The Dissertation Committee for Brandon James DeKosky certifies that this is the approved version of the following dissertation:

Decoding the Antibody Repertoire: High Throughput Sequencing of Multiple Transcripts from Single B Cells

Committee:

George Georgiou, Supervisor

Andy Ellington

Lydia Contreras

Lauren Ehrlich

Jennifer Maynard

Decoding the Antibody Repertoire: High Throughput Sequencing of Multiple Transcripts from Single B Cells

by

Brandon James DeKosky, B.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2015

Dedication

To my parents, Deborah and Robert, and my siblings, Aaron and Bryn.

Thank you for all of your love and support, for which I am forever grateful.

Acknowledgements

First and foremost I would like to thank my research advisor, George Georgiou, whose constant encouragement and support, advice, wisdom, patience, and compassion have been essential for my development as a person and as a scientist. I would also like to thank several other research mentors at UT, including Andy Ellington, Grant Willson, Brent Iverson, and Greg Ippolito for sharing their valuable time, knowledge, and experience to help out a young graduate student. Thank you as well to all of my fellow lab members and research collaborators at UT and elsewhere, including Ryan Deschner, Jason Lavinder, Jon McDaniel, Hidetaka Tanno, Erik Johnson, Oana Lungu, Kam Hon Hoi, Takaaki Kojima, Jiwon Lee, Alexa Rodin, Wissam Charab, Sebastian Schätzle, Costas Chrystostomou, Daechan Park, Joe Taft, Mark Gebhard, Mark Pogson, Scott Kerr, Yariv Wine, Bing Tan, Ellen Wirth, Megan Murrin, Moses Donkor, Kendra Garrison, Candice Lamb, Nicholas Marshall, Elizabeth Marshall, Johnny Blazeck, Peter Allen, Christien Kluwe, Danny Boutz, Andrew Horton, Brandon Rawlings, Bo Wang, Alec Rezigh, and Paula Koziol. Thanks especially to Sai Reddy for training me during my first days in the lab, and thank you to the rest of the BIGG lab members for all of your helps, tips, tricks, and time at the Institute, and for making the lab such an incredibly positive and exciting place. I will remain forever grateful to Scott Hunicke-Smith at the UT GSAF for teaching me the basics of scripting for bioinformatic analysis, and to Scott Hunicke-Smith, Jessica Wheeler, and Dhivya Arassapan for all the high-throughput sequencing you have done for us. Finally, thank-you to those who provided key advice and collaboration from afar, including Cory Berkland, Milind Singh, Nathan Dormer,

John Mascola, Peter Kwong, Danny Douek, Jeff Gray, Dennis Burton, and Patrick Wilson. Thank you to all my family and friends, who have kept me sane throughout all the ups and downs in the research lab and beyond. Thank you to the faculty who served as members of my thesis committee (George Georgiou, Andy Ellington, Lydia Contreras, Lauren Ehrlich, and Jennifer Maynard) for your help and advice on the committee. Finally, thank you to my funding sources, the Donald D. Harrington Foundation, the Hertz Foundation, the Cockrell School of Engineering, and the National Science Foundation for their generous support of my research.

Decoding the Antibody Repertoire: High Throughput Sequencing of Multiple Transcripts from Single B Cells

Brandon James DeKosky, Ph.D.
The University of Texas at Austin, 2015

Supervisor: George Georgiou

Next-generation (high throughput) DNA sequencing of immunoglobulin variable region and T-cell receptor gene repertoires is providing critical information for understanding adaptive immune responses and for diagnostic and therapeutic applications[1–4]. However, existing immune repertoire sequencing technologies yield data on only one of the two chains of immune receptors and thus cannot provide information on the identity of immune receptor pairs encoded by individual B or T lymphocytes[5–7]. This work directly addressed these limitations by developing two new technologies for sequencing the complementary DNA (cDNA) of multiple mRNA transcripts from isolated single cells with very high throughput. In these methods, cells are sequestered into individual compartments and lysed *in situ* to capture single-cell mRNA onto magnetic beads, then the magnetic beads are used as template for RT-PCR reactions inside emulsion droplets that physically link cDNA of multiple transcripts for subsequent analysis by high-throughput DNA sequencing. Experimental throughput of over $2x10^{6}$ cells in a single day with antibody heavy and light chain pairing accuracy greater than 97% was demonstrated with *in vitro* expanded human B cells. These new single-cell sequencing technologies were also applied for rapid discovery of new human antibodies and for analysis of the human immune response to vaccination. Finally we applied the techniques developed here to gain new insights regarding development of the antibody repertoire through high-throughput and high-resolution examination of naïve and memory B-cell compartments in healthy human donors.

Table of Contents

List of Tables

List of Figures

Abbreviations

APC: Antigen presenting cell

BCR: B-cell receptor

bNAb: Broadly neutralizing antibody

CHO: Chinese hamster ovary

DC: Dendritic cell

ELISA: Enzyme-linked immunosorbent assay

ELISPOT: Enzyme-linked immunosorbent spot assay

FACS: Fluorescence activated cell sorting

Fc: Fragment crystallizable

HEp-2: Human epithelial cell line 2

IgA: Immunoglobulin A

IgG: Immunoglobulin G

IgM: Immunoglobulin M

LPS: Lipopolysaccharide

MHC: Major histocompatibility complex

PBMC: Peripheral blood mononuclear cell

PCR: Polymerase chain reaction

RT-PCR: Reverse transcription polymerase chain reaction

SASA: Solvent-accessible surface area

scFv: Single chain fragment variable

SHM: Somatic hypermutation

TCR: T-cell receptor

VH: Heavy chain variable region

VL: Light chain (kappa or lambda) variable region

1. BACKGROUND

1.1 Antibodies and Antibody Repertoire Development

Antibodies are Y-shaped molecules produced by the immune system of many vertebrates to recognize and neutralize foreign pathogens with high specificity. An antibody molecule is comprised of two major portions: the constant region (or Fc region), which is identical among different antibody molecules having different specificities, and the variable region, which comprises a unique sequence for each antibody and is responsible for antigen recognition (Figure 1.1). Different antibody variable regions confer different antibody binding specificities, and binding specificity is encoded by the unique antibody amino acid sequence (which in turn is encoded by the respective immunoglobulin genes). Binding to molecular targets occurs at the exposed ends of the variable region.

The antibody variable region can be further subdivided into framework regions and complementarity-determining regions (Figure 1.1). The relatively conserved framework regions (FRs) consist of antiparallel β strands which form a β-sandwich structure called the immunoglobulin fold[8]. Within the antibody variable region are several areas of higher variability which are termed the complementarity-determining regions (CDRs), or hypervariable loops. The CDRs contain much higher variation across different antibodies than the more conserved FRs. The precise area of the variable region where antibody binding occurs is called the paratope, and the complementary binding region on an antibody’s molecular target is termed the epitope. CDRs often comprise the antibody paratope.

In addition to the specificity conferred by the antibody variable region, antibody functionality is further determined by the type of antibody constant region (Figure 1.1), termed the antibody class or isotype. Five major classes of antibody constant regions exist (IgD, IgM, IgG, IgA, and IgE), and several contain subclasses as well (e.g. IgG1, IgG2, IgA1, IgA2, etc.) Each antibody class and sub-class performs different functions in terms of conferring protection against pathogen challenge. For example, IgM is expressed early in B cell development, is secreted as a pentamer, and excels at activating immune complement (or the killing of pathogenic cells via activation of innate immune response mechanisms), whereas human IgG is expressed later in B cell development, is secreted as a monomer and is a less potent activator of complement. An overview of antibody isotypes is presented in Table 1.1.

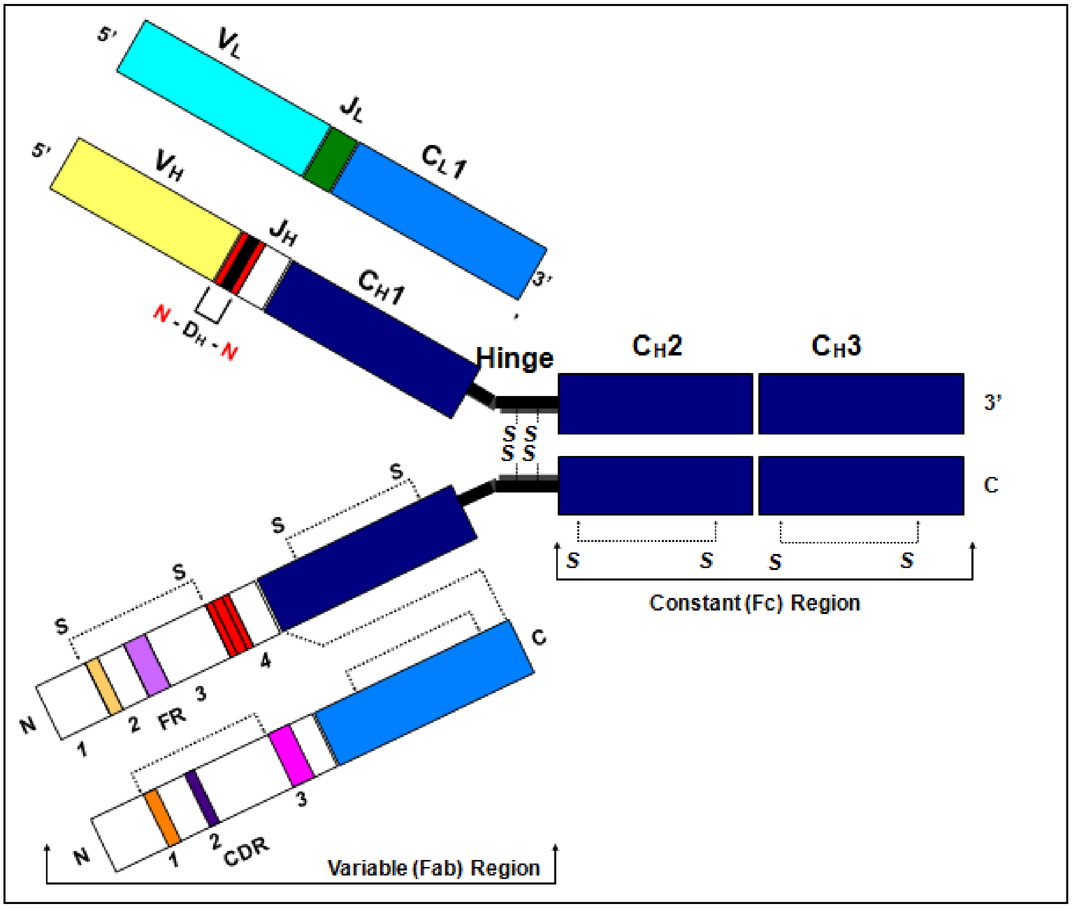


Figure 1.1 An overview of antibody structure. Antibodies are Y-shaped molecules which consist of a variable region that imparts binding specificity and a constant region that determines the functionality of the antibody molecule. The variable region is further subdivided into framework regions (FRs) which are somewhat conserved across antibodies, as well as complimentarity-determining regions (CDRs) which are highly variable across different antibodies. The variable and constant regions are comprised of both heavy chain and light chain genes. The heavy chain variable region is comprised of the recombined V_H-D_H-J_H genes, whereas the light chain variable region is composed of recombined V_L-J_L genes. Reproduced with permission of the Nature Publishing Group[2].

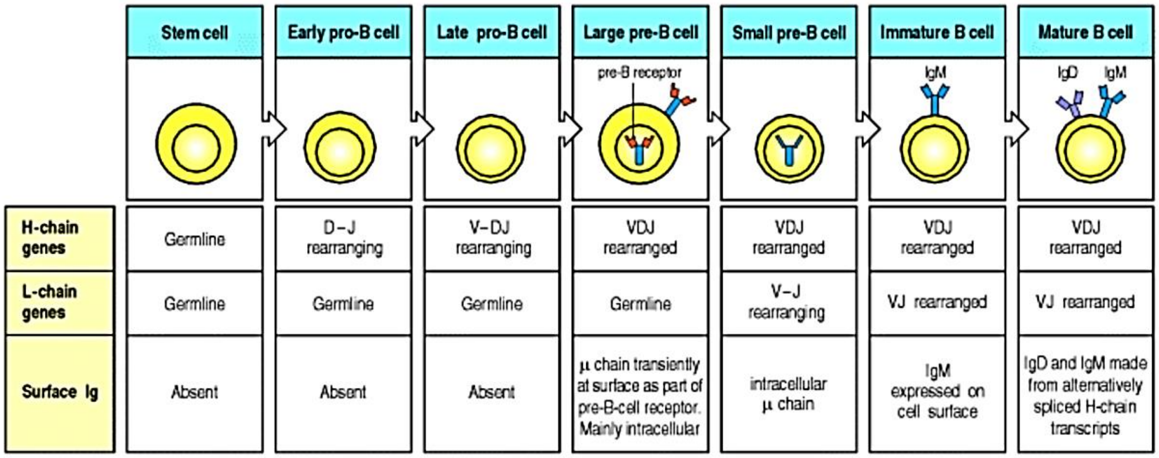
Isotype	**Secreted form**	**Molecular weight (kDa)**	**Serum conc. (mean adult mg/mL)**	**Serum half-life (days)**
IgD	monomer	188	0.03	3
IgM	pentamer	970	1.5	10
IgG1	monomer	146	9	21
IgG2	monomer	146	3	20
IgG3	monomer	165	1	7
IgG4	monomer	146	0.5	21
IgA1	dimer	160	3.0	6
IgA2	dimer	160	0.5	6
IgE	monomer	188	5.E-05	2

Table 1.1 Overview of antibody constant regions, adapted from[9].

Membrane-bound B-cell receptors (BCR), the form of antibodies that is expressed on the surface of B cells, are assembled via somatic recombination of antibody V-, (D-,) and J-genes within developing B cells in the bone marrow (Figures 1.1 and 1.2)[2,9]. Antibody variable region gene recombination is mediated by the recombination activating genes (RAG-1 and RAG-2). The RAG enzymes recognize nucleotide patterns in the genome known as recombination signal sequences (RSS)[9]. Expression of RAG genes is highest during the pro-B and pre-B stages when V-(D)-J recombination occurs. RAG recombination can also generate junctional diversity at single-stranded DNA break sites via p-nucleotide addition and junctional trimming. Another critical enzyme for B cell development is terminal deoxynucelotidyl transferase (TdT), which is a specialized polymerase that is expressed predominantly during heavy chain rearrangement at the pro-B stage, and some expression can also occur in pre-B cells. TdT adds non-templated nucleotides to the D-J and V-J junctions, thereby dramatically increasing the junctional diversity formed by recombination. The higher expression of TdT during heavy chain formation causes much broader variation in the length of heavy chain CDR3 loops (which comprise the V_H-D_H-J_H genes) because TdT adds a random number of nucleotides in each junction (typically, 0-30 nt per junction). Low expression of TdT in pre-B cells and the lack of a D-gene causes light chain CDR3 loop lengths to be fairly restricted[9].

The B cell receptor heavy chain variable region is comprised of three genes: a V_H gene, D_H gene, and a J_H gene, and the site of heavy chain recombination is fully contained by the CDR-H3 (Figure 1.1). In contrast, the light chain comprises a V_L gene and a J_L gene, and the V_L-J_L junction is contained within the CDR-L3 (Figure 1.1). After a stem cell differentiates into a pro-B cell, the pro-B cell rearranges first the heavy chain D_H-J_H junction, then recombines the D_H-J_H product with a V_H-gene to form the V_H-D_H-J_H region, with all recombination events mediated by RAG expression. Recombination is stopped when the newly formed heavy chain paired with a surrogate light chain are both expressed on the cell surface. (The surrogate light chain covers the newly-generated heavy chain variable region to prevent aberrant B-cell receptor binding.) This important step ensures that a productive, in-frame heavy chain junction without stop codons has been generated because RAG and TdT have no innate mechanism to maintain amino acid reading frames and thus only one in every three V-(D-)J recombinations will yield a productive sequence. After surface expression of the heavy chain, TdT expression sharply decreases and the B cell is now denoted as a pre-B cell. At this stage, pre-B cells with newly formed heavy chains undergo a pre-B expansion that consists of up to eight cell divisions that create multiple B cell clones with the same heavy chain[10] (observed directly in mice, inferred in humans). Next, light chain V_L-J_L recombination occurs in the expanded pre-B cells in a similar manner as heavy chains recombined using the RAG enzyme, but with very low levels of TdT expression that lead to generally restricted light chain CDR-L3 lengths[9]. As in pro-B cells, the final checkpoint to continue development after the pre-B stage consists of surface immunoglobulin display, this time with the somatically generated heavy and light chains together. RAG expression and light chain recombination ceases, and the cell is termed an immature B cell.

Next, immature B cells undergo an immune tolerance checkpoint to ensure that autoreactive antibodies do not continue in the antibody maturation pathway, as the process of random gene recombination without functional selection can generate self-reactive proteins. Tolerance checkpoints are thought to occur by clonally deleting B cells encoding BCRs that bind to self-proteins, or alternatively, receptor editing can rescue self-reactive B-cells at this stage[11]. A known hallmark of autoreactivity is the expression of a very long heavy chain CDR-H3 loop, and BCRs with long CDR-H3s are preferentially deleted from the repertoire during the pre-B tolerance checkpoint[12]. After passing tolerance checkpoints, the B-cell expresses both IgD and IgM constant regions as B-cell receptors on the cell surface and is considered a mature, or naïve, B cell. The newly generated mature B cell is now ready for positive selection in germinal center reactions of the spleen and lymph nodes in the course of an immune response.



	Stem cell	Early pro-B cell	Late pro-B cell	Large pre-B cell	Small pre-B cell	Immature B cell	Mature B cell
H-chain genes	Germline	D–J rearranging	V–DJ rearranging	VDJ rearranged	VDJ rearranged	VDJ rearranged	VDJ rearranged
L-chain genes	Germline	Germline	Germline	Germline	V–J rearranging	VJ rearranged	VJ rearranged
Surface Ig	Absent	Absent	Absent	μ chain transiently at surface as part of pre-B-cell receptor. Mainly intracellular	intracellular μ chain	IgM expressed on cell surface	IgD and IgM made from alternatively spliced H-chain transcripts

Figure 1.2 The sequential stages of B cell development. Central immune tolerance checkpoint occurs after immature B cells are formed, and autoreactive B cells are clonally deleted or edited before continuing to become mature (naïve) B cells. Copyright 2011 from Janeway's Immunobiology, 8th Edition by Murphy. Reproduced by permission of Garland Science/Taylor & Francis LLC[9].

1.2 Immune responses lead to B-cell activation and antibody secretion

One of the first steps in the immune response is the ingestion of pathogenic cells by macrophages via macro-pinocytosis which then migrate to the spleen and tissue-draining lymph nodes (collectively called the secondary lymphoid organs)[9]. There, macrophages mature into specialized antigen-presenting cells (APCs) named follicular dendritic cells that present captured foreign proteins to B- and T-cell surveillance. During follicular dendritic cell antigen presentation, peptides from digested foreign proteins are presented on the cell surface in the major histocompatibility complex (MHC), a specialized cell surface protein for display of peptides to T cells. T cells survey MHC using their somatically generated T-cell receptors, or TCR. TCR are formed in a very similar process to BCR, with the notable difference that T cells do not normally express AID and therefore TCR do not exhibit somatic hypermutation or class-switch recombination[9,13]. If a T cell's TCR binds tightly to a peptide presented on MHC by APCs during the course of infection, the T cell will be given signals to expand and proliferate by the follicular dendritic cell. Known activating signals include a combination of the secreted cytokines like IL-6, IL-12, and TGF-β, and especially the B7 co-receptors which bind to CD28 (activating) and CTLA-4 (inhibiting) receptors on the T-cell surface. If a naïve T cell recognizes its peptide antigen in the absence of professional antigen-presenting cell co-receptor signals, the T cell will undergo functional inactivation or clonal deletion. The requirement for dedicated APCs to mediate T-cell activation is a key immune control mechanism that minimizes autoimmunogenic T-cell proliferation and ensures that T-cell activation occurs only as a result of the response to infection mediated by APCs.

Effector T cells comprise two major classes:

1. <u>CD8$^+$ cytotoxic T cells:</u> bind MHC class I proteins expressed on all cells in the body and and survey cytosol (endoplasmic reticulum-derived) proteins.
2. <u>CD4$^+$ helper T cells:</u> bind MHC class II proteins that are expressed only on dedicated antigen-presenting cells (e.g. follicular dendritic cells and B cells) and survey vesicular (extracellular-derived) proteins.

CD8$^+$ effector T cells (or cytotoxic T lymphocytes, CTLs) are cytotoxic cells that clonally expand and survey cells in the body for signs of infection. When CD8 T cells identify another cell expressing its targeted foreign peptide-MHC (e.g. a viral peptide from a virally infected cell) the CTL kills the target cell or induces the target cell to undergo programmed cell death and thereby eliminate the infection. As MHC class I is expressed by all cells in the body and constantly presents peptides derived from inside the cell onto the cell surface, CTLs can survey ongoing protein expression in all cells and tissues and are the main effector cell against intracellular pathogens. Alternatively, CD4$^+$ T cells are activated by follicular dendritic cells to become helper T cells and remain in secondary lymphoid organs. CD4$^+$ T cells perform a critical role in enabling the antibody (or extraceullar) immune response. The five major CD4 effector T subclasses (T_{FH}, T_H1, T_H2, T_H17, and T_{reg}) are summarized in Table 1.2.

CD4 effector subset	APC 3rd signal	Subset function
T_{FH}	IL-6	B cell activation
T_H1	IL-12 + IFN-γ	Intracellular bacterial response
T_H2	IL-4	Parasitic response (eosinophils, mast cells, IgE B cell activation)
T_H17	TGF-β + IL-6	Extracellular bacteria/fungi, induce epithelial/stromal cells to secrete cytokines for neutrophil recruitment
T_{reg}	TGF-β	Suppress T-cell activity, prevent autoimmunity

Table 1.2 Overview of CD4$^+$ T cell effector subsets. Each subset is induced from naïve T cells via a unique third signal provided by antigen presenting cells (Signal 1 comprises the TCR-peptide-MHC interaction, and Signal 2 consists mainly of CD28-B7 interactions)[9].

Naïve B cells are activated by both antigen-presenting cells and CD4$^+$ helper T cells (specifically, T_{FH} cells, see Table 1.2). B cell selection occurs in germinal center (GC) reactions of the spleen and lymph node (Figure 1.3)[9,14]. In germinal centers, B cells that encode BCR will bind to antigen presented on follicular dendritic cells. Following receptor endocytosis of the BCR bound to antigen, the B cells digest proteins and display the resulting peptides on MHC class II on the B cell surface. CD4$^+$ follicular helper T (T_{FH}) cells that have been activated by follicular dendritic cells will in turn induce those antigen-specific B cells to activate via the TCR-peptide-MHC contacts combined with interactions between the B cell co-receptor CD40 and the T-cell CD40 ligand (CD40L, or CD154). CD4 helper T-cells secrete cytokines that stimulate B cell proliferation and differentiation such as IL-4, IL-5, and IL-6.

When naïve B-cells receive the proper signals from antigen-specific T cells, those B cells are activated to undergo somatic hypermutation mediated by the enzyme activation-induced cytidine deaminase (AID) that is linked to successive rounds of cell division in the dark zone of the GC reaction[15] (Figure 1.3). Further rounds of positive selection in the light zone successively enhance antibody affinity to the antigen of interest[16]. Activated B cells can also undergo class-switch recombination, also mediated by AID, which alters the antibody class via genomic deletion of IgM/IgD constant regions to change the antibody isotype to IgG, IgA, or IgE. Positive selection in the GC light zone induces B cells to asymetrically proliferate into a combination of antigen-secreting plasmablasts, antigen-secreting plasma cells, and memory B cells, which comprise the effector cells of the antibody response[17].

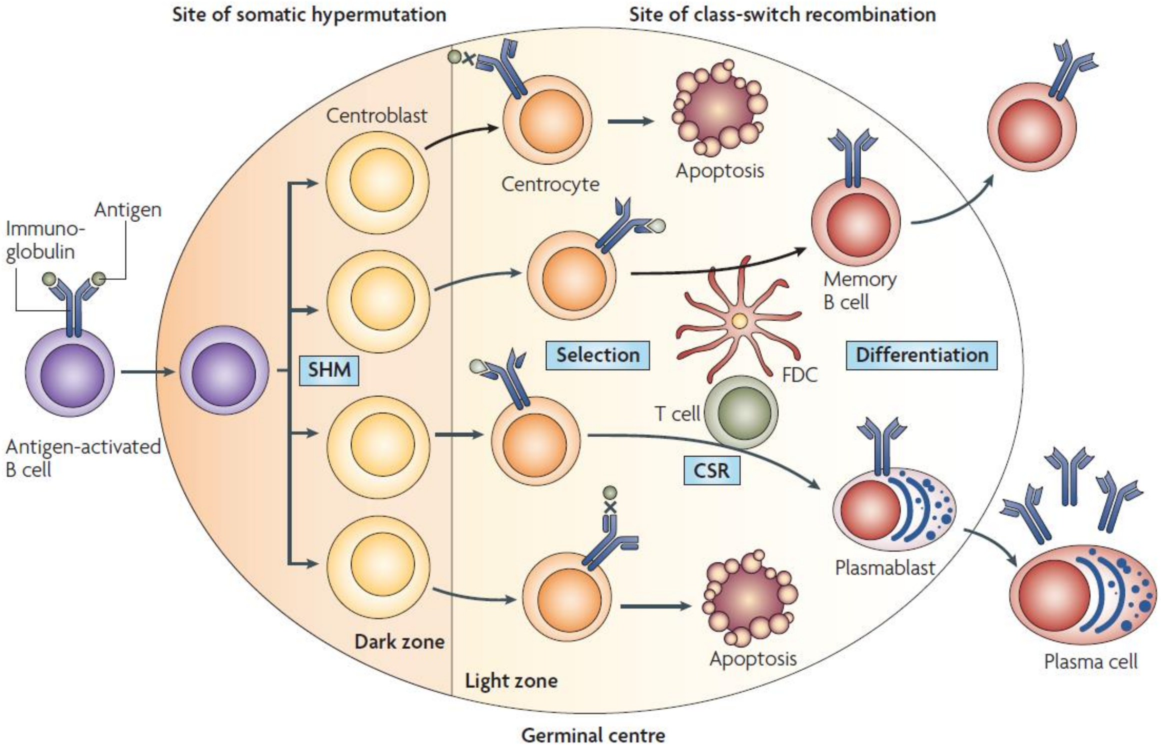


Figure 1.3 Overview of germinal center reactions where B cell selection, affinity maturation, and class-switching occurs. Naïve B cells uptake antigen from follicular dendritic cells and present peptides in MHC class II on their cell surface, and CD4 follicular helper T (T_{FH}) cells specific to those peptides induce B cell activation. B cell activation induces migration to the dark zone for clonal expansion and expression of AID for somatic hypermutation and possibly class-switch recombination, followed by return to the light zone for successive rounds of affinity maturation. B cells with disadvantageous mutations will not acquire more antigen from DCs and will apoptose, whereas B cells with improved affinity will undergo additional selection including return to the dark zone for additional proliferation, SHM, and class-switching. Activated naïve B cells mature into the effector B cells of the serum antibody response (plasmablasts, plasma cells, and memory B cells). Reproduced with permission of the Nature Publishing Group[18].

Importantly, other pathways exist for B cell activation beyond germinal center reactions. B cell maturation has also been reported at the border of the splenic T cell zone and red pulp[19], and for some antigens (termed T-independent antigens) B cell activation can occur in the absence of T-cell help. T-independent antigens often have a component that triggers a receptor of the innate immune system on the B cell surface, or alternatively can extensively cross-link IgM BCR molecules on the B cell surface[9]. While some antigens are able to induce antibody responses in the absence of T-cell help, the vast majority of proteins will not be able to activate the B cell response in the absence of $CD4^+$ T-cell assistance.

The three B effector cell subsets (plasmablasts, plasma cells, and memory B cells) each have a distinct role in fighting disease. Plasmablasts are short-term mediators of serological immunity; they secrete high levels of antibody and circulate in peripheral blood for a relatively limited amount of time (days) before undergoing apoptosis, resulting in a rapid increase in antibody concentration in serum that dissipates according to serum antibody half-life (approximately two weeks)[20,21]. Similar to plasmablasts, plasma cells (PCs) also secrete antibody but are long-lived and home to bone marrow where they can persist for many years[22–25] and continuously secrete high levels of immunoglobulin (estimated at 10,000-20,000 molecules/cell-sec)[2]. Long-term serological memory is mediated by the plasma cell populations residing in the bone marrow. Like plasma cells, memory B cells are also long-lived but do not secrete antibody. Instead, memory B cells express immunoglobulin as a B-cell receptor on their cell surface and circulate in peripheral blood (and pass through secondary lymphoid organs) until encountering its cognate antigen again. Upon re-encounter with their specific antigen, memory B cells rapidly differentiate into plasmablasts and/or plasma cells to enable long-term immune memory, and the kinetics of memory B cell activation (or secondary responses) are much faster than the first (or primary) response to a particular antigen. A comparison of various B-cell subset characteristics with relevance to antibody repertoire sequence analysis is provided in Table 1.3.

Germinal center B cell activation results in a variety of high-affinity antibodies specific to a target antigen. The resulting effector B cell clones can persist for a very long time (e.g. nearly 90 years in the case of memory B cells[26]), and the entire collection of antibodies encoded by a particular individual resulting from a lifetime of immune responses comprise that individual's antibody repertoire. Given the large number of unique B cell clones in humans (likely exceeding 2×10^6 unique antibodies in peripheral blood alone[27,28]), a comprehensive analysis of the antibody repertoire requires high-throughput data collection and analysis. Beginning in 2008, the rapidly falling costs of gene sequencing for the first time permitted economic repertoire-scale, high-resolution DNA sequence analysis of B-cell populations (Figure 1.4). These new experimental techniques, collectively known as high-throughput sequencing or next-generation DNA sequencing, are fundamentally transforming our B- and T-cell analytical methods and our understanding of adaptive immune responses.

Repertoire subset	Selection mechanisms	Isotypes	SHM?
Mature (naïve) B cell	Negative	IgM/IgD	No
Memory B cell	Positive and Negative	IgM,IgG,IgA,IgE	Often
Plasmablast/plasma cell	Positive and Negative	IgM,IgG,IgA,IgE	Often

Table 1.3 Selected characteristics of B-cell subsets relevant to antibody repertoire sequence analysis. B cells are negatively selected in central and peripheral tolerance, and positively selected for antibody affinity to antigen in GC reactions.

1.3 High Throughput Antibody Sequencing

Tremendous advances in the economics of next-generation DNA sequencing technologies over the last six years have dramatically accelerated the pace of biological research (Figure 1.4). High-throughput DNA sequencing has been a transformative method for studying adaptive immunity by permitting repertoire-scale analysis of the vast number of unique BCRs and TCRs in the adaptive immune system[2,29]. Exact measurement of total human B cell repertoire size has remained difficult to determine because of tissue sampling limitations (peripheral blood, bone marrow, and secondary lymphoid organs), combined with high-throughput sequencing error (typically ~0.5% of sequenced bases[30,31]) that contribute noise to secondary data analysis. However, lower-bound estimates of repertoire size are approximately $2x10^6$ unique B cell clones (expressing distinct BCRs) in peripheral blood alone[27,28]. Upper bounds on repertoire size are difficult to estimate but are several orders of magnitude higher, with theoretical B cell receptor diversity exceeding 10^{13} and individual limitations at around 10^{11} B cells in the human body[2,9,32,33].

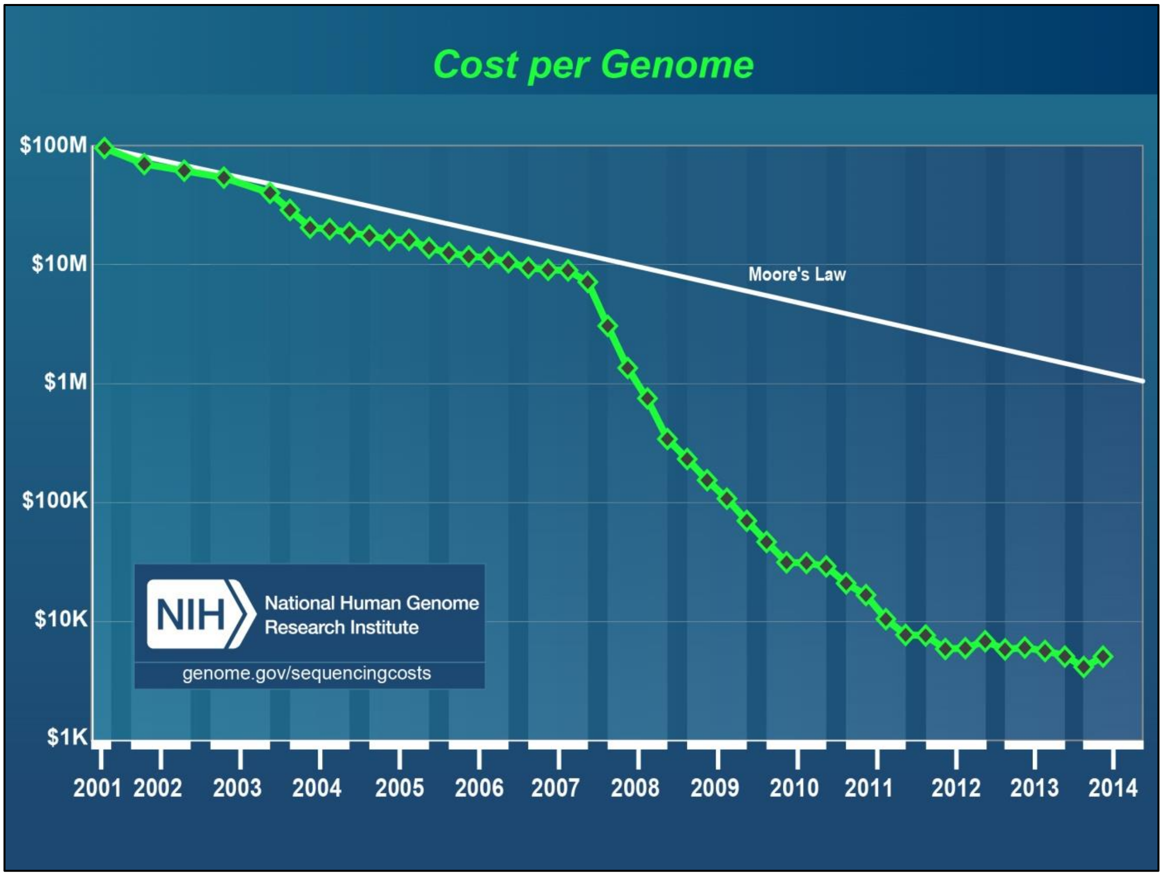


Figure 1.4 High-throughput DNA sequencing costs have dropped dramatically in recent years beginning with the introduction of the first next-generation sequencing platforms in 2008[34]. Log-scale reductions in DNA sequencing costs have directly enabled high-throughput analysis of individual antibody repertoires. New high-throughput DNA sequencing methods are fundamental drivers of recent advances in antibody repertoire sequencing and antibody discovery.

Standard immune repertoire high-throughput sequencing protocols begin with collection of 10^3-10^7 lymphocytes, followed by cell lysis and recovery of cellular mRNA. Next, mRNA is reverse transcribed and a PCR multiplex primer set which targets all known V-genes is used for PCR amplification of antibody or TCR genes. In the case of antibody analysis, the 5' PCR primers are usually designed to target V genes, while the 3' primers target either the J genes or constant regions (e.g. IgG, IgA, etc.) for sequence analysis of the entire antibody variable region (Figure 1.1)[2]. (Some protocols omit V-gene-specific primers and incorporate RACE PCR to reduce PCR amplification bias[35].) For antibody analysis, the heavy chain, kappa light chain, and lambda light chains are each amplified in separate PCR reactions. Finally, high-throughput sequencing and bioinformatic analysis are performed to quantitatively determine the composition of the input immune repertoire encoded by the cells originally isolated from experimental samples[2,31,36,37]. High-throughput repertoire sequencing has been applied in a variety of applications ranging from characterization of the repertoire in healthy and disease states[38–41], to analysis of antibody-pathogen interactions[7,42,43], to rapid antibody discovery[4,42].

Despite the tremendous recent advances, all currently available techniques for antibody repertoire analysis have one severe limitation: high-throughput antibody sequencing techniques performed to date are unable to resolve the pairing between antibody heavy and light chains. Using the high-throughput sequencing methods described above, B cell populations are lysed in bulk to collect mRNA for downstream sequence analysis. Recombined heavy and light V-(D-)J junctions are located on separate chromosomes and expressed as distinct mRNA strands, and the bulk B cell lysis required for high-throughput sequencing confounds the pairing between heavy and light mRNAs expressed by individual B cells. Next-generation sequencing techniques can sequence only one mRNA strand at a time, which further complicates efforts to preserve heavy and light chain pairing information[2,5–7]. Without the ability to analyze paired heavy and light chain sequences at single-cell resolution, the full antibody clonotype (both heavy and light chains) cannot be resolved, nor can the resulting antibody sequences be expressed to test for function nor modeled computationally.

1.4 Monoclonal Antibody Discovery Methods

The utility of serum antibodies for treating disease was first established in the late 19th century through the work of Emil von Behring, Kitasato Shibasaburo, and Emile Roux in developing serum therapies to diphtheria toxins. These early methods were based on polyclonal antibodies, or a mixture of all the different antibody specificities contained in human or animal sera. Research continued with polyclonal antibody mixtures until the 1970's when Georges Köhler and César Milstein published a method to generate hybridomas, or a B cell fused with an immortalized myeloma cell that allowed the resulting cell hybrid (called a hybridoma) to secrete antibody continuously in culture[44]. The discovery of hybridomas ushered in a new era of biotechnology as monoclonal antibodies (mAbs) against a wide variety of antigens could be isolated from mice following challenge with the antigen of interest.

In the hybridoma process, B cells and myeloma cells are fused as described above, then cells are divided into individual wells by limiting dilution and cultured as they produce and secrete antibody. Culture supernatant from each well containing secreted antibody is screened for binding to antigen via enzyme-linked immunosorbent assay (ELISA), and cells from any positive-binding wells can be retrieved, further expanded, cloned, and sequenced, while the monoclonal antibody itself can also be purified from hybridoma culture supernatant. Hybridoma methods continue to have tremendous impact; the basic techniques were outlined almost 40 years ago and are still relevant today, but hybridoma techniques have improved such that an antibody can be developed toward a particular target much more quicklky and with high reliability. In particular, recent key methods include enhancing the efficiency of hybridoma generation with human cells[45] and humanization of mouse antibodies or the development of human transgenic[46–49] or humanized[50–52] mouse models to reduce antibody immunogenicity of the resulting monoclonal antibody therapeutics. Several protocols have also permitted faster and more economical hybridoma screening to accelerate the discovery of human or humanized monoclonal antibodies[53–55]. In particular, transgenic mice have recently been used for isolating human monoclonal antibodies against human proteins (the response to self-antigen is limited in humans), and resulting mAbs can be used to agonize or block human surface receptors or target expressed oncogenes for cancer therapeutics. However despite tremendous advances in tried-and-true hybridoma technologies, mAb discovery via hybridomas remains time-consuming and expensive due to the single-cell limiting dilution needed and the time required hybridomas to adequately expand from a single cell to a cell population (several weeks). The large number of culture supernatant screens required also make hybridoma mAb discovery an extremely resource-intensive experimental technique.

