significantly. However, when analyzing the peripheral blood repertoire of Group 1 and Group 2 healthy donors, I discovered that the frequency of RF2 in the longest HCDR3s (Figure 29B, left panel; 69%) was over twice the frequency of RF2 in the shortest HCDR3s (28%). A similar pattern was seen in HIV-infected individuals (Figure 29B, right panel).

I next considered whether RF2 was selected more frequently in long HCDR3s because use of RF2 may have allowed for more efficient in-frame recombination with the highly preferred J_H6 gene. Examination of long HCDR3s in Group 1 and Group 2 healthy donors showed a similarly strong preference for RF2 (74%, $r^2=0.55$) in the longest HCDR3s of recombinants that did not use the J_H6 gene (Figure 29C, top left panel) than was seen in the total repertoire. The same trend also was seen in sequences that used J_H6 (Figure 29C, bottom left panel; 75%, $r^2=0.44$). A similar pattern of RF2 use in the presence or absence of J_H6 also was seen in cells from the HIV-infected subjects (Figure 29C, right panels). Thus,

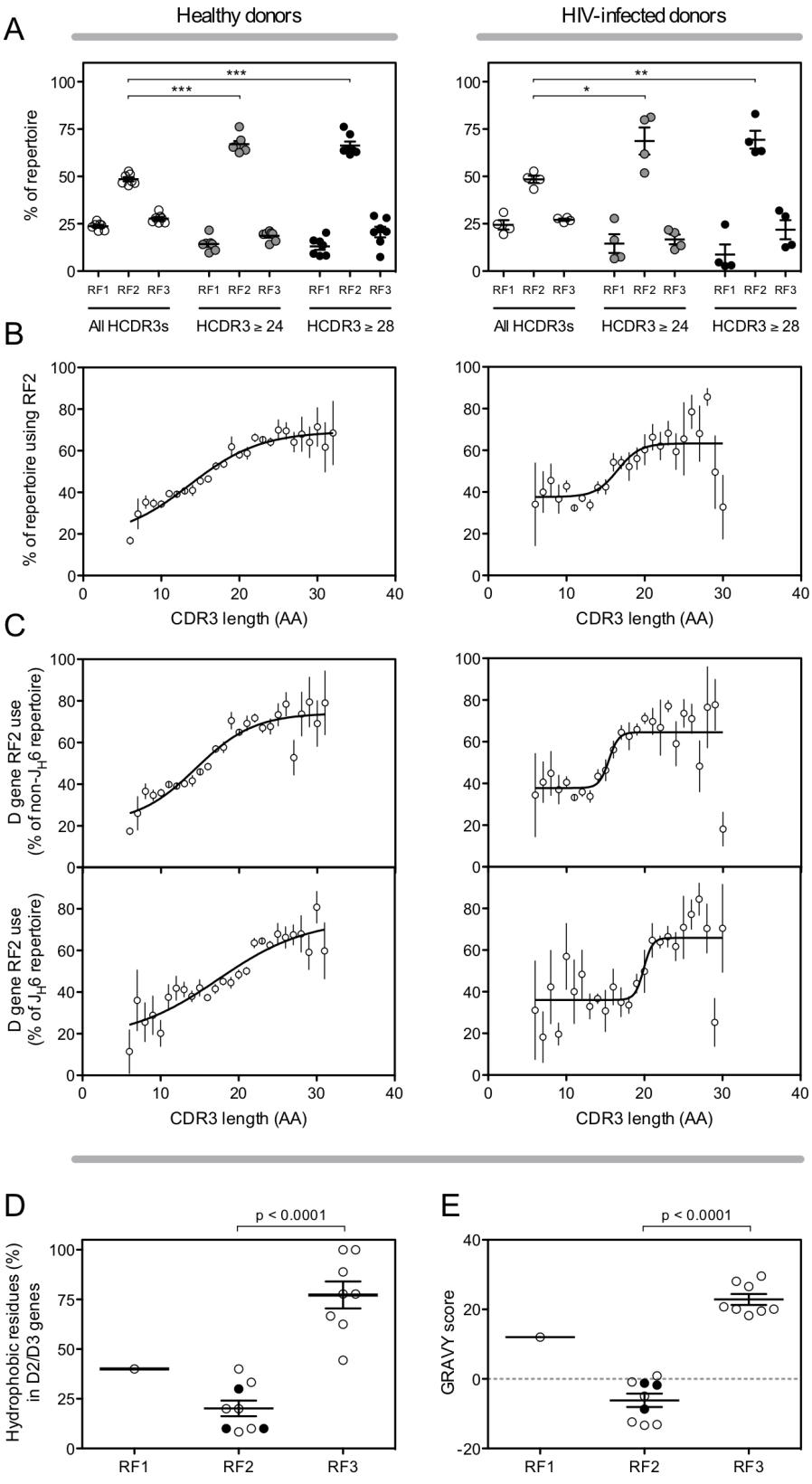

A
Healthy donors
HIV-infected donors
% of repertoire
RF1 RF2 RF3 RF1 RF2 RF3 RF1 RF2 RF3
All HCDR3s
HCDR3 ≥ 24
HCDR3 ≥ 28
B
% of repertoire using RF2
CDR3 length (AA)
C
D gene RF2 use (% of non-J_H6 repertoire)
D gene RF2 use (% of J_H6 repertoire)
CDR3 length (AA)
D
Hydrophobic residues (%) in D2/D3 genes
p < 0.0001
RF1 RF2 RF3
E
GRAVY score
p < 0.0001
RF1 RF2 RF3

Figure 29. (previous page) Long HCDR3s preferentially use reading frames (RF) that result in reduced hydrophobicity. (A) Peripheral blood antibody sequences from Group 1 and Group 2 healthy donors (left panel) or HIV-infected donors (right panel) were assembled into three HCDR3 length groups: (1) all HCDR3s; (2) HCDR3s of at least 24 amino acids; and (3) HCDR3s of at least 28 amino acids. (B) The percentage of sequences within each HCDR3 group using reading frame 2 of the diversity gene (RF2) was calculated for each HCDR3 length group. Non-linear regression analysis produced a sigmoidal curve of best fit (r^2=0.84). (C) Sequences that do not encode the joining gene J_H6 (top panel) or do encode J_H6 (bottom panel) were grouped by HCDR3 length and RF2 use within each HCDR3 length group was determined. The mean frequency ± SEM is shown. Non-linear regression analysis produced a sigmoidal curve of best fit. (D) The percentage of hydrophobic residues was calculated for each reading frame of every functional (lacking stop codons) diversity gene in the D2 and D3 germline gene families. The mean percentage ± SEM is shown for each reading frame. The RF2 hydrophobicity of the diversity genes which were shown to be increased in long HCDR3s are indicated by filled circles. The p values were determined using a Student's two-tailed t-test. (E) The grand average of hydropathicity (GRAVY) was calculated for each functional reading frame of each D2 and D3 gene. A positive GRAVY score indicates hydrophobicity, and a negative GRAVY score indicates hydrophilicity. The mean GRAVY score ± SEM is shown for each reading frame. The RF2 GRAVY scores of the diversity genes that were shown to be increased in long HCDR3s are indicated by filled circles. The p values were determined using Student's two-tailed t-test. All statistically significant differences are indicated. * = p<0.05, ** = p<0.01, *** = p<0.001

the observed RF2 preference in long HCDR3s was not primarily due to the need to form in-frame recombinations with J_H6.

Antibodies with long, hydrophobic HCDR3s often possess autoreactive properties (Crouzier et al., 1995; Aguilera et al., 2001; Wardemann et al., 2003; Haynes et al., 2005), and RF2 has been shown to be preferred in the antibody repertoire likely due to increased tyrosine frequency and lower hydrophobicity than other reading frames (Ivanov et al., 2002; Zemlin et al., 2008). Diversity genes in the D2 and D3 families are enriched for tyrosine residues, although the three D genes with the highest tyrosine content (D3-10, D3-16 and D3-22) were not among the preferred D genes (data not shown). I next analyzed the frequency of hydrophobic residues (Figure 29D) and the grand average of hydropathicity (GRAVY, Figure 29E) of each functional reading frame for the D2 and D3 families, with special attention paid to the three germline D genes that were found most often in long HCDR3s (designated by filled circles). Although the broader results validate previous data (namely that RF2 is much less hydrophobic than other reading frames), RF2 of the three preferred germline genes is not substantially less hydrophobic than RF2 of the other, non-preferred, D2/D3 gene members. Thus, although hydrophobicity and tyrosine frequency may

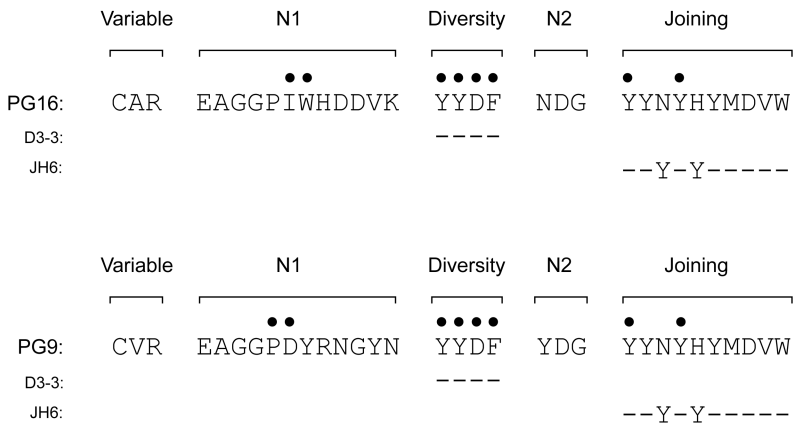


Figure 30. Amino acid residues in J_H6 and RF2 of D3-3 germline gene segments are critical to binding and neutralization of HIV by long HCDR3-containing antibodies PG9 and PG16. The HCDR3 amino acid sequences of HIV-specific mAbs PG9 and PG16 are shown aligned to the amino acid sequences of germline D3-3 and J_H6 genes. Dashes in the alignments indicate conservation with the respective PG antibody. Residues shown to be critical to binding or neutralization of HIV, defined as ≥10-fold decrease in either binding or neutralization when mutagenized are indicated by filled circles.

drive the overwhelming prominence of RF2 in the normal antibody repertoire, they do not seem to account for the strong preference for the three highly preferred D genes in the long HCDR3 repertoire.

Amino acid residues critical to binding and neutralization of HIV by broadly neutralizing antibodies PG9 and PG16 are encoded by J_H6 and D3-3 germline genes.

The broadly neutralizing HIV antibodies PG9 and PG16 contain the longest HCDR3s of any antigen-specific human monoclonal antibodies described to date (Walker et al., 2009). I considered the genetic basis for development of these unusual antibodies in light of the information presented above on the typical origin of long HCDR3s. These clonally related antibodies both possess HCDR3 regions containing 30 amino acids, which is twice the mean HCDR3 length in the total repertoire. Remarkably, both PG9 and PG16 antibodies use D3-3, one of the three D genes highly preferred in long HCDR3s, and J_H6. Further, both antibodies position the diversity gene in RF2. I performed amino acid sequence alignments with PG9 or PG16 and the corresponding germline D and J genes (Figure 30). Based on

previous studies (Pejchal et al., 2010), I identified eight critical residues in these antibodies for which a ≥ 10-fold decrease in either binding or neutralization occurred when those single amino acids were mutagenized. Interestingly, six of the eight critical residues were encoded by the germline sequence of the D and J genes. Although two additional crucial residues appear to be encoded by N-addition, it is impossible to rule out the possibility that these residues are the consequence of a post-recombination insertion event. Thus, the molecular basis for development of these most broadly neutralizing antibodies for HIV using very long HCDR3s was not a rare occurrence of unusual mutations. Instead, these antibodies were derived from a typical, almost canonical, selection of a D3-3 gene using RF2 and J_H6, and much of the high affinity of these antibodies derives from interactions mediated by unmutated germline-encoded residues.

Somatic Hypermutation-Associated Insertions and Deletions

Frequency of in-frame insertions and deletions associated with somatic hypermutation.

I separately isolated naïve, IgM memory and IgG memory B cells from four healthy individuals using flow cytometric sorting, extracted total RNA and performed RT-PCR to amplify antibody genes from those cells, and subjected the resulting amplicons to high throughput DNA sequencing. After selecting only high-quality, non-redundant antibody sequences, I obtained a total of 294,232 naïve cell sequences, 161,313 IgM memory cell sequences and 94,841 IgG memory cell sequences.

I first analyzed the variable gene regions of each sequence for the presence of insertions and deletions that did not shift the reading frame. The frequency of non-frameshift insertions (1.8% and 1.9% for IgM memory and IgG memory, respectively; Figure 31A) and deletions (2.0% and 2.6%; Figure 31B) was similar in both memory cell subsets. The frequency of both insertions and deletions was reduced significantly in the naïve subset

when compared to either IgM or IgG memory subsets. This finding is consistent with previous data suggesting that non-frameshift insertions and deletions within the variable gene are associated with the somatic hypermutation process. (Goossens et al., 1998; Wilson et al., 1998a; Bemark and Neuberger, 2003)

Biased variable gene use in sequences containing somatic hypermutation-associated insertions and deletions.

I next examined the sequences containing somatic hypermutation-associated insertions and deletions (SHA indels) for evidence of biased variable gene use. The VH4 variable gene family was much more common in the population of sequences containing insertions (57%; Figure 31C) than in the total antibody repertoire (24%), while the

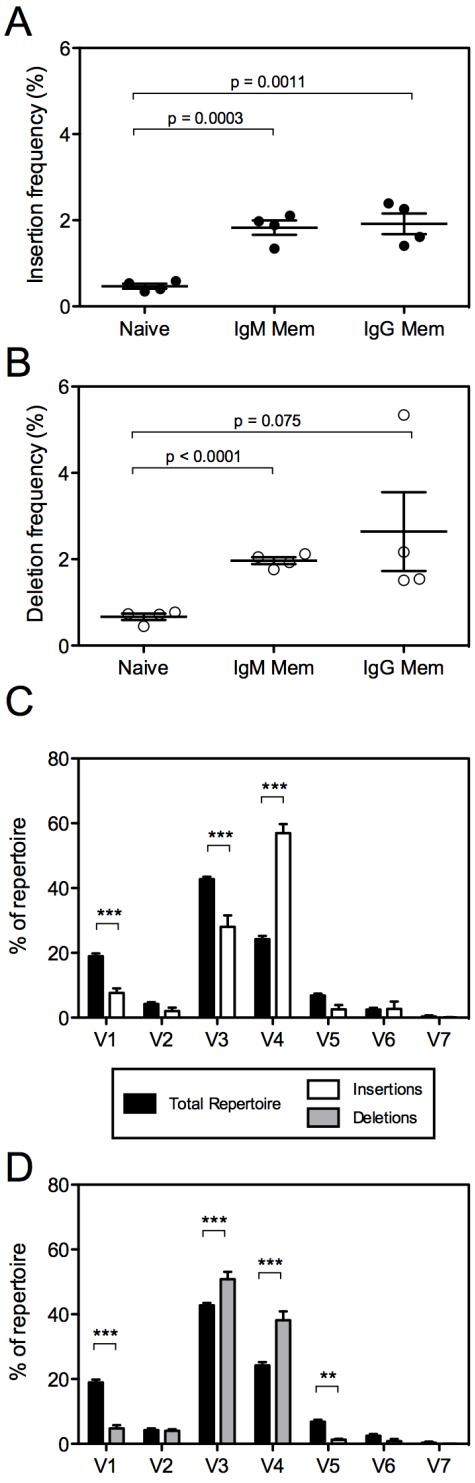


Figure 31. Frequency and variable gene use of sequences containing non-frameshift insertions or deletions. The frequency of (A) insertions or (B) deletions that were codon-length (*i.e.*, did not result in protein reading frame shift) was determined separately for the naïve, IgM memory and IgG memory subsets. The variable gene usage of VDJ gene recombinants containing insertions (C; white bars) or deletions (D; grey bars) was compared to the variable gene usage of the total repertoire (C and D; black bars).

