Microbiology Spectrum
American Society for Microbiology Press

High-Throughput DNA Sequencing Analysis of Antibody Repertoires

SCOTT D. BOYD[1] and SHILPA A. JOSHI[1]
[1]Department of Pathology, Stanford University, Stanford, CA 94305

ABSTRACT New high-throughput DNA sequencing (HTS) technologies developed in the past decade have begun to be applied to the study of the complex gene rearrangements that encode human antibodies. This article first reviews the genetic features of Ig loci and the HTS technologies that have been applied to human repertoire studies, then discusses key choices for experimental design and data analysis in these experiments and the insights gained in immunological and infectious disease studies with the use of these approaches.

INTRODUCTION

New high-throughput DNA sequencing (HTS) technologies developed in the past decade have rapidly increased the scale of data collection for all aspects of human genetics (1, 2). The complex somatic gene rearrangements of immunoglobulin (Ig) and T-cell antigen receptors (TCRs) in the adaptive immune system are particularly appropriate targets for investigation using these new technologies. The antigen specificity of adaptive human immune responses and the storage of specific immunological memory depend on the sequences of the Ig and TCR gene rearrangements expressed by B cells and T cells. Until recently, the difficulty and cost of obtaining sequence data limited the kinds of immunological research questions that could be studied. Pioneering work examining dozens to hundreds of Ig rearrangements with Sanger sequencing has revealed some overall features of the repertoires of these receptors, while physical selection and sorting of B-cell populations of interest has led to the identification of antibodies specific for a variety of infectious agents and vaccine components. However, given that a single human body contains an estimated 10^{11} B cells representing, at a minimum, millions of distinct clonal populations, experiments using Sanger sequencing were underpowered to evaluate the full scale of antibody repertoires. This chapter first reviews genetic features of Ig loci and the HTS technologies that have been applied to human repertoire studies, then discusses experimental design, data analysis choices in these experiments, and insights gained in immunological and infectious disease studies using these approaches.

ANTIBODY GENE REARRANGEMENTS

Antibodies in humans are protein complexes whose basic unit is a disulfide-linked pair of heavy chain proteins, each with an associated light chain (Fig. 1). The N-terminal regions of heavy and light chains are highly variable in their sequences, and are the antigen-binding portions of the antibody. The C-terminal regions of the proteins are termed the constant regions. Light chains are of two types, kappa (IgK) or lambda (IgL), while heavy chains (IgH) are of five major isotypes (IgM, IgD, IgG, IgA, and IgE), with four subtypes of IgG and two

Received: 15 April 2014, **Accepted:** 8 May 2014,
Published: 10 October 2014
Editors: James E. Crowe, Jr., Vanderbilt University School of Medicine, Nashville, TN; Diana Boraschi, National Research Council, Pisa, Italy; and Rino Rappuoli, Novartis Vaccines, Siena, Italy
Citation: Boyd SD, Joshi SA. 2014. High-throughput DNA sequencing analysis of antibody repertoires. *Microbiol Spectrum* 2(5):AID-0017-2014. doi:10.1128/microbiolspec.AID-0017-2014.
Correspondence: Scott D. Boyd, sboyd1@stanford.edu

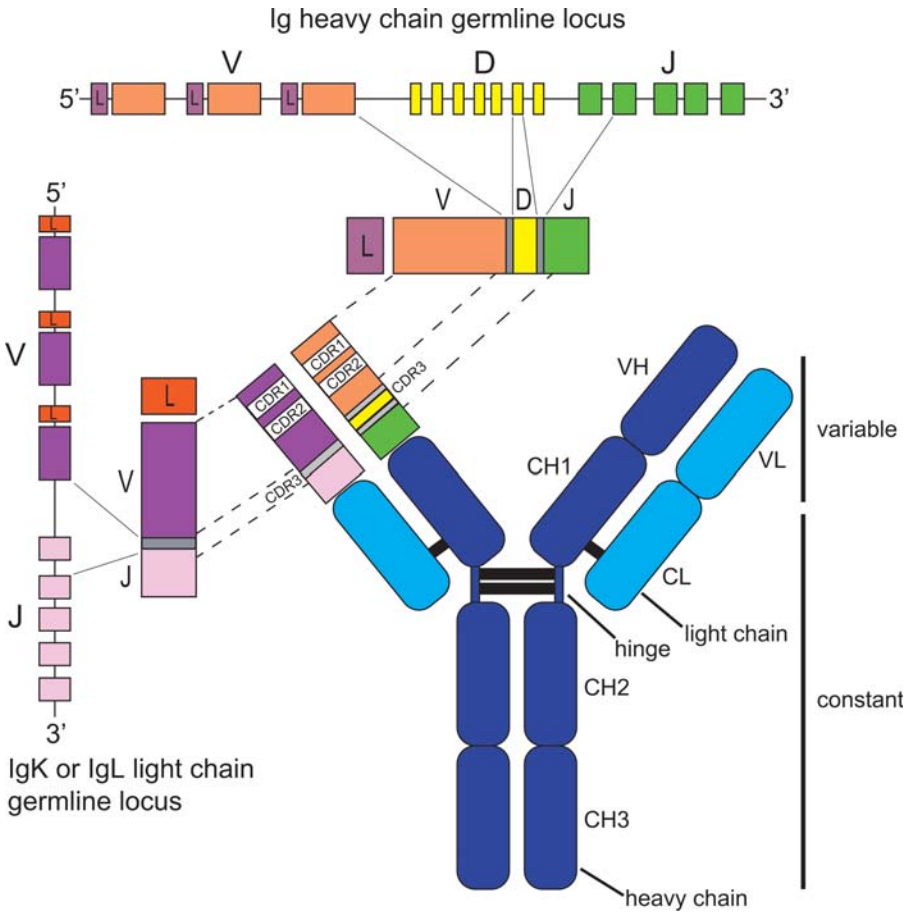

FIGURE 1 Antibody structure and genetic encoding. The germ line (unrearranged) genomic DNA configuration of the immunoglobulin heavy chain locus is depicted at the top of the figure, showing the tandem arrays of V, D, and J gene segments (not to scale). A germ line kappa or lambda light chain locus is depicted on the left-hand side, with unrearranged V and J segments. Stepwise rearrangement of the germ line DNA results in the joining of a heavy chain D and J gene segment, followed by joining of a V segment to the D-J product, to generate the DNA encoding the heavy chain variable region. In the process of rearrangement, the ends of the gene segments are subject to variable amounts of exonuclease digestion, and randomized nontemplated bases are added at the segment ends, to produce additional sequence diversity at the VDJ junctional region that encodes the complementarity-determining region 3 (CDR3) loop, which is often the region of the antibody heavy chain that has the greatest impact on antigen specificity. A similar process of V and J gene rearrangement with diversification of the VJ junction occurs in the light chain locus, to produce the rearranged light chain gene. The constant regions of the heavy and light chains (domains CH1, CH2, and CH3 for the heavy chain, and CL for the light chain) are encoded by downstream exons that are joined to the rearranged V(D)J gene by mRNA splicing. Disulfide bridges joining protein chains in the full antibody structure are shown with black line segments. doi:10.1128/microbiolspec.AID-0017-2014.f1

subtypes of IgA. Developing B cells assemble the genes encoding the antigen-binding regions of their immunoglobulin heavy chains from germ line arrays of variable (V), diversity (D), and joining (J) gene segments, by first joining a D and J segment together, and then selecting a V segment to join to the newly generated D-J product. During joining, the ends of the gene segments are subject to exonuclease digestion, and nontemplated randomized bases (N bases) are added at the segment junctions. In the developing B cell, an initial attempt to rearrange the immunoglobulin heavy chain locus on one copy of chromosome 14 can potentially give rise to an out-of-frame product; if this happens, then the B cell attempts to rearrange the other copy of the locus. After a productive

IgH is expressed, an analogous rearrangement of V and J segments at the kappa or lambda immunoglobulin light chain locus takes place. The mechanistic details and regulation of these events have been reviewed elsewhere (3). The region of the antibody heavy or light chain structure encoded by the fusion of V, (D), and J segments forms a loop that is the most diverse region of the antibody and is termed the complementarity-determining region 3 (CDR3). CDR3 is often the most important portion of the antibody for binding to antigen, but two other loops (CDR1 and CDR2) encoded by the variable gene segment can also contribute significantly to antigen specificity.

