Springer Theses
Recognizing Outstanding Ph.D. Research

Brandon DeKosky

Decoding the Antibody Repertoire

High Throughput Sequencing of Multiple Transcripts from Single B Cells

Springer

Springer Theses

Recognizing Outstanding Ph.D. Research

Aims and Scope

The series "Springer Theses" brings together a selection of the very best Ph.D. theses from around the world and across the physical sciences. Nominated and endorsed by two recognized specialists, each published volume has been selected for its scientific excellence and the high impact of its contents for the pertinent field of research. For greater accessibility to non-specialists, the published versions include an extended introduction, as well as a foreword by the student's supervisor explaining the special relevance of the work for the field. As a whole, the series will provide a valuable resource both for newcomers to the research fields described, and for other scientists seeking detailed background information on special questions. Finally, it provides an accredited documentation of the valuable contributions made by today's younger generation of scientists.

Theses are accepted into the series by invited nomination only and must fulfill all of the following criteria

- They must be written in good English.
- The topic should fall within the confines of Chemistry, Physics, Earth Sciences, Engineering and related interdisciplinary fields such as Materials, Nanoscience, Chemical Engineering, Complex Systems and Biophysics.
- The work reported in the thesis must represent a significant scientific advance.
- If the thesis includes previously published material, permission to reproduce this must be gained from the respective copyright holder.
- They must have been examined and passed during the 12 months prior to nomination.
- Each thesis should include a foreword by the supervisor outlining the significance of its content.
- The theses should have a clearly defined structure including an introduction accessible to scientists not expert in that particular field.

More information about this series at http://www.springer.com/series/8790

Brandon DeKosky

Decoding the Antibody Repertoire

High Throughput Sequencing of Multiple Transcripts from Single B Cells

Doctoral Thesis accepted by
The University of Texas at Austin, USA

Springer

Author
Dr. Brandon DeKosky
Department of Chemical and Petroleum Engineering
Department of Pharmaceutical Chemistry
The University of Kansas
Lawrence, KS
USA

Supervisor
Dr. George Georgiou
Department of Chemical Engineering,
Department of Biomedical Engineering,
Institute for Cell and Molecular Biology
The University of Texas at Austin
Austin, TX
USA

ISSN 2190-5053 ISSN 2190-5061 (electronic)
Springer Theses
ISBN 978-3-319-58517-8 ISBN 978-3-319-58518-5 (eBook)
DOI 10.1007/978-3-319-58518-5

Library of Congress Control Number: 2017939625

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

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.

Supervisor's Foreword

Antibodies are a critical feature of our adaptive immune system, and the majority of vaccines function by eliciting antibody responses for rapid neutralization and clearance of infectious agents. Antibodies are derived from B cells, which each encode a unique antibody sequence. B cells first develop in bone marrow and then migrate to secondary lymphoid organs where they become "trained" to synthesize antibodies. This process of B-cell development and clonal selection results in mature antibodies that have been iteratively optimized for recognition of target molecules. B cells then undergo further developmental changes that lead them to become plasma cells, which are antibody-producing factories that reside predominantly in bone marrow. The diversity of antibodies encoded by mammalian B cells is immense and may exceed 10^8 distinct molecular sequences in a human.

Given the critical role of antibody-mediated protection, determination of antibody sequences (or the antibody repertoire) of B-cell populations is of fundamental importance in immunology and biotechnology. Antibody repertoire sequencing is now being applied to discover novel antibody therapeutics, to better understand B-cell development, and to elucidate disease mechanisms. There is a pressing need for technologies that enable sequence characterization of antibody repertoires at high throughput and with high accuracy given their very high diversity, and also for methods to analyze the large datasets generated by high-throughput technologies. Earlier scientists developed techniques for high-throughput nucleic acid sequencing of the two chains of the full antibody molecule separately. However, high-throughput determination of complete antibody variable region sequences (i.e., both heavy and light chains paired together) had not been achieved. Indeed, sequencing of antibody heavy and light chains together *en masse* was pursued for over twenty years with very little success due to the experimental difficulties posed by single-cell analysis of multiple genes. To directly address these issues, DeKosky and colleagues reported here the very first experimental workflow capable of determining paired antibody heavy and light chain sequences from large B-cell populations, as well as associated computational advances to analyze such datasets.

Initial experiments permitted analysis of over 10^4 individual cells per experiment, which was later expanded to $>10^6$ B cells using simple and easily implemented approaches.

The very high practical significance of high-throughput paired antibody heavy and light chain sequencing has already been demonstrated, being utilized in 11 original research publications over the past three years and with an accelerating number of reports in progress. In one example, high-throughput characterization of complete antibody variable regions enabled genetic and computational structure analysis of antibody repertoires from distinct B-cell compartments, revealing new hallmarks of functional antibody selection during B-cell development [1]. In other studies, paired heavy and light chain sequencing was used to discover the identities of the protective antibodies that circulate in human serum [2, 3]. Paired heavy:light sequencing was also applied to identify the earliest known complete antibody sequence of the VRC26 human HIV broadly neutralizing antibody [4] and to quantify vaccine-based elicitation of HIV broadly neutralizing antibody precursors [5]. Building on these and other reports, the experimental platforms developed here have enabled an entire suite of new research opportunities and will be applied to advance immune research and improve human health for many years to come.

Austin, USA Dr. George Georgiou, Ph.D.
December 2016

References

1. DeKosky BJ et al (2016) Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. Proc Natl Acad Sci 113:E2636–E2645
2. Lavinder JJ et al (2014) Identification and characterization of the constituent human serum antibodies elicited by vaccination. Proc Natl Acad Sci 111:2259–2264
3. Lee J et al (2016) Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. Nat Med 22:1456–1464
4. Doria-Rose NA et al (2014) Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. Nature 509:55–625.
5. Tian M et al (2016) Induction of HIV neutralizing antibody lineages in mice with diverse precursor repertoires. Cell 166:1471–1484

Preface

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. 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. 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, and the magnetic beads are then 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. We demonstrated experimental throughput of over 2×10^6 cells in a single day, with antibody heavy and light chain pairing accuracy greater than 97% as measured with in vitro expanded human B cells. These new single-cell sequencing technologies were then 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 the development of the antibody repertoire using a high-throughput and high-resolution examination of naïve and memory B-cell compartments in healthy human donors.

Lawrence, USA — Brandon DeKosky

Parts of this thesis have been published in the following journal articles:

1. DeKosky, B. J., Ippolito, G. C., Deschner, R. P., Lavinder, J. J., Wine, Y., Rawlings, B. M., Varadarajan, N., Giesecke, C., Dorner, T., Andrews, S. F., Wilson, P. C., Hunicke-Smith, S. P., Willson, C. G., Ellington, A. D. & Georgiou, G. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nature Biotechnology* **31,** 166–169 (2013).
2. DeKosky, B. J., Kojima, T., Rodin, A., Charab, W., Ippolito, G. C., Ellington, A. D. & Georgiou, G. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nature Medicine* **21,** 86–91 (2015).
3. DeKosky, B. J., Lungu, O. I., Park, D., Johnson, E. L., Charab, W., Chrysostomou, C., Kuroda, D., Ellington, A. D., Ippolito, G. C., Gray, J. J. & Georgiou, G. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proceedings of the National Academy of Sciences, U.S.A.* **113**, E2636–E2645 (2016).

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 to all of my fellow laboratory 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 laboratory, and thank you to the rest of the BIGG laboratory members for your help, tips, tricks, and time at the Institute, and for making the laboratory 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 high-throughput sequencing. Finally, thank you to those who provided 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 through the ups and downs in the research laboratory and beyond. Many thanks 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. 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 this research.

Contents

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
IgD	Immunoglobulin D
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

List of Figures

List of Tables

Chapter 1
Background

1.1 Antibodies and Antibody Repertoire Development

The last 30 years of the biotechnology revolution have to a great extent been fueled by the discovery and application of monoclonal antibodies for research and therapeutic purposes [1, 2]. Antibodies are Y-shaped molecules produced by the immune system of many vertebrates to recognize and neutralize foreign pathogens or toxins with very high specificity. The highly specific nature of antibody-antigen interactions provides a means of molecular targeting that is extremely useful both as scientific reagents and as clinical therapeutics. An antibody molecule is comprised of two major portions: the constant region (or Fc region), which does not vary among distinct antibody molecules with different specificities, and the variable region, which comprises a unique sequence for each antibody and is the region responsible for antigen recognition (Fig. 1.1). Binding to molecular targets occurs at the exposed ends of the variable region. Different antibody variable regions confer distinct antibody binding specificities conferred by their unique antibody amino acid sequences, which are in turn derived from somatic recombination of the immunoglobulin genes and subsequent clonal selection that occurs during B-cell development.

The antibody variable region can be further subdivided into framework regions and complementarity-determining regions (Fig. 1.1). The relatively conserved framework regions (FRs) consist of antiparallel β strands which form a β-sandwich structure called the immunoglobulin fold [3]. Within the antibody variable region are distinct areas of increased 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, whereas the binding region on an antibody's molecular target is termed the epitope. CDRs often comprise the antibody paratope.

B. DeKosky, *Decoding the Antibody Repertoire*,
Springer Theses, DOI 10.1007/978-3-319-58518-5_1

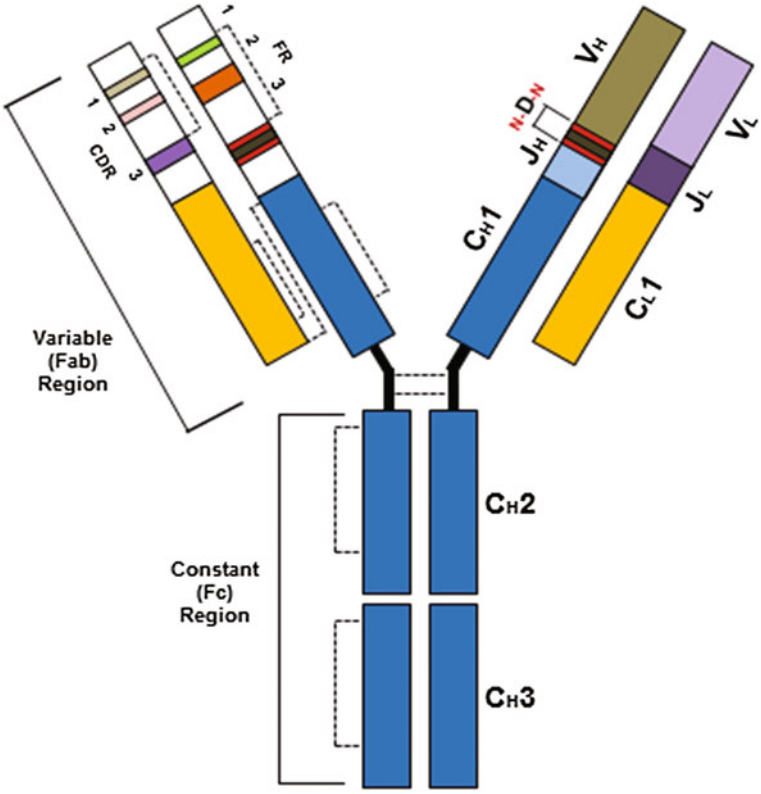

Fig. 1.1 An overview of antibody structure. Antibodies are Y-shaped molecules which consist of a variable region that imparts binding specificity and a constant region that determines the functionality of the antibody molecule. The variable region is subdivided into framework regions (FRs) which are somewhat conserved across antibodies, as well as complimentarity-determining regions (CDRs) which are highly variable across different antibodies; dashed lines indicate the location of disulfide bonds. 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

In addition to specificity conferred by the antibody variable region, antibody functionality is further determined by the characteristics of the antibody constant region (Fig. 1.1), termed the antibody class or isotype. Five major classes of antibody constant regions exist (IgD, IgM, IgG, IgA, and IgE), and several of these contain subclasses (e.g. IgG1, IgG2, IgA1, IgA2, etc.) Each antibody class and sub-class possesses a unique functional profile for conferring protection against pathogens and toxins. 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

Table 1.1 Overview of antibody constant regions, adapted from [4]

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

monomer, and is a less potent activator of complement. An overview of antibody isotypes is presented in Table 1.1.

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 (Figs. 1.1 and 1.2) [2, 4]. 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

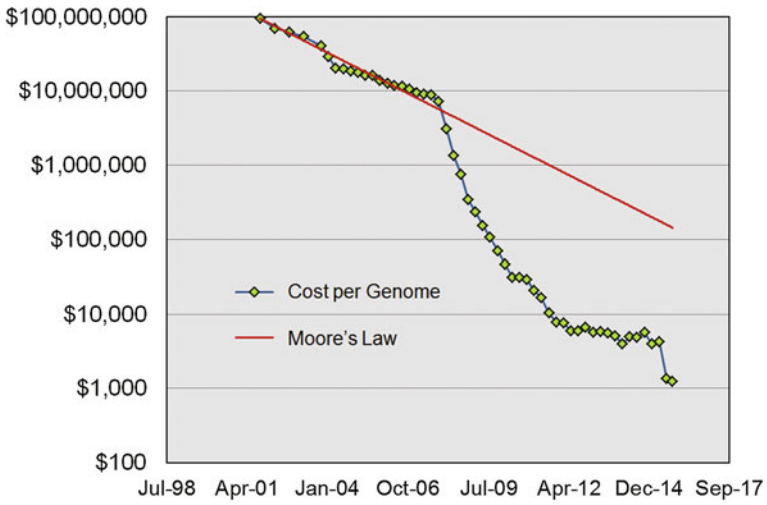

Fig. 1.2 High-throughput DNA sequencing costs have dropped dramatically in recent years beginning with the introduction of the first next-generation sequencing platforms in 2008 [32]. Log-scale reductions in DNA sequencing costs have directly enabled high-throughput analysis of individual antibody repertoires. New high-throughput DNA sequencing methods are fundamental drivers of recent advances in antibody repertoire sequencing and antibody discovery

sequences (RSS) [4]. 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 major enzyme in B-cell development is terminal deoxynucelotidyl transferase (TdT), which is a specialized polymerase 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 D-J and V-J junctions, thereby dramatically increasing the junctional diversity formed by recombination. The higher expression of TdT during heavy chain formation leads to 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). Lower expression of TdT in pre-B cells and the lack of a D-gene causes light chain CDR3 loop lengths to be shorter and more narrowly distributed than heavy chain CDR3 loops [4].

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 (Fig. 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 (Fig. 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 these recombination events mediated by RAG expression. Recombination is stopped when the newly formed heavy chain and a paired surrogate light chain are both expressed on the cell surface. (The surrogate light chain covers the newly-generated heavy chain variable region to inhibit aberrant B-cell receptor binding.) This important step ensures that a productive, in-frame heavy chain junction without stop codons has been generated; 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 decreases sharply and the B-cell is now denoted as a pre-B cell. At this stage, pre-B cells with newly formed heavy chains undergo pre-B expansion that consists of up to eight cell divisions that create multiple B-cell clones with the same heavy chain [5], which has been observed directly in mice and inferred in humans. Next, light chain V_L-J_L recombination occurs in 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 leads to generally restricted light chain CDR-L3 lengths [4]. As in pro-B cells, the final developmental checkpoint after the pre-B stage consists of B-cell receptor surface display, this time with the somatically generated heavy and light chains paired together. RAG expression and light chain recombination ceases, and the cell is now termed an immature B-cell.

Next, immature B-cells undergo an immune tolerance checkpoint to ensure that autoreactive antibodies do not advance in the antibody maturation pathway, as the process of random gene recombination without functional selection can generate self-reactive antibodies. Tolerance checkpoints are thought to occur via clonal deletion of B-cells encoding BCRs that bind to self-proteins, or alternatively,

receptor editing can rescue self-reactive B-cells at this stage [6]. A known hallmark of potential autoreactivity is the expression of a very long heavy chain CDR-H3 loop, and BCRs with long CDR-H3 s are preferentially deleted from the repertoire during the pre-B tolerance checkpoint [7]. After passing tolerance checkpoints, B-cells express both IgD and IgM constant regions as B-cell receptors on the cell surface and are then considered mature, or naïve, B-cells. The newly generated mature B-cells are ready for positive selection in germinal center reactions of the spleen and lymph nodes over the course of an immune response.

1.2 Adaptive Immune Responses Lead to B-Cell Activation and Antibody Secretion

Major reactions of the adaptive immune response occur in the the spleen and tissue-draining lymph nodes (collectively called the secondary lymphoid organs) [4]. There, specialized antigen-presenting cells (APCs) named dendritic cells and follicular dendritic cells present captured foreign proteins and present antigens for B- and T-cell surveillance. In the process of 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 highly analogous 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 [4, 8]. 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. 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 (also known as anergy) or clonal deletion. The requirement for dedicated APCs to mediate T-cell activation is a major immune control mechanism that minimizes autoimmunogenic T-cell proliferation and helps to ensure that T-cell activation occurs only as a result of the response to true infections.

Effector T cells comprise two major classes:

1. $CD8^+$ cytotoxic T cells: bind MHC class I proteins expressed on all cells in the body and and survey cytosol (endoplasmic reticulum-derived) proteins.
2. $CD4^+$ helper T cells: bind MHC class II proteins that are expressed on dedicated antigen-presenting cells (e.g. macrophages 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 a 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 help eliminate the infection. As MHC class I is expressed by all cells in the body and constantly presents peptides onto the cell surface derived from inside the cell, CTLs circulate throughout the body and can survey ongoing protein expression in most cells and tissues, serving as a major adaptive immune effector mechanism against intracellular pathogens. Alternatively, $CD4^+$ T cells are activated by APCs to become helper T cells, and their major role occurs in the 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.