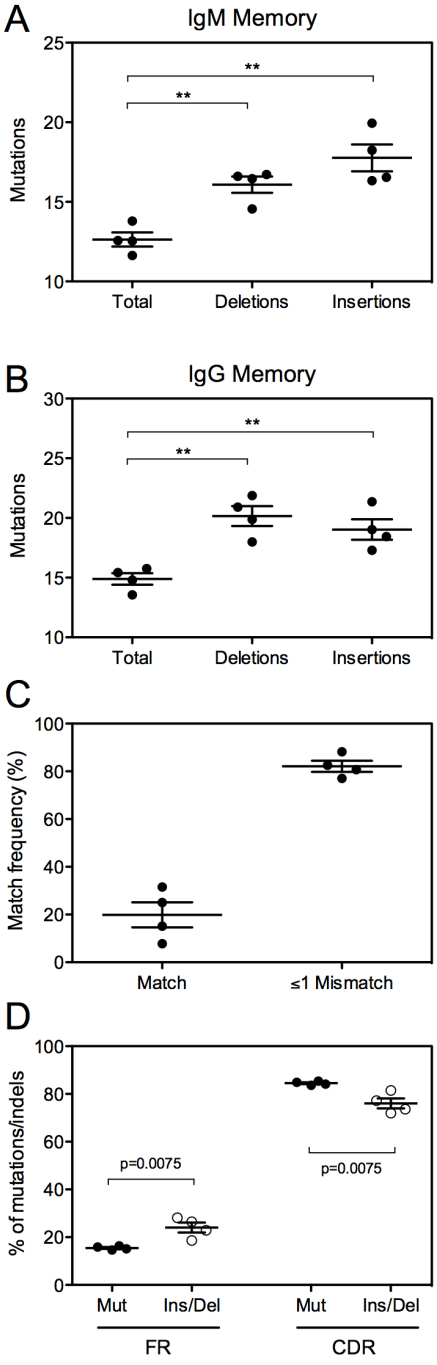
VH1 and VH3 families were observed less frequently in the insertion population (7.6% and 28%, respectively) than in the total repertoire (19% and 43%). In the population of sequences containing non-frameshift deletions, both VH3 and VH4 families (51% and 38%; Figure 31D) were more frequent than in the total repertoire (43% and 24%). The population of sequences with deletions also displayed reduced use of the VH1 family (4.8%) and VH5 family (1.3%) compared to the total repertoire (19% and 6.8%, respectively).

Antibody sequences containing SHA indels were highly mutated.

Since SHA indels are only rarely induced by the somatic hypermutation process, I hypothesized that antibody sequences containing SHA indels would display evidence of increased affinity maturation. I examined sequences containing SHA indels from both IgM memory (Figure 32A) and IgG memory (Figure 32B) subsets for evidence of increased affinity maturation. The total IgM memory subset displayed a mean mutation frequency of 12.6 mutations per sequence. Significantly higher mutation frequencies were seen in sequences from the IgM memory subset containing either SHA insertions (17.8; $p = 0.0017$) or SHA deletions (16.1; $p = 0.0022$). Sequences from the total IgG memory subset contained, on average, 14.9 mutations per antibody sequence. Much like the IgM memory subset, significantly higher mutation frequencies were seen in IgG memory sequences containing either SHA insertions (19.0; $p = 0.0056$) or SHA deletions (20.2; $p = 0.0015$).

Duplication of flanking sequence was observed in most non-frameshift SHA insertions.

For the population of sequences containing non-frameshift SHA insertions, the sequence immediately adjacent to the insertion (on either the 5′ or 3′ side of the insertion, hereafter referred to as the “flanking region”) was analyzed for homology to the sequence of the insertion. Since sequences containing insertions are highly mutated (Figure 32A-B), it is possible that additional mutations in the insertion sequence or the flanking region



accumulated following the insertion event. Therefore, flanking regions that identically matched the insertion sequence or flanking regions that contained a single mismatch were considered likely duplications. Although only 20% of insertion sequences identically matched flanking regions, 82% of sequences with insertions either identically matched or contained a single mismatch (Figure 32C), suggesting that sequence duplication was the primary mechanism of SHA insertions.

Mutations and SHA indels are differentially localized in framework and complementarity determining regions.

Although the somatic hypermutation process, which typically results in point mutations, and SHA indels have been shown to be linked, (Goossens et al., 1998;

Figure 32. Sequences containing SHA indels are highly mutated. Sequences from the IgM memory (A) or IgG memory (B) subsets were segregated into three groups: the total sequence pool for each subset (total), sequences containing insertions, or sequences containing deletions. The mean number of mutations for each of these groups was calculated for each of four healthy donor sequence pools. (C) For each insertion, the 5′ flanking sequence was analyzed for identity to the insertion sequence. The fraction of sequences with insertions that contained flanking regions that were a perfect match to the insertion sequence (match) or those that contained less than two mismatches (≤1 Mismatch) are plotted for each of four healthy donors. (D) Mutations and SHA indels (Ins/Del) were each grouped by localization in either framework (FR) or complementarity determining region (CDR). ** = $p<0.01$

Wilson et al., 1998a; Bemark and Neuberger, 2003) it is unclear whether the location of SHA indels is driven primarily by frequency of somatic hypermutation, or whether there are additional structural constraints that apply to SHA indels, but not substitutions. The somatic hypermutation process is known to preferentially target complementarity determining regions (CDRs) over framework regions (FRs) for a variety of reasons, including the increased presence of genetically encoded mutation hotspots. I analyzed the position of mutations and SHA indels (Figure 32D) and observed a significant increase in the fraction of SHA indels found in FRs and a decrease in the fraction of SHA indels found in CDRs when compared to mutations. 16% of mutations were found in FRs, while 24% of observed SHA indels were found in FRs ($p = 0.0075$). Conversely, 85% of mutations were found in CDRs, while only 76% of SHA indels were found in CDRs ($p = 0.0075$).

SHA indels revealed a hypervariable region 4 (HV4)-like region within FR3.

Sequences containing non-frameshift SHA insertions or deletions were analyzed for the position of the insertion or deletion. Insertions and deletions were grouped by codon position, and the frequency of insertions (Figure 33A) or deletions (Figure 33B) at each codon position was determined. Non-frameshift insertions and deletions were both concentrated in CDRs and the portion of FRs in close proximity to CDRs. The most common codon position for insertions was codon 35, which is in CDR1. The most common codon position for deletions was codon 57, which is in CDR2. Surprisingly, there was a cluster of codons in FR3 (codons 81-87) that contained a high frequency of deletions. This cluster of deletions was located in the middle of framework region 3a (FR3a, codon positions 78-93), which corresponds to hypervariable region 4 (HV4, also sometimes referred to as CDR4) in T cell receptors. A less prominent cluster of insertions also was seen in a similar location in FR3a.

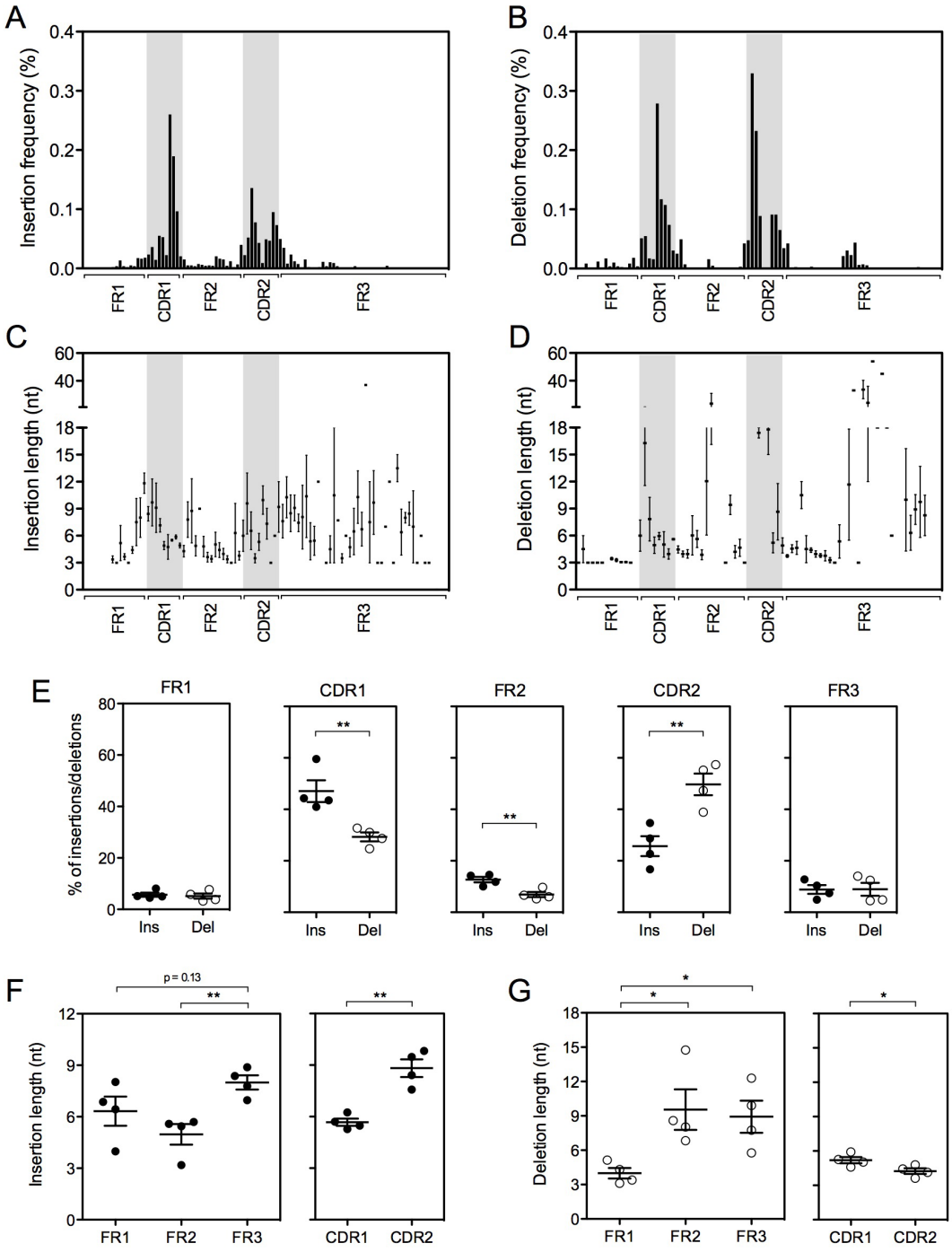


Figure 33. Genetic location and length distribution of non-frameshift insertions and deletions. The frequency of (A) insertions or (B) deletions at each codon of the variable gene reading frame was determined. The framework regions (FR) and complementarity determining regions (CDR) region were identified. The length distribution of insertions (C) or deletions (D) at each codon of the variable gene reading frame is shown. The mean value ± SEM for four donors is shown. (E) Comparison of the fraction of insertions (filled circles) or deletions (open circles) that were localized to each FR or CDR. The mean value ± SEM for four donors is shown. Comparison of the location of insertions (F) or deletions (G) for each FR or CDR is shown. The mean value ± SEM for four donors is shown.

I next performed a comparative analysis of the relative frequency of insertions and deletions located in the sequences encoding the two CDRs and three FRs that constitute the heavy chain variable (VH) gene (Figure 33E). The fraction of insertions observed in CDR1 was significantly higher than the fraction of deletions (47% of insertions were found in CDR1 while 29% of deletions were found in CDR1; $p = 0.008$), with a similar pattern seen in FR2 (13% of insertions and 7% of deletions; $p = 0.007$). In contrast, the fraction of deletions found in CDR2 was significantly higher than the fraction of insertions (50% of deletions and 26% of insertions; $p = 0.006$). There was no statistically distinguishable difference between the fraction of insertions and deletions in either FR1 or FR3.

Similar localization of insertions and deletions.

We again clustered non-frameshift insertions and deletions by codon position and calculated the mean insertion length (Figure 33C) for each codon position. As seen with insertion frequency, long insertions tended to concentrate in CDRs and in the portions of FRs that are immediately proximal to CDRs. An additional region containing a high concentration of long insertions and deletions was observed between codons 82 and 97 in FR3. Analysis of the mean insertion length of the three FRs (Figure 33F) revealed a trend toward longer insertions in FR3 when compared to FR1 ($p = 0.13$) and a significant increase in insertion length in FR3 when compared to FR2 ($p < 0.01$). Analysis of the mean insertion length of the two CDRs revealed a significant increase in insertion length in CDR2 when compared to CDR1.

Analysis of deletion length at each codon position (Figure 33D) produced results that were similar to the insertion length distribution, with increased deletion lengths found in CDR1, CDR2 and FR3. A region between codons 82-97 contained extremely long deletion events, with codon 76 displaying a mean deletion length of 54 nucleotides and codon 78 displaying a mean deletion length of 45 nucleotides. Interestingly, the location of the region

of long FR3 deletions corresponds to the location of increased FR3 deletion frequency. While the distribution of long insertions and deletions was largely similar in pattern, there was a short region between codons 51 and 55 in FR2 that contained very long deletions, and there was no corresponding region within FR2 for which long insertions were observed. Analysis of the mean deletion length of the three FRs (Figure 33G) revealed significantly longer deletions in FR2 and FR3 when compared to FR1 ($p < 0.05$). We also observed a small but significant increase in the deletion length in CDR1 when compared to CDR2. As with insertions, the CDR with lower alteration frequency (CDR2 for insertions, CDR1 for deletions) contained a significantly longer mean insertion or deletion length.

Structural display of insertion and deletion frequency and length distribution revealed

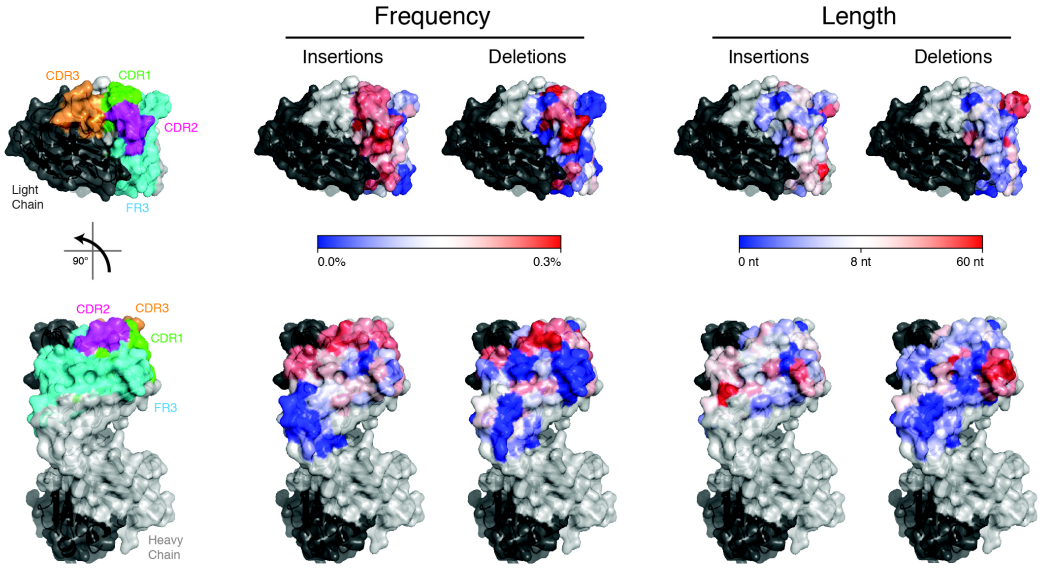


Figure 34. Structural location of non-frameshift insertions and deletions. A space-filling representation of the high resolution structure of the representative human Fab del2D1 determined by x-ray crystallography (citation) is shown at left. The del2D1 antibody light chain is colored dark grey. The del2D1 antibody heavy chain CDR1 (green), CDR2 (magenta), CDR3 (orange) and FR3 (cyan) regions are indicated, with the remaining heavy chain regions colored light grey. The insertion and deletion frequency were determined for each codon position in the variable gene. In the middle two panels, the surface of del2D1 is colored to indicate insertion or deletion frequency. The mean insertion and deletion length were calculated for each codon position in the variable gene. In the right two panels, the surface of del2D1 is colored to indicate mean insertion or deletion length in nucleotides (nt).

regions of antibody structural plasticity.