The genetic complexity of the antibody repertoire in an individual's B cells prior to exposure to antigen derives from the combinatorial assembly of V, (D), and J segments, the diversification at the segment junctions encoding the CDR3 loops of the antibody heavy or light chain, and the pairwise combination of heavy and light chains. Earlier rough estimates of the potential number of different antibodies that can be generated by these mechanisms were greater than 10^{11} (4). Such figures are undoubtedly underestimates, because the theoretical number of unique heavy or light chains is limited only by the maximal length of the nontemplated junctional base sequences considered in the calculation. Antigen-stimulated B cells activate an additional program of antibody sequence diversification called somatic hypermutation that generates new point mutations throughout the rearranged V(D)J sequence, creating clonal offspring whose antibodies may possess increased binding affinity for the antigen. Rarer processes such as receptor editing contribute further mechanisms for antibody repertoire generation by enabling secondary rearrangements and replacement of V gene segments. At the level of human populations, additional genetic diversity of antibody repertoires results from allelic variants for immunoglobulin gene segments and structural variants within the immunoglobulin loci that delete or increase the copy number of particular gene segments.

Current understanding of the human germ line DNA sequence for the immunoglobulin heavy chain locus, and the number of V, D, and J segments it contains, remains heavily indebted to the initial sequence generated by Matsuda et al. in 1998 (5). Recent human genome-sequencing efforts have often been less helpful for evaluating immunoglobulin loci, in part, because of the widespread use of oligoclonal Epstein-Barr virus (EBV)-transformed B-cell lines with rearranged Ig genes as the source of DNA for sequencing, resulting in the loss of information about gene segments deleted during rearrangement, as well as the relatively shallow depth of sequencing performed in population-level studies (6). In contrast, HTS has very recently been used to study the germ line locus in particularly tractable human samples, such as cells from a haploid hydatidiform mole, and to survey the haplotypes in this locus in different human population groups (7). The curated sequences in the IMGT database (www.imgt.org) for the human IGH locus at 14q32 indicate that most humans, depending on their haplotypes, have 123 to 129 *IGHV* gene segments, of which 43 to 46 are able to form functional rearrangements; 27 *IGHD* gene segments (23 functional); and 9 *IGHJ* segments (6 functional) (8, 9). The IGK locus at 2p11 encodes 76 *IGKV* in most individuals (31 to 36 functional), and 5 *IGKJ* gene segments (5 functional), while the IGL locus at 22q11 contains 73 to 74 *IGLV* segments (29 to 33 functional) and 7 to 11 *IGLJ* segments (4 to 5 functional) (8, 9). Allelic and copy number variation in Ig loci in different human populations were identified in earlier literature, but the pace of identification of new variants has now accelerated, making it clear that the currently curated variants represent only a small sampling of the total variation that is likely to be present across all human groups (7, 10, 11, 12).

The genetic structure of V gene segments includes a small upstream exon encoding most of a leader peptide, followed by a short intron and a second exon encoding the rest of the leader peptide and the V segment itself. In the unrearranged human germ line loci, the V, D, and J regions are separated from each other by kilobases of sequence, and the constant regions, encoded by single exons (for kappa or lambda) or several exons (for the heavy chain isotypes) are located kilobases downstream of the J segments, with the exception of the lambda locus, where a few J segments are interspersed among the most upstream constant regions. The process of rearranging the genomic DNA at heavy and light chain loci to generate in-frame heavy and light chains results in a compact V(D)J gene that is approximately 400 bases in length.

Somatic hypermutation of human antibodies occurs primarily in the specialized microenvironments of secondary lymphoid tissues, where B cells have access to antigen, specialized dendritic and stromal cells, and T cells that stimulate B cells via soluble mediators and cell-cell contact (13). Human plasmablasts observed after acute antigenic stimulation such as influenza vaccination usually show mutation levels in the range of 5 to 15% in the *IGHV* segment, while memory B cells show somewhat lower mutation levels (14). In unusual

circumstances such as chronic exposure to viral antigens in the perturbed immune systems of HIV-infected individuals, much higher mutation levels (over 30%) can be observed in some antibody lineages (15). Mutational events are targeted relatively precisely in a region extending from upstream of, or within, the leader sequence through the V(D)J rearrangement, and taper off in the intron separating the J segments from the constant regions (16, 17). Notably, leader sequences in some loci may be less mutated than the rest of the Ig gene rearrangement, as best documented in mouse kappa light chains (17).

HIGH-THROUGHPUT DNA SEQUENCING INSTRUMENTS

Soon after the completion of the first human genome draft sequence in 2001, a technological race between several different companies and academic laboratories led to the development of a handful of competing platforms for determining DNA sequences in a more highly parallel, miniaturized, and efficient manner than was possible with Sanger sequencing. Common features among these methods were the use of miniaturized microwell plates or flow cells that could capture DNA molecules at particular spatial positions, methods of generating locally amplified template from single DNA molecules for sequencing, and a method of detection of nucleotide sequence that could occur in parallel for thousands to millions of distinct templates at the same time. The platforms that have been most widely used for Ig sequencing in the published literature have been those from Roche/454, Illumina, and Ion Torrent. A brief overview of the features of these instruments is given below. Other promising technologies, including true single-molecule sequencing approaches, have been less widely used so far.

In brief, the Roche/454 and Ion Torrent methods spatially separate and amplify single-template molecules by limiting dilution of the template in aqueous solution followed by production of an aqueous-in-oil emulsion. The template is diluted so that less than one template molecule is present on average per aqueous droplet. The aqueous phase contains the enzymes, primers, nucleotides, and buffers for PCR, as well as capture beads to which the amplified template becomes attached. After PCR, the emulsion is broken and the beads are positioned in microfabricated wells of a plate. Both the Roche/454 and Ion Torrent platforms use a sequencing-by-synthesis approach in which reaction mixtures containing only one of the deoxyribonucleotide triphosphates (dATP, dCTP, dGTP, or dTTP) are sequentially used in cycles of extension of a DNA strand complementary to the template. In 454 sequencing, incorporation of one or more nucleotides during an extension step is detected by a coupled enzymatic reaction in the sequencing plate well, in which pyrophosphate liberated from the nucleotide triphosphate added to the growing DNA strand drives the generation of photons of light by luciferase enzyme (18). The Ion Torrent platform detects nucleotide incorporation during DNA synthesis via miniaturized ion sensors in the sequencing plate that detect the release of protons occurring upon nucleotide addition (19). In both the Roche/454 and Ion Torrent protocols, homopolymer tracts of a particular nucleotide are sequenced with decreased accuracy, because it is more difficult to distinguish between the levels of signal produced by, for example, 14 incorporation events compared with 15 incorporation events in a sequencing step. Error rates are discussed in greater detail below. Current versions of these instruments give approximately one million sequences with read lengths of over 500 bases in the case of Roche/454, and 10 million sequences of 100 bases for the Ion Torrent PGM. The majority of the published literature on HTS of immunoglobulin genes has used the Roche/454 methodology, but Roche has indicated that it is no longer developing the 454 platform and will not support it in the future.

