nature biotechnology

The promise and challenge of high-throughput sequencing of the antibody repertoire

George Georgiou[1–4], Gregory C Ippolito[3,4], John Beausang[5,6], Christian E Busse[7], Hedda Wardemann[7] & Stephen R Quake[5,6,8,9]

Efforts to determine the antibody repertoire encoded by B cells in the blood or lymphoid organs using high-throughput DNA sequencing technologies have been advancing at an extremely rapid pace and are transforming our understanding of humoral immune responses. Information gained from high-throughput DNA sequencing of immunoglobulin genes (Ig-seq) can be applied to detect B-cell malignancies with high sensitivity, to discover antibodies specific for antigens of interest, to guide vaccine development and to understand autoimmunity. Rapid progress in the development of experimental protocols and informatics analysis tools is helping to reduce sequencing artifacts, to achieve more precise quantification of clonal diversity and to extract the most pertinent biological information. That said, broader application of Ig-seq, especially in clinical settings, will require the development of a standardized experimental design framework that will enable the sharing and meta-analysis of sequencing data generated by different laboratories.

A potent adaptive immune system is fundamentally reliant upon the generation of a diverse repertoire of B-lymphocyte antigen receptors (BCRs, the membrane-bound form of antibodies expressed on the surface of B cells). BCRs are assembled by somatic recombination of a large number of immunoglobulin gene segments (**Fig. 1**), and the repertoire of BCRs expressed in any given individual is continuously shaped by exposure to exogenous antigens and endogenous host factors. Existing mechanisms for BCR diversification can yield an astronomical number of possible BCRs (in theory, $>10^{13}$ in humans)[1,2]; this number exceeds the total number of B lymphocytes in the human body (~$1–2 \times 10^{11}$) (ref. 3). Because of labor and cost considerations, it is completely impractical to analyze such a diverse BCR repertoire using traditional Sanger sequencing. However, Ig-seq (a term coined by Andrew Fire, Stanford University) has allowed us to determine antibody gene repertoires at an unprecedented depth. The information gained by Ig-seq is proving invaluable for understanding antibody responses in health and disease and for diagnostic purposes. In addition, Ig-seq can be combined with other techniques, including expression and isolation of antigen-specific antibodies, sequencing of multiple RNAs from single cells[4], and proteomic analyses of antibodies in blood or secretions, to help elucidate the properties of antibodies that mediate protection against infectious diseases or, alternatively, that mediate autoimmune responses. In this Review we describe the experimental approaches and technical challenges related to high-throughput antibody gene sequencing, as well as the ways in which Ig-seq might be applied to advance our understanding of immunology and to address unmet clinical needs related to infectious diseases, immune dysregulation and cancer.

[1]Department of Chemical Engineering, University of Texas at Austin, Austin, Texas, USA. [2]Department of Biomedical Engineering, University of Texas at Austin, Austin, Texas, USA. [3]Department of Molecular Biosciences, University of Texas at Austin, Austin, Texas, USA. [4]Institute for Cell and Molecular Biology, University of Texas at Austin, Austin, Texas, USA. [5]Department of Bioengineering, Stanford University, Stanford, California, USA. [6]Howard Hughes Medical Institute, Stanford University, Stanford, California, USA. [7]Max Planck Institute for Infection Biology, Berlin, Germany. [8]Biophysics Graduate Program, Stanford University, Stanford, California, USA. [9]Department of Applied Physics, Stanford University, Stanford, California, USA. Correspondence should be addressed to G.G. (gg@che.utexas.edu).

Received 10 June 2013; accepted 4 December 2013; published online 19 January 2014; doi:10.1038/nbt.2782

Generation of the antibody repertoire

Antibodies are produced by a developmentally ordered series of somatic gene rearrangement events that occur exclusively in developing B cells and continue throughout the life of an organism. Antibodies consist of heavy (μ, α, γ, δ, ε) and light chains (κ, λ), which are linked by disulfide bonds. The intact antibody contains variable and constant domains (**Fig. 1a**). Antigen binding occurs in the variable domain, which is generated by recombination of a finite set of tandemly arranged variable (V), diversity (D) and joining (J) germline gene segments (**Fig. 1b**). This process, called VDJ recombination, often results in the addition and deletion of nucleotides at the junctions between ligated gene segments (**Fig. 1b**). More specifically, DNA exonucleases can trim the ends of the gene segments, and DNA polymerases and transferases can randomly insert templated palindromic or nontemplated nucleotides, respectively.

During B-cell development, immunoglobulin heavy (IgH) chain gene recombination typically occurs before immunoglobulin light (IgL) chain gene recombination. If both IgH and IgL genes are productively rearranged, the fully assembled antibody heterodimer is expressed on the surface of the B cell. In B cells bearing productively rearranged antibodies, the process of allelic exclusion (and locus exclusion in the case of IgL) ensures that each B cell expresses a single antibody[5]. After passage through several developmental checkpoints, newly generated mature IgM^+IgD^+ B cells form the naive B cell (and, therefore, naive antibody) repertoire. Most of the diversity in the naive

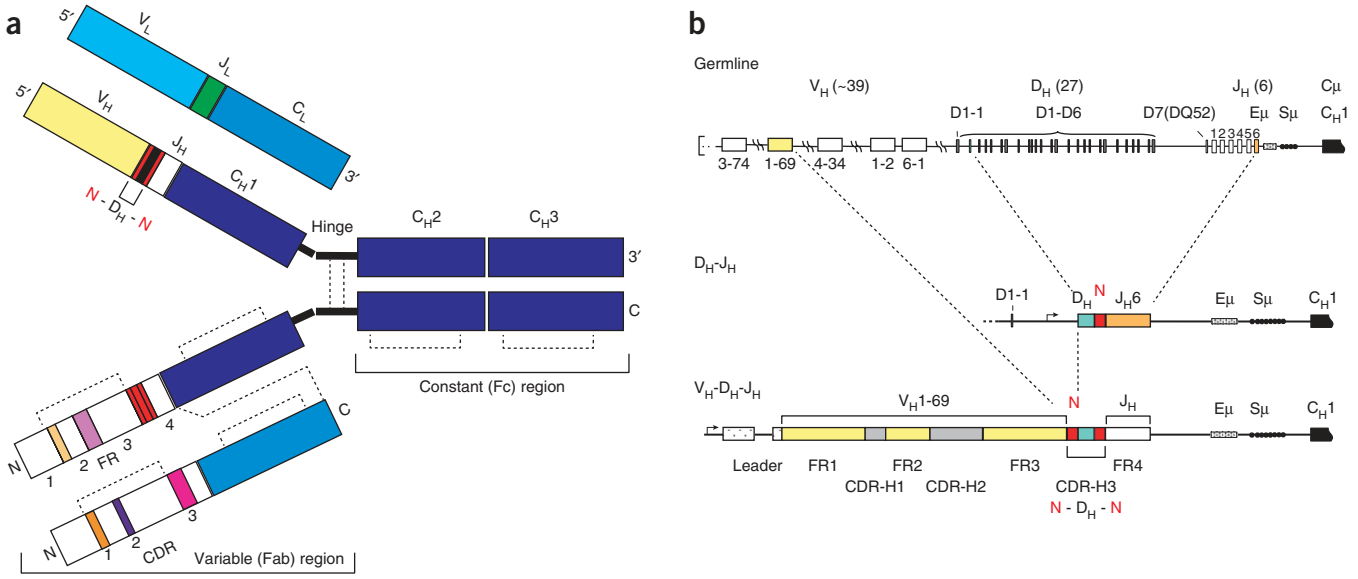

Figure 1 Antibody structure and sequence diversification mechanisms. (**a**) Schematic of IgG structure. In the top chains, domains encoded from germline V, D, J and C segments are indicated. Nontemplated N-nucleotides are shown in red. These top chains delineate the 5′ to 3′ genetic composition of the antibody. In the bottom chains, framework (FR) and complementarity-determining regions (CDRs) are indicated. These bottom chains delineate the N-terminal to C-terminal protein sequence. Dashed lines denote disulfide bonds. (**b**) Key steps in antibody diversification. The primary antibody heavy chain repertoire is created predominantly by the somatic recombination of variable (V), diversity (D) and joining (J) gene segments, and by the random nontemplated addition of N-nucleotides. The antigen-binding site of a heavy chain is formed by the juxtaposition of the hypervariable complementarity-determining regions (CDR-H1, H2 and H3) and the framework 3 region (FR3). After productive IgH rearrangement, recombination of the light chain (IgL) ensues, and the heterodimeric pairing of H and L chains forms the complete antibody of the IgM isotype that is expressed on the surface of a newly formed immature B cell. Eμ: IgM intronic enhancer; Sμ: tandem repeats critical for class-switch recombination. Numbers in parentheses refer to estimates of human germline V_H D_H and J_H segments.

antibody repertoire is concentrated at the site of IgH VDJ gene segment ligation, also known as the IgH complementarity-determining region 3 (CDR-H3) (**Fig. 1b**). Because of the combinatorial and nontemplated nature of the mechanisms that generate the CDR-H3, it is the most diverse component in terms of length and sequence of the antibody H-chain repertoire and is a principal determinant of antibody specificity[6,7]. Nonetheless, there are instances where antigen specificity is dictated solely or predominantly by the L chain.

