Trends in Immunology

CellPress

Review

Novel Approaches to Analyze Immunoglobulin Repertoires

Hedda Wardemann[1,*] and Christian E. Busse[1,*]

Analysis of immunoglobulin (Ig) repertoires aims to comprehend Ig diversity with the goal of predicting humoral immune responses in the context of infection, vaccination, autoimmunity, and malignancies. The first next-generation sequencing (NGS) analyses of bulk B cell populations dramatically advanced sampling depth over previous low-throughput single-cell-based protocols, albeit at the expense of accuracy and loss of chain-pairing information. In recent years the field has substantially differentiated, with bulk analyses becoming more accurate while single-cell approaches have gained in throughput. Additionally, new platforms striving to combine high throughput and chain pairing have been developed as well as various computational tools for analysis. Here we review the developments of the past 4–5 years and discuss the open challenges.

Trends

Advances in next-generation sequencing have ushered in a new era of immunoglobulin (Ig) repertoire analysis.

Platforms for large-scale single-cell sequencing allow us to obtain both Ig chains of a cell in the correct association and thus enable cloning and recombinant expression of the respective antibody.

Computational tools now allow the reconstitution of antibody lineages and the analysis of somatic hypermutation. However, the prediction of antigen binding directly from an antibody sequence remains elusive.

Data sharing and reporting standards are currently elaborated by the Adaptive Immune Receptor Repertoire Community to facilitate data exchange and secondary analysis by third parties.

Diversity in the Ig Repertoire

Since the seminal discovery of genetic rearrangement as the primary mechanism of Ig variability [1], immunologists have used sequence analysis to measure and interpret the vast diversity of the Ig repertoire [2]. Iterative studies aiming to estimate the number of gene **segments** available (see Glossary) to the recombination process [3,4] were successfully concluded with the availability of multiple high-resolution mammalian genomes, setting the numbers used today (Table 1). Based on these variable (V), diversity (D) and joining (J) segment counts and V(D)J joints generated by rearrangement, the potential diversity of the Ig loci both in humans and in mice has been estimated to range from 10^{12} to 10^{14}. However, these theoretical figures differ from the actual repertoire size (i.e., the receptor diversity present in an individual host at any given time) due to various limitations and biases. First, the number of B cells in the organism is limited and often several orders of magnitude lower than the number of possible rearrangements. This is especially pronounced in small model animals like mice, which harbor only around 10^8 B cells. Second, several parameters of such diversity estimates (e.g., gene segment usage, *IGH* VDJ joint variability) exhibit non-uniform distributions *in vivo*, resulting in diversity below the theoretical maximum [5]. Segment usage distributions have been shown to be influenced by a variety of factors, including the chromosomal location of a segment, chromatin state, and the quality of the recombination signal sequence (RSS) [6,7]. Third, there are structural and reactivity constrains, as cells with non-functional or autoreactive receptors will be deleted, thereby excluding such rearrangements from the repertoire. Fourth, on exposure to external antigens the repertoire will be shaped on two levels. On the cellular level, B cells encountering their cognate antigen will undergo proliferation and differentiation towards memory and **antibody-secreting cells (ASCs)**. At the molecular level, the Ig loci in these cells undergo somatic hypermutation (SHM), the secondary mechanism of Ig variability. As SHM targets positions throughout the V segment, it creates additional variants that cannot be generated by the primary mechanism (i.e., rearrangement). Given this complexity the detection of consistent and potentially predictive signals in the Ig repertoire against the background noise requires direct measurements of sufficient depth and quantitative power. Here we review the

[1]Division of B Cell Immunology, German Cancer Research Center, Heidelberg, Germany

*Correspondence:
h.wardemann@dkfz-heidelberg.de
(H. Wardemann) and
christian.busse@dkfz-heidelberg.de
(C.E. Busse).

Table 1. Segment Counts (Functional Segments in Parentheses)

Species	Locus	V	D	J	Refs
Homo sapiens	*IGH*	105 (55)	27	9 (6)	[61,62]
H. sapiens	*IGK*	76 (34)	n/a	5 (5)	[63]
H. sapiens	*IGL*	69 (30)	n/a	7 (4)	[64]
Mus musculus	*Igh*	190 (110)	12	4 (4)	[65]
M. musculus	*Igk*	140 (80)	n/a	5 (4)	[66]
M. musculus	*Igl*	4 (3)	n/a	5 (4)	[67]

The numbers shown in this table should be considered estimates, as they can vary between individual haplotypes and new alleles continue to be described [55,56].

recent experimental and computational developments in Ig repertoire sequencing that have brought us closer to the goal of making meaningful inferences for Ig repertoire data. In the context of this review, we define Ig repertoire studies as diversity assessments of the Ig heavy, kappa, and lambda V **regions**. We therefore restrict ourselves to methodologies that primarily aim to provide this information. Techniques whose primary end point is the preservation and/or cultivation of B cells are therefore not considered even if they would allow sequencing.

Common Experimental Setups

Conventional Bulk Sequencing

Bulk sequencing of B cell populations by NGS was first reported in 2009 [8]. The main considerations in this experimental approach center on the choice of tissue and cell population, the number of cells, and the type of nucleic acid to be used as the template (Table 2). Most workflows use cell subsets defined by flow cytometry as the starting material; however, as Ig rearrangement is restricted to B cells and only the rearranged locus can be amplified by PCR (Figure 1), a wide range of sample types can be used, including solid tissue blocks. Subsequently, the V(D)J joints of the targeted Ig loci are amplified by [reverse transcription (RT)-]PCR and then sequenced on Illumina HiSeq or MiSeq platforms (Box 1). The main advantages of the bulk sequencing approach are the relatively low cost and the high sensitivity for the detection of rare rearrangements due to its ability to process millions of cells. Furthermore, the experimental protocols are widely established and do not require special equipment. Finally, while early studies were focused on the V(D)J joint due to length restriction, the read-length extension of the current Illumina platforms has now facilitated the assessment of the complete V region. However, this classical setup suffers from multiple problems, which limits the use and interpretation of the data obtained (Figure 2).