In contrast, the Illumina platform, which has now become the dominant HTS method for most applications, performs template separation and amplification of diluted template DNA on the surface of a flow cell (20). The flow cell is functionalized with oligonucleotides that capture the template and prime "bridging PCR" amplifications that use linker sequences introduced at the ends of the template during library construction. The bridging PCR yields focal clusters of amplified template molecules that are covalently attached to the flow cell. The amplified clusters are then sequenced with cycles of single-base extension sequencing-by-synthesis. In this strategy, each nucleotide is labeled with a fluorophore that identifies the incorporated nucleotide and blocks further extension of the template once the nucleotide is added to the growing DNA strand. A mixture of all four nucleotide triphosphates is used in each cycle of synthesis. Nucleotide incorporation into each template cluster during each sequencing cycle can be detected by imaging the flow cell and detecting what fluorophore is present. Once the flow cell is imaged, the fluorescent labels are cleaved off and the next cycle of extension and imaging begins. Until recently, a disadvantage of the Illumina instruments was their relatively short read

lengths (initially 35 bases), but continual improvements in the methodology have led to current instruments and kits that give total read lengths of sequence that are comparable to the those of the 454 platform. The Illumina MiSeq instrument can sequence approximately 15 million templates per run, yielding up to 300 bases of sequence from each end of the template molecules, while the HiSeq 2500 model can sequence 2 billion templates, reading 125 bases from each end of the template.

EXPERIMENTAL STRATEGIES FOR SEQUENCING Ig REPERTOIRES

Sequencing Ig V(D)J rearrangements in human samples is conceptually straightforward. As a result of the close juxtaposition of V, (D), and J gene segments following genomic rearrangement, short V(D)J PCR amplicons (<500 bp, excluding the lengths of sequencing instrument linkers and barcodes) can be amplified efficiently from genomic DNA or complementary DNA from samples containing B cells. The kilobases of sequence that separate the gene segments in the unrearranged genomic DNA prevent significant amplification from non-B-cell templates in the sample. The incorporation of oligomer nucleotide "barcodes" in the primers or linkers used in library preparation can be used for sample multiplexing in a single sequencing run with any of the HTS instruments, permitting samples to be sequenced together but allowing the sequencing data to be assigned back to the samples by sorting based on the unique sample barcodes. Beyond these similarities, a number of important experimental design choices influence the kinds of interpretation that can be reliably made from sequencing libraries generated by using different approaches, as detailed below.

Cell Populations

High-quality DNA sequencing libraries of V(D)J rearrangements can be produced from templates isolated from total leukocytes, peripheral blood mononuclear cells (PBMCs), or any tissue containing B cells. In order to identify sequences that are derived from B cells that have undergone somatic mutation, the sequences in such libraries can be evaluated for the presence of mutated nucleotide positions in V, D, or J gene segments (21). However, a number of specialized B-cell subsets in humans, such as memory B cells, plasmablasts, transitional B cells, and suppressor B cells, among others, have been defined on the basis of cell surface receptors, intracellular protein expression, or other phenotypic features, and there can be additional experimental value added by using flow cytometry or other methods to sort out particular B-cell or plasma cell subsets of interest prior to generating libraries (14, 22, 23, 24). It is clear, however, that the cell subset definitions and the functions ascribed to them are subject to progressive refinement and revision over time in the immunological literature, meaning that comparisons of Ig repertoire data based on cell subsets should be made with caution unless identical protocols and definitions are used (22). As a practical matter, sorting of small cell subsets can also make it more difficult to isolate nucleic acids for sequencing with the use of standard methods and can require additional rounds of PCR to generate enough amplified material for accurate quantitation prior to sequencing. Another consideration in human studies is that sometimes samples that are precious because of the rarity of the clinical phenotype, pathogen exposure, or other immunological stimulus may only be available in the form of frozen nonviable cells or leukocyte- or PBMC-derived RNA or DNA, owing to the sample collection and storage protocols or resource limitations in clinical studies. In such cases, valuable data can still be obtained from total Ig repertoires without cell surface phenotype information, and antibody isotype expression and mutation status can be interpreted.

Targeted PCR versus 5′ RACE

The 5′ rapid amplification of cDNA ends (5′ RACE) for library preparation requires only a single primer to hybridize to a known region of the target mRNA. cDNA synthesis proceeds until the 5′ end of the target mRNA is reached, and then one of several approaches can be used to add a known, unrelated primer sequence to the 3′ end of the cDNA strand, permitting subsequent PCR amplification (25). This method has begun to be applied to generate Ig HTS libraries, and may be less subject to primer bias and multiplexed PCR artifacts that can result from using primers within the V, J, or constant regions, while also being better at amplifying heavily mutated Ig sequences where primer binding sites may be altered (26, 27). Limitations of 5′ RACE methods are the requirement for RNA as the starting material, precluding analysis of genomic DNA rearrangements, and more variable performance with RNA of suboptimal quality or limiting quantity. The alternative approach of using multiplexed primer mixtures targeting the V, D, J, or constant regions, and performing PCR from either cDNA or genomic DNA template, has been the predominant method used in the literature to date (Fig. 2). This method is subject to a few potential problems, including PCR amplification bias, as a result of different

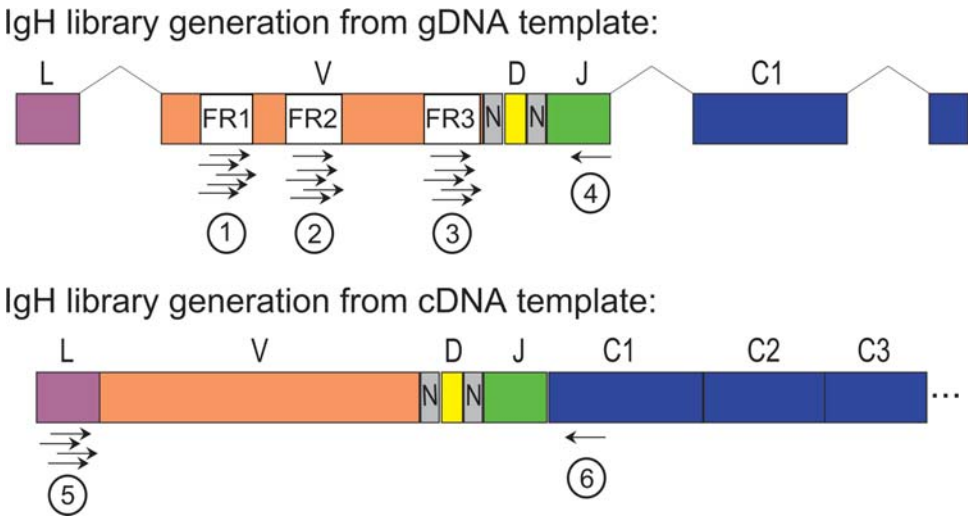

FIGURE 2 IgH library production from genomic DNA (gDNA) or complementary DNA (cDNA). The top diagram shows the rearranged gDNA encoding an antibody heavy chain. Primer sets designed to hybridize in the framework 1, 2, or 3 (FR1, FR2, FR3) regions (labeled with circled numbers 1, 2, and 3), together with a primer complementary to the J gene segments (labeled 4), can be used for PCR to amplify the VDJ gene rearrangement. Multiple primers are shown for the framework primer sets, indicating the different primers required for amplification of V segments belonging to different families. The leader peptide exon is separated from the V segment by a short intron in the genomic DNA. The lower diagram shows cDNA generated from spliced mRNA encoding a heavy chain. Primers hybridizing to the constant region (labeled 6) can be used as the initial gene-specific primer for 5′ RACE protocols (see main text), or else can be used in PCR with primers in the leader sequences (labeled 5), or framework primers, to amplify the VDJ gene rearrangements. The constant region isotype associated with the VDJ gene rearrangement can be identified in such libraries. doi:10.1128/microbiolspec.AID-0017-2014.f2