An important alternative technology to hybridoma mAb discovery is antibody isolation via the screening of combinatorial libraries. The most widely used combinatorial library discovery platform technology employs display on M13 bacterial phage, where antibody variable regions are PCR amplified from B cell populations of interest and expressed for display on the surface of bacteriophage. Then, the phage can be selected for binding to the surface of antigen-coated plates or tubes, and bound phage are eluted from tubes and amplified via re-infection of bacteria (with mutations acquired in each round). Finally after several rounds of phage panning a high-affinity monoclonal antibody can be isolated[56–62]. Phage panning has several key advantages including lower cost and the capability to affinity mature antibodies *in vitro*, however phage panning library construction requires combinatorial shuffling of heavy and light chain pairs during library generation, leading to a non-natural (synthetic) antibody library. Another major limitation to phage panning is that the resulting antibodies have not been screened by central or peripheral tolerance checkpoints in the immune system. Thus while it is a highly effective method for isolation of research and diagnostic antibodies, phage panning's inability to isolate native heavy and light chain pairings and the accrual of mutations throughout phage panning pose risks for immunogenicity and off-target binding in humans, and phage panning has limited applications for therapeutic antibody discovery[63].

A more recent method for monoclonal antibody discovery has applied high-throughput sequencing of B cell receptors to identify antibodies of interest[4,7,64]. Antibody discovery via high-throughput sequencing is rapid and efficient, and it also provides information on the entire repertoire of antibodies elicited in the individual. High-throughput sequencing of the cellular repertoire can also be used to construct a database for proteomic analysis of human serum antibodies via mass spectrometry[35,65–68]. These techniques quantify the serum antibodies generated in response to vaccination and disease and have proven useful for antibody discovery by linking antibody function (i.e. binding to a particular antigen) to the antibody sequence[65,66,69]. High-throughput and proteomic techniques for antibody discovery will become more widely used in the coming years as DNA sequencing costs decrease further.

The major limitation to any high-throughput sequencing antibody discovery approach is that, similar to phage panning, the library of heavy and light chains must be shuffled prior to antibody expression because heavy and light chain pairing information is irreversibly lost during high-throughput sequencing[5–7]. High-throughput approaches could be optimal for antibody discovery and immune repertoire analysis if a new technology were available to gather single-cell heavy and light chain pairing information at high-throughput. The state of current (low-throughput) sequence-based alternatives to high-throughput sequencing, collectively known as single-cell RT-PCR, and applications of single-cell sequencing to antibody discovery are discussed in the following section.

1.5 Single-Cell Sequencing Techniques

As mentioned above, existing immune repertoire high throughput sequencing technologies yield data on only one of the two chains of immune receptors and cannot provide information about the identity of immune receptor pairs encoded by individual B or T lymphocytes[5–7]. Because of this major limitation, lower-throughput single-cell techniques must be used when paired heavy and light chain information is required. Several experimental techniques have been employed for detection or sequencing of genomic DNA or cDNA from single cells; however these techniques are limited by low efficiency or low cell throughput (<200-500 cells) and further, they require fabrication and operation of complicated microfluidic devices[70–74]. Due to these limitations, sequence analysis of VH:VL pairs is currently performed by microtiter-well sorting of individual B cells followed by single-cell RT-PCR (scRT-PCR) and Sanger sequencing[5,12,75–77]. Once the sequence of a B cell has been isolated it can be cloned into bacteria and tested for antigen binding to a protein of interest[75,76,5,78], or alternatively each B cell can be induced to secrete antibody *in vitro* prior to screening single-cell culture supernatant by microneutralization[79]. A significant time savings can also be achieved via linkage of heavy and light chains in the RT-PCR, thereby reducing cloning steps by a factor of two[75,80,81].

Single-cell sequencing in the small volumes that are required for high-throughput experiments is severely limited by inhibition of the RT-PCR reaction by cell lysate, which poses a lower bound on microwell or droplet volume at around 5 nL/cell for one-pot cell encapsulation and RT-PCR[72]. Incomplete cell lysis and RNA degradation during thermal cell lysis can further reduce yield of linked cDNA products. Furthermore, cell lysis and mRNA recovery steps are non-trivial to perform at high-throughput and with single-cell fidelity. These experimental complications have made high-throughput sequencing of multiple mRNA transcripts from single cells a critical unsolved problem. Thus, potential solutions for sequencing of multiple mRNA strands derived from single cells would have important applications for in-depth analysis of antibody and TCR repertoires as well as provide a tremendous boost to currently available antibody discovery techniques[2,5–7].

1.6 Synopsis

This dissertation directly addresses the limitations of currently available high-throughput methods in resolving heavy and light chain pairings via the development and application of new high-throughput, single-cell sequencing technologies. In Chapter Three (*High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire*)[82], we report a new method for sequencing B cell heavy and light chains at single-cell resolution by capturing cells in micropatterned PDMS wells and lysing cells individually, with the capacity to analyze up to $5x10^4$ B cells in a single experiment. After validating our technique, we analyzed human antibody responses in healthy and vaccinated donors and applied our high-throughput paired heavy and light chain sequencing platform for human antibody discovery.

Next, Chapter Four (*In-depth determination and analysis of the human paired heavy and light chain antibody repertoire*)[83] relates a modification of the single-cell sequencing methods described in Chapter Three that provided greatly enhanced cell throughput. We constructed and validated a new flow-focusing nozzle system for single B cell isolation and analysis with the capacity to emulsify up to $3x10^6$ B cells per hour. After demonstrating that our technique was >97% accurate by analyzing technical replicates of expanded B cell populations, we characterized several previously unreported features of human antibody repertoires including a quantitative analysis of promiscuous and public light chain junctions (i.e. light chains expressed by multiple B cell clones both within and across human donors) and the high-throughput detection and sequence characterization of allelically included human B cells.

The fifth chapter of this report (*Paired VH:VL analysis of naïve B cell repertoires and comparison to antigen-experienced B cell repertoires in healthy human donors*) applied the new high-throughput techniques developed in Chapters Three and Four toward a comprehensive, high-resolution analysis of naïve and antigen-experienced B cells in the same individuals. Comprising the first high-throughput analysis of heavy and light chains to compare multiple B cell compartments, our analysis of gene usage and antibody biochemical composition generated new insights regarding antibody development, maturation, and selection processes across several human donors.

2. HIGH-THROUGHPUT SEQUENCING OF THE PAIRED HUMAN IMMUNOGLOBULIN HEAVY AND LIGHT CHAIN REPERTOIRE[1]

2.1 Rationale and Supporting Information

Currently existing immune repertoire sequencing technologies yield data on only one of the two chains of immune receptors[5–7]. Sequence analysis of VH:VL pairs is currently performed by microtiter-well sorting of individual B cells followed by single-cell RT-PCR (scRT-PCR) and Sanger sequencing[5,12,75–77,84–86]; however at most a few hundred VH:VL pairs (a number dwarfed by the enormous size of the human antibody repertoire) are identified via scRT-PCR[76,77,84,85]. Microfluidic methods for RT-PCR and the sequencing of two or more genes (for example using the Fluidigm platform[87]), have so far been limited to only 96 wells per run and require complex, proprietary instrumentation. As a result, comprehensive analysis of paired VH:VL gene family usage and somatic hypermutation frequency has been elusive.

[1] DeKosky, B. J. *et al.* High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.* **31,** 166–169 (2013). B.J.D. and G.G. developed the methodology and designed the experiments. B.J.D., G.C.I. and G.G. wrote the manuscript; B.J.D., G.C.I., R.P.D., J.J.L., Y.W., B.M.R., C.G. and S.F.A. performed the experiments; B.J.D. carried out the bioinformatic analysis; S.P.H.-S. performed Illumina sequencing; G.C.I., N.V., T.D., P.C.W., C.G.W. and A.D.E. helped design experiments; B.J.D., G.C.I., J.J.L., Y.W., S.P.H.-S., A.D.E. and G.G. analyzed the data.

Several prior studies have analyzed single cells by first isolating the cells into high-density microwell arrays[54,88–91]. Such methods have been used for phenotypic and genomic sequence analyses of cells at high throughput, however mRNA sequence analysis is complicated by inhibition of reverse transcription by cell lysate at concentrations germane to cell isolation in microwell arrays[72]. We reasoned that a system capable of selectively capture mRNA from single cells could circumvent cell lysis inhibition of the RT-PCR reaction by first permitting cell lysis under the harsh conditions which preserve mRNA (e.g. the presence of dodecyl sulfate and DTT to inactivate endogenous RNAse), and single-cell mRNA could then be purified for use as RT-PCR template. Magnetic beads could be used as the mRNA capture agent by utilizing poly(dT) oligonucleotides conjugated to the magnetic beads to bind to the polyadenylated mRNA tail. Magnetic beads are also compatible with a variety of downstream processing steps like emulsion RT-PCR, which obviates the need to intentionally release mRNA from the beads and therefore the microbeads would never need to be removed from the sample. We also reasoned that by linking VH and VL transcripts onto a single strand (similar to previously published methods[75]), we could sequence paired VH and VL chains using standard next-generation sequencing protocols.

Methodology:

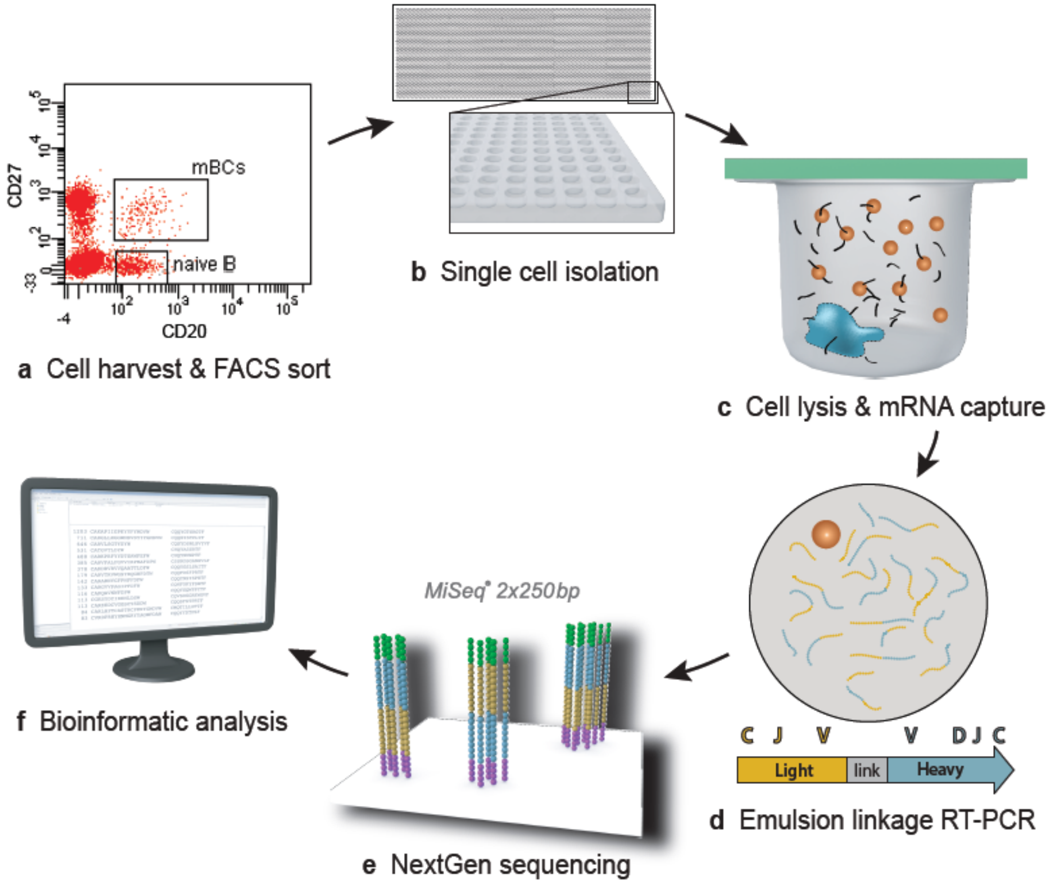


Figure 2.1 Overview of the high throughput methodology for paired VH:VL antibody repertoire analysis. (a) B-cell populations are sorted for desired phenotype (mBCs=memory B cells, naive B=naive B cells). (b) Single cells are isolated by random settling into 125 pL wells (56 μm diameter) printed in polydimethylsiloxane (PDMS) slides the size of a standard microscope slide ($1.7x10^5$ wells/slide). 2.8 μm poly(dT) microbeads are also added to the wells (average 55 beads/well). (c) Wells are sealed with a dialysis membrane and equilibrated with lysis buffer to lyse cells and anneal VH and VL mRNAs to poly(dT) beads (blue figure represents a lysed cell, orange circles depict magnetic beads, black lines depict mRNA strands. (d) Beads are recovered and emulsified for cDNA synthesis and linkage PCR to generate an ~850 base pair VH:VL cDNA product (Figure A1.1). (e) Next Generation sequencing is performed to sequence the linked strands. (f) Bioinformatic processing is used to analyze the paired VH:VL repertoire.

As shown in Figure 2.1, a population of sorted B cells is deposited by gravity into 125 pL wells molded in polydimethylsiloxane (PDMS) slides. Each slide contains 1.7×10^5 wells; four slides processed concurrently accommodate 68,000 lymphocytes at a ≥1:10 cell:well occupancy which gives at least a 95% probability of only one cell per well based on Poisson statistics. Poly(dT) magnetic beads with a diameter of 2.8 μm are deposited into the microwells at an average of 55 beads/well and the slides are covered with a dialysis membrane. Subsequently the membrane-covered slides are incubated with an optimized cell lysis solution containing 1% lithium dodecyl sulfate that results in complete cell lysis within <1 minute. The mRNA anneals to the poly(dT) magnetic beads which are then collected, washed, and emulsified with primers, reverse transcriptase, and thermostable DNA polymerase to carry out reverse transcription followed by linkage PCR (Figure A1.1). The two-step capture and amplification process (Fig. 3.1) is necessary because single-compartment cell lysis followed by RT-PCR has not proven feasible in volumes ≤5 nL due to inhibition of the reverse transcription reaction by cell lysate constituents, and because performing VH:VL linkage in emulsion droplets at the single cell level would necessitate cell entrapment, lysis, reverse transcription, and *in situ* linkage PCR that can only be performed in microfluidic devices, a strategy which requires extensive infrastructure and so far has been reported to have limited throughput at ≤300 cells per run[72]. PCR amplification as outlined in Figure A1.1 generates an ~850 base pair (bp) linked VH:VL DNA product composed of (from 5’ to 3’) the N-terminal end of CH1, the VH, a linker region, the VL and the N-terminal of Cκ or Cλ. The most informative 500 bp of this fragment which encompasses the complementarity determining regions (CDR-H3 and CDR-L3) is then sequenced on a long-read next generation sequencing platform such as the 2x250 Illumina^{TM} MiSeq (which also provides the framework region FR3 and FR4 sequences and constant region N-termini amino acid sequences that can be used for isotype assignment). If FR1 to CDR2 region sequences are also desired, the VH and VL gene repertoires are analyzed by separate 2x250 bp sequencing runs. This latter step is required because of read length limitations with existing technology; while single molecule sequencing techniques allow for longer reads the error rate is currently too high to enable robust classification of VH:VL sequences.

2.2 Results:

We employed the methodology of Figure 2.1 to determine the VH:VL repertoire of three different B-cell populations of relevance to human immunology and antibody discovery. First, we isolated IgG^{+} B cells from fresh blood donated by a healthy individual. 61,000 IgG^{+} B cells were spiked with immortalized IM-9 lymphoblast cells (to approximately 4% of total mixture) that express known VH and VL sequences as an internal control. We analyzed these cells in four PDMS slides (6.8×10^{5} total wells). After 2x250 MiSeq sequencing, we clustered the CDR-H3 regions based on 96% sequence identity, consistent with the established error rate of the MiSeq platform, to determine the number of unique clones recovered from this human sample. A total of 2,716 unique pairs were thus identified (Table A1.1). The spiked IM-9 heavy chain overwhelmingly (78-fold above background) paired with its known light chain. A heat map shows frequencies of pairing between VH and VL segments of different germline families in the class-switched IgG$^+$ cell repertoire (Figure 2.2a). A second IgG$^+$ repertoire analysis was performed using B cells from another anonymous individual; this analysis identified 2,248 unique CDR-H3 from 47,000 IgG$^+$ cells, and the IM-9 control spike again demonstrated high pairing accuracy (125-fold above background). Several V gene families (e.g. IGHV7, IGKV5, 6, and 7, IGLV4, 10, and 11) are expressed at very low frequencies in the human immune repertoire[52,92]. We detected VH:VL pairs containing these rare families, indicating that this technique can identify rare B-cell clones present at physiological levels together with much more abundant clones (e.g. the much more highly utilized IGVH3 or IGVH4 families) (Figure 2.2a). Interestingly, VH:VL germline pairing frequencies were highly correlated between the two individuals (Spearman rank correlation coefficient=0.804, $p<10^{-29}$); the most highly transcribed heavy chain genes (VH3, VH4 and VH1 families) paired most frequently with the most highly transcribed light chain genes (Vκ1, Vκ3, Vλ1 and Vλ2). However, putative differences in IgG$^+$ VH:VL germline pairing frequencies between the two individuals were also evident.

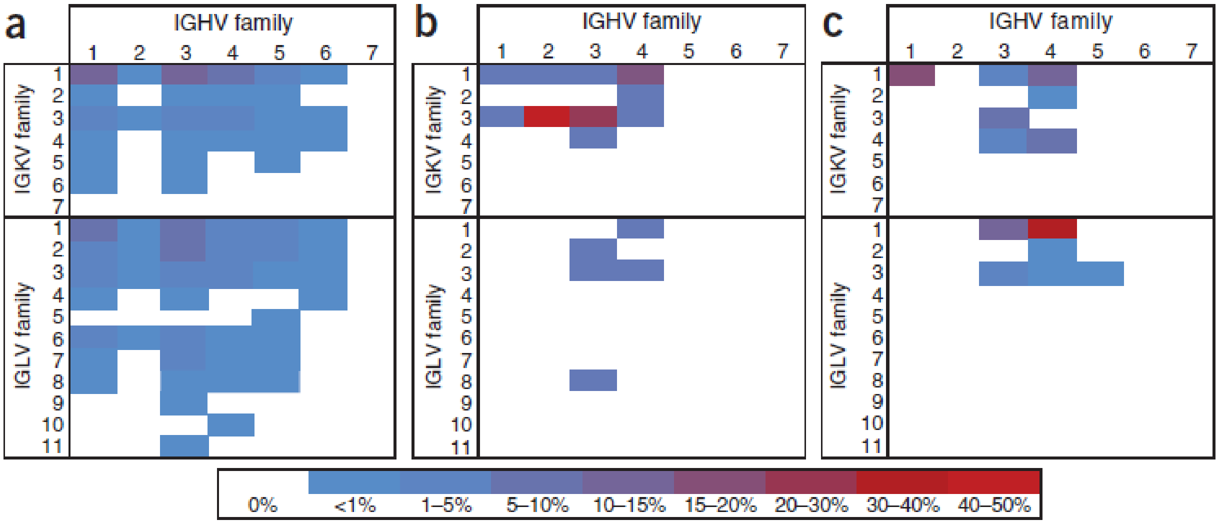


Figure 2.2 VH:VL gene family usage of unique CDR-H3:CDR-L3 pairs identified via high-throughput sequencing of cell populations from three different individuals in separate experiments using the workflow presented in Figure 2.1: (a) healthy donor peripheral IgG$^+$ B cells (n=2,716 unique CDR3 pairs), (b) peripheral tetanus toxoid (TT) specific plasmablasts, isolated seven days post-TT immunization (CD19$^+$CD3$^-$CD14$^-$CD38^{++}CD27^{++}CD20$^-$TT$^+$, n=86 unique pairs), and (c) peripheral memory B cells isolated 14 days post-influenza vaccination (CD19$^+$CD3$^-$CD27$^+$CD38^{int}, n=240 unique pairs). Each panel presents data from an independent experiment obtained from (a) 61,000 fresh B cells, (b) ~400 frozen/thawed plasmablasts, (c) 8,000 twice frozen/thawed memory B cells.

In a separate experiment, human plasmablasts (CD19$^+$CD3$^-$CD14$^-$CD38^{++}CD27^{++}CD20$^-$) from a healthy volunteer were collected 7 days after tetanus toxoid (TT) immunization, sorted for surface antigen binding and then frozen[77]. After thawing, approximately 400 recovered cells were spiked with the immortalized ARH-77 cell line as an internal control and seeded onto a single PDMS slide (1.7x10^5 total wells). In this instance, 86 unique primary CDR-H3:CDR-L3 pairs were identified, and the

ARH-77 control spike demonstrated high pairing accuracy (Fig. 3.2b, Table A1.1). We expressed ten of the identified VH:VL pairs as IgG proteins in HEK293K cells. As revealed by competitive ELISA, all ten antibodies showed specificity for TT and bound TT with high affinity (K_D ranged from 0.1 to 18 nM; Table 2.1). While certain VH chains can pair promiscuously with multiple VLs to yield functional antibodies, it is statistically implausible that 10/10 antibodies could display nM and sub-nM affinities for TT merely as a consequence of fortuitous VH:VL pairing. For comparison, 10-15% of antibodies generated by random pairing of VH genes with a small set of enriched VL genes were antigen-specific[67,68].

Antibody ID	Gene Family Assignment	Affinity (K_D)
TT1	HV3-HD1-HJ6 : KV3-KJ5	1.6 ± 0.1 nM
TT2	HV3-HD3-HJ4 : LV3-LJ1	14 ± 3 nM
TT3	HV1-HD2-HJ4 : KV3-KJ5	3.6 ± 1.8 nM
TT4	HV2-HD2-HJ4 : KV1-KJ1	2.7 ± 0.3 nM
TT5	HV4-HD2-HJ6 : KV2-KJ3	18 ± 4 nM
TT6	HV1-HD3-HJ4 : KV1-KJ2	0.57 ± 0.03 nM
TT7	HV4-HD3-HJ4 : KV1-KJ2	0.46 ± 0.01 nM
TT8	HV3-HD3-HJ4 : LV8-LJ3	2.8 ± 0.3 nM
TT9	HV4-HD2-HJ4 : KV1-KJ1	0.10 ± 0.01 nM
TT10	HV1-HD3-HJ5 : KV3-KJ5	1.6 ± 0.1 nM

Table 2.1 TT-binding affinities of IgG antibodies sequenced from TT^+ peripheral plasmablasts. Peripheral blood mononuclear cells were isolated from one healthy volunteer 7 d after TT boost immunization and TT-binding $CD19^+CD3^-CD14^-CD38^{++}CD27^{++}CD20^-$ cells were sorted and analyzed as in Figure 2.1. Genes encoding ten of the sequenced VH:VL pairs were cloned into an IgG expression vector and expressed transiently in HEK293F cells. TT-binding affinities of the resulting IgG were calculated from competitive ELISA dilution curves. Each heavy and light chain was distinct.

Finally, we compared the VH:VL pairings identified using this high-throughput approach to those identified using the established single cell sorting method[76,93]; this experiment was conducted in a double-blinded manner. Peripheral $CD19^{+}CD3^{-}CD27^{+}CD38^{int}$ memory B cells were isolated from a healthy volunteer 14 days after vaccination with the 2010-2011 trivalent FluVirin influenza vaccine[76]. For the scRT-PCR analysis, 164 single B cells were sorted into four 96-well plates, and 168 RT and 504 nested PCR reactions were performed individually to separately amplify the VH and VL (kappa and lambda) genes. DNA products were resolved by gel electrophoresis and sequenced to yield a total of 51 VH:VL pairs, of which 50 were unique. A separate B-cell aliquot from the same individual was frozen at -80°C and later thawed and processed using the new high-throughput approach described here. Two PDMS slides ($3.4x10^{5}$ total wells) were used, and the sample was spiked with IM-9 cells to confirm pairing accuracy (Table A1.1). A total of 240 unique CDR-H3:CDR-L3 pairs were recovered (Figure 2.2c). Four CDR-H3 sequences detected in the high-throughput pairing set were also observed in the single-cell RT-PCR analysis. A blinded analysis revealed that CDR-H3:CDR-L3 pairs isolated by the two approaches were in complete agreement (Table A1.3). Further, the one VH:VL pair detected in more than one of the 51 cells analyzed by single-cell RT-PCR was also detected in the aliquot processed by the new high-throughput approach (clone 2D02 was observed in two cells by scRT-PCR, Table A1.3); these findings suggest that this B-cell clone may have undergone a great deal of expansion. The 46 VH genes that were each observed only once by single-cell RT-PCR but that were not detected in the aliquot processed by our high-throughput approach presumably represent unique or very low abundance B-cell clones, as expected given the great degree of V gene diversity normally found in human peripheral memory B cells.

2.3 Discussion

In these experiments the control cell lines spiked into each aliquot of primary B cells were selected to approximate the levels of heavy chain and light chain transcription in that particular primary B-cell subpopulation. For example, the ARH-77 cell line expressed high levels of heavy chain and light chain transcripts and therefore was spiked into plasmablast populations that also express abundant heavy chain and light chain transcripts; in contrast, the IM-9 B lymphoblast cell line, which expresses lower levels of heavy chain and light chain transcripts, was spiked into memory B-cell populations. Known V_H and V_L sequences from spiked-in control cell lines were used to evaluate the frequency of non-native pairings, that is, the false discovery rate (FDR). The FDR, determined from the mispairing of spiked-in control V_H and V_L chains, was commensurate with the probability of coincident cells per well, which in turn is dictated by cell seeding density and follows Poisson statistics (Methods and Table A1.4). The FDR revealed by the mispairing frequency of control cell lines represents the upper bound of the FDR, as the control cell lines were introduced at levels over tenfold higher than the levels at which even a very highly expanded B-cell clone might be present in a biological sample in humans. Although currently V_H:V_L pairing efficiency in memory B-cell populations is relatively modest (Figure 2.2 and Table A1.2), efforts to further improve efficiency are under way.

The high-throughput V_H:V_L pairing technique described here requires one emulsion RT-PCR reaction, followed by nested PCR, sequencing and bioinformatic analysis. The entire process from B-cell isolation to the generation of V_H:V_L heat maps can be completed by a single investigator in 10 research hours over the course of four days (which includes three days for gene sequencing). For example, the work required to recover 2,716 unique V_H:V_L pairs from a sample of IgG$^+$ peripheral B cells (Figure 2.2a) was completed by a single researcher in 10 h and cost $550. Analysis of 2,700 cells using established optimal single-cell RT-PCR protocols would have required >10 weeks of effort by an experienced technician and >$25,000 in reagent and sequencing costs[11].

Because only sequences of up to 500 bp can be accurately determined with current Illumina next-generation sequencing technology, our method detects the different antibody clonotypes (antibodies comprising the same CDR-H3:CDR-L3) but cannot yet distinguish somatic variants originating from clonally related B cells that contain upstream mutations between FR1 and CDR2 regions. However, this method distinguished nearly identical but distinct CDR3 regions, as indicated by B-cell clones 2D02 and 3D05, which express light chain CDR3s that differ by only two nucleotides (Table A3.3). Rapid advances in next-generation sequencing read-length and quality will likely enable upstream somatic variant analysis in the near future.

We used PCR amplification primers targeted to the FR1 region of heavy and light chains (primers reported in Tables A1.5–A1.7). In some chronic infections (e.g., HIV), constant antigen exposure can generate antibodies that are highly mutated in all regions including the FR1 region, and therefore amplification with FR1-specific primers can bias the repertoire. In these cases, such bias can be readily circumvented by using primers that anneal to the leader peptide[94].

Finally, we note that the analysis reported here focused on the light chain that was the dominant light chain paired with a particular V_H. There are known, albeit rare, instances in mice where one heavy chain can be found paired with more than one light chain[23]. Bioinformatic analysis might discriminate between biologically relevant V_H:V_L pairs of one heavy chain with multiple light chains and false pairings that might result from multiple cells seeded into the same well of the experimental device; false pairings can be flagged because coincident cells 1 and 2 would yield the products V_H1:V_L2 and V_H2:V_L1 in addition to V_H1:V_L1 and V_H2:V_L2.

2.4 Methods

2.4.1 Preparation of microwell slides

A grid of micropillars (56 µm diameter, 50 µm height) was photolithographically patterned onto a silica wafer using SU-8 photoresist (Fisher Scientific, Pittsburgh, PA) and the silica wafer was used as a mold to print polydimethylsiloxane (PDMS) slides (Sylgard 184, Dow Corning, Midland, MI) with the dimensions of a standard microscope slide and containing approximately 170,000 wells each.

Molded PDMS slides were treated in an oxygen plasma chamber for 5 minutes to generate a hydrophilic surface. The PDMS slides were blocked in 1% bovine serum albumin (BSA) for 30 minutes and washed with deionized water followed by phosphate-buffered saline (PBS) to prepare for cell seeding.

2.4.2 Analysis of memory B-cell VH:VL pairings in response to seasonal influenza vaccination

The study was approved by the University of Chicago Institutional Review Board (IRB# 09-043-A) and the University of Texas Institutional Review Board (IRB# 2012-07-0002). A healthy 30-year-old male was vaccinated with the 2010-2011 trivalent FluVirin influenza vaccine (Novartis) and blood was drawn at Day 14 after vaccination after informed consent had been obtained. PBMCs were isolated and resuspended in DMSO/10%FCS for cryopreservation.

Frozen PBMCs were subsequently thawed and cell suspensions were stained in PBS/0.2% BSA with anti-human CD19 (HIB19, BioLegend, San Diego, CA), CD27 (O323, BioLegend), CD38 (HIT2, BioLegend), and CD3 (7D6, Invitrogen, Grand Island, NY). CD19^{+} CD3^{-} CD27^{+} CD38^{int} memory B cells were sorted using a FACSAria II sorter system (BD Biosciences, San Diego, CA). Cells were either cryopreserved in DMSO/10%FCS for subsequent high-throughput VH:VL pairing or single-cell sorted into 96-well plates containing RNAse Inhibitor Cocktail (Promega, Madison, WI) and 10mM Tris-HCl pH 8.0 for single cell PCR analysis. cDNA was synthesized from single-sorted cells using the Maxima First Strand cDNA Synthesis Kit (Fermentas, Waltham, MA) followed by amplification of the immunoglobulin variable genes using primer sets and PCR conditions previously described[76]. Variable genes were determined with in-house analysis software using the IMGT search engine[95].

Memory B cells frozen for high-throughput VH:VL pairing were thawed and recovered by centrifugation at 250x*g* for 10 minutes. Cells were resuspended in 200 uL RPMI-1640 supplemented with 1x GlutaMAX, 1x non-essential amino acids, 1x sodium pyruvate and 1x penicillin/streptomycin (Life Technologies) and incubated at 37°C for 13 hours in a 96-well plate. Recovered cells were centrifuged again at 250x*g* for 10 minutes and resuspended in 400 uL PBS, and 6 uL were withdrawn for cell counting with a hemocytometer. Approximately 8,800 cells were recovered from frozen stock. Memory B cells were then spiked with approximately 880 IM-9 cells (ATCC number CCL-159) as an internal control. Cells were resuspended over two PDMS microwell slides (340,000 wells) and allowed to settle into wells by gravity over the course of 5 minutes with gentle agitation. The cell seeding process has been calculated to be 90% efficient by measuring cell concentration in seeding buffers both pre- and post- cell seeding; thus 8,000 primary cells were analyzed in this experiment. The fraction of cells isolated in the single and multiple cell per well states was calculated using Poisson statistics:

$$P(k,\mu) = \frac{\mu^k e^{-\mu}}{k!}$$

where k equals the number of cells in a single microwell and μ is the average number of cells per well, so that the 1:39 cell:well ratio used in this experiment corresponds to 98.7% of cells deposited at an occupancy of one cell/well. 25 uL of poly(dT) magnetic beads (Invitrogen mRNA Direct Kit) were resuspended in 50 uL PBS and distributed over each PDMS slide surface, (mean of 55 poly(dT) beads per well). Magnetic beads were allowed to settle into wells by gravity for approximately 5 minutes, then a BSA-blocked dialysis membrane (12,000-14,000 MWCO regenerated cellulose, 25 mm flat width, Fisher Scientific) that had been rinsed in PBS was laid over each slide surface, sealing the microwells and trapped cells and beads inside. Excess PBS was removed from the slide and membrane surfaces using a 200 µL pipette. 500 µL of cell lysis solution (500 mM LiCl in 100 mM tris buffer (pH 7.5) with 1% lithium dodecyl sulfate, 10 mM EDTA, and 5 mM DTT) was applied to the dialysis membranes for 20 min at room temperature. Time-lapse microscopy revealed that all cells are fully lysed within 1 minute. Subsequently the slides were incubated at 4°C for 10 min at which point a Dynal MPC-S magnet was placed underneath the PDMS microwell device to hold magnetic beads inside the microwells as the dialysis membrane was removed with forceps and discarded. Working quickly, the PDMS slides were sequentially inverted in a Petri dish containing 2 mL of cold lysis solution and the magnet was applied underneath the Petri dish to force the beads out of the microwells. Subsequently 1 ml aliquots of the lysis solution containing resuspended beads were placed into Eppendorf tubes and beads were pelleted on a Dynal MPC-S magnetic rack and washed once without resuspension using 1 mL per tube of wash Buffer 1 (100 mM Tris, pH 7.5, 500 mM LiCl, 1 mM EDTA, 4°C). Beads were resuspended in wash Buffer 1, pelleted and resuspended in Wash Buffer 2 (20 mM Tris, pH 7.5, 50 mM KCl, 3 mM MgCl) and pelleted again. Finally beads were suspended in 2.85 mL cold RT-PCR mixture (Quanta OneStep Fast, VWR) containing 0.05 wt% BSA (Invitrogen Ultrapure BSA, 50 mg/mL) and primer sets for VH and VL linkage amplification (Figure A1.1, Table A1.5)[52,96]. The suspension containing the poly(dT) magnetic beads was added dropwise to a stirring IKA dispersing tube (DT-20, VWR) containing 9 mL chilled oil phase (molecular biology grade mineral oil with 4.5% Span-80, 0.4% Tween 80, 0.05% Triton X-100, v/v%, Sigma Aldrich, St. Louis, MO), and the mixture was agitated for 5 minutes at low speed. The resulting emulsion was added to 96-well PCR plates with 100 μL emulsion per well and placed in a thermocycler. The RT step was performed under the following conditions: 30 minutes at 55°C, followed by 2 min at 94°C. PCR amplification was performed under the following conditions: four cycles of 94°C for 30 s denature, 50°C for 30 s anneal, 72°C for 2 min extend; four cycles of 94°C for 30 s denature, 55°C for 30 s anneal, 72°C for 2 min extend; 22 cycles of 94°C for 30 s denature, 60°C for 30 s anneal, 72°C for 2 min extend; then a final extension step for 7 min at 72°C. After thermal cycling the emulsion was visually inspected to ensure the absence of a bulk water phase, which is a key indicator of emulsion stability. Following visual verification, the emulsion was collected and centrifuged at room temperature for 10 minutes at 16,000x*g*, the mineral oil upper phase was discarded, and 1.5 mL diethyl ether was added to extract the remaining oil phase and break the emulsion. The upper ether layer was discarded, two more ether extractions were performed and residual ether was removed in a SpeedVac for 25 minutes at room temperature. The aqueous phase was diluted 5:1 in DNA binding buffer and passed through a silica spin column (DNA Clean & Concentrator, Zymo Research, Irvine, CA) to capture the cDNA product. The column was washed twice with 300 µL wash buffer (Zymo Research Corp) and cDNA was eluted into 40 µL nuclease-free water. Finally a nested PCR amplification was performed (ThermoPol PCR buffer with Taq Polymerase, New England Biosciences, Ipswich, MA) in a total volume of 200 µL using 4 µL of eluted cDNA as template with 400 nM primers (Table A1.6) under the following conditions: 2 min initial denaturation at 94°C, denaturation at 94°C for 30 s for 39 cycles, annealing at 62°C for 30 s and extension at 72°C for 20 s, final extension at 72°C for 7 min. The approximately 850 bp linked product was extracted by agarose gel electrophoresis and sequenced using the 2x250 paired end MiSeq NextGen platform (Illumina, San Diego, CA).

2.4.3 Analysis of plasmablast VH:VL pairings in response to tetanus toxoid vaccination

One female donor underwent booster immunization against tetanus toxoid (TT)/diphteria toxoid (TD, 20 I.E. TT and 2 I.E. diphteria toxoid, Sanofi Pasteur Merck Sharpe & Dohme GmbH, Leimen, Germany) after informed consent by the Charité Universiätsmedizin Berlin had been obtained (samples were anonymously coded and study approved by the hospital's ethical approval board, number EA1/178/11, and the University of Texas at Austin Institutional Review Board, IRB# 2011-11-0095). At 7 days post tetanus toxoid immunization, EDTA blood was withdrawn and PBMC isolated by density gradient separation as described[97]. PBMCs were stained in PBS/BSA at 4°C for 15 min with anti-human CD3/CD14-PacB (clones UCHT1 and M5E2, respectively, Becton-Dickinson, BD), CD19-PECy7 (clone SJ25C1, BD), CD27-Cy5 (clone 2E4, kind gift from René van Lier, Academic Medical Centre, University of Amsterdam, The Netherlands, labelled at the Deutsches Rheumaforschungszentrum (DRFZ), Berlin), CD20-PacO (clone HI47, Invitrogen), IgD-PerCpCy5.5 (clone L27, BD), CD38-PE (clone HIT2, BD), and TT-Digoxigenin (labeled at the DRFZ) for 15 minutes at 4°C. Cells were washed and a second staining was performed with anti-Digoxigenin-FITC (Roche, labeled at the DRFZ) and DAPI was added prior to sorting. $CD19^{+}CD3^{-}CD14^{-}CD38^{++}CD27^{++}CD20^{-}TT^{+}$ plasmablasts were sorted using a FACSAria II sorter system (BD Biosciences). A portion of sorted cells were washed and cryopreserved in DMSO/10%FCS for high-throughput VH:VL pairing.

One vial containing approximately 2,000 frozen TT$^+$ plasmablasts was thawed and recovered by centrifugation at 250x*g* for 10 minutes, for which 20-30% recovery was anticipated[21]. Cells were resuspended in 300 μL RPMI-1640 supplemented with 10% FBS, 1x GlutaMAX, 1x non-essential amino acids, 1x sodium pyruvate and 1x penicillin/streptomycin (all from Life Technologies) and incubated at 37°C for 13 hours in a 96-well plate. Recovered cells were centrifuged again at 250x*g* for 10 minutes and resuspended in 400 μL PBS, and 6 μL were withdrawn for cell counting with a hemocytometer. Cells were spiked with approximately 30 ARH-77 cells as an internal control (ATCC number CRL-1621) and VH:VL transcripts were linked as described above, omitting IgM primers and using a 38-cycle nested PCR; the resulting product was submitted for 2x250 MiSeq sequencing. VH and VL chains were also amplified individually to obtain full VH and VL sequences for antibody expression. Nested PCR product was diluted 1:9 and 0.5 μL were used as template in a PCR reaction with the following conditions: 400 nM primers (Table A1.7), 2 min initial denaturation at 94°C, denaturation at 94°C for 30 s for 12 cycles, annealing at 62°C for 30 s and extension at 72°C for 15 s, final extension at 72°C for 7 min. The resulting ~450 bp VH or ~400 bp VL products were purified by agarose gel electrophoresis and submitted for 2x250 MiSeq sequencing.

Ten VH and VL pairs were selected from TT$^+$ plasmablast pairings and cloned into the human IgG expression vectors pMAZ-VH and pMAZ-VL respectively[98]. 40 μg each of circularized ligation product were co-transfected into HEK 293F cells (Invitrogen, NY, USA). Medium was harvested 6 days after transfection by centrifugation and IgG was purified by a protein-A agarose (Pierce, IL, USA) chromatography column.