To gain a better understanding of the location of insertions and deletions in the context of a fully folded antibody protein, we mapped the frequency and length distribution of both insertions and deletions onto a space-filling model of a representative antibody (Figure 34). The model we used was derived from crystallographic structural data for the human

influenza virus specific monoclonal antibody (mAb) 2D1 that we had isolated in our laboratory and previously reported (Krause et al., 2011). Insertion or deletion frequency was determined by calculating the log10 of the frequency for each codon position and represented as a blue_white_red gradient on the surface of the mAb 2D1 structure. Insertion and deletion frequency hotspots were observed at the top of the protein, with peak insertion and deletion frequencies appearing near the apex of the CDR1 and CDR2 loops. The side orientation revealed

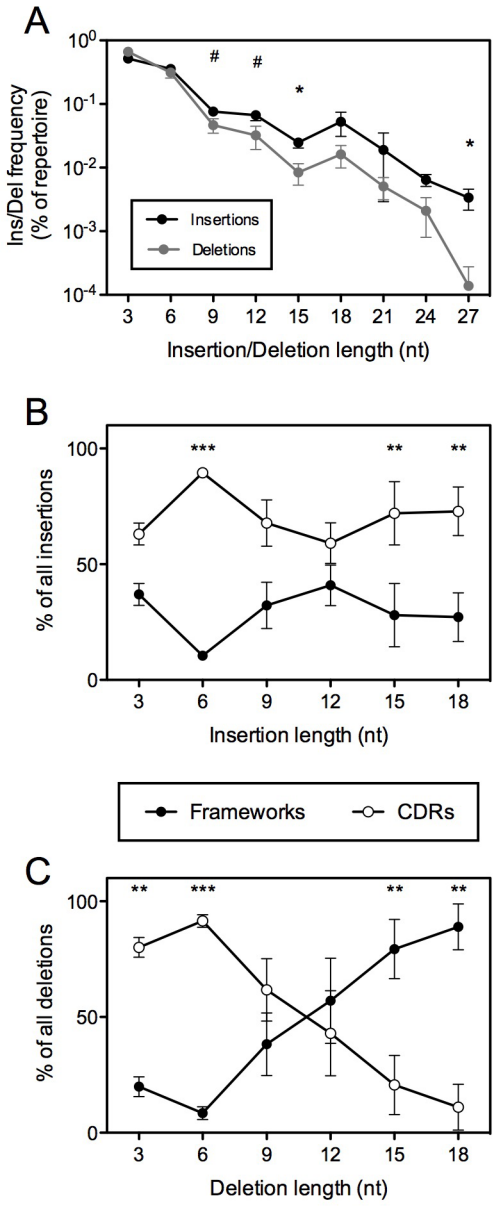


Figure 35. Difference in tolerance of long insertions and deletions in FRs and CDRs. (A) Non-frameshift insertions (black) or deletions (grey) were grouped by length and the frequency of each insertion or deletion length was calculated. The mean value ± SEM for four donors is shown. Non-frameshift insertions (B) or deletions (C) were grouped by length and the location of each insertion or deletion length to FRs (solid circles) or CDRs (open circles) was determined. The mean value ± SEM for four donors is shown.
= p<0.10, * = p<0.05, ** = p<0.01, *** = p<0.001

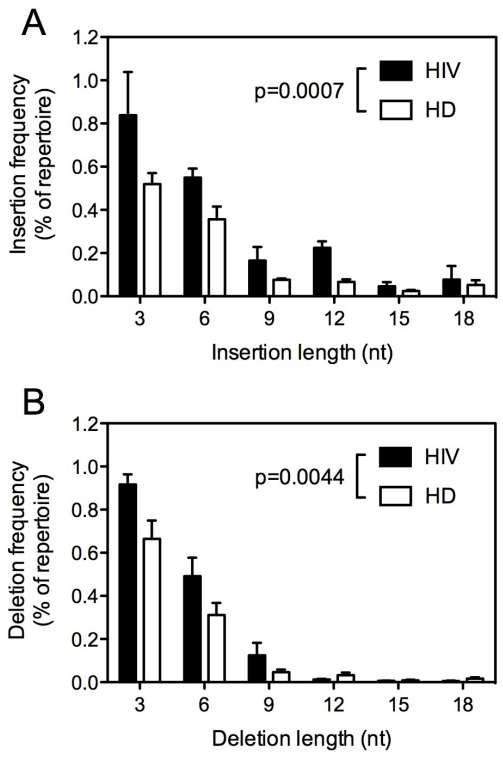


Figure 36. SHA indel frequency in HIV-infected and HIV-uninfected individuals. Sequences from HIV-infected donors (black bars; n=4) contained a higher frequency of insertions (A) or deletions (B) than HIV-uninfected healthy donors (white bars; n=4).

a reduced insertion and deletion frequency in the highly structured framework regions, with the lone framework hotspot occurring in a surface-exposed loop region of FR3.

Insertion and deletion length distribution was determined by calculating the log2 of the mean insertion length for each codon position and represented as a blue_white_red gradient on the surface of the mAb 2D1 structure. The longest insertions and deletions were focused in FR3, and were isolated to loop and short alpha-helical regions.

Long deletions were less frequent than long insertions and were tolerated poorly in CDRs.

I examined the ability of the antibody repertoire to generate and maintain sequences with long insertions and deletions. Insertions and deletions were both grouped by length (in nucleotides) and selected lengths for which we had at least 100 representative sequences with that length of insertion or deletion. The frequency of insertions and deletions was plotted for each length (Figure 35A), which revealed a significantly higher frequency of long insertions when compared to frequency of long deletions.

I next investigated whether or not there was a structural reason for the greater tolerance of long insertions over long deletions. Insertions and deletions were again grouped by length (in nucleotides) and plotted the frequency of insertions (Figure 35B) or deletions

(Figure 35C) by location in either FR or CDR. We found that both long and short insertions were concentrated in CDRs, with less than 30% of the longest insertion events occurring in FRs. In contrast, however, long deletions were highly concentrated in FRs, with 89% of the longest deletions occurring in FRs. This strong preference against long deletions in CDRs is likely due to the limited length of the CDR loops. Most CDR1 and CDR2 loops are only 8-9 amino acids long, which likely restricted the ability of these CDR loops to structurally accommodate long deletions.

SHA indels are more frequent in HIV-infected individuals

Since many bnAbs contain SHA indels, the peripheral blood antibody repertoires of four HIV-infected individuals were examined for the frequency of SHA indels and compared with the repertoires of HIV-uninfected individuals. There was a consistent increase in the frequency of SHA insertions of all lengths in HIV-infected individuals compared to HIV-uninfected individuals (Figure 36A; $p=0.0007$). There was also an increase in the frequency of SHA deletions across the spectrum of SHA deletions lengths in HIV-infected individuals compared to HIV-uninfected individuals (Figure 36B; $p=0.0044$).

Discussion

Broadly neutralizing antibodies against HIV are rarely found and often contain genetic or structural features that are unique or infrequently observed. SHA indels and exceptionally long HCDR3s, while uncommon in the normal circulating antibody repertoire, are frequently found in broadly neutralizing HIV antibodies. Much effort is currently being expended on the design of potential vaccine strategies that effectively elicit a broadly neutralizing anti-HIV antibody response. It is unclear, however, whether the designed immunogens should be constructed to induce the unique genetic features found in HIV antibodies, or whether the presence of such unique genetic features in the antibody

repertoire of uninfected individuals would enable selection of pre-existing antibodies with the required genetic features by an appropriately designed immunogen. The data presented in this chapter indicates that two frequently observed genetic features of HIV antibodies, SHA indels and exceptionally long HCDR3s, are found in the antibody repertoires of HIV-uninfected individuals.

Comparison of long HCDR3s in the HIV-infected and HIV-uninfected repertoire

Although designing a strategy for selecting or inducing antibodies with long HCDR3s might be a groundbreaking step for vaccine design, very little is known about the genetic origin of such antibodies. Two general models have been proposed for the generation of antibodies with long HCDR3s. First, it is possible that long HCDR3s are generated through the introduction of multiple short insertion events during the somatic hypermutation process (Wilson et al., 1998a; Reason and Zhou, 2006). Indeed, we have shown recently that a genetic insertion was a critical feature that mediated affinity maturation and acquisition of neutralizing potency for a human antibody that inhibits influenza virus (Krause et al., 2011), although this was a short insertion that did not cause a long HCDR3. Alternatively, it is possible that long HCDR3s could be created at the time of recombination through a combination of extensive incorporation of non-templated nucleotides during N- and P-addition and the selective use of longer germline gene segments. In this setting, the extended length of the HCDR3 would be established without the need for the affinity maturation process. In this study, I found that antibodies containing long HCDR3s are created primarily at recombination and not through affinity maturation. In addition, several genetic features were identified that are frequently seen in antibodies with long HCDR3s but are uncommon in the rest of the antibody repertoire. Finally, analysis of long HCDR3 encoding antibody sequences from HIV-infected donors produced results that closely mirrored those from healthy donors. These results suggested that antibodies with long

HCDR3s in HIV-infected individuals also typically were generated at the time of recombination.

These data show that increased HCDR3 length did not correlate with either increased mutation count or insertion frequency in the peripheral blood antibody population, although mutation count and insertion frequency were themselves strongly correlated. Further, these studies demonstrated the presence of B cells with receptors incorporating long HCDR3s in the naïve cell population. In fact, the naïve B cell subset contained a higher proportion of long HCDR3s than did the affinity-matured memory B cell population, suggesting that long HCDR3s are produced independent of the affinity maturation process. Together, the lack of correlation between increased HCDR3 length and increased number of insertions arising during affinity maturation and the presence of a sizeable fraction of long HCDR3s in the memory cell population strongly indicate that long HCDR3s are not primarily generated by repeated rounds of affinity maturation.

I also examined several events that occur during the recombination process and discovered many correlations between these molecular features of recombination and increased HCDR3 length. First, an increased number of nucleotides introduced by N- or P-addition were both found to be correlated with increased HCDR3 length. Next, germline diversity genes in the D2/D3 family and the germline joining gene J_H6 were found to be highly favored in long HCDR3s. In fact, over half of all of the longest HCDR3s used both of these germline sequence elements, while fewer than 5% of the shortest HCDR3s contained both sequence elements. Finally, somewhat surprisingly, RF2 was highly preferred in long HCDR3s. Previous work by several groups has indicated that long, hydrophobic HCDR3s often possess autoreactive properties (Crouzier et al., 1995; Aguilera et al., 2001; Wardemann et al., 2003; Haynes et al., 2005), so it seems likely that RF2 is used preferentially primarily because of the reduced hydrophobicity profile compared to other reading frames.

The observation that long HCDR3s are composed of conserved sequence elements is of critical importance for two reasons. First, these findings provide information that might be used in the design of strategies to selectively induce expansion of particular B cells encoding antibodies with long HCDR3s. Since long HCDR3s are generated using a limited set of germline gene segments, and since those germline segments are rarely used in short HCDR3s, immunogens designed to target these conserved sequence elements might induce an antibody response that is enriched in antibodies containing long HCDR3s. Second, knowledge of conserved genetic elements present in the majority of long HCDR3s provides a starting point for affinity maturation of these antibodies. For example, it has been suggested that one route to development of an HIV vaccine would be to identify structures of neutralizing epitopes and design immunogens that mimic these epitopes, with the goal of eliciting an antibody response focused on the desired neutralizing epitope (Burton, 2002; Douek et al., 2006; Burton, 2010; Walker and Burton, 2010). An alternative to this approach has been proposed, however, which consists of identifying the structural and genetic components of potently neutralizing antibodies and designing immunogens that gradually and specifically induce desired affinity maturation events that result production of broadly neutralizing antibodies. In effect, this alternative process would involve rationally guiding the affinity maturation process through the selective use of sequential immunizations (Pancera et al., 2010). This strategy would require detailed knowledge not only of the desired final product, in this case a PG9- or PG16-like broadly neutralizing antibody, but also of the genetic characteristics of the naïve predecessors of the desired broadly neutralizing antibody. The work presented here provides a substantial step toward realization of this alternative method of vaccine development. Until this study, little was known about the process of generating antibodies with long HCDR3s or the genetic characteristics of the naïve predecessors of such antibodies. I have identified conserved genetic elements in the

long HCDR3 antibody population that form a potential starting point from which rationally guided affinity maturation may begin.

In the case of mAbs PG9 and PG16, this potential is especially enticing. Since many of the residues in these mAbs that are critical for binding and neutralization are encoded in the germline gene sequence, little affinity maturation may be necessary to produce a potently neutralizing antibody. In fact, all but two residues identified in PG9 and PG16 as critical to binding and neutralization were present in the germline D3-3 or J_H6 genes, and those additional critical residues were generated by random N-addition. Accordingly, I suggest that it is highly likely that there are naïve antibodies in HIV-unexposed individuals that use D2/3 and J6 gene segments and by random happenstance of N-addition, encode many of the residues critical to PG-like neutralization. Thus, while germline reversions of PG9 and PG16 are non-neutralizing (Pancera et al., 2010), it is possible that the naive predecessors of PG-like antibodies will require only limited affinity maturation to gain neutralization capacity. If this is the case, efforts should be focused on selective induction of these rare antibodies containing long HCDR3s, rather than sequential immunization strategies to “build up” long HCDR3s with somatic insertions.

SHA indels in the HIV-uninfected repertoire

Many of the most broad and potently neutralizing HIV antibodies contain SHA indels that are critical for neutralization. However, B cells encoding antibodies with SHA indels are unusual in the peripheral circulation, with less than 2% of antibody sequences containing such insertion or deletion events. Due to their rarity, a comprehensive analysis of SHA indels has been difficult in the past. High throughput sequencing was used to determine the location and length distribution of SHA indels; for the most part, the location of SHA indels was similar to that of conventional somatic mutations. However, substantial differences in SHA indel location were identified that were likely related to structural constraints that apply

to SHA indels but do not apply to substitutions. This analysis analyses revealed regions of antibody structural plasticity, *i.e.*, regions that were able to accommodate addition or subtraction of sequence without compromising structural integrity.

With much effort being directed toward rational design of both antigens(Correia et al., 2010; Ofek et al., 2010; Wu et al., 2010a; Azoitei et al., 2011) and antibodies,(Diskin et al., 2011) it is critical to understand the regions of the antibody molecule that can withstand extensive alteration while maintaining the desired structural conformation. The broadly neutralizing HIV antibody PGT128 is exceptionally potent, and the neutralization potency requires the presence of a six amino acid deletion in HCDR2. The inducement of such a large deletion through vaccination is thought to be very difficult, however, this work has identified the frequency of such deletions in the HIV-uninfected repertoire. Although such deletions are rare, they are present, which raises the possibility of rationally designed vaccine designed to perform the presumably simpler task of selection of antibodies containing the required genetic features, rather than elicitation.

