CellPress

Review

Advanced Methodologies in High-Throughput Sequencing of Immune Repertoires

Simon Friedensohn,[1] Tarik A. Khan,[2] and Sai T. Reddy[1,*]

In recent years, major efforts have been made to develop sophisticated experimental and bioinformatic workflows for sequencing adaptive immune repertoires. The immunological insight gained has been applied to fields as varied as lymphocyte biology, immunodiagnostics, vaccines, cancer immunotherapy, and antibody engineering. In this review, we provide a detailed overview of these advanced methodologies, focusing specifically on strategies to reduce sequencing errors and bias and to achieve high-throughput pairing of variable regions (e.g., heavy-light or alpha-beta chains). In addition, we highlight recent technologies for single-cell transcriptome sequencing that can be integrated with immune repertoires. Finally, we provide a perspective on advanced immune repertoire sequencing and its ability to impact basic immunology, biopharmaceutical drug discovery and development, and cancer immunotherapy.

Trends

Benchmarking of error correction methods will become common due to available spike-in data sets.

High-throughput integration of immune repertoires with transcriptomes will become the next important milestone in repertoire sequencing.

Increases in commercially available single-cell sequencing and receptor-pairing solutions will bring this technology to less-specialized laboratories.

The Rapid Rise of High-Throughput Immune Repertoire Sequencing

The rapid rise of high-throughput **immune repertoire sequencing** (see Glossary) has led to unprecedented quantitative insight into lymphocyte diversity (Box 1) and adaptive immunity, leading to a new era of systems immunology. Here, we refer to this paradigm for T cells and B cells as T cell receptor (TCR)-seq and immunoglobulin (Ig)-seq, respectively [1,2]. In one of the first examples using TCR-seq, Robins *et al.* quantitatively measured human TCR diversity [3], while subsequent studies revealed that the clonal overlap of repertoires shared between individuals was higher than anticipated based on theoretical calculations [4–6]. These seminal studies not only provided quantitative and, in some cases, unexpected answers to longstanding questions in basic lymphocyte immunobiology, but also established the basic experimental and computational methods needed for high-throughput repertoire sequencing. Moreover, they laid the foundation for a breadth of follow-up studies, including the implementation of TCR-seq and Ig-seq for applied research.

In clinical settings, TCR-seq has been performed on patients with acute T lymphoblastic leukemia before and after treatment, revealing the receptor sequence and frequency of potential neoplastic lymphoblasts [7]. This approach was further developed as a clinical assay for the diagnosis of minimal residual disease (MRD), which describes the subset of malignant lymphocytes that remain during and after the course of treatment. TCR-seq provided equal or higher sensitivity to conventional tests (flow cytometry or PCR based) [3,7–9]. For biotechnology applications, Ig-seq has become a valuable tool for monoclonal antibody (mAb) discovery and engineering. For example, Ig-seq was performed on plasma cells of immunized mice to identify antigen-specific mAbs based on sequencing information alone, thus bypassing cost- and time-intensive screening procedures [10]. Ig-seq on human B cells has been used to identify variants of broadly neutralizing antibodies against HIV-1 [11,12]. Another innovation has been to

[1] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
[2] Pharmaceutical Development & Supplies Biologics Europe, F. Hoffman-La Roche Ltd, Basel, Switzerland

*Correspondence: sai.reddy@ethz.ch (S.T. Reddy).