Antigen affinities were determined by competitive ELISA[99] using different concentrations of IgG in a serial dilution of antigen, ranging from 100 nM to 0.05 nM in the presence of 1% milk in PBS. Plates were coated overnight at 4°C with 10 μg/mL of TT in 50 mM carbonate buffer, pH 9.6, washed three times in PBST (PBS with 0.1% Tween 20), and blocked with 2% milk in PBS for two hours at room temperature. Pre-equilibrated samples of IgG with TT antigen were added to the blocked ELISA plate, incubated for one hour at room temperature, and plates were washed 3x with PBST and incubated with 50 μl of anti-human kappa light chain-HRP secondary antibody (1:5,000, 2% milk in PBS) for ~2 min, 25°C. Plates were washed 3x with PBST, then 50 μl Ultra TMB substrate (Thermo Scientific, Rockford, IL) was added to each well and incubated at 25°C for 5 min. Reactions were stopped using equal volume of 1M H_2SO_4 and absorbance was read at 450 nm (BioTek, Winooski, VT). Each competitive ELISA replicate was fit using a four-parameter logistic (4PL) equation, with error represented as the standard deviation of 2-3 replicates for each IgG analyzed.

2.4.4 Analysis of IgG$^+$ class switched cells in healthy donor peripheral blood

PBMC from two different anonymous healthy donors (Texas Gulf Coast Regional Blood Center) were isolated from whole blood using Histopaque-1077 centrifugation gradient (Sigma-Aldrich) according to manufacturer protocols. 3.8×10^8 PBMCs were used as input for cell purification using a IgG$^+$ Memory B Cell Isolation Kit (Miltenyi Biotec, Auburn, CA) Approximately 68,000 cells (Fig. 2a) or 52,000 (Fig. A1.2) cells were spiked with 4% IM-9 immortalized B lymphoblast cells (ATCC number CCL-159) and seeded onto four or three PMS microwell slides, respectively, at a 1:10 cell:well ratio corresponding to 95.1% of cells isolated as single cells by Poisson distribution. Cell seeding was calculated to be 90% efficient and thus 61,000 cells (Figure 2.1a, Table A1.1) or 47,000 cells (Fig. A1.2, Table A1.2) were analyzed. VH:VL chains were paired and amplified as above using 35 nested PCR cycles and sequenced on the Illumina MiSeq platform..

2.4.5 Bioinformatic methods

Raw 2x250 MiSeq data were filtered for minimum Phred quality score of 20 over 50% of nucleotides to ensure high read quality in the CDR3-containing region (approximately HC nt 65-115 or LC bases 55-100). Sequence data were submitted to the International ImMunoGeneTics Information System (IMGT) for mapping to germline V(D)J genes[95]. Sequence data were filtered for in-frame V(D)J junctions and productive VH and Vκ/λ sequences were paired by Illumina read ID. CDR-H3 nucleotide sequences were extracted and clustered to 96% nt identity with terminal gaps ignored, to generate a list of unique CDR-H3s in the data set. 96% nt identity cutoff was found to be the optimal cutoff to cluster sequencing error in spiked control clones; the number of unique CDR-H3 sequences and hence the number of unique V genes reported refer to the number of clusters recovered from the sample (Table A1.1, Table A1.2).

Overwhelmingly the frequency of the top VL pair for a given VH was observed to be greater than 90%, and the top VL pair for each CDR-H3 was used for comparison with single cell RT-PCR for workflow validation (Table A1.3), consensus sequence generation and paired VH:VL expression (Table 2.1), and VH:VL gene family heat maps (Figure 2.2, Figure A1.2, Figure A1.3). In the case of separate VH and VL gene amplification for complete antibody sequencing, 2x250bp reads containing the 5' V gene FR1-CDR2 and 3' CDR2-FR4 were paired by Illumina read ID and consensus sequences were constructed from reads containing the exact CDR3 of interest. Sequence data was deposited in the NCBI Short Read Archive (SRA061316).

3. IN-DEPTH DETERMINATION AND ANALYSIS OF THE HUMAN PAIRED HEAVY AND LIGHT CHAIN ANTIBODY REPERTOIRE[2]

3.1 Introduction

The determination of immune receptor repertoires using high throughput (NextGen) DNA sequencing is rapidly becoming an indispensable tool for the understanding of adaptive immunity, antibody discovery and in clinical practice[2,29,100]. However, because the variable domains of antibody heavy and light chains (VH and VL, respectively) are encoded by different mRNA transcripts, until recently it was only possible to determine the VH and VL repertoires separately, or else paired VH:VL sequences for small numbers, and more recently, moderate numbers of cells (10^4–10^5)[82], far smaller than the ~0.7–4x10^6 B cells contained in a typical 10 ml blood draw. Thus a technology for the facile determination of the paired antibody VH:VL repertoire at great depth (i.e. >10^6 cells per analysis) and for a variety of B cell subsets is still needed for clinical research[101], antibody discovery[76,79], and for addressing a host of important questions related to the shaping of the antibody repertoire[2,5–7,102–104].

[2] DeKosky, B. J. *et al.* In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat. Med.* **21,** 89-91 (2014). B.J.D. and G.G. developed the methodology and wrote the manuscript; B.J.D., T.K., G.C.I., A.D.E. and G.G. designed the experiments; B.J.D., T.K., A.R. and W.C. performed the experiments; B.J.D. carried out the bioinformatic analysis; and B.J.D. and T.K. analyzed the data.

Several techniques have been reported for detection or sequencing of genomic DNA or cDNA from single cells; however all are limited by low efficiency or low cell throughput (< 200–500 cells) and require fabrication and operation of complicated microfluidic devices[70,72–74]. Chudakov and coworkers recently reported the use of one-pot cell encapsulation within water-in-oil emulsions, cell lysis by heating at 65°C concomitant with TCR α and β reverse transcription and finally, linking by overlap extension PCR to determine TCRα:TCRβ pairings, albeit only for TCRβV7 and with a very low efficiency (approximately 700 TCRα:TCRβ pairs recovered from $8x10^6$ PBMC)[74]. This is likely because one-pot emulsions result in a high degree of droplet size dispersity and since the RT reaction is inhibited in volumes <5 nL[72] only the small fraction of cells encapsulated within larger droplets yields cDNA for further manipulation.

Inspired by methods for the production of highly monodisperse polymeric microspheres for drug delivery purposes[105,106], we developed a new technology that enables sequencing of the paired VH:VL repertoire from millions of B cells within a few hours of experimental effort and using equipment that can be built inexpensively by any laboratory. For validation we expanded *in vitro* memory B cells isolated from human PBMCs to obtain a sample that contained multiple clones of individual B cells and showed that among aliquots (technical replicates) the accuracy of VH:VL pairing is >97%. We show that ultra-high throughput determination of the paired VH:VL repertoire provides important immunological insights such as: (i) the discovery of human light chains detected in multiple individuals that pair with a wide range of VH genes, (ii) the quantitative analysis of allelic inclusion in humans, i.e. of B cells expressing two different antibodies, and (iii) estimates of the frequencies of antibodies in healthy human repertoires that display known features of broadly neutralizing antibodies to rapidly evolving pathogens.

3.2 Results

3.2.1 DEVICE CONSTRUCTION

For facile high-throughput single-cell manipulation we assembled a simple axisymmetric flow-focusing device comprising three concentric tubes: an inner needle carrying cells suspended in PBS, a middle tube carrying a lysis solution and magnetic poly(dT) beads for mRNA capture from lysed cells, and finally an external tube with a rapidly flowing annular oil phase, all of which passed through a 140 µm glass nozzle (Figure 3.1a). The rapidly flowing outer annular oil phase focused the slower-moving aqueous phase into a thin, unstable jet that coalesced into droplets with a predictable size distribution; additionally maintaining laminar flow regime within the apparatus prevented mixing of cells and lysis solution prior to droplet formation (Figure A2.1).

To evaluate cell encapsulation and droplet size distribution, MOPC-21 immortalized B cells suspended in PBS were injected through the inner needle at a rate of 250,000 cells/min while a solution of PBS containing the cell viability dye Trypan blue (0.4% v/v) was injected through the middle tubing so that dye mixed with cells at the point of droplet formation. Resulting emulsion droplets were 73±20 µm in diameter (average±SD). Trypan blue exclusion revealed that, as expected, cells remained viable throughout the emulsification process (Figure A2.2). Replacing the Trypan blue stream with cell lysis buffer containing lithium dodecyl sulfate (LiDS) and DTT to inactivate RNases resulted in complete cell lysis as indicated by visual disappearance of cell membranes from emulsion droplets.

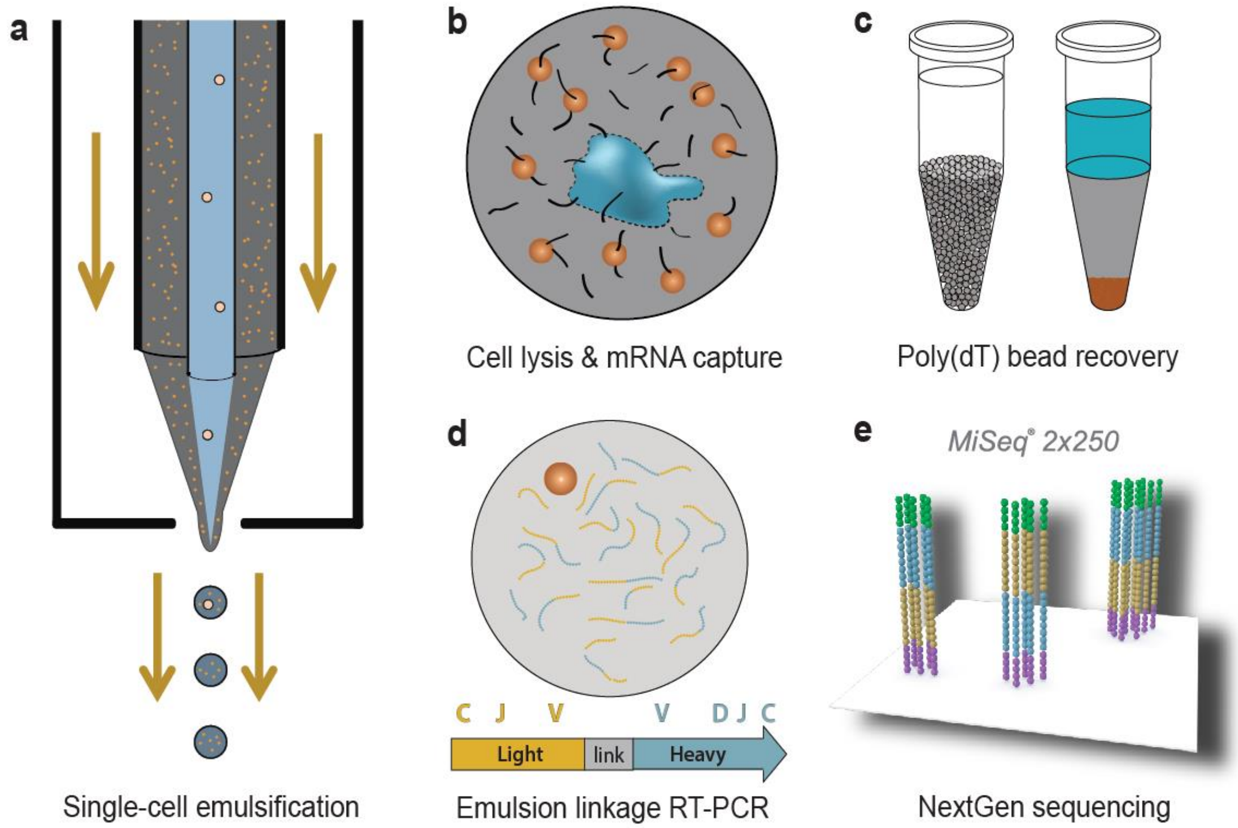


Figure 3.1 Technical workflow for ultra-high throughput VH:VL sequencing from single B cells. (**a**) An axisymmetric flow-focusing nozzle isolated single cells and poly(dT) magnetic beads into emulsions of predictable size distributions. An aqueous solution of cells in PBS (center, blue/pink circles) and cell lysis buffer with poly(dT) beads (gray/orange circles) exited an inner and outer needle and were surrounded by a rapidly moving annular oil phase (orange arrows). Aqueous streams focused into a thin jet which coalesced into emulsion droplets of predictable sizes, and cells mixed with lysis buffer only at the point of droplet formation (Fig. A2.1). (**b**) Single cell VH and VL mRNAs annealed to poly(dT) beads within emulsion droplets (blue figure represents a lysed cell, orange circles depict magnetic beads, black lines depict mRNA strands). (**c**) poly(dT) beads with annealed mRNA were recovered by emulsion centrifugation to concentrate aqueous phase (*left*) followed by diethyl ether destabilization (*right*). (**d**) Recovered beads were emulsified for cDNA synthesis and linkage PCR to generate an ~850–base pair VH:VL cDNA product. (**e**) Next-generation sequencing of VH:VL amplicons was used to analyze the native heavy and light chain repertoire of input B cells.

3.2.2 Single B cell VH:VL pairing: throughput and pairing accuracy

Human CD3$^-$CD19$^+$CD20$^+$CD27$^+$ memory B cells were isolated from PBMCs from a healthy volunteer and expanded for four days *in vitro* by stimulation with anti-CD40 antibody, IL-4, IL-10, IL-21, and CpG oligodeoxynucleotides[107]. *In vitro* expansion was performed to create a cell population containing a sufficient number of clonal B cells so that the concordance of the VH:VL repertoire in two technical replicates could be assessed. 1,600,000 *in vitro* expanded B cells were divided into two aliquots and passed through the flow-focusing nozzle at a rate of 50,000 cells/min (i.e. 16 minutes emulsification for each replicate) and processed as shown in Figure 3.1. The emulsion of lysed single cells with compartmentalized poly(dT) beads was maintained for three minutes at room temperature to allow specific mRNA hybridization onto poly(dT) magnetic beads (Figure 3.1b), then the emulsion was broken chemically (Figure 3.1c), beads were re-emulsified, and overlap extension RT-PCR was performed to generate linked VH:VL amplicons (Figure 3.1d). The resulting cDNAs were amplified by nested PCR to generate an ~850 bp VH:VL product for NextGen sequencing by Illumina MiSeq 2x250 or 2x300. Due to read length limitations of current NextGen sequencing technologies, the FRH4-(CDR-H3)-FRH3:FRL3-(CDR-L3)-FRL4 was sequenced first to reveal the pairing of the VH and VL hypervariable loops. Each of these VH:VL pairs may also comprise one or more somatic variants containing mutations within the upstream portion of the VH and VL genes. We determine the complete set of somatic variants by separate MiSeq sequencing the VH and VL portions of the paired 850 bp VH:VL amplicon followed by *in silico* gene assembly[82].

Sequence data were processed by read quality filtering, CDR-H3 clustering, VH:VL pairing, and selection for paired VH:VLs with ≥2 reads in the dataset. The clustering step resulted in high-confidence sequence data but with a lower-bound estimate of clonal diversity because clonally expanded or somatically mutated B cells with similar VH sequences collapse into a single CDR-H3 cluster. 129,097 VH:VL clusters were observed after separate analysis and clustering of Replicates 1 and 2. Of these, 37,995 CDR-H3 sequences were observed in both replicates (and hence must have originated from expanded B cells present in both technical replicates) with 36,468 paired with the same CDR-L3 across replicates revealing a VH:VL pairing precision of 98.0% (Figure 3.2, Table 3.1, Figure A2.3, see Methods). The ratio of VH:VL clusters to input cells observed (typically between 1:10 to 1:15) is a reflection of the clonality of the memory B cell population (i.e. presence of clonally related memory B cells), clustering threshold, RT-PCR efficiency and cell viability. For comparison, in our hands, sequencing the memory B cell VH repertoire by preparing amplicons directly by standard RT-PCR without pairing and using the same bioinformatic filters (sequences present at ≥2 reads, 96% clustering) resulted in a 1:6 ratio of VH clusters:input cells, which compares favorably to the yield of paired VH:VL clusters in Table 3.1. Two additional pairing analyses of somewhat smaller B cell populations from different donors were also performed (Table 3.1, Figures A2.4 and A2.5). In a separate experiment designed to verify native VH:VL pairing accuracy, plasmids encoding 11 different known human antibodies were transfected separately into HEK293 cells. Aliquots containing comparable numbers of each of the transfected cells were mixed and processed as described in Figure 3.1, and native pairings were identified for 11/11 antibodies (Table A2.1). In yet another test, approximately 260 ARH-77 immortalized human B cells[4] were mixed with 20,000 CD3$^-$CD19$^+$CD20$^+$CD27$^+$ expanded memory B cells (~100-fold excess). ARH-77 heavy and light chains were paired correctly and the ratio of correctly paired ARH-77 VH:VL reads over the top correctVH:incorrectVL, a parameter that we denote signal:topVLnoise, was 96.4:1 (2,604 correct ARH-77 VH:VL reads vs. 27 reads for the top ARH-77 VH paired with an incorrect VL, Table A2.2).

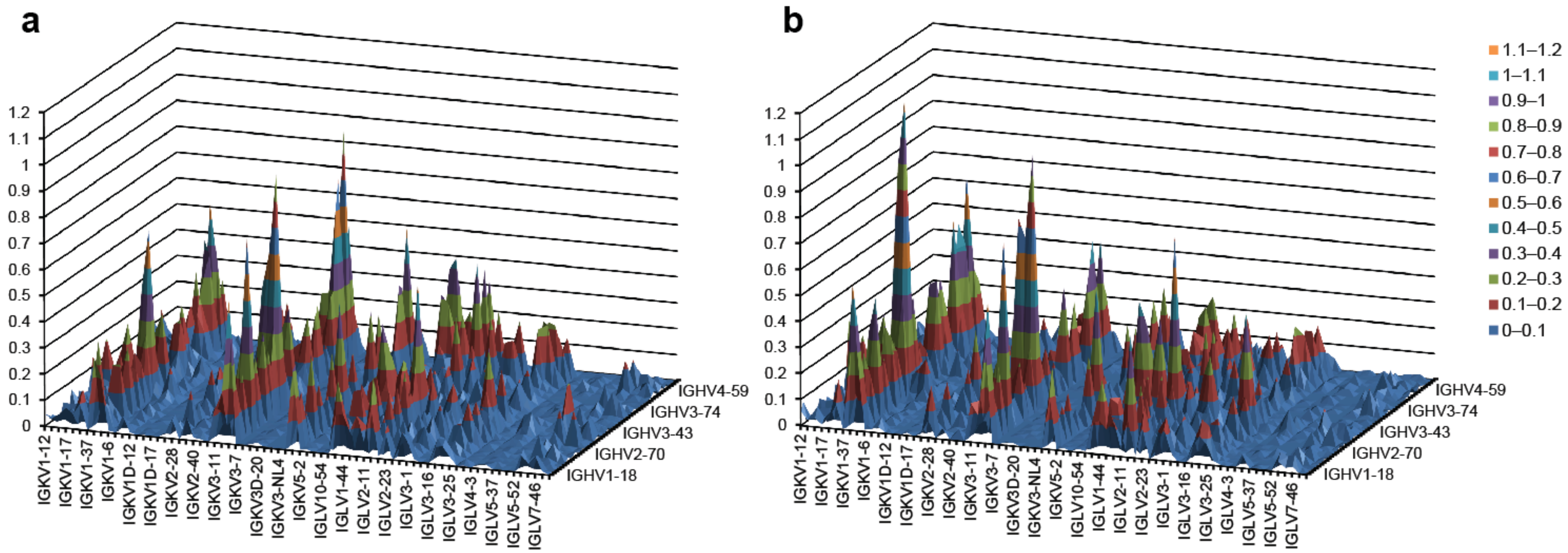


Figure 3.2 Heavy:light V-gene pairing landscape of CD3$^-$CD19$^+$CD20$^+$CD27$^+$ peripheral memory B cells in two healthy human donors. V genes are plotted in alphanumeric order; height indicates percentage representation among VH:VL clusters. (**a**) Donor 1 (*n* = 129,097), (**b**) Donor 2 (*n* = 53,679). VH:VL gene usage was highly correlated between Donors 1 and 2 (Spearman rank correlation coefficient 0.757, $p < 1\text{x}10^{-99}$). Additional heat maps are provided in Figures A2.3 and A2.4.

Human donor	V-region primer set	# cells analyzed	Emulsification rate (cells/min)	Observed VH:VL clusters	CDR-H3 detected in both replicates	CDR-H3:CDR-L3 clusters detected in both replicates	VH:VL pairing precision
Donor 1	Framework 1	1,600,000	50,000	129,097	37,995	36,468	98.0%
Donor 2	Framework 1	810,000	50,000	53,679	19,096	18,115	97.4%
Donor 3	Leader peptide	210,000	33,000	15,372	4,267	4,170	98.9%

Table 3.1 High-throughput VH:VL sequence analysis of CD3$^-$CD19$^+$CD20$^+$CD27$^+$ *in vitro*-expanded human B cells.

As discussed above, three sequencing reactions and *in silico* assembly are needed to determine the sequence of the complete linked VH:VL amplicon with Illumina MiSeq 2x250 or 2x300. Alternatively, the long-read Pacific Biosciences (PacBio) sequencing platform can be used to obtain the complete ~850bp cDNA encoding linked VH:VL sequences. However, because of its substantially lower throughput and higher cost per read, we find that despite the need for three distinct MiSeq samples compared to only one for PacBio, the former is currently much more cost-effective for deep repertoire analyses. We found PacBio sequencing to be preferable only for certain specialized applications, for example in identifying VH:VL pairs in antibodies with extensive SHM such as broadly neutralizing antibodies that arise following persistent infection with rapidly evolving viruses, most notably HIV-1. For example, we used PacBio to sequence 15,000 VH:VL amplicons from elite controller CAP256[79] and identified six variants of VRC26-class HIV broadly neutralizing antibodies within the VH:VL repertoire (Figures A2.6 and A2.7).

3.2.3 Promiscuous & public VL junctions

In contrast to the heavy chain, light chain rearrangements do not incorporate a diversity segment and exhibit restricted CDR-L3 lengths with low levels of N-addition. Light chains therefore have a much lower theoretical diversity than heavy chains and the presence of light chain sequences paired with multiple heavy chains within a single donor, referred to as "repeated" or "promiscuous" light chains, is an expected result, especially for VL junctions that are mostly germline encoded and also derive from V- and J-genes with high prevalence in human immune repertoires[108]. However the separate high-throughput sequencing of VH and VL repertoires, as has been practiced until now, cannot provide VH pairing information for a given VL and thus precludes identification and characterization of promiscuous light chains[41,109]. We observed thousands of heavy chains paired with promiscuous VL nucleotide junctions (34.9%, 29.4% and 19.6% of all heavy chains were paired with promiscuous VL junctions in Donors 1, 2, and 3, respectively). We inspected high-frequency promiscuous light chains to see if any promiscuous VL might be shared across individuals (i.e. a "public" VL). We found that highly promiscuous VLs were nearly always public: for example, of the 50 highest-frequency promiscuous VL junctions in Donor 1, 49/50 were also detected in Donors 2 and 3. Promiscuous light chains showed an average of 0.04 non-templated bases in the VL junction compared to an average of 5 non-templated bases in non-promiscuous light chains (i.e. VLs that paired with a single VH in a donor, see Figure A2.8, $p < 10^{-10}$). The lack of non-templated bases in promiscuous VL junctions indicated that promiscuity can be observed mainly in germline-encoded VL genes lacking SHM**.**

We examined in detail two representative promiscuous and public VL junctions that contained V- and J-genes with high prevalence in steady-state human immune repertoires (*KV1-39:KJ2*, 9 aa CDR-L3, *LV1-44:LJ3*, 11 aa CDR-L3, both observed at a frequency of ~1 per 1,000 VH:VL clusters)[52,92] to check for biases in VH pairing of promiscuous VL chains. *KV1-39:KJ2* and *LV1-44:LJ3* both paired with VH genes of diverse germline lineage and CDR-H3 length that reflected the overall VH gene usage in the repertoire (Figure 3.3, Spearman rank correlation coefficients: *KV1-39:KJ2* $\rho = 0.889$, $p < 10^{-21}$; *LV1-44:LJ3* $\rho = 0.847$, $p < 10^{-17}$), indicating that VL nucleotide-sequence promiscuity arises mostly from distinct VL recombination events rather than B cell activation and subsequent clonal expansion. We note that no two donors shared more than 2 VH nucleotide sequences, and no VH sequence was detected in all three donors, consistent with previous reports which showed that in contrast to VL junctions, the VH nucleotide repertoire is highly private[2,92].

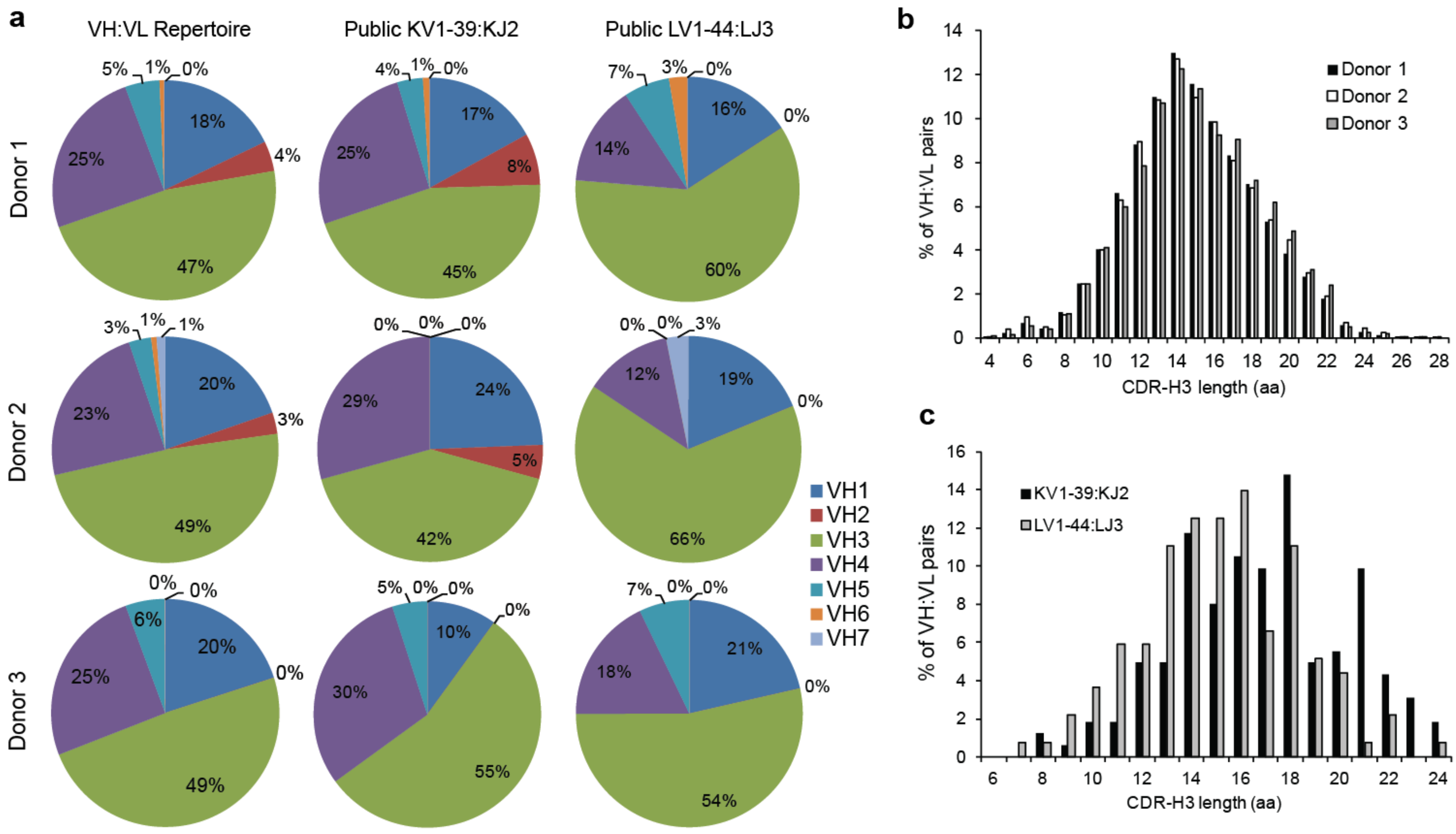


Figure 3.3 (a) VH gene family utilization in: *left* total paired VH:VL repertoires (Donor 1 *n* = 129,097, Donor 2 *n* = 53,679, Donor 3 *n* = 15,372), *center* heavy chains paired with a representative highly-ranked public and promiscuous VL observed in all three donors (*KV1-39:KJ2* 9 aa CDR-L3, tgtcaacagagttacagtaccccgtacactttt; Donor 1 *n* = 106, Donor 2 *n* = 41, Donor 3 *n* = 20), *right* heavy chains paired with a different highly-ranked public VL in all three donors (*LV1-44:LJ3* 11 aa CDR-L3, tgtgcagcatgggatgacagcctgaatggttgggtgttc; *n* = 76, *n* = 32 and *n* = 28, respectively). (**b**) CDR-H3 length distribution in VH:VL repertoires (Donor 1 *n* = 129,097, Donor 2 *n* = 53,679, Donor 3 *n* = 15,372). (**c**) CDR-H3 length distribution for all antibodies containing the two representative public VL chains from part (a).

3.2.4 Quantifying allelic inclusion in human memory B cells

Clonal selection theory postulates that each lymphocyte expresses one antibody. However, studies in mice have confirmed that this is not always the case. Allelic inclusion, the phenomenon whereby one B cell expresses two BCRs, overwhelmingly one VH gene with two different VLs, has been well-documented in mice and has been proposed to be particularly important in autoimmunity because the expression of a second BCR can dilute a pre-existing auto-reactive BCR and limit the expansion of autoreactive B cells. Similarly allelic inclusion can also provide a mechanism for autoreactive antibodies to evade central tolerance[110–114]. Almost 20 years ago A. Lanzavecchia and coworkers used FACS sorting of cells expressing both κ and λ immunoglobulin proteins on their cell surface (sIgκ$^+$/sIgλ$^+$, denoting surface-expression of both Igκ and Igλ) followed by EBV immortalization to show that sIgκ$^+$/sIgλ$^+$ allelic inclusion occurs in 0.2–0.5% of human memory B cells[115]. However, the inability to sort dual sIgκ$^+$ and dual sIgλ$^+$ human B cells and the absence of methods for the determination of the VH:VL repertoire at sufficient depth (since the frequency of allelic inclusion is low) have precluded more comprehensive determination of allelic inclusion in humans. We detected VL allelic inclusion at a rate of approximately 0.4% of VH clusters for Donor 1 and Donor 2, with dual κ/λ-transcribing B cells in approximately equal proportions to dual κ/κ- and λ/λ-transcribing B-cell clones (Figure 3.4). These heavy chains paired only with their two allelically included light chains (exact nucleotide match) in two technical replicates, and we observed that approximately 80% of these antibodies displayed somatic mutations. The somatic mutation frequency detected in allelically included VH:VL pairs was comparable to previous reports by Lanzavecchia et al. for allelically included sIgκ$^+$/sIgλ$^+$ cells (3/5 EBV-immortalized clones[115]). Also consistent with the earlier study, we observed stop codons resulting from somatic mutation that inactivated a subset of allelically included VL transcripts[115]. For the ~20% of allelically included VH that do not display SHM, we cannot rule out the possibility that these clones were derived from pre-B expansion.

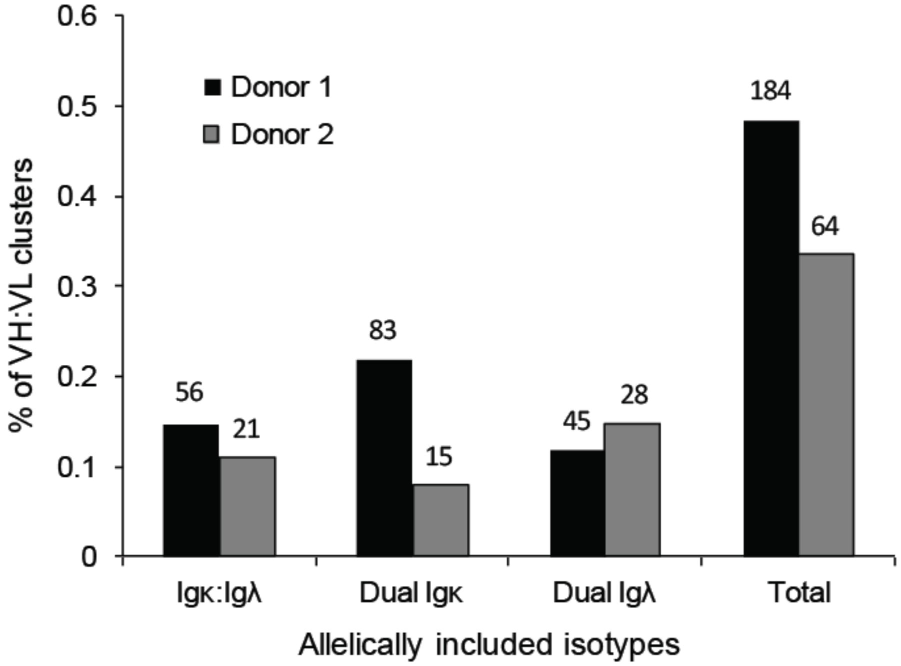


Figure 3.4 Frequency of VL transcript allelic inclusion in two donors (*n* = 184 and *n* = 64 allelically included antibodies from *n* = 37,995 and *n* = 19,096 VH:VL clusters detected across replicates in Donor 1 and Donor 2, respectively). 14 allelically included antibodies were detected in Donor 3 (8 dual κ/λ, 2 dual κ/κ, 2 dual λ/λ, *n* = 4,267 VH:VL clusters detected across replicates). Numbers above each category indicate the absolute number of observed allelically included antibodies.

3.2.5 Antibodies with gene signatures of known anti-viral bNAbs

High-resolution sequence descriptions of the immune repertoire can inform on B cell trajectories for the emergence of broadly neutralizing antibodies (bNAbs) to rapidly evolving pathogens[7,79,116,117]. Many bNAbs display highly unusual features including very long CDR-H3 and short CDR-L3 sequences[79,116,118,64], and these properties have raised the question as to whether antibodies with similar features are normally found in the repertoire of healthy donors and thus could evolve following stimulation by infection or vaccination to yield neutralizing antibodies. We found approximately 1:6,000 VH:VL clusters exhibited general characteristics of known VRC01-class anti-HIV antibodies (22, 9, and 0 for Donors 1, 2, and 3 respectively; germline *VH1-02*, a very short ≤5aa CDR-L3, and CDR-H3 length between 11 and 18 aa[64]), while antibodies with genetic characteristics of anti-influenza FI6 occurred in approximately 2–5x10^4 memory B cells (6 and 1 antibodies detected in Donors 1 and 2, respectively; *VH3-30*, *KV4-1*, 22aa CDR-H3, 9aa CDR-L3[116]).

3.3 Discussion

We have developed an easy to implement, ultra high-throughput technology for sequencing the VH:VL repertoire at relatively low cost and with high pairing accuracy. The workflow presented here permits sequence analysis of the entire population of human B cells contained in a 10 mL blood draw, or if needed, even in a unit of blood (450 ml) in a single-day experiment, an improvement orders of magnitude relative to what is feasible using robotic single-cell RT-PCR[119]. As many as 6 million B cells (or alternatively, as few as 1,000 B cells) can be analyzed per operator in a single day. Of note, the number of antibody sequences reported here (~200,000) dwarfs the entire set of <19,000 human VH:VL sequences that had been deposited in the International Nucleotide Sequence Database Collaboration (INSDC) over the past 25 years (in addition to the ~5,000 human VH:VL pairs we reported previously[82]).

The determination of the paired antibody repertoire at great depth can provide unprecedented insights on a number of medically and immunologically important issues. For example, we used HT single-cell VH:VL sequencing to detect highly-utilized promiscuous and germline-encoded VL junctions that are observed in multiple donors, to identify antibodies with bNAb-like features in HIV-1 patients[79] (Figures A2.6 and A2.7) and to quantify the frequency of bNAb-like V gene rearrangements in healthy donors, as described above. The latter is an important factor in determining whether a vaccine immunogen might be able to elicit protective immunity[64,118]. High-throughput VH:VL sequencing can also be used to search for public antibody VH:VL clonotypes[120,121] and to identify antibodies having specific features determined by computational or structural biology analyses or with relevance to pathogen neutralization[35,65,66,79,64]. In autoimmunity, high-throughput VH:VL sequencing can reveal an individual's repertoire of allelically included B cells (Figure 3.4) and the presence of B cell clones expressing antibodies containing hallmark autoimmune signatures (with respect to paratope net charge, CDR-H3 and CDR-L3 lengths, etc.) as well as other attributes of potential diagnostic and therapeutic utility[111,113,114].

3.4 Methods

3.4.1 Flow focusing apparatus:

An axisymmetric flow focusing emulsification apparatus was constructed by inserting a 26-gauge needle within 19-gauge hypodermic tubing (Hamilton Company, Reno, NV, USA) and the needle was adjusted so that the needle tip was nearly flush with the end of the hypodermic tubing (Fig. A2.1). The concentric needles were placed inside 3/8 inch OD glass tubing (Wale Apparatus, Hellertown, PA, USA) with a 140 µm orifice such that the needle exit was approximately 2 mm from the nozzle orifice. A syringe pump (KD Scientific Legato 200, Holliston, MA, USA) was used to control aqueous flow rates and a gear pump (M-50, Valco Instruments, Houston, TX, USA) was used to control oil flow rates. The flow focusing nozzle and supply lines were cleaned using 70% EtOH followed by PBS prior to all experiments.

3.4.2 Single-cell emulsification, cell viability and lysis analyses

MOPC-21 cells were resuspended at a concentration of 500,000 cells/mL in PBS and the resulting cell solution was injected through the needle at 500 µL/min. A PBS/0.4% Trypan blue solution (Sigma-Aldrich, St. Louis, MO, USA) was injected through the 19 ga hypodermic tubing at 500 µL/min and oil phase (molecular biology grade mineral oil with 4.5% Span-80, 0.4% Tween 80, 0.05% Triton X-100, v/v%, Sigma Aldrich Corp.) passed through the glass tubing at 3 mL/min. The resulting emulsion was analyzed via light microscopy (Figure A2.2); ImageJ post-processing was used to measure droplet diameters. Next the PBS/0.4% Trypan blue solution was replaced a solution of poly(dT) magnetic beads (1.0 μm diameter, New England Biosciences, Ipswich, MA, USA) pelleted and resuspended in cell lysis/binding buffer (100 mM Tris pH 7.5, 500 mM LiCl, 10 mM EDTA, 1% lithium dodecyl sulfate, 5 mM DTT) at a concentration of 45 uL magnetic bead stock/mL lysis/binding buffer to verify cell lysis. Upon emulsification no cell membranes could be observed, indicating total cell lysis (data not shown).