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 [4, 9]. Within germinal centers, B-cells that encode BCR will bind to antigen presented on follicular dendritic cells or antigen presented through other pathways. Following receptor endocytosis of the BCR in a bound complex to antigen, the B-cells digest the captured and display the resulting peptides via MHC class II on the B-cell surface. $CD4^+$ follicular helper T (T_{FH}) cells that have been activated by other APCs will in turn induce those antigen-specific B-cells to activate via TCR-peptide-MHC contacts, along with other with interactions between the B-cell co-receptor CD40 and the T cell-derived CD40 ligand (CD40L, or CD154). $CD4^+$ helper T-cells also 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 they also are activated and begin 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 [10]. Further rounds of positive selection in the light zone successively enhance antibody affinity to the

Table 1.2 Overview of $CD4^+$ T cell effector subsets. Each subset is induced to differentiate 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) [4]

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

antigen of interest [11]. Activated B-cells can undergo class-switch recombination, also mediated by AID, which alters the antibody isotype class via genomic deletion of IgM/IgD constant regions that switches the isotype to IgG, IgA, or IgE. Positive selection in the GC light zone induces B-cells to asymetrically proliferate into antigen-secreting plasmablasts, antigen-secreting plasma cells (which home to bone marrow for long-term persistence), and memory B-cells, which comprise the effector cells of the antibody response [12, 13].

Importantly, other pathways can lead to B-cell activation beyond germinal center reactions. B-cell maturation has been reported at the border of the splenic T cell zone and red pulp [14], and for some antigens (termed T-independent antigens) B-cell activation occurs 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 [4]. While some antigens are able to induce robust antibody responses in the absence of T-cell help, the vast majority of proteins will not 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 perform 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 serum antibody concentrations that dissipates in accordance with antibody half-life in serum (approximately two weeks) [15, 16]. 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 [17–23] and continuously secrete high levels of immunoglobulin (estimated at 10,000–20,000 molecules/cell-sec) [2]. Long-term serological memory is mediated by 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. Importantly, the kinetics of memory B-cell activation (or secondary responses) are much faster than the initial (or primary) responses 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.

Table 1.3 Selected characteristics of major 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

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

Germinal center B-cell activation results in multiple variants 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 [24]), and the entire collection of antibodies encoded by an individual resulting from a lifetime of immune responses comprises 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 [25, 26]), a thorough and comprehensive analysis of the antibody repertoire requires high-throughput means of sequence data collection and analysis. Beginning in 2008, the rapidly decreasing costs of gene sequencing for the first time permitted economic repertoire-scale, high-resolution DNA sequence analysis of B-cell populations (Fig. 1.2). These new experimental techniques, collectively known as high-throughput sequencing or next-generation DNA sequencing, are fundamentally transforming the major analytical methods for B- and T-cells as well as our understanding of adaptive immune responses.

1.3 High Throughput Antibody Sequencing

Tremendous advances in scale and decreases in costs of next-generation DNA sequencing technologies have dramatically accelerated the pace of biological research over the last eight years (Fig. 1.2). 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, 27]. Exact measurement of total human B-cell repertoire size is difficult to determine due to tissue sampling limitations (peripheral blood, bone marrow, and secondary lymphoid organs), combined with high-throughput sequencing error (typically $\sim$0.5% of sequenced bases [28, 29]) that contributes noise to secondary data analysis. However, lower-bound estimates of repertoire size are approximately 2×10 [6] unique B-cell clones (expressing distinct BCRs) in peripheral blood alone [25, 26]. 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, 4, 30, 31].

Standard immune repertoire high-throughput sequencing protocols begin with collection of 10^3–10^7 lymphocytes, followed by bulk 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 usually target V genes, while the 3' primers target either the J genes or constant regions (e.g. IgG, IgA, etc.) to perform sequence analysis of the entire antibody variable region with minimal coverage of the constant region (Fig. 1.1) [2]. (Some protocols omit V-gene-specific primers and incorporate RACE PCR to reduce multiplex PCR amplification bias [33]). 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 analyses are performed to quantitatively determine the composition of the input immune repertoire encoded by the cells originally isolated from experimental samples [2, 29, 34, 35]. High-throughput repertoire sequencing has been applied in a variety of applications ranging from characterization of the repertoire in healthy and disease states [36–39], to analysis of antibody-pathogen interactions [40–42], and for rapid antibody discovery [41, 43].

Despite the tremendous recent advances, all currently available techniques for antibody repertoire analysis have one severe limitation: high-throughput antibody sequencing is 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, 40, 44, 45]. Without the ability to sequence paired heavy and light chain sequences at high throughput with single-cell resolution, the full antibody clonotype (both heavy and light chains) cannot be resolved on a repertoire scale, nor can the resulting antibody sequences be expressed to test for function, nor can the antibody proteins be modeled computationally.

1.4 Next Generation Antibody Sequencing Data Analysis

The rapid growth of antibody sequencing data has also fueled rapid growth in our capabilities to analyze and interpret antibody sequence datasets. In particular, the errors introduced by Next Generation sequencing technologies greatly complicate efforts to analyze B-cell sequence data because B-cells are known undergo somatic hypermutation, creating two B-cell clones that differ by only a single nucleotide substitution in the variable region. Given that the errors (or noise) introduced by NextGen sequencing are greater than the single-nucleotide changes that can be generated via SHM, it can be very difficult to confidently assess whether two closely related antibody sequences derive from sequence error or from two true somatic variants of the same clonal lineage.

NextGen sequencing technologies exhibit an average error rate of approximately 0.5% [28, 46, 47], which consist of a mix of single base pair substitution errors and insertion/deletion errors. Different NextGen platforms exhibit distinct error rate patterns. For example, the widely-used Roche 454 sequencing platform, which is based on real-time observation of base incorporation and with pauses in between the addition of A/T/C/G bases, has difficulty determining the length of homopolymer stretches of DNA because homopolymer runs incorporate in such

rapid succession that it is difficult for the camera and software to determine the exact number of bases incorporated. Thus, 454 (along with most other real time sequencing platforms such as those sold by Pacific Biosciences) has a high rate of homopolymer insertions and deletions (indels). Indels are often trivial to correct manually, but can prove especially problematic for high-throughput analyses because they introduce frameshift errors to the amino acid translations. 454 also maintains a consistent error rate across the course of a sequence read. In contrast, the Illumina sequencing platform relies on reversible terminator chemistry that permits the observation of each base incorporation event, similar to a reversible Sanger sequencing technology. For Illumina data the probability of insertion and deletion events is very low compared to 454 data, whereas the single base substitution error rate is comparably higher. In addition, due to the accumulation of imperfections such as incomplete removal of the reversible terminator across cycles, the read quality of Illumina sequences degrades as the sequence read progresses. For example in the MiSeq platform, the first $\sim$150 bases of the read are of comparably high quality, whereas quality rapidly degrades in the final 100 bases of a 250 bp-length read. Thus, Illumina sequencing technologies require error correction that takes into account not only the overall average error rates of Illumina sequences, but also the characteristics of error introduction over the course of different sections in the sequence read.

High-confidence methods do exist for deconvoluting sequence error using molecular barcodes incorporated into reverse transcription primers [48–50]. These methods utilize a unique barcode region that is incorporated during first-strand cDNA synthesis and that is maintained throughout the following rounds of PCR. During sequence data analysis, the barcodes resulting from each RT event can be pooled, counted, and a consensus sequence, or average of all the different reads, can be generated for each barcoded RT event with a sufficient number of observations. This approach can identify the original 1st-strand cDNA sequences with very high accuracy, given sufficient coverage of the individual barcodes (minimum of 3 reads per barcode for establishing consensus.) However, in practice these methods have a limited throughput because the high number of RT events and skewing of repertoire distributions via PCR results in a low number of sequences that pass the minimum observation threshold for establishing a consensus [48–50]. Additionally, these techniques cannot deconvolute errors introduced by reverse transcription, which is an important source of sequence errors in immune repertoire sequence data.

Each antibody sequencing platform reports the estimated confidence of a particular base being correct or incorrect in the form of a quality score, similar to the Phred scores reported in DNA sequencing data via capillary electrophoresis. An essential first step in any NextGen data analysis pipeline is to filter sequence data by quality scores, which removes many of the most error-prone sequence reads in a NextGen dataset. However, it is critical to remember that sequence quality filtering is inadequate for sequence error mitigation on its own for several reasons. First, the quality score is a probabilistic measurement, which measures only the likelihood that a given base is right or wrong as calculated based on raw data collected in the sequencing platform, often using signal:noise ratios or other metrics for each

sequenced base. Filtering out the bases with the highest probability of containing errors is not the same as removing all errors from the data. Second, quality scores report only the likelihood of an error introduced by the *sequencing* process. Another major source of error introduction occurs as a result of mismatched base incorporation during reverse transcription, which cannot be identified via quality score sequence analysis as the error was introduced prior to DNA sequencing.

A very powerful tool for identifying errors is comparative frequency analysis. NextGen sequencing identifies millions of DNA sequence reads, and because the per-base error rates are low and (to a first approximation) random, the probability of a correct base sequenced is high. Thus, by investigating a number of highly similar cDNA sequences, it becomes clear that the bases observed most frequently are the highest confidence for corresponding to the true initial RNA sequence. Minimum observation cutoffs are very helpful in this regard—if a sequence has been observed only once, then it is much more likely that that particular sequence contains a sequence error than a sequence observed many times. A related technique is the establishment of consensus sequences. In this process, multiple sequence reads are aligned and "averaged" such that a final consensus sequence is constructed from a minimum of 3 NextGen sequence reads. Both minimum observation cutoffs and consensus sequence generation are extremely helpful tools to minimize the effects of sequence error in NextGen antibody sequencing datasets.

Another critical tool for Nextgen antibody sequence analysis is clustering, or the grouping of similar antibody variants by sequence similarity. Several clustering techniques and platforms have been reported [51–53]. Clustering groups highly similar sequences based on defined cutoffs and other user-defined parameters, and because sequence errors are introduced at low-levels, clustering can accurately identify antibody clones, clonotypes, and consensus sequences. However, one caveat to clustering is that it removes most low-level somatic hypermutation of a given antibody lineage, which is an important feature of any clustering analysis. Because antibody sequences derive from common genes and are highly similar, it is recommended in almost all cases to perform clustering of a high-variation subset of the full sequence (e.g., the CDR3) rather than clustering complete antibody sequences. This concentrated clustering analysis uses the highest variation region of the antibody sequence to avoid collapsing genetically similar clones with a distinct origin (i.e., different CDR3 regions) together into the same cluster.

Many different iterations of the above analysis and error correction techniques can lead to robust antibody sequence data analysis. Different techniques may be required for distinct experimental sample preparation protocols, sequencing platforms, and intended applications of the experiment. An important step in the evaluation of any bioinformatic pipeline is to evaluate performance using spike-in controls of a known sequence. These spike-in controls can use immortalized B-cell lines that express a known antibody sequence, for example the human clones IM-9 and ARH-77, and the mouse clones MOPC-315 and MOPC-21, and sometimes spike-in RNA or cDNA of a known sequence. It is important to verify that a single spike-in sequence can be collapsed to a single antibody sequence by any

experimental and bioinformatic pipeline as a test of that pipeline's ablity to identify *bona fide* antibody sequences.

For our work here and verified using spike-in controls from immortalized B-cell lines, we have found that a pipeline of quality filtering, CDR3 extraction and analysis (as the CDR3 has the highest per base variation across antibody clones), compilation by CDR3 sequence and V(D)J gene assignments, removal of single-read paired CDR-H3:CDR-L3 variants, and CDR-H3 clustering leads to a robust repertoire where most somatic variants are removed for antibody clones and where immortalized cell spike-in controls collapsed to a single sequence variant. While no methods can achieve perfection given the limitations of high NextGen error rate profiles and the unique characteristics of B-cells to undergo somatic hypermutation, the methods reported here present a highly useful compromise that allows for the annotation of antibody clusters (or lineages/clonotypes) contained in the data and provides a workable platform for robust antibody repertoire analyses.

1.5 Monoclonal Antibody Discovery Technologies

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 [1]. 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 be purified from hybridoma culture supernatant. Hybridoma methods continue to have tremendous impact. The basic techniques were outlined around 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 rapidly and with high reliability. In particular, recent key methods include enhancing the efficiency of hybridoma generation with human cells [54] and humanization of mouse antibodies or the development of human transgenic [55–58] or humanized [59–61] mouse models to reduce immunogenicity of the resulting

monoclonal antibody therapeutics in human patients. Several protocols have permitted faster and more economical hybridoma screening to accelerate the discovery of human or humanized monoclonal antibodies [62–64]. 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. Despite tremendous advances in tried-and-true hybridoma technologies, mAb discovery using hybridomas remains time-consuming and expensive due to the single-cell limiting dilution needed and the time required hybridomas to expand from a single cell to a cell population (several weeks). The large number of culture supernatant screens required also make hybridoma mAb discovery a resource-intensive experimental toolkit.

An important alternative technology to hybridoma mAb discovery is antibody isolation via in vitro screening of combinatorial libraries. The most widely used combinatorial library discovery platform technology uses scFv 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 (along with mutations acquired in each round). Finally after several rounds of phage panning a high-affinity monoclonal antibody can be isolated [65–71]. 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 therefor phage panning has more limited applications for therapeutic antibody discovery [72].

A more recent method for monoclonal antibody discovery has applied high-throughput sequencing of B-cell receptors to identify antibodies of interest [40, 43, 73]. Next Generation sequencing is revolutionizing our ability to analyze and interpret antibody repertoires because it 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 [33, 74–77]. 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 [74, 75, 78]. High-throughput and proteomic techniques for antibody discovery will become more widely used in the coming years as DNA sequencing and protein mass spectrometry costs continue to decrease.

Unfortunately the heavy and light chain pairing information is irreversibly lost during conventional high-throughput antibody sequencing [40, 44, 45], and this inability to deconvolute paired heavy and light sequences using NextGen sequencing has severely complicated efforts to rapidly discover new antibodies and applications of high-throughput sequencing for antibody discovery. Some progress has been made toward more rapid antibody discovery using NextGen sequencing. Important early work demonstrated that frequency-based pairing of highly enriched cell populations can lead to productive antibodies [43]. Additionally, phylogenetic algorithms can be effective for inferring the heavy:light pairing of highly mutated antibody lineages in some cases [42]. New antibody variants identified by heavy chain-only antibody sequencing can also provide important insights regarding antibody lineage development [40, 41, 73, 79, 80]. In addition, such sequences can be used in combination with single-cell RT-PCR data to generate new antibody variants and test antibody performance in vitro [45].

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.

1.6 Single-Cell Sequencing Techniques

As mentioned above, existing immune repertoire high throughput sequenc-ing 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 [40, 44, 45]. Due to 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 [81–85]. As a result of 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 [7, 45, 86–88]. Once the sequence of a B-cell has been isolated it can then be cloned into bacteria and tested for antigen binding to a protein of interest [45, 86, 87, 89], 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 be achieved via linkage of heavy and light chains in the RT-PCR, thereby reducing cloning steps by a factor of two [86, 90, 91].

Single-cell sequencing in the small volumes required for manageable high-throughput experiments ($\sim 10^2$ pL per reaction) 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 [83]. Incomplete cell lysis and RNA degradation during thermal cell lysis can further reduce yield of linked cDNA products using one-pot reactions. Furthermore, the 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. 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 workflows [2, 40, 44, 45].

1.7 Synopsis

This dissertation directly addresses the aforementioned 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*) [92], 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 5×10^4 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*) [93] 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 3×10^6 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 the antibody development, maturation, and selection processes across several human donors.

References

1. Kohler G, Milstein C (1975) Continuous cultures of fused cells secreting antibody of predefined specificity. Nat 256:495–497
2. Georgiou G et al (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat Biotechnol 32:158–168
3. Berg JM et al (2002) Biochemistry. Freeman, WH
4. Murphy K, Travers P, Walport M, Janeway C (2012) Janeway's immunobiology. Garland Science, US
5. Hess J et al (2001) Induction of pre-B cell proliferation after de novo synthesis of the pre-B cell receptor. Proc Natl Acad Sci 98:1745–1750
6. Lee J, Monson NL, Lipsky PE (2000) The VλJλ repertoire in human fetal spleen: evidence for positive selection and extensive receptor editing. J Immunol 165:6322–6333
7. Wardemann H et al (2003) Predominant autoantibody production by early human B-cell precursors. Sci 301:1374–1377
8. Qi Q et al (2014) Diversity and clonal selection in the human T-cell repertoire. Proc Natl Acad Sci 111:13139–13144
9. Kuppers R, Zhao M, Hansmann ML, Rajewsky K (1993) Tracing B-cell development in human germinal centers by molecular analysis of single cells picked from histological sections. EMBO J 12:4955–4967
10. Fernández D et al (2013) The proto-oncogene c-myc regulates antibody secretion and ig class switch recombination. J Immunol 190:6135–6144
11. Gitlin AD, Shulman Z, Nussenzweig MC (2014) Clonal selection in the germinal centre by regulated proliferation and hypermutation. Nat 509:637–640
12. Barnett BE et al (2012) Asymmetric B-cell division in the germinal center reaction. Sci 335:342–344
13. Klein U, Dalla-Favera R (2008) Germinal centres: role in B-cell physiology and malignancy. Nat Rev Immunol 8:22–33
14. William J, Euler C, Christensen S, Shlomchik MJ (2002) Evolution of autoantibody responses via somatic hypermutation outside of germinal centers. Sci 297:2066–2070
15. Hinton PR et al (2006) An engineered human IgG1 antibody with longer serum half-life. J Immunol 176:346–356
16. Kyu SY et al (2009) 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
17. Manz RA, Löhning M, Cassese G, Thiel A, Radbruch A (1998) Survival of long-lived plasma cells is independent of antigen. Int Immunol 10:1703–1711
18. O'Connor BP, Cascalho M, Noelle RJ (2002) Short-lived and long-lived bone marrow plasma cells are derived from a novel precursor population. J Exp Med 195:737–745
19. Amanna IJ, Carlson NE, Slifka MK (2007) Duration of humoral immunity to common viral and vaccine antigens. N Engl J Med 357:1903–1915
20. Amanna IJ, Slifka MK (2010) Mechanisms that determine plasma cell lifespan and the duration of humoral immunity. Immunol Rev 236:125–138
21. Dorner T, et al (2011) Long-lived autoreactive plasma cells drive persistent autoimmune inflammation. Nat Rev Rheumatol 7:170

