Deep Mining of Human Antibody Repertoires: Concepts, Methodologies, and Applications

Xiaolong Tian, Cheng Li, Yanling Wu,* and Tianlei Ying*  

1. Introduction

Human humoral immune response to antigens is reliant upon the tremendous diversity of B cell receptors (BCR), the membrane-bound form of immunoglobulins (Ig). Key determinants of BCR diversity include 1) nearly random rearrangement of variable (V), diversity (D), and joining (J) genes in the immunoglobulin heavy chains variable regions (V_{H}), and rearrangement of V–J genes in the immunoglobulin light chains variable regions (V_{L}), during the early stages of B cell maturation in the bone marrow;[1] 2) junctional diversity resulting from exonuclease trimmings and the random addition of palindrome (P) or non-template encoded (N) nucleotides in the process of rearrangement;[2] and 3) B cell proliferation after antigen stimulation in order to produce high-affinity antibodies and eliminate the antigen through subsequent class switch recombination (CSR) and somatic hypermutation (SHM).[3,4] It is estimated that there are over $10^{13}$ BCRs in the human antibody repertoire.[5] This astronomical size prevents a global analysis of the antibody repertoire using conventional DNA sequencing. However, advancements in high-throughput sequencing (HTS) for Ig-Seq have improved the ability to characterize the antibody repertoire on a large scale.[6] Over a decade of development, a variety of new technologies and approaches have improved Ig-Seq methods, resulting in improved understanding of B-cell immunology and accumulation of massive antibody sequence data. In this review, an overview of Ig-Seq and advanced methodologies applied to HTS of antibody repertoires will be discussed. We will describe the wide applications of Ig-Seq in basic and clinical immunology, and strategies for functional antibody discovery from antibody repertoires. Finally, we will provide our perspective on the future directions and challenges in human antibody repertoire deep mining.

2. Overview of Ig-Seq Pipeline

Although multiple strategies have been used for studying antibody repertoires, each strategy follows a similar workflow including sample collection, library preparation and sequencing, data cleaning, sequence alignment, and high-level processing (Figure 1).

For sample collection, peripheral blood mononuclear cells (PBMCs) and total B cells have been generally used because of their accessibility.[8–10] However, because antibody repertoire properties are divergent between antigen-specific B cells and plasma cells,[11] specific B cell populations or B cells located in targeted tissues should be enriched for more precise studies. Considering that as few as 2% of total B cells are circulating in
peripheral blood,

this population can only provide a narrow view of the adaptive immune response to antigens; therefore, lymph nodes, spleen, and red bone marrow are preferential as the sources of B cells if accessible.

V(D)J recombination occurs in the bone marrow, which means, for mature B cells in peripheral blood, either genomic DNA (gDNA) or mRNA can be used for the characterization of V(D)J sequences. In general, RNA-based approaches are adopted as they require only primers annealed to the conserved J or constant (C) region and template-switching primers using 5′ RACE (rapid amplification of cDNA ends) (Figure 2). This eliminates the biases that occur in DNA-based assays, which necessitate the use of degenerate V primer sets required to amplify from all VH families (Figure 2). Moreover, in the step of reverse transcription converting RNA to cDNA, unique molecular identifier (UMI) tags can be added to correct sequencing errors, and indexing barcodes can be introduced for pooled high-throughput sequencing, thus multiple samples can be sequenced in a lane to reduce costs and eliminate sequencing errors of different batches. To provide a better understanding of antibody repertoires in the context of specific B cell populations, there has been increased application of single B cell sequencing, a commercially available RNA-based sequencing technology that uses the polyA tail of mRNA to capture BCR genes. In some cases, DNA-based assays can be considered due to its accessibility, stability, and low heterogeneity, as immunoglobulin transcription varies dramatically between naive B cells and plasma cells.

The rapid development of Ig-Seq in recent years has established a variety of efficient sequencing strategies using different platforms. Once BCR genes have been sequenced, the raw sequencing data need to be filtered for high-quality reads based on the Phred quality score. The sequence errors that arise from reverse transcription and PCR, or sequencing artifacts, have been well-studied and reviewed elsewhere. Once clean data is obtained, the first step of processing is to assign each BCR gene to its reference gene. Unlike conventional DNA or RNA-Seq, V(D)J reference genes consist of a series of similar sequences from different families, which makes it difficult to perform an accurate gene assignment. Moreover, due to the occurrence of V(D)J junctional diversity and SHM, the rate of sequence alignment is usually less than 100%, and occasionally, two or more germline genes will be simultaneously called due to the nearby alignment rate.

