Abstract

Computational Identification of B Cell Clones in High-Throughput Immunoglobulin Sequencing Data

Namita Gupta

2017

Humoral immunity is driven by the expansion, somatic hypermutation, and selection of B cell clones. Each clone is the progeny of a single B cell responding to antigen, with diversified Ig receptors. The advent of next-generation sequencing technologies enables deep profiling of the Ig repertoire. This large-scale characterization provides a window into the micro-evolutionary dynamics of the adaptive immune response and has a variety of applications in basic science and clinical studies. Clonal relationships are not directly measured, but must be computationally inferred from these sequencing data. In this dissertation, we use a combination of human experimental and simulated data to characterize the performance of hierarchical clustering-based methods for partitioning sequences into clones. Our results suggest that hierarchical clustering using single linkage with nucleotide Hamming distance identifies clones with high confidence and provides a fully automated method for clonal grouping. The performance estimates we develop provide important context to interpret clonal analysis of repertoire sequencing data and allow for rigorous testing of other clonal grouping algorithms. We present the clonal grouping tool as well as other tools for advanced analyses of large-scale Ig repertoire sequencing data through a suite of utilities, Change-O. All Change-O tools utilize a common data format, which enables the seamless integration of multiple analyses into a single workflow. We then apply the Change-O suite—in concert with the nucleotide coding sequences for WNV-specific antibodies derived from single cells—to identify expanded WNV-specific clones in the repertoires of recently infected subjects through quantitative Ig repertoire sequencing analysis. The method proposed in this dissertation to computationally identify B cell clones in Ig repertoire sequencing data with high confidence is made available through the Change-O suite and can be applied to provide insight into the dynamics of the adaptive immune response.

Computational Identification of B Cell Clones in High-Throughput Immunoglobulin Sequencing Data

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Namita Gupta

Dissertation Director: Steven H Kleinstein

May 2017

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Antibodies and adaptive immunity

The immune system uses a variety of mechanisms to respond to the invasion of a foreign pathogen into the human body. The innate arm of the immune system is responsible for immediate sensing of infection and the body's initial response. This initial response is essentially the same for any type of infection and will result in the hallmark symptoms of infection: redness, swelling, soreness, and heat. These symptoms are the result of innate immune cells being recruited to the site of infection and releasing enzymes in an attempt to break down the pathogen. This brute force approach to fighting infection works on a short time-scale to curb spreading of the infection, but takes a toll on the body that is not sustainable.

The adaptive arm of the immune system works to complement the innate arm by fine-tuning the response, which requires working on a longer time-scale. While the innate cells work in the foreground at the site of infection, adaptive cells such as B and T cells work in the background to *adapt* to the specific pathogen causing the infection. This adaptation results in cells that are highly specialized to neutralize the pathogen more effectively and efficiently than the brute force innate cells. While the peak of the innate response is typically around two days after infection, the peak of the adaptive response is not until around seven days, resulting in clearing of the pathogen often by two weeks to a month after the initial infection.

The research presented in this dissertation focuses on B cells and the antibodies they produce. Antibodies have a Y-shaped structure consisting of two identical heavy chains and two identical

light chains. Each chain has a variable region, which comes in contact with antigen, and a constant region, which determines the protein's effector function. The constant region (as the name implies) is largely consistent across antibodies and provides much of the protein's structural stability in addition to determining its effector function. The aptly named variable region confers specificity of the antibody for its cognate antigen and varies widely across antibodies. The variable region consists of four relatively conserved framework regions (FWRs) that confer structural stability and three loops referred to as the complementarity determining regions (CDRs) at the tips of the protein that bind directly to antigen. The CDRs determine the specificity of the antibody, particularly CDR3, the most variable of the three regions.

1.2 What is the Ig repertoire?

Antibody formation begins with the development of B cells in the bone marrow of adults. This development requires forming a function B cell receptor (BCR) or immunoglobulin (Ig) that when secreted, is referred to as an antibody. In order to create a diverse antibody repertoire, we have evolved a process of somatically recombining gene segments in a locus to form a full gene for transcription. In the case of the heavy chain (which is most variable), this involves recombination of one of approximately 45 variable (V). 25 diverse (D), and 6 joining (J) gene segments in the Ig gene locus. These numbers, in addition to the segments of the light chain, yield over three million possible Ig molecules (Munshaw & Kepler, 2010). The combinatorial diversity is compounded by the fact that random nucleotides can be deleted or added at the junction of these segments, increasing potential diversity to over 10^7 proteins (Volpe & Kepler, 2008). The region of the Ig sequence that encompasses where the V, D, and J segments meet is by far the most variable portion of the sequence and is referred to interchangeably as the junction or Complementarity Determining Region 3 (CDR3) of the protein. B cells in circulation that have undergone V(D)J recombination but have not yet encountered antigen are called naive B cells. Of the approximately two billion B cells in the human body, the diversity has a lower bound of about two million (Boyd, Marshall, *et al.*, 2009) and a theoretical upper limit of $\sim 10^{12}$ different receptors (Munshaw & Kepler, 2010).

Once a naive B cell encounters its cognate antigen and a second signal from another immune cell, it becomes activated and moves towards a variety of fates. One fate is to begin a micro-evolutionary process to adapt to the infecting pathogen in the germinal center, where the cell will undergo many rounds of cell division (clonal expansion) to mount a response. During this expansion, the B cell

diversifies even further by replicating with a high mutation rate (approximately one point mutation per 1000 bp per cell division (McKean *et al.*, 1984; Kleinstein *et al.*, 2003)), referred to as somatic hypermutation (SHM), in the Ig gene, leading to a potential for $\sim 10^{12}$ molecules (Munshaw & Kepler, 2010). The cells with mutations that increase affinity for antigens presented during the current infection continue this maturation process and eventually differentiate to become long-lived memory or plasma B cells. This affinity maturation process is a form of accelerated evolution of the initial B cell and results in a collection of descendants referred to collectively as a B cell clone. A single B cell clone can expand during an immune response, but has been observed to be at most 0.3% of total B cells at a given time (Boyd, Marshall, *et al.*, 2009).

This antibody response is part of what determines our ability to fight off infections and is also what is utilized by most vaccines. Vaccines contain antigens intended to activate a B cell so it begins affinity maturation. Once the B cell matures, it differentiates to become an antibody factory known as a plasma cell. Plasma cells excrete the B cell receptor in large volumes as antibody. These soluble antibodies bind the foreign antigens with their variable region and the constant region determines the subsequent action the immune system will take. These actions range from phagocytosis and presentation of the antigen to other immune cells to recruiting the complement system to puncture the bacterial membrane.

The collection of circulating antibodies and B cell receptors in the body is referred to as the Ig repertoire. This repertoire is a snapshot of antigen-inexperienced naive B cells and antigen-experienced B cell clones that have undergone affinity maturation. The ability to profile the Ig repertoire using next-generation sequencing technologies (as outlined in Section 1.3) has many basic science and clinical applications, some of which are discussed in Section 1.4.

1.3 Ig repertoire sequencing technologies

Though the study of the antibodies and the response they mediate is well established, the way in which such experiments are being conducted is changing. Earlier studies utilized traditional Sanger sequencing of mRNA from B cells either from tissue micro-dissections or blood samples to look at Ig molecules, yielding at most a few hundred sequences per experiment. Scientists would use low-throughput methods to identify germline V(D)J segments, amino acids that confer specificity, and mutations from the germline. A transition to next-generation sequencing is currently ongoing in the field of Ig repertoire study. The ability to sample many more Ig sequences than previous

technologies enable new types of analyses and creates the need for automated tools to handle analysis of these large datasets.

Current high-throughput sequencing technologies can sequence billions of bases per run at costs as low as a few cents per million bases (Liu *et al.*, 2012), enabling a more complete quantification of an individual's antibody-mediated response. When sequencing Ig molecules, it is necessary for each Ig sequence to fit on a single read. The similarities between different gene segments as well as the un-templated nature of the junction region make it impossible to map shorter reads definitively to germline V, D, and J segments for each molecule. Furthermore, if the B cell has encountered antigen and undergone affinity maturation, its receptor will have accumulated mutations, which can only be identified by ensuring that the entire molecule fits on a single read.

Ig sequencing can be done using the rearranged DNA or mRNA molecules as a template (Boyd & Joshi, 2014). The Ig region is amplified and selected using internal V primers with J primers or constant region primers with 5' RACE (Benichou *et al.*, 2012). In the case of 5' RACE, the Ig leader sequence (signaling that the protein destination is the cell membrane) is also part of the amplicon. An example amplicon is shown in Figure 1.1. This amplicon would be generated using constant region primers for targeted 5' RACE. At the 5' end of the transcript, a homopolymeric sequence (TS in the Figure 1.1) is added by the terminal transferase, promoting the binding of a template switch oligonucleotide (Y. Y. Zhu *et al.*, 2001). This oligonucleotide consists of a unique molecular identifier (discussed further below) and an upstream primer for the subsequent PCR amplification steps. The amplified cDNA are then sequenced by a next-generation sequencing platform.

One of the first next-generation technologies that was used for immunoglobulin sequencing was Roche/454 pyrosequencing. Though it was slightly lower throughput than some of its competitors, it featured longer reads (Metzker. 2010). More recently, the Illumina MiSeq platform has become the preferred technology for immunoglobulin sequencing. MiSeq allows for paired-end reads, each of 150bp–300bp in length, which significantly lowers the error rate compared to pyrosequencing (Quail *et al.*, 2012). In particular, the rate of insertions and deletions is greatly decreased, as Roche/454 has 0.38 indels per 100 bases whereas MiSeq has less than 0.001 indels per 100 bases (Loman *et al.*, 2012). The lower error rate gives scientists greater power and certainty in attributing single nucleotide changes from germline gene segments to SHM rather than sequencing error. MiSeq is also much higher throughput, allowing for greater sampling of the Ig present in a biological sample.

The protocol for using this technology in Ig repertoire studies was further advanced by the

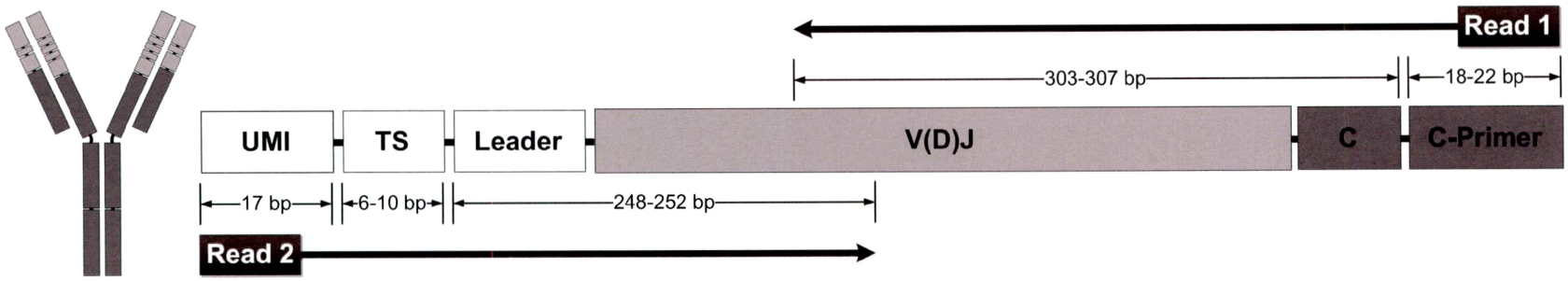

Figure 1.1: **Example Ig read configuration.** 5' RACE uses a constant region primer to capture Ig mRNA including its leader sequence upstream of the variable region. The template switch (TS) allows addition of a unique molecular identifier (UMI) and an upstream primer sequence (not shown) for further amplification. This example amplicon would be sequenced by the 2×300 paired-end Illumina MiSeq platform. Variable (light grey) and constant (dark grey) regions are shown for both the heavy and light chains in the antibody protein schematic on the left.

development of unique molecular identifiers (UMIs). These are oligo-nucleotide sequences that serve as a barcode for each mRNA molecule prior to PCR amplification (Shiroguchi *et al.*, 2012). This is a huge advancement for multiple reasons. First, we can compare reads that share the same UMI to further eliminate the effect of sequencing error using the knowledge that these reads should be identical. Second, if we see identical Ig sequences with differing UMIs, we know that these are independent copies of mRNA molecules, allowing us a new level of quantification independent of PCR amplification bias. The application of these advancements in technology to adaptive immune receptor sequencing (AIRR-Seq) provides a detailed snapshot of the adaptive immune system at a given time-point.

1.4 Applications of Ig repertoire sequencing

A large body of research has been conducted on understanding the antibody-mediated immune response using Ig AIRR-Seq data. Several applications rely on grouping sequences into B cell clones. B cell clones in sequencing data can be estimated by sequence similarity and shared CDR3 length. Clonal grouping enables quantification of inter-clonal diversity and provides insight into the evolutionary dynamics underlying affinity maturation.

One application of this type of analysis is to understand how the Ig repertoire changes with age. Older individuals have a weaker and delayed response to influenza (Y.-C. B. Wu *et al.*, 2012; Jiang, He, *et al.*, 2013) and pneumococcus (Ademokun *et al.*, 2011) vaccinations relative to younger individuals. The overall diversity of the Ig repertoire in peripheral blood and lymph nodes is lower in older individuals (Ademokun *et al.*, 2011; Jiang, He, *et al.*, 2013; Tabibian-Keissar *et al.*, 2015), characterized by having fewer clonal lineages. Furthermore, heavy chain CDR3s in older individuals tend to be longer independently of clonal expansion (*i.e.*, each B cell clone's CDR3 length is only counted once so a large clonal expansion with a long CDR3 cannot skew the distribution) (Y.-C. B. Wu *et al.*, 2012; C. Wang *et al.*, 2014). These changes in the repertoire could explain why older individuals are less able to mount successful immune responses and help to design more effective vaccines for the older population.

Influenza vaccination provides a controlled system with a well-defined timeline that enables time-course analysis of the immune response. Sampling the Ig repertoire at multiple time-points pre- and post-vaccination provides insight into the clonal dynamics underlying the adaptive immune response that can be correlated with measured antibody response to the vaccine. B cell clonal

expansions are observed at seven days post-vaccination (Ademokun *et al.*, 2011; Jackson *et al.*, 2014; Laserson *et al.*, 2014). These expansions are influenza-specific, allowing large-scale characterization of receptors that bind to influenza (Moody *et al.*, 2011; Jackson *et al.*, 2014). Expansion of B cell clones correlates with antibody response (Jackson *et al.*, 2014) and is typically seen around seven days post-vaccination (Laserson *et al.*, 2014). Clonal expansion leads to higher diversity within the clone, but different clones exhibit some convergence towards influenza specificity (Krause *et al.*, 2011). The ability to measure B cell clones during the vaccination repose has provided significant insight into the clonal dynamics of the immune response.

In addition to characterization of repertoire-level diversity, these data enable the study of the mutations within B cell clones. The large number of mutations that results from extensive clonal expansion in response to influenza vaccination increases the likelihood of forming broadly neutralizing antibodies (bnAbs), which neutralize many strains of virus (Moody *et al.*, 2011). Immunizing with an older, pandemic strain of influenza that the subjects had likely not encountered previously is more likely to lead to potential bnAbs than a seasonal vaccination (Wrammert *et al.*, 2011; Cortina-Ceballos, Godoy-Lozano, TÃ΄llez-Sosa, *et al.*, 2015). Such vaccinations also lead to similar antibody sequences observed across different individuals indicating a possible convergent evolution of bnAbs (Krause *et al.*, 2011; Vollmers *et al.*, 2013; Jackson *et al.*, 2014). Clonal expansions resulting in bnAbs have also been observed in subjects infected with HIV (J. Zhu *et al.*, 2013). Analysis of mutations within B cell clonal expansions can inform development of new and improved vaccines to elicit broad serological protection.

In addition to vaccination, profiling the B cell clones of the Ig repertoire provides insight into chronic lymphocytic leukemia (CLL). Subjects with CLL have longer CDR3s in antibodies using the VH1-69 gene segment than healthy controls (Johnson *et al.*, 1997), which may be a predictive marker for people who are at risk of this type of leukemia. High-throughput sequencing of subjects with CLL uncovered that the disease is not as monoclonal as previously thought, but the dominance of B cell clones can be polyclonal (Niklas *et al.*, 2014). Furthermore, this dominance of a few clones can be tracked in circulation by sequencing (Bashford-Rogers *et al.*, 2013). Repertoire sequencing can detect minimal residual disease with as few as 0.5 cells/μl (Boyd, Marshall, *et al.*, 2009). Infection with Epstein-Barr virus is correlated with persistent CLL clonal expansions (C. Wang *et al.*, 2014). The level of mutation of the V region of antibodies in the clonal expansions can be used to determine prognosis (Xochelli *et al.*, 2014). Development and application of Ig AIRR-Seq

has improved detection, diagnosis, and prognosis of CLL.

Many of the results stemming from Ig AIRR-Seq analyses involve characterizing either the clonal diversity of the Ig repertoire, the mutational evolutionary process of affinity maturation that occurs within a B cell clone, or both. Since clonal relationships are not directly measured by sequencing, they must be computationally inferred from the data. Thus, clonal grouping of Ig repertoire sequencing represents a fundamental step in the analysis pipeline.

1.5 Ig AIRR-Seq analysis

The first step in analyzing an Ig repertoire sequencing dataset is pre-processing of the raw reads. This includes all analysis steps leading from the raw reads to full-length Ig sequences including quality control, primer masking, annotation of reads with sequence embedded barcodes, generation of unique molecular identifier (UMI) consensus sequences, assembly of paired-end reads, etc. These steps are particularly difficult in the case of adaptive immune receptors because the receptors are not encoded directly in the genome. Standard RNA and DNA sequencing pre-processing tools rely on having a reference against which to align raw reads, so specialized software tools have been created to pre-process AIRR-Seq data (Yaari & Kleinstein, 2015).

Once full length Ig sequences have been assembled, the next step necessary for any biological insight from the data is to infer the V(D)J germline gene segments used to form each sequence. This inference enables analysis of gene segment usage, clonal grouping, somatic mutations, selection pressures, etc. Tools for this inference include IMGT/HighV-QUEST (Alamyar *et al.*, 2012), IgBLAST (Ye *et al.*, 2013), and SODA2 (Munshaw & Kepler, 2010), but each has limitations and inferring the D gene in particular remains a challenge (Munshaw & Kepler, 2010). These tools evaluate each Ig sequence individually to make a germline inference. As a way to improve and refine the germline calls, tools like TIgGER (Gadala-Maria *et al.*, 2015) leverage the information in the entire dataset to identify novel germline alleles, infer the gene segment genotype, and adjust the germline inference correspondingly.

The analysis stages following raw read processing and germline V(D)J inference are characterizing the population structure of the B cell response and detailed Ig repertoire analysis (Yaari & Kleinstein, 2015). Characterization of the B cell population structure requires identifying B cell clones and reconstructing clonal lineages. Repertoire analysis includes calculating repertoire diversity and identifying mutations resulting from SHM. An example workflow for Ig repertoire se-

quencing analysis is shown in Figure 1.2. Partitioning Ig sequences into B cell clones is fundamental to accomplishing these downstream analyses. There are various tools that can be used to complete any one of these analysis steps (Yaari & Kleinstein, 2015). However, individual tools often vary by input and output format and downloading numerous tools across different software platforms can make the entire Ig repertoire analysis pipeline quite convoluted and difficult to implement. In addition to creating software to calculate the various characteristics of the repertoire, there is a need for an cohesive suite of tools to streamline the analysis pipeline.

1.6 Motivation for this dissertation

B cells and the Ig they produce form fundamental components of the adaptive immune response. The Ig repertoire can now be sampled at a large scale using next-generation sequencing technologies. Once the Ig sequences are assembled, there are various features that can be ascertained in order to draw biological meaning, such as the germline gene segments to determine where mutations have occurred, which sequences are clonally related (*i.e.*, originate from the same V(D)J recombination event) to determine the size of a clone, or the corresponding distribution of mutations that shape affinity. These features allow us to characterize selection pressures that shape Ig maturation and to track clonal expansion and affinity maturation in a specific response to infection or vaccination.

One measurement that can be particularly important is the overall diversity of an Ig repertoire. Diversity can be indicative of health status and is often measured after bone marrow transplants to ensure the repertoire has been fully recovered. This property and others of interest depend on correctly determining which sequences are the result of B cell clonal expansions during an immune response and partitioning the repertoire accordingly. Identifying clonal groups is difficult because the sequences will not be identical, and it is a challenge to create a distance metric that reflects the underlying biology of accumulating mutations during affinity maturation. Manual inspection of sequence similarity provides little insight and is not feasible with such large datasets. Developing a computational algorithm to identify B cell clones with high confidence and creating a framework to integrate all of the Ig repertoire analysis tools are essential steps to further B cell biology.

A brief outline of the future chapters is as follows: Chapter 2 of this dissertation outlines how hierarchical clustering can identify B cell clones with high confidence in Ig AIRR-Seq data. Chapter 3 describes a framework that enables integration of several tools for Ig repertoire analysis. Chapter 4 contains applications of these analysis tools to identify novel neutralizing antibodies against West

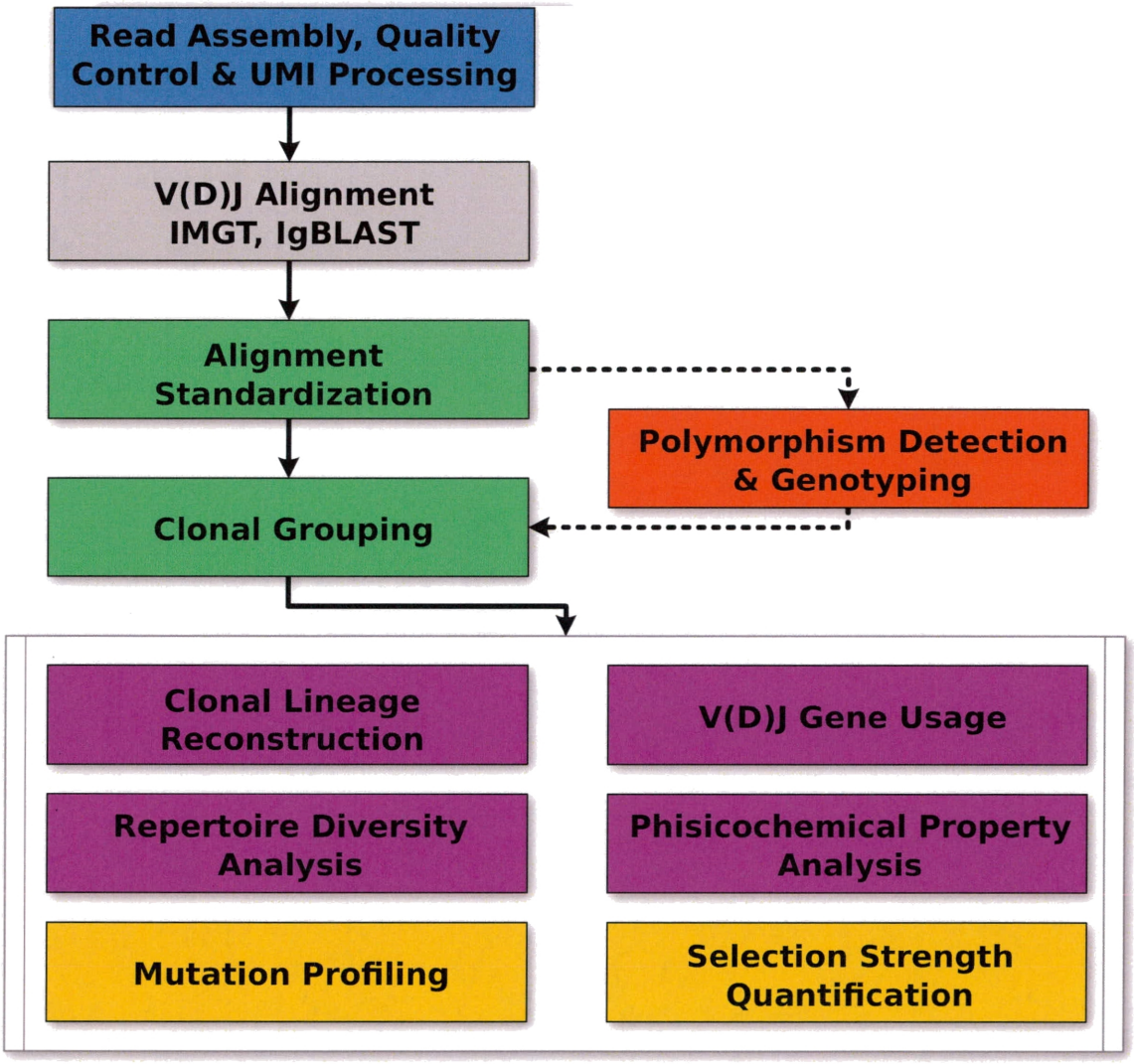

Figure 1.2: **Ig repertoire analysis workflow.** An overview of the typical steps of analyzing an Ig AIRR-Seq dataset.

Nile virus and other research topics. The concluding Chapter 5 reviews the contribution of this dissertation to the Ig repertoire sequencing field and future directions in which the work can be continued.

Chapter 2

Hierarchical clustering to identify B cell clones in AIRR-Seq data

This chapter has also been accepted for publication as:

Gupta, N. T. *et al.* Hierarchical clustering can identify B cell clones with high confidence in Ig repertoire sequencing data. *Journal of Immunology* Accepted (2017)

2.1 Introduction

The capacity of B cells to modify their antibodies or immunoglobulin (Ig) receptors to adapt in response to pathogenic challenges is a key mechanism that protects us from infection. An initial diversity of $\sim 10^7$ unique Ig molecules (Volpe & Kepler, 2008) stems from somatic recombination of gene segments in the B cell Ig gene locus compounded by stochastic nucleotide insertions and deletions at the junctions of these segments. Upon activation, these naive B cells diversify further by undergoing clonal expansion with somatic hypermutation (SHM) in the Ig gene (approximately one point mutation per 1000 bp per cell division (McKean *et al.*, 1984; Kleinstein *et al.*, 2003)) followed by selection for higher affinity B cells. This micro-evolutionary process known as affinity maturation results in B cells with diversified Ig receptors that are clonal relatives of the original activated B cell. In healthy human adults, class-switched memory B cells express Ig receptors that are $\sim 7\%$ mutated (Y.-C. B. Wu *et al.*, 2012). Our ability to profile this adaptive immune response has dramatically improved through the application of next-generation sequencing, which allows for

measurement of tens to hundreds of millions of B cell receptors (Boyd & Joshi, 2014). However, the identification of sequences that belong to the same B cell clone in these data remains a significant challenge (Yaari & Kleinstein, 2015).

Adaptive immune receptor repertoire sequencing (AIRR-Seq) is being widely used for both basic science and clinical studies (J. J. a. Weinstein *et al.*, 2009; Boyd, Marshall, *et al.*, 2009; Benichou *et al.*, 2012). Statistical properties of the repertoire, such as diversity or mutational load, are being used to gain insights into the dysregulation that occurs with aging or disease. Properly identifying clones is central to the calculation of many of these properties. For example, clone size distributions are the basis for several diversity measures, such as species richness, Shannon entropy, and the Gini-Simpson index (Yaari & Kleinstein, 2015) that parallel diversity measures in ecology (Hill, 1973). Diseases such as chronic lymphocytic leukemia are characterized by low diversity that is driven by the dominance of a small number of clones (van Dongen *et al.*, 2003), and repertoire sequencing has been used to improve minimal residual disease detection for lymphoid cancers (Boyd, Marshall, *et al.*, 2009; Logan *et al.*, 2011). Responses to drugs such as rituximab have also been measured by changes in repertoire diversity in autoimmune disease (Hershberg, Meng, *et al.*, 2014; Boletis *et al.*, 2009), characterizing treatment regimens that lead to successful remission or result in persistent clonal expansions. Decreases in repertoire diversity have been associated with aging (Y.-C. B. Wu *et al.*, 2012; Ademokun *et al.*, 2011; C. Wang *et al.*, 2014). In subjects with seasonal allergies, the IgE repertoire is the least diverse compared to other isotypes in blood and nasal biopsies, indicating a focused immune response (Y. C. B. Wu *et al.*, 2014).

Analysis of diversity within a clone also has several applications. Reconstruction of the B cell clonal lineages using methods such as maximum parsimony or likelihood (Nei & Kumar, 2000) allows tracing somatic mutations through the Ig sequences and helps in understanding the evolution of neutralizing antibodies (J. Zhu *et al.*, 2013; Tsioris *et al.*, 2015). Lineage relationships have also been used to gain insight into the mechanisms underlying isotype switching (Horns *et al.*, 2016; Looney *et al.*, 2016) and to show that B cell clones in the central nervous system in subjects with multiple sclerosis are first activated in the periphery (Stern *et al.*, 2014). Identifying clones that include sequences with known antigen specificities has also been used to reveal novel antigen-specific sequences (J. Zhu *et al.*, 2013; Jiang, He, *et al.*, 2013). Thus, clonal partitioning of AIRR-Seq data is central to a wide range of applications.

Despite its importance, there is no consensus on the best method for grouping Ig sequences

into B cell clones. Most current approaches leverage the high diversity of the junction region (*i.e.*, where the V, D, and J gene segments join) as a "fingerprint" to identify each B cell clone (Hershberg & Luning Prak, 2015). Since it is unlikely that two separate recombination events would lead to identical junctions, sequences with junction regions that are "similar enough" are determined to share a common B cell ancestor (*i.e.*, be clonally related) rather than to have arisen independently. Probabilistic models have been developed to calculate likelihood of sharing a B cell ancestor and subsequently infer clonal grouping (Kepler, 2013; Ralph & Matsen, 2016). However, these algorithms have run times that scale exponentially, which is computationally intractable for large sequencing datasets (Ralph & Matsen, 2016). In practice, most studies cluster sequences based on junction sequence similarity (Ademokun *et al.*, 2011; Stern *et al.*, 2014; Tsioris *et al.*, 2015; Jiang, He, *et al.*, 2013; Chen *et al.*, 2010; Glanville, Kuo, *et al.*, 2011; Jiang, J. a. Weinstein, *et al.*, 2011).

While many clustering approaches exist, hierarchical clustering is the most widely used framework for grouping clonally-related sequences. Hierarchical clustering requires a measure of distance between pairs of sequences, and a choice of linkage to define the distance between groups of sequences. Since hierarchical clustering produces a tree defining the relationships between all sequences, it is also necessary to specify a method to cut the hierarchy in order to identify discrete clonal groups. In practice, most studies first split the sequences using some similarity requirement on the germline gene segments (*e.g.*, identical V and J gene segments, and junction length), and then apply hierarchical clustering on the junction sequence of these smaller groups (Boyd, Marshall, *et al.*, 2009; Stern *et al.*, 2014; Tsioris *et al.*, 2015; Jiang, He, *et al.*, 2013; Chen *et al.*, 2010; Glanville, Kuo, *et al.*, 2011; Jiang, J. a. Weinstein, *et al.*, 2011; Y.-C. Wu *et al.*, 2010; Briney *et al.*, 2016). Several distance metrics have been proposed, including Hamming distance, which is simply the absolute count of differences between two amino acid (Glanville, Kuo, *et al.*, 2011; Y.-C. Wu *et al.*, 2010) or nucleotide (Jiang, He, *et al.*, 2013; Jiang, J. a. Weinstein, *et al.*, 2011) sequences, normalized edit distance (Chen *et al.*, 2010), and a metric that incorporates hot/cold-spot biases in SHM targeting (Stern *et al.*, 2014; Tsioris *et al.*, 2015). In addition to metrics defining distance between two sequences, linkage methods define how distance is calculated between groups of sequences. Different clonal grouping algorithms use single (Stern *et al.*, 2014; Tsioris *et al.*, 2015; Jiang, He, *et al.*, 2013; Jiang, J. a. Weinstein, *et al.*, 2011), average (Chen *et al.*, 2010), or complete (Ademokun *et al.*, 2011) linkage. The threshold at which the hierarchy is cut to define clusters of

clonally related sequences has also been determined in several ways. Chen *et al.*, 2010 propose a fixed threshold that is manually identified based on when the rate of cluster merging events changes for a gold standard dataset. Glanville, Kuo, *et al.*, 2011 introduced a method based on the observed bimodal distribution of distances from each sequence to its nearest neighbor. In this case, the first mode is assumed to represent sequences with clonal relatives in the data (near neighbors), while the second mode is taken to represent sequences without clonal relatives in the data (distant neighbors). The threshold is then selected to be the value that separates the two modes of this distribution (Jiang, He, *et al.*, 2013; Glanville, Kuo, *et al.*, 2011). As of yet, there has not been an in-depth evaluation of performance of hierarchical clustering-based clonal grouping algorithms including a comparison of the different distance and linkage methods on AIRR-Seq data.

In this paper, we carry out a comparative analysis of distance metrics and linkage methods for hierarchical clustering-based clonal grouping. A combination of experimental and simulation-based criteria are used to evaluate the performance of these algorithms, including estimates of specificity, sensitivity, and positive predictive value (PPV). Overall, we find that single-linkage hierarchical clustering with nucleotide Hamming distance has excellent performance, with specificity, sensitivity, and PPV all over 99%. Implementations of all clonal grouping methods, along with extensive documentation, are available through the Change-O and SHazaM packages (Gupta *et al.*, 2015) as part of the Immcantation tool suite (`http://immcantation.readthedocs.io`) for AIRR-Seq analysis.