primer annealing efficiencies to particular gene segments; the need to select primers that can be multiplexed without producing off-target primer dimers or other artifacts; and potential loss of highly mutated sequences from sequencing libraries as a result of the somatic mutation preventing primer annealing. Selection of primers for human Ig rearrangements has focused on the framework (FR) regions FR1, FR2, and FR3, which are relatively less mutated than the CDR sequences in B cells that survive to be observed in human samples, likely because mutated framework regions that destabilize the antibody structure are not compatible with the persistence of the B cell in the body (28, 29, 30). Primers targeting the leader sequences have also been used in a number of publications, particularly those studying highly mutated antibody lineages such as broadly neutralizing anti-HIV antibodies (31, 32).

Choice of Template

Ig sequencing libraries can be generated from genomic DNA template, typically with V-and J-segment primers, or from cDNA template using V-leader primers, internal V primers, J primers, or constant region primers (Fig. 2). In addition to the considerations related to potential effects of somatic mutation on primer binding, as outlined above, there are several other reasons to select either genomic DNA template, cDNA template, or both, depending on the goals of a particular experiment. One advantage of genomic DNA template is that it permits analysis of both productively rearranged Ig sequences that yield a protein product, as well as unproductively rearranged loci where a stop codon or reading frame incompatibility between the V segment and the downstream gene segments has been introduced during V(D)J rearrangement. While not giving rise to functional protein, the unproductive rearrangements provide a record of the features of the Ig locus rearrangement, such as gene-segment rearrangement frequencies, segment chewback by exonucleases, and nontemplated base addition levels prior to selection of the B cell based on expressed Ig protein. This can be useful in evaluating immunological phenotypes related to receptor gene rearrangements or B-cell selection, such as in immunodeficiency disorders (K. Roskin, submitted for publication). In addition, the

fact that there is only one productively rearranged IGH, IGK, or IGL locus per cell can be used to evaluate the sizes of B-cell clones in the cell populations studied, because multiple replicate libraries for sequencing can be generated from independent aliquots of genomic DNA template, and V(D)J rearrangements from different members of the same clone of B cells can later be identified in data analysis (21, 33). In contrast, the many copies of Ig mRNA produced within a single B cell, and the differing levels of mRNA expression for Ig genes in different subsets of B cells, such as relatively low expression in naïve cells and high expression in plasmablasts, mean that it is not possible to reliably distinguish between the presence of cells with high Ig mRNA expression versus the presence of a clone containing many cells with lower Ig mRNA expression, if only a single RNA sample is available for analysis. However, using RNA as the starting material for repertoire sequencing experiments has other important advantages, chiefly that it enables identification of the heavy chain isotype associated with a particular V(D)J rearrangement. Another potential advantage of RNA template is that it should preserve a more complete representation of the B-cell clones present in the cell sample, because each B cell can contribute multiple mRNA copies of its Ig rearrangements, so that, if some template is lost during purification owing to less than 100% yield, there will not be a linear reduction in the number of B-cell clones represented in the sample, as is the case with genomic DNA. Finally, cDNA contains more amplifiable Ig templates per nanogram of template amplified, compared with genomic DNA where the rearranged Ig loci are a tiny fraction of the total DNA quantity. Because there is a limit to the amount of DNA template that can be added to a PCR reaction, more complex libraries of Ig rearrangements therefore can be generated with fewer PCR reactions by the use of the cDNA template compared with the genomic DNA template.

Multiplexing and Chimeric Sequences

Another kind of artifact that can arise in library preparation is the chimeric sequence, usually generated when a DNA strand is incompletely extended in one PCR cycle, and can then hybridize to an unrelated template and be extended further in a subsequent cycle. If PCR amplifications are performed with several different heavy chain isotype primers in the same reaction, such chimeras may lead to errors of assignment of particular VDJ rearrangements to particular isotypes. Otherwise, they can yield apparent regions of highly increased somatic mutation in the V(D)J rearrangement, and other artifactual results. Bioinformatic filtering to remove sequences with such unusual features is one way of attempting to minimize such effects, but other approaches such as performing PCR with templates at limiting dilution in aqueous-in-oil emulsions or microtiter plates have been proposed and used as alternatives, although these dilution methods raise other challenges of sample throughput (34, 35).

Replicate Library Preparation

As noted above, the preparation of multiple libraries from a sample by using independent template aliquots (either by using a genomic DNA template, or by separating cells into separate aliquots prior to isolating RNA template) permits reliable detection of expanded clonal B-cell populations and distinction between an expanded clone of cells versus a single cell expressing high levels of Ig mRNA (21, 33). As with any experimental methodology, conducting experiments with replicate sampling and library generation helps to distinguish between real signals compared with statistical noise in the dataset and also gives a basis for assessing the robustness of any other secondary results derived from the data.

Error Correction Strategies

The combined frequency of PCR errors and sequencing errors for Ig amplicon libraries sequenced with current HTS platforms are usually less than 0.5% per base, with the exception of polynucleotide tract regions in data gathered with 454/Roche or Ion Torrent instruments. A recent comparison of error rates of benchtop HTS instruments sequencing *Escherichia coli* genomic DNA found that the insertion and deletion (indel) rate was 0.38% per base for the 454 GS Junior, 1.5% per base for the Ion Torrent PGM, and <0.001% per base for the Illumina MiSeq (36). In this comparison, the substitution error rate of the MiSeq instrument was the lowest measured, at 0.1 per 100 bases. The quality of the sequence data with all of these instruments decreases toward the distal ends of the reads. When applied to Ig templates, reported error rates for the combination of PCR and sequencing error, excluding indels, have been reported in the literature and seen in our own experiments, to be approximately 0.1 to 0.3% per base for the 454 platform and 0.1 to 0.2% for Illumina sequencing (21, 29, 36, 37, 38). One approach to detecting potential PCR or sequencing errors, and distinguishing these from true somatic mutation positions or allelic variants of gene segments, is to obtain manyfold more sequencing reads than the number of template molecules in the original sample. Errors occurring early in PCR cannot be

corrected with this approach, but, together with the evaluation of sequences derived from known templates (such as sequences cloned into plasmids), this approach can be used to correct many sequencing-derived errors, and give an upper bound of remaining PCR and sequencing errors in the data. An alternative method of detecting errors from amplification or sequencing has been adapted for Ig sequencing from earlier approaches of adding highly diverse sequence tags to nucleic acid templates prior to amplification of libraries from the templates (39, 40, 41, 42). Incorporation of a randomized sequence tag in the primer used for reverse transcription of Ig mRNA enables later comparison of the data from reads sharing the same sequence of randomized tag, and helps the inference of which variant bases are likely to be errors in such sequences (38).