22. Mahevas M, Michel M, Weill J.-C, Reynaud C-A (2013) Long-lived plasma cells in autoimmunity: lessons from B-Cell depleting therapy. Front Immunol 4
23. Halliley JL et al (2015) Long-lived plasma cells are contained within the CD19− CD38 hi CD138+ subset in human bone marrow. Immun 43:132–145
24. Yu X et al (2008) Neutralizing antibodies derived from the B-cells of 1918 influenza pandemic survivors. Nat 455:532–536
25. Boyd SD, et al (2009) Measurement and clinical monitoring of human lymphocyte clonality by massively parallel V-D-J pyrosequencing. Sci Transl Med 1:12ra23
26. Arnaout R et al (2011) High-resolution description of antibody heavy-chain repertoires in humans. PLoS ONE 6:e22365
27. Warren EH, Matsen FA, Chou J (2013) High-throughput sequencing of B- and T-lymphocyte antigen receptors in hematology. Blood 122:19–22
28. Loman NJ et al (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotech 30:434–439
29. Bashford-Rogers RJ et al (2014) Capturing needles in haystacks: a comparison of B-cell receptor sequencing methods. BMC Immunol 15:29
30. Schroeder HW Jr (2006) Similarity and divergence in the development and expression of the mouse and human antibody repertoires. Dev Comp Immunol 30:119–135
31. Apostoaei AI, Trabalka JR (2012) Review, synthesis, and application of information on the human lymphatic system to radiation dosimetry for chronic lymphocytic leukemia. SENES Oak Ridge, Inc., Tennessee
32. Wetterstrand KA. DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP). Available at: www.genome.gov/sequencingcostsdata. Accessed 27 Feb 2017
33. Wine Y et al (2013) Molecular deconvolution of the monoclonal antibodies that comprise the polyclonal serum response. Proc Natl Acad Sci 110:2993–2998
34. Menzel U et al (2014) Comprehensive evaluation and optimization of amplicon library preparation methods for high-throughput antibody sequencing. PLoS ONE 9:e96727
35. Greiff V et al (2014) Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. BMC Immunol 15:40
36. Glanville J et al (2009) Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. Proc Natl Acad Sci USA 106:20216–20221
37. Wu Y-CB, Kipling D, Dunn-Walters DK (2012) Age-related changes in human peripheral blood IGH repertoire following vaccination. Front Immunol 3
38. Schoettler N, Ni D, Weigert M (2012) 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
39. Hoi KH, Ippolito GC (2013) Intrinsic bias and public rearrangements in the human immunoglobulin V[lambda] light chain repertoire. Genes Immun 14:271–276
40. Wu X et al (2011) Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. Sci 333:1593–1602
41. Zhu J et al (2013a) De novo identification of VRC01 class HIV-1–neutralizing antibodies by next-generation sequencing of B-cell transcripts. Proc Natl Acad Sci 110:E4088–E4097
42. Zhu J et al (2013b) Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. Proc Natl Acad Sci 110:6470–6475
43. Reddy ST et al (2010) Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. Nat Biotechnol 28:965–969
44. Fischer N (2011) Sequencing antibody repertoires: the next generation. MAbs 3:17–20
45. Wilson PC, Andrews SF (2012) Tools to therapeutically harness the human antibody response. Nat Rev Immunol 12:709–719
46. Prabakaran P, Streaker E, Chen W, Dimitrov DS (2011) 454 antibody sequencing-error characterization and correction. BMC Res. Notes 4:404

47. Schirmer M et al (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res gku1341. doi:10.1093/nar/gku1341
48. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR (2013) Genetic measurement of memory B-cell recall using antibody repertoire sequencing. Proc Natl Acad Sci 110:13463–13468
49. Shugay M et al (2014) Towards error-free profiling of immune repertoires. Nat Methods 11:653–655
50. Khan TA et al (2016) Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. Sci Adv 2:e1501371
51. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinform 26:2460–2461
52. Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinform 28:3150–3152
53. Li W, Fu L, Niu B, Wu S, Wooley J (2012) Ultrafast clustering algorithms for metagenomic sequence analysis. Brief Bioinform 13:656–668
54. Yu X, McGraw PA, House FS, Crowe JE Jr (2008) An optimized electrofusion-based protocol for generating virus-specific human monoclonal antibodies. J Immunol Methods 336:142–151
55. Lonberg N et al (1994) Antigen-specific human antibodies from mice comprising four distinct genetic modifications. Nat 368:856–859
56. Mendez MJ et al (1997) Functional transplant of megabase human immunoglobulin loci recapitulates human antibody response in mice. Nat Genet 15:146–156
57. Lonberg N (2005) Human antibodies from transgenic animals. Nat Biotechnol 23:1117–1125
58. Murphy AJ et al (2014) Mice with megabase humanization of their immunoglobulin genes generate antibodies as efficiently as normal mice. Proc Natl Acad Sci 111:5153–5158
59. McCune JM et al (1988) The SCID-hu mouse: murine model for the analysis of human hematolymphoid differentiation and function. Sci 241:1632–1639
60. Hiramatsu H et al (2003) Complete reconstitution of human lymphocytes from cord blood CD34+ cells using the NOD/SCID/γcnull mice model. Blood 102:873–880
61. Ippolito GC et al (2012) 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
62. Rieger M, Cervino C, Sauceda JC, Niessner R, Knopp D (2009) Efficient hybridoma screening technique using capture antibody based microarrays. Anal Chem 81:2373–2377
63. Ogunniyi A, Story C, Papa E, Guillen E, Love J (2009) Screening individual hybridomas by microengraving to discover monoclonal antibodies. Nat Protoc 4:767–782
64. Debs BE, Utharala R, Balyasnikova IV, Griffiths AD, Merten CA (2012) Functional single-cell hybridoma screening using droplet-based microfluidics. Proc Natl Acad Sci 109:11570–11575
65. Smith G (1985) Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Sci 228:1315–1317
66. McCafferty J, Griffiths AD, Winter G, Chiswell DJ (1990) Phage antibodies: filamentous phage displaying antibody variable domains. Nat 348:552–554
67. Marks JD et al (1992) Bypassing immunization–building high-affinity human antibodies by chain shuffling. Bio-Technol 10:779–783
68. Griffiths AD et al (1994) Isolation of high-affinity human antibodies directly from large synthetic repertoires. EMBO J 13:3245–3260
69. Sblattero D, Bradbury A (2000) Exploiting recombination in single bacteria to make large phage antibody libraries. Nat Biotechnol 18:75–80
70. Mazor Y, Van Blarcom T, Carroll S, Georgiou G (2010) 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
71. D'Angelo S et al (2014) From deep sequencing to actual clones. Protein Eng Des Sel 27:301–307

72. Chan CEZ, Lim, APC, MacAry, PA, Hanson BJ (2014) The role of phage display in therapeutic antibody discovery. Int Immunol dxu082. doi:10.1093/intimm/dxu082
73. Zhou T et al (2013) Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-Class antibodies. Immun 39:245–258
74. Lavinder JJ et al (2014) Identification and characterization of the constituent human serum antibodies elicited by vaccination. Proc Natl Acad Sci 111:2259–2264
75. Boutz DR et al (2014) Proteomic identification of monoclonal antibodies from serum. Anal Chem 86:4758–4766
76. Sato S et al (2012) Proteomics-directed cloning of circulating antiviral human monoclonal antibodies. Nat Biotech 30:1039–1043
77. Cheung WC et al (2012) A proteomics approach for the identification and cloning of monoclonal antibodies from serum. Nat Biotech 30:447–452
78. Lee J et al (2016) Quantitative, molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. Nat Med, Accepted
79. Doria-Rose NA et al (2014) Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. Nature 509:55–62
80. Zhu J, et al (2012) Somatic populations of PGT135-137 HIV-1-neutralizing antibodies identified by 454 pyrosequencing and bioinformatics. Front Microbiol 3
81. Marcus JS, Anderson WF, Quake SR (2006) Microfluidic single-cell mRNA isolation and analysis. Anal Chem 78:3084–3089
82. Toriello NM et al (2008) Integrated microfluidic bioprocessor for single-cell gene expression analysis. Proc Natl Acad Sci USA 105:20173–20178
83. White AK et al (2011) High-throughput microfluidic single-cell RT-qPCR. Proc Natl Acad Sci USA 108:13999–14004
84. Furutani S, Nagai H, Takamura Y, Aoyama Y, Kubo I (2012) Detection of expressed gene in isolated single cells in microchambers by a novel hot cell-direct RT-PCR method. Anal 137:2951–2957
85. Turchaninova MA et al (2013) Pairing of T-cell receptor chains via emulsion PCR. Eur J Immunol 43:2507–2515
86. Meijer P et al (2006) Isolation of human antibody repertoires with preservation of the natural heavy and light chain pairing. J Mol Biol 358:764–772
87. Smith K et al (2009) Rapid generation of fully human monoclonal antibodies specific to a vaccinating antigen. Nat Protoc 4:372–384
88. Frölich D et al (2010) Secondary immunization generates clonally related antigen-specific plasma cells and memory B-cells. J Immunol 185:3103–3110
89. Smith K et al (2013) Fully human monoclonal antibodies from antibody secreting cells after vaccination with Pneumovax®23 are serotype specific and facilitate opsonophagocytosis. Immunobiol 218:745–754
90. Poulsen TR, Meijer P-J, Jensen A, Nielsen LS, Andersen PS (2007) Kinetic, affinity, and diversity limits of human polyclonal antibody responses against tetanus toxoid. J Immunol 179:3841–3850
91. Meijer P-J, Nielsen LS, Lantto J, Jensen A (2009) Human antibody repertoires. In: Dimitrov AS (ed), Therapeutic antibodies: Methods and protocols. New York, USA: Humana Press, 525:261–277
92. DeKosky BJ et al (2013) High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nat Biotech 31:166–169
93. DeKosky BJ et al (2015) In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. Nat Med 21:86–91

Chapter 2
High-Throughput Sequencing of the Paired Human Immunoglobulin Heavy and Light Chain Repertoire

2.1 Rationale and Supporting Information

Currently existing immune repertoire sequencing technologies yield data on only one of the two chains of immune receptors [1–3]. Sequence analysis of VH:VL pairs is therefore currently performed by microtiter-well sorting of individual B cells followed by single-cell RT-PCR (scRT-PCR) and Sanger sequencing [3–10]; 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 [6–9]. Microfluidic methods for RT-PCR and the sequencing of two or more genes (for example using the Fluidigm platform [11]), have 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.

Several prior studies analyzed single cells by first isolating the cells into high-density microwell arrays [12–16]. 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 [17]. We reasoned that a system capable of selectively capturing 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. in the presence of dodecyl

The whole content of this chapter was initially published in DeKosky, B.J. et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nat. Biotechnol. 31, 166–169 (2013), and it is reproduced here. 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.

B. DeKosky, *Decoding the Antibody Repertoire*,
Springer Theses, DOI 10.1007/978-3-319-58518-5_2

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 an mRNA capture agent by utilizing poly(dT) oligonucleotides conjugated to the magnetic beads that bind to the polyadenylated mRNA tail. Magnetic beads are 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 [5]), we could sequence paired VH and VL chains using standard next-generation sequencing protocols.

2.2 Methodology

As shown in Fig. 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 $\geq$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 min. 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 (Fig. A.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 $\leq$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 $\leq$300 cells per run [17]. PCR amplification as outlined in Fig. A.1 generates an $\sim$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 2×250 Illumina™ 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 2×250 bp sequencing

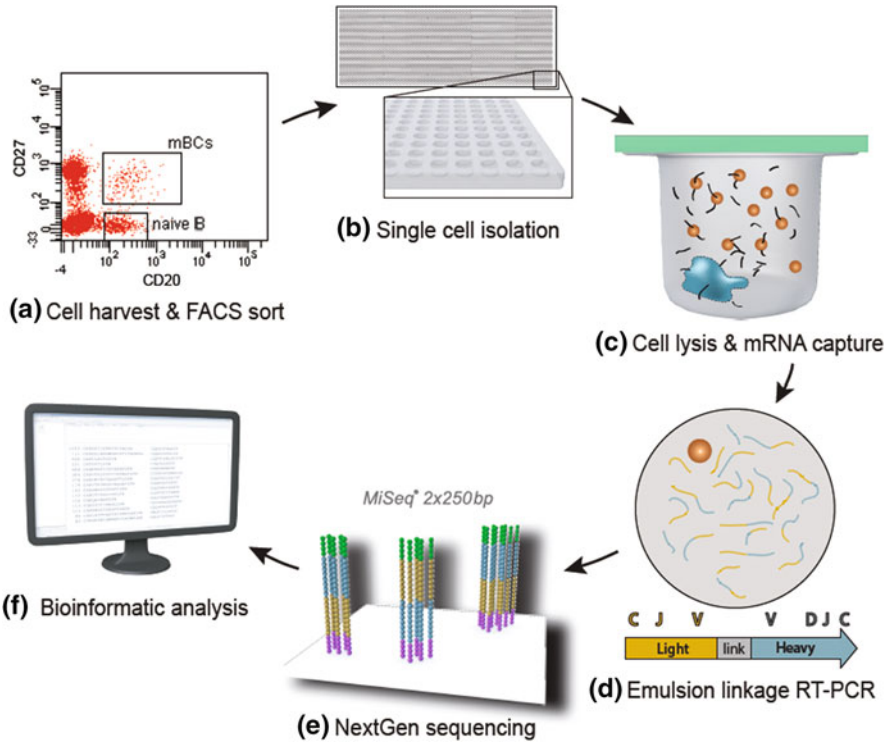

Fig. 2.1 Overview of the high throughput methodology for paired VH:VL antibody repertoire analysis. **a** B-cell populations are sorted for desired phenotype (mBCs = memory B cells, naive B = naive B cells). **b** Single cells are isolated by random settling into 125 pL wells (56 μm diameter) printed in polydimethylsiloxane (PDMS) slides the size of a standard microscope slide (1.7×10^5 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 (Fig. A.1). **e** Next Generation sequencing is performed to sequence the linked strands. **f** Bioinformatic processing is used to analyze the paired VH:VL repertoire

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.

2.3 Results

We employed the methodology of Fig. 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 2×250 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 2716 unique pairs were thus identified (Table A.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 (Fig. 2.2a). A second IgG^+ repertoire analysis was performed using B cells from another anonymous individual; this analysis identified 2248 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 [18, 19]. 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) (Fig. 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

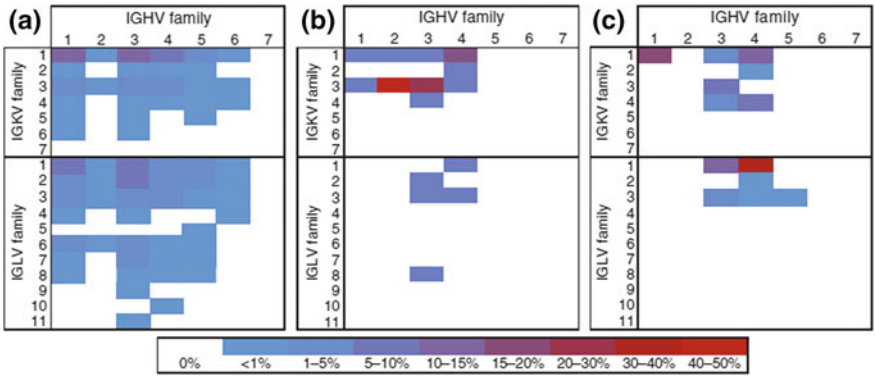

Fig. 2.2 VH:VL gene family usage of unique CDR-H3:CDR-L3 pairs identified via high-throughput sequencing of cell populations from three different individuals in separate experiments using the workflow presented in Fig. 2.1: **a** healthy donor peripheral IgG^+ B cells (n = 2716 unique CDR3 pairs), **b** 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** 8000 twice frozen/thawed memory B cells

and Vλ2). However, putative differences in IgG^+ VH:VL germline pairing frequencies between the two individuals were also evident.

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 [7]. 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 (1.7×10^5 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 A.1). We expressed ten of the identified VH:VL pairs as IgG proteins in HEK293 K 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 [20, 21].

Finally, we compared the VH:VL pairings identified using this high-throughput approach to those identified using the established single cell sorting method [6, 22]; 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 [6]. 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

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 Fig. 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

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

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 (3.4×10^5 total wells) were used, and the sample was spiked with IM-9 cells to confirm pairing accuracy (Table A.1). A total of 240 unique CDR-H3:CDR-L3 pairs were recovered (Fig. 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 A.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 A.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.

2.4 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 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 VH and VL 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 VH and VL 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 A.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 VH:VL pairing efficiency in memory B-cell populations is relatively modest (Fig. 2.2 and Table A.2), efforts to further improve efficiency are under way.

The high-throughput VH:VL 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 VH:VL 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 2716 unique VH:VL pairs from a sample of IgG + peripheral B cells (Fig. 2.2a) was completed by a single researcher in 10 h and cost $550. Analysis of 2700 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 deter-mined 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 A.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 A.5–A.7). In some chronic infections (e.g., HIV), constant anti-gen 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 using primers that anneal to the leader peptide [23].

Finally, we note that the analysis reported here focused on the light chain that was the dominant light chain paired with a particular VH. 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 VH:VL 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 VH1:VL2 and VH2:VL1 in addition to VH1:VL1 and VH2:VL2.

2.5 Methods

Methods and associated references are reported in a published version of this thesis chapter [24].