http://dx.doi.org/10.1016/j.tibtech.2016.09.010

Box 1. Generation of Immune Receptor Repertoire Diversity

Lymphocyte T cell receptors (TCR), B cell receptors (BCR), and immunoglobulins (Ig)/antibodies (the secreted form of BCRs) encode a vast functional sequence space. This diversity is confined to receptor variable regions, which facilitate the recognition and binding of the myriad pathogens and antigens encountered by a host. Besides carrying these variable regions, the BCR and TCR are structurally related. The TCR comprises two protein subunits linked by disulfide bonds (the α and the β chain in most cases, although some TCRs comprise the so-called γ/δ chains), both of which carry the aforementioned variable region as well as a constant region. Likewise, antibodies (both the membrane-bound version and the secreted form) comprise two chains, the light and the heavy chain. However, an antibody is formed by two identical heavy and two identical light chains, again linked by disulfide bonds, giving rise to its final 'Y-shaped' form. The formation of variable regions is a result of the somatic recombination, called V(D)J recombination, of distinct germline gene elements: variable (V), diversity (D), and joining (J) segments. For each segment, multiple different variants exist in the respective genomic loci. The variable regions of heavy and β chains are formed by recombining a V, D, and J segment, whereas the light and α chains only utilize a V and a J segment. In humans, these recombinations can generate a theoretical diversity of at least 5×10^{13} naïve BCRs and 10^{18} α:β TCRs [76]. However, only a subset of this diversity is expected to be physiologically present in the actual repertoire of an individual [77,78]. In addition to V(D)J recombination, B cells also undergo secondary diversification through somatic hypermutation, a highly regulated enzymatic process that introduces additional mutations in variable regions to increase the affinity of BCRs to antigen targets. In the past, empirically measuring the enormous diversity of lymphocyte repertoires was relegated to qualitative approaches, such as complementarity determining region 3 (CDR3) spectratyping or Sanger sequencing of small subsets of clones (i.e., ~10–100).

use **shotgun proteomics** workflows to obtain peptide mass spectra data of affinity-purified, polyclonal antibody proteins, whereby the sequences of these peptides are identified using a reference database generated by Ig-seq [13–15]. This approach enables a serological and molecular analysis of the antibody proteome, revealing information, such as the presence of post-translational modifications, which is not possible with Ig-seq alone [16]. In vaccine profiling, preliminary studies have shown that Ig-seq can be used to determine the dynamics and developmental lineage of B cells following seasonal influenza vaccination [17,18]. The studies described above demonstrate the gaps within applied immunology that are being overcome using immune repertoire sequencing.

While rapid progress in immune repertoire sequencing has already led to several important advances, there are still several challenges that must be overcome before these techniques fully realize their potential. Here, we review recent innovations in TCR-seq and Ig-seq that aim to solve these main challenges.

The first problem is the presence of errors and bias in sequencing data sets, which are introduced through library preparation methods [e.g., **Multiplex-PCR** or **5′ Rapid Amplification of cDNA Ends** (5′ RACE)] and next-generation sequencing (NGS) platforms themselves [19,20]. While artifacts in data sets are a concern for any NGS-based application, they are particularly troublesome for immune repertoires, given that their diversity and distribution directly impact immunological interpretations. A second limitation arises from the fact that many basic repertoire sequencing protocols rely on using bulk cell populations as input, which does not preserve information about **variable region pairing**. Given that variable region pairing is often essential for determining antigen specificity, this also limits the functional conclusions that can be made with TCR-seq and Ig-seq. To acquire true and comprehensive immune repertoire knowledge, the field must adopt error and/or bias correction methods, variable region pairing, and potentially further relevant information [e.g., (epi)genomic and transcriptomic data].

Improving Accuracy in Immune Repertoire Data

Sequencing errors and biases are generated in several ways [21]. In some cases, they can be mitigated by good laboratory practices, such as preventing degradation or contamination of the source material. However, there are inherent problems due to the underlying dependence of molecular biology-based techniques for NGS. These problems arise because the amount of starting material (genomic DNA or total RNA) is often limited and of low purity, thus requiring

Glossary

5′ Rapid amplification of cDNA ends (5′ RACE): a library preparation protocol designed to capture mRNA sequences when either the 5′ end is unknown or the usage of a multiplex forward primer set is not desirable. In a typical protocol, mRNA is reverse transcribed to cDNA using an antisense gene-specific primer. After cDNA synthesis, homopolymeric tails are added to the 3′ end of the transcript via terminal deoxynucleotidyl transferase (TdT). The resulting amplicon can now be amplified using primers targeting a gene specific region and the newly added tail. Modern protocols (SMART-RACE) use template switching, which adds the 3′ tail directly during the RT-step.

Complementary determining region 3 (CDR3): the CDRs are the main determinants of antigen binding. Each variable region carries a set of three CDRs. CDR3 is the junctional site of V(D)J recombination and, thus, is the most diverse region, and is often used to distinguish different T and B cell clones.