Paired Heavy and Light Chain Sequencing

A long-awaited experimental breakthrough in Ig repertoire studies has been the efficient high-throughput sequencing of native pairs of heavy and light chain (H+L) sequences from large numbers of individual B cells. In 2013, the proof of concept of a method of generating paired H+L sequence libraries from tens of thousands of individual B cells was described, making use of high-density microtiter plates with 125-pl well volumes to separate the cells and enable capture of mRNA from each cell on an oligo-dT functionalized polymer bead, followed by reverse transcription and joining PCR to covalently link the heavy and light chain sequences (43). While the data sets from these initial experiments contained thousands of natively paired sequences, it is likely that the protocols will improve their throughput with the use of emulsion strategies or potentially even the use of microfluidic devices (44).

DATA ANALYSIS APPROACHES

A variety of data analysis methods have been applied to Ig sequence data sets, with many laboratories developing their own pipelines of publicly available software, customized scripts and programs, and databases to manage the data. Owing to the numerous experimental strategies for library preparation being explored in the literature, and the idiosyncrasies of each group's data analysis approaches, direct comparison of reported results is challenging. This is not an insurmountable problem, so long as the experimental protocols are described in sufficient detail, and the raw sequence data are made available to other investigators who wish to evaluate the results by using their own analysis pipeline. It is not realistic to expect that all research laboratories will converge on a single experimental and data analysis approach, but full transparency about library preparation methods and software, including sharing of the scripts used for analysis, should enable appropriate verification of reported results. One overall limitation of many published articles to date has been the relatively low number of samples analyzed and the rare to absent consideration of statistical correction for multiple hypothesis testing in data sets that have many different features that can be compared between samples.

Sequence Barcode Analysis, Filtering, Primer Trimming, and Quality Score Use

As with other amplicon-sequencing experiments, data analysis of Ig libraries typically includes early steps of identifying the sequence barcodes that identify the sample of origin of a sequence. Many groups have designed barcodes that are relatively resistant to sequencing error, in that they are different from each other at two or more positions (i.e., are separated by a Hamming distance of two or more). If gene-specific primers are used in the amplification strategy and a primer is not a perfect match to the gene segment it amplified, trimming of primers is necessary to avoid spurious introduction or loss of point mutations. Quality scores for each nucleotide position are a feature calculated by all of the HTS platforms and agree fairly well with actual error rates in experiments performed with known template sequences (36). Quality scores can be used to exclude sequences of excessively poor quality from further analysis; as an alternative approach, such reads can be carried forward in the analysis and removed from consideration later if they do not align well enough to gene segments of the locus of interest.

Alignment and Parsing Programs

In the place of traditional mapping of reads to a reference scaffold, as is performed for genome or exome sequence analysis, the next step in Ig sequence analysis is determining which regions of the rearranged Ig align with germ line V, D, and J gene segments and constant region sequences in the locus that was amplified, parsing the nontemplated regions at the segment junctions and determining the positions of somatic mutation in the sequence. A number of programs have been devised for this purpose, including IgBlast, V-QUEST, and the hidden Markov model-based programs iHMMune-align and SoDA2 (45, 46, 47, 48). In response to the increasing demand for higher-throughput analysis of large

sequence data sets, the IMGT website (www.imgt.org) has introduced a new next-generation sequencing data set portal for sequence alignment and analysis, HighV-QUEST (49). Many of these published analysis programs claim superiority for their approach to sequence analysis, but, in our laboratory, evaluation of the alignments generated by IgBlast, V-QUEST, and iHMMunealign are identical or very similar for most sequences, and differ most in junctional parsing of sequences where the true answer may be unknowable. For example, IGH sequences in which the D segment is very short owing to exonuclease digestion during rearrangement, so that the true germ line D segment identity cannot be determined would be such a case. An important consideration for any of these utilities is the choice of germ line V, D, and J gene segment repertoires used for alignment, because current curated databases are not complete, so depending on the population group studied, rarer and unreported alleles can appear to carry somatic mutations as a result of misalignment to other gene segments (11, 12).

HTS Ig DATA ANALYSIS: APPLICATIONS IN INFECTIOUS DISEASE RESEARCH

A wide variety of secondary analyses can be performed on Ig repertoire data once the sequences are aligned and parsed, depending on the experimental question being asked. Often, particularly if replicate libraries have been generated from a sample, it can be helpful to collapse identical sequences into a single representative, to minimize PCR amplification biases or stochastic PCR jackpots that cause some sequences to appear in increased copy numbers in the sequencing data. Overall features of Ig repertoires, or the repertoires derived from particular B-cell subsets, can be readily determined from the parsed sequence output of any of the alignment utilities described above. Features that have been reported in many studies include gene segment usage frequencies; length, composition, and amino acid sequences of CDR3 regions; mutation levels and distributions; and heavy chain isotype associated with particular V(D)J rearrangements, among other metrics.

Recent articles have provided more detailed insights into baseline human Ig repertoires than were previously known from lower-throughput data, including the individual specific usage frequencies of V, D, and J gene segments, and the presence of copy number variants and allelic variants within particular haplotypes (11, 29, 50). Glanville et al. observed a strong influence of the germ line genome on V, D, and J gene segment usage (probably combining the effects of receptor rearrangement frequencies, and selection acting on the expressed protein) in repertoires sequenced from two pairs of identical twins (51). HTS has also enabled better measurement of the frequencies of uncommon features in Ig repertoires, such as rearrangements using two D segments in the heavy chain and indels generated as a by-product of somatic hypermutation (52, 53, 54). The distinct features of the repertoires of different B-cell subsets, such as class-switched or IgM-expressing memory cells, compared with naïve B cells, at the level of gene segment usage and junctional features, have been reported by several groups (21, 24, 30, 55). Analysis of somatic mutation patterns in Ig rearrangements, and detection of evidence of selection, has also been enhanced in part by the availability of larger data sets for evaluation, and has prompted the development of new tools and interfaces to facilitate analysis of larger sequence sets (56, 57). The influx of much larger datasets of Ig sequences that often contain many representatives of clonally expanded B-cell lineages has also stimulated new approaches for inferring the relationships of descent and genetic inheritance between individual cells within the clone (58).

Determining whether sequences derived from the same B-cell clone are present in different samples of a longitudinal time course, or different tissue sites, sample types, or B-cell subsets, is often of interest and can be approached in several different ways. For initial analyses, we often use a clone definition for comparing two heavy chain sequences requiring that the V and J segments be the same, or of the same subgroup, and that the CDR3 sequence be the same length and match at 80% of the nucleotides. Depending on the experimental question being asked, different thresholds for clonal similarity may be more appropriate. For example, in searching for members of a clonal lineage that matures from the germ line sequence state to a highly mutated state (as in the case of some broadly neutralizing antibodies against HIV), a more permissive definition of CDR3 similarity could be used, and then the putative clone members could be tested for other inherited features such as somatic hypermutation positions. Ensuring that the conclusions drawn in a study are not overly sensitive to minor differences in clonal definitions is a conservative and prudent approach. A number of groups have made striking progress in recent years by using HTS to track the clonal evolution and mutational and selection histories of B-cell lineages making antibodies specific for particular pathogens or vaccine components. Such analyses have been performed most extensively in studies of

HIV, identifying putative germ line ancestral sequences of highly mutated antibody lineages that acquire virus neutralization breadth only after many months to years of chronic infection, and identifying shared sequence features of antibodies that bind to particular epitopes of HIV (31, 32, 58, 59, 60, 61, 62, 63, 64). Similar approaches may offer insights into many other infectious diseases and chronic or recurrent infections.