3.4.3 VH:VL PAIRING OF TECHNICAL REPLICATES OF EXPANDED MEMORY B CELLS

PBMC were isolated from donated human whole blood after informed consent had been obtained (Gulf Coast Regional Blood Center, Houston, TX) and non-B cells were depleted via magnetic bead sorting (Miltenyi Biotec, Auburn, CA). B cells were stained with anti-CD20-FITC (clone 2H7, BD Biosciences, Franklin Lakes, NJ, USA), anti-CD3-PerCP (HIT3a, BioLegend, San Diego, CA, USA), anti-CD19-v450 (HIB19, BD), and anti-CD27-APC (M-T271, BD). $CD3^{-}CD19^{+}CD20^{+}CD27^{+}$ memory B cells were incubated four days in the presence of RPMI-1640 supplemented with 10% FBS, 1× GlutaMAX, 1× non-essential amino acids, 1× sodium pyruvate and 1× penicillin/streptomycin (Life Technologies) along with 10 μg/mL anti-CD40 antibody (5C3, BioLegend), 1 μg/mL CpG ODN 2006 (Invivogen, San Diego, CA, USA), 100 units/mL IL-4, 100 units/mL IL-10, and 50 ng/mL IL-21[107] (PeproTech, Rocky Hill, NJ, USA, Table A2.3). Memory B cells were then resuspended in PBS at a concentration of

100k/mL and passed through the innermost, 26-gauge needle of the flow focusing device at 500 µL/min. poly(dT) magnetic beads (1.0 µm diameter, New England Biosciences, Ipswich, MA, USA) were pelleted and resuspended in cell lysis/binding buffer (100 mM Tris pH 7.5, 500 mM LiCl, 10 mM EDTA, 1% lithium dodecyl sulfate, 5 mM DTT) at a concentration of 45 µL magnetic bead stock/mL lysis/binding buffer. The cell lysis/beads mixture was passed through the 19-gauge hypodermic tubing at 500 µL/min while oil phase (molecular biology grade mineral oil with 4.5% Span-80, 0.4% Tween 80, 0.05% Triton X-100, v/v%, Sigma Aldrich Corp.) was passed through the outermost glass tubing at 3 mL/min. The emulsified stream was collected into a series of 2 mL Eppendorf tubes, and each tube was maintained at room temperature for three minutes before being placed on ice for a minimum of twenty minutes. Tubes were centrifuged at 16,000 x *g* for 5 minutes at 4°C, and the upper mineral oil layer was discarded. 200 µL cold water-saturated diethyl ether was added to break the emulsion in each tube and the tubes were centrifuged again at 16,000 x *g* for 5 minutes at 4°C to pellet the beads; for larger emulsion volumes (Donor 1) emulsion collected in Eppendorf tubes was pooled into a 50 mL conical for centrifugation at 5,000 rpm for 7 min at 4°C, and the emulsion was broken with an equal volume of cold diethyl ether after removal of the upper mineral oil layer. Magnetic beads were withdrawn using a pipette, pelleted, washed once in cold wash buffer (100 mM Tris pH 7.5, 500 mM LiCl, 1 mM EDTA) and then resuspended in 2 mL cold lysis/binding buffer (100 mM Tris pH 7.5, 500 mM LiCl, 10 mM EDTA, 1% LiDS, 5 mM DTT). Beads were washed and resuspended in OE RT-PCR mixture as in previous reports[82]. The OE RT-PCR mixture bead suspension was emulsified and thermally cycled, cDNA was extracted, and a nested PCR was performed as reported previously. Nested PCR product was electrophoresed to purify ~850 bp linked transcripts and sequenced via Illumina 2 x 250 bp sequencing. VH:VL analysis for Donor 3 was performed using the same procedure with 0.5% LiDS in cell lysis buffer (as opposed to 1% LiDS as above) and leader peptide primers were used to test VH:VL pairing precision with a different human primer set (Table A2.4)[12]. Analysis with a known ARH-77 VH:VL control cell spike was also performed using sorted and *in vitro* expanded memory B cells from one human donor, as described above. This study was approved by the University of Texas at Austin Institutional Review Board (2012-08-0031) and Institutional Biosafety Committee (2010-06-0084) and blood was drawn after informed consent had been obtained. Cells were expanded *in vitro* as described above and ~260 ARH-77 immortalized cells were spiked into a sample of ~20,000 expanded memory B cells in 3 mL PBS; single-cell VH:VL analysis was then performed as above using 1% LiDS and Framework 1 primers.

HEK293 cells were transfected with known antibodies as reported previously[82] (Table A2.1). Pacific Biosciences single-molecule real-time sequencing was performed with VH:VL amplicons derived from week 119 CD27$^+$ peripheral B cells of the CAP256 donor[79] using Framework 1 region primers.

3.4.4 Bioinformatic analysis

Raw Illumina sequences were quality-filtered, mapped to V-, D-, and J- genes and CDR3's extracted using the International Immunogenetics Information System (IMGT)[95]. Sequence data were filtered for in-frame V(D)J junctions and productive V_H and $V_{\kappa\cdot\lambda}$ sequences were paired by Illumina read ID and compiled by exact CDR3 nucleotide and V(D)J gene usage match. CDR-H3 nucleotide sequences were extracted and clustered to 96% nt identity with terminal gaps ignored (USEARCH v5.2.32[122]) and resulting VH:VL pairs with ≥2 reads comprised the list of VH:VL clusters for each data set. To determine the complete VH:VL sequence of the entire approx. 850 nt amplicon by Illumina Miseq 2x250 or 2x300, the separate VH and VL amplicons were sequenced as separate samples and CDR3 junction regions were used as anchors for consensus sequences of each VH:VL pair[82].

Pacific Biosciences sequencing data were pre-processed with the PacBio SMRT Analysis suite followed by IMGT analysis of paired VH and VL regions. Complete VRC26-class antibody variable region sequences were identified from PacBio circular consensus sequences via BLAST search (Table A2.6 and A2.7). Allelically included sequences for Donors 1, 2, and 3 were detected using an iterative loop that identified heavy chain clusters paired with multiple light chains which were encoded by distinct Vκλ and Jκλ genes, and pairings were cross-confirmed in both replicates.

DNA sequences can be downloaded from the NCBI Short Read Archive (SRA) under study accession number SRP047462, and computer source code is available in GitHub repository NMED-NT69968. Randomization was performed by random number generation and statistical comparisons performed using a single-factor analysis of variance followed by a Tukey's HSD post-hoc test with two-tailed p-values (IBM SPSS v.20, Table A2.8).

Precision, *P*, (also called positive predictive value, *PPV*) is calculated from the number of true positives (TP) and false positives (FP)[123]:

$$P = \frac{TP}{TP + FP}$$

True VH:VL pairs cannot be known for an entire human antibody repertoire, but we can approximate true positives and false positives with matched vs. mismatched VH:VL pairs in technical replicates. We assume that a VH paired with the same VL in both replicates is a true positive in replicates 1 and 2 (TP_{1and2}), conversely, we assume any VH paired with a different VL in the two replicates is a false positive in at least one replicate (FP_{1or2}). Thus the aggregate precision of two independently analyzed replicates 1 and 2 is:

$$P_{1and2} = \frac{TP_{1and2}}{(TP_{1and2} + FP_{1or2})}$$

Probability theory dictates that the joint probability of independent events is equal to the product of the independent event probabilities; additionally, pairing precision of technical replicates 1 and 2 (P_1 and P_2, respectively) are assumed to be equal as a property of technical replicates.

$$P_{1and2} = P_1 \times P_2 = P^2$$

We can combine the previous two equations and solve for *P*, the VH:VL pairing precision of a single analysis.

$$P_{1and2} = P^2 = \frac{TP_{1and2}}{(TP_{1and2} + FP_{1or2})}$$

$$P = \sqrt{\frac{TP_{1and2}}{TP_{1and2} + FP_{1or2}}}$$

For a hypothetical experiment: if 950 VH sequences were observed with matching VL in both groups (TP_{1and2} = 950) while 50 VH sequences displayed disparate VL in the two groups, VH:VL pairing precision P is estimated by $P = (950/1{,}000)^{0.5}$ = 97.5%.

4. PAIRED VH:VL ANALYSIS OF NAÏVE B CELL REPERTOIRES AND COMPARISON TO ANTIGEN-EXPERIENCED B CELL REPERTOIRES IN HEALTHY HUMAN DONORS

4.1 Introduction

Obtaining an extensive sequence database for the naïve B cell repertoire is of great importance for immunology and medical research because the ensemble of sequences that comprise the naïve repertoire will ultimately dictate whether an organism has the ability to recognize a particular chemical species. Only when an antibody in the naïve repertoire can bind to antigen even with very low affinity, it becomes possible for the respective B cell to undergo stimulation and differentiation to ultimately produce soluble antibodies. The comprehensive interrogation and analysis of the naïve repertoire requires that both the natively paired VH and VL gene sequences (i.e. originating from individual B cells) are determined at high throughput. However, limitations of standard high-throughput antibody repertoire sequencing technologies had so far limited comprehensive descriptions of paired heavy and light chain variable sequences (VH:VL) in human naïve B cells. Earlier studies of human naïve B cell repertoires employed single-cell RT-PCR (Chapter 1) and as a result throughput was limited to at most approximately 10^3 total B cell clones per experiment[124–126]. More recently, high-throughput DNA sequencing methods have been applied for analysis the naïve B cell repertoire[92,127,128]. However, due to the limitations of high-throughput sequencing that lead to loss of antibody heavy and light chain information (Chapter 1), these recent high-throughput analyses were unable to analyze the complete heavy and light chain antibody clonotypes. In particular, high-throughput VH:VL sequencing and structural analysis of both naïve and antigen-experienced B cell repertoires may uncover key differences across B cell subsets and lead to new insight on antibody clonal selection and development mechanisms. Detailed knowledge of gene usage and biochemical or structural features for effective antibody development in healthy donors may also inform design of advanced immunogens for novel vaccine approaches[118,129,130] or enhance our understanding of aberrant antibody development and its role in autoimmune disease[40,131,24,132].

To address the aforementioned technical issues in high-throughput antibody sequencing we recently reported a new method that obtains paired heavy and light chain sequence information at very high throughput from single B cells (up to $5x10^6$ input cells per sample)[83]. Here, we apply this technique to analyze naïve B cell repertoires in three separate healthy human donors and compared our results to previously reported data generated from antigen-experienced populations of the same blood samples. In doing so, we observed distinct universal patterns of gene usage, biochemical metrics, and antibody structure that have never before been reported. These data highlight the extensive similarities and key differences across human donors that comprise the adaptive immune response of our species.

4.2 Results

4.2.1 VH:VL GENE USAGE ACROSS B CELL SUBSETS

We analyzed the peripheral blood antibody repertoire of three human donors using the single B cell VH:VL sequencing technique reported and described in Chapter Four[83]. Peripheral blood mononuclear cells (PBMC) isolated from three human donors were sorted by FACS to discriminate $CD3^{-}19^{+}20^{+}27^{-}$ naïve B cell and $CD3^{-}19^{+}20^{+}27^{+}$ antigen-experienced B cell populations. Our sequencing method incorporated single-cell sequestration and lysis in emulsion droplets to capture single-cell heavy and light chain mRNA onto poly(dT) magnetic beads, followed by emulsion overlap extension RT-PCR that linked heavy and light chain sequences to generate a single heavy:light cDNA strand for high-throughput VH:VL DNA sequencing and analysis[83]. We recovered a total of 55,355 unique naïve B cell sequences after VH:VL pairing using the technology described in Chapter Four and compared these data to 123,941 distinct CDR-H3:CDR-L3 pairs recovered from antigen-experienced cells in the same donors (Table 4.1; antigen-experienced raw data was previously reported and reprocessed for the present study[83]). Naïve and antigen-experienced heavy/light paired V-gene usage were visualized by surface plots (Figure 4.1), and paired V-gene usage in antigen-experienced repertoires was further subdivided by heavy chain isotype for analysis (Figures A3.1 and A3.2).

Donor	**CD3$^-$19$^+$20$^+$27$^-$ Naïve**	**CD3$^-$19$^+$20$^+$27$^+$ Ag-Exp**
1	13,780	34,692
2	26,372	89,249
3	15,203	-
Total	**55,355**	**123,941**

Table 4.1 Paired heavy:light B-cell receptor sequences recovered for naïve and antigen-experienced (Ag-Exp) B cell subsets after 96% clustering and quality filtering of sequence data (see Methods). Antigen-experienced raw data was re-processed alongside naïve B cell data for consistency[83].

We applied Pearson hierarchical clustering to analyze the VH:VL repertoires for similarity in V-gene usage (Figure 4.2), which yielded several striking results:

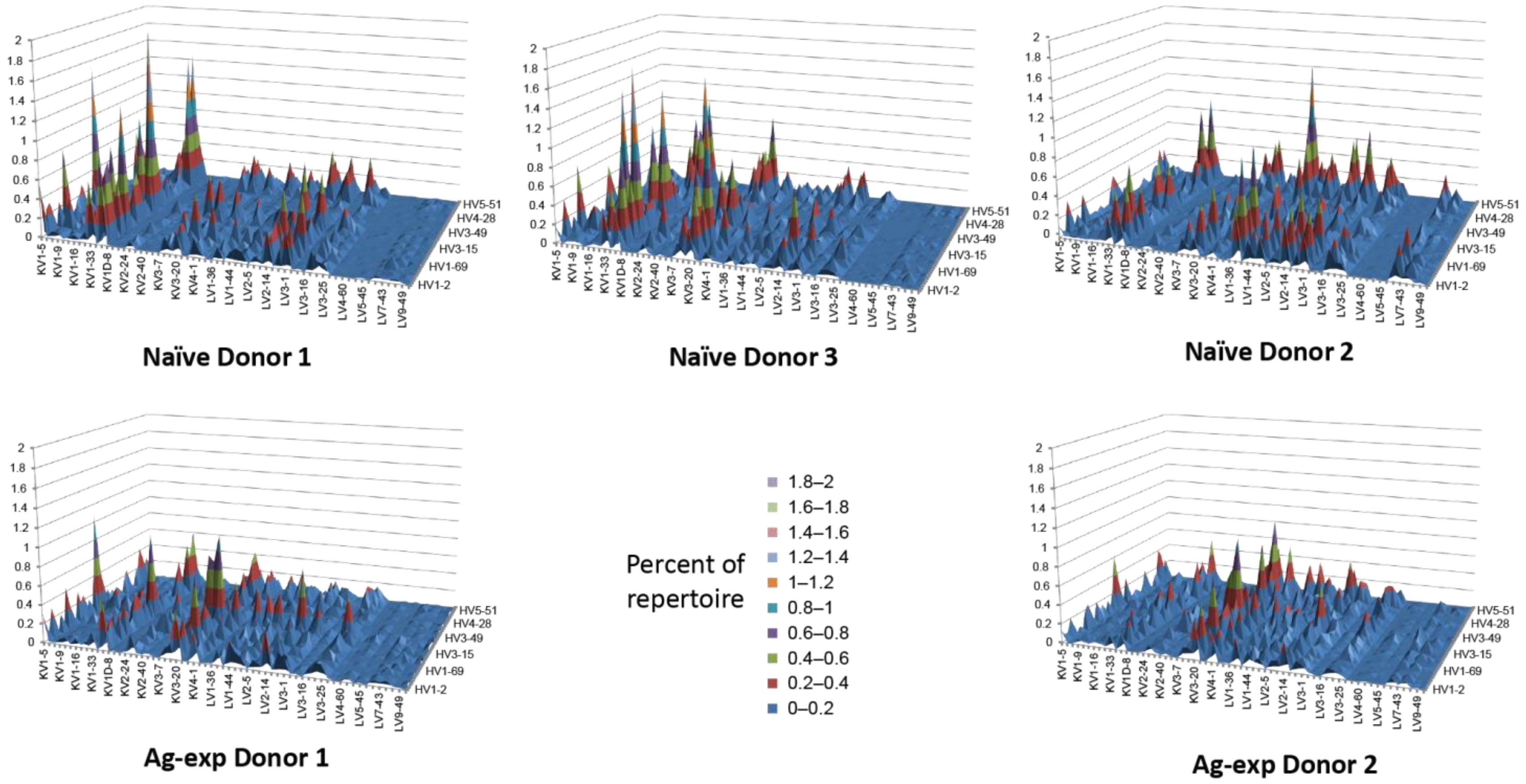


Figure 4.1 Paired heavy/light V-gene usage surface maps of sequenced antibody repertoires. Consistent trends in gene usage were readily observed, and the antibody repertoires of each donor and subset were distinct. Statistical analysis of VH:VL gene usage data presented here was performed with Pearson hierarchical clustering (Figure 4.2).

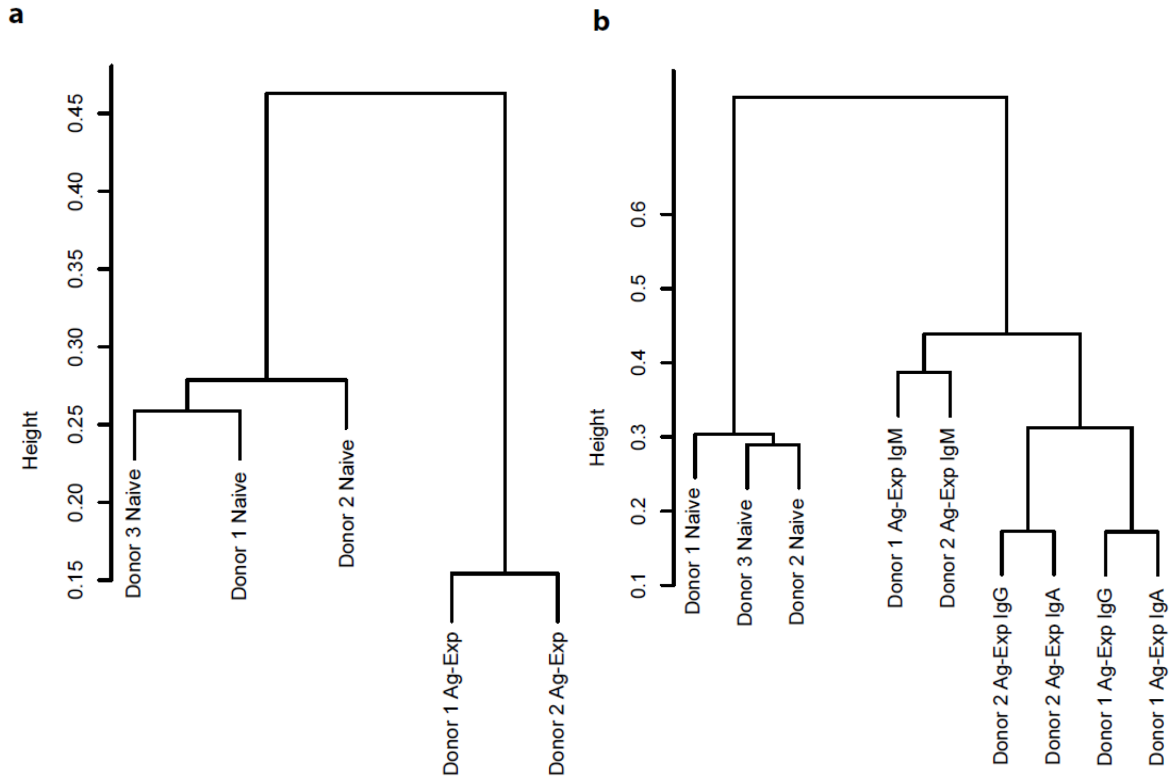


Figure 4.2 Clustergrams resulting from Pearson hierarchical cluster analysis of paired heavy and light chain V-gene usage in sequenced donor repertoires. Panel (a) compares naïve repertoires to antigen-experienced repertoires, whereas panel (b) compares naïve repertoires to each of three antigen-experienced repertoire heavy chain isotype subsets (IgM, IgA, and IgG). Relative distance is indicated by line heights connecting different groups.

i. Naïve repertoires of different donors clustered together, and antigen-experienced repertoires of different donors also clustered together as a separate grouping (Figure 4.2a). These data indicated that V-gene usage in the naïve repertoire of a given donor was more similar to the naïve repertoire of *other* donors than it was to the antigen-experienced repertoire in the same individual.

ii. Upon further subclassification of antigen-experienced repertoires by heavy chain isotype, we found that IgM repertoires across donors clustered as one grouping, and class-switched repertoires (IgG, IgA) clustered as another grouping (Figure 4.2b). Importantly, IgM repertoires were found to cluster in between the naïve and class-switched repertoires, which may be a direct consequence of IgM memory B cells serving as a transitional stage between mature naïve B cells and class-switched antigen-experienced B cells.

iii. In contrast to naïve and ag-exp IgM B cell subsets, the ag-exp IgG and IgA subsets revealed more individuality across donors. Figure 4.2b shows that the class-switched IgG and IgA repertoires of Donor 1 were more similar to each other than to the complementary IgG or IgA repertoire in Donor 2.

We also observed higher variability in IgK and IgL gene usage in naïve repertoires compared to antigen-experienced repertoires (Figure A3.3). The variation in IgK:IgL usage may have caused a broader spread among the naïve repertoire cluster compared to the antigen-experienced cluster (Figure 4.2a).

A longstanding question in antibody development is whether certain heavy/light V-gene combinations are favored or disfavored relative to what might be expected due to random VH:VL pairing given overall VH- and VL-gene expression in the repertoire. Presumably, any “holes” in VH:VL gene usage compared to overall VH:VL frequency expectation (i.e. compared to random VH:VL pairing expectation based on the fraction expression of each heavy and light V-gene) might indicate structural mismatch between heavy and light chain V-genes that prevent successful display of B-cell receptors, or other functional implications (e.g. holes in naïve repertoire heavy/light V-gene pairing could arise from heavy/light V-gene structural mismatch or particularly autoimmunogenic combinations). On the other hand, "peaks" in VH:VL gene usage could indicate a highly effective VH:VL gene pair (e.g. in ag-exp cells, a gene usage peak could result from VH:VL pairings that are protective against common pathogens). We applied several statistical techniques in search of "holes" and "peaks" in all sequenced repertoires (namely linear model-based t-tests[133], DESeq[134], and Student's t-test). Importantly, all methods revealed no holes nor peaks in the repertoire with statistical significance given the sample sizes obtained in the present study. While no statistical significance was observed across all three donors, several putative holes and peaks were observed within individual donors in both naïve and antigen-experienced subsets. VH:VL pairing peaks and holes that were unique to each individual could have occurred by any combination of genetic variation, and environmental exposure, and sampling error, and the relative contributions of these three factors is unknown.

Finally, we analyzed heavy/light V-gene combinations for V-gene pairs that were statistically enriched or depleted in antigen-experienced repertoires compared to the naïve repertoires using linear model-based t-tests[133]. We found a total of 29 statistically significant heavy/light V-gene pairings with an adjusted *p*-value less than 0.05 that indicated differences in gene expression between naïve and antigen-experienced groups (listed in Table A3.1). Interestingly, many of these V-genes observed to be differentially expressed in the paired heavy/light naïve and ag-exp repertoires were also observed with differential V-gene expression by separate heavy and light chain sequencing[92,127,128].

4.2.2 CDR3 LENGTH ANALYSIS

Prior studies have shown that CDR3 length decreases during B cell maturation into naïve B cells[12] and again increases only slightly in naïve compared to antigen-experienced repertoires[127], and also single-cell RT-PCR studies found no strong correlation between heavy and light chain lengths[124,135]. We determined the CDR-H3 and CDR-L3 lengths of naïve and antigen-experienced repertoires and compared our results to previous studies. First, we observed that the average CDR-H3 length was slightly lower in antigen-experienced repertoires compared to naïve repertoires, however the averages and medians were very similar. Importantly, the CDR-H3 length distribution was markedly narrower in antigen-experienced compared to naïve B cells, and these differences were significant by the Kolmogorov–Smirnov test for probability distribution analysis (p < 10^{-14}**,** Figure 4.3a). In contrast, CDR-L3 length distributions showed more similarity across repertoires (Figure 4.4b), and broader error bars in naïve CDR-L3 lengths resulted from variability in IgK and IgL gene usage in naïve repertoires (Figure A3.3) because IgK and IgL possess distinct CDR-L3 length distributions (maxima at 9aa in IgK versus 11aa in IgL).

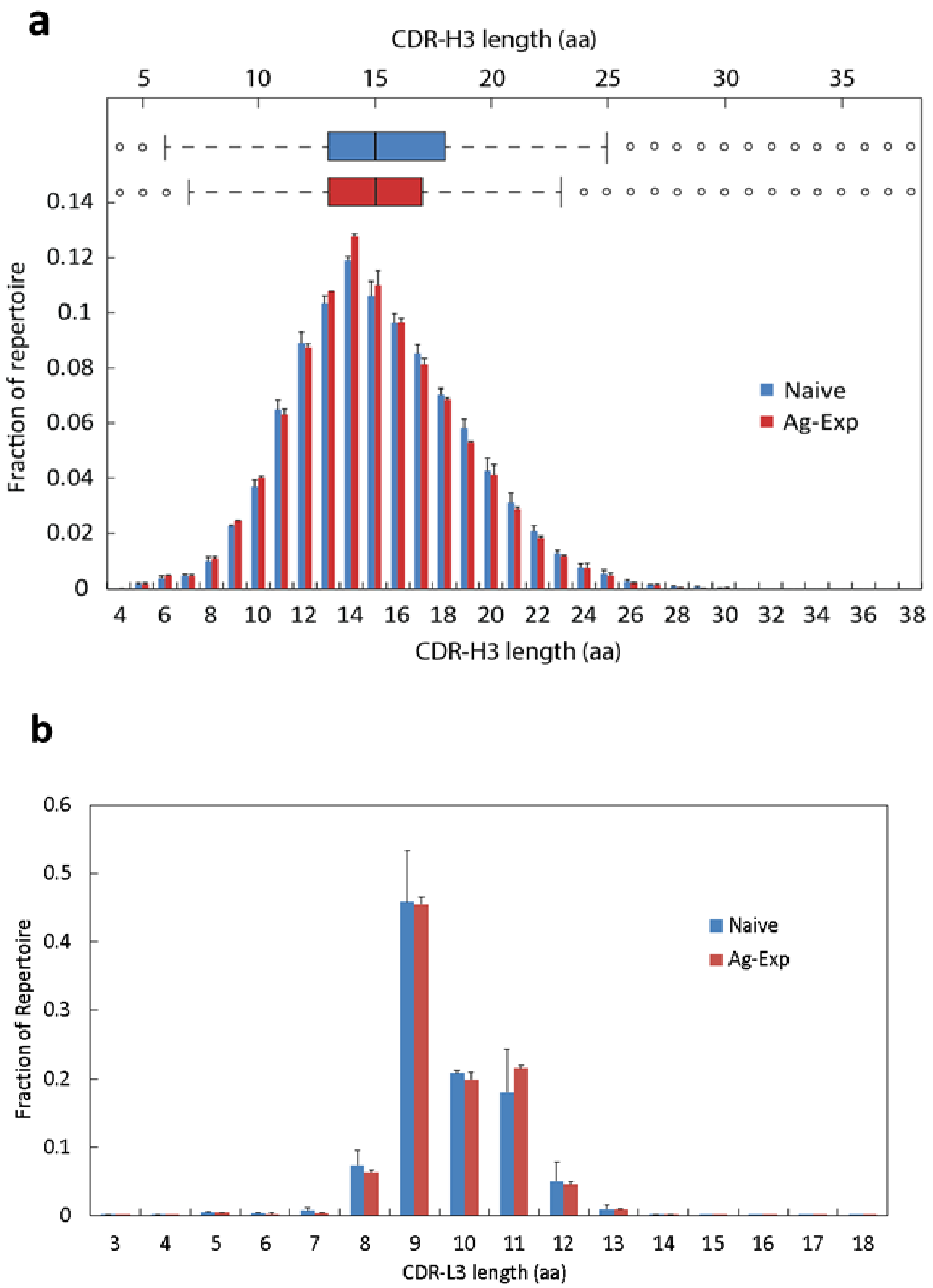


Figure 4.3 Distribution histograms (average ± standard deviation) for (a) CDR-H3 amino acid length, and (b) CDR-L3 amino acid length, averaged across all three donors.

Finally, we observed no strong correlation in paired heavy and light chain CDR3 loop lengths, similar to previous reports[124,135] (Figure A3.4), suggesting that antibody heavy and light chain sequences pair randomly with respect to CDR3 length.

4.2.3 CDR3 CHARGE

Previous reports have observed increases in average CDR-H3 positive charges in antigen-experienced repertoires compared to naïve repertoires[127], and we inspected the CDR3 charge distributions in our data to see if paired heavy and light chain CDR3 displayed similar charge features compared to these heavy chain-only observations. Both naïve and antigen-experienced repertoires exhibited charge distributions with maxima at neutral charge, with 90% of the repertoire falling between +2 and -2 by total CDR3 loop charge (Figure 4.4a). We observed that the antigen-experienced repertoire was enriched for positive charges compared the naïve repertoires on an overall CDR-H3:CDR-L3 basis (Figure 4.4a), on a heavy chain-only basis (Figure 4.4b), and on a light chain basis (Figure 4.4c); differences in charge distribution for all three naïve:antigen-experienced repertoire comparisons in Figure 4.4 were statistically significant by the K-S test ($p < 10^{-14}$). B-cell receptors with charge extremes were also slightly more prevalent in antigen-experienced repertoires (±6, 5, and 4 in the CDR-H3:CDR-L3, CDR-H3, and CDR-L3, respectively) however small sample size at charge extremes makes these observations subject to high sampling error. We also observed distinct differences in CDR-L3 charge between kappa and lambda light chain repertoire subsets. In the naïve repertoire, the majority of kappa and lambda light chain CDR3 loops were close to neutral charge, however naïve kappa light chains showed a median charge of zero compared to lambda light chains with a median charge of negative one (Figure A3.5). Interestingly, kappa light chains were strongly selected for enhanced positive charge in antigen-experienced repertoires, whereas lambda light chains showed enhancement of charge extremes (both positive and negative) in antigen-experienced repertoires. Importantly, we note that Donor 1 and Donor 2 both displayed enhanced positive CDR3 charges in antigen-experienced repertoires when compared to their naive repertoires (Figure A3.6). The enhancement of positive charge in both heavy and light CDR3 loops concurrently is presented as a heat map in Figure A3.7.

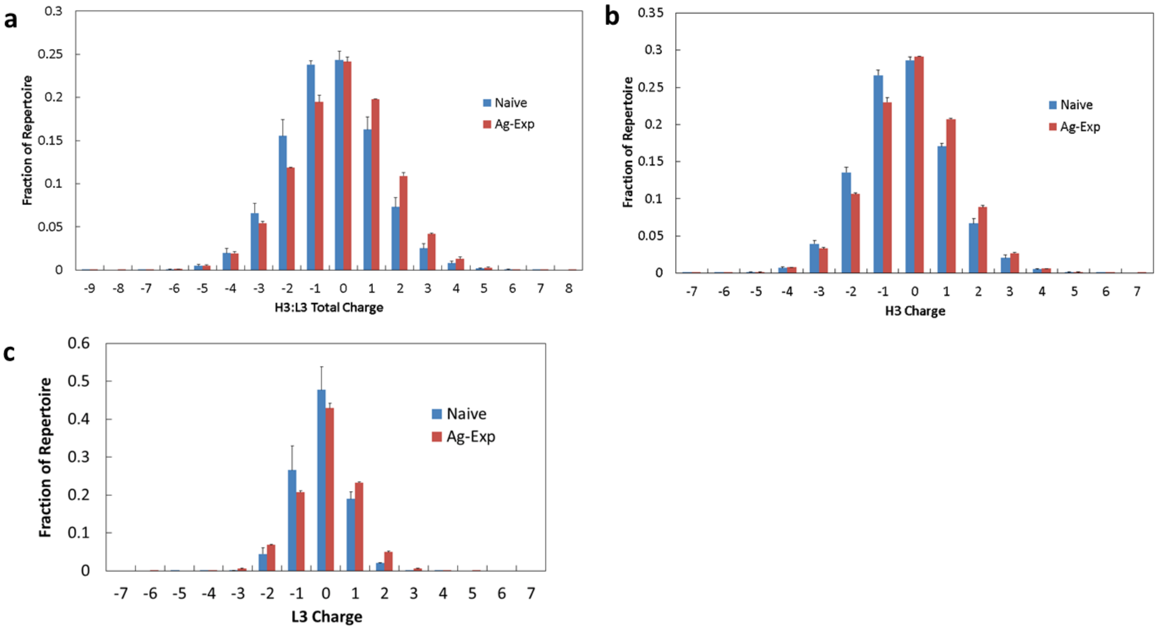


Figure 4.4 CDR3 charge distribution for naïve and antigen-experienced repertoires (average ± standard deviation) in (a) total CDR-H3 and CDR-L3 charge, (b) CDR-H3 charge, and (c) CDR-L3 charge. Differences in charge distribution between naïve and antigen-experienced repertoires for all three panels were statistically significant by the K-S test ($p < 10^{-14}$)

In terms of CDR3 charge we again observed distinct patterns across the lambda and kappa light chain repertoires. In the naïve repertoire, the majority of kappa and lambda CDR-L3 were close to neutral charge, however kappa CDR-L3 had a median charge of zero compared to lambda CDR-L3 which showed a median charge of negative one. Antigen-experienced kappa light chains were also more strongly selected for positive charges, compared to the somewhat weaker charge selection and enhancement of both positive and negative charge extremes in lambda light chains (Figure A3.5).

4.2.4 CDR3 HYDROPHOBICITY

We analyzed the average hydrophobicity per residue (avg H-index) in paired CDR-H3 and CDR-L3 loops[136]. We found that mean hydrophobicity decreased in antigen-experienced BCR relative to naïve BCR on a heavy chain basis (0.0476 ± 0.002 for naïve, compared to 0.0389 ± 0.006 for antigen-experienced). However, the CDR-H3 hdyrophobicity distribution peak shifted toward a moderately positive H-index in antigen-experienced repertoires, indicating enrichment of moderately hydrophobic CDR-H3s in antigen-experienced repertoires (average H-index between 0 and 0.1, Figure 4.5). These trends were replicated in both Donor 1 and Donor 2, and especially highly hydrophobic CDR-H3 were depleted in the antigen-experienced repertoires while lower hydrophobicity CDR-H3s were well-tolerated.

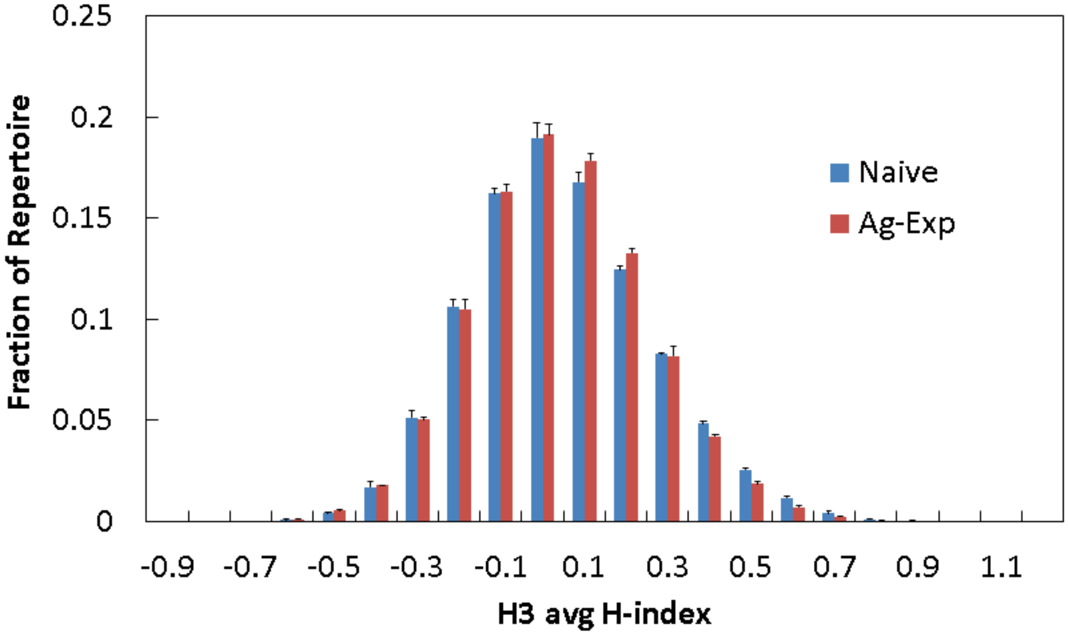


Figure 4.6 CDR-H3 loop average hydrophobicity (avg H-index ± standard deviation) distributions in naïve and antigen-experienced repertoires.

Regarding CDR-L3 hydrophobicity we again observed strong differences between the kappa and lambda repertoires. Overall, lambda light chain CDR3s were much more hydrophobic than kappa CDR3s, and these patterns were consistent across donors (Figure 4.7). We also observed that kappa light chains were strongly selected for enhanced hydrophobity indices and to a much greater extent than lambda light chains within the same donor (Figure 4.7). As lambda light chain distributions showed little differences between naïve and antigen-experienced repertoires apart from a broadening of the H-index distribution, we observed that kappa CDR-L3 are under stronger H-index positive selection pressure than lambda light chains, perhaps because of the much lower starting hydrophobicity of kappa light chains.

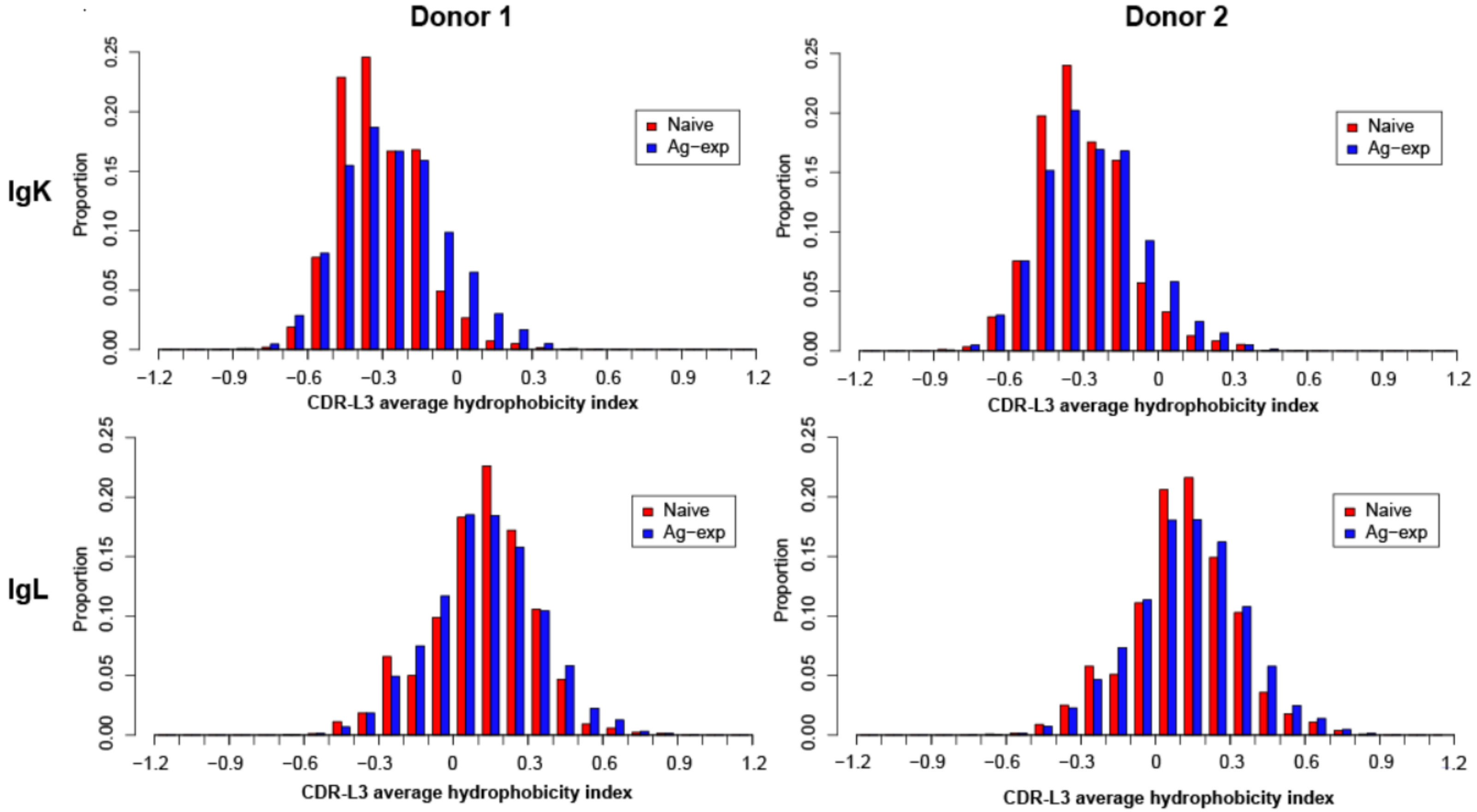


Figure 4.7 CDR-L3 loop average hydrophobicity indices in naïve and antigen-experienced antibody repertoires for Donor 1 (*left*) and Donor 2 (*right*), subdivided by IgK (*top*) and IgL (*bottom*). Kappa and lambda repertoires exhibited distinct CDR-L3 average hydrophobicity distributions (top compared to bottom graphs), and kappa light chains showed enhanced CDR-L3 hydrophobicity in antigen-experienced repertoires. All four naïve repertoires were statistically significant from antigen-experienced repertoires in terms of CDR-L3 average H-index by the K-S test ($p < 10^{-12}$); *n* for the above repertoires is provided in Table A3.2.

