SEQUENCE, STRUCTURE, AND FUNCTION RELATIONSHIPS OF HUMAN ANTIBODIES

by

Jessica Ann Strnad

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Microbiology and Immunology

December 15, 2018

Nashville, Tennessee

Approved:

Mark R. Denison, M.D.

James E. Crowe, Jr., M.D.

Jens Meiler, Ph.D.

John Anthony Capra, Ph.D.

Jonathan M. Irish, Ph.D

Andrew J. Link, Ph.D.

James W. Thomas, II, M.D.

This work is dedicated to my parents,
whose unwavering support made this possible.

ACKNOWLEDGEMENTS

Foremost, I would like to acknowledge the guidance of my mentors, James E. Crowe, Jr. and Jens Meiler. Without your willingness to collaboratively train interdisciplinary students and your vast combined experience, I would not have had the opportunity to execute these studies. Thank you for taking a chance on me and for your continued support throughout the years; I know it took longer than any of us expected.

I would like to acknowledge the support and mentorship of my committee members: Mark Denison, Tom Thomas, Jonathan Irish, Andy Link and Tony Capra. It has been a privilege working with each of you, and your insightful discussion during meetings drove this work to greater success. Your encouragement to focus on my future and advice in career decisions has been invaluable in ushering me on to the next step.

This dissertation would not have been possible without support from the following funding sources: the Virology Training Program (T32 AI 89554, NIAID/NIH); the Vanderbilt Trans-Institutional Program (TIP) "Integrating Structural Biology with Big Data for Next Generation Vaccines"; grants from the National Institute of Allergy and Infectious Disease (National Institute of Health) titled "Molecular Determinants of Cross-Reactive Antibody Response to Influenza in Humans" (R01 AI 106002) and "Structure Based Design of Antibodies and Vaccines" (U19 AI 117905); and the Defense Threat Reduction Agency (Department of Defense) grant "Molecular and Structural Basis of Fine Specificity of Antiviral Antibodies" (HDTRA1-10-1-0067).

Work of this magnitude is not accomplished in isolation, and I would be remiss if I did not acknowledge the many scientists that assisted me throughout the years. In no particular order, I would like to thank: Jordan Willis, for inspiring this co-mentored training program, for

challenging me, and for teaching me everything I know; Bryan Briney, for introducing me to next generation sequencing; Natalie Thornburg, whose mentorship as an excellent scientist and woman in STEM I greatly treasure; Gopal Sapparapu, for sage advice, and for creating an invaluable core resource for recombinant protein expression within the Crowe lab; Alex Sevy, for critical insights without which much of this work would not have succeeded; Nina Bozhanova, Amandeep Sangha, and David Nannemann, for insightful discussions as fellow Antibody Interface members, and for guidance as postdoctoral researchers; Rocco Moretti. Sam and Stephanie DeLuca, for always being willing to answer Rosetta questions thoroughly; Jinhui Dong, who taught me crystallography and, in the process, taught me how to train scientists; Andre Branchizio, Ross Troseth and Taylor Jones, for graciously listening to my stubborn opinions about data management, and for supporting our databases and software pipelines; Sandhya Bangaru and Iuliia Gilchuk, for collaborating with me and for educating me as experienced influenza scientists; Erica Parrish and Rachel Nargi, who trained me in new experimental techniques and who directly produced materials used in this work; Mahsa Majedi, Morgan Scarlett-Jones and Walter Reichard, for supporting the molecular biology underpinning this work; Robin Bombardi, for countless discussions about next generation sequencing, for steady guidance, and for constantly challenging me to find ways to do the impossible; and to everyone in the Vanderbilt Vaccine Center and the Meiler Lab, for sharing your laboratory, equipment and reagents with me for the last many years.

Graduate school is a gauntlet, and supporting a scientist through this time takes many forms. I would like to thank: Meg Fox for her enduring friendship and for modeling academic success, as her example encourages me to be a better scientist; Monique Bennett, Gabriela Alvarado, Brian Bender, Hannah King, Vidisha Singh, and Jennifer Pickens, who have given me their time,

attention, and listening ear; Lauren Williamson, who reminded me not to take myself too seriously; and Rob Carnahan, Pavlo Gilchuk, Scott Smith, and Matt Vogt, who have exemplified leadership and who have provided additional mentoring.

I must thank my family: my parents especially, who have supported this wild dream without hesitation for many years; my sister Amanda and her family, who cheer for my successes and support me wholeheartedly; Shane Finn, who has shared the greatest gift with me, and who partners with me in the great challenge of parenting; and my son, Elijah, who reminds me there is more to life than work, who has allowed me to discover the world anew through his eyes, and who brings me unspeakable joy. My very special gratitude goes out to the nurses and doctors of Vanderbilt's NICU, who worked miracles and cared for my family for 140 days. I would specifically like to thank Kristina and Brooke, our primary nurses, who chose to dedicate their incredible nursing skills to the care of my son. Their presence by his side allowed me to focus on returning to work at a time I thought it would be impossible.

Loving gratitude goes to my many friends who have fed me, consoled me, and celebrated with me. You have been my Nashville family, and you made this city feel like home.

We did it! Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

Introduction

I.1 Introduction

The human adaptive immune system is mediated in part by B cells, which produce antibodies to protect the body from infection. Antibodies are protein molecules responsible for recognizing and binding pathogenic targets (*i.e.*, antigens) to mediate effective neutralization of the microorganism. In this thesis, I will show how an understanding of antibody sequence and structure can be leveraged to predict the function of an antibody, particularly in the context of the human antibody response to influenza virus.

This first chapter introduces the fields of B cell immunology, particularly focusing on how B cells develop and the mechanisms that produce antibodies. I will also introduce influenza A virus as a target for antibody responses. Finally, I will discuss the emerging technologies that I used to study the human antibody repertoire, specifically next generation sequencing and computational structural modeling using Rosetta.

Chapter II of this thesis focuses on the impact that these new sequencing technologies had on studies of the human antibody repertoire. Our recent ability to analyze large populations of antibody sequences has begun to improve our understanding of the mechanisms of diversification, the size of the repertoire, and methods of repertoire regulation conserved between individuals. In the second chapter I will also discuss populations of human antibodies produced in response to influenza vaccine that I identified from repertoires by sequence, which are predicted to be clonally related to known anti-influenza antibodies.

The third chapter discusses techniques for computational structural modeling of antibodies.

Prior to this work, protein loop modeling techniques were limited in the length of loop that could be accurately predicted. The structures of protein loops on human antibodies are critically important for antibody function, and many of these loops are longer than existing technologies could model at the start of this work. The third chapter discusses how knowledge-based restraints that I calculated from analysis of a conserved structural motif in the most critical of the antibody protein loops improved the accuracy of antibody modeling in Rosetta.

Chapter IV describes the development of a novel method for structure-based discovery of antibodies from next generation sequencing data. It was known at the beginning of this work that antibodies that share structural features have similar function, however no techniques existed to leverage this understanding for the discovery of novel antibodies. I developed a method that paired next generation sequencing and structural similarity predictions in Rosetta to discover novel anti-influenza antibodies from human donors.

In the final chapter, I summarize these studies and discuss the sequence-structure-function relationships of antibodies as a cohesive concept. An understanding of these relationships may be applied to future work in this field, and I propose additional experiments and applications of these technologies that will benefit therapeutic discovery and vaccination efforts to come.

I.2 B cell development

B cells arise from hematopoietic stem cells in the bone marrow. During the course of their development each B cell produces a unique immunoglobulin (Ig) molecule made up of heavy and light chain (HC, LC) proteins. Immunoglobulin can be either retained on the surface of the B cell as part of the B cell receptor (BCR), or be secreted as soluble antibodies (Abs). Successful surface expression of immunoglobulin is necessary for development of the B cell, and expression

of the immunoglobulin heavy and light chains mark the distinct stages of B cell development.

Pluripotent hematopoietic stem cells undergo stages of differentiation to commit to the B cell lineage; first to early lymphocyte progenitor cells, which may further differentiate to either T- or B-lymphocytes, and later to pro-B cells, the earliest defined B-lineage cells. Over the course of development from pro-B cells to pre-B cells, and later to immature B cells, human differentiating B cells undergo stages of genetic rearrangement within their immunoglobulin loci. These gene rearrangements are referred to as V(D)J recombination, and result in the creation of functional immunoglobulin heavy and light chain genes.

The organization of the immunoglobulin loci allows for a vast diversity of antibody proteins. An understanding of the loci provides important insight into the further study of antibody structure and function. Immunoglobulin molecules are made up of two different protein subunits, referred to as the heavy and light chains. In humans, light chains may be produced from either the kappa or lambda loci located on chromosomes 2 or 22, respectively. The heavy chain is produced from a third loci located on chromosome 14. Each chain can be subdivided into two regions referred to as the variable region (V region, also referred to in the literature as the V domains) and the constant region (C region or C domains). The V region is so named due to its inherent variability in the final immunoglobulin molecule; within each loci this region is made up of multiple gene segments referred to as the variable (V), diversity (D; only present in the heavy chain), and joining (J) genes, of which one of each recombines to create the final exon encoding the V domain. Further consequences of V(D)J recombination on the diversification and function of the antibody repertoire are discussed in the next section.

Over the course of B cell development, these immunoglobulin loci are rearranged following a prescribed order determined by expression of enzyme complexes and by unique features of the

immunoglobulin gene sequences. Expression of the V(D)J recombinase complex triggers immunoglobulin loci rearrangement via recognition of recombination signal sequences (RSSs) adjacent to V(D)J gene segments, cleavage of intervening genetic material between two RSSs resulting in hair-pinning of the DNA ends by the recombination-activating proteins (RAG), processing of the cleaved ends by terminal deoxynucleotidyl transferase (TdT), and finally joining of the processed ends by DNA repair enzymes. These mechanisms occur repeatedly, beginning with the heavy chain locus D-J joining event during the early pro-B cell phase, followed by V-DJ rearrangement in the late pro-B cell phase. The fully recombined immunoglobulin heavy chain is then expressed into a protein signaling complex on the surface of the large pre-B cell, creating a checkpoint that ensures successful gene recombination, protein expression and protein folding of the heavy chain. The process continues similarly for the light chain, which undergoes only V-J rearrangement. Expression of a complete immunoglobulin molecule with both rearranged heavy and light chains signifies the development of the immature B cell. In the final phases of B cell development within the bone marrow, immature B cells expressing immunoglobulin as part of a functional BCR are tested for tolerance to self-antigens, referred to as central tolerance. Immature B cells that lack self-antigen recognition mature, escaping the bone marrow to circulate and function in the periphery.

I.3 Diversification of the antibody repertoire

The development of a population of B cells each encoding a novel recombined immunoglobulin molecule, referred to henceforth as the antibody repertoire, is a key element of acquired immunity. The central tolerance checkpoint at the end of B cell development results in removal of antibodies from the peripheral repertoire that are self- or otherwise non-specific (*e.g.*,

'sticky' antibodies able to broadly bind to other molecules). It is generally understood that the resulting antibodies in the peripheral repertoire are able to bind only one or a very small number of molecules. Therefore a diverse antibody repertoire, each member specific to a particular pathogenic protein (*i.e.*, antigen), is a necessary component for the prevention or resolution of disease caused by most viruses (Crotty & Ahmed, 2004).

Diversity in the naïve antibody repertoire is mediated by three principal mechanisms that are illustrated in Figure 1: (1) random pairing of heavy and light chains to form the antigen-binding site in the immunoglobulin molecule; (2) combinatorial diversity generated by V(D)J recombination, which together with heavy and light chain pairing results in approximately 2.3 x 10^6 possible combinations; and (3) junctional diversity generated by P- and N-nucleotide addition or deletion at recombination sites during V(D)J processing by TdT, which theoretically results in 10^{11} different antibody specificities.

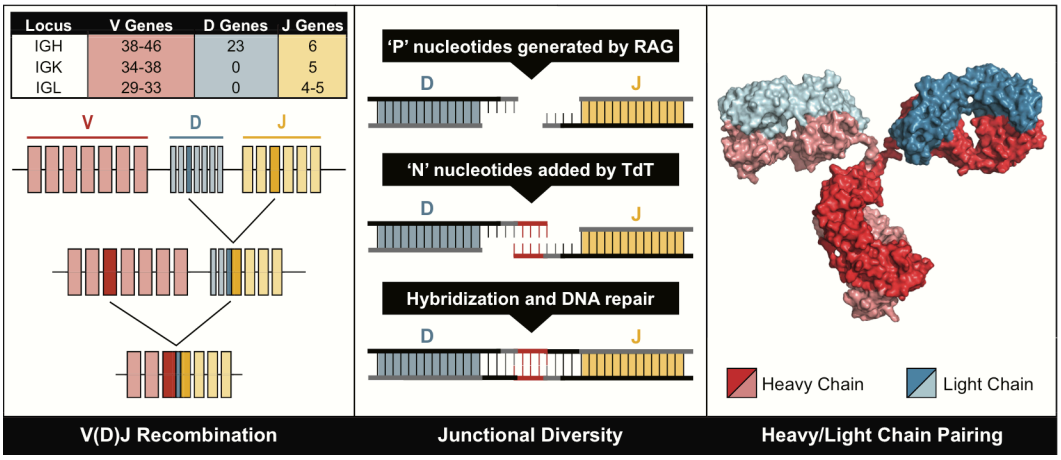


Figure 1. Diversity in the antigen-combining site of the B cell receptor repertoire (and thus also in the corresponding secreted antibody repertoire) is mediated by three principal molecular mechanisms, illustrated in the three panels, V(D)J Recombination, Junctional Diversity, and Heavy/Light Chain Pairing.

Once developed in the bone marrow, the antibody repertoire patrols peripheral tissues, further developing and differentiating in response to antigen exposure. Somatic hypermutation, a fourth mechanism of diversification, introduces point mutations into the rearranged

immunoglobulin molecule after B cell activation by antigen. Additional functional diversity in secreted antibodies is conferred by differences between C region isotypes determined by a process called class switching. The C region of an immunoglobulin determines the valency of the antibody, and enables other functions such as complement fixation or interaction with various C region receptors, which play important roles in pathogen neutralization.

It is critical to acknowledge that the antibody repertoire is a constantly changing population shaped by expansion and contraction events in response to infection. Upon B cell recognition of antigen, activation and stimulation by T helper cells in the germinal center reaction encourages differentiation into one or more specialized B cell types. The first of these are plasmablasts, which expand prolifically and produce antibody for short periods of time before dying or developing further. Surviving germinal center B cells develop into smaller populations of long-lived memory B cells (which persist in the repertoire for perhaps as long as a lifetime), or terminally differentiated plasma B cells (which cease proliferating and instead expend energy producing vast amounts of soluble antibody). Each of these B cell types, and their associated diverse antibody population, play a key role in the immune response to infection.

I.4 Influenza virus

In opposition to the adaptive immune system, viruses have developed mechanisms of diversification that allow escape from immune recognition. A prime example is influenza, an infectious disease caused by zoonotic RNA viruses. Despite continued vaccination efforts, influenza A virus (IAV) continues to cause high rates of annual disease as it escapes immunological resistance through subtype changes (*i.e.,* antigenic shift) and point mutation (*i.e.,* antigenic drift) mediated primarily by changes to its envelope proteins. Furthermore, studies

published at the time this work began showed that experimentally generated mutations in IAV envelope proteins confer respiratory transmission between humans, revealing the potential for development of highly pathogenic human IAVs with pandemic possibility. Improving our understanding of the human immune response to IAVs and identifying novel antibodies specific for IAV became the two key goals for this work.

Influenza A virus is a member of *Orthomyxoviridae,* a family of viruses characterized by having an envelope derived from the host cell membrane that incorporates viral glycoproteins and non-glycosylated proteins, encapsulating a genome consisting of numerous linear negative-sense single stranded RNA segments. In IAV, the major envelope glycoproteins are hemagglutinin (HA), which is responsible for host cell binding and entry into cells, and neuraminidase (NA), involved in viral egress from infected cells via enzymatic cleavage of sialic acid. These two proteins are the means by which IAVs are classified into subtypes; the 18 known HAs and 11 known NAs have many combinations, and result in the naming of IAV subtypes such as H1N1 and H3N2, which are the two IAV subtypes currently circulating in humans.

While antibodies specific to both HA and NA have been identified following human infection with IAV, the work presented in the following chapters focuses on the human antibody response to HA. HA is a homotrimeric glycoprotein with a molecular mass greater than 180 kDa, dependent upon the number and complexities of each N-linked glycosylation added to the protein surface (Sriwilaijaroen & Suzuki, 2012). The HA trimers project off the surface of the IAV virion as ‘spikes’, held in place by type I transmembrane domains. Each HA monomer is expressed as a precursor protein referred to as HA0. During viral maturation, each HA0 is cleaved by proteases to form the fusion-capable HA1 and HA2 molecules, which remain linked by a disulfide bond and continue to form a stable homotrimer at physiological conditions. The

HA molecule is subdivided into two domains related to the cleavage of HA1 and HA2. The head domain is formed by HA1, is distal to the virion surface, and is responsible for binding to the host cell via the receptor binding site (RBS). The stem domain is primarily made up of HA2, although contains a small portion of HA1, forms a cylindrical stalk that connects the head domain to the transmembrane domain, and undergoes considerable structural rearrangement during fusion.

Infection with IAV is mediated first by HA binding to specific terminal sialic acids found on host cell receptors. While sialic acids are commonly found on animal tissues, the distribution of particular sialic acid forms is specific to both animal species and tissues within those species. Human IAVs preferentially bind to sialic acids that attach to galactose via α2-6 linkages, which are commonly found on tissues in the human upper respiratory tract. Upon binding of sialic acid, IAV virions are taken up into cells by receptor-mediated endocytosis. As the endosome acidifies, cleaved HA undergoes a profound pH-dependent structural rearrangement to reveal the fusion domain. While the stages of this structural rearrangement are unclear, it is known that it results in host endosomal membrane fusion with the viral membrane, releasing the genetic components of the virus into the host cell, completing host cell infection.

A number of antibody-dependent, HA-mediated IAV neutralization mechanisms were known at the beginning of this work. Broadly, these mechanisms can be divided into two categories; head-binding antibodies, which were the larger known population of human anti-HA antibodies at the start of this work and which primarily function to block HA binding to host sialic acid receptors, and stem-binding antibodies, which are thought to interrupt the structural changes necessary for viral membrane fusion. Evidence suggests that stem-binding antibodies are often broadly specific to HA, in that one monoclonal antibody (mAb) specific to the HA stem is

capable of binding to many IAV strains within a subtype, and even to HA stems across subtypes (Avnir et al., 2014; Ekiert et al., 2009). This is due to the higher amount of sequence and structural conservation in the stem, caused by the necessity of functional conservation of this domain. A computational structural modeling experiment published by Sarel Fleishman in the Baker group at the beginning of this thesis work showed that small proteins could be engineered to mimic known antibody-HA stem interactions, and that these engineered proteins bind HA with low-nanomolar affinity (Fleishman et al., 2011). Therefore, I have chosen instead to focus these studies on antibodies against the head domain of IAV HA.

The head domain of IAV HA is far less conserved than the stem domain, even within a given HA subtype. This membrane-distal portion of the HA molecule is less involved in gross structural rearrangements during fusion, therefore only the receptor binding site (RBS) responsible for binding to host sialic acid receptors must be conserved for viral function. Sialic acid is a small molecule, and the functional footprint of this RBS pocket on the HA head domain is only 175 $Å^2$. IAV-neutralizing human antibodies are often specific for the RBS and function by physically blocking access to the host receptor (Kadam & Wilson, 2018). However, the average footprint of an antibody on the surface of an antigen is 1103 $Å^2$, much larger than the sialic acid binding site itself, and viral mutations that occur around the edge of the RBS pocket effectively escape antibody recognition while conserving sialic acid receptor binding (Ramaraj, Angel, Dratz, Jesaitis, & Mumey, 2012). Additional antigenic sites have been mapped on the HA head domain using mouse antibodies, and it is known that these sites also undergo considerable selective pressure to evade the antibody repertoire (Caton, Brownlee, Yewdell, & Gerhard, 1982; Wiley, Wilson, & Skehel, 1981). A few broadly-neutralizing human antibodies targeting the HA head domain were known at the beginning of this thesis, providing a starting point for the

development of methods to discover or engineer similar monoclonal antibodies (Krause, Tsibane, Tumpey, Huffman, Basler, et al., 2011a; Whittle, Zhang, Khurana, King, Manischewitz, Golding, Dormitzer, Haynes, Walter, Moody, Kepler, Liao, & Harrison, 2011a; R. Xu et al., 2013). One such antibody, CH65, is discussed in further detail in Chapter IV.

I.5 Antibody repertoire sequencing

Prior to the time of this work, immunologists understood diversification of B cell populations specific to particular foreign antigens involved a burst of diversification within a clone of B cells in the activated germinal center, followed by a selection for survival of the highest affinity clone and drastic loss of related somatic variants with lower affinity. Although this 'single winner' model correctly describes the typical panel of B cell clones isolated from experimental studies using isolation of hybridomas and mAbs, the technical approach to isolation of mAbs likely biases such studies toward the isolation of only the most avidly binding antibodies. Emerging techniques using high-throughput DNA and RNA sequence analysis are increasingly revealing that this paradigm is not correct. Instead human B cell repertoires maintain very large populations of somatic variants within clones (Krause, Tsibane, Tumpey, Huffman, Briney, et al., 2011b); see Figure 2. It may seem metabolically wasteful and counter-intuitive that the immune system would allow hundreds or thousands of related clones to persist in circulation when many of those variants possess many fewer somatic mutations than the most mature clones, and thus by inference are likely to have lower affinity of binding for the inciting epitope. This may be rationalized, however, if persisting diversity in the B cell repertoire allows the subject to respond to antigenic variation in the target, such as antigenic drift in IAV. Dealing with the enormous sequence and structural plasticity of IAV HA likely requires an equivalent

breadth of diversity in the responding antibody repertoire. Therefore, recent observations that human antibody repertoires engage pathogens with large clonal families of highly related antibodies make sense from a strategic standpoint for the immune system. Studying the diverse antibody response to antigen as a swarming population instead of as a one-to-one, specific interaction informs our understanding of disease and immunity in a new way.