Quantitation or normalization of the contribution of clonally expanded B-cell populations to an observed repertoire can be performed from multiple-replicate library data, using a modified form of the Gini-Simpson index adapted for HTS data; we have termed this a "clonality score" or "coincidence index," in recent articles. This measure can distinguish between individuals responding to acute Dengue virus infection from those who are convalescent or uninfected (21, 33). The presence of expanded B-cell clones that persist in the circulation of individuals, as measured with a similar clonality score from samples collected a year apart, is significantly associated with EBV infection status in healthy subjects (21). We have also observed that this measure of clonality is closely correlated with the increase in numbers of plasmablasts observed in the blood 7 days after vaccination with inactivated trivalent influenza vaccine, and that the clonality score is correlated with seroconversion following vaccination. Other recent studies of vaccination for influenza or pneumococcus have highlighted sequences that appear at higher copy number in sequencing libraries following vaccination, which may correspond to expanded clones stimulated by the vaccine, or cells producing higher levels of Ig mRNA, or both (37, 38, 65).

Estimation of the size of the B-cell repertoire has been attempted in a number of prior studies, but is subject to high degrees of uncertainty and underbias when extrapolated from peripheral blood samples that typically contain millions of B cells, representing less than 0.1% of the approximately 100 billion B cells in an individual's body (29). Parametric approaches that make assumptions about the distribution of clone sizes in the repertoire may be particularly unreliable, but even nonparametric estimates such as the lower bound of the "Chao 2" metric are likely to underestimate the minimal size of the repertoire (Y. Liu, personal communication, and reference 66). In our view, currently published estimates of naïve B-cell and T-cell receptor diversity are underestimates of the true values; it will await larger data sets collected with multiple independent samples from each human subject to arrive at better estimates in the future.

In light of the enormous size of the potential antibody repertoire, it has been an open question whether different humans raise similar antibodies in response to the same infectious or vaccination challenge. Previous studies using Sanger sequencing identified some circumstances where highly similar "convergent" antibodies could be detected in different people, such as immunization with *Haemophilus influenzae* type b polysaccharide vaccine or pneumococcal vaccine (67, 68). HTS of Ig repertoires has greatly increased the ability to detect such convergent sequences, and pathogen-specific antibody rearrangements have now been detected in patients infected with Dengue virus and HIV (32, 33, 63) as well as in subjects vaccinated with trivalent inactivated influenza (69). It appears that the stimulation of convergent antibodies that are specific for particular pathogens is a widespread phenomenon. Once more data are gathered for a variety of different pathogens and antigens, it may become possible to read an individual's history of pathogen exposure from the Ig sequences in their memory B-cell repertoire.

CONCLUSIONS AND FUTURE DIRECTIONS

The application of HTS methods to the study of B-cell responses and immunoglobulin structure-function relationships is now well started, and most of the cost limitations and technological shortcomings of the earlier generations of HTS instruments have been overcome. Well-designed experimental protocols and data analysis approaches promise to provide unprecedentedly accurate, extensive, and exact tracking of B-cell populations of biological or medical interest in any immunological context. Initial studies in infectious disease topics have provided new knowledge about the numbers and diversity of B-cell clones stimulated by infection or vaccination by a variety of agents, and have been especially useful in documenting the pathways of mutation and selection followed by antibodies responding to HIV. Clear and explicit documentation of the experimental methods used and data analysis approaches applied to Ig HTS data sets, along with sharing of primer sequences, scripts for analysis, and other important experimental features along with raw and processed data, will help to ensure the reproducibility and broader scientific value of studies in this area. As with any other research area generating and analyzing large and complex molecular datasets, Ig sequence analysis will benefit from a greater emphasis on the use of sample sets with larger numbers of individual patients or subjects, collection of longitudinal repeat sampling from individuals

when possible, appropriate statistical analysis to correct for multiple hypothesis testing, and validation of results with training and test sets of data. A particularly promising area is the coupling of HTS studies of immune response genetics (Ig and TCR sequencing, as well as analysis of other host genetic features) with equally powerful genetic analysis of the populations of pathogens seeking to infect the host and evade the immune system. The coming years should offer a ringside seat to researchers intent on observing these fascinating battles.

ACKNOWLEDGMENT

Conflict of interest: We disclose no conflicts.

REFERENCES

1. Gonzaga-Jauregui C, Lupski JR, Gibbs RA. 2012. Human genome sequencing in health and disease. *Annu RevMed* **63:**35–61.

2. Boyd SD. 2013. Diagnostic applications of high-throughput DNA sequencing. *Annu Rev Pathol* **8:**381–410.

3. Jung D, Giallourakis C, Mostoslavsky R, Alt FW. 2006. Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. *Annu Rev Immunol* **24:**541–570.

4. Janeway CAJ, Travers P, Waport M, Shlomchik MJ. 2001. *Immunobiology: The Immune System in Health and Disease*, 5th ed. Garland Science, New York.

5. Matsuda F, Ishii K, Bourvagnet P, Kuma K, Hayashida H, Miyata T, Honjo T. 1998. The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *J Exp Med* **188:**2151–2162.

6. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:**1061–1073.

7. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, Willsey AJ, Joy JB, Scott JK, Graves TA, Wilson RK, Holt RA, Eichler EE, Breden F. 2013. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am J Hum Genet* **92:**530–546.

8. Lefranc MP. 2011. IMGT, the International ImMunoGeneTics Information System. *Cold Spring Harb Protoc* **2011:**595–603.

9. Lefranc MP, Lefranc G. 2001. *The Immunoglobulin FactsBook*. Academic Press, New York, NY.

10. Wang Y, Jackson KJ, Gaeta B, Pomat W, Siba P, Sewell WA, Collins AM. 2011. Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. *Immunogenetics* **63:**259–265.

11. Boyd SD, Gaeta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, Jones CD, Simen BB, Hanczaruk B, Nguyen KD, Nadeau KC, Egholm M, Miklos DB, Zehnder JL, Collins AM. 2010. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J Immunol* **184:**6986–6992.

12. Wang Y, Jackson KJ, Sewell WA, Collins AM. 2008. Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. *Immunol Cell Biol* **86:**111–115.

13. Klein U, Dalla-Favera R. 2008. Germinal centres: role in B-cell physiology and malignancy. *Nat Rev Immunol* **8:**22–33.

14. Wrammert J, Smith K, Miller J, Langley WA, Kokko K, Larsen C, Zheng NY, Mays I, Garman L, Helms C, James J, Air GM, Capra JD, Ahmed R, Wilson PC. 2008. Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature* **453:**667–671.

15. Mascola JR, Haynes BF. 2013. HIV-1 neutralizing antibodies: understanding nature's pathways. *Immunol Rev* **254:**225–244.

16. Lebecque SG, Gearhart PJ. 1990. Boundaries of somatic mutation in rearranged immunoglobulin genes: 5′ boundary is near the promoter, and 3′ boundary is approximately 1 kb from V(D)J gene. *J Exp Med* **172:**1717–1727.