References

1. Wu X et al (2011) Focused EVOLUTION of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. Science 333:1593–1602
2. Fischer N (2011) Sequencing antibody repertoires: the next generation. MAbs 3:17–20
3. Wilson PC, Andrews SF (2012) Tools to therapeutically harness the human antibody response. Nat Rev Immunol 12:709–719
4. Wardemann H et al (2003) Predominant autoantibody production by early human B cell precursors. Science 301:1374–1377
5. Meijer P et al (2006) Isolation of human antibody repertoires with preservation of the natural heavy and light chain pairing. J Mol Biol 358:764–772
6. Smith K et al (2009) Rapid generation of fully human monoclonal antibodies specific to a vaccinating antigen. Nat Protoc 4:372–384
7. Frölich D et al (2010) Secondary immunization generates clonally related antigen-specific plasma cells and memory B cells. J Immunol 185:3103–3110
8. Tanaka Y et al (2010) 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
9. Scheid JF et al (2011) Differential regulation of self-reactivity discriminates between IgG(+) human circulating memory B cells and bone marrow plasma cells. Proc Natl Acad Sci USA 108:18044–18048
10. Li G-M et al (2012) Pandemic H1N1 influenza vaccine induces a recall response in humans that favors broadly cross-reactive memory B cells. Proc Natl Acad Sci USA 109:9047–9052
11. Sanchez-Freire V, Ebert AD, Kalisky T, Quake SR, Wu JC (2012) Microfluidic single-cell real-time PCR for comparative analysis of gene expression patterns. Nat Protoc 7:829–838
12. Ogunniyi A, Story C, Papa E, Guillen E, Love J (2009) Screening individual hybridomas by microengraving to discover monoclonal antibodies. Nat Protoc 4:767–782
13. Lindström S, Hammond M, Brismar H, Andersson-Svahn H, Ahmadian A (2009) PCR amplification and genetic analysis in a microwell cell culturing chip. Lab Chip 9:3465–3471
14. Tokimitsu Y et al (2007) Single lymphocyte analysis with a microwell array chip. Cytometry A 71:1003–1010
15. Yamamura S et al (2005) Single-cell microarray for analyzing cellular response. Anal Chem 77:8050–8056
16. Tajiri K et al (2007) Cell microarray analysis of antigen specific B cells: single cell analysis of antigen receptor expression and specificity. Cytometry A 71:961–967
17. White AK et al (2011) High-throughput microfluidic single-cell RT-qPCR. Proc Natl Acad Sci USA 108:13999–14004
18. Ippolito GC et al (2012) 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
19. Glanville J et al (2011) Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. Proc Natl Acad Sci USA 108:20066–20071
20. Sato S et al (2012) Proteomics-directed cloning of circulating antiviral human monoclonal antibodies. Nat Biotech 30:1039–1043
21. Cheung WC et al (2012) A proteomics approach for the identification and cloning of monoclonal antibodies from serum. Nat Biotech 30:447–452
22. Wrammert J et al (2008) Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. Nature 453:667–671
23. Scheid JF et al (2011) Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. Science 333:1633–1637
24. DeKosky BJ et al (2013) High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nat Biotech 31:166–169

Chapter 3
In-Depth Determination and Analysis of the Human Paired Heavy and Light Chain Antibody Repertoire

3.1 Introduction

The determination of immune receptor repertoires using high throughput (NextGen) DNA sequencing has rapidly become an indispensable tool for the understanding of adaptive immunity, antibody discovery and in clinical practice [1–3]. 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 to moderate numbers of B cells (10^4–10^5) [4], far smaller than the ~ 0.7–4×10^6 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 [5], antibody discovery [6, 7], and for addressing a host of important questions related to the shaping of the antibody repertoire [2, 8–13].

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 [14–17]. 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 8×10^6 PBMC) [17]. This is likely because one-pot emulsions have a high

The whole content of this chapter was initially published in 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), and it is reproduced here. 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.

B. DeKosky, *Decoding the Antibody Repertoire*,
Springer Theses, DOI 10.1007/978-3-319-58518-5_3

degree of droplet size dispersity and since the RT reaction is inhibited in volumes <5 nL [15] 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 [18, 19], 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 (Fig. 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 (Fig. B.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 (Fig. B.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.

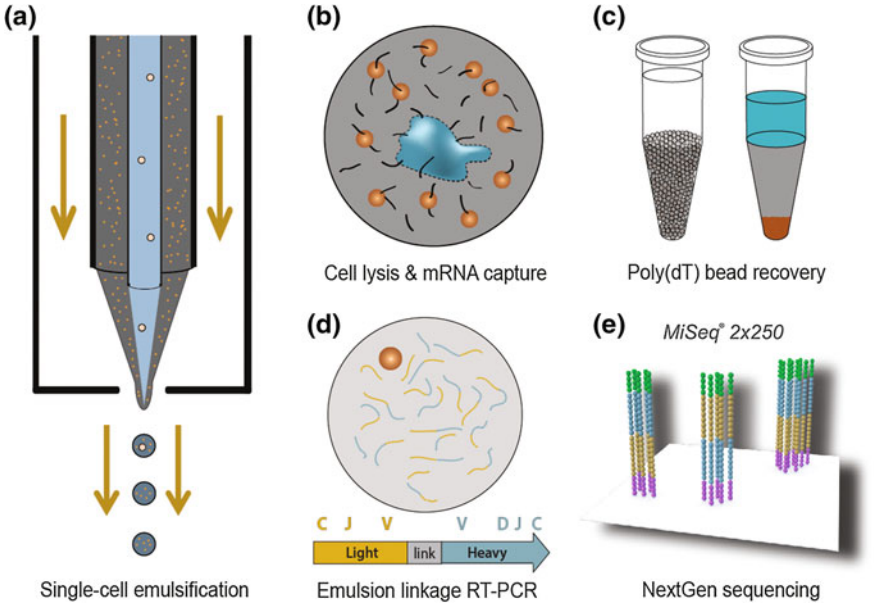

Fig. 3.1 Technical workflow for ultra-high throughput VH:VL sequencing from single B cells. **a** An axisymmetric flow-focusing nozzle isolated single cells and poly(dT) magnetic beads into emulsions of predictable size distributions. An aqueous solution of cells in PBS (center, *blue/pink circles*) 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. B.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 [20]. 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 min emulsification for each replicate) and processed as

shown in Fig. 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 (Fig. 3.1b), then the emulsion was broken chemically (Fig. 3.1c), beads were re-emulsified, and overlap extension RT-PCR was performed to generate linked VH:VL amplicons (Fig. 3.1d). The resulting cDNAs were amplified by nested PCR to generate an $\sim$850 bp VH:VL product for NextGen sequencing by Illumina MiSeq 2 $\times$ 250 or 2 $\times$ 300. 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 [4].

Sequence data were processed by read quality filtering, CDR-H3 clustering, VH:VL pairing, and selection for paired VH:VLs with $\geq$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% (Fig. 3.2, Table 3.1, Fig. B.3, see Methods). The ratio of VH:VL clusters to input cells observed (typically between 1:10 and 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

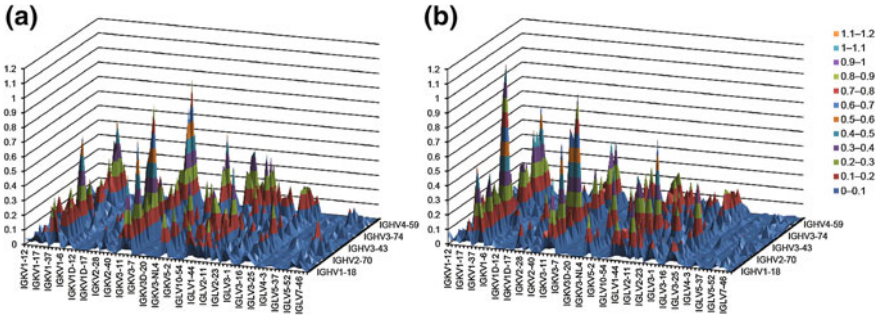

Fig. 3.2 Heavy:light V-gene pairing landscape of $CD3^{-}CD19^{+}CD20^{+}CD27^{+}$ peripheral memory B cells in two healthy human donors. V genes are plotted in alphanumeric order; height indicates percentage representation among VH:VL clusters. **a** Donor 1 (n = 129,097). **b** Donor 2 (n = 53,679).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

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

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	4267	4170	98.9

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, Figs. B.4 and B.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 Fig. 3.1, and native pairings were identified for 11/11 antibodies (Table B.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 ($\sim$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 (2604 correct ARH-77 VH:VL reads vs. 27 reads for the top ARH-77 VH paired with an incorrect VL, Table B.2).

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 2×250 or 2×300. Alternatively, the long-read Pacific Biosciences (PacBio) sequencing platform can be used to obtain the complete $\sim$850 bp 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 [7] and identified six variants of VRC26-class HIV broadly neutralizing antibodies within the VH:VL repertoire (Figs. B.6 and B.7).

3.2.3 Promiscuous and 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 [21]. 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 [22, 23]. 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 Fig. B.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 $\sim$1 per 1000 VH:VL clusters) [24, 25] 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 (Fig. 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, 25].

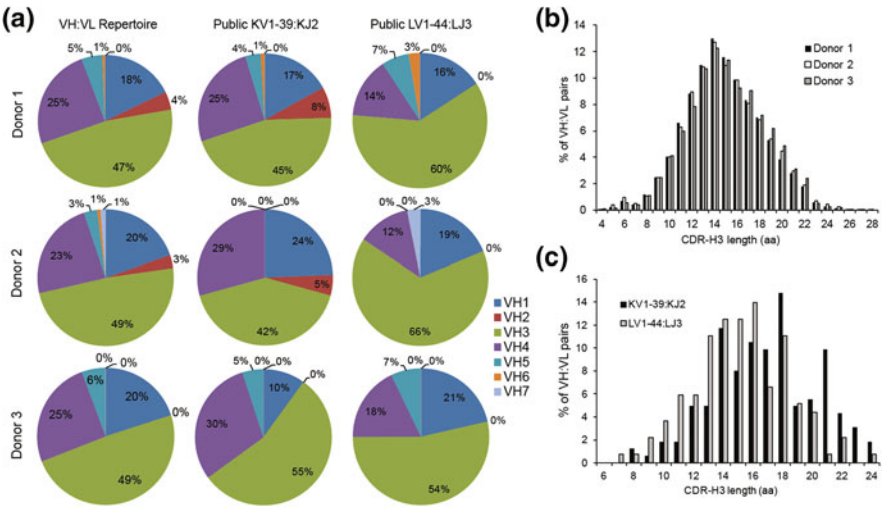

Fig. 3.3 **a** VH gene family utilization in: *left* total paired VH:VL repertoires (Donor 1 n = 129,097, Donor 2 n = 53,679, Donor 3 n = 15,372), *center* heavy chains paired with a representative highly-ranked public and promiscuous VL observed in all three donors (*KV1-39: KJ2* 9 aa CDR-L3, tgtcaacagagttacagtaccccgtacactttt; 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 [26–30]. Almost 20 years ago A. Lanzavecchia and coworkers used FACS sorting of cells expressing both κ and λ immunoglobulin proteins on their cell surface ($sIg\kappa^+/sIg\lambda^+$, denoting surface-expression of both Igκ and Igλ) followed by EBV immortalization to show that $sIg\kappa^+/sIg\lambda^+$ allelic inclusion occurs in 0.2–0.5% of human memory B cells [31]. However, the inability to sort dual $sIg\kappa^+$ and dual $sIg\lambda^+$ 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

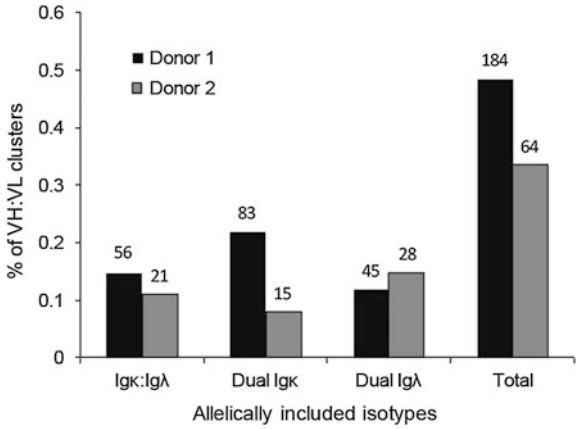

Fig. 3.4 Frequency of VL transcript allelic inclusion in two donors (n = 184 and n = 64 allelically included antibodies from n = 37,995 and n = 19,096 VH:VL clusters detected across replicates in Donor 1 and Donor 2, respectively). 14 allelically included antibodies were detected in Donor 3 (8 dual κ/λ, 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

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 (Fig. 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 $sIg\kappa^+/sIg\lambda^+$ cells (3/5 EBV-immortalized clones [31]). Also consistent with the earlier study, we observed stop codons resulting from somatic mutation that inactivated a subset of allelically included VL transcripts [31]. 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.

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, 10, 32, 33]. Many bNAbs display highly unusual features including very long CDR-H3 and short CDR-L3 sequences [7, 32, 34, 35],

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 1:6000 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 $\leq$5aa CDR-L3, and CDR-H3 length between 11 and 18 aa [35]), while antibodies with genetic characteristics of anti-influenza FI6 occurred in approximately 2–5 $\times$ 10^4 memory B cells (6 and 1 antibodies detected in Donors 1 and 2, respectively; *VH3-30*, *KV4-1*, 22aa CDR-H3, 9aa CDR-L3 [32]).

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 [36]. As many as 6 million B cells (or alternatively, as few as 1000 B cells) can be analyzed per operator in a single day. Of note, the number of antibody sequences reported here ($\sim$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 $\sim$5000 human VH:VL pairs we reported previously [4]).

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 [7] (Figs. B.6 and B.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 [34, 35]. High-throughput VH:VL sequencing can also be used to search for public antibody VH:VL clonotypes [37, 38] and to identify antibodies having specific features determined by computational or structural biology analyses or with relevance to pathogen neutralization [7, 35, 39–41]. In autoimmunity, high-throughput VH:VL sequencing can reveal an individual's repertoire of allelically included B cells (Fig. 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 [27, 29, 30].

3.4 Methods

Methods and associated references are reported in a published version of this thesis chapter [42].

References

1. Warren EH, Matsen FA, Chou J (2013) High-throughput sequencing of B-and T-lymphocyte antigen receptors in hematology. Blood 122:19–22
2. Georgiou G et al (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat Biotechnol 32:158–168
3. Logan AC et al (2011) High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. Proc Natl Acad Sci USA 108:21194–21199
4. DeKosky BJ et al (2013) High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nat Biotech 31:166–169
5. Sasaki S et al (2011) Limited efficacy of inactivated influenza vaccine in elderly individuals is associated with decreased production of vaccine-specific antibodies. J Clin Invest 121:3109–3119
6. Smith K et al (2009) Rapid generation of fully human monoclonal antibodies specific to a vaccinating antigen. Nat Protoc 4:372–384
7. Doria-Rose NA et al (2014) Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. Nat 509:55–62
8. Wilson PC, Andrews SF (2012) Tools to therapeutically harness the human antibody response. Nat Rev Immunol 12:709–719
9. Fischer N (2011) Sequencing antibody repertoires: the next generation. MAbs 3:17–20
10. Wu X et al (2011) Focused Evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. Sci 333:1593–1602
11. Finn JA, Crowe JE Jr (2013) Impact of new sequencing technologies on studies of the human B cell repertoire. Curr Opin Immunol 25:613–618
12. Finco O, Rappuoli R (2014) Designing vaccines for the Twenty-First Century society. Front Immunol 5
13. Newell EW, Davis MM (2014) Beyond model antigens: high-dimensional methods for the analysis of antigen-specific T cells. Nat Biotechnol 32:149–157
14. Marcus JS, Anderson WF, Quake SR (2006) Microfluidic Single-Cell mRNA isolation and analysis. Anal Chem 78:3084–3089
15. White AK et al (2011) High-throughput microfluidic single-cell RT-qPCR. Proc Natl Acad Sci USA 108:13999–14004
16. Furutani S, Nagai H, Takamura Y, Aoyama Y, Kubo I (2012) Detection of expressed gene in isolated single cells in microchambers by a novel hot cell-direct RT-PCR method. Analyst 137:2951–2957
17. Turchaninova MA et al (2013) Pairing of T-cell receptor chains via emulsion PCR. Eur J Immunol 43:2507–2515
18. Berkland C, Kim KK, Pack DW (2001) Fabrication of PLG microspheres with precisely controlled and monodisperse size distributions. J Controlled Release 73:59–74
19. Berkland C, Pollauf E, Pack DW, Kim K (2004) Uniform double-walled polymer microspheres of controllable shell thickness. J Controlled Release 96:101–111
20. Recher M et al (2011) IL-21 is the primary common γ chain-binding cytokine required for human B-cell differentiation in vivo. Blood 118:6824–6835

21. Jackson KJL, Kidd MJ, Wang Y, Collins AM (2013) The shape of the lymphocyte receptor repertoire: lessons from the B cell receptor. Front Immunol 4:1–12
22. Jackson KJL et al (2012) Divergent human populations show extensive shared IGK rearrangements in peripheral blood B cells. Immunogenet 64:3–14
23. Hoi KH, Ippolito GC (2013) Intrinsic bias and public rearrangements in the human immunoglobulin V[lambda] light chain repertoire. Genes Immun 14:271–276
24. Ippolito GC et al (2012) 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
25. Glanville J et al (2011) Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. Proc Natl Acad Sci USA 108:20066–20071
26. Pelanda R (2014) Dual immunoglobulin light chain B cells: trojan horses of autoimmunity? Curr Opin Immunol 27:53–59
27. Liu S et al (2005) 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
28. Rezanka LJ, Kenny JJ, Longo DL (2005) 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
29. Casellas R et al (2007) Igκ allelic inclusion is a consequence of receptor editing. J Exp Med 204:153–160
30. Andrews SF et al (2013) Global analysis of B cell selection using an immunoglobulin light chain–mediated model of autoreactivity. J Exp Med 210:125–142
31. Giachino C, Padovan E, Lanzavecchia A (1995) kappa+ lambda+ dual receptor B cells are present in the human peripheral repertoire. J Exp Med 181:1245–1250
32. Corti D et al (2011) A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza a hemagglutinins. Sci 333:850–856
33. Wrammert J et al (2011) Broadly cross-reactive antibodies dominate the human B cell response against 2009 pandemic H1N1 influenza virus infection. J Exp Med 208:181–193
34. Jardine J et al (2013) Rational HIV immunogen design to target specific germline B cell receptors. Sci 340:711–716
35. Zhou T et al (2013) Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-Class antibodies. Immun 39:245–258
36. Busse CE, Czogiel I, Braun P, Arndt PF, Wardemann H (2014) Single-cell based high-throughput sequencing of full-length immunoglobulin heavy and light chain genes. Eur J Immunol 44:597–603
37. Parameswaran P et al (2013) Convergent antibody signatures in human dengue. Cell Host Microbe 13:691–700
38. Jackson KJL et al (2014) Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. Cell Host Microbe 16:105–114
39. Wine Y et al (2013) Molecular deconvolution of the monoclonal antibodies that comprise the polyclonal serum response. Proc Natl Acad Sci 110:2993–2998
40. Lavinder JJ et al (2014) Identification and characterization of the constituent human serum antibodies elicited by vaccination. Proc Natl Acad Sci 111:2259–2264
41. Boutz DR et al (2014) Proteomic identification of monoclonal antibodies from serum. Anal Chem 86:4758–4766
42. DeKosky BJ et al (2015) In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. Nat Med 21:86–91

Chapter 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 high 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, can it then be 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 have limited prior efforts to obtain 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 as described in Chap. 1, and as a result throughput was limited to at most approximately 10^3 total B cell clones per experiment [1–3]. More recently, high-throughput DNA sequencing methods have been applied for analysis the naïve B cell repertoire [4–6]. However, due to the limitations of high-throughput sequencing that lead to loss of paired antibody heavy and light chain information (also described in Chap. 1), these recent high-throughput analyses were unable to detect the complete heavy and light chain antibody clonotypes. Thus, we hypothesized that 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 generate new insights regarding 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 the design of advanced immunogens for novel vaccine approaches [7–9] or

B. DeKosky, *Decoding the Antibody Repertoire*,
Springer Theses, DOI 10.1007/978-3-319-58518-5_4

enhance our understanding of aberrant antibody development and its role in autoimmune disease [10–13].