4.2.5 PUBLIC HEAVY AND LIGHT CHAIN SEQUENCES

Previous reports highlighted the existence of promiscuous and public light chain CDR3 antibody sequences in human immune repertoires due the lower diversity encoded by light chain VL junctions[83,108]. Consistent with these earlier studies, we found widespread promiscuous and public VL junctions in naïve repertoires by both amino acid and nucleotide sequences. For example, 68.4 ± 4.5 percent of naïve BCR were encoded by a promiscuous nucleotide CDR-L3 junction (i.e. a VL junction also expressed by at least one other BCR in the same donor), whereas 78.5 ± 4.0 percent were encoded by a promiscuous amino acid CDR-L3 junction. SHM dramatically reduced the fraction of promiscuous VL junctions observed in antigen-experienced repertoires (30.2 ± 3.9 percent by nucleotide basis, 46.9 ± 5.7 by amino acid basis), however a significant fraction of antigen-experienced light chains were still encoded by promiscuous VL sequences in the ag-exp repertoires. Despite widespread promiscuity observed in VL nucleotide sequences, we observed very few public CDR-H3 nucleotide sequences across individuals (three among all naïve donors), similar to previous reports[2,92].

While public CDR-H3 nucleotide sequences are extremely rare across individuals, public CDR-H3 amino acid sequences are known to occur as a result of similar antibody responses to common antigens via vaccination or infection[78,120,121,131]. We observed a total of 23 exact amino acid match CDR-H3 amino acid junctions among naïve repertoires, and 63 amino acid exact match CDR-H3 among antigen-experienced repertoires. We expected public amino acid CDR-H3 lengths to be significantly shorter than that of the overall repertoire because short sequences have lower theoretical diversity and therefore a higher probability of repetition by random occurrence, and indeed this was the case (Figure 4.8). However, while VH, VL, and JL gene usage was very different in naïve public CDR-H3 aa pairs, we observed a marked increase in convergent gene usage across individuals in the antigen-experienced repertoires (Figure 4.9). The relatively high frequency of cognate light chain gene usage convergence in the public antigen-experienced CDR-H3 aa sequences (59% matching VL genes across antigen-experienced antibodies encoding the same CDR-H3) suggested a functional selection for these public, antigen-experienced BCR. In contrast, the mostly incongruent VL gene usage that we observed among the naïve public CDR-H3 set (13% matching VL genes across naïve antibodies encoding the same CDR-H3) suggested random or "incidental" sequence convergence of the heavy chain CDR-H3 amino acid sequences without functional selection. Importantly, distinct patterns of N/P addition, SHM, and distinct CDR-L3 lengths across donors (despite identical light chain V- and J- gene usage) demonstrated that the convergent public VH:VL clonotypes in antigen-experienced repertoires derived from biologically distinct V-D-J and V-J recombination events (Table A3.2).

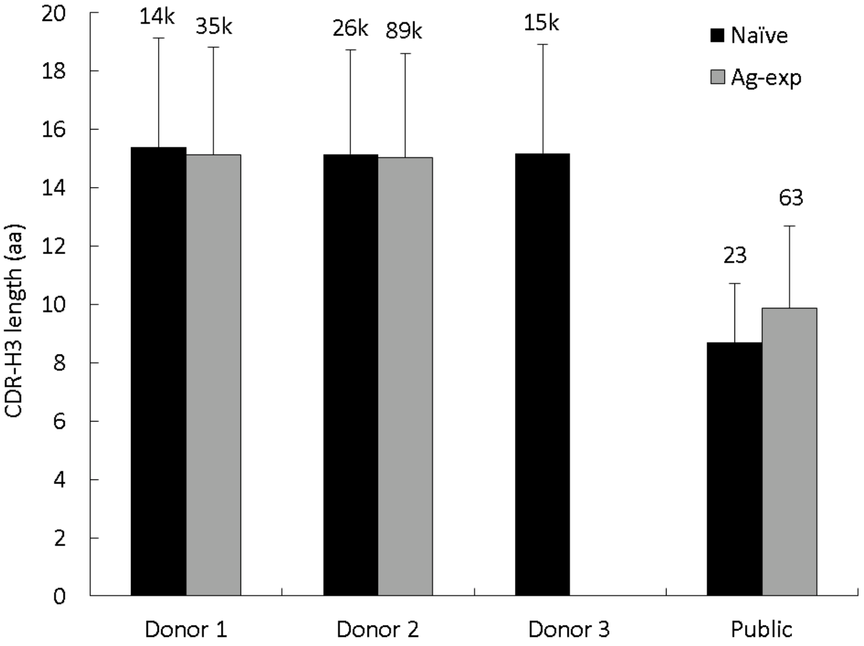


Figure 4.8 CDR-H3 length comparisons between overall repertoires and public CDR-H3 amino acid sequences (average ± standard deviation). Values above each column indicate the total number of CDR-H3 in each group.

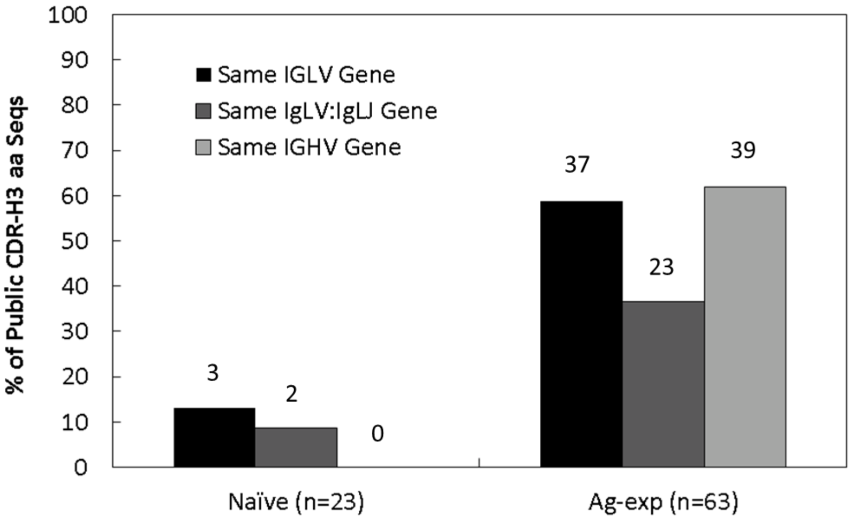


Figure 4.9 Gene usage comparisons between public CDR-H3 amino acid. Values above each column indicate the total number of public CDR-H3 in each group.

4.3 Discussion

Our high-throughput paired heavy and light chain analysis of the human naïve antibody repertoire yielded extensive information on the composition of the human naïve repertoire and the universal shifts in heavy:light antibody repertoire characteristics as naïve B cells mature into antigen-experienced cells. We found VH:VL gene usage was distinct among the various B cell subsets, even across multiple human donors (Figure 4.2a), indicating that universal mechanisms direct paired V-gene usage throughout the transition from naïve to antigen-experienced IgM to class-switched B cells. We also observed that IgM naïve and antigen-experienced repertoires showed higher VH:VL gene usage similarity across donors than to other B cell subsets isolated in the same individual, whereas V-gene usage in class-switched IgG and IgA repertoires was more similar within donors than across donors (Figure 4.2b). This result suggested that the naïve repertoire and IgM antigen-experienced repertoires of different donors have similar composition, whereas environmental exposure (which is different for each individual) manifests most strongly in the class-switched repertoire. As expected, we found that IgM VH:VL gene usage clustered in between naïve and antigen-experienced cells, which was highly consistent with the fact that IgM antigen-experienced cells can be considered a transitional stage between the naïve and class-switched repertoires. Our analysis of intra- and inter-donor V-gene usage confirmed prior reports that highlighted universal regulation across antibody repertoire subsets[137] and provided additional information regarding paired heavy/light usage across multiple donors and B cell subsets.

We also inspected naïve and antigen-experienced repertoires for evidence of preferential pairings or prohibited pairings between heavy and light chain V genes. Though preferential V-gene pairings were detected within individual repertoires, we found no statistically significant preferential heavy and light chain pairs across human donors. Structural mismatch between certain heavy and light chains V genes could still occur with a magnitude of effect smaller than detection limits for the approx. 55,000 naïve sequences analyzed in this experiment. Heavy/light pairing bias may also have predominantly affected only those V-genes with low expression in the repertoire and therefore with low total observations, or pairing bias observed in individual repertoires may have resulted from sampling variation or genetic and environmental differences across individuals. In summary, our data suggested that heavy and light chain V-genes pair randomly – and for the most part, successfully – according to overall V-gene prevalence in B cell repertoires.

We also observed that the biochemical composition of heavy and light chains was altered during the process of B-cell selection. In particular, CDR-H3 length distribution narrowed slightly (consistent with previous reports[127]) while CDR-L3 lengths remained largely unchanged after maturation to antigen-experienced B cells (Figure 4.3), indicating that the highly utilized CDR3 loop lengths are optimal for binding to most antigens. Both heavy and light chain CDR3 charges increased in antigen-experienced repertoires (Figure 4.4), and slight differences in charge composition were observed between kappa and lambda repertoires (Figure A3.5). We hypothesize that the general increases in antibody charges may better enable binding to negatively charged bacterial membranes for antibacterial antibodies. Importantly, we found that increases in H3 and L3 charge often occurred concurrently (Figure A3.7), suggesting similar selection pressure toward enhanced positive charges on both CDR-H3 and CDR-L3 loops. We also observed that kappa light chains appeared to be more strongly selected for positive charge in the antigen-experienced repertoire compared to lambda light chains, despite the fact that kappa light chains are overall more positively charged in the baseline naïve repertoire (Figure A3.7). These results suggest that kappa CDR3 loops may be slightly more effective at carrying positive charges and/or binding to negatively charged epitopes than lambda light chains.

We also observed that total hydrophobicity of CDR-H3:CDR-L3 loops increased slightly, which may help enhance antibody binding affinity in the antigen-experienced repertoire. Hydrophobicity in kappa light chain CDR3s was strongly increased in antigen-experienced repertoires compared to only minor changes is lambda light chain CDR3 hydrophobicity as a result of positive B cell selection (Figure 4.7). These data highlighted the important differences between kappa and lambda light chains – namely, that kappa light chains are overall much less hydrophobic in the CDR3 and that selection pressures causes a general increase in kappa hydrophobicity, whereas lambda light chains start slightly more hydrophobic and are under somewhat less selection pressure for hydrophobicity.

We noted several key differences in CDR3 length, charge, and hydrophobicity between kappa and lambda light chains, and these differences may serve important functional purposes. For example, receptor editing is known to mitigate self-targeting of autoimmunogenic antibodies[111,113,114]. By having two different light chain gene sets – each with a distinct distributions of properties in terms of CDR3 length, charge, and hydrophobicity – the immune system can have a greater likelihood of altering binding specificity by editing the light chain isotype. These differences between kappa and lambda repertoires may also have functional importance in allelic inclusion because kappa and lambda allelically-included antibodies are more likely to show different binding specificities given their distinct biochemical composition[83,111,113–115]. These distinctions observed for kappa and lambda light chains across multiple B cell subsets and healthy human donors highlight the potentially unique and complementary roles that kappa and lambda light chains may play in the immune repertoire.

Finally we note that strong similarities were observed across human donors: the antibody repertoires recovered from different donors displayed convergence in heavy/light V-gene usage and biochemical metrics (e.g. loop length, charge, hydrophobicity, etc.), and a large fraction of light chains were also shared across donors and within donors. We found that approximately 70 percent of naïve light chains were encoded by a promiscuous nucleotide sequence while nearly 80 percent of naïve light chains were encoded by a promiscuous amino acid sequence, and even around 50 percent of antibodies in antigen-experienced repertoires were encoded by public CDR-L3 amino acid sequences. These data demonstrated that light chain high-throughput sequencing without heavy chain pairing information is inadequate for accurate characterization of the true light chain repertoire, as traditional high-throughput sequencing cannot distinguish between identical light chains that are encoded by multiple B cell clonotypes. We also observed a small number of public heavy chain CDR3 amino acid sequences in both naïve and antigen-experienced repertoires. While naïve public CDR-H3 often exhibited disparate VH-gene and VL-gene usage (though nearly always the same JH genes), the antigen-experienced public CDR-H3 showed a strong enhancement in convergent gene usage (Figure 4.9). The convergent gene usage in public antigen-experienced CDR-H3 suggested that population-wide functional selection may result in similar public antibodies generated across multiple individuals, as suggested by previous reports[78,120,121,131].

4.4 Methods

4.4.1 Ethics statement

Informed consent was obtained from anonymous donors prior to experiments by the Gulf Coast Regional Blood Center (Houston, TX). This study was approved by the University of Texas at Austin Institutional Biosafety Committee (2010-06-0084).

4.4.2 Cell Isolation and VH:VL Pairing

PBMC were isolated from donated human whole blood and non-B cells were depleted via magnetic bead sorting (Miltenyi Biotec, Auburn, CA). B cells were stained with anti-CD20-FITC (clone 2H7, BD Biosciences, Franklin Lakes, NJ, USA), anti-CD3-PerCP (HIT3a, BioLegend, San Diego, CA, USA), anti-CD19-v450 (HIB19, BD), and anti-CD27-APC (M-T271, BD). $CD3^{-}CD19^{+}CD20^{+}CD27^{-}$ naïve B cells were analyzed for VH:VL sequences immediately following FACS sorting. $CD3^{-}CD19^{+}CD20^{+}CD27^{+}$ antigen-experienced B cells (comprised of mostly memory B cells with a small number of peripheral plasmablasts) were incubated four days in the presence of RPMI-1640 supplemented with 10% FBS, 1× GlutaMAX, 1× non-essential amino acids, 1× sodium pyruvate and 1× penicillin/streptomycin (LifeTechnologies) along with 10 µg/mL anti-CD40 antibody (5C3, BioLegend), 1 µg/mL CpG ODN 2006 (Invivogen, San Diego, CA, USA), 100 units/mL IL-4, 100 units/mL IL-10, and 50 ng/mL IL-21 (PeproTech, Rocky Hill, NJ, USA)[41] prior to high-throughput VH:VL sequencing. High-throughput emulsion-based VH:VL sequencing was performed as reported previously[83]. Briefly, cells were isolated into emulsion droplets along with poly(dT) magnetic beads for mRNA capture using a flow-focusing nozzle apparatus. Droplets contained lithium dodecyl sulfate and DTT to lyse cells and inactivate proteins, and mRNA released from lysed cells was captured by the poly(dT) sequences on magnetic beads. The emulsion was broken chemically and beads were collected, washed, and used as template for emulsion overlap extension RT-PCR which linked heavy and light chain transcripts into a single, linked cDNA construct for high-throughput sequencing. All VH:VL pairing analyses used primers targeting the Framework 1 antibody gene regions[82].

4.4.3 Bioinformatic Analysis

Raw Illumina sequences were quality-filtered, mapped to V-, D-, and J- genes and CDR3s extracted using both the International Immunogenetics Information System (IMGT)[95] and NCBI IgBlast software[138] with a CDR3 motif identification algorithm[52]. Most antibody sequences were successfully mapped by both algorithms (96% of all sequenced antibodies), and IMGT gene assignments were given priority over IgBlast assignments. Sequence data were filtered for in-frame V(D)J junctions and productive V_H and $V_{\kappa,\lambda}$ sequences were paired by Illumina read ID and compiled by exact CDR3 nucleotide and V(D)J gene usage match. CDR-H3 nucleotide sequences were extracted and clustered to 96% nt identity with terminal gaps ignored (USEARCH v5.2.32[122]), with a minimum of one nucleotide mismatch permitted during CDR-H3 clustering regardless of sequence length. Resulting VH:VL pairs with ≥2 reads comprised the preliminary list of VH:VL clusters for each data set. For determining germline identity in the FR3 region, all FR3 reads associated with the VH and VL in a given VH:VL pair were clustered by 90% identity using USEARCH[122], and the largest of the resulting clusters were analyzed by alignment of these representative FR3 sequences with IMGT to determine percent homology to known germline genes. Naïve antibody sequences were additionally filtered to include only those sequences with >98% germline identity in the FR3 region, similar to previous reports[92]. Amino acid sequence hydrophobicity was determined by the normalized version of the Kyte-Doolittle hydrophobicity index[136]; antibody isotypes were determined by analyzing constant region sequences.

4.4.4 Statistical analysis

R (version 3.1.1) was used for hierarchical clustering (function "hclust"), the Kolmogorov-Smirnov test (function "ks.test"), and the identification of differentially paired genes (package "limma" version 3.14.4)[133,139]. For hierarchical clustering the fractional frequency of V-gene pairs was multiplied by a scaling factor of 100,000. After discarding gene pairs with zero fraction, the fractions were log2-transformed and normal distributions were generated. Distance between samples was measured by Pearson correlation with complete-linkage as the agglomerative method. For Kolmogorov-Smirnov (K-S) test, raw values such as charge, length, hydrophobicity were used to compare probability distributions across experimental groups. Although the Linear

Models for Microarray Data method (limma) was originally developed to identify differentially expressed genes in microarray data, the algorithm is also applicable to quantitative PCR or RNA-Seq that provides a matrix composed of genes and expression values, and the linear model-based test is stable for experiments with a small number of replicates in that it borrows information across genes. Before running limma, gene pairs with zero usage were removed and quantile normalization was performed to normalize the difference in distribution of values among samples. P-values for multiple comparisons were corrected with the Benjamini-Hochberg procedure. Differentially paired gene cut-offs were established at a fold change of 2 and an adjusted *p*-value of 0.05.

5. CONCLUSIONS AND FUTURE PERSPECTIVES

We developed and validated a new technology for high throughput sequencing of paired antibody heavy and light chains at very high cell throughput. This work was undertaken specifically to address a critical deficiency in currently available high throughput antibody sequencing techniques, namely that the pairing information of heavy and light chains is irreversibly lost during traditional high-throughput sequencing. We first reported a microarray-based technique with capacity for up to 10^5 cells per analysis[82], then translated the same general workflow into emulsion droplets for interrogation of up to 10^7 single B cells per operator in a single day[83]. We determined the VH:VL pairing accuracy of our technique to be >97%[83] and applied our technology for antibody discovery[65,79,82,140], proteomic analysis of vaccine responses[65,69], mechanistic investigation of HIV broadly neutralizing antibody development[79], and to generate novel immunological insight related to the composition and development of antibody repertoires in healthy human donors[83,141].

Importantly, our method for sequencing multiple mRNAs from single cells has a number of applications beyond the sequencing of antibody heavy and light chain pairs. Future efforts in this area will incorporate additional mRNAs of interest, for example pairing of transcription factors implicated in B-cell development such as Blimp-1[39] to determine both B cell maturity and VH:VL sequences in a single experiment without the need for fluorescence-activated cell sorting (FACS). As FACS is an expensive and time-consuming task that requires skilled operators, the ability to omit FACS for separate analysis of cell subsets may enable faster, cheaper, and therefore more effective investigations of human adaptive immune repertoires. We will also analyze paired antibody VH:VL sequences of cells with surface expression of a particular phenotype, for example B-cell receptor affinity to an antigen of interest for extremely rapid and high-throughput antibody discovery.

Our accessible technology for sequencing the paired antibody VH:VL repertoire enables rapid interrogation of the immune response and can be applied to investigate B-cell responses in a variety of clinical and research settings. In particular, the suite of new techniques presented here will greatly enhance high-throughput, high-resolution analysis of human vaccine responses, providing new ways to test vaccine efficacy and thereby inform vaccine design. The high-throughput identification and cloning of paired VH:VL antigen-specific antibodies from responding B cells may also enable rapid generation of novel diagnostic, therapeutic or prophylactic antibodies. As DNA sequencing technologies continue to progress, low-cost high-throughput single-cell antibody sequencing can enable paired antibody repertoire analysis at great depth in large study cohorts and clinical patients and in turn provide unprecedented insights into humoral responses associated with vaccine development, autoimmunity, infectious diseases and other human disease states.

6. PUBLICATION LIST

6.1 PEER-REVIEWED ARTICLES RELATED TO DISSERTATION

Accepted

1. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, Varadarajan N, Giesecke C, Dörner T, Andrews SF, Wilson PC, Hunicke-Smith SP, Willson CG, Ellington AD, Georgiou G, "High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire," *Nature Biotechnology*, 31(2): 166-169, 2013
2. Doria-Rose NA*, Schramm CA*, Gorman J*, Moore PL*, Bhiman JN, DeKosky BJ, Ernandes MJ, Georgiev IS, Kim HJ, Pancera M, Staupe RP, Altae-Tran HR, Bailer RT, Crooks ET, Cupo A, Druz A, Garrett NJ, Hoi KH, Kong R, Louder MK, Longo NS, McKee K, Nonyane M, O'Dell S, Roark RS, Rudicell RS, Schmidt SD, Sheward DJ, Soto C, Wibmer CK, Yang Y, Zhang Z, NISC Comparative Sequencing, Mullikin JC, Binley JM, Sanders RW, Wilson IA, Moore JP, Ward AB, Georgiou G, Williamson C, Abdool Karim SS, Morris L**, Kwong PD**, Shapiro L**, Mascola JR** "Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies," *Nature* 509(7498): 55–62, 2014
3. Lavinder JJ, Wine Y, Giesecke C, Ippolito GC, Horton AP, Lungu OI, Hoi KH, DeKosky BJ, Murrin EM, Wirth MM, Ellington AD, Dörner T, Marcotte EM, Boutz DR, Georgiou G "Identification and characterization of the constituent human serum antibodies elicited by vaccination," *Proceedings of the National Academy of Sciences* 111(6): 2259–2264, 2014
4. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, Georgiou G, "In-depth determination and analysis of the human paired heavy and light chain antibody repertoire," *Nature Medicine,* 21(1): 86-91, 2014

In Preparation

1. DeKosky BJ*, Lungu OI*, Johnson E, Ippolito GC, Kuroda D, Charab W, Hoi KH, Ellington AD, Gray J, Georgiou G, "Insights in the development of the human antibody repertoire via high-throughput VH:VL sequencing and structural modeling of naïve and memory B cell repertoires in healthy individuals"

2. Wang B, Kluwe C, Lungu OI, DeKosky BJ, Villanueva I, Reyes A, Reizigh AB, Davey R, Ellington AD, Georgiou G, "Rapid identification of high-affinity antibodies against Zaire Ebolavirus through repertoire analysis"
3. Lee J, Boutz DR, Vollmers C, DeKosky BJ, Horton AP, Hoi KH, Ippolito GC, Marcotte EM, Quake SR, Georgiou G, "Proteomic identification and characterization of monoclonal antibodies comprising the serological response to seasonal influenza vaccines"

6.2 Peer-reviewed articles prior to UT-Austin

1. DeKosky BJ, Dormer NH, Ingavle GC, Roatch CH, Lomakin J, Detamore MS, Gehrke SH, "Hierarchically designed agarose and poly(ethylene glycol) interpenetrating network hydrogels for cartilage tissue engineering," *Tissue Engineering Part C: Methods*, 16(6): 1533-1542, 2010

   →#12 most recently-cited article in TEC, Dec 2014

6.3 Patents related to dissertation

1. Georgiou G, DeKosky BJ, Ellington AD, Hunicke-Smith SP, "High throughput sequencing of multiple transcripts," June 15, 2012, Serial No. 61/660,370 (pending). PCT June 17, 2013: Serial No. PCT/US2013/046130, U.S. Application Dec 12, 2014: 14/407,849.

7. APPENDICES

Appendix 1 – Chapter 2 Supplementary Information

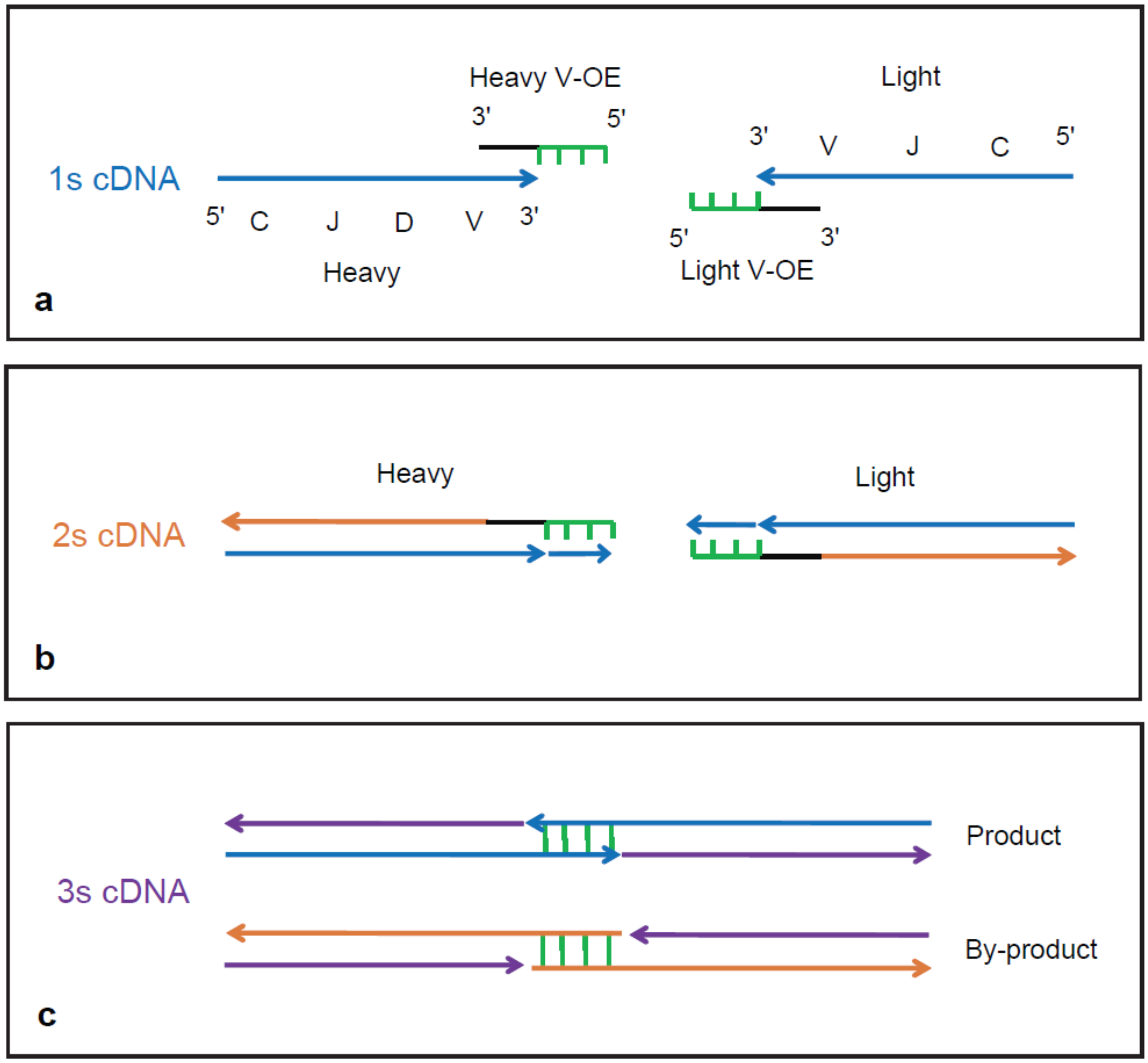


Figure A1.1 An overview of the linkage (overlap extension) RT-PCR process. a) V-region primers (black) with a 5' complementary heavy/light overlap region (green) anneal to first strand cDNA. b) Second strand cDNA is formed by 5' to 3' extension; the overlap region is incorporated into all cDNA. c) After denaturation, heavy and light chains with first strand sense anneal to generate a complete 850 bp product through 5' to 3' extension. The CDR-H3 and CDR-L3 are located near the outside of the final linked construct to allow CDR3 analysis by 2x250 paired-end Illumina sequencing. Linkage RT-PCR primer sequences are given in Table A1.5 (V-region primers denoted "fwd-OE" and constant region primers denoted "rev-OE").

Immunization	n/a	Tetanus Toxoid (TD, MSD)	Influenza (2010-11 Fluvirin)
Cell Type	IgG$^+$ B lymphocytes	Day 7 post-TT boost TT$^+$ plasmablasts	Day 14 memory B cells
Fresh Cells vs. Freeze/Thaw	Fresh	Freeze/Thaw	Freeze/Thaw
Cell:Well Ratio	1:10	1:425	1:39
% cells as single cells	95.1%	99.9%	98.7%
Unique CDR-H3 Recovered	2,716	86	240
Control Cell Spike	IM-9	ARH-77	IM-9
Accuracy Ratio[1]	78:1	650:1	942:1

[1] For known spiked cells, (reads correct VL):(reads top incorrect VL)

Table A1.1 Key statistics from several paired VH:VL repertoires. TD-tetanus toxoid/diphtheria toxoid, MSD-Merck Sharpe & Dohme

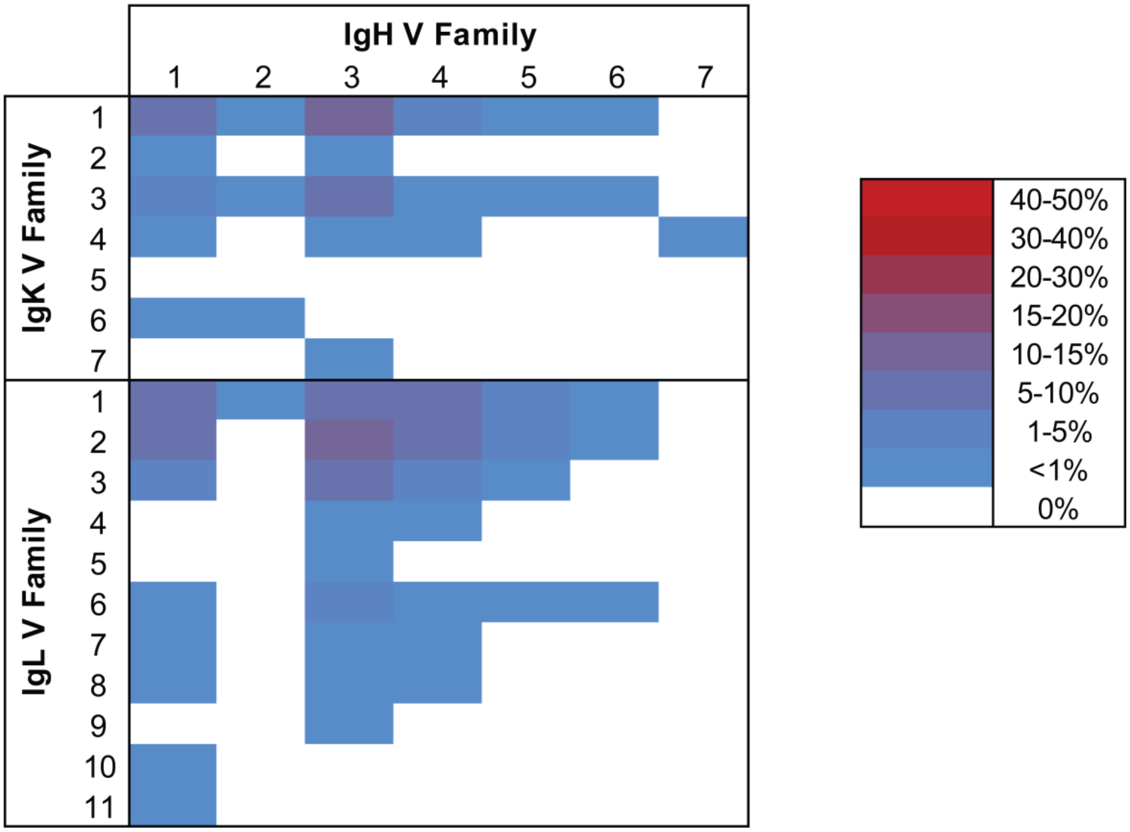


Figure A1.2 A heat map of VH:VL pairings from IgG$^+$ class-switched peripheral B cells isolated from a healthy volunteer (n=2,248). The experiment presented here is a replicate of Fig. 2a using donated blood from a different individual.

Immunization	n/a
Cell Type	IgG^{+} B lymphocytes
Fresh Cells vs. Freeze/Thaw	Fresh
Cell:Well Ratio	1:10
% cells as single cells	95.1%
Unique CDR-H3 Recovered	2,248
Control Cell Spike	IM-9
Accuracy Ratio[1]	125:1

[1] For known spiked cells, (reads correct VL):(reads top incorrect VL)

Table A1.2 Key statistics for the IgG+ VH:VL pairing experiment from a second volunteer (Fig. A1.2).

Seq ID	Isotype	CDR-H3	Paired CDR-L3[1]	Source
2D02	IgM	gcgagaggcggaaatgggcgaccctttgacaac	gcagcatgggatgacagcctgaatggttgggtg	Sanger scRT-PCR
2D02	IgM	gcgagaggcggaaatgggcgaccctttgacaac	gcagcatgggatgacagcctgaatggttgggtg	MiSeq VH:VL
3D05	IgM	gcgagaaggtactttgactac	gnagcatgggatgacagcctgaatgtttggntg	Sanger scRT-PCR
3D05	IgM	gcgagaaggtactttgactac	gcagcatgggatgacagcctgaatgtttggctg	MiSeq VH:VL
1E02	IgG1	gcgcgacatggccctgcgggaaaaagcgcgtatggttttgatatc	cagtcctatgacagcggactgaatggttatgtggtc	Sanger scRT-PCR
1E02	IgG	gcgcgacatggccctgcgggaaaaagcgcgtatggttttgatatc	cagtcctatgacaacagactgaatggttatgtggtg	MiSeq VH:VL
3A01	IgG3	gcgagagtaatagcagctcgcgaccgccggatcactcctaactactaccgccctatggacgtc	caggtgtgggatagtagtagtgaccatcaggtg	Sanger scRT-PCR
3A01	IgG	gcgagagtaatagcagctcgcgaccgccggatcactcctaattactaccgccctatggacgtc	caggtgtgggacagtagtagtgatcatcaggtg	MiSeq VH:VL

[1] The 2D02 and 3D05 CDR-L3 sequences are highly similar but differ by two bases

Table A1.3 Analysis of overlapping heavy chain sequences and paired light chain sequences identified by both single cell RT-PCR and high-throughput VH:VL pairings in a memory B cell population isolated from an individual 14 days post-vaccination with the 2010-2011 trivalent FluVirin influenza vaccine.

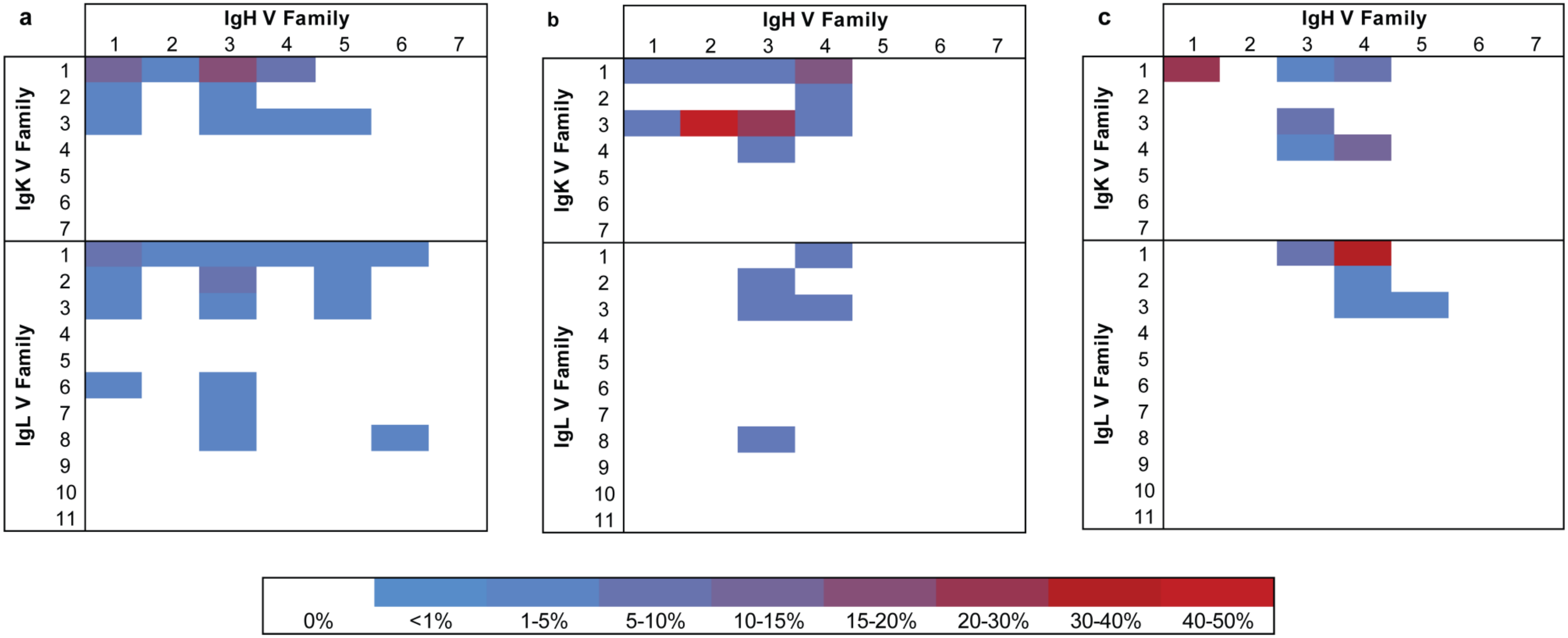


Figure A1.3 As Fig. 2b comprised the lowest sample size in Figure 2.2 (n=86 unique pairs, compared to Fig. 2a, n=2,716, and Fig. 2c, n=240) a simulation was performed to randomly select 86 VH:VL pairs from Fig. 2a and 2c and normalize all panels to 86 unique sequences. (a) healthy donor peripheral IgG$^+$ B cells, (b) day 7 tetanus-toxoid specific plasmablasts, and (c) day 14 post-influenza vaccination memory B cells. The simulation presented here facilitates comparison between panels a, b, and c.