Dozens of BCR gene annotations and high-level processing tools have been developed since 2013 (https://b-t.cr/c/wiki/15), yet the IMGT/V-QUEST and a higher-throughput version, IMGT/HighV-QUEST, remain the most widely used. These are tools of the International Immunogenetics Information System (IMGT) (http://www.imgt.org), a global reference database in immunogenetics and immunoinformatics, which contains a well-annotated set of genomic V, D, and J genes and thus provide the reference for most BCR gene assignment tools. Another tool provided by IMGT, the IMGT/StatClonotype, is operated in R software to perform a pairwise comparison of IMGT/HighV-QUEST output data. Once data sets are uploaded, IMGT clonotypes can be automatically clustered, and V(D)J genes usages statistically tested; IMGT/StatClonotype will also generate tables, and synthesize high-quality figures. It is time-saving and convenient for scientific researchers without a bioinformatics background, but only a pair of data sets can be analyzed at a time. However, the restriction of 200,000 input sequences for IMGT/HighV-QUEST is often not sufficient to cope with the millions of reads produced by typical Ig-Seq. Considering how time-consuming the ultra-high-throughput sequence processing is, the tools that have no sequence input limit and can run on personal servers are more widely used.

Finally, once the characteristics of antibody repertoires for different samples have been analyzed, the underlying biological or pathological meaning can then be extracted, which is discussed in the section “Applications of Ig-Seq.” Additionally, efforts to mine functional antibodies from antibody repertoires can be attempted, which is discussed in the section “Strategies of Functional Antibody Discovery from Antibody Repertoires.”
3. Methodologies for HTS of Antibody Repertoires

HTS refers to the deep, high-throughput, in-parallel DNA sequencing technologies developed a few decades after the Sanger DNA sequencing method first emerged in 1977,[38] and have had an enormous impact on the life sciences. Sequencing technologies are continually improved to become faster, more efficient, and cheaper, and the history, development, and applications have been well summarized.[39–42] In early studies of the antibody repertoire, only Sanger sequencing technology was available to delineate antibody sequences from hundreds of B cells.[43] Such a low throughput represents only a limited snapshot of the actual diversity of antibody repertoires. Compared to Sanger sequencing, next-generation sequencing (NGS) can provide a much broader picture of the antibody repertoire. The first NGS platform, 454 GS FLX, generated sequences up to $10^3$–$10^6$ with a read length of 400–600 bp that is sufficient to cover the full V_H or V_L region (Figure 2).[44] This platform was instrumental in the initial exploration of the human antibody repertoire. Ion Torrent technology was developed by the inventors of 454 sequencing, reaching a read length of 400 bp with $10^6$–$10^8$ reads per run, and was also applied to the investigations of full-length variable regions of antibody repertoires.[45,46] Over the past ten years, the Illumina sequencing platforms have been well-developed, providing a sequencing depth of around $10^6$–$10^{10}$,[47] thus improving the ability to study antibody diversity. Limited by a sequencing read length of no more than 300 bp, initially only the complementarity-determining region (CDR3) region can be sequenced by most Illumina platforms[48–50] (Figure 2). Nevertheless, Illumina HiSeq, Miseq, and NovaSeq 6000, and so on, make it possible to avoid losing the full-length of the variable region, owing to its 2 × 250 to 2 × 300 bp maximum read length (Figure 2). In addition, third-generation sequencing (also known as single-molecule sequencing) platforms such as single molecule real time sequencing (SMRT) developed by Pacific Biosciences and nanopore sequencing offered by Oxford Nanopore’s technology, offer the capability for single molecule real-time sequencing of average 10–15 kb read length[51] for the ultra-long antibodies explored.[52] These various platforms have their own advantages for conducting Ig-Seq, thus, sequencing strategies focused on different purposes have been flourishing (Table 1).[33,53,54]

![Figure 2. Schematic diagram of Ig-Seq targeting a) CDR3 and b) full-length V_H region.](image-url)

**Table 1. Summary of current Ig-Seq technologies.**

| Methodology                        | Application settings                         | Pitfalls                                    | Throughput      | Efficiency (antigen-specific BCRs) |
|-----------------------------------|---------------------------------------------|---------------------------------------------|-----------------|------------------------------------|
| 96 well cloning                   | Rapid antibody isolation                    | Extremely low-throughput, labor-intensive   | $10^3$–$10^4$   | ≈50%                               |
| Bulk BCR-Seq                      | Immune repertoires characterization         | Without V_H/V_L paired information, low antibody isolation efficiency | $10^5$–$10^6$   | ≈20–30%                            |
| Proteomic-based bulk BCR sequencing | Immune repertoires characterization, antibody isolation | Additional proteomic experiment, without natively V_H/V_L paired information | $10^5$–$10^6$   | ≈30–40%                            |
| Plate-based V_H/V_L paired sequencing | Antibody isolation, immune repertoires characterization | Low-throughput, labor-intensive,            | $10^5$–$10^6$   | ≈50%                               |
| Microfluidic-based single cell sequencing | Immune repertoires characterization, antibody isolation | Low-throughput, high-cost, short sample storage term | $10^5$–$10^6$   | ≈50%                               |
| Third-generation sequencing       | Novel antibody identification, display antibody library sequencing | Relatively high sequencing error rate, high-cost per cell | $10^3$–$10^4$   | ≈20–30%                            |
3.1. V<sub>H</sub>:V<sub>L</sub> Paired Ig-Seq