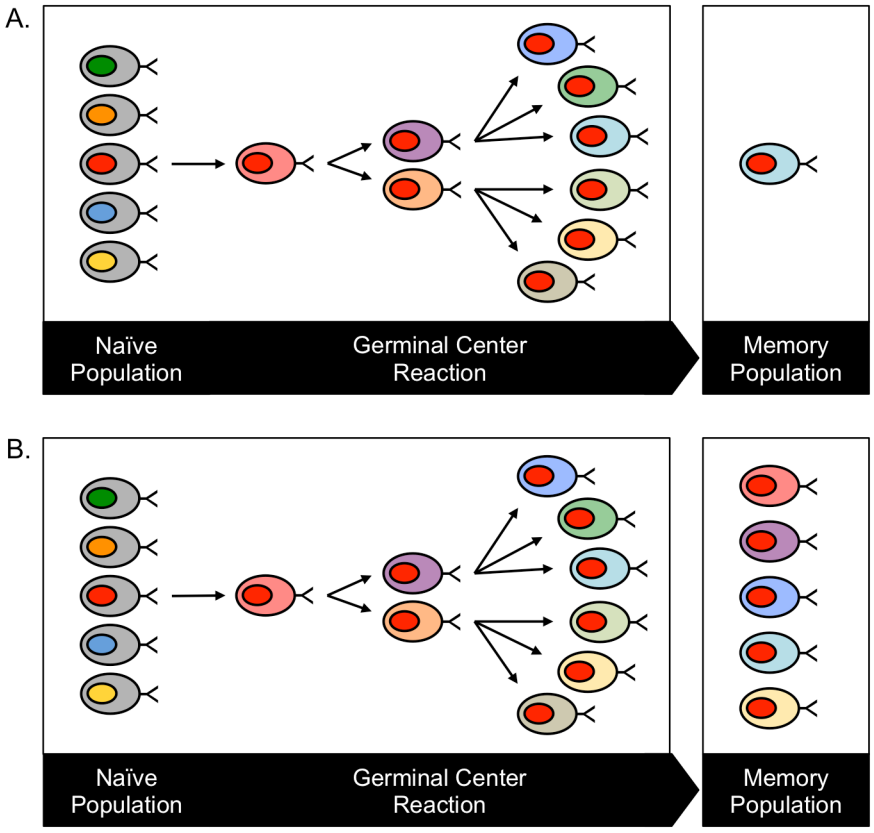


Figure 2. [A] Classical models of somatic hypermutation conceive of rapid generation of variants in the activated germinal center followed by a severe down-selection in number of variants, resulting in selection of only the clones with the most avidly binding B cell receptors for survival. [B] Newer repertoire studies using large-scale sequence analysis reveal that human B cell repertoires retain a large number of variants with diverse numbers of point mutations within clones, even in the peripheral blood.

Next generation sequencing (NGS) technologies developed in tandem with the work presented herein. These new technological advances in gene sequencing and microfluidics to provided evidence regarding the mechanisms of repertoire diversification, the size of the

antibody repertoire and methods of repertoire regulation shared by different individuals, and were the foundation upon which the applications within this thesis were developed. At the beginning of this project, the Crowe laboratory had developed an antibody sequencing method for the Roche 454 pyrosequencing platform. Using this platform, the Crowe laboratory previously studied antibody repertoire sequences from a pool of human donors to discover highly related families of antibodies responding to viral infection, novel mechanisms of antibody recombination, and very long antigen-binding antibody loops (Briney, Willis, & Crowe, 2012a; Briney, Willis, Hicar, Thomas, & Crowe, 2012c; Krause, Tsibane, Tumpey, Huffman, Briney, et al., 2011b).

Although Roche 454 pyrosequencing was capable of generating large data sets, the previous studies of antibody repertoires were still plagued with uncertainties due to under-sampling. In this body of work, I have translated our Roche 454 methods to the Illumina MiSeq and HiSeq platforms to perform even higher-throughput NGS of antibody repertoires. These Illumina platforms have become the mainstream methods for sequencing large amplicon libraries, and most of the projects presented here make use of data collected with those techniques. The NGS techniques that I developed over the course of this thesis and applications of those techniques are discussed further in Chapter II.

I.6 Modeling antibody structure with Rosetta

De novo protein structure prediction is one of the greatest challenges remaining in computational structural biology. This process models the tertiary structure of a protein from its primary amino acid sequence. Importantly, *de novo* modeling differs from template-based or homology protein modeling in that structural predictions are not based upon a previous

homologous structure. To address the challenge of predicting a protein's structure *de novo*, Rosetta uses short peptide "fragments" to piecewise assemble a complete protein structure.

Briefly, the Rosetta *de novo* protein-folding algorithm uses short peptide fragments of known proteins obtained from structures deposited in the Protein Data Bank (PDB), and inserts them into an extended-chain protein following a Monte Carlo strategy (Rohl, Strauss, Misura, & Baker, 2004b). These fragments alter the backbone conformation of the extended-chain protein, folding it toward the native tertiary structure. Finally, these low-resolution models can be filtered based on user-defined pass/fail criteria, and an energy minimization step can be applied to refine and idealize the model.

De novo protein folding relies on the assembly of short peptide fragments, and many tools are available to generate these fragment libraries. Each of these tools follow a similar protocol to select fragments: first, the primary protein sequence is used to generate secondary structure predictions; next, the sequence, secondary structure predictions and NMR data (if available) are used to pick candidate 3- and 9-amino acid fragments from the PDB; finally, these candidate fragments are scored and the best N fragments are written to a fragment library file. The ROBETTA webserver (http://robetta.bakerlab.org) is available for non-commercial use, and allows users to generate fragment libraries for academic or research purposes using a simple interface. Additionally, Gront et al. have developed the FragmentPicker that provides users with total control over the fragment picking protocol (Gront, Kulp, Vernon, Strauss, & Baker, 2011).

The TopologyBroker, a tool that allows for more complex simulations, was one of the recent improvements added to Rosetta at the time of this work (Porter, Weitzner, & Lange, 2015). The conformational space searched during a *de novo* modeling simulation is vast, and successful searches often integrate prior knowledge with sampling. In *de novo* protein folding, this prior

knowledge may be in the form of β-strand pairing constraints as well as the formation of a rigid chunk based on a structurally homologous domain, to cite two examples. Previously, protocol developers were restricted to a sequential sampling approach in which Rosetta could readily violate one set of these constraints by sampling to satisfy the other, as it was in most cases unreasonable to write a unique sampling algorithm for each new combination of constraints. The TopologyBroker was developed to create a consensus sampling approach that satisfies all of the requested constraints without requiring additional code development for each unique system; instead, the Broker provides an API that allows for plug-and-play application to generate complex sampling strategies.

De novo structure prediction is a powerful tool. As such, it is critically important to understand the limitations of the algorithm. Rosetta performs well at folding small, globular, soluble proteins as well as small, simple membrane proteins containing 80-100 residues. However, large and complex proteins present additional difficulties that are not easily overcome by *de novo* techniques. In addition to *de novo* modeling of whole proteins, Rosetta is capable of *de novo* modeling of protein loops, an often-necessary step in protein homology modeling. The accuracy of protein loop modeling declines as the loops become longer based on the increasing degrees of freedom that must be sampled. *De novo* structural prediction algorithms sample many potential folds, and it is necessary to generate large numbers of models (10,000+) in order to adequately sample the native structure. Vast computational resources are needed to generate these numbers of models, and use of distributed computational methods (such as computational clusters) is recommended.

De novo protein structure prediction algorithms are regularly assessed in the Critical Assessment of protein Structure Prediction (CASP) and reviews of Rosetta's performance in

these assessments have been widely published (Ovchinnikov et al., 2015; Raman et al., 2009). In addition to assessments of its ability to *de novo* model whole proteins, a Rosetta-based group participated in the 2011 and 2014 Antibody Modeling Assessments (AMA and AMA-II), which compares state-of-the-art structure modeling approaches to model previously unpublished antibodies (Almagro et al., 2011; 2014; Weitzner, Kuroda, Marze, Xu, & Gray, 2014). These comparative assessments reveal weaknesses and bottlenecks in existing methodologies, identifying areas for development of future improvements.

I.7 Discussion

Antibodies play a critical role in human anti-influenza immunity. Despite continued vaccination efforts, IAV continues to cause high rates of disease, and recently generated mutations in influenza HAs have been shown to confer respiratory transmission mimicking the development of highly pathogenic IAV with pandemic potential (Herfst et al., 2012; Imai et al., 2012). It is critically important that we improve our understanding of the human immune response to IAV, and I propose to do so using a hybrid-methods technique. Emerging technologies such as NGS and computational structural modeling allow us to study the human antibody repertoire as an unbiased population, examining both sequence and structural features of anti-influenza antibodies. In this hybrid-methods approach, these techniques are married to traditional experimental methods that validate antibody function. In this thesis I will describe the relationships between sequence, structure and function of human antibodies. In the course of studying these relationships, I have developed a novel structure-based antibody discovery method and applied it to find new anti-influenza antibodies with therapeutic potential.

CHAPTER II

Impact of New Sequencing Technologies on Studies of the Human Antibody Repertoire

Adapted from Finn and Crowe, Current Opinion in Immunology, 2013

II.1 Introduction

Next generation sequencing (NGS) technologies emerged in the mid-2000's with the development of Roche's 454 pyrosequencing platform. The power of this technology was immediately recognized, and the NGS commercial sector has continued to evolve at a rapid rate. Newer platforms have since replaced the Roche 454, such as the Illumina MiSeq and HiSeq platforms. Many of these modern NGS technologies developed considerably while this body of work was being performed, improving in both quality and in quantity of reads produced. This exciting era in technological advancement allowed us to study, for the first time, the diverse antibody response to antigenic exposure (either through infection or through vaccination) as a population response instead of as a one-to-one interaction. One of the earliest goals of this work was to review the available literature from the field of antibody repertoire sequencing in order to come to an understanding of the strengths and limits of each of these emerging technologies and see how these technologies could be used to study the antibody repertoire. These early studies began to change our understanding of the mechanisms of repertoire diversification, the size of the antibody repertoire, and the methods of repertoire regulation shared among individuals, and provided the foundation upon which the rest of this thesis work was developed.

II.2 Sequencing the antibody repertoire

Next generation sequencing methods can be used to determine the sequence of recombined

immunoglobulin genes amplified from primary cell or tissue samples, generating large sequence databases that allow the antibody repertoire to be studied as a population. It is possible to sequence these recombined genes isolated from genomic DNA by PCR, or more commonly from transcribed genes using cDNA made from mRNA by reverse transcription, which is then amplified by PCR. The sequences of these resulting amplicons are then determined using high-throughput DNA sequencing technologies. Many of these NGS techniques were available at the time of this work, and specifications of these methods are detailed in Table 1. Undoubtedly, the capabilities and proprietary formats of these types of technologies will continue to evolve rapidly.

Table 1. Run statistics for common next generation sequencing techniques

	Roche 454 GS FLX Titanium [1]	**Illumina MiSeq [2]**	**Illumina HiSeq 2500 [3]**
Read Length	Up to 600 bp	Up to 500 bp (250 x 2)	Up to 200 bp (100 x 2)
Output	450 Mb	7.5-8.5 Gb	540-600 Gb
Reads / Run	700,000	15 Million	3 Billion
Quality	Consensus accuracy of 99.995%	> 75% bases above Q30	> 80% bases above Q30

[1] "GS FLX+ System." 454 Life Sciences, a Roche Company. Web. Accessed 05 Aug 2013. [2] "MiSeq Benchtop Sequencer Specifications." Illumina: sequencing and array-based solutions for genetic research. Web. Accessed 05 Aug 2013. [3] "HiSeq 2500/1500 Specifications." Illumina: sequencing and array-based solutions for genetic research. Web. Accessed 05 Aug 2013.

The antibody amplicon sequences that result from NGS must be analyzed with specialized software. Several web-based approaches to antibody gene analysis were available prior to this work, such as IMGT V-QUEST, JOINSOLVER and IgBLAST, each of which identify the inferred V(D)J gene segments used during recombination via template-based alignments, as well

as provide robust annotative data for further study (Alamyar, Duroux, Lefranc, & Giudicelli, 2012; Souto-Carneiro, Longo, Russ, Sun, & Lipsky, 2004; Ye, Ma, Madden, & Ostell, 2013). These tools are limited in their ability to efficiently process large datasets, and as NGS techniques continue to improve in throughput, these inefficiencies become more problematic. In addition to these tools we have developed PyIR (unpublished), a Python wrapper for IgBLAST that distributes execution over multiple CPUs, improving our ability to process millions of raw NGS reads. Although the output of IgBLAST is parsed to a user-friendly JSON format, our library does not manipulate the original data. Additional fields have been added to the original IgBLAST output for ease of analysis and reporting. For the purposes of this thesis, all sequences discussed herein have been analyzed using PyIR.

II.3 Endogenous heavy and light chain pairs

Antibody heavy and light chain pairing is an important aspect of the diversification of the antibody repertoire, and it has been shown that antibody heavy chains are capable of pairing with many light chains (Zhu et al., 2013). Therefore, identifying the correct heavy and light chain pairing partners during repertoire sequencing is of critical importance to understand repertoire diversity. At the time of this work, technical limitations prevented large-scale sequence analysis of naturally paired heavy and light chain genes. Two principal approaches were pursued to accomplish the task of pairing heavy and light chain genes on a massive scale. The first approach aimed to pair the heavy and light chain sequences from separately sequenced repertoires using informatic approximations, while the other approach aimed to link the sequences during variable gene amplification by PCR, followed by sequence analysis of both chains in one amplicon.

Indexed sequencing protocols can be readily applied to barcode both the heavy and light

chain sequences from a single sample, after which the heavy and light chain sequences can be paired. One study paired heavy and light chain variable gene sequences according to their relative frequencies within the repertoire, with a majority (21/27, or 78%) of the pairings tested generating antigen-specific antibodies (Reddy et al., 2010). A second study found that heavy and light chain pairs could be identified often using an evolution-based analysis, wherein coevolution of the heavy and light chains resulted in correlations between both the frequency and topology of the corresponding phylogenetic tree branches (Zhu et al., 2013). In either case, although these techniques may allow isolation of antigenic binding antibodies, they do not assuredly retain the original heavy and light chain gene pairing information. Later, techniques were developed to retain the endogenous pairing information by linking the heavy and light chains during gene amplification (DeKosky et al., 2013; White et al., 2011). In one study, single B cells were lysed in isolation using a high-density microwell plate, after which mRNA transcripts were captured on magnetic beads for emulsion PCR amplification with linking primers (DeKosky et al., 2013). This process annealed the heavy and light chain complementarity determining region 3 (HCDR3 and LCDR3, respectively) sequences together into one amplicon for NGS. A similar technique was employed by another study, which used advances in microfluidics to successfully accomplish on-chip single cell RT-qPCR (White et al., 2011). While published results are limited to 300 single-cell RT-qPCR measurements per run, the success of this protocol suggests that the chip could be scaled up to more than 1000 measurements per chip. While these techniques likely highlight the future of antibody repertoire studies, the current read lengths of next generation sequencing limits the application to only HCDR3:LCDR3 paired sequences. Longer read lengths will be required to identify full-length antibody variable gene sequences that can be used to synthesize cDNA encoding the native sequence of the original antibody including all six CDRs.

II.4 Mechanisms of repertoire diversification

NGS analysis of the antibody variable gene repertoire broadens our understanding of the critical V(D)J recombination events that are central to antibody repertoire diversification. While the mechanisms of recombination activating gene (RAG)-mediated V(D)J recombination are relatively well understood, rare events that occur during V(D)J recombination have been difficult to study using individual antibodies isolated by hybridoma or single B cell sorting techniques because of the limited scale of such techniques. In contrast, rare genetic events representing additional methods of repertoire diversification are observed readily in large antibody gene repertoire sequence databases generated by NGS.

For example, V(DD)J recombination events that appear to violate the 12/23 rule of recombination, occurring when the 12-bp recombination signal sequences (RSS) flanking the D gene segment incorrectly pair to allow fusion of two D gene segments. V(DD)J recombination has been observed in both *in vitro* and *in vivo* systems, but accurate calculations of the frequencies of these events were difficult to establish in the past. Furthermore, it was unclear if the perceived V(DD)J recombinations were instead artifacts of random N-additions that simply mimicked natural D gene genomic sequences. In one study, human peripheral blood antibody repertoires collected using Roche 454 technology were analyzed using stringent criteria that revealed that V(DD)J recombination events occur in approximately 1 in 800 circulating B human cells (Briney, Willis, Hicar, Thomas, & Crowe, 2012c). A second study of human peripheral blood antibody repertoires generated with Illumina HiSeq technology found that tandem D gene sequences occur in human pro-B cells more frequently than would be expected by random chance (Larimore, McCormick, Robins, & Greenberg, 2012). Additionally, these V(DD)J

recombination events appear to be selected against during B cell development, occurring at much lower frequencies in the population of productive antibody sequences. Analysis of the larger data set produced by Illumina HiSeq found V(DD)J recombination events in approximately 1 in 25,000 B cells.

These preliminary studies offer a glimpse into the depth of information made available by antibody repertoire analysis. While rare, there are unusual recombination events that contribute to repertoire diversification with unusual structural elements, such as the formation of long complementarity determining region 3 (CDR3) loops, which are important in the recognition and neutralization of viruses such as HIV. Repertoire sequencing has delineated particular areas of structural plasticity in immunoglobulins that accommodate insertions and deletions (Briney, Willis, & Crowe, 2012b); however, the studies reveal that most long heavy chain CDR3 (HCDR3) loops are formed at the time of recombination through use of long D and J segments and extended N addition regions, not by insertions (Briney, Willis, & Crowe, 2012a).

II.5 Predicting the repertoire size

Repertoire diversification leads to the generation of a large population of unique antibody sequences. It is theorized that the human antibody repertoire may contain up to 10^{11} unmutated sequences, however laboratory studies suggest the circulating population of B cells contains far fewer sequences.

To predict the size of the circulating antibody repertoire, one study applied the 'birthday paradox' from probability theory, which concerns the probability of two people in a population of *n* random individuals sharing a birthday; the paradox being that it takes far fewer individuals than would otherwise be assumed to generate a 99% probability that two share a birthday. The

study estimated that there are minimally 2 x 10^6 unique rearrangements in the peripheral blood compartment (Boyd et al., 2009). This algorithm, however, does not estimate the upper boundary of unique sequences due to the possibility of very low copy number sequences that are not observed using the then-current sequencing techniques.

Roche 454 sequencing can be used to generate large antibody sequence databases from human PBMC samples. Using these data, the total number of productive HCDR3 sequences in each of two healthy human subjects was calculated following a simple algorithm (Arnaout et al., 2011). The number of unique sequences added to the repertoire per 1000 additional sequences was counted and found to decrease regularly, following a pattern of logarithmic decay. The point where no additional unique sequences would be observed was calculated, and that value was expanded to encompass the total blood volume of a human adult. The upper bound of the circulating human HCDR3 repertoire was estimated to be between 3 and 9 million unique sequences (Arnaout et al., 2011). As discussed previously, the three principal recognized mechanisms underlying antibody diversification result in a theorized population of 10^{11} possible antibody sequences, far more than this technique predicted in the circulating repertoire.

II.6 Global repertoire regulation across individuals

Regulatory mechanisms exist that account for the inconsistency between the theorized number of possible recombined antibody sequences and the actual number of unique sequences observed to be circulating in the blood. For example, it is known that self-reactive antibodies are removed from the population by negative selection of B cells during early B cell development. Antibody repertoire studies have shown recently that these mechanisms of regulation seem common among many individuals, suggesting that global regulatory mechanisms may be more

sophisticated than previously theorized.

One study quantitated the presence of the same HCDR3 amino acid sequence in two different individuals, also referred to as the overlap of sequences between two repertoires (Arnaout et al., 2011). Synthetic HCDR3 repertoires then were generated computationally using knowledge-based rules developed from actual human antibody repertoires. The number of sequences the two synthetic data sets shared was related directly to the mutation rate used to develop those data sets. From these data, the researchers were able to determine that the overlap between two different HCDR3 repertoires occurs significantly more frequently than would be expected by chance, supporting the possibility of a global mechanism for antibody repertoire regulation (Arnaout et al., 2011). It is now possible to conceive of several types of repertoires (see Figure 3): (1) private repertoires, derived from the clones of one donor, (2) shared (or public) repertoires, representing antibody sequences found in more than one donor, and (3) global repertoires, representing the collection of all antibody sequences in a population of subjects.

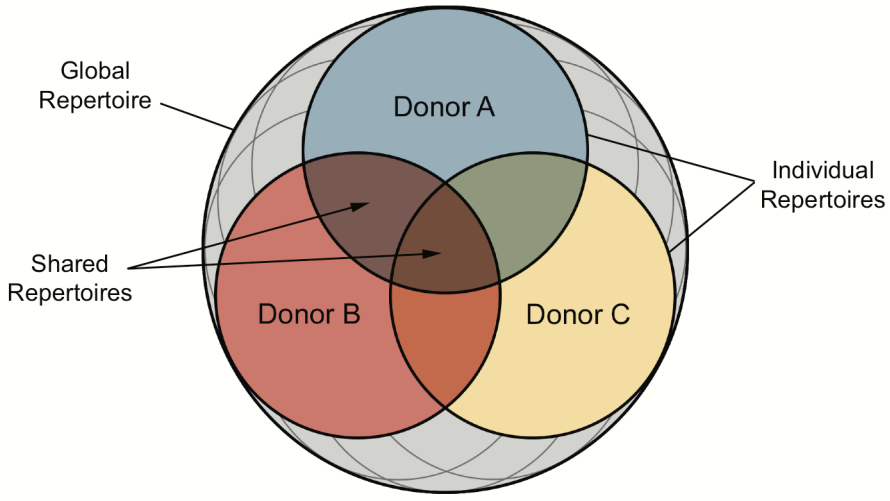


Figure 3. Types of repertoires: 1) Private, from one donor, 2) Shared, sequences found in two or more donors, and 3) Global, the sequences shared among an entire population of subjects.

Cell surface markers can be used to sort naïve and memory B cell subsets before high

throughput sequencing. In one study, such sequence data then were analyzed in the context of V(D)J recombinations to find that the hypothesized global mechanism of regulation results in increased oligoclonality in memory repertoire subsets when compared to the naïve B cell subset repertoire [17]. Furthermore, phylogenetic clustering revealed that subset repertoires cluster exclusively in an inter-donor dependent manner among four donors, revealing that the similarities between inter-donor repertoire subsets were significantly greater than the similarities between intra-donor repertoire subsets.

II.7 Identifying antibody clonal lineages

In collaboration with Sandhya Bangaru *et al.* (unpublished work), I developed a method to identify clonally related antibody sequences from NGS repertoires. Prior to this collaboration, Sandhya had discovered a novel anti-influenza antibody, FluA-20, using human hybridoma technology. FluA-20 recognizes a unique site on the HA head domain and is broadly specific to many subtypes of IAV, shown by capture ELISA against recombinant HA proteins derived from H1 (A/California/04/2009, A/Texas/36/1991), H3 (A/Hong Kong/1/1968, A/Victoria/3/1975), H7 (A/Shanghai/2/2013, A/Netherlands/219/2003) and H9 (A/Hong Kong/1073/99) subtypes (data not shown). This antibody was isolated from a donor with extensive influenza vaccination history, and its epitope may represent a new target for future influenza vaccines.

We were interested in studying the development of this antibody by tracing back the clonal lineage from which it arose. FluA-20 was isolated from a peripheral blood sample drawn day 31 after vaccination with the 2014-2015 trivalent influenza vaccine (TIV). Additional peripheral blood samples were drawn on days 0, 3, 4, 5, 6, 7, 10, 11 and 14 post-vaccination. Peripheral blood mononuclear cells (PBMCs) were isolated from these samples and frozen for future study.

After discovery of FluA-20, we prepared an antibody sequence database from these cryopreserved cells to perform clonal lineage analysis. Total RNA was isolated from 10 million PBMCs at each time point, and an antibody amplicon library was generated by RT-PCR as previously described (Bangaru et al 2016, Thornburg et al 2016). These antibody amplicon libraries were prepared using the Illumina TruSeq Library Preparation Kit (Illumina, FC-121-3001) and sequenced on an Illumina MiSeq using the PE-300 v3 reagent kit (Illumina, MS-102-3001). These raw sequences were analyzed using PyIR (unpublished), a Python wrapper for IgBLAST described in section II.2.

