RATIONAL DESIGN OF ANTIBODIES: FROM MECHANISMS OF SPECIFICITY TO NOVEL VACCINE STRATEGIES

By

Jordan R. Willis

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

CHEMICAL AND PHYSICAL BIOLOGY

August, 2014

Nashville, Tennessee

Approved:	Date:
Benjamin Spiller, Ph.D (Chair)	
Christopher Aiken, Ph.D	
Spyros Kalams, M.D.	
Jens Meiler, Ph.D (Advisor)	
James E. Crowe, Jr., M.D. (Advisor)	

For my friends and family,

who convinced me to take the red pill.

Acknowledgements

This work would not have been possible without the following people. First and foremost, I would like to thank my two mentors, James Crowe and Jens Meiler. The general consensus of the biomedical science community is that principal investigators often dislike interdepartmental collaboration. When I talk to other graduate students, I find that they have many unpleasant experiences with just a single investigator. I feel very fortunate to have two PIs with whom I get along so well. From opposite ends of the spectrums of science, they never objected to such a collaboration. Were it only a few years ago this feat may have been unattainable. Both of you are an inspiration to my scientific future. Jens with his incredible and often frustrating brilliance of the task at hand, and Jim with his scientific leadership. I could not have asked for two better bosses to nurture my scientific growth.

I would lake to thank all my colleagues in the Crowe and Meiler Labs. An early thanks goes to Kristian Kaufmann, Ralf Mueller, and Mark Hicar, as their initial guidance through my rotation was a deciding factor in my choice for the future. There are innumerable scientific colleagues who helped me with my work including Jessica Finn who helped with the high-throughput sequencing, Mark Hicar who taught me about HIV virology, Sam DeLuca who would answer every question I had about computers, Gordon Lemmon who taught me scripting, Steven Combs who guided me through ROSETTA, Natalie Thornburg for teaching me many molecular biology techniques. I would also like to thank the lab technicians who made this throughput of work possible, Rachelle Falk, Vidisha Singh, and Hannah King. Special thanks go to a dear friend and brilliant scientist Gopal Sapparapru. Your troubleshooting techniques, management of large amounts of experimental data, and guidance through countless experiments has helped make me a better scientist. I would like to make a very special mention to my very good friend and colleague Bryan Briney, who taught me almost everything I know about immunology and high-throughput sequencing. Our names appear together on many publications for good reason.

I want to thank my friends and family. Graduate school is a long arduous process and without them I would not be where I am today. First, my sister, who I cannot live up to. You have a heart of gold. My Mother, my number one fan. Her constant encouragement has helped me survive graduate school. I'm lucky to have her in my life and to have spent a majority of my time in Nashville with her close. She is a great source of inspiration in all that I do. A constant drive and loving force. My Father, my role model. His support for me has been tremendous. His drive and motivation showed me what a little hard work can do. I hope to become half of the man he is today. My friends, helping me keep sanity through graduate school. I would especially like to thank Sean Welch, who allowed me sanctuary away from graduate school. You are also a tremendous inspiration. Lastly I would like to thank my best friend in graduate school, my cat Ocho, the only constant through these last six years. I love you little buddy! Here's to many more years together.

Summary

This document is the culmination of my work on antibody design. Primarily, my target system is HIV with some work on Influenza. It is divided up into orthogonal publications, with each publication having incorporated an aspect of antibody design. Here I use the modeling suite ROSETTA whenever I mention *in silico* solutions to design problems. All computational work was done by me as well as validation with the experimental characterization.

The introduction in chapter I outlines four pieces of background knowledge that must be considered for the remaining chapters to become clear. I first detail antibodies in general, as this document only considers immunoglobulins as the design target of interest. I detail their purpose and how they are diversified to create an immunoglobulin repertoire. I next detailed the pandemic of HIV, its structure, and the current state of an efficacious vaccine, and describe the broadly neutralizing antibodies that have been characterized to date. I also describe the molecular modeling suite ROSETTA and briefly cover the ROSETTA energy function. I detail a particular application in ROSETTA known as ROSETTADESIGN, which is the focus of my thesis. I then describe how ROSETTADESIGN has been used in translational medicine. Lastly I detail the current state of the field, the difficulties in molecular modeling, the challenges in protein design, and how this work can be used to aid in these challenges. Briefly, I describe how antibody design is used to answer questions about basic science to implications for vaccine strategies. It is here that I tie antibody design with all aspects of my research.

The beginning of my research starts in chapter II. This part of my research aims to answer a basic question in immunology. Here I used ROSETTADESIGN to investigate the molecular basis for polyspecificity. It is known that a finite set of antibodies is able to accommodate a nearly limitless antigen space. Using design, I investigated which sequences are ideal to bind multiple antigens or single antigens in a protocol I used called multi-state

design.

Using the strategies and principles built upon in chapter II, chapter III focuses on a novel vaccine paradigm. Here I used antibody design to interrogate the HIV-naïve donor antibody repertoire with a simple goal in mind: Does the HIV-naïve antibody repertoire contain antibodies that resemble known broadly neutralizing antibodies? This paradigm focuses on a structural mimicry rather than a sequence identity, which not only allows me to use ROSETTADESIGN and homology modeling as a tool to investigate this hypothesis, but mandates it, showing the necessity of molecular modeling. All antibodies found in this chapter were made experimentally and characterized to corroborate computational findings.

In chapter IV, I used ROSETTADESIGN to show that the antibody PG9, which is known to be one of the most broad and potent antibodies against HIV-1 characterized to date, still has room for improvement. Mutations were returned from ROSETTADESIGN which were predicted to enhance breadth and specificity. These mutations were tested experimentally and did indeed corroborate our hypothesis that even the most broad and potent antibodies could be improved with careful computational design and analysis.

Finally, chapter V details future directions I foresee for these project. These strategies are generalizable and can be applied to any given antibody/antigen system and may even extend to any given protein-protein interface. In addition, I consider reasons for design failures, imperfect sequence recovery, and antibodies that failed to corroborate *in silico* predictions. I give an idea of many future experiments that could be used to take of advantage of the information detailed in this document. I also detail some of my other work on viral escape assessed by computational modeling and broadly neutralizing mAbs to Influenza which were in various stages of completion.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

Introduction

I.1 Antibody Overview

The lymphocytes that make up the adaptive immune system have evolved to recognize a limitless number of antigens that constitute viruses, bacteria and all foreign material to a hosts immune system (Murphy et al., 2007). The concern of this thesis is on the antigen-recognition molecules of the B-cell known as immunoglobulins (Ig). These can exist either as a membrane-anchored form on the B-cell known as the B-cell receptor (BCR) or as a secreted form with a wide range of functionality known as antibodies. The main effector function of antibodies is to bind foreign pathogens in the body, and this is the basis for the adaptive immune response. Antibodies have two separate functions. One is to bind specifically to molecules known as antigens. The other is to recruit other cells and molecules to destroy the pathogens to which immunoglobulins are bound. There are two genetic domains that make up the antibody structure and differentiate these processes (figure I.1). One is the variable domain responsible for specificity. The other is the constant domain that engages different effector functions such as complement recruitment and phagocytosis of compromised cells. Structurally, antibodies consist of two identical heavy chains, that are encoded by recombined gene segments of the heavy variable and constant domain gene segments, and two identical light chains, that are encoded by recombined copies of light chain variable and constant domains gene segments.

The variability of antibodies is what ensures that any individual with a functional immune system can produce antibodies to recognize almost any structure. The mechanisms of variability are discussed in further sections but are typically distributed to the tips of the antibody. It is important to note that B-cell bound receptors and effector functions of antibodies play important roles in the humoral immune system, but the remainder of this

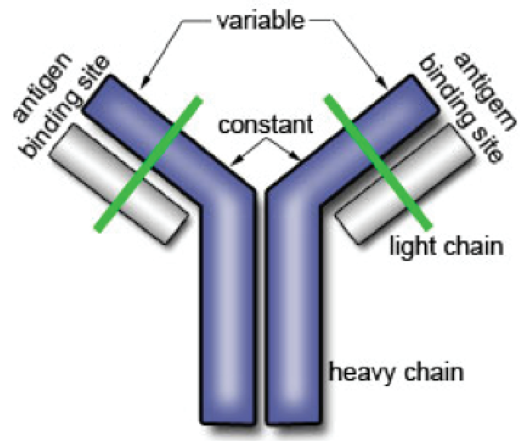


Figure I.1: Overview of antibody structure. Heavy chain is shown in blue, light chain in grey. The structure is divided into the variable portion responsible for recognition, and the constant portion responsible for effector function. The tips of the antibodies are where antigens typically bind and are therefore known as the antigen binding sites. Image reproduced from http://crdd.osdd.net/raghava/absource/abasic.html.

document will focus on the diversity and specificity of secreted antibodies of the IgG class, the most common circulating immunoglobulin.

I.1.1 Antibody Diversification

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 (IGK) are located on chromosome 22 (Brochet et al., 2008). Each of these loci consists of multiple variable (V, not to be confused with the variable region of an antibody) and joining (J) gene segments. In addition, the IGH locus also contains several diversity (D) gene segments. Sequencing of the human IGH locus revealed 55 functional V genes, 23 D genes, and six J genes (Matsuda et al., 1998; Lefranc et al., 2009).

The human variable genes (and, at the IGH locus, the diversity genes) can be grouped phylogenetically into families based on sequence similarity. Heavy chain variable genes are organized into seven families and homology within gene families is typically above 80%. The 23 functional human diversity genes are also organized into seven families. For an example variable gene, IGVH5-15*01, 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 (5 in this example), the gene number (determined by position in the germline locus, 15 in this example), 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.

I.1.1.1 Recombination to Enable Diversity

The tremendous sequence and structural diversity of antibodies can be attributed to two immunologic processes that act on antibody germline gene segments. The first is the initial recombination initiated by the recombination activating gene machinery (RAG) (Brack et al., 1978; Alt and Baltimore, 1982; Tonegawa, 1983; Schatz et al., 1989; Oettinger et al., 1990). The RAG machinery is responsible for the recombination of V, D, and J gene for the heavy chain, and the V and J gene for the light chain (figure I.2, left-panel). This process takes place to make functional B-cell receptors in the bone marrow, before antigenic stimuli. If a B-cell receptor is found to bind self-antigens of the host, it is eliminated. This clonal selection and deletion is the fundamental process for which antibodies are able to recognize foreign antigens while not attacking the host.

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) encoded proteins at the initiation of the recombination process (Hesse et al., 1989; Alt et al., 1992). 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 and Ji, 2011). After binding to one 12-bp RSS and one 23-bp RSS, the RAG complex induces single-strand DNA nicks between the coding sequence and the heptamer of each RSS, resulting in hair-

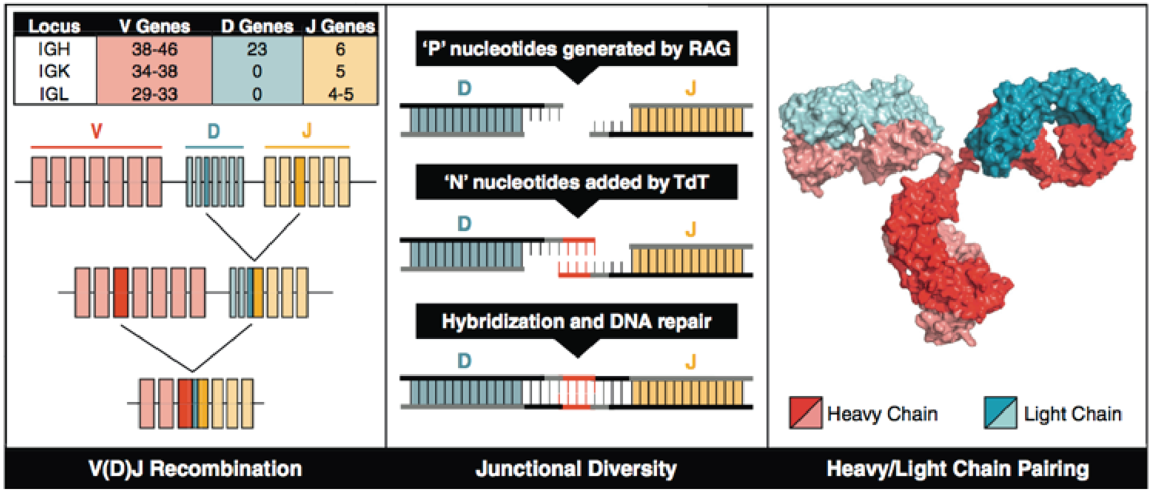


Figure I.2: Overview of Antibody Recombination. Diversity in the antigen-combining site of the B-cell receptor repertoire (and thus also in the corresponding secreted antibody repertoire) is mediated by three principal molecular mechanisms, illustrated in the three panels, left, middle, and right. Figure adapted from (Finn and Crowe, 2013).

pin 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 with another components of the RAG mutation machinery known as Artemis (Ma et al., 2002). Nucleotides may be added or removed from the coding ends, and the double strand breaks at the coding ends are joined into a single coding strand with DNA ligase IV (Lewis, 1994; Mahajan et al., 1999; Shockett and Schatz, 1999; Walker et al., 2001; Mansilla-Soto and Cortes, 2003; Roth, 2003) (figure I.2, middle panel). The newly recombined gene produces an antibody (figure I.2, right panel). Recombination of the light chain is similar but the result of a single $V_L J_L$ recombination. To establish a diverse naïve antibody repertoire, that is, antibodies that have not encountered antigens, these events of RAG-mediated recombination produce an initial repertoire of $3\ x\ 10^{11}$ different recombinations. This process happens before antigen stimulus in the bone marrow leading to the next form of antibody diversity, somatic hypermutation.

I.1.1.2 Somatic Hypermutation to Enable Diversity

Maturation of the antibody repertoire to hone specificity is known as somatic hypermutation (SHM) and is initiated by the somatic hypermutation machinery (SHMM). Somatic hypermutation is the response to antigen stimulus and occurs in various lymph tissues to diversify the antibody repertoire (Tonegawa, 1983; Brenner and Milstein, 1966).

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 lymphoid tissue, mutate the variable region of their antibody genes (figure I.3) (Li et al., 2004; MacLennan et al., 1992). 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). SHM introduces point mutations at a frequency of approximately 1 in 10^3 per base pair, which is 10^6-fold higher than the rate of spontaneous mutation in other genes (Rajewsky et al., 1987). 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., 2000, 1999). Deamination thus produces a uracil-guanine mismatch, and several 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).

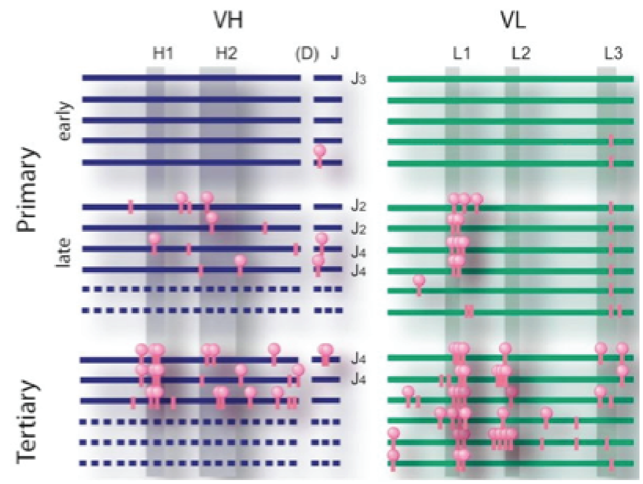


Figure I.3: Somatic mutations in response to antigen stimulus. The VH gene and VL gene are shown for various VH and VL pairs represented by blue and green lines. The CDR loops H1-H3 and L1-L3 are darkened. The pink dots represent mutations. The early response has little to no somatic mutations recapitulating naïve repertoire. The late and response starts developing mutations. A second challenge with the same antigen shows even more mutations to hone specificity. Figure adapted from (Berek and Milstein, 1988), redrawn by C., Scotti.

I.1.1.3 Implications for Antibody Structure

Antibody complementarity determining regions (CDRs, also referred to as hypervariable regions, figure I.4) are the primary region of antigen recognition that lie in loop regions of the antibody framework (figure I.4). They are preferentially targeted for affinity maturation by the SHM machinery, making them the most variable regions of the antibody gene (Padlan and Padlan, 1994). Structurally, the CDRs are largely loop-based, which makes them sufficiently flexible to incorporate the substitutions without compromising structural integrity. Framework regions (FRs) are highly structured and less able to accommodate somatic mutations (Jimenez et al., 2003).

I.2 HIV Pandemic Overview

HIV-1 is an unprecedented health problem that continues to remain a worldwide pandemic. Since the recognition of acquired immune deficiency syndrome (AIDS) in 1981 (Gottlieb et al., 1981) followed by the discovery of its causative agent, human immunodeficiency

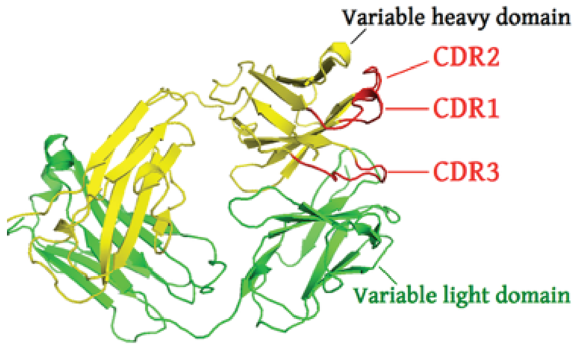


Figure I.4: Antibody structure with CDRs.The light chain in green with the LCDRs not pictured. The heavy chain is shown in yellow with the HCDRs shown in red. Figure adapted from PDB: 1IGT (Harris et al., 1997).

virus (HIV) in 1983 (Barré-Sinoussi et al., 1983), an estimated 65 million have been infected with over 25 million deaths (Hemelaar, 2012). The amount of people estimated to living with HIV is 30 million, many of whom live in the developing world (UNAIDS, 2013).

More than 40 different non-human primate species harbor SIV infections, with each species carrying a species-specific virus. Each independent zoonotic transmission can generate a different lineage. HIV type 1 (HIV-1), thought to be transmitted from chimpanzees in the Congo around 1900 (Keele, 2006), is the most common and further is split into groups M, N, O, and P. HIV-1 group M is responsible for the global pandemic and is further split into subtypes clades A-D,F-H, J and K that are tropic to specific regions. Within each subtype, variation of the amino acids vary as much as 30%. For example, clade B is the most common in North America while clade C is the most common in Sub-Saharan Africa (figure I.5). If full genome sequences are found that result from recombinations of different HIV-1 group M subtypes, they are designated circulating recombinant forms (CRFs). If they are epidemiologically linked or unique recombinant forms they are unlinked (URF) (Robertson et al., 2000).

A major contributing factor to HIV spread and defense is its potential for enormous genetic diversity. This genetic diversity stems from the error rates of the reverse-transcription

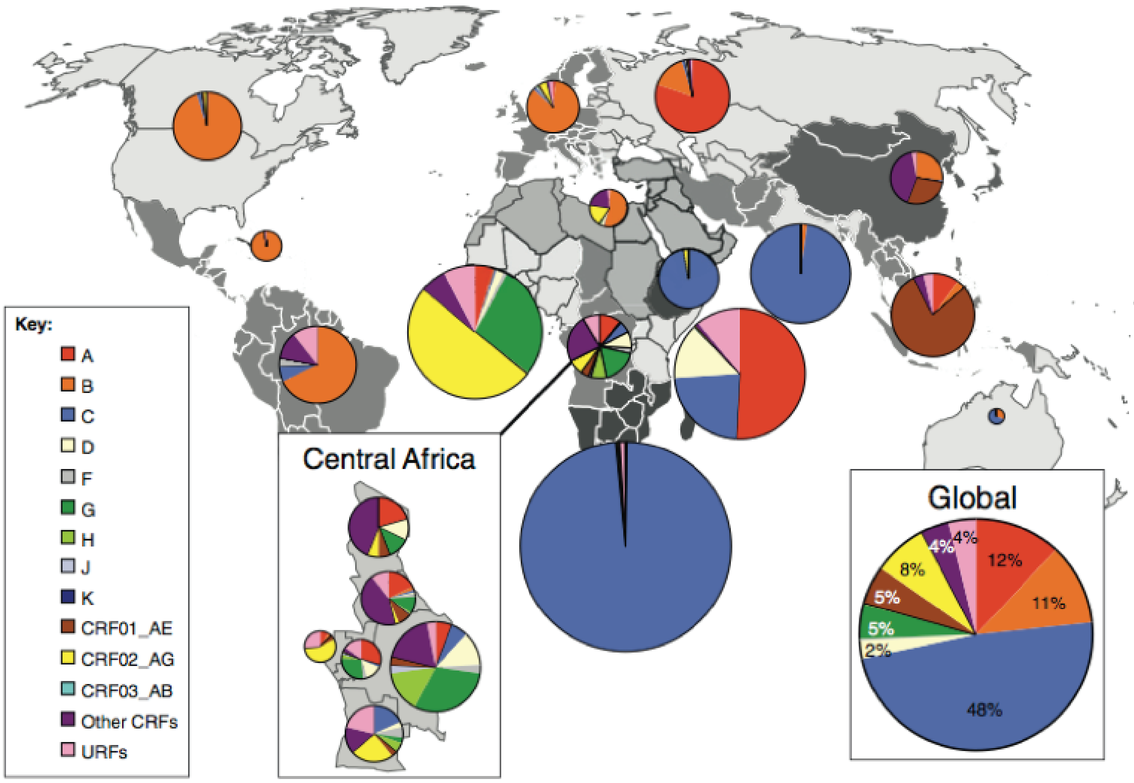


Figure I.5: Global distributions of HIV-1 subtypes. In the main figure, pie charts representing the distribution of HIV-1 subtypes and recombinants from 2004 to 2007, colored by HIV-1 subtype. Adapted from UNAIDS report 2013.

.

machinery which lacks proof-reading capabilities (Ho et al., 1995). This genetic diversity, particularly in the envelope gene (Env), gives rise to sequence divergence of up to 10% within a single individual (Korber et al., 2001). This is one of the many defense mechanisms HIV uses to evade host response and contemporary vaccination strategies.

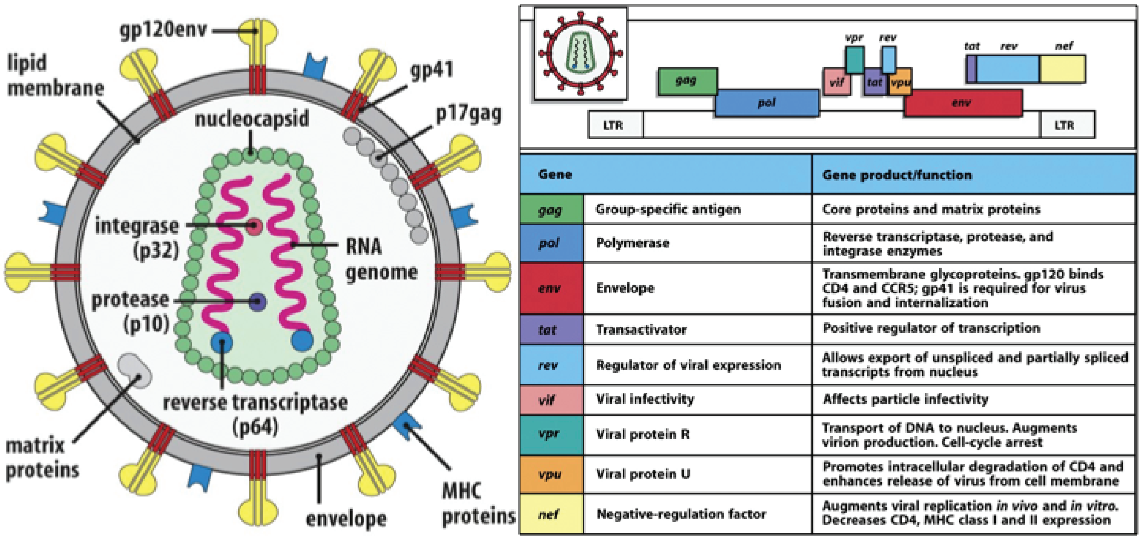
I.2.1 The HIV Virus Genome and Structure

HIV-1 is an enveloped virus containing a duplicate positive-strand RNA genome (figure I.6, left panel). The functional spike on the surface of the virion is the envelope (Env) glycoprotein and is coded by the *Env* gene (figure I.6, right panel). 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 assembled into a trimer of gp120-gp41 heterodimers that are expressed on the virion surface (Kowalski et al., 1987). 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 principal function of the glycoprotein spike is to facilitate cell entry by binding to the primary cell-surface receptor, CD4, and one of the two common co-receptors, CCR5 or 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., 2001).

The remaining genes for HIV-1 include viral enzymes such as the error-prone reverse transcriptase, the integrase that allows integration of viral DNA to the host genome, and protease, to allow cleavage of gene products into their functional subunits. There are also structural proteins that lie below the envelope that make up the inner matrix and nucleocapsid. There are also several accessory proteins, vpu, vif, vpr, p6, nef, rev, and tat which aid in combating host defense or enhancing viral fitness (Fields et al., 2007). All of these proteins serve a significant purpose to the virulence and life cycle of the HIV-1 virus, but will not be discussed further as they have low antigenicity for the induction of antibodies, the primary focus of my research. However, a list of their genes and gene products can be found in figure I.6.

I.2.2 The Viral Spike and Humoral Resistance

The failure of conventional experimental vaccine candidates to prime the immune system for a broad response against HIV-1 challenge is partially explained thorough the structural definition of the HIV-1 spike. Much of the surface is covered in carbohydrates that shield neutralizing epitopes (Binley et al., 2010). The conserved CD4 binding site is recessed and sits behind the hypervariable loops (Burton et al., 2005). The co-receptor binding



Gene		Gene product/function
gag	Group-specific antigen	Core proteins and matrix proteins
pol	Polymerase	Reverse transcriptase, protease, and integrase enzymes
env	Envelope	Transmembrane glycoproteins. gp120 binds CD4 and CCR5; gp41 is required for virus fusion and internalization
tat	Transactivator	Positive regulator of transcription
rev	Regulator of viral expression	Allows export of unspliced and partially spliced transcripts from nucleus
vif	Viral infectivity	Affects particle infectivity
vpr	Viral protein R	Transport of DNA to nucleus. Augments virion production. Cell-cycle arrest
vpu	Viral protein U	Promotes intracellular degradation of CD4 and enhances release of virus from cell membrane
nef	Negative-regulation factor	Augments viral replication *in vivo* and *in vitro*. Decreases CD4, MHC class I and II expression

Figure I.6: Simplified view of HIV structure and genome. The proteins that make up the virus structure are displayed as a schematic. The virus is coded by a duplicated RNA genome (pink) surrounding by a viral nucleocapsid proteins. The inner envelope is supported by gag protein with gp120 envelope shown as a trimer bound to gp41. These trimeric "spike" are responsible for infectivity by binding CD4-binding sites (left panel). Each gene is represented in a different color and localizes to either the nucleocapsid (green) or the outer envelope (red). Figure adapted from (Murphy et al., 2007)

site is also recessed unless CD4 has triggered a conformational shift exposing this region (Harris et al., 2011). Another defense is the relative lability of the trimeric complex (Wyatt and Sodroski, 1998). The gp120 head often sheds creating "stumps" that serve as decoy epitopes against the viral complex (Liu et al., 2008). In addition there are few functional trimeric spikes on the surface of HIV, limiting immune response to a few locations on the virion.

The biggest defense is sequence variability. Much of the antibody response is targeted to the hypervariable loops that can easily change sequence without much consequence to viral fitness. This is why the humoral response produces autologous or strain-specific neutralizing antibodies that must catch up to a constantly evolving antigenic target (Albert et al., 1990; Gray et al., 2007; Pilgrim et al., 1997; Richman et al., 2003; Sagar et al., 2006; Wei et al., 2003).

I.3 Broadly Neutralizing Antibodies to HIV

Given the major defenses of the HIV Env structure, there is a rather discouraging view for vaccine development. In fact, only four modestly neutralizing antibodies were discovered between 1991 and 2009 (Burton et al., 2005; Kwong and Mascola, 2012), two membrane proximal extracellular region (MPER) binding mAbs 2F5 and 4E10 (Zwick et al., 2001; Muster et al., 1994), a CD4 binding site neutralizing mAb b12 (Burton et al., 1994), and a complex carbohydrate binding mAbs 2G12 (Trkola et al., 1996) (table I.1, figure I.7).

It was thought that the conventional vaccine strategies could not stimulate the immune system to produce broadly neutralizing antibodies to HIV due to the extreme variability of the viruses and the capability of the virus to escape antibody responses. Recently, technologies like high throughput neutralization assays were developed that could rapidly test sera for neutralization capacity *in vitro*, allowing researchers to accurately quantify the neutralizing response of HIV infected patients (Binley et al., 2004; Blish et al., 2007; Li et al., 2005; Mascola et al., 2005; Montefiori, 2005). Several groups found that there were multiple patients who could neutralize very genetically diverse panels of HIV variants, even those variants that were not in that patients sub-type (Binley et al., 2008; Doria-Rose et al., 2010; Simek et al., 2009; Wu et al., 2006). That led to longitudinal studies to show how long it took for a broadly neutralizing response to develop. Researchers showed that this generally took anywhere from 2-4 years (Gray et al., 2011; Mikell et al., 2011; Moore et al., 2011), with earliest time points arising at 1 year (Doria-Rose et al., 2014). The question still remained if those neutralizing responses were caused by few monoclonal antibody responses or just a large polyclonal response (Gray et al., 2007; Binley et al., 2008; Li et al., 2007; Rong et al., 2009; Sather et al., 2009; Scheid et al., 2009; Tomaras et al., 2011; Walker et al., 2010).

The question was answered by the recent explosion of newly discovered broadly neutralizing antibodies isolated by multiple research groups (Corti and Lanzavecchia, 2013). It started with two new isolates, PG9 and PG16, from an African donor that led to the discovery of a completely new neutralizing epitope, which is the focus of my research (figure I.7) (Walker et al., 2009). Both PG9 and PG16 bind to a proteoglycan epitope through an extended HCDR3 structure (McLellan et al., 2011).

The discovery of PG9 and PG16 led to newly characterized antibodies using similar techniques such as microneutralization screening, high-throughput sequencing, and hybridoma technology. The Haynes laboratory characterized additional long HCDR3 antibodies that bound similarly as PG9 and PG16 but with less breadth (Bonsignori et al., 2011). There are other classes of glycan-dependent antibodies isolated by the Poignard group that bind the V3 and beta-strands that are higher potency than PG9 and PG16 (Walker et al., 2011). Other MPER antibodies have also been characterized to by the Connors group such as 10e8 that neutralizes 98% of viruses (Huang et al., 2012). Focused epitopes designed computationally also can be used to identity some of the most potent antibodies to date (the VRC series) (Wu et al., 2010a). These antibodies were identified using a designed protein scaffold of gp120 that "knocked-out" non-neutralizing epitopes. Thus, only neutralizing antibodies would be isolated upon binding. I will not elaborate further on all of the antibodies characterized to date, their method of isolation and if any longitudinal studies were used to determine their pathways of development. These characteristics are summarized in table I.1 and figure I.7.

It is interesting to note, and important for the work that will be presented here, that the broadly neutralizing antibodies to date share one of two characteristics. They are either highly somatically mutated, indicative of years of chronic infection and selective pressure, or have a very long HCDR3 (figure I.8). Both of these characteristics, long HCDRs or highly mutated genes, make it difficult to elicit such antibodies in a vaccine attempt, but will be discussed further in the upcoming chapters.

Antibody	Specificity	Breadth	V_H	SHM	HCDR3 Length	Screening Strategy
2F5	gp41 MPER	~60-70%	2-5	15.2	24	gp160 and p24 binding
4E10	gp41 MPER	~96-98%	1-69	15.6	20	gp160 and p24 binding
1EO8	gp41 MPER	~98%	3-15	22.1	22	Microneutralization
2G12	gp120 glycans	~25-30%	3-21	33.6	16	gp160 and p24 binding
PGT128	Glycans and V3 β-strand	~70-75%	4-39	27.9	21	Microneutralization
PGT127	Glycans and V3 β-strand	~50%	4-39	23.2	21	Microneutralization
PGT121	Complex type V3 N-glycans	~65-70%	4-59	21.2	26	Microneutralization
10-1074	Complex type V3 N-glycans	~55-60%	4-59	24.4	26	gp140 binding
PGT135	Glycans and V4	~30-35%	4-39	26.8	20	Microneutralization
PG9/PG16	Glycans and V1/V2	~75-80%	3-33	15.4-16.8	30	Microneutralization
CH01- CH04	Glycans and V1/V2	~50%	3-20	23.3-19.5	26	Microneutralization
PGT145	Glycans and V1/V2	~75-80%	1-8	22.8	33	Microneutralization
b12	gp120 CD4bs	~30-35%	1-3	17.3	20	Phage library
HJ16	gp120 CD4bs	~30-35%	3-30	36.7	21	EBV- immortalization
VRC01	gp120 CD4bs	~90-95%	1-2	38.7	14	Cell sorting/RT-PCR
VRC03	gp120 CD4bs	~50%	1-2	34.9	16	Cell sorting/RT-PCR
3BNC117	gp120 CD4bs	~85-90%	1-2	36.9	12	Cell sorting/RT-PCR
3BNC60	gp120 CD4bs	NA	1-2	36.9	12	Cell sorting/RT-PCR
NIH45-46	gp120 CD4bs	~90%	1-2	44	18	Cell sorting/RT-PCR
CH30- CH34	gp120 CD4bs	~80%	1-2	31.9-31.9	15	Cell sorting/RT-PCR
PGV04	gp120 CD4bs	~85-90%	1-2	38.2	16	Cell sorting/RT-PCR
3BC176	CD4i/V3	~60-70%	1-2	29.4	19	Cell sorting/RT-PCR

Table I.1: Broadly neutralizing antibody properties. Breadth refers to the amount of viruses tested that fall below 50 μg/mL. V_H is the heavy chain accessed from IMGT, SHM is the somatic hypermutation percentage of heavy chains as assessed from IMGT. Table adapted from (Corti and Lanzavecchia, 2013).

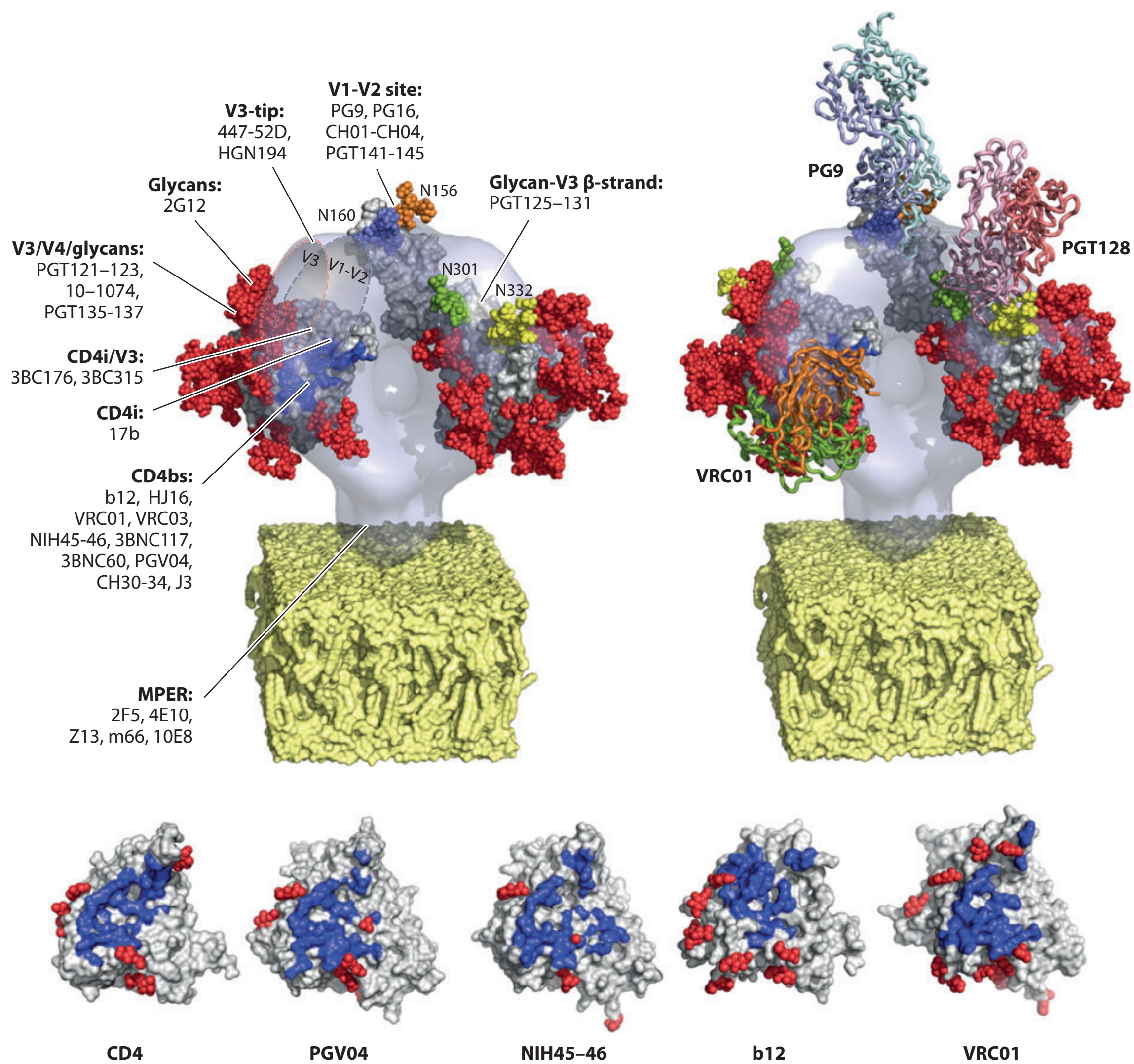


Figure I.7: Model of the HIV-1 Env trimeric glycoprotein bound to broadly neutralizing antibodies. The left panel shows the major sites targeted by broadly neutralizing antibodies. The approximate positioning of the V1/V2 and V3 loops is shown, and the CD4 footprint on the gp120 monomer is highlighted in blue. The right panel shows the Fabs of broadly neutralizing antibodies VRC01 (3NGB), PG9 (3US2), and PGT128 (3TYG) bound to gp120. Carbohydrates (oligomannose, red spheres) were modeled on the unliganded YU2 gp120 core (3TGQ) using GlyProt, with the exception of the glycans bound by PGT128 and PG9 (depicted with different colors), which were taken from the structures. The location of PG9 above the trimeric gp120 is approximate; VRC01 and PGT128 Fabs were manually positioned by with the unliganded YU2 gp120 model (approximation). The bottom panel shows in blue the footprints of CD4 (1GC1) and the CD4bs-specific antibodies PGV04 (3SE9), NIH45-46 (3U7Y), b12 (3DNL), and VRC01 (3NGB). Figure adapted from (Corti and Lanzavecchia, 2013)

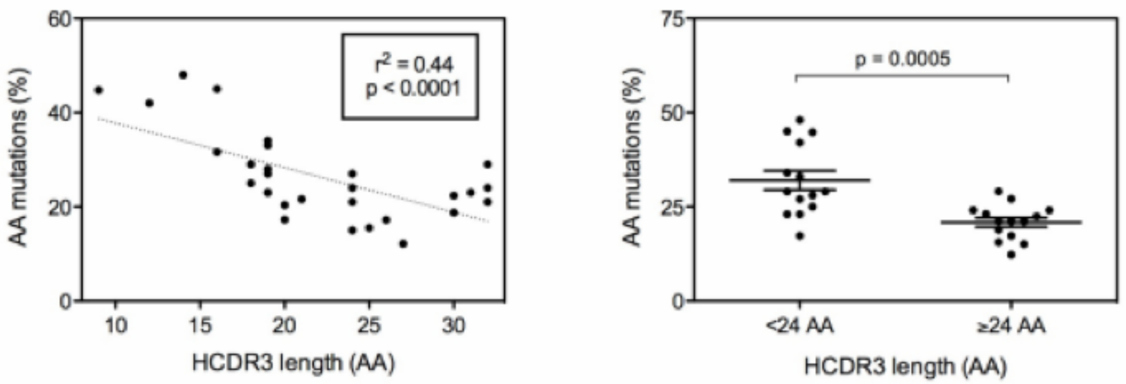


Figure I.8: Trends of HIV bNAbs. Data from a representative panel of 30 antibodies is shown. Plotted on the y-axis is the frequency of amino acid mutations of the currently characterized bNAbs. On the x-axis is the length of the heavy chain CDR3 (HCDR3). A negative correlation exists between the frequency of mutations and the HCDR3 length ($r^2 = 0.44$, left panel). When long HCDR3s (>24 AA) are binned against canonical length HCDR3s (<24AA), there is a statistical significance between the frequencies of amino acid mutation ($p = 0.0005$, right panel).

I.4 ROSETTA

Many software packages exist for the specific tasks of threading, minimization, and design. The ROSETTA software suite includes algorithms for all of these tasks and was developed for computational modeling and analysis of protein structures; further, it is free for non-commercial users. It has enabled notable scientific advances in computational biology, including *de novo* protein prediction, protein design, enzyme design, ligand docking and structure prediction of biological macromolecules and macromolecular complexes (Rohl et al., 2004; Siegel et al., 2010; Kuhlman et al., 2003; Davis and Baker, 2009; Misura et al., 2006; Davis et al., 2009; Das and Baker, 2008). The broad spectrum of applications available through ROSETTA allows for multiple computational problems to be addressed in one software framework. To aid in the understanding of ROSETTA-specific language, a supplementary glossary has been provided in Appendix VI.1.

One of the most common applications of ROSETTA is protein structure prediction via *de novo* folding and comparative modeling (Kaufmann et al., 2010; Rohl et al., 2004). *De novo* folding can be used to predict a protein's tertiary structure when only the primary sequence of a protein is known. However, to date, ROSETTA has been shown to successfully fold only small, soluble proteins (fewer than 150 amino acids), and it performs best if the proteins are mainly composed of secondary structural elements (Meiler and Baker, 2003). Structures of helical membrane proteins between 51 and 145 residues were predicted to within 4Å of the native structure (Yarov-Yarovoy et al., 2006), but only very small proteins (up to 80 residues) have been predicted to atomic-detail accuracy (Bradley et al., 2005a,b; Das et al., 2007). Accurate prediction of larger and/or more complex proteins can be achieved with the addition of experimental data, such as NMR chemical shifts and electron tomography maps (Rohl, 2005; Lange et al., 2012; Lange and Baker, 2012).

Another application, protein threading, refers to the tolerance of a tertiary fold given PDB coordinates. The ROSETTA scoring function evaluates how well a sequence can "'tolerate" a structure. It is often referred to as the "inverse folding problem". The known template structure of which a sequence will be threaded reduces almost all-conformational space by providing a protein backbone scaffold. Threading has played a major role in aiding experimental design and the interpretation of experimental results. Results can be used to help predict structure-function relationships (Kaufmann et al., 2009), and aid in designing proteins for binding pathogens (Azoitei et al., 2011; Correia et al., 2010, 2011a,b; Schief et al., 2009), determining thermostable proteins (Stranges and Kuhlman, 2013; Kuhlman et al., 2002; Der and Kuhlman, 2011), and aid in the determination of target residues for site-directed mutagenesis (Keeble et al., 2008; Fortenberry et al., 2011).

I.4.1 The ROSETTA Energy Function

All of the applications described above rely on a metric to score predictive models. This metric in ROSETTA is referred to as the ROSETTA energy function. The scoring function in ROSETTA is derived empirically through analysis of observed geometries of a subset of proteins in the PDB. We call this scoring function a knowledge-based scoring function, since it relies on previous knowledge of observed structures. The measurements include, but are not limited to, radius of gyration, packing density, distance/angle between hydrogen bonds and distance between two polar atoms. The measurements are converted into an energy function through Bayesian statistics (Simons et al., 1997; Metropolis et al., 1953).

The scoring function in ROSETTA can be separated into two main categories: centroid-based scoring and all-atom scoring. The former is used for *de novo* folding and initial rounds of loop building (Rohl, 2005; Simons et al., 1997, 1999b). The side chains are represented as "super-atoms", or "centroids", which limit the degrees of freedom to be sampled while preserving some of the chemical and physical properties of the side chain. Although this centroid-based scoring function is important for *de novo* folding, the folding protocol is not covered within the scope of this document.

The all-atom scoring function represents side chains in atomic detail. Similarly to the centroid-based scoring function, the all-atom scoring function comprises weighted individual terms that are summed to create a total energy for a protein. Most of the scoring terms are derived from knowledge-based potentials. The scoring function contains Newtonian physics based terms, including a 6-12 Lennard-Jones potential and a solvation potential. The 6-12 Lennard-Jones potential is split into two terms, an attractive term (fa_-atr) and a repulsive term (fa_rep), for all van der Waals interactions (Kuhlman and Baker, 2000; Neria et al., 1996). The solvation potential (fa_sol) models water implicitly and penalizes the burial of polar atoms (Lazaridis and Karplus, 1999). Interatomic electrostatic interactions are captured through a pair potential (fa_pair) (Simons et al., 1999b), and an orientation-dependent hydrogen bond potential for long-range and short-range hydrogen bonding (hbond_sc, hbond_lr_bb, hbond_sr_bb, and hbond_bb_sc, respectively) (Gordon et al., 1999; Wedemeyer and Baker, 2003). In addition to the electrostatic terms, the ROSETTA all-atom scoring function contains terms that dictate side chain conformations according to the Dunbrack rotamer library (fa_dun) (Dunbrack and Karplus, 1993; Dunbrack and Cohen, 1997), preference for a specific amino acid given a pair of phi/psi angles (p_aa_pp), and preference for the phi/psi angles in a Ramachandran plot (rama) (Rohl et al., 2004; Wedemeyer and Baker, 2003; Ramachandran et al., 1963).

I.4.2 ROSETTA Energy Minimization

When new sequences are threaded, or rebuilt onto target protein structures, it is often necessary to go through a round of energetic minimization. The protein undergoes all-atom refinement using the ROSETTA all-atom scoring function to yield an a protein model (Bradley et al., 2005a). Both threading and docking in ROSETTA involve an all-atom refinement of the protein. The protocol used for structural refinement, visually described in figure I.9, is often referred to as “relax”. The goal of the relax protocol is to explore the local conformational space and to energetically minimize the protein. During this process, local interactions are improved by iterative side-chain repacking, in which new side chain conformations, or “rotamers”, are selected from the Dunbrack library (Dunbrack and Karplus, 1993); and by gradient-based minimization of the entire model, in which the energy of the model is minimized as a function of the score. These small structural changes are evaluated according to the all-atom scoring function and are sampled in a Metropolis Monte Carlo method (Metropolis et al., 1953). The “relax” protocol has been shown to markedly lower the overall energy of the ROSETTA model and is essential to achieving atomic detail accuracy (Das and Baker, 2008; Bradley et al., 2005b; Rohl, 2005).

345678567845678745687 89-=969

I.4.3 ROSETTA Design

Protein design seeks to determine an amino acid sequence that folds into a given protein structure or performs a given function. The ROSETTADESIGN algorithm is an iterative process that energetically optimizes both the structure and sequence of a protein (Kuhlman et al., 2003). ROSETTADESIGN alternates between rounds of fixed backbone sequence optimization and flexible backbone energy minimization. During the sequence optimization step, a Monte Carlo simulated annealing search is used to sample the sequence space. Ev-

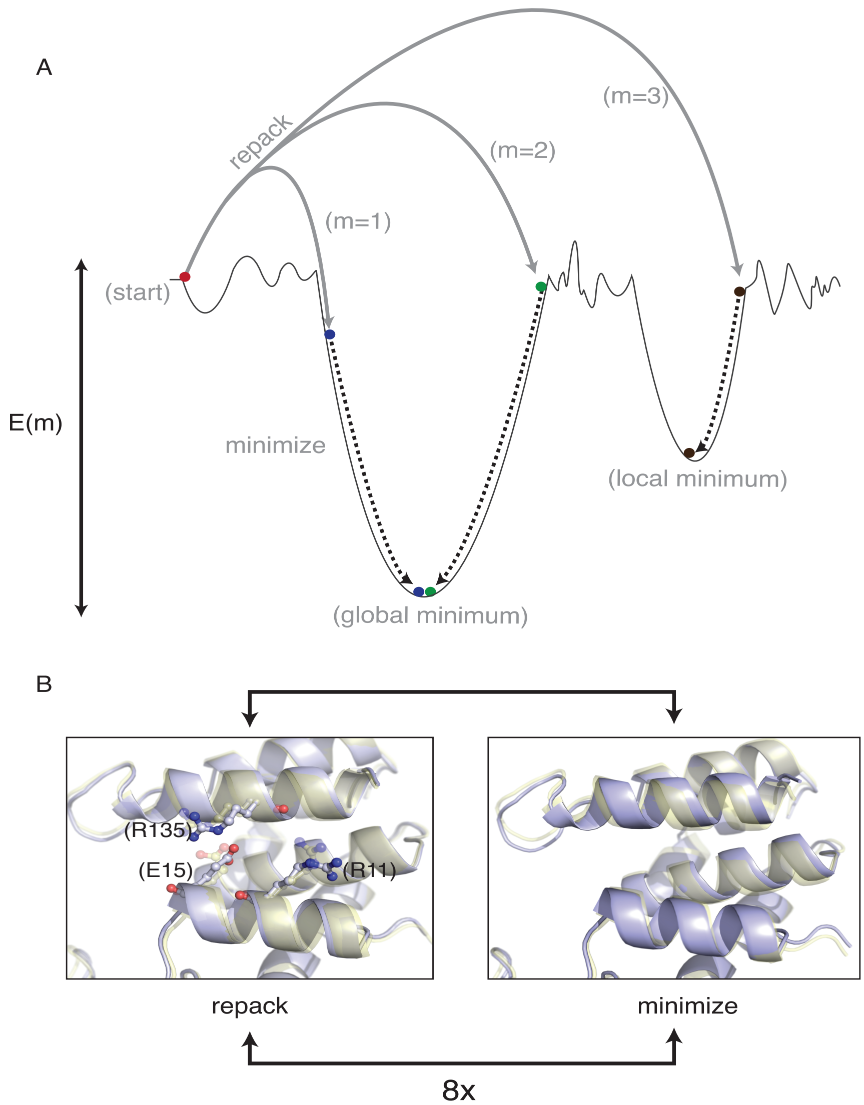


Figure I.9: Refinement via Relax. Simplified energy landscape of a protein structure. The relax protocol combines small backbone perturbations with side-chain repacking. The coupling of Monte Carlo sampling with the Metropolis selection criterion allows for sampling of diverse conformations on the energy landscape. The final step is a gradient-based minimization of all torsion angles to move the model into the closest local energy minimum (A). Comparison of structural perturbations introduced by the repack and minimization steps. During repacking, the backbone of the input models fixed, whereas side-chain conformations from the rotamer library are sampled (Dunbrack and Karplus, 1993). Comparison of the initial (transparent yellow) and final (light blue) models reveals conservation of the R135 rotamer but changes to the R11 and E15 rotamers. Minimization affects all angles and changes the backbone conformation (B).

ery amino acid is considered at each position in the sequence, and rotamers are picked from to the Dunbrack library (Dunbrack and Karplus, 1993). After each round of Monte Carlo sequence optimization, the backbone is relaxed to accommodate the designed amino acids. The practical uses of ROSETTADESIGN can be divided into five basic categories: design of novel folds. Redesign of existing proteins, design to enhance knowledge of structure, enzyme design, and design applied to translational medicine.

I.4.3.1 Design of Novel Folds

The ROSETTADESIGN method was implemented by Kuhlman and colleagues (Kuhlman and Baker, 2000). The method has been used for the *de novo* design of a fold that was not (yet) represented in the PDB (Kuhlman et al., 2003). A starting backbone model consisting of a five-strand β-sheet and two packed helices was constructed with the ROSETTA *de novo* protocol using distance constraints derived from a two-dimensional sketch. The sequence was designed iteratively with five simulation trials of 15 cycles each. The final sequence was expressed, and the structure was determined using X-ray crystallography. The experimental structure has an all-atom deviation to the predicted structure of <1.1Å.

I.4.3.2 Redesign of Existing Proteins

When nine globular proteins were stripped of all side chains and then redesigned using ROSETTADESIGN, the average sequence recovery was 35% for all residues (Dantas et al., 2003). In four of nine cases, the protein stability improved as measured by chemical denaturation. The structure of a redesigned human protein was determined experimentally. ROSETTADESIGN was then used to systematically identify mutations of carboxypeptidase that would improve the stability of the protein. All of the tested mutants were more stable than the wild-type protein, with the top-scoring mutant having a reduction of free energy of 5.2 kcal/mol.

I.4.3.3 Design to Enhance Knowledge of Structure

Protein design approaches have enhanced our knowledge of how protein sequence relates to protein structure. For instance, the finding that designed protein sequences are highly similar to the native sequence suggests that native protein sequences are optimal for their structure (Kuhlman and Baker, 2000). Babor and Kortemme investigated the antibody sequence-structure relationship using ROSETTADESIGN. They demonstrated that native sequences of antibody HCDR3 loops are optimal for conformational flexibility (Babor and Kortemme, 2009). The authors collected pairs of unbound and antigen-bound antibody structures. They used multiconstraint design to find low-scoring sequences that were consistent with both unbound and bound structures. The sequences predicted by multi-constraint design were more similar to the germline sequences than the sequences predicted to preferentially bind either the unbound or bound conformations.

I.4.3.4 Enzyme Design

The ROSETTAMATCH algorithm starts from the protein backbone and attempts to build toward the specified transition state geometry (Zanghellini et al., 2006). In this method, all possible active site positions are defined for the protein scaffold, and rotamers from the Dunbrack library are placed at each sequence position in the catalytic site. The sequence of the area surrounding the catalytic site is then designed. Recently, the ROSETTAMATCH algorithm was used to design enzymes that catalyze the retro-aldol reaction (Jiang et al., 2008). The degrees of freedom in the transition state, the orientation of the active site side chains, and the conformations of the active site side chains were simultaneously optimized. Of 72 models tested, a total of 32 were found to have catalytic activity as much as four orders of magnitude greater than that of an uncatalyzed reaction. Two of the active enzymes were crystallized. The experimental structures shared a high degree of similarity with the computational design although the loop regions surrounding the catalytic site showed significant differences from the model.

Computationally designed functional Kemp elimination catalysts using ROSETTAMATCH have also been designed. Quantum chemical predictions were used to generate an idealized transition state model, and ROSETTAMATCH was used to search for backbone configurations that would support the predicted transition state (Röthlisberger et al., 2008).

I.4.3.5 Design Applied to Translational Medicine

The successes of the ROSETTADESIGN algorithm in predicting new sequences that optimize binding and answer questions about protein structure led to its application to more biomedical applications such as vaccine design and protein therapeutics. Fleishman *et al.* used the paratope of an antibody to find hot-spot positions that neutralized influenza. Using these positions, they designed a protein that would properly present a mimic of the paratope. The crystal structure of the designed protein structure indicated that it did indeed present a functional paratope while functional studies confirmed its neutralization capacity (Fleishman et al., 2011b).

The works of the Schief group have expanded design to explore novel scaffolding approaches to be used as immunogens. Using ROSETTADESIGN, they presented the epitope to broadly neutralizing antibodies 2F5 and 4E10 to HIV, which elicited this class of antibody in animal models (Correia et al., 2010; Ofek et al., 2010). In addition, the research group used ROSETTADESIGN to target potently neutralizing antibodies against the CD4 binding site while eliminating binding to non-neutralizing antibodies that bind to decoy epitopes (Wu et al., 2010a). More recently, this group has used design to mimic an epitope to respitory syncitial virus (RSV) that is now being tested in animal models (Correia et al., 2014).