2.2 Materials and methods

2.2.1 Human B cell receptor repertoire sequencing data

Three B cell receptor repertoire sequencing datasets (Healthy, Dengue and WNV) were used to measure the performance of clonal grouping methods. The "Healthy" dataset was composed of sequences from peripheral blood mononuclear cells (PBMCs) isolated from healthy adult subjects (n = 27) as previously described (C. Wang *et al.*, 2014). The "Dengue" dataset was composed of sequences from PBMCs isolated from subjects with acute Dengue infection (n = 42) as described previously (Parameswaran *et al.*, 2013). The "WNV" dataset was composed of sequences from PBMCs and sorted plasma, memory, and naive B cells isolated from subjects recently infected with WNV (n = 7) as previously described (Tsioris *et al.*, 2015). In each case, processed sequencing data was obtained from the authors. Germline gene segments were inferred for each sequence by using IMGT/HighV-QUEST (Alamyar *et al.*, 2012). The "Healthy" dataset was run through IMGT/HighV-QUEST on December 21, 2014, "Dengue" was run through IMGT/HighV-QUEST March 12, 2015, and the "WNV" dataset was run through IMGT/HighV-QUEST on March 21, 2014. Sequences identified as non-functional by IMGT/HighV-QUEST were removed using the changeo-clt toolkit version 0.2.0 (Gupta *et al.*, 2015).

Two additional B cell receptor repertoire sequencing datasets from healthy adult subjects were used as a source of naive B cell receptor sequences for the lineage simulations. The first was composed of sequences from PBMCs and sorted naive B cells isolated from healthy control subjects (n = 4) as part of a study of Myasthenia Gravis described in Vander Heiden, et al. (Submitted). This dataset is available on SRA (Accession number: SRP081539). The second was composed of sequences from total RNA isolated from blood samples of healthy adult subjects (n=3) as part of an influenza vaccination study described in Laserson *et al.*, 2014. In this case the samples, which were originally sequenced using Roche 454, were re-sequenced using Illumina MiSeq and published for the first time here (see details below). Identical sequences from the same sample were counted once, but identical sequences from different samples were counted independently. This dataset is available on SRA (BioProject ID: PRJNA349143).

2.2.2 Library preparation and BCR sequencing of healthy subject sequences from Laserson *et al.*, 2014

The blood samples collected in the influenza vaccination study by Laserson *et al.*, 2014 were re-sequenced using the Illumina MiSeq platform as previously described (Tsioris *et al.*, 2015; Di Niro, S.-J. Lee, *et al.*, 2015). Briefly, RNA was reverse-transcribed into cDNA using a biotinylated oligo dT primer. An adaptor sequence was added to the 3' end of all cDNA, which contains the Illumina P7 universal priming site and a 17-nucleotide unique molecular identifier (UMI). Products were purified using streptavidin-coated magnetic beads followed by a primary PCR reaction using a pool of primers targeting the IGHA, IGHD, IGHE, IGHG, IGHM, IGKC and IGLC regions, as well as a sample-indexed Illumina P7C7 primer. The immunoglobulin-specific primers contained tails corresponding to the Illumina P5 sequence. PCR products were then purified using AMPure XP beads. A secondary PCR was then performed to add the Illumina C5 clustering sequence to the end of the molecule containing the constant region. The number of secondary PCR cycles was tailored to each sample to avoid entering plateau phase, as judged by a prior quantitative PCR analysis. Final products were purified, quantified with Agilent Tapestation and pooled in equimolar proportions, followed by high-throughput paired-end sequencing on the Illumina MiSeq platform. For sequencing, the Illumina 600 cycle kit was used with the modifications that 325 cycles was used for read 1, 6 cycles for the index reads, 300 cycles for read 2 and a 10% PhiX spike-in to increase sequence diversity.

2.2.3 Read processing of healthy subject sequences from Laserson *et al.*, 2014

MiSeq reads were demultiplexed using Illumina software. Positions with less than Phred quality 5 were masked with Ns. Isotype-specific primers and unique molecular barcodes (UMI) were identified in the amplicon and trimmed using pRESTO (Vander Heiden *et al.*, 2014) MaskPrimers-cut. Read 1 and read 2 consensus sequences were generated separately for each mRNA from reads grouped by UMI, which represent PCR replicates arising from a single initiating mRNA molecule. UMI read groups were aligned with MUSCLE (Edgar, 2004), and pRESTO was used to construct a consensus sequence using BuildConsensus, requiring $\geq 60\%$ of called PCR primer sequences agree for the read group, maximum nucleotide diversity of 0.1, using majority rule on indel positions, and masking

alignment columns with low posterior (consensus) quality. Paired end consensus sequences were then stitched in two rounds. First, ungapped alignment of each read pair's consensus sequence termini was optimized using a Z-score approximation and scored with a binomial p-value as implemented in pRESTO AssemblePairs-align. For read pairs failing to stitch this way, stitching was attempted using the human BCR germline V exons to scaffold each read prior to stitching or gapped read-joining, using pRESTO AssemblePairs-reference. Positions with posterior consensus quality less than Phred 5 were masked again with Ns. All pRESTO tools used were version 0.5.1 in conjunction with Python 3.4. Germline gene segments were inferred using IgBLAST version 1.4.0 (Ye *et al.*, 2013) with the IMGT/GENE-DB (Giudicelli *et al.*, 2005) reference sequences from June 7, 2014 and output was parsed with changeo-clt (Gupta *et al.*, 2015) MakeDb version 0.3.0. Duplicate sequences were collapsed and only those heavy chain sequences with at least two reads supporting the sequence were retained for further analysis.

2.2.4 Simulation of B cell clonal lineages

Each simulated clone was generated by introducing mutations into an experimentally observed naive B cell receptor sequence according to an observed lineage tree topology (*i.e.*, branching pattern). Lineage tree topologies were previously derived based on sequencing data from lymph node samples collected as part of a published study of Multiple Sclerosis (Stern *et al.*, 2014). The set of 7103, 4066, 8244, and 14782 lineage topologies from four subjects (referred to here as R1, R2, R3 and R4, respectively) were each used as the basis for 10 simulations resulting in 40 total simulated datasets. To generate a simulated dataset, the root of each lineage was randomly chosen (without replacement) from a large pool of un-mutated sequences from healthy subjects obtained from Vander Heiden, et al. (Submitted) and (Laserson *et al.*, 2014) (described above). Mutations were then added to the sequence in order to match the experimentally-observed mutation counts of each branch in the lineage tree according to the human S5F (hS5F) targeting model (Yaari, Vander Heiden, *et al.*, 2013). The simulated sequences then had germline gene segments inferred using IgBLAST version 1.4.0 (Ye *et al.*, 2013) with the IMGT/GENE-DB (Giudicelli *et al.*, 2005) reference sequences from May 2, 2016 and output was parsed with changeo-clt (Gupta *et al.*, 2015) MakeDb version 0.3.3.

2.2.5 Distance metrics

Hamming distance was defined as the absolute count of letter changes between nucleotide junction sequences (ham) or amino acid junction sequences (aa). The 5-mer distance metrics were all based on the hS5F targeting and substitution models described in (Yaari, Vander Heiden, *et al.,* 2013), which estimates: (1) the relative probability of a nucleotide position being targeted for somatic mutation, and (2) the probability of mutating to each of the three other possible nucleotides, based on the two nucleotides up- and downstream. This probability p was transformed into a distance d using the formula: $d = -log_{10}(p)$. The distance between two junction sequences was defined to be the sum of distances between each nucleotide position. For a given mutation between two junction sequences, the hS5F-min model took its distance to be the minimum of mutating from nucleotide n_1 to n_2 and from n_2 to n_1 at that position. The hS5F-avg model took the distance of the mutated position to be the average of mutating from n_1 to n_2 and from n_2 to n_1. The human S1F (hS1F) model is equivalent to hS5F-min , but used the human symmetric substitution matrix based on the single mutated nucleotide described in (Yaari, Vander Heiden, *et al.,* 2013). The m1n model is equivalent to hS5F-min, but used the mouse symmetric substitution matrix based on the single mutated nucleotide described in (Shapiro *et al.,* 2003). When normalizing by length, these distances were divided by the length of the junction region.

2.2.6 Implementation of clonal grouping algorithms

Clonal grouping algorithms were implemented and are made available in the change-o-clt toolkit (Gupta *et al.,* 2015) (version 0.3.1 or newer). Sequences were first grouped by shared V gene, J gene and junction length. Within these groups of sequences, hierarchical clustering was performed using the bygroup subcommand of DefineClones.py with the specified distance metric and linkage type. The resulting hierarchy was then trimmed into flat clusters at a fixed threshold determined using an automated method based on analyzing the "distance-to-nearest" profile. For each sequence, the distance to its nearest un-identical junction was calculated using the SHazaM R package (Gupta *et al.,* 2015) (version 0.1.3 or newer). The ideal bandwidth for the fourth derivative kernel density estimate of these distances was then estimated using the unbiased cross-validation method (Wand & Jones, 1995) of the fourth derivative of the kernel density estimate (Darlington, 1970; Hansen, 2004) from the kedd R package (Guidoum, 2015) (version 1.0.3). This bandwidth was used to calculate a binned kernel density estimate of the distances with a Gaussian kernel using the KernS-

mooth R package (version 2.23-15). The minimum between the two modes of the resulting bimodal distribution of distances was then calculated by finding the first value at which the first derivative was zero while the second derivative was positive, indicating a local minimum following a local maximum. If such a minimum were not found an error would have been returned, but this was not the case for any of the analyses herein.

2.2.7 Specificity, positive predictive value, and sensitivity

Performance was characterized by considering the binary classification task of defining the relationship between all pairs of sequences (s_1 and s_2) with the same junction length. These classifications were then pooled together for the entire dataset. If s_1 and s_2 were known to be unrelated (termed condition negative) but were grouped into the same cluster (termed test positive), this was counted as a false positive. If they were grouped into different clusters (termed test negative), this was a true negative. If s_1 and s_2 were known to be related (termed condition positive) but were grouped into different clusters, this was counted as a false negative. If they were grouped into the same cluster, this was counted as a true positive. These relationships are outlined in Figure 2.1.

In the case of experimental data, two sequences were known to be unrelated if they were derived from two separate individuals. Therefore, false positives were defined as sequences from different individuals being grouped together in a clone, while true negatives were defined as sequences from different individuals that were grouped into separate clones. Specificity was then calculated by dividing the number of true negative classifications by the number of condition negative classifications. In other words, specificity was defined as the fraction of pairs of unrelated sequences that were successfully inferred by the algorithm to be unrelated.

For the simulated datasets, the precise clonal membership of each sequence was known, yielding the intuitive definition of false positive and false negative classification. Positive predictive value (PPV) was calculated by dividing true positive classifications by test positive classifications. In other words, PPV was the fraction of predicted clonal relationships that were actually true. Sensitivity was calculated by dividing true positive classifications by condition positive classifications. In other words, sensitivity was defined to be the fraction of actual clonal relationships that were successfully inferred by the algorithm.

2.2.8 Shannon entropy calculation

The Shannon entropy of clonally related sequences (within clones) was calculated for true clones having at least two members. Entropy was calculated for each of the first 24 nucleotide positions of the junction within each clone and averaged across clones having junction length <30 nt and 51 nt. To calculate Shannon entropy of clonally unrelated sequences (between clones), the most mutated sequence was selected from each true clone. These mutated sequences were then placed into groups sharing the same V gene, J gene, and junction length. Groups with only one sequence were discarded. Entropy was calculated for each of the first 24 nucleotide positions of the junction within each group and averaged across groups having junction length <30 nt and 51 nt. Error bars represent standard error of the mean. The calculations were made on all of the simulated datasets pooled together.

2.3 Results

The problem of clonal grouping takes a set of B cell receptor sequences as input and returns a partition of that set into subsets (clonal groups) that each represent an independent clonal lineage. Here we investigate hierarchical clustering-based algorithms which infer a dendrogram based on pairwise sequence distances and then cut the dendrogram at a fixed distance (or "threshold") to predict groups of clonally-related sequences. To evaluate the performance of these clonal grouping algorithms, we consider three metrics: specificity, sensitivity, and positive predictive value (PPV) (Figure 2.1).

Specificity quantifies how frequently unrelated sequences are correctly separated into different clonal groups. In most experimental datasets, the exact clonal relationships between sequences are unknown. However, to estimate specificity we can take advantage of the fact that B cell clones cannot span multiple individuals (by definition). Using this knowledge, specificity is defined based on how frequently sequences from separate individuals are incorrectly inferred to be clonal relatives. This measure is used to quantify performance on three human Ig AIRR-Seq datasets (referred to as Healthy, Dengue, and WNV as detailed in Methods), each of which contains samples from multiple individuals.

Sensitivity represents inclusivity of an algorithm by measuring how often clonally related sequences are grouped together. PPV is a complementary metric that quantifies the precision of an algorithm by measuring how often inferred clonal relatives are truly clonally related. The calculations for sensitivity and PPV require knowledge of true clonal relationships and thus cannot be estimated from current human experimental datasets. For these measures, performance was evaluated using 40 simulated datasets based on four experimentally observed sets of clonal lineage structures (referred to as R1-R4 as detailed in Methods). In the following sections, we use these performance metrics on the human experimental and simulation data to evaluate the choice of distance metrics, linkage methods and threshold parameters in clustering-based clonal grouping algorithms.

2.3.1 Automated determination of clonal distance thresholds

A key step in hierarchical clustering-based clonal grouping involved choosing a threshold at which to cut the dendrogram, thus forming discrete groups of clonally-related sequences. In previous work (Glanville, Kuo, *et al.*, 2011; Jiang, He, *et al.*, 2013), this threshold has often been fixed at a single value determined by manual inspection of a histogram of nearest-neighbor distances (the so-called

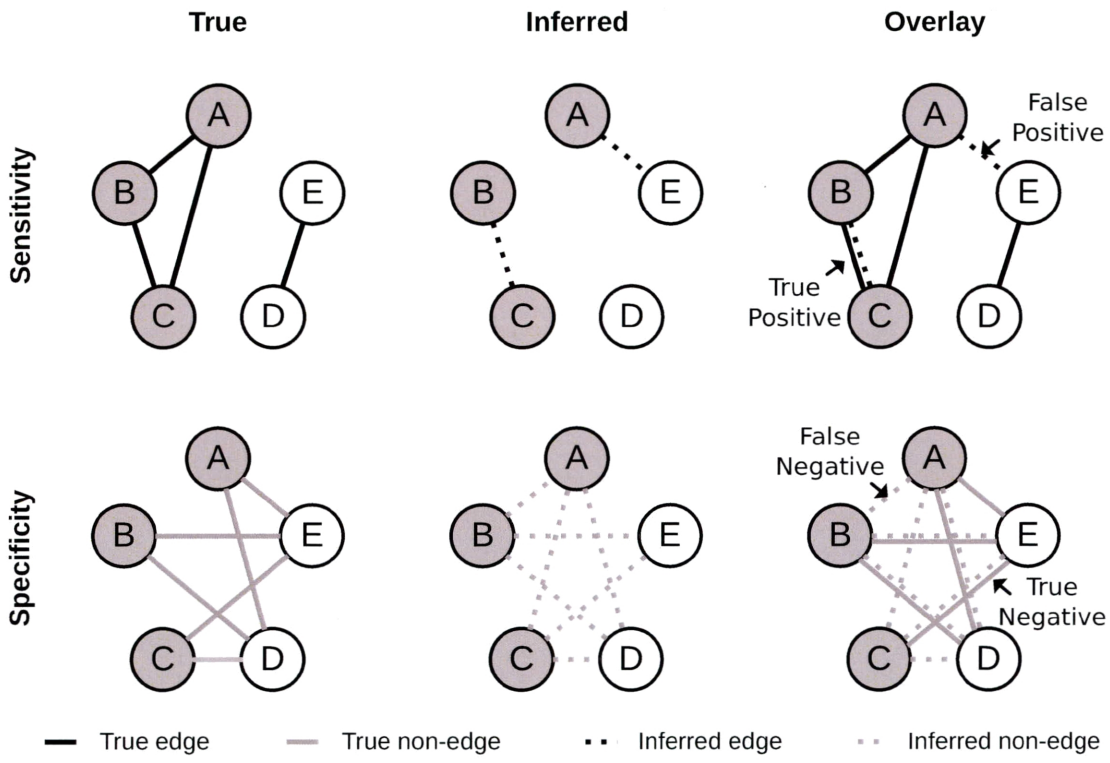

Figure 2.1: **Overview of method for calculating the performance measures: sensitivity, PPV and specificity.** Each node corresponds to a sequence whose true clonal membership is depicted by its shading (grey or white) in all panels. True relationships (leftmost panels) are shown as solid lines connecting pairs of clonally-related sequences (top) or unrelated sequences (bottom). The relationships inferred by the clonal grouping algorithm (middle panels) are also defined between pairs of sequences (dashed lines). The true and inferred edges are compared to assess performance. Sensitivity is defined by the number of true positive edges divided by the number of true edges. PPV is the number of true positive edges divided by the number of inferred edges. Specificity is number of true negative edges divided by the number of true non-edges.

"distance-to-nearest" plot (Yaari & Kleinstein, 2015)). These histograms are typically bimodal and the threshold is selected to separate these two modes (Figure 2.2A). This choice is motivated by the intuition that the smaller peak represents the distance between sequences within a clone (intra-clonal distance), while the larger peak represents the distance between sequences in different clones (inter-clonal distance). Inspection of the nearest-neighbor distance distributions for the Healthy, Dengue, and WNV experimental datasets used in this study showed that they are each clearly bimodal. However, they differed in the values that best separated the two modes (Figure 2.2B). This result indicates that the distance threshold for clonal groping is dataset-specific and must be re-computed for each study.

Manual determination of the clustering threshold is problematic because inspecting a distribution by eye is time-consuming and imprecise. We therefore sought to develop an automated analytic procedure for inferring the clustering threshold that mimics the widely used manual approach. Since the histograms generated from real data are rarely smooth, we first smooth the empirical distributions using a binned Gaussian kernel density estimator using a procedure that is well-suited for bimodal distributions (Hansen, 2004) (see Methods for details). Next, we computationally determine the minimum between the two peaks of the smoothened distribution and define this value to be the clustering threshold (Figure 2.2A). This method placed the threshold at intuitive locations in the Healthy, Dengue and WNV experimental datasets (Figure 2.2B). This method for automated determination of the clustering threshold enables efficient application of clonal grouping algorithms under many parameter settings and on many different datasets.

We next applied the automated threshold to assess the performance of clonal grouping methods on experimental and simulated datasets. Hierarchical clustering using the nucleotide-based Hamming distance metric with single linkage was an effective approach. The mean specificity of the algorithm was over 99% on experimental data (Figure 2.3A), the mean sensitivity was $\sim$ 99% on simulated datasets (Figure 2.3B), and PPV was over 99% on simulated datasets (Figure 2.3C). In contrast, using amino acid-based Hamming distance — which has been used in some previous studies (Glanville, Kuo, *et al.*, 2011; Y.-C. Wu *et al.*, 2010) — had significantly worse sensitivity (Figure 2.4). One potential shortcoming of using the Hamming distance metric is that mutations in short junctions are penalized more heavily than mutations in longer junctions. Since junction regions vary widely in length (33–81 nucleotides, 95% range from experimental datasets) and the clustering algorithm uses a fixed threshold, this bias could lead to suboptimal performance. In an

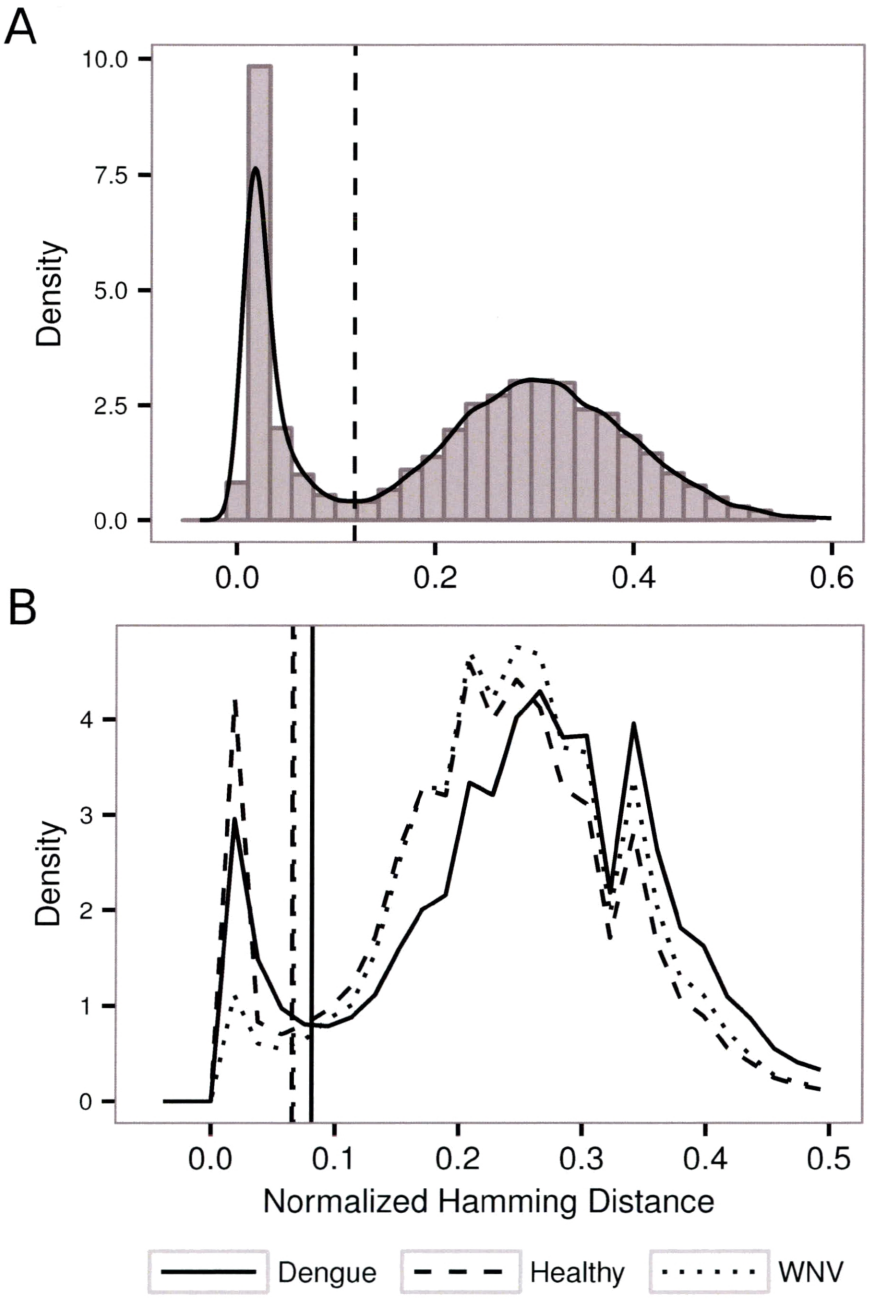

Figure 2.2: **Analysis of the "distance-to-nearest" neighbor plot to define the distance threshold for partitioning clones.** For each sequence, the length-normalized nucleotide Hamming distance to every other sequence was calculated, and the nearest (non-zero) neighbor was identified. (A) The histogram of nearest neighbor distances for a simulated dataset was fit using a density estimation of the distribution (solid line), and this fitting was then used to automatically infer a threshold that separated the two modes of the distribution (dotted vertical line). (B) Nearest neighbor distributions were calculated for the Dengue (solid line), Healthy (dashed line), and WNV (dotted line) experimental datasets. Inferred thresholds for each of these human data sets are indicated by the vertical lines.

attempt to address this issue, we and others have used a length-normalized Hamming distance metric, in which Hamming distance is divided by the length of the junction. This length-normalization had minimal effect on specificity in the experimental data (Figure 2.3A), but significantly improved sensitivity (Figure 2.3B) and PPV (Figure 2.3C) in the simulated data ($p < 10^{-4}$, paired t-test). Thus, length-normalization of the distance metric is an important step in clonal grouping algorithms.

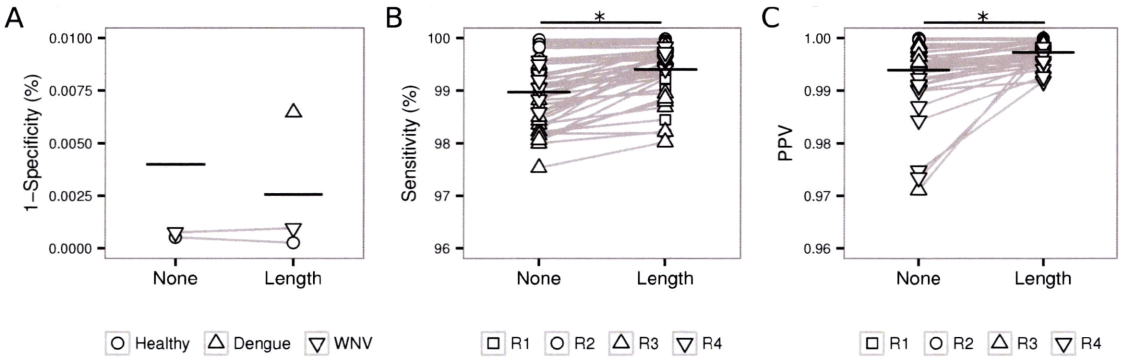

Figure 2.3: **Length normalization of the distance measure increases performance.** Single linkage hierarchical clustering was used to identify clonally-related sequences using a distance metric based on the absolute Hamming distance of the junction sequences (None), or the Hamming distance normalized by the length of the junction (Length). (A) Specificity was calculated using three human experimental data sets (Healthy, Dengue, and WNV). Sensitivity (B) and PPV (C) were calculated using 40 simulated datasets based on four experimentally observed sets of clonal lineage structures (R1-R4). Bars indicate mean performance. * $p < 0.0001$ by paired t-test.

2.3.2 Single linkage has highest sensitivity with minimal compromise of PPV

Hierarchical and other agglomerative clustering algorithms require a method for determining the distance between two sets of points (in this case, sequences). The most common linkage methods include single, average, and complete linkage (Jain & Dubes, 1988). Single linkage defines the inter-set distance as the minimum distance between all pairs of points from the given sets. This generally results in larger and more heterogeneous clusters (Jain & Dubes, 1988). Complete linkage defines the inter-set distance as the maximum distance between all pairs of points from the given sets, and generally results in smaller and more homogeneous clusters (Jain & Dubes, 1988). Average linkage defines the inter-set distance as the average distance between all pairs of points from the given sets, thus providing a compromise between single and complete linkage.

As expected, single linkage had the lowest specificity followed by average and then complete linkage (Figure 2.5A). However, these differences were small, and specificity was over 99% in all cases. A similar ranking was found for PPV, with complete and average linkage significantly improving performance relative to single linkage ($p < 10^{-4}$, paired t-test; Figure 2.5C). Once again, however, the absolute performance differences were small, with all three approaches exhibiting a mean PPV of over 99%. As specificity and PPV both reflect the accuracy of clonal grouping, we conclude that all of the linkage methods are accurate. In contrast, single linkage exhibited significantly higher sensitivity for clonal grouping relative to both average and complete linkage ($p < 10^{-4}$, paired t-test; Figure 2.5B). In this case, the sensitivity differences were large, with single linkage having a mean sensitivity of 99%, compared with 88% for average linkage and 60% for complete linkage. Overall, these results show that single linkage is significantly better at capturing the breadth of true clonal relationships, with only a modest reduction in accuracy.

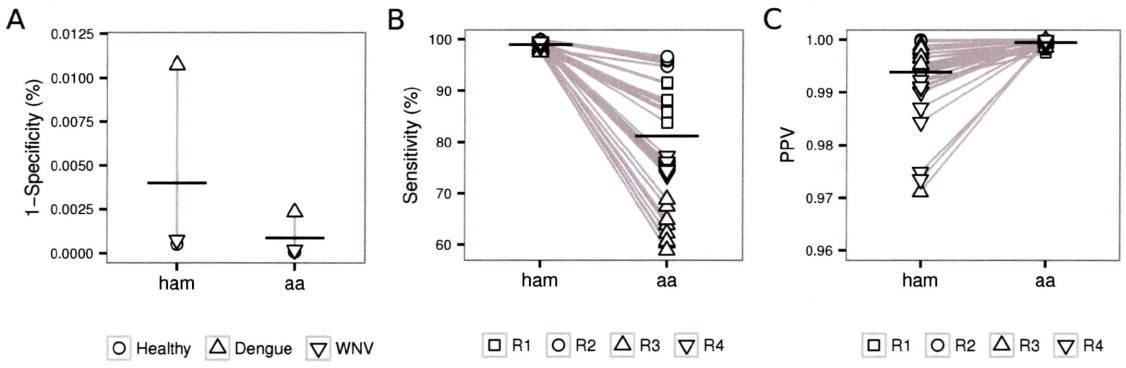

Figure 2.4: **Nucleotide Hamming distance performs much better than amino acid Hamming distance.** Single linkage hierarchical clustering was used to identify clonally-related sequences using absolute Hamming distance based on the nucleotide (ham) and amino acid (aa) sequence. (A) Specificity was calculated using three human experimental data sets (Healthy, Dengue and WNV). Sensitivity (B) and PPV (C) were calculated using 40 simulated datasets based on four experimentally observed sets of clonal lineage structures (R1-R4). Bars indicate mean performance.

2.3.3 Incorporating SHM biases does not significantly improve clonal grouping

While the Hamming distance between two sequences is quick and easy to compute, it does not account for the intrinsic targeting and substitution biases in SHM (Yaari, Vander Heiden, *et al.*, 2013). It is well established that AID and the error prone DNA repair pathways that drive B cell diversification frequently target specific DNA motifs (termed hot-spots), while others are rarely mutated (termed cold-spots). There is also a substitution bias such that transition mutations are significantly more frequent than transversions. Weighing all mutations equally undervalues the less probable mutations because two sequences are less likely to be part of a clone if they differ by mutations that occur less frequently (*i.e.*, transversion mutations at cold-spot positions).

To test whether accounting for the intrinsic biases of somatic hypermutation could improve the performance of clonal grouping algorithms, we implemented four previously proposed SHM models that account for these biases in different ways (see Methods for details). The first two models (hS5F-min and hS5F-avg) incorporate the human S5F targeting (mutability and substitution) models that incorporate the effects of the two nucleotides up- and downstream of a mutation (Yaari, Vander Heiden, *et al.*, 2013). For each pair of nucleotides (n_1 and n_2) that differ between two junctions being compared, the hS5F-avg metric assumes that each one has an equal probability of having been present in the most recent common ancestor. Thus, the distance is taken as the average of mutating from n_1 to n_2 and from n_2 to n_1. The hS5F metric assumes that the ancestral base is the one that leads to the most likely mutation, and thus uses the minimum distance at each nucleotide position. The second two models (hS1F and m1n) ignore mutability, but account for substitution bias using a model that depends only on the targeted base (*i.e.*, ignoring surrounding nucleotides). As these models are symmetric, there is no assumption of which nucleotide was ancestral. Surprisingly, we found no significant performance differences for any of the distance metrics in experimental (Figure 2.6A) or simulated datasets (Figure 2.6B,C). These results support the use of the more efficient nucleotide Hamming distance metric. Overall, we find that hierarchical clustering using length-normalized nucleotide Hamming distance with single linkage performs well, with mean sensitivity, specificity, and PPV all over 99%.

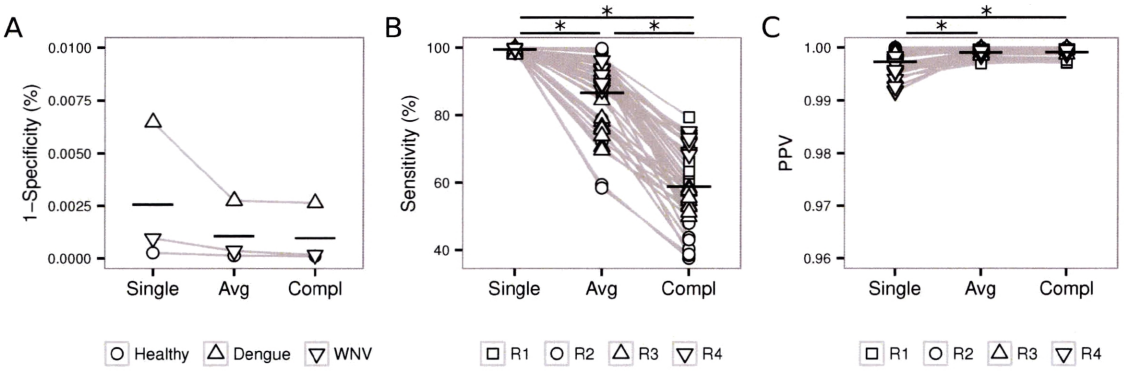

Figure 2.5: **Single linkage clustering provides the highest sensitivity, with minimal loss of specificity or PPV.** Hierarchical clustering was used to identify clonally-related sequences using length-normalized Hamming distance and Single (Single), Average (Avg) or Complete (Compl) linkage. (A) Specificity was calculated using three human experimental data sets (Healthy, Dengue and WNV). Sensitivity (B) and PPV (C) were calculated using 40 simulated datasets based on four experimentally observed sets of clonal lineage structures (R1-R4). Bars indicate mean performance. * $p < 0.0001$ by paired t-test.

2.3.4 Sequences with short junctions have high false positive rate

We next investigated the dependence of performance on junction length to better understand the source of errors in clonal grouping. Junction length was minimally correlated with specificity in the experimental datasets ($r = 0.1$, Pearson's correlation; Figure 2.7A). Similarly, there was no correlation of junction length with sensitivity in the simulated datasets ($r = 0.02$, Pearson's correlation; Figure 2.7B). In contrast, there was a strong positive correlation of junction length with PPV in the simulated datasets ($r = 0.4$, Pearson's correlation), with mean PPV of 99.1% for sequences with shorter junctions (junctions shorter than 30 nt represented by at least 0.001% of the sequences in the repertoire) compared to a mean PPV of 99.8% for sequences with longer junctions (Figure 2.7C). For these sequences with shorter junction lengths, the right peak in the nearest-neighbor distance distributions (interpreted as distances between unrelated sequences) begins to overlap the left peak (interpreted as distances between clonally-related sequences). This pattern of decreasing inter-clonal distances as junction lengths decrease was also apparent considering nearest-neighbors across individuals (Figure 2.8). Thus, it appears that the distance threshold that effectively separates clonal members with longer junctions begins to group together unrelated sequences with shorter junctions. These results raise the possibility that using a single distance threshold to separate clonal groups may not be optimal for sequences with shorter junctions.