When a B cell encounters antigen in an environment that provides requisite co-stimulatory signals and T-cell help, BCR stimulation induces B-cell proliferation. This process, known as B-cell clonal expansion, occurs primarily in highly organized areas of secondary lymphoid organs (e.g., spleen, lymph nodes and Peyer's patches[8,9]) referred to as germinal centers (**Fig. 2**). Clonal expansion is followed by somatic hypermutation of the variable domains of antibodies; this is mediated by activation-induced cytidine deaminase. B cells expressing BCRs bearing somatic mutations that increase affinity for antigen outcompete other B cells for access to antigen. As a result, the B cells bearing the highest-affinity antibodies undergo preferential expansion and survival, a process referred to as affinity maturation. Somatic hypermutation also results in sequence diversification of the CDR-H1 and CDR-H2 hypervariable regions and of the framework 3 (FR3) region, which was proposed to constitute a fourth hypervariable region of the antibody H chain[10]. Activation-induced cytidine deaminase also mediates class-switch recombination, which generates antibodies bearing different constant regions. B cells expressing somatically mutated, high-antigen-affinity BCRs can differentiate into long-lived memory B cells, capable of mediating rapid recall responses to the same antigen, or into terminally differentiated plasma cells; the latter downregulate BCR expression, establish residency in the bone marrow, gut lamina propia (and, to a smaller degree, in secondary lymphoid tissues), and secrete protective antibodies at extremely high rates estimated at 10,000–20,000 antibody molecules per second[11]. Antibody production by long-lived plasma cells in the bone marrow is postulated to proceed for very long times, possibly throughout the entire lifetime of the organism.

Diversity in the primary antibody repertoire (before exogenous antigen exposure) stems from the allelic diversity in immunoglobulin gene segments, combinatorial diversity introduced during somatic recombination, junctional diversity caused by the imprecision of the recombination process, pairing of IgH and IgL polypeptide chains, and receptor editing, wherein the existing V-gene segment is replaced with another (**Fig. 1**). In addition, V_H replacement, a process resulting from the presence of a cryptic recombination signal sequence in FR3, might influence as much as 5–12% of the human primary B-cell antibody repertoire[12]. Diversification of the post-antigen-stimulation secondary antibody repertoire stems from somatic hypermutation and class-switch recombination.

Organism age also influences the antibody repertoire[13,14]. During early ontogeny, the mammalian adult B-cell repertoire is generated in a predictable developmentally programmed fashion, whereas in advanced age humoral immune responsiveness deteriorates; this phenomenon is referred to as immunosenescence and is thought to be attributable in part to a progressive restriction of the antibody repertoire. For example, among the elderly there is an increased prevalence of autoantibodies and, at the serological level, an increased amount of either a single or a small number of serum immunoglobulins that are produced at a high level by benign outgrowths of one or more plasma cell clones[15,16].

Low-throughput analysis of the antibody repertoire

In the 1990s, Sanger sequencing enabled the determination of IgH and IgL VDJ recombinants (hereafter simply referred to as V genes)

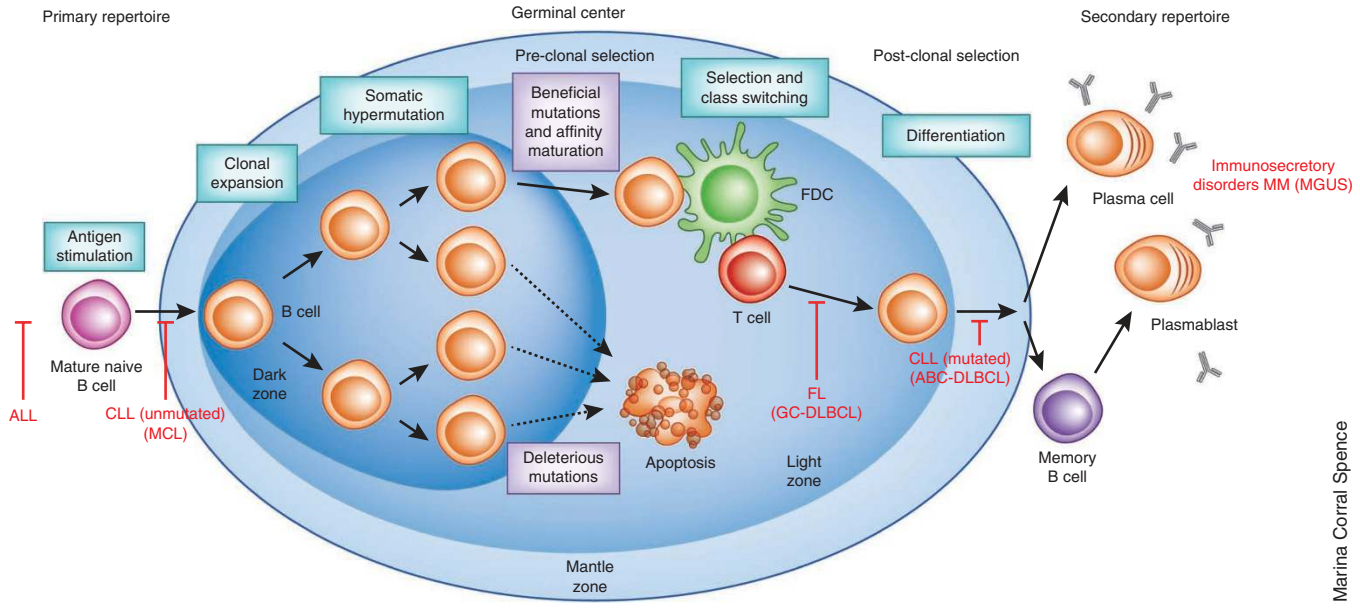

Figure 2 Key steps in the development of antigen-specific B cells. The steps of normal B-cell differentiation and diversification of the antibody repertoire are indicated in black text. Normal B cells are generated in the bone marrow, migrate to the periphery and, following developmental checkpoint selection, comprise the population of IgM^+IgD^+ mature naive B cells. When these cells are activated by cognate antigen in the presence of T-cell help, they enter a germinal center (GC) reaction where they rapidly proliferate; this results in clonal expansion and subsequent somatic hypermutation catalyzed by activation-induced cytidine deaminase. B cells bearing antibodies with high affinity for cognate antigen and that survive the GC reaction can undergo class-switch recombination to IgG, IgA or IgE isotypes and ultimately differentiate into memory B cells, antibody-secreting plasmablasts or plasma cells. After subsequent encounter with the same cognate antigen, memory B cells can proliferate or differentiate directly into antibody-secreting cells. Steps that proceed abnormally, leading to the development of human B-cell leukemias and lymphomas, are indicated in red text. ALL, acute lymphoblastic leukemia; CLL, chronic lymphocytic leukemia; MCL, mantle cell lymphoma; GC-DLBCL, germinal center diffuse large B cell lymphoma; FL, follicular lymphoma; ABC-DLBCL, activated B cell–like DLBCL; MGUS, monoclonal gammopathy of undetermined significance; MM, multiple myeloma. B-cell malignancies that have not been analyzed extensively by high-throughput sequencing are shown in parenthesis.

in typically up to a few hundred B cells per experiment[17–19]. Subsequent studies began to clone immunoglobulin genes from single B cells following limiting dilution, and to express and functionally characterize the cloned antibody proteins, thus enabling the interrogation of antibody specificity. This advance proved invaluable for the isolation of antibodies relevant to disease, especially the isolation of pathogen-neutralizing antibodies[20]. B-cell immortalization (and subsequent sequencing of the V genes in immortalized B clones) provided an alternative route for the expression of small numbers of antibodies[21–23].

The ability to interrogate the antibodies encoded by small numbers of B cells has yielded numerous important immunological insights. For example, Wardemann *et al.*[24] first used single-cell cloning and antibody expression to demonstrate that a fraction of newly generated B cells in the human bone marrow express self- and polyreactive antibodies and that their development is regulated at two independent self-tolerance checkpoints. Subsequent studies from our laboratory and from others elucidated mechanisms and features of antibodies associated with autoimmune diseases or immunodeficiency syndromes[25–28]. B-cell cloning techniques also enabled the isolation of antibodies able to neutralize numerous clinically important pathogens including severe, acute, respiratory syndrome coronavirus (SARS-CoV), influenza and HIV-1, among many others[29–35]. Understanding the mechanisms that lead to the elicitation of neutralizing antibodies is, in turn, helping researchers design more effective vaccines[20,36–38]. Nonetheless, one key limitation of low-throughput, B-cell cloning studies has been that they provide only a glimpse of a miniscule slice of the full antibody repertoire.

High-throughput sequencing of the antibody repertoire

Compared with Sanger sequencing, Ig-seq can provide a much broader picture of the antibody repertoire (**Fig. 3a**). Although conceptually simple, for its proper application, Ig-seq demands thoughtful consideration of experimental design, a detailed understanding of the sources of DNA sequence and quantification errors, the ability to delineate which V_H genes are paired with which V_L genes in each single B cell, and the use of appropriate data mining and visualization tools to make biological sense of the large amounts of information generated in such experiments.

Experimental design. The first consideration is the source of B cells. Most human antibody sequencing studies have used B cells from peripheral blood because the blood is one of the few readily accessible sources of B cells in humans (tonsils is the other one). However, it is estimated that only 2% of the $1–2 \times 10^{11}$ B cells in the human body are present in peripheral blood, compared with almost 28% in lymph nodes, 23% in the spleen and on mucosal surfaces, and 17% in the red bone marrow (*medulla ossium rubra*)[3]. Thus, the antibody repertoire in peripheral B cells provides a narrow view of the humoral response to antigen challenge.