Table 2. Choice of Template

	gDNA	mRNA/cDNA
Advantages	• Stability, allowing a wider range of samples [e.g., formalin fixed paraffin embedded (FFPE)] • Homogeneous amounts of template per rearrangement and cell	• More material • Isotype information
Disadvantages	• Only one functionally rearranged template molecule per cell, reduced sensitivity • Loss of isotype information	• Requires additional RT step • Potentially additional errors introduced by RNA polymerase and RT • Non-productive transcripts are subject to nonsense-mediated RNA decay (NMD)

The choice of the nucleic acid to be used as the primary template is dependent on the experiment, with the respective advantages and disadvantages of gDNA and mRNA being complementary. It should be noted that the strength of the individual effects is dependent on the methodology used and the experimental setup.

Glossary

Antibody-secreting cell (ASC): a functional classification for plasma cells and plasmablasts independent of their differentiation status.

B cell receptor (BCR): the protein complex on the surface of B cells comprising a membrane-bound Ig and various signal transduction components like Igα and Igβ. Note that the exact stoichiometry of the components can vary depending on the Ig isotype.

Birthday effect: the fact that possible overlaps within a set of sequences scales with the square of the sequence number, making collisions more likely than intuitively assumed; named after the observation that a random group of 23 people will more likely than not have one pair among them who share their birthday.

Domain: a structurally independent part of a protein. The Ig domain has a β-barrel structure containing a single cysteine–cysteine bond. Ig domains are typically encoded by individual exons; however, in the case of the V domain this exon does not yet exist in the germline but will be generated only on V(D)J rearrangement (also see Region).

Hamming distance: a metric for the distance between two sequences of equal length; defined as the minimum number of replacement changes in one sequence that are required to make it identical to the other.

Region: in the strict sense, the stretch of a nucleic acid that encodes a single **domain**. However, this terminology is consistently used only for the V region, while the term ‘C region’ often refers to the coding sequence for all constant domains in a molecule.

Segment: part of the genomic Ig loci that is subject to rearrangement. A segment can belong to the V, D, or J type, with V and J segments being present in all three Ig loci while D segments are found only in *IGH*. The rearranged V, D, and J segments constitute the V region, encoding the V domain.

Unique molecular identifiers (UMIs): used to tag individual mRNA transcripts; provide a reliable solution to problems of quantification and quality as they can suppress the effects of amplification bias and allow correction of PCR errors by consensus building.

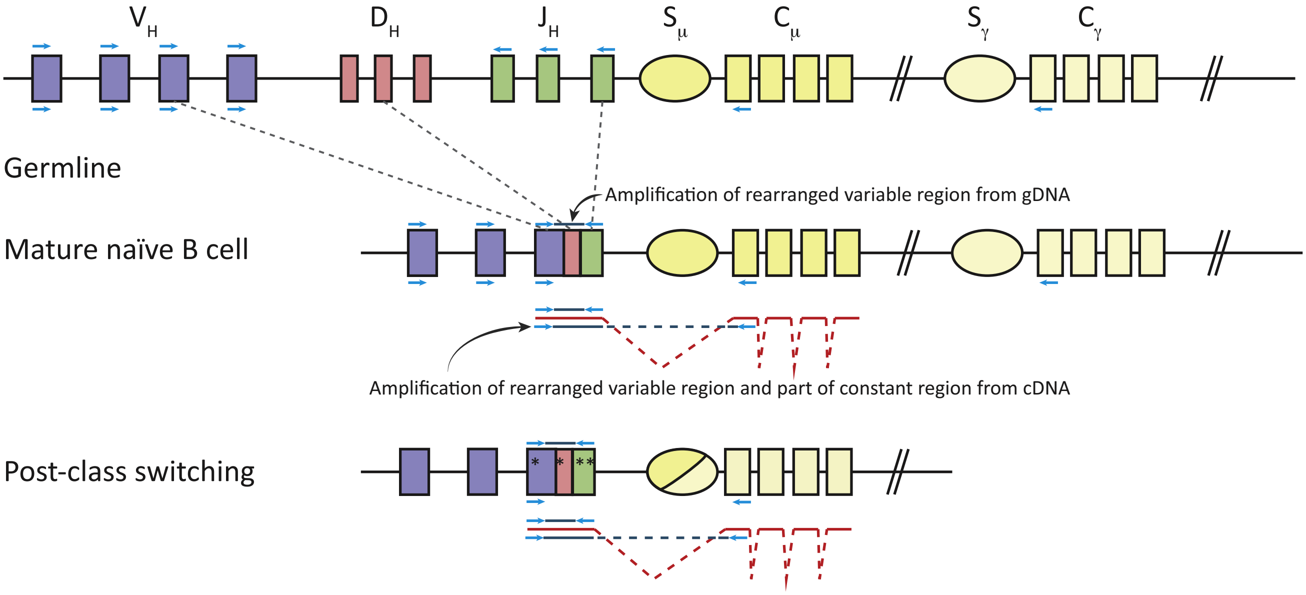
V H
D H
J H
S μ
C μ
S γ
C γ
Germline
Mature naïve B cell
Amplification of rearranged variable region from gDNA
Amplification of rearranged variable region and part of constant region from cDNA
Post-class switching
* * **
Trends in Immunology

Figure 1. Amplification Templates on IGH. Depicted here is a schematic version of an IGH locus. The top shows the locus in its germline configuration. V/J primer pairs (arrows above the line) will not produce amplicons, as the distance between the binding sites is too large. The same applies to V/C primer pairs (arrows below the line). During B cell development, V(D)J rearrangement occurs; the middle panels shows such a rearranged locus as it is found in mature naïve (MN) B cells. The red line indicates the mature mRNA produced from this locus after excision of introns (dashed). Here, the innermost V/J primer pair will amplify the VDJ joint (dashed line). Due to the distance, V/C primer pairs require cDNA (instead of gDNA) templates, in which the J-C intronic region is spliced out. At the bottom, the situation after class-switch recombination (CSR) and somatic hypermutation (SHM) (depicted as asterisks) is shown. Note that the amplicon generated by the V/C primer pair will include part of the switched constant region and is thus informative regarding the isotype. The rearrangement of the light chain loci occurs accordingly; however, these loci lack D segments and do not undergo CSR. Note that the switch between secretory and transmembrane versions of the heavy chain is not depicted here.