17. Rada C, Gonzalez-Fernandez A, Jarvis JM, Milstein C. 1994. The 5′ boundary of somatic hypermutation in a V kappa gene is in the leader intron. *Eur J Immunol* **24:**1453–1457.

18. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:**376–380.

19. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova M, Miao X, Reed B, Sabina J, Feierstein E, Schorn M, Alanjary M, Dimalanta E, Dressman D, Kasinskas R, Sokolsky T, Fidanza JA, Namsaraev E, McKernan KJ, Williams A, Roth GT, Bustillo J. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475:**348–352.

20. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara ECM, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456:**53–59.

21. Wang C, Liu Y, Xu LT, Jackson KJ, Roskin KM, Pham TD, Laserson J, Marshall EL, Seo K, Lee JY, Furman D, Koller D, Dekker CL, Davis MM, Fire AZ, Boyd SD. 2014. Effects of aging, cytomegalovirus infection, and EBV infection on human B cell repertoires. *J Immunol* **192:**603–611.

22. Jackson SM, Wilson PC, James JA, Capra JD. 2008. Human B cell subsets. *Adv Immunol* **98:**151–224.

23. Mauri C, Blair PA. 2010. Regulatory B cells in autoimmunity: developments and controversies. *Nat Rev Rheumatol* **6:**636–643.

24. Mroczek ES, Ippolito GC, Rogosch T, Hoi KH, Hwangpo TA, Brand MG, Zhuang Y, Liu CR, Schneider DA, Zemlin M, Brown EE, Georgiou G, Schroeder HW, Jr. 2014. Differences in the composition of the human antibody repertoire by B cell subsets in the blood. *Front Immunol* **5:**96. doi:10.3389/fimmu.2014.00096.

25. Frohman MA, Dush MK, Martin GR. 1988. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc Natl Acad Sci USA* **85:**8998–9002.

26. Aoki-Ota M, Torkamani A, Ota T, Schork N, Nemazee D. 2012. Skewed primary Igkappa repertoire and V-J joining in C57BL/6 mice: implications for recombination accessibility and receptor editing. *J Immunol* **188:**2305–2315.

27. Choi NM, Loguercio S, Verma-Gaur J, Degner SC, Torkamani A, Su AI, Oltz EM, Artyomov M, Feeney AJ. 2013. Deep sequencing of the murine IgH repertoire reveals complex regulation of nonrandom v gene rearrangement frequencies. *J Immunol* **191:**2393–2402.

28. van Dongen JJ, Langerak AW, Bruggemann M, Evans PA, Hummel M, Lavender FL, Delabesse E, Davi F, Schuuring E, Garcia-Sanz R, van Krieken JH, Droese J, Gonzalez D, Bastard C, White HE, Spaargaren M, Gonzalez M, Parreira A, Smith JL, Morgan GJ, Kneba M, Macintyre EA. 2003. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia* **17:**2257–2317.

29. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, Jones CD, Simen BB, Hanczaruk B, Nguyen KD, Nadeau KC, Egholm M, Miklos DB, Zehnder JL, Fire AZ. 2009. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci Transl Med* **1:**12ra23. doi:10.1126/scitranslmed.3000540.

30. Briney BS, Willis JR, McKinney BA, Crowe JE, Jr. 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.

31. Wu X, Zhou T, Zhu J, Zhang B, Georgiev I, Wang C, Chen X, Longo NS, Louder M, McKee K, O'Dell S, Perfetto S, Schmidt SD, Shi W, Wu L, Yang Y, Yang ZY, Yang Z, Zhang Z, Bonsignori M, Crump JA, Kapiga SH, Sam NE, Haynes BF, Simek M, Burton DR, Koff WC, Doria-Rose NA, Connors M, Mullikin JC, Nabel GJ, Roederer M, Shapiro L, Kwong PD, Mascola JR. 2011. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* **333:**1593–1602.

32. Scheid JF, Mouquet H, Ueberheide B, Diskin R, Klein F, Oliveira TY, Pietzsch J, Fenyo D, Abadir A, Velinzon K, Hurley A, Myung S, Boulad F, Poignard P, Burton DR, Pereyra F, Ho DD, Walker BD, Seaman MS, Bjorkman PJ, Chait BT, Nussenzweig MC. 2011. Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science* **333:**1633–1637.

33. Parameswaran P, Liu Y, Roskin KM, Jackson KK, Dixit VP, Lee JY, Artiles KL, Zompi S, Vargas MJ, Simen BB, Hanczaruk B, McGowan KR, Tariq MA, Pourmand N, Koller D, Balmaseda A, Boyd SD, Harris E, Fire AZ. 2013. Convergent antibody signatures in human dengue. *Cell Host Microbe* **13:**691–700.

34. Rubelt F, Sievert V, Knaust F, Diener C, Lim TS, Skriner K, Klipp E, Reinhardt R, Lehrach H, Konthur Z. 2012. Onset of immune senescence defined by unbiased pyrosequencing of human immunoglobulin mRNA repertoires. *PloS One* **7:**e49774. doi:10.1371/journal.pone.0049774.

35. Tan YC, Blum LK, Kongpachith S, Ju CH, Cai X, Lindstrom TM, Sokolove J, Robinson WH. 2014. High-throughput sequencing of natively paired antibody chains provides evidence for original antigenic sin shaping the antibody response to influenza vaccination. *Clin Immunol* **151:**55–65.

36. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ. 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* **30:**434–439.

37. Jiang N, He J, Weinstein JA, Penland L, Sasaki S, He XS, Dekker CL, Zheng NY, Huang M, Sullivan M, Wilson PC, Greenberg HB, Davis MM, Fisher DS, Quake SR. 2013. Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci Transl Med* **5:**171ra119. doi:10.1126/scitranslmed.3004794.

38. 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 USA* **110:**13463–13468.

39. Shiroguchi K, Jia TZ, Sims PA, Xie XS. 2012. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc Natl Acad Sci USA* **109:**1347–1352.

40. Miner BE, Stoger RJ, Burden AF, Laird CD, Hansen RS. 2004. Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. *Nucleic Acids Res* **32:**e135. doi:10.1093/nar/gnh132.

41. McCloskey ML, Stoger R, Hansen RS, Laird CD. 2007. Encoding PCR products with batch-stamps and barcodes. *Biochem Genet* **45:**761–767.

42. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. 2011. Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci USA* **108:**9530–9535.

43. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, Varadarajan N, Giesecke C, Dorner T, Andrews SF, Wilson PC, Hunicke-Smith SP, Willson CG, Ellington AD, Georgiou G. 2013. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol* **31:**166–169.

44. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. 2014. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat Biotechnol* **32:**158–168.

45. Munshaw S, Kepler TB. 2010. SoDA2: a Hidden Markov Model approach for identification of immunoglobulin rearrangements. *Bioinformatics* **26:**867–872.

46. Gaeta BA, Malming HR, Jackson KJ, Bain ME, Wilson P, Collins AM. 2007. iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics* **23:**1580–1587.

47. Giudicelli V, Brochet X, Lefranc MP. 2011. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc* **2011:**695–715.

48. Ye J, Ma N, Madden TL, Ostell JM. 2013. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* **41:**W34–W40.

49. Li S, Lefranc MP, Miles JJ, Alamyar E, Giudicelli V, Duroux P, Freeman JD, Corbin VD, Scheerlinck JP, Frohman MA, Cameron PU, Plebanski M, Loveland B, Burrows SR, Papenfuss AT, Gowans EJ. 2013. IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nat Commun* **4:**2333. doi:10.1038/ncomms3333.