Experiment	*Resting IgG+ repertoire*	*Tetanus toxoid D7 plasmablasts*
Figure	2.2a	2.2b
Cell:Well Ratio	1:10	1:425
Fraction Single Cells[1]	0.951	0.999
Spiked Clone	IM-9	ARH-77
Spike %	4.0	7.5
Estimated # spiked cells	2,830	30
Estimated Spiked Cells as Single Cells[1]	2,691	30
Estimated Spiked Cells as 2-cells-per-well[1]	139	0
Paired Reads in Dataset	287,572	30,238
Correctly Paired Spike VH:VL Reads	14,805	871
Predicted Mispairing Rate[1]	4.9%	0.1%
Spike VH : Top Non-Spike VL Mispairing Rate	1.3%	0.15%
Top Non-Spike VH : Spike VL Mispairing Rate	7.8%	0.31%
Total Recovered VH:VL Pairs From Sample	2,716	86

[1] Calculated from the Poisson distribution

Table A1.4 Statistical analysis of pairing accuracy

Conc. (nM)	Primer ID	Sequence
400	CHrev-AHX89	*CGCAGTAGCGGTAAACGGC*
400	CLrev-BRH06	*GCGGATAACAATTTCACACAGG*
40	hIgG-rev-OE-AHX89	*CGCAGTAGCGGTAAACGGC* AGGGYGCCAGGGGGAAGAC
40	hIgA-rev-OE-AHX89	*CGCAGTAGCGGTAAACGGC* CGGGAAGACCTTGGGGCTGG
40	hIgM-rev-OE-AHX89	*CGCAGTAGCGGTAAACGGC* CACAGGAGACGAGGGGGAAA
40	hIgKC-rev-OE-BRH06	*GCGGATAACAATTTCACACAGG* GATGAAGACAGATGGTGCAG
40	hIgLC-rev-OE-BRH06	*GCGGATAACAATTTCACACAGG* TCCTCAGAGGAGGGYGGGAA
40	hVH1-fwd-OE	TATTCCCATGGCGCGCCCAGGTCCAGCTKGTRCAGTCTGG
40	hVH157-fwd-OE	TATTCCCATGGCGCGCCCAGGTGCAGCTGGTGSARTCTGG
40	hVH2-fwd-OE	TATTCCCATGGCGCGCCCAGRTCACCTTGAAGGAGTCTG
40	hVH3-fwd-OE	TATTCCCATGGCGCGCCGAGGTGCAGCTGKTGGAGWCY
40	hVH4-fwd-OE	TATTCCCATGGCGCGCCCAGGTGCAGCTGCAGGAGTCSG
40	hVH4-DP63-fwd-OE	TATTCCCATGGCGCGCCCAGGTGCAGCTACAGCAGTGGG
40	hVH6-fwd-OE	TATTCCCATGGCGCGCCCAGGTACAGCTGCAGCAGTCA
40	hVH3N-fwd-OE	TATTCCCATGGCGCGCCTCAACACAACGGTTCCCAGTTA
40	hVK1-fwd-OE	GGCGCGCCATGGGAATAGCCGACATCCRGDTGACCCAGTCTCC
40	hVK2-fwd-OE	GGCGCGCCATGGGAATAGCCGATATTGTGMTGACBCAGWCTCC
40	hVK3-fwd-OE	GGCGCGCCATGGGAATAGCCGAAATTGTRWTGACRCAGTCTCC
40	hVK5-fwd-OE	GGCGCGCCATGGGAATAGCCGAAACGACACTCACGCAGTCTC
40	hVL1-fwd-OE	GGCGCGCCATGGGAATAGCCCAGTCTGTSBTGACGCAGCCGCC
40	hVL1459-fwd-OE	GGCGCGCCATGGGAATAGCCCAGCCTGTGCTGACTCARYC
40	hVL15910-fwd-OE	GGCGCGCCATGGGAATAGCCCAGCCWGKGCTGACTCAGCCMCC
40	hVL2-fwd-OE	GGCGCGCCATGGGAATAGCCCAGTCTGYYCTGAYTCAGCCT
40	hVL3-fwd-OE	GGCGCGCCATGGGAATAGCCTCCTATGWGCTGACWCAGCCAA
40	hVL-DPL16-fwd-OE	GGCGCGCCATGGGAATAGCCTCCTCTGAGCTGASTCAGGASCC
40	hVL3-38-fwd-OE	GGCGCGCCATGGGAATAGCCTCCTATGAGCTGAYRCAGCYACC
40	hVL6-fwd-OE	GGCGCGCCATGGGAATAGCCAATTTTATGCTGACTCAGCCCC
40	hVL78-fwd-OE	GGCGCGCCATGGGAATAGCCCAGDCTGTGGTGACYCAGGAGCC

Table A1.5 Overlap Extension (OE) RT-PCR primer mix

Conc. (nM)	**Primer ID**	**Sequence**
400	hIgG-all-rev-OEnested	ATGGGCCCTGSGATGGGCCCTTGGTGGARGC
400	hIgA-all-rev-OEnested	ATGGGCCCTGCTTGGGGCTGGTCGGGGATG
400	hIgM-rev-OEnested	ATGGGCCCTGGGTTGGGGCGGATGCACTCC
400	hIgKC-rev-OEnested	GTGCGGCCGCAGATGGTGCAGCCACAGTTC
400	hIgLC-rev-OEnested	GTGCGGCCGCGAGGGYGGGAACAGAGTGAC

Table A1.6 Nested PCR primers.

Conc. (nM)	Primer ID	Sequence
400	hIgG-all-rev-OEnested	ATGGGCCCTGSGATGGGCCCTTGGTGGARGC
400	hIgA-all-rev-OEnested	ATGGGCCCTGCTTGGGGCTGGTCGGGGATG
400	Linker-VHfwd-BC2	NNNNTGAAGGGGCTAGCTATTCCCATCGCGG
400	hIgKC-rev-OEnested	GTGCGGCCGCAGATGGTGCAGCCACAGTTC
400	hIgLC-rev-OEnested	GTGCGGCCGCGAGGGYGGGAACAGAGTGAC
400	Linker-VLfwd-BC2	NNNNTGAAGGGCGCCGCGATGGGAAT

Table A1.7 VH and VL Separate Amplification Primers

Appendix 2 – Chapter 3 Supplementary Information

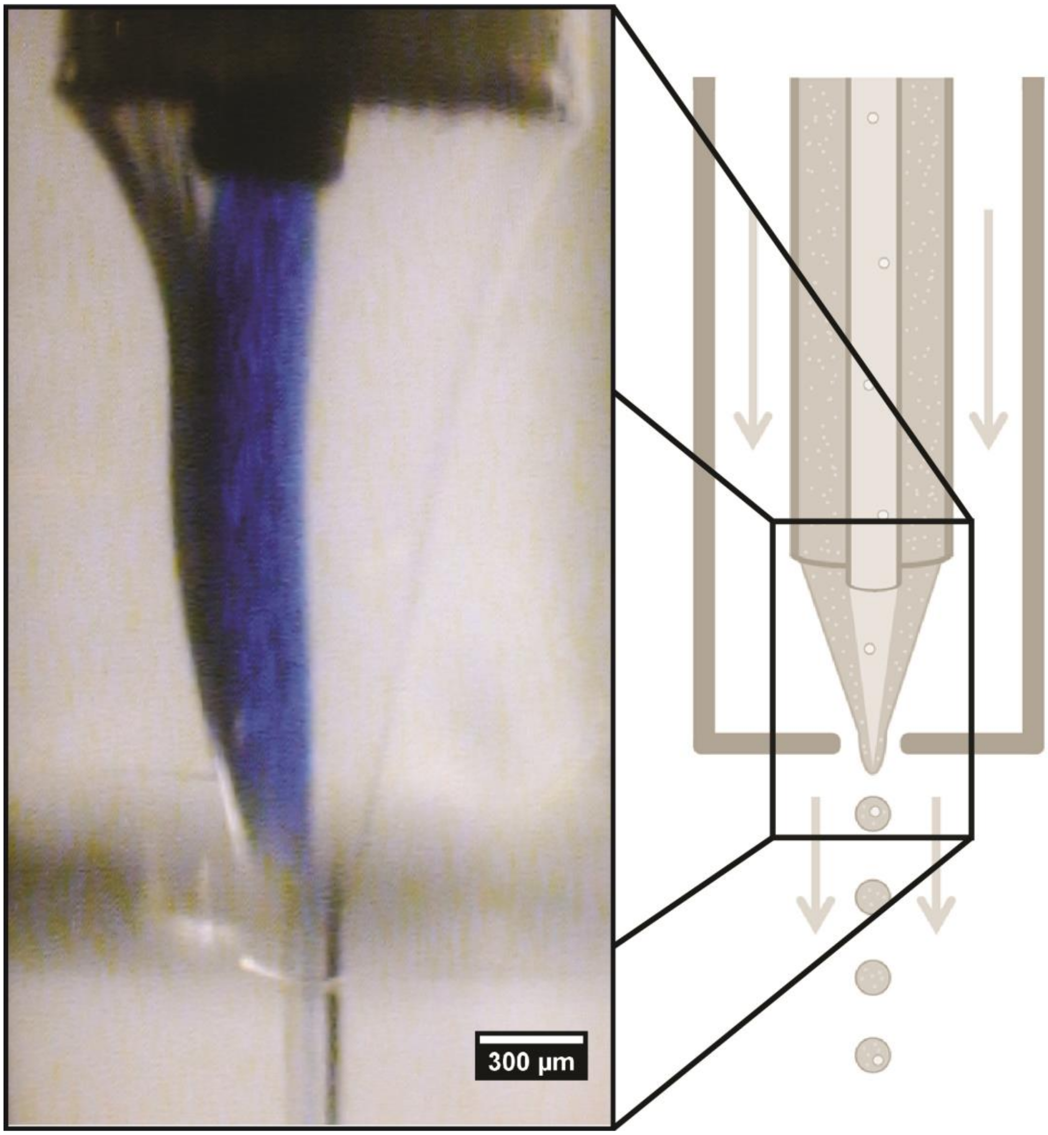


Figure A2.1 A micrograph of the axisymmetric flow-focusing nozzle during emulsion generation (left), placed in context of the diagram from Figure 2.1a (right), where PBS/0.4% Trypan blue exits the inner needle and cell lysis buffer exits the outer needle.

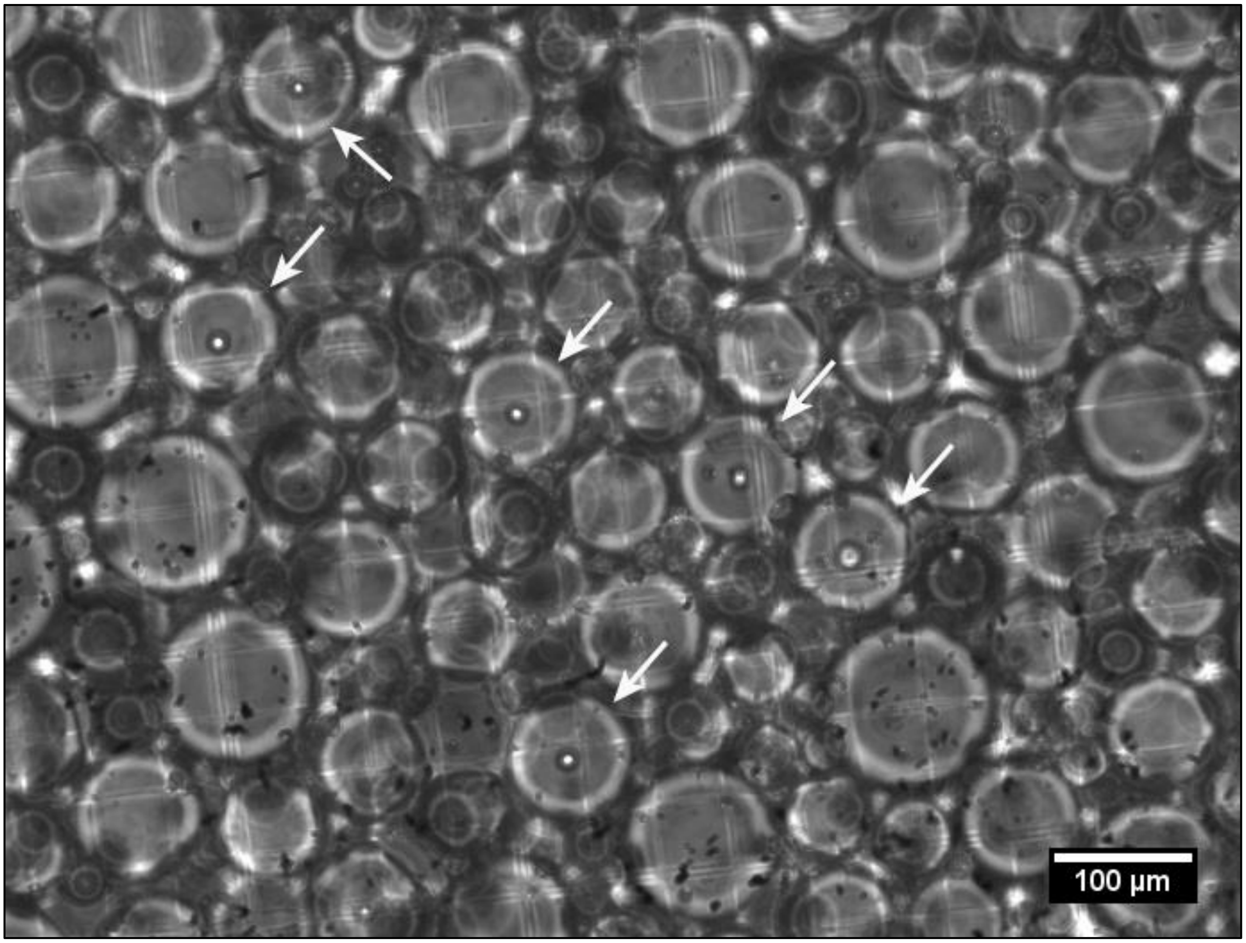


Figure A2.2 MOPC-21 immortalized B cells encapsulated in emulsion droplets. The outer aqueous stream that normally contains cell lysis buffer (Fig. 1a, gray solution) was replaced with 0.4% Trypan blue in PBS to examine cell viability throughout the flow focusing and emulsification process. Emulsified cell viability was approximately 90% and cell viability did not differ substantially from non-emulsified controls.

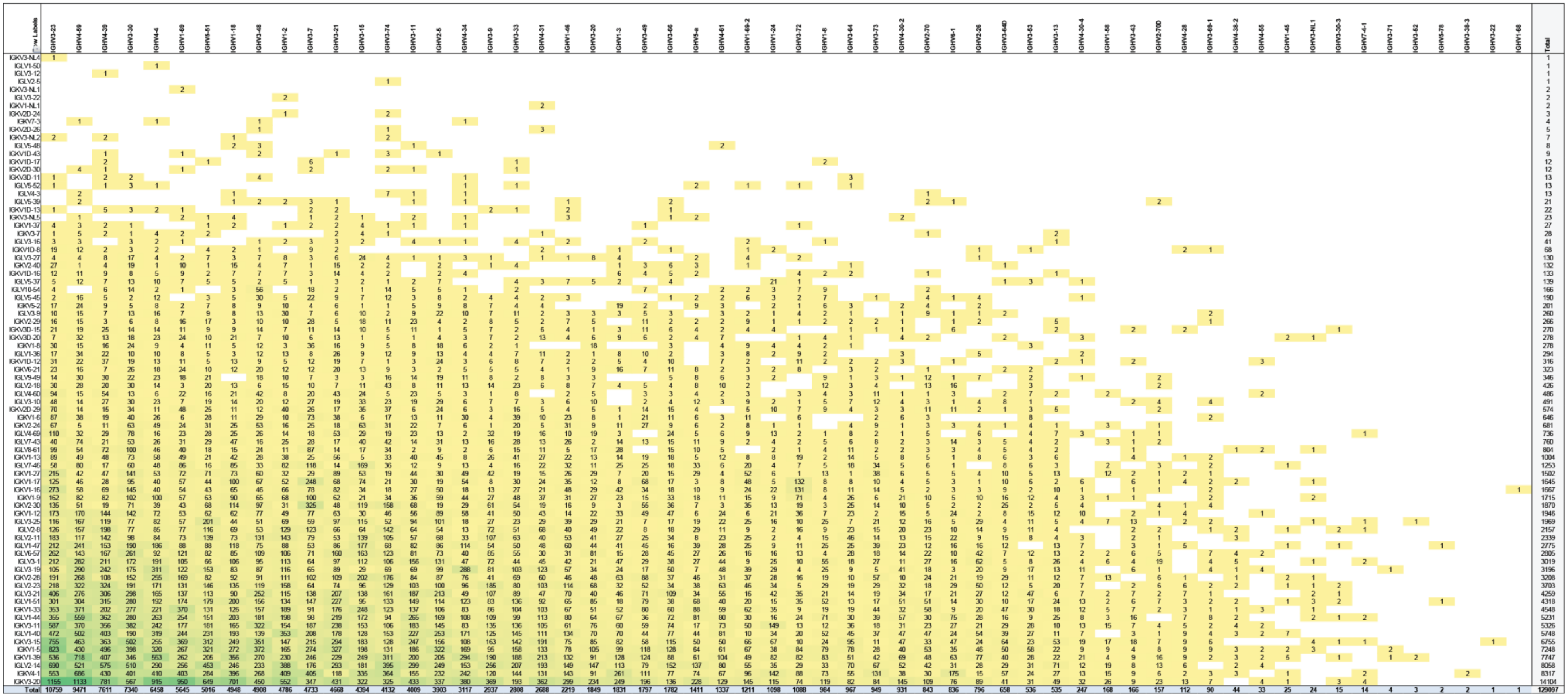


Figure A2.3 Heat map of V-gene usage for 129,097 VH:VL clusters recovered from Donor 1. Sequences were collected using primers targeting the framework 1 region; raw data is available in the online supplement.

	IGHV3-23	IGHV1-2	IGHV4-39	IGHV3-30	IGHV4-59	IGHV3-7	IGHV1-69	IGHV4-31	IGHV4-4	IGHV1-18	IGHV3-33	IGHV3-9	IGHV3-11	IGHV5-51	IGHV3-21	IGHV4-34	IGHV3-15	IGHV3-74	IGHV1-46	IGHV4-61	IGHV3-48	IGHV1-8	IGHV2-5	IGHV3-53	IGHV3-30-3	IGHV1-3	IGHV7-4-1	IGHV3-66	IGHV1-24	IGHV3-49	IGHV2-70	IGHV6-1	IGHV3-73	IGHV4-30-2	IGHV3-72	IGHV2-26	IGHV3-13	IGHV4-30-4	IGHV3-43	IGHV3-20	IGHV1-58	IGHV2-70D	IGHV3-h	IGHV4-28	IGHV4-55	IGHV1-45	IGHV3-64	IGHV4-b	IGHV3-NL1	IGHV3-64D	IGHV5-a	IGHV1-f	IGHV3-43D	IGHV5-78	IGHV3-52	IGHV3-47	Total
IGLV2-5											1																																														1
IGKV3-NL4						1																																																			1
IGKV7-3																			1														1																								2
IGLV1-41																2																																									2
IGKV1-37										2																																															2
IGLV5-48				2																																																					2
IGLV4-3	1																	1																																							2
IGKV3-NL2														4																																											4
IGKV3-NL1	2																		2																																						4
IGKV2D-30														1	3																																										4
IGLV5-52		1										1				1			1											1																											5
IGLV3-12			3										1														1																														5
IGKV1D-13		1					2											1		1							1																														6
IGKV1D-16					2								1																						2		1																				6
IGKV3D-11	3									2																		1																													6
IGKV2-29			3	2									2	1		2																																									10
IGKV3-NL5						4				2										1	1					1												2																			11
IGKV3-7	4		1		1					1			2		2					1																																					12
IGKV1D-8	9		2		1	2		2			1				1								1		2									1																							22
IGLV5-37		1		1	4	2		11			1			2		2		1	1		2																																				28
IGLV3-16			1		1	1	2	2				3	1	1			1	7	2		1								2		1	1	1						1																		29
IGKV5-2	2	5	4	2	4	3		1	5	2	1		1	2	1					5	4					2				7																											51
IGKV2-40	15	4	2			3	9		1		6		10	7	2		7		1		1		3								5																										76
IGLV5-45	10	6			1	8	4			1	5			2			2	2		1	11	13					7		2			1								1	2																79
IGLV4-60	24		7	1		6			3	1	3	12		2	7			1			2	3	1	1	1		4				2		2			1												1									85
IGKV3D-20	21	3	3	1	4	13	4	4		6	2	2	4	8	3	2	2	8	1						2						1						4				2																100
IGKV1D-12	22	8	12	8	4	5	5	7	3	1	3	1			4	2			2		3	1		2	1			2	3	9	1	1				4	1																				115
IGLV3-27	16	7	1	11		2	7		2	9	11	2	11	1	2		2	1	1	2	5			3	2	9				3	2						1				2																115
IGKV1D-17	22	4	3	6	1	14	1			6	1	5	3	6	11	2	13		5		5	1	1	1	2				2						2																						117
IGKV3D-15	18	2	8	7	5	4	2	7	3	7	5	3	5	5	4	7		2	3	3	1		2		1			2		1		2				2			4	3																	118
IGKV6-21	11	4	7	8	7	4	1	10	1	6	1	9	3	5	3		7	2	2	5	2	2	1	12			4	3	1		2		2	1			2	2		1																	131
IGLV9-49	5	5	16	5	11	2	7	6	11			3	10	4	2	3	7	3	5	5	2		3	5		3	1						8			1	2																				135
IGLV1-36	15	6	7	5	8	5	4	4		4	3	12	2	1	4	5	3		6	8	6	2	8	5	2		1	2				2				2	6	1																			139
IGLV3-9	18	7	16	12	10	2	2	3	8	2	10	2	6	4	2	10	8	5	1	2	2					1	1	1				2				3																					140
IGKV1-8	62	7	8	6	7	5	4	6	4	7	2	1	5	1	13	6			3	3		2	1		2			2		2						3	1		2	2			1														168
IGLV2-18	8	17	5	5	9	6	9	8	9	4	9	8	4	9	5	1	1	17	5	4		3		2	4	3	3	2			2	13	1							4						2											182
IGLV8-61	33	10	6	11	5	27	2	1		1	16	11		9		2	1	3	3	2	5	1	3	9	12	2		2			5	2			1											1											186
IGKV2D-29	39	20		5	8	24	10		10		6	1	7		3	13	2	5	4	1	11	2		22	4		1	21		2		1					4																				226
IGLV7-43	32	30	2	5	15	16	6	5	10	9	11	2		3	16	18	9	20	33	5	2	5	14		4	5	6	3	5	2	14	1	2		4	2																					316
IGLV3-10	18	6	10	32	7	16	8	3	5	19	15	2	11	8	14	4	40	27	6		3	13	2	3	5	1	7	4	7		3	2	4	2		2	2		4	5																	320
IGKV1-6	48	11	5	20	25	47	10	4	7	4	9	8	9	14	15	12		4	3	5	14	1	3	7	5	6	2	6	3	1	7	4			14		1			1		3	1														339
IGKV1-17	59	15	2	23	13	25	11	9		7	12	12	9	2	12	19	9	9	19	3	9	6	4	7	4	8	1		5	17	1	1	4	1	27	2	1																				368
IGLV4-69	50	27	11	15	22	9	9	11	12	29	25	18	15	4	15	8	10	2		2	11	5		21	6	2	6	16	2	1		1	3	5	1	2	2			2	1																381
IGKV2-24	41	10	4	13	14	29	11	9	7	7	24	5	16	31	8	17	9	9	10	5	33	6		7		2	8	12		13	4	6	8	1	2	3																					384
IGLV10-54	26	2	24	20	20	45	5	6	15	7	50		1	5	24	9	3	22	14	3	10	27	2	5	5	3		10		1	2	5			4		5				2		6														388
IGLV7-46	34	35	12	19	19	56	29	4	19	13	8	3	5	13	6	10	39	25	17	4	9	6	5	6	1	2	9	6		13		1	10			2		1																			441
IGKV1-13	59	11	22	48	19	18	25	14	7	12	17	18	31	10	15	13	4	6	20	10	8	7	6	7	13		8	5	5	1	3		1	2	6	4	7		5				2	2													471
IGLV2-11	51	29	37	11	15	18	16	20	21	16	11	16	6	19	12	8	7	22	11	7	4	12	3	6	4	11	9	2		17	11	17	1	2		7		2	1		9	3				1											475
IGKV2-30	61	9	6	29	7	105	33	4	1	23	7	37	29	8	12	2	33	12	6	10	17	3	4	11	2		2	5	8	10		4	2		8				4	2	2			1													519
IGKV1-27	112	32	10	36	10	25	41	9	16	21	25	21	25	18	12	6	18	4	15		4	12	7	15	8	4	5	3	9	6			2	1	2		5		3	4		2															548
IGKV1-9	83	29	18	42	23	57	15	12	27	12	16	32	16	14	28	15	7	3	15	14	11	2	6	6	10	3	2	5	3	10	6	2	4	1	32	1	2	6	7	1			1														599
IGLV6-57	60	37	44	44	36	13	21	15	16	21	12	36	31	25	33	10	31	26	7	15	2	12	1	5	13	4	6	12	9		8	1	1			2		5	16	3			2	2				3	1				1				642
IGKV1-16	119	39	33	35	8	75	28	13	13	17	13	7	12	15	20	19	25	25	20	7	18	8	9	3	16	8	2	1	16	23		1	2	4	49		4	2					2							2		1					714
IGKV1-12	149	45	51	41	44	30	28	43	43	30	34	17	34	31	48	16	18	18	9	39	22	9	21	8	5	9	4	9	3	8	6	3	4	8		8	4	4	1				3														907
IGLV2-8	103	59	65	24	34	93	25	29	26	21	39	26	24	47	30	11	44	45	51	14	24	5	13	14	5	6	8	16	5	10	8	7	2	14	1	1		3		4	2						2		1								961
IGLV3-25	83	77	53	55	41	12	29	26	30	42	57	25	38	62	23	14	85	32	26	21	6	13	14	4	13	18	24	8	1	1	9	1	11	7	2	1	5	5	6	5	2	3			2												992
IGLV3-19	56	45	118	40	103	24	54	87	60	33	61	37	36	52	29	128	7	11	25	25	9	49	10	5	4	12	29	6	16	1	13	4	4	8	2	1	3	3	7	5				8	1												1231
IGLV1-47	128	90	63	70	56	45	24	53	59	53	38	42	34	35	51	33	37	43	54	34	20	27	22	10	15	14	15	6	6	13	14	26	76	5	13	8	4	12	2	1	2	2		1			2						2				1360
IGLV2-23	88	105	103	83	57	29	57	79	31	66	68	74	30	54	30	38	43	59	34	33	38	21	21	13	31	10	24	11	24	12	18	7	12	16	1	5	2	13	1	6		2		4											2		1455
IGKV1-33	168	99	61	100	59	44	119	87	39	50	61	56	49	48	66	22	38	14	39	50	23	13	6	11	45	36	10	10	22	11	30	2	8	13	5	6	2	5	2	7	5	3	1				1									1	1547
IGLV3-1	145	86	98	77	94	35	54	48	45	65	42	88	51	61	52	20	39	57	28	54	9	61	38	31	21	10	25	22	16	11	14	9	7	11	2	40	11	3	5	10	3		2	2	3		1										1606
IGKV2-28	153	104	64	66	78	92	88	46	58	37	50	49	54	33	37	61	83	97	43	42	31	28	19	24	12	25	40	20	6	33	8	15	16	13	5	15	4	6	3	6	1					1											1666
IGLV1-51	216	55	90	117	91	65	59	67	64	70	70	65	116	71	59	87	40	39	48	42	37	31	20	24	26	4	38	9	7	27	26	14	11	16	5	16	13	3	7	10	1	5	4		2				4								1891
IGLV3-21	243	88	103	126	102	62	66	50	50	58	45	87	90	107	71	22	51	24	29	50	84	22	45	36	34	19	34	28	36	10	7	10	4	3	1	10	6	11	3	8		2	4	1						3							1945
IGLV1-40	245	83	122	86	129	36	87	101	84	63	66	82	88	64	80	58	43	20	67	45	43	27	28	36	28	30	30	22	7	9	28	4	5	25	2	5	4	7		3	4			1		1				1							1999
IGKV3-11	392	61	103	79	82	73	86	86	81	80	99	85	66	52	49	42	51	35	22	38	74	28	57	54	31	23	23	32	24	25	8	17	3	11	1	12	4	13	5	4	2			1	2			2									2118
IGKV3-15	354	83	113	148	92	70	138	74	74	57	80	59	78	77	65	87	18	25	40	33	52	17	31	19	33	33	11	44	37	10	18	6		14	2	10	16	5	5		12	3	2	2	2	5											2154
IGLV1-44	203	149	145	109	156	82	76	64	72	70	48	85	83	63	92	85	54	41	56	48	35	86	32	17	28	14	43	8	10	10	9	37	19	28	9	14	5	10	11	9	9	2	1			6								1			2234
IGKV4-1	268	168	123	130	163	143	90	66	65	109	73	52	48	100	45	153	136	123	71	75	50	50	45	39	27	75	64	39	22	31	8	55	50	15	30	1	8	5	6	4	5			1	2												2833
IGLV2-14	384	247	140	193	84	145	94	115	106	79	93	86	119	115	86	52	57	133	128	64	70	32	51	40	88	36	42	29	13	19	20	21	29	14	5	8	12	26	16	10	7	4			1	3	1		2					2			3121
IGKV1-5	618	184	227	233	183	194	247	89	151	165	194	111	136	137	217	96	104	84	96	94	93	45	111	66	54	61	16	65	32	40	30	27	9	20	26	17	23	12	10	4	19	2	5	1	4	2					2						4356
IGKV3-20	525	291	261	183	253	204	396	179	206	199	96	161	117	143	130	214	67	100	139	126	108	65	50	66	56	96	32	32	73	12	51	43	10	26	21	29	10	19	11	8	21	3	2	3	2			2			2	2					4845
IGKV1-39	478	299	270	223	361	186	176	283	212	158	128	184	166	98	125	220	140	120	140	178	104	80	78	84	65	80	36	51	68	61	74	35	33	52	15	49	29	30	23	5	5	6	3	8	4		3										5226
Total	6212	2858	2728	2673	2608	2361	2336	1882	1792	1786	1784	1764	1751	1714	1699	1689	1455	1395	1390	1239	1151	864	802	772	752	691	645	595	510	494	479	414	372	340	302	299	216	216	171	138	120	45	42	38	25	22	10	8	8	6	4	3	3	3	2	1	53679

Figure A2.4 Heat map of V-gene usage for 53,679 VH:VL clusters recovered from Donor 2. Sequences were collected using primers targeting the framework 1 region; raw data is available in the online supplement.

	IGHV4-59	IGHV3-23	IGHV4-39	IGHV3-30	IGHV3-48	IGHV5-51	IGHV1-18	IGHV1-69	IGHV3-7	IGHV4-34	IGHV3-11	IGHV3-21	IGHV1-46	IGHV1-2	IGHV3-9	IGHV3-74	IGHV3-15	IGHV4-31	IGHV4-4	IGHV1-3	IGHV3-53	IGHV1-8	IGHV3-33	IGHV3-49	IGHV5-a	IGHV3-64	IGHV3-66	IGHV3-73	IGHV3-20	IGHV4-61	IGHV3-72	IGHV3-43	IGHV3-13	IGHV1-24	IGHV4-30-2	IGHV1-58	IGHV4-30-4	IGHV3-h	IGHV3-30-3	IGHV4-28	IGHV4-55	IGHV7-4-1	IGHV6-1	IGHV4-b	IGHV1-45	IGHV3-52	Total
IGKV3-NL4																				1																											1
IGKV2-29				1																																											1
IGLV5-48									1																																						1
IGLV2-5					1																																										1
IGKV2-18																							1																								1
IGLV5-39																						1																									1
IGKV3-NL5		2																																													2
IGKV3-NL1	2																																														2
IGLV8-61									2																																						2
IGKV3D-11			1											1																																	2
IGLV3-16	1													1															1																		3
IGLV3-9					1											1					1																										3
IGLV4-60																							4																								4
IGKV1-13																													1			4															5
IGKV1D-17				2											1		2			1																											6
IGLV5-37								1				2		2			1																														6
IGKV3-7	3						1																		2																						6
IGLV3-12			2						1										1								2		1																		7
IGKV1D-16	2	4	1						2									1																													10
IGLV9-49			1	3							4					1		1							1																						11
IGKV1D-8			2		2				2	3	2									1											1																13
IGLV5-45		2	1		4				1	1		1		1		1	1					2											2														17
IGLV3-27		2		1	6		2	1				2	1							1						1	1																				18
IGKV1-8	1	2	1	5	4	2			1	2	1	4																	1																		24
IGKV2-40	2	2		1		3			2	1	7		2		3		1																			2											26
IGLV2-18	5	1		1	1	5	4		3		2	2	2			4					1	2	1			2	2			1															1		40
IGLV3-10	3	5		3	2	2	1	2	3	1	2	2	2		2		3		1		2		3				1	4	1					1								1					47
IGLV1-36	6	3	5	1	4	1	3	1	1			5	1		2	4	8	4			3	1	4		2						1					1											61
IGKV3D-15	3	7	1	3	2	2	2	3	4	3	8	1	5		3		1	3	4		3		2	2		2						1			1												66
IGKV2D-29	2	5	6	1	4	4	2	1	12	4	2	2	1	2		3	3	3	1	6				2						1	2		1														70
IGKV2-24	4	7	3	2	10	2	1	1	9	3	2	5	4	3		4	8		1		1	1	2	2			2						1														78
IGKV1D-12	13	9	7	5	4	2	1	2	2	2	3		4	2	4	5				1			5	1		4			2	4								1									83
IGKV3D-20	3	8	5	3	2	4	7	5	7	4	4	2	7	5	1	1	1		2	3	3	1							1	3			1						1		2						86
IGKV1-6	4	11	4	6	2	4	3	4	12	5	4	5	3	1	5	1		1		2		5			3				1		1			1													88
IGLV6-57	12	1	7	7	2	4	4	3	2	1	13	4		8		3	2				2				3	3	3	4	2																		90
IGLV7-43	8	8	2	10	1	3	4		2	5		2	6	4		3		2	2	3	1		1	7		4	6	1				2				4	2										93
IGKV1-16	4	19	1	17	2	4	3	5	10	3	2	4	8	4	2	1	3		2	1	1	1	1	4	3	2		2		1	8			1													119
IGLV3-25	6	5	9	2	1	17	7	3	4	1	3	5	7	4	7	9	3	1	4	5	3	3	2	1	2			3	5						3	2											127
IGLV7-46	2	8	5	9	8	9	8	8	4	8	1	5	11	5	2	4	14		2	2	1		3	3	1		3	4	1																		131
IGKV1-9	12	12	9	3	6	11	6	3	8	8	6	6	4	1	5	3		2		3					1		3				22		2														136
IGLV3-1	9	8	4	8	6	12	4	3	3	8	14	4	4	5	7	4	10	1		3	4	10	1		2	2					2		5	3	1		1										148
IGKV1-NL1	9	15	9	9	8	4	13	8	6	3	3	11	1	6	10	5	2	1	2	4		1	1		6				1	1	1			6					2					2			150
IGKV1-17	4	23	2	10	11	2	6	9	10	8	6	10	1	3		1	3		1	3		3	3	13			1	3	1		11		2			1											151
IGKV1-27	8	25	9	13	8	4	1	5	7	5	6	7	4	2	6	8	2	4		10	2	1	1		5	2	2				5	2						1									155
IGKV2-30	2	11	2	7	17	8	17	3	42	3	6		3	1	4	4	5	1		2	1			1		6		3	1		2				2	1											155
IGLV4-69	8	25	12	11	4	3	16	6	16	4	6	14	3	3	5	4	2	7	3		6		3			1						2	4			2		2	1								173
IGKV1-12	38	21	21	10	12	8	6	7	8	9	11	11	5	5	2	2	1	1	4	6	6	4	5	3	2	5	5		1	4	2				1		1										227
IGLV2-11	32	9	15	8	8	25	5	9	10	12	4	1	14	9	13	5	8	6	17	2	5	5	3	3	2	2	2	1	4				2			3			1	1							246
IGLV1-47	27	14	22	16	13	13	13	10	7	15	6	9	19	13	7	3	15	8	3	3	4	3	7	4	3	3	2	5		4	1	2	2		2	2			2								282
IGLV1-51	41	16	23	14	14	21	11	6	11	9	6	9	12	14	15	25	5	10	7	2	1	4	8	4	4		1	8	4	3	5	6	2		1			2	1		1						326
IGLV2-8	20	24	53	13	8	23	6	20	19	4	19	6	12	19	18	13	4	10	3	1	4	2	11	2	1		2	1		4	1				2				1								326
IGKV1-33	33	33	16	22	23	24	18	28	11	8	19	19	16	7	20	8	11	21	6	2	8	8	16	1	5	2		2	1	4	1	4	5	4		1	1	1	2								411
IGLV1-40	51	28	43	31	13	26	23	26	7	20	24	17	16	30	20	7	8	26	7	9	9		1	1	5	2	6	13	5	6	2	1			2	2		1			1						489
IGKV2-28	26	25	24	20	21	18	30	20	12	18	24	17	13	13	23	35	40	11	6	6	7	7	1	14	4	7	4	11	4	6	5	6	4	2	4	3	1						1				493
IGLV2-23	80	22	53	36	23	16	8	12	26	23	11	13	49	14	17	36	7	8	7	10	9	5	4	10	11	3	4	8	9	3		3	3		7	8	2						1				561
IGLV1-44	61	43	51	31	36	39	23	24	15	19	20	27	34	43	9	10	8	12	2	2	2	10	3	6	9	6	3	7	2	6	3	2	5		2	2			2			1					580
IGLV2-14	50	87	64	54	19	51	25	14	20	24	33	28	26	26	24	46	13	15	21	9	4	17	7	6	5	8	2	3	11	8		12	15	1				2		1			1				752
IGLV3-21	69	72	57	41	60	22	28	33	18	6	54	50	24	13	38	20	25	9	12	6	25	9	8	5	6	3	13	4	4	5	4	7	2	2	2			2	2		2	1					763
IGKV3-11	68	91	95	40	48	30	45	31	27	41	41	27	16	27	27	9	18	20	14	8	27	6	18	13	2	13	9	2	3	1	2	9	2	9	3	1		2									845
IGKV1-5	66	117	37	53	39	37	40	42	50	45	29	55	19	22	20	18	17	10	14	22	20	20	12	9	10	4	13	2	3	7	15	2	4	8	5		2	1		4			1				894
IGKV3-15	64	115	71	57	82	66	57	76	34	34	49	52	18	9	32	36	10	6	8	28	17	7	23	18	4	15	4		10	4		8	11	6	5		3		1	1	1						1042
IGKV4-1	102	77	56	48	27	37	53	57	68	45	17	15	42	57	22	28	55	30	24	44	14	16	8	15	16	16	4	17	8	7	3	6	2	6	8	4		2	2	1		1				1	1061
IGKV1-39	208	93	95	89	72	59	98	66	86	103	62	36	55	68	75	53	51	52	52	26	37	36	17	15	22	13	29	14	12	33	2	16	5	10	7	10	3	3	3		2						1788
IGKV3-20	202	146	110	58	79	72	90	135	57	96	57	62	65	74	56	58	38	51	62	36	19	26	22	18	17	23	16	12	23	6	10	11	8	13	6	10	8	2		6	4	1		1			1866
Total	1376	1263	1017	785	720	699	695	693	665	617	593	559	539	527	507	486	409	338	295	274	253	217	212	183	159	154	145	134	124	122	112	106	90	73	64	59	24	22	21	14	13	5	4	3	1	1	15372

Figure A2.5 Heat map of V-gene usage for 15,372 VH:VL clusters recovered from Donor 3. Sequences were collected using primers targeting the leader peptide region; raw data is available in the online supplement.