A current major challenge in protein design is the *de novo* design of a novel protein-protein interface. So far, the most successful attempts at *de novo* interface design have been relatively modest, focusing on small proteins and yielding micromolar affinity (Mandell et al., 2009; Huang et al., 2007). This small boost in affinity often requires display technology to increase potency and specificity. The ROSETTA community is well aware of these limitations and work on increasing the accuracy of predicted interface mutations, particularly around hydrogen bonding networks and explicit solvent models (Combs and Meiler, 2012).

I.5 The State of the Field

Here I can describe the state of the field in molecular modeling and antibody design. I will describe molecular modeling as the "folding-problem", the solution of which, is considered to be the holy-grail of computational modeling. In a related field which is more applicable to my work, I describe the "inverse-folding problem."

I.5.1 The Folding Problem

As described, there have been tremendous advances in computational modeling and it's role as a promising surrogate for costly experimentation. However, most work in computational modeling serves as a proof-of-principle using known model systems as benchmarks. Most, if not all computational studies at the time of writing require an experimental validation of the model using the very experimentation for which they are trying to serve as a prediction. However, experimental structures are only currently available for less than 1/1000$^{\text{th}}$ of the proteins for which sequences are known and that number is expected to increase with advances in high-throughput sequencing (Moult et al., 2014). To that end, computer-aided structure prediction is expected to play a major role in helping solve the ever growing number of structures. But, where is the state of the field?

Computational structure prediction and computational docking have been continually evaluated in the Critical Assessment of Structure Prediction (CASP) and Critical Assessment of Predicted Interactions (CAPRI) community-wide experiments. These experiments have been invaluable to the field as they have served as a metric for evaluating the various international groups attempts to tackle the unsolved problems of protein folding and protein docking. This has served as a "grand challenge" in computational biology. A physics based approach coupled with a knowledge of protein folding pathways were expected to be the answer to this question, but as so far, knowledge-based approaches like those of ROSETTA have been the front runners in these assessments. However, physics based methods are starting to bear fruit, especially for refinement stages of molecular modeling, where the global fold is known, but full-atom placement is needed to optimize bond angles, distances and structure free energy.

For small proteins (<100 AA), *ab initio* methods where a template is not used have been successful in predicting the overall topology of the structure. ROSETTA fairs particularly well in this task, as this application was what it was originally developed for (Simons et al., 1999a). For larger proteins, the view is quite bleak, as an improvement judged by CASP has only led to an increase of accurately predicted models by 24% in two decades (Kryshtafovych et al., 2014). This figure is not surprising as the search space increases exponentially as sequence space gets larger, and sampling methods like those of ROSETTA eventually hit a computational limit.

Homology modeling appears to be the more suitable method for structure prediction as it limits the search space to a certain folds. If the target and template sequence are sufficiently aligned, then many methods have shown to be accurate (as evaluated by CASP). The number of folds deposited is slowing down, which makes homology modeling a more viable solution to structure prediction.

Often, a handful of mutations are made to a known structure and it is the desired goal to model those mutations onto the existing structure and explain the effect at a molecular level. This is becoming easier to do as refinement stages of modeling progress. A few mutations often does not change the overall fold of protein molecules and their effects can be modeled accurately. This type of minimal mutational modeling is now a credible form of empirical science as it is appearing in more and more high-impact publications were a full crystal structure or solution NMR structure is not needed. I use this method here to explain observations to a traditional antibody known as PG9. This is detailed in chapter

IV.

I.5.2 The Inverse-Folding Problem

The folding problem, which tries to accurately determine a protein structure given an amino acid sequence must consider many poorly understood principles of folding kinetics and thermodynamics. The problem is trying to find an accurate globular fold in which the free energy of the folded structure is considerably more stable than the entropy cost of the fold. However, this problem has created many all-atom "force fields" that are needed to evaluate such folds. The all-atom nature of these modeling trajectories created a new class of problems in the field that could be solved with these "force fields." Rather than search for the lowest-energy structure for a given amino acid sequence, researchers are searching for the lowest-energy sequence for a given protein structure (Baker, 2014). This is known as the "inverse-folding problem" and less formally known as protein design.

The problems that are relevant to the inverse-folding problem are related to the folding-problem. They often both use the same scoring function or "force-field", so a rapid and accurate knowledge- or physics-based approach has been the focus of many research groups. It is also a matter of exponential conformations that need to be sampled. In the protein-folding problem, what is the conformations that fit a given sequence. If each the length of the amino acid sequence, A, where each amino acid can sample N conformations, then A^N conformations are available to a complete sampling strategy. As cited in "Levinthal's Paradox", this is clearly not how proteins fold, and complete computational sampling of large proteins can't be completed with current technology (Levinthal, 1969).

The sampling space needed for the inverse-protein is also similar. Instead of conformations, amino acid identities need to be sampled and scored. The same exponential sequence space formula is also applied, were a protein of length A, has 20^A sequence combinations to choose from the 20 canonical amino acids in mammalian system. The problem is compounded when several conformations or "rotamers" for each amino acid are considered,

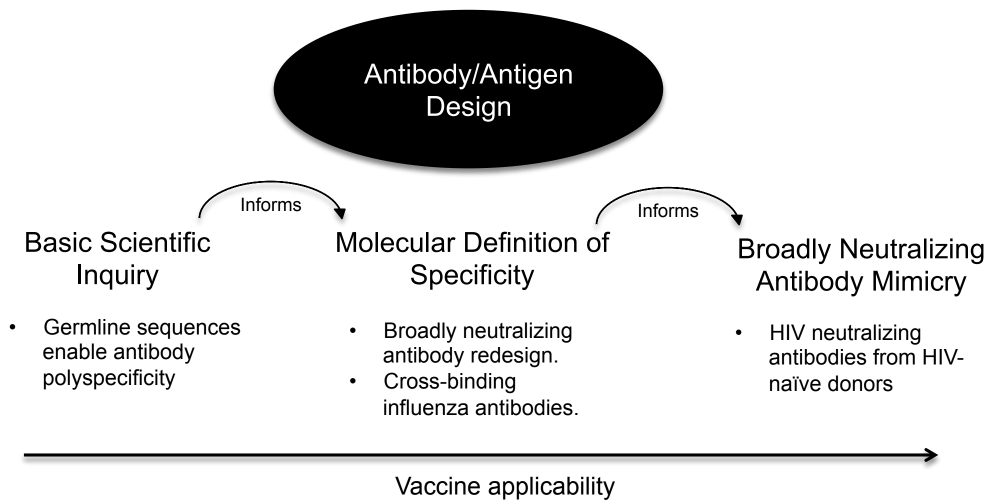


Figure I.10: Questions answered through antibody design. Vaccine applicability spans from basic scientific questions using antibody design about specificity. These principles can be used to enhance specificity. Finally, we can query the immunoglobulin repertoire using antibody design.

where 20^A, no longer constant, but defined by the amount of "rotamers" considered R as R^N. Regardless, protein-design has been far more successful at application driven research than the *ab initio* methods. This is due to the reduced sequence space that needs to be sampled. For instance, we can define far fewer positions to be "designed" rather than the entire protein. We can also only consider certain amino acid identities (for instance, consider only hydrophobics in a core and hydrophilics in solvent-space). The success of protein-design is detailed in application driven research in section I.4.3.

I.5.3 Antibody Design Summary

The goal of this thesis exists in three main sections ranging from questions of basic science to application driven antibody design. However, it is not entirely clear how these concepts tie together. In figure I.10, I show how antibody design can be used to scale up to vaccine applicability from answering basic immunologic questions. As the applicability progresses, one step informs the other. First, I will use antibody design to interrogate a very basic question about antibody specificity. That is, how does a finite set of antibody structures bind a near infinite antigen repertoire? Using principles derived from specificity, I can begin to redesign known antibodies to enhance specificity, and increase breadth. Finally, I can use antibody design and how it relates to specificity and breadth to interrogate the antibody structural repertoire. This answers questions about vaccine applicability. What does a normal structural repertoire look like? How close is it to forming a broadly neutralizing antibody that confers protection against HIV? If it is not close, how many mutations are necessary to achieving broad neutralization? All of these questions I aim to answer using antibody design with the ROSETTADESIGN application and the model system HIV.

CHAPTER II

Mechanisms of Polyspecificity

II.1 Introduction

Human antibodies are critical for eradication of viral and bacterial infections, while providing the basis for immunological memory. Antibody protein molecules are encoded by several recombined germline gene segments prior to antigen exposure. The initial set of antibodies that are generated by recombination in the bone marrow is the antigen-naïve antibody repertoire. It is of great interest to know how a finite set of such germline gene-encoded antibodies can recognize the large number of possible foreign antigens. A current hypothesis in the field suggests that antibodies encoded by germline gene segments are structurally flexible and therefore able to accommodate binding to many antigens, much like one glove fitting the shape of many hands. The phenomenon of one structure binding to many unrelated targets is known as polyspecificity. In this chapter, I will describe how I further support this hypothesis using computational design by showing the entire antibody protein variable region sequence is close to ideal for polyspecificity by mechanisms of flexibility. I will detail the computational protocol I have developed and the results that suggest how a finite set of antibody germline gene segments can encode antibodies that can engage a large number of potential antigens. Computational design of antibodies capable of binding multiple antigens may allow the rational design of antibodies that retain polyspecificity for diverse epitope binding, which will be an important to future vaccine design.

II.1.1 Three Models of Protein Binding

Antibodies are encoded by the rearrangement of variable (V), diversity (D), and joining (J) gene segments into recombined genes that encode a large but ultimately finite number of unmutated antibody structures, known as the germline repertoire (Tonegawa, 1983). There are approximately 10^4 combinations of the V, D, and J heavy chain gene segments and an estimated 10^{11} possible combinations when junctional diversity is considered (Patten et al., 1996). This number of potential antibodies is far less than the immeasurable number of epitopes on foreign antigens to which one could be exposed. The germline gene repertoire therefore encodes a finite number of starting structures in the germline repertoire that must be capable of recognizing and binding a large and diverse array of antigens (Patten et al., 1996; Schultz et al., 2002; Collins et al., 2003).

The classical protein binding mechanism was the "lock-and-key" model, where antibodies acquired somatic mutations in order to rigidify a pre-bound structure that would complement the shape of the epitope (figure II.1A). This mechanism dominated the field for many years but has to assume that one antibody optimally binds to one particular antigen (Notkins, 2004; James and Tawfik, 2003). The lock-and-key model has many shortcomings, as the "one paratope-one epitope" principle leaves little room to describe the polyspecificity phenomenon, an antibody's ability to recognize multiple unrelated antigens.

Polyspecificity has been demonstrated in a variety of biochemical and structural studies, therefore the "lock-and-key" model cannot possibly describe all antibody-antigen interactions without the existence of multiple paratopes per antibody (Schultz et al., 2002; Yin et al., 2003; James et al., 2003; Foote and Milstein, 1994). In contrast to the "lock-and-key" model, a degree of pre-bound structural flexibility are found in two models of antigen binding, the "induced-fit" and the "conformational flexibility" models. In these models, germline gene-coded antibodies retain a degree of structural plasticity in their backbone in order to bind a number of different unrelated antigens. The induced-fit model hypothesizes that upon binding conformational changes are induced to accommodate the interacting structure (figure II.1B) (Notkins, 2004; James et al., 2003).

The conformational flexibility hypothesis in protein binding suggests that an unbound protein assume a variety of conformations (conformational isomerism), a subset of which is recognized by the interacting partner (figure II.1C). For antibodies, a large body of work has attributed polyspecificity to the nature of their germline gene sequences. It has been

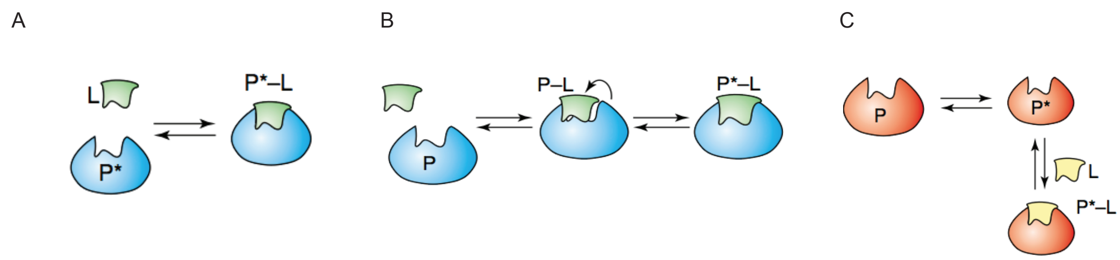


Figure II.1: Three models of protein binding. The "lock-and-key" model assumes the protein binding site in a pre-bound state is optimized for the shape of the ligand (A). The "induced-fit" mechanism allows for conformational change after the ligand had bound to optimize shape in a two-step isomerization (B). For the "conformational flexibility" model, the pre-bound structure exists in several isomers which the ligand selects the conformation that complements it's structure (C). Figure adapted from (James and Tawfik, 2003)

reported that polyspecific antibodies often retain a larger proportion of germline gene sequences than more mature, specific antibodies (Notkins, 2004; Chen et al., 1991; Crouzier et al., 1995; Harindranath et al., 1993).

II.1.2 Evidence for Conformational Flexibility

Conformational flexibility is emerging as an important hypothesis to explain both polyspecificity and changes in affinity between germline and mature antibody sequences (Schultz et al., 2002; Notkins, 2004; James and Tawfik, 2003; Yin et al., 2003; James et al., 2003; Foote and Milstein, 1994; Romesberg et al., 1998; Manivel et al., 2000; Yin et al., 2001; Nair et al., 2002; Jimenez et al., 2003; Li et al., 2003; Mohan et al., 2009; Marlow et al., 2010; Wong et al., 2011; Davies and Cohen, 1996; Mohan et al., 2003; Wedemayer et al., 1997; Zimmermann et al., 2010). The first evidence for conformational isomerism in antibodies was observed through kinetic experiments in which antibodies show a triphasic distribution that, in some cases, appears to reflect the existence of multiple isomers of the unbound antibody in solution, in the pre-equilibrium state (James et al., 2003). To determine the dynamics of the binding process, James and colleagues examined the pre-steady-state kinetics of complex formation between SPE7 and DNP-Ser, as well as between SPE7 and the hapten cross-reactants. The kinetics confirmed the existence of an equilibrium in solution between two preexisting isomers, only one of which bound the haptens. Pre-steady-state binding kinetics were analyzed by monitoring changes in SPE7's intrinsic fluorescence upon rapid mixing with ligand.

In 1997, Wedemayer and colleagues found a structural basis for conformational flexibility observed for germline antibodies (Wedemayer et al., 1997). They solved the crystal structures for a germline antibody with and without its target hapten, and mature antibody with 6 somatic mutations that bound the target hapten 30,000 times stronger than the germline counterpart. They noticed that the rigid-body deviation in the crystal structure was significant between the bound and unbound germline antibody structure indicating a degree of flexibility. In contrast, the mature antibody had less structural deviation upon hapten binding. They showed that the somatic mutations observed in the mature antibody stabilize the binding sites either directly or indirectly by locking the structure into the pre-bound conformation. This was the first indirect evidence showing that germline encoded antibodies may be more flexible than the mature sequences due to the intrinsic properties of the sequence.

More recently, structural studies along with computational tools have corroborated these findings by showing direct evidence that antibodies encoded by germline gene sequences retain flexibility in their HCDR3 loops (Wong et al., 2011; Babor and Kortemme, 2009). For example, Babor et al. redesigned germline or mature HCDR3 loops in antibodies that had been crystallized in free or antigen-bound states (Babor and Kortemme, 2009). These investigators found that germline gene-encoded HCDR3 sequences are nearly optimal for conformational flexibility. The study, while exceptional in its concept, was limited as the dataset contained many antibody/hapten (non-protein) complexes, which may not reflect the biology of interactions with larger protein targets that are more typical in foreign pathogens. Some antibodies classified as "germline" in the study were not from antigen-naïve cells. Further, that study exclusively analyzed the HCDR3 loop, not the entire variable region.

Schmidt *et al.* used molecular dynamics simulations and structural analysis to determine how mutations in the antibody variable domain enhance antigen binding to the influenza virus HA protein (Schmidt et al., 2013). In the study, they found two broadly neutralizing antibodies that have branched in lineage from a common intermediate, and an unmutated common ancestor (UCA) in which they obtained high-resolution crystal structures. They found that even though the UCA and mature antibodies have nearly identical binding configurations, the affinity for influenza for the mature antibodies was 40-fold greater than the UCA. Molecular dynamics simulations predicted that the paratope in unbound UCA was not in an optimal conformation for binding, while the mature antibodies had a higher probability of being pre-configured for the influenza HA epitope.

II.1.3 Experimental Rationale

The V_H-gene encodes the HCDR1, HCDR2, much of the immunoglobulin framework regions and the beginning of the HCDR3 loop. I hypothesized that the conformational flexibility mediating the polyspecificity of germline gene-encoded antibodies is determined at least in part by the heavy chain variable region encoded by the V_H-gene, considering it makes up a large portion of the structure. The focus of my study was to test this hypothesis using computational design. Specifically, I analyzed the somatic mutations in sets of mature antibodies that derived from the same V_H gene and for which co-crystal structures with biologically relevant target proteins were available. Sets of mature antibody-antigen complexes incorporating antibodies that derived from a common germline V_H-gene were input into the ROSETTA "multi-state" design algorithm that recovers the optimal single sequence for an antibody to bind all antigens simultaneously (Babor and Kortemme, 2009; Humphris and Kortemme, 2007; Leaver-Fay et al., 2011a). The sequences recovered using this protocol would be considered inherently flexible and polyspecific, since they are predicted to accommodate binding to diverse antigens using a structurally diverse set of antibody conformational states. In contrast, I also tested monospecificity for each antibody by measuring which sequences are preferred during the design towards a single antigen. This is known as the "single-state" design protocol. For any change between the preference for sequence between the multi-state design protocol that considered polyspecificity and the single-state design that considers monospecificity, recapitulates *in silico*, the process of affinity maturation.

Fundamentally, our approach compares germline and mature antibody sequences optimized in nature through evolution and maturation with sequences predicted to be optimal based on ROSETTA's energy function applied to a set of co-crystallized antibody/antigen complexes. The power of the present approach is that I predicted germline and mature sequences *in silico* without any prior knowledge of either, which is an important step towards rational antibody design. I would expect that results of this type of analysis will continue to improve as the size of the collection of conformational ensembles available in the Protein Data Bank (PDB) increases and as the accuracy of the ROSETTA energy function continues to improve.

II.2 Multi- and Single-State Design of Antigen-Antibody Complexes

I compiled panels of antigen-antibody complexes from the Protein Data Bank (PDB) in which the antibody heavy chain variable region were encoded by germline V_H-genes, designated V_H3-23, V_H1-69, or V_H5-51 (Wu et al., 2010b; Tian et al., 2008). Antigen-antibody complexes were selected only if they contained Homo sapiens or humanized antibodies and the binding partner was a protein antigen. The search of the PDB returns 10, 8 or 3 candidate complexes for V_H1-69, V_H3-23, or V_H5-51 respectively (table II.1).

For each panel I compared the mature (somatically mutated) sequence to the inferred germline gene sequence via a multiple sequence alignment (figure II.2A). The number of mutations with respect to the germline sequence range from 4 to 23 mutations with an average of 12.2. All HCDR1, HCDR2, and framework positions that differed from the germline sequence of the common V_H-gene sequence in at least one position in the multi-

PDB ID	V_H* Germline	Antibody	Ligand	V_H* Mutations
2CMR	1-69*01	D5	gp41	6
3FKU	1-69*01	F10	HA	13
3GBM	1-69*01	CR6261	HA	15
3MA9	1-69*01	8066	gp41	4
3MAC	1-69*01	8062	gp120	7
3P30	1-69*01	1281	gp41	20
1G9M	1-69*02	17b	gp120	21
2DD8	1-69*05	M396	SARS-RBD	5
2XRA	1-69*05	HK20	gp41	14
2XTJ	1-69*10	1D05	PCSK9	4
2QQN	3-23*01	anti-Nrps-1	Nrps-1	10
2R56	3-23*01	IgE	BLG	23
2VYR	3-23*01	VH9	MDM4	10
3KR3	3-23*01	DX-2647	IGF-II	8
1S78	3-23*04	Pertiuzimab	ErbB2	22
2FJG	3-23*04	G6	VEGF	15
3DVN	3-23*04	Apu2.16	Ubiquitin	18
3BN9	3-23*04	E2	MT-SP1	5
2B1A	5-51*01	2219	UG1033	17
2XWT	5-51*01	K1-70	TSHR	8
3HMX	5-51*01	Ustekinumab	IL-12	12

Table II.1: Antibody-antigen test set. Details of the 10, 8, and 3 complexes for V_H1-69, V_H3-23, and V_H 5-51 respectively. The antibodies bind a diverse set of antigens but each share a common germline across a test set. The V_H mutation count of amino acid mutations away from their inferred germline gene. *Predicted from IMGT

ple sequence alignment were included in the computational design simulations as “variable positions”. Note that my study explicitly excluded positions that remained unchanged as no claims can be made with respect to the relevance of these positions for conformational flexibility or polyspecificity. My analysis is limited to antibody regions encoded by the V_H-gene as only this region can be unambiguously aligned within each set of antibodies. Therefore, I excluded D-gene and J-gene that encode HCDRH3, and antibody light chain positions. The identity and conformation in all variable positions to identify the sequence and conformation that return minimal energy for the given protein backbone of the antibody/antigen complex (Kuhlman and Baker, 2000). In this work, I used multi-state design

(Leaver-Fay et al., 2011a) to find a single sequence that minimized energy with all antigens within each V_H gene-encoded group. To reduce noise in the outcome of the computations, 100 simulations were executed, and results are displayed using WebLogo representation (Crooks et al., 2004) (figure II.2 C).

For a concrete example, consider position 31 (PDB numbering, boxed in II.2 C). This position, encoded by V_H5-51, diverged from a germline serine residue in the sequence for all three complexes. Complexes 2B1A and 2XWT (PDB code) possess an aspartate residue in this position acquired by somatic mutation, while 3HMX has a threonine in the same position. The multi-state design protocol selected the germline residue serine as the energetically most favorable residue out of all 20 possible genetically encoded amino acids when interaction with all three structurally diverse antigens is required. The experiment was repeated as three separate "single-state design" experiments (figure II.2B, right side) to predict the sequences and conformations that minimized interaction energy for each antigen individually. The resulting sequences were compared to both the inferred germline and the mature sequence (figure II.2C). In this experiment position 31 is predicted as an aspartate for complexes 2B1A and 2XWT, and as a threonine for 3HMX, the mature amino acid sequence (data not shown).

For this work, it was important to convert the outcome to a statistical quantitative analysis. Each design outcome is compared to the mature or germline sequence, by computing a bit-score "recovery" measure. The results can either recover towards germline, mature or neither sequences. The bit-score computation I used is described in the methods of the appendix section. The advantage of the bit-score measure in comparison to a more simplistic percentage-recovery is that it analyzes the relative probabilities of all twenty amino acids in a particular sequence position, not just the probability of the correct one. It thereby arrives at an accurate measure of "surprise" of seeing a certain outcome, a normalized measure in information theory that can be readily compared between different experiments. In our experiment high bit-score for the germline sequence indicated that among the 100 designed sequences, germline gene-encoded residues were chosen in a large number of instances (figure II.2D). To facilitate comparison across complexes that have a different number of designed entities considered, I determined the sum bit-scores over all designed positions and normalized the score to fall between 0 and 1 by division with the maximum bit-score that could be achieved, *i.e.*, every amino acid position designed towards a germline or mature sequence.

II.3 Specificity Inferred by Sequence Design

The results of the multi-state design simulations returned sequences that resembled germline gene-encoded sequences more often than mature sequences. This finding was remarkable as no information about germline sequences was input into the simulation. I found that the designed sequences gave normalized bit-scores of 0.54, 0.60, and 0.43 for germline genes V_H1-69, V_H3-23, or V_H5-51, respectively. In contrast, statistically significant reduced bit-scores of 0.48, 0.45, or 0.26 ($p < 0.0001$) were observed when comparing the designed sequences with the mature genes (figure II.3A). The single-state redesign of mature antibodies for binding to their associated antigen gave normalized bit-scores of 0.47, 0.43, or 0.28 for comparison with germline gene-encoded sequences and 0.57, 0.54, or 0.53 for comparison with mature sequences of V_H1-69, V_H3-23 or V_H5-51, respectively. In this design experiment, a proclivity to recover the somatically mutated mature sequences was observed (figure II.3A). Given that a normalized bit-score is the preference for each design experiment to match a certain sequence profile, a high bit-score to germline sequence indicates the output matching the germline profile, while a high bit-score to the mature sequence indicates a preference for the mature profile, each design experiment outcome can be measured as a difference in bit-scores (mature - germline). With this definition, a preference for mature sequence gave a positive Δbit-score, while a preference for germline residues gave a negative Δbit-score for a given complex - *i.e.*, the Δbit-score provided an *in silico* predicted metric for antibody optimization for affinity to a specific antigen versus

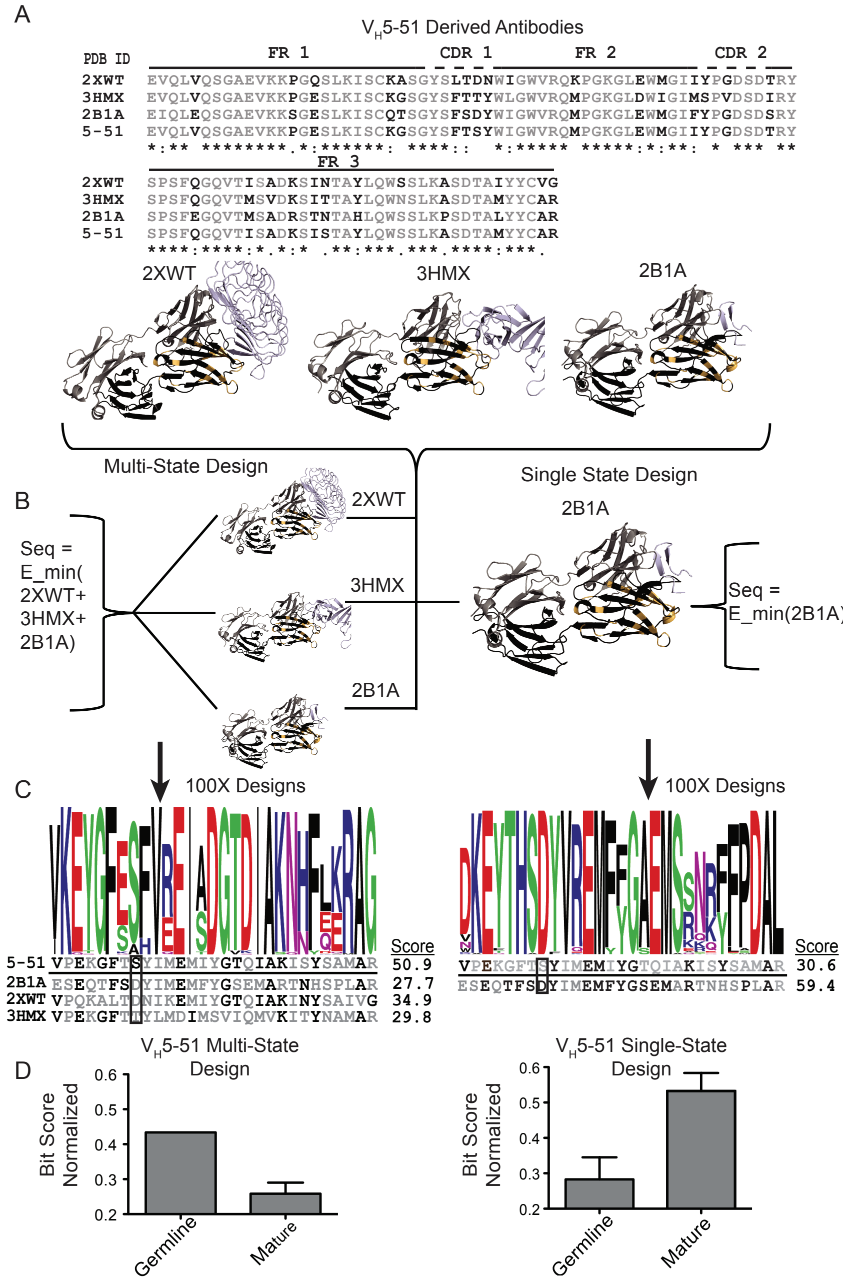


Figure II.2: Multi-state and single-state design methodology. Position candidates were chosen for design if the position differed from the germline sequence in at least one mature complex (A). Co-crystal structures for each complex are shown with designed positions highlighted in gold. Single- and multi-state design schemes are shown (B) Each position in the sequence logo corresponds to a position conserved for design (C). Bit-scores were determined quantitatively by measuring the frequency of a letter at each position (D).

polyspecificty. I observed positive values for single-state design and negative values for multi-state design, indicating a preference for the mature or germline sequences, respectively (figure II.3B, $p < 0.0001$).

II.4 Affinity Maturation Correlates with Predicted Affinity

The number of somatic mutations can be used as a measure of the maturity of an antibody (Briney et al., 2012). Hence, I asked the question if the Δbit-score, the change in proclivity for a germline or mature sequence, correlated with affinity, *i.e.*, if tendency to recover mature versus germline sequences increased as antibody maturation progressed. Such a correlation would indicate that as antibodies mature, features of the germline sequence critical for polyspecificity are replaced with features critical to recognize one target antigen. Figure II.3 C shows the somatic mutation percentage of antibodies in each complex as a metric for "*in vivo* maturation" correlated with the Δbit-score as a metric for "*in silico* predicted optimization for affinity versus polyspecificty". For positive Δbit-scores, the mature sequence was preferred, indicating a preference for specificity. For negative values, the germline sequence, and hence polyspecificity was preferred. The correlation coefficient for the "*in vivo* affinity maturation" and "*in silico* predicted optimization for affinity vs. polyspecificity" was 0.83.

II.5 Backbone Conformational Space for Germline Sequences

Torsional phi-psi angles in the protein backbone were compared across the sets of experimental structures for positions that recovered to germline sequence for multi-state design and those positions that recovered to a non-germline sequence. I found that positions that converted back to germline in multi-state design, *i.e.*, positions critical for conformational flexibility according to the simulation, had a deviation of $19.6° \pm 2.0°$ across beta-sheet phi-psi torsion angles. Sequence positions that did not recover to a germline gene-encoded amino acid had a reduced deviation $15.5° \pm 1.5°$ for beta-sheet backbone torsion angles ($p = 0.099$) (figure II.4 A-C). Considering the limited range for beta-sheet backbone torsion

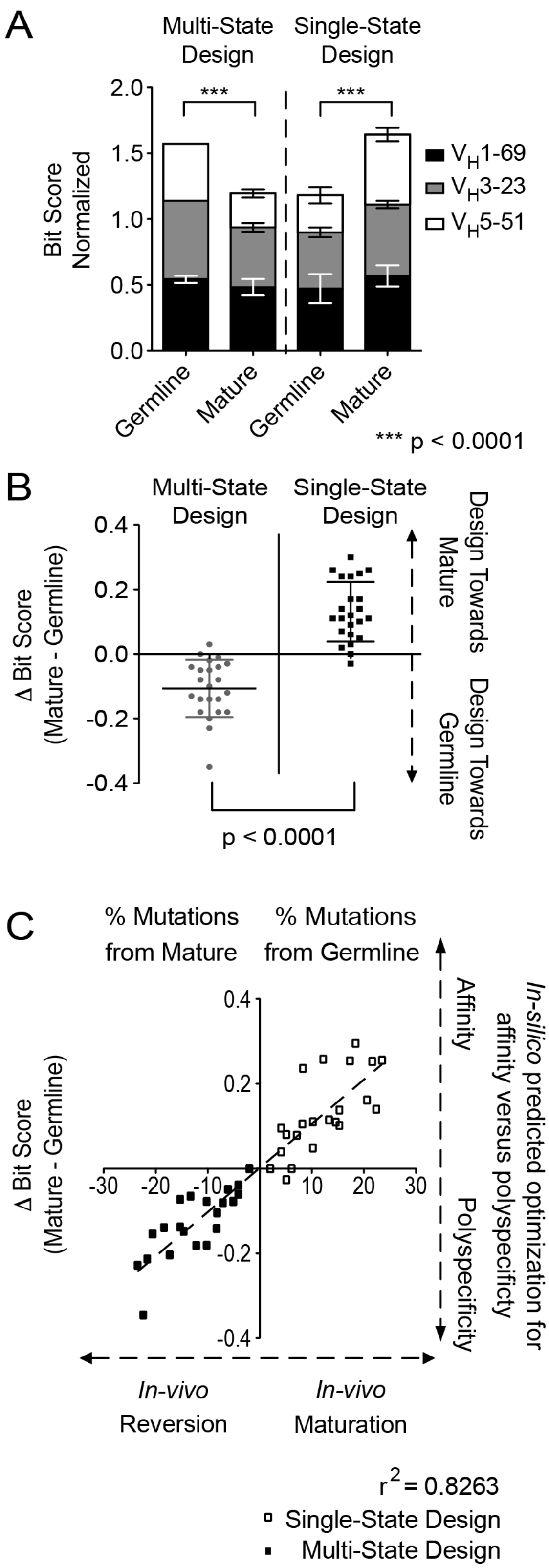


Figure II.3: Multi-state designs toward the germline sequence. Antibodies encoded by the same inferred germline V_H gene preferred germline sequences when considered in the multi-state design, inferring a more flexible combining site. The bar graph shows the bit-score for each of the three different inferred germline groups and then the sum of the scores in a grouped bar. A perfect design would have a normalized bit-score of 1.0, and summated score of 3.0 for three germline groups. Multi-state design preferred germline sequences for all complexes, while in contrast single-state design preferred mature sequences (A, $p<0.0001$). The change in bit-score is determined to be the proclivity to either the mature (positive score) or the germline (negative score) sequence. Each complex was assigned a change in bit-score. The change in proclivity between design protocols was significant (B, $p< 0.0001$). Each complex was scored against mature and germline sequences and a difference was calculated (Δbit-score). Positive numbers returned showed a proclivity towards mature sequences, while a negative score suggested a design toward germline. A tight correlation was observed ($r^2=0.8263$) for the *in silico* predicted optimization for specificity versus polyspecificity (Δbit-score) and the *in vivo* maturation process (C, plotted as the mutation percentage away from V_H gene sequence).

angles, I don't expect large deviations. For reference, all framework residue beta-sheets in antibody-antigen complexes across my dataset have an average phi-psi deviation of 18.7° ± 0.9°.

II.6 Impact of Residue Environment on Specificity

Figure II.5 maps each amino acid position encoded by the V_H gene segment onto the immunoglobulin fold using a custom Collier de Perles representation, as described by Ruiz and Lefranc (Ruiz and Lefranc, 2002). I modified the output to distinguish positions by location in the interface with the antigen and the degree of burial. I correlated these metrics to the bit-score at a per-residue level. Each residue given is in IMGT numbering.

For multi-state design (figure II.5 A-C), 33 out of a possible 46 of the designed interface residues (72%) contributed to polyspecificity, *i.e.*, recovered to germline sequence with a normalized bit-score > 0. Remarkably, also 41 out of 77 residues outside the interface (53%) recovered to germline. Residues 25, 40 and 105, far removed from the interface, recovered perfectly (normalized bit-score = 1) in at least two of the three germline gene test sets. These residues are highly buried, with a neighbor count score of 13.3 ± 0.5. The intermediately packed residues 17, 51, 70, and 71, with an average neighbor score of 8.6 ± 2.2 neighbors, were predicted to contribute to polyspecificity, even though they lie in distal positions from the antigen-binding site. The interface residues 35, 63, 64, and 82 were found to contribute to polyspecificity in two out of the three germline gene test sets. A conserved serine, which was found in all three germline sequences at position 36 in the CDR1, was the only residue identified as critical for polyspecificity in all three germline genes.

In contrast, for single-state design, it is more difficult to deduce overall trends for any specific position as the paratope is altered in each antibody and the recognized epitopes cover diverse structural space. Generally, when each complex was considered individually, 214 designed interface residues recovered to their mature sequence out of a possible

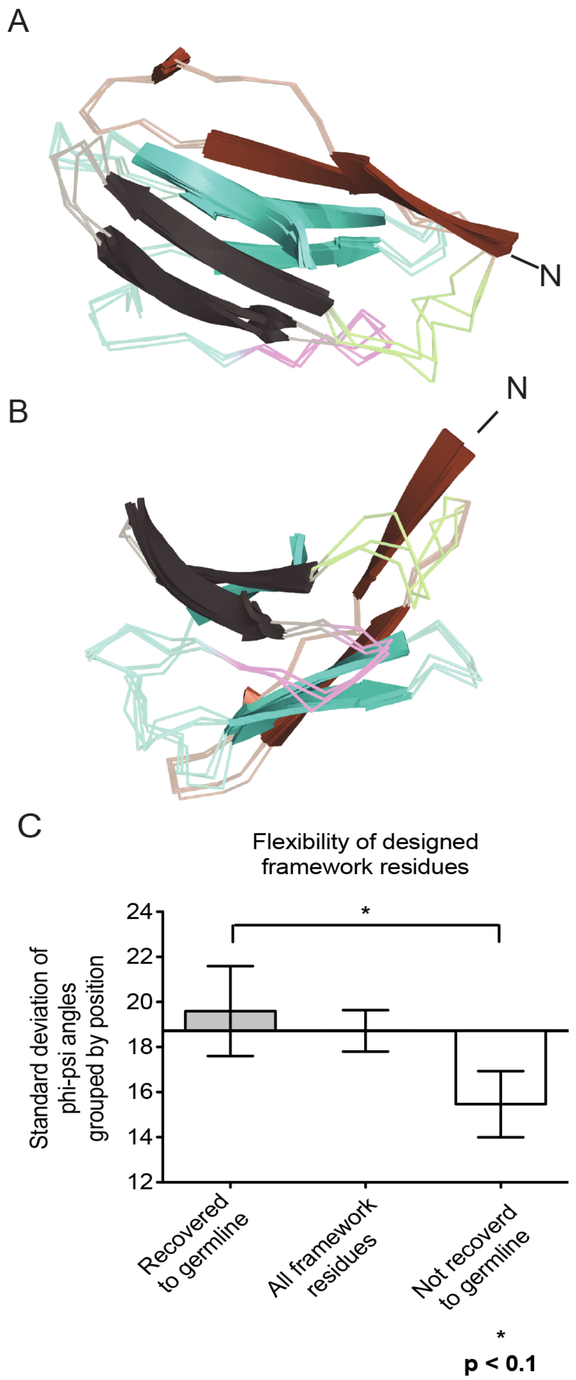


Figure II.4: Phi-psi variances for framework residues. The degree of structural variation of the framework residues were measured as the standard deviation of the phi and psi angles over each residue position. Side view of immunoglobulin fold for V_H5-51 complexes aligned by framework residues. Beta-sheets included in the analysis are shown as a cartoon representation, while loop regions are in a transparent ribbon representation. Framework 1 is shown in brown, HCDR1 in green, framework 2 in black, HCDR 2 in magenta, and framework 3 in cyan (A). Same as (A) but top down view (B). The standard deviations of the phi-psi angles of each framework position were binned into either a residue that was found to be critical for polyspecificity (recovered to germline) or a residue that was not recovered to germline in multi-state design. For each position, the phi-psi angles were averaged, and the standard error of the mean was calculated. An average of 19.6° ± 2.0° for germline recovered residues and 15.47° ± 1.5° for non-germline recovered residues supporting our hypothesis that residues which enable polyspecificity alter beta-sheet packing to a greater degree than residues that do not. The axis is normalized to 18.7° ± 0.9°, the average deviation for all beta-sheet framework positions (C).

340 designed amino acids, indicating their importance for recognition of, and affinity for binding to, the specific antigen (63%, figure II.5 D-F). When non-interface residues were considered, 411 out of a possible 699 designed residues recovered to their mature sequence (59%).

Residues that were found to be critical for polyspecificity, *i.e.*, reverted to germline in multi-state design, differed substantially for each germline gene test set considered. For the V_H1-69 gene derived antibodies, all of the residues in the HCDR2 loop contributed to binding interactions in the single-state but not the multi-state design. In contrast only G63 and T64 residues contributed in the multi-state case but not in single-state designs. Residue L50 was recovered in all single-state complexes but was not critical for multi-state design. For the V_H3-23 gene, residues A55 and Y66 were not recovered in multi-state design but were found to be important for high affinity in single-state design. For the V_H5-51 complexes, non-interface residues P15, M53 and A80 were not recovered in multi-state design but were found to be critical in single-state design. HCDR2 was found to be critical in single-state design for all V_H5-51 complexes.

II.7 Mature Sequence Bias

To understand some of the trends described above more quantitatively, I determined for each residue in each antibody/antigen complex if it was part of the interface, *i.e.*, directly engaging the antigen. For this purpose the change in neighbor count between unbound antibody and bound antibody/antigen complex score was measured, and positions with a change larger than 1.0 were classified as “interacting residues”. Next, I counted how often a residue position appeared in the interface within each set of antibody/antigen complexes. Positions were binned as occurring in the ensemble interface never, once, two-four times, or more than four times and average bit-scores were compared (figure II.6). I found a general trend for interface ensemble size correlating with interface ensembles sampled. For the set of structures derived from V_H3-23, which contained a total of 8 complexes, I found that

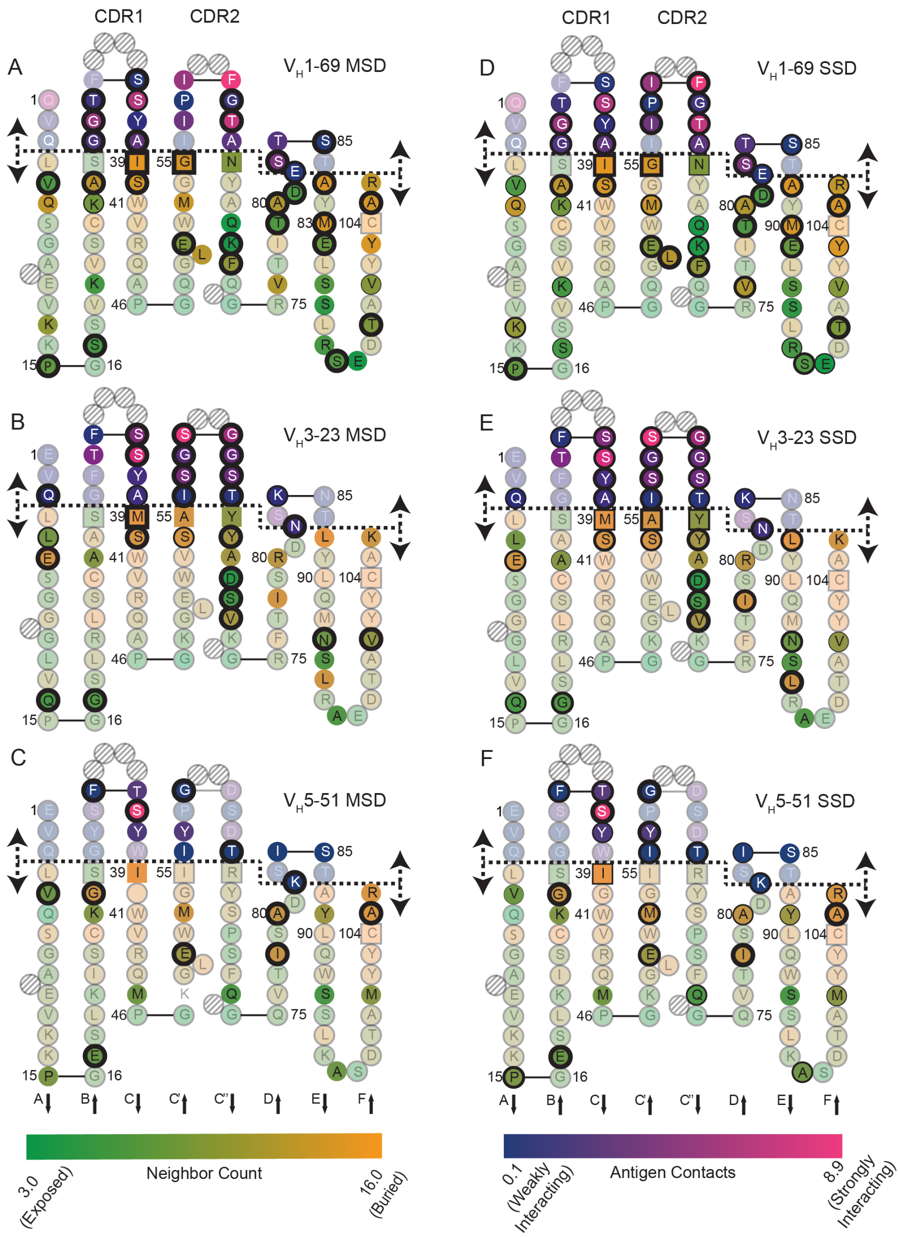


Figure II.5: Colliers de Perles reepresentation of V_H gene segments. The 98 amino acids present in V_H1-69, V_H3-23, or V_H5-51 are shown in a Collier de Perles two-dimensional representation and numbered according to the IMGT numbering scheme 37. Hatched circles are missing residues according to the IMGT numbering scheme and are shown to make graphs consistent. Square boxes represent the boundary between framework and CDR loops. A dashed line is shown for the interface. Interface residues are colored with a blue-pink gradient indicating a numerical antigen contact score defined by a change in neighbors between the free and bound complex. Non-interface residues are colored with a green-orange gradient according to their degree of burial defined through a neighbor count. A, B, C show the germline sequence represented in the immunoglobulin fold with the thickness of each line representing the design bit-score for that position relative to the germline sequence for multi-state design. D, E, F the thickness of the line corresponds to the mature sequence bit-score averaged over each complex.

residue positions that are never found in the interface gave an average bit-score of 2.3 ± 0.4. If a residue position was found only in one interface, the average bit-score dropped to 1.2 ± 1.1. As residues were found more frequently at the interface between 2-4 complexes, and 5-8 complexes, the average bit-score increased to 2.5 ± 0.8 and 3.6 ± 0.7 respectively. For the 10 V_H1-69 complexes, an average bit-score of 2.3 ± 0.3 was observed for residues that were never found in the interface. If a residue was only found in the interface once, the average bit-score dropped to 1.9 ± 1.0. For interface occurrences between 2-4 and 5-8, I found the average bit-score to increase to 2.6 ± 0.7 and drop to 0.8 ± 0.4 respectively. Due to the limited number of residues occurring in multiple interfaces, a significant change in bit-score between each grouping was not observed for V_H1-69 (p= 0.1844) and V_H3-23 residue positions (p=0.2007).

II.8 Evolutionary Sequence Bias

I expected the result of multi-state design to deviate from germline in cases where alternate amino acids are compatible with the conformational space and binding modes observed in the ensemble of structures. Alternative amino acids might be tolerated but are not observed in evolution - "evolutionary sequence bias". To test this hypothesis, I reverted each position back to germline and compared the energetic change with the favored residue returned by multi-state design. Using reference energies, ROSETTA facilitates the direct comparison for energies between different residue types (Kuhlman et al., 2003). For complexes derived from V_H5-51, all positions in which the germline residue were not chosen in at least 10% of the 100 simulated models were forced into the germline identity (figure II.7 A, x-axis). The difference in average energy of the germline sequence at that position from the average energy of the residue returned by multi-state design was calculated (y-axis). For each position, if positive values were returned for all three complexes, ROSETTADESIGN would most likely place a non-germline amino acid at that position. If negative values were returned for all three complexes, ROSETTA would most likely place a germline amino acid at

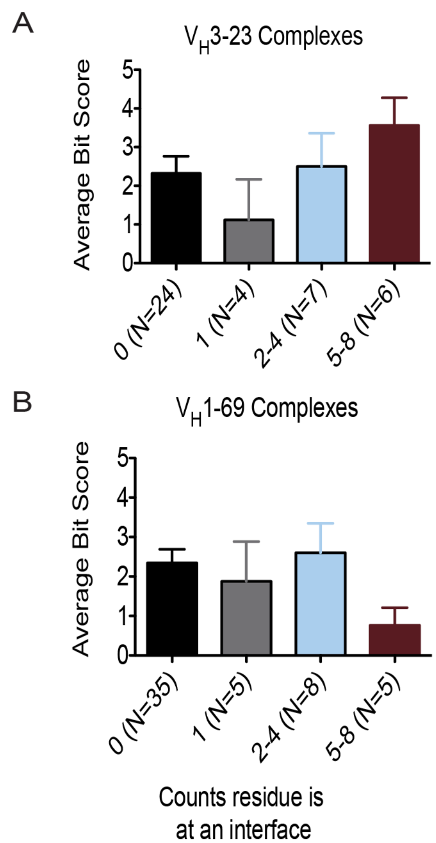


Figure II.6: Interface occurrences affect germline sequence recovery. For V_H3-23 (A) and V_H1-69 complexes (B), I binned each residue position into how many times it occurred in an interface (interface ensembles). Most designed positions never occurred in an interface. As their occurrences became more frequent, I observed a trend for increasing the recovered germline residue. This trend fell off for V_H1-69 complexes (B) for positional occurrences between 5-8 interfaces.

that position. I found that, in most cases, the energetic contribution of the designed amino acid is not significantly more stabilizing than the germline amino acid, *i.e.* the germline sequence is tolerated as well. Only positions 52, 76, 88, and 98 gave a significant energy increase for the germline sequence in at least one complex. Changes in energy were classified as significant if larger than 0.7 ROSETTA energy units (REU, horizontal dashed line). This threshold was derived from the average difference in energy between the germline and mature residue (0.7 $\pm$ 0.2 REU, data not shown). For Figure II.7B, a multiple sequence alignment is given as a reference, where each position that was considered in multi-state de-

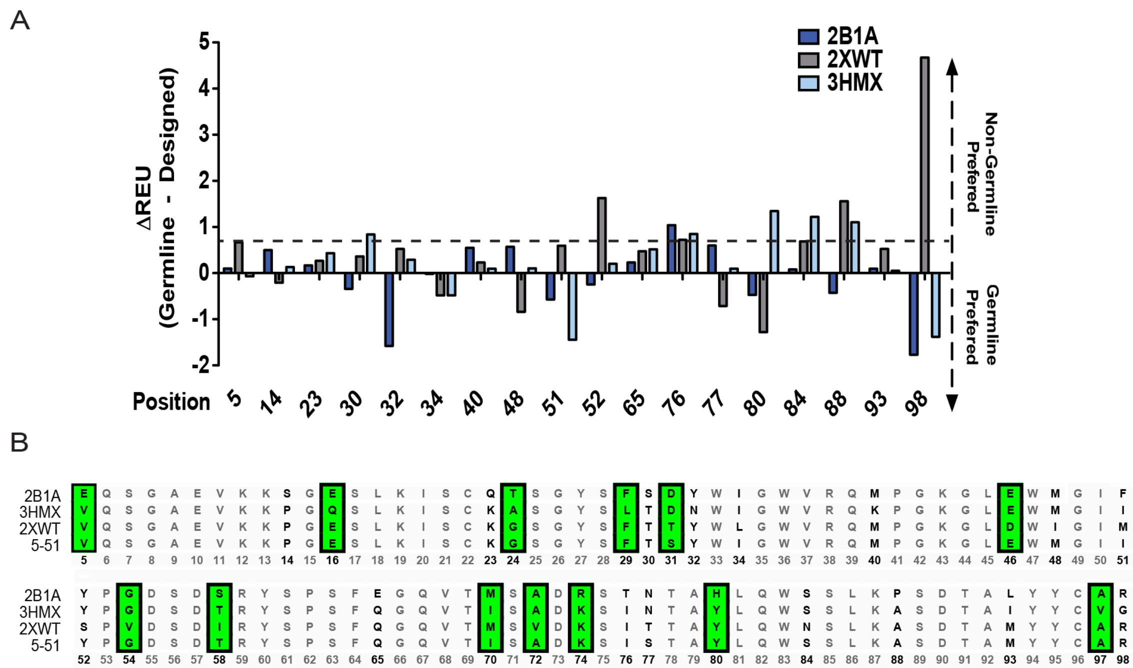


Figure II.7: ROSETTA multi-state design solutions. I evaluated a complete germline reversion of V_H5-51 sequence versus the sequences output by multi-state design. (A) Consideration of positions in which the multi-state design algorithm chose a non-germline amino acid for at least 10% of the models where evaluated. The difference in energy of the germline sequence and the multi-state design solution sequence is shown for each position. Bars above 0 represent the multi-state design sequence preferred while bars below the line represent the germline amino acid preference. The horizontal dashed line at 0.7 ROSETTA energy units (REU) shows the average energy difference between the germline and mature sequence and is represented as a marker for sequence tolerance. (B) The multiple sequence alignment for each V_H5-51 complex is shown and compared with the germline sequence. Sequences highlighted in bold were considered for design. Sequences highlighted in green are positions in which the multi-state design algorithm chose the germline amino acid as the design solution. The numbers in the bottom row are the alignment-numbering scheme of each position and correspond to the position numbers in (A).

sign is highlighted in bold while each position that recovered well to the germline sequence is highlighted in green.

II.9 Discussion

II.9.1 Limitations of Computation

I recognize several important limitations of my study:

1. I assumed that the ROSETTADESIGN protocol determined the optimal sequence for any given design challenge. While it has been demonstrated that ROSETTADESIGN typically recovers close-to-optimal sequences (Kuhlman and Baker, 2000), inaccuracies in the scoring function and limitations in the sampling algorithm will introduce errors. In the future, this limitation could be reduced by improvements applied to the energy function and comparing the results obtained with complementary energy functions. I assumed that the finite and small set of antibody conformations observed in the set of co-crystallized mature antibodies completely describes the conformational flexibility of the germline gene-encoded antibody (finite ensemble bias). While I used the largest ensembles available (10, 8, and 3 antigen-antibody complexes), this assumption must be wrong, introducing a bias. For example, assume there is a sequence position that is part of the paratope in only one of the n complexes. In this antibody, a somatic mutation occurred at this position greatly increasing affinity to the antigen. The somatically mutated amino acid is however compatible with all other n-1 complex structures. In such a scenario the multi-state design algorithm will recover the somatically mutated instead of the germline amino acid. Here, I found that as a residue was more often part of the paratope, it became more likely to be recovered to the germline sequence. This finding might be due to the fact that a critical conformation that the germline antibody needs to adopt was not represented in the ensemble (for framework residues) or the epitope needed for recognition by a critical germline amino acid represented in the ensemble.

2. I assumed that the germline gene-encoded antibodies were able to adopt the conformations of each of the mature antibodies derived from it. This assumption is important, as crystal structures of the "true" germline antibody in complex with the antigen are generally not available. While this assumption is expected to be correct for the majority of cases, notable exceptions are discussed in the literature (Yin et al., 2003; Li et al., 2003; Wedemayer et al., 1997; Sethi et al., 2006).

3. It is not guaranteed that only the germline amino acid is compatible with all conformations adopted by mature antibodies. Rather, it is likely that for some positions alternative amino acids are plausible or even better in realizing the conformational flexibility needed. The germline sequence observed in nature is optimized in evolution and clearly works, but does not need to be perfect in all positions. In such a scenario, multi-state design could return amino acids that deviate from germline (evolutionary sequence bias). Conversely, the mature sequences observed in the co-crystal structures are not guaranteed to be the perfect sequence for high affinity. In some positions a somatic mutation might have introduced a better amino acid but is not the "true" best option. Some somatic mutations might have occurred by chance and do not contribute to affinity maturation. Some positions might not have experienced somatic mutations but still favorable mutations exist. In all these cases I expect the single-state design to deviate from the mature sequence observed in the co-crystal structure (evolutionary sequence bias).