Consensus sequence: the sequence across all residues that is the most similar sequence with respect to a set of sequences. In general, this sequence is found by selecting the most frequent nucleotide or amino acid in each column of a multiple sequence alignment.

Droplet emulsions: a liquid in which the droplet of another liquid is dispersed. Both liquids are normally immiscible. For most single-cell applications, the dispersed phase is water or an aqueous solution, whereas the continuous phase consists of an organic liquid, such as mineral oil.

Hamming and Levenshtein distances: both describe the distances between sequence strings. The Hamming distance is used for strings of equal length and describes the numbers of substitutions or errors that can transform one string into another. The Levenshtein distance, also known as edit distance, is a generalization of the Hamming distance for strings of unequal length. It describes the number of insertions, deletions and substitutions that are needed to transform one string into another.

Box 2. Next-Generation Sequencing Platforms Used for Immune Repertoire Sequencing

Several NGS systems have been used for immune repertoire sequencing; these include instruments from IonTorrent [39], Pacific Biosciences [60,61], Roche/454 Life Sciences [3], and Illumina [27,46,49]. Similar to the fields of genomics and transcriptomics, Illumina-based sequencing has become the most widely used technology in immune repertoire sequencing [2]. Each of these NGS platforms vary in the amount and type of errors that are added during base calling. For example, while the MiSeq instrument of Illumina is reported to have a relatively low error rate, a recent study indicated that the percentage of error-free reads of a 150 bp, paired-end sequencing run was close to only 80% [36]. The number of erroneous reads is likely to be even greater for immune repertoire data sets, because longer sequencing runs are necessary to sequence full-length variable regions (~350–420 bp). Although other instruments show higher error rates than MiSeq, their usage might be preferable under certain circumstances. Longer read lengths (>1 kb) generated by Pacific Biosciences technology is useful for receptor pairing protocols, which physically link variable regions together [59,61], while the fast sequencing turnover achieved with semiconductor technology developed by IonTorrent is potentially interesting for clinical applications [36]. The recent improvements in nanopore sequencing could obtain ultra-long reads (>100 kb), which may facilitate researchers to screen the complete naïve or rearranged immune loci more efficiently [79]. This development could lead to the discovery of novel immune receptor diversification mechanisms, such as recently reported chromosomal integrations into variable regions [80].

reverse transcriptases and/or DNA polymerases that are unable to perfectly transcribe or amplify the original material. While the use of high-fidelity enzymes (such as newly engineered proofreading reverse transcriptases [22]) and careful experimental design can partially reduce error rates, nucleotide misincorporation will still occur and propagate due to the exponential nature of PCR [23,24] and further errors can be added through the sequencing process itself (Box 2). Additionally, in many cases, multiplex PCR is conducted, where degenerate primer sets target either the beginning of the variable region or the leader region (further upstream of each V-segment) [25]. This process introduces considerable bias towards preferentially amplified sequences [26]. As a means to circumvent these steps, several protocols have been adapted for TCR-seq and Ig-seq based on template-switching reactions that allow a single forward primer to be used (e.g., 5′ RACE) [27–30]. However, a recent study on amplification efficiency of TCR variable regions demonstrated that even when using a single primer and 5′ RACE, preferential and biased amplification still occurred [19], most likely due to the nature of the template-switching reaction.

The accumulation of errors in repertoire sequencing leads to an artificial inflation of observed clones or variable regions. Prudent filtering of sequencing reads based on quality scores is a viable strategy to remove some false-positive reads and improve diversity estimates [31]. If **paired-end sequencing** data are generated (from Illumina sequencing), merging both reads based on their overlap increases the quality of the overlapping region [32], which usually occurs in the clonal-identifying **complementarity determining region 3** (CDR3) of variable regions. Removing scarce clones, which only have few reads mapping to them, has been suggested as a generic strategy to remove falsely identified sequences [4]. Besides these general guidelines, several more sophisticated methods for correcting for substitution errors and, to a lesser extent, also for amplification bias have been applied to TCR-seq and Ig-seq (Table 1) (for more details, please refer to a recent review [33]).