50. Kidd MJ, Chen Z, Wang Y, Jackson KJ, Zhang L, Boyd SD, Fire AZ, Tanaka MM, Gaeta BA, Collins AM. 2012. The inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements. *J Immunol* **188:**1333–1340.

51. Glanville J, Kuo TC, von Budingen HC, Guey L, Berka J, Sundar PD, Huerta G, Mehta GR, Oksenberg JR, Hauser SL, Cox DR, Rajpal A, Pons J. 2011. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc Natl Acad Sci USA* **108:**20066–20071.

52. Briney BS, Willis JR, Crowe JE, Jr. 2012. Location and length distribution of somatic hypermutation-associated DNA insertions and deletions reveals regions of antibody structural plasticity. *Genes Immun* **13:**523–529.

53. Briney BS, Willis JR, Hicar MD, Thomas JW, II, Crowe JE, Jr. 2012. Frequency and genetic characterization of V(DD)J recombinants in the human peripheral blood antibody repertoire. *Immunology* **137:**56–64.

54. Larimore K, McCormick MW, Robins HS, Greenberg PD. 2012. Shaping of human germline IgH repertoires revealed by deep sequencing. *J Immunol* **189:**3221–3230.

55. Wu YC, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. 2010. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* **116:**1070–1078.

56. Uduman M, Yaari G, Hershberg U, Stern JA, Shlomchik MJ, Kleinstein SH. 2011. Detecting selection in immunoglobulin sequences. *Nucleic Acids Res* **39:**W499–W504.

57. Yaari G, Uduman M, Kleinstein SH. 2012. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. *Nucleic Acids Res* **40:**e134. doi:10.1093/nar/gks457.

58. Sok D, Laserson U, Laserson J, Liu Y, Vigneault F, Julien JP, Briney B, Ramos A, Saye KF, Le K, Mahan A, Wang S, Kardar M, Yaari G, Walker LM, Simen BB, St John EP, Chan-Hui PY, Swiderek K, Kleinstein SH, Alter G, Seaman MS, Chakraborty AK, Koller D, Wilson IA, Church GM, Burton DR, Poignard P. 2013. The effects of somatic hypermutation on neutralization and binding in the PGT121 family of broadly neutralizing HIV antibodies. *PLoS Pathog* **9:**e1003754. doi:10.1371/journal.ppat.1003754.

59. Liao HX, Chen X, Munshaw S, Zhang R, Marshall DJ, Vandergrift N, Whitesides JF, Lu X, Yu JS, Hwang KK, Gao F, Markowitz M, Heath SL, Bar KJ, Goepfert PA, Montefiori DC, Shaw GC, Alam SM, Margolis DM, Denny TN, Boyd SD, Marshal E, Egholm M, Simen BB, Hanczaruk B, Fire AZ, Voss G, Kelsoe G, Tomaras GD, Moody MA, Kepler TB, Haynes BF. 2011. Initial antibodies binding to HIV-1 gp41 in acutely infected subjects are polyreactive and highly mutated. *J Exp Med* **208:**2237-2249.

60. Kwong PD, Mascola JR. 2012. Human antibodies that neutralize HIV-1: identification, structures, and B cell ontogenies. *Immunity* **37:**412–425.

61. Liao HX, Lynch R, Zhou T, Gao F, Alam SM, Boyd SD, Fire AZ, Roskin KM, Schramm CA, Zhang Z, Zhu J, Shapiro L, Mullikin JC, Gnanakaran S, Hraber P, Wiehe K, Kelsoe G, Yang G, Xia SM, Montefiori DC, Parks R, Lloyd KE, Scearce RM, Soderberg KA, Cohen M, Kamanga G, Louder MK, Tran LM, Chen Y, Cai F, Chen S, Moquin S, Du X, Joyce MG, Srivatsan S, Zhang B, Zheng A, Shaw GM, Hahn BH, Kepler TB, Korber BT, Kwong PD, Mascola JR, Haynes BF. 2013. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* **496:**469–476.

62. Zhou T, Zhu J, Wu X, Moquin S, Zhang B, Acharya P, Georgiev IS, Altae-Tran HR, Chuang GY, Joyce MG, Do Kwon Y, Longo NS, Louder MK, Luongo T, McKee K, Schramm CA, Skinner J, Yang Y, Yang Z, Zhang Z, Zheng A, Bonsignori M, Haynes BF, Scheid JF, Nussenzweig MC, Simek M, Burton DR, Koff WC, Mullikin JC, Connors M, Shapiro L, Nabel GJ, Mascola JR, Kwong PD. 2013. Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-class antibodies. *Immunity* **39:**245–258.

63. Zhu J, Wu X, Zhang B, McKee K, O'Dell S, Soto C, Zhou T, Casazza JP, Mullikin JC, Kwong PD, Mascola JR, Shapiro L. 2013. De novo identification of VRC01 class HIV-1-neutralizing antibodies by next-generation sequencing of B-cell transcripts. *Proc Natl Acad Sci USA* **110:**E4088–E4097.

64. Doria-Rose NA, Schramm CA, Gorman J, Moore PL, Bhiman JN, Dekosky BJ, Ernandes MJ, Georgiev IS, Kim HJ, Pancera M, Staupe RP, Altae-Tran HR, Bailer RT, Crooks ET, Cupo A, Druz A, Garrett NJ, Hoi KH, Kong R, Louder MK, Longo NS, McKee K, Nonyane M, O'Dell S, Roark RS, Rudicell RS, Schmidt SD, Sheward DJ, Soto C, Wibmer CK, Yang Y, Zhang Z, Nisc Comparative S, Mullikin JC, Binley JM, Sanders RW, Wilson IA, Moore JP, Ward AB, Georgiou G, Williamson C, Abdool Karim SS, Morris L, Kwong PD, Shapiro L, Mascola JR. 2014. Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature* **509:**55–62.

65. Ademokun A, Wu YC, Martin V, Mitra R, Sack U, Baxendale H, Kipling D, Dunn-Walters DK. 2011. Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages. *Aging Cell* **10:**922–930.

66. Chao A. 1987. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43:**783–791.

67. Zhou J, Lottenbach KR, Barenkamp SJ, Lucas AH, Reason DC. 2002. Recurrent variable region gene usage and somatic mutation in the human antibody response to the capsular polysaccharide of Streptococcus pneumoniae type 23F. *Infect Immun* **70:**4083–4091.

68. Lucas AH, McLean GR, Reason DC, O'Connor AP, Felton MC, Moulton KD. 2003. Molecular ontogeny of the human antibody repertoire to the Haemophilus influenzae type B polysaccharide: expression of canonical variable regions and their variants in vaccinated infants. *Clin Immunol* **108:**119–127.

69. Jackson KJL, Liu Y, Roskin KM, Glanville J, Hoh RA, Seo K, Marshall EL, Gurley TC, Moody MA, Haynes BF, Walter EB, Liao H, Albrecht RA, García-Sastre A, Chaparro-Riggers J, Rajpal A, Pons J, Simen BB, Hanczaruk B, Dekker CL, Laserson J, Koller D, Davis MM, Fire AZ, Boyd SD. 2014. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. *Cell Host Microbe* doi:10.1016/j.chom.2014.05.013.