To determine if using multiple distance thresholds could improve performance, we assessed precision (PPV) and recall (sensitivity) across a range of distance thresholds using sequences of varying junction lengths. We selected the shortest junction length with at least 0.001% of total sequences (24 nt), the overall mean junction length (51 nt), and the longest junction length with a distinguishable spread in precision across distance thresholds (81 nt) as example junction lengths with which to assess performance. When considering all junction lengths as one group, the automated threshold appears close to optimal in trading off between PPV and sensitivity with both over 99% (Figure 2.9A). The same holds true when considering separately the mean and longer junction lengths of 51 nt (Figure 2.9C) and 81 nt (Figure 2.9D) respectively. Interestingly, the single threshold chosen on the entire data set still provided a near optimal trade-off in performance for sequences with shorter (24 nt) junctions, although peak sensitivity was lower for some of the simulated repertoires (Figure 2.9B). Thus, using a junction length-specific threshold is unlikely to improve performance.

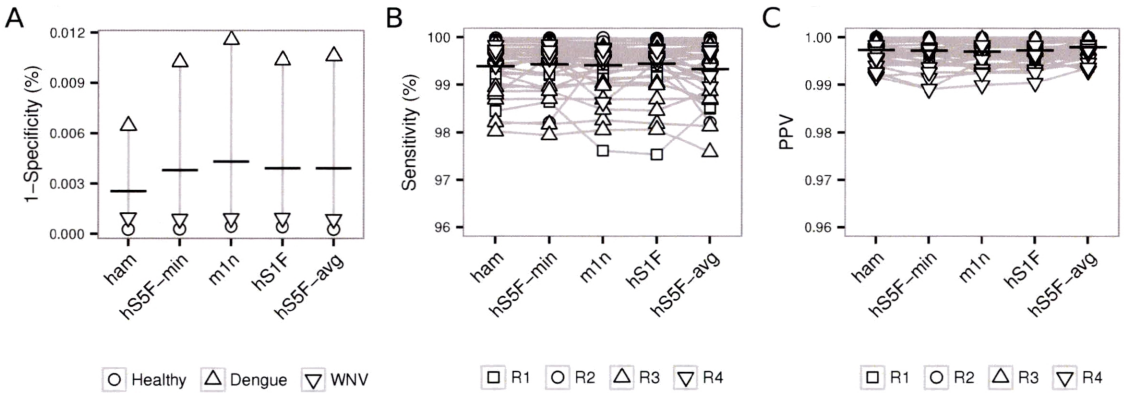

Figure 2.6: **Including hot- and cold-spot biases in the distance measure does not significantly impact the performance of clonal grouping.** Single linkage hierarchical clustering was used to identify clonally-related sequences using length-normalized nucleotide Hamming distance (ham), as well as other distance metrics that incorporated varying SHM biases as described in Materials and Methods: hS5F-min, m1n, hS1F, and hS5F-avg. (A) Specificity was calculated using three human experimental data sets (Healthy, Dengue and WNV). Sensitivity (B) and PPV (C) were calculated using 40 simulated datasets based on four experimentally observed sets of clonal lineage structures (R1-R4). Bars indicate mean performance.

The inability to separate unrelated sequences with shorter junction lengths implies a lack of diversity between clones. Indeed, sequences with short junctions had a lower nucleotide diversity than sequence with longer junctions (Figure 2.10). In other words, unrelated sequences with short junctions were more similar on a per nucleotide basis than unrelated sequences with longer junctions. This difference in diversity was not spread evenly across the junction region, but only became apparent after the first ~ 7 nt of the junction region, which are generally derived directly from the V gene segment (Giudicelli *et al.*, 2005). As expected, this was in contrast to nucleotide diversity within clones, which was low across all junction lengths (Figure 2.10). Overall, these results show that sequences with shorter junctions have a lower diversity than expected (given their length), making it difficult to separate clonally related and unrelated sequences.

Although sequences with longer junctions can be grouped into clones with relatively high sensitivity and PPV, false positive assignments are still present. One reason underlying these errors is the use of the IGHJ6 gene, which is over-represented in false positives with junctions at least 30 nt in length ($p < 10^{-3}$, Chi-squared test). The IGHJ6 gene extends an extra ten nucleotides into the junction region relative to all other IGHJ genes (Giudicelli *et al.*, 2005) and clones that use this J gene would thus be more similar to each other than clones using other J genes.

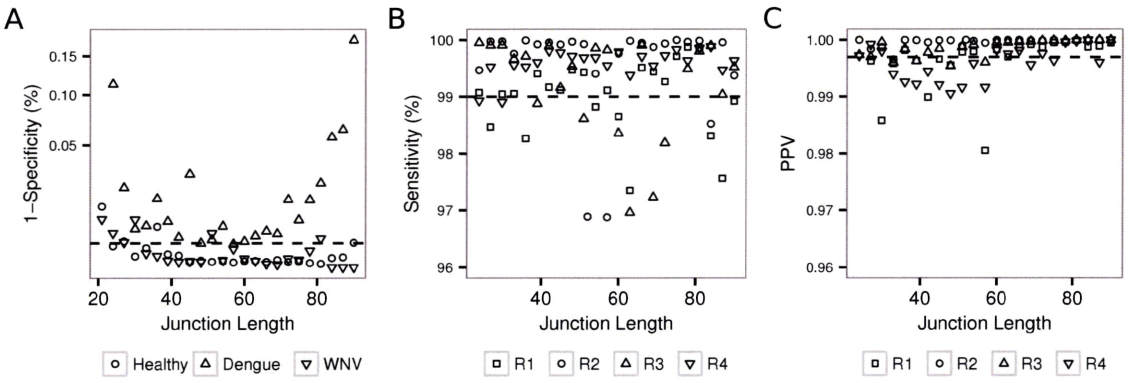

Figure 2.7: **PPV is decreased among sequences with smaller junction lengths.** Single linkage hierarchical clustering was used to identify clonally-related sequences using length-normalized nucleotide Hamming distance. (A) Specificity was calculated using three human experimental data sets (Healthy, Dengue and WNV). Sensitivity (B) and PPV (C) were calculated using 40 simulated datasets based on four experimentally observed sets of clonal lineage structures (R1-R4). Horizontal dashed lines are shown at an arbitrary value in each panel to highlight trends.

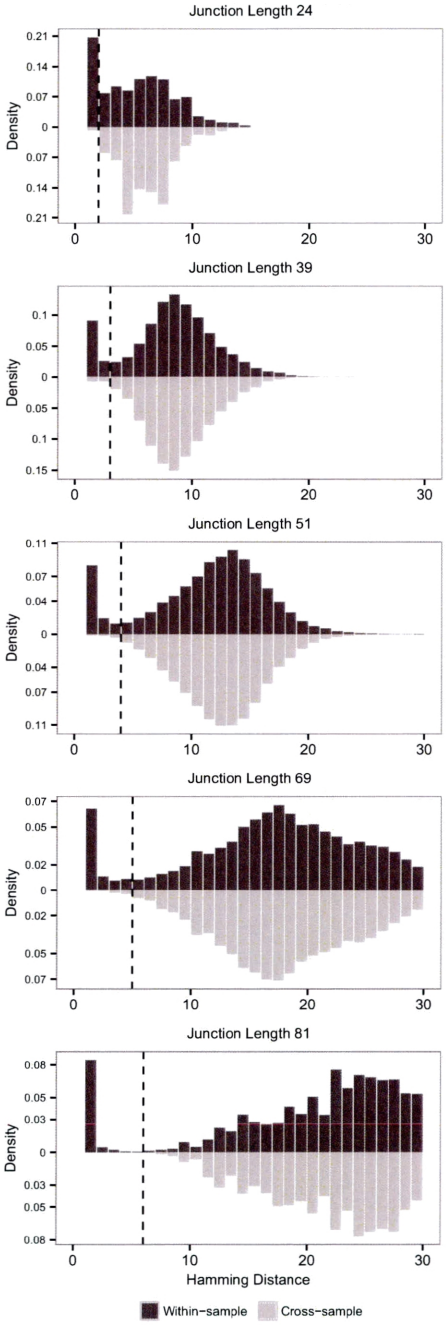

Figure 2.8: **Peaks in the "distance-to-nearest" neighbor distribution begin to converge at small junction lengths.** Nucleotide Hamming distance to nearest neighbor distributions were calculated for sequences with junctions of length 24, 39, 51, 69, and 81 nucleotides from the Healthy experimental dataset. Nearest neighbors were defined using sequences within the same subject (dark grey bars), or by using sequences from all other subjects (light grey bars). The single distance threshold inferred using length-normalized Hamming distance on the entire Healthy dataset was multiplied by the corresponding junction length and shown by the dashed line in each distribution.

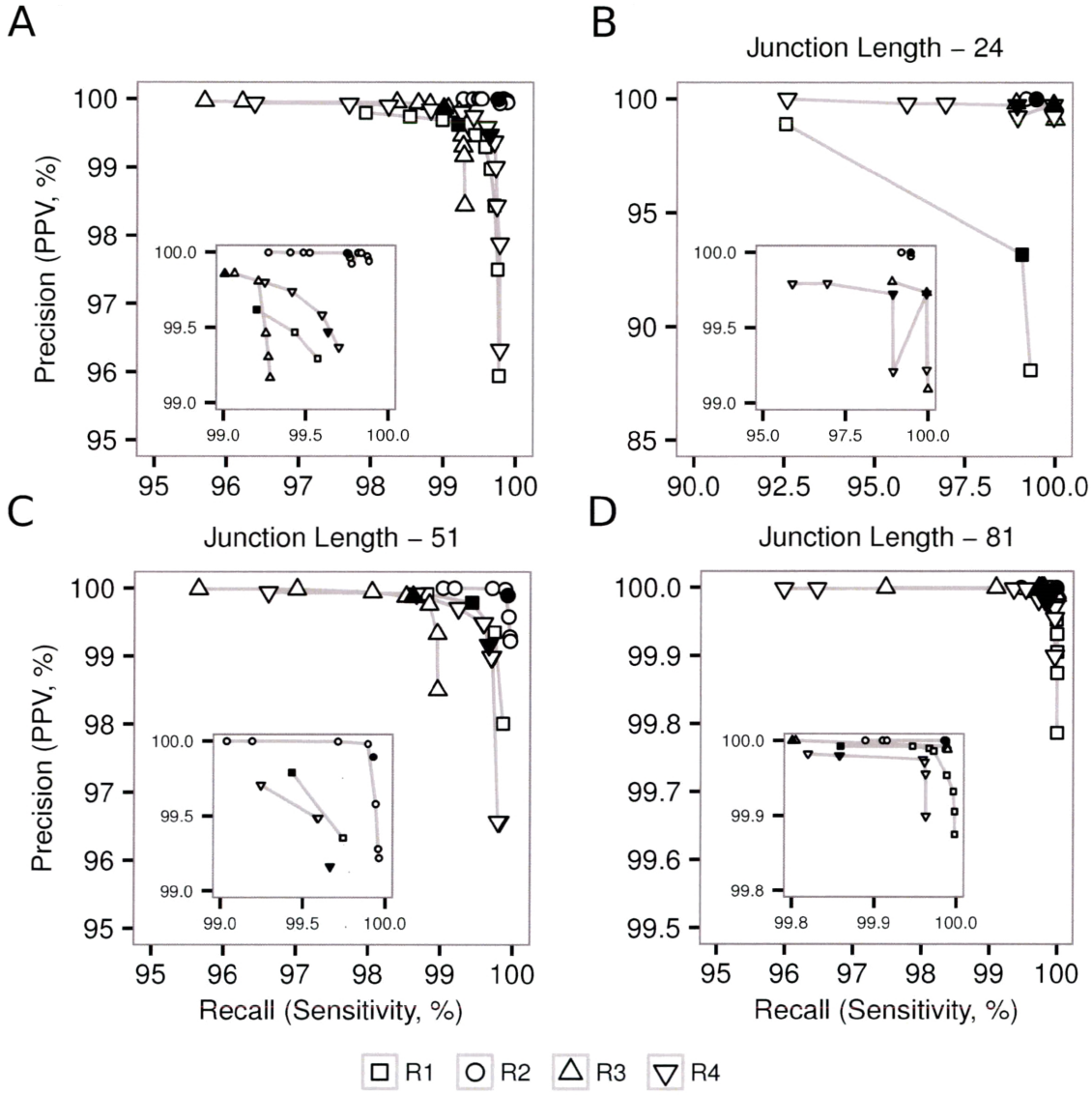

Figure 2.9: **A single distance threshold is near-optimal for all junction lengths.** Single linkage hierarchical clustering with length-normalized nucleotide Hamming distance was used to identify clonally-related sequences in 40 simulated datasets based on four experimentally observed sets of clonal lineage structures (R1-R4). Precision-recall curves were generated by varying the distance threshold from 0.10 to 0.20 at intervals of 0.01. The precision (PPV) and recall (sensitivity) of each run were averaged across the ten simulations of each repertoire in (A) sequences of all lengths, and sequences with junctions of length (B) 24, (C) 51 and (D) 81 nucleotides. The performance of the algorithm run with the inferred threshold is shown by filled points in all panels. Insets show the same data zoomed in on the upper right of the plot.

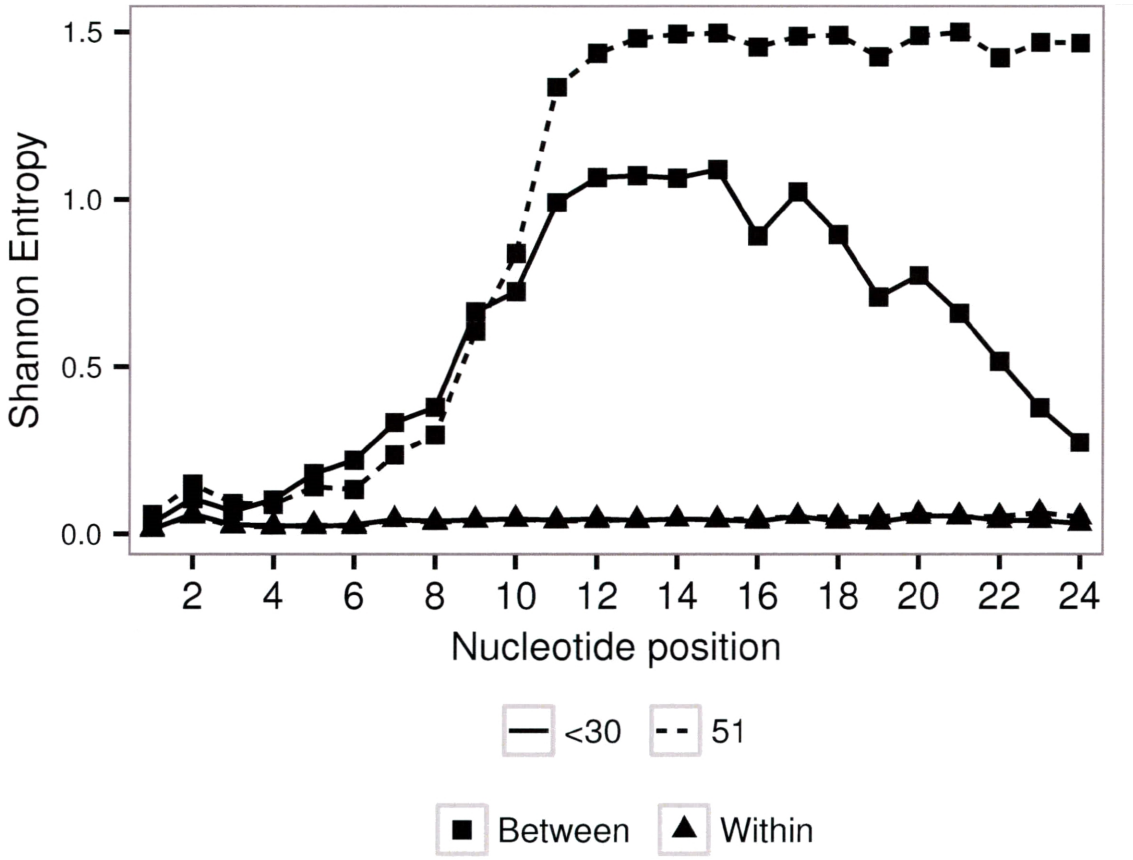

Figure 2.10: **Unrelated sequences with shorter junctions have lower entropy per nucleotide.** Shannon entropy was calculated for each of the first 24 junction nucleotides of clonally related (triangle) and unrelated (square) sequences with junctions <30 nt in length (solid lines) and 51 nt in length (dashed lines). Error bars are standard error of the mean entropy across all sequences in the given group.

2.4 Discussion

AIRR-Seq enables large-scale characterization of the Ig repertoire with the potential for significant basic science and clinical insights. Effective population-level analysis of these data often relies on first identifying groups of clonally related sequences. While hierarchal clustering-based approaches are widely applied in current studies, estimates of their performance and the tradeoffs inherent in the choice of distance or linkage method are lacking. In this study, we carry out an in-depth comparison of hierarchical clustering-based clonal grouping algorithms using an automated analysis pipeline, along with experimental and simulated validation datasets. The analysis pipeline has three stages. First, sequences are separated by V gene, J gene, and junction length. Second, sequences in these groups are assembled into a hierarchy as defined by the distance metric and linkage method. Finally, the hierarchy is partitioned into discrete clones at a fixed distance threshold. While previous applications of this framework relied on a manual process to choose the distance threshold, we minimized human imprecision by developing an automated method to select a customized threshold for any dataset based on analysis of the "distance-to-nearest" distribution.

Quantitative evaluation of the clonal grouping methods was based on a combination of human experimental and simulated data using the common performance measures of specificity, sensitivity and positive predictive value (PPV) (Ralph & Matsen, 2016; Chen *et al.*, 2010; Briney *et al.*, 2016). Experimental data was used to estimate specificity based on the fact that, by definition, B cell clones cannot span different individuals. Sensitivity and related measures like PPV cannot be estimated from human experimental data, since current approaches do not allow unequivocal identification of members of a clone. However, some murine model systems now allow identification of individual clones through Brainbow color labelling of individual B cells prior to affinity maturation (Tas *et al.*, 2016). In this study, we used simulated data — where all clonal relationships are known explicitly — to estimate sensitivity and PPV along with specificity. Simulations have previously been used to validate performance of a probabilistic clonal grouping algorithm (Ralph & Matsen, 2016) and to benchmark other repertoire analysis tools (Safonova *et al.*, 2015; Yaari, Uduman, *et al.*, 2012). Unlike previous approaches, our validation framework does not rely on an underlying model of clonal expansion and affinity maturation. Rather, the lineage tree topologies are taken from experimental datasets and are overlaid with new root sequences and somatic mutation patterns to more closely mimic observed repertoire structures. Furthermore, sensitivity and PPV were calculated on the dataset as a whole, which is less biased by clone size than the per-read averages calculated in

previous studies (Ralph & Matsen, 2016).

Hierarchical clustering methods relate sequences to each other, but do not split sequences into discrete clusters. Thus, a critical step in clonal grouping based on hierarchical clustering is determining the threshold used to partition the sequences into clusters (each representing a single clone). Previous approaches used a fixed threshold such as 80% nucleotide similarity (Jiang, J. a. Weinstein, *et al.*, 2011), or relied on manual inspection of the "distance-to-neares" plot to generate a study-specific threshold (Glanville, Kuo, *et al.*, 2011). "Distance-to-nearest" plots are generally bimodal, with the two peaks interpreted as clonally related (small distance peak) and unrelated (larger distance peak) sequences. Previously the clustering threshold has been determined manually by looking for the distance that best separates these two peaks. However, this process is time-consuming, can be subjective, and there are often multiple possible thresholds that provide equivalent separation between the peaks. To minimize human bias and to enable rapid evaluation of a range of parameter choices for this study, we developed an automated method to mimic the manual approach. Other general methods that have been used for determining the number of clusters in other types of data include the silhouette (Rousseeuw, 1987) or v-fold cross validation (Statsoft, 2013), but these require many rounds of clustering for optimization and are computationally intractable for the large size of AIRR-Seq datasets. The gap statistic (Tibshirani *et al.*, 2001) is also not applicable since it requires a null distribution of expected within-cluster dispersion, which is unknown and would require several assumptions to simulate for Ig sequences. Thus, the automated threshold inference based on the "distance-to-nearest" plot proved most feasible for the data type and is supported by biological intuition.

Hierarchical clustering is an agglomerative (or "bottom up") method. Each sequence starts as its own cluster, and the closest pair of clusters is merged together until all sequences are connected. Closeness is defined by a distance metric. Many previous studies used Hamming distance (Jiang, J. a. Weinstein, *et al.*, 2011; Jiang, He, *et al.*, 2013), which simply counts the number of differences between two junction sequences. Others attempted to incorporate the intrinsic biases of somatic hypermutation to account for the presence of hot- and cold-spots (Stern *et al.*, 2014; Tsioris *et al.*, 2015). Here we found that incorporating the targeting and substitution biases of SHM into the distance metric did not significantly improve performance compared to nucleotide Hamming distance. It is possible that more sophisticated distance measures could play a more important role under conditions different from those investigated here. For example, when the mutation

frequency is low, a different metric may better capture the importance of each individual mutation in determining the separation between clones. However, the current results suggest that the additional assumptions and computational cost of more complex distance metrics are unlikely to provide substantial performance improvements.

While distance is measured between pairs of individual sequences, the linkage method defines how to calculate the closeness between clusters that contain multiple sequences. We evaluated the tradeoffs in the most common linkage methods: single, average and complete. Single linkage is generally considered to be the most inclusive and we found that it provides the best overall performance with specificity, sensitivity, and PPV all over 99%. However, the appropriate choice of linkage may depend on the biological question being addressed. Complete linkage offers a higher PPV, but at the cost of a significant loss of sensitivity. This may be appropriate for research questions that are highly dependent on the accuracy of calling sequences as part of the same clone. For example, studies that attempt to link small numbers of antigen-specific sequences with clonal relatives or establish migration patterns between compartments with infrequent overlaps may benefit from the high confidence in each clonal connection provided by complete linkage. Nevertheless, the high absolute performance of single linkage should be acceptable for most studies.

The specificity, sensitivity, and PPV of single linkage clustering with Hamming distance are all over 99%. However, the errors that are made by this algorithm are not random. We found that Ig sequences incorrectly grouped together as clonally-related had disproportionately short junction regions (here defined as less than 30 nt). Since the V gene extends into the junction region by approximately seven nucleotides (Giudicelli *et al.*, 2005), a higher fraction of the nucleotides in short junctions would be expected to have limited diversity compared with longer junctions, potentially limiting the ability to distinguish between clones. This could be particularly problematic when using a length normalized distance metric, which we showed was critical to achieve acceptable specificity. However, our analysis showed that the problem went beyond the V segment constituting a higher fraction of junction nucleotides. Clonally unrelated sequences with short junctions have less entropy on a per nucleotide basis compared to similar sequences with longer junctions. This lack of inter-clonal diversity could be due to a lower mutation frequency, the use of a restricted set of D genes, or fewer untemplated nucleotide additions between the germline gene segments. However, the entropy of clonally related sequences was comparable between short and longer junctions, suggesting that a uniformly lower mutation frequency is not responsible for the lower diversity in short junctions.

Current algorithms for inferring germline gene segments still struggle with inference of the D gene (Munshaw & Kepler, 2010), making it difficult to determine if the underlying cause of low diversity is due to D gene usage bias, fewer untemplated nucleotide additions, or another mechanism.

Despite the diversity differences between shorter and longer junctions, using a separate threshold to partition sequences with different junction lengths did not improve performance. As precision-recall curves showed, the single threshold selected by analyzing the entire dataset as a whole almost always optimized the trade-off between sensitivity and PPV for all junction lengths. While a few repertoires did have an alternate threshold with slightly improved performance, these thresholds were not evident from the "distance-to-nearest" distributions. It is possible a method other than hierarchical clustering could better separate clones with shorter junctions, but this would be a minor improvement as absolute performance of the single linkage hierarchical clustering with Hamming distance was high.

False positive clonal assignments still occur among sequences with longer junctions, but these appear to have a different underlying cause. In this case the lack of nucleotide diversity can be explained, at least in part, by an over-representation of the IGHJ6 gene. This gene extends an extra ten nucleotides into the junction region (Giudicelli *et al.*, 2005), causing sequences to appear more similar than others using different J genes. It is possible that a separate analysis of these sequences may improve performance. One possibility for better separating clones that do not have sufficient diversity in the junction region is to require shared mutations in the V or J region, although this would penalize clones that have few mutations overall. Likelihood-based approaches, such as Cloanalyst (Kepler, 2013) or partis (Ralph & Matsen, 2016), may help to increase confidence in clones with short junctions or those using IGHJ6, although these approaches are too computationally intensive to use on full AIRR-Seq data sets. While it has been suggested that partis improves performance relative to hierarchal clustering (Ralph & Matsen, 2016), this study did not use dataset-specific distance thresholds and thus likely dramatically underestimated the performance of the clustering-based method.

The comparative analysis presented here suggests clear tradeoffs in the choice of distance and linkage methods. However, it is possible that different tradeoffs would become apparent in data with different clone size distributions, mutation frequencies, etc. The simulation data used to measure sensitivity and PPV were based on lineage tree topologies drawn from only four underlying repertoires. The simulations also assume that Ig sequences maintain the same junction length during

clonal evolution, an assumption that was also made in the clustering algorithm. However, recent research indicates that a small percentage of SHM events may lead to changes in junction length within a clone (Yeap *et al.*, 2015). Insertions/deletions may be present in the junction due to sequencing errors, but the inclusion of UMIs followed by computational approaches for sequencing error-correction can reduce this impact (Yaari & Kleinstein, 2015; Shugay *et al.*, 2014). Few clonal grouping methods deal with junction length differences, and while these effects are also not accounted for in the current study, their influence on performance is expected to be small. Another possible source of bias in the performance on experimental data is the potential presence of so-called "public clones," or highly similar sequences across individuals. Such sequences may skew specificity estimates that were approximated on publicly available human experimental datasets based on the frequency of inferred groups that spanned individuals. Furthermore, this specificity measure depends on the frequency of highly similar sequences found across individuals, which may differ from the frequency of highly similar sequences found within an individual by chance. Future studies could benefit from using a larger number of AIRR-Seq datasets that span age, tissue, disease state, etc. in addition to simulations based on a larger number of underlying experimental Ig repertoires.

In summary, computational methods for grouping Ig sequences into B cell clones is a critical part of AIRR-seq studies, and allows for understanding the structure and affinity maturation of the Ig repertoire. Here we developed a framework for comparative analysis of clonal grouping approaches and determined that single linkage hierarchical clustering with length-normalized nucleotide Hamming distance performs well on both human experimental and simulated datasets. This algorithm is available as part of the Change-O and SHazaM packages (Gupta *et al.*, 2015) in our Immcantation tool suite (`http://immcantation.readthedocs.io`).

Chapter 3

Ig AIRR-Seq Analysis Toolkit

This chapter has also been published as:

Gupta, N. T. *et al.* Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* **31,** 3356–3358 (2015)

My contribution to the following project includes writing code in changeo-clt and `shazam` and the extensive documentation and integration for the entire Change-O workflow.

3.1 Introduction

Large-scale characterization of immunoglobulin (Ig) repertoires is now feasible due to dramatic improvements in high-throughput sequencing technology. Repertoire sequencing is a rapidly growing area, with applications including: detection of minimum residual disease, prognosis following transplant, monitoring vaccination responses, identification of neutralizing antibodies, and inferring B cell trafficking (Robins, 2013; Stern *et al.*, 2014). We previously developed the Repertoire Sequencing Toolkit (pRESTO) for producing assembled and error-corrected reads from high-throughput lymphocyte receptor sequencing experiments (Vander Heiden *et al.*, 2014), which may then be fed into existing methods for alignment against V(D)J germline databases (*e.g.*, IMGT/HighV-QUEST (Alamyar *et al.*, 2012), IgBLAST (Ye *et al.*, 2013)). However, extracting measures of biological and clinical interest from the resulting germline annotated repertoire remains a time-consuming and error-prone process that is often dependent upon custom analysis scripts.

Some centralized workflows are available for Ig repertoire analysis, but are limited in flexibility of individual tasks. IGGalaxy (Moorhouse *et al.*, 2014) parses the output of IgBLAST (Ye *et*

al., 2013) and IMGT/HighV-QUEST and provides a visual report summarizing the results, but provides limited downstream analyses. Similarly, ImmunExplorer parses IMGT/HighV-QUEST (Alamyar *et al.*, 2012) output files and has a built-in clonality and diversity calculator with limited parameter options and no further analyses such as mutation calling. ImmuneDiveRsity (Cortina-Ceballos, Godoy-Lozano, Samano-Sanchez, *et al.*, 2015) includes steps from raw read processing to clone identification in a single executable script, but requires using the built-in IgBLAST (Ye *et al.*, 2013) for V(D)J germline inference and also has limited parameter options. Similar to ImmuneDiveRsity (Cortina-Ceballos, Godoy-Lozano, Samano-Sanchez, *et al.*, 2015), LymAnalyzer (Y. Yu *et al.*, 2016) includes several analysis steps from raw read processing to clone identification and lineage trees. However, as with the other pipelines, the user has limited control over the parameters of the underlying algorithms, it is not possible to substitute other algorithms for any single step of the pipeline, and the output file(s) are not formatted in such a way as to easily allow further downstream analyses. Although there is some appeal to having a single script to run a full Ig repertoire analysis workflow, analysis tasks vary widely depending on the biological question and experimental design at hand.

More recently, ImmuneDB (Rosenfeld *et al.*, 2016) was released as an Ig AIRR-Seq analysis framework consisting of Python tools that interact with IG sequences stored in a MySQL database. ImmuneDB (Rosenfeld *et al.*, 2016) does offer modularity, but any other tools that are used must be configured to interact with the MySQL database for input and output. Here, we introduce Change-O, a suite of modular Python and R (R-Core, 2015) based utilities that cover a range of complex analysis tasks for Ig repertoire sequencing data. Change-O offers both modularity and a centralized tab-delimited format for storing Ig sequence information that can easily be viewed with standard spreadsheet applications and read by many programming languages including Python and R (R-Core, 2015).

3.2 Features

The Change-O suite is composed of four software packages: a collection of Python commandline tools (changeo-ctl) and three separate R (R-Core, 2015) packages (`alakazam`, `shazam`, `tigger`) (Table 3.1). Data is passed to Change-O utilities in the form of a tab-delimited text file. Each utility identifies the relevant input data based on standardized column names, and adds new columns to the file with the output information to be carried through to the next analysis step. Change-O provides tools to import data from the frequently used IMGT/HighV-QUEST (Alamyar *et al.*, 2012) tool, as well as a set of utilities to perform basic database operations, such as sorting, filtering, and modifying annotations. The more computationally expensive components have built-in multiprocessing support. Each utility includes detailed help documentation and optional logging to track errors. Example workflow scripts are provided on the website, which can easily be modified by adding, removing, or reordering analysis steps to meet different analysis goals. As detailed below, several repertoire analyses may be carried out, depending on the nature of the study.

3.2.1 Inference of novel alleles and individual genotype

Germline segment assignment tools, such as IMGT/HighV-QUEST, work by aligning each sequence against a database of known alleles. However, this process will fail for sequences that utilize previously undetected alleles. In this case, the sequence will be assigned to the closest known allele and any polymorphisms will be incorrectly identified as somatic mutations. To address this problem, the Change-O suite includes the TIgGER (Tool for Immunoglobulin Genotype Elucidation) method (Gadala-Maria *et al.*, 2015). TIgGER determines the complete set of variable region gene segments carried by an individual and identifies novel alleles, yielding a repertoire of germline alleles personalized to an individual, and adjusts the germline variable region gene assignments based on this individual Ig genotype. This process significantly improves the quality of germline assignments, thus increasing the confidence of downstream analysis dependent upon mutation profiles.

3.2.2 Partitioning sequences into clonally related groups

Identifying sequences that are descended from the same B cell (clonal groups) is important to virtually all Ig repertoire analyses. Clonal group sizes and lineage structures provide information on the underlying response, and clonally related sequences cannot be treated independently in

statistical analyses and models. Although the in-depth evaluation and comparison of hierarchical clustering-based clonal grouping methods were discussed in the previous chapter proposed a high confidence algorithm, Change-O provides several options for partitioning sequences into clones. Users can choose between several published hierarchical clustering-based clonal grouping methods (Ademokun *et al.*, 2011; Chen *et al.*, 2010; Glanville, Zhai, *et al.*, 2009; Stern *et al.*, 2014; Tsioris *et al.*, 2015) with distance metrics including Hamming distance as well as distance that employ several published SHM hot/cold-spot targeting models (D. S. Smith *et al.*, 1996; Yaari, Vander Heiden, *et al.*, 2013; Stern *et al.*, 2014), multiple linkage methods, and a user-defined distance threshold. Users may also specify the region of the sequence to be used to calculate distance between Ig receptors. Change-O also includes tools to help the user determine a dataset-specific distance threshold based on distance patterns in the repertoire (Glanville, Zhai, *et al.*, 2009).

3.2.3 Quantification of repertoire diversity

To assess repertoire diversity, Change-O provides an implementation of the general diversity index (${}^{q}D$) proposed by Hill (Hill, 1973), which encompasses a range of diversity measures as a smooth curve over a single varying parameter q. Special cases of this general index of diversity correspond to the most popular diversity measures: species richness ($q = 0$), the exponential Shannon-Weiner index (as $q \to 1$), the inverse of the Simpson index ($q = 2$), and the reciprocal abundance of the largest clone (as $q \to +\infty$). Resampling strategies are also provided to perform significance tests and allow comparison across samples with varying sequencing depth (Y. C. B. Wu *et al.*, 2014; Stern *et al.*, 2014).

3.2.4 Generation of B cell lineage trees

Lineage trees provide a means to trace the ancestral relationships of cells within a clone. This information has been used to estimate mutation rates (Kleinstein *et al.*, 2003), infer B cell trafficking patterns (Stern *et al.*, 2014), and to trace the accumulation of mutations that drive affinity maturation (Y.-C. B. Wu *et al.*, 2012; Uduman, Shlomchik, *et al.*, 2014). Change-O provides a tool for generating lineage trees using `PHYLIP`'s maximum parsimony algorithm (Felsenstein J., 2005), with modifications to meet the requirements of an Ig lineage tree (Barak *et al.*, 2008; Stern *et al.*, 2014). Trees may be viewed and exported into different file formats using the `igraph` (Csardi, 2006) R package.