Second, it is important to consider whether to use genomic DNA (gDNA) or mRNA for immunoglobulin sequencing analyses (**Fig. 3a**). Whether or not one should use gDNA or mRNA depends on what question is being asked. Sequencing gDNA facilitates estimation of the clonality of a given Ig sequence (in other words, the number of B cells expressing that antibody) because the number of sequence reads will, in general, be proportional to the number

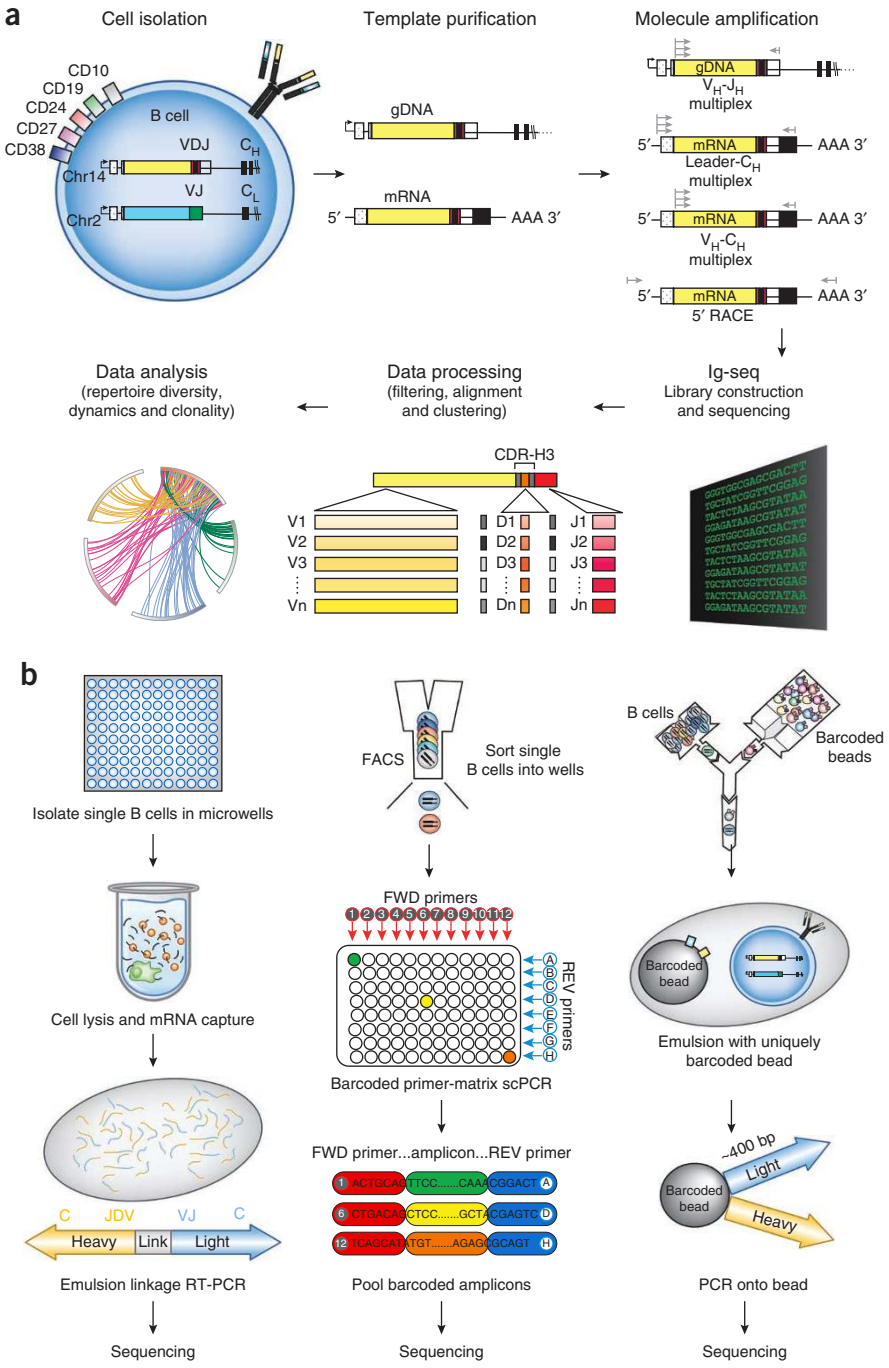

Figure 3 Methods for high-throughput sequencing of the Ig sequence repertoire. (**a**) Schematic of steps involved in high-throughput sequencing of Ig genes from bulk B-cell populations of B-cell subsets sorted according to expression of indicated cell-surface markers. Either genomic DNA (gDNA) or mRNA can be used as template, and the choice of template influences the number and location of primers used for subsequent PCR amplification. gDNA amplification is performed using primers complementary to the rearranged V-region gene (VDJ recombinant); amplification of cDNA is performed either using a 5′ primer pool complementary to the leader peptides or FR1s of V-gene segments, and a single 3′ C_H1 (or Cκ,Cλ if amplifying light chain genes) primer, or alternatively by 5′ RACE. Although throughput is high, in bulk analysis information regarding which V_H and V_L chains were paired in the same cell is lost, as cells are lysed in bulk and V_H and V_L genes are amplified in separate reactions. (**b**) Schematic of single-cell immunoglobulin repertoire sequencing methods, which preserve information about endogenous V_H:V_L pairs. Left panel: B-cell lysis and mRNA capture in picoliter well arrays. Middle panel: single-cell PCR following limited B-cell dilution and amplification using barcoded primers. Right panel: microfluidic barcoding of V_H and V_L cDNAs. Ig, immunoglobulin.

of gDNA template molecules (assuming no primer biases, as discussed below). On the other hand, using mRNA as a template can provide an estimate of the relative expression level of various immunoglobulin sequences in the repertoire. However, because immunoglobulin transcription varies dramatically (up to 100-fold) between naive B cells and plasma cells (**Fig. 2**)[39], using unsorted bulk B cells from peripheral blood as the source of mRNA makes it challenging to deduce cellular clonal frequencies. When unsorted peripheral blood mononuclear cells are used as a source of mRNA, the degree of somatic hypermutation in sequencing data sets may be employed *ipso facto* to distinguish V-gene transcripts derived from antigen-experienced versus naive B cells[40]. Alternatively, cell-surface markers can be used to sort B-cell subsets of interest before mRNA isolation[41,42]. Further complicating comparisons between repertoires from different studies, recent work[43–46] has unveiled an unexpectedly high degree of polymorphism in the human *IGH* locus. Understanding the degree of genetic variation is critical for assigning VDJ segment usage, estimating somatic hypermutation and comparing antibody responses among individuals.

Another experimental design consideration relates to the large number of B cells in humans (~2–4 × 10^9 B cells in the blood alone). The minimal sampling depth needed to cover the antibody repertoire to an extent sufficient to answer a particular biological question needs to be considered. Clearly the sequencing depth must be greater than the number of B cells in the sample (which means that enumeration of B cells by fluorescence-activated cell sorting is essential). Further, the required sequencing depth is also dependent on the approach used to minimize PCR and base-calling errors (see below).

Distinguishing error from true biological variation. There are two sources of sequence errors: those that arise from sample preparation (reverse transcription and PCR) and those inherent to the DNA sequencing platform.

Relative clonal frequencies can be estimated from gDNA by preparing and sequencing libraries from multiple aliquots from the same sample (technical replicates)[41,47]. In this approach, each aliquot is used to generate a barcoded library, and the resulting libraries from different aliquots are sequenced. The barcode prevents artifacts due to the contamination of one library with DNA from a different aliquot. Because antibody genes found in multiple libraries can arise only if

the respective clonal B cells are present in each of the starting aliquots, this method can reveal clonal expansions[41,47]. However, using gDNA as the templates is not without complications. First, amplification of VDJ segments from gDNA necessitates the use of primer sets that anneal to all the individual germline V-gene segments (**Fig. 3a**). Another drawback of gDNA-derived antibody gene libraries is that they contain productive and nonproductive VDJ rearrangements (the latter are also present in cDNA libraries but at a much lower frequency due to the decay of nonsense RNAs)[48,49]. Lastly, the lower concentration of template in gDNA necessitates a greater number of PCR cycles; this increases error frequencies and further confounds quantification.

Using mRNA as the starting material enables amplification with reverse transcription and 5′RACE (5′ rapid amplification of cDNA ends) with 3′ primers that anneal to the constant region of IgH or IgL, thus circumventing the need for complex V-gene-specific primer sets (**Fig. 3a**). If starting with mRNA, errors introduced by reverse transcriptase can be minimized using commercial high-fidelity retroviral reverse transcriptases or thermostable group II intron reverse transcriptases[50].

Regardless of whether gDNA or cDNA is used as the template, PCR introduces amplification artifacts owing to the differential amplification of some DNA templates over others (even in 5′RACE), base misincorporation and template switching; the latter results in chimeras from the joining of fragments encoded by two or more template DNAs. Nucleotide misincorporation by PCR cannot generally be distinguished from most types of base-calling errors introduced during sequencing, but the latter generally occur at higher frequency and hence they are a greater concern. Chimeras resulting from template switching generate sequences that either cannot be assigned to a germline V-gene segment by standard VDJ identification algorithms, or are interpreted as having an artifactually very high rate of somatic hypermutation. The presence of chimeras makes discernment of true V_H gene replacement events particularly challenging. However, the fact that gene replacement by definition has to involve recombination to upstream V_H gene segments in the chromosome (because downstream V_H gene segments will have been deleted during the primary recombination event) can be used to assist the identification of true V_H replacement events. Quantification biases introduced by PCR may be minimized by amplifying and then sequencing the same sample with two or more different primer sets, followed by informatics comparison of the similarities in the respective datasets[51]; another possible solution is using emulsion PCR to sequester individual DNA molecules for amplification[52].