4. The imperfect nature of the ROSETTA scoring function will not yield 100% agreement with natural phenomenon (Kuhlman and Baker, 2000). Importantly, water coordination can often be important in antibody-antigen binding sites. However, ROSETTA is currently being developed to include tools with explicit solvent models (Lemmon et al., 2012).

It is important to understand these biases and limitations to arrive at an accurate interpretation of the results. Given these known limitations, I expected imperfect agreement of *in silico* predicted and natively observed mature and germline gene-encoded antibody sequences. Nevertheless, I found a remarkably high correspondence of residues designed for polyspecificity in a blinded fashion and the amino acids encoded by germline genes.

II.9.2 Interpretation

Germline gene-encoded sequences for commonly used V_H segments are hypothesized to possess high conformational flexibility making them ideal for binding diverse antigens, *i.e.*, being polyspecific. During antibody maturation, somatic mutations are introduced that increase affinity for a specific target in part by adding attractive interaction to the antigen (increasing enthalpic gain) and in part by locking the conformation critical for interaction with the specific antigen (reducing entropic cost). Here I tested this hypothesis by analyzing three sets of antibodies, each derived from a commonly used V_H gene and each co-crystallized with a protein target in its antigen-specific binding conformation.

I chose to not directly compare conformational flexibility for germline and mature antibodies. While this approach may be feasible in general through predicting the accessible conformational space using molecular dynamics (Wong et al., 2011), it is challenging to achieve complete sampling of large conformational spaces that include the entire immunoglobulin framework. To circumvent this problem, I chose to solve the inversely related protein design problem, which was to study amino acid sequences that are consistent with the conformational space seen in antibody/antigen co-crystal structures. This approach is complementary and potentially superior as it replaces sampling of the large conformational space in antibody backbone regions with solving the better understood ranking of amino acid sequences, given a certain antibody/antigen complex conformation. Specifically, I employed multi-state design to find single amino acid sequences that were compatible with the multiple conformations of antigen combining sites.

Computational tools to design multi-specific proteins were first described by pioneering work in the Kortemme laboratory (Babor and Kortemme, 2009; Humphris and Kortemme, 2007). In parallel, Leaver-Fay and colleagues developed a general algorithm for multi-state design in the ROSETTA framework, in which they designed one protein to interact with non-native targets (Leaver-Fay et al., 2011a). I used the latter tool to design antibody sequences that are optimal for facilitating interactions to:

1. Multiple and diverse antigens, or

2. A single specific antigen.

In the absence of *a priori* knowledge of the germline or mature sequences or the mechanism of antibody maturation through somatic mutations, multi-state design of one antibody to recognize several target proteins recovered sequences similar to those encoded by the inferred germline gene segment. When designing the same antibody to recognize one specific target, the sequence recapitulated the mature antibody sequence. This trend correlated tightly with the divergence of the mature sequence from the inferred germline sequence, *i.e.*, the more somatic mutations an antibody contained, the more reversions to germline needed in order to facilitate interactions to multiple antigens.

Use of a computational tool to approach questions regarding polyspecificity as a function of protein sequence is advantageous, as the ROSETTADESIGN algorithm is able to rapidly enumerate the effect of multiple simultaneous mutations in sequence space for the entire heavy chain variable region. This task is quite difficult if not impossible to complete experimentally at this scale. In this manner, conformational flexibility in the framework regions, HCDR1, and HCDR2 can be tested in a holistic model. All mutated positions in the V_H gene segment were considered simultaneously, including the effect of interactions between different domains in the antibody, thus revealing the role of interface and non-interface residues in both poly- and monospecificity. Because this approach considers multiple antibodies of variable conformation at once, each with a distinct binding mode, the multi-state design algorithm predicts sequences that are inherently flexible and capable of adopting the diverse set of conformations needed to bind to multiple antigens.

Harindranath and colleagues demonstrated that polyspecific antibodies were encoded largely by germline gene sequences (Harindranath et al., 1993). Romesberg and Spiller presented structural evidence for flexibility in germline gene-encoded sequences (Romesberg et al., 1998). In addition, Schmidt et al. correlated mature sequence to rigidity of the paratope (Schmidt et al., 2013). Taken together, these data suggest conformational flexibility coupled with pre-sampled conformations of the target binding site as the underlying mechanism for polyspecificity (Wedemayer et al., 1997). Here, I used a multi-state design algorithm to assess the contribution of the V_H gene segment to specifying an antibody with conformational flexibility, preorganization, and polyspecificity. I found that this property is largely attributed to antibody sequences in the germline gene repertoire, since designing antibodies for polyspecificity, sequences recovered germline gene-encoded sequences, while designing antibodies for monospecificity to a single target, returned sequences similar to the mature antibody. This trend increased in strength the higher the number of somatic mutations that had accumulated, *i.e.*, the further optimized the antibody had become. Importantly, the effect is not limited to the HCDR3, which often contributes much to antibody specificity. I obtained the same finding to be clearly measurable throughout HCDRs 1 and 2 as well as the immunoglobulin frameworks. I found each germline V_H gene to encode a set of amino acids that enabled polyspecificity in a distinct manner. These positions were present not only in the paratope, but also in the buried or semi-buried positions of the immunoglobulin frameworks (figure II.5). I expect, that with an increasing number of antibody-antigen complexes in the PDB it will become easier to discern general trends.

I conclude that conformational flexibility in the beta-sheet framework is critical for changing critical regulators of the conformation of the paratope - *i.e.*, the takeoff and landing angles of HCDR loops, thereby enabling the paratope of germline antibodies to assume multiple conformations. Accordingly, I find that residues that contribute the most to polyspecificity contain larger deviations of their phi-psi torsion angles (figure II.4). During antibody maturation, mutations in these positions likely lock in the target-specific framework conformation, reducing the entropic cost of target binding. Somatic mutations in the paratope, for example within HCDR1 and HCDR2, can directly increase affinity to a target (enthalpic contribution to free energy), or lock in a conformation that recognizes the target (entropic contribution to free energy). I found that on average 62% of residues in the paratope and 42% of residues in the framework were important for changing the binding pattern of the antibody from polyspecificity to recognition of one specific target (figure II.5).

I identified at least four specific scenarios in which current datasets are limiting for informing design efforts:

The first scenario involves a framework position that does not interact with the epitope in any of our tested complexes. For this position, the germline residue, and only the germline residue, is capable of adopting the phi-psi angles in order to accommodate the flexibility needed for the binding site. Multi-state design likely designs in the germline residue for each simulation. I then observe agreement between *in silico* design and natively observed sequence for a majority of the designed positions (figure II.3).

The second scenario involves a framework position that also lies distal from the epitope. In this scenario, the germline residue but also other amino acids are compatible with the observed conformations since they both contain properties to adopt the phi-psi angles necessary to accommodate the flexible binding site. For this scenario, I expect ROSETTA's multi-state design algorithm to pick one of the compatible amino acids, not necessarily the germline gene-encoded one. This outcome can occur either because the conformational ensemble is incomplete or because of the evolutionary sequence bias. I find that both biases contribute to ambiguity. Residues that are never found in the interface give modest recovery to germline sequences being either "hit-or-miss" (finite ensemble bias, figure II.6), and residues that are reverted to an amino acid different from that encoded in the germline are not significantly better in energy score than the germline encoded amino acid (evolutionary sequence bias, figure II.7)

The third scenario concerns residues that are at part of the paratope in only one instance. If the mature residue forms critical interactions that minimize the free energy of binding in this one complex, while in all other complexes the residue is not part of the paratope and the mature amino acid seen for the one complex is compatible with the backbone confirmation, ROSETTA will choose the mature residues from the first complex also in multi-state design mode. I observed this trend, especially for V_H3-23 complexes. If a residue was found in only one interface (figure II.6), that position tended to have a low recovery to the germline sequence.

The last scenario deals with positions that are part of the paratope multiple times and that experience frequent somatic mutations. As positions are found to be more frequently in interface ensembles, the germline recovery increases as these positions become more important to facilitating direct interactions with their antigen (figure II.6). These residues contribute to polyspecificity by being the preferred residue in interaction with multiple antigens, rather than facilitating binding by altering beta-sheet packing.

II.10 Conclusions and Future Directions

These results suggest that the naturally occurring antibody maturation process can be recapitulated or reversed at least partially *in silico*, opening exciting new avenues for antibody engineering work. Further, my results suggest the applicability of multi-state design to engineer polyspecific antibodies, exploring another important strategy for designing broadly neutralizing antibody therapeutics. Traditional antibody engineering approaches emphasize isolating monoclonal antibodies that are highly specific for a given antigen, relying on display techniques in which emphasis typically is placed only on HCDR loop design. The method described here considers the entire antibody variable region during design, including critical framework residues that allow for conformational flexibility and contribute to polyspecificity. Considering that I found that up to 64% of framework and HCDR residues may contribute to binding and specificity, computational design will be invaluable to rapidly enumerate the large sequence and structural space of residues that can contribute to breadth of binding diverse targets.

CHAPTER III

HIV Neutralizing Antibodies in HIV-Naïve Donors

III.1 Introduction

The induction of broadly neutralizing HIV-specific antibodies is likely to be a critical component of the mechanisms of protection for an effective HIV vaccine. In the work presented here, I used a novel approach for examining the heavy chain complementarity determining region 3 (HCDR3) repertoire of HIV-naïve donors to interrogate the structural properties of long HCDR3 loops. Some broadly neutralizing HIV-neutralizing human antibodies possess long HCDR3 regions, which typically are created at the time of original antibody gene recombination rather than during somatic hypermutation. I sought to determine if antibodies with long, structured HCDR3s that are present in the naïve B-cell repertoire of HIV-naïve donors confer neutralizing properties to those antibodies similar to that of previously isolated HIV-neutralizing antibodies such as PG9 and PG16, which target the HIV envelope protein variable loops 1 and 2. Using ultra-deep nucleotide sequence analysis of HCDR3 regions in antibody gene replicons for 64 different HIV-naïve donors, I obtained approximately 25,000 unique sequences that encoded HCDR3s of 30 amino acids in length, the same size as broadly neutralizing antibody PG9. The modeling suite ROSETTA was used to assess the ability of these 30 amino acid length HCDR3 sequences to form a loop with structure similar to that of antibody PG9. The PG9 backbone template then was threaded rapidly with sequences from HIV-naïve donors to evaluate the energetic state of naturally occurring long HCDR3s when tested in a PG9-like conformation. The sequences encoding HCDR3s with the most favorable predicted energy states in this conformation then were redesigned *in silico* to optimize binding with minimal changes, simulating the process of somatic hypermutation. The sequences that were found to mimic the binding energy and HCDR3 structure of PG9 were synthesized and characterized experimentally for ability to bind to HIV envelope (Env) protein and neutralize HIV infectious virions. I found 12 unique antibodies that present PG9 HCDR3 mimicry with 0 to 7 mutations away from their respective HIV-naïve wild-type sequence. This work, using a robust new bioinformatics and modeling pipeline, suggests that HIV-naïve donors may possess naïve B-cells encoding antibodies with long HCDR3s that can neutralize HIV prior to infection. Expanding and preserving these unique naïve B-cells from the naïve repertoire may represent an important new strategy for HIV vaccine priming.

III.1.1 Potential Paradigm Shifts in Vaccinology

Elicitation of broadly neutralizing antibodies (bNAbs) against human immunodeficiency virus type 1 (HIV-1) has been one of the greatest challenges in modern vaccinology (Ackerman and Alter, 2013). A bNAb response occurs in 10-30% of those infected, with about 1% of those possessing "elite" neutralization breadth (Simek et al., 2009). These antibodies typically arise 1 year after infection at the earliest and peak at approximately 3-4 years after infection (van Gils et al., 2009). Recent advances in donor selection from cohorts of chronically infected patients, novel screening methods, and antibody isolation technologies, have allowed identification of dozens of new antibodies to the envelope protein gp120/gp41 of HIV-1 (Kwong and Mascola, 2012). These antibodies bind at least five major epitopes on gp120 including the CD4 binding site (Wu et al., 2011), the gp120 outer domain (Trkola et al., 1996), the V3 loop and V3 complex glycans (Mouquet et al., 2012), the V1/V2 loop and surrounding glycans (Walker et al., 2009), or the membrane proximal region (MPER) on gp41 (Huang et al., 2012). While the various bNAbs of interest were isolated using different strategies, they often share certain unusual and distinctive genetic features, such as very large numbers of somatic mutations or very long heavy chain complimentarity determining region 3 (HCDR3) structures (Corti and Lanzavecchia, 2013).

Although studies based on bNAb isolation have revealed new targets for immunogen design and have been useful for experimental therapeutic efficacy (Klein et al., 2012), the design of a vaccine against HIV-1 that induces sterilizing or protective immunity has remained elusive (McCoy and Weiss, 2013). For example, the RV144 trial that tested a vaccine strategy using priming immunization with a recombinant canarypox vector followed by a bivalent gp120 protein boosting immunization revealed 31% efficacy in the first year for this regimen (Rerks-Ngarm et al., 2009). While the RV144 trial results did suggest modest efficacy, the waning of protection after one year suggested major improvement in immunogenicity is still needed. An obstacle in vaccine development for elicitation of bNAbs is the limited understanding of the maturation process from germline to hypermutated variable gene segments or the process for eliciting secretion of antibodies with long HCDR3s, which are characteristic of a number of mature bNAbs (Haynes et al., 2012b).

One recent study conducted by Doria-Rose and colleagues at the Vaccine Research Center (VRC) structurally characterized twelve somatically related V1/V2 binding mAbs that share characteristics with PG9 and PG16, including a long protruding HCDR3 loop (Doria-Rose et al., 2014). They found that the unmutated common ancestor was present at 30-38 weeks post-infection and was able to weakly neutralize superinfecting virus. This finding is a great step in corroborating my rationale for eliciting these types of V1/V2 binding mAbs as a strategy for vaccine development.

Here, I attempted to study the occurrence of antibodies in the naïve B-cell repertoire with long HCDR3s that might mediate neutralization of HIV, prior to exposure to HIV antigens. I examined the HCDR3 repertoire of a large number of HIV-naïve donors in context of the HCDR3 structure of the V2/V3-binding antibody PG9. This antibody was isolated initially from a donor in the IAVI African Protocol G cohort using high-throughput B-Cell microneutralization assays to identify functional B-Cell clones (Walker et al., 2009). These antibodies possess unusual HCDR3 structures characterized by a 30 amino acid long (IMGT numbering) hammerhead-structured HCDR3 region that protrudes about $20/AA$ from the surface of the rest of the combining site (Pancera et al., 2010; Pejchal et al., 2010; McLellan et al., 2011). Initial studies using point mutations identified the HCDR3 as the primary structural feature that mediated neutralization (Pancera et al., 2010), and this finding was confirmed later by atomic-resolution structural studies of the mAbs in complex with V1/V2 scaffolds (McLellan et al., 2011; Pancera et al., 2013).

Using ultra-deep sequencing technology along with a robust molecular modeling pipeline, I identified long HCDR3 sequences from the B-cells of HIV-naïve donors that were predicted to mimic the unusual structure of mAb PG9. First, I identified 30 amino acid length HCDR3 sequences, which matched the amino length of the PG9 HCDR3. Using pilot comparative modeling experiments in the software package ROSETTA (Leaver-Fay et al., 2011b; Kaufmann et al., 2010), I created position specific structure scoring matrices (P3SMs) to score the large number of HCDR3 sequences returned from HCDR3 sequencing rapidly without the need for a full molecular modeling trajectory to assess the ability of the sequences to achieve a PG9-like loop structure. The sequences were ranked by predicted tolerance for assuming the structure of the HCDR3 of PG9 in the PG9/CAP45 complex. Sequences that returned scores close to those of PG9 were characterized experimentally by synthesis and expression of recombinant antibodies, followed by testing for binding to HIV antigens and neutralization of HIV. I found 12 antibodies that mimics the function of PG9 with at least one mAb that neutralized HIV.

Here, I used a structural convergence paradigm, as I found the sequences that mimicked PG9 function by searching for antibodies with the ability to form a similar structure, rather than clones with a high level of genetic identity or homology. The results suggest that a vaccine strategy based on elicitation of PG9-like antibodies may provide several advantages in rational vaccine design for HIV, since PG9-like HCDR3 loops are found in the HCDR3 repertoire of HIV-naïve individuals. Furthermore, the number of somatic mutations necessary to achieve broad and potent neutralization is less than the extraordinary number of mutations typical of many conventional bNAbs (Klein et al., 2013). Taken together, the targeting of B-Cell receptors with sequences that encode PG9-like HCDR3 structures in the repertoire of HIV-naïve individuals may provide a critical target in immunogen design

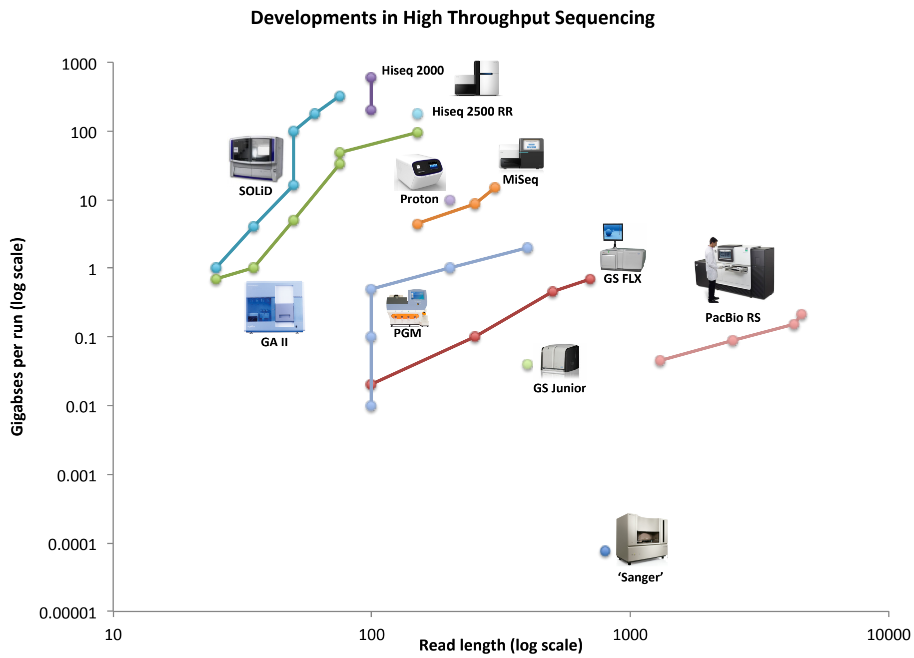


Figure III.1: Current sequencing technologies. On the x-axis is the current read length for each sequencing platform. The y-axis is the bases per run. Each point is a new iteration of that platforms sequencing read length and coverage. HiSeq has the most coverage with relatively short read lengths. Figure adapted from (Nederbragt, 2012)

for HIV vaccine development.

III.1.2 Ultra High-Throughput Sequencing

Through the remainder of this section I will refer to next-generation sequencing as high-throughput sequencing (HTS). These can be divided into technologies that often produce a smaller number of reads but allow a much longer read length. This approach includes Roche 454-pyrosequencing ©, Illumina MiSeq ©, IonTorrent ©, and PacBio © platforms. These high-throughput sequencing technologies give a throughput from 0.7 to 95 gigabases (figure III.1 and table III.1) (Nederbragt, 2012). In contrast, some platforms fall into what I call ultra high-throughput sequencing (UHTS), including those of the Illumina HiSeq © and SoLiD © platform, which tends to give shorter reads but achieves higher number of reads, from 320-600 gigabases (figure III.1, table III.1). The advances in HTS spurred an initial investigation to examine the healthy donor repertoire in my laboratory.

Platform	Instrument	Year	Reads per run	Read length (mode or average)	Bases per run (gigabases)
Sanger	3730xl	ND	96	800	0.0000768
454	GS20	2005	200,000	100	0.02
454	GS FLX	2007	400,000	250	0.1
454	GS FLX Titanium	2009	1,000,000	500	0.45
454	GS FLX+	2011	1,000,000	700	0.7
454	GS Junior	2010	100,000	400	0.04
IonTorrent	PGM 314 chip	2011	100,000	100	0.01
IonTorrent	PGM 316 chip	2011	1,000,000	100	0.1
IonTorrent	PGM 318 chip	2011	5,000,000	100	0.5
IonTorrent	PGM 318 chip	2012	5,000,000	200	1.0
IonTorrent	PGM 318 chip V2	2013	5,000,000	400	2.0
IonTorrent	Proton PI	2012	50,000,000	200	10.0
Illumina	GA	2008	28,000,000	35	1.0
Illumina	GA II	ND	100,000,000	50	5.0
Illumina	GAIIx	2009	440,000,000	75	33.0
Illumina	GAIIx	2011	640,000,000	75	48.0
Illumina	GAIIx	2012	640,000,000	150	95.0
Illumina	HiSeq 2000	2010	2,000,000,000	100	200.0
Illumina	HiSeq 2000	2011	3,000,000,000	100	600.0
Illumina	HiSeq 2500 RR	2012	600,000,000	150	180.0
Illumina	MiSeq	2011	30,000,000	150	4.5
Illumina	MiSeq	2012	30,000,000	250	8.5
Illumina	MiSeq	2013	30,000,000	300	15.0
SOLiD	3	ND	320,000,000	50	16.0
SOLiD	4	ND	2,000,000,000	50	100.0
SOLiD	5500xl	2011	3,000,000,000	60	180.0
SOLiD	5500xl W	2013	3,000,000,000	75	320.0
PacBio	RS C1	2011	36,000	1,300	0.045
PacBio	RS C2	2012	36,000	2,500	0.090
PacBio	RS C2 XL	2012	36,000	4,300	0.155
PacBio	RS II C2 XL	2013	47,000	4,600	0.216

Table III.1: Figure adapted from (Nederbragt, 2012)

III.1.3 Long HCDR3s are Established at Original Recombination

Using first generation HTS technology, we first observed long HCDR3 sequences in the healthy HIV-naïve repertoire (Briney et al., 2012). Although long HCDR3s were known to exist (Zemlin et al., 2003; Ivanov et al., 2005) they are primarily found in chronically HIV-infected subjects (Walker et al., 2009; Spurrier et al., 2011; Choe et al., 2003; Walker et al., 2011). This led to a multiple hypothesis model about the origin of long HCDR3s in chronically infected patients (figure III.2). Given that long HCDR3 antibodies against HIV are highly mutated relative to other antibodies, it could be hypothesized that insertions and deletions (indels) due to somatic mutations could be responsible for the 'elongation' of HCDR3s. These antibodies could be elicited by a chronic infection or heterogeneous prime-boost vaccine strategy, both of which will exhibit selective pressure for antibody maturation from germline sequence (figure III.2). This long HCDR3 origin model was known as the 'mutational' model. The alternate hypothesis was that long HCDR3s were established in the bone marrow at the time of original recombination and were found in the circulating repertoire at low frequency, or were selected against due to known issues associated with autoimmunity (Aguilera et al., 2001; Wardemann et al., 2003; Ichiyoshi and Casali, 1994; Crouzier et al., 1995). This is known as the "original recombination" model.

These two models for the origin of long HCDR3s were the original focus for this project and are detailed in specific aims of the 2012 thesis of Bryan Briney from the Crowe Laboratory. The models were tested by comparing immunoglobulin sequence repertoires for long HCDR3s to those of canonical length HCDR3s (Briney, 2012). I can further test these models based on two main assumptions about antibody maturation process. One, all antibodies are generated in the bone marrow from the germline repertoire and affinity mature in the lymph tissue (Tonegawa, 1983; Murphy et al., 2007). Thus, an antibody that is recently recombined in the bone marrow would have fewer mutations than antibodies that have gone through multiple rounds of affinity maturation. Figure III.4 A (top panel) shows an analysis of peripheral B-cell mean number of mutations as a function of HCDR3 length. Longer HCDR3s tend to have lower levels of somatic mutations. Figure III.3 B (top panel) specifically shows the correlation for IgM and IgG class types that have undergone positive and negative selection.

The second assumption is that the affinity maturation process is associated with insertion and deletion events that are likely to be caused by somatic hypermutation associated proteins (Reason and Zhou, 2006; Wilson et al., 1998a,b). Figure III.4 A and III.4 B (lower panels) show a negative correlation to insertion-deletion associated (indel) percentage,suggesting further evidence supporting the 'original recombination' model. Finally, a statistically significant change in long HCDR3 percentage between naïve B-cells and class switched IgM and IgG B-cells indicates that these antibody sequences are present at origi-

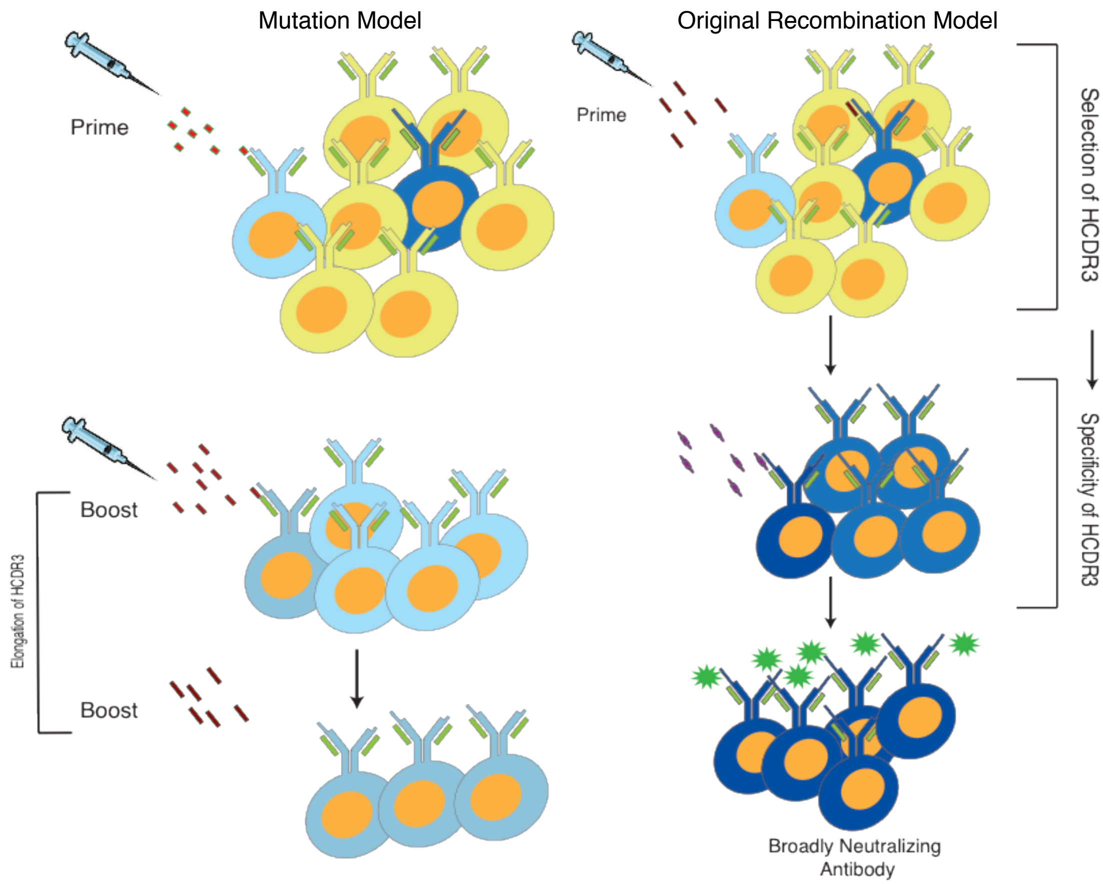


Figure III.2: Origins of long HCDR3 models. Two models are proposed to explain the origin of HCDR3s. In the mutation model (left), B-cells (yellow) with canonical length HCDR3s are 'elongated' through selective pressure on B-cells (blue) by chronic infection or repeated rounds of vaccine boosts to trigger affinity maturation to an evolving antigen. This repeated round of exposure creates insertions into the long HCDR3. The original recombination model (left) assumes the long HCDR3 is in low frequency in the naïve population (yellow). Antigens from a vaccine or chronic infection select the infrequent B-cell and expand its clonal population and refine affinity.

nal recombination and are selected against during maturation.

Although the original goal for the work done by Bryan Briney was to determine the genetic origin of HCDR3s, it confirmed a population, albeit a very small one, of very long HCDR3 sequences that are the same length as many of the exceptional broadly neutralizing antibodies against HIV-1. This fortuitous conclusion, along with rapidly advancing throughput of HTS technology, led to the experimental rationale.

III.1.4 Rationale and Experimental Design

My conclusions from examining the HIV-naïve donor repertoires made me consider various existing approaches to HIV vaccine design. Traditional vaccinology assumes that one can make a germline antibody into a broadly neutralizing antibody by a series of prime-boost steps to achieve an optimal level of mutations to produce a protective response to HIV challenge. Indeed, many great strides have been made in the field (Liao et al., 2013; Javier Guenaga and Wyatt, 2011; Kwong and Wilson, 2009; Doria-Rose et al., 2012). However, the other unusual characteristics about broadly neutralizing antibodies to HIV is many of them possess very long HCDR3s, where almost all of the contact and therefore mechanism of neutralization, stem from the long HCDR3 (McLellan et al., 2011; Pejchal et al., 2010; Klein et al., 2013; Zhou et al., 2007).

With knowledge that long HCDR3s are present in HIV-naïve donors, and that they are established at the time or recombination, I hypothesized that it was possible for these long HCDR3 sequences to mimic some of the long HCDR3-driven broadly neutralizing antibodies. I choose the complex of PG9 and a scaffolded epitope V1/V2 to be my template (McLellan et al., 2011). The goal was to see if I could mimic the 30-length HCDR3 in PG9 to neutralize HIV (PG9 mimicry). I chose PG9 with the following rationale:

1. The co-crystal structure had recently been elucidated (figure III.3) (McLellan et al., 2011).

2. The long HCDR3 accounts for neutralization and functionality with few contact residues in other regions of the antibody (Pancera et al., 2010; Pejchal et al., 2010).

3. Variants with germline reversions of the framework still retains neutralization ability (Klein et al., 2013).

4. The RV144 trial, the first trial to show substantial vaccine efficacy, revealed an increase in V1/V2 binding antibodies (the binding region of PG9) (Haynes et al., 2012a).

5. Interactions with the V1/V2 region are structure-dependent and sequence independent, *i.e.* backbone-backbone hydrogen bonding across beta-sheets (McLellan et al., 2011).

6. There is a 9 mutation difference between PG9 and its sister antibody PG16 at the HCDR3 paratope, showing a structural as well as functional convergence irrespective of sequence similarity (Walker et al., 2009; Pancera et al., 2010; Pejchal et al., 2010).

Bryan Briney's work with first generation high-throughput sequencing yielded a low frequency of sequences with a length of 30 amino acids (0.4%). Building on this observation, figure III.5 proposes an experimental plan that could test a large number of sequences for PG9 mimicry. In order to make the proposal feasible, I needed to increase the number of donors sequenced and the amount of B-cell coverage in a new round of next-generation high-throughput sequencing. I also knew that the coverage necessary was going to produce 500 million to 1 billion sequences, far more data than could be supported on my existing data analysis architecture.

In collaboration with Jessica Finn who did the sequencing, I planned to implement custom databases and search algorithms to collect the sequences into a functional antibodyome. For prediction of PG9 mimicry, I also needed to develop a structural prediction scheme based on the package ROSETTA that would rapidly be able to predict whether a sequence could tolerate the PG9 configuration. I also planned for a reasonable number of recombinant antibodies to be made in the laboratory to test for experimental validation using binding and neutralization analysis.

III.2 Ultra High-Throughput Sequencing of HCDR3s

As explained in the rationale, it was necessary to obtain many more sequences than were available using the coverage of 454-pyrosequencing. I decided to use the HiSeq platform, as the throughput at the time was 10^9 reads of 150 base pairs (bp). Although, the sequencing read of 150 bp was not long enough to cover the entire antibody variable gene, it would

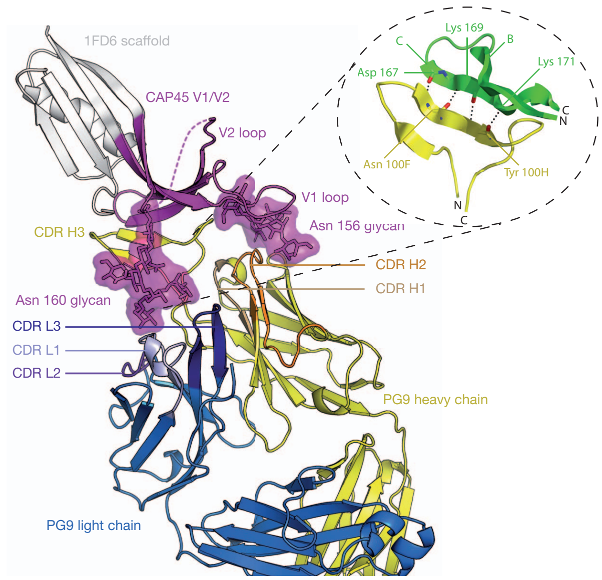


Figure III.3: PG9 in complex with V1/V2 scaffold. The crystal structure adapted from (McLellan et al., 2011). Beta-sheet interactions at the interface are highlighted.

provide coverage of the HCDR3 portion of a recombined gene, sufficient for my experimental goals. I estimated that 0.4% of the sequences would be greater than 28-length at 3 standard deviations from the mean HCDR3 length. This approach in theory, would deliver 400,000 recombined very-long HCDR3 reads.

Using the scheme found in figure III.6, I indexed antibody gene repertoires from 64 different healthy donors with indexes based on primer design by Bryan Briney and Jessica Finn in the Crowe laboratory. First, I obtained white blood cells from 64 HIV-naïve donors through the American Red Cross. No further information was obtained about each donor except for hepatitis B, C and HIV negative results. The mRNA was purified from the peripheral blood mononuclear cells (PBMC), and subjected to an RT-PCR and PCR. The HiSeq run was done in the VANderbilt Technologies for Advanced GEnomics (VANTAGE). The raw data was reconstructed with paired-end algorithms and run through PyIg (unpublished). This program called on V, D, and J gene segments for the HCDR3 region and stored them to a database custom built for large amounts of information. The statistics for the HiSeq run are found in table III.2. The full methodology can be found in appendix

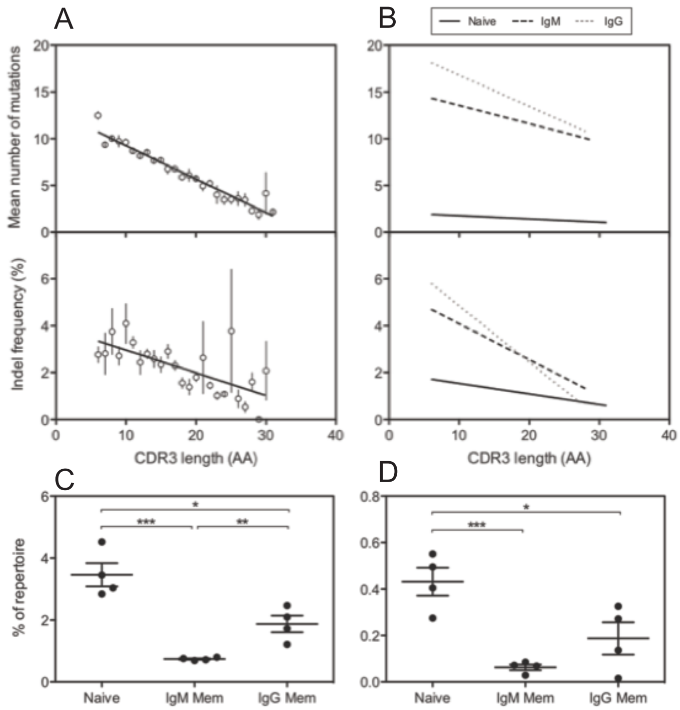


Figure III.4: Maturation sequence markers and HCDR3 length. Insertion and deletion (indel) frequency and mutation frequency is associated with affinity maturation and shows a negative correlation with HCDR3 length (A). The correlation becomes more pronounced in peripheral blood classed-switched B-cells, IgM and IgG (B). The percentage of the repertoire with long HCDR3s (>24 amino acids, C) and very long HCDR3s (>28 amino acids, D) shows a statistically significant change from naïve B-cells to class-switched B-cells (*** – $p < 0.001$, * – $p < 0.1$).

VI.3.

Next, I queried my database to get a distribution for the frequency of HCDR3 lengths. Without removing any redundancies of amino acid sequences, I binned each length and got a distribution, referred to as "all sequences" (figure III.7 A). I then removed all redundancies within each donor and plotted them as a function of length, referred to as "donor unique" (figure III.7 A, B). Finally, I made a distribution that pooled all the sequences together and removed all HCDR3 redundancies, referred to as "total unique" (figure III.7 A-C). The redundancies found in all donors combined, subtracted by all unique sequences, are by shared in at least one donor. For example, for length of HCDR3 equal to 1, I have 174 occurrences when I add up "donor unique". That is, for donor 1, I have X amount

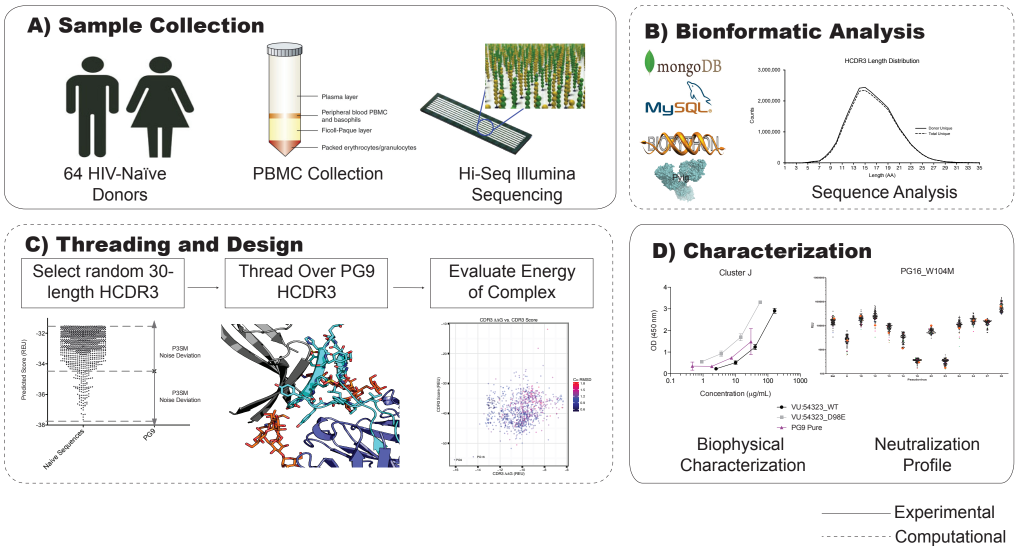


Figure III.5: The methodology can be split into four subsections that are a combination of computational (dashed-line) and experimental work (solid-line). HIV-naïve donor blood is collected from 64 adult donors and the HCDR3 is sequenced on the HiSeq Illumina platform (A). The raw sequences are reconstructed and analyzed against germline databases using custom software. The sequences are parsed and stored in optimized databases to handle the large quantity of antibody sequences (B). HCDR3 sequences are chosen by length and tested for PG9 mimicry using the ROSETTA software suite. Iterative rounds of minimization, docking, and design, followed by rigorous statistical analysis allows for a robust prediction of potential candidates from the HIV-naïve donor repertoire that may neutralize HIV (C). A tractable number of sequences are synthesized and tested experimentally through biophysical characterization and neutralization studies against HIV-1 (D).

of unique sequences among that donor added to X amount of sequences for donor 2, etc. However, when I pooled all the sequences of length 1 first, and then removed redundancies, I arrived at 17 "unique sequences". That means there are 137 amino acid sequences found in one or more donors. An easier way to view it is to plot the percentage of sequences at a given length that are shared among one or more donor (figure III.7 D).

The mean length of HCDR3 sequences in my dataset was 16.40 ± 4.03. This finding was in agreement with our previous work using a much smaller dataset (Briney et al., 2012). Although there is no standard definition of a 'long HCDR3', I arbitrarily chose the cutoff to be two or three standard deviations from the mean for long or very long HCDR3s, respectively (24AA, 28AA, figure III.7 C). My recombination software (PyIg)

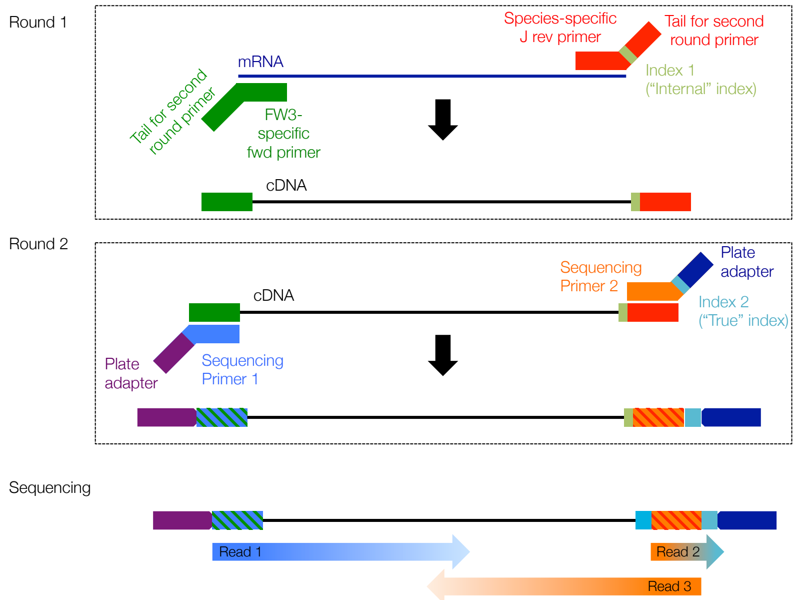


Figure III.6: RT-PCR in round 1 allows addition of an internal index to categorize donors. The cDNA is then subjected to a nested round 2 PCR where the necessary HiSeq plate adaptors and sequencing regions are added that are used by Illumina. The sequencing makes two pair-end reads that are later reconstructed.

also can assign V, D, and J gene segments using the BLAST algorithm (Ye et al., 2013). I observed similar trends in D and J gene family usage as a function of length. D_H3 and J_H6 family increased as length of HCDR3 increased (figure III.7 F, G). This finding makes intuitive sense as these are the longest gene segments in their respective families. For V-gene families, I did not observe a difference in germline gene usage as a function of length (figure III.7 E). This trend also was reported with our previous work in which I sequenced the entire antibody segment, however, I can't be absolutely confident in my assignments of V-gene considering I only have on average 20 bp of the V-gene. I can also narrow down individual genes for the D-gene segment and observed an increase in D_H2-2, D_H2-15, D_H3-3, D_H3-10 and D_H3-22 as HCDR3 length increased (figure III.7H).

III.3 Addition of Non-Canonical Amino Acids in ROSETTA

The ROSETTA software suite was initially developed for *ab initio* folding of small globular proteins using fragment-based search methods and knowledge-based potentials to score

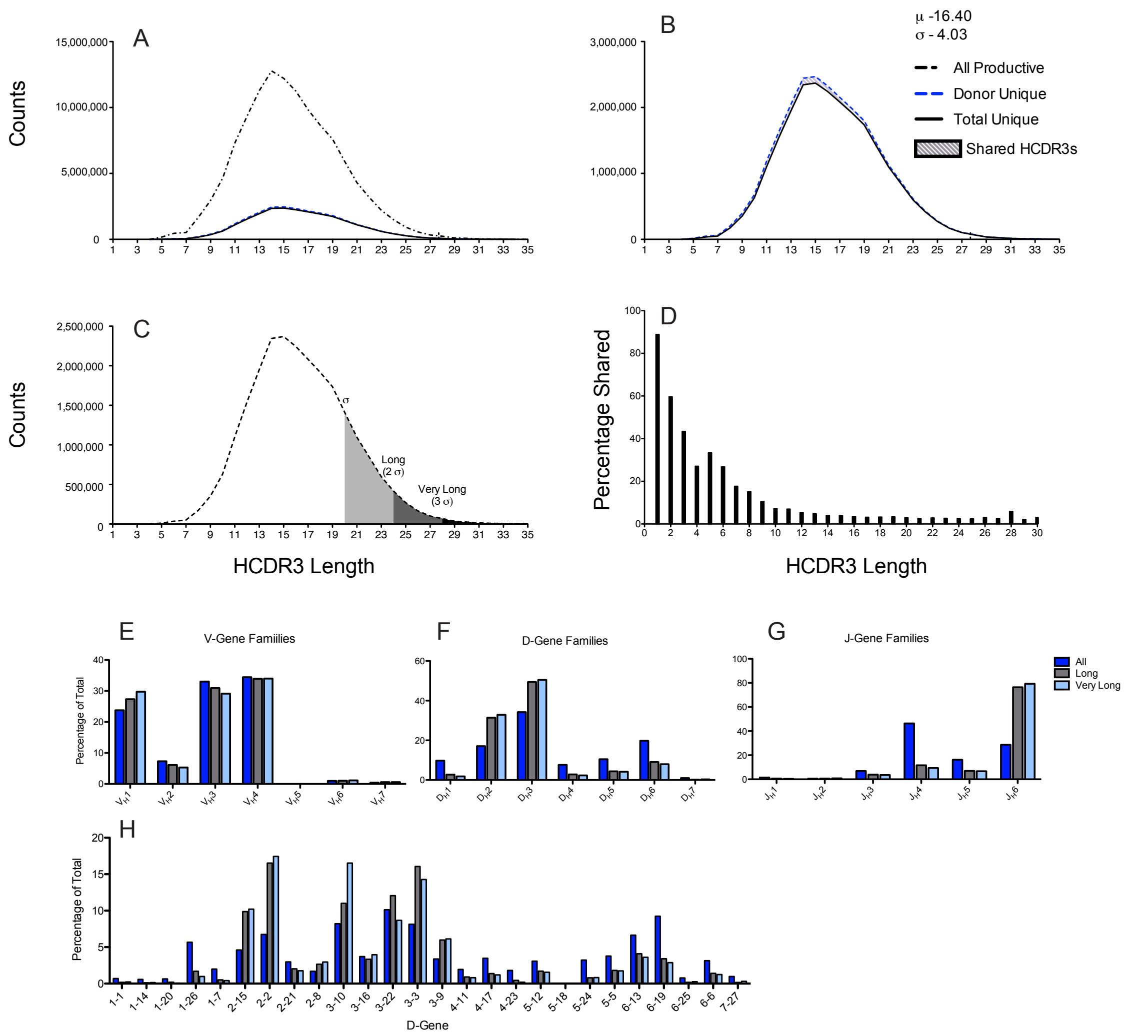


Figure III.7: Length distribution for all productive sequences (black dashed line), unique among within each donor (blue dashed) and unique among every donor (black solid) (A). Zoomed in distribution of "total unique" and "donor unique" to see overlap. Sequences that are unique between donors are shaded in grey (B). "Total unique" with standard deviations shown at 20, 24, and 28 (C). Percentage of shared sequence as a function of HCDR3 length. "Shared sequence" is defined as being found in two or more donors (D). V gene families, D Gene families and J Gene family frequencies for all sequences, long, and very long HCDR3 sequences (E, F, G respectively). D-gene families can further be divided into individual gene frequencies for all, long, and very long HCDR3 sequences (H).

models (Simons et al., 1997). Because of this, the code for ROSETTA made protein amino acid residues the smallest object. All proteins would be made up of residue objects, and all building blocks would be made up of parameters that describe a residue object. This made

Metric	Counts
Raw Reads	514,312,664
Joined Reads	460,931,435
Unique Reads	167,667,706
Recombinant Reads	159,609,585
Productive CDR3s	118,440,255
Unique CDR3s by AA	23,357,390
30 Length CDR3s	74,457
Unique 30 Length CDR3s	24,917

Table III.2: Sequencing results from the HiSeq of 64 donors. Raw reads indicate the amount of reads that passed VANTAGE quality metrics. Joined reads are reads that found a paired end partner and could be joined together. Unique reads removed duplicates. Productive HCDR3s are those reads that do not contain a stop codon. Unique CDR3s are those HCDR3s that are not duplicated by amino acid sequence. 30 length HCDR3s are those sequences that are 30 amino acids long by IMGT numbering. Unique HCDR3 sequences are those sequences that are not duplicated in any donor by amino acids.

sense at the time of ROSETTA's inception given its primary purpose, but as the scoring function was being developed for a wide variety of molecular modeling tasks, the residue based code became difficult to implement for non-residue type molecules, *i.e.* drug-type ligands, glycans, and post-translational modifications.

The PG9 complex relied on complex and high-mannose type glycans that were bound to asparagine residues at HIV Env protein position N160 and N156 (HIV strain HXBc2 numbering) (McLellan et al., 2011). Removal of these residues abrogated binding to HIV envelope (Doores et al., 2010). Thus, it was necessary to include both of these glycans in my predictions. It also was necessary to encode these glycans as non-canonical amino acid residues rather than post-translational modification due to the residue object requirement of ROSETTA score function. I based the protocol of adding the two non-canonical amino acids after the work of Renfrew and colleagues who developed a generalized protocol for implementing non-canonical amino acids into ROSETTA (Renfrew et al., 2012). The protocol used a series of steps that I followed loosely that involved extraction of the residue, minimization of the residue using quantum mechanics simulations, and description of the new amino acid types as a series of parameters that ROSETTA can recognize. I first had to benchmark the non-canonical amino acid types before I could use them in the protocol. To do this, I ran minimization and design against the glycan and determined if my best scoring models were the closest to the native structure found in the PDB. The best scoring model according to total energy was aligned with the native structure. I saw minimal deviation from the native structure indicating that for good scoring models, the glycan conformation is preserved (figure III.8 A). A general trend between the glycan score and the total score as a function of the glycan RMSD (the deviation from the native conformation) was observed, indicating that the glycan deviation follows the ROSETTA scoring function (figure III.8 B). I also observed that PG9 HCDR3 amino acids were ideal for antibody-antigen complex during redesign of the PG9-antigen (figure III.8 C).

III.4 High-Throughput Threading of HCDR3 Sequences

I took 4,000 random HCDR3 sequences from an available dataset of 26,417 (table III.2). These sequences were "threaded" over the PG9 HCDR3 backbone and energetically minimized. I did not include the antigen in this first set of modeling experiments as my first goals were to establish a high-throughput prediction model of PG9-like antibodies and not necessarily anti-HIV gp120 specific neutralizing antibodies. This approach was for an experimental contingency in case my modeling experiments predicted that none of the sequences could bind gp120 antigen.

The threading protocol removes amino acid sequence identity of the HCDR3 loop and replaces, or "threads", one of the 4,000 random HCDR3 sequences from my dataset. After the amino acid identity has been replaced, the PG9 antibody is energetically minimized and scored with the ROSETTA scoring function which is detailed in chapter I.

Considering each model took approximately 1.2 CPU hours to complete at the computing cluster (https://vanderbilt.accre.com) at the time of writing, all 26,417 models would have taken 31,700 CPU hours to complete. Considering ROSETTA takes approximately 1,000-10,000 models to determine energy landscape (Simons et al., 1999a; Bradley et al.,

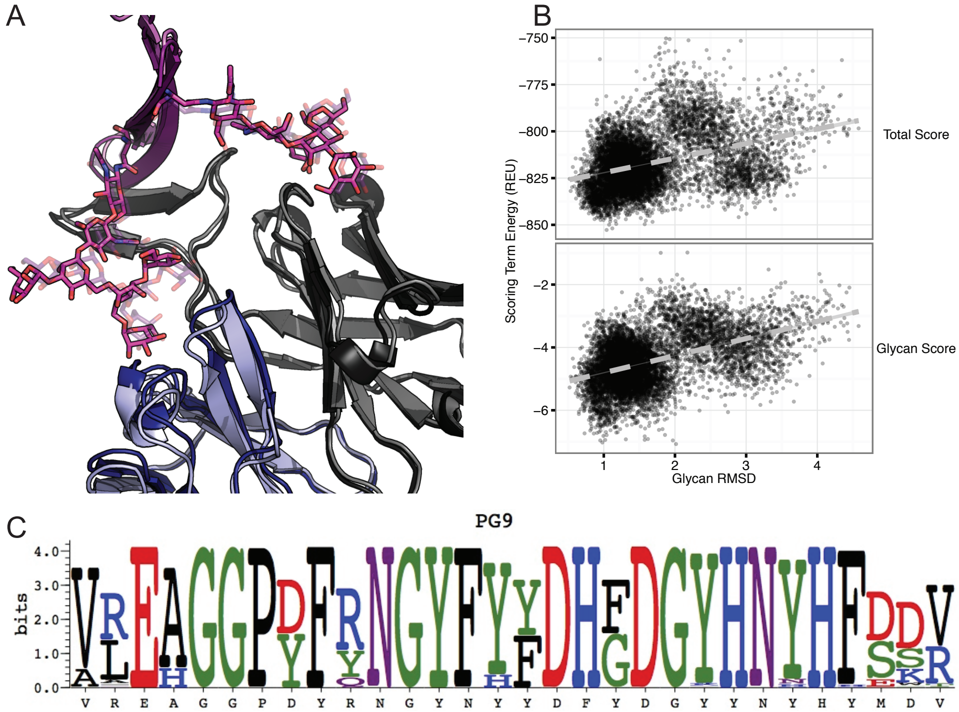


Figure III.8: Glycan addition and benchmarking template. Energetically minimized loop of the PG9 HCDR3 overlaid on the native PG9-antigen complex (PDB-3U4E) (McLellan et al., 2011). Native PG9-antigen is shown in darker colors while the redesigned, energetically minimized and re-docked structures are shown in lighter colors. Light chain is shown in blue, heavy chain grey, and antigen in magenta. The two glycans are shown in stick representation. The native glycan positions are shown in transparent stick conformation (A). The score of the glycan and the score of the model are shown as a function of the glycan RMSD. On the y-axis are ROSETTA scores, and on the x-axis is the glycan RMSD from native structure found in the PDB. The top panel is the total energy of the complex compared with the glycan RMSD while the bottom compares the glycan score to the glycan RMSD. A general trend is shown between the glycan deviation from native conformation and an increase in score (B). Sequence logo of redesigned HCDR3 loop of PG9 with glycans present. The x-axis shows PG9 native sequence (C).

2005a, 2003; Das et al., 2007; Raman et al., 2009), the time needed for completion was approaching CPU times from 3.17 x 10^7 - 3.17 x 10^8 hours. I could utilize approximately 600 processors running in parallel, cutting down CPU times to 52,834 hours or 6 years. These numbers were unrealistic at the time of writing, so I decided to optimize heuristics methods that could be accomplished with a far fewer number of models. I chose 4,000 random sequences with 50 models as a strict cutoff to maximize data output within the capabilities of my computational resources.

To evaluate the threading protocol, I scored my models using the ROSETTA scoring function and plotted it against the deviation from PG9's native structure (figure III.9 A). As a control, PG9 and PG16 also were put into the threading protocol to evaluate the scoring function's ability to make distinctions between poor scoring and favorable scoring models. Considering I know that PG9 and PG16 sequences form PG9's backbone conformation, I expected a very low deviation from PG9's crystal structure conformation. I also expected ROSETTA's scoring function to score these sequences as favorable relative to many of the random HCDR3s that were evaluated in the protocol. Indeed, I observed that PG9 and PG16 consistently ranked as the most favorable in terms of score and conformation and confirmed the accuracy precision of my threading protocol (figure III.9 A).