3.2.5 Analysis of somatic hypermutation hot/cold-spot motifs

SHM is a process that operates in activated B cells, and introduces point mutations into the DNA coding for the Ig receptor at a very high rate ($\approx 10^{-3}$ per base-pair per division) (McKean *et al.*, 1984; Kleinstein *et al.*, 2003). Accurate background models of SHM are critical, since SHM displays intrinsic hot/cold-spot biases (Yaari, Vander Heiden, *et al.*, 2013). Change-O provides utilities for estimating the mutability and substitution rates of DNA motifs from large-scale Ig sequencing data to construct hot/cold-spot motif models, including the ability to generate models based solely on silent mutations and thus avoid the confounding influence of selection (Yaari, Vander Heiden, *et al.*, 2013). These tools can be used to build models of SHM targeting, and gain insight into the relative contributions of AID and different error-prone repair pathways in SHM.

3.2.6 Analysis of selection pressure

Change-O includes an implementation of the BASELINe (Yaari, Uduman, *et al.*, 2012) method for quantifying selection pressure in Ig sequences. BASELINe quantifies deviations in the frequency of replacement mutations compared to a background model of SHM. To quantify selection, users may use published background models (D. S. Smith *et al.*, 1996; Yaari, Vander Heiden, *et al.*, 2013), or infer the background from their own data using the SHM model building tools in Change-O.

Table 3.1: **Summary of Change-O features.**

Package	**Analysis Tasks**
changeo-clt	Parsing of V(D)J assignment output
	Basic database manipulation
	Multiple alignment of sequence records
	Assignment of sequences into clonal groups
	Calculation of CDR3 physiochemical properties
alakazam	Clonal diversity analysis
	Lineage reconstruction
shazam	SHM hot/cold-spot modeling
	Quantification of selection pressure
tigger	Inference of novel germline alleles
	Construction of personalized germline repertoires

3.3 Conclusion

Change-O is a suite of utilities implementing a wide range of B cell repertoire analysis methods. Together these tools allow researchers to quickly implement advanced analysis pipelines for analyzing large data sets generated by repertoire sequencing experiments. A simple tab-delimited file with standardized column names allows for communication between the utilities, and can easily be viewed using standard spreadsheet applications. This format also allows research groups the flexibility to incorporate other analysis tools into their in-house analysis pipelines by simply adding additional columns of information to the central file. Change-O, along with pRESTO (Vander Heiden *et al.*, 2014), provide key components of an analytical ecosystem that enables sophisticated analysis of high-throughput Ig repertoire sequencing datasets. Both are available with extensive documentation as part of the Immcantation framework (`immcantation.readthedocs.io`)

Chapter 4

Applications of clonal grouping to WNV infection and other diseases

This chapter has been adapted from the following publications:

- My contribution to Section 4.1 includes all of the computational analysis of the Ig sequences obtained from both single-cell and next-generation sequencing from:

  Tsioris, K., Gupta, N. T. *et al.* Neutralizing antibodies against West Nile virus identified directly from human B cells by single-cell analysis and next generation sequencing. *Integrative Biology* **7**, 1587-1597 (2015)

- My contribution to Section 4.2 consists of computational analysis of the Ig sequences obtained from laser capture and microdissection from:

  Di Niro, R., Lee, S.-J., *et al.* Salmonella Infection Drives Promiscuous B Cell Activation Followed by Extrafollicular Affinity Maturation. *Immunity* **43,** 120–131 (2015)

- My contribution to Section 4.3 consists of computational analysis of the Ig sequences obtained from laser capture and microdissection from:

  Di Niro, R., Mesin, L., *et al.* High abundance of plasma cells secreting transglutaminase 2-specific IgA autoantibodies with limited somatic hypermutation in celiac disease intestinal lesions. *Nature Medicine* **18** (ed Nat) 441–5 (2012)

4.1 West Nile virus

4.1.1 Introduction

West Nile virus (WNV) is a mosquito-borne, enveloped positive-strand RNA virus that can lead to severe neurological disease. The virus belongs to the family flaviviridae, which includes yellow fever, hepatitis C virus, and dengue viruses (Suthar *et al.*, 2013; Colpitts *et al.*, 2012). The emergence of WNV in North America was first documented in 1999 in New York, USA. WNV has now become established throughout the USA and has spread into Canada, Mexico, and the Caribbean. Reports from the CDC indicate infection of more than 41,000 people to date, including more than 1,700 fatalities (Montgomery & K. O. Murray, 2015). The cumulative incidence of WNV infection may reach 3 million people (Colpitts *et al.*, 2012; Petersen *et al.*, 2012). Currently, there is no approved treatment or vaccine against WNV (Suthar *et al.*, 2013; Colpitts *et al.*, 2012; Kyle Austin & Dowd, 2014).

Passive administration of neutralizing antibodies (NAbs) is one possible route to treat viral infections (Marasco & Sui, 2007; Klein *et al.*, 2013) and could have therapeutic value in the context of severe flavivirus infection (de Alwis & de Silva, 2014; de Alwis, S. a. Smith, *et al.*, 2012; Dowd & Pierson, 2011; de Jong *et al.*, 2014). WNV-specific NAbs have been derived by phage and yeast display libraries from both humans and mice (Oliphant, Engle, *et al.*, 2005; Pierson *et al.*, 2007; Throsby *et al.*, 2006; Vogt *et al.*, 2009; Gould *et al.*, 2005). Murine antibodies have shown potency in mouse models, but the epitopes targeted by this class of antibodies comprise only a minor component of the neutralizing response in humans. These antibodies have limited utility as therapeutics to date (Oliphant, Nybakken, *et al.*, 2007). Other disadvantages of library-based methods to derive NAbs are the random pairing of heavy- and light-chains, which obscures the natural humoral response, involves time-consuming assays, and identifies antibodies with limited neutralization function (Hammers & Stanley, 2014). In contrast, recent studies have identified a large number of highly potent human immunodeficiency virus-1 (HIV-1) specific NAbs directly derived from HIV-infected patients using flow cytometry to sort memory B cells (MBCs) based on their affinity to HIV antigens (Scheid, Mouquet, Feldhahn, *et al.*, 2009; Scheid, Mouquet, Ueberheide, *et al.*, 2011). For potential vaccine strategies, WNV-specific NAbs directly derived from humans could also reveal information about naturally targeted epitopes on the virus. NAbs, however, are only a part of the humoral immune response to WNV. For a better understanding of

antibody-mediated mechanisms involved, such as disease outcome or persistence of antibodies and virus (K. Murray *et al.*, 2010), our analysis of the humoral response should span from single cells to the level of the antibody repertoire.

In contrast to HIV or influenza, the prevalence of WNV is low and many cases are undiagnosed, making it challenging to assemble a large WNV cohort. Despite these obstacles, we effectively discovered and evaluated four novel NAbs against WNV directly from rare WNV-specific human B cells isolated from a set of recently infected and post-convalescent subjects. For this purpose, we combined microengraving (Love *et al.*, 2006; Ogunniyi, Thomas, *et al.*, 2014; Ogunniyi, Story, *et al.*, 2009), an integrated multiparameter single-cell analysis method and next-generation sequencing (NGS). We analyzed activated memory B cells (MBCs) and antibody-secreting cells (ASCs) from blood on a single-cell level and evaluated the relationship of the parental WNV-specific clones within the circulating repertoire of B cells (Figure 4.1). Despite a low frequency of WNV-specific B cells (mean <24 Ag– events per 100,000 PBMCs), these methods for integrated analysis allowed us to obtain NAbs and enabled analysis of the humoral response to WNV on a single-cell and repertoire level. The single-cell analysis revealed rare but persistent WNV-specific MBCs and ASCs in post-convalescent subjects. Furthermore, the results presented here indicate that the antibody response is independent of an asymptomatic vs. symptomatic disease outcome, as we have noted previously (Qian *et al.*, 2014). Using the nucleotide coding sequences for WNV specific antibodies discovered from single cells, we also revealed expanded WNV specific clones in the repertoires of recently infected subjects through NGS and bioinformatic analysis.

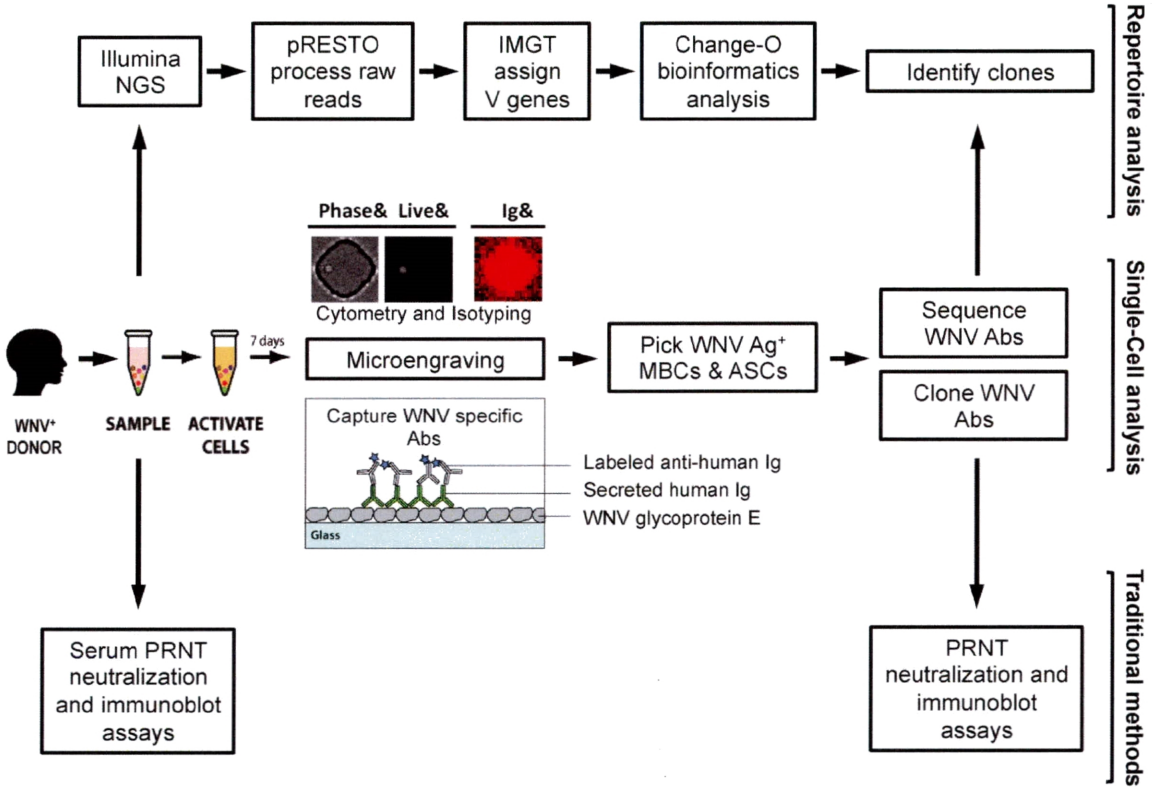

Figure 4.1: **Analysis workflow.** Blood is collected from individuals with a history of infection with WNV and PBMCs are isolated for further analysis. In parallel subject serum is analyzed using neutralization and immunoblot assays to determine the individual's overall response to WNV. PBMC samples are analyzed by microengraving, an integrative single-cell analysis process. WNV-specific antibodies are captured on a protein microarray from single antibody-secreting cells (ASCs) and single stimulated memory B cells (MBCs). Subsequently, individual WNV-specific MBCs and ASCs identified by microengraving are recovered and the sequences coding the variable region of the antibody heavy and light chains of their corresponding antibodies are obtained. Paired heavy and light chain coding sequences are used to clone the WNV-specific antibodies. The cloned antibodies are evaluated by neutralization and immunoblot assay. In parallel, the subject antibody repertoire is analyzed by next generation sequencing (NGS). The NGS data is processed by an integrated bioinformatics pipeline (pRESTO and Change-O) to identify clones of WNV-specific antibodies.

4.1.2 Materials and Methods

4.1.2.1 Recruitment of human subjects and sample validation

Human subjects exhibiting various WNV infection phenotypes were recruited from a well-characterized cohort of volunteers. Blood was obtained with written informed consent under the guidelines of the Human Investigations Committee of Baylor College of Medicine, which approved the study. Investigation of coded samples was approved by Yale University and Massachusetts Institute of Technology in compliance with HIPAA guidelines. The diagnosis of WNV infection was determined following CDC guidelines (Nolan *et al.*, 2013) and subjects were stratified by CDC definitions to a severe neuroinvasive phenotype, mild fever-only phenotype, or asymptomatic infection as described previously (Nolan *et al.*, 2013). Infection was validated by qualitative rapid nucleic acid test at the blood bank (cobas®TaqScreen West Nile virus Test, Roche Molecular Systems, Pleasanton, CA), positive immunoblot as described previously (Qian *et al.*, 2014), or IgM ELISAs. Subjects (n = 11) were 45.5% female and 90.9% white and included asymptomatic, mild, and severe subjects. Subjects from the asymptomatic, mild, and severely infected groups were not statistically different for age, gender, or race in this study.

4.1.2.2 Single-cell analysis by microengraving

Actively secreting and memory B cell populations in PBMC samples obtained from subjects were analyzed by microengraving (Love *et al.*, 2006; Ogunniyi, Thomas, *et al.*, 2014) to determine the distributions of secreted antibody isotypes and the frequencies of WNV E specific antibodies. There was a range of cells from our donors with a median of 149[thin space (1/6-em)]090 antibody secreting cells (range 51[thin space (1/6-em)]888-375[thin space (1/6-em)]028) and a median of 145[thin space (1/6-em)]601 memory B cells (range 56[thin space (1/6-em)]862-385[thin space (1/6-em)]436). Frozen PBMCs were thawed and maintained in complete RPMI media containing IL-6 (20 ng/mL, Peprotech) and Chk2 inhibitor II (2 μg/mL, Sigma-Aldrich). After resting or stimulation, cells were stained for viability (Calcein violet AM) and for expression of CD19 (Brilliant Violet® 605), CD20 (Alexa Fluor® 488), CD27 (PerCP-efluor® 710), CD38 (PerCP-efluor® 710) and CD138 (APC), loaded into nanowells, then imaged on an epifluorescence microscope. Separate microarrays were generated by sealing nanowells with capture slides functionalized with donkey anti-human Ig (25 μg/mL, Jackson Immunoresearch) or WNV E protein (50 μg/mL, Aviva Systems Biology). We used a flu specific antibody secreting cell line (X. Yu *et al.*, 2008) to determine non-

specific binding events for subsequent background frequency subtraction (Ogunniyi, Thomas, *et al.*, 2014). Isotype-specific information was obtained by staining microarrays with a panel of anti-human IgG1, IgA, IgG3 and IgM detection antibodies (BD Biosciences and Invitrogen). Microarrays from WNV E-coated slides were probed with mouse anti-human Igκ and Igλ (BD Biosciences) detection antibodies. Microarrays were scanned on a GenePix® 4200AL Autoloader (Molecular Devices), and median fluorescence intensities (MFIs) from bound species were extracted with GenePix® Pro 6.0 software. Isotype-specific data was analyzed as previously described (Ogunniyi, Thomas, *et al.*, 2014). Antibodies were considered WNV-E specific if microarray elements were from wells with viable, CD19+/CD20+/- cells that also satisfied the following criteria: % saturation $\leq$ 2, signal-to-noise ratio $\geq$ 1, $\geq$ 40% of pixels above background +1 standard deviation, CV $\leq$ 70 and background corrected MFI $\geq$ 1500. A fraction of cells from such wells were recovered with an automated micromanipulator (AVISO CellCellector™, ALS GmbH), and heavy and light chain variable genes were amplified by single-cell RT-PCR (Ogunniyi, Thomas, *et al.*, 2014). In short, the cells were recovered using a glass capillary mounted on an automated micromanipulator and a microscope. The glass capillary was placed over the nanowell to recover the desired MBC or ASC by applying a vacuum. The cell was subsequently released into a well of a 96 well plate containing water. To subsequently recover antibody heavy and light chain variable genes, a nested PCR was performed using the primers and protocol previously published by X. Wang & Stollar, 2000. These procedures have been described in more detail by Ogunniyi, Thomas, *et al.*, 2014.

4.1.2.3 Expression and validation of WNV specific antibodies

The protocols to clone, express and validate antigen specific antibodies has been previously described (Ogunniyi, Thomas, *et al.*, 2014; Tiller, Meffre, *et al.*, 2008). In brief, paired heavy and light chain coding sequences were cloned into vectors for expression as human IgG1. We utilized human embryonic kidney (HEK) 293 T cells (ATCC) to transiently express the cloned antibodies. Antibodies in culture supernatants were evaluated for binding on a custom protein microarray consisting of WNV E protein. Microarrays were scanned on GenePix® 4200AL and analyzed with GenePix® Pro 6.0. Subsequently, the expressed WNV specific antibodies were evaluated by immunoblot and PRNT assay as described above.

4.1.2.4 Library preparation and next generation sequencing

To evaluate B cell subsets from recently infected subjects listed, fresh PBMCs were sorted to isolate plasma cells, naive and memory B cells, and RNA was isolated for sequencing. The number of sorted B cells from each subject was a median of 536[thin space (1/6-em)]792 (range 51[thin space (1/6-em)]232-2[thin space (1/6-em)]108[thin space (1/6-em)]051). UID barcoded NGS libraries (Vander Heiden *et al.*, 2014) were prepared by AbVitro, Inc from 250 ng of the extracted RNA of both sorted B cells as well as unsorted PBMCs. NGS libraries were subsequently sequenced at the MIT BioMicrocenter on the Illumina Miseq platform using the Illumina 2×300 bp sequencing kit according to the manufacturer specifications. PBMCs from post-convalescent subjects were directly processed without sorting, and libraries were sequenced using the Illumina 2×150 bp sequencing kit.

4.1.2.5 High-throughput antibody repertoire sequence analysis

Preprocessing was carried out using pRESTO version 0.4 (http://clip.med.yale.edu/pRESTO) (Vander Heiden *et al.*, 2014) and involved removing primer sequences, filtering based on sequence quality and annotating sequences with sample information. Following preprocessing, V(D)J germline segments were determined using IMGT/HighV-QUEST (Alamyar *et al.*, 2012). Functional V(D)J sequences of each subject were combined with WNV-specific sequences and then Change-O (Gupta *et al.*, 2015) was used to divide the sequences into clonally-related groups by a two-step procedure: (1) sequences were partitioned based on common V gene, J gene, and junction region length, and (2) within these larger groups, sequences differing from one another by a distance of less than 0.01 were defined as clones. Distance was measured as the number of point mutations weighted by a nucleotide substitution probability previously described (Yaari, Vander Heiden, *et al.*, 2013). These data will be available through controlled access from the database of Genotypes and Phenotypes (dbGaP).

Clones for antibodies MIT185 and MIT186 as portrayed in Figure 4.6 were identified by manual inspection of Figure 4.6. To generate Figure 4.6, for each WNV-specific query sequence, the NGS data was filtered for sequences with the same V and J germline gene segment. For each of these sequences, similarity to the query (y-axis) is the Needleman-Wunsch distance on the entire sequence without end-gap penalties divided by the length of the alignment. The distance from germline (x-axis) is the Hamming distance (with ambiguous characters not counting as a mismatch) between

the sequence and its germline divided by the number of non-N nucleotides in the sequence.

The smallest observed WNV-specific clone composed 0.01% of the repertoire. Selection pressures were quantified using BASELINe (Yaari, Uduman, *et al.,* 2012; Uduman, Yaari, *et al.,* 2011) with the S5F somatic hypermutation targeting model (Yaari, Vander Heiden, *et al.,* 2013).

4.1.2.6 Generation and analysis of lineage trees

Lineage trees were constructed for each clonal group by removing indel positions and then using maximum parsimony with the dnapars application of PHYLIP (Felsenstein J., 2005). This was followed by recursively replacing inferred ancestors in each tree with descendants having a Hamming distance of zero from their inferred parent. The trunk length of a clone is defined to be the branch length from the germline root node to the most recent common ancestor. Comparisons were made to Ig repertoire sequencing of three individuals who received an influenza vaccination and had blood drawn at several time-points (8 days, 2 days, and 1 hour pre-vaccination; 1 hour, 1 day, 3 days, 7 days, 2 weeks, 3 weeks, and 4 weeks post-vaccination). Clones were defined in the same way as described above. For the vaccination data, trunk lengths were only measured for expanded clones, or clones that at some time-point compose at least 0.05% of the Ig repertoire. Comparisons between trunk lengths were made using a two-tailed Student's t-test.

4.1.3 Results

4.1.3.1 Identification of study subjects with high serum neutralizing antibody titers

To examine developing humoral responses to WNV, we recruited a cohort of recently infected subjects during the 2012 WNV outbreak in Houston, Texas (Table 4.1). Blood samples were collected at two time points, 1.9 ± 0.8 months post onset (range 0.8-2.9; n=7) and 3.2 ± 1.2 months (range 2-4.7; n=6, one subject lost to follow-up) after infection. Both time points correspond to samples of the recently infected period following exposure to WNV. To identify post-convalescent subjects with high serum neutralizing antibody titers, we screened our cohort of more than 160 WNV subjects who we have monitored since 2002 (K. O. Murray *et al.*, 2013). We initially selected individuals from asymptomatic and severe infection groups with a history of exposure less than 2 years after infection with WNV (total n=17, enrolled n=4). In total, we identified 11 subjects from our cohort of varying age, gender, disease outcome, and disease stage for our study.

4.1.3.2 West Nile virus neutralizing antibodies identified by single-cell analysis

Using microengraving (Ogunniyi, Thomas, *et al.*, 2014; Ogunniyi, Story, *et al.*, 2009) (Figure 4.1), we recovered more than 90 WNV-specific single MBCs and ASCs, 31 of which had both heavy (Table 4.2)and light chains (Table 4.3). From those cells, we recovered 19 antibody-coding sequences where both heavy (V1, V3, V4 V region family) and light (V1, V2, V3 V region family) chains were present after PCR amplification. These 19 paired antibody-coding sequences were used for cloning and further evaluation (Table 4.4). We found no significant difference ($p<0.05$) in the amount of selection pressure between MBCs and ASCs by analyzing the mutations in the complementarity-determining (CDR) and framework (FWR) regions (Yaari, Uduman, *et al.*, 2012). Of the 19 recovered antibodies, twelve expressed with sufficient amounts of protein for further evaluation by immunoblot, plaque reduction neutralization assay (PRNT), and protein binding assays to determine their relative affinities (Table 4.5). As controls, we used serum from WNV positive (HS+) and WNV negative (HS-) subjects, and an antibody (2G12) with known affinity to HIV-1. Of the WNV-specific antibodies, five were from post-convalescent and seven from recently infected subjects. Three expressed antibodies showed reactivity to WNV envelope (E) protein by immunoblot, and four also showed efficient neutralization of WNV. Interestingly, three of the neutralizing antibodies did not show reactivity to WNV E protein by immunoblot, suggesting that they recognize conformation-dependent epitopes. Antibody MIT89 was positive both by immunoblot and PRNT.

Table 4.1: **Clinical data for West Nile virus-infected subjects.**

ID	Age at onset/ blood draw*	Disease status	Diagnosis	WNV E IgG immunoblot	PRNT serum titer
PC1	48*	Post-conv.	WNM	+++++	1:80
PC2	44*	Post-conv.	Asymp	++++	1:640
PC3	26*	Post-conv.	WNF	+++	1:80
PC4	21*	Post-conv.	WNF	+++++	1:80
RI1	67	Recent	WNE	++	1:80
RI2	30	Recent	Asymp	+++++	1:40
RI3	64	Recent	WNE	+++++	1:40
RI4	65	Recent	Asymp	+++	1:40
RI5	37	Recent	WNE	+++++	1:40
RI6	31	Recent	WNF	+++++	1:80
RI7	59	Recent	Asymp	+++++	1:40

* Age of the donor at the time of blood draw. WNE, encephalitis; WNF, febrile illness; WNM, meningitis; Asymp, asymptomatic.
PRNT indicates the dilution of serum needed to achieve 90% reduction of WNV growth in Vero cells.
Immunoblot scale maximum is five +++++, corresponding to the reactivity with a known positive WNV patient serum.

Furthermore, we observed a positive trend between the relative antibody affinities for WNV E protein and WNV neutralization. All neutralizing antibodies were recovered exclusively from post-convalescent subjects.

4.1.3.3 Next generation sequencing of B cell repertoires reveals clonal expansion in individuals recently infected with West Nile virus

To gain a better understanding of the overall variance among the B cells in circulation, we sequenced the B cell heavy chains from PBMCs, naive, and memory B cell populations of seven recently infected subjects (Table 4.6). The distributions of isotypes determined by NGS of PBMCs and MBCs showed no significant differences between asymptomatic and symptomatic recently infected subjects, confirming the results measured from individual cells by microengraving (Figure 4.2). The distribution of clone sizes in the sequencing data are shown for each subject in Figure 4.3. The degree of mutation in the sequences from recently infected subjects is shown in Figure 4.4. To identify WNV-specific clones in the repertoire, the individual WNV-specific antibody sequences obtained by the single-cell analysis ("queries") were combined with the NGS data and groups of clonally-related sequences were automatically identified using Change-O as described in methods (Gupta *et al.*, 2015). This method identified three expanded clones that included one of the query sequences. To identify additional WNV-specific clones, the sequences from the NGS were plotted based on their level of mutation and similarity to each of the query antibody sequences (Figures 4.5 & 4.6). The three WNV-specific clones were clearly identifiable as outlier groups of sequences with higher similarity to the queried WNV-specific antibody sequences compared to other sequences with a similar level of mutation (Figures 4.5B & 4.6). In addition, manual inspection of these plots identified two additional WNV-specific clones. The WNV-specific clones showed evidence of expansion, and sequences were found both among PBMCs and memory B cells with both IgA and IgG isotypes. Overall, five WNV-specific clones were identified through integration of the single-cell analysis and NGS data.

Since WNV is a recent pathogen to North America, the WNV-specific clones we identified are likely a result of a primary immune response originating from naive B cells, rather than a recall response from memory B cells. Therefore, we hypothesized that the WNV-specific clonal expansion we observed would have initiated from less affinity-matured cells than those that generate antigen-specific clones after a recall response to re-occurring infections or vaccinations to viruses such as

Table 4.2: **West Nile virus specific antibody heavy chains from single-cell analysis.**

			Heavy Chain					
Clone No.	**Subject**	**Well ID**	**V Region**	**D Region**	**J Region**	**No. of Mismatches**	**CDR3**	**CDR3 Length**
MIT 86	PC3	1363 0301	IGHV2-26*01	IGHD3-22*01	IGHJ4*02	0	CARVNSSGYYFDYW	14
MIT 87	PC3	2167 0307	IGHV4-59*01	IGHD3-22*01	IGHJ5*02	0	CARGYYDSTRDWFDPW	16
	PC4	0324 0603	IGHV2-26*01	IGHD2-2*02	IGHJ6*02	14	CARAPSEYAAMDVW	14
MIT 88	PC4	1561 0703	IGHV3-64*05	IGHD3-3*01	IGHJ5*02	11	CVKDLTGAIFGVISELAVW	19
MIT 89	PC4	1739 0202	IGHV1-2*02	IGHD3-10*02	IGHJ4*02	15	CARGRGVDVMSPLYDNW	17
MIT 90	PC4	2362 0402	IGHV3-23*04	IGHD2-15*01	IGHJ5*02	17	CGKDPRRDCSDIKCNFGPNWFDPW	24
	PC4	1969 0205	IGHV1-69*01	IGHD2-21*02	IGHJ4*02	31	CATNSDWTFDHW	12
MIT 91	PC1	0429 0602	IGHV1-8*01	IGHD5-5*01	IGHJ5*02	0	CARGLVDTAMVTFDPW	16
MIT 92.93	PC1	0845 0107	IGHV5-51*03	IGHD3-3*01	IGHJ6*02	0	CARHFTQGYDFWSGYYTDYYGMDVW	25
MIT 94	PC2	1048 0207	IGHV1-18*01	IGHD5-12*01	IGHJ5*02	30	CARLHYHLDQGWIDPW	16
MIT 95	PC2	0332 0304	IGHV3-33*01	IGHD5-24*01	IGHJ2*01	0	CARWGGRRDGYINWYFDLW	19
MIT 96	PC2	2037 0106	IGHV1-2*02	IGHD6-19*01	IGHJ6*02	1	CARGPPPYYYYGMDVW	16
MIT 180	RI7B	1839 0402	IGHV3-49*04	IGHD2-2*01	IGHJ6*02	9	CTRDLEVAAANLAEYYYGMDVW	22
MIT 181	RI7A	0641 0702	IGHV4-61*01	IGHD3-22*01	IGHJ4*02	0	CASLAGDHYYDRWYYFDYW	19
MIT 182	RI7A	0751 0301	IGHV3-20*01	IGHD3-3*02	IGHJ4*02	13	CARDEGPGAFGYW	13
MIT 183	RI7A	1067 0602	IGHV3-64*05	IGHD2-2*03	IGHJ4*02	3	CVKDQGIGYCSSTSCRWAAAGTWGVFDYW	29
MIT 184	RI7A	1612 0203	IGHV1-69*01	IGHD2-15*01	IGHJ6*02	7	CARARSVAATPNDYFYYGMDVW	22
MIT 185	RI1	0907 0101	IGHV4-30-2*01	IGHD2-2*01	IGHJ6*03	34	CALARGLGSCIGTSCYAPYYYFMDLW	26
MIT 186	RI1	1743 0301	IGHV4-34*01	IGHD1-26*01	IGHJ6*03	38	CVRVMPDLGLARWAEKFYNFLSGDAKEYYYMDVW	34
MIT 187	RI1	1440 0403	IGHV4-31*03	IGHD3-10*01	IGHJ6*02	17	CARRYSQRFGEFFPTTVDMDVW	22
	RI3	2343 0501	IGHV4-4*07	IGHD3-22*01	IGHJ3*02	0	CARDYVAVITSPTFDIW	17
	RI3	2319 0604	IGHV1-2*02	IGHD3-16*02	IGHJ6*03	5	CARGLRGTTRRHYYYYYYMDVW	22
	RI3	1672 0201	IGHV2-5*02	IGHD2-21*01	IGHJ4*02	29	CAHSYIATPDYW	12
	RI3	0123 0304	IGHV4-59*01	IGHD3-3*01	IGHJ6*01	0	CARGDYDFWSGYLFDIW	17
	RI2	0947 0203	IGHV3-30*18	IGHD7-27*01	IGHJ3*02	27	CAKALLGSDSDVVLTDDAFHTW	22
	RI4	0320 0106	IGHV4-31*03	IGHD2-2*02	IGHJ4*02	11	CARARYTSQSFDSW	14
	RI6	0521 0703	IGHV3-7*03	IGHD3-3*01	IGHJ5*02	11	CARMFYDFWSGYYMDVW	17
	RI6	1134 0405	IGHV3-33*01	IGHD6-13*01	IGHJ6*02	0	CARDQGVTAAGTLYYYYYGMDVW	23
	RI6	0931 0207	IGHV2-26*01	IGHD3-22*01	IGHJ4*02	0	CARITYYYDSSGYPDYW	17
	RI6	1357 0401	IGHV4-39*01	IGHD1-26*01	IGHJ4*02	14	CARQKYSGSYIEYW	14
	RI7_R	0443 0602	IGHV3-48*103	IGHD2-21*02	IGHJ6*02	18	CARGREVAGDHYGMDVW	17

* only cloned antibodies labeled as MIT ### expressed protein

Table 4.3: **West Nile virus specific antibody light chains from single-cell analysis.**

			Light Chain				
Clone No.	**Subject**	**Well ID**	**V Region**	**J Region**	**No. of Mismatches**	**CDR3**	**CDR3 Length**
MIT 86	PC3	1363 0301	IGKV1D-33*01	IGKJ4*01	2	CQQYDNLPLTF	11
MIT 87	PC3	2167 0307	IGKV1-39*01	IGKJ1*01	2	CQQSYSTRTF	10
	PC4	0324 0603	IGKV3-20*01	IGKJ2*01	5	CQQYGSSYTF	10
MIT 88	PC4	1561 0703	IGKV1D-12*01	IGKJ4*01	15	CQQAYSFPLTF	11
MIT 89	PC4	1739 0202	IGLV2-23*03	IGLJ3*02	16	CCSFAGGGTWVF	12
MIT 90	PC4	2362 0402	IGKV1-5*01	IGKJ2*01	7	CQQFNSYPNTF	11
	PC4	1969 0205	IGLV2-23*02	IGLJ2*01	20	CCSYAGSTYNALF	13
MIT 91	PC1	0429 0602	IGKV1-9*01	IGKJ5*01	1	CQQLNSYLPITF	12
MIT 92,93	PC1	0845 0107	IGKV1D-33*01	IGKJ4*01	1	CQQHDNLPPQ#LTF	14
MIT 94	PC2	1048 0207	IGKV3-20*01	IGKJ1*01	19	CQQYGSSPWTF	11
MIT 95	PC2	0332 0304	IGKV3-11*01	IGKJ5*01	1	CQQRSNWPPITF	12
MIT 96	PC2	2037 0106	IGLV9-49*01	IGLJ3*02	0	CGADHGSGSNFV*PE#VF	18
MIT 180	RI7B	1839 0402	IGKV2-30*01	IGKJ1*01	6	CMQGTHWRTF	10
MIT 181	RI7A	0641 0702	IGKV1-39*01	IGKJ4*01	0	CQQSYSTPPLTF	12
MIT 182	RI7A	0751 0301	IGKV1-9*01	IGKJ2*01	6	CQQLNSYPYTF	11
MIT 183	RI7A	1067 0602	IGKV3-20*01	IGKJ4*01	1	CQQYGSSPPLTF	12
MIT 184	RI7A	1612 0203	IGLV3-1*01	IGLJ3*01		CQAWDSSTAWVF	12
MIT 185	RI1	0907 0101	IGLV1-44*01	IGLJ2*01	23	CAAWDVSLRGVVF	13
MIT 186	RI1	1743 0301	IGKV3-20*01	IGKJ1*01	35	CQQYSGSQWTF	11
MIT 187	RI1	1440 0403	IGKV2-28*01	IGKJ2*02	12	CMQGLQTPFTF	11
	RI3	2343 0501	IGLV2-23*02	IGLJ3*02	0	CCSYAGSSTFWVF	13
	RI3	2319 0604	IGLV3-21*03	IGLJ2*01	2	CQVWDTSSDLVVF	13
	RI3	1672 0201	IGLV2-14*03	IGLJ3*02		CSSHASGDTLVF	12
	RI3	0123 0304	IGLV2-14*03	IGLJ1*01	0	CSSYTSSSTHYVF	13
	RI2	0947 0203	IGKV3-15*01	IGKJ4*01	18	CQQYNNWPPLTF	12
	RI4	0320 0106	IGKV3-20*01	IGKJ2*02	4	CQQYGSSFPYSF	12
	RI6	0521 0703	IGKV3-20*01	IGKJ4*01	8	CQQYVKSPLTF	11
	RI6	1134 0405	IGKV1-17*01	IGKJ1*01	2	CLQHNSYPWTF	11
	RI6	0931 0207	IGKV1-5*03	IGKJ1*01	5	CQQYNSNSGTF	11
	RI6	1357 0401	IGLV1-40	IGLJ3*02	7	CQSYDSSLSGSGVF	14
	RI7_R	0443 0602	IGLV1-51*01	IGLJ2*01	1	CGTWDSSLSAVVF	13