Loss of Chain-Pairing Information

Due to the pooling of a huge number of cells, information about the Ig heavy:light chain association present within individual cells is lost. However, many advanced applications that go beyond the use of *IGH* CDR3 as a cellular identifier benefit from the sequence information of both chains for in-depth analysis of clonal relationships, or even require it (e.g., for functional assessment of recombinant antibodies).

Non-uniformity of Ig Expression

The various stages of B cell development and differentiation show substantial variability in Ig expression levels (i.e., Ig transcripts per cell), with differences of several orders of magnitude. It is therefore difficult to quantitatively infer the cellular composition of a sample when bulk sequencing is performed using mRNA as a template. This is further complicated by the potential inadvertent contamination of large low-Ig-expressing populations by single high expressers (e.g., ASCs). Mitigation of this problem is possible using genomic DNA (gDNA) as the template but comes at the cost of losing isotype information, as the genomic distance between *IGH* J and the first constant region exon is too large to be amplified efficiently.

Biased Amplification

Ig loci can contain several dozens of functional V segments (Table 1), which exhibit sequence identities between each other that range from 50% up to 100%. Since every primer used for PCR amplification has a unique binding profile for the respective V segment [9], the overall amplification factor for the individual segment cannot be reliably predicted, especially when mixtures of multiple primers are used. Furthermore, PCR amplification does often not follow the ideal exponential trajectory but is affected by binary stochastic effects determining the onset of

Box 1. NGS Platforms

While sequencing technology is still rapidly evolving, the following three platforms are by far the most prominent.

Illumina is the current standard platform for NGS. It immobilizes individual DNA molecules on a surface and amplifies them to generate microscopically confined spots (termed 'clusters'). These clusters are then sequenced using reactivatable dye-terminator chemistry. For repertoire sequencing applications, sequencing is typically performed from both ends of the molecule ('paired-end sequencing'). It is important to note that while the two sequences obtained in this process can always be associated with each other (as they come from the same physical location on the surface), they do not have to overlap. Therefore, complete coverage of V regions became possible only with the advent of 2 × 250- and 2 × 300-bp MiSeq platforms.

The Roche/454 platform was discontinued by the manufacturer in mid-2016; nevertheless, it remains important as there are numerous datasets available that were generated with this technique. Roche/454 relies on capture beads and emulsion PCRs to generate a clonal context; the beads are then sequenced in the individual well of a picotiter plate by pyrosequencing. This technology provided relatively long reads (600–700 bp), which facilitated the sequencing of complete V regions, a unique feature until the 2 × 300-bp MiSeq platform was introduced. The main downside of Roche/454 was its tenfold higher cost per megabase pair. Also, like all pyrosequencing platforms, Roche/454 exhibits difficulties in sequencing longer homopolymer stretches (5+ bp). Although homopolymers are not frequently found in V regions, the tendency of the platform to introduce INDELs, thereby causing frameshifts, has to be accounted for during analysis. While another pyrosequencing technology (IonTorrent) has also been used for repertoire studies [60], it does not provide any genuine advantages versus Roche/454 besides the smaller package size and continued availability.

Pacific Bioscience (PacBio) is a more recent, technologically distinct platform based on single-molecule sequencing. Its unique capability to obtain very long reads (>10 kb) is in contrast to the high base error rate (10%) and the limited throughput of only several ten thousand reads per run.

amplification and by saturation effects at the end. This results in the inability to determine the original amount of template based on read counts.

PCR Artifacts

The PCR amplification in bulk setups is affected by two potential problems: substitution errors and the production of chimeric amplicons. Substitution errors are, in general, often observed in PCR-based protocols. However, the potential presence of SHM in Ig sequences presents a specific challenge for analysis: while consensus building can remove true SHM from the analysis, uncorrected PCR errors will be falsely interpreted as SHMs. Chimeric amplicons are fusions of fragments of independent amplicons, which are assumed to arise due to overlapping conserved regions. In humans and mice the majority of these events can be typically identified by diverging V segment assignments for different parts of the sequence. However, this approach faces its limits with highly related amplicons or in species that use gene conversion to diversify their repertoire [10].

Manual Single-Cell Analysis

Sequencing of Ig genes from single cells was developed in the mid-1990s [11,12]. Although this approach was initially limited to the processing of several dozen cells, these early reports already recognized its central advantages: the preservation of chain-pairing information and the possibility of linking additional information to the cell of origin (e.g., flow cytometry data). The possibility to obtain full-length paired chains facilitated the extension of these techniques by the subsequent cloning and expression of recombinant monoclonal antibodies [13]. This allows single-cell analyses to provide a high-dimensional snapshot of a cell including its Ig sequence, cellular phenotype, and antibody reactivity. Additionally, due to the lack of competition (usually there is only one dominant transcript per cell and locus) and the experimental conditions that allow the PCR to run into saturation, single-cell PCR (scPCR) is not affected by amplification bias. Additional refinements [14–16] made it feasible to handle several hundred cells in parallel. Nevertheless, the high cost and low total number of cells entailed remain the central drawback of the classical single-cell approach. It should be noted