Error Correction via Replicate Sequencing

One approach for error correction relies on using technical replicates (e.g., sequencing libraries independently prepared from the same source material), after which sequencing results obtained from each replicate are then validated against each other [21]. Greiff *et al.* used this procedure to establish a reliability detection cutoff for excluding erroneous reads in Ig-seq data [25,34]. Ig-seq reads were ranked by abundance and compared across replicate data sets; only the set of clones (defined by their CDR3) that were present in the respective replicate data sets at a specified fraction (e.g., 95%) were retained. This filtering procedure reduced the number of uniquely detected clones by 10- to 40-fold, while still retaining approximately 95% of all sequencing reads. Similar numbers have been reported by Warren *et al.* [4]. However, when

Immune repertoire sequencing: lymphocyte receptor variable regions are sequenced with modern high-throughput methods as opposed to older, low-throughput methods, such as Sanger-sequencing. Both RNA and DNA can be used as input material for different (sequencing) library preparations.

Multiplex PCR: instead of separating several PCRs into single tubes, a multiplex PCR is carried out as one reaction with all of the amplification primers for each individual reaction present. In immune repertoire sequencing, this principle is applied to amplify all of the possible variable regions generated by somatic recombination. Depending on the library preparation protocol, a combination of forward primers targeting either leader region or the variable region and a set of reverse primers targeting the end of the variable region or constant regions are utilized.

Shotgun proteomics: an analytical workflow in proteomics. Mixtures of proteins are first digested and denatured. The resulting fragments are then analyzed using liquid chromatography coupled to mass spectrometry (LC-MS), and the mass profiles at each chromatography peak are compared with a known database to identify the peptide fragments and, ultimately, the proteins contained in the initial mixture.

Split-pool synthesis: a concept borrowed from combinatorial chemistry and used in several new high-throughput single-cell sequencing technologies to create many uniquely barcoded microparticles that can be used to tag the genome or transcriptome of single cells. In this process, the particles are split in several pools and different nucleotides (or oligonucleotides) are covalently attached to the particle in each pool. Afterwards, the particles are pooled and split again and new (oligo) nucleotides are attached. This process is repeated multiple times to ensure uniquely barcoded particles.

Variable region pairing: TCRs and BCRs (antibodies) comprise two types of polypeptide chain, which are the variable-$\alpha$and -β chain for TCRs (or $\gamma\delta$chains in a subset of T cells and the variable-light (V_L) and –heavy (V_H) for BCRs. During bulk immune repertoire sequencing, this pairing information is lost because the chains are expressed from separate genomic loci.

Table 1. Methods for Correcting Immune Repertoire Sequencing Data

Method	Notes	Error correction	Bias correction	Refs
MiXCR	Corrects sequences via clustering of the whole variable region; is primarily designed as an alignment tool and works with both T and B cell sequences; can be used if no other method of error correction is possible or when correcting already-sequenced libraries	Yes	No	[40]
IMSEQ	Corrects sequences via clustering of CDR3s; is primarily designed as an alignment tool and works with both T and B cell sequences; can be used if no other method of error correction is possible or when correcting already sequenced libraries	Yes	No	[37]
IgRepertoireConstructor	Corrects sequences via clustering of the whole variable region; annotates V- and J-genes; only works with B cell sequences; can be used if no other method of error correction is possible or when correcting already sequenced libraries	Yes	No	[38]
Replicate sequencing	Can be used with any wet-lab protocol and analysis platform; since replicate sequencing requires no changes, it is an easy and fast method of removing errors; however, rare variants might not be observed across libraries	Yes	No	[25,34]
Unique molecular identifiers (UIDs)	Require a specialised wet-lab protocol and analysis methods; studies that require robust results should utilize UIDs, because they correct for error and bias (in the case of multiplexed or 5′ RACE-based library preparations methods); UIDs have become the highest standard in immune repertoire sequencing and other fields	Yes	Yes	[27,42–44,49,50,70]

applying replicate filtering for the entire variable [V(D)J] region, considerably fewer reads were retained (around 50%), highlighting that errors accumulate across the whole sequencing amplicon. These errors are especially problematic in Ig-seq because they cannot be distinguished from naturally occurring and biologically relevant mutations introduced by somatic hypermutation. While replicate sequencing helps to identify reliable full-length variable regions, the loss of up to 50% of the data and the need to at least double the sequencing depth via replicates increase the costs of this approach. Rare sequences that are not shared across replicates will also not be found. Furthermore, the existence of reproducible hotspot PCR and sequencing errors is now well known, suggesting that certain erroneous sequences derived from highly abundant clones are present across replicates [27,35,36].