* only cloned antibodies labeled as MIT ### expressed protein

Table 4.4: **Cloned West Nile virus specific antibody sequences.**

		Heavy Chain				Light Chain		
Clone No.	**Subject**	**V Region**	**D Region**	**J Region**	**CDR3**	**V Region**	**J Region**	**CDR3**
MIT 86	PC3	V2-26*01	D3-22*01	J4*02	CARVNSSGYYFDYW	V1D-33*01	J4*01	CQQYDNLPLTF
MIT 87	PC3	V4-59*01	D3-22*01	J5*02	CARGYYDSTRDWFDPW	V1-39*01	J1*01	CQQSYSTRTF
MIT 88	PC4	V3-64*05	D3-3*01	J5*02	CVKDLTGAIFGVISELAVW	V1D-12*01	J4*01	CQQAYSFPLTF
MIT 89	PC4	V1-2*02	D3-10*02	J4*02	CARGRGVDVMSPLYDNW	V2-23*03	J3*02	CCSFAGGGTWVF
MIT 90	PC4	V3-23*04	D2-15*01	J5*02	CGKDPRRDCSDIKCNFGPNWFDPW	V1-5*01	J2*01	CQQFNSYPNTF
MIT 91	PC1	V1-8*01	D5-5*01	J5*02	CARGLVDTAMVTFDPW	V1-9*01	J5*01	CQQLNSYLPITF
MIT 92	PC1	V5-51*03	D3-3*01	J6*02	CARHFTQGYDFWSGYYTDYYGMDVW	V4-1*01	J1*01	CQQYYSTPPTF
MIT 93	PC1	V5-51*03	D3-3*01	J6*02	CARHFTQGYDFWSGYYTDYYGMDVW	V1-39*01	J1*01	CQQSYSTRTF
MIT 94	PC2	V1-18*01	D5-12*01	J5*02	CARLHYHLDQGWIDPW	V3-20*01	J1*01	CQQYGSSPWTF
MIT 95	PC2	V3-33*01	D5-24*01	J2*01	CARWGGRRDGYINWYFDLW	V3-11*01	J5*01	CQQRSNWPPITF
MIT 96	PC2	V1-2*02	D6-19*01	J6*02	CARGPPPYYYYGMDVW	V9-49*01	J3*02	CGADHGSGSNFV*PE#VF
MIT 180	RI7	V4-61*01	D3-22*01	J4*02	CASLAGDHYYDRWYYFDYW	V1-39*01	J4*01	CQQSYSTPPLTF
MIT 181	RI7	V3-20*01	D3-3*02	J4*02	CARDEGPGAFGYW	V1-9*01	J2*01	CQQLNSYPYTF
MIT 182	RI7	V3-64*05	D2-2*03	J4*02	CVKDQGIGYCSSTSCRWAAAGTWGVFDYW	V3-20*01	J4*01	CQQYGSSPPLTF
MIT 183	RI7	V1-69*01	D2-15*01	J6*02	CARARSVAATPNDYFYYGMDVW	V3-1*01	J3*01	CQAWDSSTAWVF
MIT 184	RI7	V3-49*04	D2-2*01	J6*02	CTRDLEVAAANLAEYYYGMDVW	V2-30*01	J1*01	CMQGTHWRTF
MIT 185	RI1	V4-30-2*01	D2-2*01	J6*03	CALARGLGSCIGTSCYAPYYYFMDLW	V1-44*01	J2*01	CAAWDVSLRGVVF
MIT 186	RI1	V4-34*01	D1-26*01	J6*03	CVRVMPDLGLARWAEKFYNFLSGDAKEYYYMDVW	V3-20*01	J1*01	CQQYSGSQWTF
MIT 187	RI1	V4-31*03	D3-10*01	J6*02	CARRYSQRFGEFFPTTVDMDVW	V2-28*01	J2*02	CMQGLQTPFTF

Table 4.5: **Characteristics of West Nile virus neutralizing antibodies.**

Antibody ID	Subject ID	Disease state	WNV E IgG Immunoblot	PRNT Neutralization*	Relative affinity (Ag/Ig) to WNV E protein
MIT 87	PC3	Post-conv.		+	+++
MIT 89	PC3	Post-conv.	+	+	+++
MIT 90	PC4	Post-conv.		+	++++
MIT 91	PC1	Post-conv.			+++
MIT 95	PC2	Post-conv.		+	+++++
MIT 180	RI7	Recent infect.			+++++
MIT 181	RI7	Recent infect.			++++
MIT 182	RI7	Recent infect.			+
MIT 183	RI7	Recent infect.	++		
MIT 184	RI7	Recent infect.			+
MIT 185	RI1	Recent infect.			no data
MIT 187	RI1	Recent infect.	+		no data
2G12	NA	HIV	NA	NA	
HS+ (1:20)			+++	+	
HS- (1:20)					

* Neutralization cut off value: 90%; 2G12, HIV specific antibody control; HS, Human serum controls

Table 4.6: **Summary of next generation sequencing and clone analysis for each subject.**

ID	Disease State	Raw Reads	Antibody sequences	Assigned to clones	WNV+ antibodies	Number of WNV clones	WNV clone sizes
RI1	Recent infect.	2233926	23788	13056	3	3	5,22,33
RI2	Recent infect.	1900765	75683	52588	1	0	NA
RI3	Recent infect.	1972153	63132	45571	5	1	5
RI4	Recent infect.	1462649	53144	26618	2	0	NA
RI5	Recent infect.	1385024	55486	37869	2	0	NA
RI6	Recent infect.	2491337	79991	58256	10	0	NA
RI7	Recent infect.	1613256	70044	39773	9	1	11
PC1	Post-conv.	1723558	61668	59402	5	0	NA
PC2	Post-conv.	2156022	26078	24996	5	0	NA
PC3	Post-conv.	1893813	143829	135139	2	0	NA
PC4	Post-conv.	1898589	21389	20479	6	0	NA

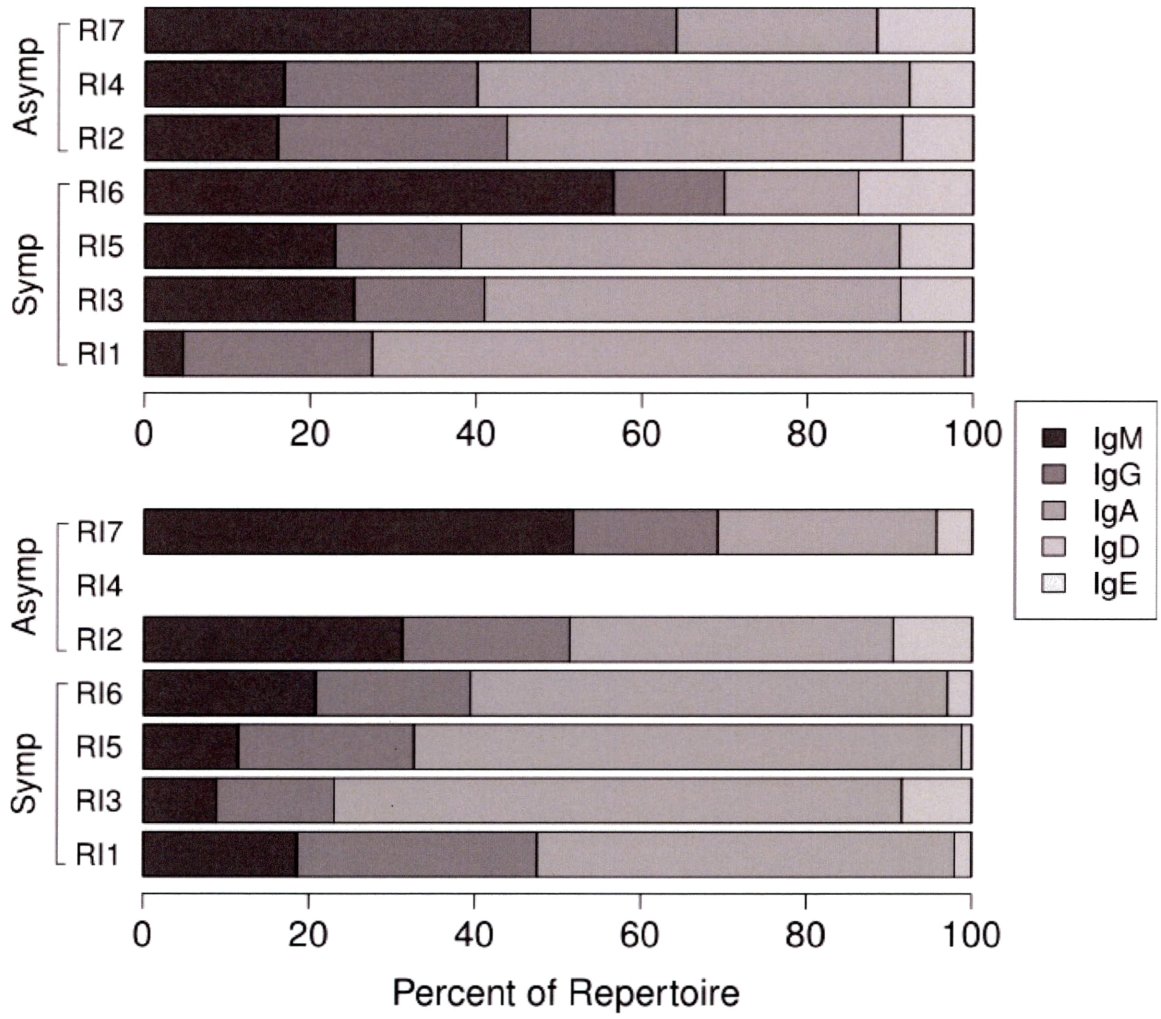

Figure 4.2: **No difference in isotype distribution between asymptomatic and symptomatic recently infected subjects.** Distribution of isotypes from PBMCs (upper) and Memory B cells (lower) NGS Ig sequencing by subject.

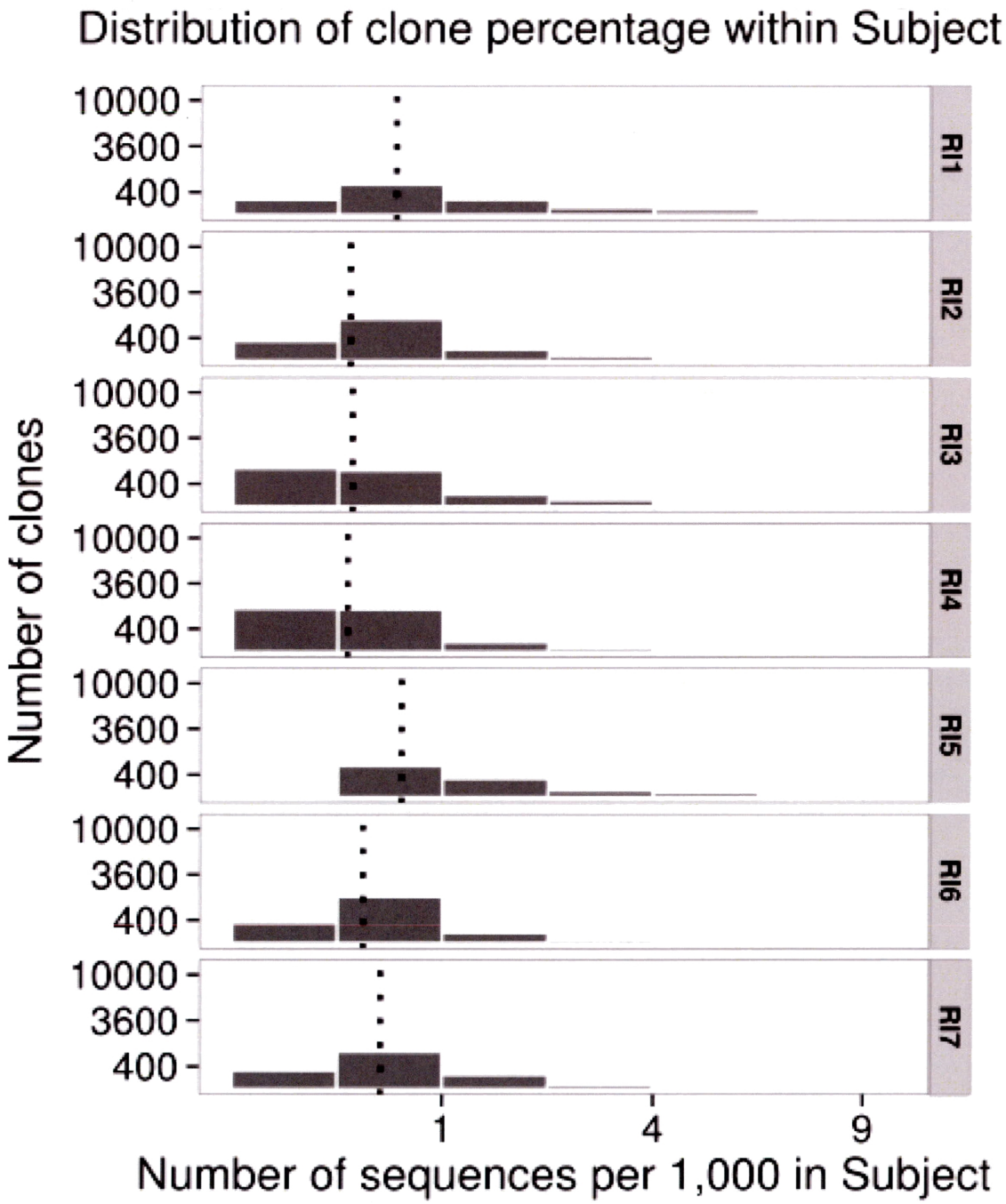

Figure 4.3: **Distribution of antibody repertoire clone sizes.** Histogram of clone sizes normalized by sequencing depth in each subject, dotted line shows median of distribution.

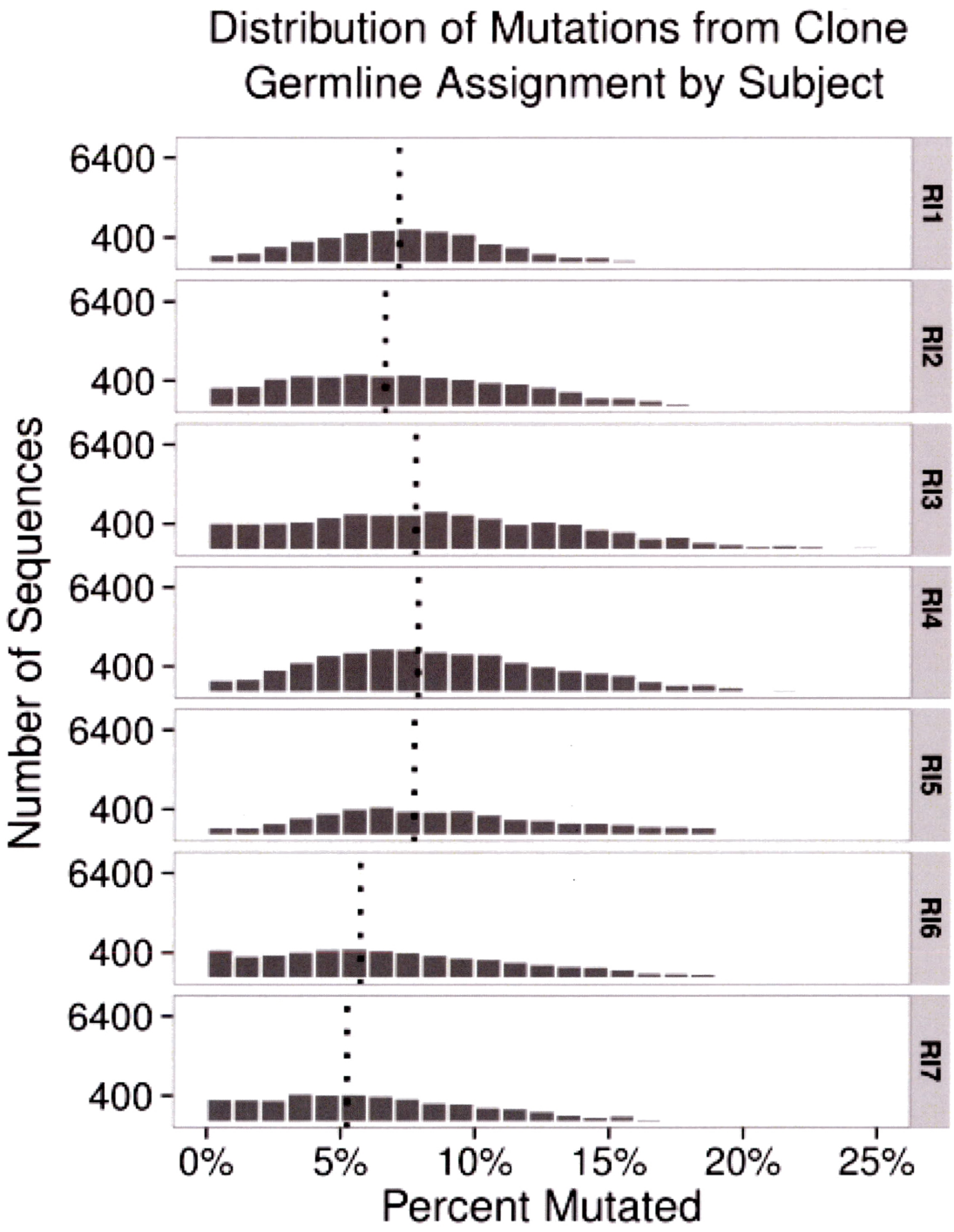

Figure 4.4: **Distribution of antibody repertoire mutation levels.** Histogram of mutation levels (V segment up to start of CDR3) from the germline sequence of each clone in Ig sequences, dotted line shows median of distribution.

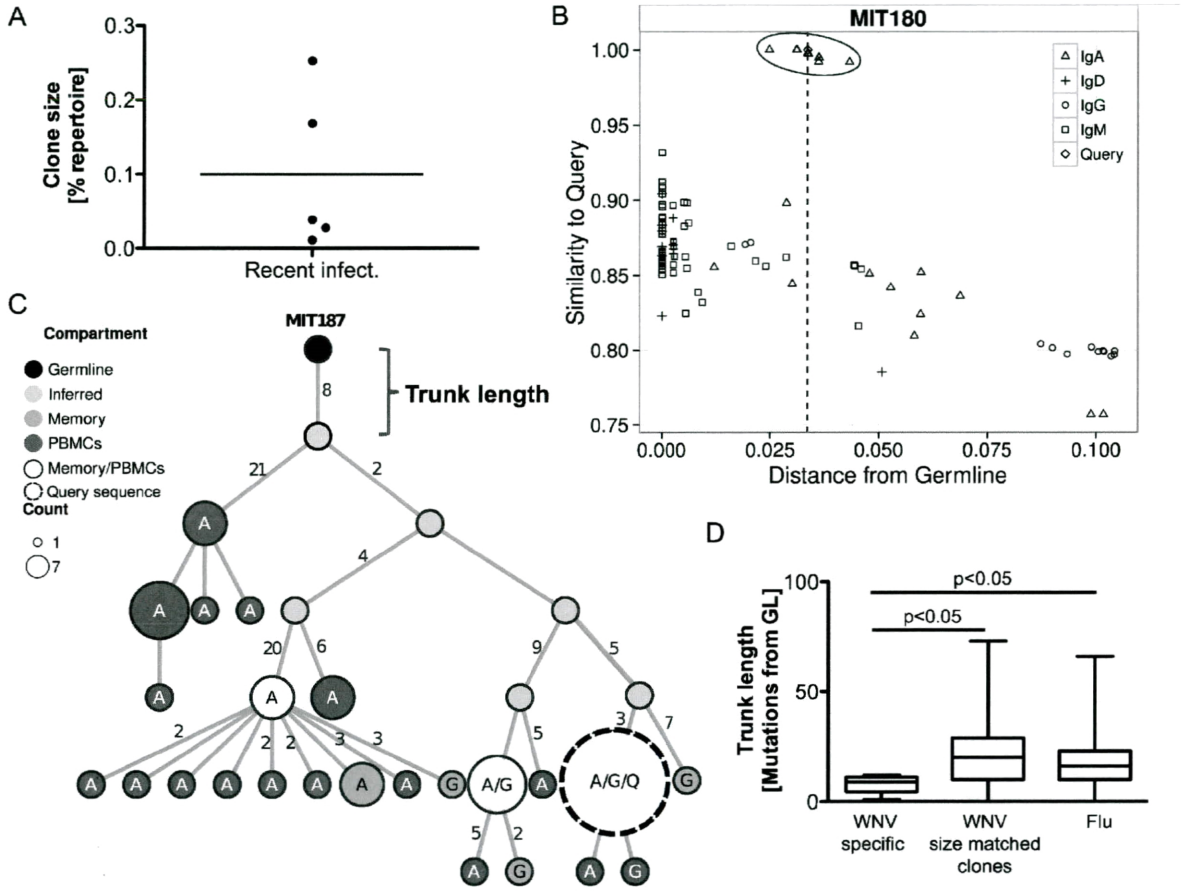

Figure 4.5: **Trunk length analysis of West Nile virus-specific clones in the immunoglobin repertoire indicates unlikely previous exposure to the virus.** (A) Size of WNV-specific clones found in recently infected subjects. (B) Representative plot of sequence similarity of heavy chain (same V and J segment) identified within the repertoire as a function of distance of the sequence from germline (x-axis) and the corresponding "query" WNV-specific sequence (y-axis). An outlier cluster representing a putative WNV-specific clone is circled. (C) Maximum parsimony lineage tree for one WNV-specific clone (MIT187). Each node represents a unique sequence, with size representing the number of duplicate reads observed. Edge lengths correspond to the number of mutations between sequences. Shading of the node represents the compartment in which the sequence was found. The nodes are labeled with the isotypes of the observed sequence (A: IgA, G: IgG, Q: WNV-specific query sequence from single-cell screening). (D) Plot of observed trunk lengths (number of mutations between germline and most recent common ancestor of clone). Emerging WNV-specific clones have significantly fewer trunk mutations (Student's t-test) compared to size-matched clones found in the same subjects, and to expanded clones (at least 0.05% of the repertoire) from subjects who received an influenza vaccination.

influenza. To investigate this hypothesis, we approximated the sequence of the initiating B cell for each clone as the most recent common ancestor in a maximum parsimony lineage tree (Figures 4.5C & 4.7). Maximum parsimony lineage trees minimize the number of mutation events in each clone and infer sequences that may have existed between observed antibodies and the germline sequence. The trunk length of the lineage tree (*e.g.*, the number of mutations in the most recent common ancestor compared with the germline sequence) approximates the maturation state of the initiating B cell for each clone.

In comparison to other similarly sized B cell clones in the recently infected WNV subjects, the trunks in WNV-specific clones had significantly ($p < 0.05$) fewer mutations (Figure 4.5D). It is possible, however, that some of the similarly-sized clones in this cohort are WNV-specific, which could confound the comparison. To address this issue, we also compared the trunk lengths in WNV-specific clones with expanded (at least 0.05% of the repertoire) clonal lineages from three responses following influenza vaccination, obtained from publicly available data 35. We found that the WNV-specific clones once again had trunks that were significantly ($p < 0.05$) less mutated than those found from the influenza response. The close mutational distance of sequences that give rise to WNV-specific B cell clones to their respective germlines supports our hypothesis that these subjects have not previously been exposed to WNV and have experienced primary affinity maturation rather than a recall response. In the case of two clones (MIT187 and MIT180) (Figure 4.7), the query sequence was not a terminal leaf; other sequences with additional mutations were observed in the lineage tree. Antibodies both more and less mutated than the query sequences could also show high affinity and neutralizing activity to WNV. This general approach of using specific antibodies found by microengraving as query sequences to reveal similar antibodies at the repertoire level could have wide applications for a larger scale search for therapeutic neutralizing antibodies specific to many other diseases.

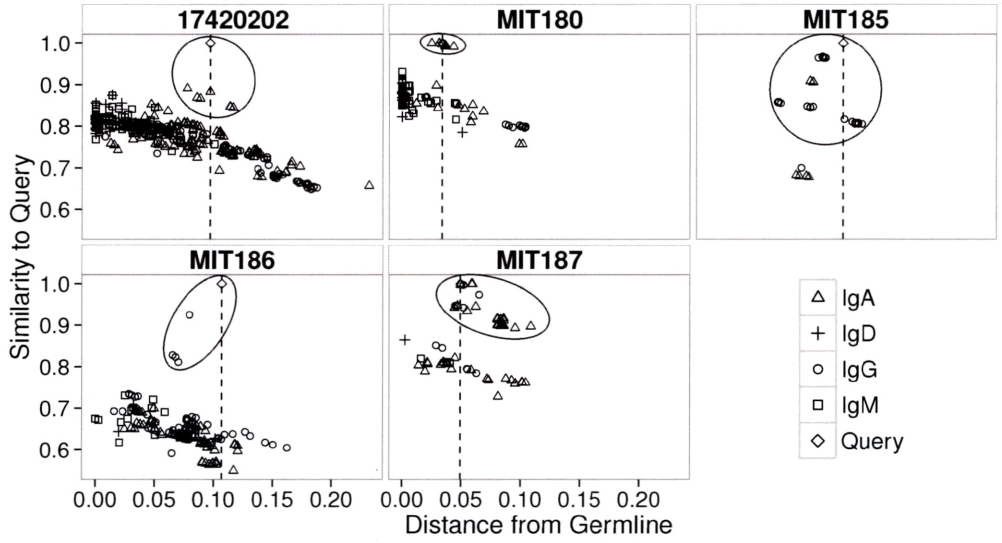

Figure 4.6: **Query plots of West Nile virus-specific clones found in Ig repertoire.** Plots of sequence similarity of heavy chain (same V and J segment) identified within the repertoire as a function of distance of the sequence from germline (x-axis) and the corresponding WNV-specific "query" sequence (y-axis). Outlier clusters representing putative WNV-specific clones are circled. The shape of each point indicates the isotype of the Ig sequence. 17420202, MIT180, and MIT187 were identified by automated clonal grouping; MIT185 and MIT186 were identified by manual inspection.

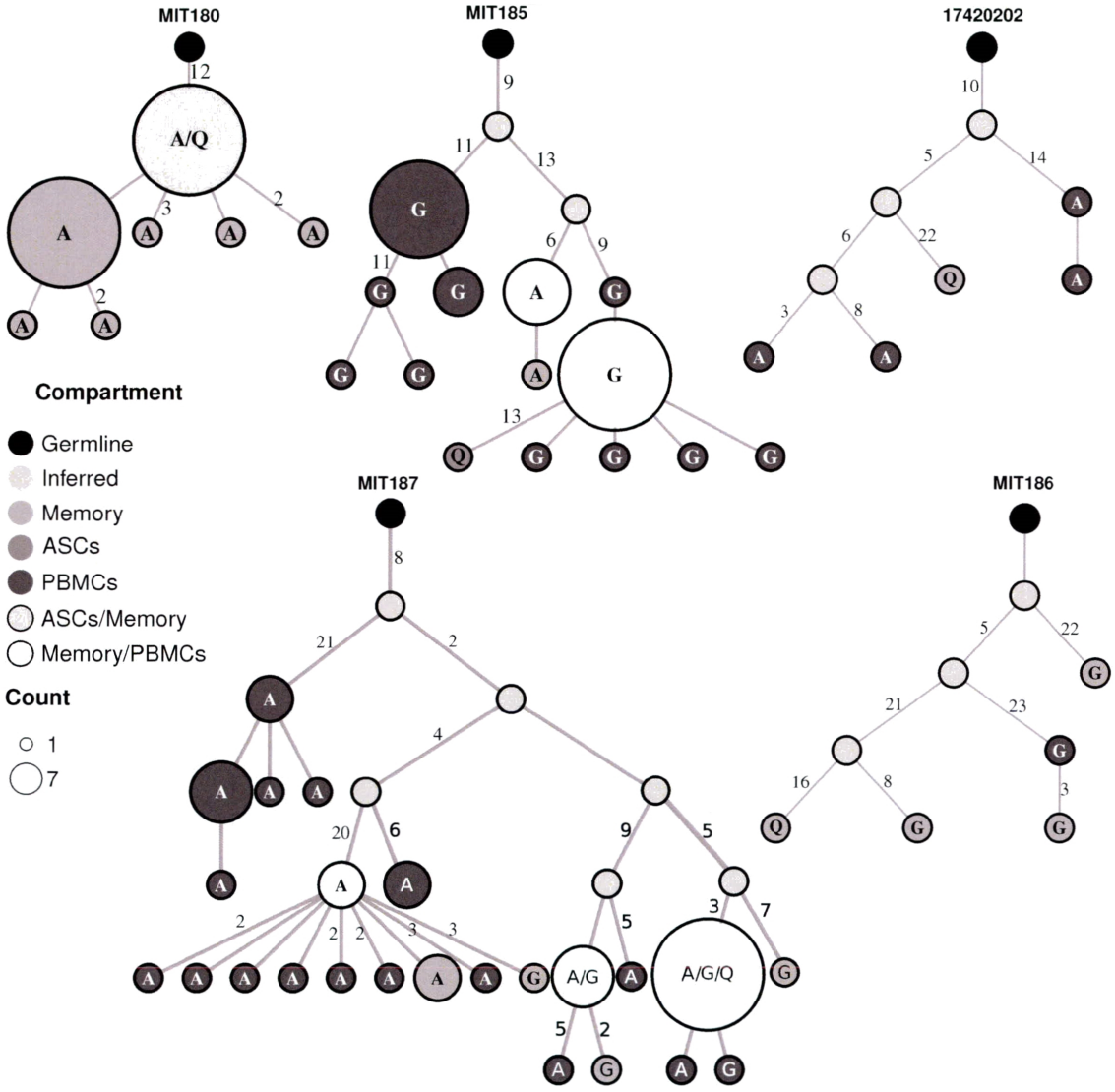

Figure 4.7: **Lineage trees of West Nile virus-specific clones found in Ig repertoire.** Maximum parsimony lineage trees of WNV-specific clones. Each node represents a unique sequence, with size correlating to number of duplicate reads observed. Edge lengths correspond to the number of mutation between sequences (unlabeled edges are one mutation). Shading of the node represents the compartment in which the sequence was found. The node label determines the type of sequence observed (A: IgA, G: IgG, Q: WNV-specific query sequence from single-cell screening).

4.1.4 Discussion

In this study, we analyzed PBMCs from individuals with recent or post-convalescent WNV infections. We effectively identified WNV-specific antibodies by single-cell analysis. We next cloned and expressed the antibodies identified by single-cell analysis and performed PRNT, to identify WNV-specific NAbs. Four of the NAbs identified also showed high affinity to WNV E protein as determined by protein microarray. In contrast, only one NAb was also positive in the immunoblot assay. These results may indicate that three of the discovered NAbs recognize conformation dependent epitopes on the WNV E protein that could have been denatured during the immunoblot assay 36. Similarly, antibodies to the related Flavivirus family member dengue virus, a quaternary epitope was required for neutralization of the virus (de Alwis, S. a. Smith, *et al.*, 2012), indicating that denaturing of the higher order antibody protein structure could result in a loss of function. We also observed that neutralizing activity was positively correlated with increased antibody affinity to WNV E protein. This trend suggests that the development of NAbs in WNV infection requires a high degree of affinity maturation. In support of this hypothesis, we found NAbs exclusively in the post-convalescent subjects, which could indicate that the detected antibodies developed later during the course of infection, and that they require a prolonged exposure to antigen to promote affinity maturation. Chronic or repeated exposure to viral antigens such as HIV-1 can elicit affinity maturation beyond the degree observed here (Scheid, Mouquet, Feldhahn, *et al.*, 2009).

In contrast to previously identified NAbs derived using murine phage and yeast display libraries (Oliphant, Engle, *et al.*, 2005; Pierson *et al.*, 2007; Throsby *et al.*, 2006; Vogt *et al.*, 2009; Gould *et al.*, 2005), the antibodies here were directly derived from human B cells obtained from individuals previously infected with WNV. These antibodies therefore reflect the natural selection and maturation of WNV-specific NAbs in humans. The challenge here exists in identifying the rare B cells producing NAbs. For instance, in infections with the dengue virus, a related flavivirus, only a small fraction of the antibodies expressed strongly neutralized the virus (de Alwis & de Silva, 2014) although recent studies have identified an epitope that may be relevant for effective neutralization (Dejnirattisai *et al.*, 2014). This outcome highlights the importance of utilizing single-cell analysis methods to identify rare B cells of interest. Future studies should further elucidate the epitopes and mechanisms of action of the NAbs identified here. In addition, developing and validating a small animal model and testing these antibodies could provide evidence for potential therapeutic use.