I performed a search for sequences clonally related to FluA-20 (*i.e.*, "siblings"). From the database of annotated antibody sequences obtained from this donor, I first filtered for antibody sequences with $V_H4\text{-}61/J_H4$ lineage. The HCDR3 region of these sequences was pairwise aligned to the HCDR3 of FluA-20 using a PAM30 matrix, with penalties for gap opening and gap extension of -14 and -3, respectively. HCDR3 sequences with a Hamming distance of <= 3 to FluA-20 were selected as siblings and the 'full length' nucleotide and amino acid sequence was queried from our database for further analysis. We identified siblings to FluA-20 in blood samples from four time points: days 5, 6, 11 and 14 post-vaccination with TIV. We inferred that the majority of these siblings arose from one common ancestor, and clustered into three major groups (designated Cluster A, B and C) that differ by point mutations across the VH gene region.

I constructed a network graph from the aligned, full-length sequences to visualize the relationships between these clusters (see Figure 4). Identical sequences were grouped into single nodes, and edges were drawn between two nodes if their Hamming distance was the lowest compared to all other nodes. Nodes denoting the inferred unmutated common ancestor (UCA) and the germline $V_H4\text{-}61/J_H4$ sequence were added manually. This network was visualized using

Cytoscape and manually adjusted for visual clarity (to prevent nodes from overlapping edges to which they are not connected, and to shorten distances between nodes that are closely related). The network analysis of these sequences revealed that FluA-20 arose from blasting cells present at day 6 that also were observed at day 14.

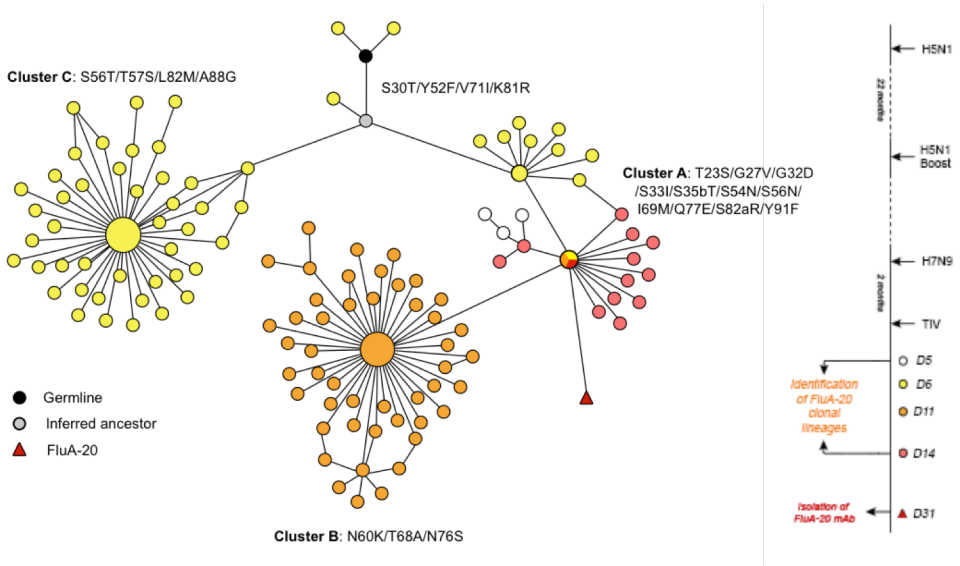


Figure 4. Network analysis of sequences clonally related to FluA-20. Nodes represent unique sequences observed in our database, with the size of the node correlating to the count of replicate sequences observed. The colour of each node denotes the time point at which it was found; white for day 5, yellow for day 6, orange for day 11 and red for day 14. The black node represents the $V_H4\text{-}61/J_H4$ germline sequence and the gray node represents an inferred unmutated common ancestor (UCA). The maroon, triangle-shaped node represents FluA-20. Edges drawn between nodes show that those sequences are more closely related to each other than to any other sequence. Edge distances are arbitrary and used only to visually clarify the graph.

To validate that these siblings were indeed related to FluA-20, we recombinantly expressed several of the identified antibodies as Fab fragments and Sandhya assessed their binding to recombinant HA via ELISA. Three of these sibling antibodies, Sib 2, Sib 3 and Sib 45, have very similar activity and breadth as FluA-20, despite mutations in the paratope. We identified two sibling antibodies in a cluster that had mutated more than FluA-20, Sib 28 and Sib 48, whose paratope mutations abrogated binding to some subtypes of H3 and H5, but not others. Finally, Sib 7 and Sib 33 lost activity to all 13 HAs tested, likely due to the addition of deleterious

somatic mutations (see Appendix 1).

In addition to testing somatic variants of FluA-20 identified by clonal lineage analysis, we expressed the FluA-20 UCA as recombinant Fab or IgG. Despite reverting 17 somatic mutations in the heavy chain variable gene and 12 mutations in the light chain variable gene, these recombinant UCA antibodies retained substantial binding breadth (see Appendix 1). As expected, however, both FluA-20 and the somatically mutated sibling antibodies we discovered displayed an increase in binding potency and breadth.

This study shows that NGS data can be leveraged to study the development of a broadly protective anti-influenza antibody. Clonal lineage assessment revealed that FluA-20, a monoclonal antibody discovered using human hybridoma technology, may have arisen from a naïve B cell carrying a low-affinity, broadly specific anti-HA antibody with a sequence similar to the inferred UCA. Based on our analysis, it is suggested that this B cell expanded as a plasmablast through days 5 and 6, undergoing somatic hypermutation to generate a population of somatic variants with differing functions that persisted at least until day 31 in the circulating repertoire. The knowledge gained from this study is helpful in the development of future IAV vaccines targeted at eliciting broadly protective anti-influenza antibodies.

II.8 Discussion

Recent and ongoing development of high-throughput amplicon sequence analysis techniques is providing a new and detailed view of the complexity and composition of human B cell repertoires. These technologies will continue to evolve, with the most likely next leap the acquisition of the ability to link heavy and light chain sequences at high throughput with high facility. Proteomics sequencing of the expressed antibody repertoire in serum is on the horizon

(Cheung et al., 2012). Robust computational methods for modeling the structure and function of antibodies, such as Rosetta, are starting to provide important insights and are a major focus of this thesis work (Willis, Briney, Deluca, Crowe, & Meiler, 2013). Early views of the human B cell repertoire suggest that the size of antigenic-specific clones that persist after exposure to foreign pathogens is much larger than previously thought. In contrast, and perhaps paradoxically, the size of the total antibody repertoire in circulating B cells may be orders of magnitude smaller than predicted. Future studies will need to address how large epitope-specific 'swarms' of somatic variants can be maintained in a repertoire of relatively small and fixed size without compromising responses to future exposures. Finally, the high level of concordance of the structure and size of repertoires between individuals suggests that there are strong regulatory programs that we understand only in part.

CHAPTER III

Improving Antibody Loop Modeling with Restraints

Adapted from Finn et al., PLOS ONE, 2016

III.1 Introduction

The field of antibody-mediated immunity has long benefited from structural studies of protein-protein interactions, in most cases through the determination of co-crystal structures of antibodies in complex with their antigens. Such studies often reveal the molecular mechanism of pathogen neutralization (Hashiguchi et al., 2015; Hong et al., 2013; Y. Li et al., 2011; Whittle, Zhang, Khurana, King, Manischewitz, Golding, Dormitzer, Haynes, Walter, Moody, Kepler, Liao, & Harrison, 2011a). However, the size and complexity of the antibody repertoire coupled with the substantial resources needed for experimental structure determination prohibit such studies on a comprehensive scale. B cell development leads to the generation of a large population of unique antibody proteins, and it is theorized that this diverse antibody repertoire may contain 10^{11} or more different protein sequences (Glanville et al., 2009; Trepel, 1974). Recent studies determined that the circulating antibody repertoire contains at least 10^{6} unique sequences, a number still far too large for comprehensive experimental structural studies (Arnaout et al., 2011; Boyd et al., 2009).

Analysis of antibody structures determined by X-ray crystallography revealed conservation of structural features even in the regions of the antibody with the most sequence diversity, the six complementarity determining region (CDR) loops, which are responsible for antigen binding. Three of these loops are contributed by the heavy chain component of the fragment variable (Fv) domain of the antibody (HCDRs), and three are contributed by the light chain Fv domain

(LCDRs). Two studies have identified robust rules that define canonical structures for five of the six CDR loops (Morea, Tramontano, Rustici, Chothia, & Lesk, 1998; North, Lehmann, & Dunbrack, 2011). However, the HCDR3 defies classification attempts. The HCDR3 is encoded by the junction of three gene segments (V, D and J genes) connected by random nucleotide additions or deletions that are not encoded in the antibody germline gene segments, but rather introduced by the host enzyme terminal deoxynucleotidyl transferase during antibody gene recombination. The HCDR3 is therefore significantly more diverse in sequence length and composition than the other CDR loops, which are encoded by either a single gene segment (heavy and light chain CDRs 1 and 2) or by a simplified junction (LCDR3) (Fanning, Connor, & Wu, 1996; Finn & Crowe, 2013; Tonegawa, 1983). As a result a large and diverse conformational space is observed for HCDR3s. Accordingly, HCDR3 is often especially important for antigen recognition and binding as has been revealed in previous structural studies (Weitzner, Dunbrack, & Gray, 2015).

The Rosetta software suite for macromolecular modeling can *de novo* predict the structure of a protein or portions thereof. The tertiary structure of a protein is determined from its primary sequence by pairing effective sampling techniques with knowledge-based energy functions. These energy functions for the most part assume that optimal geometries within proteins can be derived from a statistical analysis of the available structural information stored in the Protein Data Bank (Kaufmann, Lemmon, Deluca, Sheehan, & Meiler, 2010; Simons, Kooperberg, Huang, & Baker, 1997). Similar approaches are used during comparative modeling, when structurally divergent regions (typically loops) of otherwise homologous proteins must be predicted (Rohl, Strauss, Chivian, & Baker, 2004a). Rosetta is capable of predicting antibody structures with low root mean square deviation (RMSD) to experimental structures outside the

HCDR3; however accurately modeling the HCDR3 loop remains a challenge (Almagro et al., 2011; 2014; Sircar, Kim, & Gray, 2009; Weitzner et al., 2014).

In an effort to classify canonical structures of the HCDR3 loop, prior work has subdivided it into two domains: the less diverse "torso" and the more variable "head" (Morea et al., 1998; North et al., 2011) (see Figure 5). Two major families of canonical torso structures have been identified, and are referred to as "bulged" and "non-bulged" torsos (North et al., 2011). In this study, the geometries of the bulged torso domain have been used to develop restraints that restrict the sampling space of the HCDR3 torso and result in more native-like models when *de novo* modeling the entire HCDR3 loop.

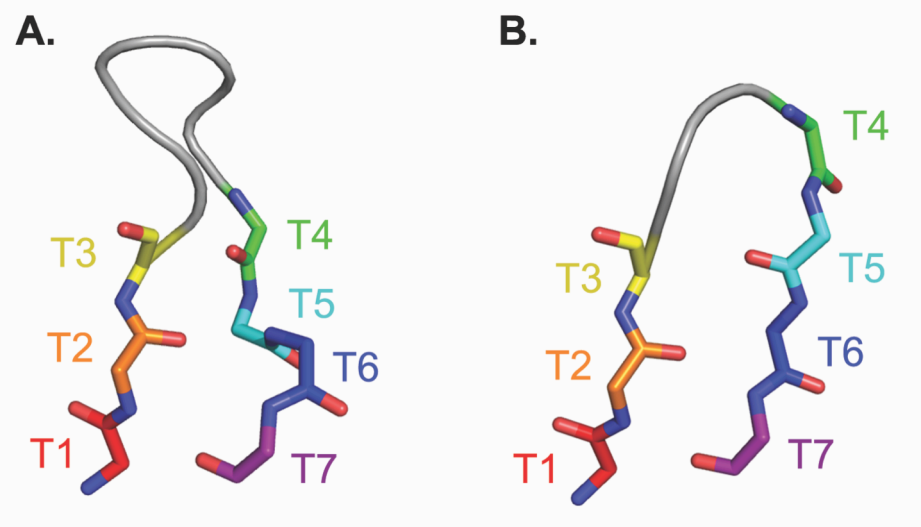


Figure 5. Defining the HCDR3 torso. The torso is defined as the first three and last four residues of the HCDR3 loop, numbered from T1 to T7. Main chain atoms are shown for bulged (panel A; PDBID 1UYW) and non-bulged (panel B; PDBID 2J88) torsos. In many (but not all) bulged torsos, a side-chain interaction between T2 and T6 causes the C-terminal side of the torso to bulge outward; the lack of such an interaction in non-bulged torsos leaves the beta-strand structure intact.

Previous studies have used restraints to model the bulged HCDR3 torso, following rules previously described by Shirai *et al*. wherein a pseudodihedral angle restraint was calculated from the Cα atoms of residues T5, T6, T7 and the following initial residue of Framework 4 to

define the bulged or non-bulged torso (Almagro et al., 2014; Shirai et al., 2014; Shirai, Kidera, & Nakamura, 1996; Weitzner et al., 2014; Whitelegg & Rees, 2000). Weitzner *et al*. utilized RosettaAntibody implemented within the Rosetta 3 framework to predict the structures of 11 previously unpublished antibody structures for the second antibody modeling assessment (AMA-II) (Weitzner et al., 2014). The longest HCDR3 loop in AMA-II contained 16 residues, and was predicted by the RosettaAntibody team with an RMSD of 3.70Å to the native HCDR3 loop. Shirai et al. also competed in AMA-II, and used their torso restraint rules to filter results generated by a pipeline that includes both Spanner and OSCAR for loop structure prediction; in comparison to the RosettaAntibody team described above, their best model for the longest HCDR3 loop had an RMSD of 3.29Å to the native HCDR3 loop (Shirai et al., 2014).

In this study, a novel set of restraints was tested on 28 previously crystallized apo human antibodies with HCDR3 loops of increasing length and structural complexity. We expect that these restraints will improve modeling of antibodies for which no structural information is available, providing a means by which comprehensive structural studies of antibodies may be accomplished.

III.2 Results

Measuring bulged and non-bulged torso dihedral angles

An annotated list of antibodies was used to cull experimentally derived structures from the Protein Data Bank (PDB), expanding upon the list published by North et al. (North et al., 2011). Following the IMGT conventions for defining the HCDR3, where the first HCDR3 residue occurs immediately following the V-gene residue Cys104 and the last HCDR3 residue occurs immediately preceding the J-gene residue Trp118, the torso is defined as the first three and the

last four residues of the HCDR3 (Lefranc et al., 2003; North et al., 2011). Accordingly, torso domain regions were pulled from these structures as two short peptide fragments (T1-T3 and T4-T7) and clustered using Rosetta at a threshold of 2 Å to separate bulged and non-bulged torsos. Previous studies identified a sequence motif (Arg or Lys at T2 and Asp at T6) that contributes to bulged torso formation in some but not all cases; these key residues were conserved in our bulged cluster, with 80% of bulged structures presenting Arg or Lys at T2, 73% presenting Asp at T6, and 65% retaining the complete T2/T6 sequence motif (see Appendix 2) (Morea et al., 1998; North et al., 2011). We found that germline-encoded regions of the antibody sequence often contribute these critical residues, as the end of the V gene segment contributes the first two to three torso residues while the J gene segment contributes the last four torso residues. The T2/T6 sequence motif that is often found in bulged torsos is present in 73% of naïve V and 92% of J germline gene allele segments (see Appendix 2.

The φ and ψ angles of the seven torso residues of each antibody structure were measured, with key differences between bulged and non-bulged torsos identified in the ψ angles of residues T4 and T6 (see Table 2). However, upon further study of previously defined torso clusters we observed that the ψ angle of T4 is able to form two distinct conformations in both bulged and non-bulged torso clusters, and the T4 ψ angle does not distinguish between bulged and non-bulged torso clusters; the differences we observed when comparing all bulged antibodies to all non-bulged antibodies were due to the limited sample size of structures available in the PDB for these sub-conformations (see Appendix 2) (North et al., 2011). This is in contrast to for example T5, where a larger standard deviation is observed but still a statistically significant preference for a smaller ψ angle in a bulged torso exists. Average φ and ψ angles were calculated as follows:

$$atan2\left(\frac{\sum \sin \alpha}{n}, \frac{\sum \cos \alpha}{n}\right) \quad (1)$$

An approximate standard deviation was found using the following equations. For the vector *v*:

$$\underset{v}{\rightarrow} = \left(\frac{\sin\alpha}{n}, \frac{\cos\alpha}{n}\right) \qquad (2)$$

Approximate standard deviation is calculated using:

$$\sqrt{2 \times \left[1 - \underset{v}{\rightarrow}\right]} \qquad (3)$$

It is worth noting that straightforward average and standard deviation calculations are insufficient when handling circular values such as dihedral angles.

Table 2. Bulged and non-bulged dihedral angle measurements.

Torso Residue	Bulged		Non-bulged	
	φ	ψ	φ	ψ
T1	-145 ± 9	148 ± 12	-146 ± 12	145 ± 16
T2	-101 ± 22	142 ± 13	-109 ± 20	136 ± 26
T3	-107 ± 32	137 ± 33	-119 ± 44	138 ± 51
T4	-121 ± 49	161 ± 48	-82 ± 49	3 ± 59
T5	-95 ± 35	98 ± 26	-126 ± 43	136 ± 53
T6	-87 ± 18	-30 ± 26	-118 ± 34	129 ± 24
T7	-126 ± 14	134 ± 10	-125 ± 19	136 ± 11

The average and standard deviation of φ and ψ angles were calculated from existing human and mouse antibody crystal structures available in the PDB. Torso structures were clustered as bulged (n = 218) and non-bulged (n = 38) using a cluster radius of 2 Å.

Derivation of restraints for bulged torso conformation

It has been observed that Rosetta rarely samples the bulged torso conformation when modeling HCDR3 loops (Weitzner et al., 2015). Due to this limitation, coupled with the greater amount of experimentally derived structural data available for bulged torsos than non-bulged torsos and the fact that bulged torsos are more prevalent in the human antibody repertoire, we chose to focus on developing restraints to improve modeling of HCDR3 loops with bulged

torsos. Rosetta uses a defined format to read in experimentally derived restraints. We used our measurements to generate dihedral angle restraints following a circular harmonic scoring function. Since the ψ angle measurement of T4 varies by 180 degrees between known bulged torso clusters, this measurement was omitted from our calculated restraints (see Appendix 2).

Modeling HCDR3 loops using bulged torso restraints

Following the protocol capture outlined in Supplemental Information, these restraints were used to model and score the HCDR3 loops from 28 benchmark antibodies whose structures had been previously determined by X-ray crystallography (see Table 3). These 28 benchmark structures represent HCDR3 lengths from 11 to 26 residues, with a mean length of 16 residues, spanning a range regularly observed in human antibody repertoires that also have a mean HCDR3 length of 16 amino acids (Briney, Willis, Hicar, Thomas, & Crowe, 2012d). Each of the benchmark antibodies was crystallized in the absence of an antigen (*i.e.*, apo) in order to avoid attempts to model conformations achieved by induced fit with a binding partner.

Restraints function as a penalty during Rosetta's scoring protocol, *i.e.*, a positive energy value is added when a dihedral angle leaves the allowed range. In this case, models formed with native-like bulged torso dihedral angles would have no (or only a very small) penalty from the restraint term, whereas models that deviated from the bulged torso dihedral angles would be penalized with a positive energy score. When restraints were applied during modeling, we observed a higher density of low-scoring, low-RMSD models (see Figure 6C, blue circles, n=26) than when modeling without restraints (see Figure 6A, blue circles, n=2). These low-scoring, low-RMSD models are defined as scoring in the top 10% of models, with Cα RMSD16 to the native structure of ≤ 2 Å (represented as blue circles, whereas models scoring below the top 10%

Table 3. Experimentally derived antibodies used to benchmark bulged torso restraints.

PDB ID	HCDR3 Length	Resolution (Å)	Source
1WT5	11	2.10	Humanized
2G75	11	2.28	Human
4G5Z	11	1.83	Human
3QRG	12	1.70	Human
4G6K	12	1.90	Humanized
4LLU	12	2.16	Human
1FVC	13	2.20	Humanized
3HI5	13	2.50	Human
4HFW	13	2.60	Human
4FQH	14	2.05	Human
4NM4	14	2.65	Human
8FAB	14	1.80	Human
3G6A	15	2.10	Human
3TNM	15	1.85	Human
3W9D	15	2.32	Human
1AQK	16	1.84	Human
1DQL	16	2.60	Human
1OM3	16	2.20	Human
1U6A	17	2.81	Human
3AAZ	17	2.20	Humanized
4M5Y	17	1.55	Human
3INU	18	2.50	Human
3QEH	18	2.59	Human
4F58	18	2.49	Human
1HZH	20	2.70	Human
4LKC	22	2.20	Human
1RHH	24	1.90	Human
4FNL	26	2.30	Human

28 high-resolution antibody structures solved by X-ray crystallography were used to benchmark the bulged torso restraints. Each of these antibody structures was solved in the absence of antigen (*i.e.*, apo structures) and all residues in the HCDR3 loops were resolved.

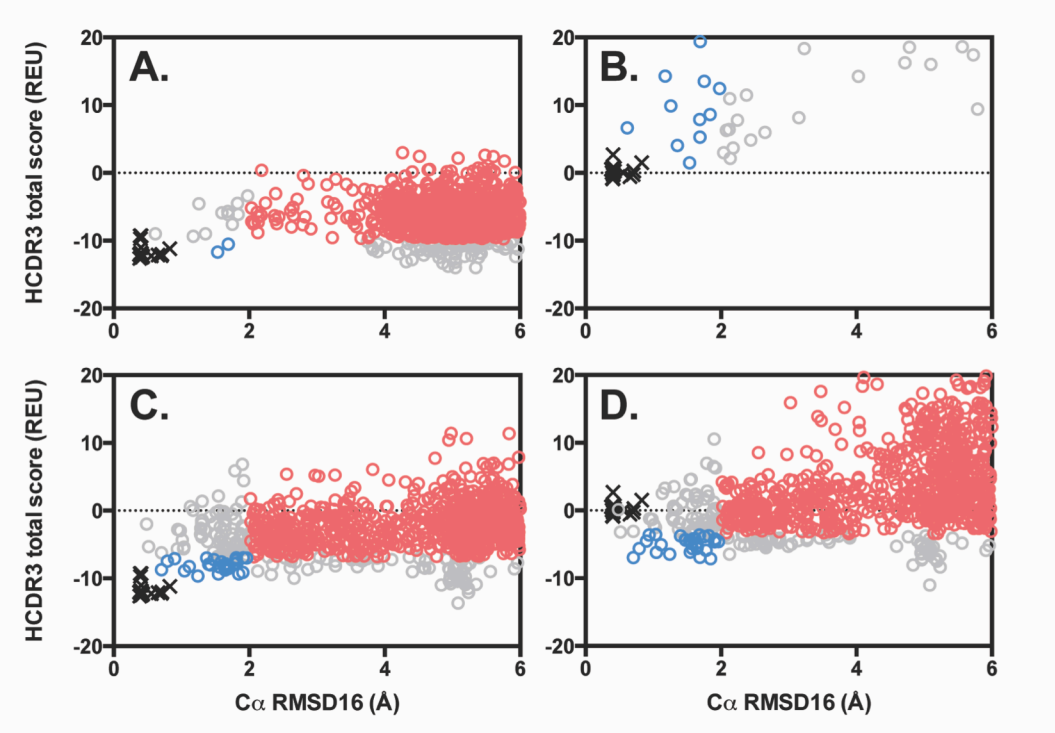


Figure 6. Bulged torso restraints improve native-like HCDR3 sampling and recovery. Using Rosetta LoopModel, 1,000 models of the benchmark antibody 4G5Z (circles) were generated with or without bulged restraints and these models were then scored with or without bulged restraints (panel A, modeled and scored without restraints; panel B, modeled without but scored with restraints; panel C, modeled with but scored without restraints; panel D, modeled and scored with restraints). The native crystal structure 4G5Z was also minimized using Rosetta FastRelax, generating 20 structures (black x's). The total HCDR3 score (in Rosetta Energy Units, or REU) is shown versus the Cα root mean square deviation of the HCDR3 loop, normalized to that of a protein loop containing 16 residues (RMSD16, in Å) to the native crystal structure. Models with scores ranked in the top 10% and RMSD16 ≤ 2 Å have been colored blue, while models with scores ranked below the top 10% and RMSD16 > 2 Å have been colored red. Improved native-like HCDR3 sampling is observed as a greater density of low RMSD16 models (blue circles) in comparison to Panel A, while improved model recovery is defined as a greater correlation between RMSD16 and score (colored vs. gray circles) in comparison to Panel A, as seen in panels C and D.

of models with Cα RMSD16 > 2 Å are represented as red circles in Figure 6 and Appendix 2. When restraints were applied during scoring but not during modeling (see Figure 6B) we found that the resulting models incur substantial restraint penalties due to non-native-like sampling of the torso domain, however the correlation between score and RMSD16 is improved. During application of this protocol wherein a native structure is unavailable, the ability to identify native-like models by score alone is extremely valuable. When applying restraints during both modeling and scoring, Rosetta generates a model population where an increased number of native-like structures correlate with low scores (see Figure 6D, blue circles, n=30) as compared to experiments modeled and scored without restraints (see Figure 6A, blue circles, n=2). Finally, we found that the application of these restraints results in more models whose backbone structures agree with bulged torso measurements defined in the literature (n=719 with restraints, n=33 without restraints; see Appendix 2) (Shirai et al., 1996; Weitzner et al., 2015).