In contrast, the randomly selected 4,000 HIV-naïve HCDR3 sequences tended to group into three separate categories: 1) Those sequences which maintain the HCDR3 structure of PG9 but scored poorly (figure III.9 B), 2) sequences that scored well but collapsed or deviated away from PG9's native conformation (figure III.9 D), and finally 3) sequences that scored well and retained PG9's native conformation (figure III.9 C). These three groups gave a diverse sequence-energy landscape to be carried on into my heuristics in order to evaluate the remaining 22,000 HCDR3 sequences.

III.5 Heuristics to Rapidly Score HCDR3 Sequences

Using the selected 4,000 HIV-naïve HCDR3 sequences, I randomly chose half of the sequences to be in the benchmark set or the training set. For the training set, 2,000 sequences were evaluated with the ROSETTA scoring function. For each amino acid position 96-125 (PDB numbering), I filled a position structure specific scoring matrix (P3SM). For each position, and each amino acid identity, I gave an initial score of 0.0 to baseline yielding a 20 x 30 matrix, and averaged each amino acid identity score at each position (figure III.10 A,B). The matrix can be visualized as a heat map with favorable scores in blue and unfavorable

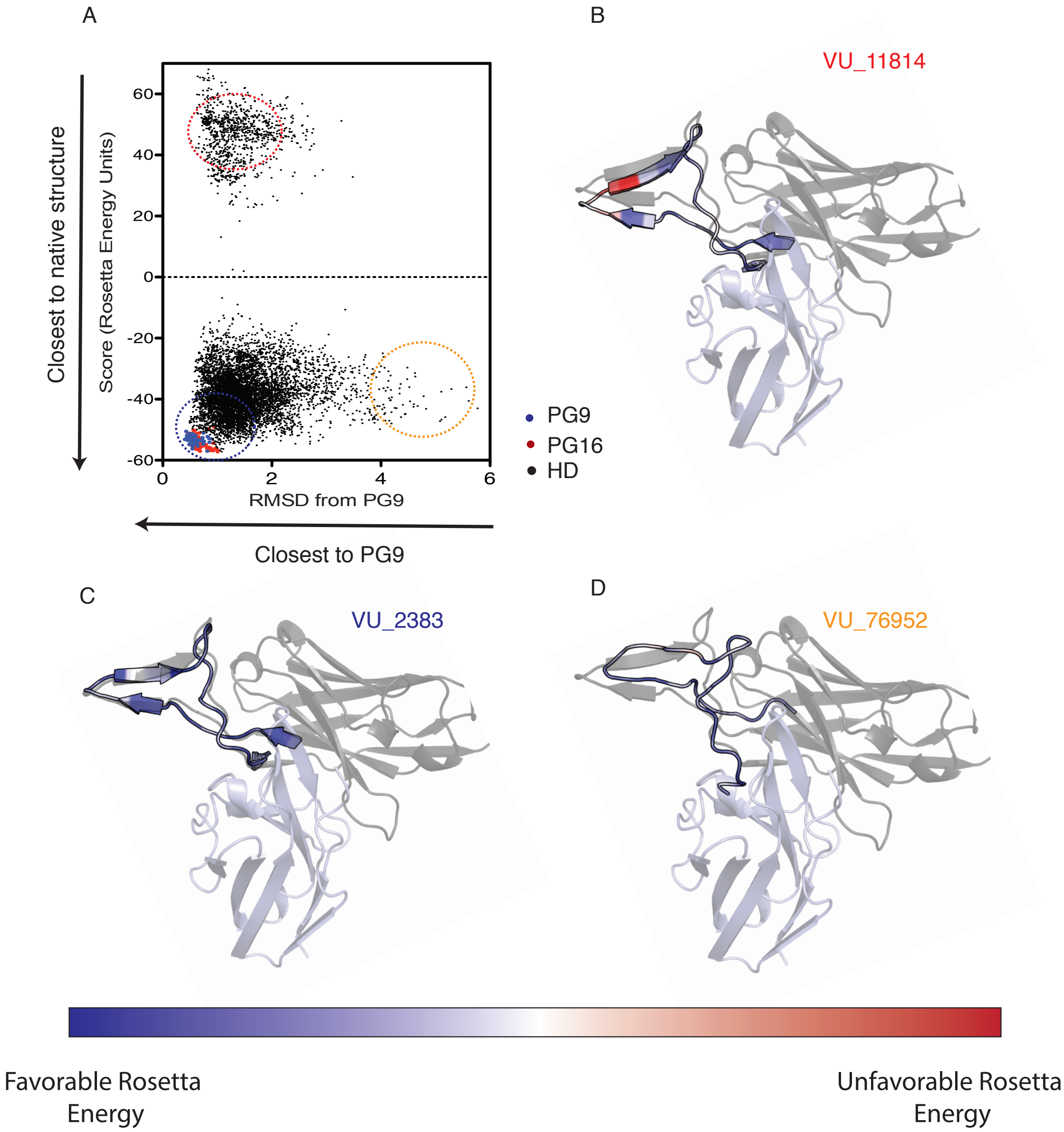


Figure III.9: The threading protocol for PG9 was evaluated for structural mimicry using the ROSETTA scoring function. The models scores were evaluated and plotted against the root mean squared deviation (RMSD) from the native PG9 HCDR3 structure. Lower scoring models are closer to native structure (y-axis) while models closer to PG9 HCDR3 structure have a lower RMSD (x-axis). Repeated measures of PG9 and PG16 were grouped close to native structure and close to PG9 structure (blue and red points) while HIV-naïve donor antibody sequences are shown are labeled HD (black points) (A). For figures B-D, HCDR3 structures were aligned to native PG9 (shown in transparency). The HCDR3s are colored by their ROSETTA score with blue being a favorable scoring residue and red being an unfavorable score. There were three different outcomes observed for the threading experiment. Sequences that retain PG9's structure but produce an unfavorable score (B, dashed red circle in figure A), sequences which produced a favorable score but collapsed away from PG9 native structure (D, dashed orange circle in figure A), and sequences that scored favorably and adopted PG9 conformation (C, blue dashed circle in figure A).

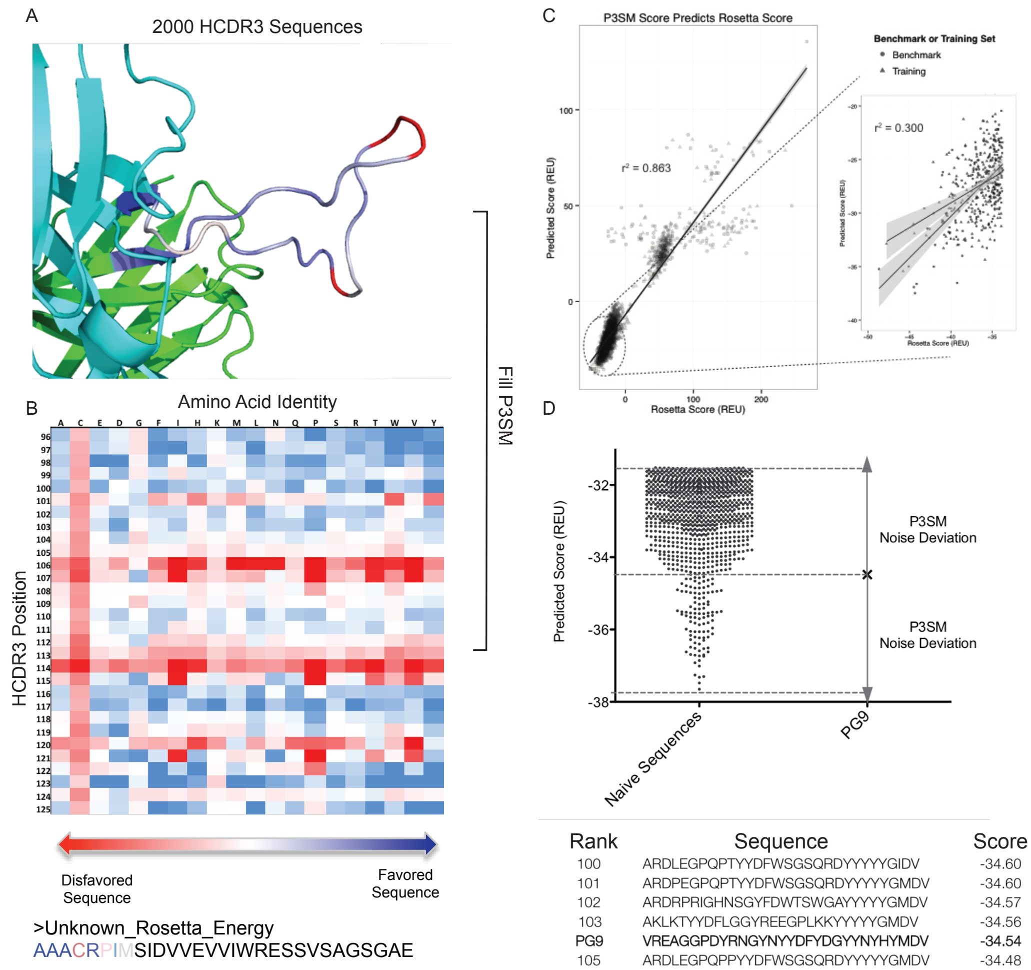
scores in red (figure III.10 B).

The P3SM became my heuristic to rapidly score other sequences that were not run through the computationally expensive threading protocol. I could validate my heuristic by evaluating how well the P3SM predicted ROSETTA energies by applying it to the other 2,000 sequences in the benchmark set. The P3SM may not give absolute ROSETTA energies but should provide a relative ranking compared to other HCDR3 sequences. I observed a 0.863 r^2-value for a correlation between actual ROSETTA energies and P3SM predicted energies. For the top 10% energies evaluated by the ROSETTA energy function, my correlation fell to 0.300 (figure III.10C).

To determine the relative noise of the P3SM, I ranked the scores and determined the position of PG9. It was found in 104^{th} position out of all 26,417 sequences (0.3%). The difference in the PG9 P3SM score and the top scoring sequence was 3.82 ROSETTA energy units (REU). Thus, the final noise tolerance of my matrix was -34.84 $\pm$ 3.82 (-31.02 to 38.66) that encompassed approximately 1000 sequences (figure III.10 D). I then defined my P3SM heuristic to be able to accurately pick out the top 1000 out of 26,417 sequences (3.79%).

III.6 Docking and Minimization of HCDR3 Variants

With 1,000 candidate sequences from the original sequence pool of 26,417, I ran the threading protocol again in the presence of the antigen and complex glycans that were added from earlier sections (see addition of non-canonical amino acids). The threading protocol was modified slightly to include all atom constraints, and an additional docking step and 100 models generated for each sequence (100,000 models). The full protocol is detailed in the Appendix section VI.3 with a protocol capture in section VI.5. In addition to PG9 deviations and total scores, I also evaluated binding energies, which are the change in free energy ($\Delta\Delta$G). I defined the $\Delta\Delta$G as follows:



Rank	Sequence	Score
100	ARDLEGPQPTYYDFWSGSQRDYYYYYGIDV	-34.60
101	ARDPEGPQPTYYDFWSGSQRDYYYYYGMDV	-34.60
102	ARDRPRIGHNSGYFDWTSWGAYYYYYGMDV	-34.57
103	AKLKTYYDFLGGYREEGPLKKYYYYYGMDV	-34.56
PG9	VREAGGPDYRNGYNYYDFYDGYYNYHYMDV	-34.54
105	ARDLEGPQPPYYDFWSGSQRDYYYYYGMDV	-34.48

Figure III.10: Heuristics predict HCDR3 Sequences that mimic PG9 structure. 2,000 random models were evaluated at each amino acid position. A cartoon representation of the HCDR3 loop is shown with each position colored with a blue-red gradient according to favorable to unfavorable amino acid identities, respectively (A). Each position was initially assigned a score of 0.0 and enumerated through averaging each amino acid identity with the score to fill a position structure specific scoring matrix (P3SM) from 2,000 HCDR3 sequences. The matrix was then used to rapidly score the other 2,000 models to predict a ROSETTA score (B). A simple linear regression was applied to the actual ROSETTA energy score of a sequence and it's predicted score to give a coefficient of determination (r^2) of 0.863. When only the top 10% by actual ROSETTA score were considered the coefficient of determination dropped to 0.330 (C). Using the P3SM, PG9 scored 104th. Calculating the difference in score between the top scoring sequence and PG9 left a noise tolerance of 3.82 ROSETTA energy units (REU). Calculating $\pm$ 3.82 REU of PG9 left 1,000 HCDR3 sequences that fell within the noise tolerance range of the P3SM (D).

$$\Delta\Delta G = \Delta G_{\text{Bound}} - \Delta G_{\text{Unbound}}$$

Where,

$$\Delta G_{\text{Bound}} = \text{RosettaScore}_{\text{Complex}}$$

and

$$\Delta G_{\text{Unbound}} = \text{RosettaScore}_{\text{Separated}}$$

I decomposed the binding energies, total energies, and deviations into several metrics to evaluate the models. Total energy, binding energy, and deviation (Cα - RMSD) for just the HCDR3 portion of the model, the binding energy contribution by both of the glycans (N156 $\Delta\Delta$G and N160$\Delta\Delta$G, HIV strain HxBC2 numbering), and the total binding energy for the entire model. Initially each of these metrics was evaluated individually to see where they ranked or in pairs by plotting them against each other (figure III.11, left panel). Favorable sequences were near PG9 in the plot. I also plotted HCDR3 metrics as heat maps and score models qualitatively (figure III.11, right panel). Both of these methods were inefficient, as each metric produces a different rank ordering of the HCDR3 sequences. Therefore, I sought to combine these six metrics into one score to easily compare where each sequence ranked in comparison to PG9. To combine the metrics, I assigned a Z-score to each metric to find out where that model ranked. The Z-score was weighted for each scoring metric and averaged to produce a weighted Z-score that was used to efficiently rank sequences with one score using the following equation:

$$\text{Weighted-Z} = \frac{\sum_i^N w_i \times Z_i}{N}$$

where w_i is the weight of each Z-score statistic i, Z_i is the z score for the statistic i, and N

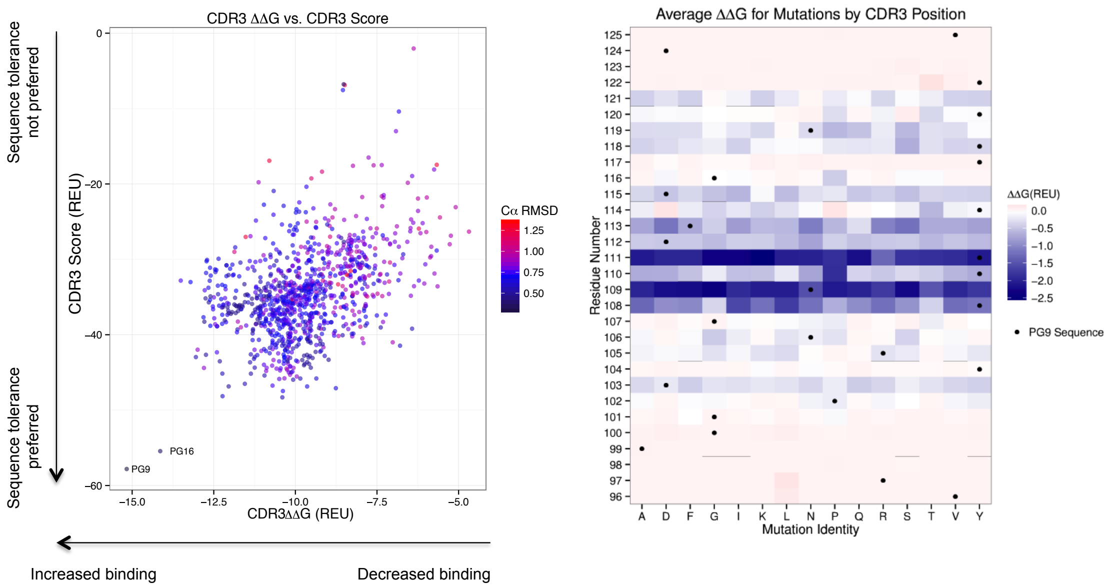


Figure III.11: Scatter plots and heat maps for P3SM threading analysis. Each metric was plotted on a separate axis and sequences which were found to score well were found in the lower left of the graph. For example, HCDR3 binding energy is plotted against HCDR3 score. PG9 and PG16 were found to have favorable binding energy and score (left panel). A heat map was generated for each metric for residues in the HCDR3. For example, contribution to binding energy for each amino acid identity was plotted as a heat map. The red-blue scale is for neutral to favorable reactions (right panel).

is the total number of statistics. The addition of the weights allowed optimization during *ad hoc* analysis to ensure PG9 and PG16 were the most favorable weighted Z-score. The final weights used in the protocol were:

1. Total ΔΔG — 3.0
2. HCDR3 ΔΔG — 1.0
3. HCDR3 Total Score — 1.0
4. HCDR3 C_{α}RMSD — 0.5
5. N156ΔΔG — 0.5
6. N160ΔΔG — 0.5

I chose the top 80 HCDR3 HIV-naïve donor antibody sequences to carry on to experimental steps, as this number was my upper limit to synthesize.

III.7 Clustering of HIV-Naïve Sequences

Rather than synthesize all 80 HIV-naïve donor antibody sequences that were predicted to be closest in structure to PG9, I instead considered that several of the sequences were related siblings to each other. Indeed, upon clustering at a threshold 85% amino acid similarity (4 mutations), the sequences clustered into nine different groups, while five sequences formed independent groups (figure III.12).

I next aligned the nucleotide and amino acid sequences to find sequence similarities among antibodies in each of my PG9 clustering groups. Surprisingly, only the beginning and end of the sequences corresponding to the base of the HCDR3 displayed a high degree of similarity (figure III.12 A-B). The similarity arises from the in-frame J_H-6 gene for most long HCDR3 sequences, with the exception of cluster I and IG2, which used J_H-4 and had evidence for significant J gene exonuclease activity, respectively (table III.3). I also detected some nucleotide conservation for positions 27-36 corresponding to amino acid 9-12, which resulted in a semi-conserved SSGY motif (figure III.12 A-B). Rather than synthesize and express each of the 80 variants, I chose to synthesize one member from each cluster and one of the sequences that did not cluster. These sequences were selected based on their ROSETTA weighted Z-score metric within each cluster (table III.4).

III.8 Design of Top PG9-Mimicry Candidates

I realized that the models for these wild-type sequences sometimes contained clashes or caused other strain in the HCDR3 loop that was reflected as a difference in normalized Z-scores (table III.4) or as an energetic gap between the predicted energies of PG9 and the top 80 variants selected (figure III.13 A). Because of this observation, I decided to run a dock-design protocol that would relieve clashes and strains for unfavorable amino acids. I also imposed a filter that gave bonuses for amino acids that were native to the starting sequence. In this way, I was able achieve a higher level of PG9 mimicry by lowering the energetic gap while retaining as many wild-type amino acids as possible (figure III.13).

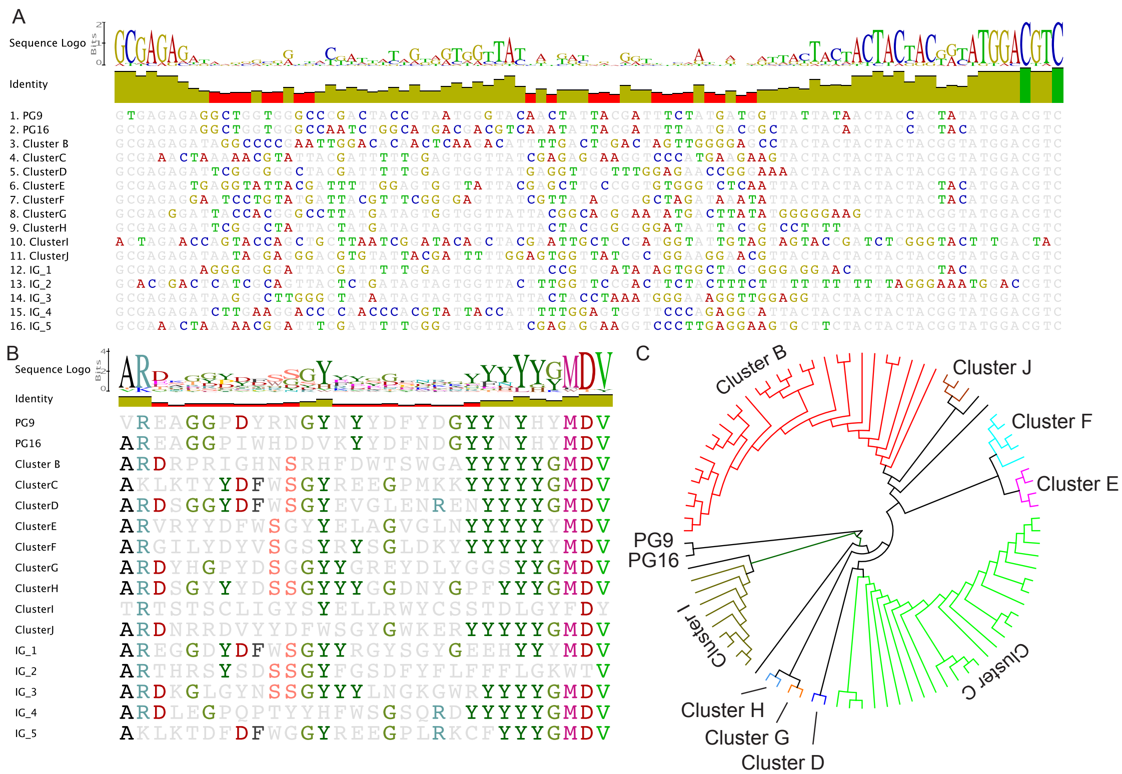


Figure III.12: PG9-mimicry candidates cluster into groups. The consensus nucleotide sequence were aligned for each cluster B-J. There was little sequence similarity in the junctions except for the V_H and J_H gene. Sequence logo representations are shown above the sequence to detect conservation. Five independent group sequences are also shown (A). The same as (A) translated. Conserved elements are shown in the J_H region due to conserved use of the J_H6 gene segment (B). The cladogram for each HCDR3 amino acid sequence shows how the top 80 sequences clustered into 9 groups with some groups having multiple sequences that differ by 1-4 amino acid mutations but were derived from the same lineage (C).

I chose a variant from each cluster whose sequence recovery (the percentage of designed sequences that were wild-type) and carefully analyzed each predicted mutation suggested by ROSETTA. I ranked the mutations based on a fitness score. I defined the fitness score as the change in total score for the mutation from the wild-type added to the change in binding energy for the mutation compared to wild-type. If the fitness was found to be significant, the mutation was confirmed by visual inspection in PyMoL (*i.e.*, if ROSETTA predicts more hydrogen bonds, are there more hydrogen bonds).

Briefly, I will explain the rationale for choosing the mutations for cluster B. Variant

Cluster	Members	V_H	V_D	V_J	V-D Length	D-J Length	D 5'-Exo	D 3'-Exo	J-Exo
B	29	V3-07*01	D3-09*01	J6*02	21	15	3	10	2
C	25	V4-34*01	D3-03*01	J6*02	10	24	0	6	2
D	2	V1-46*01	D3-03*01	J6*02	9	25	4	6	3
E	4	V1-02*02	D3-03*01	J6*03	4	23	0	4	0
F	5	V4-34*01	D3-16*02	J6*03	9	15	4	2	0
G	2	V1-02*02	D3-22*01	J6*02	16	16	7	3	10
H	2	V4-04*02	D3-22*01	J6*02	6	23	1	0	8
I	9	V4-34*12	D2-02*01	J4*02	54	9	4	12	2
J	3	V3-07*01	D3-03*01	J6*02	14	14	0	6	1
IG1	1	V3-07*01	D3-03*01	J6*03	8	29	2	4	9
IG2	1	V2-70*01	D3-22*01	J6*02	10	48	3	7	25
IG3	1	V1-02*02	D3-22*01	J6*02	11	21	6	0	5
IG4	1	V1-03*01	D3-03*02	J6*02	17	9	0	8	0
IG5	1	V4-34*01	D3-03*01	J6*02	12	29	2	6	7
PG16	1	V3-33*05	D3-03*01	J6*03	34	9	1	18	2
PG9	1	V3-33*05	D3-03*01	J6*03	34	0	1	2	9

Table III.3: Each of the nine clusters and independent sequences (IG1-5) and their representative V, D, and J genes are shown. V-D and D-J lengths are the nucleotide lengths of those junctions. D 5'-, D 3'-, and J-Exo were the amount of nucleotides excised in the junctions to make a productive recombination. Point mutations in the junction are not shown.

VU42252 was chosen as the representative candidate for design since it provided the best sequence recovery to the wild-type sequence as well as beneficial fitness (figure III.14 A). Each mutation is plotted as a function of its fitness and grouped together by position (figure III.14 B). If there was a large decrease in energy from the wild-type sequence, it corresponded to an increase in fitness for a given mutation. Those mutations that had a favorable change in fitness with a magnitude of greater than 1.5 ROSETTA energy units were manually inspected with PyMoL (figure III.14 C,D). Two such mutations are shown for position 106 (figure III.14 C, PDB numbering) and position 120 (figure III.14 D, PDB numbering). For position 106, the wild-type serine was not preferred since it leaves a large exposed gap between the antigen face and the antibody interface. Upon mutation to an asparagine, the gap was filled in. Additionally the asparagine makes new hydrogen bond contacts with the antigen. This finding was predicted to benefit the binding energy and stabilization of the complex. Position 120 has a wild-type tyrosine that repels against the large steric bulk of the glycan. A mutated asparagine at this position allows an inter-

HCDR3 hydrogen bonds to stabilize the loop while also making additional hydrogen bonds to the glycan. It is important to note that visual inspection using PyMol of these mutations were reflected in the ROSETTA scoring function.

For cluster B, I chose six different positions to be mutated alone or in combination that produced five different candidate variants. This proposed a panel of six different variants to validate for cluster B, one wild-type sequence, and five mutational variants that ranged from 3-10 mutations. The remaining clusters were analyzed similarly and in total, 10 wild-type sequences were chosen for experimental characterization, and 74 different sequences that were mutational variants of the representative cluster sequence were chosen.

Cluster	**Weighted Z-Score**
Cluster B	-1.24
Cluster C	0.28
Cluster D	0.11
Cluster E	0.15
Cluster F	-0.34
Cluster F	-0.20
Cluster G	0.31
Cluster H	0.11
Cluster I	-2.54
Cluster J	1.48
PG9	-4.80
PG16	-1.87
IG 4	-1.46

Table III.4: Weighted scores of PG9-mimic clusters. The top scoring sequence for each cluster are shown with weighted the Z-score.

III.9 Synthesis and Screening of PG9-Mimics

For the 84 variants, I chose a synthesis strategy that would allow us to swap in HCDR3 sequences with unique cloning sites rather than resynthesize the framework and HCDR1-2 sequences redundantly. It was to this end that I introduced two unique cloning sites at the 5′ end and the 3′ end of the HCDR3 sequence. Using these cloning sites, small gene fragments could be synthesized rapidly and cost-effectively. I did not expect that all of

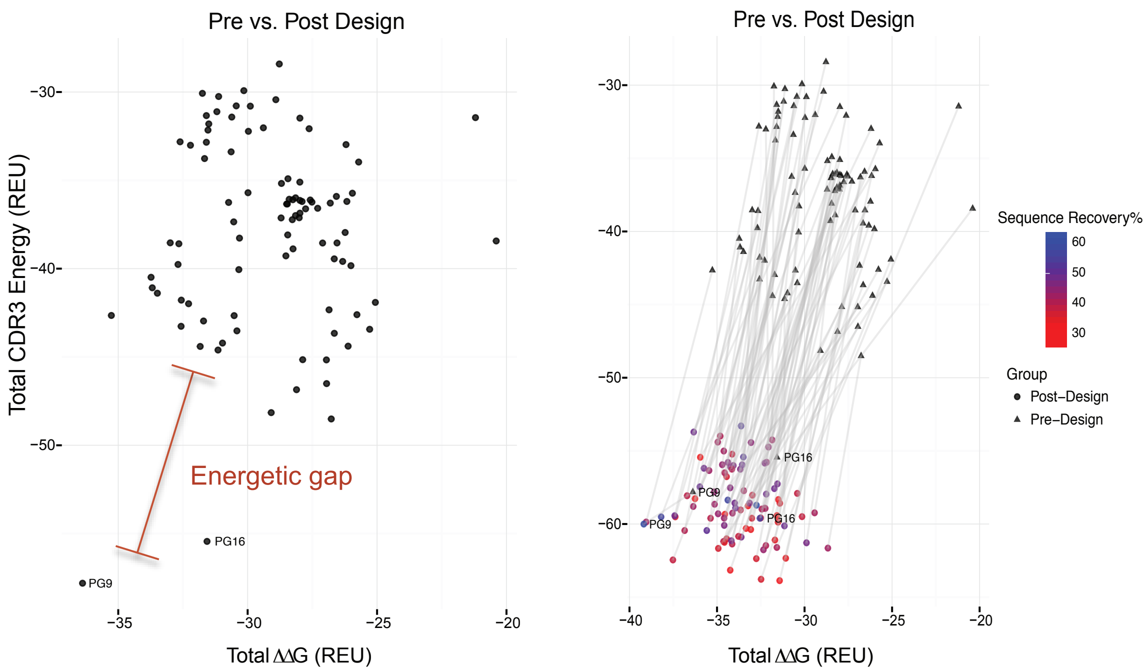


Figure III.13: Energetic barriers for complete PG9-mimicry. The total HCDR3 energy and binding energy are plotted for the top 80 sequences selected by normalized Z-score as well as PG9 and PG16. There was a significant energetic gap between these sequences and complete PG9 mimicry (A). After redesign, the sequences approached the energy of PG9 and PG16 (shown as connecting triangles to circles). The blue-red scale is the sequence recovery percentage, indicating how much of the wild-type sequence was retained.

the 84 variants would express or bind HIV gp120 and decided to screen each candidate for binding and expression on a small scale. I made 30 mL test cultures for recombinant antibody expression to screen the supernatant for the presence of IgG antibody and its ability to bind gp120 protein. According to the literature cited by McLellan *et al.*, only certain gp120 monomers bind to PG9 using enzyme-linked immunosorbent assays (ELISA), while PG16 binds gp120 monomers very weakly (McLellan et al., 2011). Therefore, I combined fifteen different gp120 protein variants into an antigen cocktail to maximize the chances of binding to each of my 84 PG9-mimics. This cross screening for antibody expression and antigen binding was designed to pre-emptively select which of my 84 candidates should be carried on to upscaled expression.

Figure III.15 plots my screening metrics for expression and binding. I binned each

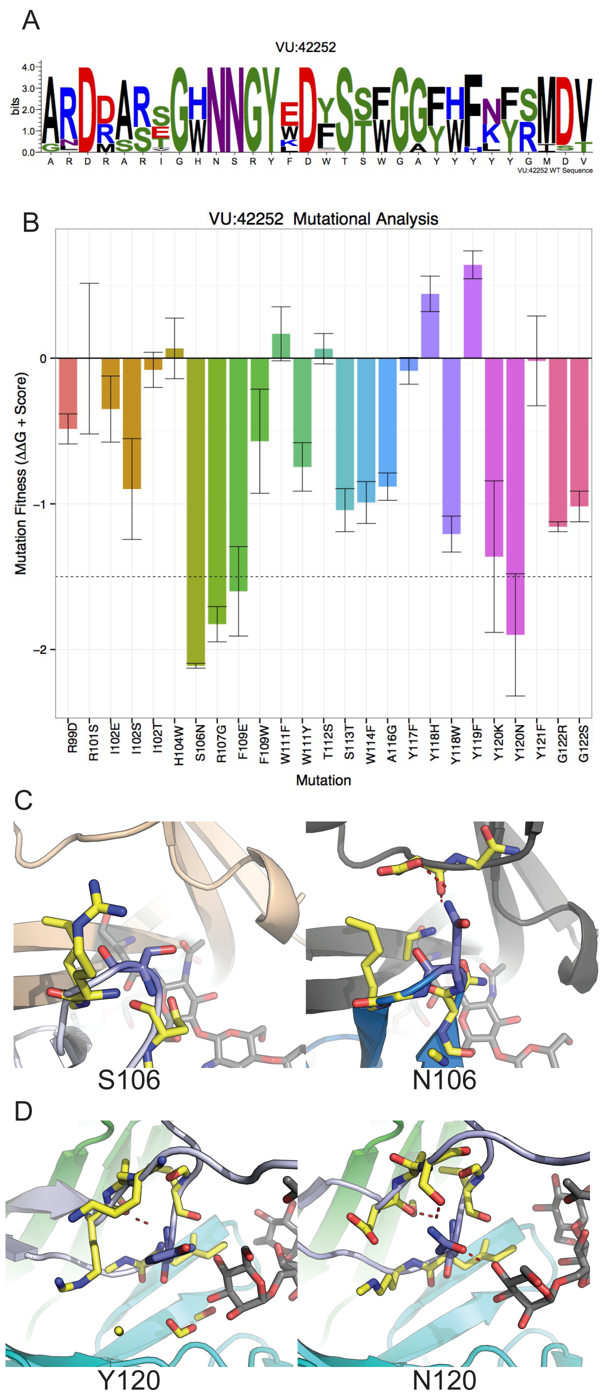


Figure III.14: Mutation analysis of cluster B. A sequence logo representation of the cluster B variant with the best sequence recovery is shown. The wild-type sequence of the cluster B candidate named VU42252 is plotted on the x-axis. The preferred mutations for each position in the HCDR3 are shown on the y-axis with the height of each letter corresponding to ROSETTA's preference for that mutation (A). Each mutation that was predicted to benefit fitness is plotted by position. The more preferred mutations correspond to a more negative number. A cutoff threshold of 1.5 ROSETTA energy units is shown as a dashed line to indicate mutations that were considered for experimental characterization after manual inspection (B). Manual inspection of the mutation at position 106 from the wild-type serine (left panel) to asparagine (right panel). The antigen is shown in beige or gray, for the wild-type or mutation, respectively. The surrounding residues are colored in yellow. The HCDR3 loop is colored blue. The asparagine hydrogen bonds to the antigen, which favors binding energy and stabilization of the HCDR3 loop (C). Manual inspection of position 120 is plotted with the HCDR3 loop in light blue, the heavy chain in green, and the light chain in cyan. The glycans are shown in stick representation. The designed asparagine compared to the wild-type tyrosine makes inter-HCDR3 stabilizing hydrogen bonds as well as additional bonds with the glycan (C).

candidate to either be further characterized, or not. I based this decision on an expression cutoff of at least 300 μg/L and a binding of at least 1 OD in ELISA at maximal concentration. For some antibodies that did not express well but showed some binding activity, I chose to carry on to further experimentation. The results are summarized in Table III.5 with 35 variants to be expressed at a larger scale. For these, I choose either a 300 mL expression (10-fold expansion of volume) or a 1L (33-fold expansion of volume) expression based on initial expression levels. These antibodies were purified and carried on to biophysical characterization and neutralization studies. With the exception of cluster I, there was at least one candidate from every cluster that would be further characterized (table III.5).

Cluster	Total Candidates	Wild-Type Expressed[1]	Mutants Expressed[2]	Wild-Type Bound[3]	Mutants Bound[4]
B	6	+	5/5	++	3/5
C	5	++	3/4	-	2/3
D	7	+++	6/6	+++	4/6
E	7	-	6/6	N/A	3/6
F	6	-	2/5	N/A	2/2
G	8	+++	7/7	-	4/7
H	7	+++	4/6	+++	4/4
I	6	-	0/5	N/A	0/0
J	6	+++	4/5	+++	1/4
IG	26	+	23/25	-	8/23
Total	**84**	**7/10**	**60/74**	**4/7**	**31/60**

Table III.5: Expression and binding statistics. Each cluster is shown with its total number of candidates that we attempted to validate expression and binding studies. We record if the wild-type sequence expressed and bound, as well as the number of mutant sequences that expressed and bound.

Wild-type expression[1] : +–> 300 μg/L, ++ – > 1 mg/L, +++ – > 5 mg/L

Number of expressed mutants[2] . Positive if they expressed > 300 μg/L

Wildtype binding[3] : + –> 1 OD, ++ – > 2O D, +++ – > 3OD

Mutants bound[4] : Positive if OD > 1.0

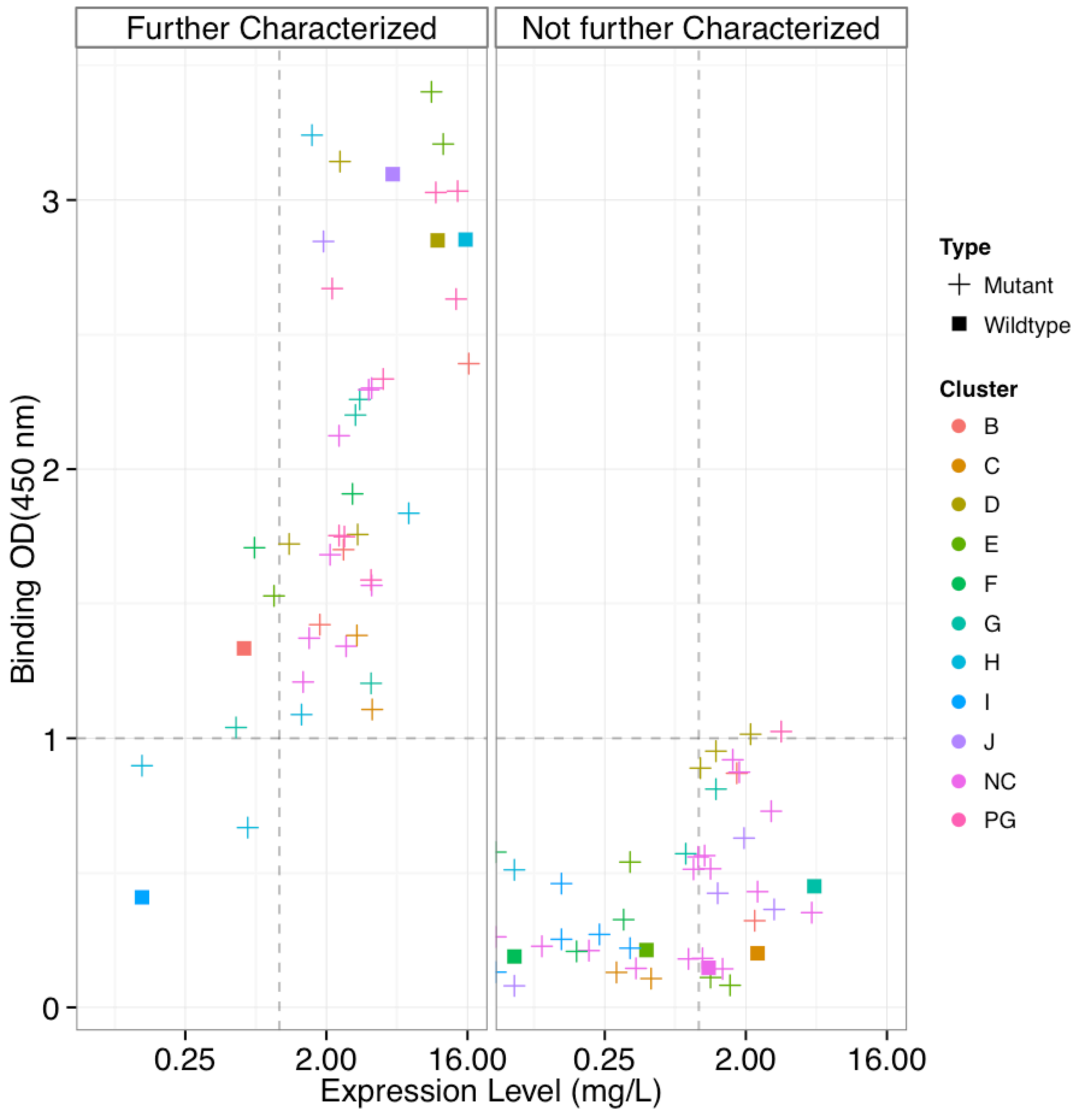


Figure III.15: Expression and binding of 84 variants. Supernatants were tested for binding for all 84 variants and plotted on the y-axis. The x-axis is the level of expression for each variant. Plus symbols denote wild-type sequence, squares are mutants. I color by cluster (NC is the non-clustered or independent group). Only variants that bound tightly and expressed well were considered for upscaled expression and further characterization.

III.10 Biophysical Characterization of PG9-Mimics

I chose to use 8 gp120 monomers based on the ability of PG9 to bind them using my ELISA experimental conditions. These monomers contained 5 clade B variants (BaL.01, SC422661.8, 6535.3, RHPA4259.7, and TRJO4551.58), 2 clade C variants (CAP45.2.00.G3 and ZM109F.PB) and 1 laboratory adapted SHIV chimera (HXBc2P3.2). I found that PG9 bound these 8 candidate monomers based on previous literature (McLellan et al., 2011) and my own pilot studies (chapter IV). I coated each plate in a 384-well format with gp120 monomer and tested binding to the purified 35 variants I analyzed based on my expression/binding sieve analysis in the previous section. Due to replication of these experiments, and consequently the amount of protein used, 4 variants were lost, leaving a total of 31 variants analyzed. I calculated effective concentration at half-maximal binding (EC_{50}) in ELISA of each variant against 8 gp120 monomers. I started my dilution at a very high concentration in order to detect weakly binding antibodies. I set my cutoff for binding at EC_{50} less than 100 μg/mL.

Not surprisingly, most variants did not bind with the same breadth and potency of PG9. In fact, 16 of the variants characterized did not reach my threshold of binding for any of the monomers tested (figures III.16, III.17). However, I found at least one gp120 monomer that bound fifteen of my variants with an EC_{50} less than 100 μg/mL. I find the easiest monomers to bind were BaL.01, ZM109F.PB, and CAP45.2.00.G3. My broadest antibody was VU43171_6MUT from cluster C, which bound 7 out of 8 gp120 monomers (figures III.16, III.17). My most avid antibody was VU28693_5MUT from cluster E, which bound ZM109 and BaL.01 at a concentration of 1 μg/mL. For comparison, PG9 binds ZM109 or BaL.01 with an EC_{50} of 0.03 μg/mL or 0.06 μg/mL, respectively. I also found a completely wild-type antibody, that is, an antibody with no design modifications from ROSETTA, which bound with an EC_{50} to BaL.01 or ZM109 at 3.7 μg/mL or 3.6 μg/mL, respectively.

III.11 Neutralization of HIV by HIV-Naïve Donor Antibodies by PG9-Mimics

From the literature, I know that changes in the HCDR3 loop sequence can dramatically alter specificity (Pancera et al., 2010; Pejchal et al., 2010; Pancera et al., 2013). For instance, making an antibody variant PG9 HCDR3 and a PG16 backbone was able to neutralize additional viruses, while one with PG16 HCDR3 and PG9 backbone also was able to pick

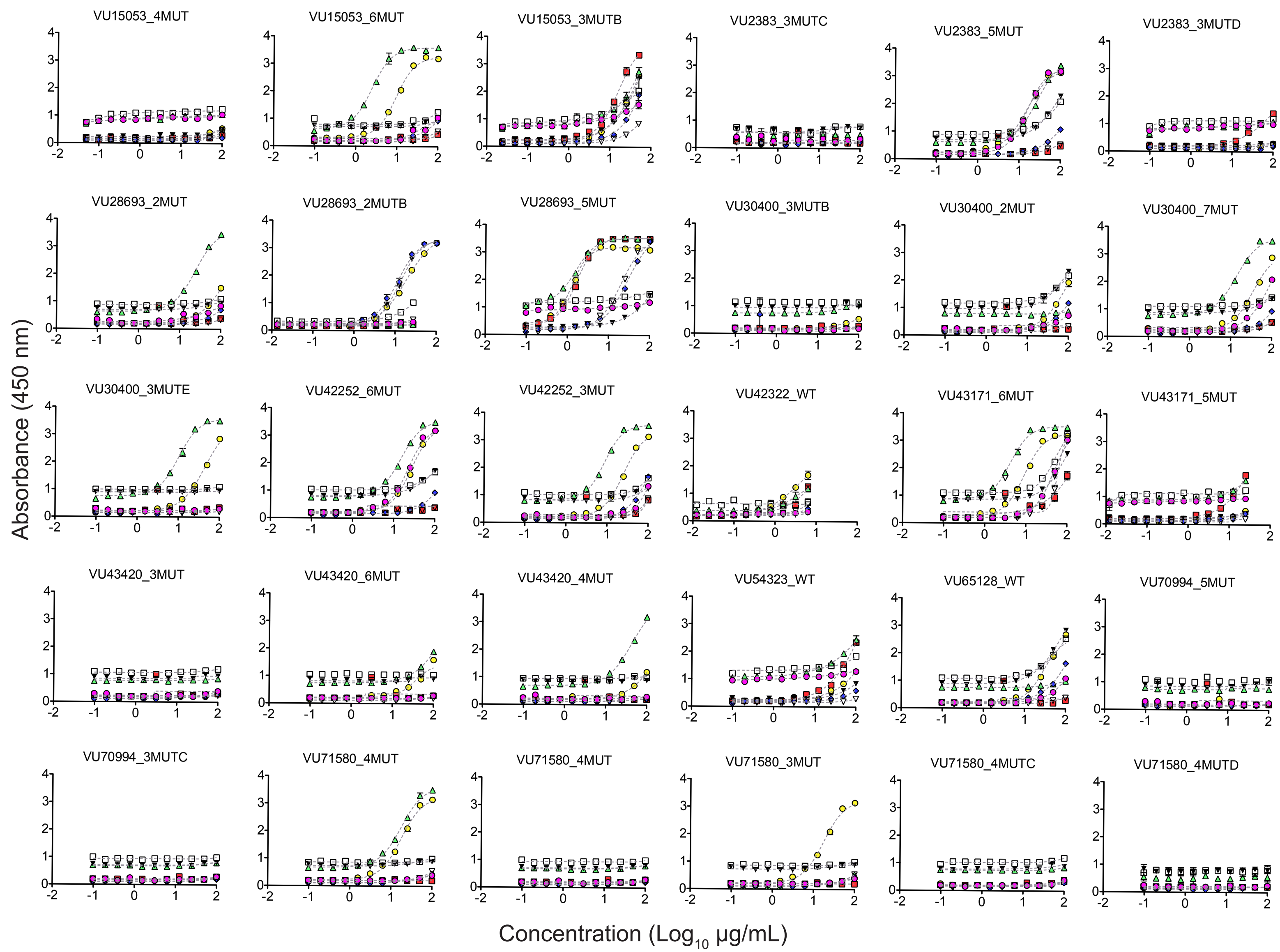


Figure III.16: Expressed candidate ELISA binding. The y-axis is the absorbance for each ELISA binding curve, x-axis is the log of concentration. Each panel represents a different antibody out of the 31 variants tested. Different monomers are represented by symbols. PG9 wild-type (PG9*wt*) is shown at the bottom right as a reference.

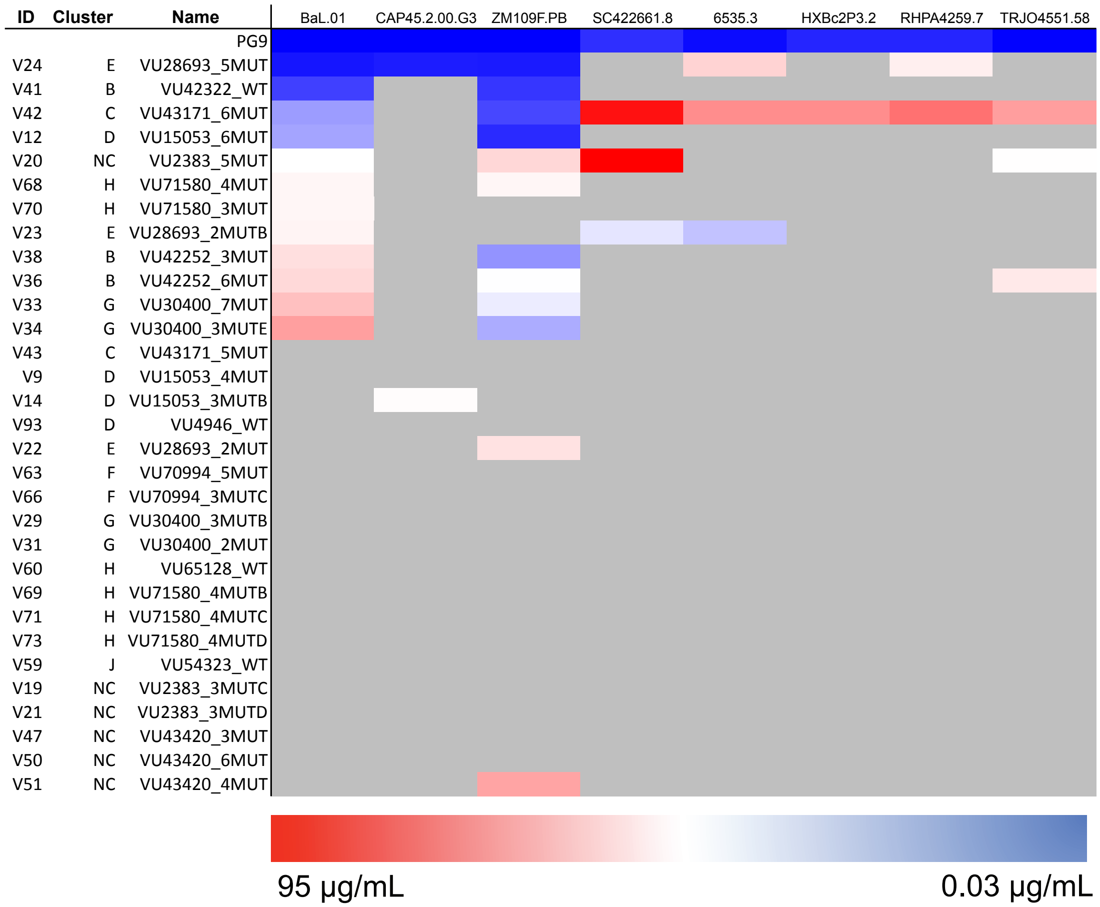


Figure III.17: Heat-map of ELISA binding to gp120 monomers. Each cell represents the effective concentration at half-maximal binding for each variant bound to 8 representative viruses. The red-white-blue scale is a gradient from the maximum EC_{50} , to the minimal, 95 μg/mL to 0.03 μg/mL, respectively. I show the ID, a unique identifier, the cluster each antibody falls in, and the name of the antibody that indicates how many mutations from the wild-type sequence were present. The antibodies are sorted by their strength of binding to BaL.01.

up additional breadth (Pejchal et al., 2010). The exact molecular mechanisms that allow PG9 to bind most gp120 monomers and PG16 weakly bind gp120 monomers are currently unknown(McLellan et al., 2011). Both PG9 and PG16 also do not bind gp120 monomers there are able to neutralize indicating their preference for a trimer specific conformation. It was because of this observation that I decided to test 31 out of my 35 HCDR3 variants for neutralization against pseudoviruses that present native trimer (Montefiori, 2005). I sent our antibody variants to a collaborator, Dr. David Montefiori, who designed and performed multiple neutralization assays to test recombinant HIV variants (DeCamp et al., 2014).

It would stand to reason that the relatively large EC_{50} observed in the binding studies would result in equally large IC_{50} values from the neutralization assays. It was because of this observation that I decided to start the assay concentrations of 100 μg/mL for the neutralization screen. This decision made my neutralization screen protein-limited and at the time of this writing, I could only test activity to two viruses, RHPA and RHPA.N160A. From figure III.17, it is evident that these viruses are difficult to neutralize and we are currently pursuing studies of much more neutralization susceptible viral strains from the literature and figure III.17. Regardless, VU30400_7MUT from group G neutralized RHPA at 49.3 μg/mL, while PG9 neutralized at 6.22 μg/mL. Both mAbs lost the ability to neutralize RHPA.N160A, a knockout mutation for this class of antibody that removes the large glycan at position 160, indicating its functional mimicry of PG9 (Doores et al., 2010).

III.12 Analysis of Mutations with ROSETTA

Lastly, I wanted to analyze the binding antibodies for sequence conservation to see if there were any trends. Indeed, figure III.18 shows antibodies that bound BaL.01 with an EC_{50} less than 55 μg/mL. The neck of the HCDR3 loop showed great sequence conservation for residues 95-97 with a sequence of amino acids VRE or VRD conserved sequence, and residues 118-125 with a YYYYMDV motif, the wild-type sequence of an unmutated J_H6 gene. These trends were observed for long HCDR3 class of antibodies in recent clustering studies (North et al., 2011). I was more interested in conserved sequence elements that are predicted to be at the antigen interface. For those residues, there was great sequence divergence and only a few conserved elements were observed. I observed an aromatic residue at position 104, an asparagine, glycine, and tyrosine and for positions 106-108. These sequences fall in a critical hairpin loop that is necessary to make the crucial turn in order to align the beta-sheets necessary for making contact with the antigen. Indeed, ROSETTA preferred very few residues at this position due to the non-canonical torsions that were adopted.

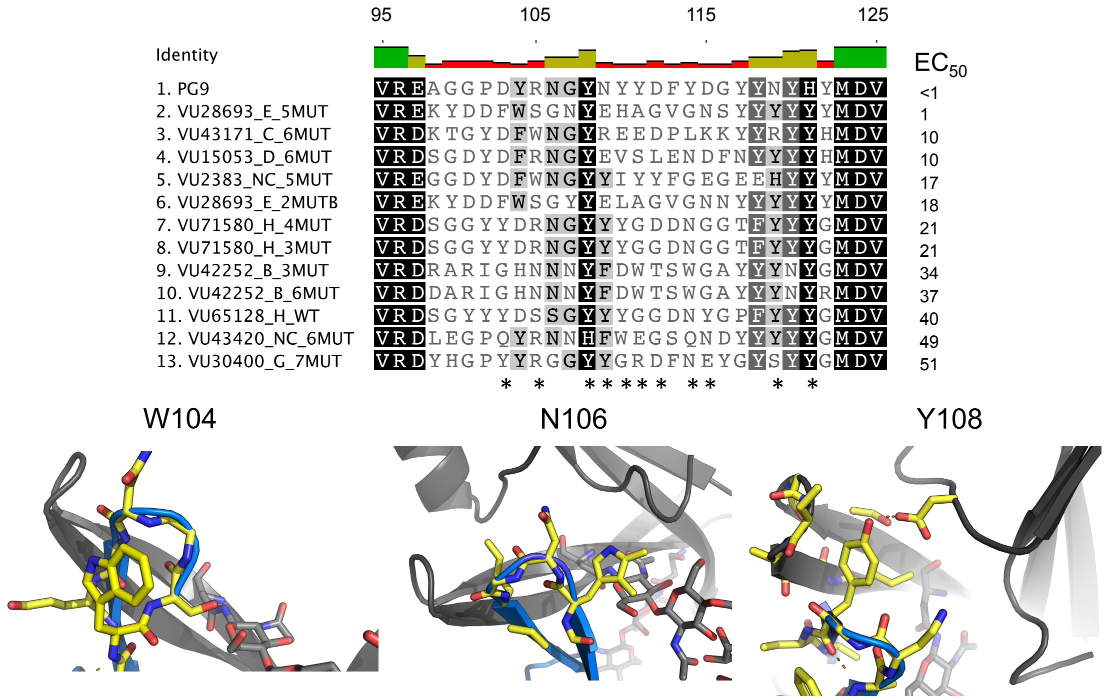


Figure III.18: Mutational analysis of HIV-naïve binding. HCDR3 sequences are shown for variants that had EC$_{50}$ values less than 55μg/mL and ranked 1-13 according to binding. PG9 bound at the lowest concentration and served as a reference. Conservation is shown as an identity plot above the sequence alignment. Sequences are colored white-black according to their consensus conservation. The asterisks below the sequence alignment show contact residues according to the native crystal structure. Representative structures of conserved elements including an aromatic at position 104, an asparagine at position 106, and a tyrosine at position 108 are shown with the antigen in black, the HCDR3 in blue, and the interacting residues shown in yellow stick conformation. The aromatic residue buries a hydrophobic residue against the loop creating $\pi-\pi$ stacking against the nitrogen backbone. N106 was preferred at the position where torsion was necessary to adopt the loop beta-turn. A tyrosine at position 108 hydrogen bonds with a glutamate on the antigen face stabilizing the structure.