Though previous longitudinal studies of the immune response to influenza vaccination indicate

that clonal expansion occurs seven days post-vaccination (Laserson *et al.*, 2014; Jiang, He, *et al.*, 2013; Vollmers *et al.*, 2013), the recently infected subjects in this study had been infected with WNV at least four weeks prior to the first blood draw. The expansion of WNV-specific clones in comparison with post-convalescent subjects exposed to WNV years prior to sampling implies that the antibody-mediated response may persist longer than in the case of other infections or vaccines. Chronic infections have been shown to lead to expanded B cell clones in the case of cytomegalovirus (CMV) or clones with higher levels of mutations in the case of Epstein-Barr virus (EBV) (C. Wang *et al.*, 2014).

In summary, we effectively identified novel human-derived NAbs and analyzed the humoral response to WNV infection by integrating single-cell analysis, repertoire next generation sequencing, and conventional methods. Detecting persistent antigen-specific MBCs and ASCs in the periphery of post-convalescent subjects can reveal the origin of these antibodies and potentially provide clues about the persistence of WNV in the body. We believe that WNV provides an excellent model to examine primary naive responses of human B cells, given the disease's well-defined clinical history for the onset of infection, the low occurrence of infection by this or others flaviviruses (e.g., dengue) in North America, which allows the characterization of naive responses to a recently introduced pathogen with unlikely re-exposure. Our results comparing the distances of inferred clonal founder cells from their respective germlines support the hypothesis that WNV-specific antibodies develop from naive B cells during a primary infection, rather than a recall response from an affinity matured memory B cells. These results highlight the importance of analyzing the immune response to infection with integrated single-cell data, serum analysis, and whole repertoire analysis. These methods to analyze the humoral response can also be applied to other infectious diseases research and in vaccine trials.

4.2 Salmonella

4.2.1 Introduction

The immune response to microorganisms is an interplay between aspects of innate and adaptive immunity. Successful pathogens often have multiple mechanisms to evade or subvert the immune response. Furthermore, pathogens contain molecular patterns that stimulate a wide variety of innate immune pattern-recognition receptors, whether expressed on innate cells such as macrophages,

dendritic cells (DCs), or on adaptive cells such as B lymphocytes. These pathways and innate receptor ligands in turn shape adaptive immunity.

Currently, our knowledge of B cell responses is mostly based on artificial systems that lack these natural innate immune cues. Instead, they use non-replicating antigens given in adjuvant. The "canonical" response described in these models includes a rapid transient extrafollicular (EF) plasmablast (PB) response followed by germinal center (GC) appearance (Shlomchik & Weisel, 2012). While some pathogen responses follow this progression (*e.g.*, the response to influenza (Coro *et al.*, 2006; Moyron-Quiroz *et al.*, 2004; Onodera *et al.*, 2012; Rothaeusler & Baumgarth, 2010)), there is increasing evidence that in many other infections this is not the case. During Ehrlichia muris infection splenic GC formation is suppressed (Racine *et al.*, 2010). Similarly, Borrelia burgdorferi interferes with the B cell response by affecting its quality and kinetics, delaying GC appearance and instead stimulating immunoglobulin M (IgM) antibody-forming cells (AFCs) (Hastey *et al.*, 2012).

Though the immune response to Salmonella enterica serovar Typhimurium (STm), a facultative intracellular gram negative bacterium, is relatively well studied (Dougan *et al.*, 2011), information on the B cell response is limited. This is a major omission, considering that STm is a clinically relevant microorganism and that live attenuated strains have been proposed and are in phase I clinical trials as vectors for vaccines (Kong *et al.*, 2012). Furthermore, STm and related serovars are a major cause of infectious diarrhea in the developed world and they are also responsible for serious disseminated infections in Africa and Asia. It is highly homologous to Salmonella Typhi, and considered a murine model for the study of this pervasive human pathogen. The B cell response to STm can be protective in both mice and humans, via antibodies or other mechanisms (Nanton *et al.*, 2012).

STm induces a massive extrafollicular AFC response in the spleen, while GC formation is greatly delayed (Cunningham *et al.*, 2007). Both T-dependent (TD) and T-independent (TI) components contribute to the response (Gil-Cruz et al., 2009). The mechanisms that shape this type of B cell response remain to be elucidated, whereas parameters of virulence and protection have received greater attention. Deletion of the signaling adaptor MyD88 appeared to favor, rather than inhibit, STm virulence (Arpaia *et al.*, 2011; Barr *et al.*, 2010; Neves *et al.*, 2010).

A number of studies have addressed the targets of the B cell response, yet overall these remain poorly defined. LPS, outer membrane proteins (OMPs) and possibly flagellin are identified as primary Ags of the switched Ab response (Bobat *et al.*, 2011; Calderon *et al.*, 1986; Cunningham

et al., 2007; Ortiz *et al.*, 1989; Singh *et al.*, 1992). Recently, some of the authors of the present work have screened immune sera on antigen (Ag) microarrays, thus identifying antibody (Ab) signatures of human and murine Salmonellosis (S. Lee *et al.*, 2012). Serum signatures can partly describe the status of the Ab response, but they do not reveal its ontogeny; moreover, serum Ab profiles might be discordant with memory or effector cell specificities (Guan *et al.*, 2009). Knowing antigenic targets is certainly important for vaccine design, yet further research is necessary to understand the underlying mechanisms of response and protection; for instance, to explain why vaccines to Salmonella have only moderate, transient efficacy (McGregor *et al.*, 2013).

Here we focused both on defining the specificities of the B cell response and addressing why it follows an extrafollicular pathway rather than a GC one. Our initial hypothesis was that the massive plasmablast response was polyclonal and non-specific, owing to innate immune receptor stimulation of B cells. Initial evidence indicated that the response was apparently non-specific. However, a series of experiments using a variety of approaches ultimately revealed a process in which very low affinity, yet specific, B cells — found at unexpectedly high precursor frequency — join the initial proliferative plasmablast response, and in the absence of developed GCs eventually acquired somatic mutations which in turn led to sufficient affinity maturation for the ultimate detection of conventional "specificity" for the immunizing bacteria. These results reveal an unappreciated pathway of response to a gram-negative bacterial pathogen and in addition lead to a revised view of the nature of clonal selection, specificity, affinity, and humoral immune response evolution.

4.2.2 Materials and methods

4.2.2.1 Mouse strains

The following mice strains were used: wild type C57Bl/6J and BALB/c (NCI or Jackson Laboratories); MyD88-/- BALB/c (Adachi *et al.*, 1998) or MyD88-/- C57Bl/6 (from Drs. Bockenstedt and Goldstein at Yale University); IL-1R-/- C57Bl/6 (from Dr. Flavell at Yale University); Tlr2-/- Tlr4-/- nrampwt C57Bl/6 and nrampwt C57Bl/6 (from Dr. Barton at UCSF); asc -/- Balb/c (Sutterwala *et al.*, 2006); Tlr9-/- C57Bl/6 (Hemmi *et al.*, 2000); TCRβ-/- C57Bl/6 (Jackson Laboratories); Vκ8-R+/+ and Jκ-/- BALB/c. B1-8+/+ and B1-8+/+ Jκ-/- C57Bl/6 (Sonoda *et al.*, 1997).

4.2.2.2 Mice, bacteria, and infection procedures

The mice strains used in this study were bred under specific pathogen free conditions in the animal facility at Yale University. The AroA- attenuated Salmonella Typhimurium strain SL3261 (Hoiseth & Stocker, 1981) was kindly provided by Roy Curtiss III, Arizona State University. For infection, 105 bacteria (unless otherwise specified) in PBS were injected i.p.. The bacterial burden was assessed by plating serial dilutions of tissue homogenates, prepared by 1mm-beads disruption in a FastPrep-24 instrument (MP Biomedicals). All mouse work was according to protocols approved by the Yale Institutional Animal Care and Use Committee.

4.2.2.3 Laser capture and microdissection

7 μm spleen sections were prepared from OCT-frozen tissues on the membrane-coated PEN slides (Leica). To detect plasmablast patches, slides were blocked with 10% FCS and 0.1% BSA in PBS and immunohistochemistry was performed using a rat anti-CD138-biotin antibody, followed by a goat antirat-biotin antibody, and then HRP-conjugated streptavidin. To detect GC-like structures, slides were blocked with 10% rat serum, 1% BSA and 0.05% Tween-20, and the following antibodies were used: PNA-biotin followed by AP-conjugated streptavidin (Southern Biotech), and anti-IgD-FITC followed by HRP-conjugated anti-FITC (Millipore). The slides were developed with 3-amino-9-ethylcarbazole or Fast Blue BB base solution (HRP and AP, respectively), washed extensively in water to remove salt, and allowed to dry until dissected. Microdissections were performed using a Leica LMD6500 instrument equipped with an optical microscope. Dissected patches were collected in the cap of PCR microtubes in 10μl of digestion buffer (50mM Tris-HCl, 50mM KCl, 0.63mM

EDTA, 0.22% Igepal, 0.22% Tween20, 0.8mg/ml proteinase K). Patches were digested at 55°C for two hours, then at 90°C for 5 minutes, and used for PCR amplification of antibody genes.

Amplification of antibody genes by PCR was done as follows. A primary PCR was performed by adding 40μl of PCR reaction mix, which used the high fidelity Pfu Ultra II Fusion polymerase (Agilent). A second, nested PCR was performed in 50μl using 0.5μl of the primary PCR as template. The following primers were used: for the primary PCR, a shorter version of the 5' MsVHE primer described in (Tiller, Busse, *et al.*, 2009), called 5' MsVHE-short (5'- GGGAATTCGAG-GTGCAGCTGCAG -3'), was used together with a mix of 4 antisense primers mapping in the JH region (3' SalI P-mJH01/02/03/04) as reported in (Tiller et al., 2009). For the nested PCR, the full-length sense primer 5' MsVHE was used together with a mix of 4 nested JH antisense primers that were newly designed: 5'- TGGTCCCTGTGCCCCAGACATCG -3', 5'- GTGGTGCCTTGGCC-CCAGTAGTC -3', 5'- AGAGTCCCTTGGCCCCAGTAAGC -3' and 5'- GAGGTTCCTTGACC-CCAGTAGTC -3'. The resulting PCR products were cloned and sequenced using the Zero Blunt TOPO PCR Cloning kit for sequencing (Life Technologies) per the manufacturer's protocol.

4.2.2.4 Analysis of microdissected sequences

Raw reads were filtered in several steps to identify and remove low quality sequences. Conservative thresholds were applied in all cases, to increase the reliability of mutation calls, at the potential expense of excluding some real mutations. Preprocessing was carried out using pRESTO (Vander Heiden *et al.*, 2014). as follows:

1. Reads with a mean Phred quality score below 20 were removed.

2. Reads without valid constant region primer or template switch sequences were removed, with a maximum primer match error rate of 0.2 and a maximum template switch error rate of 0.5. Both template switch additions and constant region primer sequences were deleted from the reads. The isotype of each read was assigned according to its constant region primer match.

3. Reads with identical unique molecule identifiers (UIDs) were collapsed into a single consensus sequence for each UID. UID read groups with nucleotide diversity scores (Nei & Li, 1979) exceeding 0.1 or majority isotype frequency under 0.6 were discarded. In cases where multiple isotypes were identified in a single UID read group, the consensus sequence was based only upon the subset of reads in the UID read group assigned to the majority isotype.

4. UID consensus sequence mate-pairs were then assembled into full length Ig sequences with a maximum allowed error rate of 0.3 and p-value threshold of 0.01.

5. Duplicate full length sequences were discarded, with the exception that duplicate sequences derived from different biological samples and/or assigned to different isotypes were retained. Each sequence was assigned an mRNA copy number value based on the total number of UIDs having an identical sequence. Following preprocessing, V(D)J germline segments were assigned using IMGT/HighV-QUEST (Alamyar *et al.*, 2012).

Post-processing of IMGT/HighV-QUEST output and clonal grouping was performed as follows:

1. Non-functional sequences were removed from further analysis.

2. Functional V(D)J sequences were assigned into clonal groups by first partitioning sequences based on common IGHV gene, IGHJ gene, and junction region length. Within these larger groups, sequences differing from one another by a weighted distance of less than 3 within the junction region were defined as clones. Distance was measured as the number of point mutations weighted by a symmetric version of the nucleotide substitution probability previously described (D. S. Smith *et al.*, 1996). A distance of 3 corresponds to three transition mutations or one of the least likely mutations.

3. Lineage trees were constructed for each clonal group via maximum parsimony using the dnapars application of PHYLIP (Felsenstein J., 2005).

4. Inferred ancestors in each tree were recursively replaced with descendants having the same sequence as their inferred parent.

4.2.3 SHM takes place in follicles and at extrafollicular sites

Given evidence for extensive SHM, it was compelling to investigate where it takes place. The appearance of specific AFC occurring prior to GC formation would be consistent with extra-GC SHM and Ag-driven affinity maturation. Although SHM does not canonically happen at EF sites, it does occur in murine spleen in the autoimmune setting (William *et al.,* 2002), showing that GCs are not strictly necessary for this process; however, a physiological counterpart to extra-GC mutation has to date not been revealed in a murine pathogen-specific response. The rapid appearance of isotype-switched plasmablasts post-infection shows that activation-induced cytidine deaminase (AID) is activated very early in the response among extrafollicular B cells, making it plausible to think that SHM could also be occurring.

To resolve this, we first searched for GC-like structures that might have been an alternative site for SHM. The apparent contrast between the lack of GCs as assessed by immunofluorescence and the flow cytometric detection of cells with a GC phenotype, the frequencies of which admittedly did not rise above baseline until day 22, could be reconciled by the finding, starting 3 weeks after infection, of scattered small PNA-positive cell aggregates in some sections. The cells in these small clusters expressed the PNA target weakly, as they were only found by overexposing PNA; they also expressed low, but detectable, amounts of the GC marker Bcl6. These clusters, which we refer to as "GC-like," were found at an atypical site — the interface between the B and T cell zones — and did not appear to develop past this stage into proper GCs for the duration of our studies, up to 6 weeks after infection. A similar response was observed in BALB/c mice, with lack of proper GCs and a remarkable EF plasmablast response. Together these data suggest that GCs could potentially form at late time points, but do not mature. We considered whether innate immune signals in this context could in fact suppress GC formation. Indeed, when BALB/c mice deficient for MyD88 were infected, fully developed GCs were observed as early as 10 days after infection, accompanied by EF plasmablasts, and were more numerous by day 15. However, this was not observed in C57BL/6 MyD88-deficient animals, indicating that additional factors may suppress GC formation in a strain-specific manner.

To address whether SHM takes place in "GC-like" structures and/or at EF sites, we performed laser capture microdissection (LCM) and V region sequencing of both structures (example, Figure 4.8A) at 3 weeks post-infection. Picks comprised 20 to 30 cells and in the case of plasmablasts, whenever possible, the same patches were taken from two or three consecutive slides. We dissected

both adjacent and distant patches (Table 4.7). 14 plasmablast picks yielded a PCR product that was then cloned, followed by sequencing of V gene inserts in multiple colonies derived from each product. 79% of the unique sequences obtained showed some somatic mutations, averaging 3.2 mutations per mutated sequence (Tables 4.8 and 4.9). When clonally related sequences (with the same VDJ rearrangement) were found, then a lineage tree was built that describes the evolution of the clone. 11 of 14 picks gave clones to generate lineage trees. Such trees demonstrate that ongoing V region diversification was taking place among the few cells captured in each such microdissected patch. Figure 4.8B-D shows representative trees (see also Figure 4.9); in some cases (Figure 4.8B,C) the trees were fairly simple, with two or three clonally related sequences that were exclusively derived from the same pick. In other cases, there was higher complexity, with sequences from different, though adjacent, picks (Figure 4.8D).

The mutations observed in the lineage trees were authentic and not a result of PCR error: the rates are far higher than that expected from PCR error using high-fidelity polymerases (William *et al.*, 2002). Moreover, the presence of shared mutations is not expected from PCR error, but is expected from clonal expansion and selection of authentic SHM. The isolation of the same mutation from different picks of the same geographic patch (*i.e.*, taken from different serial sections), and hence different PCR amplifications fully excludes PCR error as the explanation. Rather, when one considers that it takes only 4 divisions to create a cluster of 16 cells, and that SHM in GCs introduces about 0.25 mutations/V region/division, then the extent of mutation is consistent with this high rate, as there should be roughly 1 difference between each sequence (Kleinstein *et al.*, 2003). In parallel, we dissected and sequenced some of the GC-like structures. A similar degree of SHM was observed in these structures, with 70% of the sequences being mutated with an average of 2.3 mutations per mutated sequence (Table 4.8). Overall, these data show that there is robust diversification through SHM, and as extensively shown and discussed below, that this is likely to occur locally and determine affinity maturation.

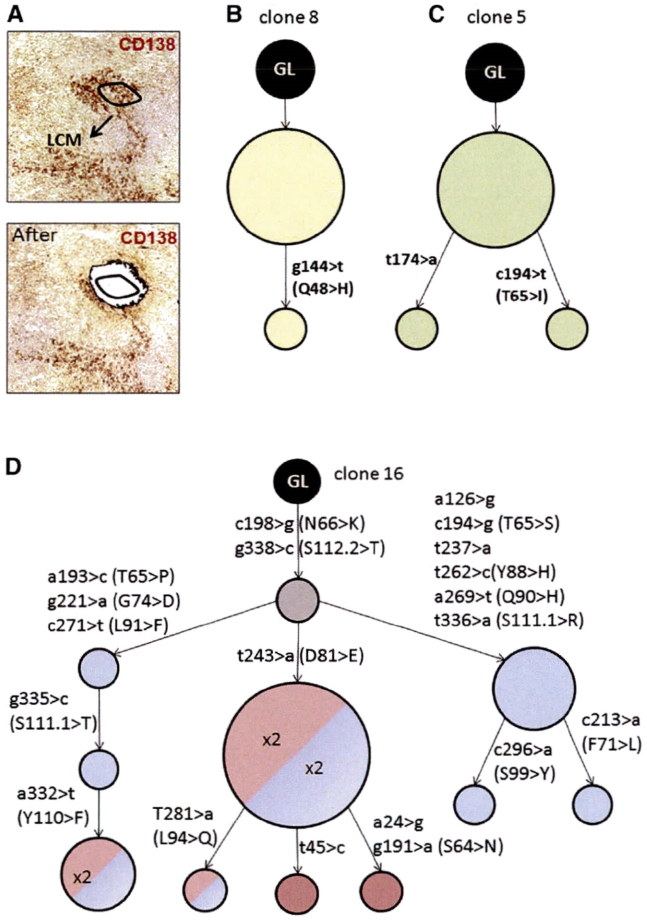

Figure 4.8: **Ongoing SHM Takes Place in Plasmablast Patches.** (A) Example of a plasmablast patch pick by laser capture and microdissection (LCM); the same patch is shown before and after the pick. The black line indicates the area of $\sim$ 20 cells that was dissected. The plasmablast patches are identified with anti-CD138 staining by immunohistochemistry. (B-D) Examples of three clonal trees of different complexity derived from the analysis of the Ab gene sequences obtained by LCM. The size of each node indicates the number of identical sequences found. In (D), a more complex tree is shown that was composed of sequences that derived from several nearby picks; different colors of the nodes denote different, but adjacent, picks from which the sequence was derived, while the number within the node indicates in how many serial slides (always from the same patch) the same sequence was found. The gray circle indicates an inferred intermediate. The position of the mutated nucleotides and aminoacids (in the case of replacement mutations) are shown along the branches. See also Figure 4.9 and Tables 4.7, 4.8, and 4.9.

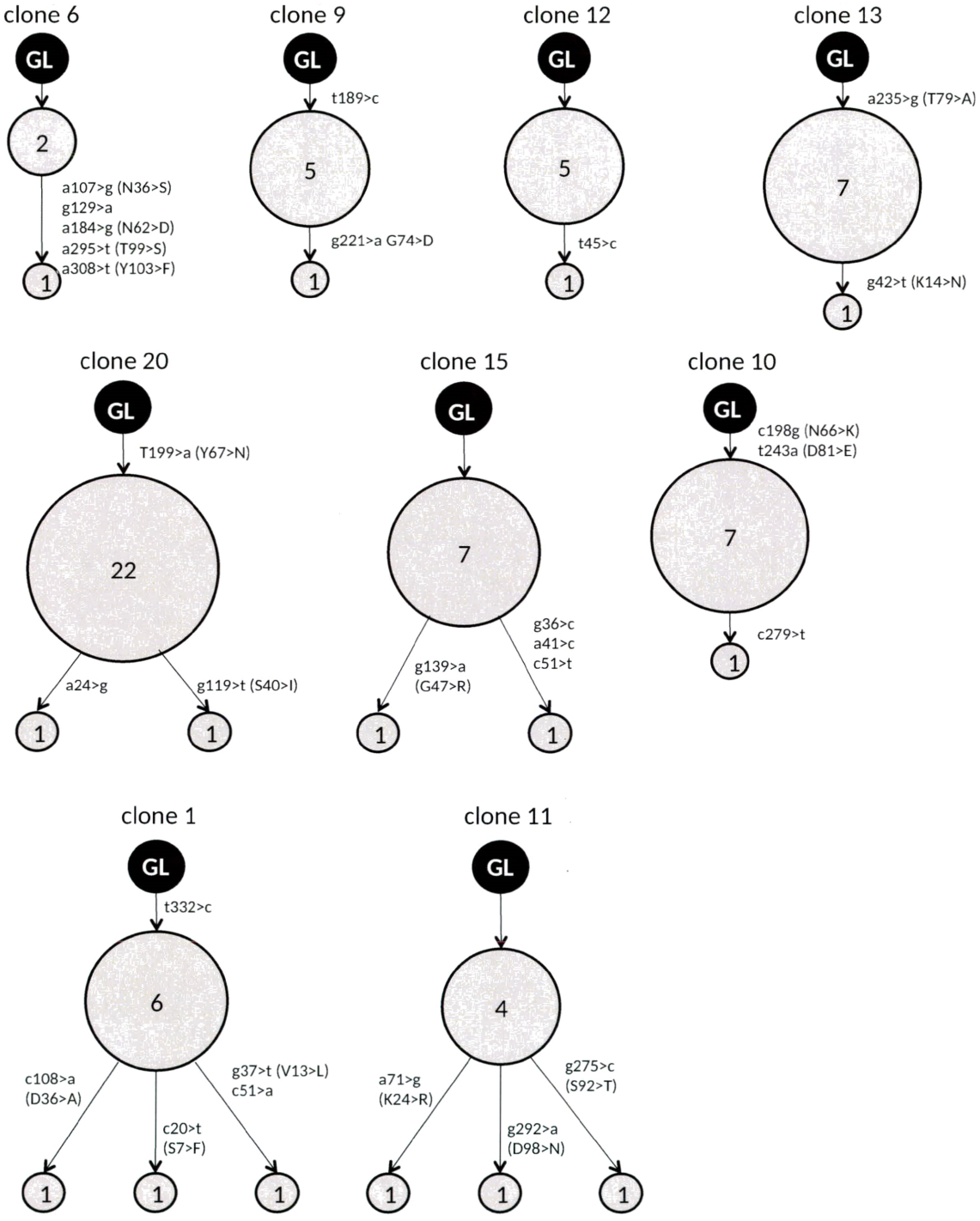

Figure 4.9: **Clonal lineage trees.** Additional trees generated from data obtained from laser capture and microdissection. See Tables 4.7, S3 and S4 for additional details. The size of the node is representative of the number of identical sequences that were obtained, which are also shown by the number inside each node.

Table 4.7: **Summary of the picks from which sequences were obtained.**

	Mouse 1				Mouse 2		
Plasmablast Patches	**A**	**B (adjacent to C)**	**C (adjacent to B)**	**Consecutive Slide**	**D**	**E**	**F**
Pick #	1	2	3	#1			6
	7	8	9	#2	10	11	
	13	14	15	#3	16	17	
"GC-like" structures	**GC 1**	**GC 2**	**GC 3**		**GC 4**		
Pick #	19	20	21		22		

Table 4.8: **Summary of the data obtained from experiments with laser capture and microdissection.** For GC-like structures, a total of 4 picks yielded sequences: 20 unique sequences, 14 mutated sequences with a total of 33 mutations, for an average of 1.6 mutation/sequence or 2.3 mutations/mutated sequence. For plasmablast patches, 14 picks yielded 38 unique sequences, 30 mutated sequences with 95 mutations, for an average of 2.5 mutation/sequence or 3.2 mutations/-mutated sequence.

Mouse 1, clones from "GC-like" structures; 15 unique sequences								
Clone	**GC-like**	**Pick**	**# Uniq seqs**	**V-gene**	**J-gene**	**Junction Length**	**# Uniq muts**	**Avg muts**
2	GC 2	20	1	IGHV14-2*01	IGHJ2*01	27	1	1
3	GC 2	20	1	IGHV14-4*01	IGHJ2*01	42	0	0
4	GC 3	21	1	IGHV1-26*01	IGHJ2*01	42	1	1
6	GC 1	19	2	IGHV14-3*01	IGHJ1*03	48	5	2.5
9	GC 1	19	2	IGHV14-4*01	IGHJ3*01	42	2	1.5
12	GC 3	21	2	IGHV1-66*01	IGHJ2*01	39	1	0.5
13	GC 1	19	2	IGHV1-80*01	IGHJ3*01	42	2	1.5
17	GC 3	21	1	IGHV1-81*01	IGHJ2*01	45	0	0
18	GC 3	21	1	IGHV14-1*01	IGHJ2*01	51	0	0
19	GC 3	21	1	IGHV14-4*01	IGHJ2*01	54	0	0
21	GC 3	21	1	IGHV14-4*01	IGHJ3*01	21	0	0
Mouse 1, clones from plasmablast patches; 19 unique sequences								
Clone	**Patch**	**Pick**	**# Uniq seqs**	**V-gene**	**J-gene**	**Junction Length**	**# Uniq muts**	**Avg muts**
5	A	13	3	IGHV14-2*01	IGHJ2*01	42	2	0.7
14	A	1/13	1	IGHV1-81*01	IGHJ2*01	36	2	2
15	A	7	3	IGHV1-82*01	IGHJ2*01	36	4	1.3
16	B/C	2/3/8/9/14/15	10	IGHV1-9*01	IGHJ4*01	54	20	4.7
24	C	15	1	IGHV14-3*01	IGHJ2*01	45	0	0
25	C	3	1	IGHV1-19*01	IGHJ2*01	39	0	0
Mouse 2, clones from "GC-like" structures; 5 unique sequences								
Clone	**GC-like**	**Pick**	**# Uniq seqs**	**V-gene**	**J-gene**	**Junction Length**	**# Uniq muts**	**Avg muts**
20	GC 4	22	3	IGHV1-81*01	IGHJ2*01	36	5	3.7
22	GC 4	22	1	IGHV1-4*01	IGHJ2*01	27	4	4
23	GC 4	22	1	IGHV1-85*01	IGHJ4*01	45	2	2
Mouse 2, clones from plasmablast patches; 19 unique sequences								
Clone	**Patch**	**Pick**	**# Uniq seqs**	**V-gene**	**J-gene**	**Junction Length**	**# Uniq muts**	**Avg muts**
1	E	17	4	IGHV14-1*01	IGHJ2*01	33	5	2
7	D	10	1	IGHV1-19*01	IGHJ2*01	39	1	1
8	E	11	2	IGHV14-4*01	IGHJ2*01	27	1	0.5
10	E	11	2	IGHV1-9*01	IGHJ4*01	54	3	2.5
11	D	16	4	IGHV1-62-2*01	IGHJ2*01	51	8	1.8
26	D	16	1	IGHV14-1*01	IGHJ2*01	48	2	2
28	E	17	1	IGHV14-1*01	IGHJ2*03	33	1	1
29	E	17	1	IGHV14-1*01	IGHJ3*01	33	0	0
30	E	11	1	IGHV14-4*01	IGHJ2*03	27	0	0
31	E	11	1	IGHV14-4*01	IGHJ4*01	54	10	10
32	F	6	1	IGHV1-82*01	IGHJ2*01	33	0	0

Table 4.9: **Summary of the data obtained from experiments with laser capture and microdissection.** LCM data as in Table 4.8, displayed according to the pick from which the sequences were obtained.

Mouse 1, clones from "GC-like" structures; 15 unique sequences				
Pick	**Clone**	**V-gene**	**J-gene**	**Junction Length**
19	6	IGHV14-3*01	IGHJ1*03	48
	9	IGHV14-4*01	IGHJ3*01	42
	13	IGHV1-80*01	IGHJ3*01	42
20	2	IGHV14-2*01	IGHJ2*01	27
	3	IGHV14-4*01	IGHJ2*01	42
	4	IGHV1-26*01	IGHJ2*01	42
21	12	IGHV1-66*01	IGHJ2*01	39
	17	IGHV1-81*01	IGHJ2*01	45
	18	IGHV14-1*01	IGHJ2*01	51
	19	IGHV14-4*01	IGHJ2*01	54
	21	IGHV14-4*01	IGHJ3*01	21
Mouse 1, clones from plasmablast patches; 19 unique sequences				
Pick	**Clone**	**V-gene**	**J-gene**	**Junction Length**
1	14	IGHV1-81*01	IGHJ2*01	36
2	16	IGHV1-9*01	IGHJ4*01	54
3	16	IGHV1-9*01	IGHJ4*01	54
	25	IGHV1-19*01	IGHJ2*01	39
7	15	IGHV1-82*01	IGHJ2*01	36
8	16	IGHV1-9*01	IGHJ4*01	54
9	16	IGHV1-9*01	IGHJ4*01	54
13	5	IGHV14-2*01	IGHJ2*01	42
	14	IGHV1-81*01	IGHJ2*01	36
14	16	IGHV1-9*01	IGHJ4*01	54
15	24	IGHV14-3*01	IGHJ2*01	45
	16	IGHV1-9*01	IGHJ4*01	54
Mouse 2, clones from "GC-like" structures; 5 unique sequences				
Pick	**Clone**	**V-gene**	**J-gene**	**Junction Length**
22	20	IGHV1-81*01	IGHJ2*01	36
	22	IGHV1-4*01	IGHJ2*01	27
	23	IGHV1-85*01	IGHJ4*01	45
Mouse 2, clones from plasmablast patches; 19 unique sequences				
Pick	**Clone**	**V-gene**	**J-gene**	**Junction Length**
6	32	IGHV1-82*01	IGHJ2*01	33
10	7	IGHV1-19*01	IGHJ2*01	39
11	8	IGHV14-4*01	IGHJ2*01	27
	10	IGHV1-9*01	IGHJ4*01	54
	30	IGHV14-4*01	IGHJ2*03	27
	31	IGHV14-4*01	IGHJ4*01	54
16	11	IGHV1-62-2*01	IGHJ2*01	51
	26	IGHV14-1*01	IGHJ2*01	48
17	1	IGHV14-1*01	IGHJ2*01	33
	28	IGHV14-1*01	IGHJ2*03	33
	29	IGHV14-1*01	IGHJ3*01	33

4.3 Celiac disease

4.3.1 Introduction

Celiac disease (CD) is a multifactorial disorder characterized by an intestinal inflammatory response to ingested cereal gluten proteins (Sollid, 2000). The human leukocyte antigen association and the central role of CD4+ T cells in the pathogenesis are thoroughly investigated (Jabri & Sollid, 2006). T cells of the lesion recognize gluten peptides that are deamidated in vivo by the enzyme transglutaminase 2 (TG2) (Molberg *et al.*, 1998; Wal *et al.*, 2016). Notably, the great majority of CD patients develop an autoantibody response, with TG2 itself being the main autoantigen (Dieterich *et al.*, 1997). It is not known whether these antibodies have a role in the pathophysiology of CD, yet anti-TG2 immunoglobulin A (IgA) antibodies are increasingly used as diagnostic tool (Rostom *et al.*, 2005) and in the follow-up of the treatment, as upon commencement of a gluten-free diet (GFD) autoantibodies disappear from serum within months (Sugai *et al.*, 2010). Anti-TG2 antibodies are produced by plasma cells (PCs) localized in the lamina propria of the intestinal mucosa (Marzari *et al.*, 2001; Picarelli *et al.*, 1996; Di Niro, Mesin, *et al.*, 2012), but PCs localized elsewhere may also contribute to the antibody production. Korponay-Szabo *et al.*, 2004 developed a double-color immunofluorescence method that allowed visualizing antibody deposits in correspondence with the subepithelial TG2 layer in the small intestine, as well as in other tissues. This method builds on colocalization between IgA and the TG2 protein, the latter usually identified by means of the CUB7402 murine monoclonal antibody. Importantly, this has been proposed as a tool for early diagnosis12 on affected individuals without signs of villous atrophy (Kaukinen *et al.*, 2005; Tosco *et al.*, 2008) or with negative/borderline serology (Salmi *et al.*, 2006). Recently, we reported a flow cytometry-based method to describe and enumerate TG2-specific PCs from freshly obtained biopsies of the small intestine of CD patients (Di Niro, Mesin, *et al.*, 2012). We found that in untreated CD patients the frequency of these cells was exceptionally high, and that they produced antibodies with limited somatic hypermutation but nonetheless reasonable affinity. In the same work we reported an immunofluorescence-based method to visualize TG2-specific PCs on cryosections. As this material can be stored for prolonged periods of time and requires relatively little amount of specimen, studies on cryosections provide a simple yet very valid alternative for the analysis of intestinal humoral responses. Even though significant discoveries of the anti-TG2 antibody response have been done, its importance in the diagnostic workup of CD and the unknown

role in the pathogenesis of the disease suggest that more research on this topic is necessary. In this work, we expand our previous findings on the intestinal anti-TG2 response by investigating and describing the location of TG2-specific PCs and antibodies, the phenotype of these cells as well as the niche supporting them, and the kinetics of their disappearance upon commencement of a GFD, and we discuss a mechanism for a possible pathogenic involvement. Our findings provide a new tool to analyze the antibody response in CD and in general the specificity of antibody responses directly in affected tissues.