The results of modeling the 28 benchmark HCDR3 loops with or without bulged torso restraints can be found in Figures 7-9. We observed changes in both conformational sampling and in model discretion by score when restraints were applied. To analyze improvements in conformational sampling, models were ranked by RMSD16 to their native structure (see Figure 7) and to study changes in scoring discretion the models were ranked by HCDR3 score (see Figure 8). Finally, models were clustered using a package called Calibur and the best cluster by average HCDR3 score was analyzed (see Figure 9).

Bulged HCDR3 restraints improve native-like conformational sampling

Modeling with bulged torso restraints improved native-like conformational sampling (the number of models with RMSD16 below 2 Å) in 26 out of 28 benchmark cases (see Figure 7); in

the remaining case of benchmark antibody 1RHH with an HCDR3 loop of 24 residues, no models below 2 Å were observed when modeling with or without restraints, and in the case of 4FNL with an HCDR3 loop of 26 residues, 2 models below 2 Å were observed when modeling without restraints, compared to no models sampled below 2 Å when modeling with restraints. On average, 90 models below 2 Å were generated with restraints, compared to only 12 models below 2 Å without restraints. The best RMSD sampled using bulged torso restraints was below 1 Å in 18 out of 28 cases with restraints, compared to 10 out of 28 cases without restraints. The average difference in the best RMSD sampled was 0.33 Å lower when restraints were applied during modeling. Furthermore, the average RMSD16 of the most native-like 10% of models (when ranked by RMSD16) is below 1 Å in 11 out of 28 cases when restraints are applied, compared to just 1 of 28 cases without restraints, revealing improved depth of high-resolution native-like sampling.

State-of-the-art computational methods to construct loop regions in proteins work reliably until about eight residues, and provide good results from some loops up to twelve residues. Beyond this limit, the conformational space often becomes too large to be sampled exhaustively. Many HCDR3 loops are longer and specialized methods are needed to limit the conformational space. Our analyses describe better sampling of native-like structures during modeling of these diverse HCDR3 loops when our torso restraints are used, with qualitative changes in performance observed at 14 and 18 amino acids.

		Without Restraints			With Restraints		
PDB ID	Length	Best RMSD16 sampled	Average RMSD16 of top 10 by RMSD16	Models below 2Å RMSD16	Best RMSD16 sampled	Average RMSD16 of top 10 by RMSD16	Models below 2Å RMSD16
1WT5	11	0.72	1.24	31	0.74	0.84	235
2G75	11	0.37	1.11	23	0.48	0.55	449
4G5Z	11	0.61	1.44	13	0.48	0.80	100
3QRG	12	0.72	0.84	26	0.61	0.69	167
4G6K	12	0.93	1.19	50	0.92	1.09	201
4LLU	12	1.16	1.20	44	0.59	0.75	296
1FVC	13	0.81	1.27	37	0.94	1.00	245
3HI5	13	1.37	1.58	22	0.95	1.40	90
4HFW	13	2.07	2.24	0	0.64	0.80	99
4FQH	14	0.75	1.98	4	0.46	0.70	25
4NM4	14	0.61	2.03	3	0.64	0.91	38
8FAB	14	1.39	1.52	24	0.95	1.16	146
3G6A	15	1.10	1.77	10	0.77	0.87	99
3TNM	15	0.76	1.80	8	0.76	1.12	30
3W9D	15	1.44	1.92	5	1.33	1.53	34
1AQK	16	1.39	1.78	8	0.96	1.24	65
1DQL	16	1.47	1.91	6	1.44	1.73	22
1OM3	16	1.84	2.14	1	1.10	1.39	31
1U6A	17	2.11	2.40	0	0.91	2.17	1
3AAZ	17	1.40	1.92	5	1.26	1.59	24
4M5Y	17	0.88	1.34	19	0.68	0.94	79
3INU	18	1.73	2.15	2	1.35	1.73	16
3QEH	18	1.90	2.30	2	1.76	2.05	4
4F58	18	2.15	2.48	0	1.45	2.02	2
1HZH	20	1.71	2.07	4	1.32	1.80	9
4LKC	22	2.04	2.38	0	0.71	1.66	7
1RHH	24	2.62	2.84	0	2.32	2.57	0
4FNL	26	1.75	2.33	2	2.09	2.25	0
Average	**16**	**1.35**	**1.83**	**12**	**1.02**	**1.33**	**90**

Figure 7. Torso restraints improve sampling of bulged HCDR3 loops. For each benchmark antibody structure, 1,000 models were generated with or without bulged torso restraints. The number of models below 2 Å RMSD16 to the native structure, the best RMSD16 sampled, and the average RMSD16 of the best 10 models ranked by RMSD16 are provided. For RMSD16-containing cells, blue shading represents RMSD16 ≤ 1 Å; yellow shading represents RMSD16 between 1 and 2 Å; red represents RMSD16 > 2 Å. For cells containing the number of models below 2 Å, blue shading represents ≥ 100 models; yellow shading represents ≥ 10 models; red shading represents fewer than 10 models.

Bulged HCDR3 restraints improve scoring discretion

The ability to identify native-like HCDR3 loops by score when *de novo* modeling using Rosetta is of critical importance. Unfortunately, we found the predictive ability of Rosetta's scoring function in the absence of restraints to be lacking; when ranking models by HCDR3 score, only 2 of 28 benchmark cases resulted in a top-scoring model with RMSD16 < 2 Å (see Figure 8). However when restraints were applied, ranking models by score resulted in 7 of 28 cases with an RMSD16 below 2 Å and two of those with RMSD16 below 1 Å (antibody 3QRG, 12 amino acids long and 4FQH, 14 amino acids long). On average, the RMSD16 of the best scoring model improved by 0.84 Å when restraints were used during modeling and scoring. Because restraints improve sampling, there was also a marked improvement in the average RMSD16 of the top 10 models ranked by score; when restraints are applied, the average is below 2 Å in 9 out of 28 cases, but no results below 2 Å were found when restraints were not used. On average, there is an improvement of 1.22 Å in the average RMSD16 of the top 10 models ranked by score. The average rank of the first model below 2 Å is 17 when restraints are applied and in 8 of 28 cases the first-ranking model is below 2 Å, compared to only 2 out of 28 cases resulting in a first-ranking model below 2 Å and an average rank of 82 when restraints are not used. Altogether these analyses reveal that the bulged torso restraints improve scoring discretion of native-like structures, but that further improvement to the scoring of HCDR3 loops is needed (Das & Baker, 2008).

		Without Restraints			With Restraints		
PDB ID	Length	RMSD16 of best scoring model	Average RMSD16 of top 10 by score	Rank of first model <2Å RMSD16	RMSD16 of best scoring model	Average RMSD16 of top 10 by score	Rank of first model <2Å RMSD16
1WT5	11	3.39	3.63	3	3.18	2.31	9
2G75	11	3.63	3.93	40	1.54	1.45	1
4G5Z	11	6.04	5.49	26	5.08	4.32	7
3QRG	12	4.05	2.56	3	0.85	1.15	1
4G6K	12	1.33	2.35	1	1.39	1.35	1
4LLU	12	3.50	3.34	23	2.52	1.99	4
1FVC	13	1.40	2.96	1	2.03	2.05	10
3HI5	13	2.23	3.67	66	1.80	1.72	1
4HFW	13	3.75	3.42	N/A	3.31	2.01	2
4FQH	14	4.26	4.66	27	0.47	1.83	1
4NM4	14	2.59	4.23	15	2.89	1.73	1
8FAB	14	4.21	4.02	4	1.71	1.59	1
3G6A	15	5.18	4.71	44	2.03	2.12	6
3TNM	15	3.14	3.13	2	3.29	3.18	15
3W9D	15	2.89	3.72	12	2.62	2.50	12
1AQK	16	4.48	4.23	25	1.81	1.94	1
1DQL	16	2.09	3.70	2	3.03	2.90	33
1OM3	16	3.48	3.70	281	2.65	2.87	51
1U6A	17	3.14	3.76	N/A	2.93	2.94	183
3AAZ	17	3.11	3.31	7	3.58	2.77	8
4M5Y	17	2.33	2.07	3	2.43	2.54	8
3INU	18	3.47	4.00	656	4.70	3.70	5
3QEH	18	2.48	3.63	176	2.49	3.12	16
4F58	18	5.01	4.75	N/A	4.69	3.28	19
1HZH	20	3.87	3.59	412	2.02	3.26	14
4LKC	22	4.82	4.12	N/A	3.08	3.44	39
1RHH	24	5.78	4.39	N/A	4.32	3.75	N/A
4FNL	26	3.86	3.97	46	3.33	3.00	N/A
Average	**16**	**3.55**	**3.75**	**82**	**2.71**	**2.53**	**17**

Figure 8. Torso restraints improve recovery of native-like bulged HCDR3 loops. For each benchmark antibody structure, 1,000 models were generated with or without bulged torso restraints. The number of models below 2 Å RMSD16 to the native structure, best RMSD16 sampled, average RMSD16 of the best 10 models ranked by RMSD16, RMSD16 of the best model ranked by Rosetta score, average RMSD16 of the top 10 models ranked by Rosetta score, and the rank of the first model below 2 Å when sorted by Rosetta score are provided. For RMSD16-containing cells, blue shading represents RMSD16 ≤ 1 Å; yellow shading represents RMSD16 between 1 and 2 Å; red represents RMSD16 > 2 Å. For rank-containing cells, blue shading represents rank 1; yellow shading represents ranks 2 to 10; red shading represents ranks > 10.

Clustering bulged HCDR3 loop models

Using the clustering package Calibur (S. C. Li & Ng, 2010), we analyzed the HCDR3 models generated with and without bulged restraints (see Figure 9). Only clusters containing >1% of models (10 or more) were considered. For models made based on structures with 20 or more amino acids in the HCDR3 loop, no sufficiently large clusters were found. For the other benchmark structures, clusters were sorted by average cluster HCDR3 score, with the lowest average HCDR3 score being chosen as the "correct" cluster. This approach to selecting the "correct" conformation is common when *de novo* modeling HCDR3 loops, as the native structure of the loop is not known outside of benchmark studies. When restraints were used during modeling, the rank of the cluster size (how large a cluster is compared to other clusters) improved in 18 out of 24 cases over experiments where restraints were not used. When restraints were applied during modeling, the average RMSD16 of the correct cluster improved in 21 out of 24 cases. The average RMSD16 for the best cluster by score was top-ranking in 9 out of 24 cases when restraints were applied during modeling, compared to just 3 out of 24 cases when restraints were not used, which reveals the predictive power of our scoring metrics when restraints are applied.

		Without Restraints				With Restraints			
PDB ID	Length	Best Average Cluster Score	Cluster Size (Rank)	Average Cluster RMSD16 (Rank)	Best RMSD16 in Cluster	Best Average Cluster Score	Cluster Size (Rank)	Average Cluster RMSD16 (Rank)	Best RMSD16 in Cluster
1WT5	11	-10.24	32 (5)	3.99 (7)	2.52	-3.92	178 (3)	1.93 (1)	0.75
2G75	11	-8.88	38 (4)	4.31 (6)	3.79	-4.84	527 (1)	1.49 (1)	0.48
4G5Z	11	-7.91	11 (7)	3.61 (2)	3.10	-0.36	370 (2)	1.98 (1)	0.37
3QRG	12	-8.30	15 (12)	3.13 (5)	2.01	-9.45	132 (1)	0.91 (1)	0.61
4G6K	12	-7.40	124 (2)	1.78 (1)	0.77	-8.13	129 (1)	1.17 (1)	0.85
4LLU	12	-7.95	43 (6)	3.12 (3)	2.62	-5.94	12 (13)	1.83 (4)	1.66
1FVC	13	-11.02	40 (8)	2.68 (2)	1.62	-7.68	404 (1)	1.91 (1)	0.94
3HI5	13	-9.88	64 (4)	1.92 (1)	1.21	-6.94	86 (1)	1.54 (1)	0.84
4HFW	13	-10.36	17 (11)	4.34 (14)	3.58	-5.18	11 (11)	3.53 (7)	3.02
4FQH	14	-11.26	14 (18)	4.64 (11)	3.99	-10.94	18 (12)	1.20 (1)	0.43
4NM4	14	-8.80	14 (12)	4.69 (14)	4.17	-3.68	16 (11)	2.58 (2)	1.80
8FAB	14	-11.70	11 (16)	4.29 (10)	3.61	-8.81	37 (4)	1.66 (2)	1.36
3G6A	15	-11.33	32 (7)	4.11 (9)	3.29	-5.45	238 (1)	2.17 (3)	0.77
3TNM	15	-14.16	10 (18)	2.91 (2)	2.51	-9.52	12 (9)	3.34 (5)	2.95
3W9D	15	-10.26	10 (6)	3.80 (5)	3.46	-6.39	24 (9)	2.81 (3)	2.04
1AQK	16	-10.82	19 (1*)	4.07 (1*)	3.60	-8.98	39 (1)	2.33 (2)	1.90
1DQL	16	-12.76	11 (1*)	2.84 (1*)	2.42	-9.44	66 (3)	2.70 (4)	1.74
1OM3	16	-10.91	19 (8)	3.27 (5)	2.71	-7.07	10 (21)	2.94 (7)	2.18
1U6A	17	-16.88	18 (7)	3.08 (2)	2.42	-3.53	71 (2)	3.29 (5)	2.75
3AAZ	17	-12.46	12 (1*)	4.25 (1*)	3.96	-12.60	29 (4)	2.52 (3)	1.41
4M5Y	17	-17.06	11 (16)	2.02 (1)	1.09	-13.59	113 (2)	1.92 (1)	0.71
3INU	18	-15.69	16 (3)	3.58 (2)	2.95	-13.19	26 (2)	4.26 (10)	3.35
3QEH	18	-12.27	11 (4)	4.33 (4)	4.05	-12.13	11 (1*)	3.57 (1*)	2.99
4F58	18	N/A	N/A	N/A	N/A	-7.89	13 (12)	2.88 (2)	2.40
Average	**15**	**-11.11**	**28 (9)**	**3.48 (5)**	**2.77**	**-7.54**	**111 (6)**	**2.30 (3)**	**1.54**

Figure 9. Cluster analysis of bulged HCDR3 loop modeling. Calibur was used to cluster the 1,000 models generated with or without bulged torso restraints for each antibody, using a threshold of 2.0. Clusters containing less than 1% of the total models were omitted from analysis; models generated for benchmark antibodies 4F58, 1HZH, 4LKC, 1RHH and 4FNL did not produce any large clusters upon analysis (N/A). Average Rosetta score was calculated for each cluster, and the cluster with the lowest average score was selected as the "correct" cluster. The size of this correct cluster (and it's rank among cluster sizes), its average RMSD16 to the native structure (and rank among average RMSD16 measurements) are provided. Cells containing rank data are shaded blue if the value represents the top rank, yellow for ranks 2-3, and red for ranks >3; if only one cluster (1*) was found, the cell is shaded gray. For RMSD16-containing cells, blue shading represents RMSD16 ≤ 1 Å; yellow shading represents RMSD16 between 1 and 2 Å; red represents RMSD16 > 2 Å. Values were omitted from column averages if ≤1 cluster was found.

III.3 Discussion

There is a growing body of work surrounding canonical structures of antibody CDR loops, first described by Chothia and colleagues and updated as recently at 2011 by the Dunbrack group (Morea et al., 1998; North et al., 2011). These groups have shown that that five of the six CDR loops take on canonical structures, and that the remaining HCDR3 forms only a few canonical classes of structure in its torso domain. Our work builds upon this background, and has led to the development of knowledge-based structural restraints from available crystal structures of HCDR3 loops with bulged torsos. We have shown that these restraints can be used to restrict the sampling space Rosetta searches during *de novo* loop modeling, limiting the torso domain to the φ and ψ angles of these residues that have been experimentally observed. These torso restraints improve native-like structure sampling and score-based differentiation of native-like HCDR3 models. We have also shown that such structural restraints improve Rosetta's ability to model longer HCDR3 loops than previously possible, extending the range of the technique to cover more biologically relevant HCDR3 loop lengths.

While this study focuses on benchmarking new knowledge-based restraints against antibodies whose structures have been experimentally determined, the true value of these restraints is in their ability to improve *de novo* antibody modeling. Such antibody structural predictions are a more rapid approach than experimental structural techniques, and can improve our understanding of host-pathogen interactions, provide insight into mechanisms of viral infection, and may lead to new monoclonal antibody therapeutics or vaccine candidates. Combined with our prior understanding of canonical CDR loops, which had made it possible to homology model much of the functional surface of the antibody (the "paratope") using Rosetta, we can now predict the remaining HCDR3 which is critical in many antibody-antigen

interactions. The central dogma of structural biology, that structure dictates function, lets us expect that improved accuracy in modeling HCDR3 will lead to improved accuracy in modeling antibody/antigen interactions which in turn leads to improved prediction of antibody function. We recognize that further experiments would be needed to prove this. Finally, upcoming advances in antibody sequencing, including the ability to sequence endogenously paired heavy and light chains, will provide the last critical insight in antibody modeling; we must now come to understand restrictions at the heavy and light chain interface that alter the paratope, and incorporate such restrictions into our structural predictions.

Although we have applied this approach to improving human antibody modeling, we recognize that this approach to structural restraint development is applicable to many other protein families in which structurally diverse surface loops with key functional importance are supported upon more structurally restricted framework regions (Das & Baker, 2008). Obvious examples include proteins with the PDZ domain and peptidase C1 domain protein families, which were found to use bulged HCDR3-like loops to recognize and bind their substrates (Weitzner et al., 2015). Finally, we have shown that knowledge-based structural restraints can be calculated easily and applied to improve modeling of novel loops not previously solved by experimental techniques, provided enough experimentally derived structural data is available for framework regions of functional loops in other protein families, and that canonical classes of those regions can be defined.

CHAPTER IV

Structure-Based Discovery of Human Anti-Influenza Antibodies

IV.1 Introduction

The amount of available antibody repertoire sequence information is expanding rapidly, but our ability to predict the function of antibodies from sequence alone is limited. Here, we describe a sequence-to-function prediction method that couples structural data for a single antibody/antigen complex with repertoire data. We used a position-specific structure-scoring matrix (P3SM) incorporating structure-prediction scores from Rosetta to identify antibody variable loops that have predicted structural similarity to the influenza virus-specific human antibody CH65. While a conventional sequence similarity search failed to identify new influenza antibodies, the P3SM approach identified new members of this class. Recombinant antibody expression, crystallography, and virus inhibition assays showed that the HCDR3 loops of the newly identified antibodies possessed similar structure and antiviral activity as the comparator CH65. This approach enables rapid discovery of new human antibodies with desired structure and function using cDNA repertoires that are obtained readily with current amplicon sequencing techniques.

Functional annotation of the emerging B and T cell immunome data will require development of new methods for predicting protein function from sequence. In some cases, when germline gene segment sequences encode a particular amino acid motif that binds an antigen in a canonical way, antibody specificity can be inferred because of germline gene usage. Examples include V_H*1-69*-encoded influenza hemagglutinin stem antibodies, V_H*1-02*-encoded HIV-1 CD4 binding site antibodies, or V_H*4-34*-encoded autoreactive antibodies that bind to

polylactosamine (Navis et al., 2014; Sui et al., 2009; Thompson et al., 1991). However, methods that predict the functional properties of encoded amino sequences without regard to the V_H gene segment origin or evolutionary gene history of an antibody are lacking. We interrogated a large repository of human antibody variable gene sequences from healthy individuals to identify antibodies with similar specificity and function to the influenza H1 HA-specific human antibody CH65 (Whittle, Zhang, Khurana, King, Manischewitz, Golding, Dormitzer, Haynes, Walter, Moody, Kepler, Liao, & Harrison, 2011b). Although a sequence motif has been described to assess CH65-like functionality, sequence-based similarity or motif searches of our antibody gene repertoires failed to identify clones that bound to influenza HA in a manner similar to CH65 (Schmidt, Therkelsen, et al., 2015b). The central dogma of structural biology states that sequence determines structure determines function, thus we hypothesized that an antibody search strategy that predicts structural similarity of antibody sequences and evaluates their fitness to bind antigen will have an increased sensitivity compared to a pure sequence-based search. In effect, such a structure-based search weights each sequence position based on the predicted consequences on structure and binding, while in a sequence-based search all sequence positions are weighted equally. To test this hypothesis, using only antibody sequences, we deployed a novel method to make predictions of antibody protein structures to identify a class of anti-influenza antibodies with members that shared structural features with the comparator antibody, bound to the same epitope on influenza HA, and mediated potent virus inhibition. By integrating an efficient sequence-based prediction method for the structure and function of antibodies, coupled with experimental data using rapid recombinant antibody expression, we accelerated discovery of diverse members of this antibody structural class. This sophisticated yet efficient method of predicting structural and functional networks of antibodies from sequences allows

rapid, targeted discovery of new antiviral antibodies and will facilitate improved understanding of the diversification of function in antibody structural families.

Antibody/antigen interactions are mediated by shape complementarity and biochemical properties of side chains in the interface. It is generally expected that more than one antibody amino acid sequence can achieve a given structural solution needed to interact with an antigen, but currently there are no efficient methods for predicting members of such structural classes based on antibody gene sequence alone. We recently developed a novel structure-based antibody discovery method that uses a P3SM to predict structural homologs rapidly from antibody gene sequences (Willis et al., 2016). Using Rosetta comparative modeling and a linear regression method to predict the thermostability and interaction energy of antibody heavy chain complementarity determining region 3 (HCDR3) loops with a target antigen, we screened a large database of antibody variable gene sequences. We selected sequences whose HCDR3 structure was predicted to be similar to the previously characterized monoclonal antibody CH65, regardless of their amino acid sequence. We found that when expressed in the context of the original antibody framework, these newly identified HCDR3 sequences functioned with similar specificity and affinity as CH65, even though they were unrelated in sequence and could not be identified by sequence similarity only. Crystal structures of the new antibodies identified by the P3SM search validated that the HCDR3 structures possessed a high degree of structural similarity to that of the original CH65 antibody on which the structural prediction was based.