Interestingly, SHA indels were more frequently observed in HIV-infected individuals than in HIV-uninfected individuals. While it is possible that HIV infection specifically selects for and enhances the frequency of antibodies containing SHA indels, it is also possible that the increase in SHA indels is a side effect of chronic infection and is not specific to HIV. Chronic infections involve repeated exposure to antigen and likely result in repeated stimulation of the same clones of antigen-specific B cells. Since SHA indels are associated with the somatic hypermutation process, it is possible that the increase in SHA indels seen in HIV-infected donors was simply an artifact of repeated stimulation of a small number of HIV-specific B cell clones. To discriminate any HIV-specific effect from effects that are generally caused by chronic infection, further analysis of the antibody repertoires of individuals with chronic infections other than HIV (Hepatitis C, for example) must be performed.

CHAPTER V

STRUCTURAL HOMOLOGS OF THE BROADLY NEUTRALIZING HIV ANTIBODY PG9 IDENTIFIED IN THE PERIPHERAL BLOOD OF UNINFECTED INDIVIDUALS

Introduction

The clonally-related broadly neutralizing antibodies PG9 and PG16 contain extremely long HCDR3s with which they are able to penetrate the glycan shield on gp120 and interact directly with the gp120 protein (Figure 37). Both antibodies target a conserved epitope on the V1/V2 and V3 loops and interact directly with carbohydrates on the surface of gp120 (McLellan et al., 2011). Serum antibodies targeting glycan-dependent epitopes were found in approximately half of the tested individuals with broadly neutralizing activity, and depletion of these antibodies resulted in loss of neutralization potency (Gray et al.,

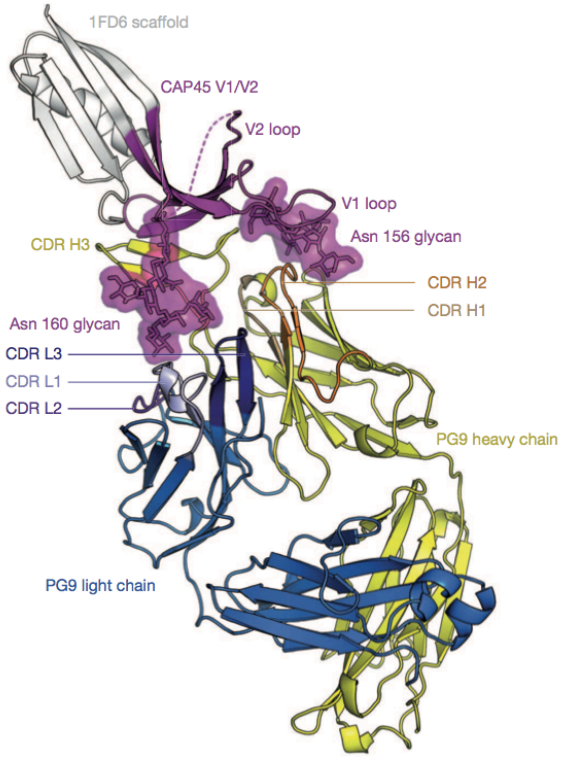


Figure 37. Crystal structure of PG9 in complex with Cap45 V1/V2 scaffold protein. The crystal structure of PG9 (heavy chain in blue, light chain in yellow) in complex with the CAP45 V1/V2 scaffold protein revealed the binding mode of PG9. The extremely long HCDR3 loop is able to reach past glycans at positions N156 and N160 (light purple) to contact the CAP V1/V2 protein (purple). Although most of the gp120 protein has been replaced by the 1FD6 scaffold (light grey), the V1/V2 loop region is in a position consistent with the virion particle located at the top of the page. While the HCDR3 reaches past glycans to interact directly with conserved regions of the polyprotein, light chain CDRs make direct contact with the glycan at position N160.

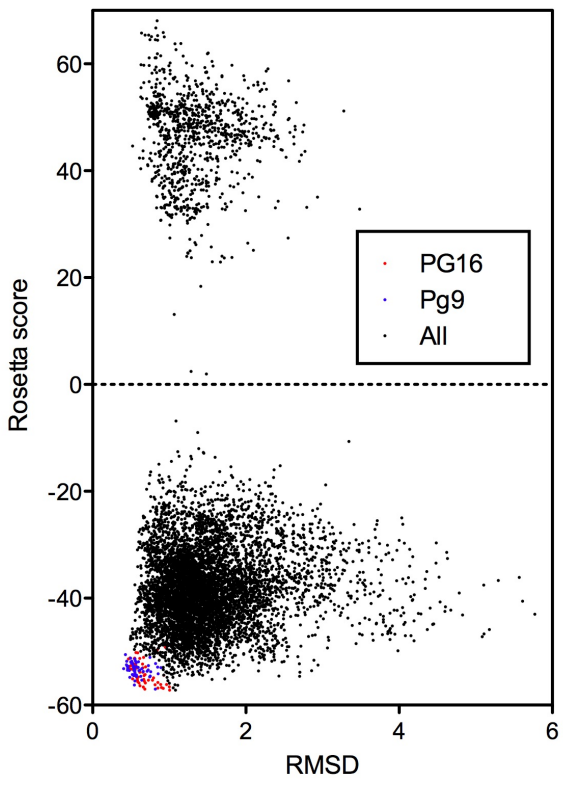


Figure 38. Most healthy donor HCDR3s are not predicted to accommodate a PG9-like structure. The RMSD and Rosetta Energy Score was plotted for each decoy of PG9 (blue), PG16 (red) and each healthy donor antibody sequence with a 30 amino acid long HCDR3 (black).

2009; Walker et al., 2010). PG9 and PG16 are able to more potently neutralize older virus isolates (isolated between 1985-1989) than VRC01, which targets the CD4bs, and contemporary virus isolates (2003-2006) are more resistant to neutralization by VRC01 than to neutralization by PG9 and PG16 (Euler et al., 2011). In addition to the exceptional length of the HCDR3, the loop contains unique secondary structural elements that are mediated by a complex hydrogen-bonding network. This secondary structure, referred to by some as an “ax-head” or “hammer-head” structure, is critical to the binding interaction and appears to be stable in both the antigen bound state and in the unbound state (Pancera et al., 2010; McLellan et al., 2011).

Antibodies containing long HCDR3s have previously been shown to be present in the peripheral blood repertoire of HIV-uninfected individuals (Briney et al., 2012a). However, neutralization by PG9 and PG16 depends not only on the presence of a long HCDR3, but also on the unique secondary structure within the HCDR3 loop. In collaboration with my computational colleague, Jordan Willis, I examined the peripheral blood repertoire to identify the presence of antibodies that contain structural homology to the unique HCDR3s of PG9 and PG16.

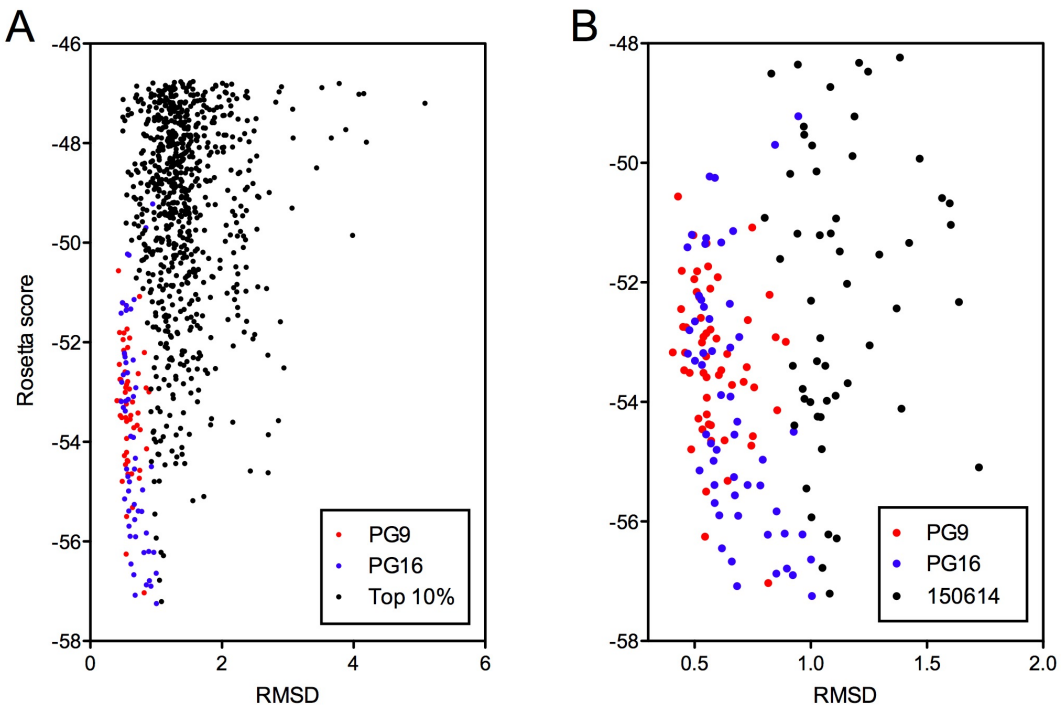


Figure 39. Long HCDR3s from HIV-uninfected donors are predicted to adopt a PG-like structure. Rosetta Energy Score and RMSD of (A) the top-scoring 10% of decoys or (B) all decoys of the single top scoring antibody sequence.

Results

The majority of 30 amino acid HCDR3s are not structural homologs of PG9 or PG16

Using ultra-high throughput HCDR3 sequencing, I isolated approximately 5.5 million unique HCDR3s from four individuals. The HCDR3 repertoire was examined for the presence of HCDR3s with a length of 30 amino acids, and 2200 were identified. Each HCDR3 was then threaded by Jordan onto a crystal structure of PG9 bound to a V1/V2 epitope scaffold by Jordan using the Rosetta software suite. To avoid capture of any particular modeling run in a local energy minimum as opposed to a global minimum, each sequence was threaded multiple times. Each individual run of a single sequence is referred to as a decoy (as in, multiple decoys were run for each sequence). Each decoy was given a

Rosetta Energy Score (RES), which is similar to change in Gibbs' Free Energy in that a lower score indicates an energetically favorable conformation. In addition to threading HCDR3 sequences from HIV-uninfected donors, Jordan also threaded the HCDR3s of PG9 and PG16. These positive controls allowed the definition of a threshold RES for which adoption of the PG9 conformation was likely. Based on the RES and root mean square deviation (RMSD, a measure of the spatial similarity of the final predicted structure to the PG9 crystal structure), the overwhelming majority of HIV-uninfected HCDR3s are predicted to be unable to conform to the PG9 HCDR3 structure (Figure 38).

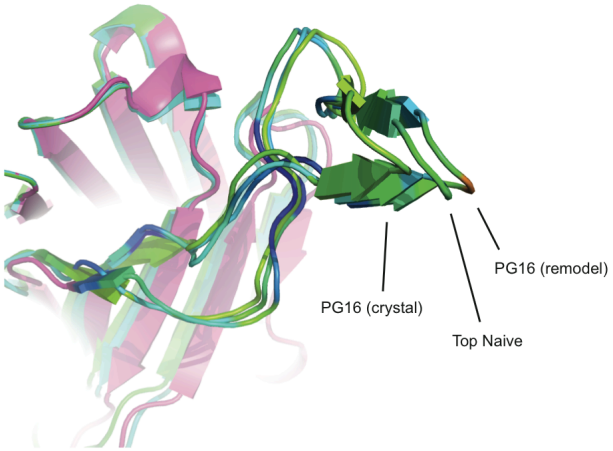


Figure 40. Structural homology between PG16 and naïve B cell sequences isolated from HIV-uninfected donors. The PG16 crystal structure, threading of the PG16 sequence onto the PG16 crystal structure (remodel) and the top scoring naïve sequence (Top Naïve) are shown. Each structure is colored according to the Rosetta Energy Score, with green indicating favorable RES and red indicating unfavorable RES.

Several HCDR3s from HIV-uninfected individuals are predicted to conform to the PG9 crystal structure

Limiting the plot of antibody decoys to just the to lowest scoring 10% of sequences revealed several sequences that several sequences approach the RES and RMSD values of PG9 and PG16 (Figure 39A). In fact, the best scoring sequence (the lowest average RES) has RES scores that are comparable to PG9 and PG16, although the RMSD is slightly higher (Figure 39B). While there is little homology between the HCDR3 sequence of the top scoring sequence and the HCDR3 sequence of PG9 or PG16, but the predicted structure of the antibody is very similar to that of PG16 (Figure 40).

The three best scoring HCDR3 sequences from HIV-uninfected individuals were used to generate chimeric antibody sequences in which the HCDR3 of PG16 was replaced the HCDR3 sequence from the HIV-uninfected individual. Of the three chimeric antibodies that were created, two appeared to contain somatically related HCDR3s and were unable to be expressed. The third sequence was expressed as a full-length IgG1 molecule paired with the PG16 light chain, but was not able to bind any of a panel of HIV isolates. This was not unexpected, however, since the homology of the healthy donor HCDR3 with PG16 HCDR3 extended only to the secondary structure of the HCDR3 loop, not to the individual amino acid side chains.

Discussion

In close collaboration with Jordan Willis, who performed all of the computation detailed in this chapter, I have discovered the presence of antibodies in the HIV-uninfected peripheral blood antibody repertoire that are predicted to be highly structurally homologous to the unique HCDR3 structure of the broadly neutralizing antibodies PG9 and PG16. Not only were these antibodies identified in the HIV-infected donors, but they were identified in the naïve B cell population, indicating that no somatic mutation is necessary to produce antibodies with predicted structural similarity to PG9 and PG16. Not only does this discovery provide useful information about the genetic origin and development of broadly neutralizing HIV antibodies, it also raises the possibility of a novel method of vaccine development. Most currently available vaccines are based on some form of viral mimicry, whether through the use of inactivated or attenuated virus, through the use of viral subunits, or through the use of scaffolded viral epitopes. Various forms of HIV viral mimicry have been tested, and each effort has been unsuccessful at generating potently neutralizing antibodies and protection from infection. Detailed knowledge of the HIV-uninfected repertoire and identification of

potential bnAb precursors in the HIV-uninfected repertoire may allow an alternate method of vaccine development in which immunogens are designed to target bnAb precursors and guide them toward the desired neutralizing antibody. While it is possible that the immunogens designed to target bnAb precursors may mimic the virus in some form, it is also possible, even likely, that immunogens designed to rationally direct the affinity maturation process have little homology to the original pathogen.

One could envision a multi-step vaccination procedure in which a panel of immunogens is introduced in sequence, with each immunogen designed to elicit a handful of somatic mutations in the progenitor antibody sequence. Each stage of the vaccination process would produce intermediate antibodies that become closer to the desired bnAb response with continued vaccination. Intriguingly, since less somatically mutated progenitors of bnAbs have been shown to be less broadly neutralizing than the bnAb itself (Wu et al., 2011), it is possible that the partially affinity matured intermediate antibodies that are elicited by the multi-step vaccination procedure may provide partial protection from infection even before the entire vaccination regimen is complete.

CHAPTER VI

DISCUSSION

Since the start of the HIV/AIDS outbreak over two decades ago, intense energy and extraordinary resources have ben invested in the search for a protective vaccine. All early attempts were either ineffective or, in the worst case, enhanced the possibility of infection. Recently, the RV144 trial in Thailand was modestly successful, but large gains in effectiveness must be made before a vaccine is suitable for widespread distribution.

The tremendous global success with other viral vaccines raises the question as to why HIV vaccine development has been so difficult. Many of the difficulties lie in distinct properties of this virus compared with other viruses. Foremost among these is HIV's enormous sequence diversity. Because of an error-prone reverse transcriptase, a high propensity for recombination, and an extremely rapid turnover in vivo, HIV's capacity for mutation and adaptation is enormous (Korber et al., 2001). Viruses even within the same clade may differ by up to 20%, and in places such as Africa where there are multiple prominent clades, circulating viruses can differ within the highly variable envelope protein by up to 38% (Korber et al., 2001; Walker and Korber, 2001). Indeed, the amount of HIV diversity within a single infected individual can exceed the variability generated over the course of a global influenza epidemic, the latter of which results in the need for a new vaccine each year. With more than 33 million people currently infected with HIV, and the need for a vaccine that simultaneously protects against all potential exposures, HIV sequence diversity alone represents a staggering challenge.