The IGHV gene is located on chromosome 14,[53] and the IGKV and IGLV are located on chromosome 2 and chromosome 22 respectively,[54,63] which makes it difficult to acquire V<sub>H</sub>:V<sub>L</sub> paired amplicons. As a result, efforts in the antibody repertoire researches are usually concentrated solely on IGHV diversity. Initial studies in V<sub>H</sub>:V<sub>L</sub> paired antibody repertoire sequencing relied on single B cell cloning by Sanger sequencing of the individual V<sub>H</sub> and V<sub>L</sub> genes in 96-well plates,[64] enabling the isolation of pathogen-neutralizing antibodies.[65] However, this process is inherently low-throughput, expensive, and yields a very limited set of antibody sequences. Wan et al.[9] took the first step in the V<sub>H</sub>:V<sub>L</sub> paired HTS sequencing. Rabbits were immunized with human progesterone receptor A/B (PR A/B) peptides, and 80,000 V region sequences from the immunized rabbits' antibody repertoires were sequenced using the Roche 454 platform. This generated a reference database to enable the protein sequence search of immunoglobulin V region matches. Once high-confidence and enriched antibodies were antigen-enriched and identified by liquid chromatography-mass spectrometry (LC-MS/MS), these heavy and light chain sequences were then expressed combinatorially in a matrix. Although functional monoclonal antibodies were obtained, the native pairing of IGH and IGK/IGL was disrupted, thus access to V<sub>H</sub>:V<sub>L</sub> paired functional antibodies by randomly arranging and combining results resulted in a large workload and low probability.

To address the challenges in acquiring the natively paired heavy and light chains, Busse et al.[66] combined single B cell V<sub>H</sub> and V<sub>L</sub> gene amplicons using a 2D, bar-coded primer matrix. In brief, forward and reverse primers that incorporated unique identification tags were used for single B cell PCR, and according to the unique tag combinations, each IGH and IGK/IGL sequences can be traced back to the original B cell. Nevertheless, using this approach to perform HTS Ig-Seq requires massive numbers of bar-coded primers and subsequent PCR.

An additional approach for sequencing native V<sub>H</sub>:V<sub>L</sub> pairs relies on single B cell lysis and mRNA capture in a high-density microwell plate.[67] In this study, by depositing single B cells into microwells and lysed in situ, mRNA is immediately captured on poly-dT magnetic beads, and V<sub>H</sub>:V<sub>L</sub> paired amplicons were generated via reverse transcription and emulsion V<sub>H</sub>:V<sub>L</sub> linkage PCR procedures. Finally, about 10% unique V<sub>H</sub>:V<sub>L</sub> pairs of 7 x 10⁴ B cells were sequenced by Illumina MiSeq 2 x 250 bp platform. This approach opened the door to automated endogenous antibody V<sub>H</sub>:V<sub>L</sub> paired HTS. As an “upgraded version” of this method, Georgiou laboratory has recently developed an improved V<sub>H</sub>:V<sub>L</sub> paired HTS method.[68] Using an axisymmetric flow focusing device, millions of B cells were separated into microdroplets and lysed, then single-cell mRNA was captured on poly-dT magnetic beads, and V<sub>H</sub>:V<sub>L</sub> paired amplicons with an overlap extension PCR. Illumina MiSeq 2 x 300 bp platform was used for sequencing. With this method, the throughput of V<sub>H</sub>:V<sub>L</sub> paired sequencing is unprecedentedly improved. More conveniently, the restriction sites contained in the 32 bp length linkers between V<sub>H</sub>:V<sub>L</sub> make it easily to be subcloned into an antigen binding fragment (Fab) expression vector. Therefore, this method is bifunctional in antigen-specific antibody discovery. Given the indispensability of antibody light chain in exercising humoral immunity functions, natively paired V<sub>H</sub> and V<sub>L</sub> genes may be expected in most antibody repertoire studies.

SMART-Seq2 is a proper plate-based single-cell sequencing technology in V<sub>H</sub>:V<sub>L</sub> paired sequencing which generates full-length V<sub>H</sub> and V<sub>L</sub> cDNA libraries from each individual cell.[69] In this method, single cells are sorted and lysed in the microtiter plates, where mRNA is isolated and reverse transcribed. Next, DNA tagmentation reaction and adapter ligation are performed, and the subsequent enrichment PCR are used to construct sequencing libraries. The use of a dual-index strategy developed by Illumina makes it possible that up to 96 samples can be pooled and sequenced on a single lane of an Illumina sequencer.[70] However, no early multiplexing of the methods deeply limits the V<sub>H</sub>:V<sub>L</sub> paired sequencing throughput, and the additional PCR purification step may lead to the loss of material.