4.3.2 Materials and methods

4.3.2.1 Patient material

Adult CD patients were diagnosed according to standard criteria, including human leukocyte antigen genotyping, anti-TG2 serum titer, and histological analysis of small intestinal biopsy (AGA Institute, 2006). The same criteria were used to exclude CD diagnosis in controls. Ethical approval for study of Norwegian subjects was obtained from the Regional Ethics Committee in South-Eastern Norway (project S-97201). Each study subject gave written informed consent. Duodenal biopsy specimens were obtained by gastroduodenoscopy. Tissue sections of biopsies of seven CD patients with active disease, from three CD patients treated with a GFD and two healthy control subjects were studied. In addition, biopsies from five untreated CD patients and five healthy controls were processed and used for flow cytometry analysis and PCs of biopsies from two untreated CD patients were processed for high-throughput sequencing. Ethical approval for study of the Finnish subjects was obtained from the Ethics Committee of Tampere University Hospital, and each subject provided written informed consent. Samples from 15 CD patients and 4 nonceliac controls suffering from dyspepsia were studied. From these subjects small-bowel mucosal biopsies were taken either with an adult-size Watson capsule from the proximal jejunum or upon endoscopy with forceps from the distal duodenum. For each subject, part of the samples were snap-frozen and embedded in optimal cutting temperature compound (OCT, Tissue-Tec, Miles, Elkhart, IN) for storing at -70 °C, whereas the rest of the biopsy specimens were paraffin embedded. Serum IgA class antibodies against TG2 were detected by enzyme-linked immunosorbent assay using human recombinant TG2 as antigen, with a cutoff line of 5.0 U/ml (Celikey, Phadia, Freiburg, Germany).

4.3.2.2 Immunohistochemistry and laser capture and microdissection

For laser capture and microdissection, 7 μm thick cryosections from specimens of patients were cut on 2 μm PEN-membrane slides (Leica, Buffalo Grove, IL). Sections were fixed in acetone for 10 min and stored at -80°C. For laser capture and microdissection, a Leica LMD6500 instrument equipped with an optical microscope was used, and sections were therefore stained by immunohistochemistry. Briefly, slides were thawed and rinsed in PBS. Primary reagents (mouse anti-CD138 or recombinant human biotinylated TG2 produced in E. coli) were diluted in 1% bovine serum albumin in PBS and added to the slide for 45 min at RT in a moist chamber. Slides were extensively rinsed and secondary reagents (goat anti-mouse IgG-HRP, 1:500, Santa Cruz (Dallas, TX), or SA-HRP,

1:500, Southern Biotech) in 1% bovine serum albumin in PBS were added for 30 min at RT. After extensive washing with PBS, development was performed by adding 3-amino-9-ethylcarbazole substrate (Sigma, St Louis, MO) and incubating for 2 to 20 min until proper development occurred. Slides were carefully washed with water to remove all salts, dried at RT, and then stored at 4°C until necessary. Dissections were performed in the caps of 0.5 ml tubes, to which 10 μl of digestion buffer (50 mm Tris-HCl, 50 mm KCl, 0.63 mm EDTA, 0.22% Igepal, 0.22% Tween-20, and 0.8 mg/ml proteinase K) were added. After dissection of plasma cell patches in the caps, these were closed and the tubes spinned. The samples were added with a drop of mineral oil, digested at 55°C for 2 h, followed by inactivation of proteinase K at 90°C for 5 min, and used for PCR amplification of antibody genes directly from genomic DNA.

4.3.2.3 PCR and analysis of sequences from plasma cell picks

A seminested PCR was performed in 50 μl using the high-fidelity PfuTurbo DNA polymerase enzyme (Agilent, Santa Clara, CA) as per the manufacturer's instructions. Primers used in the primary and secondary PCRs were as described by Kuppers, 2004. The VH1, VH3, and VH5 primers were used as a mix in the primary PCR, and individually in the secondary PCR. Antisense primers in the J regions were used as originally described in the protocol (Kuppers, 2004). PCR products were purified and cloned in the Zero Blunt TOPO PCR cloning kit for sequencing (Life Technologies) as per the manufacturer's instructions, and individual colonies were sequenced at the Keck DNA sequencing facility at Yale.

Analysis of the sequences was performed as follows: preprocessing was carried out using pRESTO (Vander Heiden *et al.*, 2014) as previously described (Stern *et al.*, 2014). V(D)J germline segments were determined using IMGT/HighV-QUEST (Alamyar *et al.*, 2012) and divided into clonally related groups based on common V gene, J gene, and junction region length as previously described (Stern *et al.*, 2014). Lineage trees were constructed for each clonal group with the dnapars application of PHYLIP (Felsenstein J., 2005).

4.3.3 Analysis of PCs dissemination in the gut mucosa by laser capture and microdissection

To investigate whether intestinal PC populations expand locally or at a distant site, we performed laser capture and microdissection followed by sequencing of immunoglobulin heavy chain variable (IGHV) genes on patches with such cells from intestinal biopsies of CD patients. To this end, we adapted a published set of primers for a seminested PCR of antibody genes on genomic DNA. Because of the available microscope configuration, the procedure was optimal for dissecting sections stained by immunohistochemistry under transmitted light. Preliminary experiments indicated that staining for the CD138 marker was optimal under these conditions, and we used this marker to stain sections from samples known to have high frequency of TG2-specific PCs. We dissected patches comprising 20 to 30 PCs (an example is shown in Figure 4.10a) from one patient with active CD. In order to increase the variability of the analysis, the same patch was picked from three consecutive slides. We amplified genomic DNA and obtained a PCR product from all of the picks, and built and sequenced libraries. Importantly, several IGHV-5 sequences with few or no mutations were found; IGHV-5 genes had on average 2.2 mutations (n=9) as compared with other genes that averaged 11.4 mutations/sequence (n=33). Although not a formal demonstration, this strongly suggests specificity for TG2 of some of the dissected cells; in fact, as us and others have reported (Di Niro, Mesin, *et al.*, 2012; Benckert *et al.*, 2011) IGHV-5 genes with low degree of somatic hypermutation are otherwise rare among the intestinal PC compartment.

Interestingly, in some cases clonal sequences were found that showed shared and unique mutations, thus raising the possibility that such clones mutate and evolve locally. We further extended our analysis and compared sequences found in different patches that were sampled from an additional patient with an active disease. We dissected four PC patches (named A, B, C, and D) that were not adjacent to each other. The same patch was dissected from several consecutive slides. Libraries were obtained from 14 picks, yielding 53 unique sequences representing 26 different clones; 18 were represented by one single sequence and 8 by 2 or more and allowed us to build lineage trees. Table 4.10 provides a summary of the eight clonal trees that were built with sequences from this patient: whereas a few simple trees could be built from clonal sequences that were found only within one of the patches (example in Figure 4.10b, with sequences derived from one single patch, clone ID 23 according to Table 4.10), in many other cases larger trees could be built where several different patches contributed clones (example in Figure 4.10c, with sequences from three different

patches, clone ID 22 according to Table 4.10).

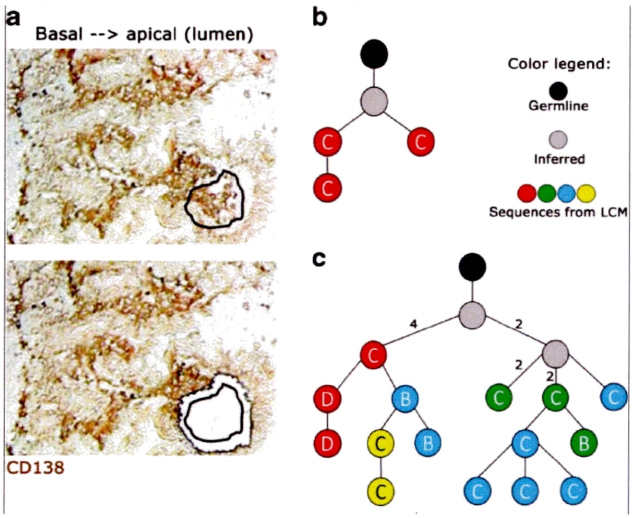

Figure 4.10: **Laser capture and microdissection (LCM) of plasma cell patches rich in transglutaminase 2 (TG2)-specific ones.** Sequences are obtained from PCR performed on genomic DNA from dissected patches in the lamina propria. (a) Example of a pick comprising 20 plasma cells. (b, c) Examples of clonal trees built from clonally related sequences derived from the same patch ((b) clone ID 23 fromgod my e Table 1) or from three different patches ((c) clone ID 22 from Table 1) of the same patient. Data are summarized in Table 4.10. The letter inside the node indicates the patch from which the sequence comes, whereas different colors indicate sequences derived from different, consecutive cryosections. The number next to the line connecting the inferred sequence (gray node) and actual observed sequences (colored nodes) represents the number of mutations by which the two connected nodes differ.

Table 4.10: **Laser capture and microdissection.** Clonal trees of immunoglobulin heavy chain variable (IGHV) genes of the plasma cells (PCs) built from picks. Clones 22 and 23 were used for the representative trees shown in Figure4.10b,c. Mutations include both silent and replacement ones.

CLONE ID	Number of unique sequences	V-GENE:	Number of unique mutations	Average number of mutations	Found in patch(es)
2	3	VH5-51	2	0.7	C, D
4	2	VH5-51	9	5.5	C
7	2	VH5-51	8	5	C
14	2	VH3-66	10	9.5	C
21	3	VH5-51	13	6.7	B, C
22	15	VH5-51	21	5.2	B, C, D
23	3	VH5-51	3	1.3	C
24	2	VH5-a	6	5.5	A, D

Chapter 5

Conclusions and Future Directions

5.1 Summary

Humoral immunity is driven by the expansion, somatic hypermutation, and selection of B cell clones. Each clone consists of the descendants of a single B cell that responded to antigen, with Ig receptors that are diversified due to SHM. The collection of naive B cells and mature clones can now be profiled using next-generation sequencing technologies. This large-scale characterization provides a window into the micro-evolutionary dynamics of the adaptive immune response and has a variety of applications in basic science and clinical studies. Measurements from the repertoire such as diversity or mutational load can provide insight into the dysregulation that occurs with aging or disease, identify mutation patterns that lead to broadly neutralizing antibodies, and predict successful response to vaccination. Clonal relationships are not directly measured, but must be computationally inferred from these sequencing data. While several hierarchical clustering-based methods have been proposed, they vary in distance and linkage methods and have not yet been rigorously compared. In Chapter 2 we used a combination of human experimental and simulated data to characterize the performance of hierarchical clustering-based methods for partitioning sequences into clones. We found that single linkage clustering has high performance, with sensitivity and positive predictive value (PPV) both over 99%, whereas other linkages result in a significant loss of sensitivity. Surprisingly, distance metrics that incorporate the biases of somatic hypermutation did not outperform simple Hamming distance. Although errors were more likely in sequences with short junctions, using the entire dataset to choose a single distance threshold for clustering was

near optimal. Our results suggest that hierarchical clustering using single linkage with nucleotide Hamming distance identifies clones with high confidence and provides a fully automated method for clonal grouping. The performance estimates we developed provide important context to interpret clonal analysis of repertoire sequencing data and allow for rigorous testing of other clonal grouping algorithms.

There are several other downstream Ig repertoire analyses in addition to grouping clonally related sequences, such as reconstructing B cell lineages, calculating repertoire diversity, identifying somatic hypermutations, and quantifying selection pressures. Various tools can be used to complete any one of these analysis steps (Yaari & Kleinstein, 2015), but cobbling together separate tools with differing input and output formats into a workflow can present difficulties. In addition to creating software to calculate the various characteristics of the repertoire, there is a need for an cohesive suite of tools that uses a standardized data format to streamline the analysis pipeline. In Chapter 3 we presented a suite of utilities, Change-O, which provides tools for advanced analyses of large-scale Ig repertoire sequencing data. Change-O includes tools for determining the complete set of Ig variable region gene segment alleles carried by an individual (including novel alleles), partitioning of Ig sequences into clonal populations, creating lineage trees, inferring somatic hypermutation targeting models, measuring repertoire diversity, quantifying selection pressure, and calculating sequence chemical properties. All Change-O tools utilize a common data format, which enables the seamless integration of multiple analyses into a single workflow. Change-O, in addition to the pre-processing suite pRESTO, is freely available for non-commercial use as part of the Immcantation framework (`immcantation.readthedocs.io`).

The Change-O tools, particularly clonal grouping, can be applied in many ways to gain biological insight, examples of which were presented in Chapter 1. In Section 1.1, we studied a cohort of subjects infected with West Nile virus (WNV), an emerging mosquito-borne disease that can lead to severe neurological illness and currently has no available treatment or vaccine. We used microengraving, an integrated single-cell assay, to analyze recently infected and post-convalescent subjects and efficiently identified four novel WNV neutralizing antibodies. We also assessed the humoral response to WNV on a single-cell and repertoire level by integrating AIRR-Seq into our analysis. Using the nucleotide coding sequences for WNV-specific antibodies derived from single cells, we revealed the ontogeny of expanded WNV-specific clones in the repertoires of recently infected subjects through quantitative AIRR-Seq analysis. This analysis also indicated that the humoral response

to WNV did not depend on an anamnestic response, due to an unlikely previous exposure to the virus. This innovative and integrative approach to analyze the evolution of neutralizing antibodies from natural infection on a single-cell and repertoire level can also be applied to vaccine studies, and could aid the development of therapeutic antibodies and our understanding of other infectious diseases.

In addition to a combination approach to identify novel neutralizing antibodies, in Sections 4.2 and 4.3 we explore implications of identifying clonal expansions from laser capture and microdissection of different tissue sites. The traditional understanding of the B cell response includes a rapid transient extrafollicular plasmablast response followed by formation of germinal centers, wherein affinity maturation occurs (Shlomchik & Weisel, 2012). The B cell response to Salmonella typhimurium occurs massively at extrafollicular sites, without notable germinal centers. In Section 4.2 we used laser microdissection and Ig sequencing to show that SHM occurred efficiently at extrafollicular sites leading to affinity maturation that in turn led to detectable STm Ag-binding. These results suggest a revised vision of how clonal selection and affinity maturation operate in response to Salmonella. Clonal selection initially is promiscuous, activating cells with virtually undetectable affinity, yet SHM and selection occur during the extrafollicular response yielding higher affinity, detectable antibodies. Localized dissection can also provide insight into diseases that affect mucosa, such as celiac disease, which is characterized by autoantibodies to transglutaminase 2 in gut mucosal tissue. In Section 4.3, we used laser microdissection of plasma cell patches followed by Ig sequencing to identify clonally related sequences across several sites, indicating that intestinal plasma cell populations expand in distinct areas of the mucosa. These results shed new light on the processes underlying the B-cell response in celiac disease.

This dissertation has not only proposed and made available tools for streamlined AIRR-Seq analysis, but has applied these tools to gain meaningful biological insight into a variety of disease contexts.

5.2 Future Directions

Significant progress has been made in the development of AIRR-Seq in the past decade, but the field is still burgeoning, both in experimental protocols and creation of new and improved analysis tools. Currently, typical analysis workflows consist of a series of linear steps as seen in Figure 1.2. Some of the earlier steps could benefit from the knowledge resulting from later steps to improve

performance. Inference of the V(D)J germline alleles is one of the first steps after pre-processing of the raw sequencing reads. Inferences made by commonly used tools such as IgBLAST (Ye *et al.*, 2013) or IMGT/HighV-QUEST (Alamyar *et al.*, 2012) are computed based on alignment results of each sequence individually. However, knowledge of clonal groups in the data could help to improve alignment, as clonally related sequences would share the same germline V(D)J alleles. An initial rough prediction of germline genes — not necessarily even alleles — would be enough to group the sequences into B cell clones. A multiple alignment of clonal relatives with the germline alleles would lead to a more informed prediction than considering each sequence individually. In addition, the most recent common ancestor of the clonal lineage could be a better estimate of the unmutated Ig sequence originating the lineage than any of the observed sequences. This sequence, which can be statistically inferred (Kepler, 2013), could improve germline inference even further. Using more biological knowledge could lead to more confident D allele calls, which are currently difficult to infer accurately (Munshaw & Kepler, 2010). Improved D calls would enable better characterization of the CDR3 region such as which nucleotides stem from the N/P untemplated additions vs. the D gene. This additional knowledge could help in furthering understanding the formation of successful neutralizing antibodies.

Although Chapter 2 demonstrated that clustering can group clonally related Ig sequences with high confidence, it still calls thousands of erroneous clonal relationships in a typical dataset that can be corrected. Hierarchical clustering-based clonal grouping fails more frequently on Ig sequences with short junctions. Probabilistic models, while too computationally intensive to be feasible on entire AIRR-Seq datasets, may be able to improve performance on these short junctions. A hybrid approach combining clustering and probabilistic inference could improve clonal grouping performance relative to either approach alone. The hierarchical clustering-based clonal grouping algorithm outlined in Chapter 2 uses nucleotide Hamming distance metric between junctions, which assumes that all clonally related junctions are the same length. Recent research indicates that in a small percentage of cases, SHM may lead to insertions or deletions in the junction within a B cell clone (Yeap *et al.*, 2015). This and the possibility of sequencing errors should be taken into consideration to improve clonal grouping. A distance metric that allows for varying junction lengths such as Levenshtein distance may improve performance in these situations. This may be particularly relevant when using sequencing protocols that are known to have errors in the form of indels.

In addition to algorithmic refinements to clonal grouping, evaluation of performance can also

be improved. The simulated datasets used to evaluate sensitivity and PPV of clonal grouping in Chapter 2 were based on a limited number of underlying repertoire structures. The simulations also assume that Ig sequences maintain the same junction length during clonal evolution, which is not always the case (Yeap *et al.*, 2015). Recent advancements in experimental protocols enable labeling of individual B cells and tracking of clonal lineages (Tas *et al.*, 2016). The labeled clones can be isolated and sequenced to provide experimental gold standard datasets in which clonal relationships are known *a priori*. The datasets could then be used to evaluate performance of clonal grouping algorithms while avoiding the assumptions underlying simulations.

There is also untapped potential in biological application of these clonal grouping methods. Vaccination provides a unique opportunity for studying the immune response as the timeline is well-controlled and peak antibody response is known to be approximately seven days post-vaccination. Many studies of influenza vaccination have characterized changes in the Ig repertoire throughout the adaptive response (Y.-C. B. Wu *et al.*, 2012; Jiang, He, *et al.*, 2013; Vollmers *et al.*, 2013; Laserson *et al.*, 2014; Jackson *et al.*, 2014). The assumption is often made that measurable clonal expansions from day seven samples represent influenza-specific clones (Laserson *et al.*, 2014), but has not yet been verified experimentally. It is also established that influenza vaccination has lower efficacy in older individuals (Sasaki *et al.*, 2011). Previous studies have noted that older adults have fewer lineages that are more mutated relative to younger adults in response to vaccination (Jiang, He, *et al.*, 2013), however no such comparison has been made at the level of B cell clones. Comparison of mutation patters within expanded influenza-specific clones between responders and non-responders could provide insight into what contributes to a successful vaccination response.

Chapter 6

Acknowledgments

Completing this dissertation has been a long and at times arduous journey. I am immeasurably grateful for all of the help and support I have received along the way.

I would first like to express my sincerest gratitude to my advisor, Steven Kleinstein. I was drawn to Yale because of my interest in his work in computational immunology. Upon becoming a student, I learned of his reputation as a top-notch mentor. He helped me choose a project that best aligned with my interests, was available for weekly one-on-one meetings to discuss as much or as little as I desired, provided detailed feedback on all of my written and oral presentations, and fostered a wonderful lab environment. In addition to shaping the scientist I have become, he has truly guided my development into a working professional.

I would also like to thank my thesis committee members, Eric Meffre and Yuval Kluger for providing guidance in how best to accomplish the aims of my project. Yuval always had his door open and was willing to chat with me about anything from applied mathematics to coordinating calendars. I am also grateful to Chris Love for his support throughout our collaboration and during my job search and Kevin O'Connor for being genuinely kind and willing to offer advice.

I feel so fortunate to have been part of the Computational Biology and Bioinformatics program at Yale. In particular, I would like to extend a special thank you to our registrar, Lisa Sobel, for being a friend and counselor to me and helping me navigate the intricacies of graduate school administration on countless occasions. I am also thankful for the students of CBB, past and present, who helped keep me grounded and find my place in New Haven. Specifically, I'd like to thank Stefan Avey, Tara Bancroft (honorary CBB), Jieming Chen, Yuwei Cheng, Daniel Gadala-Maria,

Vijay Garla, Gadareth Higgs, Xiu Huang, Tingting Jiang, Mohammed Khan, Lucas Lochovsky, Thai Binh Luong, Mate Nagy, Emmett Sprecher, Jason Vander Heiden, and Gili Zilberman. The Kleinstein lab has always been a wonderful environment for having fun while working. I thank Rob Amezquita, Chris Bolen, Damian Fermin, Daniel Gadala-Maria, Susanna Marquez, Hailong Meng, Juilee Thakar, Mohamed Uduman, Jason Vander Heiden, and Gur Yaari for many fun-filled discussions about science, life, and how to debug each.

Outside of academia, I have been fortunate enough to have made deeply meaningful relationships that have been an unending source of love and emotional support. For this I extend my heartfelt gratitude to Deb Ayeni, Geetanjoli Banerjee, Meenakshi Chatterjee, Jieming Chen, Brian Coppedge, Mohammed Khan, Clark Reddy, and my boyfriend, Ryan Powles.

This dissertation is dedicated to my family, without whom none of this would have been possible. My brothers, Piyush and Gaorav Gupta, have been my idols since I was a little girl, always showing me my best possible self ten years in the future. I cannot thank them enough for loving me unconditionally and always having my back. My sisters-in-law, Charlotte Kuperwasser and Yuliya Pylayeva, have been my role models and shining examples of women in science. I am so grateful to them for being big sisters to me and for their love, support, and insight into how to become a successful professional woman myself. Finally, I thank my mother, who is such an inspiration to me and is the embodiment of the strong, successful woman I hope to become.

Namita Gupta

New Haven

2017

References

1. Adachi, O., Kawai, T., Takeda, K., Matsumoto, M., Tsutsui, H., Sakagami, M., Nakanishi, K. & Akira, S. Targeted Disruption of the MyD88 Gene Results in Loss of IL-1- and IL-18-Mediated Function. *Immunity* **9,** 143–150 (1998).

2. Ademokun, A., Wu, Y.-C. C., Martin, V., Mitra, R., Sack, U., Baxendale, H., Kipling, D. & Dunn-Walters, D. K. Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages. *Aging Cell* **10,** 922–30 (2011).

3. AGA Institute, M. F. AGA Institute Medical Position Statement on the Diagnosis and Management of Celiac Disease. *Gastroenterology* **131,** 1977–80 (2006).

4. Alamyar, E., Duroux, P., Lefranc, M. P. & Giudicelli, V. IMGTÂő 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* **882,** 569–604 (2012).

5. Arpaia, N., Godec, J., Lau, L., Sivick, K. E., McLaughlin, L. M., Jones, M. B., Dracheva, T., Peterson, S. N., Monack, D. M. & Barton, G. M. *TLR signaling is required for salmonella typhimurium virulence* **5,** 675–688 (Cell 144, 2011).

6. Barak, M., Zuckerman, N. S., Edelman, H., Unger, R. & Mehr, R. IgTree: creating Immunoglobulin variable region gene lineage trees. *Journal of Immunological Methods* **338,** 67 74 (2008).

7. Barr, T. a., Brown, S., Mastroeni, P. & Gray, D. *TLR and B cell receptor signals to B cells differentially program primary and memory Th1 responses to Salmonella enterica.* **5,** 2783–9 (J Immunol 185, 2010).

8. Bashford-Rogers, R. J. M., Palser, A. L., Huntly, B. J., Rance, R., Vassiliou, G. S., Follows, G. A. & Kellam, P. Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome research* **23,** 1874–1884 (2013).

9. Benckert, J., Schmolka, N., Kreschel, C., Zoller, M. J., Sturm, A., Wiedenmann, B. & Wardemann, H. The majority of intestinal IgA+ and IgG+ plasmablasts in the human gut are antigen-specific. *The Journal of Clinical Investigation* **121,** 1946–55 (2011).

10. Benichou, J., Ben-Hamo, R., Louzoun, Y. & Efroni, S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* **135,** 183–91 (2012).

11. Bobat, S., Flores-Langarica, A., Hitchcock, J., Marshall, J. L., Kingsley, R. A., Goodall, M., Gil-Cruz, C., Serre, K., Leyton, D. L., Letran, S. E., Gaspal, F., Chester, R., Chamberlain, J. L., Dougan, G., Lopez-Macias, C., Henderson, I. R., Alexander, J., Maclennan, I. C. M. & Cunningham, A. F. Soluble flagellin, FliC, induces an Ag-specific Th2 response, yet promotes T-bet-regulated Th1 clearance of Salmonella typhimurium infection. *European Journal of Immunology* **41,** 1606–1618 (2011).

12. Boletis, J. N., Marinaki, S., Skalioti, C., Lionaki, S. S., Iniotaki, A. & Sfikakis, P. P. Rituximab and mycophenolate mofetil for relapsing proliferative lupus nephritis: A long-term prospective study. *Nephrology Dialysis Transplantation* **24,** 2157–2160 (2009).

13. Boyd, S. D. & Joshi, S. A. High-Throughput DNA Sequencing Analysis of Antibody Repertoires. *Microbiology Spectrum* **2,** AID–0017–2014 (2014).

14. Boyd, S. D., Marshall, E. L., Merker, J. D., Maniar, J. M., Zhang, L. N., Sahaf, B., Jones, C. D., Simen, B. B., Hanczaruk, B., Nguyen, K. D., Nadeau, K. C., Egholm, M., Miklos, D. B., Zehnder, J. L. & Fire, A. Z. Measurement and Clinical Monitoring of Human Lymphocyte Clonality by Massively Parallel V-D-J Pyrosequencing. *Science Translational Medicine* **1,** 12ra23–12ra23 (2009).

15. Briney, B. *et al.* Clonify: unseeded antibody lineage assignment from next-generation sequencing data. *Scientific Reports* **6,** 23901 (2016).

16. Calderon, I., Lobos, S. R., Rojas, H. A., Palomino, C., RodrÃŋguez, L. H. & Mora, G. C. Antibodies to porin antigens of Salmonella typhi induced during typhoid infection in humans. *Infection and Immunity* **52,** 209–212 (1986).

17. Chen, Z., Collins, A. M., Wang, Y. & Gata, B. A. Clustering-based identification of clonally-related immunoglobulin gene sequence sets. *Immunome Research* **6,** S4 (2010).

18. Colpitts, T. M., Conway, M. J., Montgomery, R. R. & Fikrig, E. West Nile Virus: biology, transmission, and human infection. *Clinical Microbiology Review* **25,** 635–48 (2012).

19. Coro, E. S., Chang, W. L. W. & Baumgarth, N. Type I IFN receptor signals directly stimulate local B cells early following influenza virus infection. *Journal of Immunology* **176,** 4343–51 (2006).

20. Cortina-Ceballos, B., Godoy-Lozano, E. E., Samano-Sanchez, H., Aguilar-Salgado, A., Velasco-Herrera, M. D. C., Vargas-ChÃąvez, C., VelÃązquez-RamÃŋrez, D., Romero, G., Moreno, J., TÃľllez-Sosa, J. & MartÃŋnez-Barnetche, J. Reconstructing and mining the B cell repertoire with ImmunediveRsity. *mAbs* **7,** 516–524 (2015).

21. Cortina-Ceballos, B., Godoy-Lozano, E. E., TÃľllez-Sosa, J., Ovilla-MuÃśoz, M., SÃąmano-SÃąnchez, H., Aguilar-Salgado, A., GÃşmez-Barreto, R. E., Valdovinos-Torres, H., LÃşpez-MartÃŋnez, I., Aparicio-Antonio, R., RodrÃŋguez, M. H. & MartÃŋnez-Barnetche, J. Longitudinal analysis of the peripheral B cell repertoire reveals unique effects of immunization with a new influenza virus strain. *Genome Medicine* **7,** 124 (2015).

22. Csardi, G. N. T. The igraph software package for complex network research. *InterJournal Complex Systems* **1695,** 1695 (2006).

23. Cunningham, A. F., Gaspal, F., Serre, K., Mohr, E., Henderson, I. R., Scott-Tucker, A., Kenny, S. M., Khan, M., Toellner, K.-M., Lane, P. J. L. & MacLennan, I. C. M. *Salmonella induces a switched antibody response without germinal centers that impedes the extracellular spread of infection.* **10,** 6200–6207 (J Immunol 178, 2007).

24. Darlington, R. Is kurtosis really âĂIJpeakedness?âĂİ. *The American Statistician* (1970).

25. De Alwis, R., Smith, S. a., Olivarez, N. P., Messer, W. B., Huynh, J. P., Wahala, W. M. P. B., White, L. J., Diamond, M. S., Baric, R. S., Crowe, J. E. & de Silva, A. M. Identification of human neutralizing antibodies that bind to complex epitopes on dengue virions. *Proceedings of the National Academy of Sciences* **109,** 7439–7444 (2012).

26. De Alwis, R. & de Silva, A. M. Measuring antibody neutralization of dengue virus (DENV) using a flow cytometry-based technique. *Dengue* **1138,** 27–39 (2014).

27. De Jong, Y. P., Dorner, M., Mommersteeg, M. C., Xiao, J. W., Balazs, A. B., Robbins, J. B., Winer, B. Y., Gerges, S., Vega, K., Labitt, R. N., Donovan, B. M., Giang, E., Krishnan, A., Chiriboga, L., Charlton, M. R., Burton, D. R., Baltimore, D., Law, M., Rice, C. M. & Ploss, A. Broadly neutralizing antibodies abrogate established hepatitis C virus infection. *Science Translational Medicine* **6,** 254ra129–254ra129 (2014).

28. Dejnirattisai, W., Wongwiwat, W., Supasa, S., Zhang, X., Dai, X., Rouvinsky, A., Jumnainsong, A., Edwards, C., Quyen, N. T. H., Duangchinda, T., Grimes, J. M., Tsai, W.-y., Lai, C.-y., Wang, W.-k., Malasit, P., Farrar, J., Simmons, C. P., Zhou, Z. H., Rey, F. A., Mongkolsapaya, J. & Screaton, G. R. A new class of highly potent, broadly neutralizing antibodies isolated from viremic patients infected with dengue virus. *Nature Immunology* **16,** 170–177 (2014).

29. Di Niro, R., Lee, S.-J., Vander Heiden, J. A., Elsner, R. A., Trivedi, N., Bannock, J. M., Gupta, N. T., Kleinstein, S. H., Vigneault, F., Gilbert, T. J., Meffre, E., McSorley, S. J. & Shlomchik, M. J. Salmonella Infection Drives Promiscuous B Cell Activation Followed by Extrafollicular Affinity Maturation. *Immunity* **43,** 120–131 (2015).

30. Di Niro, R., Mesin, L., Zheng, N.-Y., Stamnaes, J., Morrissey, M., Lee, J.-H., Huang, M., Iversen, R., du PrÃl', M. F., Qiao, S.-W., Lundin, K. E. A., Wilson, P. C. & Sollid, L. M. High abundance of plasma cells secreting transglutaminase 2-specific IgA autoantibodies with limited somatic hypermutation in celiac disease intestinal lesions. *Nature Medicine* **18** (ed Nat) 441–5 (2012).

31. Dieterich, W., Ehnis, T., Bauer, M., Donner, P., Volta, U., Riecken, E. O. & Schuppan, D. Identification of tissue transglutaminase as the autoantigen of celiac disease. *Nature Medicine* **3,** 797–801 (1997).

32. Dougan, G., John, V., Palmer, S. & Mastroeni, P. *Immunity to salmonellosis* **1,** 196–210 (Immunological reviews 240, 2011).

33. Dowd, K. A. & Pierson, T. C. Antibody-mediated neutralization of flaviviruses: A reductionist view. *Virology* **411,** 306–315 (2011).

34. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32,** 1792–1797 (2004).

35. Felsenstein J. *PHYLIP (Phylogeny Inference Package) version 3.6.* Seattle, Washington, 2005.

36. Gadala-Maria, D., Yaari, G., Uduman, M. & Kleinstein, S. H. Automated analysis of high-throughput B cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proceedings of the National Academy of Sciences* **112,** E862–70 (2015).

37. Giudicelli, V., Chaume, D. & Lefranc, M.-P. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Research* **33,** D256–61 (2005).

38. Glanville, J., Kuo, T. C., von BÃijdingen, H.-C., Guey, L., Berka, J., Sundar, P. D., Huerta, G., Mehta, G. R., Oksenberg, J. R., Hauser, S. L., Cox, D. R., Rajpal, A. & Pons, J. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proceedings of the National Academy of Sciences* **108,** 20066–71 (2011).

39. Glanville, J., Zhai, W., Berka, J., Telman, D., Huerta, G., Mehta, G. R., Ni, I., Mei, L., Sundar, P. D., Day, G. M. R., Cox, D., Rajpal, A. & Pons, J. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proceedings of the National Academy of Sciences* **106,** 20216–21 (2009).

40. Gould, L. H., Sui, J., Foellmer, H., Oliphant, T., Wang, T., Ledizet, M., Murakami, A., Noonan, K., Lambeth, C., Kar, K., Anderson, J. F., de Silva, A. M., Diamond, M. S., Koski, R. a., Marasco, W. a. & Fikrig, E. Protective and therapeutic capacity of human single-chain Fv-Fc fusion proteins against West Nile virus. *Journal of Virology* **79,** 14606–13 (2005).

41. Guan, Y., Sajadi, M. M., Kamin-Lewis, R., Fouts, T. R., Dimitrov, A., Zhang, Z., Redfield, R. R., DeVico, A. L., Gallo, R. C. & Lewis, G. K. *Discordant memory B cell and circulating anti-Env antibody responses in HIV-1 infection.* **10,** 3952–3957 (Proceedings of the National Academy of Sciences of the United States of America 106, 2009).