A further hurdle to HIV vaccine development is that HIV is an infection of the immune system, specifically targeting CD4+ T lymphocytes. Within days of exposure, massive infection and loss of memory CD4+ T cells ensues, particularly within the gut-associated

lymphoid tissue (Brenchley et al., 2004; Mehandru et al., 2004) where most of these cells reside. The loss of these cells, which are critical for coordinating effective immune responses, results in considerable immune impairment within the first weeks of HIV infection. Further, bacterial translocation across a damaged intestinal mucosa may even help to drive ongoing CD4+ T cell activation and facilitate viral replication, which is most efficient in activated cells (Brenchley et al., 2006).

Yet another challenge is that HIV has evolved strategies to avoid immune elimination. HIV rapidly establishes a latent reservoir of infected lymphocytes by integration of its genetic material into the host chromosome. This represents an enormous obstacle because this is an irreversible process that occurs immediately after infection and ends only with the death of the infected cell (Chun et al., 1997). The virus is immunologically silent in this latent reservoir, but production of infectious virus particles may be subsequently initiated if cells become activated at a later time. The stability of this reservoir means that lifelong infection of the host is maintained, even in the face of potent anti-HIV medication.

One potentially promising avenue of immunogen design is to produce epitope mimics determined from structural studies of antibody-antigen interactions. In the case of many of the most broad and potently neutralizing HIV antibodies, however, unique genetic or structural features are present that may be difficult to elicit through vaccination. An easier path may involve selection of antibodies in the HIV-uninfected antibody repertoire that already contain the desired unique elements. In this case, the immunogens may be designed not to mimic the epitope, but to select likely progenitors to the bnAb. Thus, the immunogen may be potentially designed even in cases where the crystal structure of the antibody-antigen complex is difficult or impossible to obtain. The first step toward determining the potential of this vaccine strategy, however, is to evaluate whether antibodies containing the desired unique features of HIV bnAbs exist within the HIV-uninfected antibody repertoire. In this report, I focused on the identification of three unique elements

that are commonly found in HIV bnAbs: long HCDR3s, SHA indels and unique secondary structure within the HCDR3 loop.

Long HCDR3s were identified in the HIV-uninfected human repertoire and were found to utilize a select subset of germline genes. These same germline genes contain many of the critical residues for binding and neutralization by the broadly neutralizing antibodies PG9 and PG16. This indicates not only that potential precursors to PG9 and PG16 are present in the HIV-uninfected repertoire, but that the majority of long HCDR3s in the HIV-uninfected repertoire are built with the same genetic building blocks as PG9 and PG16. SHA indels had previously been identified in the healthy donor peripheral blood repertoire, but I analyzed the position and length distribution of SHA indels on a much larger sequence set. As SHA indels are infrequent, study of the distribution of these events has been difficult. When attempting to identify the frequency of a particular SHA indel found in a bnAb, knowledge of the approximate frequency of all SHA indels in the healthy donor repertoire, as was known before this report, is insufficient. To accurately identify the potential precursor frequency of the bnAb of interest, the frequency of sequences containing the appropriate length SHA indel at the appropriate codon position is necessary. In the course of this study, I have assembled a comprehensive database of the position and location of SHA indels in the healthy donor peripheral blood repertoire. Finally, we analyzed the HIV-uninfected antibody repertoire for the presence of sequences that display predicted structural homology to the unique HCDR3 of PG9 and PG16. We isolated a handful of sequences with such predicted structural homology, although these sequences contained few of the required residues for proper interaction with HIV. Nevertheless, the presence of structural homologs to PG9 and PG16 suggests the possibility that an immunogen can be designed to select antibodies based on this structural homology and induce a relatively small set of somatic mutations in order to gain binding and neutralization capacity.

METHODS

Sample Preparation and Sorting

Peripheral blood was obtained from healthy adult donors following informed consent, under a protocol approved by the Vanderbilt Institutional Review Board. Mononuclear cells from the blood of four donors were isolated by density gradient centrifugation with Histopaque 1077 (Sigma). Prior to staining, B cells were enriched by paramagnetic separation using microbeads conjugated with antibodies to CD19 (Miltenyi Biotec). Cells from particular B cell subsets were sorted as separate populations on a high speed sorting cytometer (FACSAria III; Becton Dickinson) using the following phenotypic markers, naïve B cells: $CD19^+/CD27^-/IgM^+/IgG^-/CD14^-/CD3^-$, IgM memory B cells: $CD19^+/CD27^+/IgM^+/IgG^-/CD14^-/CD3^-$ and IgG memory B cells: $CD19^+/CD27^+/IgM^-/IgG^+/CD14^-/CD3^-$. Total RNA was isolated from each sorted cell subset using a commercial RNA extraction kit (RNeasy; Qiagen) and stored at -80°C until analysis.

Tissue-specific total RNA and mRNA

Purified polyA+ mRNA (lymph node) or total tissue RNA (all other samples) from the tissues of healthy human subjects was obtained from a commercial source (Clontech).

cDNA synthesis and PCR amplication of antibody genes

RT-PCR primers were originally described by the BIOMED-2 consortium (van Dongen et al., 2003) and were slightly modified to suit amplification for large-scale parallel pyrosequencing (454 Sequencing; 454 Life Sciences/Roche). 100 ng of each total RNA sample and 10 pmol of each RT-PCR primer were used in duplicate 50 µl RT-PCR reactions using the OneStep RT-PCR system (Qiagen). Thermal cycling was performed in a BioRad DNA Engine PTC-0200 thermal cycler using the following protocol: 50°C for 30:00, 95°C for 15:00, 35 cycles

of (94°C for 0:45, 58°C for 0:45, 72°C for 2:00), 72°C for 10:00. cDNA synthesis and amplification were verified by agarose gel electrophoresis before duplicate RT-PCR reactions were pooled. 5 µl of each pooled RT-PCR reaction was used as template for 100 µl 454-adapter PCR reactions, which were carried out in quadruplicate. 20 pmol of each 454-adapter primer and 0.25 units of AmpliTaq Gold polymerase (Applied Biosystems) were used for each reaction. Thermal cycling was performed in a BioRad DNA Engine PTC-0200 thermal cycler using the following protocol: 95°C for 10:00, 10 cycles of (95°C for 0:30, 58°C for 0:45, 72°C for 2:00), 72°C for 10:00. Amplification was verified by agarose gel electrophoresis before quadruplicate 454-adapter PCR reactions were pooled.

Amplicon Purification and Quantification

Amplicons were purified from the pooled 454-adapter PCR reactions using the Agencourt AMPure XP system (Beckman Coulter Genomics). Purified amplicons were quantified using a Qubit fluorometer (Invitrogen).

Amplicon Nucleotide Sequence Analysis

Quality control of the amplicon libraries and emulsion-based clonal amplification and sequencing on the 454 Genome Sequencer FLX Titanium system were performed by the W. M. Keck Center for Comparative and Functional Genomics at the University of Illinois at Urbana-Champaign, according to the manufacturer's instructions (454 Life Sciences). Signal processing and base calling were performed using the bundled 454 Data Analysis Software version 2.5.3 for amplicons.

Antibody Sequence Analysis

For germline gene assignments and initial analysis, the FASTA files resulting from 454 sequencing were submitted to the IMGT HighV-Quest webserver (IMGT, the international

ImMunoGeneTics information system; www.imgt.org; founder and director: Marie-Paule LeFranc, Montpellier, France). Antibody sequences returned from IMGT were considered to be “high-quality” sequences if they met the following requirements: sequence length of at least 300 nt; identified variable and joining genes; an intact, in-frame recombination; and absence of stop codons or ambiguous nucleotide calls within the reading frame.

Clustering of Antibody Repertoires

We perform agglomerative hierarchical clustering with complete linkage on both VDJ genes and individual donor subsets. First, we perform a filter that removes VDJ genes with low counts of low variation across all samples. Then we calculate pairwise distances between genes and samples using Pearson correlation. We standardize the values in the heat map to display in the range -3 to +3. Dendrograms and heatmaps were created with Matlab R2010b.

Analysis of Differential Expression of V(D)J Recombinants

We use the edgeR software (Robinson et al., 2010) to calculate differential expression between tissues. EdgeR uses the negative binomial as the appropriate distribution for count data. We obtain dispersion estimates and test differential expression using the generalized linear model (GLM) likelihood ratio test. The columns in the table show the fold change between tissues and the p-value and Benjamini and Hochberg false discovery rate.

Stringent Filtering for Putative V(DD)J recombinations

The antibody sequence region identified by IMGT as a putative diversity gene is designated here as the “match region”. The length of the match region, minus any mismatches between the match region and the germline diversity gene, is designated the “match score”. Our filtering process required the match regions to contain a maximum of one nucleotide

difference from the germline diversity gene segment. The match score, which represents the number of identically matched nucleotides between the match region and the germline diversity gene segment, was required to be at least 60% of the overall length of the germline diversity gene segment, except in the case of the short IGHD7-27 gene segment for which we required a match score of 72% of the germline diversity gene length.

Data Analysis

All statistical analyses were performed with Graphpad Prism software. Three-dimensional antibody structural models were colored using MacPyMol and custom scripts. All Circos plots were made using Circos software (www.circos.ca/software).

REFERENCES

Aguilera, I., Melero, J., Nuñez-Roldan, A., and Sanchez, B. (2001). Molecular structure of eight human autoreactive monoclonal antibodies. Immunology *102*, 273–280.

Akamatsu, Y., Tsurushita, N., Nagawa, F., Matsuoka, M., Okazaki, K., Imai, M., and Sakano, H. (1994). Essential residues in V(D)J recombination signals. J Immunol *153*, 4520–4529.

Akira, S., Okazaki, K., and Sakano, H. (1987). Two pairs of recombination signals are sufficient to cause immunoglobulin V-(D)-J joining. Science *238*, 1134–1138.

Alt, F., Blackwell, T., and Yancopoulos, G. (1987). Development of the primary antibody repertoire. Science *238*, 1079–1087.

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

Alt, F.W., Oltz, E.M., Young, F., Gorman, J., Taccioli, G., and Chen, J. (1992). VDJ recombination. Immunology Today *13*, 306–314.

Arnaout, R., Lee, W., Cahill, P., Honan, T., Sparrow, T., Weiand, M., Nusbaum, C., Rajewsky, K., and Koralov, S.B. (2011). High-resolution description of antibody heavy-chain repertoires in humans. PLoS ONE *6*, e22365.

Azoitei, M.L., Correia, B.E., Ban, Y.-E.A., Carrico, C., Kalyuzhniy, O., Chen, L., Schroeter, A., Huang, P.-S., McLellan, J.S., Kwong, P.D., et al. (2011). Computation-guided backbone grafting of a discontinuous motif onto a protein scaffold. Science *334*, 373–376.

Barbas, C.F., Collet, T.A., Amberg, W., Roben, P., Binley, J.M., Hoekstra, D., Cababa, D., Jones, T.M., Williamson, R.A., and Pilkington, G.R. (1993). Molecular profile of an antibody response to HIV-1 as probed by combinatorial libraries. J Mol Biol *230*, 812–823.

Bemark, M., and Neuberger, M.S. (2003). By-products of immunoglobulin somatic hypermutation. Genes Chromosomes Cancer *38*, 32–39.

Binley, J.M., Cayanan, C.S., Wiley, C., Schülke, N., Olson, W.C., and Burton, D.R. (2003). Redox-triggered infection by disulfide-shackled human immunodeficiency virus type 1 pseudovirions. J Virol *77*, 5678–5684.

Binley, J.M., Lybarger, E.A., Crooks, E.T., Seaman, M.S., Gray, E., Davis, K.L., Decker, J.M., Wycuff, D., Harris, L., Hawkins, N., et al. (2008). Profiling the specificity of neutralizing antibodies in a large panel of plasmas from patients chronically infected with human immunodeficiency virus type 1 subtypes B and C. J Virol *82*, 11651–11668.

Binley, J.M., Wrin, T., Korber, B., Zwick, M.B., Wang, M., Chappey, C., Stiegler, G., Kunert, R., Zolla-Pazner, S., Katinger, H., et al. (2004). Comprehensive cross-clade neutralization analysis of a panel of anti-human immunodeficiency virus type 1 monoclonal antibodies. J Virol *78*, 13232–13252.

Boyd, S.D., Gaëta, B.A., Jackson, K.J., Fire, A.Z., Marshall, E.L., Merker, J.D., Maniar, J.M., Zhang, L.N., Sahaf, B., Jones, C.D., et al. (2010). Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. J Immunol *184*, 6986–6992.

Boyd, S.D., Marshall, E.L., Merker, J.D., Maniar, J.M., Zhang, L.N., Sahaf, B., Jones, C.D., Simen, B.B., Hanczaruk, B., Nguyen, K.D., et al. (2009). Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Sci Transl Med *1*, 12ra23.

Brack, C., Hirama, M., Lenhard-Schuller, R., and Tonegawa, S. (1978). A complete immunoglobulin gene is created by somatic recombination. Cell *15*, 1–14.

Brenchley, J.M., Price, D.A., Schacker, T.W., Asher, T.E., Silvestri, G., Rao, S., Kazzaz, Z., Bornstein, E., Lambotte, O., Altmann, D., et al. (2006). Microbial translocation is a cause of systemic immune activation in chronic HIV infection. Nat Med *12*, 1365–1371.

Brenchley, J.M., Schacker, T.W., Ruff, L.E., Price, D.A., Taylor, J.H., Beilman, G.J., Nguyen, P.L., Khoruts, A., Larson, M., Haase, A.T., et al. (2004). CD4+ T cell depletion during all stages of HIV disease occurs predominantly in the gastrointestinal tract. J Exp Med *200*, 749–759.

Brenner, S., and Milstein, C. (1966). Origin of antibody variation. Nature *211*, 242–243.

Briney, B., Willis, J.R., and Crowe, J.E. (2012a). Human Peripheral Blood Antibodies with Long HCDR3s Are Established Primarily at Original Recombination Using a Limited Subset of Germline Genes. PLoS ONE *7*, e36750.

Briney, B.S., Willis, J.R., McKinney, B.A., and Crowe, J.E. (2012b). High-throughput antibody sequencing reveals genetic evidence of global regulation of the naïve and memory repertoires that extends across individuals. Genes Immun.

Brochet, X., Lefranc, M.-P., and Giudicelli, V. (2008). IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. Nucl Acids Res *36*, W503–W508.

Burton, D.R. (2002). Antibodies, viruses and vaccines. Nat Rev Immunol *2*, 706–713.

Burton, D.R. (2010). Scaffolding to build a rational vaccine design strategy. Proc Natl Acad Sci USA *107*, 17859–17860.

Burton, D.R., Barbas, C.F., Persson, M.A., Koenig, S., Chanock, R.M., and Lerner, R.A. (1991). A large array of human monoclonal antibodies to type 1 human immunodeficiency virus from combinatorial libraries of asymptomatic seropositive individuals. Proc Natl Acad Sci USA *88*, 10134–10137.

Burton, D.R., Desrosiers, R.C., Doms, R.W., Koff, W.C., Kwong, P.D., Moore, J.P., Nabel, G.J., Sodroski, J., Wilson, I.A., and Wyatt, R.T. (2004). HIV vaccine design and the neutralizing antibody problem. Nat Immunol *5*, 233–236.

Burton, D.R., Pyati, J., Koduri, R., Sharp, S.J., Thornton, G.B., Parren, P.W., Sawyer, L.S., Hendry, R.M., Dunlop, N., and Nara, P.L. (1994). Efficient neutralization of primary isolates

of HIV-1 by a recombinant human monoclonal antibody. Science *266*, 1024–1027.