Several of the complications underlying DNA sequencing error have been recently reviewed[53–57] so here we focus only on the issues most pertinent to Ig-seq, and on recent technological advances aimed at increasing sequencing accuracy of antibody repertoires. Sequencing errors relate to the particular DNA sequencing technology used in each experiment and encompass incorrect base assignment, insertion/deletions (collectively known as indels) and ambiguous base calls. Pyrosequencing-based technologies (Roche 454 and IonTorrent) are dominated by indels, whereas dye-labeled reversible terminator technology (Illumina HiSeq and MiSeq) is dominated by substitution errors[55,58–61]. Indels generated by pyrosequencing methods arise at frequencies around 5×10^{-3} and can be computationally accounted for with various degrees of success[56,62,63]. According to a recent analysis, the frequency of base substitutions varies from 0.3% to 0.9%, depending on the platform[61]. Overall, the Illumina platform is most suitable for Ig-seq applications because of its combination of relatively low base-calling error rates and relatively low cost.

It should be noted that the importance of sequencing errors depends on the objective of the antibody repertoire sequencing experiment. For example, if the objective is to generate CDR-H3 length distribution statistics or V-J segment use statistics, then sequencing errors are less of a problem, as meaningful information can be obtained using clustering algorithms that group together highly homologous sequences and minimize the effects of sequencing error[51,64–67]. For other applications where sequence accuracy is critical, a number of recently developed approaches can help reduce sequencing errors. For example, up to tenfold higher sequencing accuracy can be attained using circularized V-gene DNA that is sequenced multiple times using the PacBio platform[68]; however, this accuracy comes at a higher cost and much lower throughput. Various DNA barcoding techniques, whereby a nucleotide barcode is appended to each DNA template molecule before PCR amplification, can also improve sequencing accuracy[69–71]. Barcoding enables more accurate quantification of the DNA template molecules in the original library (by counting barcodes instead of sequence reads), and error correction (by generating consensus reads having the same barcode). In a recent application of barcoding to Ig-seq, the Quake laboratory[72] achieved 95% identity in the V-gene repertoire (for sequences present at >5 reads) between two technical replicates. Another method[73] that relies on barcoding and sequencing both strands can achieve very high accuracy ($<10^{-8}$ error rate), although at the expense of throughput.

Identifying endogenous V_H:V_L pairs. As discussed above, until recently, native V_H:V_L pairs could be identified only after single-cell cloning by limiting dilution and Sanger sequencing of the individual V_H and V_L genes. This process is inherently low throughput, expensive (due to high reagent usage) and yields a very limited set of antibody sequences. A modest increase in single-cell cloning throughput (from ~500 to 2,000 B cells) was achieved using overlap extension PCR to produce single V_H:V_L amplicons (as opposed to separate V_H and V_L cDNAs) for Sanger sequencing[74]. In a study from the Quake laboratory, Weinstein *et al.*[4] sorted ~200 mouse B cells into microwells, followed by quantitative RT-PCR on a microfluidic chip using primers specific for genes of interest that also contained sequencing adapters, thus enabling the correlation between expression of multiple genes and the degree of somatic hypermutation in the antibody heavy and light chain genes.

Additional approaches for sequencing endogenous V_H:V_L pairs were more recently developed in the Wardemann and Georgiou laboratories. Busse *et al.*[75] used a two-dimensional, bar-coded primer matrix to combine single-cell V_H and V_L gene amplification with high-throughput sequencing; this increases throughput up to a total of 50,000 individual B cells (**Fig. 3b**). Importantly, this approach also enabled the facile cloning of IgH and Igκ,λ sequences into expression vectors for further antibody characterization. DeKosky *et al.*[76] developed a V_H:V_L pairing technology that relies on sequestering single B cells into subnanoliter volume wells, lysing the cells, capturing RNA on poly-dT beads and generating amplicons encoding linked V_H:V_L segments by emulsion overlap extension PCR (**Fig. 3b**). Yields of up to $6–7 \times 10^3$ unique V_H:V_L pairs from 7×10^4 activated memory B cells in a one-day experiment, with >96% validated pairing accuracy, have been reported. This method was adapted to detect co-expression of V genes with transcription factors of interest (e.g., *BLIMP1*) in antibody-secreting cells (B.J. DeKosky, personal communication).

A much higher-throughput method (>2×10^6 B cells per experiment) under development in our laboratory relies on cell encapsulation within controlled microdroplet diameter emulsions (an important consideration because reverse transcriptase is inhibited

in droplets of <5 nl volume) (unpublished data). Alternatively, single cells have been encapsulated in water-in-oil emulsion together with uniquely barcoded beads using a microfluidic device. After reverse transcriptase and PCR, every RNA molecule originating from a single cell is effectively loaded onto the uniquely barcoded beads. PCR products are sequenced on the Illumina platform and correctly paired V_H and V_L sequences are identified by virtue of their shared barcode[77] (**Fig. 3c**). Given the rapid pace of technology advancement, it may be expected that in the very near future all antibody repertoire analyses will report natively paired V_H and V_L genes.

Bioinformatic analysis of antibody sequences. Several established methods for VDJ assignment and CDR-H3 identification are available (e.g., IMGT/V-Quest, IgBLAST or iHMMune-align) and faster, more precise algorithms continue to be developed[54,57,78,79]. Estimation of the size of the antibody repertoire can be accomplished by rarefaction analysis, maximum entropy and Poisson log-normal distribution models[51,80,81]. There are also many approaches for clustering V genes likely to have originated from the expansion and somatic hypermutation of a single B cell encoding an unmutated antibody precursor[57]. However, clustering becomes particularly problematic as the sequencing depth and repertoire diversity increases. Likewise, inference of V_H evolution (i.e., how somatically mutated V genes arise from precursor sequences initially present in the naive primary repertoire) is a major challenge[67,82]. Finally, it should be noted that researchers working in this area generally use custom-made bioinformatics pipelines. The lack of standardization and shared computational resources makes it extremely difficult to carry out meta-analyses of published data generated in different laboratories; this is a key issue that will need to be addressed as the field matures. Interchangeable data formats deposited in a central database, validated open-source algorithms for data analysis and standards for Ig-seq experimental description analogous to the minimum information about a microarray experiment[83] are critically needed[84].

Applications of antibody repertoire sequencing

Ig-seq is finding a wide range of basic and applied immunology applications.

Antibody discovery. Screening of large combinatorial libraries by ribosomal, phage, bacterial or yeast display is widely employed for the isolation of antibodies capable of binding virtually any ligand[85–87]. Combinatorial libraries are typically generated by the random pairing of very large (typically $>10^7$ each) ensembles of V_H and V_L genes isolated from mammals (naive or immune libraries, depending on whether the animal was immunized); alternatively one can use synthetic libraries in which a single, or a small set, of V_H and V_L genes are diversified by mutagenesis of the CDRs. Screening involves sequential rounds of binding to antigen, a process that progressively restricts the diversity of the library to very few antibody clones with the requisite affinity and specificity. We and others have used high-throughput sequencing as a means of evaluating the initial diversity of antibodies encoded in a library[64,88,89] and to determine how this diversity declines as binders to antigen are progressively enriched[89–92]. During library screening, antibody diversity is reduced as antigen-binding clones are enriched over the background of unrelated antibodies in the starting library. However, the expression of some antibodies having high affinity often adversely affects the growth of the cells that encode them (*nota bene*: antibodies with poor affinity are equally likely to have an adverse effect on cell growth, but obviously such clones are of no interest in library screening). As a result, the respective antibody genes are enriched during early rounds of screening for antigen binding but are then progressively depleted because they are outcompeted by faster growing cells expressing lower quality antibodies[87,90,92,93]. High-throughput sequencing of libraries following one or two rounds of selection for antigen binding has therefore been used to rescue clones encoding high-affinity antibodies that could not have been discovered otherwise.

Three different approaches that exploit antibody repertoire sequencing analysis for antibody isolation directly from animals or humans, without library screening, have been developed. First, in work from the Georgiou laboratory, Reddy *et al.*[94] observed that in mice the antibody repertoire encoded by $CD138^+$ antibody-secreting B cells in the bone marrow (bone marrow plasma cells) becomes highly polarized seven days after secondary immunization with antigen, with the most abundant V_H and V_L genes representing 2–30% of all V-gene sequences. To discover antigen-specific antibodies, Reddy *et al.*[94] paired V_H and V_L genes based on their relative rank frequency in the respective repertoires and thus obtained antibodies with nanomolar affinity for antigen. Subsequently, Saggy *et al.*[95] showed that V_H genes present at high frequencies in the splenic V-gene repertoire of immunized mice encode antibodies capable of binding the immunizing antigen. However, the types of bone marrow or spleen samples used in these mouse analyses are rarely available from humans.

A different approach, pioneered by Shapiro, Kwong and coworkers, builds on the observation that the co-evolution of heavy and light chains in broadly neutralizing HIV-1–specific antibodies is reflected in the matching topology of V_H and V_L phylogenetic trees as gleaned from Ig-seq[63,96]. V_H and V_L genes within matching branches of the respective phylogenetic trees, that is, displaying similar patterns of mutation accumulation, were paired to yield novel broadly neutralizing antibodies. This method is particularly useful for the identification of antibodies with high somatic hypermutation levels (such as those elicited by persistent infection) where a deep phylogenetic tree can be constructed.

Ig-seq has also enabled the proteomic identification of antibodies from biological fluids. Even though the main effector function of B cells is the secretion of antibodies into blood or mucosal and respiratory epithelia, the composition of the antibody repertoire in these bodily fluids remained elusive. We and others realized that shotgun liquid chromatography–tandem mass spectrometry (LC-MS/MS) proteomic identification of the antibodies that comprise the humoral response requires a matched personal antibody sequence database, obtained by Ig-seq, to interpret the MS/MS spectra (**Fig. 4**)[97–99]. In this manner, Polakiewicz and coworkers[98,99] identified high-affinity, antigen-specific antibodies from immunized animals, from a human vaccinated with hepatitis B and from a cytomegalovirus-infected volunteer[98,99]. In more recent studies, we determined the monoclonal antibody repertoire that comprises the serum polyclonal response in rabbits and in humans following vaccination (unpublished data and ref. 66). The advent of antibody discovery by proteomic mining of the serum antibody response has in turn opened the way for the isolation of biologically relevant antibodies from convalescent patients. Finding the serum antibodies that were responsible for the resolution of disease states in patients is likely to be of great relevance for drug discovery because such antibodies will have already been established to be of therapeutic value.