The high-throughput single-cell BCR sequencing (BCR-Seq) technology uses 10X Genomics Chromium, a droplet-based technique, and could obtain auto paired V<sub>H</sub> and V<sub>L</sub> sequences from tens of thousands single B cells in one run, has outperformed in Ig-Seq.[28-54,61-72] After a B cell sorting, each inputted cell is encapsulated with a 10X barcoded gel bead in a single partition by the 10X Chromium Controller. Within each nanoliter-scale partition, mRNA from cells undergoes reverse transcription to generate cDNA, where all cDNA from individual cells share a common 10X barcode, thus allowing the access of V<sub>H</sub> and V<sub>L</sub> paired sequences (Figure 3). The resulting 10X barcoded library is compatible with standard NGS short-read sequencing on Illumina sequencers, and the full-length V(D)J sequences are assembled.[73]

High-throughput single-cell RNA-Seq methods allow paired V<sub>H</sub> and V<sub>L</sub> derived from thousands of cells to be captured and sequenced quickly, but only generate short reads from one end of a cDNA template, which limits the reconstruction of highly diverse full-length antibody variable regions.[74] A novel method, termed Repertoire and Gene Expression by Sequencing (RAGE-Seq) was developed to overcome this limitation.[75] In brief, droplet-based single-cell BCR-Seq is used to generate a barcoded cDNA library, which is split and simultaneously subjected to short-read sequencing for 3’ expression profiling and targeted capture using custom probes followed by long-read sequencing. The short-read sequencing is used to generate highly accurate cell-barcode sequences, which permit demultiplexing of the long-read data. Demultiplexed long-reads are subjected to de novo assembly and error correction to generate full-length V<sub>H</sub> and V<sub>L</sub> sequences. Transcriptome profiles generated from short-read sequencing can then be linked to the antigen-receptor sequence for each individual cell.

3.2. Antigen-Specific Ig-Seq

Ig-Seq is a powerful tool for interrogating immune responses to infection and vaccination reflected by the total antibody repertoire, but it provides limited information about the antigen specificity of the sequenced BCRs. The antigen enriched antibodies identified by the combination of LC-MS/MS and Ig-Seq as described above[56] is feasible in the study of antigen-specific immunity functions, natively paired V<sub>H</sub> and V<sub>L</sub> genes may be expected in most antibody repertoire studies.
BCRs sequencing. Additionally, recent technical improvements make single-cell sequencing an invaluable tool for describing complex cell populations, as well as T cells or B cells, which facilitates the bulk antibody sequencing of antigen-enriched B cells. As discussed above, single-cell sequencing advancement in bulk VH:VL paired sequencing proved invaluable in the antibody repertoire deep mining. Since the input cell count restriction of the single-cell sequencing platform necessitates a presorting for the millions of sampled B cells, antigen-specific B cells is thus widely adopted for further sequencing. For example, using fluorescence-labeled OVA (Ovalbumin) conjugated to a secondary antibody, antigen-specific B cells from lymph nodes of immunized animals were OVA-positive-selected by FACS, and then VH and VL were sequenced using 10X Genomics platform and natively paired according to the identical UMIs. More recently, antigen-specific single-cell Ig-Seq was successfully applied to the development of antibodies against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). In this study, biotinylated SARS-CoV-2 RBD (receptor-binding domain) recombinant proteins were coupled with streptavidin dynabeads for antigen-specific B cell enrichment, using the 10X Genomics platform, similar procedures were carried out to determine the selected paired VH and VL sequences, and potent neutralized antibodies were finally identified (Figure 3).

Using a 10X Genomics Chromium Controller platform, Setliff et al. have developed an antigen-specific Ig-Seq method, LIBRA-Seq (linking B cell receptor to antigen specificity through sequencing). First, fluorescently labeled, distinctive DNA-barcode antigens were used to sort antigen-positive B cells by fluorescence-activated cell sorting (FACS). Then, sorted single B cells were lysed, and cDNA was reverse transcribed sharing a common barcode. After barcoded cDNA from all cells was mixed freely, target VH and VL libraries were enriched using PCR, and the libraries were sequenced (Figure 3). Finally, each antigen specific paired VH and VL sequences were distinguished using their adjunct barcodes. This novel assay enables the detection of monoclonal antibodies specific to panels of diverse antigens, thus facilitating the rapid identification of cross-reactive antibodies and cost-effective antibody development for multiple antigenic pathogenic diseases. The LIBRA-Seq technology has been commercialized and will potentially be a practical tool to delineate antigen-specific antibody repertoires that may serve as therapeutic agents or vaccine templates. Predictably, as a more precise antibody repertoire is generated using novel antigen-specific sequencing strategies, subsequent applications of this technology will be encouraged.