Burton, D.R., Stanfield, R.L., and Wilson, I.A. (2005a). Antibody vs. HIV in a clash of evolutionary titans. Proc Natl Acad Sci USA *102*, 14943–14948.

Burton, D.R., Stanfield, R.L., and Wilson, I.A. (2005b). Antibody vs. HIV in a clash of evolutionary titans. Proc Natl Acad Sci USA *102*, 14943–14948.

Calarese, D.A., Scanlan, C.N., Zwick, M.B., Deechongkit, S., Mimura, Y., Kunert, R., Zhu, P., Wormald, M.R., Stanfield, R.L., Roux, K.H., et al. (2003). Antibody domain exchange is an immunological solution to carbohydrate cluster recognition. Science *300*, 2065–2071.

Capra, J.D., and Kehoe, J.M. (1975). Hypervariable regions, idiotypy, and the antibody-combining site. Adv Immunol *20*, 1–40.

Cardoso, R.M.F., Zwick, M.B., Stanfield, R.L., Kunert, R., Binley, J.M., Katinger, H., Burton, D.R., and Wilson, I.A. (2005). Broadly Neutralizing Anti-HIV Antibody 4E10 Recognizes a Helical Conformation of a Highly Conserved Fusion-Associated Motif in gp41. Immunity *22*, 163–173.

Casali, P., Pal, Z., Xu, Z., and Zan, H. (2006). DNA repair in antibody somatic hypermutation. Trends Immunol. *27*, 313–321.

Celada, F., and Seiden, P.E. (1996). Affinity maturation and hypermutation in a simulation of the humoral immune response. Eur. J. Immunol. *26*, 1350–1358.

Changela, A., Wu, X., Yang, Y., Zhang, B., Zhu, J., Nardone, G.A., O'Dell, S., Pancera, M., Gorny, M.K., Phogat, S., et al. (2011). Crystal structure of human antibody 2909 reveals conserved features of quaternary structure-specific antibodies that potently neutralize HIV-1. J Virol *85*, 2524–2535.

Chen, C., Nagy, Z., Prak, E., and Weigert, M. (1995). Immunoglobulin heavy chain gene replacement: A mechanism of receptor editing. Immunity *3*, 747–755.

Choe, H., Li, W., Wright, P.L., Vasilieva, N., Venturi, M., Huang, C.-C., Grundner, C., Dorfman, T., Zwick, M.B., Wang, L., et al. (2003). Tyrosine sulfation of human antibodies contributes to recognition of the CCR5 binding region of HIV-1 gp120. Cell *114*, 161–170.

Choi, Y.W., Herman, A., Digiusto, D., Wade, T., Marrack, P., and Kappler, J. (1990). Residues of the variable region of the T-cell-receptor beta-chain that interact with S. aureus toxin superantigens. Nature *346*, 471–473.

Chun, T.W., Carruth, L., Finzi, D., Shen, X., DiGiuseppe, J.A., Taylor, H., Hermankova, M., Chadwick, K., Margolick, J., Quinn, T.C., et al. (1997). Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. Nature *387*, 183–188.

Corbett, S. (1997). 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*, 587–597.

Correia, B.E., Ban, Y.-E.A., Holmes, M.A., Xu, H., Ellingson, K., Kraft, Z., Carrico, C., Boni,

E., Sather, D.N., Zenobia, C., et al. (2010). Computational design of epitope-scaffolds allows induction of antibodies specific for a poorly immunogenic HIV vaccine epitope. Structure *18*, 1116–1126.

Crotty, S., and Ahmed, R. (2004). Immunological memory in humans. Semin. Immunol. *16*, 197–203.

Crouzier, R., Martin, T., and Pasquali, J.L. (1995). 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*, 4526–4535.

Cuomo, C.A., Mundy, C.L., and Oettinger, M.A. (1996). DNA sequence and structure requirements for cleavage of V(D)J recombination signal sequences. Molecular and Cellular Biology *16*, 5683–5690.

David, D., Demaison, C., Bani, L., and Theze, J. (1996). Progressive decrease in VH3 gene family expression in plasma cells of HIV-infected patients. Int. Immunol. *8*, 1329–1333.

David, D., Demaison, C., Bani, L., Zouali, M., and Theze, J. (1995). Selective variations in vivo of VH3 and VH1 gene family expression in peripheral B cell IgM, IgD and IgG during HIV infection. Eur. J. Immunol. *25*, 1524–1528.

de Wildt, R.M., van Venrooij, W.J., Winter, G., Hoet, R.M., and Tomlinson, I.M. (1999). Somatic insertions and deletions shape the human antibody repertoire. J Mol Biol *294*, 701–710.

Dhillon, A.K., Donners, H., Pantophlet, R., Johnson, W.E., Decker, J.M., Shaw, G.M., Lee, F.-H., Richman, D.D., Doms, R.W., Vanham, G., et al. (2007). Dissecting the neutralizing antibody specificities of broadly neutralizing sera from human immunodeficiency virus type 1-infected donors. J Virol *81*, 6548–6562.

Di Noia, J., and Neuberger, M.S. (2002). Altering the pathway of immunoglobulin hypermutation by inhibiting uracil-DNA glycosylase. Nature *419*, 43–48.

Difilippantonio, M.J., McMahan, C.J., Eastman, Q.M., Spanopoulou, E., and Schatz, D.G. (1996). RAG1 mediates signal sequence recognition and recruitment of RAG2 in V(D)J recombination. Cell *87*, 253–262.

Diskin, R., Marcovecchio, P.M., and Bjorkman, P.J. (2010). Structure of a clade C HIV-1 gp120 bound to CD4 and CD4-induced antibody reveals anti-CD4 polyreactivity. Nat Struct Mol Biol *17*, 608–613.

Diskin, R., Scheid, J.F., Marcovecchio, P.M., West, A.P., Klein, F., Gao, H., Gnanapragasam, P.N.P., Abadir, A., Seaman, M.S., Nussenzweig, M.C., et al. (2011). Increasing the potency and breadth of an HIV antibody by using structure-based rational design. Science *334*, 1289–1293.

Doores, K.J., and Burton, D.R. (2010). Variable loop glycan dependency of the broad and potent HIV-1-neutralizing antibodies PG9 and PG16. J Virol *84*, 10510–10521.

Doores, K.J., Fulton, Z., Huber, M., Wilson, I.A., and Burton, D.R. (2010). Antibody 2G12

recognizes di-mannose equivalently in domain- and nondomain-exchanged forms but only binds the HIV-1 glycan shield if domain exchanged. J Virol *84*, 10690–10699.

Douek, D.C., Kwong, P.D., and Nabel, G.J. (2006). The rational design of an AIDS vaccine. Cell *124*, 677–681.

Dörner, T., Foster, S., Farner, N., and Lipsky, P. (1998). Somatic hypermutation of human immunoglobulin heavy chain genes: targeting of RGYW motifs on both DNA strands. Eur. J. Immunol. *28*, 3384–3396.

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

Ekiert, D.C., Bhabha, G., Elsliger, M.-A., Friesen, R.H.E., Jongeneelen, M., Throsby, M., Goudsmit, J., and Wilson, I.A. (2009). Antibody recognition of a highly conserved influenza virus epitope. Science *324*, 246–251.

Euler, Z., Bunnik, E.M., Burger, J.A., Boeser-Nunnink, B.D.M., Grijsen, M.L., Prins, J.M., and Schuitemaker, H. (2011). Activity of broadly neutralizing antibodies, including PG9, PG16 and VRC01, against recently transmitted subtype B HIV-1 variants from early and late in the epidemic. J Virol.

Garcia, K., Degano, M., Stanfield, R., Brunmark, A., Jackson, M., Peterson, P., Teyton, L., and Wilson, I. (1996). An alpha beta T cell receptor structure at 2.5 angstrom and its orientation in the TCR-MHC complex. Science *274*, 209–219.

Goossens, T., Klein, U., and Küppers, R. (1998). Frequent occurrence of deletions and duplications during somatic hypermutation: implications for oncogene translocations and heavy chain disease. Proc Natl Acad Sci USA *95*, 2463–2468.

Gorny, M.K., Conley, A.J., Karwowska, S., Buchbinder, A., Xu, J.Y., Emini, E.A., Koenig, S., and Zolla-Pazner, S. (1992). Neutralization of diverse human immunodeficiency virus type 1 variants by an anti-V3 human monoclonal antibody. J Virol *66*, 7538–7542.

Gorny, M.K., Wang, X.-H., Williams, C., Volsky, B., Revesz, K., Witover, B., Burda, S., Urbanski, M., Nyambi, P., Krachmarov, C., et al. (2009). Preferential use of the VH5-51 gene segment by the human immune response to code for antibodies against the V3 domain of HIV-1. Mol Immunol *46*, 917–926.

Gray, E.S., Taylor, N., Wycuff, D., Moore, P.L., Tomaras, G.D., Wibmer, C.K., Puren, A., Decamp, A., Gilbert, P.B., Wood, B., et al. (2009). Antibody specificities associated with neutralization breadth in plasma from human immunodeficiency virus type 1 subtype C-infected blood donors. J Virol *83*, 8925–8937.

Han, S., Dillon, S.R., Zheng, B., Shimoda, M., Schlissel, M.S., and Kelsoe, G. (1997). V(D)J recombinase activity in a subset of germinal center B lymphocytes. Science *278*, 301–305.

Hartley, O., Klasse, P.J., Sattentau, Q.J., and Moore, J.P. (2005). V3: HIV's switch-hitter. Aids Res Hum Retrov *21*, 171–189.

Haynes, B.F., and Montefiori, D.C. (2006). Aiming to induce broadly reactive neutralizing antibody responses with HIV-1 vaccine candidates. Expert Rev Vaccines *5*, 347–363.

Haynes, B.F., Fleming, J., St Clair, E.W., Katinger, H., Stiegler, G., Kunert, R., Robinson, J.E., Scearce, R.M., Plonk, K., Staats, H.F., et al. (2005). Cardiolipin polyspecific autoreactivity in two broadly neutralizing HIV-1 antibodies. Science *308*, 1906–1908.

Henderson, K.A., Streltsov, V.A., Coley, A.M., Dolezal, O., Hudson, P.J., Batchelor, A.H., Gupta, A., Bai, T., Murphy, V.J., Anders, R.F., et al. (2007). Structure of an IgNAR-AMA1 complex: targeting a conserved hydrophobic cleft broadens malarial strain recognition. Structure *15*, 1452–1466.

Hesse, J., Lieber, M., Mizuuchi, K., and Gellert, M. (1989). V(D)J Recombination - a Functional Definition of the Joining Signals. Genes Dev. *3*, 1053–1061.

Hessell, A.J., Poignard, P., Hunter, M., Hangartner, L., Tehrani, D.M., Bleeker, W.K., Parren, P.W.H.I., Marx, P.A., and Burton, D.R. (2009). Effective, low-titer antibody protection against low-dose repeated mucosal SHIV challenge in macaques. Nat Med *15*, 951–954.

Hessell, A.J., Rakasz, E.G., Tehrani, D.M., Huber, M., Weisgrau, K.L., Landucci, G., Forthal, D.N., Koff, W.C., Poignard, P., Watkins, D.I., et al. (2010). Broadly neutralizing monoclonal antibodies 2F5 and 4E10 directed against the human immunodeficiency virus type 1 gp41 membrane-proximal external region protect against mucosal challenge by simian-human immunodeficiency virus SHIVBa-L. J Virol *84*, 1302–1313.

Huang, C.-C., Venturi, M., Majeed, S., Moore, M.J., Phogat, S., Zhang, M.-Y., Dimitrov, D.S., Hendrickson, W.A., Robinson, J.E., Sodroski, J., et al. (2004). Structural basis of tyrosine sulfation and VH-gene usage in antibodies that recognize the HIV type 1 coreceptor-binding site on gp120. Proc Natl Acad Sci USA *101*, 2706–2711.

Ichihara, Y.Y., Matsuoka, H.H., and Kurosawa, Y.Y. (1988). Organization of human immunoglobulin heavy chain diversity gene loci. Embo J *7*, 4141–4150.

Ippolito, G.C., Schelonka, R.L., Zemlin, M., Ivanov, I.I., Kobayashi, R., Zemlin, C., Gartland, G.L., Nitschke, L., Pelkonen, J., Fujihashi, K., et al. (2006). Forced usage of positively charged amino acids in immunoglobulin CDR-H3 impairs B cell development and antibody production. J Exp Med *203*, 1567–1578.

Ivanov, I., Link, J., Ippolito, G., and Schroeder, H.J. (2002). The Antibodies (CRC Press).

Ivanov, I.I., Schelonka, R.L., Zhuang, Y., Gartland, G.L., Zemlin, M., and Schroeder, H.W. (2005). Development of the expressed Ig CDR-H3 repertoire is marked by focusing of constraints in length, amino acid use, and charge that are first established in early B cell progenitors. J Immunol *174*, 7773–7780.

Jackson, S.M., Wilson, P.C., James, J.A., and Capra, J.D. (2008). Human B cell subsets. Adv Immunol *98*, 151–224.

Jiang, X., Burke, V., Totrov, M., Williams, C., Cardozo, T., Gorny, M.K., Zolla-Pazner, S., and Kong, X.-P. (2010). Conserved structural elements in the V3 crown of HIV-1 gp120. Nat Struct Mol Biol *17*, 955–961.

Kabat, E.A., Wu, Te, T., Gottesman, K.S., and Foeller, C. (1992). Sequences of Proteins of Immunological Interest (Diane Books Publishing Company).

Kelsoe, G. (1994). B cell diversification and differentiation in the periphery. J Exp Med *180*, 5–6.

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

Kiyoi, H., Naoe, T., Horibe, K., and Ohno, R. (1992). Characterization of the immunoglobulin heavy chain complementarity determining region (CDR)-III sequences from human B cell precursor acute lymphoblastic leukemia cells. J. Clin. Invest. *89*, 739–746.

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

Koralov, S.B., Novobrantseva, T.I., Hochedlinger, K., Jaenisch, R., and Rajewsky, K. (2005). Direct in vivo VH to JH rearrangement violating the 12/23 rule. J Exp Med *201*, 341–348.

Koralov, S.B., Novobrantseva, T.I., Königsmann, J., Ehlich, A., and Rajewsky, K. (2006). Antibody repertoires generated by VH replacement and direct VH to JH joining. Immunity *25*, 43–53.

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

Kowalski, M., Potz, J., Basiripour, L., Dorfman, T., Goh, W.C., Terwilliger, E., Dayton, A., Rosen, C., Haseltine, W., and Sodroski, J. (1987). Functional regions of the envelope glycoprotein of human immunodeficiency virus type 1. Science *237*, 1351–1355.

Krause, J.C., Ekiert, D.C., Tumpey, T.M., Smith, P.B., Wilson, I.A., and Crowe, J.E. (2011). An insertion mutation that distorts antibody binding site architecture enhances function of a human antibody. MBio *2*, e00345–10.

Kurosawa, Y., and Tonegawa, S. (1982). Organization, structure, and assembly of immunoglobulin heavy chain diversity DNA segments. J Exp Med *155*, 201–218.

Kwong, P.D., Wyatt, R.T., Robinson, J., Sweet, R.W., Sodroski, J., and Hendrickson, W.A. (1998). Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Nature *393*, 648–659.

Lee, A.I., Fugmann, S.D., Cowell, L.G., Ptaszek, L.M., Kelsoe, G., and Schatz, D.G. (2003). A functional analysis of the spacer of V(D)J recombination signal sequences. PLoS Biol. *1*, E1.

Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., et al. (2009). IMGT, the international ImMunoGeneTics information system. Nucl Acids Res *37*, D1006–D1012.

Lewis, S.M. (1994). The mechanism of V(D)J joining: lessons from molecular, immunological, and comparative analyses. Adv Immunol *56*, 27–150.

Li, H., Llera, A., and Mariuzza, R. (1998). Structure-function studies of T-cell receptor superantigen interactions. Immunol Rev *163*, 177–186.