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 5×10^6 input cells per sample) [14]. 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 [14]. 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 [14]. 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 [14]). Naïve and antigen-experienced heavy/light paired V-gene usage were visualized by surface

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 [14]

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

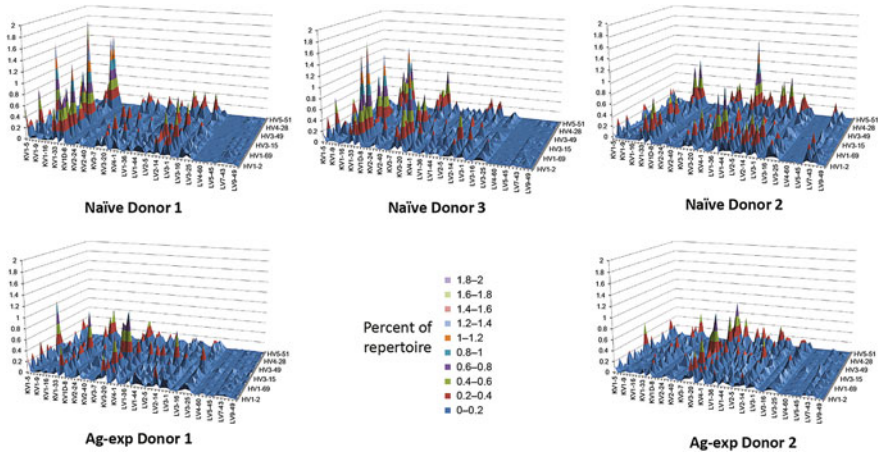

Fig. 4.1 Paired heavy/light V-gene usage surface maps of sequenced antibody repertoires. Consistent trends in gene usage were readily observed, and the antibody repertoires of each donor and subset were distinct. Statistical analysis of VH:VL gene usage data presented here was performed with Pearson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

plots (Fig. 4.1), and paired V-gene usage in antigen-experienced repertoires was further subdivided by heavy chain isotype for analysis (Figs. C.1 and C.2).

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

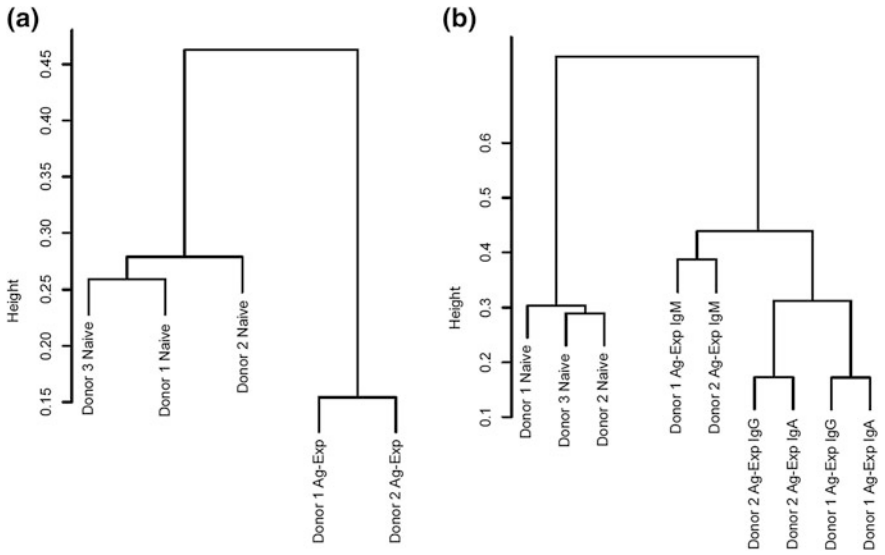

Fig. 4.2 Clustergrams resulting from Pearson hierarchical cluster analysis of paired heavy and light chain V-gene usage in sequenced donor repertoires. Panel **a** compares naïve repertoires to antigen-experienced repertoires, whereas panel **b** compares naïve repertoires to each of three antigen-experienced 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 (Fig. 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 (Fig. 4.2b). Importantly, IgM repertoires were found to cluster between 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 antigen-experienced IgM B cell subsets, the antigen-expeienced 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 (Fig. C.3). The variation in IgK:IgL usage may have caused a broader spread among the naïve repertoire cluster compared to the antigen-experienced cluster (Fig. 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 antigen-experienced 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 [15], DESeq [16], 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 [15]. 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 C.1). Interestingly, many of these V-genes observed to be differentially expressed in the paired heavy/light naïve and antigen-experienced repertoires were also observed with differential V-gene expression by separate heavy and light chain sequencing [4–6].

4.2.2 CDR3 Length Analysis

Prior studies have shown that CDR3 length decreases during B cell maturation into naïve B cells [17] and again increases only slightly in naïve compared to antigen-experienced repertoires [4]. Also, single-cell RT-PCR studies found no strong correlation between heavy and light chain lengths [1, 18]. 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 ($p < 10^{-14}$, Fig. 4.3a). In contrast, CDR-L3 length distributions showed more similarity across repertoires (Fig. 4.4b), and broader error bars in naïve CDR-L3 lengths resulted from variability in IgK and IgL gene usage in naïve repertoires (Fig. C.3) because IgK and IgL possess distinct CDR-L3 length distributions (maxima at 9aa in IgK versus 11aa in IgL).

Finally, we observed no strong correlation in paired heavy and light chain CDR3 loop lengths, similar to previous reports [1, 18] (Fig. C.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 [4], and we inspected CDR3 charge distributions to see if paired heavy and light chain CDR3 displayed similar charge features compared to prior 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 (Fig. 4.4a). We observed that the antigen-experienced

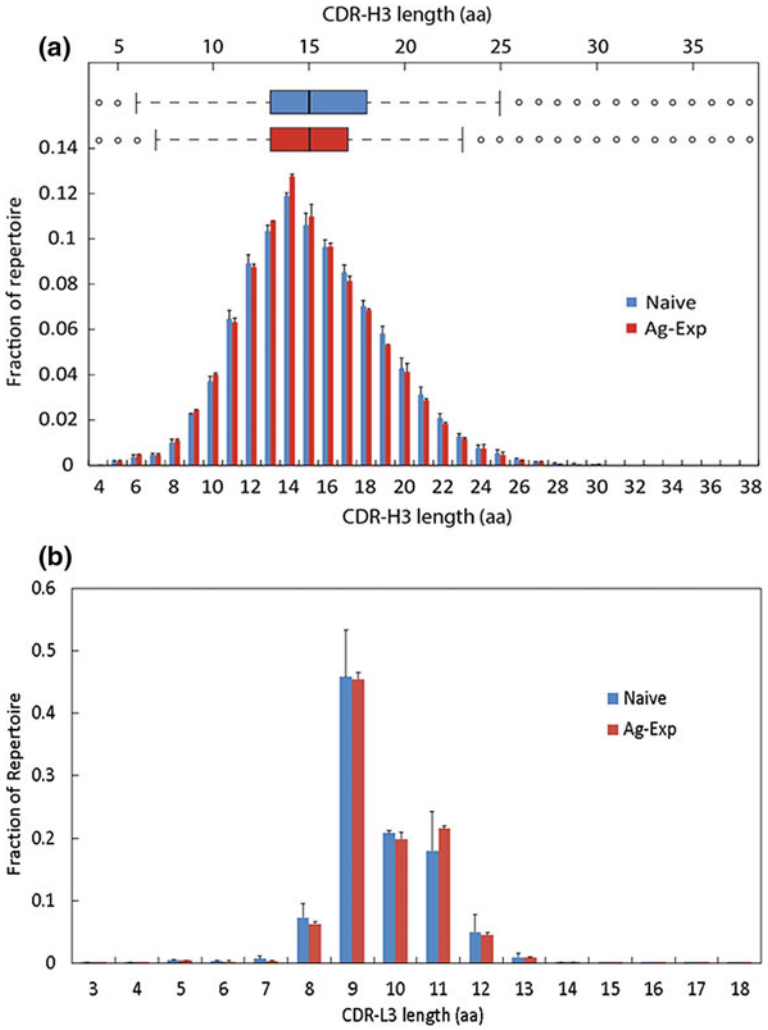

Fig. 4.3 Distribution histograms (average ± standard deviation) for **a** CDR-H3 amino acid length, and **b** CDR-L3 amino acid length, averaged across all three donors

repertoire was enriched for positive charges compared to naïve repertoires on an overall CDR-H3:CDR-L3 basis (Fig. 4.4a), on a heavy chain-only basis (Fig. 4.4b), and on a light chain basis (Fig. 4.4c); differences in charge distribution for all three naïve:antigen-experienced repertoire comparisons in Fig. 4.4 were statistically significant by the K-S test ($p < 10^{-14}$). B-cell receptors with charge extremes were 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

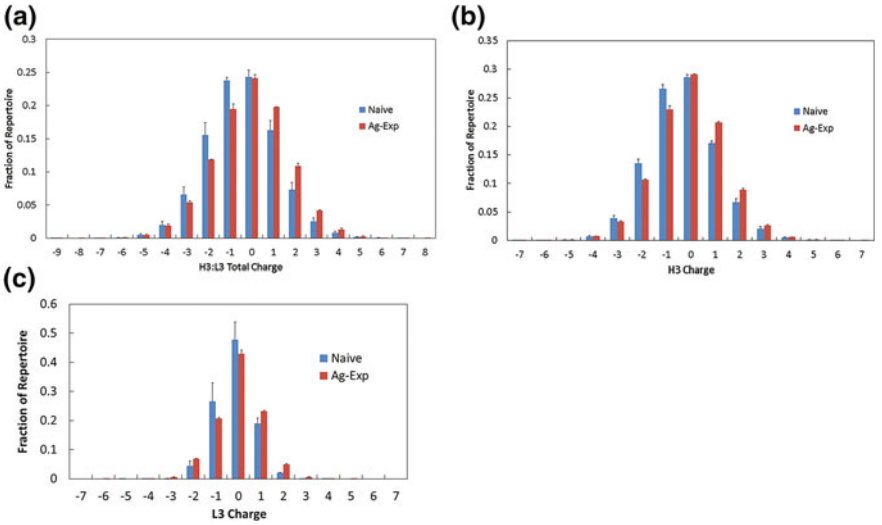

Fig. 4.4 CDR3 charge distribution for naïve and antigen-experienced repertoires (average ± standard deviation) in **a** total CDR-H3 and CDR-L3 charge **b** CDR-H3 charge, and **c** CDR-L3 charge. Differences in charge distribution between naïve and antigen-experienced repertoires for all three panels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

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 (Fig. C.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 (Fig. C.6). The enhancement of positive charge in both heavy and light CDR3 loops concurrently is presented as a heat map in Fig. C.7.

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 (Fig. C.5).

4.2.4 CDR3 Hydrophobicity

We analyzed the average hydrophobicity per residue (avg H-index) in paired CDR-H3 and CDR-L3 loops [19]. 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-H3 s in antigen-experienced repertoires (average H-index between 0 and 0.1, Fig. 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-H3 s were well-tolerated.

Regarding CDR-L3 hydrophobicity we again observed strong differences between kappa and lambda repertoires. Overall, lambda light chain CDR3 s were much more hydrophobic than kappa CDR3 s, and these patterns were consistent across donors (Fig. 4.6). We also observed that kappa light chains were strongly selected for enhanced hydrophobicity indices and to a much greater extent than lambda light chains within the same donor (Fig. 4.6). 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.

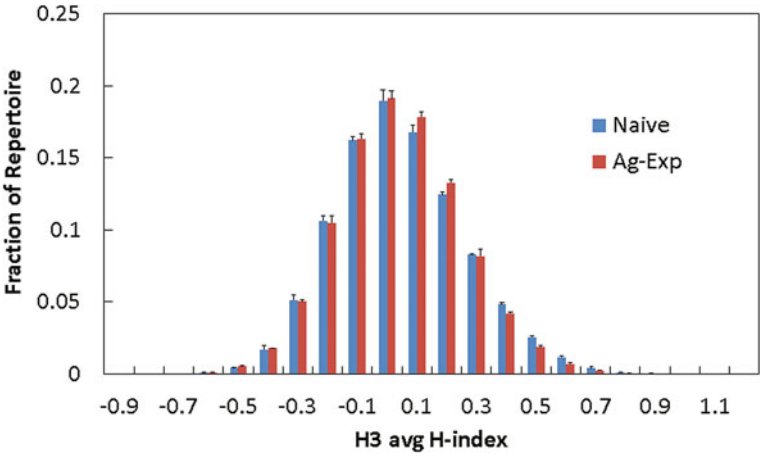

Fig. 4.5 CDR-H3 loop average hydrophobicity (avg H-index ± standard deviation) distributions in naïve and antigen-experienced repertoires

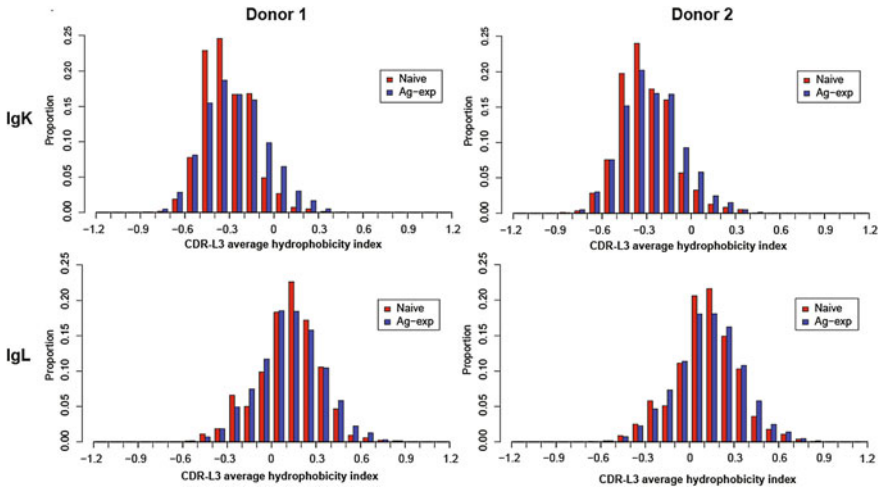

Fig. 4.6 CDR-L3 loop average hydrophobicity indices in naïve and antigen-experienced antibody repertoires for Donor 1 (*left*) and Donor 2 (*right*), subdivided by IgK (*top*) and IgL (*bottom*). Kappa and lambda repertoires exhibited distinct CDR-L3 average hydrophobicity distributions (*top* compared 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 C.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 [14, 20]. 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% 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% 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% 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 antigen-experienced 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 [5, 21].

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 [11, 22–24]. We observed a total of 23 exact amino acid match CDR-H3 amino acid junctions among

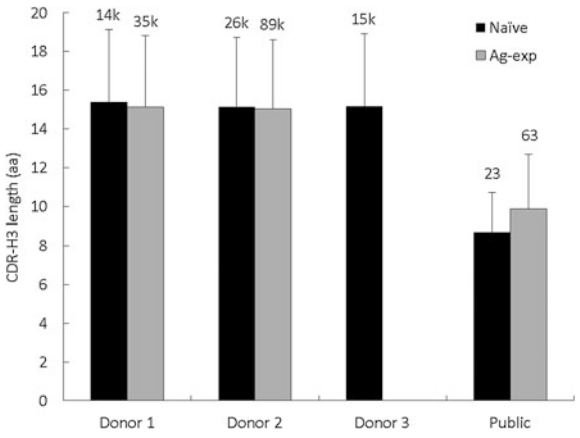

Fig. 4.7 CDR-H3 length comparisons between overall repertoires and public CDR-H3 amino acid sequences (average ± standard deviation). Values above each column indicate the total number of CDR-H3 in each group

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 (Fig. 4.7). 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 (Fig. 4.8). 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

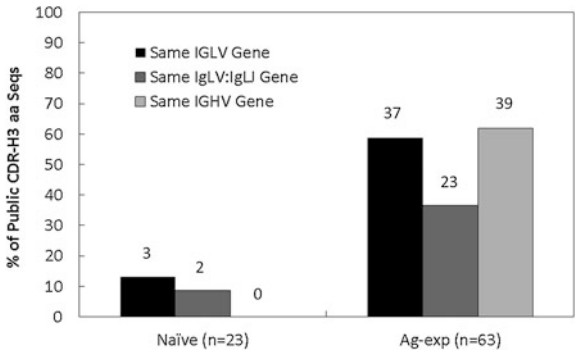

Fig. 4.8 Gene usage comparisons between public CDR-H3 amino acid. Values above each column indicate the total number of public CDR-H3 in each group

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 C.2).