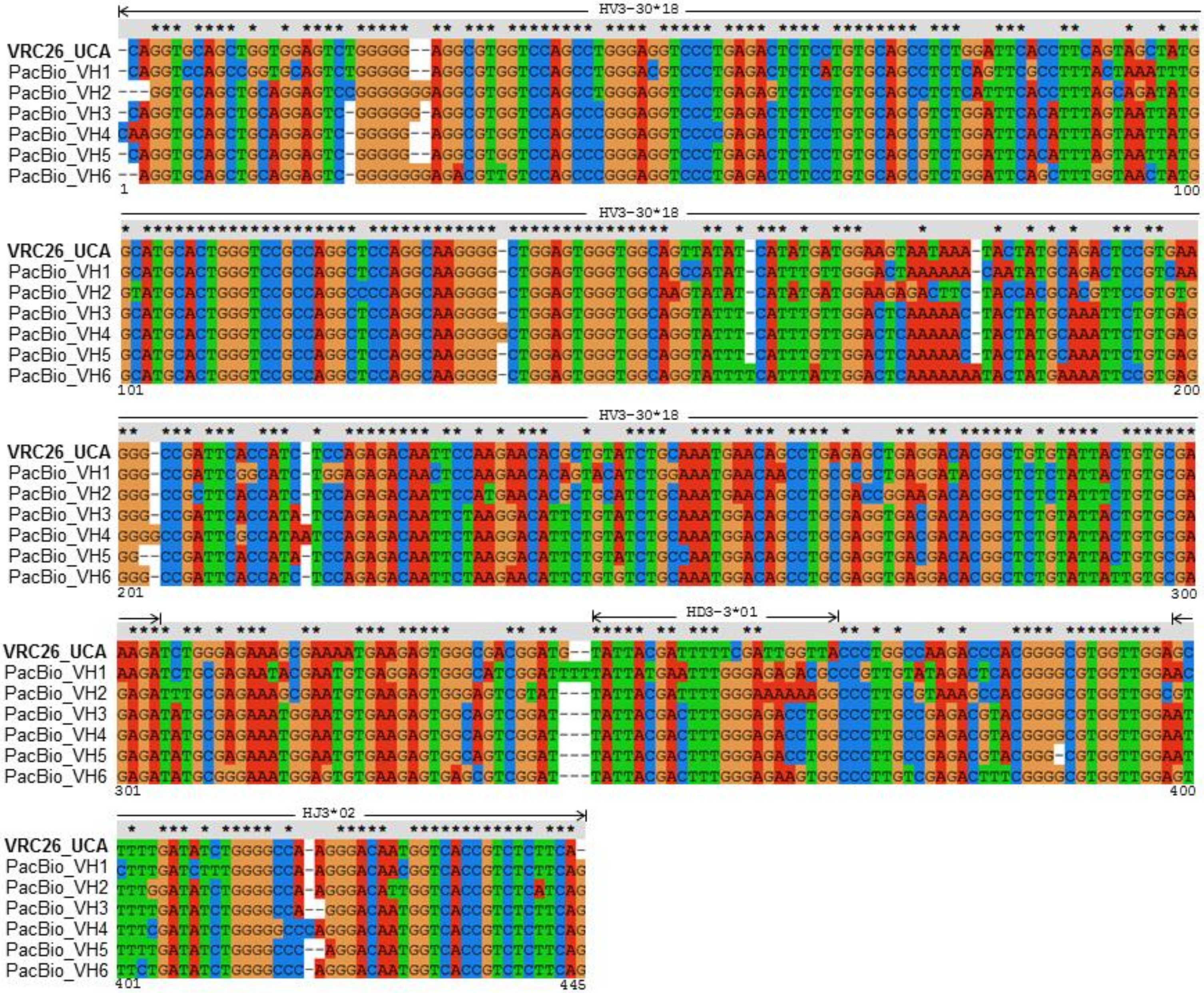


Figure A2.6 VH alignment of the six VRC26 HIV broadly neutralizing antibody variants recovered by PacBio sequencing of complete ~850bp VH:VL amplicons. Sequences were recovered from CD27$^+$ peripheral B cells of the CAP256 donor and aligned to the VRC26 VH unmutated common ancestor (UCA, Doria-Rose et al., *Nature* 2014). Corresponding light chain variants are shown in Figure A2.7.

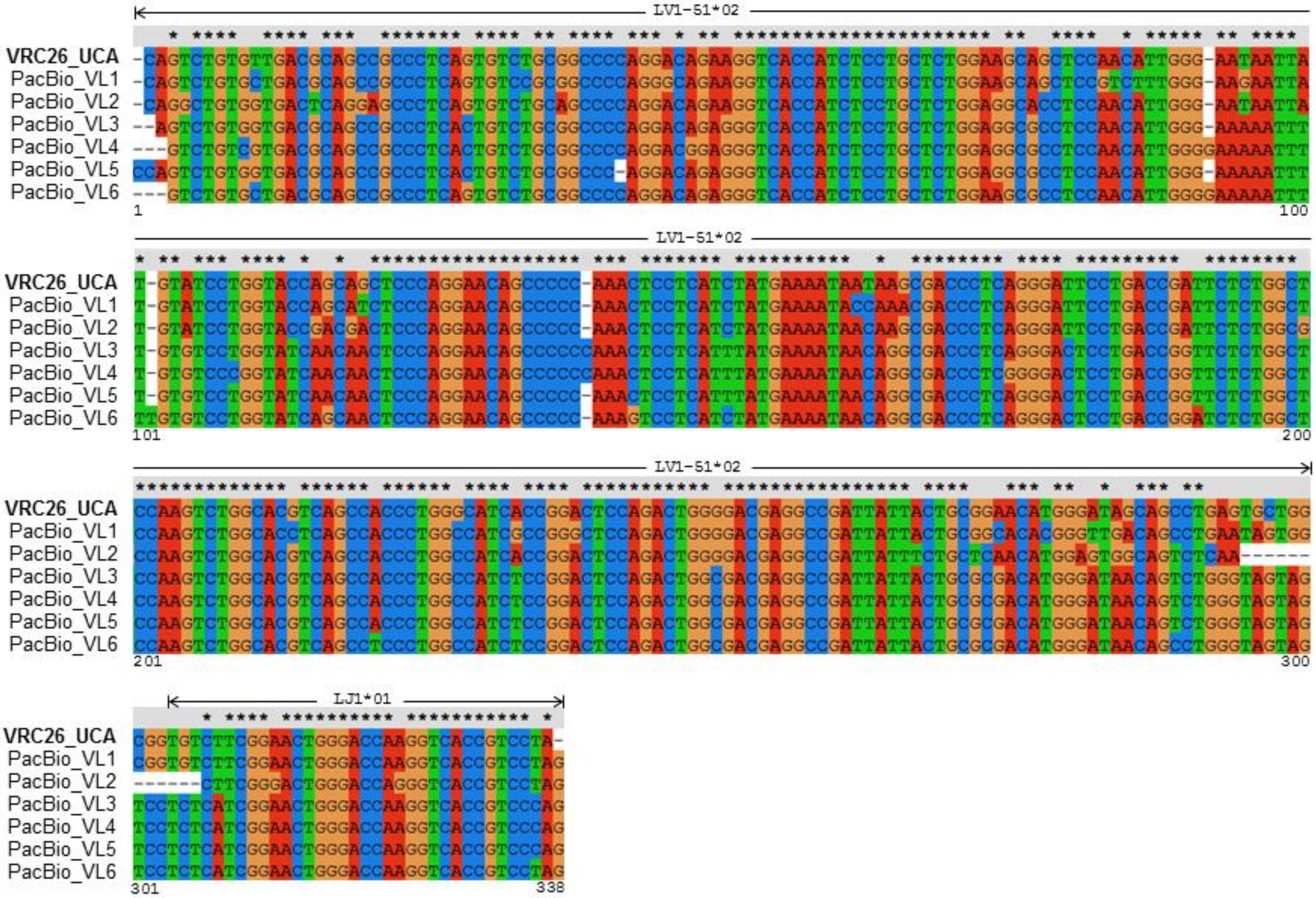


Figure A2.7 VL alignment of the six VRC26 HIV broadly neutralizing antibody variants recovered by PacBio sequencing of complete ~850bp VH:VL amplicons. Sequences were recovered from CD27$^+$ peripheral B cells of the CAP256 donor and aligned to the VRC26 VL unmutated common ancestor (UCA, Doria-Rose et al., *Nature* 2014). Corresponding heavy chain variants are shown in Figure A2.6.

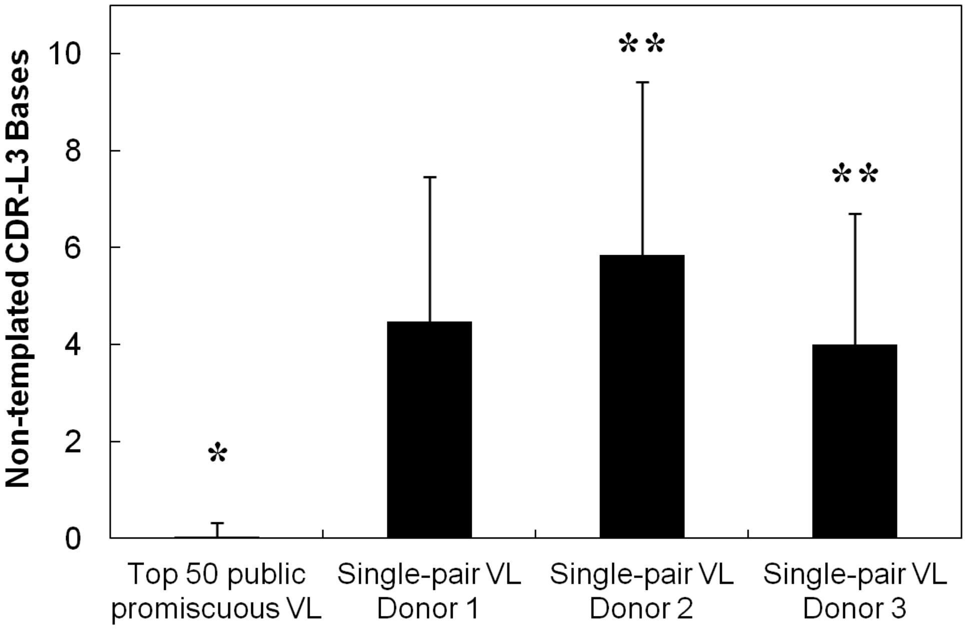


Figure A2.8 Comparison of the number of non-templated bases (sum of somatic mutations and non-templated insertions) in the top 50 public, promiscuous VL nucleotide junctions shared by Donors 1, 2, and 3 to 50 randomly selected VL junctions paired with only a single heavy chain in the Donor 1, Donor 2, or Donor 3 repertoires (mean±s.d.). Statistical significance noted where $p < 0.05$ (* $p < 10^{-10}$ compared to all other groups, ** $p = 0.0043$).

		Heavy Chain											
		1H	**2H**	**3H**	**4H**	**5H**	**6H**	**7H**	**8H**	**9H**	**10H**	**11H**	**Total**
Light Chain	**1L**	**1,842**	4	20	13	18	16	39	20	49	6	4	**2,031**
	2L	0	**4,916**	34	31	59	41	102	127	146	28	8	**5,492**
	3L	0	2	**6,251**	9	38	25	116	60	118	13	2	**6,634**
	4L	21	27	75	**14,592**	81	158	348	189	397	75	51	**16,014**
	5L	5	15	97	41	**16,204**	99	192	231	277	86	19	**17,266**
	6L	2	12	92	37	64	**16,427**	358	180	404	62	23	**17,661**
	7L	9	13	218	72	112	180	**21,315**	203	1,320	78	45	**23,565**
	8L	4	39	85	71	242	145	365	**32,393**	506	79	72	**34,001**
	9L	4	29	182	105	116	186	1,335	323	**35,391**	109	46	**37,826**
	10L	12	24	944	189	1,597	1,080	3,519	1,898	4,291	**8,535**	98	**22,187**
	11L	32	66	1,153	272	1,258	1,655	6,405	6,567	6,185	555	**14,126**	**38,274**
	Total	**1,931**	**5,147**	**9,151**	**15,432**	**19,789**	**20,012**	**34,094**	**42,191**	**49,084**	**9,626**	**14,494**	**220,951**

Table A2.1 VH:VL pairing analysis of a mixture of HEK293 cells transfected with 11 different known antibodies. The maximum read count for each row and column is highlighted; 11/11 antibodies were identified and paired correctly in this control experiment. Read count variation was expected due to varying transfection & expression efficiency for the 22 distinct heavy and light chain plasmids, and antibody clones #10 and 11 exhibited notable VH-VL imbalance by total read counts. The signal:topVLnoise ratio (the relevant parameter for native pair assignment, see Table A2.2) averaged 35:1 overall and 87:1 if noise from light chains 10 and 11 (which showed VH-VL imbalance, see total VH and VL reads) was excluded.

Estimated input human B cells	20,000
Estimated ARH-77 spiked cells	260
VH:VL Reads after CDR3 clustering	403,897
Recovered CDR-H3:CDR-L3 Clusters	1,751
Correct ARH-77 VH:VL Reads (Signal)	2,604
ARH-77 Top Incorrect VL Reads (topVLnoise)	27
ARH-77 2nd-Ranked Incorrect VL Reads	19
ARH-77 3rd-Ranked Incorrect VL Reads	16
ARH-77 Signal:topVLnoise Ratio[*]	96.4

[*] The key metric for VH:VL pair assignment (see main text)

Table A2.2 Accuracy statistics for human VH:VL paired analysis with an ARH-77 immortalized cell line control spike.

Sample	*FACS Count* **Fresh Bmems**	*Hemocytometer Count* **After 4d Activation**
Donor 1	1.8 million	1.6 million viable
Donor 2	1.1 million	1.3 million viable
Donor 3	347k	300k viable
ARH-77 spike experiment	87k	20k viable

Table A2.3 Memory B cell counts before and after *in vitro* activation. Values must be considered rough estimates due to varying contributions of hemocytometer sampling, centrifugation/recovery cell loss, and cell death, stasis, and expansion over four days *in vitro*.

Conc (nM)	Primer ID	Primer Sequence
40	VH1_LP	tattcccatcgcggcgcACAGGTGCCCACTCCCAGGTGCAG
40	VH3_LP	tattcccatcgcggcgcAAGGTGTCCAGTGTGARGTGCAG
40	VH4/6_LP	tattcccatcgcggcgcCCCAGATGGGTCCTGTCCCAGGTGCAG
40	VH5_LP	tattcccatcgcggcgcCAAGGAGTCTGTTCCGAGGTGCAG
40	hVλ1for_LP	gcgccgcgatgggaataNNNNNNNNNNNNNNNNTCTGCTCGAGTTCGGTCAGGTCCTGGGCCCAGTCTGTGCTG
40	hVλ2for_LP	gcgccgcgatgggaataNNNNNNNNNNNNNNNNTCTGCTCGAGTTCGGTCAGGTCCTGGGCCCAGTCTGCCCTG
40	hVλ3for-2_LP	gcgccgcgatgggaataNNNNNNNNNNNNNNNNTCTGCTCGAGTTCGGTCAYWCTGCACAGGCTCTGTGACCTCCTAT
40	hVλ4/5for_LP	gcgccgcgatgggaataNNNNNNNNNNNNNNNNTCTGCTCGAGTTCGGTCAGGTCTCTCTCSCAGCYTGTGCTG
40	hVλ6for_LP	gcgccgcgatgggaataNNNNNNNNNNNNNNNNTCTGCTCGAGTTCGGTCAGTTCTTGGGCCAATTTTATGCTG
40	hVλ7for_LP	gcgccgcgatgggaataNNNNNNNNNNNNNNNNTCTGCTCGAGTTCGGTCAGGTCCAATTCYCAGGCTGTGGTG
40	hVλ8for_LP	gcgccgcgatgggaataNNNNNNNNNNNNNNNNTCTGCTCGAGTTCGGTCAGAGTGGATTCTCAGACTGTGGTG
40	hVκ1/2for_LP	gcgccgcgatgggaataNNNNNNNNNNNNNNNNTCTGCTCGAGTTCGGTCAATGAGGSTCCCYGCTCAGCTGCTGG
40	hVκ3for_LP	gcgccgcgatgggaataNNNNNNNNNNNNNNNNTCTGCTCGAGTTCGGTCACTCTTCCTCCTGCTACTCTGGCTCCCAG
40	hVκ4for_LP	gcgccgcgatgggaataNNNNNNNNNNNNNNNNTCTGCTCGAGTTCGGTCAATTTCTCTGTTGCTCTGGATCTCTG

Table A2.4 Leader peptide overlap extension primers.

Appendix 3 – Chapter 4 Supplementary Information

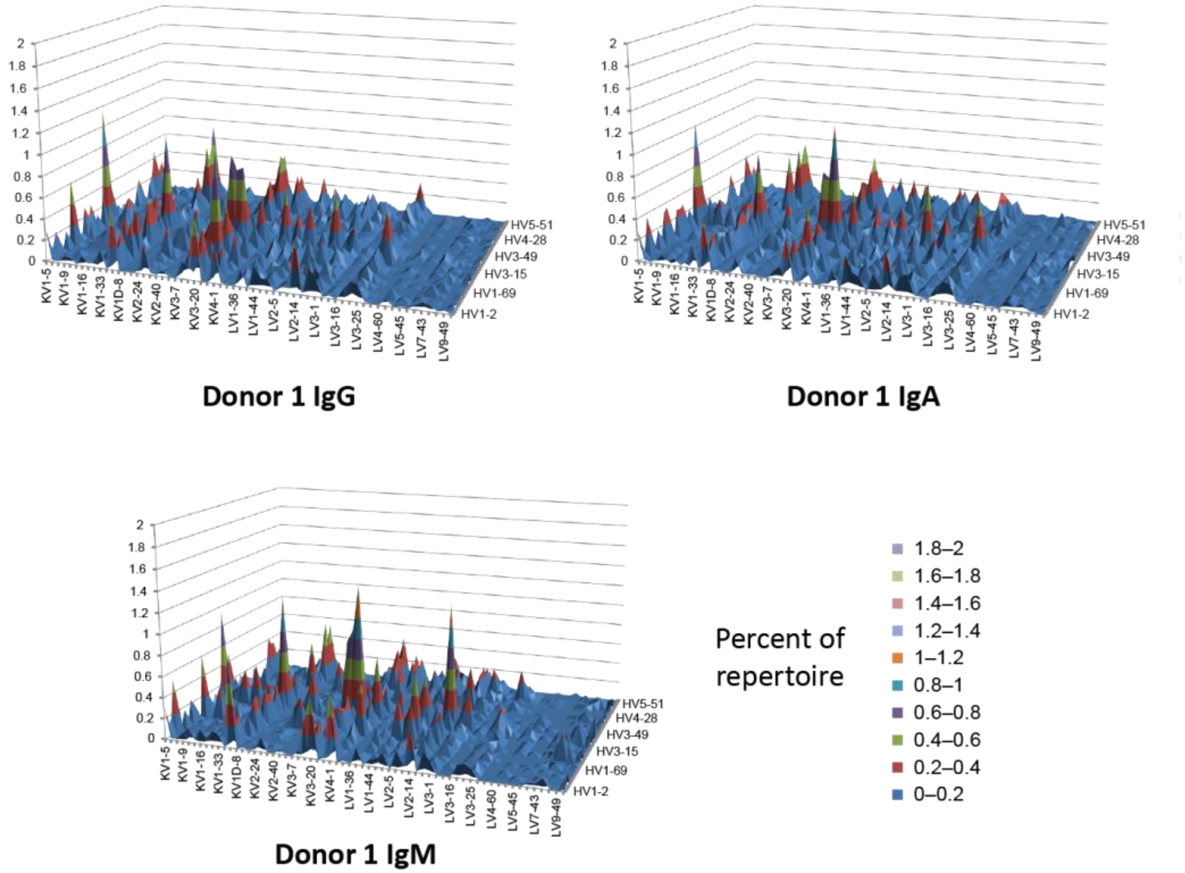
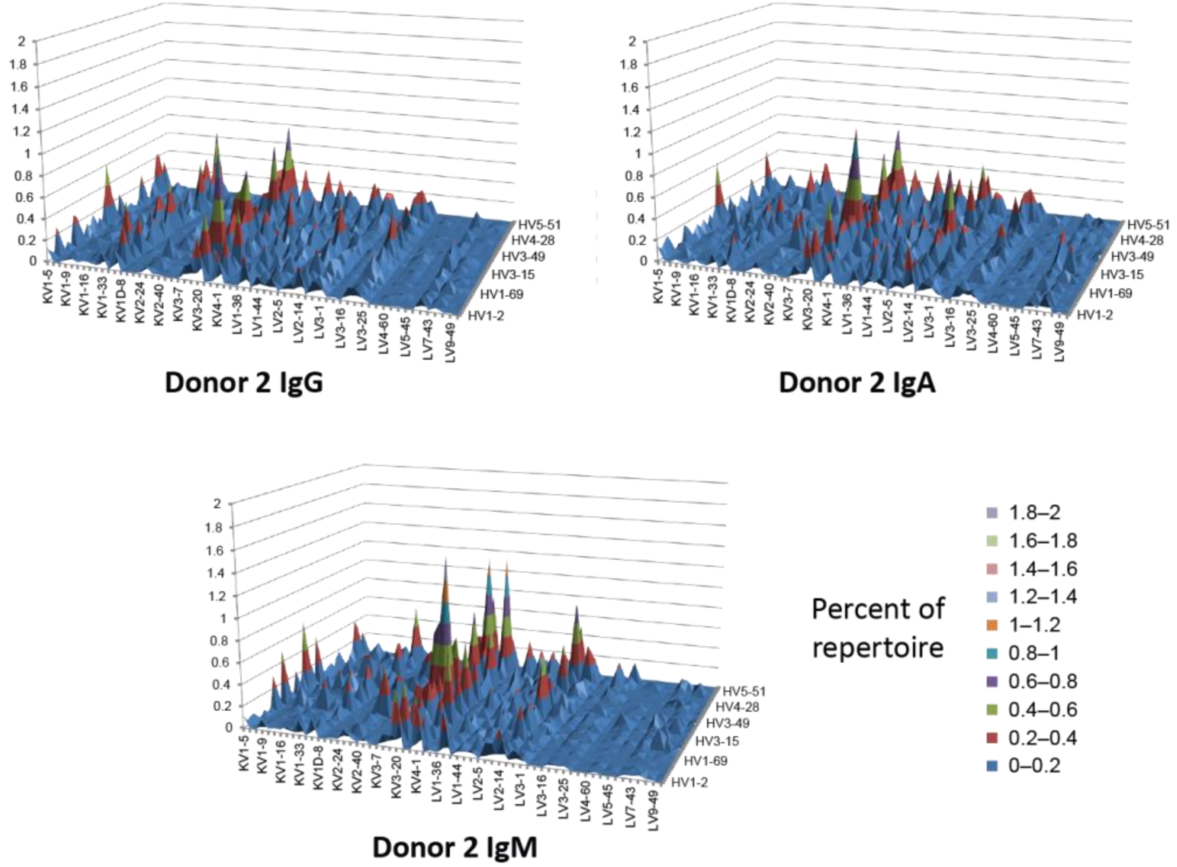


Figure A3.1 V-gene pairing surface plot for B-cell receptors observed in Donor 1, separated by heavy chain isotype.



Donor 2 IgG

Donor 2 IgA

Donor 2 IgM

Figure A3.2 V-gene pairing surface plot for B-cell receptors observed in Donor 2, separated by heavy chain isotype.

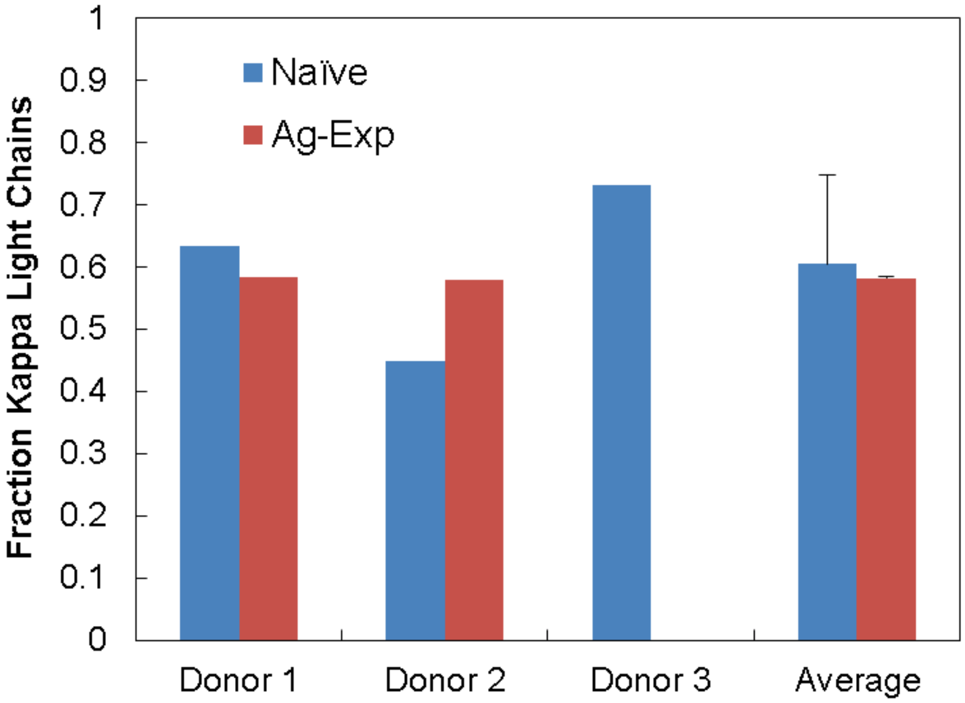
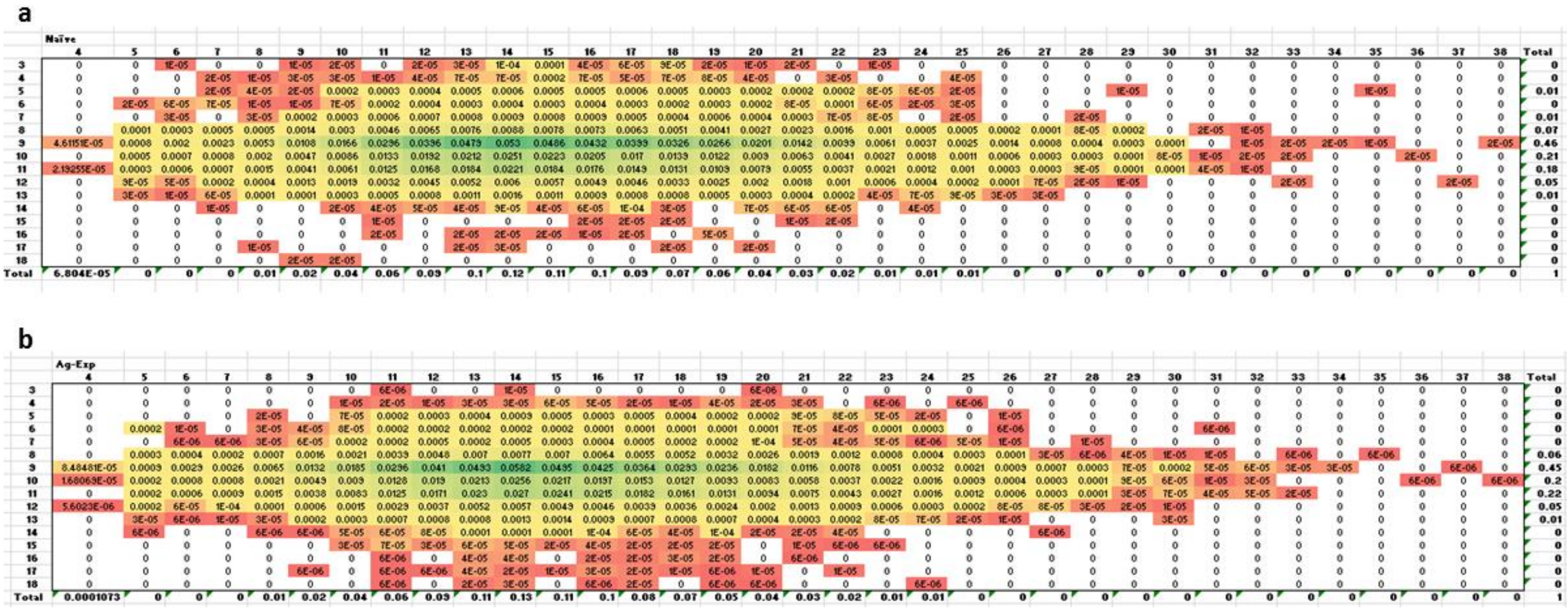


Figure A3.3 Fraction IgK light chain gene usage across B cell subsets for Donors 1, 2, and 3. Error bars indicate standard deviation for averaged values.



a

Naïve

	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	Total
3	0	0	1E-05	0	0	1E-05	2E-05	0	2E-05	3E-05	1E-04	0.0001	4E-05	6E-05	9E-05	2E-05	1E-05	2E-05	0	1E-05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	2E-05	1E-05	3E-05	3E-05	1E-05	4E-05	7E-05	7E-05	0.0002	7E-05	5E-05	7E-05	8E-05	4E-05	0	3E-05	0	0	4E-05	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	2E-05	4E-05	2E-05	0.0002	0.0003	0.0004	0.0005	0.0006	0.0005	0.0005	0.0006	0.0005	0.0003	0.0002	0.0002	0.0002	8E-05	6E-05	2E-05	0	0	0	1E-05	0	0	0	0	0	1E-05	0	0	0	0.01
6	0	2E-05	6E-05	7E-05	1E-05	1E-05	7E-05	0.0002	0.0004	0.0003	0.0004	0.0003	0.0004	0.0003	0.0002	0.0003	0.0002	8E-05	0.0001	6E-05	2E-05	3E-05	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	3E-05	0	3E-05	0.0002	0.0003	0.0006	0.0007	0.0008	0.0009	0.0008	0.0009	0.0005	0.0004	0.0006	0.0004	0.0003	7E-05	8E-05	0	2E-05	0	0	2E-05	0	0	0	0	0	0	0	0	0	0	0.01
8	0	0.0001	0.0003	0.0005	0.0005	0.0014	0.003	0.0046	0.0065	0.0076	0.0088	0.0078	0.0073	0.0063	0.0051	0.0041	0.0027	0.0023	0.0016	0.001	0.0005	0.0005	0.0002	0.0001	8E-05	0.0002	0	2E-05	1E-05	0	0	0	0	0	0	0.07
9	4.61151E-05	0.0008	0.002	0.0023	0.0053	0.0108	0.0166	0.0296	0.0396	0.0479	0.053	0.0486	0.0432	0.0399	0.0326	0.0266	0.0201	0.0142	0.0099	0.0061	0.0037	0.0025	0.0014	0.0008	0.0004	0.0003	0.0001	0	1E-05	2E-05	2E-05	1E-05	0	0	2E-05	0.46
10	0	0.0005	0.0007	0.0008	0.002	0.0047	0.0086	0.0133	0.0192	0.0212	0.0251	0.0223	0.0205	0.017	0.0139	0.0122	0.009	0.0063	0.0041	0.0027	0.0018	0.0011	0.0006	0.0003	0.0003	0.0001	8E-05	1E-05	2E-05	2E-05	0	0	2E-05	0	0	0.21
11	2.19255E-05	0.0003	0.0006	0.0007	0.0015	0.0041	0.0061	0.0125	0.0168	0.0184	0.0221	0.0184	0.0176	0.0149	0.0131	0.0109	0.0079	0.0055	0.0037	0.0021	0.0012	0.001	0.0003	0.0003	3E-05	0.0001	0.0001	4E-05	1E-05	0	0	0	0	0	0	0.18
12	0	9E-05	5E-05	0.0002	0.0004	0.0013	0.0019	0.0032	0.0045	0.0052	0.006	0.0057	0.0049	0.0046	0.0033	0.0025	0.002	0.0018	0.001	0.0006	0.0004	0.0002	0.0001	7E-05	2E-05	1E-05	0	0	0	2E-05	0	0	0	2E-05	0	0.05
13	0	3E-05	1E-05	6E-05	0.0001	0.0001	0.0003	0.0005	0.0008	0.0011	0.0016	0.0011	0.0009	0.0008	0.0008	0.0005	0.0003	0.0004	0.0002	4E-05	7E-05	3E-05	3E-05	3E-05	0	0	0	0	0	0	0	0	0	0	0	0.01
14	0	0	0	1E-05	0	0	2E-05	4E-05	5E-05	4E-05	9E-05	4E-05	6E-05	1E-04	3E-05	0	7E-05	6E-05	6E-05	0	4E-05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	1E-05	0	0	0	0	2E-05	2E-05	2E-05	0	0	1E-05	2E-05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	2E-05	0	2E-05	2E-05	2E-05	1E-05	2E-05	0	5E-05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	1E-05	0	0	0	0	2E-05	3E-05	0	0	0	2E-05	0	2E-05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	2E-05	2E-05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	6.804E-05	0	0	0	0.01	0.02	0.04	0.06	0.09	0.1	0.12	0.11	0.1	0.09	0.07	0.06	0.04	0.03	0.02	0.01	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	1

b

Ag-Exp

	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	Total
3	0	0	0	0	0	0	0	6E-06	0	0	1E-05	0	0	0	0	0	6E-06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	1E-05	2E-05	1E-05	3E-05	3E-05	6E-05	5E-05	2E-05	1E-05	4E-05	2E-05	3E-05	0	6E-06	0	6E-06	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	2E-05	0	7E-05	0.0002	0.0003	0.0004	0.0003	0.0005	0.0003	0.0005	0.0004	0.0002	0.0002	9E-05	8E-05	5E-05	2E-05	0	1E-05	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0.0002	1E-05	0	3E-05	4E-05	8E-05	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	7E-05	4E-05	0.0001	0.0003	0	6E-06	0	0	0	0	6E-06	0	0	0	0	0	0	0	0
7	0	0	6E-06	6E-06	3E-05	6E-05	0.0002	0.0002	0.0005	0.0002	0.0005	0.0003	0.0004	0.0005	0.0002	0.0002	1E-04	5E-05	4E-05	5E-05	6E-06	5E-05	1E-05	0	1E-05	0	0	0	0	0	0	0	0	0	0	0
8	0	0.0003	0.0004	0.0002	0.0007	0.0016	0.0021	0.0039	0.0048	0.007	0.0077	0.007	0.0064	0.0055	0.0052	0.0032	0.0026	0.0019	0.0012	0.0008	0.0004	0.0003	0.0001	3E-05	6E-06	4E-05	1E-05	1E-05	0	6E-06	0	6E-06	0	0	0	0.06
9	8.48481E-05	0.0009	0.0029	0.0026	0.0065	0.0132	0.0185	0.0296	0.041	0.0493	0.0582	0.0495	0.0425	0.0364	0.0293	0.0236	0.0182	0.0116	0.0078	0.0051	0.0032	0.0021	0.0009	0.0007	0.0003	7E-05	0.0002	5E-05	6E-05	3E-05	3E-05	0	0	6E-06	0	0.45
10	1.68069E-05	0.0002	0.0008	0.0008	0.0021	0.0049	0.009	0.0128	0.019	0.0213	0.0256	0.0217	0.0197	0.0153	0.0127	0.0093	0.0083	0.0058	0.0037	0.0022	0.0016	0.0009	0.0004	0.0003	0.0001	9E-05	6E-05	1E-05	3E-05	0	0	0	6E-06	0	6E-06	0.2
11	0	0.0002	0.0006	0.0009	0.0015	0.0038	0.0083	0.0125	0.0171	0.023	0.027	0.0241	0.0215	0.0182	0.0161	0.0131	0.0094	0.0075	0.0043	0.0027	0.0016	0.0012	0.0006	0.0003	0.0001	3E-05	7E-05	4E-05	5E-05	2E-05	0	0	0	0	0	0.22
12	5.6023E-06	0.0002	6E-05	1E-04	0.0001	0.0006	0.0015	0.0029	0.0037	0.0052	0.0057	0.0049	0.0046	0.0039	0.0036	0.0024	0.002	0.0013	0.0009	0.0006	0.0003	0.0002	8E-05	8E-05	3E-05	2E-05	1E-05	0	0	0	0	0	0	0	0	0.05
13	0	3E-05	6E-06	1E-05	3E-05	0.0002	0.0003	0.0007	0.0008	0.0008	0.0013	0.0014	0.0009	0.0007	0.0008	0.0007	0.0004	0.0003	0.0002	8E-05	7E-05	2E-05	1E-05	0	0	0	3E-05	0	0	0	0	0	0	0	0	0.01
14	0	6E-06	0	0	6E-06	6E-06	5E-05	6E-05	8E-05	0.0001	0.0001	0.0001	1E-04	6E-05	4E-05	1E-04	2E-05	2E-05	4E-05	0	0	0	0	6E-06	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	3E-05	7E-05	3E-05	6E-05	5E-05	2E-05	4E-05	2E-05	2E-05	2E-05	0	1E-05	6E-06	6E-06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	6E-06	0	4E-05	4E-05	0	2E-05	2E-05	3E-05	2E-05	0	6E-06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	6E-06	0	6E-06	6E-06	4E-05	2E-05	1E-05	3E-05	2E-05	1E-05	6E-06	1E-05	0	1E-05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	6E-06	0	2E-05	3E-05	0	6E-06	2E-05	0	6E-06	6E-06	0	0	0	6E-06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	0.0001073	0	0	0	0.01	0.02	0.04	0.06	0.09	0.11	0.13	0.11	0.1	0.08	0.07	0.05	0.04	0.03	0.02	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure A3.4 CDR-H3:CDR-L3 length heat maps of (a) naïve donor repertories, and (b) antigen-experienced donor repertoires.

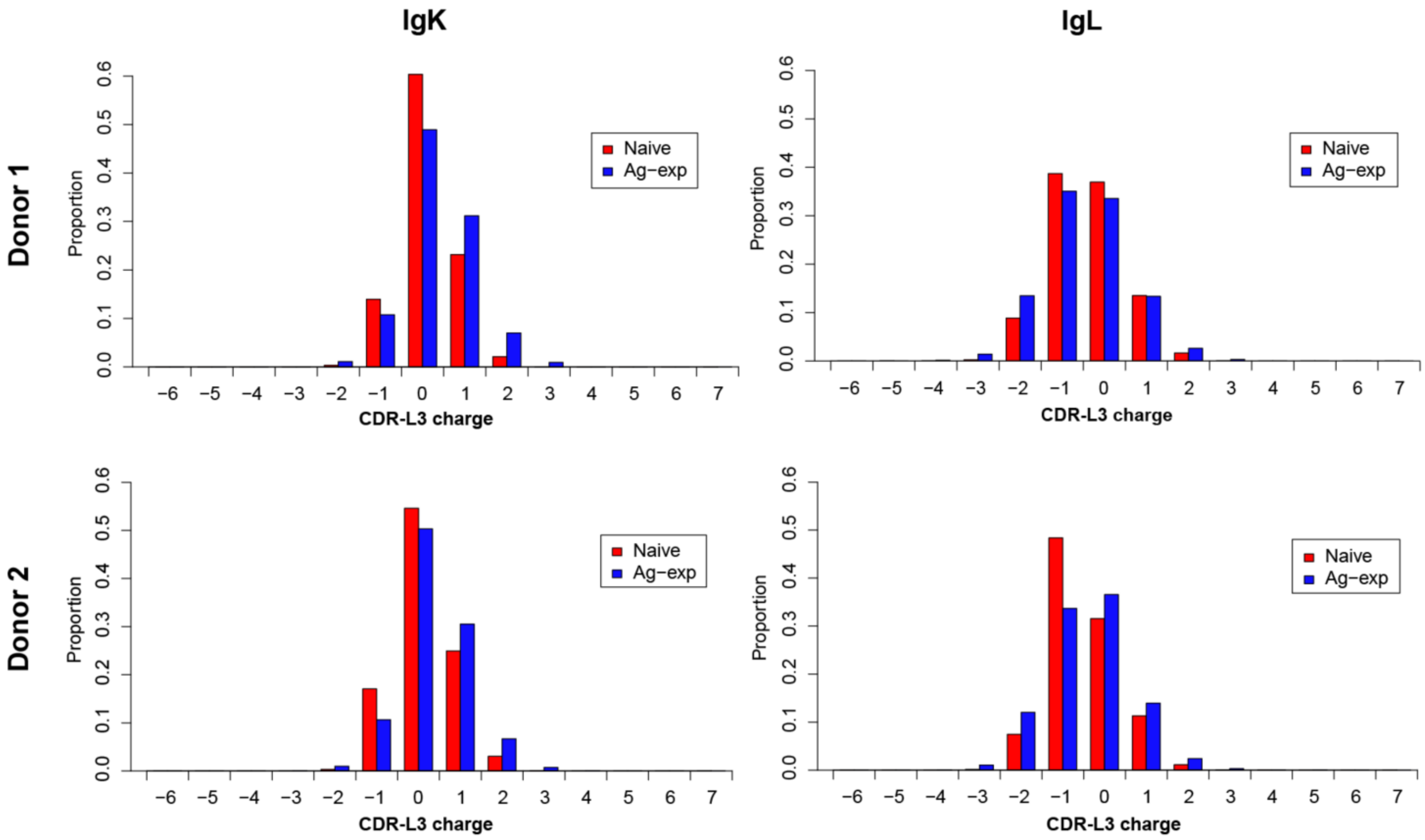


Figure A3.5 CDR-L3 loop charge in naïve and antigen-experienced antibody repertoires for Donor 1 (*left*) and Donor 2 (*right*), subdivided by IgK (*top*) and IgL (*bottom*). This figure demonstrates that kappa and lambda repertoires exhibit distinct CDR-L3 charge distributions (top compared to bottom graphs). All naïve repertoires were statistically significant from antigen-experienced repertoires in the same group by the K-S test (D1-K, D2-K, D2-L $p < 10^{-14}$, D1-L $p < 10^{-10}$); n for all distributions are provided in Table A3.2.