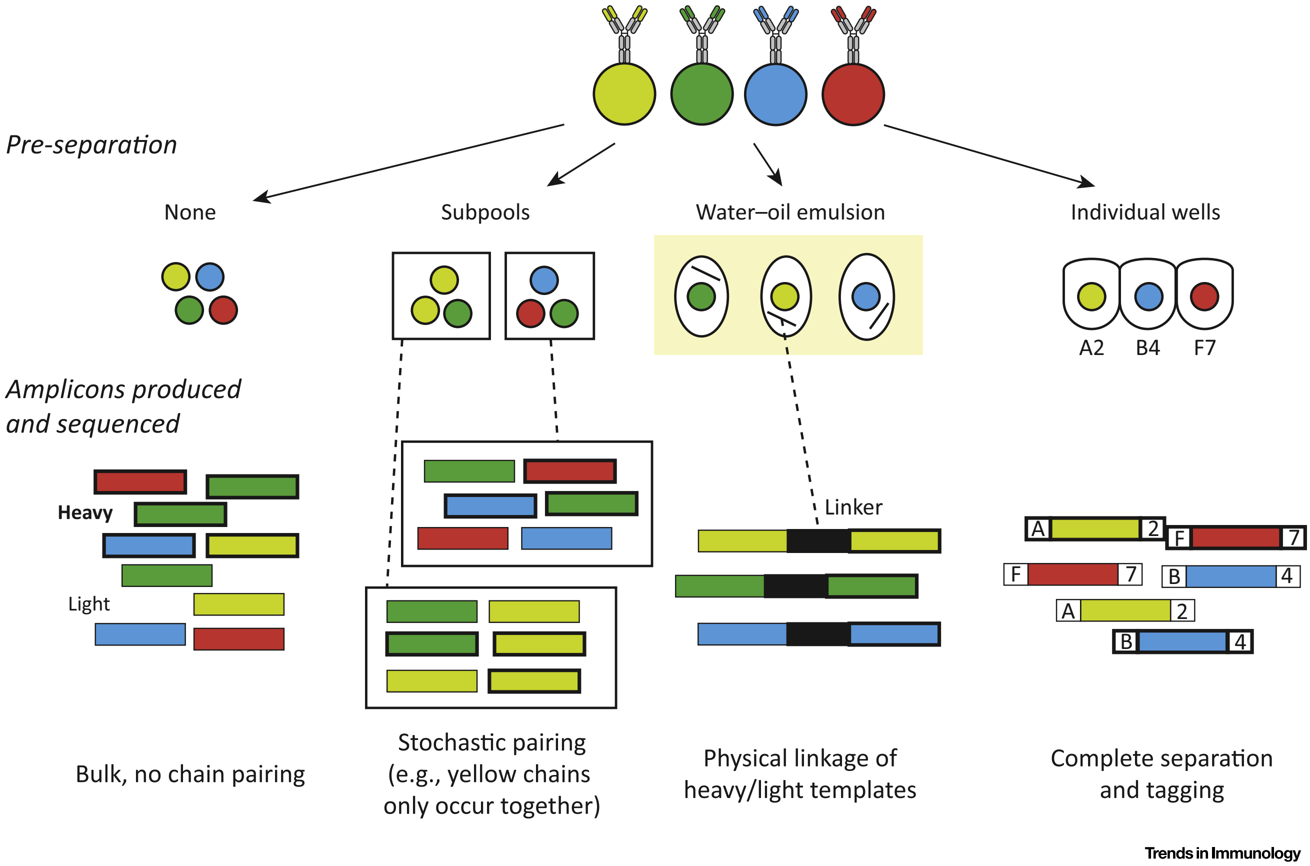
Pre-separation
None
Subpools
Water–oil emulsion
Individual wells
A2
B4
F7
Amplicons produced and sequenced
Heavy
Light
Linker
A
2
F
7
B
4
Bulk, no chain pairing
Stochastic pairing (e.g., yellow chains only occur together)
Physical linkage of heavy/light templates
Complete separation and tagging
Trends in Immunology

Figure 2. Schematic Sample Processing on Different Platforms. This figure depicts the differences in sample processing between platforms. Detailed discussion can be found in the Common Setups and Recent Developments sections. Bulk: No pre-separation is performed; nucleic acids are directly extracted from the sample. No additional labeling is performed during amplification [exception: unique molecular identifier (UMI) protocols]. Sequencing reads can be distinguished between heavy and light chains (indicated here with thick or light borders, respectively) but association is not possible. Stochastic pairing: Subpools of cells are created based on the expected clonal frequency. Processing is identical to bulk but the pools are kept separate. Chain association can be inferred by the consistent presence and absence of the associated chains in the reads obtained from a subpool (example: filled). Physical linkage: Cells are singled out in water-in-oil emulsion and subsequently lysed. The heavy and light chain transcripts within an individual droplet are joined, then the emulsion is broken and the templates are amplified. The paired-end reads obtained from the individual amplicons provide the associated chains, although the complete sequence might not be available. Complete separation: Single cells are isolated by flow cytometry. During amplification, barcodes are included in the amplicon that encode the position of the well within the experiment. Amplicons are mixed only before sequencing and reads can then be demultiplexed using the barcodes.

that while Sanger sequencing is the primary cost- and rate-determining step, it is also required to unambiguously assign a sequence to its cell of origin – information that would be lost on multiplexing. The second technical challenge faced by scPCR protocols is the binary outcome of the amplification reaction, the efficiency of which is substantially influenced by the amount of Ig transcripts in the cell. There have also been many considerations of the design of primer sets, mainly due to concerns that high levels of SHM might inhibit primer binding resulting in reduced scPCR efficiency [17]. However, the part of FWR1 adjacent to the signal sequence has been shown to be relatively conserved and to exhibit low levels of SHM even in antibodies that have undergone extensive hypermutation [18]. This part can thus be utilized as a reliable primer-binding site, allowing the efficient amplification of highly mutated antibody genes. In contrast to primers located further 3′ on the V segment, FWR1 primers do not only allow detection of the vast majority of SHMs present in a V segment, but furthermore provide full-length amplicons that can be directly used for cloning and recombinant antibody production.

Recent Developments

Experimental Approaches

Addressing the Amplification Bias

As discussed above, one central shortcoming of conventional bulk sequencing has been its inability to reliably infer the initial amount of template nucleic acid in a sample from the number of sequence reads observed for a given rearrangement. As quantification is a central aim in many repertoire analyses, various approaches have been developed to address this issue. It should, however, be noted that the problem of non-uniform Ig expression is not mitigated by any of them.

Unique Molecular Identifiers (UMIs)

UMIs are inserted during the RT of RT-PCR by tagging all possible templates before amplification, thereby facilitating quantification unaffected by amplification bias [19]. If reads from both ends of a molecule are combined [9,20], even short UMIs contain sufficient information to allow grouping of reads according to their source cDNA. This facilitates consensus building that suppresses PCR errors and the exclusion of chimeric molecules. However, currently available techniques require template switching during the RT step, which is an inefficient process that could lead to the preferential loss of low-expression transcripts. It should further be noted that although comparable primer extension/adaptor ligation procedures have been reported for recombined genomic *Igh* templates [7], there are currently no protocols utilizing UMIs in combination with gDNA [21].