Understanding immune repertoire development. The use of Ig-seq to map out the global antibody repertoire was first accomplished by Quake and coworkers in the zebrafish, which exhibit the basic features of adaptive immunity but have fewer possible VDJ combinations than

mice or humans[51,100]. Ig-seq has subsequently been used to analyze the repertoires of a variety of other species, including mice, chickens and cattle[51,66,68,94,101–107].

More importantly, however, Ig-seq is providing unprecedented insights into the mechanisms responsible for shaping the human naive BCR repertoire. As noted above, Ig-seq analysis of the V_H repertoire revealed the high degree of allelic diversity in the human immunoglobulin locus[43,45,108]. Analysis by our laboratory and by others of germline V_H, Vκ and Vλ segment usage and frequencies of recombination between particular V-D and D-J segments in the naive BCR repertoire revealed a marked skewing that in turn then shapes the repertoire in mature, antigen-experienced B cells[7,43,47,49,64,102,109–112]. A single D-segment reading frame is overwhelmingly preferred in VDJ recombination, inversions of D segments are extremely rare[113] and the CDR-H3 region exhibits universal categorical constraints with respect to amino acid composition, average hydrophobicity, charge distribution and length. Antibodies with long or charged CDR-H3s are of great clinical interest because, although they are more likely to be autoreactive, they often enable binding to occluded sites on pathogens and mediate pathogen neutralization. In the human repertoire, long CDR-H3s arise by D-D joining (which occurs at frequencies from 0.125% to >0.5% in the naive repertoire but is reduced in antigen-experienced B cells) and by extensive nucleotide addition combined with preferential usage of longer D segments and of the longer J6 segment in the germline[47,49,114,115]. The conservation of hydrophobicity, which results in the selection of a neutral CDR-H3 and against hydropathically extreme (charged or hydrophobic) antibody-binding sites, is a common theme across vertebrate evolution[7,116]. Collectively, these findings indicate that the CDR-H3 loop in humans explores vast sequence space, but within boundaries[40,53,64,102,105].

Recent studies suggest determinism in VDJ recombination frequencies in the primary antibody repertoire. VDJ segment usage in the antibody repertoire of homozygotic twins is indistinguishable, suggesting that it is determined by genetic factors[40]. Nonetheless, it is important to note that diversification in the CDR-H3 results in a highly private repertoire that shows very little overlap among individuals. Further, the human naive antibody repertoire that developed in immunocompromised mice engrafted with human $CD34^+$ hematopoietic stem cells[105] showed V-J usage patterns and, even more strikingly, checkpoint depletion of antibodies displaying autoreactivity signatures similar to those in humans[24].

Important insights are also being gleaned from sequencing of antibody repertoires encoded by B cells at different anatomical sites. One particularly noteworthy study analyzed the murine intestinal IgA repertoire[117]. Another study of mouse mucosal B-cell repertoires in the intestinal lamina propria raised the intriguing possibility that, contrary to long-standing dogma, the bone marrow might not be the exclusive site of adult B-cell development. This was because B cells in the gut expressed Rag-1 recombinase, displayed a pre-B or immature B phenotype and importantly, encoded a distinctive V_L repertoire; these features are consistent with B-cell development and selection *in situ* in the lamina propria[118]. The IgG V_H repertoire in the cerebrospinal fluid has also been examined[119] and found to have differences in somatic hypermutation compared to that in the peripheral blood, further suggesting that B-cell maturation might have occurred independently in the central nervous system or the periphery.

Several studies have attempted to estimate the total diversity of the human antibody repertoire by Ig-seq of peripheral B cells[40,47,102]. This task is very challenging owing to the large size of the antibody repertoire, variable sampling depth, transcriptional differences among B-cell subsets, sequencing errors and, last but not least, the fact that peripheral blood contains <2% of all B cells[3]. The overwhelming majority of published studies[105,120] as well as our unpublished data are consistent with the notion that, on the one hand, the V_H gene repertoire is highly private (unique to an individual), although a small number of CDR-H3 appear to be shared (in other words, they are stereotypical or public) among different individuals. However, a sizable fraction of shared sequences are found in light chains, due to the lower diversity of the V_L gene repertoires[121,122]. It is intriguing that shared IgL chain genes also tend to be the most abundantly expressed[123].

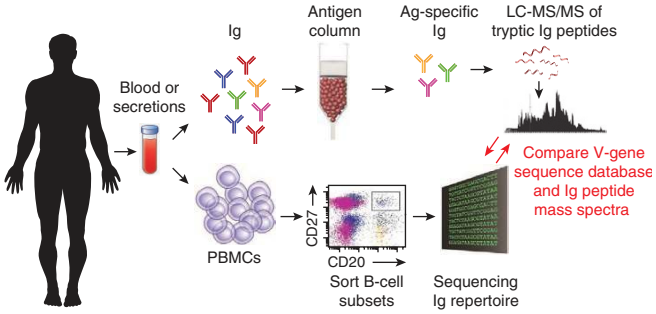

Figure 4 Deconvoluting the serum antibody repertoire. B cells from peripheral blood or other tissues are sorted and subjected to high-throughput immunoglobulin V-gene sequencing, resulting in generation of a personal antibody sequence database. Antigen-specific antibodies from serum are then isolated by affinity chromatography, digested into peptides, and subjected to LC-MS/MS analysis. The MS/MS data are interpreted using the antibody sequence database, thereby allowing identification of CDRH3-derived peptides and the genes encoding the repertoire of antigen-specific antibodies in the serum. Ig, immunoglobulin.

Infectious diseases. Ig-seq is also providing insights into the adaptive immune responses elicited by pathogen challenge or vaccination. Pathogen challenge can affect the BCR repertoire of responding B as well as the naive repertoire. Many pathogens produce superantigens, which are proteins that bind to certain antibody V domains, resulting in BCR cross-linking and subsequent B-cell deletion. Predictably, superantigen exposure results in a skewed naive antibody repertoire[103]. Surprisingly, however, depletion of V genes bound by superantigen was not observed in the naive repertoire of transgenic mice constitutively expressing superantigen[48]. Interestingly, skewed naive B-cell antibody repertoire was also reported for patients with chronically evolving hepatitis C infection[124]. Changes in the overall antibody repertoire are also evident following vaccination or infection[120,125–127]. Notably, Boyd, Fire and coworkers[125] observed convergent antibody signatures (stereotyped CDR-H3 sequences) in patients experiencing acute dengue infection. This observation raises the possibility that Ig-seq aimed at detecting stereotypical responses may be used as a diagnostic tool for predicting infectious disease severity.

Another exciting potential application of Ig-seq is the identification of V genes clonally related to those encoded by protective antibodies isolated from an individual; one can then infer the antibody lineages that led to the evolution of these protective antibodies, starting from an unmutated common ancestor IgH germline sequence. This approach is particularly relevant to understanding the evolution of broadly neutralizing antibodies (bNAbs) during infection with rapidly evolving viral pathogens such as influenza and HIV-1 (refs. 67,96,127–130). Time-ordered sequencing of evolving virus populations and antibody responses in the same host, together with isolation of bNAbs, is helping delineate the dynamic between

adaptive immune responses that exert selective pressure on the virus and the emergence of viral escape mutations[131]. Tracing the evolutionary paths that lead to the generation of bNAbs is also critical for the design of immunogens and vaccination schedules that will elicit an immune response by first activating naive, germline, antibody-expressing B cells and then steering B clonal selection toward an affinity maturation pathway that leads to the production of bNAbs[36,38]. Multidonor analysis of VRC01-class anti-HIV-1 bNAbs confirmed that the elicitation of such antibodies from a single ancestor B cell indeed occurs in multiple individuals[132]. Ig-seq is also useful for evaluating how innate immune responses elicited by adjuvants, such as toll-like receptor 4 (TLR4) or TLR7/8 agonists, affect the diversity of the antibody response and possibly antibody functionality[133].

Ig-seq might also be applied to answer the long-standing question of why people in certain age groups, usually the elderly, exhibit higher susceptibility to infectious disease and/or are less well-protected by vaccination. For example, by analyzing antibody lineage structure, isotype and mutational load in the V_H repertoire of volunteers in various age groups before and after influenza vaccination, our laboratory detected a higher IgM mutational load before vaccination and a lower degree of repertoire diversity after vaccination in elderly individuals[120]. Another study also detected a smaller degree of IgM and IgA CDR-H3 diversification in elderly individuals before and after receiving the influenza or the 23-valent pneumococcal vaccine[126].

Immune dysregulation. We expect that deep sequencing of the antibody repertoire in patients with autoimmunity or primary immunodeficiency will provide important mechanistic insights that in turn may guide the development of appropriate therapies. However, very few analyses of antibody repertoires in patients with immune dysregulation have been published thus far; we believe this dearth reflects the very recent development of Ig-seq. That said, a few studies have been reported. In multiple sclerosis, the cerebrospinal fluid V_H gene repertoire is biased and shows strong evidence of B-cell activation[119]. It will be interesting to determine whether the activated B cells in the cerebrospinal fluid are reactive towards KIR4.1, a recently discovered dominant antigen in multiple sclerosis[134]. In idiopathic IgG4-associated cholangitis (an autoimmune disease associated with abnormal levels of IgG4 in serum), Ig-seq revealed the presence of large clonal IgG4 expansions in affected tissues and peripheral blood; these clonal expansions disappeared after corticosteroid treatment, indicating that determination of IgG4 clonality is a distinguishing feature of the disease and therefore constitutes a useful tool for differential diagnosis[135].