42. Guidoum, A. kedd: Kernel estimator and bandwidth selection for density and its derivatives. *R package version 1.0.3* (2015).

43. Gupta, N. T., Vander Heiden, J. A., Uduman, M., Gadala-Maria, D., Yaari, G. & Kleinstein, S. H. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* **31,** 3356–3358 (2015).

44. Hammers, C. M. & Stanley, J. R. Antibody Phage Display: Technique and Applications. *The Journal of Investigative Dermatology* **134,** e17 (2014).

45. Hansen, B. Bandwidth selection for nonparametric distribution estimation. *manuscript, University of Wisconsin* (2004).

46. Hastey, C. J., Elsner, R. A., Barthold, S. W. & Baumgarth, N. *Delays and diversions mark the development of B cell responses to Borrelia burgdorferi infection.* **11,** 5612–5622 (J Immunol 188, 2012).

47. Hemmi, H., Takeuchi, O., Kawai, T., Kaisho, T., Sato, S., Sanjo, H., Matsumoto, M., Hoshino, K., Wagner, H., Takeda, K. & Akira, S. A Toll-like receptor recognizes bacterial DNA. *Nature* **408,** 740–5 (2000).

48. Hershberg, U. & Luning Prak, E. T. The analysis of clonal expansions in normal and autoimmune B cell repertoires. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **370,** 20140239– (2015).

49. Hershberg, U., Meng, W., Zhang, B., Haff, N., St Clair, E. W., Cohen, P. L., McNair, P. D., Li, L., Levesque, M. C. & Luning Prak, E. T. Persistence and selection of an expanded B-cell clone in the setting of rituximab therapy for Sjogren's syndrome. *Arthritis Research & Therapy* **16,** R51 (2014).

50. Hill, M. Diversity and evenness: a unifying notation and its consequences. *Ecology* **54,** 427–432 (1973).

51. Hoiseth, S. K. & Stocker, B. Aromatic-dependent Salmonella typhimurium are non-virulent and effective as live vaccines. *Nature* **291,** 238–239 (1981).

52. Horns, F., Vollmers, C., Croote, D., Mackey, S. F., Swan, G. E., Dekker, C. L., Davis, M. M. & Quake, S. R. Lineage tracing of human B cells reveals the in vivo landscape of human antibody class switching. *eLife* **5** (2016).

53. Jabri, B. & Sollid, L. M. Mechanisms of disease: immunopathogenesis of celiac disease. *Nature Clinical Practice. Gastroenterology & Hepatology* **3,** 516–525 (2006).

54. Jackson, K. J. L., Liu, Y., Roskin, K. M., Glanville, J., Hoh, R. A., Seo, K., Marshall, E. L., Gurley, T. C., Moody, M. A., Haynes, B. F., Walter, E. B., Liao, H.-X., Albrecht, R. A., GarcÃŋa-Sastre, A., Chaparro-Riggers, J., Rajpal, A., Pons, J., Simen, B. B., Hanczaruk,

B., Dekker, C. L., Laserson, J., Koller, D., Davis, M. M., Fire, A. Z. & Boyd, S. D. Human Responses to Influenza Vaccination Show Seroconversion Signatures and Convergent Antibody Rearrangements. *Cell Host & Microbe* **16,** 105–14 (2014).

55. Jain, A. K. & Dubes, R. C. *Algorithms for Clustering Data* (Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988).

56. Jiang, N., He, J., Weinstein, J. A., Penland, L., Sasaki, S., He, X.-S., Dekker, C. L., Zheng, N.-Y., Huang, M., Sullivan, M., Wilson, P. C., Greenberg, H. B., Davis, M. M., Fisher, D. S. & Quake, S. R. Lineage structure of the human antibody repertoire in response to influenza vaccination. *Science Translational Medicine* **5,** 171ra19 (2013).

57. Jiang, N., Weinstein, J. a., Penland, L., White, R. a., Fisher, D. S. & Quake, S. R. Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proceedings of the National Academy of Sciences* **108,** 5348–5353 (2011).

58. Johnson, T. A., Rassenti, L. Z. & Kipps, T. J. Ig VH1 genes expressed in B cell chronic lymphocytic leukemia exhibit distinctive molecular features. *Journal of Immunology* **158,** 235–46 (1997).

59. Kaukinen, K., PerÃd'aho, M., Collin, P., Partanen, J., Woolley, N., Kaartinen, T., Nuutinen, T., Halttunen, T., MÃd'ki, M. & Korponay-Szabo, I. Small-bowel mucosal transglutaminase 2-specific IgA deposits in coeliac disease without villous atrophy: a prospective and randomized clinical study. *Scandinavian Journal of Gastroenterology* **40,** 564–572 (2005).

60. Kepler, T. B. Reconstructing a B-cell clonal lineage. I. Statistical inference of unobserved ancestors. *F1000Research* **103,** 1–15 (2013).

61. Klein, F., Mouquet, H., Dosenovic, P., Scheid, J. F., Scharf, L. & Nussenzweig, M. C. Antibodies in HIV-1 vaccine development and therapy. *Science* **341,** 1199–204 (2013).

62. Kleinstein, S. H., Louzoun, Y. & Shlomchik, M. J. Estimating hypermutation rates from clonal tree data. *Journal of Immunology* **171,** 4639–4649 (2003).

63. Kong, W., Brovold, M., Koeneman, B. a., Clark-Curtiss, J. & Curtiss, R. Turning self-destructing Salmonella into a universal DNA vaccine delivery platform. *Proceedings of the National Academy of Sciences* **109,** 19414–9 (2012).

64. Korponay-Szabo, I. R., Halttunen, T., Szalai, Z., Laurila, K., Kiraly, R., Kovacs, J. B., Fesus, L. & Maki, M. In vivo targeting of intestinal and extraintestinal transglutaminase 2 by coeliac autoantibodies. *Gut* **53,** 641–648 (2004).

65. Krause, J. C., Tsibane, T., Tumpey, T. M., Huffman, C. J., Briney, B. S., Smith, S. a., Basler, C. F. & Crowe, J. E. Epitope-specific human influenza antibody repertoires diversify by B cell intraclonal sequence divergence and interclonal convergence. *Journal of immunology* **187,** 3704–3711 (2011).

66. Kuppers, R. in *B Cell Protocols* 225–238 (Humana Press, Totowa, NJ, 2004).

67. Kyle Austin, S. & Dowd, K. A. B cell response and mechanisms of antibody protection to west nile virus. *Viruses* **6,** 1015–1036 (2014).

68. Laserson, U. *et al.* High-resolution antibody dynamics of vaccine-induced immune responses. *Proceedings of the National Academy of Sciences* **111,** 4928–4933 (2014).

69. Lee, S., Liang, L., Juarez, S., Nanton, M. R., Gondwe, E. N. & Msefula, C. L. Identification of a common immune signature in murine and human systemic Salmonellosis. *Proceedings of the National Academy of Sciences* **109,** 4998–5003 (2012).

70. Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. & Law, M. Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology* **2012,** 251364 (2012).

71. Logan, A. C., Gao, H., Wang, C., Sahaf, B., Jones, C. D., Marshall, E. L., Buno, I., Armstrong, R., Fire, A. Z., Weinberg, K. I., Mindrinos, M., Zehnder, J. L., Boyd, S. D., Xiao, W., Davis, R. W. & Miklos, D. B. High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *Proceedings of the National Academy of Sciences* **108,** 21194–21199 (2011).

72. Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J. & Pallen, M. J. Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology* **30,** 434–9 (2012).

73. Looney, T. J., Lee, J. Y., Roskin, K. M., Hoh, R. A., King, J., Glanville, J., Liu, Y., Pham, T. D., Dekker, C. L., Davis, M. M. & Boyd, S. D. Human B-cell isotype switching origins of IgE. *Journal of Allergy and Clinical Immunology* **137,** 579–586 (2016).

74. Love, J. C., Ronan, J. L., Grotenbreg, G. M., van der Veen, A. G. & Ploegh, H. L. A microengraving method for rapid selection of single cells producing antigen-specific antibodies. *Nature Biotechnology* **24,** 703–707 (2006).

75. Marasco, W. A. & Sui, J. The growth and potential of human antiviral monoclonal antibody therapeutics. *Nature Biotechnology* **25,** 1421–1434 (2007).

76. Marzari, R., Sblattero, D., Florian, F., Tongiorgi, E., Not, T., Tommasini, a., Ventura, a. & Bradbury, a. Molecular dissection of the tissue transglutaminase antoantibody response in celiac disease. *Journal of Immunology* **166,** 4170–4176 (2001).

77. McGregor, A. C., Waddington, C. S. & Pollard, A. J. *Prospects for prevention of Salmonella infection in children through vaccination.* **3,** 254–62 (Current opinion in infectious diseases 26, 2013).

78. McKean, D., Huppi, K., Bell, M., Staudt, L., Gerhard, W. & Weigert, M. Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin. *Proceedings of the National Academy of Sciences* **81,** 3180–3184 (1984).

79. Metzker, M. L. Sequencing technologies - the next generation. *Nature Reviews Genetics* **11,** 31–46 (2010).

80. Molberg, O., Mcadam, S. N., Korner, R., Quarsten, H., Kristiansen, C., Madsen, L., Fugger, L., Scott, H., NorÃľn, O., Roepstorff, P., Lundin, K. E., Sjostrom, H. & Sollid, L. M. Tissue transglutaminase selectively modifies gliadin peptides that are recognized by gut-derived T cells in celiac disease. *Nature Medicine* **4,** 713–7 (1998).

81. Montgomery, R. R. & Murray, K. O. Risk factors for West Nile virus infection and disease in populations and individuals. *Expert Review of Anti-Infective Therapy* **13,** 1–9 (2015).

82. Moody, M. A. *et al.* H3N2 influenza infection elicits more cross-reactive and less clonally expanded anti-hemagglutinin antibodies than influenza vaccination. *PloS one* **6,** e25797 (2011).

83. Moorhouse, M. J., van Zessen, D., IJspeert, H., Hiltemann, S., Horsman, S., van der Spek, P. J., van der Burg, M. & Stubbs, A. P. ImmunoGlobulin galaxy (IGGalaxy) for simple determination and quantitation of immunoglobulin heavy chain rearrangements from NGS. *BMC Immunology* **15,** 59 (2014).

84. Moyron-Quiroz, J. E., Rangel-Moreno, J., Kusser, K., Hartson, L., Sprague, F., Goodrich, S., Woodland, D. L., Lund, F. E. & Randall, T. D. *Role of inducible bronchus associated lymphoid tissue (iBALT) in respiratory immunity* **9,** 927–934 (Nature medicine 10, 2004).

85. Munshaw, S. & Kepler, T. B. SoDA2: A Hidden Markov Model approach for identification of immunoglobulin rearrangements. *Bioinformatics* **26,** 867–872 (2010).

86. Murray, K. O., Garcia, M. N., Yan, C. & Gorchakov, R. Persistence of detectable immunoglobulin M antibodies up to 8 years after infection with West Nile virus. *American Journal of Tropical Medicine and Hygiene* **89,** 996–1000 (2013).

87. Murray, K., Walker, C., Herrington, E., Lewis, J. A., McCormick, J., Beasley, D. W. C., Tesh, R. B. & FisherâĂŘHoch, S. Persistent Infection with West Nile Virus Years after Initial Infection. *The Journal of Infectious Diseases* **201,** 2–4 (2010).

88. Nanton, M. R., Way, S. S., Shlomchik, M. J. & McSorley, S. J. *B cells are essential for Protective Immunity against Salmonella Independent of Antibody Secretion* **12,** 5503–5507. arXiv: NIHMS150003 (J Immunol 189, 2012).

89. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences* **76,** 5269–73 (1979).

90. Nei, M. & Kumar, S. *Molecular evolution and phylogenetics* 333 (Oxford University Press, 2000).

91. Neves, P., Lampropoulou, V., Calderon-Gomez, E., Roch, T., Stervbo, U., Shen, P., Kuhl, A. A., Loddenkemper, C., Haury, M., Nedospasov, S. A., Kaufmann, S. H. E., Steinhoff, U., Calado, D. P. & Fillatreau, S. Signaling via the MyD88 adaptor protein in B cells suppresses protective immunity during salmonella typhimurium infection. *Immunity* **33,** 777–790 (2010).

92. Niklas, N., PrÃűll, J., Weinberger, J., Zopf, A., Wiesinger, K., Krismer, K., Bettelheim, P. & Gabriel, C. Qualifying high-throughput immune repertoire sequencing. *Cellular immunology* **288,** 31–38 (2014).

93. Nolan, M. S., Schuermann, J. & Murray, K. O. West Nile Virus Infection among Humans, Texas, USA, 2002âĂŞ2011. *Emerging Infectious Diseases* **19,** 137 (2013).

94. Ogunniyi, A. O., Story, C. M., Papa, E., Guillen, E. & Love, J. C. Screening individual hybridomas by microengraving to discover monoclonal antibodies. *Nature Protocols* **4,** 767–782 (2009).

95. Ogunniyi, A. O., Thomas, B. A., Politano, T. J., Varadarajan, N., Landais, E., Poignard, P., Walker, B. D., Kwon, D. S. & Love, J. C. Profiling human antibody responses by integrated single-cell analysis. *Vaccine* **32,** 2866–2873 (2014).

96. Oliphant, T., Engle, M., Nybakken, G. E., Doane, C., Johnson, S., Huang, L., Gorlatov, S., Mehlhop, E., Marri, A., Chung, K. M., Ebel, G. D., Kramer, L. D., Fremont, D. H. & Diamond, M. S. Development of a humanized monoclonal antibody with therapeutic potential against West Nile virus. *Nature Medicine* **11,** 522–30 (2005).

97. Oliphant, T., Nybakken, G. E., Austin, S. K., Xu, Q., Bramson, J., Loeb, M., Throsby, M., Fremont, D. H., Pierson, T. C. & Diamond, M. S. Induction of epitope-specific neutralizing antibodies against West Nile virus. *Journal of Virology* **81,** 11828–39 (2007).

98. Onodera, T., Takahashi, Y., Yokoi, Y., Ato, M., Kodama, Y., Hachimura, S., Kurosaki, T. & Kobayashi, K. *Memory B cells in the lung participate in protective humoral immune responses to pulmonary influenza virus reinfection.* **7,** 2485–90 (Proceedings of the National Academy of Sciences of the United States of America 109, 2012).

99. Ortiz, V., Isibasi, A., Garcia-Ortigoza, E. & Kumate, J. *Immunoblot detection of class-specific humoral immune response to outer membrane proteins isolated from Salmonella typhi in humans with typhoid fever* **7,** 1640–1645 (Journal of clinical microbiology 27, 1989).

100. Parameswaran, P., Liu, Y., Roskin, K. M. M., Jackson, K. K. K. L., Dixit, V. P. P., Lee, J.-Y., Artiles, K. L., Zompi, S., Vargas, M. J. J., Simen, B. B. B., Hanczaruk, B., McGowan, K. R. R., Tariq, M. A. a., Pourmand, N., Koller, D., Balmaseda, A., Boyd, S. D. D., Harris, E. & Fire, A. Z. Z. Convergent Antibody Signatures in Human Dengue. *Cell Host & Microbe* **13,** 691–700 (2013).

101. Petersen, L. R., Carson, P. J., Biggerstaff, B. J., Custer, B., Borchardt, S. M. & Busch, M. P. Estimated cumulative incidence of West Nile virus infection in US adults, 1999âĂŞ2010. *Epidemiology and Infection* **141,** 1–5 (2012).

102. Picarelli, A., Maiuri, L., Frate, A., Greco, M., Auricchio, S. & Londei, M. Production of antiendomysial antibodies after in-vitro gliadin challenge of small intestine biopsy samples from patients with coeliac disease. *Lancet* **348,** 1065–1067 (1996).

103. Pierson, T. C., Xu, Q., Nelson, S., Oliphant, T., Nybakken, G. E., Fremont, D. & Diamond, M. S. The Stoichiometry of Antibody-Mediated Neutralization and Enhancement of West Nile Virus Infection. *Cell Host & Microbe* **1,** 135–145 (2007).

104. Qian, F., Thakar, J., Yuan, X., Nolan, M., Murray, K. O., Lee, W. T., Wong, S. J., Meng, H., Fikrig, E., Kleinstein, S. H., Thakar, J., Yuan, X., Wong, S. J., Meng, H., Fikrig, E., Kleinstein, S. H. & Montgomery, R. R. Immune Markers Associated with Host Susceptibility to Infection with West Nile Virus. *Viral Immunology* **27,** 39–47 (2014).

105. Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P. & Gu, Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13,** 341 (2012).

106. Racine, R., Jones, D. D., Chatterjee, M., McLaughlin, M., Macnamara, K. C. & Winslow, G. M. *Impaired germinal center responses and suppression of local IgG production during intracellular bacterial infection.* **9,** 5085–5093 (J Immunol 184, 2010).

107. Ralph, D. & Matsen, F. A. Likelihood-based inference of B-cell clonal families. *arXiv,* 1–46 (2016).

108. R-Core, T. *R: A Language and Environment for Statistical Computing* Vienna, Austria, 2015.

109. Robins, H. Immunosequencing: applications of immune repertoire deep sequencing. *Current Opinion in Immunology* (2013).

110. Rosenfeld, A. M., Meng, W., Luning Prak, E. T. & Hershberg, U. ImmuneDB: A system for the analysis and exploration of high-throughput adaptive immune receptor sequencing data. *Bioinformatics,* btw593 (2016).

111. Rostom, A., Dube, C., Cranney, A., Saloojee, N., Sy, R., Garritty, C., Sampson, M., Zhang, L., Yazdi, F., Mamaladze, V., Pan, I., MacNeil, J., Mack, D., Patel, D. & Moher, D. The diagnostic accuracy of serologic tests for celiac disease: a systematic review. *Gastroenterology* **128,** S38–S46 (2005).

112. Rothaeusler, K. & Baumgarth, N. B-cell fate decisions following influenza virus infection. *European Journal of Immunology* **40,** 366–377 (2010).

113. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20,** 53–65 (1987).

114. Safonova, Y., Lapidus, A. & Lill, J. IgSimulator: a versatile immunosequencing simulator. *Bioinformatics,* 1–2 (2015).

115. Salmi, T. T., Collin, P., Korponay-Szabo, I. R., Laurila, K., Partanen, J., Huhtala, H., KirÃąly, R., Lorand, L., Reunala, T., Maki, M. & Kaukinen, K. Endomysial antibody-negative coeliac disease: clinical characteristics and intestinal autoantibody deposits. *Gut* **55,** 1746–1753 (2006).

116. Sasaki, S. *et al.* Limited efficacy of inactivated influenza vaccine in elderly individuals is associated with decreased production of vaccine-specific antibodies. *The Journal of clinical investigation* **121,** 3109–19 (2011).

117. Scheid, J. F., Mouquet, H., Feldhahn, N., Seaman, M. S., Velinzon, K., Pietzsch, J., Ott, R. G., Anthony, R. M., Zebroski, H., Hurley, A., Phogat, A., Chakrabarti, B., Li, Y., Connors, M., Pereyra, F., Walker, B. D., Wardemann, H., Ho, D., Wyatt, R. T., Mascola, J. R., Ravetch, J. V. & Nussenzweig, M. C. Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals. *Nature* **458,** 636–640 (2009).

118. Scheid, J. F., Mouquet, H., Ueberheide, B., Diskin, R., Klein, F., Oliveira, T. Y. K., Pietzsch, J., Fenyo, D., Abadir, A., Velinzon, K., Hurley, A., Myung, S., Boulad, F., Poignard, P., Burton, D. R., Pereyra, F., Ho, D. D., Walker, B. D., Seaman, M. S., Bjorkman, P. J., Chait, B. T. & Nussenzweig, M. C. Sequence and Structural Convergence of Broad and Potent HIV Antibodies That Mimic CD4 Binding. *Science* **333,** 1633–1637 (2011).

119. Shapiro, G. S., Ellison, M. C. & Wysocki, L. J. Sequence-specific targeting of two bases on both DNA strands by the somatic hypermutation mechanism. *Molecular Immunology* **40,** 287–95 (2003).

120. Shiroguchi, K., Jia, T. Z., Sims, P. A. & Xie, X. S. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proceedings of the National Academy of Sciences* **109,** 1347–52 (2012).

121. Shlomchik, M. J. & Weisel, F. *Germinal center selection and the development of memory B and plasma cells* **1,** 52–63 (Immunological reviews 247, 2012).

122. Shugay, M., Britanova, O. V., Merzlyak, E. M., Turchaninova, M. a., Mamedov, I. Z., Tuganbaev, T. R., Bolotin, D. a., Staroverov, D. B., Putintseva, E. V., Plevova, K., Linnemann, C.,

Shagin, D., Pospisilova, S., Lukyanov, S., Schumacher, T. N. & Chudakov, D. M. Towards error-free profiling of immune repertoires. *Nature Methods* **11,** 653–5 (2014).

123. Singh, S. P., Upshaw, Y., Abdullah, T., Singh, S. R. & Klebba, P. E. *Structural relatedness of enteric bacterial porins assessed with monoclonal antibodies to Salmonella typhimurium OmpD and OmpC* **6,** 1965–1973 (Journal of bacteriology 174, 1992).

124. Smith, D. S., Creadon, G., Jena, P. K., Portanova, J. P., Kotzin, B. L. & Wysocki, L. J. Di- and trinucleotide target preferences of somatic mutagenesis in normal and autoreactive B cells. *Journal of Immunology* **156,** 2642–52 (1996).

125. Sollid, L. M. Molecular basis of celiac disease. *Annual Review of Immunology* **18,** 53–81 (2000).

126. Sonoda, E. *et al.* B cell development under the condition of allelic inclusion. *Immunity* **6,** 225–33 (1997).

127. Statsoft, I. *Electronic Statistics Textbook* Tulsa, OK, 2013.

128. Stern, J. N. H., Yaari, G., Vander Heiden, J. A., Church, G. M., Donahue, W. F., Hintzen, R. Q., Huttner, A. J., Laman, J. D., Nagra, R. M., Nylander, A., Pitt, D., Ramanan, S., Siddiqui, B. A., Vigneault, F., Kleinstein, S. H., Hafler, D. A. & O'Connor, K. C. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Science Translational Medicine* **6,** 248ra107–248ra107 (2014).

129. Sugai, E., Nachman, F., VÃąquez, H., GonzÃąlez, A., Andrenacci, P., Czech, A., Niveloni, S., Mazure, R., Smecuol, E., Cabanne, A., MauriÃśo, E. & Bai, J. C. Dynamics of celiac disease-specific serology after initiation of a gluten-free diet and use in the assessment of compliance with treatment. *Digestive and Liver Disease* **42,** 352–358 (2010).

130. Suthar, M. S., Diamond, M. S. & Gale, M. West Nile virus infection and immunity. *Nature Reviews Microbiology* **11,** 115–28 (2013).

131. Sutterwala, F. S. *et al.* Critical role for NALP3/CIAS1/Cryopyrin in innate and adaptive immunity through its regulation of caspase-1. *Immunity* **24,** 317–27 (2006).

132. Tabibian-Keissar, H., Hazanov, L., Schiby, G., Rosenthal, N., Rakovsky, A., Michaeli, M., Shahaf, G. L., Pickman, Y., Rosenblatt, K., Melamed, D., Dunn-Walters, D., Mehr, R. & Barshack, I. Aging affects B-cell antigen receptor repertoire diversity in primary and secondary lymphoid tissues. *European Journal of Immunology,* n/a–n/a (2015).

133. Tas, J. M. J., Mesin, L., Pasqual, G., Targ, S., Jacobsen, J. T., Mano, Y. M., Chen, C. S., Weill, J.-C., Reynaud, C.-A., Browne, E. P., Meyer-Hermann, M. & Victora, G. D. Visualizing antibody affinity maturation in germinal centers. *Science (New York, N.Y.)* **351,** 1048–54 (2016).

134. Throsby, M., Geuijen, C., Goudsmit, J., Bakker, A. Q., Korimbocus, J., Kramer, R. A., Clijsters-van der Horst, M., de Jong, M., Jongeneelen, M., Thijsse, S., Smit, R., Visser, T. J., Bijl, N., Marissen, W. E., Loeb, M., Kelvin, D. J., Preiser, W., ter Meulen, J. & de Kruif, J. Isolation and characterization of human monoclonal antibodies from individuals infected with west nile virus. *Journal of Virology* **80,** 6982–6992 (2006).

135. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63,** 411–423 (2001).

136. Tiller, T., Busse, C. E. & Wardemann, H. Cloning and expression of murine Ig genes from single B cells. *Journal of Immunological Methods* **350,** 183–193 (2009).

137. Tiller, T., Meffre, E., Yurasov, S., Tsuiji, M., Nussenzweig, M. C. & Wardemann, H. Efficient generation of monoclonal antibodies from single human B cells by single cell RT-PCR and expression vector cloning. *Journal of Immunological Methods* **329,** 112–124 (2008).

138. Tosco, A., Maglio, M., Paparo, F., Rapacciuolo, L., Sannino, A., Miele, E., Barone, M. V., Auricchio, R. & Troncone, R. Immunoglobulin A anti-tissue transglutaminase antibody deposits in the small intestinal mucosa of children with no villous atrophy. *J Pediatr Gastroenterol Nutr* **47,** 293–298 (2008).

139. Tsioris, K., Gupta, N. T., Ogunniyi, A. O., Zimnisky, R. M., Qian, F., Yao, Y., Wang, X., Stern, J. N. H., Chari, R., Briggs, A. W., Clouser, C. R., Vigneault, F., Church, G. M., Garcia, M. N., Murray, K. O., Montgomery, R. R., Kleinstein, S. H. & Love, J. C. Neutralizing antibodies against West Nile virus identified directly from human B cells by single-cell analysis and next generation sequencing. *Integrative Biology* **7,** 1587–1597 (2015).

140. Uduman, M., Shlomchik, M. J., Vigneault, F., Church, G. M. & Kleinstein, S. H. Integrating B cell lineage information into statistical tests for detecting selection in Ig sequences. *Journal of Immunology* **192,** 867–74 (2014).

141. Uduman, M., Yaari, G., Hershberg, U., Stern, J. A., Shlomchik, M. J. & Kleinstein, S. H. Detecting selection in immunoglobulin sequences. *Nucleic Acids Research* **39,** W499–504 (2011).

142. Van Dongen, J. J. M., Langerak, A. W., Bruggemann, M., Evans, P. A. S., Hummel, M., Lavender, F. L., Delabesse, E., Davi, F., Schuuring, E., Garcia-Sanz, R., van Krieken, J. H. J. M., Droese, J., Gonzalez, D., Bastard, C., White, H. E., Spaargaren, M., Gonzalez, M., Parreira, A., Smith, J. L., Morgan, G. J., Kneba, M. & Macintyre, E. A. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia* **17,** 2257–2317 (2003).

143. Vander Heiden, J. A., Yaari, G., Uduman, M., Stern, J. N. H., O'Connor, K. C., Hafler, D. A., Vigneault, F. & Kleinstein, S. H. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* **30,** 1930–1932 (2014).

144. Vogt, M. R., Moesker, B., Goudsmit, J., Jongeneelen, M., Austin, S. K., Oliphant, T., Nelson, S., Pierson, T. C., Wilschut, J., Throsby, M. & Diamond, M. S. Human monoclonal antibodies against West Nile virus induced by natural infection neutralize at a postattachment step. *Journal of Virology* **83,** 6494–507 (2009).

145. Vollmers, C., Sit, R. V., Weinstein, J. A., Dekker, C. L. & Quake, S. R. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proceedings of the National Academy of Sciences* **110,** 13463–8 (2013).

146. Volpe, J. M. & Kepler, T. B. Large-scale analysis of human heavy chain V(D)J recombination patterns. *Immunome Research* **4,** 3 (2008).

147. Wal, Y. V. D., Kooy, Y., Veelen, P. V., PeÃśa, S., Mearin, L., Papadopoulos, G., Wal, Y. V. D., Kooy, Y., Veelen, P. V. & Pen, S. Cutting Edge: Selective Deamidation by Tissue Transglutaminase Strongly Enhances Gliadin-Specific T Cell Reactivity. *Journal of Immunology* **161,** 1585–8 (2016).

148. Wand, M. P. & Jones, M. C. *Kernel Smoothing, Vol. 60 of Monographs on statistics and applied probability* (Chapman and Hall, London, 1995).

149. Wang, C., Liu, Y., Xu, L. T., Jackson, K. J. L., Roskin, K. M., Pham, T. D., Laserson, J., Marshall, E. L., Seo, K., Lee, J.-Y., Furman, D., Koller, D., Dekker, C. L., Davis, M. M., Fire, A. Z. & Boyd, S. D. Effects of aging, cytomegalovirus infection, and EBV infection on human B cell repertoires. en. *Journal of Immunology* **192,** 603–11 (2014).

150. Wang, X. & Stollar, B. Human immunoglobulin variable region gene analysis by single cell RT-PCR. *Journal of Immunological Methods* **244,** 217–225 (2000).

151. Weinstein, J. J. a., Jiang, N., White, R. a. R., Fisher, D. D. S. & Quake, S. R. S. High-throughput sequencing of the zebrafish antibody repertoire. *Science (New York. N.Y.)* **324,** 807–10 (2009).

152. William, J., Euler, C., Christensen, S. & Shlomchik, M. J. *Evolution of Autoantibody Responses via Somatic Hypermutation Outside of Germinal Centers* **5589,** 2066–2070 (Science 297, 2002).

153. Wrammert, J., Koutsonanos, D., Li, G.-M., Edupuganti, S., Sui, J., Morrissey, M., McCausland, M., Skountzou, I., Hornig, M., Lipkin, W. I., Mehta, A., Razavi, B., Del Rio, C., Zheng, N.-Y., Lee, J.-H., Huang, M., Ali, Z., Kaur, K., Andrews, S., Amara, R. R., Wang, Y., Das, S. R., O'Donnell, C. D., Yewdell, J. W., Subbarao, K., Marasco, W. a., Mulligan, M. J., Compans, R., Ahmed, R. & Wilson, P. C. Broadly cross-reactive antibodies dominate the human B cell response against 2009 pandemic H1N1 influenza virus infection. *The Journal of experimental medicine* **208,** 181–193 (2011).

154. Wu, Y.-C. B., Kipling, D. & Dunn-Walters, D. K. Age-Related Changes in Human Peripheral Blood IGH Repertoire Following Vaccination. *Frontiers in Immunology* **3,** 193 (2012).

155. Wu, Y.-C., Kipling, D., Leong, H. S., Martin, V., Ademokun, A. a. & Dunn-Walters, D. K. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* **116,** 1070–8 (2010).

156. Wu, Y. C. B., James, L. K., Vander Heiden, J. A., Uduman, M., Durham, S. R., Kleinstein, S. H., Kipling, D. & Gould, H. J. Influence of seasonal exposure to grass pollen on local and peripheral blood IgE repertoires in patients with allergic rhinitis. English. *Journal of Allergy and Clinical Immunology* **134,** 604–612 (2014).

157. Xochelli, A., Agathangelidis, A., Kavakiotis, I., Minga, E., Sutton, L. A., Baliakas, P., Chouvarda, I., Giudicelli, V., Vlahavas, I., Maglaveras, N., Bonello, L., Trentin, L., Tedeschi, A., Panagiotidis, P., Geisler, C., Langerak, A. W., Pospisilova, S., Jelinek, D. F., Oscier, D., Chiorazzi, N., Darzentas, N., Davi, F., Ghia, P., Rosenquist, R., Hadzidimitriou, A., Belessi, C., Lefranc, M.-P. & Stamatopoulos, K. Immunoglobulin heavy variable (IGHV) genes and alleles: new entities, new names and implications for research and prognostication in chronic lymphocytic leukaemia. *Immunogenetics,* 61–66 (2014).

158. Yaari, G. & Kleinstein, S. H. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Medicine* **7,** 121 (2015).

159. Yaari, G., Uduman, M. & Kleinstein, S. H. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. *Nucleic Acids Research* **40,** e134 (2012).

160. Yaari, G., Vander Heiden, J. A., Uduman, M., Gadala-Maria, D., Gupta, N., Joel, J. N., O'Connor, K. C., Hafler, D. A., Laserson, U., Vigneault, F. & Kleinstein, S. H. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Frontiers in Immunology* **4,** 358 (2013).

161. Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research* **41,** 1–7 (2013).

162. Yeap, L.-S., Hwang, J. K., Du, Z., Meyers, R. M., Meng, F.-L., JakubauskaitÄŰ, A., Liu, M., Mani, V., Neuberg, D., Kepler, T. B., Wang, J. H. & Alt, F. W. Sequence-Intrinsic Mechanisms that Target AID Mutational Outcomes on Antibody Genes. *Cell* **163,** 1124–1137 (2015).

163. Yu, X., Tsibane, T., McGraw, P. A., House, F. S., Keefer, C. J., Hicar, M. D., Tumpey, T. M., Pappas, C., Perrone, L. A., Martinez, O., Stevens, J., Wilson, I. A., Aguilar, P. V., Altschuler, E. L., Basler, C. F. & Crowe, J. E. Neutralizing antibodies derived from the B cells of 1918 influenza pandemic survivors. *Nature* **455,** 532–6 (2008).

164. Yu, Y., Ceredig, R. & Seoighe, C. LymAnalyzer: A tool for comprehensive analysis of next generation sequencing data of T cell receptors and immunoglobulins. *Nucleic Acids Research* **44,** e31 (2016).

165. Zhu, J., Ofek, G., Yang, Y., Zhang, B., Louder, M. K., Lu, G., McKee, K., Pancera, M., Skinner, J., Zhang, Z., Parks, R., Eudailey, J., Lloyd, K. E., Blinn, J., Alam, S. M., Haynes, B. F., Simek, M., Burton, D. R., Koff, W. C., Mullikin, J. C., Mascola, J. R., Shapiro, L. & Kwong, P. D. 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* **110,** 6470–5 (2013).

166. Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R. & Siebert, P. D. Reverse transcriptase template switching: A SMART approach for full-length cDNA library construction. *BioTechniques* **30,** 892–897 (2001).