### 3.3. Third-Generation Sequencing

Generally, VH or VL gene with an approximate size of 400 bp can be covered by NGS platforms. However, challenges remain for the conventional NGS sequencing of the single-chain variable fragment (scFv) and Fab from combinatorial libraries due to their approximately 750 and 1700 bp lengths, respectively. In this regard, the third-generation sequencing platforms are needed for the long-read antibodies exploring. As for the most widely accepted and used SMRT sequencing platform, when a sample of DNA or cDNA is loaded to a chip, it is diffused into a zero-mode waveguide (ZMW), which allows the observation of only one nucleotide at a time. The four dNTPs are fluorescently linked: as each nucleotide interacts with the DNA polymerase active site, fluorescence is emitted and detected by ZMW. Finally, the fluorescent signals are translated into the sequences by base calling.

Using third-generation sequencing, full-length phage displayed scFv antibody repertoire was characterized, with novel exons and splice variants in the human antibody heavy chain identified, and ultralong CDR3 in bovine antibody heavy chains were detected. Compared with NGS, the unique mechanism of third-generation sequencing makes it possess...
of each individual is in the range of 0.9–1.7 × 10^9. Experiments revealed that circulating heavy-chain clonotypes were shared between two subjects, and 0.3% of antibodies co-exist among all three subjects. In this regard, the repertoire size in twins is probably overrated. In elderly subjects, an increased number of B cells with long heavy chain CDR3 regions, accumulation of highly mutated IgM and IgG, and persistent clonal B cell populations were presented in the blood. Interestingly, the finding that V, D, and J gene usages are similar in young and elderly individuals, is inconsistent with the study of a memory B cell IgM repertoire. In that study, the elderly individuals demonstrated a higher usage of IGHV3 and IGHJ4 and lower IGHD1-IGHJ4 recombination than that of the young individuals. In neonates, V, D, and J gene usage in the IgM repertoire showed no difference compared with adults, but the length of heavy chain CDR3 was shorter than that in the healthy human repertoire. Notably, the most prominent difference was the occurrence of an N2 addition (neonates: 64.87% vs adults: 85.69%). Meanwhile, the identification of relatively high frequencies of shared antibody clonotypes between adults and neonates indicated that these clonotypes have persisted for decades.

Gender disparity in immune responses is a well-studied phenomenon, and sex hormone-dependent activation of T cells was shown to play an important role. In B cells, a total of 24 IGHV genes, 9 IGHD genes, and 5 IGHJ genes presented sex-based differences but found no bias in V(D)J gene recombination. Ig-Seq has also enabled the ability to uncover the extent to which genetic background affects the characteristics of an individual’s antibody repertoire. Focusing on monozygotic twins, it was determined that the overall antibody repertoire is profoundly determined by the germ-line genome as both IgM and IgG repertoires showed high similarity in twins. In response to an antigenic stimulus, antibodies showed individual-specific selection. Rubelt et al. further confirmed that heritable mechanisms affect the antibody repertoire, as well as T cell receptors (TCR) repertoire.

4.2. Comparison among Different Populations

Human neonates have not developed a mature immune system to respond to various antigens, and elderly individuals exhibit a decline in immune function, thus age is believed to be a key factor in the humoral immunity. The underlying molecular mechanisms regarding changes in immune function can be partially revealed by Ig-Seq. In elderly subjects, an increased number of B cells with long heavy chain CDR3 regions, accumulation of highly mutated IgM and IgG, and persistent clonal B cell populations were presented in the blood. Interestingly, the finding that V, D, and J gene usages are similar in young and elderly individuals, is inconsistent with the study of a memory B cell IgM repertoire. In that study, the elderly individuals demonstrated a higher usage of IGHV3 and IGHJ4 and lower IGHD1-IGHJ4 recombination than that of the young individuals. In neonates, the V, D, and J gene usage in the IgM repertoire showed no difference compared with adults, but the length of heavy chain CDR3 was shorter than that in the healthy human repertoire. Notably, the most prominent difference was the occurrence of an N2 addition (neonates: 64.87% vs adults: 85.69%). Meanwhile, the identification of relatively high frequencies of shared antibody clonotypes between adults and neonates indicated that these clonotypes have persisted for decades.