Cancer. B-cell leukemias, lymphomas and multiple myeloma are malignancies that arise at different stages of B-cell development (**Fig. 2**). As such, BCRs on malignant B cells constitute a biomarker for the abundance of the malignant cell population. Ig-seq of the V-gene repertoire in peripheral B cells, bone marrow samples, tumors and even blood-borne free DNA has been used both for disease detection and for delineating the degree to which antibody evolution and diversification in malignant cells correlates with disease progression or relapse[47,58,136–140]. For example, Ig-seq of V genes in peripheral blood facilitated detection of cancerous cells and minimal residual disease following treatment of B-cell chronic lymphocytic leukemia (CLL), the most common leukemia in adults[47,58,137]. The use of Ig-seq to determine the V-gene repertoire *en masse* and detect the presence of antibodies encoded by leukemic clones (whose V-gene sequences are established from analyses of cancer cell samples before initiation of treatment) circumvents the need to develop a personalized PCR assay for the CLL clonotype in each patient to determine whether relapse has occurred. Ig-seq was also used to detect minimal residual disease in pediatric patients with B-cell acute lymphoblastic leukemia (B-ALL)[138] and as a marker of non-Hodgkins lymphoma[136].

Ig-seq also revealed that B-ALL patients display various degrees of clonotypic diversity, which arises predominantly from V_H gene replacement and appears to be related to relapse frequency[138,139]. A small degree of B-cell clonal heterogeneity was also observed in CLL (with the degree varying based on whether the disease originated from unmutated or somatically hypermutated B cells)[58,137]. In contrast, the malignant clonotype in multiple myeloma, a disease which arises from terminally differentiated plasma cells in the bone marrow (**Fig. 2**) that lack active mechanisms of antibody diversification, displayed little evidence of heterogeneity[140]. Of note, in CLL and other hematologic malignancies, the same (stereotypic) CDR-H3 is detected in many patients[141]. The wider application of high-throughput sequencing for the detection of malignancy-stereotypic clonotypes, which may be present at a low frequency in subjects that do yet not display clinical disease, may prove to be a useful, early diagnostic tool. Lastly, Ig-seq may also be applied to understand gammopathies, which are conditions that result in abnormally high levels of antibodies in serum; these are more prevalent in the elderly and can develop into multiple myeloma[15,16].

Conclusions

The humoral immune system has evolved to encode an astonishing diversity of antibodies that collectively comprise the antibody repertoire and provide a potent arsenal of recognition reagents (or anticipatory receptors) that can recognize virtually any organic macromolecule of biological significance. The B lymphocyte and its 'anticipatory' receptor, a term used to emphasize the ability of the naive repertoire to bind biologically relevant antigens, is as ancient as the last common vertebrate ancestor (500 million years ago)[142]. The extraordinary complexity of the vertebrate adaptive immune system has been likened to the Titan brothers Epimetheus (hindsight) and Prometheus (foresight) of Greek mythology[143]. The germline conservation of antibody genes carries the imprint of ancient adaptations to pathogens humans were exposed to during evolution and, hence, is Epimethean in scope; on the other hand, the immunologic foresight represented by the repertoire encoded by antigen-experienced, mature B cells within an individual enables adaptations to future pathogen challenges and is a Promethean and anticipatory feature. The advent of high-throughput DNA sequencing has enabled the determination of the antibody gene repertoire at unprecedented depth that was inconceivable, even half-a-dozen years ago. We can begin to decipher both the Epimethean and the Promethean compartments of humoral immunity and how each is shaped by the other. Technologies to improve sequence precision and data analysis are being developed at a breakneck pace, reshaping our understanding of many important aspects of B-cell immunology and increasingly affecting clinical diagnosis, antibody drug discovery and vaccine development. However, realizing the full impact of Ig-seq will require the implementation of standards for experimental annotation and data analysis, as well as the creation of databases facilitating deposition and sharing of these important data.

ACKNOWLEDGMENTS

We thank B. DeKosky, J. Lavinder & K. Hon Hoi for reading this manuscript; J. Lavinder and C. Chrysostomou for help with the figures; and A. Fire for coining the term Ig-seq. We are also very grateful to the *Nature Biotechnology* Editorial staff for their meticulous editing and for numerous useful suggestions. This work was supported by grants from the Clayton Foundation, Defense Threat Reduction Agency and US National Institutes of Health (G.G.).