Other than those sequence elements, I saw great sequence diversity, particularly in the contact residues predicted from the crystal structure and homology models (figure III.18).

III.13 Conclusions and Future Directions

A protective vaccine against HIV-1 will likely elicit broadly neutralizing serum antibody response (Mascola et al., 1999; Burton et al., 2012; Hessell et al., 2009b, 2010, 2009a;

Hessell and Haigwood, 2012). There are a limited number of neutralizing targets for these bNAbs, which include the CD4-binding site, the V3-loop, the V1/V2-loop, the membrane proximal region (MPER), and the outer N332 glycans (Walker et al., 2010). Here, I aimed to target the V1/V2 loop due to the ubiquity of patients infected with HIV to target this region with neutralizing antibodies (Walker et al., 2010; Gray et al., 2009; Lynch et al., 2011; Georgiev et al., 2013). As discussed in the introduction, the RV144 trial, the only HIV vaccine trial to date to show modest efficacy, showed that the principal correlate of protection was the elicitation of V1/V2 binding mAbs and selective pressure on the V2 region of HIV Env (Rolland et al., 2012; Haynes et al., 2012a).

Recently, a genetic pathway for development of V1/V2 binding antibodies with long HCDR3s has been elucidated for potent bNabs (Doria-Rose et al., 2014). A patient from the CAPRISA cohort was studied longitudinally since initial HIV infection, and the researchers found a modestly potent neutralizing antibody at week 58 post-infection by an antibody with an HCDR3 length of 35 amino acids. In parallel the researchers had also been taking PBMC samples at various time points throughout infection to find the genetic pathway for the development of these neutralizing antibodies. Using pyrosequencing, they found an unmutated antibody at week 30-38 without V_H or V_L gene mutations. Longitudinal sequencing analysis showed this antibody mutates away from the unmutated common ancestor (UCA) and picking up potency, with a total of 14 mutations in the HCDR3 regions (40% mutation) at 58 weeks, but relatively is relatively germline in the other regions (16%). The studies showed 54% mutations from the UCA in the HCDR3 with a neutralization breadth up to 47%. This study showed the tremendous range of sequence diversity that can converge onto one epitope while maximizing breadth and potency. This study is exceptional in its design but does leave some unanswered questions. Firstly, the investigators only derive their antibodies from V_H3-30 gene encoded sequences. This approach leaves out a tremendous amount of potential recombinations that could become neutralizers. This study also focused on one UCA phylogeny and patient. Although the UCA was said to be available at the original time of recombination, there is no evidence that this antibody is present in completely HIV-naïve individuals as it was detected at 30-38 weeks post-infection. This finding is not a true HIV-naïve study as it attempts to characterize the developmental pathways of V1/V2 binding antibodies after infection.

I interrogated the long HCDR3 repertoire prior to infection rather than post-infection. I used 64 different donors, maximizing my sequence pool and diversity. I also used a deeper sequencing method to get the depth necessary to characterize such a broad repertoire. My elegant combination of computational design with bioinformatically-driven heuristics allowed me to interrogate the tremendous sequence diversity of the HIV-naïve antibody donor repertoire. I aimed to answer two simple questions: do HIV-naïve donors possess long HCDR3 sequences that potentially bind and neutralize HIV? If not, will a minimal number of mutations allow them to bind V1/V2 epitopes and neutralize HIV?

The approach to answering these questions involved a four-part strategy that married computational and experimental methods in order to investigate the HIV-naïve donor repertoire. First, I used deep sequencing to accumulate a vast pool of HCDR3 sequences. I then used bioinformatics analysis with new algorithms to determine 30 amino acid length length HCDR3 sequences. Using ROSETTA, I used these sequences to establish a heuristic that would let us rapidly evaluate 30 amino acid length HCDR3 sequences for their ability to form a PG9-type loop. This allowed us to trim down my vast sequence pool to a manageable number of HCDR3 sequences to study experimentally. ROSETTADESIGN allowed me to simulate the process of somatic mutation by applying minimal designs to my sequences in order to enhance potency and breadth.

I experimentally characterized 84 variants, a combination of 10 clusters returned from the computational predictions, and 74 combinations of mutations predicted to enhance binding. Of those, I trimmed the number down to 31 due to a lack of expression or binding on initial screening. Of the 31 antibodies, I performed ELISA experiments on 8 representative gp120 monomers finding that a total of thirteen mAb including two wild-type sequences had an EC_{50} less than 50 μg/mL, well within my tolerance to be considered a binding antibody. For my neutralization work, one antibody variant neutralized a Tier-2 virus with a only 7-fold lower potency than PG9. I expect that many more of my variants will neutralize under optimal experimental conditions. Neutralizing a Tier-2 virus with an antibody from an HIV-naïve donor is ambitious for my studies of HIV-naïve binders and I expect that a Tier-1 virus, may be potentially easier to neutralize.

My work has several implications for vaccine design as it demonstrates that the repertoire of multiple HIV-naïve donors contains antibodies with long HCD3s that have the ability to bind gp120. This finding demonstrates how close an HIV-naïve donor is to actually mimicking a mAb that is known to be a broad and potent V1/V2 binding antibody. It was long thought that a repeated vaccination schedule that would gradually induce the necessary somatic mutations would be needed to recapitulate the broad and potent antibodies that bind the CD4-binding site. Here I show that an average HIV-naïve donor is closer to producing a broadly neutralizing mAb than initially hypothesized (Mikell et al., 2011). This potential paradigm shift in vaccine design would aim to prime for these B-cells with long HCDR3s and then boost for specificity, offering protection from HIV-1 challenge.

CHAPTER IV

Redesign of A Long HCDR3 Antibody

IV.1 Introduction

Recent studies described the isolation of a number of human monoclonal antibodies (mAbs) with broad and potent neutralizing activity, many of which exhibit unusual features (Bonsignori et al., 2011; McLellan et al., 2011; Walker et al., 2009, 2011). As discussed in the chapter I, broadly neutralizing antibodies to HIV generally contain high levels of somatic mutations or exceptionally long HCDR3 lengths. The V2/V3 neutralizing class of anti-HIV antibodies which includes PG9, PG16, CH01, CH04, PGT141 and PGT145 all have a long heavy chain complementarity determining region 3 (HCDR3) and possess unique structural elements that interact with complex protein and glycan features reaching past a large bulk of complex and high mannose glycans to interact with a short segment termed strand-C5. These antibodies share similar neutralization sensitivity including glycan knockouts and strand-C point mutations that interact with interface residues (Doria-Rose et al., 2012; Doores et al., 2010). For PG9 and its clonally related sibling PG16, crystal structures have been solved in complex with V1/V2 gp120 showing that these antibodies both engage the epitope with their HCDR3 loop in a similar ways with the exception of glycan interactions (Pancera et al., 2013). While PG16 prefers hybrid type glycans at position N173 (HIV variant HXBc2 numbering), PG9 has little dependence.

IV.1.1 Experimental Rationale

As an extension of my work in chapter III that considers antibodies with exceptionally long HCDR3s, I chose to pursue a redesign study of the broadly neutralizing antibody PG9. This allowed me to ask relatively simple questions that may have broad and far-reaching implications for antibody and vaccine design. Is the native sequence of PG9 optimal for binding and neutralization potency? PG9 and PG16 converge on structure and binding modes but they are encoded by different sequences. Therefore, I hypothesized that the HCDR3 loop of PG9 could be redesigned to achieve improved affinity of binding, increased potency, and breadth of neutralization for diverse HIV strains.

There has already been precedent for chimeric antibodies of PG9/PG16 where a motif from PG16 responsible for the recognition of complex type glycans was transposed onto PG9 in order to increase potency and breadth. This approach allowed PG9, which initially had no preference for complex type glycans at position 173 (HIV variant HxBC2), to bind those glycan types with stronger affinity, while retaining PG9s ability to bind high mannose type antibodies. This chimeric antibody extended the breadth of PG9 with a small subset of mutations on the light chain CDR3 loop termed PG9-RSH (Pancera et al., 2013).

In addition, NIH45-46, a broadly neutralizing mAb that shows structural mimicry for CD4 and closely resembles VRC01, was mutated by one amino acid in the HCDR2 loop (Scheid et al., 2011; Diskin et al., 2011). The mutation was not designed computationally; rather, the investigators aligned the bound structure of NIH45-46 with CD4 and observed that CD4 had a hydrophobic burial of a phenylalanine residue at position 54 (figure IV.1 A). This interaction was recapitulated in another antibody, VRC03, with a tryptophan residue at position 54. The wild-type amino acid G54 did not fully recapitulate the CD4 interaction as it left a large gap between the gp120 outer domain and the HCDR2 (figure IV.1 A). The investigators predicted that a mutation to either a hydrophobic residue, such as the phenylalanine of CD4, or the tryptophan used by VRC03 would increase potency by increasing the mimicry of CD4. Indeed, a mutation to tryptophan at position 54 (NIH45-46$_{\mathrm{G54W}}$) increased neutralization potency for a majority of the viral strains tested and up to 2,000-fold for one of the strains (figure IV.1 B,C).

Using the high-throughput sequencing data obtained in chapter III, I predicted I could map the energy landscape of the HCDR3 structure using the ROSETTA scoring function. That is, we sought to look at amino acid sequences at all positions of the HCDR3 and determine their overall level of fitness for each position. Does PG9 contain the optimal

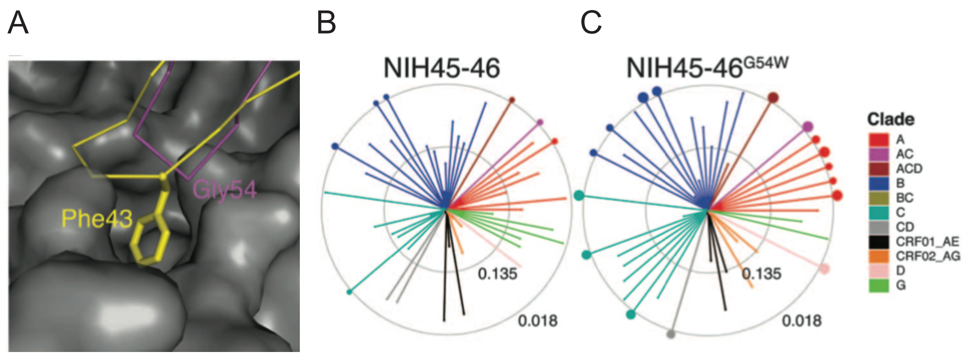


Figure IV.1: Rational design of NIH45-46 to increase neutralization potency. The gp120 bridging sheet is shown as a surface representation with CD4 shown in yellow and NIH45-46 shown in purple (A). Spider plots showing the neutralization profile for NIH45-46 and point mutant NIH45-46G54W are shown. The length of the line corresponds inversely with the IC_{50} value. Each circle represents a ten-fold change in IC_{50} (B,C). Figure adapted from (Diskin et al., 2011)

sequence for the HCDR3 loop? If I didn't see a complete recovery of PG9 sequence, I could then predict that other sequence combinations or point mutations exists that enhance fitness of the HCDR3 and may increase breadth and potency to HIV. Again, these mutations would then be carried over to the laboratory to be tested experimentally with binding and neutralization assays.

IV.2 Mapping the Energy Landscape of PG9

I retrieved the atomic resolution structure of the complex of mAb PG9 with the CAP45.2.00.G3 variant V1/V2 scaffold from the Protein Data Bank (PDB ID:3U4E) (McLellan et al., 2011). A large number of naturally-occurring 30-amino-acid (30 AA) length HCDR3 antibody sequences was identified in antibody gene repertoires from high-throughput sequencing of antibody amplicons from RNA of B-cells of HIV-negative donors. Retrieval of this dataset is discussed at great length in chapter III. A heat map of amino acid occurrences is displayed in figure IV.2 A for 30-length HCDR3s. Diversity among the repertoires is seen for all positions 98-118. The sequence conservation at the $5'$ and $3'$ ends of the HCDR3 sequences, 96-98 and 118-125, respectively, are due to the ARD motif that make up the

$5'$ end of a canonical neck of a long HCDR3 loop or the J_H6 template sequence, which was seen in a majority of long HCDR3 sequences (North et al., 2011; Briney et al., 2012). Between these two stretches of sequence conservation, I observed large sequence diversity. Glycine, tyrosine and serine are generally tolerated at all positions, while proline, lysine and methionine are found less frequently between positions 99-117 (figure IV.2A). This phenomenon is well established in loop unstructured regions connecting beta-sheets in antibodies (Minuchehr and Goliaei, 2005; De et al., 2005). This propensity for a structure to tolerate a diverse set of amino acid sequences was the focus of the current study. The idea is that there is tremendous sequence space to be explored in 30-length HCDR3s that may further enhance breadth and specificity.

My methods are described fully in the appendix VI.3, but use the same general protocol as described in Chapter III. I used the software suite ROSETTA to determine the ability of diverse 30-AA HCDR3 sequences to tolerate the structure of the hammerhead configuration of the HCDR3 of PG9 by threading 4,000 naturally-occurring unique sequences over the PG9 HCDR3 structure. Once the sequences are threaded, I scored them by evaluating the ROSETTA scoring function for each position. The contribution to the total score of each position (the sum of all scoring terms of ROSETTA), which can be thought of as thermal stability, and the contribution to the binding energy (the total score in complex subtracted from the total score in a separated interface) of each position were evaluated. These scores are best viewed as heat maps (figure IV.2 B,C). These two metrics, total score and binding energy, can be summed to determine what I define as the mutational "fitness". In this way, I trimmed the tremendous sequence space as viewed in the heat map in figure IV.2 A, to a focused sequence space containing mutations that advantageously affect either thermal stability or binding energy. These metrics are shown in the heat maps of figure IV.2 A-B, offer an advantage structure-based metrics in exploring design. As expected, PG9 itself scored as the most fit sequence for a majority of amino acid positions for the HCDR3 (dots plotted in figure IV.2).

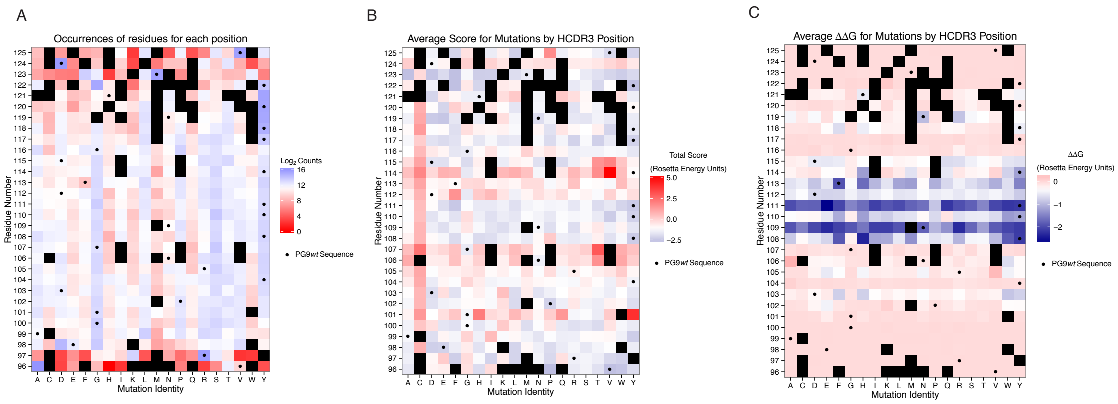


Figure IV.2: Amino acid usage and energy landscape of PG9. Mutation identity is plotted on the x-axis with each of the positions in the 30-length HCDR3 on the y-axis. The usage of each amino acid is shown in a $\log_2$ blue-red scale counted from 26,422 HCDR3 sequences (A). 4,000 randomly selected sequences were chosen and their individual score from the ROSETTA energy function is shown on a blue-white scale (B). The contribution of the same 4,000 randomly selected sequences contribution to binding energy is shown on a blue-white scale (C). For A-C, the PG9 native sequence is shown as a dot.

IV.3 Redesign of PG9

Rather than pick the amino acids that had the best fitness for each position, I allowed a complete redesign of the PG9 HCDR3 loop using ROSETTADESIGN. My reasons for choosing this method rather than a simple matrix lookup generated in the previous section were two fold:

1. Complete redesign can account for cooperative mutations. Consider position 99 that has a wild-type alanine for PG9. My heat maps for the energy landscape predicted that there are many more favorable mutations I could make including an aspartic acid, asparagine or tyrosine. However, I are unaware if the new mutations have the potential to be cooperative. That is, do the aspartic acid, asparagine, or tyrosine require neighboring mutations to be fully stable? The complete redesign allowed me to account for cooperative mutations while recapitulating the energy landscape predicted in section IV.2.

2. Using a combination of filters and movers based on my specific design goals, I pre-

vented ROSETTA from designing amino acids too far away from the original PG9 sequence, position, and structure. I can also tell the ROSETTA scoring function to optimize for binding energy, thermal stability, or a combination thereof. This information would be lost on a matrix lookup (Fleishman et al., 2011a; Kaufmann et al., 2010; Kuhlman and Baker, 2000).

Again, the full design protocol is detailed in the in the Appendix (Chapter VI.3) and follows the same basic structure as the redesign of sequences detailed in Chapter III. I designed 1,000 decoys allowing small docking perturbations and minimal backbone movement. I filtered the design to optimize for binding energy. That is, only obtain sequences if they were better than binding energy of PG9*wt*.

The easiest way to view the sequences returned from the PG9 redesign is with a sequence logo representation (figure IV.3 A). The x-axis is the PG9*wt* sequence while the height of the letters at each position, measured in bit, measure ROSETTA's preference for that amino acid given the nature of the design challenge. As expected from the observations from the energy landscape (figure 4.2), the original PG9 sequence was returned for a majority of the positions, considering the evolutionary sequence bias of the PG9 structure (nature optimizes sequences for the PG9 structure). Regardless, any time an amino acid was returned in 10% or more of the models, I further inspected the design fitness of the mutation.

I measured the design fitness as a sum of the difference in total energy from wild-type sequence and the difference in binding energy from the wild-type sequence ($\Delta\Delta G$ + total score). For some of the positions, multiple amino acids were suggested by ROSETTA rather than the wild-type amino acid sequence (figure IV.3 A,B). For most positions, the design fitness was negligible, falling above the noise threshold (figure IV.3 B, dashed-line). However design at antibody amino acid positions 104, 109, 115, 120, and 123 (PDB numbering) suggested alternative amino acids that were predicted to benefit HCDR3 fitness for the antibody-V1/V2 interaction (figure IV.3 B).

I wanted to make sure that each of these mutations made intuitive sense upon examination of the structure. I viewed each mutation in context and compared it to the wild-type amino acid. My aim was to determine if this mutation was an artifact and to confirm that it was non-cooperative with other mutations, that is, the mutation enhances fitness alone and not in cooperation with many other mutations made to the sequence. This is because I wanted to retain as much of the PG9*wt* sequence as possible. These visual inspections are shown in figure IV.3 B in the table, with a full justification given. If a mutation was found to be non-cooperative, and still enhanced fitness, it was considered for experimental characterization (green squares), N109Y and D115N met these criteria. I also considered N109L and a cooperative mutation A99S and Y120N as they showed strong stabilization through inter-HCDR3 loop hydrogen bonding. I also made one combinatorial mutation that included the double cooperative mutant A99S-Y120N and two single mutations D115N and N109L. This variant is simply referred to as PG9_4MUT (figure IV.3C). I did not pursue further evaluation of designs that appeared to compromise the structural integrity of the HCDR3 loop by visual inspection.

IV.4 Experimental Characterization of PG9 Variants

For the five mutational variants of PG9, I used a similar cloning strategy as described in Chapter III that takes advantage of unique cloning sites between the HCDR3 $5'$ and $3'$ ends. Each of the variants expressed well, and protein concentration was not a limiting factor. I began by testing all the variants against a 15 antigen mix of gp120 Env proteins to qualitatively measure binding. In this preliminary study, the double mutant, A98S-Y120N produced a significantly lower signal than that of PG9, and this variant was not considered further.

PG9*wt* does bind to some gp120 monomers (although PG9*wt* also can neutralize HIV variants for which it does not bind to monomer). I used a panel of representative gp120 monomers from HIV clades B and C to perform screening for binding of PG9 variants to

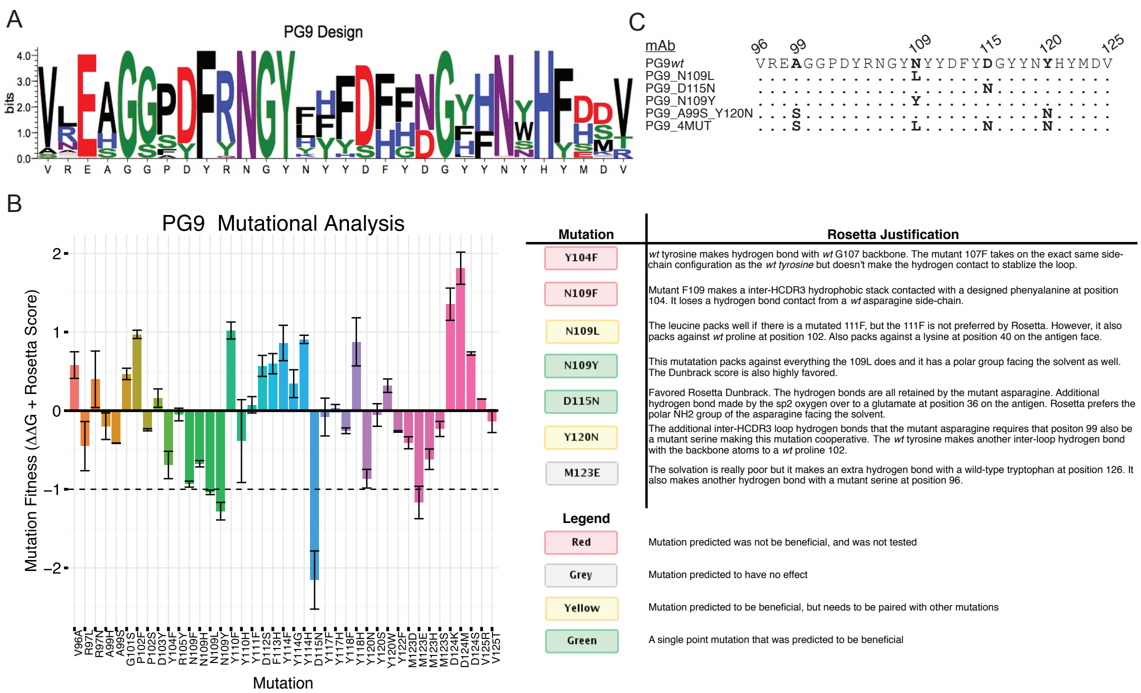


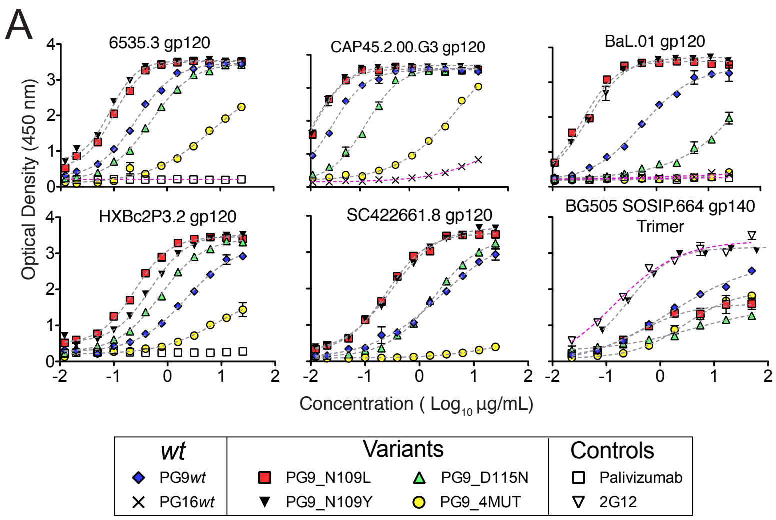
Figure IV.3: Redesign of PG9 HCDR3. For 1,000 designed models, the sequences returned are best viewed as a sequence logo. The x-axis is the PG9*wt* sequence while the y-axis represents the preference, measured in bit, of the amino acid identities, identified by the height of the letter (A). For sequences that were returned greater than 10% of the time, I manually inspected their fitness as a measure of binding energy and thermal stability (y-axis). Some positions had more than one amino acid favored and are grouped by color. ROSETTA noise is plotted as a dashed line at -1 REU. The more negative a mutation is, the more it is beneficial to the PG9 complex. Each mutation is visually inspected and justified (table). They are either a single point mutations that benefits, a cooperative mutation that benefits, no change in fitness, and detriment to fitness, as green, yellow, grey, and red, respectively (B). The final mutations that are chosen to be carried out experimentally are three point mutations and two combinations thereof (C).

Env (Li et al., 2006, 2005). The results were in good agreement with previous studies of the binding of PG9*wt* to gp120 monomers (McLellan et al., 2011). For these PG9 variants, I calculated half-maximal effective concentration (EC_{50}) values. For each gp120 monomer tested, the PG9 variants N109L and N109Y exhibited 2.3-14.2 fold stronger binding than did PG9*wt* (figure IV.4), while PG9 variant D115N exhibited comparable binding energies to PG9*wt*. PG9_4MUT exhibited 2-100 fold reduced binding. This finding is most likely due to the A98S-Y120N mutation that I had previously determined as deleterious.

I also determined the EC_{50} for binding of these PG9 variants to a recombinant form of native gp140 trimer that is recognized by PG9, termed BG505-SOSIP.66419-21. In these assays, both PG9 variants N109L and N109Y exhibited 3.5- or 5.9-fold stronger binding respectively than PG9*wt*. In addition to the stronger binding affinity, the variant N109Y bound to trimer with a complete sinusoidal curve and a strong maximum signal mimicking the binding profile of the glycan-specific mAb 2G12, which is optimal for binding to the trimer (Sanders et al., 2013) (figure IV.4A). The extreme change in maximal signal is intriguing, and may suggest changes in valency of P9 to the trimer although this has not been confirmed (Julien et al., 2013).

I next tested the panel of redesigned PG9 variants and PG9*wt* for neutralizing activity against a panel of viruses displaying PG9-susceptible or -resistant HIV Env molecules, using a TZM-bl neutralization assay (Montefiori, 2009). The PG9 variant N109Y exhibited increased neutralization potency for all viruses tested, including viral variants for which PG9*wt* did not have activity (i.e., had neutralization concentration >33 μg/mL) (figure IV.4B). Remarkably, PG9 variant N109Y neutralized at 3.72 μg/mL an HIV strain with the N160A mutation that removes the glycan at that position that is required for binding of PG9*wt*6. The PG9 variant N109L also exhibited an increase in potency against HIV strains compared to PG9*wt*, although not at the same level as the PG9 variant N109Y. In all assays tested, N109Y and N109L consistently had enhanced breadth and potency.

The magnitude of the improvement to neutralization was modest in some cases, but the improvement was consistent over a wide variety of HIV strains and showed significant p-values ($p < 0.05$) for 10 out of the 15 virues for PG9_N109Y and 4 out of the 15 viruses for N109L (table IV.1). Using a meta-analysis for all p-values tested gave a p-value of $5.44\ x\ 10^{-15}$ and $7.36\ x\ 10^{-4}$ indicating a strong statistical significance observed for the increase in potency of neutralization for PG9_N109Y and PG9_N109L respectively. I also found a decrease in potency to be statistically significant for 8 out of the 15 viruses tested for PG9_D115N and a combined p-value of $2.64\ x\ 10^{-8}$.



B

Binding (EC50, µg/mL)

HIV Strain	PG9 *wt*	PG9 N109L	PG9 N109Y	PG9 D115N	PG9 4MUT
BaL.01	0.54	0.04	0.05	>	>
CAAN5342.A2	1.20	0.47	0.44	1.06	2.33
CAP45.2.00.G3	0.03	0.01	0.01	0.13	9.45
HXBc2P3.2	2.41	0.25	0.49	0.74	7.78
RHPA4259.7	0.66	0.13	0.16	1.32	15.70
SC422661.8	2.48	0.25	0.31	1.89	24.87
TRJO4551.58	0.17	0.05	0.07	0.39	>
ZM109F.PB	0.02	0.01	0.00	0.04	2.27
6535.3	0.26	0.10	0.09	0.49	7.74
BG505 SOSIP.664	1.48	0.42	0.25	3.63	2.86
7165	>	>	>	>	>
AC10.0.29	>	>	>	>	>
QH0692.42	>	>	>	>	>
PVO.4	>	>	>	>	>
REJO4541.67	>	>	>	>	>
THRO4156.18	>	>	>	>	>
TRO.11	>	>	>	>	>
YU2	>	>	>	>	>

Neutralization (IC50, µg/mL)

HIV Strain	PG9 *wt*	PG9 N109L	PG9 N109Y	PG9 D115N	PG9 4MUT
WITO4160	0.03	0.01	0.01	0.05	0.07
X1632_S2_B10	0.47	0.24	0.13	4.20	19.20
X2278_C2_B6	0.07	0.01	0.02	0.35	2.70
Ce703010217	0.02	0.01	0.01	0.03	0.04
BG505.N332	0.04	0.02	0.02	0.11	0.28
CNE55	1.13	1.80	0.65	8.60	25.20
SC422661.8	0.11	0.04	0.05	1.40	11.10
Du422.1	1.90	0.15	0.44	2.80	5.00
TH023.6	1.80	0.30	0.20	13.70	>
R2184.c04	0.28	0.49	0.11	17.20	>
TRO.11	>	11.03	3.10	>	>
SC22.3C2	>	>	12.90	>	>
Ce1086_B2	>	>	18.60	>	>
398_F1_F5_20	>	>	10.80	>	>
TH023.6 /N160A.5	>	>	3.72	>	>

Figure IV.4: Representative binding curves are shown with the optical density at 450 nm shown on the y-axis plotted against the $\log_{10}$ concentration in μg/mL on the x-axis (A). All EC_{50} values were calculated from the curves like the ones shown in (A) as well as the neutralization IC_{50} against a 15 virus panel (B) *p*-values for Neutralization t-tests

HIV Strain	PG9_N109Y vs. PG9*wt*	PG9_N109L vs. PG9*wt*	PG9_D115N vs. PG9*wt*
WITO4160	0.0003	0.2238	0.0027
X1632_S2_B10	0.0800	0.2999	0.0432
X2278_C2_B6	0.0049	0.0156	0.0605
Ce703010217	0.0655	0.0671	0.0148
BG505.N332	0.1371	0.2350	0.0284
CNE55	0.2997	0.3538	0.0021
SC422661.8	0.0049	0.0087	0.0011
Du422.1	0.0048	0.0192	0.1132
TH023.6	0.0462	0.0913	0.0161
R2184.c04	0.0088	0.3112	0.0015
TRO.11	0.0047	0.0102	ND
SC22.3C2	0.8890	0.2221	0.4173
Ce1086_B2	0.0430	ND	ND
398_F1_F5_20	0.0002	ND	ND
TH023.6 /N160A.5	0.0001	ND	ND
Fishers Combined Probability	5.44E-15	7.36E-04	2.64E-08

Legend	
(light olive)	< .05
(light green)	< .01
(dark green)	< .001

Table IV.1: Statistical tests for neutralization breadth of PG9 variants. The IC_{50} values between each neutralization assay for PG9_N109Y, PG9_N109L, and PG9_D115N were compared with a student's non-parametric t-test against the IC_{50} value for PG9. They are shown as p-values for each viral variant. A total p-value for each antibody is shown as a Fisher's combined probability.

IV.5 Models to Corroborate Experimental Outcome

I sought to develop a predictive model to determine the molecular basis for the increased potency and breadth of these PG9 variants using the ROSETTA scoring function. I generated three different mutants D115N, N109L and N109Y, which were compared to that of PG9*wt* using the ROSETTA scoring function (figure IV.5 A). I analyzed the top 25 models for each of the scoring metrics shown. my scoring metrics were binding contribution for the HCDR3, binding for the full complex, total score for the bound and unbound structure ($\Delta\Delta G$). For each metric calculated, I observed statistically significant improvements in

HCDR3 stabilization for N109L or N109Y ($p < 0.01$ or $p < 0.001$, respectively), but none of my other metrics.

Upon examination of the predicted structures of the top scoring models, I found the antibody position 109 was located on an antiparallel beta-sheet at the apical tip of the HCDR3 forming a hydrophobic pocket near the interface of the antigen and apical tip (figure IV.5 B). The pocket is formed by antibody residues Y104, Y110 and P102 of the antibody heavy chain (PDB numbering). In addition, D167 on the antigen face makes contact with this position. Examination of the structure revealed that the bulk of the hydrophobic amino acid at this position of the pocket contributes to stabilization of the preferred structure of HCDR3. The small hydrophobic bulk of the asparagine fills the pocket, but as the bulk increased to a leucine and then a tyrosine, the predictive model suggested a further stabilization of the HCDR3 loop. In addition, the polar group on the end of the designed tyrosine at position 109 points into solvent space, recapitulating the effect of the polar head of the original asparagine PG9*wt*.

It is important to note that since I calculate the stabilization as the total energy of the HCDR3 loop, it can be dissected into the individual scoring terms given by ROSETTA. These are both described in the chapters I and the appendix VI. I have broken the total score down for the HCDR3 loop in figure IV.6. There is little deviation for most of the scoring terms, however, both the attractive force term, the solvation term in the ROSETTA scoring function are improved for the N109L and N109Y mutations. In addition the N109Y shows a favored π-π term. These interactions are accounted for in the model as the N109 position is between a large hydrophobic bulk. In addition, position N109Y also achieves a more favorable π-π stacking interaction with residue Y110 compared to PG9*wt* as the aromatic ring of the designed tyrosine can stack with position 110.

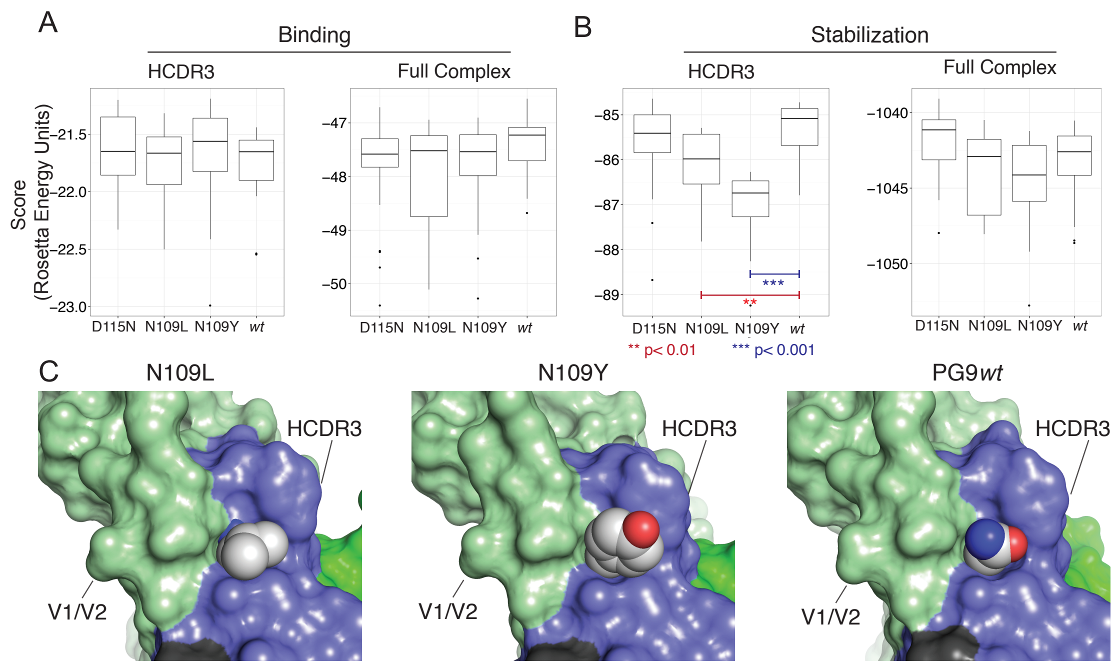


Figure IV.5: The top 25 models for each binding metric were analyzed. The x-axis indicates each of the variants and the y-axis is the ROSETTA energy units. The metrics are decomposed by binding energy (A) and thermal stabilization (B). Just the HCDR3 is considered (left) or the entire complex (right). Surface representation of position 109. Green is the antigen labeled V1/V2, blue is the HCDR3 loop, dark green is the N160 glycan. Each mutation of interest is shown as a sphere representation that is adjusted to fit the ROSETTA atom radius. Spheres are colored by atom type with oxygen in red, nitrogen blue, and carbon in grey. Hydrogens are removed for clarity.

IV.6 Discussion

These results have important implications for antibody and vaccine design. The studies reveal the power of ROSETTA computational modeling to design antibodies with improved function using structural predictions. Remarkably, the improvements in neutralizing potency and breadth observed here for PG9 variants were achieved not by altering interface residues, but rather by increased stability of HCDR3 loops discovered using a holistic model to determine stability of an antigen-antibody complex. This finding is consistent with recent mutagenesis experiments showing that non-contact residues are essential for antigen recognition by many broadly neutralizing antibodies to HIV (Klein et al., 2013). Non-contact residues in antibody frameworks contribute to high affinity binding by facil-

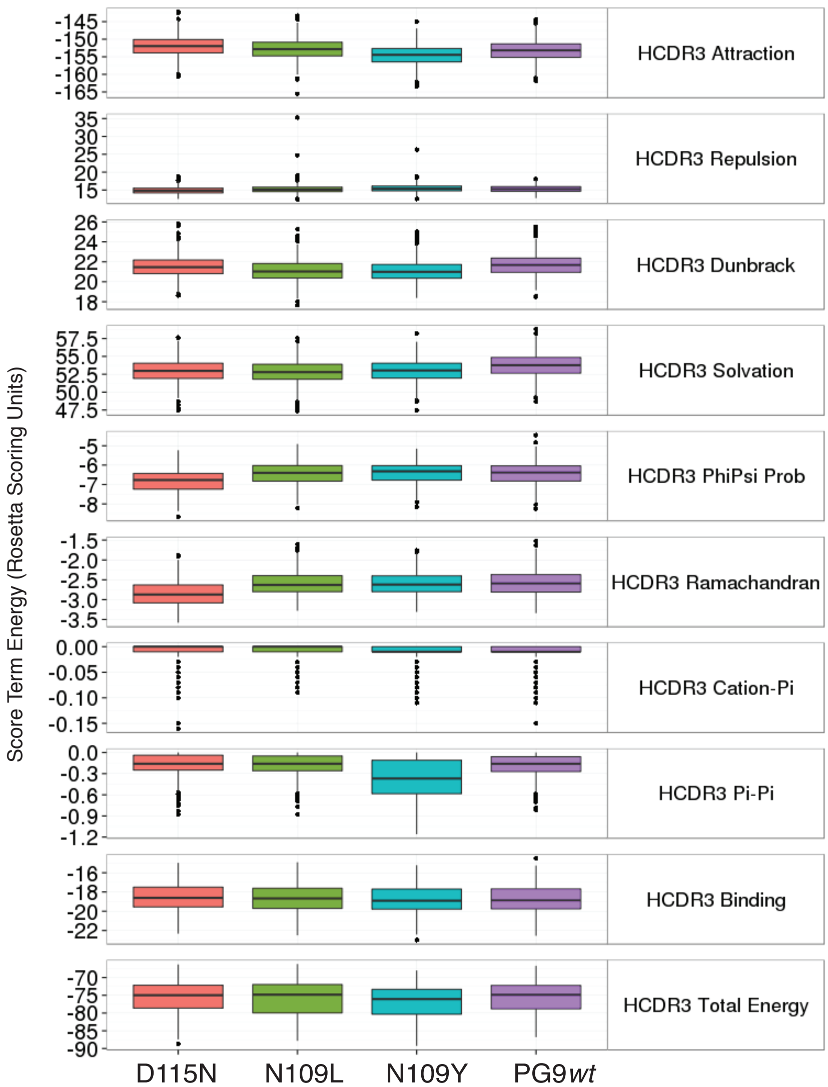


Figure IV.6: Decomposed scoring terms for PG9 variants. The contribution of individual scoring terms to the total energy score for the HCDR3 loop for each mutation. The predictive model used 1000 simulations for each variant. Each scoring term for ROSETTAis shown in the y-axis panel. The y-axis value is the score for that energy term.

itating formation and stability of a pre-configured, low energy binding site (Willis et al., 2013; Manivel et al., 2000; Marlow et al., 2010; Wedemayer et al., 1997; Schmidt et al., 2013). Optimally configured binding sites form ordered paratopes that pay a smaller entropic penalty upon forming antibody-antigen complexes (Marlow et al., 2010).

This work shows the efficiency of combining ROSETTADESIGN computational experiments with expert knowledge and wet laboratory validation experiments. For a HCDR3 loop of 30-AAlength, 600 single point mutants are possible, and the number of variants with more than one mutation is enormous. From this large potential set of mutated antibodies, ROSETTADESIGN identified a focused panel of candidate PG9 variants, from which a small subset was considered favorable, and two of five experimentally tested variants exhibited enhanced potency and breadth of neutralization. The computational experiments provided tremendous enrichment for variants with improved binding, but as expected was not completely accurate. For example, although the model suggested that D115N would have the greatest increase in fitness (figure 4.3) this variant was not improved in activity. The negative result is important, as ROSETTA often predicts design failures, and their exploration is fundamental in improving the ROSETTA algorithm and scoring function for the most accurate representations of experimental observation. This computational-experimental feedback has been instrumental to my work and will be the target of my future directions.

With the combination of high-throughput sequencing, rapid threading, and experimental feedback, I complete a robust bioinformatics pipeline that can rapidly test antibodies for improvement based solely on their *in silico* predictions. The results here suggest that there probably is a diversity of antibodies with long and structured HCDR3s that fit the PG9 topology in nature with HIV neutralizing activity that have yet-to-be discovered. I hypothesize this conclusion from three parts of evidence:

1. Examination of the energy landscape of PG9 suggests that there are mutations that are predicted to be better suited for the PG9 topology.

2. PG9 and PG16 diverge in sequence but converge on a structural topology and have approximately identical specificities and potencies (McLellan et al., 2011; Pejchal et al., 2010; Pancera et al., 2010).

3. I have discovered point mutations in PG9 that enhance breadth and specificity.

These yet to be discovered antibodies may possess higher HIV inhibitory activity and breadth than the antibodies that are currently in hand. Additional antibody exploration efforts may be worthwhile to identify antibodies of interest with which to design epitope mimetic vaccines, as has been successfully recently implemented (Correia et al., 2014; Jardine et al., 2013).

IV.7 Conclusions and Future Directions

Two observations may be critical to explore in this current work. The maximum signal difference in the binding assay for the BG505-SOSIP trimer between my variants and the PG9*wt* is worth further exploration. Julien and colleagues observed that PG9 recognizes the trimer asymmetrically (Julien et al., 2013). There are three epitopes displayed on the apical tip of the gp120 trimer, yet PG9 only has a 1:3 valency of binding to HIV Env. The molecular mechanism for this trimeric preference is unavailable due to the high resolution of the structure reported in the study, but the model suggests that PG9 may interact with the adjacent N160 glycan. Could the mutation cause a change in valency in binding? This change would explain to the maximal signal change in my binding assay. Future experiments could replicate the study performed by Julien *et al.* either with high-resolution gel-filtration or isothermal titration calorimetry.

Another observation is the glycan independence of the binding of the N109Y variant. It was originally thought that the binding of any V1/V2 binding antibody would depend on glycans at position N160 and N156/N1706 considering they not only block the recessed C-strand epitope, but make considerable binding contributions for both PG16 and PG9 (McLellan et al., 2011; Pancera et al., 2013). It was demonstrated that these glycans are needed for recognition and specificity, as mutational experiments completely abrogated neutralization. I was able to replicate that finding, however, my variants still neutralized HIV glycan knockout viruses, albeit, at much lower potency (3.52 μg/mL). It is worth replicating this "glycan independence" with many more viral species that have been mutagenized to knockout glycans. I have already begun to pursue this aim.

Finally, I can attempt to repeat the application of this entire technology to mutagenesis of PG9's sibling, PG16. I already have the mutational candidates, and the antibody cDNAs are being synthesized at the time of this writing. It is important to keep in mind that PG16 specific to the trimeric-Env, so variants can only be tested with neutralization experiments or if I synthesize stable trimer (Sanders et al., 2013).

CHAPTER V

Conclusions and Future Directions

V.1 Chapter II - Multi-state design and polyspecificity

I expand on the limitations of ROSETTADESIGN by looking at the imperfect agreement between what we expect ROSETTADESIGN to recovery and what was actually observed. I bin these into three categories, mature sequence bias, evolutionary sequence bias, and incomplete ensemble bias. Section II.7 accounts for mature sequence bias, which may be the most abstract concept. When a residue is "important" for most of the antibody-antigen complexes, it should be recovered often. However if it only is important in a few complexes, then any given residue can meet the design challenges for the complexes in which it is not important.

Section II.8 refers to evolutionary sequence bias. Here we may see imperfect agreement with the design and actual antibody sequence simply because the antibody has not achieved maximal potency. ROSETTADESIGN may have selected better amino acid sequences for the epitope. Given enough time to for complete antibody evolution, the sequence may have matched. This is founded on the principle that ROSETTADESIGN gives the best sequence for any design challenge.

Section II.9.1 gives the limits of computation including the finite ensemble size that states there are not enough PDBs to compensate for "structural space" that ROSETTADESIGN samples. This section also details some of the more classic limitations including the limits of the ROSETTA scoring function and limitations of the sampling algorithms. In the future this may be ameliorated with more structures being deposited to the PDB, and improvements to the scoring function including explicit solvent modeling.

The power of chapter II is in the future directions. While we use multi-state design to interrogate the flexibility of the germline repertoire, the end goal was always to design

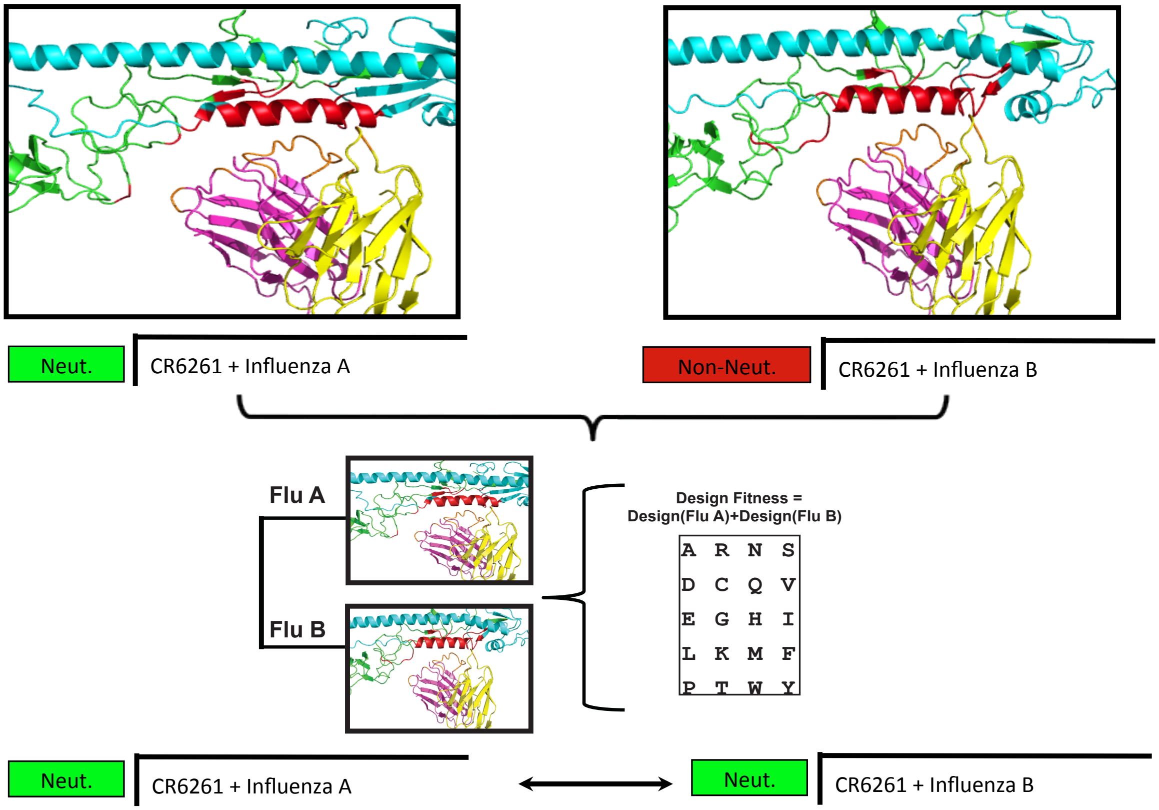


Figure V.1: Multi-state design of broadly neutralizing influenza antibodies. Initially, the antibody CR6261 only binds and neutralizes only Influenza A subtypes. Using multi-state design I plan to make minimal mutations at the interface that allow binding to Influenza type B while retaining binding to Influenza type A. This principle allows me to design in cross-specificity.

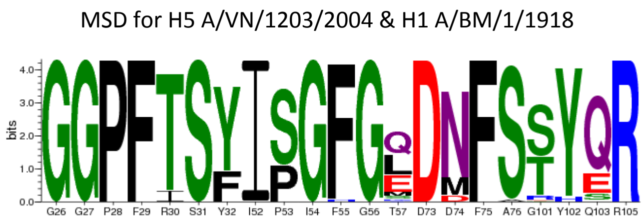
cross-binding antibodies. Many pathogens which evade traditional vaccination do so by evolving multiple serotypes or antigenic variants that encompass a tremendous sequence space. HIV, Influenza and dengue virus (DENV) are such examples of antibodies that use antigenic variation as a means to evade a broadly protective immunogenic response. It is to these pathogens that a broad cross-reactive response will be critical (Corti and Lanzavecchia, 2013; Lanzavecchia and Sallusto, 2009; Corti et al., 2011; Simonelli et al., 2013). Multi-state design will be invaluable in the design of these antibodies. Consider V.1, here we take a known cross-reactive group 1 antibody against influenza known as CR6261 (Throsby et al., 2008) which does not bind to Influenza type B.

I even took this idea into the lab looking at special cases for influenza antibody. I considered the antibody CR6261 as it was bound to the stem portion of influenza (Corti et al., 2011). At the stem region, their is less conformational diversity. I hypothesized that this epitope loses neutralization affinity due to point mutations at the interface rather than large conformational shifts that are evident in the head region. This would be easier for ROSETTADESIGN to recover. First, I wanted to create a proof-of-concept by seeing if we can enhance specificity to already known binders. I chose H1/South Carolina/1918 and H5/Vietnam/2004 pandemic strains as both had a crystal structure bound to CR6261. Using multi-state design, I told ROSETTADESIGN to enhance binding to both variants. Figure V.2 shows the design process. The sequence logo in (A) shows the amino acids preferred at the interface. For (B), we analyzed the fitness of each mutation as either having beneficial or deleterious effects. If it enhanced fitness for both H1 and H5, we made the mutations in the laboratory. They indeed did enhance binding to both variants compared to wild-type CR6261 (figure V.2 C.).

Of course this proof-of-concept can extend well beyond influenza. My plan was to use this to get a serotype specific antibody to bind (and therefore, potentially neutralize) cross-serotype, cross-group, and cross influenza sub-type. However, many labs are trying to design cross-reactive antibodies. One advantage of multi-state design is that if you fine-tune specificity you do not have to lose specificity to your original epitope. For example, in a paper by Simonelli *et al.* (Simonelli et al., 2013), they used computational modeling along with NMR restraints to place give a molecular definition of an anti-DENV against all four DENV serotypes. They then used their knowledge of rationale design to fine tune specificity to each serotype individually. However, upon making one mutation specific to a given serotype often abrogated binding to the others. Therefore it is possible to enhance breadth at the cost of specificity. This type of problem is ideal for multi-state design as it can fine tune specificity without predicted loss of binding to your original epitope.

In related work, structural viral homologs could also be used in multi-state design. For example, an antibody found in chronically exposed repository infected patients bound

A



C

Fab Construct	Binding Affinity K_D (nM)	
	1918 H1 HA	VN/1203 H5 HA
CR6261	534	2.23
CR6261 R30I	2.14	42.6
CR6261D74N	1.75	6.76
CR6261 R30I/ D74N	0.38	0.59

B

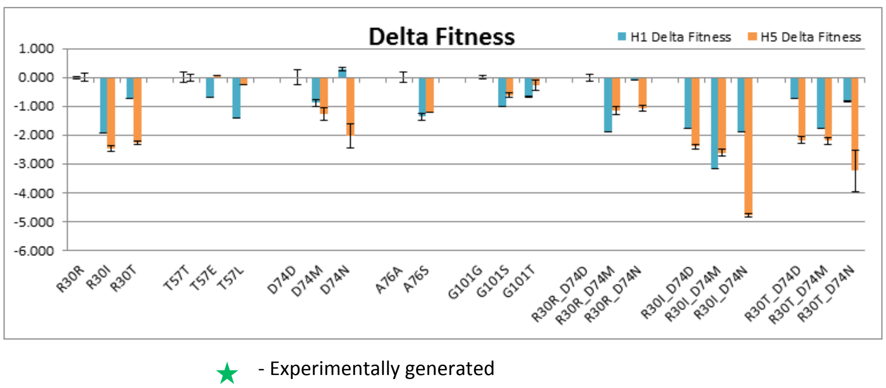


Figure V.2: Preliminary for MSD proof-of-concept. A sequence logo for all positions considered for redesign at the interface. Higher letters indicate ROSETTA's proclivity for a certain amino acid sequence (A). Fitness analysis of each position is a sum of stability and binding energy. Lower energies indicate better fitness. Ideally, both H1 and H5 would have decreased fitness energy. Stars indicate mutations that were made experimentally (B). The binding affinities from the suggested mutations from ROSETTA. The double mutant gives the best binding affinity (C).

and neutralized four different paramyxoviruses, human respiratory syncytial virus (HRSV), human metapneumovirus (HMPV), bovine RSV (BRSV) and pneumonia virus of mice (PVM) (Corti et al., 2013). This antibody was found by screening a common structural homolog (prefusion protein F) against anti-HRSV screened from over 200 donors to narrow down their search. I hypothesize that this type of structural information can be used in multi-state design to make *de novo* designed cross-neutralizers much the way nature selected cross-neutralizers from these patients.

Finally, there is much use for this algorithm's fitness function. While I describe antibody design in the context of *designing for* a certain antigen, it may be beneficial to *design* *against*. I can modify the scoring function in such a way where we select mutations that will *design for* one antigen, and *design against* others. This allows any fine tuning of specificity changes that are needed against antibodies while not compromising the structural integrity of the immunoglobulin fold. I can imagine this will be an invaluable tool for designing against antigens that are found to be related to autoimmunity.