4.4. Novel Mechanisms in Antibody Repertoire Development Uncovered by Ig-Seq

Susumu Tonegawa discovered that V(D)J recombination is the genetic mechanism behind the diversity of human antibodies. Subsequently, other mechanisms that contribute to the formation of antibody repertoires were discovered, such as somatic hypermutation, class switch, junctional diversity, and the 12/23 rule. Recently, new mechanisms revealed how B cells exploit cohesin-mediated chromatin loop extrusion to
create new antibodies via V(D)J recombination\textsuperscript{[106]} and class switch.\textsuperscript{[107]} Our lab has found a significantly higher occurrence and longer length of the N2 insertion in adults compared to neonates, suggesting a potential molecular mechanism behind the differences in diversity among antibody repertoires as humans age.\textsuperscript{[99]} Furthermore, in addition to the common V(D)J recombination, V(DD)J recombinations have been discovered, and were detected with a frequency of 1 in 800 circulating B cells.\textsuperscript{[108]} The ultra-long heavy chain CDR3s resulting from J recombination, V(DD)J recombinations have been discovered, and were detected with a frequency of 1 in 800 circulating B cells.\textsuperscript{[108]} The ultra-long heavy chain CDR3s resulting from J recombination, V(DD)J recombinations have been discovered, and were detected with a frequency of 1 in 800 circulating B cells.\textsuperscript{[108]} The ultra-long heavy chain CDR3s resulting from J recombination, V(DD)J recombinations have been discovered, and were detected with a frequency of 1 in 800 circulating B cells.\textsuperscript{[108]} The ultra-long heavy chain CDR3s resulting from J recombination, V(DD)J recombinations have been discovered, and were detected with a frequency of 1 in 800 circulating B cells.

3. Strategies of Functional Antibody Discovery from Antibody Repertoires

In addition to the hybridoma technique, B cell immortalization and microneutralization,\textsuperscript{[111]} large-scale screening of libraries by phage, ribosome, yeast, or bacterial display are widely applied in antibody isolation.\textsuperscript{[112,113,61,114]} Using these technologies, we and others have successfully developed functional antibodies against various targets.\textsuperscript{[115–117]} For instance, human single-domain neutralizing antibodies targeting diverse epitopes of SARS-CoV-2 were recently identified by using phage-display based technology.\textsuperscript{[118]} Flow cytometry-based single-cell sorting and cloning is another effective tool and has also successfully been employed in the development of neutralizing antibodies against SARS-CoV-2.\textsuperscript{[119]} These traditional techniques are reliable, but time-consuming and labor-intensive. Also, single-cell sorting and cloning are limited to analysis of a small fraction of the human antibody repertoire. In contrast, application of Ig-Seq is able to capture the entire authentic human antibodies in circulating B cells. Different approaches that exploit antibody repertoire sequencing analysis for functional antibody isolation directly from humans or animals, without library screening, have been developed (Table 2).

Even without rounds of selection for antigen binding and the VH and VL paired information, bulk BCR sequencing has been used to identify many functional antibodies. By sequencing the antibody variable region genes, the most abundant VH and VL sequences are paired and the recombining antibodies verified with nanomolar affinity for antigens.\textsuperscript{[126]} Combined with phage display technology, the VH clones picked from the antibody repertoire can be paired with the corresponding VL library with potent affinity against seasonal influenza.\textsuperscript{[120]} Based upon sequence similarity to the HIV-1 broadly neutralizing antibody,\textsuperscript{[121]} VH and VL genes generated by bulk BCR sequencing within matching branches in the phylogenetic trees were able to be paired and yield novel HIV-1 broadly neutralizing antibodies.\textsuperscript{[122]} Setliff et al.\textsuperscript{[57]} showed that a public antibody clonotype shared among the naïve VH IgG repertoires and antigen-specific VH:VL paired IgG repertoires encodes antibody with potent binding activity to gp120.

Customized antibody sequences databases produced by Ig-Seq have enabled the proteomic identification of antibodies from serum. In this manner, high-affinity antibodies for human PR A/B,\textsuperscript{[9]} hepatitis B (HBV),\textsuperscript{[58]} and tetanus toxoid (TT),\textsuperscript{[59,67]}

| Methodology | Discovery criteria | Paired VL source | Pathogen | Function | Ref. |
|-------------|-------------------|------------------|----------|----------|------|
| Bulk BCR sequencing | Most abundant VH clones | Most abundant VH | H3N2 | Potent binding | [55] |
| Bulk BCR sequencing | Most abundant clones | Most abundant VL | Ovalbumin, Cls, bright | Potent binding | [56] |
| Bulk BCR sequencing | Phylogenetic relationship to discovered antibody | Matched phylogenetic similarity to discovered antibody | HIV-1 | Potent neutralization | [57] |
| Bulk BCR sequencing + DNA-barcoded VH:VL paired sequencing | Public antibody clonotypes | Supplemental DNA barcoded VH:VL paired sequencing | HIV-1 | Potent binding | [58] |
| Proteomic-based bulk BCR sequencing | Antigen-specific sequences with high confidence | Most abundant VL | PR A/B | Potent binding | [9] |
| V\textsubscript{H}/V\textsubscript{L} linkage PCR-based paired sequencing | Surface antigen binding plasmablasts | / | TT | Potent binding | [59] |
| Proteomic-based V\textsubscript{H}/V\textsubscript{L} paired sequencing | Antigen-specific sequences with high confidence | / | TT | Potent binding | [60] |
| Single-cell BCR sequencing | Vaccination elicited over 50-fold expansion memory B cells | / | Influenza | Broad neutralization | [27] |