Li, Y., O'Dell, S., Walker, L.M., Wu, X., Guenaga, J., Feng, Y., Schmidt, S.D., McKee, K., Louder, M.K., Ledgerwood, J.E., et al. (2011). Mechanism of neutralization by the broadly neutralizing HIV-1 monoclonal antibody VRC01. J Virol.

Li, Y.Y., Migueles, S.A.S., Welcher, B.B., Svehla, K.K., Phogat, A.A., Louder, M.K.M., Wu, X.X., Shaw, G.M.G., Connors, M.M., Wyatt, R.T.R., et al. (2007). Broad HIV-1 neutralization mediated by CD4-binding site antibodies. Nat Med *13*, 1032–1034.

Li, Z., Woo, C.J., Iglesias-Ussel, M.D., Ronai, D., and Scharff, M.D. (2004). The generation of antibody diversity through somatic hypermutation and class switch recombination. Genes Dev. *18*, 1–11.

Lu, M., Blacklow, S.C., and Kim, P.S. (1995). A trimeric structural domain of the HIV-1 transmembrane glycoprotein. Nat Struct Biol *2*, 1075–1082.

Lutz, J., Müller, W., and Jäck, H.-M. (2006). VH replacement rescues progenitor B cells with two nonproductive VDJ alleles. J Immunol *177*, 7007–7014.

MacLennan, I.C. (1994). Germinal centers. Annu Rev Immunol *12*, 117–139.

MacLennan, I.C., Liu, Y.J., and Johnson, G.D. (1992). Maturation and dispersal of B-cell clones during T cell-dependent antibody responses. Immunol Rev *126*, 143–161.

Mahajan, K.N., Gangi-Peterson, L., Sorscher, D.H., Wang, J., Gathy, K.N., Mahajan, N.P., Reeves, W.H., and Mitchell, B.S. (1999). Association of terminal deoxynucleotidyl transferase with Ku. Proc Natl Acad Sci USA *96*, 13926–13931.

Mansilla-Soto, J., and Cortes, P. (2003). VDJ Recombination: Artemis and Its In Vivo Role in Hairpin Opening. J Exp Med *197*, 543–547.

Mascola, J.R., Lewis, M.G., Stiegler, G., Harris, D., VanCott, T.C., Hayes, D., Louder, M.K., Brown, C.R., Sapan, C.V., Frankel, S.S., et al. (1999). Protection of Macaques against pathogenic simian/human immunodeficiency virus 89.6PD by passive transfer of neutralizing antibodies. J Virol *73*, 4009–4018.

Mascola, J.R., Stiegler, G., VanCott, T.C., Katinger, H., Carpenter, C.B., Hanson, C.E., Beary, H., Hayes, D., Frankel, S.S., Birx, D.L., et al. (2000). Protection of macaques against vaginal transmission of a pathogenic HIV-1/SIV chimeric virus by passive infusion of neutralizing antibodies. Nat Med *6*, 207–210.

Matsuda, F., Ishii, K., Bourvagnet, P., Kuma, K.I., Hayashida, H., Miyata, T., and Honjo, T. (1998). The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. J Exp Med *188*, 2151–2162.

McBlane, J.F., van Gent, D.C., Ramsden, D.A., Romeo, C., Cuomo, C.A., Gellert, M., and

Oettinger, M.A. (1995). Cleavage at a V(D)J recombination signal requires only RAG1 and RAG2 proteins and occurs in two steps. Cell *83*, 387–395.

McLellan, J.S., Pancera, M., Carrico, C., Gorman, J., Julien, J.-P., Khayat, R., Louder, R., Pejchal, R., Sastry, M., Dai, K., et al. (2011). Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9. Nature *480*, 336–343.

Mehandru, S., Poles, M.A., Tenner-Racz, K., Horowitz, A., Hurley, A., Hogan, C., Boden, D., Racz, P., and Markowitz, M. (2004). Primary HIV-1 infection is associated with preferential depletion of CD4+ T lymphocytes from effector sites in the gastrointestinal tract. J Exp Med *200*, 761–770.

Montalbano, A., Ogwaro, K.M., Tang, A., Matthews, A.G.W., Larijani, M., Oettinger, M.A., and Feeney, A.J. (2003). V(D)J recombination frequencies can be profoundly affected by changes in the spacer sequence. J Immunol *171*, 5296–5304.

Motoyama, N., Okada, H., and Azuma, T. (1991). Somatic mutation in constant regions of mouse lambda 1 light chains. Proc Natl Acad Sci USA *88*, 7933–7937.

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

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

Muster, T., Guinea, R., Trkola, A., Purtscher, M., Klima, A., Steindl, F., Palese, P., and Katinger, H. (1994). Cross-neutralizing activity against divergent human immunodeficiency virus type 1 isolates induced by the gp41 sequence ELDKWAS. J Virol *68*, 4031–4034.

Muster, T., Steindl, F., Purtscher, M., Trkola, A., Klima, A., Himmler, G., Rüker, F., and Katinger, H. (1993). A conserved neutralizing epitope on gp41 of human immunodeficiency virus type 1. J Virol *67*, 6642–6647.

Nadel, B., Tang, A., Lugo, G., Love, V., Escuro, G., and Feeney, A.J. (1998). Decreased frequency of rearrangement due to the synergistic effect of nucleotide changes in the heptamer and nonamer of the recombination signal sequence of the V kappa gene A2b, which is associated with increased susceptibility of Navajos to Haemophilus influenzae type b disease. J Immunol *161*, 6068–6073.

Nemazee, D. (2006). Receptor editing in lymphocyte development and central tolerance. Nat Rev Immunol *6*, 728–740.

Neuberger, M.S. (2008). Antibody diversification by somatic mutation: from Burnet onwards. Immunol Cell Biol *86*, 124–132.

Neuberger, M.S., and Milstein, C. (1995). Somatic hypermutation. Curr Opin Immunol *7*, 248–254.

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

Ofek, G., Guenaga, F.J., Schief, W.R., Skinner, J., Baker, D., Wyatt, R.T., and Kwong, P.D. (2010). Elicitation of structure-specific antibodies by epitope scaffolds. Proc Natl Acad Sci USA *107*, 17880–17887.

Ofek, G., Tang, M., Sambor, A., Katinger, H., Mascola, J.R., Wyatt, R.T., and Kwong, P.D. (2004). Structure and mechanistic analysis of the anti-human immunodeficiency virus type 1 antibody 2F5 in complex with its gp41 epitope. J Virol *78*, 10724–10737.

Pancera, M., McLellan, J.S., Wu, X., Zhu, J., Changela, A., Schmidt, S.D., Yang, Y., Zhou, T., Phogat, S., Mascola, J.R., et al. (2010). Crystal structure of PG16 and chimeric dissection with somatically related PG9: structure-function analysis of two quaternary-specific antibodies that effectively neutralize HIV-1. J Virol *84*, 8098–8110.

Pantophlet, R., and Burton, D.R. (2006). GP120: target for neutralizing HIV-1 antibodies. Annu Rev Immunol *24*, 739–769.

Pantophlet, R., and Ollmann Saphire, E. (2003). Fine mapping of the interaction of neutralizing and nonneutralizing monoclonal antibodies with the CD4 binding site of human immunodeficiency virus type 1 gp120. Journal of ….

Papavasiliou, F., Casellas, R., Suh, H., Qin, X.F., Besmer, E., Pelanda, R., Nemazee, D., Rajewsky, K., and Nussenzweig, M.C. (1997). V(D)J recombination in mature B cells: a mechanism for altering antibody responses. Science *278*, 298–301.

Parren, P.W., Marx, P.A., Hessell, A.J., Luckay, A., Harouse, J., Cheng-Mayer, C., Moore, J.P., and Burton, D.R. (2001). Antibody protects macaques against vaginal challenge with a pathogenic R5 simian/human immunodeficiency virus at serum levels giving complete neutralization in vitro. J Virol *75*, 8340–8347.

Pejchal, R., Doores, K.J., Walker, L.M., Khayat, R., Huang, P.-S., Wang, S.-K., Stanfield, R.L., Julien, J.-P., Ramos, A., Crispin, M., et al. (2011). A potent and broad neutralizing antibody recognizes and penetrates the HIV glycan shield. Science *334*, 1097–1103.

Pejchal, R., Gach, J.S., Brunel, F.M., Cardoso, R.M., Stanfield, R.L., Dawson, P.E., Burton, D.R., Zwick, M.B., and Wilson, I.A. (2009). A conformational switch in human immunodeficiency virus gp41 revealed by the structures of overlapping epitopes recognized by neutralizing antibodies. J Virol *83*, 8451–8462.

Pejchal, R., Walker, L.M., Stanfield, R.L., Phogat, S.K., Koff, W.C., Poignard, P., Burton, D.R., and Wilson, I.A. (2010). Structure and function of broadly reactive antibody PG16 reveal an H3 subdomain that mediates potent neutralization of HIV-1. Proc Natl Acad Sci USA *107*, 11483–11488.

Pham, P., Bransteitter, R., Petruska, J., and Goodman, M.F. (2003). Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. Nature *424*, 103–107.

Phung, Q.H., Winter, D.B., Cranston, A., Tarone, R.E., Bohr, V.A., Fishel, R., and Gearhart,

P.J. (1998). Increased hypermutation at G and C nucleotides in immunoglobulin variable genes from mice deficient in the MSH2 mismatch repair protein. J Exp Med *187*, 1745–1751.

Poignard, P., Saphire, E.O., Parren, P.W., and Burton, D.R. (2001). gp120: Biologic aspects of structural features. Annu Rev Immunol *19*, 253–274.

Potter, K.N., Li, Y., and Capra, J.D. (1996). 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*, 2982–2988.

Prabakaran, P., Chen, W., Singarayan, M.G., Stewart, C.C., Streaker, E., Feng, Y., and Dimitrov, D.S. (2011). Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations. Immunogenetics.

Purtscher, M., Trkola, A., Gruber, G., Buchacher, A., Predl, R., Steindl, F., Tauer, C., Berger, R., Barrett, N., and Jungbauer, A. (1994). A broadly neutralizing human monoclonal antibody against gp41 of human immunodeficiency virus type 1. Aids Res Hum Retrov *10*, 1651–1658.

Raaphorst, F.M., Raman, C.S., Tami, J., Fischbach, M., and Sanz, I. (1997). Human Ig heavy chain CDR3 regions in adult bone marrow pre-B cells display an adult phenotype of diversity: evidence for structural selection of DH amino acid sequences. Int. Immunol. *9*, 1503–1515.

Rada, C., and Milstein, C. (2001). The intrinsic hypermutability of antibody heavy and light chain genes decays exponentially. Embo J *20*, 4570–4576.

Rada, C., Ehrenstein, M.R., Neuberger, M.S., and Milstein, C. (1998). Hot spot focusing of somatic hypermutation in MSH2-deficient mice suggests two stages of mutational targeting. Immunity *9*, 135–141.

Radic, M., and Zouali, M. (1996). Receptor editing, immune diversification, and self-tolerance. Immunity *5*, 505–511.

Rajewsky, K., Förster, I., and Cumano, A. (1987). Evolutionary and somatic selection of the antibody repertoire in the mouse. Science *238*, 1088–1094.

Ramsden, D.A., Baetz, K., and Wu, G.E. (1994). Conservation of sequence in recombination signal sequence spacers. Nucl Acids Res *22*, 1785–1796.

Ramsden, D.A., McBlane, J.F., van Gent, D.C., and Gellert, M. (1996). Distinct DNA sequence and structure requirements for the two steps of V(D)J recombination signal cleavage. Embo J *15*, 3197–3206.

Reason, D.C., and Zhou, J. (2006). Codon insertion and deletion functions as a somatic diversification mechanism in human antibody repertoires. Biol. Direct *1*, 24.

Reth, M., Gehrmann, P., Petrac, E., and Wiese, P. (1986). A novel VH to VHDJH joining mechanism in heavy-chain-negative (null) pre-B cells results in heavy-chain production.

Nature *322*, 840–842.

Retter, M.W., and Nemazee, D. (1998). Receptor editing occurs frequently during normal B cell development. J Exp Med *188*, 1231–1238.

Richman, D.D., Wrin, T., Little, S.J., and Petropoulos, C.J. (2003). Rapid evolution of the neutralizing antibody response to HIV type 1 infection. Proc Natl Acad Sci USA *100*, 4144–4149.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

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

Roth, D.B. (2003). Restraining the V(D)J recombinase. Nat Rev Immunol *3*, 656–666.

Roth, D.B., Menetski, J.P., Nakajima, P.B., Bosma, M.J., and Gellert, M. (1992). V(D)J recombination: broken DNA molecules with covalently sealed (hairpin) coding ends in scid mouse thymocytes. Cell *70*, 983–991.

Sadofsky, M.J. (2001). The RAG proteins in V(D)J recombination: more than just a nuclease. Nucl Acids Res *29*, 1399–1409.

Salzwedel, K., Smith, E.D., Dey, B., and Berger, E.A. (2000). Sequential CD4-coreceptor interactions in human immunodeficiency virus type 1 Env function: soluble CD4 activates Env for coreceptor-dependent fusion and reveals blocking activities of antibodies against cryptic conserved epitopes on gp120. J Virol *74*, 326–333.

Sanders, R.W., Venturi, M., Schiffner, L., Kalyanaraman, R., Katinger, H., Lloyd, K.O., Kwong, P.D., and Moore, J.P. (2002). The mannose-dependent epitope for neutralizing antibody 2G12 on human immunodeficiency virus type 1 glycoprotein gp120. J Virol *76*, 7293–7305.

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

Saphire, E.O., Parren, P.W., Pantophlet, R., Zwick, M.B., Morris, G.M., Rudd, P.M., Dwek, R.A., Stanfield, R.L., Burton, D.R., and Wilson, I.A. (2001). Crystal structure of a neutralizing human IGG against HIV-1: a template for vaccine design. Science *293*, 1155–1159.

Sather, D.N., Armann, J., Ching, L.K., Mavrantoni, A., Sellhorn, G., Caldwell, Z., Yu, X., Wood, B., Self, S., Kalams, S., et al. (2009). Factors Associated with the Development of Cross-Reactive Neutralizing Antibodies during Human Immunodeficiency Virus Type 1 Infection. J Virol *83*, 757–769.

Scamurra, R.W., Miller, D.J., Dahl, L., Abrahamsen, M., Kapur, V., Wahl, S.M., Milner, E.C., and Janoff, E.N. (2000). Impact of HIV-1 infection on VH3 gene repertoire of naive human B

cells. J Immunol *164*, 5482–5491.

Scanlan, C.N., Offer, J., Zitzmann, N., and Dwek, R.A. (2007). Exploiting the defensive sugars of HIV-1 for drug and vaccine design. Nature *446*, 1038–1045.

Scanlan, C.N., Pantophlet, R., Wormald, M.R., Ollmann Saphire, E., Stanfield, R., Wilson, I.A., Katinger, H., Dwek, R.A., Rudd, P.M., and Burton, D.R. (2002). The broadly neutralizing anti-human immunodeficiency virus type 1 antibody 2G12 recognizes a cluster of alpha1-->2 mannose residues on the outer face of gp120. J Virol *76*, 7306–7321.

Schatz, D.G. (2004). V(D)J recombination. Immunol Rev *200*, 5–11.

Schatz, D.G., Oettinger, M.A., and Baltimore, D. (1989). The V(D)J recombination activating gene, RAG-1. Cell *59*, 1035–1048.

Schatz, D.G., Oettinger, M.A., and Schlissel, M.S. (1992). V(D)J recombination: molecular biology and regulation. Annu Rev Immunol *10*, 359–383.

Scheid, J.F., Mouquet, H., Kofer, J., Yurasov, S., Nussenzweig, M.C., and Wardemann, H. (2011). Differential regulation of self-reactivity discriminates between IgG+ human circulating memory B cells and bone marrow plasma cells. Proc Natl Acad Sci USA *108*, 18044–18048.