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Schroeder, H.W. Jr. Similarity and divergence in the development and expression of the mouse and human antibody repertoires. *Dev. Comp. Immunol.* **30**, 119–135 (2006).
2. Janeway, C.A. *Immunobiology* (Garland Science, 2004).
3. Apostoaei, A.J. & Trabalka, J.R. *Review, Synthesis, and Application of Information on the Human Lymphatic System to Radiation Dosimetry for Chronic Lymphocytic Leukemia* (SENES Oak Ridge, Inc., Oak Ridge, TN, 2010).
4. Weinstein, J.A., Zeng, X., Chien, Y.H. & Quake, S.R. Correlation of gene expression and genome mutation in single B-cells. *PLoS ONE* **8**, e67624 (2013).
5. Brady, B.L., Steinel, N.C. & Bassing, C.H. Antigen receptor allelic exclusion: an update and reappraisal. *J. Immunol.* **185**, 3801–3808 (2010).
6. Xu, J.L. & Davis, M.M. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* **13**, 37–45 (2000).
7. Ippolito, G.C. *et al.* Forced usage of positively charged amino acids in immunoglobulin CDR-H3 impairs B cell development and antibody production. *J. Exp. Med.* **203**, 1567–1578 (2006).
8. McHeyzer-Williams, M., Okitsu, S., Wang, N. & McHeyzer-Williams, L. Molecular programming of B cell memory. *Nat. Rev. Immunol.* **12**, 24–34 (2012).
9. Victora, G.D. & Nussenzweig, M.C. Germinal centers. *Annu. Rev. Immunol.* **30**, 429–457 (2012).
10. Capra, J.D. & Kehoe, J.M. Variable region sequences of five human immunoglobulin heavy chains of the VH3 subgroup: definitive identification of four heavy chain hypervariable regions. *Proc. Natl. Acad. Sci. USA* **71**, 845–848 (1974).
11. Hibi, T. & Dosch, H.M. Limiting dilution analysis of the B cell compartment in human bone marrow. *Eur. J. Immunol.* **16**, 139–145 (1986).
12. Zhang, Z. VH replacement in mice and humans. *Trends Immunol.* **28**, 132–137 (2007).
13. Schroeder, H.W. Jr., Hillson, J.L. & Perlmutter, R.M. Early restriction of the human antibody repertoire. *Science* **238**, 791–793 (1987).
14. Boyd, S.D., Liu, Y., Wang, C., Martin, V. & Dunn-Walters, D.K. Human lymphocyte repertoires in ageing. *Curr. Opin. Immunol.* **25**, 511–515 (2013).
15. Varettoni, M. *et al.* Clues to pathogenesis of Waldenstrom macroglobulinemia and immunoglobulin M monoclonal gammopathy of undetermined significance provided by analysis of immunoglobulin heavy chain gene rearrangement and clustering of B-cell receptors. *Leuk. Lymphoma* **54**, 2485–2489 (2013).
16. Kyle, R.A. *et al.* Prevalence of monoclonal gammopathy of undetermined significance. *N. Engl. J. Med.* **354**, 1362–1369 (2006).
17. Ehlich, A., Martin, V., Muller, W. & Rajewsky, K. Analysis of the B-cell progenitor compartment at the level of single cells. *Curr. Biol.* **4**, 573–583 (1994).
18. Klein, U., Rajewsky, K. & Kuppers, R. Human immunoglobulin (Ig)M+IgD+ peripheral blood B cells expressing the CD27 cell surface antigen carry somatically mutated variable region genes: CD27 as a general marker for somatically mutated (memory) B cells. *J. Exp. Med.* **188**, 1679–1689 (1998).
19. Kuppers, R., Zhao, M., Hansmann, M.L. & Rajewsky, K. Tracing B cell development in human germinal centres by molecular analysis of single cells picked from histological sections. *EMBO J.* **12**, 4955–4967 (1993).
20. Wilson, P.C. & Andrews, S.F. Tools to therapeutically harness the human antibody response. *Nat. Rev. Immunol.* **12**, 709–719 (2012).
21. Tiller, T. *et al.* Autoreactivity in human IgG+ memory B cells. *Immunity* **26**, 205–213 (2007).
22. Corti, D. *et al.* A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. *Science* **333**, 850–856 (2011).
23. Corti, D. & Lanzavecchia, A. Broadly neutralizing antiviral antibodies. *Annu. Rev. Immunol.* **31**, 705–742 (2013).
24. Wardemann, H. *et al.* Predominant autoantibody production by early human B cell precursors. *Science* **301**, 1374–1377 (2003).
25. Meffre, E. & Wardemann, H. B-cell tolerance checkpoints in health and autoimmunity. *Curr. Opin. Immunol.* **20**, 632–638 (2008).
26. Scheid, J.F. *et al.* Differential regulation of self-reactivity discriminates between IgG+ human circulating memory B cells and bone marrow plasma cells. *Proc. Natl. Acad. Sci. USA* **108**, 18044–18048 (2011).
27. Di Niro, R. *et al.* High abundance of plasma cells secreting transglutaminase 2-specific IgA autoantibodies with limited somatic hypermutation in celiac disease intestinal lesions. *Nat. Med.* **18**, 441–445 (2012).
28. Amara, K. *et al.* Monoclonal IgG antibodies generated from joint-derived B cells of RA patients have a strong bias toward citrullinated autoantigen recognition. *J. Exp. Med.* **210**, 445–455 (2013).
29. Traggiai, E. *et al.* An efficient method to make human monoclonal antibodies from memory B cells: potent neutralization of SARS coronavirus. *Nat. Med.* **10**, 871–875 (2004).
30. Wrammert, J. *et al.* Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature* **453**, 667–671 (2008).
31. Yu, X. *et al.* Neutralizing antibodies derived from the B cells of 1918 influenza pandemic survivors. *Nature* **455**, 532–536 (2008).
32. Walker, L.M. *et al.* Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. *Science* **326**, 285–289 (2009).
33. Scheid, J.F. *et al.* Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals. *Nature* **458**, 636–640 (2009).
34. Beltramello, M. *et al.* The human immune response to Dengue virus is dominated by highly cross-reactive antibodies endowed with neutralizing and enhancing activity. *Cell Host Microbe* **8**, 271–283 (2010).
35. Muellenbeck, M.F. *et al.* Atypical and classical memory B cells produce Plasmodium falciparum neutralizing antibodies. *J. Exp. Med.* **210**, 389–399 (2013).
36. Kwong, P.D. & Mascola, J.R. Human antibodies that neutralize HIV-1: identification, structures, and B cell ontogenies. *Immunity* **37**, 412–425 (2012).
37. Burton, D.R. *et al.* A Blueprint for HIV Vaccine Discovery. *Cell Host Microbe* **12**, 396–407 (2012).
38. Haynes, B.F., Kelsoe, G., Harrison, S.C. & Kepler, T.B. B-cell-lineage immunogen design in vaccine development with HIV-1 as a case study. *Nat. Biotechnol.* **30**, 423–433 (2012).
39. Klein, U., Kuppers, R. & Rajewsky, K. Evidence for a large compartment of IgM-expressing memory B cells in humans. *Blood* **89**, 1288–1298 (1997).
40. Glanville, J. *et al.* Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc. Natl. Acad. Sci. USA* **108**, 20066–20071 (2011).
41. Maecker, H.T. *et al.* New tools for classification and monitoring of autoimmune diseases. *Nat. Rev. Rheumatol.* **8**, 317–328 (2012).
42. Kaminski, D.A., Wei, C., Qian, Y., Rosenberg, A.F. & Sanz, I. Advances in human B cell phenotypic profiling. *Front. Immunol.* **3**, 302 (2012).
43. Boyd, S.D. *et al.* Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J. Immunol.* **184**, 6986–6992 (2010).
44. Wang, Y. *et al.* Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. *Immunogenetics* **63**, 259–265 (2011).
45. Kidd, M.J. *et al.* 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 (2012).
46. Watson, C.T. *et al.* 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 (2013).
47. Boyd, S.D. *et al.* Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci. Transl. Med.* **1**, 12ra23 (2009).
48. Aoki-Ota, M., Torkamani, A., Ota, T., Schork, N. & Nemazee, D. Skewed primary Igkappa repertoire and V-J joining in C57BL/6 mice: implications for recombination accessibility and receptor editing. *J. Immunol.* **188**, 2305–2315 (2012).
49. Larimore, K., McCormick, M.W., Robins, H.S. & Greenberg, P.D. Shaping of human germline IgH repertoires revealed by deep sequencing. *J. Immunol.* **189**, 3221–3230 (2012).
50. Mohr, S. *et al.* Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA* **19**, 958–970 (2013).
51. Weinstein, J.A., Jiang, N., White, R.A. III, Fisher, D.S. & Quake, S.R. High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324**, 807–810 (2009).
52. Rubelt, F. *et al.* Onset of immune senescence defined by unbiased pyrosequencing of human immunoglobulin mRNA repertoires. *PLoS ONE* **7**, e49774 (2012).
53. Prabakaran, P., Streaker, E., Chen, W. & Dimitrov, D.S. 454 antibody sequencing - error characterization and correction. *BMC Res. Notes* **4**, 404 (2011).
54. Benichou, J., Ben-Hamo, R., Louzoun, Y. & Efroni, S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* **135**, 183–191 (2012).
55. Baum, P.D., Venturi, V. & Price, D.A. Wrestling with the repertoire: the promise and perils of next generation sequencing for antigen receptors. *Eur. J. Immunol.* **42**, 2834–2839 (2012).
56. Bolotin, D.A. *et al.* Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. *Eur. J. Immunol.* **42**, 3073–3083 (2012).
57. Mehr, R., Sternberg-Simon, M., Michaeli, M. & Pickman, Y. Models and methods for analysis of lymphocyte repertoire generation, development, selection and evolution. *Immunol. Lett.* **148**, 11–22 (2012).
58. Campbell, P.J. *et al.* Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc. Natl. Acad. Sci. USA* **105**, 13081–13086 (2008).
59. Nguyen, P. *et al.* Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. *BMC Genomics* **12**, 106 (2011).
60. Kircher, M., Heyn, P. & Kelso, J. Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics* **12**, 382 (2011).
61. Loman, N.J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **30**, 434–439 (2012).
62. Michaeli, M., Noga, H., Tabibian-Keissar, H., Barshack, I. & Mehr, R. Automated cleaning and pre-processing of immunoglobulin gene sequences from high-throughput sequencing. *Front. Immunol.* **3**, 386 (2012).
63. Zhu, J. *et al.* Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proc. Natl. Acad. Sci. USA* **110**, 6470–6475 (2013).