IV.2 Results

Criteria for selecting a representative antibody

We sought to identify a structural class of antibodies, starting with a representative antibody

having an available co-crystal structure with its target antigen and known virus neutralizing function as a target for our structure-based search. The influenza HA-specific human antibody CH65 binds to many influenza virus type A subtype H1 HA proteins, and this antibody was crystallized in complex with the H1 HA (PDB ID: 5UGY) from A/Solomon Islands/3/2006 (designated here as SI06) (Whittle, Zhang, Khurana, King, Manischewitz, Golding, Dormitzer, Haynes, Walter, Moody, Kepler, Liao, & Harrison, 2011b). A genetically related antibody with a differing junctional sequence, designated CH67 (PDB ID: 4HKX), also was described (Schmidt et al., 2013). These antibodies possess interesting structural features that determine their function, namely a dipeptide motif on the tip of HCDR3 that interacts with the HA protein in a way that directly mimics the atomic features of the HA/sialic acid interaction. In addition, the CH65 and CH67 HCDR3 hypervariable loops exhibit a high level of thermostability, and preconfigure the paratope to bind HA, reducing entropic cost for an optimal interaction (H. Xu et al., 2015).

Sequence-based similarity searches have restricted efficacy

We used an existing database of antibody variable domain sequences from the Vanderbilt Vaccine Center Biorepository that were obtained by next generation amplicon sequencing of peripheral blood B cells from human subjects. The antibodies CH65 and CH67 are encoded by the antibody germline V_H and J_H genes *IGHV1-2* and *IGHJ6* and they possess a HCDR3 length of 19 amino acids. Our antibody variable gene database contained ~67,000 unique junctional sequences using *IGHV1-2/IGHJ6* and HCDR3 length of 19. We did not identify any HCDR3 sequences in our database with high sequence identity to CH65. We chimerized the junctional sequences with the highest ranked sequence similarity on CH65 and displayed these antibodies

as yeast surface-display scFv, however, we did not detect binding for any of those antibodies to the SI06 HA head domain protein in a flow cytometric yeast binding assay (see Table 4).

Table 4. CH65-like HCDR3s identified by sequence similarity did not bind SI06 HA.

ID	HCDR3 Sequence	Muts	Gaps	scFv Exp	scFv Bind
CH65	ARGGLEPRSVDYYYYGMDV	-	-	59%	91%
21415	ARGALEPRSQYYYYYGMDV	3	0	58%	0%
18723	ARGALEPRSRYYYYYGMDV	3	0	57%	0%
8890	ARGHLEPR-GDYYYYGMDV	2	1	58%	0%
11101	ARGGLEGR-V-YYYYGMDV	1	2	49%	0%
10116	ARG-LEPG-VDYYYYGMDV	1	2	55%	0%

From our human antibody sequencing database, the HCDR3 loop from all sequences sharing *IGHV1-2/IGHJ6* germline gene use were pairwise aligned to the native CH65 HCDR3 sequence. The number of mutations (muts) and sequence gaps were quantified, as shown. Sequences with the lowest total number of mutations and gaps were expressed as chimerized HCDR3 on yeast surface-display CH65 scFv. The percentage of cells expressing scFv as well as the percentage of scFv(+) cells that bound to SI06 HA in a flow cytometric assay are described. Although all of the sequences were expressed on the surface of yeast, none of the sequence-based HCDR3 siblings bound SI06 HA.

Filtering sequences using a position-specific structure scoring matrix

We next used a rapid method to screen the nucleotide sequence databases for predicted antibody structural similarity and antigen binding affinity using the P3SM method, a Rosetta-based heuristic model. Performance of this method varied between antibody/antigen systems, but we found that calculating the P3SM using linear ridge regression is a robust, reproducible way to map Rosetta Energy Units (REU) to a position-specific scoring matrix (see Figure 10).

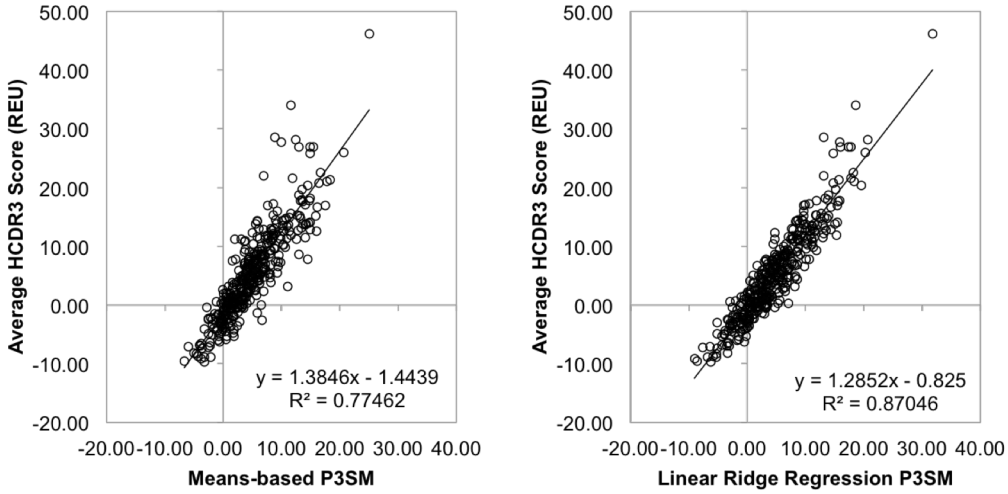


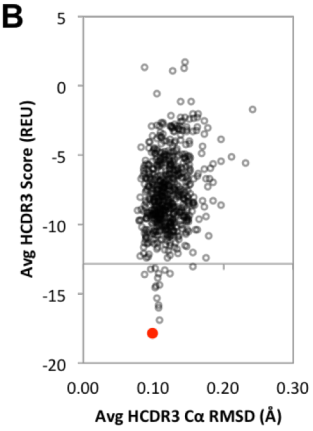
Figure 10. Linear ridge regression analysis improves the correlation between P3SM and Rosetta HCDR3 score. In the past, we calculated P3SMs using a means-based algorithm to determine the weighted score at each position (A). Both means-based and linear ridge regression P3SMs (B) correlate well with Rosetta homology modeling scores when searching for CH65-like antibodies, however linear-ridge regression analysis has been shown to be more robust across all antibody-antigen pairs (data not shown) and is now our preferred method.

We randomly selected 400 HCDR3 sequences from our antibody gene dataset and modeled each sequence ten times over each of the three CH65 Fab structures in the asymmetric unit of the co-crystal structure (PDB ID 5UGY). Using linear ridge regression analysis of the per-residue Rosetta scores for the best 5 out of 10 models for each sequence/Fab pair, we calculated the P3SM in which each cell of the matrix contained the weighted score for an amino acid at the given position (see Figure 11A). Next, we scored each HCDR3 sequence in our dataset with this P3SM and selected the top 600 rank-ordered sequences for further analysis (see Figure 11B). These 600 HCDR3 sequences were modeled by threading onto the CH65/SI06 HA complex crystal structure (PDB ID 5UGY) using Rosetta, and we filtered these hits by predicted thermostability and binding energy to identify 15 sequences for experimental validation (see Figure 11C). These HCDR3 sequences failed to score better than wild-type CH65 and lacked the Val106-Asp107 dipeptide motif previously described to be critical for CH65-HA binding. These

A

##	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
312	0.51	0.00	0.00	0.35	0.00	0.79	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.17	0.16	3.68	0.00	0.00
313	-0.32	0.00	0.00	0.00	0.00	0.65	0.00	2.25	1.35	0.00	0.00	0.00	0.87	-0.68	2.03	-1.00	0.05	0.14	0.00	0.00
314	-2.19	0.47	0.26	-0.72	-0.08	-1.93	2.11	3.44	1.09	0.30	-0.46	-0.19	1.84	-1.06	-0.47	-1.09	0.00	3.83	0.14	0.04
315	-1.53	0.60	-0.25	-0.44	0.90	-2.09	0.38	1.53	0.02	-0.83	-0.75	-0.50	5.84	-0.29	-0.45	0.07	0.88	0.92	1.14	0.18
316	0.48	0.09	-0.83	0.62	0.09	-0.28	0.91	-0.40	0.90	-0.95	-0.63	-0.03	5.64	-0.97	1.98	0.50	-0.72	-0.83	-0.81	0.56
317	-0.51	-0.95	1.05	0.54	0.74	0.79	-0.99	0.75	-0.25	1.10	-0.04	0.07	-0.57	0.11	0.06	0.90	-0.78	0.10	1.69	1.54
318	-2.97	-0.54	1.17	1.98	-1.50	-1.13	-0.43	2.06	0.00	3.86	0.56	-0.29	-3.70	0.65	3.56	-3.11	-0.99	2.62	2.64	0.89
319	0.06	0.18	-1.02	0.07	1.75	0.90	-0.43	1.15	-1.26	0.67	0.56	0.21	0.54	-0.19	-0.66	1.15	0.03	-0.39	1.72	0.31
320	0.63	0.96	-0.77	1.11	0.60	1.55	-0.50	0.14	-0.24	-0.13	-1.00	-0.64	3.22	0.87	-0.54	-1.19	-1.19	0.72	1.10	0.64
321	-0.40	-0.56	-1.07	-0.78	1.62	-0.72	1.18	1.26	1.09	-1.25	-0.47	1.36	0.87	0.05	1.54	0.17	-1.62	0.22	0.92	1.91
322	-0.26	-0.79	-0.10	-0.54	0.86	0.64	0.53	-0.39	1.27	-0.46	0.08	-0.31	2.66	1.70	0.40	-0.39	-0.06	-0.27	-0.95	1.73
323	1.48	0.05	-0.85	-0.32	-0.77	0.76	-0.36	0.20	0.09	1.55	0.68	-0.25	7.12	-0.37	-1.02	-0.55	-1.13	0.39	-0.64	-0.73
324	-1.90	-0.59	-0.84	0.48	0.92	0.81	0.76	2.01	0.23	2.45	-0.75	-1.02	-0.75	0.00	2.20	-0.92	0.74	0.77	2.10	-1.37
325	0.37	-0.14	0.18	1.19	-1.18	0.10	-1.28	0.33	-0.32	2.80	0.22	0.23	8.17	-1.61	-0.65	-0.82	1.29	-1.08	-0.52	-1.96
326	0.87	0.00	-0.40	0.15	-0.31	2.59	0.76	0.09	0.00	0.09	0.00	-0.84	0.38	-0.77	1.90	1.33	1.72	-0.57	-0.40	-1.27
327	0.87	0.11	1.61	-0.52	-0.23	0.00	0.20	0.00	0.60	0.00	0.00	0.00	0.51	0.00	-0.48	0.96	1.60	0.74	0.00	-0.62
328	0.49	0.00	0.00	0.00	0.50	0.13	0.00	-0.15	-0.17	1.18	-0.02	0.00	0.00	1.12	-0.07	0.00	0.66	-0.15	1.80	0.00
329	0.00	0.00	1.68	0.45	0.00	-0.05	0.11	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.13	0.00	1.72	1.23	0.00	0.00
330	0.36	0.00	-0.33	0.00	1.60	0.69	0.00	0.36	0.00	0.13	0.00	0.00	0.00	0.00	0.00	1.72	0.00	0.64	0.00	0.15

B



C

ID	HCDR3 Sequence	P3SM Score	Apo HCDR3 REU	Complex HCDR3 REU	HCDR3 Ca RMSD	Complex DDG	scFv Expression	scFv Binding
CH65	ARGGLEPRSVDYYYYGMDV	-10.35	-13.97	-17.85	0.0993	-20.25	59%	91%
1203*	ARAGWVPTTTSYYYYGMDV	-12.35	-12.56	-16.92	0.1087	-19.27	53%	0%
16996*	ARAGWVPTTTYYYYYGIDF	-9.39	-13.77	-16.00	0.1071	-17.68	46%	0%
8200*	ARAGWVPTTTYYYFYGMDV	-9.45	-12.55	-15.80	0.1070	-18.25	46%	0%
7969*	ARAGWVPTTTYYYYYGMDV	-10.23	-13.02	-15.31	0.1057	-18.08	51%	0%
18444*	ARAGWVPTTTYYYYYGMDG	-10.19	-12.87	-14.59	0.1031	-17.43	46%	0%
4492*	ARAGWVPTTTYYYYYGMGV	-11.97	-13.62	-14.46	0.1035	-17.81	49%	0%
11225*	ARAGWVPTTTYYYYYGVDV	-10.37	-13.01	-14.36	0.1094	-17.16	47%	0%
3083	ARGGWEPAADIYYYYGMDV	-9.24	-13.10	-13.86	0.1060	-16.99	54%	0%
10961	ARGGMTATTTNYYYYGMDV	-11.97	-10.74	-13.58	0.0874	-17.84	55%	0%
17038	ASGAIFPTNSYYYYYGMDV	-9.04	-15.39	-13.57	0.1078	-16.26	48%	0%
12374	ATTSQHSSSVLYYYYGMDV	-7.75	-13.10	-13.54	0.1241	-20.20	48%	0%
13458	ASAQATAGTLSYYYYGMDV	-11.18	-12.54	-13.26	0.1162	-15.86	50%	0%
9987	ARGGIAARRTHYYYYGMDV	-10.67	-12.11	-13.21	0.0964	-22.03	53%	0%
14857	ARGGVVPAATHYYYYGMDV	-9.34	-13.07	-13.16	0.1021	-17.99	56%	0%
3342	ATGLDTAMVTFCYYYYMDV	-9.13	-13.30	-13.04	0.1561	-14.69	52%	0%

Figure 11. The position-specific structure scoring matrix (P3SM) rapidly identifies potential structural homologs to CH65. In (A), linear ridge regression analysis of per-residue Rosetta energy scores from homology modeling simulations of 400 HCDR3 sequences over the CH65 crystal structure 5UGY determined a weight for each amino acid at each HCDR3 position (PDB residues 312-330), resulting in our position-specific structure scoring matrix (P3SM). In (B), the top 1000 sequences rank-ordered by P3SM score were each homology modeled on the CH65/SI06 complex structure, and the average score versus the root mean square deviation of the HCDR3 loop is shown. The native CH65 sequence is shown in red for comparison. Limited HCDR3 loop deviation was observed and sequences with good scores (low REU) retained the native CH65 structure during modeling. In (C), the best 15 homology modeled sequences rank-ordered by HCDR3 score were selected for further analysis of the antibody/antigen interaction. Sequence IDs denoted by an asterisk (*) belong to a cluster of clonally related sequences identified from a single donor. Apo (uncomplexed) HCDR3 scores as well as the complex DDG were calculated by separating the antibody from the antigen and rescoring the antibody while allowing limited side-chain minimization. These sequences were expressed as chimeric HCDR3 on yeast surface-display CH65 scFv, and the percentage of cells expressing scFv as well as the percentage of scFv(+) cells that bound to SI06 HA in a flow cytometric assay are listed.

15 HCDR3 sequences were expressed in the framework of CH65 on the surface of yeast as scFv, but as expected none bound to the SI06 HA protein when screened using a flow cytometric binding assay.

Within these 15 antibodies selected by our P3SM and filtering, however, we identified a cluster with similar HCDR3 sequences (see Figure 11C, denoted by asterisk). This cluster of sequences belongs to a population of related sequences observed in the sequence repertoires of one of the donors with repertoire sequence data obtained at four separate time points between 2004 and 2005. These samples are among the oldest in our database and were collected near in time to the discovery of the CH65 antibody, which was identified in a sample from 2008 after vaccination with the seasonal trivalent influenza vaccine (Schmidt, Do, et al., 2015a). In early 2009, significant changes were introduced to circulating H1s due to genetic reassortment between human and swine influenza viruses, and CH65 does not bind to the 2009 pandemic H1N1 virus HA. In keeping with our focus on CH65-like antibodies, we chose to narrow our study to these pre-2009 HCDR3 sequences.

First, we expressed two members of this family as antigen-binding Fabs using the full heavy chain variable domain sequence from our database to see if the lack of binding affinity we observed in yeast surface display was caused by the chimerization of HCDR3 sequences onto the CH65 framework in the scFv format. The database of antibody sequences we used does not contain linked heavy and light chain sequences, and the native light chain pairing for these antibodies is unknown, therefore we paired these heavy chains with either the CH65 or CH67 light chain. The resulting Fabs for those sequences identified using our P3SM bound SI06 HA exclusively when paired with the CH65 light chain, albeit at significantly lower affinity than wild-type Fab (see Figure 12).

A

Heavy Chain	CH65 EC50 (µg/mL)	CH67 EC50 (µg/mL)
wt	0.05	0.67
1203	13.30	> 50
7969	28.38	> 50

B

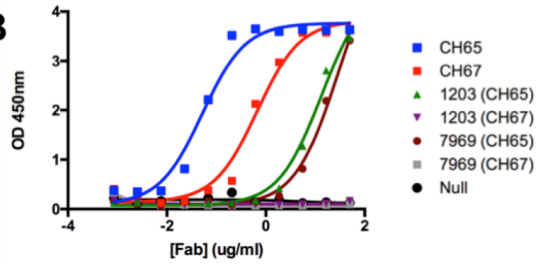


Figure 12. Two HCDR3 sequences identified using P3SM were chimerized into either CH65 or CH67 recombinant Fab backgrounds, and binding to SI06 HA is assessed as half-maximal binding (EC_{50} values, A) calculated from representative binding curves (B). The sequences identified using our method showed preferential bias to the CH65 Fab background.

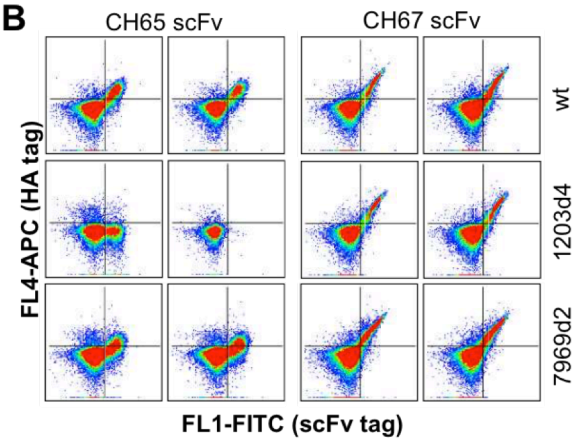
Minimal *in silico* affinity maturation rescues antibody function

To determine if minimal mutations to the HCDR3 of the chimeric antibodies rescued antigen binding, we performed *in silico* affinity maturation using Rosetta Design (see Figure 13A). Although a limited number of mutations improved the P3SM and Rosetta HCDR3 scores of many of these sequences, upon visual inspection of the models these mutations were not expected to contribute significantly to antibody/antigen binding (data not shown). Three mutations significantly improved the Rosetta score of sequence 1203, although only one altered the critical binding dipeptide (W101L, V102H and S107E). Rosetta converged on a similar solution for sequence 7969 (W101L, V102H and Y107D). These sequences were expressed as chimerized CH65 or CH67 scFv, and we observed native-like binding to SI06 HA for many of the designed HCDR3 sequences in the CH67 background (see Figure 13B, C). This bias toward better performance of the HCDR3s in the CH67 background was unexpected and may be caused by interactions between the HCDR3 loops and the CH67 light chain variable domain, or an overall improved fit between the CH67 variable domains in the scFv format.

A

Harrison	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116
IMGT	104	105	106	107	108	109	110	111	111.1	111.2	111.3	112.3	112.2	112.1	112	113	114	115	116	117	118
Kabat	92	93	94	95	96	97	98	99	100	100A	100B	100C	100D	100E	100F	100G	100H	100I	101	102	103
CH65	C	A	R	G	G	L	E	P	R	S	V	D	Y	Y	Y	Y	G	M	D	V	W
1203	C	A	R	A	G	W	V	P	T	T	T	S	Y	Y	Y	Y	G	M	D	V	W
1203d4	C	A	R	A	G	L	H	P	T	T	T	E	Y	Y	Y	Y	G	M	D	V	W
7969	C	A	R	A	G	W	V	P	T	T	T	Y	Y	Y	Y	Y	G	M	D	V	W
7969d2	C	A	R	A	G	L	H	P	T	T	T	D	Y	Y	Y	Y	G	M	D	V	W

B



C

	CH65 scFv		CH67 scFv	
HCDR3	Exp	Bind	Exp	Bind
CH65	21%	81%	24%	84%
CH67	22%	82%	23%	82%
1203d4	23%	3%	25%	88%
7969d2	21%	41%	30%	93%

Figure 13. *In silico* affinity maturation rescued function of P3SM-identified HCDR3 loops. (A) Sequence alignment of native and affinity matured HCDR3 sequences, with residues numbered in Harrison, IMGT and Kabat formats for reference. The dark red residues (C96 and W116) are the canonical flanking residues of the HCDR3 loop. Blue shading represents residues that differ from the wildtype CH65 sequence, while bold red highlighting represents residues that were mutated by Rosetta design. (B) These sequences were expressed as chimeric HCDR3 on yeast surface-display CH65 scFv. scFv expression was measured by tagging with FITC-conjugated anti-V5 antibody, while binding of scFv to biotinylated SI06 HA was measured by tagging with APC-conjugated streptavidin. Duplicate flow cytometric experiments are shown for each HCDR3 sequence in both CH65 and CH67 scFv backgrounds. (C) The percentage of cells expressing scFv as well as the percentage of scFv(+) cells that bound to SI06 HA in a flow cytometric assay are shown. Unexpectedly, the designed HCDR3 loops showed a preference for the CH67 scFv background.

To characterize the binding affinity and antiviral function of these designed HCDR3 loops further, we chimerized the HCDR3 loops on mammalian cell-expressed forms of CH65 or CH67 Fab proteins. We measured virus inhibition function using hemagglutinin inhibition (HAI) and virus microneutralization assays. The experiments for determining EC_{50} ($n = 2$) and IC_{50} ($n = 3$) were conducted four and three times independently, and representative data is presented in Table 5. Many of the designed HCDR3 sequences had comparable binding to HA as that of wild-type CH65 and CH67. These sequences also mediated similar levels of antiviral function in HAI and neutralization assays. The exception was HCDR3 1203d4, which showed a bias toward superior performance in the CH67 Fab background, but which performed poorly in the CH65 Fab background.

Table 5. *In silico* affinity maturation of P3SM-identified sequences rescues wildtype function.