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 (Fig. 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 (Fig. 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 [25] 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 V-gene 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 approximately 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 [4]) while CDR-L3 lengths remained largely unchanged after maturation to antigen-experienced B cells (Fig. 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 (Fig. 4.4), and slight differences in charge composition were observed between kappa and lambda repertoires (Fig. C.5). We hypothesize that the general increases in antibody charges may better enable binding to negatively charged bacterial membranes for anti-bacterial antibodies. Importantly, we found that increases in H3 and L3 charge often occurred concurrently (Fig. C.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 (Fig. C.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 CDR3 s was strongly increased in antigen-experienced repertoires compared to only minor changes in lambda light chain CDR3 hydrophobicity as a result of positive B cell selection (Fig. 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 [26–28]. 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 [14, 26–29]. 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% of naïve light chains were encoded by a promiscuous nucleotide sequence while nearly 80% of naïve light chains were encoded by a promiscuous amino acid sequence, and even around 50% 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 (Fig. 4.8). 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 [11, 22–24].

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 [14]. 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 [30].

4.4.3 Bioinformatic Analysis

Raw Illumina sequences were quality-filtered, mapped to V-, D-, and J- genes and CDR3 s extracted using both the International Immunogenetics Information System (IMGT) [31] and NCBI IgBlast software [32] with a CDR3 motif identification algorithm [33]. 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 [34]), 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 [34], 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 [5]. Amino acid sequence hydrophobicity was determined by the normalized version of the Kyte-Doolittle hydrophobicity index [19]; 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) [15, 35]. 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.

References

1. Brezinschek H-P, Foster SJ, Dörner T, Brezinschek RI, Lipsky PE (1998) Pairing of variable heavy and variable κ chains in individual naive and memory B cells. J Immunol 160:4762–4767
2. Bräuninger A, Goossens T, Rajewsky K, Küppers R (2001) Regulation of immunoglobulin light chain gene rearrangements during early B cell development in the human. Eur J Immunol 31:3631–3637
3. Tian CX et al (2007) Evidence for preferential Ig gene usage and differential TdT and exonuclease activities in human naive and memory B cells. Mol Immunol 44:2173–2183
4. Wu YC et al (2010) High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. Blood 116:1070–1078
5. Glanville J et al (2011) Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. Proc Natl Acad Sci USA 108:20066–20071
6. Mroczek ES et al (2014) Differences in the composition of the human antibody repertoire by B cell subsets in the blood. Front Immunol 5:96
7. McGuire AT et al (2013) Engineering HIV envelope protein to activate germline B cell receptors of broadly neutralizing anti-CD4 binding site antibodies. J Exp Med 210:655–663
8. Jardine J et al (2013) Rational HIV immunogen design to target specific germline B cell receptors. Sci 340:711–716
9. McLellan JS et al (2013) Structure-based design of a fusion glycoprotein vaccine for respiratory syncytial virus. Sci 342:592–598
10. Schoettler N, Ni D, Weigert M (2012) 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

11. Lindop R et al (2011) Molecular signature of a public clonotypic autoantibody in primary Sjögren's syndrome: a 'forbidden' clone in systemic autoimmunity. Arthritis Rheum 63:3477–3486
12. Dorner T, et al (2011) Long-lived autoreactive plasma cells drive persistent autoimmune inflammation. Nat Rev Rheumatol 7:170
13. Di Niro R et al (2012) High abundance of plasma cells secreting transglutaminase 2-specific IgA autoantibodies with limited somatic hypermutation in celiac disease intestinal lesions. Nat Med 18:U441–U204
14. DeKosky BJ et al (2015) In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. Nat Med 21:86–91
15. Smyth GK (2005). In: Gentleman R, Carey VJ, Huber W, Irizarry RA, Dudoit S (eds) Bioinformatics and computational biology solutions using r and bioconductor, Springer, New York, pp 397–420
16. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11:R106
17. Wardemann H et al (2003) Predominant autoantibody production by early human B cell precursors. Sci 301:1374–1377
18. de Wildt RM, Hoet RM, van Venrooij WJ, Tomlinson IM, Winter G (1999) Analysis of heavy and light chain pairings indicates that receptor editing shapes the human antibody repertoire. J Mol Biol 285:895–901
19. Eisenberg D (1984) Three-dimensional structure of membrane and surface proteins. Annu Rev Biochem 53:595–623
20. Jackson KJL, Kidd MJ, Wang Y, Collins AM (2013) The shape of the lymphocyte receptor repertoire: lessons from the B cell receptor. Front Immunol 4:1–12
21. Georgiou G et al (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat Biotechnol 32:158–168
22. Parameswaran P et al (2013) Convergent antibody signatures in human dengue. Cell Host Microbe 13:691–700
23. Jackson KJL et al (2014) Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. Cell Host Microbe 16:105–114
24. Smith K et al (2013) Fully human monoclonal antibodies from antibody secreting cells after vaccination with Pneumovax®23 are serotype specific and facilitate opsonophagocytosis. Immunobiol 218:745–754
25. Briney BS, Willis JR, McKinney BA, Crowe JE (2012) High-throughput antibody sequencing reveals genetic evidence of global regulation of the naive and memory repertoires that extends across individuals. Genes Immun 13:469–473
26. Liu S et al (2005) 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
27. Casellas R et al (2007) Igκ allelic inclusion is a consequence of receptor editing. J Exp Med 204:153–160
28. Andrews SF et al (2013) Global analysis of B cell selection using an immunoglobulin light chain–mediated model of autoreactivity. J Exp Med 210:125–142
29. Giachino C, Padovan E, Lanzavecchia A (1995) kappa+ lambda+ dual receptor B cells are present in the human peripheral repertoire. J Exp Med 181:1245–1250
30. DeKosky BJ et al (2013) High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nat Biotech 31:166–169
31. Brochet X, Lefranc M-P, Giudicelli V (2008) IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. Nucleic Acids Res 36:W503–W508
32. Ye J, Ma N, Madden TL, Ostell JM (2013) IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res 41:W34–W40

33. Ippolito GC et al (2012) 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
34. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinform 26:2460–2461
35. Smyth GK (2004) 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

Chapter 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 [1], 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 [2]. We determined the VH:VL pairing accuracy of our technique to be >97% [2] and applied our technology for antibody discovery [1, 3–5], proteomic analysis of vaccine responses [3, 6], mechanistic investigation of HIV broadly neutralizing antibody development [4], and to generate novel immunological insight related to the composition and development of antibody repertoires in healthy human donors [2, 7].

A primary future direction for paired heavy:light sequencing is to analyze the features of various B cell subsets. The work presented here analyzed the sequence and structural differences between naïve and antigen-experienced B cells [7], uncovering a number of unreported features regarding the development and maturation of the human B cell repertoire in healthy adults. However, many more distinct B cells have been identified that remain to be analyzed, including earlier developmental stages (e.g. immature B cells) and antigen-experienced subsets such as plasmablasts, plasma cells (including long-lived and short-lived plasma cell subsets), “double-negative” or tissue-like memory, and more recently identified subsets which are still being refined and developed [8–12]. The high-throughput sequence and structural analyses of these B cell populations may reveal their distinct developmental origins and functional contributions to effective (or occasionally, ineffective) adaptive immunity. The antigen-presenting role of B cells has been recognized to have greater importance in recent years [13], and paired heavy:light sequencing may reveal additional insights for the antigen-presenting influence of B cells on T cell responses as well. A major topic of interest is in the identification of

B. DeKosky, *Decoding the Antibody Repertoire*,
Springer Theses, DOI 10.1007/978-3-319-58518-5_5

the B cell precursor subset that leads to long-lived plasma cells [12, 14, 15], which has the potential to significantly accelerate the development of highly effective vaccines. Enhanced understanding of B cell development will help us design vaccines with more effective long-term protection by enhanced elicitation of long-lived plasma cells.

Another critical application of paired heavy:light sequencing is in the sequence analysis of vaccine and disease responses. Recent reports have identified convergent genetic signatures that were elicited in response to influenza vaccination [16], dengue infection [17], HIV infection [18, 19], and other diseases [20, 21]. However, most of these signatures rely on the heavy chain identification only, which results in poor predictive capacity and an inability to experimentally test identified antibodies for binding specificity. In other cases, such as VRC01-class HIV broadly neutralizing antibody identification, the gene signatures are based on paired heavy:light sequences which cannot be detected at high-throughput by any other technology [18]. Paired heavy:light sequence information will help to better elucidate the genetic convergence among antibody repertoires of distinct individuals, as it provides a much higher resolution for the identification of genetically similar antibodies than separate heavy-only and light-only antibody sequencing can provide.

Similar to the identification of genetic convergence in vaccines and disease responses, a number of genetic similarities have been observed in autoimmune antibodies [20–23]. Paired heavy:light sequencing in the context of autoimmunity may also provide enhanced resolution of autoimmune B cell features. The additional information contained in paired heavy:light sequences could one day enable diagnosis of the precise mechanism of autoimmunity based on a single blood sample. Furthermore, the confident identification of autoimmune mechanistic targets would highlight potential insights as to how autoimmunity develops and enable targeted strategies for antigen-specific therapies.

Paired heavy:light sequencing has a reasonable efficiency (approximately one antibody VH:VL cluster per 15-20 input cells, as reported in Chap. 4), however it is likely that significant opportunities for improvement remain as the technology matures. Several features of the paired heavy:light sequencing workflow could be further optimized. Perhaps most importantly, the RT-PCR enzyme mix could be further improved. Recently developed protocols in molecular biology have generated enhanced reverse transcription and PCR enzymes [24], and these proteins or comparative analysis of multiple RT-PCR kits may enhance the yield of emulsion PCR reactions. Additionally, further optimization of the mRNA capture step may be possible. Variation of the number of beads, size of emulsion droplets, and lysis buffer compositions could reveal an optimized protocol to enhance the amount of mRNA captured using paired heavy:light sequencing. Additional non-specific B cell stimulation technologies could also increase the amount of immunoglobulin mRNA transcribed to further enhance antibody sequence recovery [4]. Finally, advances in sequence analysis throughput and/or error rates could improve the yield of data and bioinformatic algorithms and enhance the amount of information that can be obtained from a single paired heavy:light sequencing run. The paired

heavy:light sequencing workflow currently operates with acceptable efficiency to obtain far greater yield than any other technology and address a wide range of scientific problems; still, the above opportunities for improvement would provide even greater utility and a broader range of applications for this antibody sequencing platform.

A further important application beyond the analysis of antibody repertoires is in the high-throughput sequencing of paired T cell receptor (TCR) β chain:α chain sequences. Paired β:α sequencing is an analogous problem to heavy:light sequencing, with a few key differences: (i) T cells have lower T-cell receptor expression levels than most B cells, (ii) the TCR genes are much more diverse than B cell genes, (iii) T cells do not show somatic hypermutation, and (iv) T cells show a higher rate of multiple α chains than B cells with allelically included light chains. These features make T cell analysis somewhat more difficult experimentally due to the lower expression levels and higher size of multiplex primer libraries, and will require similar bioinformatic protocols with slight differences as compared to antibody bioinformatic analysis (iii, iv). Paired β:α T cell receptor sequencing is underway with promising results, and the ability to analyze paired β:α repertoires will dramatically improve our ability to understand T cell responses in the setting of infectious diseases, vaccination, and autoimmunity.

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. Importantly, the poly(dT)-based single-cell capture method collects all mRNAs at the single-cell level, which includes far more than the immunoglobulin variable region gene sequences, and future efforts in this area will incorporate additional mRNAs of interest. For example, one could pair heavy chain sequences with 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. Another approach is to use barcoded beads or hydrogel droplets for analysis of the entire transcriptome of single cells at very high throughput [25, 26]. 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.

In summary, our accessible technology for sequencing the paired antibody VH:VL repertoire has enabled 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 is enhancing high-throughput, high-resolution analysis of human vaccine responses, providing new ways to test vaccine efficacy and inform vaccine design. The high-throughput identification and cloning of paired VH:VL antigen-specific antibodies from responding B cells will enable rapid generation of novel diagnostic, therapeutic or prophylactic antibodies and catalyze further high-impact research in the origins and

development of humoral immunity. 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.

References

1. DeKosky BJ et al (2013) High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nat Biotech 31:166–169
2. DeKosky BJ et al (2015) In-depth determination and analysis of the human paired heavy—and light-chain antibody repertoire. Nat Med 21:86–91
3. Lavinder JJ et al (2014) Identification and characterization of the constituent human serum antibodies elicited by vaccination. Proc Natl Acad Sci USA 111:2259–2264
4. Doria-Rose NA et al (2014) Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. Nature 509:55–62
5. Wang B et al (2015) Facile discovery of a diverse panel of Anti-Ebola Virus antibodies by immune repertoire mining. Sci Rep 5:13926
6. Lee J et al (2016) Quantitative, molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. Nat Med (Accepted)
7. DeKosky BJ et al (2016) Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. Proc Natl Acad Sci USA 113:E2636–E2645
8. Wirths S, Lanzavecchia A (2005) ABCB1 transporter discriminates human resting naive B cells from cycling transitional and memory B cells. Eur J Immunol 35:3433–3441
9. Jackson SM, Wilson PC, James, JA, Capra JD (2008) In: Frederick W, Alt KFATHFMJWU, Emil RU (eds) Advances in immunology. Academic Press, vol 98, pp 151–224
10. Moir S et al (2008) Evidence for HIV-associated B cell exhaustion in a dysfunctional memory B cell compartment in HIV-infected viremic individuals. J Exp Med 205:1797–1805
11. Kaminski DA, Wei C, Qian Y, Rosenberg AF, Sanz I (2012) Advances in human B cell phenotypic profiling. Front Immunol 3:302
12. Halliley JL et al (2015) Long-lived plasma cells are contained within the $CD19^{-}CD38^{hi}CD138^{+}$ Subset in Human Bone Marrow. Immunity 43:132–145
13. Nanton MR, Way SS, Shlomchik MJ, McSorley SJ (2012) Cutting edge: B cells are essential for protective immunity against salmonella independent of antibody secretion. J Immunol 189:5503–5507
14. Amanna IJ, Carlson NE, Slifka MK (2007) Duration of humoral immunity to common viral and vaccine antigens. N Engl J Med 357:1903–1915
15. Amanna IJ, Slifka MK (2010) Mechanisms that determine plasma cell lifespan and the duration of humoral immunity. Immunol Rev 236:125–138
16. Jackson KJL et al (2014) Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. Cell Host Microbe 16:105–114
17. Parameswaran P et al (2013) Convergent antibody signatures in human dengue. Cell Host Microbe 13:691–700
18. Zhou T et al (2013) Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-Class antibodies. Immunity 39:245–258
19. Gorman J et al (2016) Structures of HIV-1 Env V1V2 with broadly neutralizing antibodies reveal commonalities that enable vaccine design. Nat Struct Mol Biol 23:81–90

20. Lindop R et al (2011) Molecular signature of a public clonotypic autoantibody in primary Sjögren's syndrome: a 'forbidden' clone in systemic autoimmunity. Arthritis Rheum 63:3477–3486
21. Rounds WH et al (2014) The antibody genetics of multiple sclerosis: comparing next-generation sequencing to sanger sequencing. Front Neurol 5:166
22. Schoettler N, Ni D, Weigert M (2012) 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
23. Kalinina O et al (2014) Light chain editors of anti-DNA receptors in human B cells. J Exp Med 211:357–364
24. Ellefson JW et al (2016) Synthetic evolutionary origin of a proofreading reverse transcriptase. Science 352:1590–1593
25. Klein AM et al (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161:1187–1201
26. Macosko EZ et al (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161:1202–1214

Appendix A
Chapter 2 Supplementary Information

See Figs. A.1, A.2, A.3 and Tables A.1, A.2, A.3, A.4, A.5, A.6, A.7.