Schelonka, R.L., Tanner, J., Zhuang, Y., Gartland, G.L., Zemlin, M., and Schroeder, H.W. (2007). Categorical selection of the antibody repertoire in splenic B cells. Eur. J. Immunol. *37*, 1010–1021.

Schelonka, R.L., Zemlin, M., Kobayashi, R., Ippolito, G.C., Zhuang, Y., Gartland, G.L., Szalai, A., Fujihashi, K., Rajewsky, K., and Schroeder, H.W. (2008). Preferential use of DH reading frame 2 alters B cell development and antigen-specific antibody production. J Immunol *181*, 8409–8415.

Schlissel, M., Constantinescu, A., Morrow, T., Baxter, M., and Peng, A. (1993). Double-strand signal sequence breaks in V(D)J recombination are blunt, 5'-phosphorylated, RAG-dependent, and cell cycle regulated. Genes Dev. *7*, 2520–2532.

Schroeder, H.W., and Cavacini, L. (2010). Structure and function of immunoglobulins. J Allergy Clin Immunol *125*, S41–S52.

Schroeder, H.W., and Wang, J.Y. (1990). Preferential utilization of conserved immunoglobulin heavy chain variable gene segments during human fetal life. Proc Natl Acad Sci USA *87*, 6146–6150.

Schroeder, H.W., Zemlin, M., Khass, M., Nguyen, H.H., and Schelonka, R.L. (2010). Genetic control of DH reading frame and its effect on B-cell development and antigen-specifc antibody production. Crit. Rev. Immunol. *30*, 327–344.

Shapiro, G., and Wysocki, L. (2002). DNA target motifs of somatic mutagenesis in antibody genes. Crit. Rev. Immunol. *22*, 183–200.

Shockett, P.E., and Schatz, D.G. (1999). DNA hairpin opening mediated by the RAG1 and RAG2 proteins. Molecular and Cellular Biology *19*, 4159–4166.

Siebenlist, U.U., Ravetch, J.V.J., Korsmeyer, S.S., Waldmann, T.T., and Leder, P.P. (1981). Human immunoglobulin D segments encoded in tandem multigenic families. Nature *294*, 631–635.

Simek, M.D., Rida, W., Priddy, F.H., Pung, P., Carrow, E., Laufer, D.S., Lehrman, J.K., Boaz, M., Tarragona-Fiol, T., Miiro, G., et al. (2009). Human immunodeficiency virus type 1 elite neutralizers: individuals with broad and potent neutralizing activity identified by using a high-throughput neutralization assay together with an analytical selection algorithm. J Virol *83*, 7337–7348.

Smith, K., Garman, L., Wrammert, J., Zheng, N.-Y., Capra, J.D., Ahmed, R., and Wilson, P.C. (2009). Rapid generation of fully human monoclonal antibodies specific to a vaccinating antigen. Nat Protoc *4*, 372–384.

Spurrier, B., Sampson, J.M., Totrov, M., Li, H., O'Neal, T., Williams, C., Robinson, J.E., Gorny, M.K., Zolla-Pazner, S., and Kong, X.-P. (2011). Structural analysis of human and macaque mAbs 2909 and 2.5B: implications for the configuration of the quaternary neutralizing epitope of HIV-1 gp120. Structure *19*, 691–699.

Stanfield, R.L., Gorny, M.K., Williams, C., Zolla-Pazner, S., and Wilson, I.A. (2004). Structural rationale for the broad neutralization of HIV-1 by human monoclonal antibody 447-52D. Structure *12*, 193–204.

Starcich, B.R., Hahn, B.H., Shaw, G.M., McNeely, P.D., Modrow, S., Wolf, H., Parks, E.S., Parks, W.P., Josephs, S.F., and Gallo, R.C. (1986). Identification and characterization of conserved and variable regions in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS. Cell *45*, 637–648.

Steen, S., Gomelsky, L., and Roth, D. (1996). The 12/23 rule is enforced at the cleavage step of V(D)J recombination in vivo. Genes Cells *1*, 543–553.

Stiegler, G., Kunert, R., Purtscher, M., Wolbank, S., Voglauer, R., Steindl, F., and Katinger, H. (2001). A potent cross-clade neutralizing human monoclonal antibody against a novel epitope on gp41 of human immunodeficiency virus type 1. Aids Res Hum Retrov *17*, 1757–1765.

Stijlemans, B., Conrath, K., Cortez-Retamozo, V., Van Xong, H., Wyns, L., Senter, P., Revets, H., De Baetselier, P., Muyldermans, S., and Magez, S. (2004). Efficient targeting of conserved cryptic epitopes of infectious agents by single domain antibodies. African trypanosomes as paradigm. J Biol Chem *279*, 1256–1261.

Taki, S., Meiering, M., and Rajewsky, K. (1993). Targeted insertion of a variable region gene into the immunoglobulin heavy chain locus. Science *262*, 1268–1271.

Throsby, M., van den Brink, E., Jongeneelen, M., Poon, L.L.M., Alard, P., Cornelissen, L., Bakker, A., Cox, F., van Deventer, E., Guan, Y., et al. (2008). Heterosubtypic neutralizing monoclonal antibodies cross-protective against H5N1 and H1N1 recovered from human IgM+ memory B cells. PLoS ONE *3*, e3942.

Tian, C., Luskin, G.K., Dischert, K.M., Higginbotham, J.N., Shepherd, B.E., and Crowe, J.E. (2007). Evidence for preferential Ig gene usage and differential TdT and exonuclease

activities in human naïve and memory B cells. Mol Immunol *44*, 2173–2183.

Tian, C., Luskin, G.K., Dischert, K.M., Higginbotham, J.N., Shepherd, B.E., and Crowe, J.E. (2008). Immunodominance of the VH1-46 antibody gene segment in the primary repertoire of human rotavirus-specific B cells is reduced in the memory compartment through somatic mutation of nondominant clones. J Immunol *180*, 3279–3288.

Tonegawa, S. (1983). Somatic generation of antibody diversity. Nature *302*, 575–581.

Trkola, A., Purtscher, M., Muster, T., Ballaun, C., Buchacher, A., Sullivan, N., Srinivasan, K., Sodroski, J., Moore, J.P., and Katinger, H. (1996). Human monoclonal antibody 2G12 defines a distinctive neutralization epitope on the gp120 glycoprotein of human immunodeficiency virus type 1. J Virol *70*, 1100–1108.

Vale, A.M., Tanner, J.M., Schelonka, R.L., Zhuang, Y., Zemlin, M., Gartland, G.L., and Schroeder, H.W. (2010). The peritoneal cavity B-2 antibody repertoire appears to reflect many of the same selective pressures that shape the B-1a and B-1b repertoires. J Immunol *185*, 6085–6095.

van Dongen, J.J.M., Langerak, A.W., Brüggemann, M., Evans, P.A.S., Hummel, M., Lavender, F.L., Delabesse, E., Davi, F., Schuuring, E., García-Sanz, R., et al. (2003). Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. Leukemia *17*, 2257–2317.

van Gent, D.C., Ramsden, D.A., and Gellert, M. (1996). The RAG1 and RAG2 proteins establish the 12/23 rule in V(D)J recombination. Cell *85*, 107–113.

Veazey, R.S., Shattock, R.J., Pope, M., Kirijan, J.C., Jones, J., Hu, Q., Ketas, T., Marx, P.A., Klasse, P.J., Burton, D.R., et al. (2003). Prevention of virus transmission to macaque monkeys by a vaginally applied monoclonal antibody to HIV-1 gp120. Nat Med *9*, 343–346.

Wagner, S.D., Milstein, C., and Neuberger, M.S. (1995). Codon bias targets mutation. Nature *376*, 732.

Walker, B.D., and Korber, B.T. (2001). Immune control of HIV: the obstacles of HLA and viral diversity. Nat Immunol *2*, 473–475.

Walker, J.R., Corpina, R.A., and Goldberg, J. (2001). Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. Nature *412*, 607–614.

Walker, L.M., and Burton, D.R. (2010). Rational antibody-based HIV-1 vaccine design: current approaches and future directions. Curr Opin Immunol *22*, 358–366.

Walker, L.M., Huber, M., Doores, K.J., Falkowska, E., Pejchal, R., Julien, J.-P., Wang, S.-K., Ramos, A., Chan-Hui, P.-Y., Moyle, M., et al. (2011). Broad neutralization coverage of HIV by multiple highly potent antibodies. Nature *477*, 466–470.

Walker, L.M., Phogat, S.K., Chan-Hui, P.-Y., Wagner, D., Phung, P., Goss, J.L., Wrin, T., Simek, M.D., Fling, S., Mitcham, J.L., et al. (2009). Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. Science *326*, 285–289.

Walker, L.M., Simek, M.D., Priddy, F., Gach, J.S., Wagner, D., Zwick, M.B., Phogat, S.K., Poignard, P., and Burton, D.R. (2010). A limited number of antibody specificities mediate broad and potent serum neutralization in selected HIV-1 infected individuals. PLoS Pathog *6*.

Wang, Y.-H., and Diamond, B. (2008). B cell receptor revision diminishes the autoreactive B cell response after antigen activation in mice. J. Clin. Invest. *118*, 2896–2907.

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

Watson, L.C., Moffatt-Blue, C.S., McDonald, R.Z., Kompfner, E., Ait-Azzouzene, D., Nemazee, D., Theofilopoulos, A.N., Kono, D.H., and Feeney, A.J. (2006). Paucity of V-D-D-J rearrangements and VH replacement events in lupus prone and nonautoimmune TdT-/- and TdT+/+ mice. J Immunol *177*, 1120–1128.

Wei, X., Decker, J., Wang, S., Hui, H., and Kappes, J. (2003). Antibody neutralization and escape by HIV-1. Nature.

Weiss, C.D. (2003). HIV-1 gp41: mediator of fusion and target for inhibition. Aids Rev *5*, 214–221.

Weitkamp, J.H., Kallewaard, N.L., Bowen, A.L., LaFleur, B.J., Greenberg, H.B., and Crowe, J.E. (2005). VH1-46 is the dominant immunoglobulin heavy chain gene segment in rotavirus-specific memory B cells expressing the intestinal homing receptor alpha4beta7. J Immunol *174*, 3454–3460.

Wiesendanger, M., Kneitz, B., Edelmann, W., and Scharff, M.D. (2000). Somatic hypermutation in MutS homologue (MSH)3-, MSH6-, and MSH3/MSH6-deficient mice reveals a role for the MSH2-MSH6 heterodimer in modulating the base substitution pattern. J Exp Med *191*, 579–584.

Wilson, P.C., de Bouteiller, O., Liu, Y., Potter, K., Banchereau, J., Capra, J.D., and Pascual, V. (1998a). Somatic hypermutation introduces insertions and deletions into immunoglobulin genes. J Exp Med *187*, 59–70.

Wilson, P.C., Liu, Y., Banchereau, J., Capra, J.D., and Pascual, V. (1998b). Amino acid insertions and deletions contribute to diversify the human Ig repertoire. Immunol Rev *162*, 143–151.

Wu, X., Yang, Z.-Y., Li, Y., Hogerkorp, C.-M., Schief, W.R., Seaman, M.S., Zhou, T., Schmidt, S.D., Wu, L., Xu, L., et al. (2010a). Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. Science *329*, 856–861.

Wu, X., Zhou, T., Zhu, J., Zhang, B., Georgiev, I., Wang, C., Chen, X., Longo, N.S., Louder, M., McKee, K., et al. (2011). Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. Science *333*, 1593–1602.

Wu, Y.-C., Kipling, D., Leong, H.S., Martin, V., Ademokun, A.A., and Dunn-Walters, D.K. (2010b). High-throughput immunoglobulin repertoire analysis distinguishes between human

IgM memory and switched memory B-cell populations. Blood *116*, 1070–1078.

Xiang, S.-H., Doka, N., Choudhary, R.K., Sodroski, J., and Robinson, J.E. (2002). Characterization of CD4-induced epitopes on the HIV type 1 gp120 envelope glycoprotein recognized by neutralizing human monoclonal antibodies. Aids Res Hum Retrov *18*, 1207–1217.

Zemlin, M., Schelonka, R.L., Ippolito, G.C., Zemlin, C., Zhuang, Y., Gartland, G.L., Nitschke, L., Pelkonen, J., Rajewsky, K., and Schroeder, H.W. (2008). Regulation of repertoire development through genetic control of DH reading frame preference. J Immunol *181*, 8416–8424.

Zhang, Z., Burrows, P.D., and Cooper, M.D. (2004). The molecular basis and biological significance of VH replacement. Immunol Rev *197*, 231–242.

Zhang, Z., Zemlin, M., Wang, Y.-H., Munfus, D., Huye, L.E., Findley, H.W., Bridges, S.L., Roth, D.B., Burrows, P.D., and Cooper, M.D. (2003). Contribution of Vh gene replacement to the primary B cell repertoire. Immunity *19*, 21–31.

Zheng, N.-Y., Wilson, K., Jared, M., and Wilson, P.C. (2005). Intricate targeting of immunoglobulin somatic hypermutation maximizes the efficiency of affinity maturation. J Exp Med *201*, 1467–1478.

Zhou, J., Lottenbach, K.R., Barenkamp, S.J., and Reason, D.C. (2004). Somatic hypermutation and diverse immunoglobulin gene usage in the human antibody response to the capsular polysaccharide of Streptococcus pneumoniae Type 6B. Infect. Immun. *72*, 3505–3514.

Zhou, T., Georgiev, I., Wu, X., Yang, Z.-Y., Dai, K., Finzi, A., Kwon, Y.D., Scheid, J.F., Shi, W., Xu, L., et al. (2010). Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. Science *329*, 811–817.

Zhou, T., Xu, L., Dey, B., Hessell, A.J., Van Ryk, D., Xiang, S.-H., Yang, X., Zhang, M.-Y., Zwick, M.B., Arthos, J., et al. (2007). Structural definition of a conserved neutralization epitope on HIV-1 gp120. Nature *445*, 732–737.

Zolla-Pazner, S., Zhong, P., Revesz, K., Volsky, B., Williams, C., Nyambi, P., and Gorny, M.K. (2004). The cross-clade neutralizing activity of a human monoclonal antibody is determined by the GPGR V3 motif of HIV type 1. Aids Res Hum Retrov *20*, 1254–1258.

Zouali, M. (1996). Nonrandom features of the human immunoglobulin variable region gene repertoire expressed in response to HIV-1. Appl Biochem Biotech *61*, 149–155.

Zwick, M.B., Komori, H.K., Stanfield, R.L., Church, S., Wang, M., Parren, P.W.H.I., Kunert, R., Katinger, H., Wilson, I.A., and Burton, D.R. (2004a). The long third complementarity-determining region of the heavy chain is important in the activity of the broadly neutralizing anti-human immunodeficiency virus type 1 antibody 2F5. J Virol *78*, 3155–3161.

Zwick, M.B., Labrijn, A.F., Wang, M., Spenlehauer, C., Saphire, E.O., Binley, J.M., Moore, J.P., Stiegler, G., Katinger, H., Burton, D.R., et al. (2001). Broadly neutralizing antibodies targeted to the membrane-proximal external region of human immunodeficiency virus type 1

glycoprotein gp41. J Virol *75*, 10892–10905.

Zwick, M.B., Parren, P.W.H.I., Saphire, E.O., Church, S., Wang, M., Scott, J.K., Dawson, P.E., Wilson, I.A., and Burton, D.R. (2003). Molecular features of the broadly neutralizing immunoglobulin G1 b12 required for recognition of human immunodeficiency virus type 1 gp120. J Virol *77*, 5863–5876.

Zwick, M.B., Saphire, E.O., and Burton, D.R. (2004b). gp41: HIV's shy protein. Nat Med *10*, 133–134.