Antibody	**HCDR3 Sequence**	**EC_{50} (μg/mL)**	**KD (nM)**	**HAI (μg/mL)**	**IC_{50} (μg/mL)**
CH65	ARGGLEPRSVDYYYYGMDV	0.0410	17.75 ± 0.08	6.25	0.7726
CH65:1203d4 Chimera	ARAGLHPTTTEYYYYGMDV	0.9172	116.2 ± 2.55	25.0	3.026
CH65:7969d2 Chimera	ARAGLHPTTTDYYYYGMDV	0.1159	36.52 ± 0.04	3.13	0.9642
CH67	ARAGLEPRSVDYYFYGLDV	0.5756	45.49 ± 0.46	12.5	0.6464
CH67:1203d4 Chimera	ARAGLHPTTTEYYYYGMDV	0.2568	55.55 ± 1.34	12.5	0.8325
CH67:7969d2 Chimera	ARAGLHPTTTDYYYYGMDV	0.1014	26.22 ± 0.02	6.25	0.5787
CH65 V106D	ARGGLEPRSDDYYYYGMDV	> 50	*n.d.*	*n.d.*	> 40
EEEV-16	ARADGYNFDY	> 50	*n.d.*	*n.d.*	> 40

Half-maximal binding determined by ELISA (EC50), binding affinity calculated using Octet BLI (KD), hemagglutinin inhibition (HAI) and half-maximal viral neutralization (IC50) of the designed HCDR3 sequences chimerized to CH65 or CH67 Fab are shown. Two negative controls were included in the study; CH65 V106D is a loss-of-function point mutant of the wildtype CH65 sequence, while EEEV-16 is a recombinant Fab not specific to influenza.

Experimental confirmation of structural similarity

Once we confirmed that these chimeric Fabs had similar function to CH65 and CH67, we sought to validate that our method correctly predicted their structural similarity. We successfully crystallized and solved structures for three of our four constructs as apo protein (*i.e.*, uncomplexed with HA; see Figure 14 and Table 6). Our structural analysis aimed to confirm that, like CH65 and CH67, these chimeras possess structural pre-configuration of the HCDR3 loop before binding the HA interface, taking on the conformation observed in the native crystal structures. Indeed, our three new structures exhibited HCDR3 conformations highly consistent with the available apo crystal structures of CH65 and CH67. When superimposed, chimeric HCDR3s 1203d4 and 7969d2 in the CH65 background had RMSD to the native CH65 HCDR3 of 0.198 Å and 0.219 Å, respectively. For chimeric HCDR3 7969d2 in the CH67 background, the RMSD to native CH67 HCDR3 was 0.685 Å; the larger deviation is caused mostly by a backbone movement at residue T104, which differs from the position of the native CH67 residue R104.

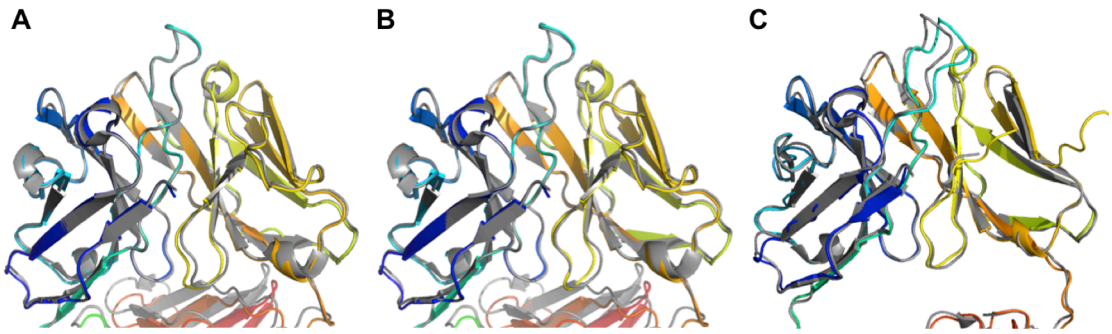


Figure 14. X-ray crystallography confirms structural homology of P3SM-selected antibody sequences. (A) CH65:7969d2 in rainbow, aligned to the variable domain of CH65 (PDBID 4WUK) in gray. (B) CH65:1203d4 in rainbow, aligned to the variable domain of CH65 (PDBID 4WUK) in gray. (C) CH67:1203d4 in rainbow, aligned to the variable domain of CH67 (PDBID 4HKB) in gray. In each structure, the HCDR3 loop is shown in cyan.

Table 6. Data collection and refinement statistics for P3SM-selected protein crystals.

Data collection			
Crystal	CH65:7969d2	CH65:1203d4	CH67:1203d4
PDB ID	6DLA	6DLB	6DL8
Wave Length (Å)	0.97750	0.97750	0.97750
Space group	$P2_12_12_1$	$P2_12_12_1$	P1
Unit cell dimensions			
a, b, c (Å)	57.4, 67.2, 130.4	57.5, 67.3, 130.8	72.2, 73.7, 76.9
α, β, γ	90.0, 90.0, 90.0	90.0, 90.0, 90.0	62.9, 88.9, 62.1
Resolution (Å)	46.80 – 2.00	46.91 – 2.20	40.53 – 3.80
Unique reflections	34481 (4897)	26563 (3819)	11641 (1661)
Redundancy	6.6 (6.5)	6.5 (6.3)	1.8 (1.7)
Completeness (%)	99.5 (98.6)	100.0 (100.0)	98.1 (95.3)
R_{merge} (%)	10.1 (39.1)	10.4 (61.4)	5.9 (11.3)
I/σ(I)	10.4 (4.0)	10.7 (3.1)	10.1 (6.5)
Refinement statitics			
R_{factor}	17.77	19.03	23.29
R_{free}	20.65	22.24	28.37
R.m.s.d. (bond) (Å)	0.0072	0.0057	0.0021
R.m.s.d. (angle) (deg)	0.836	0.599	0.581
Ramachandran plot			
Favored (%)	98.38	98.38	94.02
Allowed (%)	1.62	1.62	5.74
Outliers (%)	0.00	0.00	0.24

IV.3 Discussion

In this work, we present a framework for large-scale structure similarity prediction and functional assignment of human antibodies. Using next generation immunome data. We used this framework to identify new members of the CH65 antibody structural class, which binds to influenza HA in a manner similar to the native host ligand, sialic acid. The method predicts similarity of structure and function of antibody loops even though the loops diverge in sequence and genetic origin. Such structure-based functional assignment of Abs represents a new approach to human therapeutic monoclonal antibody discovery based on immunome sequencing data.

CHAPTER V

Concluding Remarks and Future Directions

V.1 Review

At the beginning of this thesis, I hypothesized that an understanding of antibody sequence and structure could be leveraged to predict the function of an antibody, and I sought to do so particularly in the context of identifying human antibodies targeting influenza HA. Emergent technologies such as next generation sequencing and computational structural modeling of antibodies aided these understandings, and were incorporated into methods I developed over the course of this work. These methods were successful at identifying novel anti-influenza antibodies using either sequence- or structure-based approaches, and we have shown that these antibodies would not have been discovered using methods that existed previously.

Sequence-based approaches, which identify functionally related antibodies by classifying members of clonal lineages from antibody repertoires, were made possible due to continued development of next generation sequencing techniques. These techniques and the antibodies discovered using them were discussed thoroughly in Chapter II. In brief, antibodies that are related by germline gene use and that have few amino acid mutations in the HCDR3 region (the V(D)J junctional site) are theorized to have arisen from the same naïve B cell. Upon recognition of antigen by this naïve B cell, somatic hypermutation and proliferation in the germinal center expands the single cell to a population of functionally-related, although not sequence-identical, daughter B cells. Previous techniques, including immortalization of human B cells using hybridoma technology, did not have the throughput necessary to identify multiple members of a specific B cell lineage within a single human donor. Next generation sequencing, on the other

hand, has allowed researchers for the first time to study millions of antibody sequences from many individuals in an economical and timely manner. I leveraged this technology to track, over time, the changing antibody repertoire of an individual from whom we had discovered an interesting anti-influenza monoclonal antibody (FluA-20) via hybridoma technology. We found that FluA-20 arose from a population of cells blasting on days 5 and 6 following vaccination with the 2014-2015 TIV. Furthermore, we found that recombinantly expressed members of this antibody's lineage were able to similarly bind a broad panel of influenza HAs despite mutations in the paratope, some paratope mutations altered breadth or abrogated binding to HA, and that the inferred UCA antibody retained some function despite the reversal of 29 amino acid mutations in the variable domain.

While we have shown that sequence-based discovery of human antibodies with targeted function is informative, we hoped to broaden our ability to define the function of "orphan" antibodies in our database; those for whom no clonally related antibodies had been previously discovered. Similar to Crick's famous "Central Dogma" (DNA makes RNA and RNA makes protein), the "Central Dogma of genomics and structural biology" states that an amino acid sequence determines a protein's structure, and that the structure of that protein determines its function. This is supported by evidence that antibodies bind to their antigenic partners via reversible non-covalent interactions that involve shape complementarity, electrostatic interactions, hydrogen bonds, and hydrophobic interactions. Shape complementarity is generated by the backbone dihedral angles of the amino acids in the CDR loops, while the latter three features are generated primarily by interactions of specific side chain residues within the protein-protein interface. Therefore we theorized that predicting the structures of these orphan antibodies from our sequence database would be possible, and that these structural predictions would bring

us closer to determining antibody function.

When I began this thesis work, it was known that five of the six CDR loops on the surface of the antibody took on canonical structures that could be determined by sequence alone. The remaining HCDR3 loop defied canonical definition, leading many researchers to determine the structure using *de novo* prediction techniques. *De novo* modeling of protein loops was restricted to relatively short loops containing 12 or fewer amino acids. We had determined, through next generation sequencing, that the average human HCDR3 loops were 16 amino acids long, with the longest HCDR3 loops containing more than 30 residues, far outside the possibilities of previous *de novo* modeling techniques. By analyzing the available crystallographic data deposited in the Protein Data Bank, which had been collected from hundreds of human antibodies, I was able to calculate a set of structural restraints that improves both sampling and scoring of HCDR3 loops during *de novo* antibody modeling in Rosetta. These results are discussed in detail in Chapter III of this work, but in brief, we were able to extend the capabilities of Rosetta *de novo* modeling to accurately predict HCDR3 loops containing upward of 17 amino acids.

While this study improved our ability to predict antibody structure, it also provided critical information about sequence-structure relationships and restrictions inherent in human (and perhaps other species') antibodies. We found that the "bulged torso" structure conserved in many antibodies is conferred via the germline genes used to form the junction. The torso domain is made up from the final two to three amino acids encoded by the V gene, and by the first three to four amino acids encoded by the J gene. Unless altered by exonucleases during V(D)J recombination, or by somatic hypermutation, these genes strictly encode for amino acid residues that will form the bulged beta structure that makes up this conserved domain.

Finally to bring together our understanding of antibody sequence, structure, and function, we

developed a novel structure-based antibody discovery method, discussed in Chapter IV. This method leverages structural similarity predicted from sequence to identify functionally related classes of antibodies. No method of structure-based antibody discovery existed before the development of this technology; the closest comparison is structure-based design algorithms, which could predict functional antibody sequences based on a desired structure, however these sequences were potentially non-native and may not naturally occur in a human repertoire. Using this technique, we were able to identify antibody sequences from a human donor that were not predicted to emulate the function of a target anti-influenza antibody by sequence similarity alone. The identified antibodies were, upon minimal *in silico* affinity maturation, capable of binding the target influenza HA with near-native affinity, and functioned similarly in hemagglutinin inhibition as well as neutralization assays. Furthermore, the recent influenza vaccination or infection status of the human donor was unknown, providing increased support for the ability of this technique to identify orphan antibodies that would not be otherwise obvious from sequence data alone.

V.2 Concluding Remarks

It is critically important that we further our understanding of the human immune system, particularly in regards to the sequence-structure-function relationships of antibodies. The ability to generate specific antibody responses, via diversification of the antibody repertoire, is a key element of acquired immunity and is necessary to effectively clear viral infections such as influenza. Our recent ability to study the antibody repertoire as a population, rather than individual members studied discretely, is due to emerging technologies such as next generation sequencing and computational structural modeling. Continued efforts to develop these

technologies have led to an increased ability to identify potential therapeutic antibodies. These efforts have also improved our understanding of the mechanisms of antibody repertoire regulation within and between individuals, which will inform future vaccination strategies. Finally, while applied to influenza virus for the duration of this work, these techniques are target-agnostic and able to be used to study human antibody responses to any number of infectious or immunogenic targets.

V.3 Future Directions

It remains both important and timely to continue leveraging new technologies in the study of influenza virus immunity. This year we mark the one hundredth anniversary of the 1918 influenza pandemic, one of the deadliest global health crises in human history. To date, our greatest weapon against another influenza pandemic is the seasonal influenza vaccine. However, data from recent years has shown that the seasonal influenza vaccine is less effective when circulating viruses are antigenically distinct from vaccine strains, revealing that the pandemic potential of influenza is unabated by these efforts.

One approach to combating influenza virus is the development of a "universal influenza vaccine" (UIV). The goal of such a vaccine is to trigger a broad, lasting antibody response capable of neutralizing many or all subtypes of influenza. This approach typically focuses on antibody responses to regions of influenza HA that are conserved between subtypes due to their importance in host receptor binding or viral fusion. In February of 2018, the National Institute of Allergy and Infectious Diseases (NIAID) released their plan for the development of a UIV that includes continued study and characterization of the antibodies elicited by influenza infection or vaccination. In addition, NIAID has called for increased study of rationally designed influenza

vaccines, to include novel immunogens and alternative vaccine delivery techniques.

I propose that a third critical aspect of antibody-mediated immunity must be studied to aid efforts in developing new vaccines, including UIV; global mechanisms of regulation that restrict the human antibody repertoire. The mechanisms that generate a diverse antibody repertoire are theorized to result in 10^{11} different antibody sequences, which are then subjected to strict selection criteria during negative selection in the bone marrow, and less understood selection criteria in the periphery. Antibody repertoire studies have shown that these mechanisms of regulation seem common among individuals, suggesting that global regulatory mechanisms may be more sophisticated than previously theorized. These studies were described in Chapter II, but in brief: one study of synthetic HCDR3 repertoires showed that the overlap between two different HCDR3 repertoires occurs significantly more frequently than expected by chance, and another study showed that some regulatory mechanism results in increased clonality as B cells progress from naïve to memory subsets. In a third study, phylogenetic clustering revealed that B cell subset repertoires (*i.e.,* naïve or memory subsets) cluster more closely to similar subsets in other donors (inter-donor) than they do to other subsets within the same donor (intra-donor). These findings support the existence of global mechanisms of regulation that are shared between individuals.

The population of circulating B cells resulting from such regulatory mechanisms has been shown to contain a limited number of unique sequences. Due to convergent evolution the likes of which we observed with our structure-based discovery method, even fewer unique structural solutions exist. These restrictions suggest that some antibody-mediated binding solutions may not naturally occur because they are eliminated from the antibody repertoire during negative selection or via later regulatory mechanisms, limiting our ability to successfully elicit these

responses by vaccination.

One initial study that may be performed with existing next generation sequencing technology involves sequence-based analysis of early B cell populations that are undergoing development in the bone marrow. At two critical stages, immunoglobulin loci are transcribed and expressed on the surface of the developing B cell. These checkpoints confirm proper gene recombination, structural stability of the resulting protein, and regular signaling function. In the first checkpoint, the recombined heavy chain loci is expressed on the surface of large pre-B cells in complex with the surrogate light chain made up of VpreB and λ5. VpreB expression can be used as a surface marker to differentiate large pre-B cells from other B cells in the bone marrow, allowing this population of cells to be sorted and sequenced independently. Sequence analysis of this population of cells would reveal the heavy chain recombinations that are able to be formed and expressed discrete from light chain recombination and pairing, which occurs later in B cell development. This study could reveal sequence restraints imposed on antibodies by V(D)J recombination machinery in the context of the cellular environment. The second checkpoint, which confirms expression and function of both heavy and light chain as a complete immunoglobulin molecule, occurs at the immature B cell stage prior to cells migrating to the periphery. These cells can be sorted using light chain as a surface marker, as this is the first and only time light chain is expressed during B cell development in the bone marrow. Studying the differences between this immature B cell population and the naïve B cell repertoire in peripheral tissues could identify specific sequences that are removed by regulatory mechanisms such as, but not limited to, negative selection of self-specific antibodies or selection against incompatible heavy-light chain pairs.

It is known that very flexible protein loops, referred to as loops with high plasticity, are able

to bind a broader range of targets in a less-specific manner than loops with more organized structure. I hypothesize that, in addition to sequence restraints imposed by V(D)J recombination machinery and structural restraints imposed by heavy-light chain pairing, additional structural restraints exclude certain antibody sequences from the naïve repertoire because stable antibody loops are required for normal B cell development. Secondary protein structure is known to stabilize long protein loops, as structural elements such as alpha helices and beta strands order the atoms in a protein backbone, limiting flexibility. Following the sequencing experiments proposed above, I recommend an additional study that measures the predicted secondary structure of HCDR3 loop sequences using available techniques such as PSI-PRED or JUFO. The existence of a positive correlation between HCDR3 loop length and secondary structure after negative selection would support the hypothesis that long loops must be stabilized to improve antibody function. In addition to secondary structural analysis, the tools developed within this work provide, for the first time, the ability to perform *de novo* tertiary structural predictions of most human HCDR3 loops with atomic-resolution accuracy. The presence of a subset of long, disordered, unstructured HCDR3 loops in early B cell populations in the bone marrow, but not in peripheral populations, would suggest that such loops are non-specific and potentially auto-reactive and are actively removed by negative selection.

These experiments are aimed at defining some of the global regulatory mechanisms that restrict the human antibody repertoire. Such restrictions are currently poorly understood, and increased study of these mechanisms will improve our knowledge of the scope and range of the human antibody repertoire. Future vaccination efforts, including development of the critical UIV, will be aided by an improved understanding of the naïve antibody repertoire that can be targeted and manipulated by vaccination. These studies are greatly aided by modern technologies such as

next generation sequencing and computational structural modeling, which must continue to be supported and developed for increased throughput and accuracy.

BIBLIOGRAPHY

Alamyar, E., Duroux, P., Lefranc, M.-P., & Giudicelli, V. (2012). IMGT(®) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods in Molecular Biology (Clifton, N.J.)*, *882*(Chapter 32), 569–604. http://doi.org/10.1007/978-1-61779-842-9_32

Almagro, J. C., Beavers, M. P., Hernandez-Guzman, F., Maier, J., Shaulsky, J., Butenhof, K., et al. (2011). Antibody modeling assessment. *Proteins*, *79*(11), 3050–3066. http://doi.org/10.1002/prot.23130

Almagro, J. C., Teplyakov, A., Luo, J., Sweet, R. W., Kodangattil, S., Hernandez-Guzman, F., & Gilliland, G. L. (2014). Second antibody modeling assessment (AMA-II). *Proteins*, *82*(8), 1553–1562. http://doi.org/10.1002/prot.24567

Arnaout, R., Lee, W., Cahill, P., Honan, T., Sparrow, T., Weiand, M., et al. (2011). High-resolution description of antibody heavy-chain repertoires in humans. *PloS One*, *6*(8), e22365. http://doi.org/10.1371/journal.pone.0022365

Avnir, Y., Tallarico, A. S., Zhu, Q., Bennett, A. S., Connelly, G., Sheehan, J., et al. (2014). Molecular signatures of hemagglutinin stem-directed heterosubtypic human neutralizing antibodies against influenza A viruses. *PLoS Pathogens*, *10*(5), e1004103. http://doi.org/10.1371/journal.ppat.1004103

Boyd, S. D., Marshall, E. L., Merker, J. D., Maniar, J. M., Zhang, L. N., Sahaf, B., et al. (2009). Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Science Translational Medicine*, *1*(12), 12ra23.

Briney, B. S., Willis, J. R., & Crowe, J. E. (2012a). Human peripheral blood antibodies with long HCDR3s are established primarily at original recombination using a limited subset of germline genes. *PloS One*, *7*(5), e36750. http://doi.org/10.1371/journal.pone.0036750

Briney, B. S., Willis, J. R., & Crowe, J. E. (2012b). Location and length distribution of somatic hypermutation-associated DNA insertions and deletions reveals regions of antibody structural plasticity. *Genes and Immunity*, *13*(7), 523–529. http://doi.org/10.1038/gene.2012.28

Briney, B. S., Willis, J. R., Hicar, M. D., Thomas, J. W., & Crowe, J. E. (2012c). Frequency and genetic characterization of V(DD)J recombinants in the human peripheral blood antibody repertoire. *Immunology*, *137*(1), 56–64. http://doi.org/10.1111/j.1365-2567.2012.03605.x

Caton, A. J., Brownlee, G. G., Yewdell, J. W., & Gerhard, W. (1982). The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). *Cell*, *31*(2 Pt 1), 417–427.

Cheung, W. C., Beausoleil, S. A., Zhang, X., Sato, S., Schieferl, S. M., Wieler, J. S., et al.

(2012). A proteomics approach for the identification and cloning of monoclonal antibodies from serum. *Nature Biotechnology*, *30*(5), 447–452. http://doi.org/10.1038/nbt.2167

Crotty, S., & Ahmed, R. (2004). Immunological memory in humans. *Seminars in Immunology*, *16*(3), 197–203. http://doi.org/10.1016/j.smim.2004.02.008

Das, R., & Baker, D. (2008). Macromolecular modeling with rosetta. *Annual Review of Biochemistry*, *77*(1), 363–382. http://doi.org/10.1146/annurev.biochem.77.062906.171838

DeKosky, B. J., Ippolito, G. C., Deschner, R. P., Lavinder, J. J., Wine, Y., Rawlings, B. M., et al. (2013). High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nature Biotechnology*, *31*(2), 166–169. http://doi.org/10.1038/nbt.2492

Ekiert, D. C., Bhabha, G., Elsliger, M.-A., Friesen, R. H. E., Jongeneelen, M., Throsby, M., et al. (2009). Antibody recognition of a highly conserved influenza virus epitope. *Science (New York, N.Y.)*, *324*(5924), 246–251. http://doi.org/10.1126/science.1171491

Fanning, L. J., Connor, A. M., & Wu, G. E. (1996). Development of the immunoglobulin repertoire. *Clinical Immunology and Immunopathology*, *79*(1), 1–14.

Finn, J. A., & 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. http://doi.org/10.1016/j.coi.2013.09.010

Fleishman, S. J., Whitehead, T. A., Ekiert, D. C., Dreyfus, C., Corn, J. E., Strauch, E.-M., et al. (2011). Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science (New York, N.Y.)*, *332*(6031), 816–821. http://doi.org/10.1126/science.1202617

Glanville, J., Zhai, W., Berka, J., Telman, D., Huerta, G., Mehta, G. R., et al. (2009). Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(48), 20216–20221. http://doi.org/10.1073/pnas.0909775106

Gront, D., Kulp, D. W., Vernon, R. M., Strauss, C. E. M., & Baker, D. (2011). Generalized fragment picking in Rosetta: design, protocols and applications. *PloS One*, *6*(8), e23294. http://doi.org/10.1371/journal.pone.0023294

Hashiguchi, T., Fusco, M. L., Bornholdt, Z. A., Lee, J. E., Flyak, A. I., Matsuoka, R., et al. (2015). Structural basis for Marburg virus neutralization by a cross-reactive human antibody. *Cell*, *160*(5), 904–912. http://doi.org/10.1016/j.cell.2015.01.041

Herfst, S., Schrauwen, E. J. A., Linster, M., Chutinimitkul, S., de Wit, E., Munster, V. J., et al. (2012). Airborne transmission of influenza A/H5N1 virus between ferrets. *Science (New York, N.Y.)*, *336*(6088), 1534–1541. http://doi.org/10.1126/science.1213362

Hong, M., Lee, P. S., Hoffman, R. M. B., Zhu, X., Krause, J. C., Laursen, N. S., et al. (2013). Antibody recognition of the pandemic H1N1 Influenza virus hemagglutinin receptor binding site. *Journal of Virology*, *87*(22), 12471–12480. http://doi.org/10.1128/JVI.01388-13

Imai, M., Watanabe, T., Hatta, M., Das, S. C., Ozawa, M., Shinya, K., et al. (2012). Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets. *Nature*, *486*(7403), 420–428. http://doi.org/10.1038/nature10831

Kadam, R. U., & Wilson, I. A. (2018). A small-molecule fragment that emulates binding of receptor and broadly neutralizing antibodies to influenza A hemagglutinin. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(16), 4240–4245. http://doi.org/10.1073/pnas.1801999115

Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H., & Meiler, J. (2010). Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry*, *49*(14), 2987–2998. http://doi.org/10.1021/bi902153g

Krause, J. C., Tsibane, T., Tumpey, T. M., Huffman, C. J., Basler, C. F., & Crowe, J. E. (2011a). A broadly neutralizing human monoclonal antibody that recognizes a conserved, novel epitope on the globular head of the influenza H1N1 virus hemagglutinin. *Journal of Virology*, *85*(20), 10905–10908. http://doi.org/10.1128/JVI.00700-11

Krause, J. C., Tsibane, T., Tumpey, T. M., Huffman, C. J., Briney, B. S., Smith, S. A., et al. (2011b). Epitope-specific human influenza antibody repertoires diversify by B cell intraclonal sequence divergence and interclonal convergence. *Journal of Immunology (Baltimore, Md. : 1950)*, *187*(7), 3704–3711. http://doi.org/10.4049/jimmunol.1101823

Larimore, K., McCormick, M. W., Robins, H. S., & Greenberg, P. D. (2012). Shaping of human germline IgH repertoires revealed by deep sequencing. *Journal of Immunology (Baltimore, Md. : 1950)*, *189*(6), 3221–3230. http://doi.org/10.4049/jimmunol.1201303

Lefranc, M.-P., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., et al. (2003). IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Developmental and Comparative Immunology*, *27*(1), 55–77.

