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

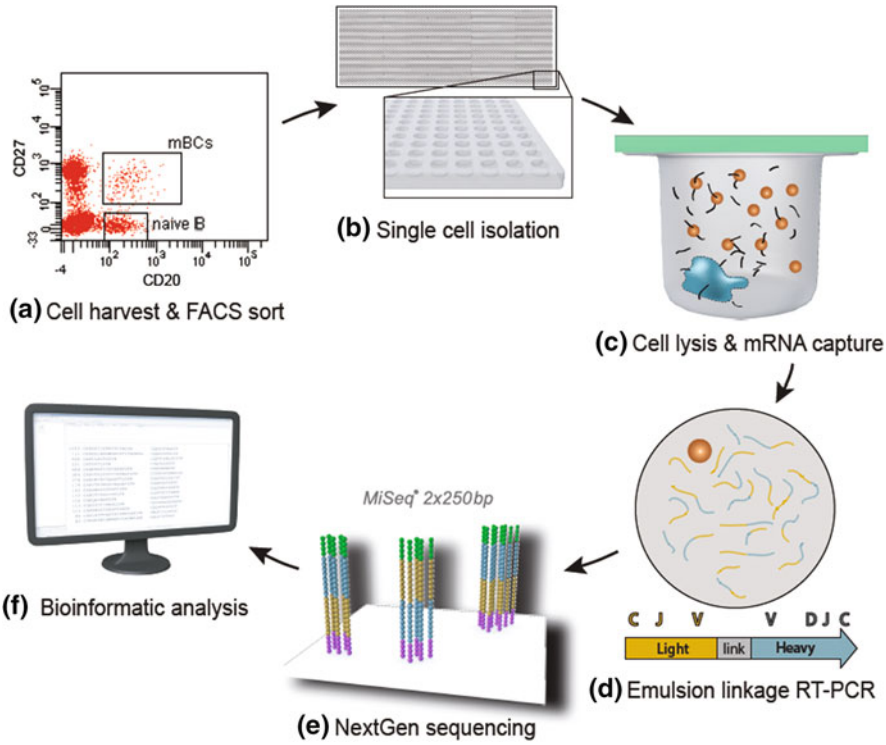

Fig. 2.1 Overview of the high throughput methodology for paired VH:VL antibody repertoire analysis. **a** B-cell populations are sorted for desired phenotype (mBCs = memory B cells, naive B = naive B cells). **b** Single cells are isolated by random settling into 125 pL wells (56 μm diameter) printed 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

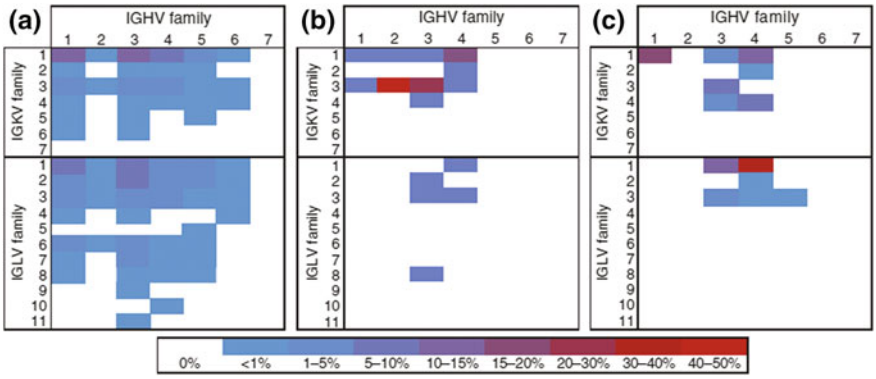

Fig. 2.2 VH:VL gene family usage of unique CDR-H3:CDR-L3 pairs identified via high-throughput sequencing of cell populations from three different individuals in separate experiments using the workflow presented in Fig. 2.1: **a** healthy donor peripheral IgG^+ B cells (n = 2716 unique CDR3 pairs), **b**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

Springer

http://www.springer.com/978-3-319-58517-8

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

DeKosky, B.

2017, XXVIII, 87 p. 34 illus., Hardcover
ISBN: 978-3-319-58517-8