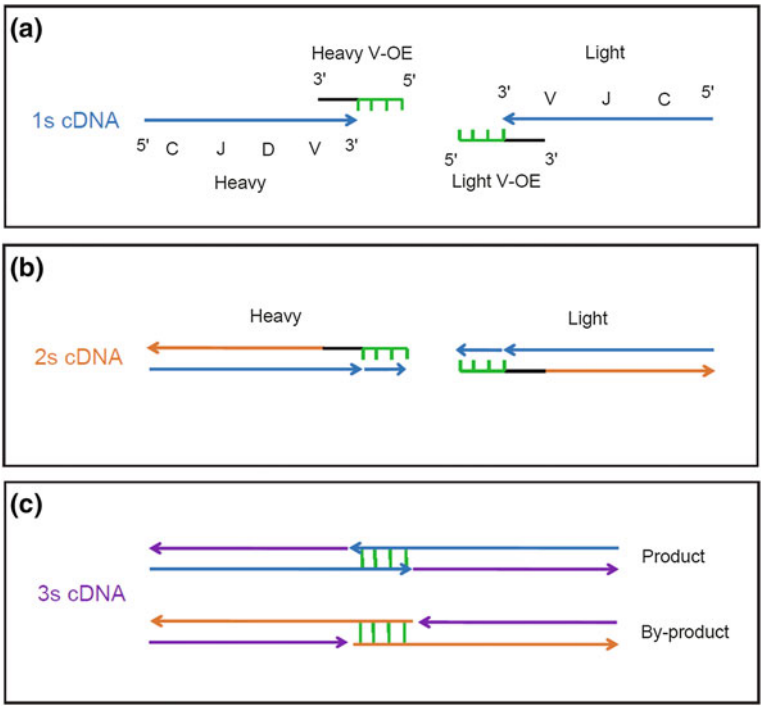

Fig. A.1 An overview of the linkage (overlap extension) RT-PCR process. **a** V-region primers (*black*) with a 5′ complementary heavy/light overlap region (*green*) anneal to first strand cDNA. **b** Second strand cDNA is formed by 5′–3′ extension; the overlap region is incorporated into all cDNA. **c** 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 2 × 250 paired-end Illumina sequencing. Linkage RT-PCR primer sequences are given in Table A.5 (V-region primers denoted "fwd-OE" and constant region primers denoted "rev-OE")

B. DeKosky, *Decoding the Antibody Repertoire*,
Springer Theses, DOI 10.1007/978-3-319-58518-5

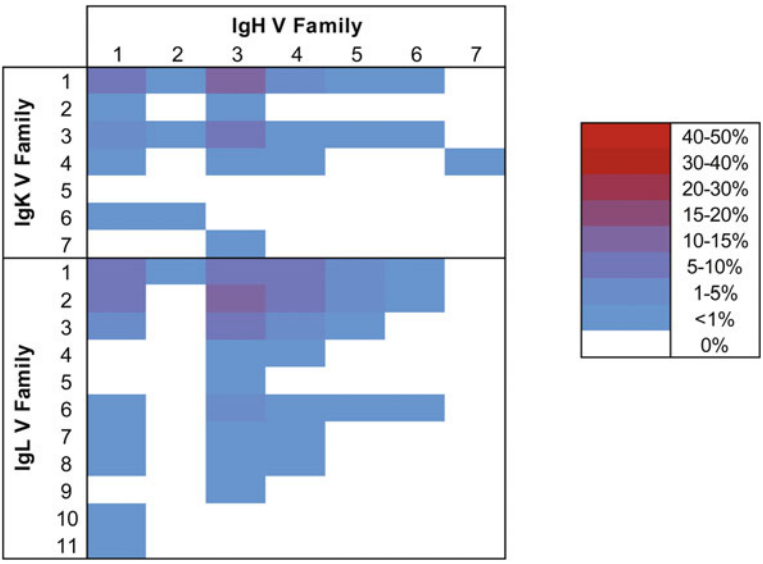

Fig. A.2 A heat map of VH:VL pairings from IgG$^+$ class-switched peripheral B cells isolated from a healthy volunteer (n = 2248). The experiment presented here is a replicate of Fig. 2.2a using donated blood from a different individual

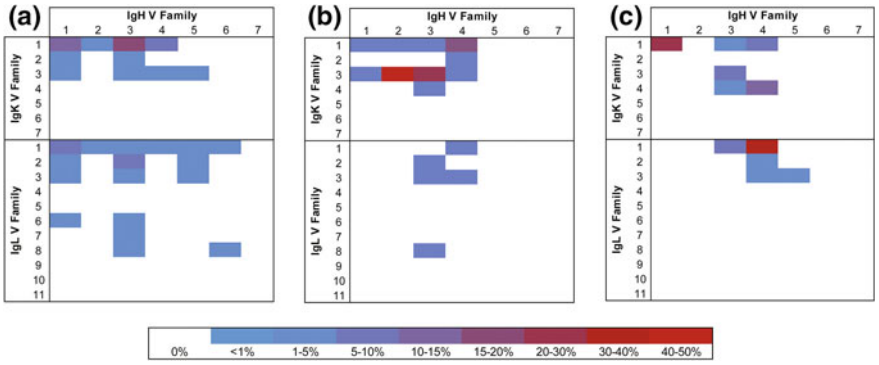

Fig. A.3 As Fig. 2.2b comprised the lowest sample size in Fig. 2.2 (n = 86 unique pairs, compared to Fig. 2.2a, n = 2716, and Fig. 2.2c, n = 240) a simulation was performed to randomly select 86 VH:VL pairs from Fig. 2.2a, c and normalize all panels to 86 unique sequences. **a** healthy donor peripheral 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**

Table A.1 Key statistics from several paired VH:VL repertoires

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 versus 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	2716	86	240
Control cell spike	IM-9	ARH-77	IM-9
Accuracy ratio[a]	78:1	650:1	942:1

TD-tetanus toxoid/diphtheria toxoid, MSD-Merck Sharpe and Dohme
[a]For known spiked cells, (reads correct VL):(reads top incorrect VL)

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

Immunization	n/a
Cell type	IgG^+ B lymphocytes
Fresh cells versus Freeze/Thaw	Fresh
Cell:Well ratio	1:10
% Cells as single cells	95.1%
Unique CDR-H3 recovered	2248
Control cell spike	IM-9
Accuracy ratio[a]	125:1

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

Table A.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

Seq ID	Isotype	CDR-H3	Paired CDR-L3[a]	Source
2D02	IgM	gcgagaggcggaaatgggcgacccctttgacaac	gcagcatgggatgacagcctgaatggttgggtg	Sanger scRT-PCR
2D02	IgM	gcgagaggcggaaatgggcgacccctttgacaac	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

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

Table A.4 Statistical analysis of pairing accuracy

Experiment	Resting IgG+ repertoire	Tetanus toxoid D7 plasmablasts
Figure	2.2a	2.2b
Cell:Well ratio	1:10	1:425
Fraction single cells[a]	0.951	0.999
Spiked clone	IM-9	ARH-77
Spike %	4.0	7.5
Estimated # spiked cells	2830	30
Estimated spiked cells as single cells[a]	2,691	30
Estimated spiked cells as 2-cells-per-well[a]	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	2716	86

[a]Calculated from the poisson distribution

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

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

(continued)

Table A.5 (continued)

Conc. (nM)	Primer ID	Sequence
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 A.6 Nested PCR primers

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 A.7 VH and VL separate amplification 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

Appendix B
Chapter 3 Supplementary Information

See Figs. B.1, B.2, B.3, B.4, B.5, B.6, B.7, B.8 and Tables B.1, B.2, B.3, B.4.

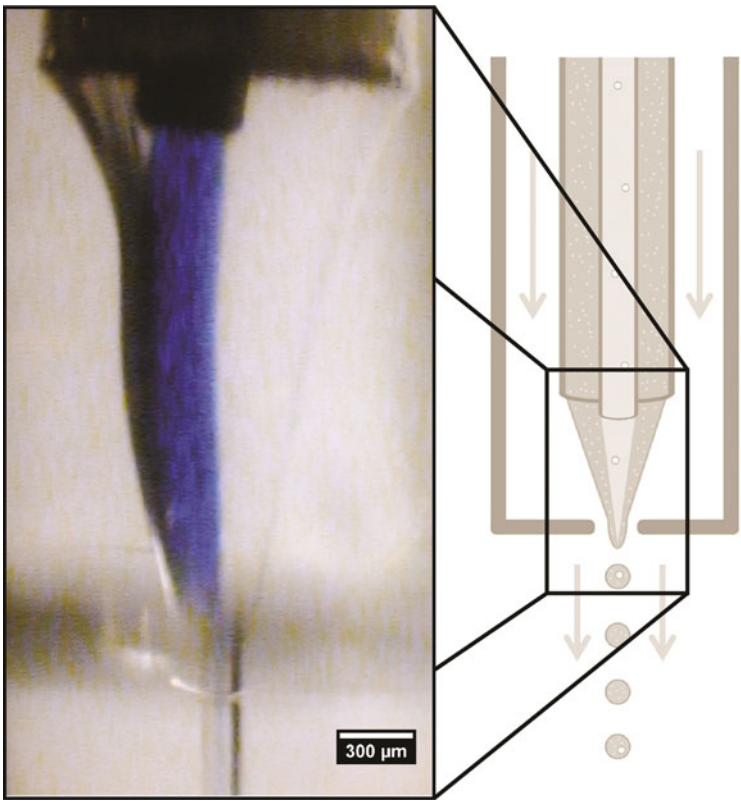

Fig. B.1 A micrograph of the axisymmetric flow-focusing nozzle during emulsion generation (*left*), placed in context of the diagram from Fig. 2.1a (*right*), where PBS/0.4% Trypan blue exits the inner needle and cell lysis buffer exits the outer needle

B. DeKosky, *Decoding the Antibody Repertoire*,
Springer Theses, DOI 10.1007/978-3-319-58518-5

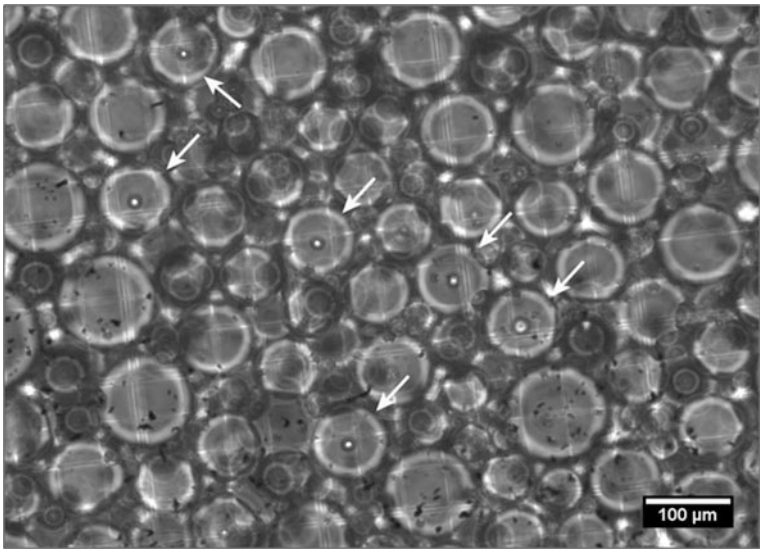

Fig. B.2 MOPC-21 immortalized B cells encapsulated in emulsion droplets. The outer aqueous stream that normally contains cell lysis buffer (Fig. 3.1a, *gray solution*) was replaced with 0.4% Trypan blue in PBS to examine cell viability throughout the flow focusing and emulsification process. Emulsified cell viability was approximately 90% and cell viability did not differ substantially from non-emulsified controls

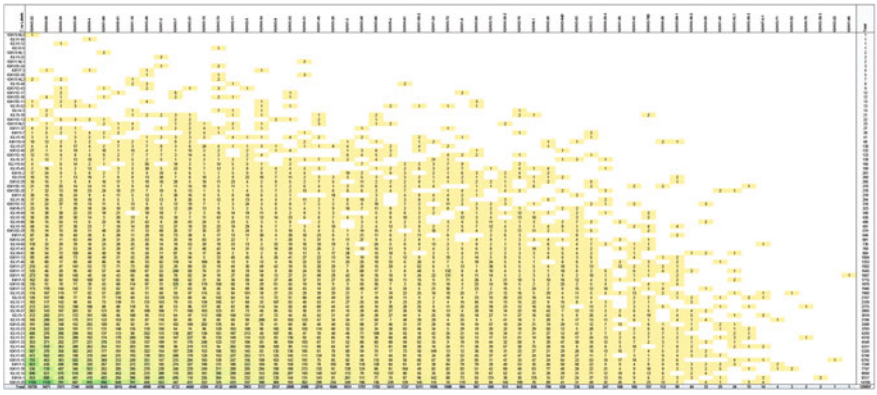

Fig. B.3 Heat map of V-gene usage for 129,097 VH:VL clusters recovered from Donor 1. Sequences were collected using primers targeting the framework 1 region; raw data is available in the online supplement

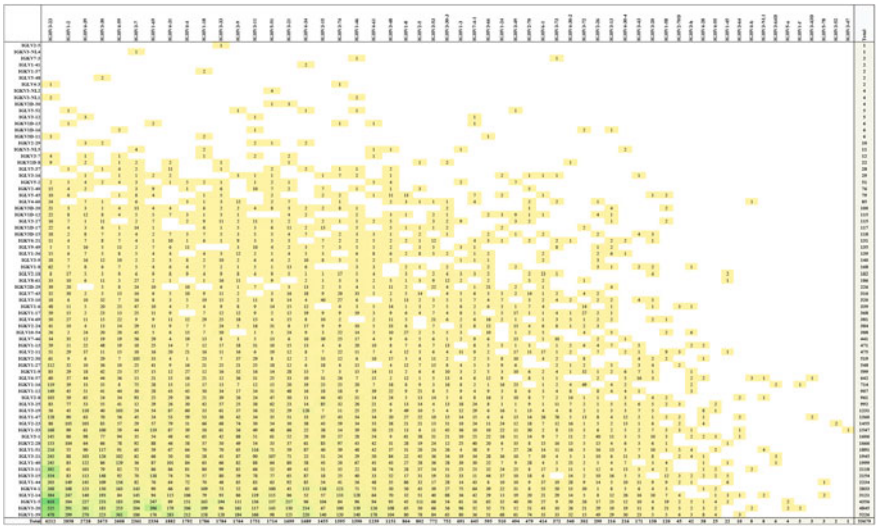

Fig. B.4 Heat map of V-gene usage for 53,679 VH:VL clusters recovered from Donor 2. Sequences were collected using primers targeting the framework 1 region; raw data is available in the online supplement

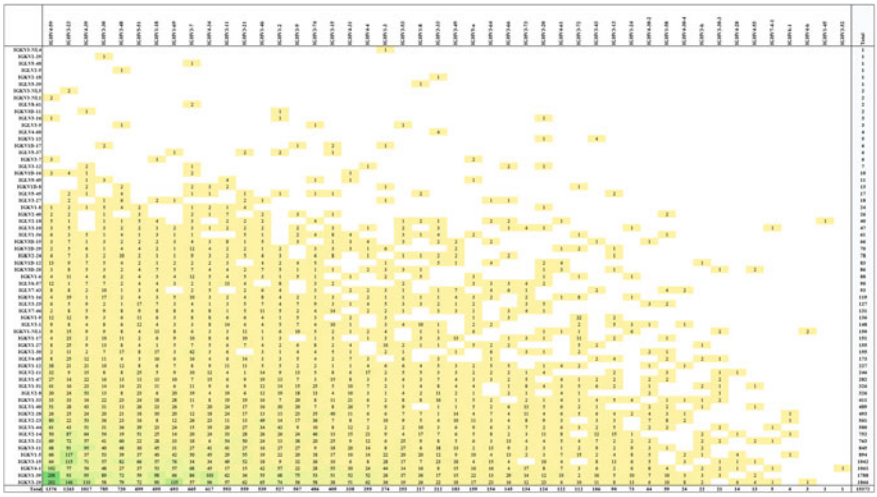

Fig. B.5 Heat map of V-gene usage for 15,372 VH:VL clusters recovered from Donor 3. Sequences were collected using primers targeting the leader peptide region; raw data is available in the online supplement

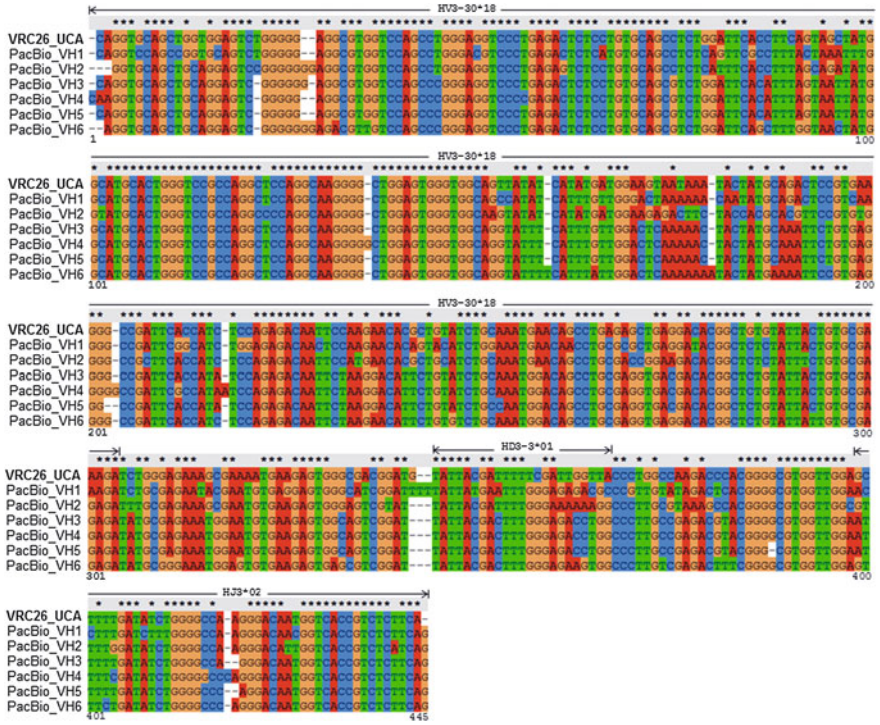

Fig. B.6 VH alignment of the six VRC26 HIV broadly neutralizing antibody variants recovered by PacBio sequencing of complete ~850bp VH:VL amplicons. Sequences were recovered from $CD27^{+}$ peripheral B cells of the CAP256 donor and aligned to the VRC26 VH unmutated common ancestor (UCA, Doria-Rose et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