Quantitative Bulk PCR

An alternative approach is the reduction of amplification bias by both experimental and computational means [22–24]. The cornerstone of this platform is a synthetic library of V-J segments, which is used to optimize the primer set for low amplification bias and to normalize the read counts after sequencing. The approach therefore does not require tagging, which simplifies the switching between mRNA and gDNA templates. However, the lack of a unique terminal sequence tag prevents the detection of false chimeric transcripts. Furthermore, the performance of primer sets on alleles not present in the synthetic library is unknown.

Addressing the Pairing Problem

As noted above, not all biological questions will profit from additional pairing information. Experiments aiming to identify clonal expansion can often be restricted to Ig heavy sequences as the entropy of *IGH* CDR3 is substantially higher than that of light chain rearrangements and thus usually sufficient for clonal identification. However, for experiments trying to assess receptor reactivity (both computationally and by cloning), light chain information is essential. Therefore, several paths have been taken to enhance the throughput of classical single-cell sequencing approaches. While all technologies described below provide chain-pairing information, they differ substantially in terms of throughput, the ability to link additional single-cell data, and their commercial availability.

Stochastic Pairing

The basic principle of this approach is to reconstruct the original Ig chain pairing of expanded clones under conditions of limiting dilution [25]. To this end, pools of cells are produced that are several times smaller than the inverse of the expected frequency of the clone of interest and sequenced by conventional bulk sequencing protocols. As most pools will contain either no or only one cell of the clone of interest, chain linkage can be inferred from the consistent presence or absence of both chains in these pools. Beside the cell pools, stochastic pairing methods do not require changes in the experimental setup and can therefore be combined with quantification strategies as discussed above [25]. It is straightforward to select pool sizes if clonal frequencies are in a narrow range (i.e., order of magnitude) and sampling depth can be high if clones are rare and thus a single pool can comprise a large number of cells. However, samples with broad distributions of clonal frequency will require more sophisticated pooling schemes. The central limitation of this method is its intrinsic dependency on clonal expansion and the possibility of resampling. It therefore cannot be used for highly diverse populations like mature naïve B cells, which exhibit a minimal degree of clonal expansion. Finally, it should be noted that there is a detection limit for unmutated light chains: Due to their low diversity, pools of substantially more than 10^4 cells will contain the vast majority of all rearrangements. Thus, clonal light chains with low frequencies cannot be discriminated from this background.

Physical Linkage

In this diverse set of methods, a single-cell context is temporarily created during the processing of cells. This is achieved by physical means within nanotiter plates [26] or by microfluidic encapsulation [27] so that transcripts that originate from the same cell are physically linked into a single amplicon. Subsequently, the compartments are broken up again and the amplicons are subjected to bulk sequencing. The advantage of these technologies is their high throughput, which allows processing of several million cells. While these techniques provide pairing information even on non-expanded cells, it is not possible to perform consensus-based error correction on such data as it cannot be determined whether the reads originate from the same cell or from several clonally related and SHM-diversified cells. In addition, they require microfluidic devices, the construction of which been described in detail [27] but which are not yet commercially available.

Complete Separation

These methods are advanced implementations of classical single-cell setups optimized for higher throughput. Like the classical protocols, they create the single-cell context early (typically by flow cytometry) and maintain it throughout the experiment but allow multiplexed sequencing on NGS platforms. This is achieved by tagging each individual amplicon with a DNA barcode, which encodes the physical location of a single cell within the set of assay plates [28–30]. Due to this mapping to an individual well/cell, these methods allow the building of consensus sequences from reads mapped to the same location, which greatly enhances the sequence quality. As the classical single-cell setups, these protocols allow linkage of additional data (e.g., flow cytometry) to an individual cell. In addition, they preserve clonal material (both cDNA and PCR amplicons), resulting in significant cost reductions for downstream cloning applications compared with gene synthesis.

Single-Cell Transcriptomics

This holistic approach will yield a full transcriptional snapshot of the cell, which as a byproduct also provides the antigen–receptor sequences [31]. While it facilitates extremely high-dimensional profiling of the analyzed cell, the cost is substantially higher than for Ig-specific techniques, thus limiting sample sizes to several hundred cells. Depending on the further development of sequencing costs, this approach might become an attractive alternative in the future.

Computational

As in many other fields dealing with NGS data, computational approaches have become critical for the analysis of Ig sequence data [32]. The main focus of work in this area has been on four closely related problems of genetic inference and on structural prediction, as discussed below. Issues of quality control and preprocessing of reads have recently been reviewed [33].

Segment Usage and V(D)J Joint Annotation

BLAST-related algorithms are the most widely used methods for inferring utilized V and J segments [34]. However, their alignment-based approach performs suboptimally in annotation of the V(D)J joint due to its short D segments, non-templated nucleotides, and potential SHMs. In addition, BLAST algorithms cannot quantify the likelihood of alternative outcomes. To overcome these problems, various tools like iHMMune, SoDA2, partis, and repgenHMM have been developed that use probabilistic approaches and thereby attempt to find better algorithmic representations of the biological processes [35–38].

Clusters and Phylogeny

Clustering of clonally related cells and inference of their phylogeny are often performed in combination and aim to identify cells that are derived from the same common ancestor. Basic clustering procedures are mainly implemented as grouping algorithms using V and J segment identity, CDR3 length, and CDR3 **Hamming distance** as grouping criteria. While these procedures are fast and computationally inexpensive, they can be either too permissive or too stringent, resulting in false or missing cluster members. By contrast, phylogenetic inference methods are computationally expensive but can utilize the additional information of shared hypermutated sites. Therefore, multiple tools use a combination of the two approaches to quickly restrict the data set to potentially interesting groups and then perform the inference procedure for each individual group [39]. Finally, it should further be noted that several assumptions of classical phylogenetics (e.g., substitution rates, no mixing of ancestral and extant species) are not necessarily met by the processes in the germinal center. Whether and to what extent this affects the inference process remains to be demonstrated.

SHM and Selective Pressure

The SHMs observed in memory populations are the result of activation-induced cytidine deaminase (AID) in combination with error-prone repair mechanisms and fixation due to proliferation. Here, SHM models taking the sequence context of a mutation into account have been successful in making reliable predictions under a neutral model (i.e., the absence of selection pressure), so that substantial deviations between model and data can be attributed to selective pressure [5,40].