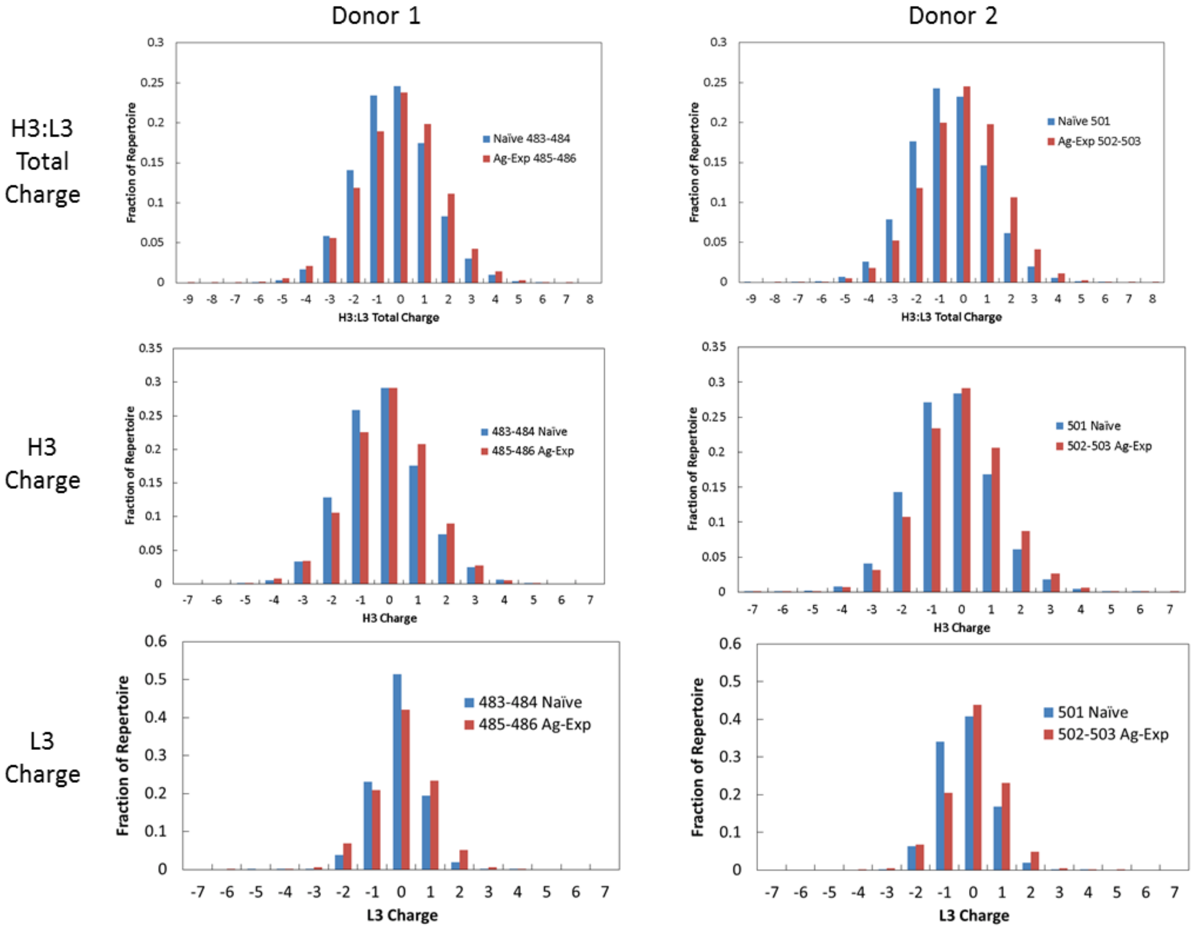
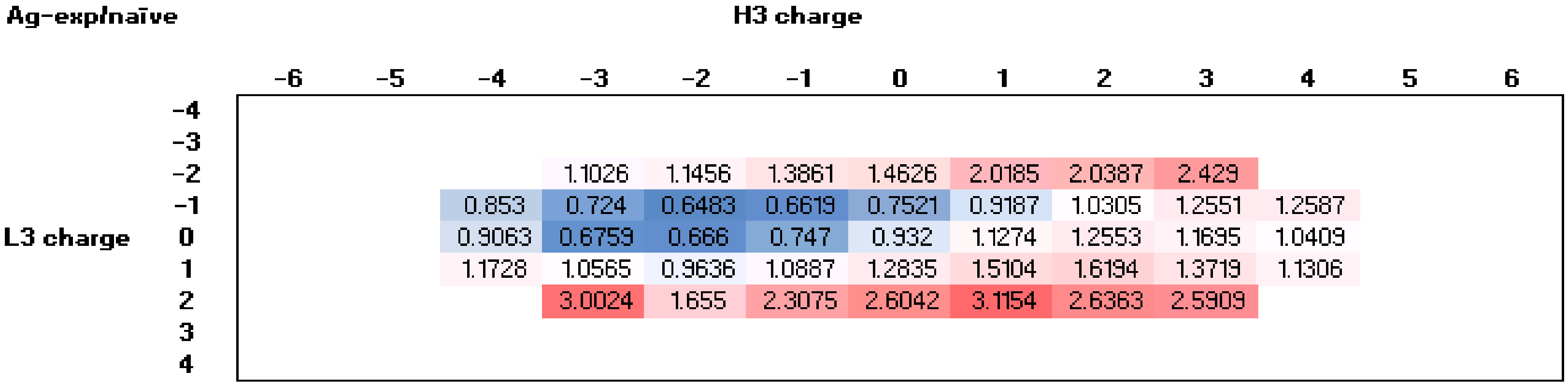


Figure A3.6 Charge distributions for naïve and antigen-experienced repertoires of Donors 1 and 2, further subdivided by CDR-H3:CDR-L3 total charge (top), CDR-H3 charge (middle), and CDR-L3 charge (lower).



Ag-exp/naïve		H3 charge												
		-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
	-4													
	-3													
	-2				1.1026	1.1456	1.3861	1.4626	2.0185	2.0387	2.429			
	-1			0.853	0.724	0.6483	0.6619	0.7521	0.9187	1.0305	1.2551	1.2587		
L3 charge	0			0.9063	0.6759	0.666	0.747	0.932	1.1274	1.2553	1.1695	1.0409		
	1			1.1728	1.0565	0.9636	1.0887	1.2835	1.5104	1.6194	1.3719	1.1306		
	2				3.0024	1.655	2.3075	2.6042	3.1154	2.6363	2.5909			
	3													
	4													

Figure A3.7 Relative representation ratio heat map of CDR-H3:CDR-L3 charge combinations across naïve and antigen-experienced repertoires. Values represent the ratio of antigen-experienced:naïve repertoire fractional representation for a given H3:L3 charge combination; red and blue shading represents relative increases and decreases in representation in antigen-experienced compared to naïve repertoires, respectively.

ID	logFC	AveExpr	t	P.Value	adj.P.Val
HV3-33:KV1-8	-4.625	2.934	-7.087	5.08E-06	2.68E-03
HV6-1:KV1-33	-4.510	3.723	-6.968	6.14E-06	2.68E-03
HV4-34:KV1-8	-4.127	3.796	-5.959	3.31E-05	8.46E-03
HV3-74:KV4-1	3.986	3.933	5.789	4.47E-05	8.46E-03
HV3-74:KV2-28	4.151	3.044	5.742	4.85E-05	8.46E-03
HV6-1:LV3-19	-3.624	2.692	-5.238	1.20E-04	1.38E-02
HV3-74:LV2-8	3.510	2.410	5.139	1.45E-04	1.38E-02
HV1-69:KV1-8	-3.952	3.334	-5.098	1.56E-04	1.38E-02
HV3-7:KV4-1	3.595	4.312	5.096	1.57E-04	1.38E-02
HV3-15:KV2-28	3.455	3.317	5.043	1.73E-04	1.38E-02
HV1-18:KV1-8	-3.454	3.035	-5.041	1.74E-04	1.38E-02
HV3-74:LV1-51	3.289	2.490	4.857	2.45E-04	1.59E-02
HV4-59:KV1-8	-3.292	4.021	-4.827	2.59E-04	1.59E-02
HV6-1:LV1-40	-3.251	3.659	-4.813	2.67E-04	1.59E-02
HV1-24:LV2-14	-3.250	4.539	-4.798	2.74E-04	1.59E-02
HV1-58:KV1-33	-3.039	2.682	-4.530	4.58E-04	2.49E-02
HV4-34:LV3-1	-2.907	4.976	-4.337	6.65E-04	3.14E-02
HV5-51:KV1-8	-3.339	2.720	-4.307	7.04E-04	3.14E-02
HV1-3:KV4-1	3.324	3.625	4.291	7.26E-04	3.14E-02
HV1-58:LV3-1	-2.873	2.283	-4.258	7.75E-04	3.14E-02
HV3-30:KV1-8	-2.917	3.613	-4.248	7.90E-04	3.14E-02
HV6-1:LV3-1	-3.043	3.618	-4.236	8.09E-04	3.14E-02
HV1-58:KV1-39	-3.768	3.078	-4.224	8.27E-04	3.14E-02
HV4-61:KV3-20	2.963	3.795	4.061	1.14E-03	3.99E-02
HV1-69:LV3-1	-2.638	5.963	-4.059	1.14E-03	3.99E-02
HV3-21:KV1-8	-2.767	3.706	-3.985	1.32E-03	4.35E-02
HV2-26:KV1-33	-2.606	3.537	-3.977	1.35E-03	4.35E-02
HV1-46:KV1-33	-2.554	5.407	-3.910	1.54E-03	4.79E-02

Table A3.1 Statistically significant differentially expressed heavy/light V-gene pairs with adjusted p < 0.05 between naïve and antigen-experienced antibody repertoires (abbreviations: *logFC* log2 fold change between conditions, *AveExpr* log2 average expression across all observed values, *t* moderated t-statistic, *P.Value* associated *p*-value, *adj.P.val* adjusted *p*-value.)

Isotype	Cell subset	Donor 1	Donor 2
IgK	Naïve	8,521	11,312
	Ag-Exp	19,755	49,844
IgL	Naïve	4,919	13,860
	Ag-Exp	14,067	36,126

Table A3.2 Recovered IgK and IgL pairs for Donor 1 and Donor 2 among naïve and antigen-experienced subsets. These data are represented graphically in Figure 4.6.

Donor	CDR-H3_aa	CDR-L3_aa	CDR-H3_nt	CDR-L3_nt	Gene usage
1	ARSSRGAAAGFDY	GTWDSSLSA-VV	gcgaga**AGCTCGAGGGG**agcagcagc**CGGT**tttgactac	ggaacatgggacagcagcctgagtgct---gtggta	HV4-59:D6-13:J4/λV1-51:J2
3	ARSSRGAAAGFDY	GTWDSSLSAGVV	gcgaga**TCTTCCCGCGG**agcagcagctggttttgactac	ggaacatgggatagcagcctgagtgctgg**G**gtggta	HV4-59:D6-13:J4/λV1-51:J2
2	ARSSRGAAAGFDY	(LQNAGVTGT)	gcgaga**TCTTCACGCGGG**gcagcagctggttttgactac	(ctgcagaatgctggc**GTAACGG**ggacg)	HV4-59:D6-13:J4/κV3-20:J1
2	ANEIRPNDY	LISYTDARIWV	gcgaacga**AATAAGAC**ctaatgactac	ttaatctcctatactgatgctcg**AAT**ttgggtg	HV3-23:D3-3:J4/λV7-46:J3
3	ANEIRPNDY	LLSYGDV--WV	gcgaacga**GATCCGACCCA**atgattat	ttgctctcctacggtgatg------tttgggtg	HV3-23:D1-14:J4/λV7-46:J3
1	ARARGYSYGYSDY	QKYNSAPALT	gcgagag**CTC**gtggatacagctatggttactctgactac	caaaagtataacagtgcccc**CGC**gctcact	HV1-8:D5-18:J4/κV1-27:J4
3	ARARGYSYGYSDY	QKYNSAPP-T	gcgagag**CCC**gtggatacagctatggttactctgactac	caaaagtataacagtgcccctcc---cact	HV3-48:D5-18:J4/κV1-27:J4
1	ARGHYGLDV	QQYGSSPIT	gcgagag**GAC**actacggtttggacgtc	cagcaatatggtagctcaccgatcacc	HV3-11:-:J6/κV3-20:J5
2	ARGHYGLDV	QQYGSSSRT	gcgagag**GTC**actacggtttggacgtc	cagcagtatggtagctcatctc**GA**acg	HV3-7:-:J6/κV3-20:J1
1	ARGEDYYYGMDV	QSIDTDGTHGVV	gccaggggga**AG**actactactacggtatggacgtc	cagtcaattgacaccgatggtactcat**GG**tgtggta	HV3-11:D3-16:J6/λV3-25:J2
2	ARGEDYYYGMDV	QSADSSGTY-VV	gcgcgaggtga**AG**actactactacggtatggacgtc	caatcagcagacagcagtggcactta---tgtggta	HV3-11:D2-21:J6/λV3-25:J2

Table A3.3 Selected nucleotide sequences for public ag-exp CDR-H3 amino acid BCR. Donor 3 in this analysis corresponds to Donor 3 in a previous study[83]. Non-templated bases are in bold/caps and deviations from germline underlined. Distinct nucleotide sequences and CDR-L3 lengths indicated distinct heavy and light recombination events.

8. REFERENCES

1. Reddy, S. T. & Georgiou, G. Systems analysis of adaptive immunity by utilization of high-throughput technologies. *Curr. Opin. Biotechnol.* **22,** 584–589 (2011).
2. Georgiou, G. *et al.* The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* **32,** 158–168 (2014).
3. Mathonet, P. & Ullman, C. The application of next generation sequencing to the understanding of antibody repertoires. *Front. Immunol.* **4,** 265 (2013).
4. Reddy, S. T. *et al.* Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat. Biotechnol.* **28,** 965–969 (2010).
5. Wilson, P. C. & Andrews, S. F. Tools to therapeutically harness the human antibody response. *Nat. Rev. Immunol.* **12,** 709–719 (2012).
6. Fischer, N. Sequencing antibody repertoires: the next generation. *MAbs* **3,** 17–20 (2011).
7. Wu, X. *et al.* Focused Evolution of HIV-1 Neutralizing Antibodies Revealed by Structures and Deep Sequencing. *Science* **333,** 1593–1602 (2011).
8. Berg, J. M. *et al. Biochemistry*. (W H Freeman, 2002).
9. Murphy, K. W., Mark; Janeway, Charles; Travers, Paul J. *Janeway's immunobiology*. (Garland Science, 2012).
10. Hess, J. *et al.* Induction of pre-B cell proliferation after de novo synthesis of the pre-B cell receptor. *Proc. Natl. Acad. Sci. U.S.A.* **98,** 1745–1750 (2001).
11. Lee, J., Monson, N. L. & Lipsky, P. E. The VλJλ Repertoire in Human Fetal Spleen: Evidence for Positive Selection and Extensive Receptor Editing. *J. Immunol.* **165,** 6322–6333 (2000).
12. Wardemann, H. *et al.* Predominant autoantibody production by early human B cell precursors. *Science* **301,** 1374–1377 (2003).
13. Qi, Q. *et al.* Diversity and clonal selection in the human T-cell repertoire. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 13139–13144 (2014).
14. Kuppers, R., Zhao, M., Hansmann, M. L. & Rajewsky, K. Tracing B-cell development in human germinal centers by molecular analysis of single cells picked from histological sections. *Embo J.* **12,** 4955–4967 (1993).
15. Fernández, D. *et al.* The Proto-Oncogene c-myc Regulates Antibody Secretion and Ig Class Switch Recombination. *J. Immunol.* (2013). doi:10.4049/jimmunol.1300712
16. Gitlin, A. D., Shulman, Z. & Nussenzweig, M. C. Clonal selection in the germinal centre by regulated proliferation and hypermutation. *Nature* **509,** 637–640 (2014).
17. Barnett, B. E. *et al.* Asymmetric B Cell Division in the Germinal Center Reaction. *Science* **335,** 342–344 (2012).
18. Klein, U. & Dalla-Favera, R. Germinal centres: role in B-cell physiology and malignancy. *Nat. Rev. Immunol.* **8,** 22–33 (2008).
19. William, J., Euler, C., Christensen, S. & Shlomchik, M. J. Evolution of Autoantibody Responses via Somatic Hypermutation Outside of Germinal Centers. *Science* **297,** 2066–2070 (2002).

20. Hinton, P. R. *et al.* An Engineered Human IgG1 Antibody with Longer Serum Half-Life. *J. Immunol.* **176,** 346–356 (2006).
21. Kyu, S. Y. *et al.* Frequencies of human influenza-specific antibody secreting cells or plasmablasts post vaccination from fresh and frozen peripheral blood mononuclear cells. *J. Immunol. Methods* **340,** 42–47 (2009).
22. Manz, R. A., Löhning, M., Cassese, G., Thiel, A. & Radbruch, A. Survival of long-lived plasma cells is independent of antigen. *Int. Immunol.* **10,** 1703–1711 (1998).
23. O'Connor, B. P., Cascalho, M. & Noelle, R. J. Short-lived and Long-lived Bone Marrow Plasma Cells Are Derived from a Novel Precursor Population. *J. Exp. Med.* **195,** 737–745 (2002).
24. Dorner, T. *et al.* Long-lived autoreactive plasma cells drive persistent autoimmune inflammation. *Nat. Rev. Rheumatol.* **7,** 170+ (2011).
25. Mahevas, M., Michel, M., Weill, J.-C. & Reynaud, C.-A. Long-Lived Plasma Cells in Autoimmunity: Lessons from B-Cell Depleting Therapy. *Front. Immunol.* **4,** (2013).
26. Yu, X. *et al.* Neutralizing antibodies derived from the B cells of 1918 influenza pandemic survivors. *Nature* **455,** 532–536 (2008).
27. Boyd, S. D. *et al.* Measurement and Clinical Monitoring of Human Lymphocyte Clonality by Massively Parallel V-D-J Pyrosequencing. *Sci. Transl. Med.* **1,** 12ra23 (2009).
28. Arnaout, R. *et al.* High-Resolution Description of Antibody Heavy-Chain Repertoires in Humans. *PLoS ONE* **6,** (2011).
29. Warren, E. H., Matsen, F. A. & Chou, J. High-throughput sequencing of B- and T-lymphocyte antigen receptors in hematology. *Blood* **122,** 19–22 (2013).
30. Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **30,** 434–439 (2012).
31. Bashford-Rogers, R. J. *et al.* Capturing needles in haystacks: a comparison of B-cell receptor sequencing methods. *BMC Immunol.* **15,** 29 (2014).
32. Schroeder Jr, H. W. Similarity and divergence in the development and expression of the mouse and human antibody repertoires. *Dev. Comp. Immunol.* **30,** 119–135 (2006).
33. Apostoaei, A. I. & Trabalka, J. R. Review, synthesis, and application of information on the human lymphatic system to radiation dosimetry for chronic lymphocytic leukemia. *SENES Oak Ridge Inc* (2012).
34. *National Human Genome Research Institute* Permission granted for presentation and teaching purposes at <http://www.genome.gov/sequencingcosts/>
35. Wine, Y. *et al.* Molecular deconvolution of the monoclonal antibodies that comprise the polyclonal serum response. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 2993–2998 (2013).
36. Menzel, U. *et al.* Comprehensive Evaluation and Optimization of Amplicon Library Preparation Methods for High-Throughput Antibody Sequencing. *PLoS ONE* **9,** e96727 (2014).
37. Greiff, V. *et al.* Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. *BMC Immunol.* **15,** 40 (2014).

38. Glanville, J. *et al.* Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 20216–20221 (2009).
39. Wu, Y.-C. B., Kipling, D. & Dunn-Walters, D. K. Age-related changes in human peripheral blood IGH repertoire following vaccination. *Front. Immunol.* **3,** (2012).
40. Schoettler, N., Ni, D. & Weigert, M. B cell receptor light chain repertoires show signs of selection with differences between groups of healthy individuals and SLE patients. *Mol. Immunol.* **51,** 273–282 (2012).
41. Hoi, K. H. & Ippolito, G. C. Intrinsic bias and public rearrangements in the human immunoglobulin V[lambda] light chain repertoire. *Genes Immun.* **14,** 271–276 (2013).
42. Zhu, J. *et al.* De novo identification of VRC01 class HIV-1–neutralizing antibodies by next-generation sequencing of B-cell transcripts. *Proc. Proc. Natl. Acad. Sci. U.S.A.* **110,** E4088–97 (2013).
43. Zhu, J. *et al.* Mining the antibodyome for HIV-1–neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 6470-6475 (2013).
44. Kohler, G. & Milstein, C. CONTINUOUS CULTURES OF FUSED CELLS SECRETING ANTIBODY OF PREDEFINED SPECIFICITY. *Nature* **256,** 495–497 (1975).
45. Yu, X., McGraw, P. A., House, F. S. & Crowe Jr, J. E. An optimized electrofusion-based protocol for generating virus-specific human monoclonal antibodies. *J. Immunol. Methods* **336,** 142–151 (2008).
46. Lonberg, N. *et al.* Antigen-specific human antibodies from mice comprising four distinct genetic modifications. *Nature* **368,** 856–859 (1994).
47. Mendez, M. J. *et al.* Functional transplant of megabase human immunoglobulin loci recapitulates human antibody response in mice. *Nat. Genet.* **15,** 146–156 (1997).
48. Lonberg, N. Human antibodies from transgenic animals. *Nat. Biotechnol.* **23,** 1117–1125 (2005).
49. Murphy, A. J. *et al.* Mice with megabase humanization of their immunoglobulin genes generate antibodies as efficiently as normal mice. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 5153–5158 (2014).
50. McCune, J. M. *et al.* The SCID-hu mouse: murine model for the analysis of human hematolymphoid differentiation and function. *Science* **241,** 1632–1639 (1988).
51. Hiramatsu, H. *et al.* Complete reconstitution of human lymphocytes from cord blood CD34+ cells using the NOD/SCID/γcnull mice model. *Blood* **102,** 873–880 (2003).
52. Ippolito, G. C. *et al.* Antibody Repertoires in Humanized NOD-scid-IL2R gamma(null) Mice and Human B Cells Reveals Human-Like Diversification and Tolerance Checkpoints in the Mouse. *PLoS ONE* **7,** e35497 (2012).
53. Rieger, M., Cervino, C., Sauceda, J. C., Niessner, R. & Knopp, D. Efficient Hybridoma Screening Technique Using Capture Antibody Based Microarrays. *Anal. Chem.* **81,** 2373–2377 (2009).

54. Ogunniyi, A., Story, C., Papa, E., Guillen, E. & Love, J. Screening individual hybridomas by microengraving to discover monoclonal antibodies. *Nat. Protoc.* **4,** 767–782 (2009).
55. Debs, B. E., Utharala, R., Balyasnikova, I. V., Griffiths, A. D. & Merten, C. A. Functional single-cell hybridoma screening using droplet-based microfluidics. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 11570–11575 (2012).
56. Smith, G. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228,** 1315–1317 (1985).
57. McCafferty, J., Griffiths, A. D., Winter, G. & Chiswell, D. J. Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* **348,** 552–554 (1990).
58. Marks, J. D. *et al.* Bypassing immunization - building high-affinity human antibodies by chain shuffling. *Bio-Technol.* **10,** 779–783 (1992).
59. Griffiths, A. D. *et al.* Isolation of high-affinity human antibodies directly from large synthetic repertoires. *Embo J.* **13,** 3245–3260 (1994).
60. Sblattero, D. & Bradbury, A. Exploiting recombination in single bacteria to make large phage antibody libraries. *Nat. Biotechnol.* **18,** 75–80 (2000).
61. Mazor, Y., Van Blarcom, T., Carroll, S. & Georgiou, G. Selection of full-length IgGs by tandem display on filamentous phage particles and Escherichia coli fluorescence-activated cell sorting screening. *FEBS J.* **277,** 2291–2303 (2010).
62. D'Angelo, S. *et al.* From deep sequencing to actual clones. *Protein Eng. Des. Sel.* **27,** 301–307 (2014).
63. Chan, C. E. Z., Lim, A. P. C., MacAry, P. A. & Hanson, B. J. The role of phage display in therapeutic antibody discovery. *Int. Immunol.* dxu082 (2014). doi:10.1093/intimm/dxu082
64. Zhou, T. *et al.* Multidonor Analysis Reveals Structural Elements, Genetic Determinants, and Maturation Pathway for HIV-1 Neutralization by VRC01-Class Antibodies. *Immunity* **39,** 245–258 (2013).
65. Lavinder, J. J. *et al.* Identification and Characterization of the Constituent Human Serum Antibodies Elicited by Vaccination. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 2259–2264 (2014).
66. Boutz, D. R. *et al.* Proteomic Identification of Monoclonal Antibodies from Serum. *Anal. Chem.* **86,** 4758–4766 (2014).
67. Sato, S. *et al.* Proteomics-directed cloning of circulating antiviral human monoclonal antibodies. *Nat. Biotechnol.* **30,** 1039–1043 (2012).
68. Cheung, W. C. *et al.* A proteomics approach for the identification and cloning of monoclonal antibodies from serum. *Nat. Biotechnol.* **30,** 447–452 (2012).
69. Lee, J. *et al.* Proteomic identification and characterization of monoclonal antibodies comprising the serological response to seasonal influenza vaccines.
70. Marcus, J. S., Anderson, W. F. & Quake, S. R. Microfluidic Single-Cell mRNA Isolation and Analysis. *Anal. Chem.* **78,** 3084–3089 (2006).
71. Toriello, N. M. *et al.* Integrated microfluidic bioprocessor for single-cell gene expression analysis. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 20173–8 (2008).

72. White, A. K. *et al.* High-throughput microfluidic single-cell RT-qPCR. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 13999–14004 (2011).
73. Furutani, S., Nagai, H., Takamura, Y., Aoyama, Y. & Kubo, I. Detection of expressed gene in isolated single cells in microchambers by a novel hot cell-direct RT-PCR method. *Analyst* **137,** 2951–2957 (2012).
74. Turchaninova, M. A. *et al.* Pairing of T-cell receptor chains via emulsion PCR. *Eur. J. Immunol.* **43,** 2507–2515 (2013).
75. Meijer, P. *et al.* Isolation of human antibody repertoires with preservation of the natural heavy and light chain pairing. *J. Mol. Biol.* **358,** 764–772 (2006).
76. Smith, K. *et al.* Rapid generation of fully human monoclonal antibodies specific to a vaccinating antigen. *Nat. Protoc.* **4,** 372–384 (2009).
77. Frölich, D. *et al.* Secondary Immunization Generates Clonally Related Antigen-Specific Plasma Cells and Memory B Cells. *J. Immunol.* **185,** 3103–3110 (2010).
78. Smith, K. *et al.* Fully human monoclonal antibodies from antibody secreting cells after vaccination with Pneumovax®23 are serotype specific and facilitate opsonophagocytosis. *Immunobiology* **218,** 745–754 (2013).
79. Doria-Rose, N. A. *et al.* Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature* **509,** 55–62 (2014).
80. Poulsen, T. R., Meijer, P.-J., Jensen, A., Nielsen, L. S. & Andersen, P. S. Kinetic, Affinity, and Diversity Limits of Human Polyclonal Antibody Responses against Tetanus Toxoid. *J. Immunol.* **179,** 3841–3850 (2007).
81. Meijer, P.-J., Nielsen, L. S., Lantto, J. & Jensen, A. in (ed. Dimitrov, A. S.) **525,** 261–277 (Humana Press, 2009).
82. DeKosky, B. J. *et al.* High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.* **31,** 166–169 (2013).
83. DeKosky, B. J. *et al.* In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat. Med.* **21,** 89-91 (2014).
84. Tanaka, Y. *et al.* Single-Cell Analysis of T-Cell Receptor Repertoire of HTLV-1 Tax-Specific Cytotoxic T Cells in Allogeneic Transplant Recipients with Adult T-Cell Leukemia/Lymphoma. *Cancer Res.* **70,** 6181–6192 (2010).
85. Scheid, J. F. *et al.* Differential regulation of self-reactivity discriminates between IgG(+) human circulating memory B cells and bone marrow plasma cells. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 18044–18048 (2011).
86. Li, G.-M. *et al.* Pandemic H1N1 influenza vaccine induces a recall response in humans that favors broadly cross-reactive memory B cells. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 9047-9052 (2012)
87. Sanchez-Freire, V., Ebert, A. D., Kalisky, T., Quake, S. R. & Wu, J. C. Microfluidic single-cell real-time PCR for comparative analysis of gene expression patterns. *Nat. Protoc.* **7,** 829–838 (2012).
88. Lindström, S., Hammond, M., Brismar, H., Andersson-Svahn, H. & Ahmadian, A. PCR amplification and genetic analysis in a microwell cell culturing chip. *Lab. Chip* **9,** 3465–3471 (2009).

89. Tokimitsu, Y. *et al.* Single lymphocyte analysis with a microwell array chip. *Cytometry A* **71,** 1003–1010 (2007).

90. Yamamura, S. *et al.* Single-cell microarray for analyzing cellular response. *Anal. Chem.* **77,** 8050–8056 (2005).

91. Tajiri, K. *et al.* Cell microarray analysis of antigen specific B cells: Single cell analysis of antigen receptor expression and specificity. *Cytometry A* **71,** 961–967 (2007).

92. Glanville, J. *et al.* Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 20066–20071 (2011).

93. Wrammert, J. *et al.* Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature* **453,** 667–71 (2008).

94. Scheid, J. F. *et al.* Sequence and Structural Convergence of Broad and Potent HIV Antibodies That Mimic CD4 Binding. *Science* **333,** 1633–1637 (2011).

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

96. Lim, T. S. *et al.* V-gene amplification revisited - An optimised procedure for amplification of rearranged human antibody genes of different isotypes. *New Biotechnol.* **27,** 108–117 (2010).

97. Mei, H. E. *et al.* Blood-borne human plasma cells in steady state are derived from mucosal immune responses. *Blood* **113,** 2461–9 (2009).

98. Mazor, Y., Barnea, I., Keydar, I. & Benhar, I. Antibody internalization studied using a novel IgG binding toxin fusion. *J. Immunol. Methods* **321,** 41–59 (2007).

99. Friguet, B., Chaffotte, A. F., Djavadi-Ohaniance, L. & Goldberg, M. E. Measurements of the true affinity constant in solution of antigen-antibody complexes by enzyme-linked immunosorbent assay. *J. Immunol. Methods* **77,** 305–319 (1985).

100. Logan, A. C. *et al.* High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 21194–21199 (2011).

101. Sasaki, S. *et al.* Limited efficacy of inactivated influenza vaccine in elderly individuals is associated with decreased production of vaccine-specific antibodies. *J. Clin. Invest.* **121,** 3109–3119 (2011).

102. Finn, J. A. & Crowe Jr., J. E. Impact of new sequencing technologies on studies of the human B cell repertoire. *Curr. Opin. Immunol.* **25,** 613–618 (2013).

103. Finco, O. & Rappuoli, R. Designing Vaccines for the Twenty-First Century Society. *Front. Immunol.* **5,** (2014).

104. Newell, E. W. & Davis, M. M. Beyond model antigens: high-dimensional methods for the analysis of antigen-specific T cells. *Nat. Biotechnol.* **32,** 149–157 (2014).

105. Berkland, C., Kim, K. K. & Pack, D. W. Fabrication of PLG microspheres with precisely controlled and monodisperse size distributions. *J. Controlled Release* **73,** 59–74 (2001).

106. Berkland, C., Pollauf, E., Pack, D. W. & Kim, K. Uniform double-walled polymer microspheres of controllable shell thickness. *J. Controlled Release* **96,** 101–111 (2004).
107. Recher, M. *et al.* IL-21 is the primary common γ chain-binding cytokine required for human B-cell differentiation in vivo. *Blood* **118,** 6824–6835 (2011).
108. Jackson, K. J. L., Kidd, M. J., Wang, Y. & Collins, A. M. The shape of the lymphocyte receptor repertoire: lessons from the B cell receptor. *Front. Immunol.* **4,** 1–12 (2013).
109. Jackson, K. J. L. *et al.* Divergent human populations show extensive shared IGK rearrangements in peripheral blood B cells. *Immunogenetics* **64,** 3–14 (2012).
110. Pelanda, R. Dual immunoglobulin light chain B cells: Trojan horses of autoimmunity? *Curr. Opin. Immunol.* **27,** 53–59 (2014).
111. Liu, S. *et al.* Receptor Editing Can Lead to Allelic Inclusion and Development of B Cells That Retain Antibodies Reacting with High Avidity Autoantigens. *J. Immunol.* **175,** 5067–5076 (2005).
112. Rezanka, L. J., Kenny, J. J. & Longo, D. L. Dual isotype expressing B cells [κ(+)/λ(+)] arise during the ontogeny of B cells in the bone marrow of normal nontransgenic mice. *Cell. Immunol.* **238,** 38–48 (2005).
113. Casellas, R. *et al.* Igκ allelic inclusion is a consequence of receptor editing. *J. Exp. Med.* **204,** 153–160 (2007).
114. Andrews, S. F. *et al.* Global analysis of B cell selection using an immunoglobulin light chain–mediated model of autoreactivity. *J. Exp. Med.* **210,** 125–142 (2013).
115. Giachino, C., Padovan, E. & Lanzavecchia, A. kappa+lambda+ dual receptor B cells are present in the human peripheral repertoire. *J. Exp. Med.* **181,** 1245–1250 (1995).
116. Corti, D. *et al.* A Neutralizing Antibody Selected from Plasma Cells That Binds to Group 1 and Group 2 Influenza A Hemagglutinins. *Science* **333,** 850–856 (2011).
117. Wrammert, J. *et al.* Broadly cross-reactive antibodies dominate the human B cell response against 2009 pandemic H1N1 influenza virus infection. *J. Exp. Med.* **208,** 181–193 (2011).
118. Jardine, J. *et al.* Rational HIV Immunogen Design to Target Specific Germline B Cell Receptors. *Science* **340,** 711–716 (2013).
119. Busse, C. E., Czogiel, I., Braun, P., Arndt, P. F. & Wardemann, H. Single-cell based high-throughput sequencing of full-length immunoglobulin heavy and light chain genes. *Eur. J. Immunol.* **44,** 597–603 (2014).
120. Parameswaran, P. *et al.* Convergent Antibody Signatures in Human Dengue. *Cell Host Microbe* **13,** 691–700 (2013).
121. Jackson, K. J. L. *et al.* Human Responses to Influenza Vaccination Show Seroconversion Signatures and Convergent Antibody Rearrangements. *Cell Host Microbe* **16,** 105–114 (2014).
122. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461 (2010).
123. Saha, S. & Raghava, G. P. S. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res.* **34,** W202–W209 (2006).

124. Brezinschek, H.-P., Foster, S. J., Dörner, T., Brezinschek, R. I. & Lipsky, P. E. Pairing of Variable Heavy and Variable κ Chains in Individual Naive and Memory B Cells. *J. Immunol.* **160,** 4762–4767 (1998).
125. Bräuninger, A., Goossens, T., Rajewsky, K. & Küppers, R. Regulation of immunoglobulin light chain gene rearrangements during early B cell development in the human. *Eur. J. Immunol.* **31,** 3631–3637 (2001).
126. Tian, C. X. *et al.* Evidence for preferential Ig gene usage and differential TdT and exonuclease activities in human naive and memory B cells. *Mol. Immunol.* **44,** 2173–2183 (2007).
127. Wu, Y. C. *et al.* High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* **116,** 1070–1078 (2010).
128. Mroczek, E. S. *et al.* Differences in the Composition of the Human Antibody Repertoire by B Cell Subsets in the Blood. *Front. Immunol.* **5,** (2014).
129. McGuire, A. T. *et al.* Engineering HIV envelope protein to activate germline B cell receptors of broadly neutralizing anti-CD4 binding site antibodies. *J. Exp. Med.* **210,** 655–663 (2013).
130. McLellan, J. S. *et al.* Structure-Based Design of a Fusion Glycoprotein Vaccine for Respiratory Syncytial Virus. *Science* **342,** 592–598 (2013).
131. Lindop, R. *et al.* Molecular signature of a public clonotypic autoantibody in primary Sjögren's syndrome: a 'forbidden' clone in systemic autoimmunity. *Arthritis Rheum.* **63,** 3477–3486 (2011).
132. Di Niro, R. *et al.* High abundance of plasma cells secreting transglutaminase 2-specific IgA autoantibodies with limited somatic hypermutation in celiac disease intestinal lesions. *Nat. Med.* **18,** 441–U204 (2012).
133. Smyth, G. K. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds. Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A. & Dudoit, S.) 397–420 (Springer New York, 2005).
134. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
135. De Wildt, R. M. T., Hoet, R. M. A., van Venrooij, W. J., Tomlinson, I. M. & Winter, G. Analysis of Heavy and Light Chain Pairings Indicates that Receptor Editing Shapes the Human Antibody Repertoire. *J. Mol. Biol.* **285,** 895–901 (1999).
136. Eisenberg, D. Three-Dimensional Structure of Membrane and Surface Proteins. *Annu. Rev. Biochem.* **53,** 595–623 (1984).
137. Briney, B. S., Willis, J. R., McKinney, B. A. & Crowe, J. E. High-throughput antibody sequencing reveals genetic evidence of global regulation of the naive and memory repertoires that extends across individuals. *Genes Immun.* (2012). at <http://dx.doi.org/10.1038/gene.2012.20>
138. Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* **41,** W34–W40 (2013).
139. Smyth, G. K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments : Statistical Applications in Genetics and Molecular Biology. *Stat. Appl. Genet. Mol. Biol.* **3,** Article 3 (2004).

140. Wang, B. *et al.* Rapid identification of high-affinity antibodies against Zaire Ebolavirus through repertoire analysis. (In preparation).

141. DeKosky, B. *et al.* Insights in the development of the human antibody repertoire via high-throughput VH:VL sequencing and structural modeling of naïve and memory B cell repertoires in healthy individuals. (In Preparation).