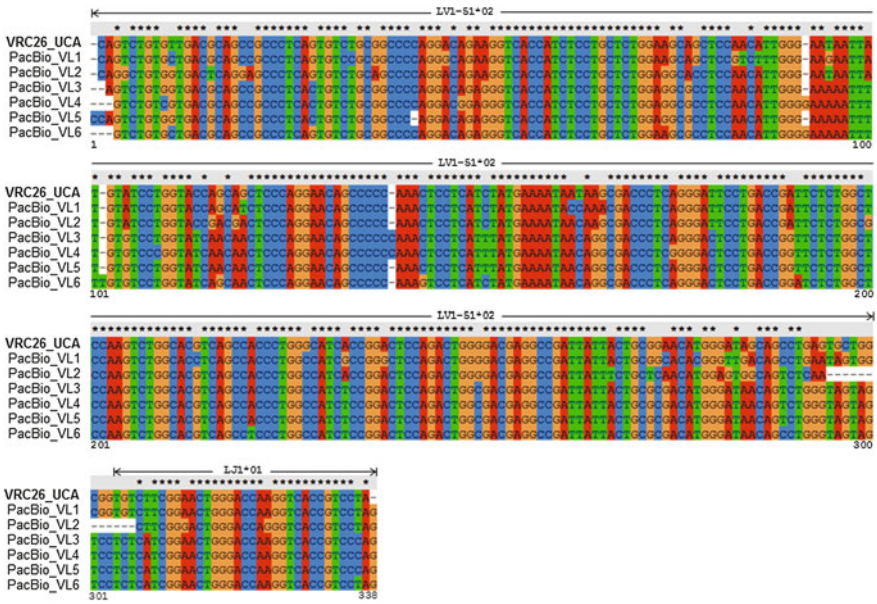

Fig. B.7 VL alignment of the six VRC26 HIV broadly neutralizing antibody variants recovered by PacBio sequencing of complete ~850bp VH:VL amplicons. Sequences were recovered from $CD27^+$ peripheral B cells of the CAP256 donor and aligned to the VRC26 VL unmutated common ancestor (UCA, Doria-Rose et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

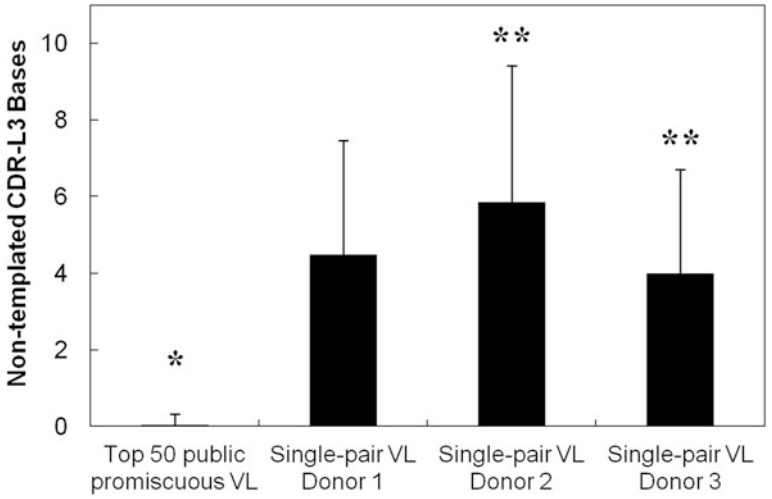

Fig. B.8 Comparison of the number of non-templated bases (sum of somatic mutations and non-templated insertions) in the top 50 public, promiscuous VL nucleotide junctions shared by Donors 1, 2, and 3–50 randomly selected VL junctions paired with only a single heavy chain in the Donor 1, Donor 2, or Donor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

Table B.1 VH:VL pairing analysis of a mixture of HEK293 cells transfected with 11 different known antibodies

		Heavy Chain											
		1H	2H	3H	4H	5H	6H	7H	8H	9H	10H	11H	Total
Light chain	1L	**1842**	4	20	13	18	16	39	20	49	6	4	2031
	2L	0	**4916**	34	31	59	41	102	127	146	28	8	5492
	3L	0	2	**6251**	9	38	25	116	60	118	13	2	6634
	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	1320	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	1335	323	**35,391**	109	46	37,826
	10L	12	24	944	189	1597	1080	3519	1898	4291	**8535**	98	22,187
	11L	32	66	1153	272	1258	1655	6405	6567	6185	555	**14,126**	38,274
	Total	1931	5147	9151	15,432	19,789	20,012	34,094	42,191	49,084	9626	14,494	220,951

The maximum read count for each row and column is highlighted; 11/11 antibodies were identified and paired correctly in this control experiment (correct VH-VL pairings displayed in bold). Read count variation was expected due to varying transfection and 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 B.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

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

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	1751
Correct ARH-77 VH:VL reads (signal)	2604
ARH-77 top incorrect VL reads (top VLnoise)	27
ARH-77 2nd-ranked incorrect VL reads	19
ARH-77 3rd-ranked incorrect VL reads	16
ARH-77 Signal:topVLnoise ratio[a]	96.4

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

Table B.3 Memory B cell counts before and after in vitro activation

Sample	FACS count	Hemocytometer count
	Fresh Bmems	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

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

Table B.4 Leader peptide overlap extension primers

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

Appendix C
Chapter 4 Supplementary Information

See Figs. C.1, C.2, C.3, C.4, C.5, C.6, C.7 and Tables C.1, C.2, C.3.

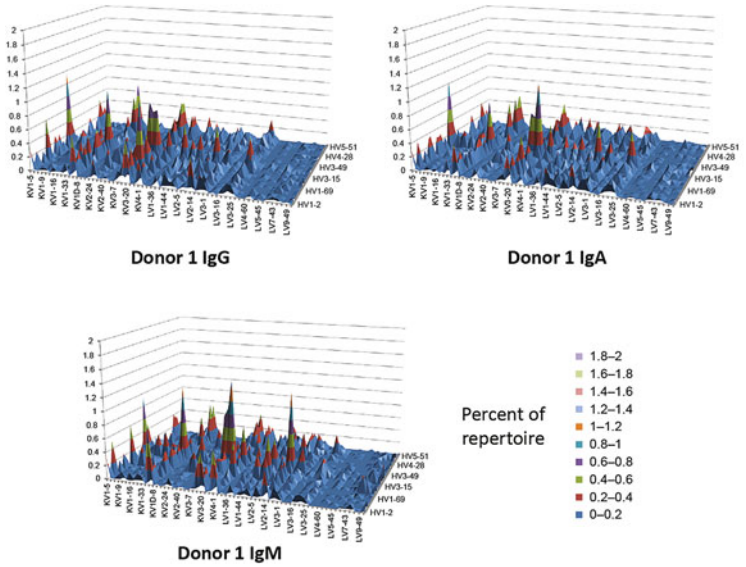

Fig. C.1 V-gene pairing surface plot for B-cell receptors observed in Donor 1, subdivided by heavy chain isotype

B. DeKosky, *Decoding the Antibody Repertoire*,
Springer Theses, DOI 10.1007/978-3-319-58518-5

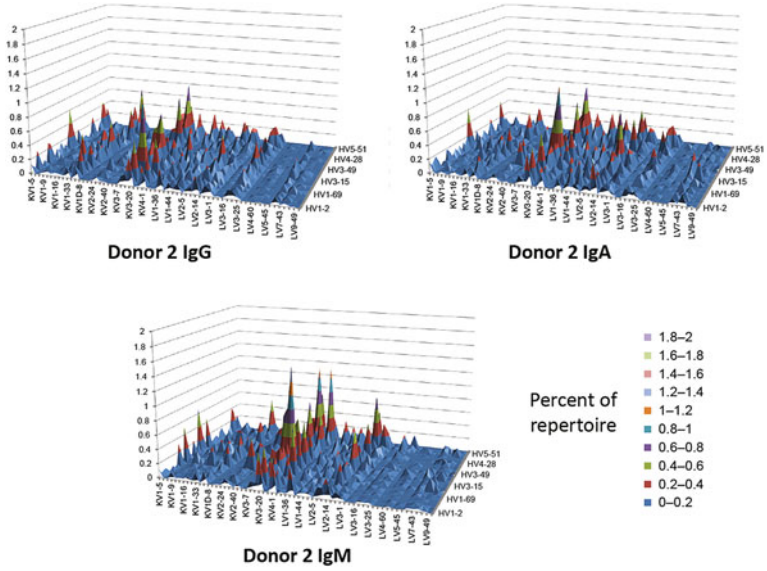

Fig. C.2 V-gene pairing surface plot for B-cell receptors observed in Donor 2, subdivided by heavy chain isotype

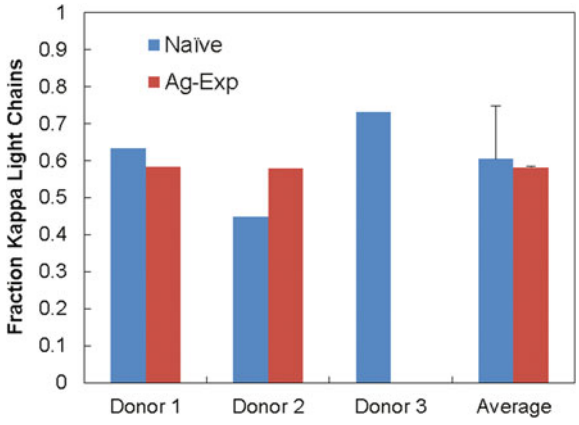

Fig. C.3 Fraction IgK light chain gene usage across B cell subsets for Donors 1, 2, and 3. Error bars indicate standard deviation for averaged values

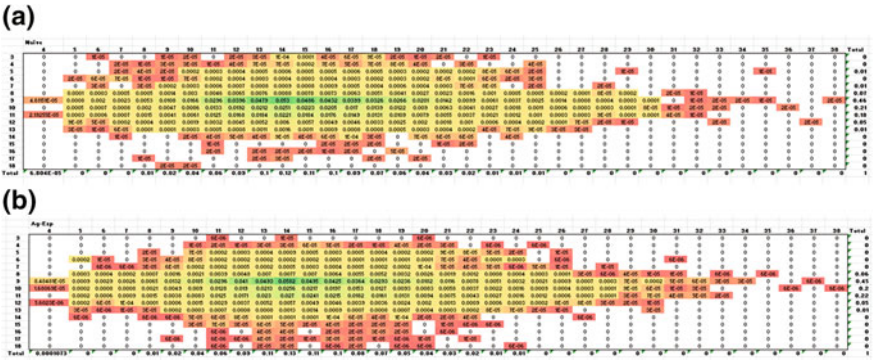

Fig. C.4 CDR-H3:CDR-L3 length heat maps of **a** naïve donor repertories, and **b** antigen-experienced donor repertoires

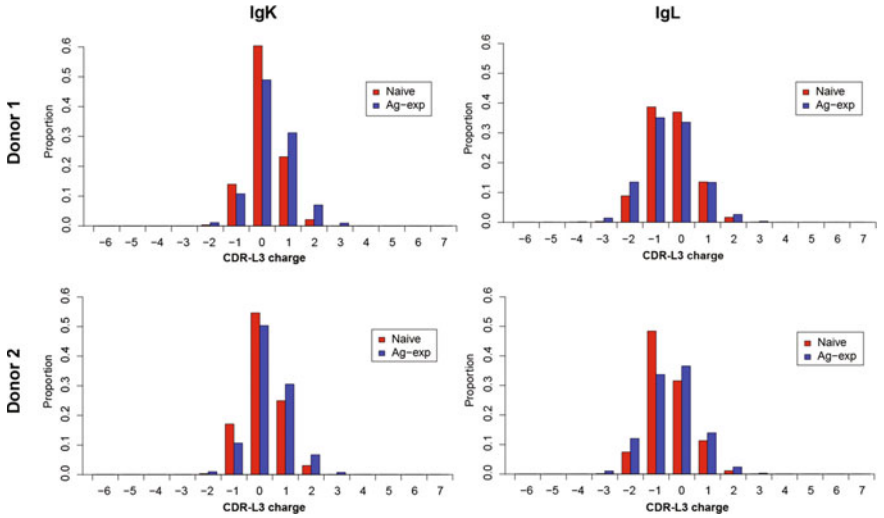

Fig. C.5 CDR-L3 loop charge in naïve and antigen-experienced antibody repertoires for Donor 1 (*left*) and Donor 2 (*right*), subdivided by IgK (*top*) and IgL (*bottom*). This figure demonstrates that kappa and lambda repertoires exhibit distinct CDR-L3 charge distributions (*top* compared to *bottom* 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 C.2

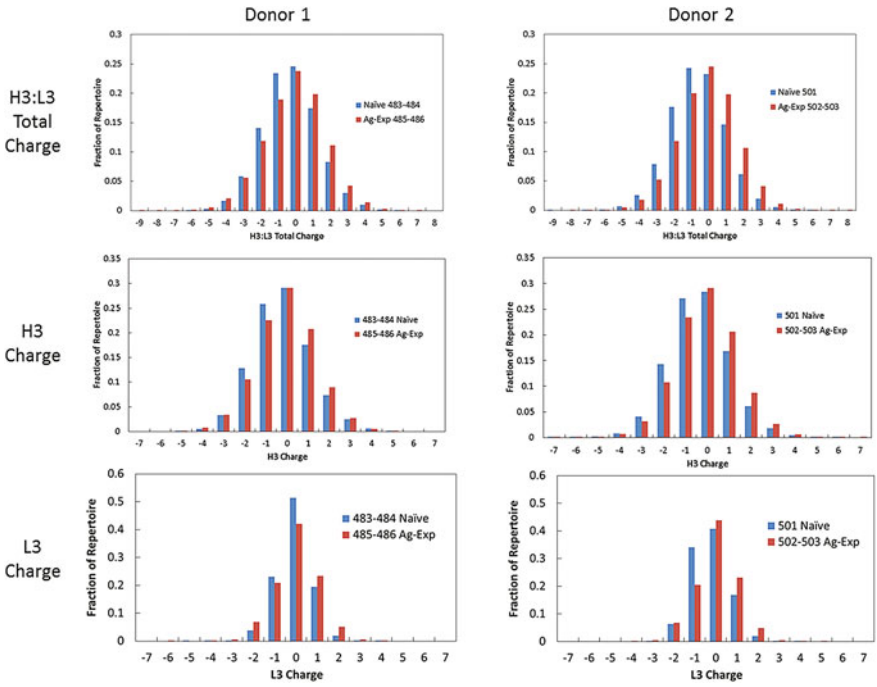

Fig. C.6 Charge distributions for naïve and antigen-experienced repertoires of Donors 1 and 2, further subdivided by CDR-H3:CDR-L3 total charge (*top*), CDR-H3 charge (*middle*), and CDR-L3 charge (*lower*)

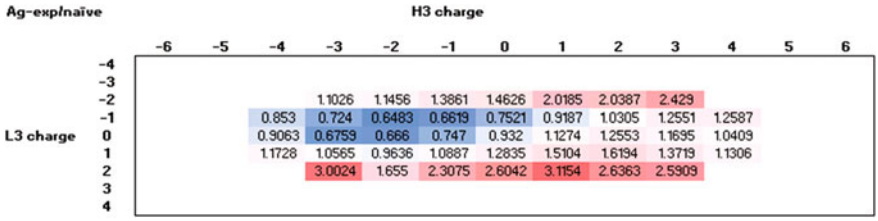

Fig. C.7 Relative representation ratio heat map of CDR-H3:CDR-L3 charge combinations across naïve and antigen-experienced repertoires. Values represent the ratio of antigen-experienced:naïve repertoire fractional representation for a given H3:L3 charge combination; red and blue shading represents relative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

Table C.1 Statistically significant differentially expressed heavy/light V-gene pairs with adjusted $p < 0.05$ between naïve and antigen-experienced antibody repertoires

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

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

Table C.2 Recovered IgK and IgL pairs for Donor 1 and Donor 2 among naïve and antigen-experienced subsets

Isotype	Cell subset	Donor 1	Donor 2
IgK	Naïve	8521	11,312
	Ag-Exp	19,755	49,844
IgL	Naïve	4919	13,860
	Ag-Exp	14,067	36,126

Table C.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 (DeKosky et al, *Nature Medicine* 2015). 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

Donor	CDR-H3_aa	CDR-L3_aa	CDR-H3_nt	CDR-L3_nt	Gene usage
1	ARSSRGAAAGFDY	GTWDSSLSA-VV	gcgagaAGCTCGAGGGGagcagcagcCGGTtttgactac	ggaacatgggacagcagcctgagtgct---gtggta	HV4-59:D6-13:J4/λV1-51:J2
3	ARSSRGAAAGFDY	GTWDSSLSAGVV	gcgagaTCTTCCGCGGagcagcagctggttttgactac	ggaacatgggatagcagcctgagtgctggGgtggta	HV4-59:D6-13:J4/λV1-51:J2
2	ARSSRGAAAGFDY	(LQNAGVTGT)	gcgagaTCTTCACGCGGGgcagcagctggttttgactac	(ctgcagaatgctggcGTAACGGggacg)	HV4-59:D6-13:J4/κV3-20:J1
2	ANEIRPNDY	LISYTDARIWV	gcgaacgaAATAAGACctaatgactac	ttaatctcctatactgatgctcgAATttgggtg	HV3-23:D3-3:J4/λV7-46:J3
3	ANEIRPNDY	LLSYGDV--WV	gcgaacgaGATCCGACCCAatgattat	ttgctctcctacggtgatg------ttgggtg	HV3-23:D1-14:J4/λV7-46:J3
1	ARARGYSYGYSDY	QKYNSAPALT	gcgagagCTCgtggatacagctatggttactctgactac	caaaagtataacagtgccccCGCgctcact	HV1-8:D5-18:J4/κV1-27:J4
3	ARARGYSYGYSDY	QKYNSAPP-T	gcgagagCCCgtggatacagctatggttactctgactac	caaaagtataacagtgcccctcc---cact	HV3-48:D5-18:J4/κV1-27:J4
1	ARGHYGLDV	QQYGSSPIT	gcgagagGACactacggtttggacgtc	cagcaatatggtagctcaccgatcacc	HV3-11:-:J6/κV3-20:J5
2	ARGHYGLDV	QQYGSSSRT	gcgagagGTCactacggtttggacgtc	cagcagtatggtagctcatctctGAacg	HV3-7:-:J6/κV3-20:J1
1	ARGEDYYYGMDV	QSIDTDGTHGVV	gccagggggaAGactactactacggtatggacgtc	cagtcaattgacaccgatggtactcatGGtgtggta	HV3-11:D3-16:J6/λV3-25:J2
2	ARGEDYYYGMDV	QSADSSGTY-VV	gcggcgaggtgaAGactactactacggtatggacgtc	caatcagcagacagcagtggcactta---tgtggta	HV3-11:D2-21:J6/λV3-v

Index

B. DeKosky, *Decoding the Antibody Repertoire*,
Springer Theses, DOI 10.1007/978-3-319-58518-5