Table 2. Functional antibody discovery using Ig-Seq.
and neutralizing antibodies against HIV-1 and human norovirus were accurately and efficiently observed. Applied with bulk BCR sequencing, vaccine-responsive V_{H} clones that expanded dramatically after vaccination were able to be identified; accompanied with single-cell BCR sequencing preserving natively paired V_{H} sequences, broadly anti-influenza antibodies were identified in seasonal influenza vaccinated adult. In more recent studies, antigen-specific single B cell sequencing has been adopted for the identification of antibodies. Using this method, HIV- and influenza-specific antibodies concurrently were identified in HIV infected subjects, OVA potent binding antibodies identified from OVA-immunized animals, and SARS-CoV-2 potent neutralizing antibodies identified from recovered patients.

6. Conclusion

The humoral immune system has evolved to encode an astonishing number of antibodies that collectively comprise the antibody repertoire to recognize virtually any antigen. Ultra-deep Ig-Seq can now enable the full picture of human circulating antibody repertoire diversity. In the last decade, Ig-Seq technologies and related bioinformatics analysis tools have been developed at a breakneck pace, and can now provide fast, unbiased, and almost error-free antibody repertoires, reshaping our understanding of many important aspects of humoral immunology and advancing antibody discovery and vaccine development.

However, major biological gaps remain in the pipeline of Ig-Seq, such as 1) the human tissue-derived B cell source limitation and antibody repertoire derived from peripheral blood lacks representativeness, which is insuperable currently; 2) low-throughput, time-consuming, and high-cost in native antibody chains pairing that is urgently in need of technological innovation; 3) importantly, the missing relationship between antibody sequence and its function currently diminish the value of Ig-Seq and make it labor-intensive to perform functional antibody discovery from antibody repertoires. Gaps remain in the use of antibody repertoire data as no standard experimental annotation and data analysis currently exist. Functional antibodies have only been identified in a small fraction of Ig-Seq studies, thus more effective approaches urgently need to be developed and commercialized. Currently, a total of 1184 BCR and TCR repertoires and 2 billion sequences are deposited in iReceptor available across multiple projects, labs, and institutions. BCR sequences can be extracted from bulk unselected RNA sequencing data using the computational tool TRUST. Many computational tools have also been developed to assemble paired V_{H} and V_{L} antibody sequences from the increasing single-cell RNA-Seq data. Moreover, novel technologies of single protein molecules in nanopores and convergent selection in antibody repertoires by deep learning may profoundly accelerate Ig-Seq development and antibody repertoire deep mining. Overall, it is expected that deep mining of the antibody repertoire will lead to more efficient and faster development of clinical diagnostics and immunological therapeutics.

Acknowledgements

This work was supported by the National Key R&D Program of China (2019YFA0904400), National Megaprojects of China for Major Infectious Diseases (2018ZX10301-403), and National Natural Science Foundation of China (81822027, 81630090, 81902108).

Conflict of Interest

The authors declare no conflict of interest.

Keywords

antibody discovery, antibody repertoire, B cell receptors, high-throughput sequencing, Ig-Seq

Received: June 12, 2020
Revised: August 5, 2020
Published online: September 2, 2020