V.2 Chapter III - Broadly neutralizing antibodies from HIV-naïve donor repertoires

In chapter III, I used antibody design to interrogate the HIV-naïve repertoire to answer a simple question for the paradigms of HIV vaccinology. How close is the naïve donor repertoire to eliciting neutralizing antibodies? I used the principles guided by both reverse and forward vaccinology (Burton et al., 2012). Reverse vaccinology principles require that the broadly neutralizing antibodies are first characterized from chronically infected patients. I used the the V1/V2 binding antibody PG9 that was discovered in a chronically infected African donor (McLellan et al., 2011; Walker et al., 2009). This is where I introduced a new paradigm into vaccinology. Rather than used structure based immunogen design from the PG9 epitope which has been characterized, I instead investigated the healthy donor repertoire. In my early work here, I had helped discover long HCDR3s in healthy donors, and it was with this information I wanted to pursue this question.

I conclude that although there are some antibodies from the HIV-naïve donor repertoire that are able to mimic PG9 hammerhead like configuration, there are very few from our population pool. Even those that we did find needed some amounts of mutation that honed specificity to make these antibodies truely binding and neutralizing.

There is an absolute limitation to this study, and that is the amount of long-HCDR3 sequences we started with. Out of one-half-billion sequences, we only were able to obtain 26,000 unique 30-length sequences that were viable in our study. That is an incredibly low amount. Along with Andy Fire, Jessica Finn and myself, we have devised new clever experiments that can potentially enrich for long HCDR3s. We have noticed that in this

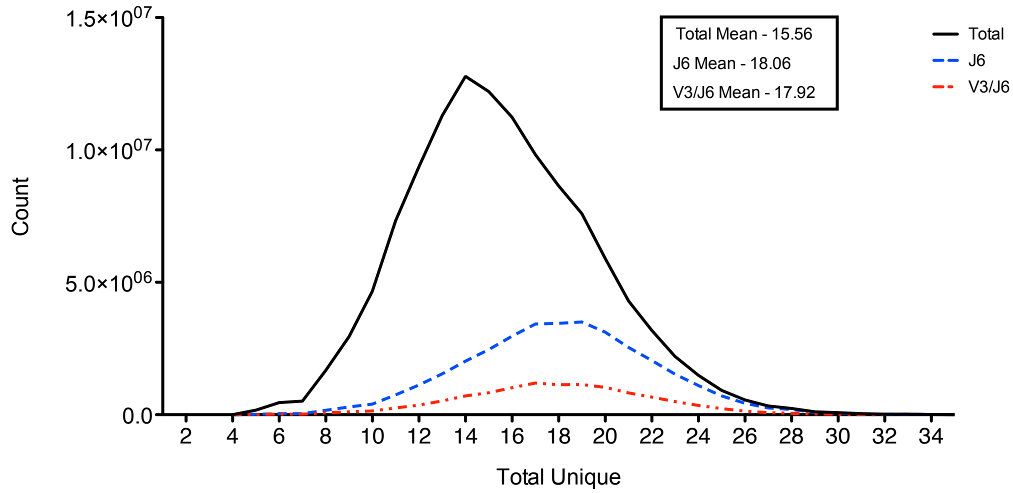


Figure V.3: HCDR3 of J_H6 gene families. A distribution curve for HCDR3 sequences is shown for all VDJ familes (black). When only J_H6 genes are considered, the mean length shifts from 15.56 to 18.06. If only J_H6 and V_H3 gene families are considered the mean shifts down to 17.92.

study and previous studies, that most antibodies with long HCDR3s contain a J_H6 gene segment (Briney et al., 2012). In addition, PG-type antibodies often use V_H3-30 genes. This makes the initial PCR rounds for gene-specific targeting a simple task. Rather than use ambiguous primer-sets for all gene families, I plan to use just specific primers to J_H6 and V_H3-30 for amplification of B-cell mRNA transcripts. I can actually predict that this will enhance the mean HCDR3 length from 15.56 to 18.06 (figure V.3). In addition, new advances in high resolution gel electrophoresis allow single base pair resolution and purification. That would allow only very-long HCDR3 sequences to be purified and enriched from the canonical length background. Using this, I could truly use this population pool in my newly created bioinformatics pipeline to test the HIV-naïvedonor sequences.

One problem that is mentioned is the potential framework bias. Although it has been demonstrated that the HCDR3 loop is responsible for a majority of the contact and by extension the mechanism of neutralization (Pejchal et al., 2010; Pancera et al., 2010), it can't be ruled out that the framework provides some contribution to binding in the LHCDR3 and HCDR2 region (McLellan et al., 2011). However, in recent studies, mutations all the way back to germline have shown that these mutations aren't absolutely necessary necessary to necessitate binding and by extension neutralization (Klein et al., 2013). In chapter II, we find at least fifty percent of the the germline framework may be necessary for binding, even those residues that lie extremely distal to the interface. This leads me to believe in an inherent framework bias in our study. I used PG9 framework as the complete sequence to the rest of the heavy chain and the light chain was unknown (the current technology is limited at the time of writing). But other studies show us that surrounding mutations of the HCDR3 loop allow it to form its native conformation (Wong et al., 2011). I want to know the effects of other germline frameworks on HIV-naïve HCDR3 loops.

I proposed a display technology solution to this problem. If I can isolate 30-length (or longer) HCDR3 sequences from the B cell repertoire, I can engineer cloning sites into them as I have done for the PG9 framework. However, each cloning site will be slightly upstream of the HCDR3 site. In this manner we can use any germline framework to test each HIV-naïve conformation. In this way, we can test inherit framework bias from each germline framework on the healthy HCDR3 sequence. For the amount of HCDR3 sequences I get (roughly estimate 100,000 using the technique described above), I can test this on each V_H gene family, giving 5.2 million different combinations of antibodies I can test and remove framework bias. In addition, I can combine germline light chain repertoires to increase our combinations. I feel like this would be like an incredibly useful experiment to find antibodies from HIV-naïve donors with the absolute minimum amount of mutations. In this way I can get an absolute threshold to describe what a vaccine will need to elicit using a very inexpensive technology.

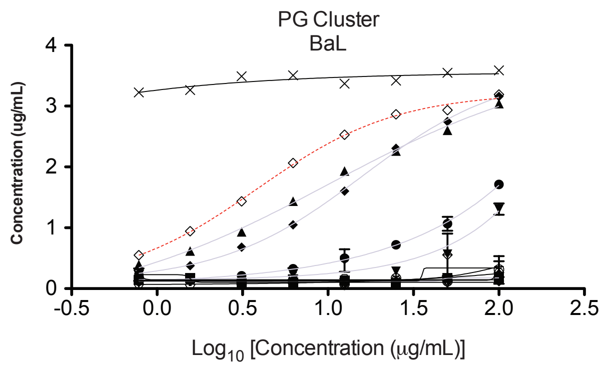
V.3 Chapter IV - Broadly neutralizing antibody redesign

I indicate some of the issues with redesign and highlighted them in the publication that accompanies chapter IV. For instance, ROSETTADESIGN, indicated that D115N would be the most beneficial mutation. When I characterized this mutation experimentally, I found this mutation to actually hinder binding. This is due to some of the limitations of ROSETTA that have been discussed at length, especially in chapter II where I discuss the limits of the ROSETTA scoring function. This is being addressed by a large number of ROSETTA collaborators in the Commons with more accurate representation of hydrogen bondings and explicit solvent models.

The excitement of future endeavors with the redesign of antibodies is the most exciting and probably one of the most straight forward future directions I have. For one, this type of design which we applied to PG9 can be applied to any antibody, whether it be broadly neutralizing or modestly potent. ROSETTA allows us to tailor the scoring function to *design for* stability or binding energy, which we have found to be a necessary component to increasing potency and breadth of contemporary antibodies.

One step I have taken with future directions, is applying the same methodology to PG9s sister antibody PG16. In figure V.4, an initial pilot study using the same redesign methodology has been performed. As is evident, some of the PG16 variants are able to bind BaL gp120 monomers with greater affinity. There is a huge caveat at the current time of this study. These PG16 HCDR3 loops were put on the PG9 backbone with the exception of PG16*wt*. The PG9 framework may be responsible for the antibody binding BaL gp120 monomer. I have begun taking these constructs and putting them on PG16 backgrounds to see if there is difference in PG9/PG16 backbone. If there is, that gives plausibility to framework bias as discussed above.

The future for antibody redesign is staggering, especially in context of evolutionary sequence bias as introduced above. Researchers isolate antibodies in the middle of their evolutionary cycle where they may not be optimized for potency and breadth, only the ability to bind and not necessarily to a threshold. With the redesign principles that we have gained insight into with PG9 redesign, it opens up an exciting new avenues for any antibodies, and aids in understanding the principles that need to be considered for the grand challenge in antibody design, the *de novo* design of a nM binder.



	Variant	EC_{50} (μg/mL)	OD_{max}
PG16 Variants	PG1	Ambiguous	1.71
	VUPG16_1MUT	Ambiguous	0.17
	VUPG16_1MUTB	10.29	3.04
	VUPG16_1MUTC	Ambiguous	1.31
	VUPG16_2MUT	15.91	3.16
Controls	VRC01	Ambiguous	3.59
	RSV/MPV	Ambiguous	0.22
	MPV	Ambiguous	0.32
	RSV	Ambiguous	0.36
	MPX	Ambiguous	0.12
	VAX	Ambiguous	0.27
	DENV	Ambiguous	0.18

Figure V.4: An initial pilot study with PG16 HCDR3 variants on a PG9 backbone. All the variants have a PG9 backbone except for PG16*wt* which may explain the strength in binding. Various negative and positive controls are shown for other viral species.

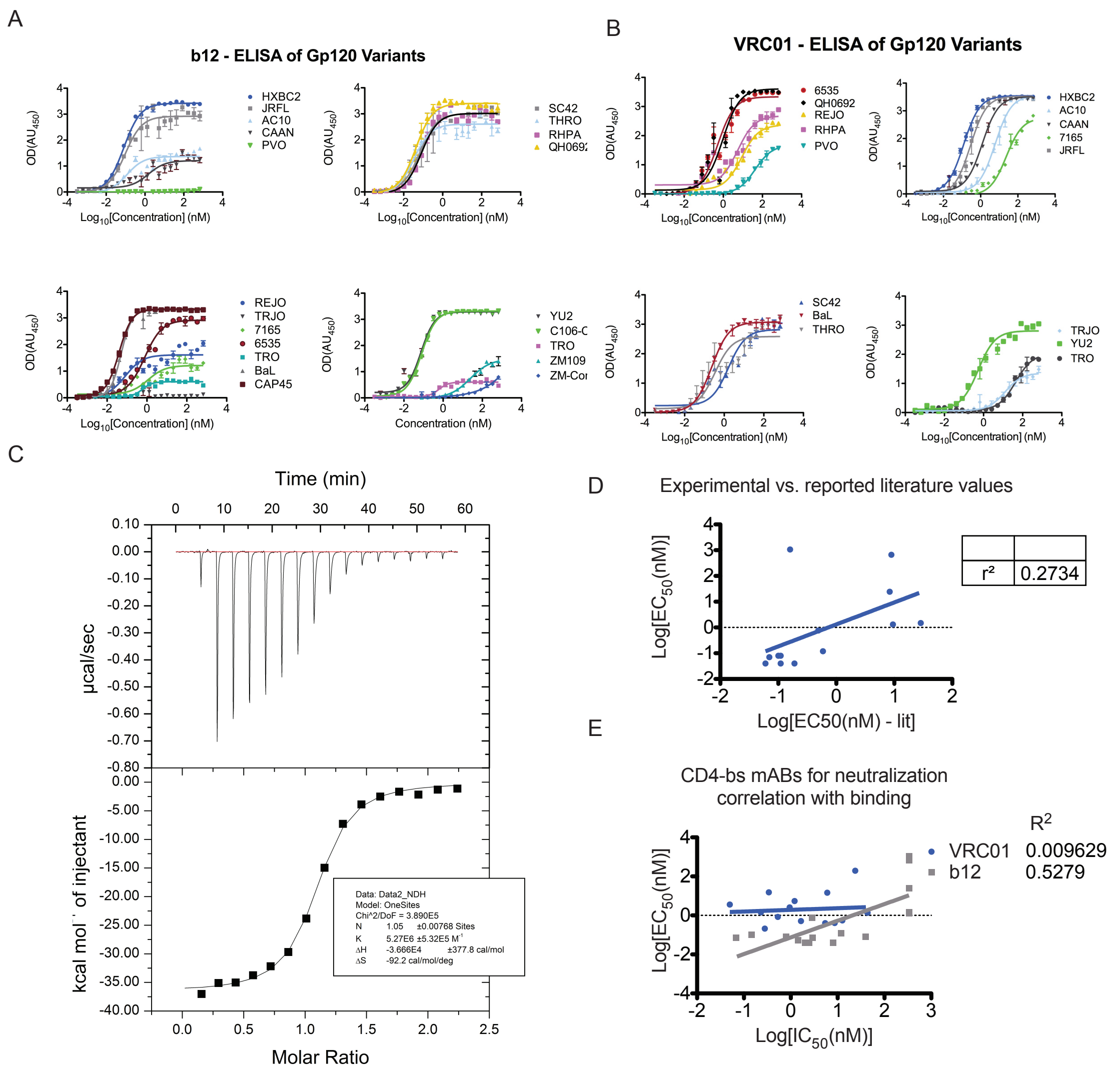
V.4 Other Applications of Antibody Design

My very first specific aim was to establish a correlation between *in silico* binding energy and experimental binding energy. This was the start of a very tedious process of making diverse HIV antigen to test two broadly neutralizing antibodies VRC01, and b12 (Wu et al., 2009; Li et al., 2011). As I began reading the literature, most notably a study done by Kwon and colleagues (Kwon et al., 2012), I noticed that the VRC01 bound gp120 was different from the b12 bound structure where the b12 bound structure is in the pre-CD4 bound state and VRC01 is in the CD4-bound state. I hypothesized that, VRC01 would be entropically limited because it would always need to change the conformation of gp120 in order to bind. However when it does so, it exposes an extremely conserved epitope.

To first test this hypothesis, I wanted a baseline of gp120 binding using ELISAs and EC_{50} as a metric. Figure V.5 panel A and B show these curves for many variants of gp120 for b12 and VRC01, respectively. I then started to use ITC to get a better metrics of how each antibody was binding (figure V.5 C), and I found that VRC01 made many hydrogen bonding contacts, but gave up huge entropic penalties. Much work is needed on this front as an accurate ITC for each variant is absolutely necessary to confirm this hypothesis, but the initial work of creating constructs and protein has been done. It is also important to note that the reported literature values may be different from mine which again underlines the importance of in house experiments.

With the EC_{50} values, I can plot an *in silico* binding energy vs. the experimental binding energy (figure V.5) and find that there is no correlation for VRC01 but good correlation for b12. This result is intriguing as the conformational shift is not accounted for in ROSETTA and there are probably many other states that gp120 samples. A mechanism of resistance has been worked out with brute force (Li et al., 2011), but I would like to confirm *in silico*.

These results are very preliminary and would best be guided by ITC experiments for the rest of the variants, but much of the work has been done.



Name	ΔG (kcal/mol)	ΔH (kcal/mol)	ΔST (kcal/mol)
VRC01 Fab/Bal gp120	-9.2	-36.6	27.4
VRC01 Fab/YU2 (from literature*)	-9.87	-48.6	38.7

Figure V.5: CD4 binding mAbs mechanisms of escape. Binding ELISAS against clade B and C gp120 monomers for CD4 binding antibody b12 (A) and VRC01 (B). Isothermal titration calorimetry of VRC01 antibody against b12 highlighting enthalpic domination (C). Experimental vs. literature binding values show a weak correlation (D). Computational predicted binding energy for homology modeling of VRC01 and gp120 variants and b12 and gp120 variants correlates well for only b12.

CHAPTER VI

Appendix

VI.1 Appendix I - ROSETTA Glossary

All-atom - in the case of sampling, synonymous with fine movements and often including side chain information; also referred to as high-resolution

Benchmark - another word for a test of a method, scoring function, algorithm, etc. by comparing results from the method to accepted methods/models

Binary file - a file in machine-readable language that can be executed *in silico*

BioPython - a set of tools for biological computing written and compatible with Python

Build - to compile the source code so it may be used as a program

Centroid - in ROSETTA centroid mode, side chains are represented as unified spheres centered at the residues center of mass

Cluster center - the geometric center of a cluster, or group, of models

Clustering - grouping models with similar structure together

Comparative model - a protein model where the primary sequence from one protein (target) is placed, or threaded, onto the three dimensional coordinates of a protein of known structure (template)

Cyclic coordinate descent (CCD) - based on robotics, CCD loop closure is used to build loops in ROSETTA by fragment assembly and close loops by decreasing the gap between two termini in three-dimensional space

De novo - from the sequence; also called *ab initio*, with no experimental guidance

Directory - synonymous with a folder, usually contains one or more files or other folders

Distance matrix - a matrix containing the pairwise distances for every point in a set of points

Dunbrack rotamer library - a set of likely side chain conformations for the twenty canonical amino acids based on protein structures in the Protein Data Bank (PDB)

Executable - binary file used to execute the program

Force field/Scoring function/Energy function/Potential - often used interchangeably; a means of assessing the energy of the generated models

Fragment - in ROSETTA folding and loop building, a set of three-dimensional coordinates corresponding to a given amino acid sequence fragment

Database - also called the fragment library, contains all the interchangeable data needed for ROSETTA

Gap - in sequence alignment, a gap is inserted when the sequences are of low homology; usually appear as a dash (-); the gaps form a sequence alignment correspond to areas where loops are built during comparative modeling

GDT/GDT_TS - global distance test (total score); a measure of similarity between two protein structures having the same amino acid sequence; the largest set of residues Cα-atoms in the model structure falling within a defined distance cutoff of their position in the experimental structure

Gradient-based minimization - also known as minimization by steepest descent; in this case, a means of energy minimization in which one takes steps proportional to the negative of the gradient of the function (energy) at the current point

High-resolution - in the case of sampling, synonymous with fine movements and often including side chain information

Homology model - a more specific type of comparative model where the protein sequence of interest (target) is a homolog of the protein of known structure (template)

Interface delta - the interface delta score is defined as the contribution to the total score for which the presence of the ligand is responsible

Kinematic loop closure (KIC) - robotics-inspired loop closure algorithm which analytically determines all mechanically accessible conformations for torsion angles of a peptide chain using polynomial resultants

Knowledge-based - in the case of ROSETTA, based on information obtained from structures found in the PDB

Libraries - in computing, a collection of code and data (classes and functions) used by a piece of software and is often used in software development

Ligand - the part of the structure that binds to a protein to serve some biological purpose

Low-resolution - a somewhat subjective term, in the case of sampling, synonymous with coarse movements of the protein and/or ligand backbone and side chains; the individual atoms of low-resolution structures or models cannot be resolved, or observed

Metropolis criterion - often combined with the Monte Carlo sampling algorithm; allows for generation of an ensemble that represents a probability distribution

Model - in the case of this protocol, a structure generated by ROSETTA; sometimes called a decoy

Monte Carlo sampling - a randomized and repetitive computational sampling method

Mover - a generic class that takes as input a pose and performs some modification on that pose; for example, a mover might take in a pose and rotate every residue

Namespace - in computer science, an abstract container holding a logical grouping of unique identifiers or symbols; in ROSETTA, examples of namespaces are loops, relax, etc.

Native-like - close to the experimentally determined structure; a model that is native-like usually has an RMSD to the experimentally determined structure of < 2Å

Options file - often called a flags file; a file containing ROSETTA options that can be passed to a ROSETTA executable after the @ symbol; can be easier to use than passing ROSETTA options over the command line

Pack/repack - in ROSETTA, side chains are packed/repacked by switching out rotamers and scoring them using the ROSETTA scoring function

Pose - in ROSETTA protocol, a three-dimensional conformation of the ligand, protein, or ligand/protein complex at any given time-point

Python - interpreted, object-oriented, high-level programming language http://www.python.org/

Relax - in ROSETTA, an iterative protocol of side chain repacking and gradient-based minimization; often referred to as full-atom (or all-atom) refinement

Robetta ROSETTAstructure prediction server (http://robetta.bakerlab.org/) freely available to not-for-profit users

RosettaCommons - a group of more than twenty labs that develop the ROSETTA software suite

REU - arbitrary energy units specific to the ROSETTA scoring function

RosettaScripts - also called “the scripter” or ROSETTAXML; an XML-like language that allows for specifying modeling tasks in ROSETTA

Rotamer - rotational conformer of an amino acid or ligand side chain

SCons - a tool for constructing software from its source code http://www.scons.org/

Script - in computer programming, a script is a sequence of instructions that is interpreted or carried out by another program rather than by the computer processor (as a compiled program is)

Source code - human-readable files that are the implementation of the program; are written in C++ in ROSETTA

Target - in comparative, or homology, modeling, the protein for which we are generating a model; the target sequence is the primary sequence of the protein for which we want to make a model

Template - in comparative modeling, the protein of known structure on which the target is threaded

Threading - placing the primary sequence of one protein (target) on the three-dimensional coordinates of a protein of known structure (template) based on a sequence alignment loop building

XML - Extensible Markup Language; in this case, used to write protocols to pass to

VI.2 Appendix II - ROSETTA Scoring Terms

Scoring Term	Explanation of Scoring Term
Attraction	The Van de Walles scoring term to indicate how much attraction residues have on each other
Dunbrack	A statistical probability score indicating how often a side-chain configuration has been seen in the protein data bank (PDB)
Repulsion	The Van De Walles scoring term to indicate how much repulsion residues have on each other
Solvation	How well are hydrophobics packed away from solvent and hydrophillic groups are facing solvent
Ramachandran	A statistical probability of how well ϕ-ψ angles fit into the Ramachandran plot as a function of secondary-structure
Total	A summation of all individual scoring terms to get a total score
ΔΔG	The change in total energy score when residues are moved out of complex
ϕ-ψ Prob	A statistical probability score of how well a side-chain configuration has been seen given a ϕ-ψ angle in the PDB
Cation-π	A score encompassing how the configuration of positive cation at the end of a charged residues interact with pi orbitals
π-π	A score encompassing how two π orbitals interact
HCDR3 Stabilization	The total score of residues only found in the HCDR3
Full Complex Stabilization	The total score of all residues representing a free energy of the model
HCDR3 Binding	The contribution to ΔΔG by residues found in the HCDR3
Full Complex Binding	The ΔΔG for the entire complex

Table VI.1

VI.3 Appendix III - Materials and Methods

VI.3.1 Chapter II - Materials and Methods

VI.3.1.1 Selection of Antigen-Antibody Complexes

Diverse antigen-antibody complexes were collected from the Protein Data Bank (PDB; www.pdb.org) in which antibodies in different complexes were derived from the same predicted heavy chain variable gene segment. Candidate complexes were queried from the protein databank using the IMGT-3D structural query editor for immune system receptors (Kaas et al., 2004). PDB structures were used as design candidates if they met the following criteria: 1) the antibody was encoded by a V_H1-69, V_H3-23, or V_H5-51 gene segment, 2) the structure contained a human immunoglobulin, and 3) the ligand type was a protein complex. The search yielded 10, 8, or 3 antibody-antigen complexes encoded by the heavy chain variable gene segments V_H1-69, V_H3-23, or V_H5-51, respectively. Nature of the antigen and antibody isotype were not considered in the selection as the 21 complexes represent an exhaustive search of the PDB for these gene-segments. The gene segments were aligned using the ClustalW2 multiple sequence alignment algorithm (Larkin et al., 2007). Each input structure was energetically minimized using the ROSETTAscoring function but constrained to PDB input backbone coordinates (Das et al., 2007).

VI.3.1.2 Multi-state Design of Antigen-Antibody Complexes

Three design experiments were performed, one for each of the three germline segments (V_H1-69, V_H3-23, or V_H5-51) using the multi-state design mode of the ROSETTAalgorithm and scoring functions. I adapted a generalized multi-state design protocol that was described in detail previously that perform design on multiple antibody-antigen complexes at once (Leaver-Fay et al., 2011a). Briefly, each computational design experiment computed an optimal sequence predicted to define a low-energy structure. In the multi-state design experiments, an energetic consensus sequence for all of the states was predicted, rather than treating each state as a separate entity. The energy for a given sequence was computed and designated the "design fitness" for all states. The corresponding amino acids were derived from the alignment (*e.g.*, heavy chain amino acid 5 on complex A corresponded to heavy chain amino acid 5 on complex B). The details of the multi-state algorithm is described elsewhere (Leaver-Fay et al., 2011a).

VI.3.1.3 Single-State Design of Antigen-Antibody Complexes

Single-state design was performed using the ROSETTA multi-state application. The algorithm was altered so that only one complex was considered for each of the 10, 8, or 3 design experiments with V_H1-69, V_H3-23, or V_H5-51 complexes, respectively.

VI.3.1.4 Design Analysis of Multiple- or Single-State Design

For each design experiment, 100 independent design trajectories were calculated. Sequence logos then were generated using the Berkley web-logo server (http://weblogo.berkeley.edu/) (Crooks et al., 2004). Information for each sequence logo can be extrapolated as follows extending the work of Schneider *et al.* (Schneider and Stephens, 1990). For each variable position, the probability of seeing each of the 20 naturally encoded amino acids p_i was computed and compared with the background probability p_b = 1/20 = 5%. To quantify the deviation of the observed probability from the background probability I compute the self-information for each of the 20 amino acids as $I_i = p_i \times \log_2(20 \times p_i)$ in 'bit'. If the amino acid occurs as often as expected from the background probability, I_i is zero. Ii becomes larger if the amino acid is over-represented and approaches 4.32 if p_i = 100%. A total bit-score for the sequence design was obtained by summing all individual bit-scores for each amino acid. The bit-scores for the target sequence then were analyzed, and statistics were computed using Prism software version 5.0 (GraphPad Software). For comparisons between germline sequence and mature sequence within the same design experiment, a Wilcoxon matched pairs test (non-normal, paired t-test) was used to compute the p-value at 99% confidence level. For comparison between design experiments, a student's paired t-test was used to compute the p-value at 99% confidence level.

VI.3.1.5 Amino Acid Environment

The neighbor vector algorithm quantitatively determines the surface-exposure of a given residue and is described by Durham and colleagues elsewhere (Durham et al., 2009). Briefly, each C_β is computed to a vector and each vector is given a score based on the number and orientation of each $C_\beta$in the proximity. The weight of each neighbor falls of as a function of distance. For interface scores, the change in neighbor vector was used, where the neighbor vector score of the amino acids in the unbound antibody is subtracted from the neighbor vector scores of the complex. Interface residues would have a large change in neighbors and proportional to the change in neighbor vector score.

VI.3.1.6 Phi-psi Angle Calculations

All V_H framework residues were grouped by complex. For each residue, phi-psi angles and secondary structure classification were determined using DSSP (Kabsch and Sander, 1983). For each residue position across all complexes considered in design, the standard deviation of the phi-psi angles was calculated if they were included in the beta-sheet framework. A student's t-test was performed between the standard deviations between residue positions that recovered to germline (bit-score > 1), or did not recover to germline (bit-score < 1). For a reference, a deviation for all framework beta-sheet positions was also calculated for all residues even if they were not included in the design protocol.

VI.3.2 Chapter III - Materials and Methods

VI.3.2.1 RNA Extraction

Peripheral blood mononuclear cells were isolated from 64 HIV-uninfected individuals (HIV-naïve) by processing leukoreduction filters as previously described (Weitkamp and Crowe, 2001). Briefly, RC2D leukoreduction filters were obtained from the American Red Cross and were backwashed with 35 mL of sterile PBS with 10mM EDTA. The resulting PBMC suspension was overlaid onto 15 mL of HistoPaque 1077 and centrifuged at 600 RCF for 25 minutes. The buffy coat was removed and washed twice with fresh PBS with 10mM EDTA. Total RNA was isolated from 10 million PBMCs using the RNeasy kit according to the manufacturer's standard operating procedure.

VI.3.2.2 cDNA Synthesis, PCR Amplification and Purification

cDNA was synthesized from 100 ng of total RNA and 10 pmol of each RT-PCR Illumina-adapter primers in duplicate 50 μL RT-PCR reactions using the OneStep RT-PCR system. The RT-PCR reactions were performed in a BioRad DNA Engine PTC-0200 thermal cycler running 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 was confirmed on a 1% E-Gel EX. After which duplicate reactions were pooled. 2 μL of each cDNA sample and 20 pmol of each indexed Illumina-adapter primer were used to template 100 μL PCR amplification reactions in duplicate using the AmpliTaq Gold polymerase system. Thermal cycling was performed using the following protocol:Âă95°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. Amplicons were purified from the PCR reaction mix using the Agencourt AMPure XP system following the standard protocol, and duplicate reactions were pooled during the final elution. The removal of primers and correct amplicon size was confirmed on a 1% E-Gel EX. Each amplicon sample was quantified using a Qubit fluorometer and the Quant-iT® dsDNA HS Assay Kit and 8 indexed amplicon samples were pooled for each of the 8 lanes on the Illumina HiSeq flowcell.

VI.3.2.3 Illumina HiSeq Protocol

The amplicon libraries underwent quality control by running on the Agilent Bioanalyzer High Sensitivity DNA assay to confirm the final library size and on the Agilent Mx3005P qPCR machine using the KAPA Illumina library quantification kit to determine concentration. For each library a 2 nM stock was created and denatured with NaOH. 12 pM of denatured libraries were loaded on the Illumina cBot for cluster generation on a paired-end flow cell. The flow cell was then loaded onto the Illumina HiSeq 2000 utilizing v3 chemistry and HTA 1.8. The raw sequencing reads in BCL format were processed through CASAVA-1.8.2 for FASTQ conversion and demultiplexing. The RTA chastity filter was used and only the pass filter reads were retained for further analysis.

VI.3.2.4 Paired-End Read Assembly and Junction Analysis

FASTQ paired end reads were input into PANDAseq assembler software to produce a single sequence that was indexed by donor and position (Bartram et al., 2011). Each sequence was uploaded to a custom database using the MongoDB framework that carried donor, position, sequence, and Phred quality score. The resulting sequences were concatenated and converted to FASTA format using BioPython SeqIO module (Cock et al., 2009). Heavy chain CDR3 (HCDR3) junctions were analyzed using custom software. The software was modified to run in parallel on a high throughput computing cluster and to condense output to a minimum number of fields. The software was also modified to output the junction results in JSON format. The sequences were analyzed with BioPython to remove sequence ambiguity in each donor. The JSON files were then uploaded to the custom database using MongoDB framework. The two databases were linked by their donor id and position.

VI.3.2.5 30 Length HCDR3 Selection and Position Specific Structure Scoring Matrix (P3SM) Generation

The custom database was queried for 30-length HCDR3 amino acid sequences generating > 26,000 unique sequences. 4,000 random sequences were selected for the pilot analysis in order to generate a custom position specific structure score matrix (P3SM) for PG9 HCDR3 structure. PG9 in complex with scaffolded template CAP45 (PDB ID: 3U4E) was used as a starting structure. The structure was stripped of waters and heavy chain and light chain constant regions. For the first round pilot, I also removed the CAP45 complex. Next, I used ROSETTASCRIPTS application available with the software suite from the ROSETTACOMMONS (www.rosettacommongs.org) to thread and minimize the random HCDR3 sequences from HIV-naïvedonors (Fleishman et al., 2011a). 50 decoys of each sequence were allowed to energetically minimize after threading yielding 200,000 models. 2,000 sequences (100,000 models) were used to fill the 30 by 20 P3SM using ROSETTAper amino-acid energies of the HCDR3 loop. The remaining 2,000 sequences were used to benchmark the P3SM protocol.

VI.3.2.6 Selecting Sequences from the P3SM Heuristic for Validation

After benchmark validation, the random 4,000 sequences were used in a final construction of a P3SM. Rapid prediction of score for each of the 26,000 HIV-naïve HCDR3 sequences were calculated using the P3SM. PG9's sequence scored 112nd out of 26,000 giving a noise tolerance of -3.82 REU (The top scoring sequence subtracted from the PG9 Score). Using $\pm$ 3.82 as my noise tolerance from PG9's score, 1,000 candidate sequences were selected to be further evaluated in complex.

VI.3.2.7 Sequence Tolerance Evaluated by Rosetta Design in Complex

The top 1,000 candidate sequences evaluated by the P3SM were carried on to a separate ROSETTA protocol. This protocol evaluated sequence tolerance in complex with CAP45 antigen and surrounding glycans. N-linked glycan 156 and 160 (HXBC2 numbering) were both included in the complex input to ROSETTA as a non-canonical amino acid using the method described in Renfrew et al. (Renfrew et al., 2012). After determining proper binding orientation with PG9, the entire complex was threaded with HIV-naïve sequences. High-resolution docking perturbations were allowed but highly constrained to initial orientation using standard ROSETTA constraints files. I generated 100 models for each naïve sequence and calculated a binding energy for each complex as:

$$\Delta\Delta \mathrm{G} = \Delta \mathrm{G}_{\mathrm{Bound}} - \Delta \mathrm{G}_{\mathrm{Unbound}}$$

were,

$$\Delta \mathrm{G}_{\mathrm{Bound}} = \mathrm{RosettaScore}_{\mathrm{Complex}}$$

and

$$\Delta \mathrm{G}_{\mathrm{Unbound}} = \mathrm{RosettaScore}_{\mathrm{Separated}}$$

In addition, the protocol was run a second time with sulfated tyrosines at positions 100G and 100H (Kabat numbering) if a tyrosine appeared at those positions in the HIV-naïve sequences. Complex energies and interface binding metrics were parsed into a MySQL database for further analysis using included scripts from BioPython.

VI.3.2.8 Bootstrapping with Complex Energies

The energy of each model evaluated in complex was reapplied to the P3SM and again ran through each HIV-naïve donor sequence to predict a Rosetta energy using the same methodology as described. The bootstrapped models were included in the rest of the protocol.

VI.3.2.9 HIV-Naïve Complex Energy Evaluation

To filter naïvesequences into a realistic number to synthesize I evaluated multiple metrics. To weight each sequence, Z-Scores were assigned for the following score term metrics. HCDR3 total energy, HCDR3 Cα-RMSD, HCDR3 $\Delta\Delta$G(the contribution to binding energy from just the HCDR3), total ΔΔG, ASN156 ΔΔG, and ASN158 ΔΔG. The Z-score is a measure of how many standard deviations a scoring metric fell from the mean. In terms of energy, all negative Z-Scores are preferred. When a Z-score was assigned for each HIV-naïvecomplex sequence, an average weighted Z-score was calculated using the following equation:

$$\text{Weighted-Z} = \frac{\sum_i^N w_i \times Z_i}{N}$$

Weights (w_i) for each score term in the equation: total ΔΔG -3.0, HCDR3 Cα-RMSD -0.5, HCDR3-ΔΔG -1.0, HCDR3 Score - 1.0, ASN156 ΔΔG - 0.5, and ASN158 ΔΔG - 0.5. This comprehensive metric can be used to rank-order each complex. In addition I used PG9 as a positive control and determined how many standard deviations away each of the HIV-naïve complex scoring terms were from PG9's score using the following equation:

$$\text{Compare Score} = \frac{\bar{X}_{ScoringTerm} - \bar{X}_{PG9ScoringTerm}}{\sigma_{PG9ScoringTerm}}$$

The compare score can then be weighted using the previous equation using the same weights to give one comprehensive metric to rank-order each HIV-naïve sequence. The top 50 sequences were selected based on the average of the weighted compare score and weighted Z-score. 32 additional models were included from the bootstrapped protocol in the final results yielding 82 candidate HIV-naïve sequences for experimental characterization.

VI.3.2.10 Clustering Analysis

The sequences were clustered with ClustalW2 built in clustering algorithm after a multiple sequence alignment. The ClustalW plugin was used from the Genious Software suite (http://www.geneious.com/). The dendrogram was manually inspected and clusters were assigned yielding 10 candidate sequence groups for experimental characterization.

VI.3.2.11 Design Analysis for Sequence Tolerance

Using the ROSETTADESIGN algorithm, the HIV-naïve sequences tested for recovery using a small energetic bonus for favoring the native sequence (Kuhlman and Baker, 2000). I applied a filter to minimize score and binding energy while favoring the native sequence. 100 models were generated using this protocol. After analysis, the sequence recovery was added to the Z-score metrics and the compare score using a weight of -2.0 (negative weight for favoring positive deviations) and reevaluated. Within each cluster, the HIV-naïve sequence with the highest recovery and lowest Z-Score was further evaluated. For each mutated position, if a mutation was seen in greater that 10% of the models and gave an energetic bonus of greater than 1.5 ROSETTA Energy Units (REUs), it was manually inspected using PyMOL and compared with the native sequence along with the native PG9 sequence from the native crystal complex (PDB ID: 3U4E).

VI.3.2.12 Antibody Expression

To prepare HIV-naïve PG9 variants and PG9 variant point mutations, I used recombinant expression in mammalian cells as previously described (Xu et al., 2010). Briefly, the MAb PG9 heavy- and light-chain genes were cloned into the pEE6.4 and pEE12.4 vectors, respectively. A BsiWI and XhoI cloning site were generated at AA position 95 and 110 (Kabat numbering), respectively. Using the unique cloning sites, the HIV-naïve HCDR3 sequences were synthesized and cloned into the PG9 backbone. The DNA was co-transfected at a 1:1 heavy-light chain ratio into HEK 293F using polyethylenimine transfection reagent at a ratio 2:1 of PEI to DNA. 30 mL of culture was used for each variant and supernatant was collected on day 3.

CAP45 gp120 was cloned into pCNA3.4 using HindIII and EcoRI restriction sites. A CD5 signal peptide and 8X HIS tag was cloned onto the $5'$ and $3'$ end respectively. The DNA was transfected into HEK293F using polyethylenimine at a ratio of 2:1. On day 7, the supernatant was collected and purified with a 5 mL Talon cobalt HIS affinity column according to the manufactures specifications. The protein was concentrated using centrifugal units with a 100 kD cutoff.

VI.3.2.13 PG9/HIV Naïve Variant Antiboy Characterization

ELISA plates were coated with 2 μg/mL of goat-anti-human (H+L) unlabeled antibody in PBS Buffer at 4° overnight. The wells were washed with 0.05% Tween and PBS Buffer all of the following steps. Using 2% powdered milk and 1% goat serum, the wells were blocked for 2 hours at room temperature. 200 μL of supernatant collected from expression were applied to each well and allowed to complex with the capture antibody for 1 hour at 37°. Starting at 25 μg/mL, 100 μL CAP45 gp120 was serially diluted at 1:3 in duplicate and allowed to bind for 1 hour at 37°. 100 μL of mAb b12 was used diluted at 1 μg/mL in blocking buffer and allowed to incubate for 1 hour at 37°. 100 μL of 1:5,000 of goat-anti-human labeled with horseradish peroxidase secondary was added to each well and allowed to incubate for 1 hour at 37°. 100 μL of 3,3′,5,5′-tetramethylbenzidine was added to each well. The reaction was stopped with 1N HCL and read at 450 nM absorbance. The EC_{50} of each HIV-naïve variant was compared with PG9 positive control.

VI.3.2.14 Statistics and Graph Generation

All statistics were calculated in the R-programming language (http://www.r-project.org) or GraphPad package. All graphs were generated in GraphPad package or the ggplot2 library (http://ggplot2.org) in the R-programming language.

VI.3.3 Chapter IV - Materials and Methods

VI.3.3.1 Position Specific Scoring Matrix to Determine the Tolerance of Diverse Sequences to the Hammerhead Structure of PG9

We obtained large numbers of human PBMCs from 64 otherwise healthy HIV-negative subjects by recovering cells from leuko-reduction filters obtained from the Nashville, TN Red Cross. Bryan Briney extracted total RNA from white blood cells retained in the filters, then performed RT-PCR amplification of expressed antibody heavy chain genes using primers designed to amplify all human heavy chain antibody sequences (Briney et al., 2012). I determined the sequences of the HCDR3 region of the amplicons using HiSeq next generation sequencing (Illumina) according to the manufacture's instructions. Amplifying and sequencing 64 donors separately yielded a total of $5.14\ x\ 10^8$ HCDR3 sequences. A subset of 4,000 randomly selected 30-amino acid length HCDR3 sequences was used to determine what amino acids were tolerated by antibodies in the hammerhead configuration of the PG9*wt* HCDR3 by threading each sequence over the backbone coordinates of PG9*wt* using ROSETTA. The backbone was energetically minimized with iterative rounds of small docking perturbations. Scores of each amino acid were input into a custom position specific scoring matrix (PSSM). The matrix then was used to rapidly compute the remaining 30 length amino acids predicted score given by ROSETTA.

VI.3.3.2 Redesign of PG9 HCDR3

Using the ROSETTADESIGN algorithm, iterative rounds of design, docking, and minimization were applied to each position in the HCDR3 with a small energetic bonus applied to recovery of the native sequence (Kuhlman and Baker, 2000). 100 models were generated using this protocol (see protocol capture). For each mutated position, if a mutation was seen in greater that 10% of the models and gave an energetic bonus of greater than 1.0 ROSETTA energy Units, it was manually inspected using PyMOL and compared with the native sequence along with the native PG9 sequence from the native crystal complex (PDB

ID-3U4E) (McLellan et al., 2011).

VI.3.3.3 Antibody and gp120 Expression

To prepare HIV-naïve PG9 variants and PG9 variant point mutations, I used recombinant expression in mammalian cells as previously described (Xu et al., 2010). Briefly, the mAb PG9 heavy- and light-chain genes were cloned into the pEE6.4 and pEE12.4 vectors, respectively (Lonza). BsiWI and XhoI cloning sites were generated at AA position 95 and 130, respectively. HIV-naïve HCDR3 sequences were synthesized, and cloned into the PG9 backbone (GeneArt) using the unique cloning sites. The DNA was co-transfected at a 1:1 heavy-light ratio into FreeStyle 293-F cells (Life Technologies) using 25 kDa linear polyethylenimine (PEI, Polysciences Inc.) transfection reagent at a ratio 2:1 of PEI to DNA. 30 mL of culture was used for each variant and supernatant was collected on day 5 and purified on a protein G column (GE).

Each gp120 was cloned into pCDNA3.4 (Life Technologies) using HindIII and EcoRI restriction sites. A CD5-signal peptide and 8x His tag was cloned onto the 5' and 3' end, respectively. The DNA was transfected into FreeStyle 293-F cells using PEI at a ratio of 2:1 (Life Sciences). On day 5, the supernatant was clarified and the protein purified on a 5 mL HisTALON cobalt column (Clontech) according to the manufacturers specifications. The protein was concentrated using Amicon Ultra centrifugal filters with a 100 kD cutoff (Millipore, Billerica, MA) and further purified on a Superdex column (GE) using size exclusion. BG505 SOSIP.664 trimer was received as a gift from John Moore.

VI.3.3.4 PG9 Variant Characterization

ELISA plates were coated with 3 μg/mL of gp120 and incubated overnight at 4°C. The wells were washed with phosphate buffered saline with 0.05% Tween (PBS-T) in all of the following steps. The uncoated sites on the wells were blocked with 2% skim milk and 1% goat serum in PBS-T for 2 hours at room temperature. All antibodies were diluted serially in two-fold starting from 25 μg/mL for 24 dilutions. Horseradish peroxidase-conjugated goat-anti-human IgG was added to each well and allowed to incubate for 1 hour at 37°C and color developed with 3,3,5′,5′-tetramethylbenzidine (Thermo). The reaction was stopped with 1N HCl and read at 450 nM. The EC_{50} of each PG9 variant was compared with PG9 positive control.

For BG505 SOSIP.664 Trimer, ELISAs were performed according the protocol as previously described (Sanders et al., 2013). Maxisorp 96-well plates (Nunc) were coated overnight with mAb D7324 (Aalto Bioreagents) at 5 mg/mL in 0.1 M $NaHCO_3$, pH 8.6 (100 μL/well). After the washing and blocking steps, purified, D7324-tagged BG505 Env proteins were added at 800 ng/mL in PBS and 2% milk for 2 h at ambient temperature and the unbound Env proteins were washed away. PG9 and PG9-variants were diluted to 25 μg/mL in PBS with 10% sheep serum/2% milk and diluted serially 2-fold and allowed to incubate for 2 h at room temperature followed by 3 washes with PBS-T. Horseradish peroxidase-conjugated goat-anti-human IgG was added for 1 h at a 1:3,000 dilution (final concentration 0.33 mg/mL) in 10% sheep serum/2% milk, followed by 5 washes with PBS-T. Color development and optical density measurement was done as above.

VI.3.3.5 Neutralization Assays

Neutralization was measured as a function of reductions in luciferase (Luc) reporter gene expression after a single round of infection in TZM-bl cells as described (Montefiori, 2009; Simek et al., 2009). This assay has been formally optimized and validated and was performed in compliance with Good Clinical Laboratory Practices (Sarzotti-Kelsoe et al., 2013). TZM-bl cells were obtained from the NIH AIDS Research and Reference Reagent Program, as contributed by John Kappes and Xiaoyun Wu. Briefly, virus at a dose of 50,000-150,000 relative luminescence units (RLU) equivalents was incubated with serial 3-fold dilutions of test sample in duplicate in a total volume of 150 μL for 1 hr at 37°C in 96-well flat-bottom culture plates. Freshly trypsinized cells (10,000 cells in 100 μL of growth medium containing 75 μg/mL DEAE dextran) were added to each well. One set of

control wells received cells + virus (virus control) and another set received cells only (background control). After a 48 hour incubation, 100 μL of cells was transferred to a 96-well black solid plates (Costar) for measurements of luminescence using the Britelite Luminescence Reporter Gene Assay System (PerkinElmer Life Sciences). Neutralization titers are the dilution at which RLU were reduced by 50% compared to virus control wells after subtraction of background RLUs. Assay stocks of molecularly cloned Env-pseudotyped viruses were prepared by transfection in 293T cells and were titrated in TZM-bl cells as described (Li et al., 2005). Additional details of the assay and all supporting protocols may be found at http://www.hiv.lanl.gov/content/nab-reference-strains/html/home.htm.
All of the Env-pseudotyped viruses used for these assays exhibited a Tier 2 neutralization phenotype except for TH023.6 and TH023.6/N160A.5, which exhibited a tier 1A phenotype. The Envs for these pseudoviruses were derived from genetic subtypes A (398_F1_F5_20), B (WITO4160.33, X2278_C2_B6, SC422661.8, TRO.11, SC22.3C2.LucR.T2A.ecto), C (Ce703010217, Du422.1, Ce1086_B2), G (X1632_S2_B6) and CRF01_AE (CNE55, R2184.c04).

VI.3.3.6 Statistics and Graph Generation

All statistics were calculated in the R-programming language (http://www.r-project.org) or Prism package (GraphPad) through the Ipython interface www.ipython.org. All graphs were generated in Prism package or the ggplot2 library (http://ggplot2.org) in the R-programming language.

VI.4 Appendix IV - Experimental Standard Operating Procedures

VI.4.1 Antibody Synthesis From Crystal Structures

I will detail the process of gene synthesis for the Crowe Lab Lonza vector system using a workable example.

VI.4.1.1 Full Heavy Chain Variable

Using PG9 Heavy Chain as a working example.

Get sequence from crystal structure

Using the PDB ID: 3U36 I get the heavy chain sequence in FASTA format:

```
>PG9_crystal_structure_3U36
QRLVESGGGVVQPGSSLRLSCAASGFDFSRQGMHWVRQAPGQGLEWVAFI
KYDGSEKYHADSVWGRLSISRDNSKDTLYLQMNSLRVEDTATYFCVREAG
GPDYRNGYNYYDFYDGYYNYHYMDVWGKGTTVTVSSASTKGPSVFPLAPS
SKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYS
LSSVVTVPSSSLGTQTYICNVNHKPSNTKVDKKVEPKSCDKGLEVLFQ
```

Truncate to variable regions

The variable region starts with EVQ or EQL and usually ends with TVSS. This is a bit subjective, but for this purpose, it does not really matter since I will prepend a sequence:

```
>PG9_variable_domain
QRLVESGGGVVQPGSSLRLSCAASGFDFSRQGMHWVRQAPGQGLEWVAFIK
YDGSEKYHADSVWGRLSISRDNSKDTLYLQMNSLRVEDTATYFCVREAGGP
DYRNGYNYYDFYDGYYNYHYMDVWGKGTTVTVSS
```

Reverse translate

Use a reverse translator to get nucleotide sequences.

```
>PG9_variable_nucleotide
CAGCGCCTGGTGGAAAGCGGCGGCGGCGTGGTGCAGCCGGGCAGCAGCCT
GCGCCTGAGCTGCGCGGCGAGCGGCTTTGATTTTAGCCGCCAGGGCATGC
ATTGGGTGCGCCAGGCGCCGGGCCAGGGCCTGGAATGGGTGGCGTTTATT
AAATATGATGGCAGCGAAAAATATCATGCGGATAGCGTGTGGGGCCGCCT
GAGCATTAGCCGCGATAACAGCAAAGATACCCTGTATCTGCAGATGAACA
GCCTGCGCGTGGAAGATACCGCGACCTATTTTTGCGTGCGCGAAGCGGGC
GGCCCGGATTATCGCAACGGCTATAACTATTATGATTTTTATGATGGCTA
TTATAACTATCATTATATGGATGTGTGGGGCAAAGGCACCACCGTGACCG
TGAGCAGC
```

Prepend 5′ region

I will prepend a 5’ smaI site (CCCGGG) and a portion of the leader sequence to keep it in frame (TCTGGCT). The leader sequence will be kept in frame and cleaved.

```
>PG9_with_smaI
(CCCGGG)TCTTGGGCTCAGCGCCTGGTGGAAAGCGGCGGCGGCGTGGTG
CAGCCGGGCAGCAGCCTGCGCCTGAGCTGCGCGGCGAGCGGCTTTGATTT
TAGCCGCCAGGGCATGCATTGGGTGCGCCAGGCGCCGGGCCAGGGCCTGG
AATGGGTGGCGTTTATTAAATATGATGGCAGCGAAAAATATCATGCGGAT
AGCGTGTGGGGCCGCCTGAGCATTAGCCGCGATAACAGCAAAGATACCCT
GTATCTGCAGATGAACAGCCTGCGCGTGGAAGATACCGCGACCTATTTTT
GCGTGCGCGAAGCGGGCGGCCCGGATTATCGCAACGGCTATAACTATTAT
GATTTTTATGATGGCTATTATAACTATCATTATATGGATGTGTGGGGCAA
AGGCACCACCGTGACCGTGAGCAGC
```

Append 3′ region

I then add on an ApaI restriction site (GGGCCC) along with additional nucleotides (GCCGGTACCAA) to keep it in frame.

```
>PG9_with_smaI/ApaI
(CCCGGG)TCTTGGGCTCAGCGCCTGGTGGAAAGCGGCGGCGGCGTGGTGCA
GCCGGGCAGCAGCCTGCGCCTGAGCTGCGCGGCGAGCGGCTTTGATTTTAGC
CGCCAGGGCATGCATTGGGTGCGCCAGGCGCCGGGCCAGGGCCTGGAATGGG
TGGCGTTTATTAAATATGATGGCAGCGAAAAATATCATGCGGATAGCGTGTG
GGGCCGCCTGAGCATTAGCCGCGATAACAGCAAAGATACCCTGTATCTGCAG
ATGAACAGCCTGCGCGTGGAAGATACCGCGACCTATTTTTGCGTGCGCGAAG
CGGGCGGCCCGGATTATCGCAACGGCTATAACTATTATGATTTTTATGATGG
CTATTATAACTATCATTATATGGATGTGTGGGGCAAAGGCACCACCGTGACC
GTGAGCAGCGCCGGTACCAA(GGGCCC)
```

Order Product

This will be the final product ordered. It is very important that you optimize for **mammalian** systems. You also do not introduce the following sites: HindIII, EcoRI, SmaI, ApaI. In addition, do not remove the SmaI or ApaI sites that I just added.

VI.4.1.2 Full Lambda Chain Variable

Using PG9 Lambda Chain as a working example.

Get sequence from crystal structure

Using the PDB ID: 3U36 I get the light chain sequence in FASTA format:

```
>PG9_crystal_structure_3U36_light_chain
QSALTQPASVSGSPGQSITISCQGTSNDVGGYESVSWYQQHPGKAPKVVIY
```

```
DVSKRPSGVSNRFSGSKSGNTASLTISGLQAEDEGDYYCKSLTSTRRRVFG
TGTKLTVLGQPKAAPSVTLFPPSSEELQANKATLVCLISDFYPGAVTVAWK
ADSSPVKAGVETTTPSKQSNNKYAASSYLSLTPEQWKSHKSYSCQVTHEGS
TVEKTVAPTECS
```

Truncate to variable regions

The variable region starts with QSAL and usually ends with GQP. This is a bit subjective, but for this purpose, it does not really matter since I will prepend a sequence:

```
>PG9_light_chain_variable
QSALTQPASVSGSPGQSITISCQGTSNDVGGYESVSWYQQHPGKAPKVVIY
DVSKRPSGVSNRFSGSKSGNTASLTISGLQAEDEGDYYCKSLTSTRRRVFG
TGTKLTVLGQP
```

Reverse translate

Use a reverse translator to get nucleotide sequences. They will be optimized so it does not matter.

```
>PG9_LC_variable_nucleotide
CAGAGCGCGCTGACCCAGCCGGCGAGCGTGAGCGGCAGCCCGGGCCAGAG
CATTACCATTAGCTGCCAGGGCACCAGCAACGATGTGGGCGGCTATGAAA
GCGTGAGCTGGTATCAGCAGCATCCGGGCAAAGCGCCGAAAGTGGTGATT
TATGATGTGAGCAAACGCCCGAGCGGCGTGAGCAACCGCTTTAGCGGCAG
CAAAAGCGGCAACACCGCGAGCCTGACCATTAGCGGCCTGCAGGCGGAAG
ATGAAGGCGATTATTATTGCAAAAGCCTGACCAGCACCCGCCGCCGCGTG
TTTGGCACCGGCACCAAACTGACCGTGCTGGGCCAGCCG
```

Prepend 5′ region

I will prepend a 5′ SalI site (GTCGAC) and a nucleotide (T) to keep it in frame.

```
>PG9_with_SalI
(GTCGAC)TCAGAGCGCGCTGACCCAGCCGGCGAGCGTGAGCGGCAGCCCGG
GCCAGAGCATTACCATTAGCTGCCAGGGCACCAGCAACGATGTGGGCGGCTA
TGAAAGCGTGAGCTGGTATCAGCAGCATCCGGGCAAAGCGCCGAAAGTGGTG
ATTTATGATGTGAGCAAACGCCCGAGCGGCGTGAGCAACCGCTTTAGCGGCA
GCAAAAGCGGCAACACCGCGAGCCTGACCATTAGCGGCCTGCAGGCGGAAGA
TGAAGGCGATTATTATTGCAAAAGCCTGACCAGCACCCGCCGCCGCGTGTTT
GGCACCGGCACCAAACTGACCGTGCTGGGCCAGCCG
```

Append 3′ region

I then add on an NotI restriction site (GCGGCCGC) with no additional nucleotides.

```
>PG9_with_SalI/NotI
(GTCGAC)TCAGAGCGCGCTGACCCAGCCGGCGAGCGTGAGCGGCAGCCCGG
GCCAGAGCATTACCATTAGCTGCCAGGGCACCAGCAACGATGTGGGCGGCTA
TGAAAGCGTGAGCTGGTATCAGCAGCATCCGGGCAAAGCGCCGAAAGTGGTG
ATTTATGATGTGAGCAAACGCCCGAGCGGCGTGAGCAACCGCTTTAGCGGCA
GCAAAAGCGGCAACACCGCGAGCCTGACCATTAGCGGCCTGCAGGCGGAAGA
TGAAGGCGATTATTATTGCAAAAGCCTGACCAGCACCCGCCGCCGCGTGTTT
GGCACCGGCACCAAACTGACCGTGCTGGGCCAGCCG(GCGGCCGC)
```

Order Product

This will be the final product ordered. It is very important that you optimize for **mammalian** systems. You also do not want to avoid the following sites. HindIII, EcoRI, SalI and NotI. In addition, do not remove the SalI or NotI sites that I just added.