Common Ancestors

To understand antibody maturation pathways, inference of both the unmutated common ancestor (UCA) and the most recent common ancestor (MRCA) is of substantial relevance, as the original germline configuration can often no longer be sampled from a donor. The crucial problem of UCA prediction is inference of the germline configuration of *IGH* CDR3. This requires a combination of the tools described above, as it needs to model the likelihood of a given unmutated CDR3 and the overlying SHM process and match it with the inferred phylogenetic tree. By contrast, the MRCA can be computed by overlaying the UCA with all mutations that are shared among the observed sequences. Determining the MRCA can be useful in experimental settings that need to distinguish between memory recall and *de novo* recruitment of naïve cells into the response.

Antibody Structure and Binding Properties

Structural predictions of antibodies are predominantly made using RosettaAntibody [41], which uses a hybrid approach, crafting all regions with the exception of CDR3, which is simulated. It currently does not allow the simulation of antibody–antigen interaction. While structural features responsible for a common binding behavior have been identified [42], it should be stressed that such studies still require measurements of actual antibodies and that full *in silico* prediction remains elusive.

Insights

The use of Ig repertoire analysis has become especially widespread in human immunology. Here, vertical study designs face various technical and legal restrictions that limit the use of exogenous labels to track cells. In addition, they are often limited to peripheral blood as the sample material. Repertoire analysis offers a unique opportunity for cellular tracking by using Ig sequences as molecular barcodes. It can thus provide unprecedented insight by assessing not only the size and phenotype of B cell populations but also their complete clonal composition over time. In addition, properties like isotype and hypermutation patterns can be directly derived from the obtained sequences. Due to these features, repertoire analysis has facilitated a plethora of investigations ranging from studies of general developmental processes to more applied questions like aging or response to infection [20,43–46]. As the topic of repertoire analysis in the context of vaccination has recently received an excellent in-depth review [47], we focus here on insights concerning basic immunology gained from such studies. It is often assumed that it should be straightforward to detect responses against the same antigen shared between individuals. This so-called 'public repertoire' has been well studied with haptens in mice and is typically characterized by a single dominant 'signature' clonotype. However, whether this concept can be readily transferred to responses in outbred human populations against non-haptenated antigens is less clear.

Here several findings of recent Adaptive Immune Receptor Repertoire (AIRR) studies are noteworthy. First, various studies have now indicated that segment usage distributions are stable within an individual and conserved between the naïve and memory compartments. While usage distributions also show a general conservation between non-related individuals, the conservation is more pronounced in monozygotic twins [48,49]. Second, there are now several studies showing that the primary immune response is not as selective as typically assumed in the classical clonal selection model (i.e., not only B cells with high affinity will be recruited into the germinal center) [50,51,52]. Together this suggests that clonotypes that have been observed to produce high-affinity responses in individual A but are rare in the naïve repertoire of individual B are not likely to become detectable in B, as other clonotypes with higher initial frequencies will prevail. While this does not invalidate the usefulness of hapten studies in investigating GC reactions, the structural solutions in these models are likely to be more restricted than plain protein antigens. Therefore a 'signature' might not be readily detectable between individuals unless clonal responses are normalized against the baseline repertoire of a donor (see Outstanding Questions). Importantly, this baseline is not limited to mature naïve cells but should also include the memory populations, as these have been observed to be recruited into the response even under conditions of primary immunization with a novel antigen [50]. This is supported by other reports that the adaptive immune system is not stateless with regard to unknown antigens, as previous exposure to non-related antigens can alter responses in the long term [53]. Furthermore, repertoire comparisons between individuals should always critically evaluate sampling depth and the expected entropy in the data to avoid false-positive random matches due to the phenomenon of the '**birthday effect**'. A 'public response' should be implied if and only if the same clonotype recognizes the same epitope and this shared reactivity has been experimentally confirmed. Such an interpretation has to be compared with the abovementioned random overlaps of frequently used segments in antibodies targeting non-related epitopes. Finally, the experimental design should carefully control for erroneous assignments in multiplex setups, as this could result in false-positive repertoire overlaps between samples [54].

Open Challenges

The Unknowns of Individual Germline Diversity

The prevalent approach to the identification of unmutated V, D, and J segments is the alignment of the query sequence against a reference germline database (GLDB). Thus, the quality of the utilized GLDB is critical to data evaluation, including downstream analysis steps like clustering that depend on the identified gene segments. For an individual species, the ideal GLDB should be both complete and accurate; that is, contain all existing segments and only existing segments. Building such a GLDB for outbred populations like humans is complicated for multiple reasons. First, sequence data from large consortia (e.g., HGP, 1kGP) have to be handled with caution as they are often generated using EBV-transformed lymphoblasts as the gDNA source. As these cell lines are typically derived from B cells, one or both of the *IGH* alleles will be rearranged in all or 60% of the lines, respectively. In addition, memory B cells can bring hypermutated sequences into the sequence pool. Second, the size and highly repetitive nature of Ig loci can constitute a problem for assemblies based on short-read technologies (Box 1); therefore, the global architecture of the locus is rarely captured. However, this information is crucial for building haplotypes and distinguishing between paralogs and orthologs of gene segments. Third, while donor-specific bacterial artificial chromosomes (BACs) have been successfully used to address questions of *IGH* locus variability [55], such efforts are currently not feasible on a larger scale. Two recent developments should be noted here. On the one hand, long-read sequencing technologies like PacBio (Box 1) are now commercially available and could provide the read length required for locus assemblies that do not require BAC cloning. On the other hand, it has been shown that bulk sequencing data can be used for direct inference of germline alleles as well as for the assignment of allelic variants to haplotypes [5,56,57]. In combination these approaches could potentially help to address this complex issue over the next years.

Data Sharing and Secondary Analyses

As both computational tools and reference databases undergo rapid development, it will become increasingly difficult to compare studies based on the data analysis performed in the original report. While it is unrealistic to expect that a single uniform mode of analysis could ever meet all of the diverse requirements of primary analysis, there is a huge potential for studies performing identical secondary analyses on many published datasets. This will require open access to the data and sufficient documentation of the experimental parameters. This task has recently been started to be approached by the AIRR Community (http://www.airr-community.org).