Li, S. C., & Ng, Y. K. (2010). Calibur: a tool for clustering large numbers of protein decoys. *BMC Bioinformatics*, *11*(1), 25. http://doi.org/10.1186/1471-2105-11-25

Li, Y., O'Dell, S., Walker, L. M., Wu, X., Guenaga, J., Feng, Y., et al. (2011). Mechanism of neutralization by the broadly neutralizing HIV-1 monoclonal antibody VRC01. *Journal of Virology*, *85*(17), 8954–8967. http://doi.org/10.1128/JVI.00754-11

Morea, V., Tramontano, A., Rustici, M., Chothia, C., & Lesk, A. M. (1998). Conformations of the third hypervariable region in the VH domain of immunoglobulins. *Journal of Molecular Biology*, *275*(2), 269–294. http://doi.org/10.1006/jmbi.1997.1442

Navis, M., Tran, K., Bale, S., Phad, G. E., Guenaga, J., Wilson, R., et al. (2014). HIV-1 receptor binding site-directed antibodies using a VH1-2 gene segment orthologue are activated by Env trimer immunization. *PLoS Pathogens*, *10*(8), e1004337. http://doi.org/10.1371/journal.ppat.1004337

North, B., Lehmann, A., & Dunbrack, R. L. (2011). A new clustering of antibody CDR loop conformations. *Journal of Molecular Biology*, *406*(2), 228–256. http://doi.org/10.1016/j.jmb.2010.10.030

Ovchinnikov, S., Kinch, L., Park, H., Liao, Y., Pei, J., Kim, D. E., et al. (2015). Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife*, *4*, e09248. http://doi.org/10.7554/eLife.09248

Porter, J. R., Weitzner, B. D., & Lange, O. F. (2015). A Framework to Simplify Combined Sampling Strategies in Rosetta. *PloS One*, *10*(9), e0138220. http://doi.org/10.1371/journal.pone.0138220

Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., et al. (2009). Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins*, *77 Suppl 9*(S9), 89–99. http://doi.org/10.1002/prot.22540

Ramaraj, T., Angel, T., Dratz, E. A., Jesaitis, A. J., & Mumey, B. (2012). Antigen-antibody interface properties: composition, residue interactions, and features of 53 non-redundant structures. *Biochimica Et Biophysica Acta*, *1824*(3), 520–532. http://doi.org/10.1016/j.bbapap.2011.12.007

Reddy, S. T., Ge, X., Miklos, A. E., Hughes, R. A., Kang, S. H., Hoi, K. H., et al. (2010). Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nature Biotechnology*, *28*(9), 965–969. http://doi.org/10.1038/nbt.1673

Rohl, C. A., Strauss, C. E. M., Chivian, D., & Baker, D. (2004a). Modeling structurally variable regions in homologous proteins with rosetta. *Proteins*, *55*(3), 656–677. http://doi.org/10.1002/prot.10629

Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., & Baker, D. (2004b). Protein structure prediction using Rosetta. *Methods in Enzymology*, *383*, 66–93. http://doi.org/10.1016/S0076-6879(04)83004-0

Schmidt, A. G., Do, K. T., McCarthy, K. R., Kepler, T. B., Liao, H.-X., Moody, M. A., et al. (2015a). Immunogenic Stimulus for Germline Precursors of Antibodies that Engage the Influenza Hemagglutinin Receptor-Binding Site. *Cell Reports*, *13*(12), 2842–2850. http://doi.org/10.1016/j.celrep.2015.11.063

Schmidt, A. G., Therkelsen, M. D., Stewart, S., Kepler, T. B., Liao, H.-X., Moody, M. A., et al. (2015b). Viral receptor-binding site antibodies with diverse germline origins. *Cell*, *161*(5),

1026–1034. http://doi.org/10.1016/j.cell.2015.04.028

Schmidt, A. G., Xu, H., Khan, A. R., O'Donnell, T., Khurana, S., King, L. R., et al. (2013). Preconfiguration of the antigen-binding site during affinity maturation of a broadly neutralizing influenza virus antibody. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(1), 264–269. http://doi.org/10.1073/pnas.1218256109

Shirai, H., Ikeda, K., Yamashita, K., Tsuchiya, Y., Sarmiento, J., Liang, S., et al. (2014). High-resolution modeling of antibody structures by a combination of bioinformatics, expert knowledge, and molecular simulations. *Proteins*, *82*(8), 1624–1635. http://doi.org/10.1002/prot.24591

Shirai, H., Kidera, A., & Nakamura, H. (1996). Structural classification of CDR-H3 in antibodies. *FEBS Letters*, *399*(1-2), 1–8.

Simons, K. T., Kooperberg, C., Huang, E., & 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. http://doi.org/10.1006/jmbi.1997.0959

Sircar, A., Kim, E. T., & Gray, J. J. (2009). RosettaAntibody: antibody variable region homology modeling server. *Nucleic Acids Research*, *37*(Web Server issue), W474–9. http://doi.org/10.1093/nar/gkp387

Souto-Carneiro, M. M., Longo, N. S., Russ, D. E., Sun, H.-W., & Lipsky, P. E. (2004). Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER. *Journal of Immunology (Baltimore, Md. : 1950)*, *172*(11), 6790–6802.

SRIWILAIJAROEN, N., & SUZUKI, Y. (2012). Molecular basis of the structure and function of H1 hemagglutinin of influenza virus. *Proceedings of the Japan Academy. Series B, Physical and Biological Sciences*, *88*(6), 226–249. http://doi.org/10.2183/pjab.88.226

Sui, J., Hwang, W. C., Perez, S., Wei, G., Aird, D., Chen, L.-M., et al. (2009). Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. *Nature Structural & Molecular Biology*, *16*(3), 265–273. http://doi.org/10.1038/nsmb.1566

Thompson, K. M., Sutherland, J., Barden, G., Melamed, M. D., Randen, I., Natvig, J. B., et al. (1991). Human monoclonal antibodies against blood group antigens preferentially express a VH4-21 variable region gene-associated epitope. *Scandinavian Journal of Immunology*, *34*(4), 509–518.

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

Trepel, F. (1974). Number and distribution of lymphocytes in man. A critical analysis. *Klinische Wochenschrift*, *52*(11), 511–515.

Weitzner, B. D., Dunbrack, R. L., & Gray, J. J. (2015). The origin of CDR H3 structural diversity. *Structure (London, England : 1993)*, *23*(2), 302–311. http://doi.org/10.1016/j.str.2014.11.010

Weitzner, B. D., Kuroda, D., Marze, N., Xu, J., & Gray, J. J. (2014). Blind prediction performance of RosettaAntibody 3.0: grafting, relaxation, kinematic loop modeling, and full CDR optimization. *Proteins*, *82*(8), 1611–1623. http://doi.org/10.1002/prot.24534

White, A. K., VanInsberghe, M., Petriv, O. I., Hamidi, M., Sikorski, D., Marra, M. A., et al. (2011). High-throughput microfluidic single-cell RT-qPCR. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(34), 13999–14004. http://doi.org/10.1073/pnas.1019446108

Whitelegg, N. R., & Rees, A. R. (2000). WAM: an improved algorithm for modelling antibodies on the WEB. *Protein Engineering*, *13*(12), 819–824.

Whittle, J. R. R., Zhang, R., Khurana, S., King, L. R., Manischewitz, J., Golding, H., Dormitzer, P. R., Haynes, B. F., Walter, E. B., Moody, M. A., Kepler, T. B., Liao, H.-X., & Harrison, S. C. (2011a). Broadly neutralizing human antibody that recognizes the receptor-binding pocket of influenza virus hemagglutinin. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(34), 14216–14221. http://doi.org/10.1073/pnas.1111497108

Wiley, D. C., Wilson, I. A., & Skehel, J. J. (1981). Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature*, *289*(5796), 373–378.

Willis, J. R., Briney, B. S., Deluca, S. L., Crowe, J. E., & Meiler, J. (2013). Human germline antibody gene segments encode polyspecific antibodies. *PLoS Computational Biology*, *9*(4), e1003045. http://doi.org/10.1371/journal.pcbi.1003045

Willis, J. R., Finn, J. A., Briney, B., Sapparapu, G., Singh, V., King, H., et al. (2016). Long antibody HCDR3s from HIV-naïve donors presented on a PG9 neutralizing antibody background mediate HIV neutralization. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(16), 4446–4451. http://doi.org/10.1073/pnas.1518405113

Xu, H., Schmidt, A. G., O'Donnell, T., Therkelsen, M. D., Kepler, T. B., Moody, M. A., et al. (2015). Key mutations stabilize antigen-binding conformation during affinity maturation of a broadly neutralizing influenza antibody lineage. *Proteins*, *83*(4), 771–780. http://doi.org/10.1002/prot.24745

Xu, R., Krause, J. C., McBride, R., Paulson, J. C., Crowe, J. E., & Wilson, I. A. (2013). A recurring motif for antibody recognition of the receptor-binding site of influenza hemagglutinin. *Nature Structural & Molecular Biology*, *20*(3), 363–370. http://doi.org/10.1038/nsmb.2500

Ye, J., Ma, N., Madden, T. L., & Ostell, J. M. (2013). IgBLAST: an immunoglobulin variable

domain sequence analysis tool. *Nucleic Acids Research*, *41*(Web Server issue), W34–40. http://doi.org/10.1093/nar/gkt382

Zhu, J., Ofek, G., Yang, Y., Zhang, B., Louder, M. K., Lu, G., et al. (2013). Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(16), 6470–6475. http://doi.org/10.1073/pnas.1219320110

APPENDIX 1

Supplemental Information for Chapter II

With unpublished data from Sandhya Bangaru

Materials and Methods

Expression of soluble HA proteins. Sequences encoding the HA genes of interest were optimized for mammalian cell expression, and cDNAs were synthesized (Genscript) as soluble trimeric constructs as described previously (Bangaru et al., 2016). HA protein was expressed by transient transfection of 293F cells with polyethylenimine (PEI) transfection reagent and grown in expression medium (Freestyle 293 Expression Medium; Invitrogen, 12338). Cell supernatants were harvested after 7 days, filtered sterilized with a 0.4 μm filter and recombinant protein purified with HisTrap TALON FF crude columns (GE Healthcare Life Sciences).

Next-generation DNA sequence analysis of expressed antibody variable genes. Total RNA was extracted from 10 million PBMCs. A one-step RT-PCR was performed for 25 cycles using heavy-chain BIOMED-2 variable antibody gene-specific primers (van Dongen et al., 2003) and the OneStep SuperScript III with Platinum® Taq High Fidelity kit (Invitrogen, 11304011). The Illumina-specific adapters were added using the Illumina TruSeq Library Preparation Kit (Illumina, FC-121-3001) according to the manufacturer's recommendations. The final amplicon libraries were sequenced on an Illumina MiSeq instrument using the MiSeq PE-300 v3 reagent kit (Illumina, MS-102-3001). Sequence analysis was performed using IG-BLAST v1.4, and results were parsed to MongoDB for further study.

Identifying clonally related sequences. From a database of annotated antibody sequences obtained from this donor, we queried HCDR3s with VH4-61/JH4 lineage. These HCDR3 sequences were pairwise aligned to the HCDR3 of FluA20 using a PAM30 matrix, with penalties for gap opening and gap extension of -14 and -3, respectively. HCDR3 sequences with a Hamming distance of <= 3 to FluA20 were selected as siblings and the 'full length' nucleotide and amino acid sequences were queried from our database for further analysis.

Visualizing clonally related sequences. A network graph was built from the aligned, full length sequences queried as described previously. Identical sequences were clustered into single nodes, and edges were drawn between two nodes if their Hamming distance was the lowest compared to all other nodes. Nodes denoting the inferred common ancestor and the germline VH4-61/JH4 sequence were manually added. This network was visualized using Cytoscape and manually adjusted for visual clarity (to prevent nodes from overlapping edges to which they are not connected, and to shorten distances between nodes that are closely related).

Fab and IgG cloning, expression and purification for binding assays. FluA-20 and sibling Fab and IgG were expressed in 293F mammalian cells. The heavy and light chains of the Fab were cloned independently into the phCMV3 vector and fused with the N-terminal IgK secretion signal peptide. A His6 tag was added to the C-terminus of the Fab heavy chain. Recombinant DNAs for both heavy and light chains were purified separately and co-transfected into 293F cells. The cells were cultured for 6-7 days at 37C, while shaking at 125 r.p.m. Secreted Fabs were purified Ni- NTA Superflow (Qiagen), monoS chromatography (GE Healthcare).

Determination of half maximal effective concentration (EC50) for binding. To determine EC50 concentrations for binding, we performed ELISA using 384-well plates that were coated overnight at 2 μg/mL with the recombinant HA protein of interest. The plates then were blocked with 50 μL of 5% non-fat dry milk, 2% goat serum and 0.1% Tween-20 in PBS for 1 h at RT. The plates were washed and three-fold dilutions of the mAb starting from 10 μg/mL were added to the wells and incubated for an hour. The plates were washed and 25 μL of 1:4,000 dilution of anti-human IgG alkaline phosphatase conjugate (Meridian Life Science, W99008A) was added. After a final wash, 25 μL of phosphatase substrate solution (1 mg/mL p-nitrophenol phosphate in 1 M Tris aminomethane) was added to the plates, incubated for 20 minutes and the optical density values were measured at 405 nm wavelength on a BioTek plate reader. The plates were washed 3 times between each step with PBS containing 0.1% Tween-20. Each dilution was performed in quadruplicate, and the EC50 values were calculated in Prism software (GraphPad) using non-linear regression analysis. The experiment was conducted twice independently.

Additional Figures

Subtype	Strain	Binding EC_{50} (μg/mL) for mutant for indicated chain							
		FluA-20	Sib 2	Sib 3	Sib 7	Sib 28	Sib 33	Sib 45	Sib 48
H1	A/Solomon Islands/03/2006	0.05	0.07	0.04	**NB**	0.09	**NB**	0.04	0.08
	A/Texas/36/1991	0.05	0.09	0.05	**NB**	0.08	**NB**	0.04	0.06
H2	A/Singapore/1/1957	0.18	0.14	0.07	**NB**	0.29	**NB**	0.08	0.42
H3	A/Hong Kong/1/1968	0.04	0.08	0.06	**NB**	0.07	**NB**	0.04	0.07
	A/Texas/50/2012	0.09	0.09	0.05	**NB**	**>**	**NB**	0.07	**NB**
	A/Switzerland/9715293/2013	0.55	0.32	0.16	**NB**	**NB**	**NB**	0.34	**NB**
H5	A/Indonesia/5/2005	6.05	0.43	1.21	**NB**	**NB**	**NB**	0.46	**NB**
H7	A/Netherlands/219/2003	0.05	0.11	0.07	**NB**	0.09	**NB**	0.05	0.11
	A/Shanghai/2/2013	0.12	0.10	0.10	**NB**	0.49	**NB**	0.12	0.57
H9	A/HongKong/1073/99	0.41	0.22	0.09	**NB**	0.33	**NB**	0.12	0.82
H12	A/duck/Alberta/60/1976	0.05	0.07	0.04	**NB**	0.08	**NB**	0.04	0.12
H14	A/mallard duck/Astrakhan/263/1982	0.95	0.35	0.19	**NB**	1.72	**NB**	0.26	**NB**
H15	A/shearwater/Western Australia/2576/1979	0.08	0.16	0.10	**NB**	0.38	**NB**	0.10	1.20

Figure 1. Binding of FluA-20 sibling antibodies to HAs derived from different subtypes (Sandhya Bangaru, unpublished data). > indicates EC_{50} values 10-fold higher than FluA-20 and NB indicates that no binding was observed at antibody concentrations below 10 μg/mL.

Figure 2. Sequence of FluA-20 and the unmutated common ancestor (UCA) of FluA-20 are aligned (Sandhya Bangaru, unpublished data). Mutated residues are colored as red and a unique disulfide bond in CDR H3 is highlighted in yellow. The key residues Asp98 (H), Tyr100a (H), Tyr48 (L), and Gln55 (L) that were later identified to be critical for the interaction with HA originate from the UCA (in red circles).

		Group 1														Group 2									
Subtype		H1				H2	H5		H6	H8	H9		H12	H13	H16	H3			H4	H7			H10	H14	H15
HA from indicated strain		A/California/2009	A/Texas/36/1991	A/Fort Monmouth/1/1947	A/Solomon Islands/03/2006	A/Singapore/1/1957	A/Vietnam/1203/2004	A/Indonesia/05/2005	A/Taiwan/2/2013	A/turkey/Ontario/6118/1967	A/turkey/Wisconsin/1/1966	A/Hong Kong/1073/99	A/duck/Alberta/60/1976	A/gull/Maryland/704/1977	A/black-headedgull/Sweden/4/1999	A/Hong Kong/1/1968	A/Texas/50/2012	A/Switzerland/9715293/2013	A/duck/Czechoslovakia/1956	A/New York/107/2003	A/Shanghai/2/3013	A/Netherlands/219/2003	A/chicken/Germany/N/1949	A/mallard duck/Astrakhan/263/1982	A/shearwater/WesternAustralia/25 76/1979
EC50 (ng/mL)	FluA-20	8	4	12	5	7	283	85	147	1740	63	9	51	70	>	6	4	19	13	808	66	29	31	13	15
	rFluA-20	31	46	20	47	178	>	6046	229	959	193	407	50	>	>	35	86	548	18	>	117	54	129	71	136
	UCA	140	48	39	36	70	>	>	831	4327	359	342	67	>	>	54	1304	2365	25	>	150	83	361	132	73

Figure 3. Binding EC50 (ng/mL) for FluA-20, recombinant FluA-20 (rFluA-20) and unmutated common ancestor of FluA-20 (FluA-20-UCA) to HAs derived from different strains representing group 1 (green) and group 2 (blue) IAVs (Sandhya Bangaru, unpublished data). The table is displayed in purple-white color scale corresponding to strong-weak binding, respectively. The > symbol indicates no binding observed at concentrations less than 10 μg/mL.

APPENDIX 2

Supplemental Information for Chapter III

Adapted from Finn et al., PLOS ONE, 2016

Materials and Methods

Calculating bulged and non-bulged torso dihedral angles. A collection of antibody heavy chain variable domains was manually curated from the PDB, building upon a published list. The torso residues of these structures were extracted from the PDB files and were clustered using Rosetta Cluster with a cluster radius of 2 Å to separate bulged and non-bulged antibody torsos. φ and ψ dihedral angles of the seven torso residues were found using Biopython, with average and approximate standard deviation calculated using Equations 1 and 2.

Generating HCDR3 loop models. The complete protocol for generating the HCDR3 loop models using Rosetta is described in the following Rosetta Protocol Capture. In brief, structure files for each benchmark antibody were downloaded from the PDB and were cleaned such that only a single variable domain remained. Input files for loop modeling were generated with the assistance of a suite of python scripts, and fragments were selected using the fragment picker. Centroid loop modeling was accomplished using cyclic coordinate descent (CCD), followed by a kinematic closure (KIC) full-atom refinement.

HCDR3 torso sequence analysis. Sequences of the seven torso residues were taken from each of the PDB files of the bulged antibody torso cluster found above and used to generate a WebLogo using the default webserver settings. A second WebLogo was generated using the

sequences of the torso residues taken from the IMGT human V_H and J_H gene segments.

Rosetta Protocol Capture

All Rosetta protocols were performed using version 6d19a9e478a3fc1cf369591953624a66990855ae (2013-11-15 14:37:19).

I *De novo* modeling of bulged HCDR3 loops without restraints

I.1 Prepare a PDB input file. Typically, this is accomplished by removing unnecessary chains, waters and non-protein molecules (e.g. gold) leaving behind one asymmetric unit containing a heavy and light chain. Cleaned PDB files are then renumbered using the renumber_pdb.py script:

```
/path/to/Rosetta/tools/protein_tools/scripts/pdb_renumber.py --norestart
XXXX.pdb XXXX_renum.pdb
```

I.2 Generate a FASTA sequence file from the PDB file.

```
/path/to/Rosetta/tools/protein_tools/scripts/get_fasta_from_pdb.py
XXXX_renum.pdb H > XXXX_.fasta
```

I.3 Generate a loops file (XXXX_.loops) containing one line as follows:

```
LOOP [residue before HCDR3 loop begins] [residue after HCDR3 loop ends] 0 0 0

LOOP 96 108 0 0 0
```

I.4 Generate fragments.

There are two ways to prepare files for the fragment picker: either by using the make_fragments.pl script, or by running the Robetta webserver (http://www.robetta.org/). For ease of use, the instructions below describe preparing these files using Robetta.

Submit a job to the Robetta Fragment Server by clicking the “Submit” link under Fragment Libraries from the main Robetta page (http://www.robetta.org/). Enter your registered username, the target name, and the FASTA file in the provided fields. For benchmarking purposes, select “Exclude Homologues”. Click “Submit” to place the job in queue. Jobs typically complete in less than 1 hour once they become active.

Once complete, download the following files from the webserver for use by the fragment picker:

```
XXXX_.checkpoint, XXXX_.psipred_ss2, XXXX_jufo_ss, XXXX_.homolog_vall
```

Finally, update the fragment_picker_quota.options file to point to the correct input files, then run:

```
/path/to/Rosetta/main/source/bin/fragment_picker.default.linuxgccrelease
@fragment_picker_quota.options
```

I.5 Prepare the model_wo_rest.options file to point to the correct input files, then run Rosetta LoopModel:

```
/path/to/Rosetta/main/source/bin/loopmodel.default.linuxgccrelease
@model_wo_rest.options -out:prefix ex1- >& OUT1.log &
```

II *De novo* modeling of bulged HCDR3 loops with restraints

II.1 As before, prepare a PDB input file, a FASTA sequence file, a loops file and generate fragments (Protocol I steps 1-4).