VI.4.1.3 Designing a swappable vector

This was how I made a PG9 vector with a swappable HCDR3 sequence. This particular protocol will work for any HCDR3 sequence that uses J_H6, but can be extended to any family based on the constant FR4 sequences that are needed.

Get HCDR3 and FR4 Sequence

First thing I need is the HCDR3 sequence. Considering that I'm only adding restriction sites (RE) to the HCDR3 region, I will only consider that in the example. Make sure this sequence contains CVR (the beginning of the HCDR3) and TVSS (the end of FR4). Those two regions will contain the restriction sites when we are done. Then we can back translate.

```
TTTTGcgtacgCGAAGCGGGCGGCCCGGATTATCGCAAC
--F--C--V--R--E--A--G--G--P--D--Y--R--N

GGCTATAACTATTATGATTTTTATGATGGCTATTATAAC
--G--Y--N--Y--Y--D--F--Y--D--G--Y--Y--N

TATCATTATATGGATGTGTGGGGCAAAGGCACCACCGTG
--Y--H--Y--M--D--V--W--G--K--G--T--T--V

ACCGTctcgagC
--T--V--S--S
```

The restriction sites are shown in lower case in the HCDR3 region. BsIWI (cgtacg) and XhoI (ctcgag). It is just a matter of swapping that sequence into the Lonza vector.

Order Construct

```
cccgggTCTTGGGCTCAGCGCCTGGTGGAAAGCGGCGGCGG
CGTGGTGCAGCCGGGCAGCAGCCTGCGCCTGAGCTGCGCGG
CGAGCGGCTTTGATTTTAGCCGCCAGGGCATGCATTGGGTG
CGCCAGGCGCCGGGCCAGGGCCTGGAATGGGTGGCGTTTAT
TAAATATGATGGCAGCGAAAAATATCATGCGGATAGCGTGT
GGGGCCGCCTGAGCATTAGCCGCGATAACAGCAAAGATACC
CTGTATCTGCAGATGAACAGCCTGCGCGTGGAAGATACCGC
GACCTATTTTTGcgtacgCGATAGCGGCGGCTATGATTTTT
GGAGCGGCTATGAAGTGGGCCTGGAAAACCGCGAAAACTAT
TATTATTATGGCATGGATGTGTGGGGCAAAGGCACCACCGT
GACCGTctcgagCGCCGGTACCAAgggccc
```

This construct can be ordered now with the two restriction sites flanking the HCDR3 sequence. When you order the construct, keep the restriction sites BsiWI, XhoI, ApaI, HindIII, EcoRI, and SmaI. You can clone this full vector with SmaI and ApaI.

VI.4.1.4 Synthesizing HCDR3 Only

If the swappable construct with BsiWI and ApaI sites was made, you can simply order the HCDR3 sequence using the technique below. Here is a random 30-length sequence (2383:160514).

Get HCDR3 Sequence of Interest

We will use the IMGT definition of HCDR3.

```
>2383:160514
CAREGGDYDFWSGYYRGYSGYGEEHYYYMDVW
```

Cut Off Leading Sequences

This is usually a CVR or CAR. We don't need these sequences since the vector has them.

```
>2383:160514_truncated
EGGDYDFWSGYYRGYSGYGEEHYYYMDVW
```

Reverse Translate

```
 >2383:160514_nucleotide
GAAGGCGGCGATTATGATTTTTGGAGCGGCTATTATCGCGGCTAT
AGCGGCTATGGCGAAGAACATTATTATTATATGGATGTGTGG
```

Add 5′ Sequence

Adding the BsiWI sequence (CGTACG) along with a nucleotide (C) to keep it in frame.

```
 >2383:160514_nucleotide
CGTACGCGAAGGCGGCGATTATGATTTTTGGAGCGGCTATTATCG
CGGCTATAGCGGCTATGGCGAAGAACATTATTATTATATGGATGT
GTGGGGCAAA
```

Add 3′ Sequence

Have to add a lot of the constant region (GGCAAAGGCACCACCGTGACCGT) since it takes a several nucleotides to get to the 5’ XhoI site (CTCGAG).

```
>2383:160514_BsiWI_XhoI
CGTACGCGAAGGCGGCGATTATGATTTTTGGAGCGGCTATTATCG
CGGCTATAGCGGCTATGGCGAAGAACATTATTATTATATGGATGT
GTGGGGCAAAGGCACCACCGTGACCGTCTCGAG
```

Order Sequence

When you order the construct, keep the restriction sites BsiWI, XhoI, ApaI, HindIII, EcoRI, and SmaI. You can clone this into the HCDR3 swap vector with BsiWI and XhoI.

VI.4.1.5 Restriction Map - All Constructs

For reference figure VI.1.

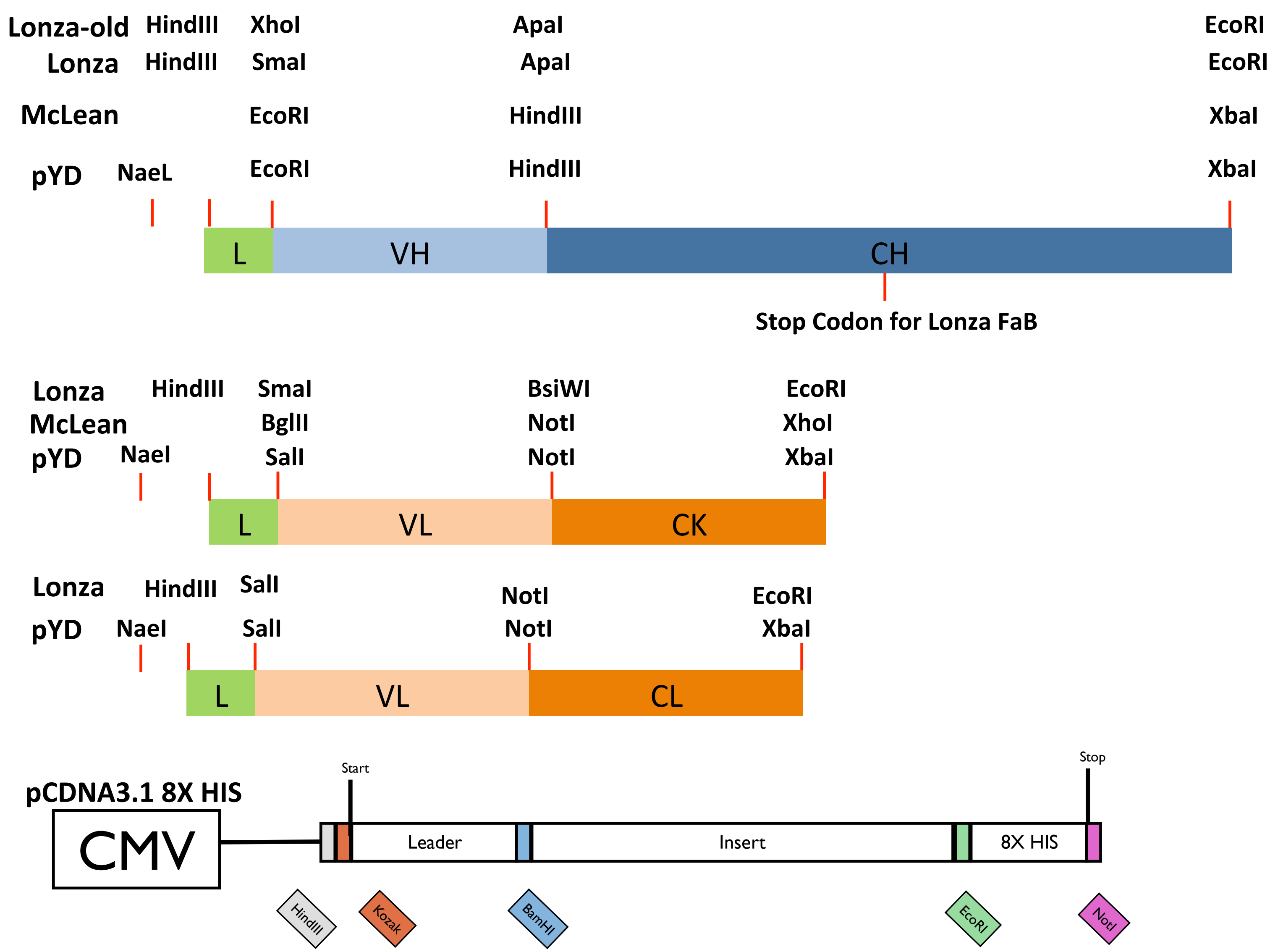

Lonza-old HindIII XhoI ApaI EcoRI
Lonza HindIII SmaI ApaI EcoRI
McLean EcoRI HindIII XbaI
pYD NaeL EcoRI HindIII XbaI
L
VH
CH
Stop Codon for Lonza FaB
Lonza HindIII SmaI BsiWI EcoRI
McLean BglII NotI XhoI
pYD NaeI SalI NotI XbaI
L
VL
CK
Lonza HindIII SalI NotI EcoRI
pYD NaeI SalI NotI XbaI
L
VL
CL
pCDNA3.1 8X HIS
CMV
Start
Stop
Leader
Insert
8X HIS
HindIII
Kozak
BamHI
EcoRI
NotI

Figure VI.1

Template B

	1	2	3	4	5	6	7	8	9	10	11	12
A	CC	VC	Dil 4	Dil 4	Dil 4	Dil 4	Dil 4	Dil 4	Dil 4	Dil 4	Dil 4	Dil 4
B	CC	VC	Dil 3	Dil 3	Dil 3	Dil 3	Dil 3	Dil 3	Dil 3	Dil 3	Dil 3	Dil 3
C	CC	VC	Dil 2	Dil 2	Dil 2	Dil 2	Dil 2	Dil 2	Dil 2	Dil 2	Dil 2	Dil 2
D	CC	VC	Dil 1	Dil 1	Dil 1	Dil 1	Dil 1	Dil 1	Dil 1	Dil 1	Dil 1	Dil 1
E	CC	VC	Dil 4	Dil 4	Dil 4	Dil 4	Dil 4	Dil 4	Dil 4	Dil 4	Dil 4	Dil 4
F	CC	VC	Dil 3	Dil 3	Dil 3	Dil 3	Dil 3	Dil 3	Dil 3	Dil 3	Dil 3	Dil 3
G	CC	VC	Dil 2	Dil 2	Dil 2	Dil 2	Dil 2	Dil 2	Dil 2	Dil 2	Dil 2	Dil 2
H	CC	VC	Dil 1	Dil 1	Dil 1	Dil 1	Dil 1	Dil 1	Dil 1	Dil 1	Dil 1	Dil 1
			Samples 1 & 2		*Samples 3 & 4*		*Samples 5 & 6*		*Samples 7 & 8*		*Samples 9 & 10*	

Figure VI.2

VI.4.2 HIV Neutralization Assay

Pre-Neutralization Assay - -Make growth media (GM) also called TZMbl media. (DMEM high glucose 1X + 110 mg/mL NaPy + 1X Pen-Strep + 10% heat inactivated FBS) -If you are testing serum, make sure to heat inactivate.

Neutralization Assay Using the plate setup found in figure VI.2.

1. Put 150 μL of GM into column 1
2. Put 100 μL of GM in the rest of the wells
3. Start with 7.5 μg/mL of antibody in the Dil 1 wells (Row H, columns 3-12). This well be serially diluted 3 fold and give a final dilution of 0.2 μg/mL.
4. Fill columns 3-12, row H up to 150 μL with GM.
5. Use 50 μL of these rows to do a serial 3-fold dilution. Discard the 50 μL out of the remaining row to ensure only 100 μL s left.
6. Make a viral stock of 10 mL of GM to 4,000 $TCID_{50}$. Viruses must be titered (see other protocols).
7. Add 50 μL of 4,000 $TCID_{50}$ stock to columns 2-12. Make sure you don't do column 1.
8. Cover and incubate at 37°C for 2 hours.

While the virus and antibodies are incubating. Prepare the cells to be added at the end of the incubation.

1. TZMbl cells should be confluent. Decant cell culture and wash with sterile PBS.

2. Add 5 mL of tryspin (to T150 flask) and incubate for 1 minute.
3. Aspirate trypsin and incubate cells at 37 °C for five minutes.
4. Add 10 mL of growth media and count cells.
5. Dilute cells in 500 mL falcon tube to 100,000/mL with GM.
6. If it has been two hours, take the virus and antibody plates and add 100 μLof cells to every well.
7. Cover and incubate at 37 °C for 48-72 hours.
8. From every well aspirate 100 μL of GM.
9. Add 100 μL of Bright Glow Luciferase reagent and mix 10 times in every row. Incubate for 2 minutes.
10. Record the luminescence on a luminometer.
11. If background is low enough, IC_{50} can be recorded from automatic PRISM non-linear regression analysis after $\log_2$ transformation of concentrations.

VI.5 Appendix V - Computational Standard Operating Procedures

Here I will detail the computational procedures including running ROSETTA and analyzing data with scripts that are often available in the Meiler Lab Scripts repository or available on request. Also a lot of the procedures are detailed in IPython Notebooks and will also be available on request. **Using ROSETTA version 80616601370**

VI.5.1 Chapter I - Multi-State Design

Here I will detail how I ran ROSETTADESIGN for multi-state design and how I analyzed the results. I will use the simplified example of IGH_V5-51 which only contains three sets of molecular structures. There is a great protocol capture for complex procedures attached to the publication by Andrew Leaver-Fay showing how to *design for* and *design against* in multiple states (Leaver-Fay et al., 2011a).

VI.5.1.1 Running ROSETTA Multi-State Design

To run multi-state design, I have to prepare several files.

- Entity File - A file containing the amount of residue positions to design as well as instructions for the packer to behave on all proteins. For example, we could want the packer only to use certain rotamers around the interface. This could be handled with the entity file.
- Correlation File - Tells how residues correlate to each other. For example, residue 1 on protein A should be designed with residue 2 on protein B etc.
- Secondary Residue File - This is a residue file as defined in the documentation, but will only instruct the packer to operate on each protein individually. Every state must have it's own secondary residue file.
- Fitness File - A master file containing all other files as well as instructions for the fitness function.

Clean PDB - First, I can download all three protein PDBs with the clean_pdb.py script. Clean PDB supports the following syntax:

```
clean_pdb.py <PDB_ID> <CHAINS>
```

We only want the asymmetric unit in the crystal structure, so it helps to manually inspect the PDBs. We need one heavy chain, one light chain, and one antigen. I just go to www.pdb.org to find these chain codes.

```
clean_pdb.py 2B1A HLP
clean_pdb.py 2XWT ABC
clean_pdb.py 3HMX HLB
```

Although not absolutely necessary, it makes it easier to label the chain IDs the same. All heavy chains have H, light L, and antigen A. There is a change PDB id script which allows us to quickly rename chain IDs. The script takes the following syntax.

```
set_pdb_chain_id.py old_chain new_chain input output
```

I have looked through all the PDBs and figured out which names to change.

```
set_pdb_chain_id.py P A 2B1A_HLP.pdb 2B1A_HLA.pdb
set_pdb_chain_id.py A H 2XWT_ABC.pdb 2XWT_HBC.pdb
set_pdb_chain_id.py B L 2XWT_HBC.pdb 2XWT_HLC.pdb
set_pdb_chain_id.py C A 2XWT_HLC.pdb 2XWT_HLA.pdb
set_pdb_chain_id.py B A 3HMX_HLB.pdb 3HMX_HLA.pdb
```

To ensure the starting structures use the correct numbering scheme, we should renumber each chain starting with 1.

```
renumber_pdb.py 2B1A_HLA.pdb 2B1A_clean.pdb
renumber_pdb.py 2XWT_HLA.pdb 2XWT_clean.pdb
renumber_pdb.py 3HMX_HLA.pdb 3HMX_clean.pdb
```

Now we can remove all temporary files. Only *_clean.pdb files should remain in the working directory. The next thing to do would be to find all positions with at least one difference. This requires manual inspection of the alignment. For the V_H gene, there are 29 amino acid positions that will differ from germline in at least one position. These positions will be considered. Given that data, we can construct the entity residue file.

```
#The entity.resfile
#The number of positions to design
29
#Allow all amino acids
#except cystine and use rotamer libraries 1,2
#and aromatic 2.
ALLAAxc EX 1 EX 2 EX ARO 2
#beginning of residue file
start
```

The correlation file maps how each residue in each file should map to the others. There will be three correlation files, one for each state. Since each amino acid lines up, i.e. design position 5 in 2B1A with position 5 in 3HMX, all the correlation files will be the same. Here is an example of one correlation file.

```
#all.corr
#The first column is the entity,
#the second is the residue number for that state,
#the last is the chain.
1 5 H
2 14 H
3 16 H
4 23 H
5 24 H
6 29 H
7 30 H
8 31 H
9 32 H
```

```
10 34 H
11 40 H
12 46 H
13 48 H
14 51 H
15 52 H
16 54 H
17 58 H
18 65 H
19 70 H
20 72 H
21 74 H
22 76 H
23 77 H
24 80 H
25 84 H
26 88 H
27 93 H
28 97 H
29 98 H
```

A secondary residue file is also needed in case any extra packing tasks are needed to be supplied to each state. For example, we could tell one PDB state to design around the interface in single state design mode while everything in the correlation file designs together. I do not require extra design tasks for this protocol so all secondary residue files will be the same. For example,

```
#tells the packer to use all natural side
#chain configurations for everything
#that is not being designed.
NATRO
#the input side chain is allowed
use_input_sc
#start the residue file
start
```

We next to create a states file that has the PDB, correlation file and secondary resfile names in it. Name them 2B1A.states, 2XWT.states, and 3HMX.states. They should look like the following when opened.

```
#2B1A.states
input_files/2B1A_clean.pdb input_files/all.corr input_files/
    all.2res

#2XWT.states
input_files/2XWT_clean.pdb input_files/all.corr input_files/
    all.2res
```

```
#3HMX.states
input_files/3HMX.pdb input_files/all.corr input_files/all.2
   res
```

Lastly, a fitness file needs to be constructed to tell multi-state how to design. I call this file fitness.daf and it points to the locations of the states files.

```
#initialize the states and what the states file name is
STATE_VECTOR A input_files/2B1A.states
STATE_VECTOR B input_files/2XWT.states
STATE_VECTOR C input_files/3HMX.states


#tell design to minimize energy for each state
SCALAR_EXPRESSION best_A = vmin( A )
SCALAR_EXPRESSION best_B = vmin( B )
SCALAR_EXPRESSION best_C = vmin( C )

#Fitness – design to minimize all energies simultaneously
FITNESS best_A + best_B + best_C
```

Running Rosetta Multi-State Design

The ROSETTAexecutable is called mpi_msd.mpi.<operatingsystem>. It must be compiled in MPI mode as each state is assigned to a processor. The command line takes the following options.

- entity_resfile - The resfile that we created in the input portion
- fitness_file - The fitness file we created in the input portion
- ms::pop_size - How many sequences to keep in memory at once (100 is a good number)
- ms::generation - How many sequence generations should MSD go through see (Leaver-Fay et al., 2011a) to find see how the genetic algorithm selects sequences.
- ms::numresults - How many results to output. Will output top N sequences.
- ms:fraction_by_recombination - How often should a cross-over even take place between sequences in the population. Read (Leaver-Fay et al., 2011a) for details on the genetic algorithm.
- database - The location of the database.

I construct an options file with all those options (options.txt) that looks like this.

```
–entity_resfile entity.resfile
–fitness_file fitness.daf
```

```
-ms
    -pop_size 100
    -generations 435
    -numresults 100
    -fraction_by_recombination .04
-database my/rosetta/database/location/
```

Finally we can run ROSETTA using the following command after starting MPD.

```
mpd && mpiexec -n 4 /my/rosetta/location/mpi_msd.mpi.
   myoperatingsystem \
@input_files/options.txt
```

Warning!
ROSETTA may complain about some of the comments (anything starting with #) not being recognized, if so, just remove it from the file
The output will be 300 files, 100 for each of the states. We only need to analyze 100 files considering that the designed entities will be the same for all three files. For example, position 5 will be the same for 2B1A, 2XWT, and 3HMX.

VI.5.1.2 Analysis of MSD Output

I wrote a design analysis script called design_analysis.py which encompasses many design analysis tools. I will only go into the functionality that is necessary to use, but you can read the options file for more use.

```
design_analysis.py --help
>>                Design Analysis
-----------------------------------------------------------------

This script is intented to encompass the entire
   functionality
of design analysis. Everything you could want to do with
   design
is called upon in this script. The most basic functionality
   is
to pass a list of pdbs or get a position matrix of
   occurences
count of just one line.

The functionality extends from there by giving bitscores,
changes in energy, giving position specific scoring matrices
of your design,giving a customizable sequence logo. This is
   a
combination of many scripts and classes.

optional arguments:
  -h, --help show this help message and exit
```

```
Necessary:
  PDB files have to be included
  *.pdb  The PDB files to be analyzed

Recommended Options:
  Will give you a more complete analysis based on a res file
     ,
  and a native pdb to compare it to.

  --native_pdb N_PDB, -p N_PDB
        The native pdb file to compare against
  --corr corr.corr, -c corr.corr
        Get the results defined only in the corr file
  --res resfile.resfile, -r resfile.resfile
        Get the results defined only in the residue file

Output Options: Please read carefully:
  These arguments change file name, which file is printed,
  which is output to a dictionary,and give verbose printing

  --verbose, -v
        everything printed to a file will also be shown
        on the screen
  --prefix PREFIX, -P PREFIX
        The prefix for what all the output files will be
  --score_files O_FILES [O_FILES ...], -s
    What do you want output to a file?
    Can list as a space seperated (eg -s n d nd):
    a - full analysis dict
    d - give analysis of just designed residues
    n - just the native residues scores are shown
    nd - just the native residues of the residues designed
            Defaults to full analysis dictionary
  -b Should the output be in bit score?
        Defaults to occurences instead of bitscore
  -S If you specify a native file and a design file,
        it will give you an output of the stats of the
           design

Rosetta Energy Analysis:
  Options for outputing options about energy scores,
  the dictionaries analyzed depend on what you asked
  for using the -s output options flag
```

```
  --rosetta, -t  This option will output a .csv file of the
  model,chain,residue,residue number,and rosetta scores.

Bit Score Options:
  options for bitscore metric for each designed residue

  -n    do you want the bit scores to be
        normalized by the shannon entropy

Sequence Logos Options:
  These options handle the sequence logos that can be output
  from the design analysis script, and uses the api of
     weblogo
  to do so.

  --seq, -l
    Turn on Sequence Logos for all the dictionaries
    you supply given in an .eps file
  --path LOGO\_PATH, -lp LOGO\_PATH
    What is the path to weblogo software?
    Defaults to meilerlab enviroment
  --format {eps,jpg,png,png_print,pdf,jpeg,svg,logodata}
    What format do you want the sequence logo in?
  --units {bits,nats,kt,kJ/mol,kcal/mol,probability}
    What do you want the units of the sequence
    logo to be in? Defaults to bits.
  --stacks S_STACKS
    How many sequences per line in the logo, default=after
    forty letters it will go to a new line.
  --stack_width S_STACK_WIDTH
    How wide is each stack in the logo. Value of 25 is
       useful
    for x-axis labels >3 characters and 30
    for labels as 'sequence_numbers'.
  --title S_TITLE
    The title of your sequence logo
  --x_label S_X_AXIS
    What do you want the x axis titled?
  --y_axis_height S_Y_HEIGHT
    How high do you want the Y axis,
    currently 4.32 which is the maximum
    acheivable score in a unbiased design
  --y_label S_Y_LABEL
    Title of Y-Axis
  --error_bars S_Y_ERROR
```

```
    Do you want error bars turned on, YES/NO?
  --fine_print S_FINE
    Fine Print
  --color_scheme
    {auto,chemistry,charge,classic,
    hydrophobicity,monochrome}
    The color scheme of the sequence logo.
    Defaults to Classic
  --labels {sequence,numbers,sequence_numbers}
    The x-axis labels can either take on the
    native residues sequence given with a native
    pdb file or the numbering of the pdb residue.
  --debug
    Get the full command line of what was put into weblo
```

It's obvious that this analysis can do a lot, but I will stick with the basics. First a new input that is completely germline. We can do this with the packer. There is a script called make_-res.py which will make a residue file from a FASTA file. I use this to take the germline sequence and thread it over one of my inputs. Then that input can be used as our template.

```
>IGHV5-51*01
EVQLVQSGAEVKKPGESLKISCKGSGYSFTSYWIGWVRQMPGKGLEW
MGIIYPGDSDTRYSPSFQGQVTISADKSISTAYLQWSSLKASDTAMY
YCAR
```

Then we can use the make_res.py script.

```
make_res.py IGHV5-51.fasta > IGHV5-51.res
```

This residue file can then be used to mutate one of our templates back to germline.

```
/path/to/rosetta/bin/fixbb.default.<operating_system> \
-s 2B1A_clean.pdb \
-resfile IGHV5-51.res -database /path/to/database \
-o 2B1A_germline.pdb
```

Now we can use the analysis script from the analysis directory

```
../analysis/design_analysis -p ../input_files/2B1A_clean.pdb
    \
-S -b -s nd -c ../input_files/all.corr
../output_files/msd_output_*A*.pdb

>>
#score vs mature
Total Bit Score of Design ===> 28.4534
Total Shannon Entropy of Design ===> 115.9577
Normalized Bit Score for design ===> 0.2454
```

and

```
../scripts/design_analysis -p analysis/2B1A_germline.pdb \
-S -b -s nd -c ../input_files/all.corr
../output_files/msd_output_*A*.pdb

>>
#score vs mature
Total Bit Score of Design ===> 50.2318
Total Shannon Entropy of Design ===> 115.9577
Normalized Bit Score for design ===> 0.4337
```

This gives a design score towards 0.35 for the germline sequence and 0.24 for the mature sequence. Everything can be repeated this exact way. For single state design you can keep the fitness file exactly the same, but remove the other states.

```
#initialize the states and what the states file name is
STATE_VECTOR A 2b1a.states
STATE_VECTOR B 2xwt.states
STATE_VECTOR C 3hmx.states


#tell design to minimize energy for each state
SCALAR_EXPRESSION best_A = vmin( A )
SCALAR_EXPRESSION best_B = vmin( B )
SCALAR_EXPRESSION best_C = vmin( C )

#Fitness - design to minimize all energies simultaneously
FITNESS best_A
```

And run the exact same procedures.

VI.5.2 Chapter II - Database and Design

This section accompanies chapter II. I will go over uploading the sequences to a database, selecting the correct sequences, threading, and finally design.

VI.5.2.1 Sequence Analysis

The methods section detailed how I actually sequence the amplicons from 64 healthy donors, but processing them takes quite a bit of computational work. The VANTAGE core at Vanderbilt returns the sequencing runs as two paired end reads in FASTQ format. They must be “stiched” together to make one read. For example, for donor 10, there are “2185-RC-10_1.fastq” for the forward read and “2185-RC-10_2.fastq” for the reverse read. I use a stitching algorithm called “pandaseq” to process theses (Bartram et al., 2011). These commands are incredibly simple to run.

```
/usr/local/bin/pandaseq -f 2185-RC-10_1.fastq -r 2185-RC-10
    _2.fastq -T 23 > donor_10.fasta
```

Pandaseq will automatically output to the fasta format which is convenient for the next step. Here I use PyIg, my own sequence aligner against Ig mAbs that's based on IgBLAST (Ye et al., 2013). I will probably publish this soon when it is more stable. For human IgG's it works incredibly simple to use.

```
./PyIg
usage: igblast [-h] -q query.fasta [-d DB_PATH] [-i
    INTERNAL_DATA]
               [-a AUX_PATH] [-y {Ig,TCR,custom}] [-or {
                   human,mouse}]
               [-nV NUM_V] [-nD NUM_D] [-nJ NUM_J] [-dgm
                   D_GENE_MATCHES]
               [-s {imgt,kabat}] [-x EXECUTABLE] [-o OUT] [-
                   t TMP]
               [-e E_VALUE] [-w WORD_SIZE] [-pm
                   PENALTY_MISMATCH]
               [-nP NUM_PROCS] [-op OUTPUT_OPTIONS] [-z] [-c
                   ] [-j]

optional arguments:
  -h, --help            show this help message and exit

Necessary:
  These have to be included

  -q query.fasta, --query query.fasta
                        The fasta file to be input into
                           igBlast

Database Paths:
  -d DB_PATH, --db_path DB_PATH
                        The database path to the germline
                           repertoire
  -i INTERNAL_DATA, --internal_data INTERNAL_DATA
                        The database path to internal data
                           repertoire
  -a AUX_PATH, --aux_path AUX_PATH
                        The auxilariay path that contains
                           the frame origins of the germline
                            genes for each repertoire.
                         Helps produce translation and other
                             metrics

IgBlast Sprecific:
  IgBlast Specific Options with a Default
```

```
  -y {Ig,TCR,custom}, --type {Ig,TCR,custom}
                        Is this an IG or TCR recombination
  -or {human,mouse}, --organism {human,mouse}
                        The organism repeortire to blast
                           against
  -nV NUM_V, --num_v NUM_V
                        How many V-genes to match?
  -nD NUM_D, --num_d NUM_D
                        How many D-genes to match?
  -nJ NUM_J, --num_j NUM_J
                        How many J-genes to match?
  -dgm D_GENE_MATCHES, --d_gene_matches D_GENE_MATCHES
                        How many nuclodtieds in the D-gene
                           must match to call it a hit
  -s {imgt,kabat}, --domain {imgt,kabat}
                        Which classification system do you
                           want

General Settings:
  -x EXECUTABLE, --executable EXECUTABLE
                        The location of the executable,
                           default is /usr/bin/igblastn
  -o OUT, --out OUT     output file prefix
  -t TMP, --tmp TMP     temporary directory to store files
     in.
                        Defaults to ./tmp
  -e E_VALUE, --e_value E_VALUE
                        Real value for excpectation value
                           threshold in blast.
                        Put in scientific notation
  -w WORD_SIZE, --word_size WORD_SIZE
                        Word size for wordfinder algorithm
  -pm PENALTY_MISMATCH, --penalty_mismatch PENALTY_MISMATCH
                        Penalty for nucleotide mismatch
  -nP NUM_PROCS, --num_procs NUM_PROCS
                        How many do you want to split the
                           job across, default is the number
                            of processors

Outputting Options:
  -op OUTPUT_OPTIONS, --output_options OUTPUT_OPTIONS
                        Open this file and comment out
                           options you don't want in your
                           final file.
```

```
                          The first column is the name of the
                             option.
                          The second column is used by the
                             parser and should not be changed.
  -z, --zip               Zip up all output files
  -c, --concatenate       Turn off automatic concatenation and
      deletion of temporary files.
                          Files are split up at the beginning
                             to run across multiple processors
  -j, --json              Use the JSON output option that will
      format the text driven igblast output to a json
     document.
                          Defaults to CSV
```

I'll keep it simple for the protocol capture, but want to show how robust PyIg can be. Assume you are in the PyIg directory under PyIg/src/

```
python2.7 execution.py -q donor_10.fasta -d datafiles/
   database/ -i internal_data/ -a datafiles/optional_file/ -
   y Ig -or human -nV 1 -nD 1 -nJ 1 -s imgt -o donor_10 -nP
   23 -z -j
```

Each option is listed in the help. Make sure you see which each one does. For instance -nP needs to be carefully considered since it is the number or processors it will be run on. The output to this command line is a json file that will be uploaded to a Mongo database. One entry (with default output settings) looks like this.


```
{
    "_id" : "donor_10_1000",
    "raw_seq" : "TGGAGCTGAGCAGCCTGAGATCTGAGGACACGGCCGT
    ATATTACTGTGCGAAAGAACTATATGATAGTAGTGGTTATTACTACTTCC
    TGCCTTCTTACTACTACTACGGTATGGACGTCTGGGGCCAAGGGACCACG
    GTCACCGTCTCCTCAGGTAAG",
    "d_region" : "TATGATAGTAGTGGTTATTACTAC",
    "cdr3_aa" : "AKELYDSSGYYYFLPSYYYYGMDV",
    "fw4_aa" : "WGQGTTVTVSSGK",
    "full_seq_aa" : "AKELYDSSGYYYFLPSYYYYGMDVWGQGTTVTVSSGK",
    "cdr3" : "GCGAAAGAACTATATGATAGTAGTGGTTATTACTACTTCCTGCCTT
    CTTACTACTACTACGGTATGGACGTCTG",
    "top_d" : "IGHD3-22*01",
    "v_d_junction" : "ACTA",
    "top_j" : "IGHJ6*02",
    "cdr3_aa_length" : 24,
    "fw4" : "TGGGGCCAAGGGACCACGGTCACCGTCTCCTCAGGTAAG",
    "d_j_junction" : "TTCCTGCCTTCT",
    "d_or_j_junction" : "",
    "top_v" : "IGHV1-69*06",
```

```
    "full_seq" : "GCGAAAGAACTATATGATAGTAGTGGTTATTACTACTTCCT
    GCCTTCTTACTACTACTACGGTATGGACGTCTGTGGGGCCAAGGGACCACGGTCA
    CCGTCTCCTCAGGTAAG",
    "full_seq_aa" : "AKELYDSSGYYYFLPSYYYYGMDVWGQGTTVTVSSGK",
    "d_bit_score" : 48.3,
    "d_evalue" : "3e-10.0",
    "d_alignment_length" : 24,
    "d_query_seq" : "TATGATAGTAGTGGTTATTACTAC",
    "d_subject_seq" : "TATGATAGTAGTGGTTATTACTAC",
    "d_percent_identity" : 100,
    "d_percent_positives" : 100,
    "d_mismatches" : 0,
    "d_positives" : 24,
    "d_identical" : 24,
    "d_subject_length" : 31,
    "d_score" : 24,
}
```

To download and install mongo, consult the manual (http://docs.mongodb.org/manual/installation/). This installation will also include all the tools that making uploading files very easy. The output of PyIg should be “donor_10.json.gz”. We can then upload this json file to mongodb using the “mongoimport” application.

Note:
I assume you have an instance of MongoDB running.

```
gunzip donor_10.json.gz & mongoimport -d test_database -c
   protocol_capture --file donor_10.json
```

Here I have made a database called test_database and a collection called protocol_capture. First thing to do is remove redundancies within mongo. To do that, I can make a simple index on the “Input Sequence” key and drop duplicates.

```
db.protocol_capture.ensureIndex({'Input Sequence':1},{unique
    :true, dropDups:true})
```

The next thing I want to do is remove non-productive sequences. PyIg outputs a productive field, using mongo, I can tell the database to drop any “document” that contains a stop codon in the HCDR3.

```
db.protocol_capture.remove({"Productive CDR3":"False"})
```

For ROSETTA, I want only the thirty length HCDR3s. I can use “mongoexport” for this query along with an ‘awk’ statement to get out the 30 length fasta files.

```
mongoexport -d test_database -c protocol_capture -q '{"CDR3
   AA Length":30}' --csv -fields "Sequence ID","CDR3 AA" |
   awk -F "," '{gsub("\"","",$1);gsub("\"","", $2);print(">"
   $1"\n"$2)}' > 30_length.fasta
```

This fasta file will be used in the remainder of the protocol. I will not go over all the different types of analysis I can do with this database for the purpose of this protocol.

VI.5.2.2 PSSM Heuristics

Using the fasta file “30_length.fasta” generated in the previous section, a quick script written in python to get random 2,000 sequences is shown below.

```
from Bio import SeqIO
import random
handle = open('30_length.fasta')
records = SeqIO.parse(handle, "fasta")
dictionary = {}
for seq in records:
   dictionary[seq.id] = str(seq.seq)
random_dict = random.sample(dictionary.items(),2000)

with open('2000\_sequences.fasta') as f:
    for seq in random_dict:
        f.write(``>{0}\n{1}\n''.format(seq[0],seq[1]))
```

Using ROSETTADESIGN, I will generate 2,000 resfiles that tell the packer to mutate the HCDR3 into each of the entries in the fasta file. To do this, there is a script available from the Meiler lab called “fasta_to_resfile.py” which will generate the resfiles necessary.

```
fasta_into_res.py random_2000.fasta 95 126 H 0
```

The rest of the protocol uses ROSETTASCRIPTS to do the design. For ROSETTASCRIPTS, you need an xml file, options file, and command line. The xml file is a scripting file that tells ROSETTA specific set of instructions (here I will name it threading.xml). The first step is relatively simple only doing a design and a full-atom minimization (called relax).

```
<dock_design>
    <SCOREFXNS>
    </SCOREFXNS>
    <FILTERS>
    </FILTERS>
    <TASKOPERATIONS>
     <InitializeFromCommandline name=ifcl/>
     <ReadResfile name=rr filename=%%resfiles%% />
    </TASKOPERATIONS>
        <MOVERS>
         <PackRotamersMover name=pr scorefxn=score12
            task_operations=ifcl,rr/>
         <FastRelax name=fr task_operations=ifcl/>
        </MOVERS>
        <APPLY_TO_POSE>
        </APPLY_TO_POSE>
        <PROTOCOLS>
```

```
        <Add mover_name=pr />
        <Add mover_name=fr />
      </PROTOCOLS>
</dock_design>
```

The only caveat here is that we specify the resfile as a variable so the protocol does not have to be hard-coded. The command line will specify each resfile to give to the job. First, an option file must be produced as a simple text file.

```
-in
    -path
        -database /my/rosetta/database/path
-out
    -pdb_gz
-parser
    -protocol threading.xml
-s input_pg9_no_antigen.pdb
-nstruct 100
```

Lastly the command line which will include the resfile as an option (named run.csh).

```
#!/bin/csh
set res = $1
set out = `basename $res .resfile`
mpiexec -n 101 my/rosetta/pathrosetta_scripts.mpistatic.
   linuxgccrelease @flags.txt -out:prefix $out -parser:
   script_vars resfiles=${res} > out.log
```

And to run the command I simply use an awk script to generate all the commands.

```
ls *resfile | awk '{system( "run.csh "$1 ) }'
```

Note: This should be run on a computer cluster as 100 processors are needed in the above protocol

The next step is to run the output PDB files and generate a position-specific scoring matrix. This is easily accomplished with the "create_pssm_from_threading.py" script that is also found in the Meiler lab scripts repository. A simple command to generate a PSSM.

```
./create_pssm_from_threading.py -g -r resfiles/donor\_10\
   _991403.resfile -n 2000.p3sm *.pdbs
```

The output .p3sm can now be used to predict the top N sequences from the 30_length.fasta generated earlier in the protocol.

```
./create_pssm_from_threading.py -r resfiles/donor\_10\
   _2832895.resfile -s -p 2000.p3sm 30\_length.fasta
```

This generates a file called "scored_fasta.output". I use awk and some other gnu commands to get the top 1000 scored fasta files.

```
sort -nk 3 scored_fasta.output | head -n 1000 | awk '{print
   (">"$1"\n"$2)}' > top1000.fasta
```

Finally, I can make all the resfiles using the same command as before.

```
fasta_into_res.py top1000.fasta 95 125 H 0
```

For the full design protocol using these sequences and resfiles. See the next section on PG9 redesign (subsection VI.5.3).

VI.5.3 Chapter III - PG9 Design

This protocol capture will detail the how to use ROSETTADESIGN to predict mutations that enhance specificity. This accompanies the manuscript Willis *et al.* ***Nature Med.*** (submitted). It assumes that you have a ROSETTA license from www.rosettacommons.org.

VI.5.3.1 Preparing the input files

Using PG9/CAP45 complex, I have prepared a ROSETTA compatible file called PG9_input.pdb. This has spcecial identifiers for the glycans that ROSETTA's database can understand. To create your own glycan input, an excellent protocol capture is provided in an accompanying manuscript by Doug Renfrew (Renfrew et al., 2012).

The design protocol used runs through the following steps.

- Favor native residue - gives bonuses to sequences which match PG9*wt*
- Design/minimize/dock iteratively
- Constrain movements so glycans retain input position
- Relax the energy of the structure
- Re-dock
- Score binding energy and structure energy

For this redesign we need several input files. The XML script, options file, residue file, and constraint file. The most complex of which, the XML file, informs Rosettaof our protocol. Use the following .xml file which is found under:

```
/input_files/threading_design.xml
```

XML-File

```
<dock_design>
  <SCOREFXNS>
    Redefine scoring function to take in constraints
    <scorewts weights=talaris2013>
      <Reweight scoretype=atom_pair_constraint weight=0.5/>
    </scorewts>
    <scoredockwts weights=talaris2013 patch=docking>
```

```
    <Reweight scoretype=atom_pair_constraint weight=0.5/>
  </scoredockwts>
</SCOREFXNS>
  <FILTERS>
    DDG filter for design - will design until this
    is satisfied
    <Ddg name=ddg chain_num=2,3 repack=1
      scorefxn=talaris2013 threshold=-20/>
    When docking or minimizing, won't violate atom-pairs
       defined
    in glycan_constraints
    <ScoreType name=atom_pair_constraint scorefxn=
       scoredockwts
      score_type=atom_pair_constraint threshold=100/>
  </FILTERS>
<TASKOPERATIONS>
  <InitializeFromCommandline name=ifcl/>
  <ReadResfile name=rrd filename="input_files/
     normal_design.resfile"/>
</TASKOPERATIONS>
<MOVERS>
  Gives bonuses for input residues
  <FavorSequenceProfile name=fsp scaling="prob"
    weight=1.5 use_current=1
    matrix="IDENTITY" scorefxns=scorewts/>
  Penalizes movements to far away from
  glycan
  <ConstraintSetMover name=pair_on
    cst_file="input_files/glycan_constraints.cst"/>
  Design step that takes in residue file
  <PackRotamersMover name=pr_design scorefxn=scorewts
    task_operations=rrd,ifcl/>
  Turns of penalization
  <ConstraintSetMover name=pair_off cst_file=none/>
  Docks protein around interface
  <Docking name=dock score_high=scoredockwts fullatom=1
    local_refine=1 jumps=1 task_operations=ifcl/>
  Docks until MonteCarlo criterion is satisfied
  <GenericMonteCarlo name=gmc_dock mover_name=dock
    filter_name=atom_pair_constraint drift=0/>
  Minimize energy of protein
  <MinMover name=min scorefxn=scorewts chi=1 bb=1 jump=1/>
  Design protocol
  <ParsedProtocol name=design_pp>
    <Add mover=pair_on/>
```

```
      <Add mover=pr_design/>
      <Add mover=gmc_dock/>
      <Add mover=min/>
      <Add mover=pair_off/>
    </ParsedProtocol>
    Run design until binding energy threshold is satisfied
    <GenericMonteCarlo name=gmc_design
      mover_name=design_pp filter_name=ddg drift=False/>
    Relax protein
    <FastRelax name=fr scorefxn=scorewts
      task_operations=ifcl/>
    Get DDG
    <ddG name=per_ddg  chain_name="H,L"
      scorefxn=talaris2013 per_residue_ddg=1/>
    </MOVERS>
    <APPLY_TO_POSE>
    </APPLY_TO_POSE>
    <PROTOCOLS>
      Ordered list of steps for the protocol
      each defined in the mover or filter definitions
       <Add mover_name=fsp/>
       <Add mover_name=gmc_design/>
       <Add mover_name=pair_on/>
       <Add mover_name=fr/>
       <Add mover_name=pair_off/>
       <Add mover_name=per_ddg/>
       <Add filter_name=ddg/>
    </PROTOCOLS>
</dock_design>
```

The behavior of the these instructions is described fully in (Fleishman et al., 2011a). They are divided up into a set of movers, filters and task of operations. All of the movers and filters along with their options are explained at the Rosetta Commons users guide (https://www.rosettacommons.org/docs/latest/RosettaScripts.html).

Options-File

The options file are passed to the application. Defines output and input options as well as other options which can't be defined in the XML file.

```
-s input_files/pg9_input.pdb #input PDB
-nstruct 200 #the number of output models to generate
-docking
    -sc_min #minimize side chains during docking
-parser:protocol input_files/threading_design.xml
-ex1 #Rotmer library 1
-ex2 #Rotmer library 2
```

```
-ex1aro #Rotamer aromatic 1
-out:path
    -pdb ./output_files/
    -score /dev/null/
```

Each option is explained with a # comment.

Residue File
The residue file tells the packer how to design the protein. The first line lets the packer use the side chains of the input PDB even if they are not in the rotamer libraries. The "NATAA" lines tells the packer to use input amino acid for everything not defined under start. In other words it will only design everything under start. The first column is the residue number, the second is the chain, and "ALLAA" tells the packer to use all amino acid identities at this position. For complete documentation of the resfile, visit https://www.rosettacommons.org/docs/latest/resfiles.html

```
USE_INPUT_SC
NATAA
EX 1 EX 2
start
96 H ALLAA
97 H ALLAA
98 H ALLAA
99 H ALLAA
100 H ALLAA
101 H ALLAA
102 H ALLAA
103 H ALLAA
104 H ALLAA
105 H ALLAA
106 H ALLAA
107 H ALLAA
108 H ALLAA
109 H ALLAA
110 H ALLAA
111 H ALLAA
112 H ALLAA
113 H ALLAA
114 H ALLAA
115 H ALLAA
116 H ALLAA
117 H ALLAA
118 H ALLAA
119 H ALLAA
120 H ALLAA
121 H ALLAA
```

```
122 H ALLAA
123 H ALLAA
124 H ALLAA
125 H ALLAA
```

Constraint File
The constraint file ensures that the glycan's are involved in binding. The torsional angles of the glycan can cause major structural perturbations.

```
AtomPair NZ 57H O31 29G BOUNDED 0 2.57 0.2 0.5 tag
AtomPair O 54H O32 29G BOUNDED 0 3.71 0.2 0.5 tag
AtomPair O 55H OS1 29G BOUNDED 0 4.30 0.2 0.5 tag
AtomPair ND2 73H O71 29G BOUNDED 0 4.02 0.2 0.5 tag
AtomPair OD1 52L O81 33G BOUNDED 0.5 2.96 0.2 0.5 tag
AtomPair OG 34L O81 33G BOUNDED 0.5 2.28 0.2 0.5 tag
AtomPair NH2 30H NZ 41G BOUNDED 0 4.85 0.2 0.5 tag
```

The constrain file syntax is found in the documentation - (https://www.rosettacommons.org/docs/latest/constraint-file.html). Briefly, I define two atoms with the input crystal structure distances. If these are violated, then there is an energetic penalty.

VI.5.3.2 Running ROSETTA

To run ROSETTA, I use an application called ROSETTASCRIPTS (Fleishman et al., 2011a). Since we have defined all the input files. Running the application only requires passing the options file.

```
my/path/to/rosetta/source/bin/rosetta_scripts.
    myoperatingsystem @input_files/threading.txt -database my
    /path/to/rosetta/database/
```

This protocol generates 200 models each taking approximately 1 hour to complete. It is best to run this protocol on a computational cluster with each node producing a separate model (-nstruct 1). All files are output into a directory output models/. There are 200 pre-generated models for analysis.

VI.5.3.3 Analyzing Models

There are three scripts in the /analysis folder that are used to analyze the mutations. Score_-vs_rmsd full.py will give all the models energies as well as how much they deviated from the original structure. Get_per_ddg.py will give all of the binding energies decomposed by residues. Scores_decomposed_by_resfile.py will decompose the energies of HCDR3 loop. They are each run using the following.

```
score_vs_rmsd_full.py âĹŠm âĹŠn ../input_files/pg9_input.pdb
    âĹŠo s_v_rmsd âĹŠr ../input_files/normal_design.resfile
    ../output_files/âĹŮ.pdb
```

```
get_per_ddg.py –m –o per_ddg ../output_files/âĹŮ.pdb

scores_decompose_by_resfile.py âĹŠm âĹŠo HCDR3 âĹŠr ../
    input_files/normal_design.res
```

These will yield a series of data files that can be uploaded to a database or in a spreadsheet viewer. The complex queries I used to check energies between wt and mutations are beyond the scope of a protocol capture. But you can contact jwillis0720@gmail.com if you need additional guidance.

References

Ackerman, M. and Alter, G. (2013). Mapping the journey to an HIV vaccine. *The New England Journal of Medicine*, 369(4):389–391.

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

Albert, J., Abrahamsson, B., Nagy, K., Aurelius, E., Gaines, H., Nyström, G., and Fenyö, E. M. (1990). Rapid development of isolate-specific neutralizing antibodies after primary HIV-1 infection and consequent emergence of virus variants which resist neutralization by autologous sera. *AIDS (London, England)*, 4(2):107–112.

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. *PNAS*, 79(13):4118–4122.

Alt, F. W., Oltz, E. M., Young, F., Gorman, J., Taccioli, G., and Chen, J. (1992). VDJ recombination. *Immunology today*, 13(8):306–314.

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., Baker, D., Strong, R. K., and Schief, W. R. (2011). Computation-guided backbone grafting of a discontinuous motif onto a protein scaffold. *Science (New York, NY)*, 334(6054):373–376.

Babor, M. and Kortemme, T. (2009). Multi-constraint computational design suggests that native sequences of germline antibody H3 loops are nearly optimal for conformational flexibility. *Proteins*, 75(4):846–858.

Baker, D. (2014). Protein folding, structure prediction and design. *Biochemical Society transactions*, 42(2):225–229.

Barré-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vézinet-Brun, F., Rouzioux, C., Rozenbaum, W., and Montagnier, L. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science (New York, NY)*, 220(4599):868–871.

Bartram, A. K., Lynch, M. D. J., Stearns, J. C., Moreno-Hagelsieb, G., and Neufeld, J. D. (2011). Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end illumina reads. *Applied and environmental microbiology*, 77(11):3846–3852.

Berek, C. and Milstein, C. (1988). The dynamic nature of the antibody repertoire. *Immunological reviews*, 105:5–26.

Binley, J. M., Ban, Y.-E. A., Crooks, E. T., Eggink, D., Osawa, K., Schief, W. R., and Sanders, R. W. (2010). Role of complex carbohydrates in human immunodeficiency virus type 1 infection and resistance to antibody neutralization. *Journal of Virology*, 84(11):5637–5655.

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., Wood, B., Nathe, C., Richman, D., Tomaras, G. D., Bibollet-Ruche, F., Robinson, J. E., Morris, L., Shaw, G. M., Montefiori, D. C., and Mascola, J. R. (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. *Journal of Virology*, 82(23):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., Petropoulos, C. J., and Burton, D. R. (2004). Comprehensive cross-clade neutralization analysis of a panel of anti-human immunodeficiency virus type 1 monoclonal antibodies. *Journal of Virology*, 78(23):13232–13252.

Blish, C. A., Nedellec, R., Mandaliya, K., Mosier, D. E., and Overbaugh, J. (2007). HIV-1 subtype A envelope variants from early in infection have variable sensitivity to neutralization and to inhibitors of viral entry. *AIDS (London, England)*, 21(6):693–702.

Bonsignori, M., Hwang, K.-K., Chen, X., Tsao, C.-Y., Morris, L., Gray, E., Marshall, D. J., Crump, J. A., Kapiga, S. H., Sam, N. E., Sinangil, F., Pancera, M., Yongping, Y., Zhang, B., Zhu, J., Kwong, P. D., O'dell, S., Mascola, J. R., Wu, L., Nabel, G. J., Phogat, S., Seaman, M. S., Whitesides, J. F., Moody, M. A., Kelsoe, G., Yang, X., Sodroski, J., Shaw, G. M., Montefiori, D. C., Kepler, T. B., Tomaras, G. D., Alam, S. M., Liao, H.-X., and Haynes, B. F. (2011). Analysis of a clonal lineage of HIV-1 envelope V2/V3 conformational epitope-specific broadly neutralizing antibodies and their inferred unmutated common ancestors. *Journal of Virology*, 85(19):9998–10009.

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

Bradley, P., Chivian, D., Meiler, J., Misura, K. M. S., Rohl, C. A., Schief, W. R., Wedemeyer, W. J., Schueler-Furman, O., Murphy, P., Schonbrun, J., Strauss, C. E. M., and Baker, D. (2003). Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins*, 53 Suppl 6:457–468.

Bradley, P., Malmström, L., Qian, B., Schonbrun, J., Chivian, D., Kim, D. E., Meiler, J., Misura, K. M. S., and Baker, D. (2005a). Free modeling with Rosetta in CASP6. *Proteins*, 61 Suppl 7:128–134.

Bradley, P., Misura, K. M. S., and Baker, D. (2005b). Toward high-resolution de novo structure prediction for small proteins. *Science (New York, NY)*, 309(5742):1868–1871.

Brenner, S. and Milstein, C. (1966). Origin of antibody variation. *Nature*, 211(5046):242–243.

Briney, B. S. (2012). *The Development and Genetic Origin of Broadly Neutralizing HIV Antibodies*. PhD thesis, Vanderbilt University, Nashville.

Briney, B. S., Willis, J. R., and Crowe, J. E. (2012). Human peripheral blood antibodies with long HCDR3s are established primarily at original recombination using a limited subset of germline genes. *PloS one*, 7(5):e36750.

Brochet, X., Lefranc, M.-P., and Giudicelli, V. (2008). IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic acids research*, 36(Web Server issue):W503–8.

Burton, D. R., Ahmed, R., Barouch, D. H., Butera, S. T., Crotty, S., Godzik, A., Kaufmann, D. E., McElrath, M. J., Nussenzweig, M. C., Pulendran, B., Scanlan, C. N., Schief, W. R., Silvestri, G., Streeck, H., Walker, B. D., Walker, L. M., Ward, A. B., Wilson, I. A., and Wyatt, R. (2012). A Blueprint for HIV Vaccine Discovery. *Cell host & microbe*, 12(4):396–407.

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 (New York, NY)*, 266(5187):1024–1027.

Burton, D. R., Stanfield, R. L., and Wilson, I. A. (2005). Antibody vs. HIV in a clash of evolutionary titans. *PNAS*, 102(42):14943–14948.

Casali, P., Pal, Z., Xu, Z., and Zan, H. (2006). DNA repair in antibody somatic hypermutation. *Trends in immunology*, 27(7):313–321.

Chen, C., Stenzel-Poore, M. P., and Rittenberg, M. B. (1991). Natural auto- and polyreactive antibodies differing from antigen-induced antibodies in the H chain CDR3. *Journal of immunology (Baltimore, Md : 1950)*, 147(7):2359–2367.

Choe, H., Li, W., Wright, P. L., Vasilieva, N., Venturi, M., Huang, C.-c., Grundner, C., Dorfman, T., Zwick, M. B., Wang, L., Rosenberg, E. S., Kwong, P. D., Burton, D. R., Robinson, J. E., Sodroski, J. G., and Farzan, M. (2003). Tyrosine sulfation of human antibodies contributes to recognition of the CCR5 binding region of HIV-1 gp120. *Cell*, 114(2):161–170.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, 25(11):1422–1423.

Collins, A. M., Sewell, W. A., and Edwards, M. R. (2003). Immunoglobulin gene rearrangement, repertoire diversity, and the allergic response. *Pharmacology & therapeutics*, 100(2):157–170.

Combs, S. and Meiler, J. (2012). Partial Covalent Interactions in Protein Design. *Protein science : a publication of the Protein Society*, 21:230–231.