The Effector Repertoire

Currently, AIRR studies are predominantly focused on assessing the repertoire present in the cellular components of the adaptive immune system. This often disregards the fact that the main effects of B cells are mediated by soluble antibodies in the various bodily fluids. While some cell populations, like plasma cells, can easily be predicted to secrete antibodies, the situation is more complex for most other antigen-experienced populations. For example, recent evidence indicates that populations previously assumed to be anergic contain secretory transcripts [16]. It is therefore interesting to see that there are attempts to correlate sequencing information with serum antibody titers [58,59]. The main limitation here is that it is currently technically challenging to perform mass spectrometry on antibody derived peptides without nucleotide reference sequences and that the respective ASCs might not be present in the peripheral blood. Even if this problem could be solved, the lack of linkage between the individual peptides complicates the interpretation, as both the intra- (e.g., CDR2 to CDR3) and inter- (i.e., heavy:light chain pairing) chain relations remain unknown.

Concluding Remarks

Given the large diagnostic potential of repertoire studies, future technical development is expected to be significantly influenced by the type of additional information that can be used to leverage the plain sequencing data (e.g., cell type classification). Here microfluidic devices will gain prominence if they succeed in providing sampling depth comparable with the current methods. Single-cell approaches are likely to capitalize on their ability to provide high-dimensional, in-depth information on smaller populations of interest (e.g., antigen-binding cells). As the field moves towards maturity on a technical level, data interpretation and comparison will

Outstanding Questions

Are peripheral blood mononuclear cells (PBMCs) a reliable proxy of all immunological processes? Most studies in human subjects use PBMCs, which offer a simple, low-invasive, and repeatable mode of sampling. However, many steps of B cell activation occur in secondary lymphoid organs, not in the peripheral circulation. It therefore remains to be determined whether and how a given immunological process manifests in PBMCs and how stable these signatures are.

How to compare the repertoires of different donors? The repertoire is shaped by multiple components (e.g., heredity, historic exposure, current exposure), so how can the noise in interindividual comparisons be reduced? Does this require normalization against the mature naïve B cell compartment?

Can antibody reactivity be predicted from sequence data? Although NGS offers unparalleled throughput it does not provide any affinity data and current recombinant expression techniques do not (yet) deliver the throughput required for large-scale screening of antibodies. While *in silico* models are often presented as an alternative, they are computationally expensive themselves but might be up for the task in the near future.

move to the forefront again. To this end the establishment of standards for data annotation, deposition, and sharing will be critical. Importantly, these standards will not only have to address technical and biological aspects but will also have to deal with issues of intellectual property as well as the privacy of personal health information.

References

1. Honjo, T. *et al.* (1974) Organization of immunoglobulin genes: reiteration frequency of the mouse kappa chain constant region gene. *Proc. Natl. Acad. Sci. U. S. A.* 71, 3659–3663
2. Georgiou, G. *et al.* (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* 32, 158–168
3. Weigert, M. and Riblet, R. (1977) Genetic control of antibody variable regions. *Cold Spring Harb. Symp. Quant. Biol.* 41, 837–846
4. Behlke, M.A. *et al.* (1985) T-cell receptor beta-chain expression: dependence on relatively few variable region genes. *Science* 229, 566–570
5. Elhanati, Y. *et al.* (2015) Inferring processes underlying B-cell repertoire diversity. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* Published online July 20, 2015. http://dx.doi.org/10.1098/rstb.2014.0243
6. Kidd, M.J. *et al.* (2016) DJ pairing during VDJ recombination shows positional biases that vary among individuals with differing IGHD locus immunogenotypes. *J. Immunol.* 196, 1158–1164
7. Bolland, D.J. *et al.* (2016) Two mutually exclusive local chromatin states drive efficient V(D)J recombination. *Cell Rep.* 15, 2475–2487
8. Weinstein, J.A. *et al.* (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science* 324, 807–810
9. Khan, T.A. *et al.* (2016) Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci. Adv.* 2, e1501371
10. Lavinder, J.J. *et al.* (2014) Systematic characterization and comparative analysis of the rabbit immunoglobulin repertoire. *PLoS One* 9, e101322
11. Brezinschek, H.P. *et al.* (1995) Analysis of the heavy chain repertoire of human peripheral B cells using single-cell polymerase chain reaction. *J. Immunol.* 155, 190–202
12. Kantor, A.B. *et al.* (1997) An unbiased analysis of V_H-D-J_H sequences from B-1a, B-1b, and conventional B cells. *J. Immunol.* 158, 1175–1186
13. Wardemann, H. *et al.* (2003) Predominant autoantibody production by early human B cell precursors. *Science* 301, 1374–1377
14. Smith, K. *et al.* (2009) Rapid generation of fully human monoclonal antibodies specific to a vaccinating antigen. *Nat. Protoc.* 4, 372–384
15. Tiller, T. *et al.* (2009) Cloning and expression of murine Ig genes from single B cells. *J. Immunol. Methods* 350, 183–193
16. Muellenbeck, M.F. *et al.* (2013) Atypical and classical memory B cells produce *Plasmodium falciparum* neutralizing antibodies. *J. Exp. Med.* 210, 389–399
17. Scheid, J.F. *et al.* (2011) Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science* 333, 1633–1637
18. Murugan, R. *et al.* (2015) Direct high-throughput amplification and sequencing of immunoglobulin genes from single human B cells. *Eur. J. Immunol.* 45, 2698–2700
19. Kivioja, T. *et al.* (2011) Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* 9, 72–74
20. Vollmers, C. *et al.* (2013) Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 110, 13463–13468
21. Turchaninova, M.A. *et al.* (2016) High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nat. Protoc.* 11, 1599–1616
22. Robins, H.S. *et al.* (2009) Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. *Blood* 114, 4099–4107
23. Carlson, C.S. *et al.* (2013) Using synthetic templates to design an unbiased multiplex PCR assay. *Nat. Commun.* 4, 2680
24. DeWitt, W.S. *et al.* (2016) A public database of memory and naive B-cell receptor sequences. *PLoS One* 11, e0160853
25. Howie, B. *et al.* (2015) High-throughput pairing of T cell receptor α and β sequences. *Sci. Transl. Med.* 7, 301ra131
26. DeKosky, B.J. *et al.* (2013) High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.* 31, 166–169
27. McDaniel, J.R. *et al.* (2016) Ultra-high-throughput sequencing of the immune receptor repertoire from millions of lymphocytes. *Nat. Protoc.* 11, 429–442
28. Busse, C.E. *et al.* (2014) Single-cell based high-throughput sequencing of full-length immunoglobulin heavy and light chain genes. *Eur. J. Immunol.* 44, 597–603
29. Lu, D.R. *et al.* (2014) Identifying functional anti-*Staphylococcus aureus* antibodies by sequencing antibody repertoires of patient plasmablasts. *Clin. Immunol.* 152, 77–89
30. Han, A. *et al.* (2014) Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nat. Biotechnol.* 32, 684–692
31. Stubbington, M.J.T. *et al.* (2016) T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* 13, 329–332
32. Greiff, V. *et al.* (2015) Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends Immunol.* 36, 738–749
33. Yaari, G. and Kleinstein, S.H. (2015) Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med.* 7, 121
34. Ye, J. *et al.* (2013) IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* 41, W34–W40
35. Gaëta, B.A. *et al.* (2007) iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics* 23, 1580–1587
36. Munshaw, S. and Kepler, T.B. (2010) SoDA2: a hidden Markov model approach for identification of immunoglobulin rearrangements. *Bioinformatics* 26, 867–872
37. Ralph, D.K. and Matsen, F.A. (2016) Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation. *PLoS Comput. Biol.* 12, e1004409
38. Elhanati, Y. *et al.* (2016) repgenHMM: a dynamic programming tool to infer the rules of immune receptor generation from sequence data. *Bioinformatics* 32, 1943–1951
39. Briney, B. *et al.* (2016) Clonify: unseeded antibody lineage assignment from next-generation sequencing data. *Sci. Rep.* 6, 23901
40. Gupta, N.T. *et al.* (2015) Change-o: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* 31, 3356–3358
41. Sircar, A. *et al.* (2009) RosettaAntibody: antibody variable region homology modeling server. *Nucleic Acids Res.* 37, W474–W479
42. Laffy, J.M.J. *et al.* (2016) Promiscuous antibodies characterised by their physico-chemical properties: from sequence to structure and back. *Prog. Biophys. Mol. Biol.* Published online September 14, 2016. http://dx.doi.org/10.1016/j.pbiomolbio.2016.09.002
43. Horns, F. *et al.* (2016) Lineage tracing of human B cells reveals the *in vivo* landscape of human antibody class switching. *Elife* 5, e16578
44. Bagnara, D. *et al.* (2015) A reassessment of IgM memory subsets in humans. *J. Immunol.* 195, 3716–3724
45. Martin, V. *et al.* (2015) Ageing of the B-cell repertoire. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* Published online September 5, 2015. http://dx.doi.org/10.1098/rstb.2014.0237