64. Glanville, J. *et al.* Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci. USA* **106**, 20216–20221 (2009).
65. Chen, Z., Collins, A.M., Wang, Y. & Gaeta, B.A. Clustering-based identification of clonally-related immunoglobulin gene sequence sets. *Immunome Res.* **6** (suppl. 1), S4 (2010).
66. Wine, Y. *et al.* Molecular deconvolution of the monoclonal antibodies that comprise the polyclonal serum response. *Proc. Natl. Acad. Sci. USA* **110**, 2993–2998 (2013).
67. Zhu, J. *et al.* Somatic populations of PGT135–137 HIV-1-neutralizing antibodies identified by 454 pyrosequencing and bioinformatics. *Frontiers Microbiol.* **3**, 315 (2012).
68. Larsen, P.A. & Smith, T.P. Application of circular consensus sequencing and network analysis to characterize the bovine IgG repertoire. *BMC Immunol.* **13**, 52 (2012).
69. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. USA* **108**, 9530–9535 (2011).
70. Fu, G.K., Hu, J., Wang, P.H. & Fodor, S.P. Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proc. Natl. Acad. Sci. USA* **108**, 9026–9031 (2011).
71. Shiroguchi, K., Jia, T.Z., Sims, P.A. & Xie, X.S. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc. Natl. Acad. Sci. USA* **109**, 1347–1352 (2012).
72. Vollmers, C., Sit, R.V., Weinstein, J.A., Dekker, C.L. & Quake, S.R. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc. Natl. Acad. Sci. USA* **110**, 13463–13468 (2013).
73. Schmitt, M.W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. USA* **109**, 14508–14513 (2012).
74. Meijer, P.J. *et al.* Isolation of human antibody repertoires with preservation of the natural heavy and light chain pairing. *J. Mol. Biol.* **358**, 764–772 (2006).
75. Busse, C.E., Czogiel, I., Braun, P., Arndt, P.F. & Wardemann, H. Single-cell based high-throughput sequencing of full-length immunoglobulin heavy and light chain genes. *Eur. J. Immunol.* doi:10.1002/eji.201343917 (22 October 2013).
76. DeKosky, B.J. *et al.* High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.* **31**, 166–169 (2013).
77. Church, G.M., Vigneault, F., Laserson, U. & Bachelet, I. High-throughput immune sequencing. US patent US20130296535 (2013).
78. Lakhani, K.R. *et al.* Prize-based contests can provide solutions to computational biology problems. *Nat. Biotechnol.* **31**, 108–111 (2013).
79. Kidd, B.A., Peters, L.A., Schadt, E.E. & Dudley, J.T. Unifying immunology with informatics and multiscale biology. *Nat. Immunol.* **15,** 118–127 (2014).
80. Rempala, G.A., Seweryn, M. & Ignatowicz, L. Model for comparative analysis of antigen receptor repertoires. *J. Theor. Biol.* **269**, 1–15 (2011).
81. Mora, T., Walczak, A.M., Bialek, W. & Callan, C.G. Jr. Maximum entropy models for antibody diversity. *Proc. Natl. Acad. Sci. USA* **107**, 5405–5410 (2010).
82. Yaari, G., Uduman, M. & Kleinstein, S.H. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. *Nucleic Acids Res.* **40**, e134 (2012).
83. Edgar, R. & Barrett, T. NCBI GEO standards and services for microarray data. *Nat. Biotechnol.* **24**, 1471–1472 (2006).
84. Brusic, V., Gottardo, R., Kleinstein, S.H., Davis, M.M. & the HIPC Steering Committee. Computational resources for high-dimension immune analysis from the Human Immunology Project Consortium. *Nat. Biotechnol.* doi:10.1038/nbt.2777 (19 January 2014).
85. Bradbury, A.R., Sidhu, S., Dubel, S. & McCafferty, J. Beyond natural antibodies: the power of *in vitro* display technologies. *Nat. Biotechnol.* **29**, 245–254 (2011).
86. Chan, A.C. & Carter, P.J. Therapeutic antibodies for autoimmunity and inflammation. *Nat. Rev. Immunol.* **10**, 301–316 (2010).
87. Scott, A.M., Wolchok, J.D. & Old, L.J. Antibody therapy of cancer. *Nat. Rev. Cancer* **12**, 278–287 (2012).
88. Ge, X., Mazor, Y., Hunicke-Smith, S.P., Ellington, A.D. & Georgiou, G. Rapid construction and characterization of synthetic antibody libraries without DNA amplification. *Biotechnol. Bioeng.* **106**, 347–357 (2010).
89. Mahon, C.M. *et al.* Comprehensive interrogation of a minimalist synthetic CDR-H3 library and its ability to generate antibodies with therapeutic potential. *J. Mol. Biol.* **425**, 1712–1730 (2013).
90. Fischer, N. Sequencing antibody repertoires: the next generation. *MAbs* **3**, 17–20 (2011).
91. Ravn, U. *et al.* Deep sequencing of phage display libraries to support antibody discovery. *Methods* **60**, 99–110 (2013).
92. Ravn, U. *et al.* By-passing in vitro screening–next generation sequencing technologies applied to antibody display and in silico candidate selection. *Nucleic Acids Res.* **38**, e193 (2010).
93. Larman, H.B., Xu, G.J., Pavlova, N.N. & Elledge, S.J. Construction of a rationally designed antibody platform for sequencing-assisted selection. *Proc. Natl. Acad. Sci. USA* **109**, 18523–18528 (2012).
94. Reddy, S.T. *et al.* Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat. Biotechnol.* **28**, 965–969 (2010).
95. Saggy, I. *et al.* Antibody isolation from immunized animals: comparison of phage display and antibody discovery via V gene repertoire mining. *PEDS* **25**, 539–549 (2012).
96. Zhu, J. *et al.* 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 (2013).
97. Lavinder, J.J. *et al.* Proteomic identification of antibodies. US patent 20130178370. (2013).
98. Cheung, W.C. *et al.* A proteomics approach for the identification and cloning of monoclonal antibodies from serum. *Nat. Biotechnol.* **30**, 447–452 (2012).
99. Sato, S. *et al.* Proteomics-directed cloning of circulating antiviral human monoclonal antibodies. *Nat. Biotechnol.* **30**, 1039–1043 (2012).
100. Jiang, N. *et al.* Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proc. Natl. Acad. Sci. USA* **108**, 5348–5353 (2011).
101. Ota, M. *et al.* Regulation of the B cell receptor repertoire and self-reactivity by BAFF. *J. Immunol.* **185**, 4128–4136 (2010).
102. Arnaout, R. *et al.* High-resolution description of antibody heavy-chain repertoires in humans. *PLoS ONE* **6**, e22365 (2011).
103. Gronwall, C., Kosakovsky Pond, S.L., Young, J.A. & Silverman, G.J. *In vivo* VL-targeted microbial superantigen induced global shifts in the B cell repertoire. *J. Immunol.* **189**, 850–859 (2012).
104. Wang, F. *et al.* Reshaping antibody diversity. *Cell* **153**, 1379–1393 (2013).
105. Ippolito, G.C. *et al.* Antibody repertoires in humanized NOD-scid-IL2Rgamma(null) mice and human B cells reveals human-like diversification and tolerance checkpoints in the mouse. *PLoS ONE* **7**, e35497 (2012).
106. Castro, R. *et al.* Teleost fish mount complex clonal IgM and IgT responses in spleen upon systemic viral infection. *PLoS Pathog.* **9**, e1003098 (2013).
107. Wu, L. *et al.* Fundamental characteristics of the immunoglobulin VH repertoire of chickens in comparison with those of humans, mice, and camelids. *J. Immunol.* **188**, 322–333 (2012).
108. Wang, Y. *et al.* IgE sequences in individuals living in an area of endemic parasitism show little mutational evidence of antigen selection. *Scand. J. Immunol.* **73**, 496–504 (2011).
109. Wu, Y.C. *et al.* High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* **116**, 1070–1078 (2010).
110. Prabakaran, P. *et al.* Origin, diversity, and maturation of human antiviral antibodies analyzed by high-throughput sequencing. *Front. Microbiol.* **3**, 277 (2012).
111. Vollmers, C., Sit, R.V., Weinstein, J.A., Dekker, C.L. & Quake, S.R. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc. Natl. Acad. Sci. USA* **110**, 13463–13468 (2013).
112. Briney, B.S., Willis, J.R., Hicar, M.D., Thomas, J.W. II & Crowe, J.E. Jr. Frequency and genetic characterization of V(DD)J recombinants in the human peripheral blood antibody repertoire. *Immunology* **137**, 56–64 (2012).
113. Benichou, J. *et al.* The restricted DH gene reading frame usage in the expressed human antibody repertoire is selected based upon its amino acid content. *J. Immunol.* **190**, 5567–5577 (2013).
114. Briney, B.S., Willis, J.R., McKinney, B.A. & Crowe, J.E. Jr. 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 (2012).
115. Briney, B.S., Willis, J.R. & Crowe, J.E. Jr. Location and length distribution of somatic hypermutation-associated DNA insertions and deletions reveals regions of antibody structural plasticity. *Genes Immun.* **13**, 523–529 (2012).
116. Zemlin, M. *et al.* Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. *J. Mol. Biol.* **334**, 733–749 (2003).
117. Lindner, C. *et al.* Age, microbiota, and T cells shape diverse individual IgA repertoires in the intestine. *J. Exp. Med.* **209**, 365–377 (2012).
118. Wesemann, D.R. *et al.* Microbial colonization influences early B-lineage development in the gut lamina propria. *Nature* **501**, 112–115 (2013).
119. von Budingen, H.C. *et al.* B cell exchange across the blood-brain barrier in multiple sclerosis. *J. Clin. Invest.* **122**, 4533–4543 (2012).
120. Jiang, N. *et al.* Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci. Transl. Med.* **5**, 171ra119 (2013).
121. Jackson, K.J. *et al.* Divergent human populations show extensive shared IGK rearrangements in peripheral blood B cells. *Immunogenetics* **64**, 3–14 (2012).
122. Hoi, K.H. & Ippolito, G.C. Intrinsic bias and public rearrangements in the human immunoglobulin Vlambda light chain repertoire. *Genes Immun.* **14**, 271–276 (2013).
123. Schoettler, N., Ni, D. & Weigert, M. B cell receptor light chain repertoires show signs of selection with differences between groups of healthy individuals and SLE patients. *Mol. Immunol.* **51**, 273–282 (2012).
124. Racanelli, V. *et al.* Antibody V(h) repertoire differences between resolving and chronically evolving hepatitis C virus infections. *PLoS ONE* **6**, e25606 (2011).
125. Parameswaran, P. *et al.* Convergent antibody signatures in human dengue. *Cell Host Microbe* **13**, 691–700 (2013).
126. Ademokun, A. *et al.* Vaccination-induced changes in human B-cell repertoire and pneumococcal IgM and IgA antibody at different ages. *Aging Cell* **10**, 922–930 (2011).
127. Krause, J.C. *et al.* Epitope-specific human influenza antibody repertoires diversify by B cell intraclonal sequence divergence and interclonal convergence. *J. Immunol.* **187**, 3704–3711 (2011).
128. Kwong, P.D., Mascola, J.R. & Nabel, G.J. Broadly neutralizing antibodies and the search for an HIV-1 vaccine: the end of the beginning. *Nat. Rev. Immunol.* **13**, 693–701 (2013).

129. Wu, X. *et al.* Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* **333**, 1593–1602 (2011).
130. Zhu, J. *et al.* Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proc. Natl. Acad. Sci. USA* **110**, 6470–6475 (2013).
131. Liao, H.X. *et al.* Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* **496**, 469–476 (2013).
132. Zhou, T. *et al.* Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-class antibodies. *Immunity* **39**, 245–258 (2013).
133. Wiley, S.R. *et al.* Targeting TLRs expands the antibody repertoire in response to a malaria vaccine. *Sci. Transl. Med.* **3**, 93ra69 (2011).
134. Srivastava, R. *et al.* Potassium channel KIR4.1 as an immune target in multiple sclerosis. *N. Engl. J. Med.* **367**, 115–123 (2012).
135. de Buy Wenniger, L.J. *et al.* IgG4+ clones identified by next-generation sequencing dominate the b-cell receptor repertoire in IgG4-associated cholangitis. *Hepatology* **57**, 2390–2398 (2013).
136. He, J. *et al.* IgH gene rearrangements as plasma biomarkers in Non- Hodgkin's lymphoma patients. *Oncotarget* **2**, 178–185 (2011).
137. Logan, A.C. *et al.* High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *Proc. Natl. Acad. Sci. USA* **108**, 21194–21199 (2011).
138. Faham, M. *et al.* Deep-sequencing approach for minimal residual disease detection in acute lymphoblastic leukemia. *Blood* **120**, 5173–5180 (2012).
139. Gawad, C. *et al.* Massive evolution of the immunoglobulin heavy chain locus in children with B precursor acute lymphoblastic leukemia. *Blood* **120**, 4407–4417 (2012).
140. Tschumper, R.C. *et al.* Comprehensive assessment of potential multiple myeloma immunoglobulin heavy chain V-D-J intraclonal variation using massively parallel pyrosequencing. *Oncotarget* **3**, 502–513 (2012).
141. Darzentas, N. & Stamatopoulos, K. Stereotyped B cell receptors in B cell leukemias and lymphomas. *Methods Mol. Biol.* **971**, 135–148 (2013).
142. Hirano, M. *et al.* Evolutionary implications of a third lymphocyte lineage in lampreys. *Nature* **501**, 435–438 (2013).
143. Silverstein, A.M. History of immunology. *Cell. Immunol.* **42**, 1–2 (1979).