[1] D. Jung, F. W. Alt, Cell 2004, 116, 299.
[2] S. V. Desiderio, G. D. Yancopoulos, M. Paskind, E. Thomas, M. A. Boss, N. Landau, F. W. Alt, D. Baltimore, Nature 1984, 311, 752.
[3] Z. Li, C. J. Woo, M. D. Iglesias-Ussel, D. Ronai, M. D. Scharff, Genes Dev. 2004, 18, 1.
[4] J. Stavnezer, J. E. Guikema, C. E. Schrader, Annu. Rev. Immunol. 2008, 26, 261.
[5] H. W. Schroeder Jr., Dev. Comp. Immunol. 2006, 30, 119.
[6] G. Georgiou, G. C. Ippolito, J. Beausang, C. E. Busse, H. Wardemann, S. R. Quake, Nat. Biotechnol. 2014, 32, 158.
[7] J. Ye, N. Ma, T. L. Madden, J. M. Ostell, Nucleic Acids Res. 2013, 41, W34.
[8] D. A. Bolotin, S. Poslavsky, I. Mitrophanov, M. Shugay, I. Z. Mamedov, E. V. Putintseva, D. M. Chudakov, Nat. Methods 2015, 12, 380.
[9] W. C. Cheung, S. A. Beausoleil, X. Zhang, S. Sato, S. M. Schieferl, J. S. Wieler, J. G. Beaudet, R. K. Ramenani, L. Popova, M. J. Comb, Nat. Biotechnol. 2012, 30, 447.
[10] A. Vaisman Mentesh, S. Rosenberg, M. Yavzori, Y. Dror, E. Fudim, B. Ungar, U. Kopylov, O. Picard, A. Kigel, S. Ben-Horin, Front. Immunol. 2019, 10, 2921.
[11] Y. Qiu, S. Stegalkina, J. Zhang, E. Boudanova, A. Park, Y. Zhou, P. Prabakaran, S. Pougatcheva, I. V. Ustyugova, T. U. Vogel, S. T. Mundle, R. Oomen, S. Delagrange, T. M. Ross, H. Kleanthous, H. Qiu, J. Virol. 2020, 94, e01035.
[12] J. D. Galson, J. Trück, E. A. Clutterbuck, A. Fowler, V. Cerundolo, A. J. Pollard, G. Lunter, D. F. Kelly, Genome Med. 2016, 8, 68.
[13] A. J. Apostoaei, J. R. Trabalka, Review, Synthesis, and Application of Information on the Human Lymphatic System to Radiation Dosimetry for Chronic Lymphocytic Leukemia, SENES Oak Ridge, Inc., Oak Ridge, TN 2010.
[14] S. M. Lewis, Adv. Immunol. 1994, 56, 27.
[15] F. Melchers, E. Ten Boekel, T. Seidl, X. C. Kong, T. Yamagami, K. Onishi, T. Shimizu, A. G. Rolink, J. Andersson, Immunol. Res. 2000, 175, 33.
[16] J. D. Freeman, R. L. Warren, J. R. Webb, B. H. Nelson, R. A. Holt, Genome Res. 2009, 19, 1817.
[17] E. Waltari, M. Jia, C. S. Jiang, H. Lu, J. Huang, C. Fernandez, A. Finzi, D. E. Kaufmann, M. Markowitz, M. Tsuji, X. Wu, Front. Immunol. 2018, 9, 628.
[18] S. G. Lin, Z. A. Bu, Z. Du, Y. Zhang, J. Hu, F. W. Alt, Proc. Natl. Acad. Sci. U.S.A. 2016, 113, 7846.

[19] T. Kivioja, A. Vähäranta, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, J. Taipale, Nat. Methods 2012, 9, 72.

[20] J. M. Heath, M. Ismail, T. Oakes, B. Chain, Briefings Bioinf. 2018, 19, 554.

[21] U. Klein, R. Küppers, K. Rajewsky, Blood 1997, 98, 1288.

[22] B. Roy, R. S. Neumann, O. Snir, R. Iversen, G. K. Sandve, Proc. Natl. Acad. Sci. U. S. A. 2014, 111, 4928.

[23] C. V. Vollmers, L. Penland, J. N. Kanbar, S. R. Quake, Trends Biotechnol. 2017, 35, 203.

[24] S. R. Quake, Nature 2010, 465, 375.

[25] M. Barba, H. Czosnek, A. Hadidi, E. Alamyar, P. Duroux, M.-P. Lefranc, V. Giudicelli, N. Thomas, J. Heather, W. Ndifon, J. Shawe-Taylor, B. Chain, V. Greiff, Nat. Methods 2018, 15, 1159.

[26] M. C. Nussenzweig, V. Costantini, P. D. Brewer-Jensen, M. L. Mallory, W. N. Voss, J. A. Reuter, D. V. Spacek, M. P. Snyder, K. E. A. Lundin, L. M. Sollid, F. Horns, C. L. Dekker, S. R. Quake, Cell. Rep. 2020, 30, 14049.

[27] C. Ziegenhain, B. Vieth, S. Parekh, A. Guillamet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann, W. Enard, Mol. Cell 2017, 65, 631.
Xiaolong Tian obtained his M.A. degree in animal genetics from the College of Animal Science and Technology, China Agricultural University in 2018. Currently he is a Ph.D. candidate under the supervision of Prof. Tianlei Ying in School of Basic Medical Science, Fudan University. His current research focuses on antibody repertoire and cancer immunotherapy.

Yanling Wu is an associate professor at School of Basic Medical Science, Fudan University. She obtained her Ph.D. degree from School of Basic Medical Science, Fudan University in 2018. Her current research interests are focused on translational medicine research in the fields of biomedicine and antibody engineering.

Tianlei Ying is a full professor at School of Basic Medical Science, Fudan University. He received his B.S. and Ph.D. degree from Fudan University. His research focuses on fully human therapeutic monoclonal for infectious diseases and cancer, the design, engineering, and clinical application of novel antibody constructs, and precision immunology.