46. Tsioris, K. *et al.* (2015) Neutralizing antibodies against West Nile virus identified directly from human B cells by single-cell analysis and next generation sequencing. *Integr. Biol.* 7, 1587–1597

47. Galson, J.D. *et al.* (2014) Studying the antibody repertoire after vaccination: practical applications. *Trends Immunol.* 35, 319–331

48. Laserson, U. *et al.* (2014) High-resolution antibody dynamics of vaccine-induced immune responses. *Proc. Natl. Acad. Sci. U. S. A.* 111, 4928–4933

49. Rubelt, F. *et al.* (2016) Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. *Nat. Commun.* 7, 11112

50. Galson, J.D. *et al.* (2016) B-cell repertoire dynamics after sequential hepatitis B vaccination and evidence for cross-reactive B-cell activation. *Genome Med.* 8, 68

51. Kuraoka, M. *et al.* (2016) Complex antigens drive permissive clonal selection in germinal centers. *Immunity* 44, 542–552

52. Tas, J.M. *et al.* (2016) Visualizing antibody affinity maturation in germinal centers. *Science* 351, 1048–1054

53. Andrews, S.F. *et al.* (2015) Immune history profoundly affects broadly protective B cell responses to influenza. *Sci. Transl. Med.* 7, 316ra192

54. Sinha, R. *et al.* (2017) Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. *bioRxiv* Published online April 9, 2017. http://dx.doi.org/10.1101/125724

55. Watson, C.T. *et al.* (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

56. Gadala-Maria, D. *et al.* (2015) Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc. Natl. Acad. Sci. U. S. A.* 112, E862–E870

57. Corcoran, M.M. *et al.* (2016) Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat. Commun.* 7, 13642

58. Wine, Y. *et al.* (2013) Molecular deconvolution of the monoclonal antibodies that comprise the polyclonal serum response. *Proc. Natl. Acad. Sci. U. S. A.* 110, 2993–2998

59. Wine, Y. *et al.* (2015) Serology in the 21st century: the molecular-level analysis of the serum antibody repertoire. *Curr. Opin. Immunol.* 35, 89–97

60. He, L. *et al.* (2014) Toward a more accurate view of human B-cell repertoire by next-generation sequencing, unbiased repertoire capture and single-molecule barcoding. *Sci. Rep.* 4, 6778

61..Pallarès, N. *et al.* (1999) The human immunoglobulin heavy variable genes. *Exp. Clin. Immunogenet.* 16, 36–60.62

62. Ruiz, M. *et al.* (1999) The human immunoglobulin heavy diversity (IGHD) and joining (IGHJ) segments. *Exp. Clin. Immunogenet.* 16, 173–184

63. Barbié, V. and Lefranc, M.P. (1998) The human immunoglobulin kappa variable (IGKV) genes and joining (IGKJ) segments. *Exp. Clin. Immunogenet.* 15, 171–183

64. Pallarès, N. *et al.* (1998) The human immunoglobulin lambda variable (IGLV) genes and joining (IGLJ) segments. *Exp. Clin. Immunogenet.* 15, 8–18

65. Johnston, C.M. *et al.* (2006) Complete sequence assembly and characterization of the C57BL/6 mouse Ig heavy chain V region. *J. Immunol.* 176, 4221–4234

66. Brekke, K.M. and Garrard, W.T. (2004) Assembly and analysis of the mouse immunoglobulin kappa gene sequence. *Immunogenetics* 56, 490–505

67. Sanchez, P. *et al.* (1991) V lambda–J lambda rearrangements are restricted within a V-J-C recombination unit in the mouse. *Eur. J. Immunol.* 21, 907–911