Correia, B. E., Ban, Y.-E. A., Friend, D. J., Ellingson, K., Xu, H., Boni, E., Bradley-Hewitt, T., Bruhn-Johannsen, J. F., Stamatatos, L., Strong, R. K., and Schief, W. R. (2011a). Computational protein design using flexible backbone remodeling and resurfacing: case studies in structure-based antigen design. *Journal of Molecular Biology*, 405(1):284–297.

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., Burke, K. Y., Bradley-Hewitt, T., Bruhn-Johannsen, J. F., Kalyuzhniy, O., Baker, D., Strong, R. K., Stamatatos, L., and Schief, W. R. (2010). Computational design of epitope-scaffolds allows induction of antibodies specific for a poorly immunogenic HIV vaccine epitope. *Structure (London, England : 1993)*, 18(9):1116–1126.

Correia, B. E., Bates, J. T., Loomis, R. J., Baneyx, G., Carrico, C., Jardine, J. G., Rupert, P., Correnti, C., Kalyuzhniy, O., Vittal, V., Connell, M. J., Stevens, E., Schroeter, A., Chen, M., Macpherson, S., Serra, A. M., Adachi, Y., Holmes, M. A., Li, Y., Klevit, R. E., Graham, B. S., Wyatt, R. T., Baker, D., Strong, R. K., Crowe, J. E., Johnson, P. R., and Schief, W. R. (2014). Proof of principle for epitope-focused vaccine design. *Nature*.

Correia, B. E., Holmes, M. A., Huang, P.-S., Strong, R. K., and Schief, W. R. (2011b). High-resolution structure prediction of a circular permutation loop. *Protein science : a publication of the Protein Society*, 20(11):1929–1934.

Corti, D., Bianchi, S., Vanzetta, F., Minola, A., Perez, L., Agatic, G., Guarino, B., Silacci, C., Marcandalli, J., Marsland, B. J., Piralla, A., Percivalle, E., Sallusto, F., Baldanti, F., and Lanzavecchia, A. (2013). Cross-neutralization of four paramyxoviruses by a human monoclonal antibody. *Nature*, 501(7467):439–443.

Corti, D. and Lanzavecchia, A. (2013). Broadly neutralizing antiviral antibodies. *Annual Review of Immunology*, 31:705–742.

Corti, D., Voss, J., Gamblin, S. J., Codoni, G., Macagno, A., Jarrossay, D., Vachieri, S. G., Pinna, D., Minola, A., Vanzetta, F., Silacci, C., Fernandez-Rodriguez, B. M., Agatic, G., Bianchi, S., Giacchetto-Sasselli, I., Calder, L., Sallusto, F., Collins, P., Haire, L. F., Temperton, N., Langedijk, J. P. M., Skehel, J. J., and Lanzavecchia, A. (2011). A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. *Science (New York, NY)*, 333(6044):850–856.

Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome research*, 14(6):1188–1190.

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. *Journal of immunology (Baltimore, Md : 1950)*, 154(9):4526–4535.

Dantas, G., Kuhlman, B., Callender, D., Wong, M., and Baker, D. (2003). A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *Journal of Molecular Biology*, 332(2):449–460.

Das, R. and Baker, D. (2008). Macromolecular modeling with Rosetta. *Annual Review of Biochemistry*, 77:363–382.

Das, R., Qian, B., Raman, S., Vernon, R., Thompson, J., Bradley, P., Khare, S., Tyka, M. D., Bhat, D., Chivian, D., Kim, D. E., Sheffler, W. H., Malmström, L., Wollacott, A. M., Wang, C., Andre, I., and Baker, D. (2007). Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. *Proteins*, 69 Suppl 8:118–128.

Davies, D. R. and Cohen, G. H. (1996). Interactions of protein antigens with antibodies. *PNAS*, 93(1):7–12.

Davis, I. W. and Baker, D. (2009). RosettaLigand docking with full ligand and receptor flexibility. *Journal of Molecular Biology*, 385(2):381–392.

Davis, I. W., Raha, K., Head, M. S., and Baker, D. (2009). Blind docking of pharmaceutically relevant compounds using RosettaLigand. *Protein science : a publication of the Protein Society*, 18(9):1998–2002.

De, S., Sur, K., and Dasgupta, S. (2005). Characterization of the nonregular regions of proteins by a contortion index. *Biopolymers*, 79(2):63–73.

DeCamp, A., DeCamp, A., Hraber, P., Hraber, P., Bailer, R. T., Bailer, R. T., Seaman, M. S., Seaman, M. S., Ochsenbauer, C., Ochsenbauer, C., Kappes, J., Kappes, J., Gottardo, R., Gottardo, R., Edlefsen, P., Edlefsen, P., Self, S., Self, S., Tang, H., Tang, H., Greene, K., Greene, K., Gao, H., Gao, H., Daniell, X., Daniell, X., Sarzotti-Kelsoe, M., Sarzotti-Kelsoe, M., Gorny, M. K., Gorny, M. K., Zolla-Pazner, S., Zolla-Pazner, S., LaBranche, C. C., LaBranche, C. C., Mascola, J. R., Mascola, J. R., Korber, B. T., Korber, B. T., Montefiori, D. C., and Montefiori, D. C. (2014). Global Panel of HIV-1 Env Reference Strains for Standardized Assessments of Vaccine-Elicited Neutralizing Antibodies. *Journal of Virology*, 88(5):2489–2507.

Der, B. S. and Kuhlman, B. (2011). Biochemistry. From computational design to a protein that binds. *Science (New York, NY)*, 332(6031):801–802.

Di Noia, J. and Neuberger, M. S. (2002). Altering the pathway of immunoglobulin hypermutation by inhibiting uracil-DNA glycosylase. *Nature*, 419(6902):43–48.

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., and Bjorkman, P. J. (2011). Increasing the potency and breadth of an HIV antibody by using structure-based rational design. *Science (New York, NY)*, 334(6060):1289–1293.

Doores, K. J., Doores, K. J., Burton, D. R., and Burton, D. R. (2010). Variable loop glycan dependency of the broad and potent HIV-1-neutralizing antibodies PG9 and PG16. *Journal of Virology*, 84(20):10510–10521.

Doria-Rose, N. A., Georgiev, I., O'dell, S., Chuang, G.-Y., Staupe, R. P., McLellan, J. S., Gorman, J., Pancera, M., Bonsignori, M., Haynes, B. F., Burton, D. R., Koff, W. C., Kwong, P. D., and Mascola, J. R. (2012). A short segment of the HIV-1 gp120 V1/V2 region is a major determinant of resistance to V1/V2 neutralizing antibodies. *Journal of Virology*, 86(15):8319–8323.

Doria-Rose, N. A., Klein, R. M., Daniels, M. G., O'dell, S., Nason, M., Lapedes, A., Bhattacharya, T., Migueles, S. A., Wyatt, R. T., Korber, B. T., Mascola, J. R., and Connors, M. (2010). Breadth of human immunodeficiency virus-specific neutralizing activity in sera: clustering analysis and association with clinical variables. *Journal of Virology*, 84(3):1631–1636.

Doria-Rose, N. A., Schramm, C. A., Gorman, J., Moore, P. L., Bhiman, J. N., Dekosky, B. J., Ernandes, M. J., Georgiev, I. S., Kim, H. J., Pancera, M., Staupe, R. P., Altae-Tran, H. R., Bailer, R. T., Crooks, E. T., Cupo, A., Druz, A., Garrett, N. J., Hoi, K. H., Kong, R., Louder, M. K., Longo, N. S., McKee, K., Nonyane, M., O'dell, S., Roark, R. S., Rudicell, R. S., Schmidt, S. D., Sheward, D. J., Soto, C., Wibmer, C. K., Yang, Y., Zhang, Z., Sequencing, N. C., Mullikin, J. C., Binley, J. M., Sanders, R. W., Wilson, I. A., Moore, J. P., Ward, A. B., Georgiou, G., Williamson, C., Abdool Karim, S. S., Morris, L., Kwong, P. D., Shapiro, L., and Mascola, J. R. (2014). Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature*.

Dunbrack, R. L. and Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein science : a publication of the Protein Society*, 6(8):1661–1681.

Dunbrack, R. L. and Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *Journal of Molecular Biology*, 230(2):543–574.

Durham, E., Dorr, B., Woetzel, N., Staritzbichler, R., and Meiler, J. (2009). Solvent accessible surface area approximations for rapid and accurate protein structure prediction. *Journal of molecular modeling*, 15(9):1093–1108.

Fields, B. N., Fields, B. N., Knipe, D. M., Knipe, D. M., Howley, P. M., and Howley, P. M. (2007). *Fields' Virology*, (v. 2).

Finn, J. A. and Crowe, J. E. (2013). Impact of new sequencing technologies on studies of the human B cell repertoire. *Current opinion in immunology*, 25(5):613–618.

Fleishman, S. J., Leaver-Fay, A., Corn, J. E., Strauch, E.-M., Khare, S. D., Koga, N., Ashworth, J., Murphy, P., Richter, F., Lemmon, G., Meiler, J., and Baker, D. (2011a). RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PloS one*, 6(6):e20161–.

Fleishman, S. J., Whitehead, T. A., Ekiert, D. C., Dreyfus, C., Corn, J. E., Strauch, E.-M., Wilson, I. A., and Baker, D. (2011b). Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science (New York, NY)*, 332(6031):816–821.

Foote, J. and Milstein, C. (1994). Conformational isomerism and the diversity of antibodies. *PNAS*, 91(22):10370–10374.

Fortenberry, C., Bowman, E. A., Proffitt, W., Dorr, B., Combs, S., Harp, J., Mizoue, L., and Meiler, J. (2011). Exploring Symmetry as an Avenue to the Computational Design of Large Protein Domains (vol 45, pg 18026, 2011). *Journal of the American Chemical Society*, 133(51):21028–21028.

Georgiev, I. S., Doria-Rose, N. A., Zhou, T., Kwon, Y. D., Staupe, R. P., Moquin, S., Chuang, G.-Y., Louder, M. K., Schmidt, S. D., Altae-Tran, H. R., Bailer, R. T., McKee, K., Nason, M., O'dell, S., Ofek, G., Pancera, M., Srivatsan, S., Shapiro, L., Connors, M., Migueles, S. A., Morris, L., Nishimura, Y., Martin, M. A., Mascola, J. R., and Kwong, P. D. (2013). Delineating antibody recognition in polyclonal sera from patterns of HIV-1 isolate neutralization. *Science (New York, NY)*, 340(6133):751–756.

Gordon, D. B., Marshall, S. A., and Mayo, S. L. (1999). Energy functions for protein design. *Current opinion in structural biology*, 9(4):509–513.

Gottlieb, M. S., Schroff, R., Schanker, H. M., Weisman, J. D., Fan, P. T., Wolf, R. A., and Saxon, A. (1981). Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. *The New England Journal of Medicine*, 305(24):1425–1431.

Gray, E. S., Madiga, M. C., Hermanus, T., Moore, P. L., Wibmer, C. K., Tumba, N. L., Werner, L., Mlisana, K., Sibeko, S., Williamson, C., Abdool Karim, S. S., Morris, L., and and the CAPRISA002 Study Team (2011). The neutralization breadth of HIV-1 develops incrementally over four years and is associated with CD4+ T cell decline and high viral load during acute infection. *Journal of Virology*, 85(10):4828–4840.

Gray, E. S., Madiga, M. C., Moore, P. L., Mlisana, K., Abdool Karim, S. S., Binley, J. M., Shaw, G. M., Mascola, J. R., and Morris, L. (2009). Broad neutralization of human immunodeficiency virus type 1 mediated by plasma antibodies against the gp41 membrane proximal external region. *Journal of Virology*, 83(21):11265–11274.

Gray, E. S., Moore, P. L., Choge, I. A., Decker, J. M., Bibollet-Ruche, F., Li, H., Leseka, N., Treurnicht, F., Mlisana, K., Shaw, G. M., Karim, S. S. A., Williamson, C., Morris, L., and CAPRISA 002 Study Team (2007). Neutralizing antibody responses in acute human immunodeficiency virus type 1 subtype C infection. *Journal of Virology*, 81(12):6187–6196.

Harindranath, N., Ikematsu, H., Notkins, A. L., and Casali, P. (1993). Structure of the VH and VL segments of polyreactive and monoreactive human natural antibodies to HIV-1 and Escherichia coli beta-galactosidase. *International immunology*, 5(12):1523–1533.

Harris, A., Harris, A., Borgnia, M. J., Borgnia, M. J., Shi, D., Shi, D., Bartesaghi, A., Bartesaghi, A., He, H., He, H., Pejchal, R., Pejchal, R., Kang, Y. K., Kang, Y. K., Depetris, R., Depetris, R., Marozsan, A. J., Marozsan, A. J., Sanders, R. W., Sanders, R. W., Klasse, P. J., Klasse, P. J., Milne, J. L. S., Milne, J. L. S., Wilson, I. A., Wilson, I. A., Olson, W. C., Olson, W. C., Moore, J. P., Moore, J. P., Subramaniam, S., and Subramaniam, S. (2011). Trimeric HIV-1 glycoprotein gp140 immunogens and native HIV-1 envelope glycoproteins display the same closed and open quaternary molecular architectures. *PNAS*, 108(28):11440–11445.

Harris, L. J., Larson, S. B., Hasel, K. W., and McPherson, A. (1997). Refined structure of an intact IgG2a monoclonal antibody. *Biochemistry*, 36(7):1581–1597.

Haynes, B. F., Gilbert, P. B., McElrath, M. J., Zolla-Pazner, S., Tomaras, G. D., Alam, S. M., Evans, D. T., Montefiori, D. C., Karnasuta, C., Sutthent, R., Liao, H.-X., DeVico, A. L., Lewis, G. K., Williams, C., Pinter, A., Fong, Y., Janes, H., DeCamp, A., Huang, Y., Rao, M., Billings, E., Karasavvas, N., Robb, M. L., Ngauy, V., de Souza, M. S., Paris, R., Ferrari, G., Bailer, R. T., Soderberg, K. A., Andrews, C., Berman, P. W., Frahm, N., De Rosa, S. C., Alpert, M. D., Yates, N. L., Shen, X., Koup, R. A., Pitisuttithum, P., Kaewkungwal, J., Nitayaphan, S., Rerks-Ngarm, S., Michael, N. L., and Kim, J. H. (2012a). Immune-correlates analysis of an HIV-1 vaccine efficacy trial. *The New England Journal of Medicine*, 366(14):1275–1286.

Haynes, B. F., Kelsoe, G., Harrison, S. C., and Kepler, T. B. (2012b). B-cell-lineage immunogen design in vaccine development with HIV-1 as a case study. *Nature Biotechnology*, 30(5):423–433.

Hemelaar, J. (2012). The origin and diversity of the HIV-1 pandemic. *Trends in Molecular Medicine*, 18(3):182–192.

Hesse, J. E., Lieber, M. R., Mizuuchi, K., and Gellert, M. (1989). V(D)J recombination: a functional definition of the joining signals. *Genes & Development*, 3(7):1053–1061.

Hessell, A. J. and Haigwood, N. L. (2012). Neutralizing antibodies and control of HIV: moves and countermoves. *Current HIV/AIDS reports*, 9(1):64–72.

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. (2009a). Effective, low-titer antibody protection against low-dose repeated mucosal SHIV challenge in macaques. *Nature medicine*, 15(8):951–954.

Hessell, A. J., Rakasz, E. G., Poignard, P., Hangartner, L., Landucci, G., Forthal, D. N., Koff, W. C., Watkins, D. I., and Burton, D. R. (2009b). Broadly neutralizing human anti-HIV antibody 2G12 is effective in protection against mucosal SHIV challenge even at low serum neutralizing titers. *PLoS Pathogens*, 5(5):e1000433.

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., and Burton, D. R. (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. *Journal of Virology*, 84(3):1302–1313.

Ho, D. D., Neumann, A. U., Perelson, A. S., Chen, W., Leonard, J. M., and Markowitz, M. (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature*, 373(6510):123–126.

Huang, J., Ofek, G., Laub, L., Louder, M. K., Doria-Rose, N. A., Longo, N. S., Imamichi, H., Bailer, R. T., Chakrabarti, B., Sharma, S. K., Alam, S. M., Wang, T., Yang, Y., Zhang, B., Migueles, S. A., Wyatt, R., Haynes, B. F., Kwong, P. D., Mascola, J. R., and Connors, M. (2012). Broad and potent neutralization of HIV-1 by a gp41-specific human antibody. *Nature*, 491(7424):406–412.

Huang, P.-S., Love, J. J., and Mayo, S. L. (2007). A de novo designed protein protein interface. *Protein science : a publication of the Protein Society*, 16(12):2770–2774.

Humphris, E. L. and Kortemme, T. (2007). Design of multi-specificity in protein interfaces. *PLoS computational biology*, 3(8):e164.

Ichiyoshi, Y. and Casali, P. (1994). Analysis of the structural correlates for antibody polyreactivity by multiple reassortments of chimeric human immunoglobulin heavy and light chain V segments. *The Journal of experimental medicine*, 180(3):885–895.

Ivanov, I. I., Ivanov, I. I., Schelonka, R. L., Schelonka, R. L., Zhuang, Y., Zhuang, Y., Gartland, G. L., Gartland, G. L., Zemlin, M., Zemlin, M., Schroeder, H. W., 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. *Journal of immunology (Baltimore, Md : 1950)*, 174(12):7773–7780.

James, L. C., Roversi, P., and Tawfik, D. S. (2003). Antibody multispecificity mediated by conformational diversity. *Science (New York, NY)*, 299(5611):1362–1367.

James, L. C. and Tawfik, D. S. (2003). Conformational diversity and protein evolution–a 60-year-old hypothesis revisited. *Trends in biochemical sciences*, 28(7):361–368.

Jardine, J., Julien, J.-P., Menis, S., Ota, T., Kalyuzhniy, O., McGuire, A., Sok, D., Huang, P.-S., Macpherson, S., Jones, M., Nieusma, T., Mathison, J., Baker, D., Ward, A. B., Burton, D. R., Stamatatos, L., Nemazee, D., Wilson, I. A., and Schief, W. R. (2013). Rational HIV Immunogen Design to Target Specific Germline B Cell Receptors. *Science (New York, NY)*.

Javier Guenaga, P. D. G. O. D. B. W. R. S. P. D. K. G. B. K. H. and Wyatt, R. T. (2011). Heterologous Epitope-Scaffold PrimeâĹűBoosting Immuno-Focuses B Cell Responses to the HIV-1 gp41 2F5 Neutralization Determinant. *PloS one*, 6(1).

Jiang, L., Althoff, E. A., Clemente, F. R., Doyle, L., Röthlisberger, D., Zanghellini, A., Gallaher, J. L., Betker, J. L., Tanaka, F., Barbas, C. F., Hilvert, D., Houk, K. N., Stoddard, B. L., and Baker, D. (2008). De novo computational design of retro-aldol enzymes. *Science (New York, NY)*, 319(5868):1387–1391.

Jimenez, R., Salazar, G., Baldridge, K. K., and Romesberg, F. E. (2003). Flexibility and molecular recognition in the immune system. *PNAS*, 100(1):92–97.

Julien, J.-P., Lee, J. H., Cupo, A., Murin, C. D., Derking, R., Hoffenberg, S., Caulfield, M. J., King, C. R., Marozsan, A. J., Klasse, P. J., Sanders, R. W., Moore, J. P., Wilson,

I. A., and Ward, A. B. (2013). Asymmetric recognition of the HIV-1 trimer by broadly neutralizing antibody PG9. *PNAS*, 110(11):4351–4356.

Kaas, Q., Ruiz, M., and Lefranc, M.-P. (2004). IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic acids research*, 32(Database issue):D208–10.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637.

Kaufmann, K. W., Dawson, E. S., Henry, L. K., Field, J. R., Blakely, R. D., and Meiler, J. (2009). Structural determinants of species-selective substrate recognition in human and Drosophila serotonin transporters revealed through computational docking studies. *Proteins*, 74(3):630–642.

Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H., and Meiler, J. (2010). Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry*, 49(14):2987–2998.

Keeble, A. H., Joachimiak, L. A., Maté, M. J., Meenan, N., Kirkpatrick, N., Baker, D., and Kleanthous, C. (2008). Experimental and computational analyses of the energetic basis for dual recognition of immunity proteins by colicin endonucleases. *Journal of Molecular Biology*, 379(4):745–759.

Keele, B. F. (2006). Chimpanzee Reservoirs of Pandemic and Nonpandemic HIV-1. *Science (New York, NY)*, 313(5786):523–526.

Klein, F., Diskin, R., Scheid, J. F., Gaebler, C., Mouquet, H., Georgiev, I. S., Pancera, M., Zhou, T., Incesu, R.-B., Fu, B. Z., Gnanapragasam, P. N. P., Oliveira, T. Y., Seaman, M. S., Kwong, P. D., Bjorkman, P. J., and Nussenzweig, M. C. (2013). Somatic mutations of the immunoglobulin framework are generally required for broad and potent HIV-1 neutralization. *Cell*, 153(1):126–138.

Klein, F., Halper-Stromberg, A., Horwitz, J. A., Gruell, H., Scheid, J. F., Bournazos, S., Mouquet, H., Spatz, L. A., Diskin, R., Abadir, A., Zang, T., Dorner, M., Billerbeck, E., Labitt, R. N., Gaebler, C., Marcovecchio, P. M., Incesu, R.-B., Eisenreich, T. R., Bieniasz, P. D., Seaman, M. S., Bjorkman, P. J., Ravetch, J. V., Ploss, A., and Nussenzweig, M. C. (2012). HIV therapy by a combination of broadly neutralizing antibodies in humanized mice. *Nature*, 492(7427):118–122.

Korber, B., Gaschen, B., Yusim, K., Thakallapally, R., Kesmir, C., and Detours, V. (2001). Evolutionary and immunological implications of contemporary HIV-1 variation. *British medical bulletin*, 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 (New York, NY)*, 237(4820):1351–1355.

Kryshtafovych, A., Fidelis, K., and Moult, J. (2014). CASP10 results compared to those of previous CASP experiments. *Proteins*, 82 Suppl 2:164–174.

Kuhlman, B. and Baker, D. (2000). Native protein sequences are close to optimal for their structures. *PNAS*, 97(19):10383–10388.

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science (New York, NY)*, 302(5649):1364–1368.

Kuhlman, B., O'Neill, J. W., Kim, D. E., Zhang, K. Y. J., and Baker, D. (2002). Accurate computer-based design of a new backbone conformation in the second turn of protein L. *Journal of Molecular Biology*, 315(3):471–477.

Kwon, Y. D., Finzi, A., Wu, X., Dogo-Isonagie, C., Lee, L. K., Moore, L. R., Schmidt, S. D., Stuckey, J., Yang, Y., Zhou, T., Zhu, J., Vicic, D. A., Debnath, A. K., Shapiro, L., Bewley, C. A., Mascola, J. R., Sodroski, J. G., and Kwong, P. D. (2012). Unliganded HIV-1 gp120 core structures assume the CD4-bound conformation with regulation by quaternary interactions and variable loops. *PNAS*, 109(15):5663–5668.

Kwong, P. D. and Mascola, J. R. (2012). Human Antibodies that Neutralize HIV-1: Identification, Structures, and B Cell Ontogenies. *Immunity*, 37(3):412–425.

Kwong, P. D. and Wilson, I. A. (2009). HIV-1 and influenza antibodies: seeing antigens in new ways. *Nature immunology*, 10(6):573–578.

Lange, O. F. and Baker, D. (2012). Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. *Proteins*, 80(3):884–895.

Lange, O. F., Rossi, P., Sgourakis, N. G., Song, Y., Lee, H.-W., Aramini, J. M., Ertekin, A., Xiao, R., Acton, T. B., Montelione, G. T., and Baker, D. (2012). Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *PNAS*, 109(27):10873–10878.

Lanzavecchia, A. and Sallusto, F. (2009). Human B cell memory. *Current opinion in immunology*, 21(3):298–304.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)*, 23(21):2947–2948.

Lazaridis, T. and Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins*, 35(2):133–152.

Leaver-Fay, A., Jacak, R., Stranges, P. B., and Kuhlman, B. (2011a). A generic program for multistate protein design. *PloS one*, 6(7):e20937.

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y.-E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D., and Bradley, P. (2011b). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in enzymology*, 487:545–574.

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

Lemmon, G., Kaufmann, K., and Meiler, J. (2012). Prediction of HIV-1 Protease/Inhibitor Affinity using RosettaLigand. *Chemical biology & drug design*, 79(6):888–896.

Levinthal, C. (1969). Levinthal: How to fold graciously - Google Scholar. *Mossbauer spectroscopy in . . .* .

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

Li, M., Gao, F., Mascola, J. R., Stamatatos, L., Polonis, V. R., Koutsoukos, M., Voss, G., Goepfert, P., Gilbert, P., Greene, K. M., Bilska, M., Kothe, D. L., Salazar-Gonzalez, J. F., Wei, X., Decker, J. M., Hahn, B. H., and Montefiori, D. C. (2005). Human immunodeficiency virus type 1 env clones from acute and early subtype B infections for standardized assessments of vaccine-elicited neutralizing antibodies. *Journal of Virology*, 79(16):10108–10125.

Li, M., Salazar-Gonzalez, J. F., Derdeyn, C. A., Morris, L., Williamson, C., Robinson, J. E., Decker, J. M., Li, Y., Salazar, M. G., Polonis, V. R., Mlisana, K., Karim, S. A., Hong, K., Greene, K. M., Bilska, M., Zhou, J., Allen, S., Chomba, E., Mulenga, J., Vwalika, C., Gao, F., Zhang, M., Korber, B. T. M., Hunter, E., Hahn, B. H., and Montefiori, D. C. (2006). Genetic and Neutralization Properties of Subtype C Human Immunodeficiency Virus Type 1 Molecular env Clones from Acute and Early Heterosexually Acquired Infections in Southern Africa. *Journal of Virology*, 80(23):11776–11790.

Li, Y., Li, H., Yang, F., Smith-Gill, S. J., and Mariuzza, R. A. (2003). X-ray snapshots of the maturation of an antibody response to a protein antigen. *Nature structural biology*, 10(6):482–488.

Li, Y., Migueles, S. A., Welcher, B., Svehla, K., Phogat, A., Louder, M. K., Wu, X., Shaw, G. M., Connors, M., Wyatt, R. T., and Mascola, J. R. (2007). Broad HIV-1 neutralization mediated by CD4-binding site antibodies. *Nature medicine*, 13(9):1032–1034.

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., Graham, B. S., Haynes, B. F., Burton, D. R., Wyatt, R. T., and Mascola, J. R. (2011). Mechanism of Neutralization by the Broadly

Neutralizing HIV-1 Monoclonal Antibody VRC01. *Journal of Virology*, 85(17):8954–8967.

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 & Development*, 18(1):1–11.

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

Liu, J., Bartesaghi, A., Borgnia, M. J., Sapiro, G., and Subramaniam, S. (2008). Molecular architecture of native HIV-1 gp120 trimers. *Nature*, 455(7209):109–113.

Lynch, R. M., Rong, R., Boliar, S., Sethi, A., Li, B., Mulenga, J., Allen, S., Robinson, J. E., Gnanakaran, S., and Derdeyn, C. A. (2011). The B cell response is redundant and highly focused on V1V2 during early subtype C infection in a Zambian seroconverter. *Journal of Virology*, 85(2):905–915.

Ma, Y., Pannicke, U., Schwarz, K., and Lieber, M. R. (2002). Hairpin Opening and Overhang Processing by an Artemis/DNA-Dependent Protein Kinase Complex in Nonhomologous End Joining and V(D)J Recombination. *Cell*, 108(6):781–794.

MacLennan, I. C. (1994). Germinal centers. *Annual review of immunology*, 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. *Immunological reviews*, 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. *Proceedings of the National Academy of Sciences of the United States of America*, 96(24):13926–13931.

Mandell, D. J., Coutsias, E. A., and Kortemme, T. (2009). Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nature methods*, 6(8):551–552.

Manivel, V., Sahoo, N. C., Salunke, D. M., and Rao, K. V. (2000). Maturation of an antibody response is governed by modulations in flexibility of the antigen-combining site. *Immunity*, 13(5):611–620.

Mansilla-Soto, J. and Cortes, P. (2003). VDJ recombination: Artemis and its in vivo role in hairpin opening. *The Journal of experimental medicine*, 197(5):543–547.

Marlow, M. S., Dogan, J., Frederick, K. K., Valentine, K. G., and Wand, A. J. (2010). The role of conformational entropy in molecular recognition by calmodulin. *Nature methods*, 6(5):352–358.

Mascola, J. R., D'Souza, P., Gilbert, P., Hahn, B. H., Haigwood, N. L., Morris, L., Petropoulos, C. J., Polonis, V. R., Sarzotti, M., and Montefiori, D. C. (2005). Recommendations for the design and use of standard virus panels to assess neutralizing antibody responses elicited by candidate human immunodeficiency virus type 1 vaccines. *Journal of Virology*, 79(16):10103–10107.

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., Lu, Y., Robb, M. L., Katinger, H., and Birx, D. L. (1999). Protection of Macaques against pathogenic simian/human immunodeficiency virus 89.6PD by passive transfer of neutralizing antibodies. *Journal of Virology*, 73(5):4009–4018.

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. *The Journal of experimental medicine*, 188(11):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(3):387–395.

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

McLellan, J. S., Pancera, M., Carrico, C., Gorman, J., Julien, J.-P., Khayat, R., Louder, R., Pejchal, R., Sastry, M., Dai, K., O'dell, S., Patel, N., Shahzad-Ul-Hussan, S., Yang, Y., Zhang, B., Zhou, T., Zhu, J., Boyington, J. C., Chuang, G.-Y., Diwanji, D., Georgiev, I., Do Kwon, Y., Lee, D., Louder, M. K., Moquin, S., Schmidt, S. D., Yang, Z.-Y., Bonsignori, M., Crump, J. A., Kapiga, S. H., Sam, N. E., Haynes, B. F., Burton, D. R., Koff, W. C., Walker, L. M., Phogat, S., Wyatt, R., Orwenyo, J., Wang, L.-X., Arthos, J., Bewley, C. A., Mascola, J. R., Nabel, G. J., Schief, W. R., Ward, A. B., Wilson, I. A., and Kwong, P. D. (2011). Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9. *Nature*.

Meiler, J. and Baker, D. (2003). Coupled prediction of protein secondary and tertiary structure. *PNAS*, 100(21):12105–12110.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of chemical physics*, 21:1087.

Mikell, I., Sather, D. N., Kalams, S. A., Altfeld, M., Alter, G., and Stamatatos, L. (2011). Characteristics of the earliest cross-neutralizing antibody response to HIV-1. *PLoS Pathogens*, 7(1):e1001251.

Minuchehr, Z. and Goliaei, B. (2005). Propensity of amino acids in loop regions connecting beta-strands. *Protein and peptide letters*, 12(4):379–382.

Misura, K. M. S., Chivian, D., Rohl, C. A., Kim, D. E., and Baker, D. (2006). Physically realistic homology models built with ROSETTA can be more accurate than their templates. *PNAS*, 103(14):5361–5366.

Mohan, S., Kourentzi, K., Schick, K. A., Uehara, C., Lipschultz, C. A., Acchione, M., DeSantis, M. E., Smith-Gill, S. J., and Willson, R. C. (2009). Association energetics of cross-reactive and specific antibodies. *Biochemistry*, 48(6):1390–1398.

Mohan, S., Sinha, N., and Smith-Gill, S. J. (2003). Modeling the binding sites of anti-hen egg white lysozyme antibodies HyHEL-8 and HyHEL-26: an insight into the molecular basis of antibody cross-reactivity and specificity. *Biophysical journal*, 85(5):3221–3236.

Montefiori, D. C. (2005). *Evaluating Neutralizing Antibodies Against HIV, SIV, and SHIV in Luciferase Reporter Gene Assays*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Montefiori, D. C. (2009). Measuring HIV neutralization in a luciferase reporter gene assay. *Methods in molecular biology (Clifton, NJ)*, 485:395–405.

Moore, P. L., Gray, E. S., Sheward, D., Madiga, M., Ranchobe, N., Lai, Z., Honnen, W. J., Nonyane, M., Tumba, N., Hermanus, T., Sibeko, S., Mlisana, K., Abdool Karim, S. S., Williamson, C., Pinter, A., Morris, L., and CAPRISA 002 Study (2011). Potent and broad neutralization of HIV-1 subtype C by plasma antibodies targeting a quaternary epitope including residues in the V2 loop. *Journal of Virology*, 85(7):3128–3141.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (CASP)–round x. *Proteins*, 82 Suppl 2:1–6.

Mouquet, H., Scharf, L., Euler, Z., Liu, Y., Eden, C., Scheid, J. F., Halper-Stromberg, A., Gnanapragasam, P. N. P., Spencer, D. I. R., Seaman, M. S., Schuitemaker, H., Feizi, T., Nussenzweig, M. C., and Bjorkman, P. J. (2012). Complex-type N-glycan recognition by potent broadly neutralizing HIV antibodies. *PNAS*, 109(47):E3268–77.

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

Muramatsu, 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. *The Journal of biological chemistry*, 274(26):18470–18476.

Murphy, K. M., Travers, P., and Walport, M. (2007). *Janeway's Immunobiology (Immunobiology: The Immune System (Janeway))*. Garland Science, 7 edition.

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. *Journal of Virology*, 68(6):4031–4034.

Nair, D. T., Singh, K., Siddiqui, Z., Nayak, B. P., Rao, K. V. S., and Salunke, D. M. (2002). Epitope recognition by diverse antibodies suggests conformational convergence in an antibody response. *Journal of immunology (Baltimore, Md : 1950)*, 168(5):2371–2382.

Nederbragt, L. (2012). development in NGS.

Neria, E., Fischer, S., and Karplus, M. (1996). Simulation of activation free energies in molecular systems. *The Journal of chemical physics*, 105(5):1902–1921.

North, B., Lehmann, A., and Dunbrack, Jr, R. L. (2011). A New Clustering of Antibody CDR Loop Conformations. *Journal of Molecular Biology*, 406(2):228–256.

Notkins, A. L. (2004). Polyreactivity of antibody molecules. *Trends in immunology*, 25(4):174–179.

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

Ofek, G., Guenaga, F. J., Schief, W. R., Skinner, J., Baker, D., Wyatt, R., and Kwong, P. D. (2010). Elicitation of structure-specific antibodies by epitope scaffolds. *PNAS*, 107(42):17880–17887.

Padlan, E. A. and Padlan, E. A. (1994). Anatomy of the antibody molecule. *Molecular immunology*, 31(3):169–217.

Pancera, M., McLellan, J. S., Wu, X., Zhu, J., Changela, A., Schmidt, S. D., Yang, Y., Zhou, T., Phogat, S., Mascola, J. R., and Kwong, P. D. (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. *Journal of Virology*, 84(16):8098–8110.

Pancera, M., Shahzad-Ul-Hussan, S., Doria-Rose, N. A., McLellan, J. S., Bailer, R. T., Dai, K., Loesgen, S., Louder, M. K., Staupe, R. P., Yang, Y., Zhang, B., Parks, R., Eudailey, J., Lloyd, K. E., Blinn, J., Alam, S. M., Haynes, B. F., Amin, M. N., Wang, L.-X., Burton, D. R., Koff, W. C., Nabel, G. J., Mascola, J. R., Bewley, C. A., and Kwong, P. D. (2013). Structural basis for diverse N-glycan recognition by HIV-1-neutralizing V1-V2-directed antibody PG16. *Nature structural & molecular biology*, 20(7):804–813.

Patten, P. A., Gray, N. S., Yang, P. L., Marks, C. B., Wedemayer, G. J., Boniface, J. J., Stevens, R. C., and Schultz, P. G. (1996). The immunological evolution of catalysis. *Science (New York, NY)*, 271(5252):1086–1091.

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. *PNAS*, 107(25):11483–11488.

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. *The Journal of experimental medicine*, 187(11):1745–1751.

Pilgrim, A. K., Pantaleo, G., Cohen, O. J., Fink, L. M., Zhou, J. Y., Zhou, J. T., Bolognesi, D. P., Fauci, A. S., and Montefiori, D. C. (1997). Neutralizing antibody responses to human immunodeficiency virus type 1 in primary infection and long-term-nonprogressive infection. *The Journal of infectious diseases*, 176(4):924–932.

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

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(1):135–141.

Rajewsky, K., Förster, I., and Cumano, A. (1987). Evolutionary and somatic selection of the antibody repertoire in the mouse. *Science (New York, NY)*, 238(4830):1088–1094.

Ramachandran, G., RamaKrishan, C., and Sasiekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7:95–99.

Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., Kim, D., Kellogg, E., DiMaio, F., Lange, O., Kinch, L., Sheffler, W., Kim, B.-H., Das, R., Grishin, N. V., and Baker, D. (2009). Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins*, 77(S9):89–99.

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. *The EMBO journal*, 15(12):3197–3206.

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

Renfrew, P. D., Choi, E. J., Bonneau, R., and Kuhlman, B. (2012). Incorporation of noncanonical amino acids into Rosetta and use in computational protein-peptide interface design. *PloS one*, 7(3):e32637.

Rerks-Ngarm, S., Pitisuttithum, P., Nitayaphan, S., Kaewkungwal, J., Chiu, J., Paris, R., Premsri, N., Namwat, C., de Souza, M., Adams, E., Benenson, M., Gurunathan, S., Tartaglia, J., McNeil, J. G., Francis, D. P., Stablein, D., Birx, D. L., Chunsuttiwat, S., Khamboonruang, C., Thongcharoen, P., Robb, M. L., Michael, N. L., Kunasol, P., Kim, J. H., and MOPH-TAVEG Investigators (2009). Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *The New England Journal of Medicine*, 361(23):2209–2220.

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. *PNAS*, 100(7):4144–4149.

Robertson, D. L., Anderson, J. P., Bradac, J. A., Carr, J. K., Foley, B., Funkhouser, R. K., Gao, F., Hahn, B. H., Kalish, M. L., Kuiken, C., Learn, G. H., Leitner, T., McCutchan, F., Osmanov, S., Peeters, M., Pieniazek, D., Salminen, M., Sharp, P. M., Wolinsky, S., and Korber, B. (2000). HIV-1 nomenclature proposal. *Science (New York, NY)*, 288(5463):55–56.

Rohl, C. A. (2005). Protein structure estimation from minimal restraints using Rosetta. *Methods in enzymology*, 394:244–260.

Rohl, C. A., Strauss, C. E. M., Chivian, D., and Baker, D. (2004). Modeling structurally variable regions in homologous proteins with rosetta. *Proteins*, 55(3):656–677.

Rolland, M., Edlefsen, P. T., Larsen, B. B., Tovanabutra, S., Sanders-Buell, E., Hertz, T., Decamp, A. C., Carrico, C., Menis, S., Magaret, C. A., Ahmed, H., Juraska, M., Chen, L., Konopa, P., Nariya, S., Stoddard, J. N., Wong, K., Zhao, H., Deng, W., Maust, B. S., Bose, M., Howell, S., Bates, A., Lazzaro, M., O'Sullivan, A., Lei, E., Bradfield, A., Ibitamuno, G., Assawadarachai, V., O'Connell, R. J., deSouza, M. S., Nitayaphan, S., Rerks-Ngarm, S., Robb, M. L., McLellan, J. S., Georgiev, I., Kwong, P. D., Carlson, J. M., Michael, N. L., Schief, W. R., Gilbert, P. B., Mullins, J. I., and Kim, J. H. (2012). Increased HIV-1 vaccine efficacy against viruses with genetic signatures in Env V2. *Nature*, 490(7420):417–420.

Romesberg, F. E., Spiller, B., Schultz, P. G., and Stevens, R. C. (1998). Immunological origins of binding and catalysis in a Diels-Alderase antibody. *Science (New York, NY)*, 279(5358):1929–1933.

Rong, R., Li, B., Lynch, R. M., Haaland, R. E., Murphy, M. K., Mulenga, J., Allen, S. A., Pinter, A., Shaw, G. M., Hunter, E., Robinson, J. E., Gnanakaran, S., and Derdeyn, C. A. (2009). Escape from autologous neutralizing antibodies in acute/early subtype C HIV-1 infection requires multiple pathways. *PLoS Pathogens*, 5(9):e1000594.

Roth, D. B. (2003). Restraining the V(D)J recombinase. *Nature reviews Immunology*, 3(8):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(6):983–991.

Röthlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S., and Baker, D. (2008). Kemp elimination catalysts by computational enzyme design. *Nature*, 453(7192):190–195.

Ruiz, M. and Lefranc, M.-P. (2002). IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. *Immunogenetics*, 53(10-11):857–883.

Sadofsky, M. J. (2001). The RAG proteins in V(D)J recombination: more than just a nuclease. *Nucleic acids research*, 29(7):1399–1409.

Sagar, M., Wu, X., Lee, S., and Overbaugh, J. (2006). Human immunodeficiency virus type 1 V1-V2 envelope loop sequences expand and add glycosylation sites over the course of infection, and these modifications affect antibody neutralization sensitivity. *Journal of Virology*, 80(19):9586–9598.

Sanders, R. W., Derking, R., Cupo, A., Julien, J.-P., Yasmeen, A., de Val, N., Kim, H. J., Blattner, C., de la Peña, A. T., Korzun, J., Golabek, M., de Los Reyes, K., Ketas, T. J., van Gils, M. J., King, C. R., Wilson, I. A., Ward, A. B., Klasse, P. J., and Moore, J. P. (2013). A next-generation cleaved, soluble HIV-1 Env Trimer, BG505 SOSIP.664 gp140, expresses multiple epitopes for broadly neutralizing but not non-neutralizing antibodies. *PLoS Pathogens*, 9(9):e1003618.

Sarzotti-Kelsoe, M., Bailer, R. T., Turk, E., Lin, C.-L., Bilska, M., Greene, K. M., Gao, H., Todd, C. A., Ozaki, D. A., Seaman, M. S., Mascola, J. R., and Montefiori, D. C. (2013). Optimization and validation of the TZM-bl assay for standardized assessments of neutralizing antibodies against HIV-1. *Journal of immunological methods*.

Sather, D. N., Armann, J., Ching, L. K., Mavrantoni, A., Sellhorn, G., Caldwell, Z., Yu, X., Wood, B., Self, S., Kalams, S., and Stamatatos, L. (2009). Factors associated with the development of cross-reactive neutralizing antibodies during human immunodeficiency virus type 1 infection. *Journal of Virology*, 83(2):757–769.

Schatz, D. G. and Ji, Y. (2011). Recombination centres and the orchestration of V(D)J recombination. *Nature reviews Immunology*, 11(4):251–263.

Schatz, D. G., Oettinger, M. A., and Baltimore, D. (1989). The V(D)J recombination activating gene, RAG-1. *Cell*, 59(6):1035–1048.

Scheid, J. F., Mouquet, H., Feldhahn, N., Seaman, M. S., Velinzon, K., Pietzsch, J., Ott, R. G., Anthony, R. M., Zebroski, H., Hurley, A., Phogat, A., Chakrabarti, B., Li, Y., Connors, M., Pereyra, F., Walker, B. D., Wardemann, H., Ho, D., Wyatt, R. T., Mascola, J. R., Ravetch, J. V., and Nussenzweig, M. C. (2009). Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals. *Nature*, 458(7238):636–640.

Scheid, J. F., Mouquet, H., Ueberheide, B., Diskin, R., Klein, F., Oliveira, T. Y. K., Pietzsch, J., Fenyo, D., Abadir, A., Velinzon, K., Hurley, A., Myung, S., Boulad, F., Poignard, P., Burton, D. R., Pereyra, F., Ho, D. D., Walker, B. D., Seaman, M. S., Bjorkman, P. J., Chait, B. T., and Nussenzweig, M. C. (2011). Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science (New York, NY)*, 333(6049):1633–1637.

Schief, W. R., Ban, Y.-E. A., and Stamatatos, L. (2009). Challenges for structure-based HIV vaccine design. *Current opinion in HIV and AIDS*, 4(5):431–440.

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 & Development*, 7(12B):2520–2532.

Schmidt, A. G., Xu, H., Khan, A. R., O'Donnell, T., Khurana, S., King, L. R., Manischewitz, J., Golding, H., Suphaphiphat, P., Carfi, A., Settembre, E. C., Dormitzer, P. R., Kepler, T. B., Zhang, R., Moody, M. A., Haynes, B. F., Liao, H.-X., Shaw, D. E., and Harrison, S. C. (2013). Preconfiguration of the antigen-binding site during affinity maturation of a broadly neutralizing influenza virus antibody. *PNAS*, 110(1):264–269.

Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, 18(20):6097–6100.

Schultz, P. G., Yin, J., and Lerner, R. A. (2002). The chemistry of the antibody molecule. *Angewandte Chemie (International ed. in English)*, 41(23):4427–4437.

Sethi, D. K., Agarwal, A., Manivel, V., Rao, K. V. S., and Salunke, D. M. (2006). Differential epitope positioning within the germline antibody paratope enhances promiscuity in the primary immune response. *Immunity*, 24(4):429–438.

Shockett, P. E. and Schatz, D. G. (1999). DNA hairpin opening mediated by the RAG1 and RAG2 proteins. *Molecular and cellular biology*, 19(6):4159–4166.

Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., St Clair, J. L., Gallaher, J. L., Hilvert, D., Gelb, M. H., Stoddard, B. L., Houk, K. N., Michael, F. E., and Baker, D. (2010). Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science (New York, NY)*, 329(5989):309–313.

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., Birungi, J., Pozniak, A., McPhee, D. A., Manigart, O., Karita, E., Inwoley, A., Jaoko, W., Dehovitz, J., Bekker, L.-G., Pitisuttithum, P., Paris, R., Walker, L. M., Poignard, P., Wrin, T., Fast, P. E., Burton, D. R., and Koff, W. C. (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. *Journal of Virology*, 83(14):7337–7348.

Simonelli, L., Pedotti, M., Beltramello, M., Livoti, E., Calzolai, L., Sallusto, F., Lanzavecchia, A., and Varani, L. (2013). Rational engineering of a human anti-dengue antibody through experimentally validated computational docking. *PloS one*, 8(2):e55561.

Simons, K. T., Bonneau, R., Ruczinski, I., and Baker, D. (1999a). Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins*, Suppl 3:171–176.

Simons, K. T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of Molecular Biology*, 268(1):209–225.

Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C., and Baker, D. (1999b). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*, 34(1):82–95.

Spurrier, B., Sampson, J. M., Totrov, M., Li, H., O'Neal, T., Williams, C., Robinson, J., 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 (London, England : 1993)*, 19(5):691–699.

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(5):637–648.

Steen, S. B., Gomelsky, L., and Roth, D. B. (1996). The 12/23 rule is enforced at the cleavage step of V(D)J recombination in vivo. *Genes to cells : devoted to molecular & cellular mechanisms*, 1(6):543–553.

Stranges, P. B. and Kuhlman, B. (2013). A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. *Protein science : a publication of the Protein Society*, 22(1):74–82.

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., Cinatl, J., ter Meulen, J., Lasters, I., Carsetti, R., Peiris, M., de Kruif, J., and Goudsmit, J. (2008). Heterosubtypic neutralizing monoclonal antibodies cross-protective against H5N1 and H1N1 recovered from human IgM+ memory B cells. *PloS one*, 3(12):e3942.

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. *Journal of immunology (Baltimore, Md : 1950)*, 180(5):3279–3288.

Tomaras, G. D., Binley, J. M., Gray, E. S., Crooks, E. T., Osawa, K., Moore, P. L., Tumba, N., Tong, T., Shen, X., Yates, N. L., Decker, J., Wibmer, C. K., Gao, F., Alam, S. M., Easterbrook, P., Abdool Karim, S., Kamanga, G., Crump, J. A., Cohen, M., Shaw, G. M., Mascola, J. R., Haynes, B. F., Montefiori, D. C., and Morris, L. (2011). Polyclonal B cell responses to conserved neutralization epitopes in a subset of HIV-1-infected individuals. *Journal of Virology*, 85(21):11502–11519.

Tonegawa, S. (1983). Somatic generation of antibody diversity. *Nature*, 302(5909):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. *Journal of Virology*, 70(2):1100–1108.

UNAIDS (2013). UNAIDS report on the global AIDS epidemic 2013. Technical Report JC2502, UNAIDS.

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(1):107–113.

van Gils, M. J., Euler, Z., Schweighardt, B., Wrin, T., and Schuitemaker, H. (2009). Prevalence of cross-reactive HIV-1-neutralizing activity in HIV-1-infected patients with rapid or slow disease progression. *AIDS (London, England)*, 23(18):2405–2414.

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(6847):607–614.

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., Mitcham, J. L., Hammond, P. W., Olsen, O. A., Phung, P., Fling, S., Wong, C.-H., Phogat, S., Wrin, T., Simek, M. D., Principal Investigators, P. G., Koff, W. C., Wilson, I. A., Burton, D. R., and Poignard, P. (2011). Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature*.

Walker, L. M., Phogat, S. K., Chan-Hui, P.-Y., Wagner, D., Phung, P., Goss, J. L., Wrin, T., Simek, M. D., Fling, S., Mitcham, J. L., Lehrman, J. K., Priddy, F. H., Olsen, O. A., Frey, S. M., Hammond, P. W., Investigators, P. G. P., Kaminsky, S., Zamb, T., Moyle, M., Koff, W. C., Poignard, P., and Burton, D. R. (2009). Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. *Science (New York, NY)*, 326(5950):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 Pathogens*, 6(8):e1001028.

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 (New York, NY)*, 301(5638):1374–1377.

Wedemayer, G. J., Patten, P. A., Wang, L. H., Schultz, P. G., and Stevens, R. C. (1997). Structural insights into the evolution of an antibody combining site. *Science (New York, NY)*, 276(5319):1665–1669.

Wedemeyer, W. J. and Baker, D. (2003). Efficient minimization of angle-dependent potentials for polypeptides in internal coordinates. *Proteins*, 53(2):262–272.

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

Weitkamp, J. H. and Crowe, J. E. (2001). Blood donor leukocyte reduction filters as a source of human B lymphocytes. *BioTechniques*, 31(3):464–466.

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. *The Journal of experimental medicine*, 191(3):579–584.

Willis, J. R., Briney, B. S., Deluca, S. L., Crowe, J. E., and Meiler, J. (2013). Human germline antibody gene segments encode polyspecific antibodies. *PLoS computational biology*, 9(4):e1003045.

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

Wilson, P. C., de Bouteiller, O., Liu, Y. J., Potter, K., Banchereau, J., Capra, J. D., and Pascual, V. (1998b). Somatic hypermutation introduces insertions and deletions into immunoglobulin V genes. *The Journal of experimental medicine*, 187(1):59–70.

Wong, S. E., Sellers, B. D., and Jacobson, M. P. (2011). Effects of somatic mutations on CDR loop flexibility during affinity maturation. *Proteins*, 79(3):821–829.

Wu, X., Parast, A. B., Richardson, B. A., Nduati, R., John-Stewart, G., Mbori-Ngacha, D., Rainwater, S. M. J., and Overbaugh, J. (2006). Neutralization escape variants of human immunodeficiency virus type 1 are transmitted from mother to infant. *Journal of Virology*, 80(2):835–844.

Wu, X., Yang, Z.-Y., Li, Y., Hogerkorp, C.-M., Schief, W. R., Seaman, M. S., Zhou, T., Schmidt, S. D., Wu, L., Xu, L., Longo, N. S., McKee, K., O'dell, S., Louder, M. K., Wycuff, D. L., Feng, Y., Nason, M., Doria-Rose, N., Connors, M., Kwong, P. D., Roederer, M., Wyatt, R. T., Nabel, G. J., and Mascola, J. R. (2010a). Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science (New York, NY)*, 329(5993):856–861.

Wu, X., Zhou, T., O'Dell, S., Wyatt, R., Kwong, P. D., and Mascola, J. (2009). Mechanism of HIV-1 Resistance to Monoclonal Antibody b12 that Effectively Targets the Site of CD4 Attachment. *Journal of Virology*.

Wu, X., Zhou, T., Zhu, J., Zhang, B., Georgiev, I., Wang, C., Chen, X., Longo, N. S., Louder, M., McKee, K., O'dell, S., Perfetto, S., Schmidt, S. D., Shi, W., Wu, L., Yang, Y., Yang, Z.-Y., Yang, Z., Zhang, Z., Bonsignori, M., Crump, J. A., Kapiga, S. H., Sam, N. E., Haynes, B. F., Simek, M., Burton, D. R., Koff, W. C., Doria-Rose, N., Connors, M., NISC Comparative Sequencing Program, Mullikin, J. C., Nabel, G. J., Roederer, M., Shapiro, L., Kwong, P. D., and Mascola, J. R. (2011). Focused Evolution of HIV-1 Neutralizing Antibodies Revealed by Structures and Deep Sequencing. *Science (New York, NY)*.

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(7):1070–1078.

Wyatt, R. and Sodroski, J. (1998). The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens. *Science (New York, NY)*, 280(5371):1884–1888.

Xu, R., Ekiert, D. C., Krause, J. C., Hai, R., Crowe, J. E., and Wilson, I. A. (2010). Structural basis of preexisting immunity to the 2009 H1N1 pandemic influenza virus. *Science (New York, NY)*, 328(5976):357–360.

Yarov-Yarovoy, V., Schonbrun, J., and Baker, D. (2006). Multipass membrane protein structure prediction using Rosetta. *Proteins*, 62(4):1010–1025.

Ye, J., Ma, N., Madden, T. L., and Ostell, J. M. (2013). IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic acids research*, 41(Web Server issue):W34–40.

Yin, J., Beuscher, A. E., Andryski, S. E., Stevens, R. C., and Schultz, P. G. (2003). Structural plasticity and the evolution of antibody affinity and specificity. *Journal of Molecular Biology*, 330(4):651–656.

Yin, J., Mundorff, E. C., Yang, P. L., Wendt, K. U., Hanway, D., Stevens, R. C., and Schultz, P. G. (2001). A comparative analysis of the immunological evolution of antibody 28B4. *Biochemistry*, 40(36):10764–10773.

Zanghellini, A., Jiang, L., Wollacott, A. M., Cheng, G., Meiler, J., Althoff, E. A., Röthlisberger, D., and Baker, D. (2006). New algorithms and an in silico benchmark for computational enzyme design. *Protein science : A publication of the Protein Society*, 15(12):2785–2794.

Zemlin, M., Klinger, M., Link, J., Zemlin, C., Bauer, K., Engler, J. A., Schroeder, Jr, H. W., and Kirkham, P. M. (2003). Expressed Murine and Human CDR-H3 Intervals of Equal Length Exhibit Distinct Repertoires that Differ in their Amino Acid Composition and Predicted Range of Structures. *Journal of Molecular Biology*, 334(4):733–749.

Zheng, N.-Y., Wilson, K., Jared, M., and Wilson, P. C. (2005). Intricate targeting of immunoglobulin somatic hypermutation maximizes the efficiency of affinity maturation. *The Journal of experimental medicine*, 201(9):1467–1478.

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., Burton, D. R., Dimitrov, D. S., Sodroski, J., Wyatt, R., Nabel, G. J., and Kwong, P. D. (2007). Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Nature*, 445(7129):732–737.

Zimmermann, J., Romesberg, F. E., Brooks, C. L. I., and Thorpe, I. F. (2010). Molecular Description of Flexibility in an Antibody Combining Site. *Journal of Physical Chemistry*, 114(21):7359–7370.

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., and Parren, P. W. (2001). Broadly neutralizing antibodies targeted to the membrane-proximal external region of human immunodeficiency virus type 1 glycoprotein gp41. *Journal of Virology*, 75(22):10892–10905.