II.2 Prepare a restraint file. A script has been provided to make restraint file formatting easy, however these files can be manually created from the values available in Table 1.

```
/HCDR3_prot_capture/scripts/maketorsoconstraints.py –b –s 97 –e 107 >
XXXX_bulged.constraints
```

II.3 Prepare the model_w_rest.options file to point to the correct input files. Note the added flags for restraint file handling.

II.4 Finally, run Rosetta LoopModel:

```
/path/to/Rosetta/main/source/bin/loopmodel.default.linuxgccrelease
@model_w_rest.options -out:prefix ex2- >& OUT2.log &
```

III Comparing models generated with and without restraints

III.1 For comparison purposes, re-score the models generated in Protocol 1 using the restraint penalties. Prepare the scoring_wrest.options file to point to the correct input files, then score the models with restraints:

```
/path/to/Rosetta/main/source/bin/score_jd2.default.linuxgccrelease
@scoring_wrest.options -s ex1-*.pdb -out:prefix wrest-
```

Additional Figures

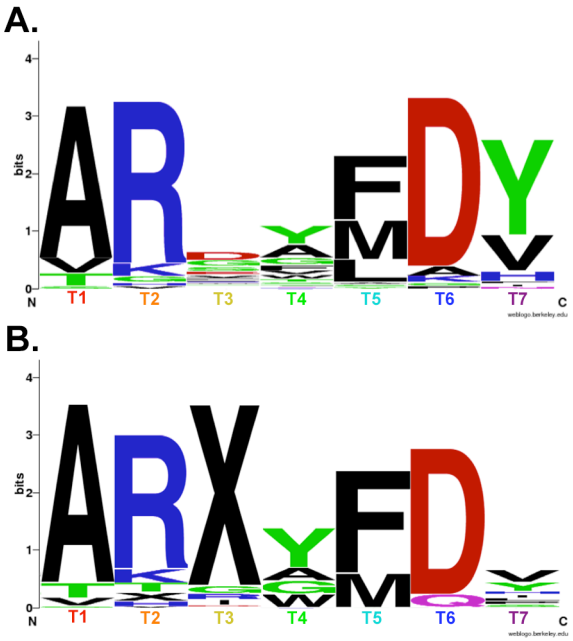


Figure 1. Bulged torso structures share similar sequences, which are germline-encoded. Previous studies identified a sequence motif in bulged torso structures, which are formed primarily via a side-chain interaction between either Arg or Lys (R/K) at T2 and Asp (D) at T6. A consensus sequence from bulged torsos culled from the PDB shows the prevalence of these residues at these positions (panel A). These residues are germline-encoded, as observed in a consensus sequence of the V_H and J_H gene segments that contribute to the torso domain (panel B).

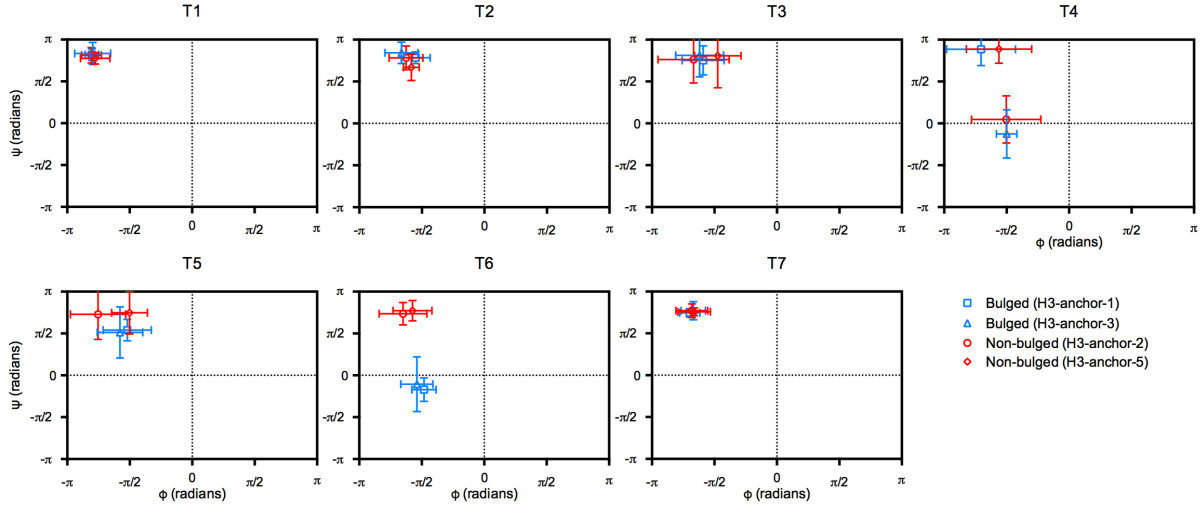
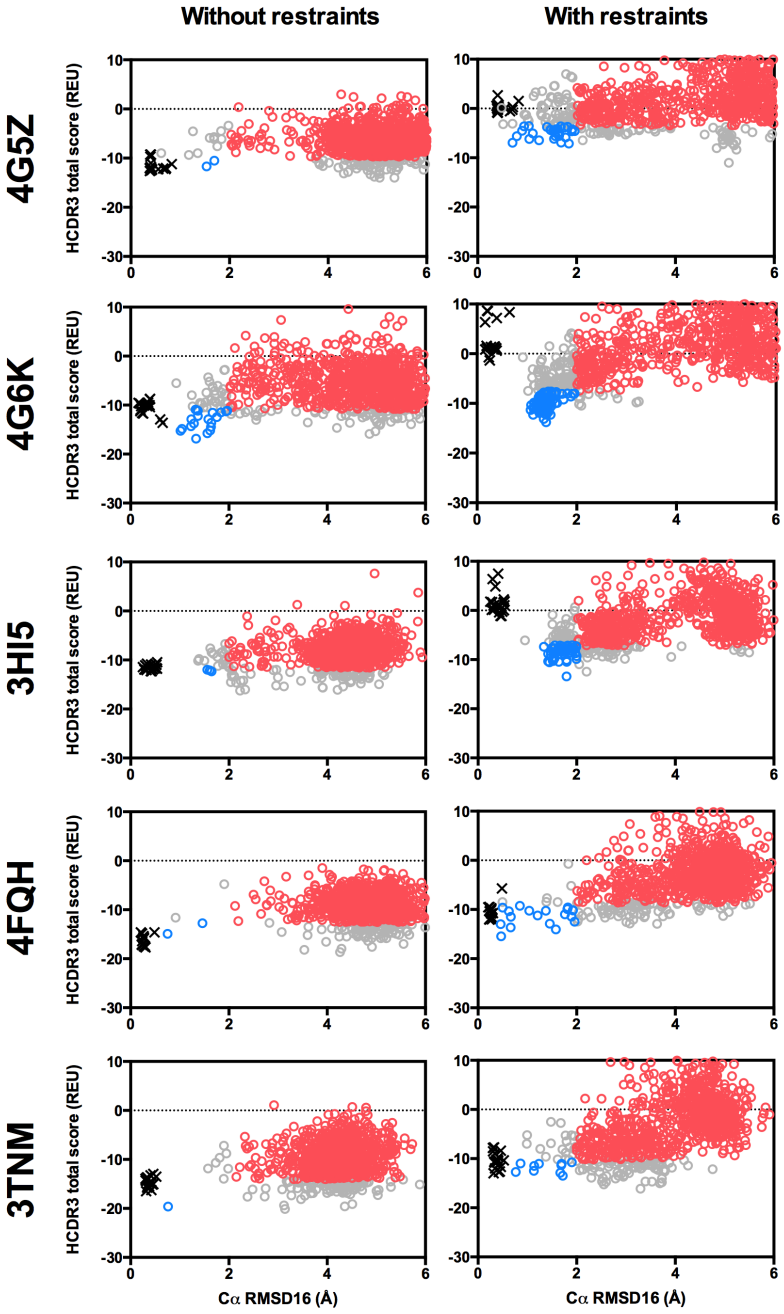


Figure 2. Average φ and ψ angles observed for each torso residue in known bulged and non-bulged clusters. North et al. defined seven canonical torso conformations from experimentally-determined antibody structures. Two of these clusters are considered bulged (H3-anchor-1 and H3-anchor-3; blue) and two are considered non-bulged (H3-anchor-2 and H3-anchor-5; red). φ and ψ angles are well defined for both bulged and non-bulged HCDR3 torso residues. Bulged and non-bulged torsos are differentiated by their ψ angle at T6. The ψ angle at T4 is bimodal for both bulged and non-bulged HCDR3 torsos, with ~180 degrees separating the two clusters within each definition.


Without restraints
With restraints
4G5Z
4G6K
3HI5
4FQH
3TNM
HCDR3 total score (REU)
Cα RMSD16 (Å)
10
0
-10
-20
-30
0
2
4
6

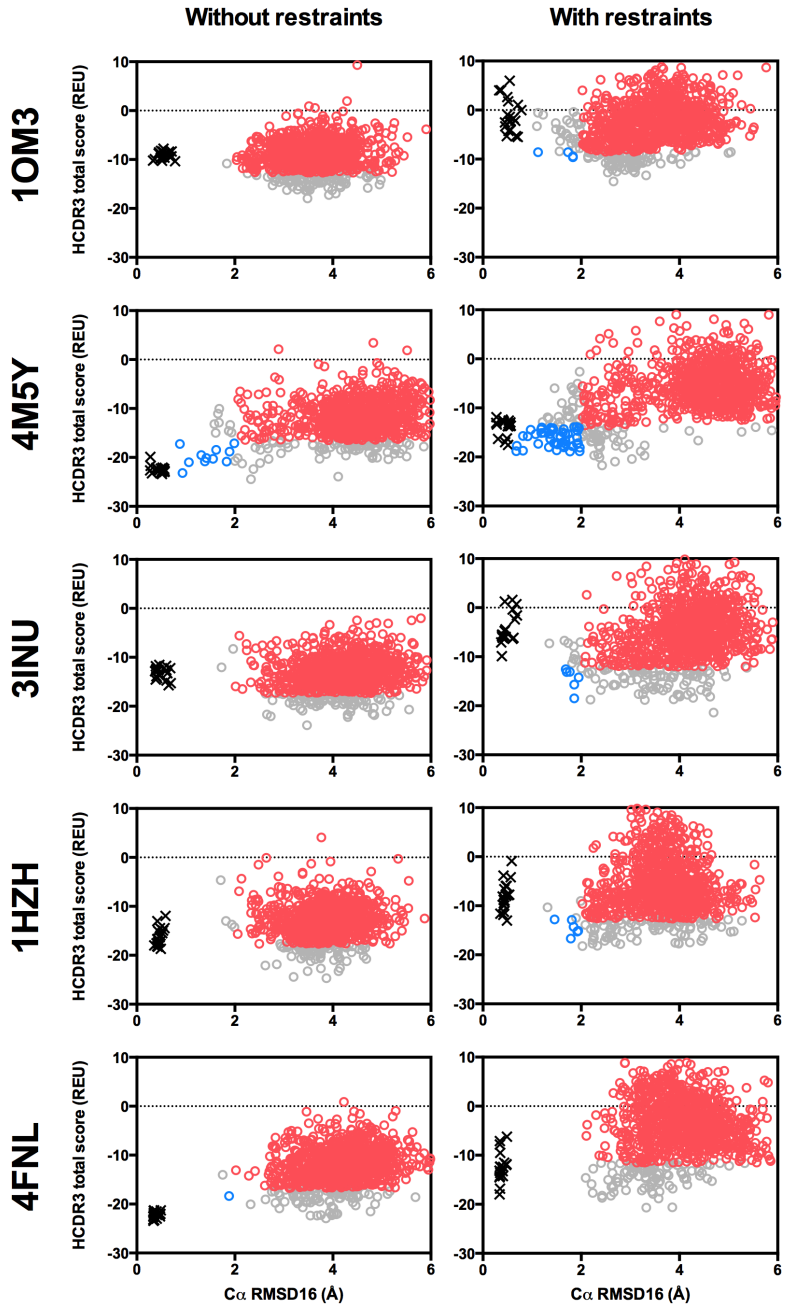


Figure 3. Bulged torso restraints improve native-like HCDR3 sampling and recovery. 1,000 models of each benchmark antibody were generated and scored with or without bulged restraints using Rosetta LoopModel (comparable to Figs 2A and 2D). Models with scores ranked in the top 10% and RMSD16 ≤ 2 Å have been colored blue, while models with scores ranked below the top 10% and RMSD16 > 2 Å have been colored red. The native crystal structure was also minimized using Rosetta FastRelax, generating 20 structures (black x's). The total HCDR3 score vs. the HCDR3 Cα RMSD16 to the native crystal structure is shown.

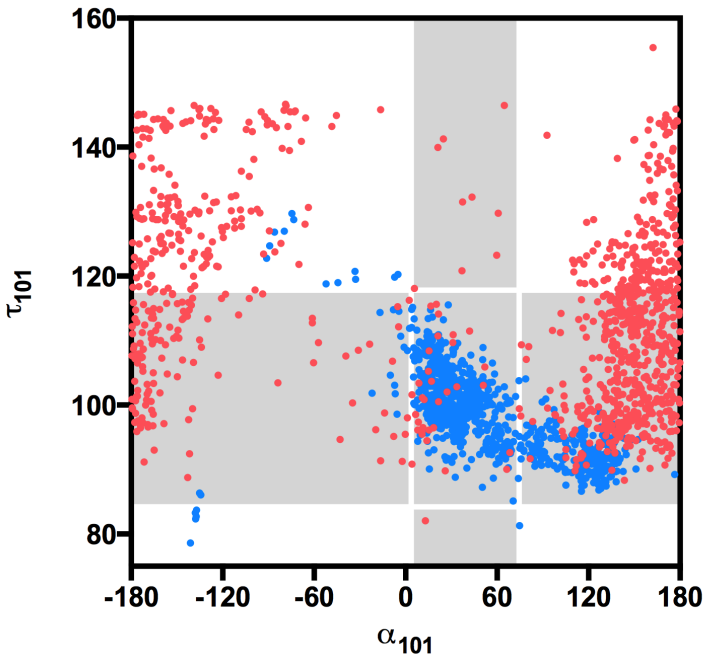


Figure 4. Bulged torso restraints improve sampling of HCDR3 torso angles. Using Rosetta LoopModel, 1,000 models of the benchmark antibody 4G5Z were generated without (red) or with (blue) bulged restraints. The τ_{101} angle and α_{101} dihedral angle defined by Weitzner et al. were calculated for each model. Gray regions of the plot denote ± 3σ of the mean angles calculated for bulged HCDR3 torsos by Weitzner et al. Improved recovery of bulged torsos was observed as a greater density of points in the center gray region when restraints were applied (n=719), versus when no restraints were applied (n=33).

APPENDIX 3

Supplemental Information for Chapter IV

Materials and Methods

Next-generation DNA sequence analysis of expressed antibody variable genes. As described previously (Appendix 1), total RNA was extracted from cryopreserved PBMCs, and a one-step RT-PCR was performed for using heavy-chain BIOMED-2 variable antibody gene-specific primers (van Dongen et al., 2003) and high-fidelity Taq polymerases. The Illumina-specific adapters were added using the Illumina TruSeq Library Preparation Kit (Illumina, FC-121-3001) according to the manufacturer's recommendations. The final amplicon libraries were sequenced on an Illumina MiSeq instrument using the MiSeq PE-300 v3 reagent kit (Illumina, MS-102-3001). Sequence analysis was performed using PyIG (unpublished), and results were parsed to MongoDB for further study.

Yeast transformation. EBY-100 yeast cells were transformed following the protocol by Gietz and Schiestl. Briefly, yeast cells were washed and resuspended in sterile water to 10^8 cells/mL. 10^7 cells were transferred to each well of a 96-well plate. The plate was centrifuged at 1500 *g* for 5 minutes to pellet the cells and the supernatant was removed by inverting the plate. Transformation mix was prepared following the published protocol. 50 µL transformation mix added to each well, and cells were resuspended by pipetting. 100 µL PEG3350 (50% w/v) was added to each well and contented mixed by pipteting. Cells were heat shocked by incubating at 42°C for one hour. The plate was centrifuged for ten minutes at 1500 *g* and the supernatant again removed by inverting the plate. 50 µL of sterile water was added to each well and the cells were

resuspended by pipetting. The transformed yeast were plated on SD-CIT.CAA selective plates and incubated at 30°C, for 72 hours.

Yeast growth and scFv expression. Transformed colonies were picked to 4 mL SD-CIT.CAA medium and grown at 30°C, shaking at 225 rpm for 24 hours. Next, yeast were pelleted by centrifugation, washed with SG-CIT.CAA expression medium, and transferred to SG-CIT.CAA. Yeast were incubated at room temperature shaking at 225 rpm for 24 to 48 hours to allow for scFv expression.

Recombinant antibody expression and purification. The heavy and light chain variable regions were cloned into Mclean Fab and lambda vector, respectively. The Fab fragment was expressed by transient co-transfection of the expression vector containing heavy chain and light chain into Expi-CHO cells. Recombinant Fab was purified from culture supernatant using a anti-CH1 CaptureSelect column. Purified Fab was measured by optical absorbance at 280 nm, and purity and integrity were analyzed by reducing and nonreducing SDS-PAGE. The purified Fab was concentrated to ~10 mg/mL for crystallization and K_D determination.

Production of recombinant soluble HA proteins. The design and expression of recombinant HA proteins for binding studies were described previously. Sequences encoding the HA genes were synthesized as soluble trimeric constructs by replacing the transmembrane and cytoplasmic domain sequences with cDNAs encoding the GCN4 trimerization domain and a His-tag at the C-terminus. Synthesized genes were subcloned into the pcDNA3.1(+) mammalian expression vector (Thermo Fisher Scientific) and expressed in FreeStyle 293-F cells (Thermo Fisher

Scientific).

Flow cytometric binding analysis of yeast surface display scFv. After inducing surface-display scFv expression in yeast cells, 5 x 10^5 cells were added to each well of a 96-well V-bottom plate. The cells were pelleted by centrifugation at 1500 *g* for 5 minutes, and washed once in PBS containing 0.05% BSA (wash buffer). The cells were resuspended in 50 µL 10 nM biotinylated SI06 HA prepared in wash buffer. The cells were allowed to incubate in antigen for one hour at room temperature, after which they were washed three times. Next, the cells were resuspended in 50 µL stain solution (1:250 FITC-conjugated V5 peptide and 0.1 µg/well APC-conjugated streptavidin prepared in wash buffer). The cells were again allowed to incubate for one hour at room temperature, after which they were pelleted and washed one time. Finally, the cells were resuspended in 250 µl wash buffer and kept on ice until analyzed by flow cytometry.

Half maximal effective concentration (EC_{50}) analysis. ELISAs were performed to obtain EC_{50} values for binding using 384-well plates coated with the HA of interest at a 2 µg/mL concentration and incubated overnight at 4°C. The plates were blocked with 5% non-fat dry milk, 2% goat serum, and 0.1% Tween-20 in PBS for one hour. Three-fold dilutions of the mAb, starting from 50 µg/mL, were added to the wells, incubated for one hour, followed by a one hour incubation of 1:4,000 dilution of goat F(ab')2 anti-human lambda light chain horseradish peroxidase conjugate (SouthernBiotech, 2072-05). The plates were washed three times between each step with PBS containing 0.1% Tween-20. 1-Step TMB Ultra-ELISA Substrate solution (Thermo Fisher) was added to the plates, incubated for ten minutes, and the optical density values were measured at 450 nm wavelength on a BioTek plate reader. Each dilution was

performed in duplicate, and the EC_{50} values were calculated in Prism software (GraphPad) using non-linear regression analysis.

Hemagglutination inhibition (HAI) assay. Neutralization potential function of Fabs was determined by HAI assay as previously described (Bangaru et al., 2016).

Half maximal inhibitory concentration (IC_{50}) analysis. 40 $TCID_{50}$ of A/Solomon Islands/3/2006 H1N1 (FR-331, IRR; Batch HA128 Immunology Core) virus was mixed with serial two-fold dilutions of Fabs starting from 40 µL/mL and incubated for 1 hour at RT. This virus-antibody mixture was added to monolayers of MDCK cells and incubated for 24 hours at 37°C with 5% CO_2. Human IgG CH65 was used as a positive control, and an unrelated antibody, EEEV-16 Fab, was used as negative control. All experiments were performed at triplicate. After 24 hours cells were fixed and virus inactivated by 80% methanol in PBS. Cells were incubated with blocking buffer PBS with 0.05% Tween-20 for 1 hour at RT, primary anti-Influenza Nucleoprotein mouse antibodies (1:6000, BEI, NR-4282) for 1 hour at RT, and secondary anti-mouse IgG AP conjugated antibodies (1:3000, Novex, DKXMO AP AFFINITY, A16014) for 1 hour at RT. Wells with virus infected cell monolayers were visualized by AP substrate. Low OD_{405} values correspond to samples without influenza virus. High OD_{405} values correspond to samples with infected by Influenza virus MDCK cells. EC_{50} values were calculated using a non-linear regression analysis by Prism v. 5.0 (GraphPad).

K_D determination. K_D values were determined by bio-layer interferometry using an Octet RED instrument (ForteBio, Inc.), as described previously. Biotinylated SI06 HA protein (10 µg/ml)

was loaded onto streptavidin-coated biosensors in kinetics buffer (1× PBS, pH 7.4, 1% bovine serum albumin [BSA], and 0.05% Tween 20) for 300 sec. For measurement of k_{on}, association of Ab was measured for 120 sec by exposing the sensors to seven concentrations of Fab (2-fold dilutions starting at 200 nM) in kinetics buffer. For measurement of k_{off}, dissociation of Ab was measured for 120 sec in kinetics buffer. Experiments were performed at 30°C. K_D was calculated as the ratio of k_{off} to k_{on} determined from binding curves of at least four concentrations for each Fab.

Crystallization and x-ray structure determination. All recombinant Fabs were concentrated to ~ 10 mg/ml in 20 mM Tris-HCl, 50 mM NaCl for crystallization trials, and Fab crystals were grown using sitting-drop vapor diffusion method at 18ºC. Crystals of Fab CH65:1203d4 were obtained in 1.8 – 2.2 M $(NH_4)_2SO_4$, 100 mM Tris-HCl pH8.3, crystals of Fab CH65:7969d2 in1.8 – 2.2 M $(NH_4)_2SO_4$, 100 mM Bis-tris pH 6.5, and crystals of Fab CH67:1203d4 in 22% - 32% PEG 1000, 100 mM HEPES pH 7.0 – 8.0. Crystals were cryo-protected in mother liquor supplemented with 20% (w/v) glycerol (CH65:1203d4 and CH65:7969d2) or 40% PEG 1000 (CH67:1203d4), flash frozen, and stored in liquid nitrogen until data collection. X-ray diffraction data for the CH65:7969d2, CH65:1203d4, CH67:1203d4 apo Fabs were collected to 2.00 Å, 2.20 Å and 3.80 Å resolutions at the Cornell High Energy Synchrotron Source (CHESS) F1 beamline. The diffraction data sets were processed and scaled with XDS and scala. The crystal structure of Fabs was determined by molecular replacement with Phaser using the variable and constant domains of Fabs in the PDB (4WUK, 4HKB) as search models for CH65:1203d4/CH65:7969d2 and CH67:1203d4 respectively. The model was iteratively rebuilt using Coot and refined in Phenix.

Accession codes: Atomic coordinates and structure factors for the crystal structures of apo Fabs CH65:7969d2, CH65:1203d4, CH67:1203d4 have been deposited in the Protein Data Bank with the accession codes 6DLA, 6DLB, and 6DL8, respectively.