Error Correction Through Sequence Clustering

The clustering of highly similar sequences has been another approach to error correction, which has the advantage of being able to use nearly all the reads generated by an NGS run. In immune repertoire sequencing, clustering is often performed by grouping together sequences based on their CDR3 [or even complete V(D)J sequence] using a distance metric such as the **Hamming or Levenshtein distance** [3,6,37,38]. To avoid grouping together highly homologous, yet distinct sequences, some algorithms also separate CDR3s into their constituent V-, D-, and J-gene segments. Clones are then not only separated by one distance metric across the whole CDR3, but also by several metrics, one for each of the junctional segments [39]. The actual sequence or clone representing each cluster is generally identified by building the **consensus sequence** of all sequences in a given cluster. While effective for TCR-seq, clustering and correcting errors solely based on CDR3 are less effective for Ig-seq data sets generated from antigen-experienced B cells. This is because somatic hypermutations lead to the generation of multiple clones that share the same CDR3 sequence but differ at various other mutational hotspots across the rest of their V(D)J region. Thus, correcting sequences based on CDR3 alone will result in overcorrection by removing unique somatic variants and reducing diversity. MiXCR, a recently developed immune repertoire annotation software, performs heuristic error correction of variable regions by multilayer clustering of highly similar sequences [40], a method that tries to model the underlying error processes experienced by a set of immune repertoire sequences. If the user-specific criteria are chosen wisely and the underlying nature of the data set allows for it, the algorithm in MiXCR is able to distinguish between true somatic variants and erroneous variants in Ig-seq. However, identifying the correct parameters is challenging and likely varies depending on the input material. It may be valuable for future studies to compare the analysis results from MiXCR to other modes of error correction, such as replicate sequencing or advanced experimental-bioinformatic workflows (see below).

Error and Bias Correction with Unique Molecular Identifiers

One of the most powerful error-correction strategies developed is based on unique molecular identifiers (UIDs) [41,42]. Unlike the above *in silico* methods, UID-based correction requires additional experimental library preparation steps, but has the distinct advantage of offering capabilities to correct both errors and amplification bias. All UID methods rely on the single-molecule tagging of RNA or DNA molecules with a degenerate nucleotide sequence (e.g., NNNNN), which forms the basis of the UID. For immune repertoire sequencing, a common approach is to add these degenerate sequences as overhangs next to a gene-specific primer, which are then used for reverse transcription (RT) of mRNA, resulting in first-strand cDNA molecules that are UID tagged [43]. Alternatively, UIDs can be incorporated in cDNA during the template-switching 5′ RACE protocol [27,44]. Following UID tagging, cDNA is used as the input for PCR amplification, and library preparation can proceed as per usual. After NGS is performed, sequencing reads are grouped together based on their UIDs. Given that single-molecule tagging was performed, it is assumed that reads sharing the same UID originated from the same mRNA molecule. Errors can then be corrected by simply building a consensus sequence from UID groups. Another advantage is that counting UIDs instead of raw reads also leads to an improved measurement of transcript abundances (i.e., less bias) [45]. One disadvantage is that incorporating UIDs increases the size of the amplicon, resulting in additional technological hurdles related to the read length. For example, while the Illumina MiSeq platform is able to generate 2 × 300 bp read lengths, the addition of a UID may lead to challenges in obtaining high-quality reads across the whole V(D)J region because read quality drops towards the end of each read. One way to circumvent this problem relies on adding UIDs to both ends of the amplicon. The subsequent use of Tn5 transposase breaks longer sequences into a mix of sequences that are anchored at either the V-segment or the constant region. Nontagmented amplicons can link both fragments via their UIDs [46]. After sequencing each library variant, the complete read is assembled bioinformatically by building the consensus sequence out of all of the sequences that carry matching UIDs. However, it has been observed that UID-based consensus building is not able to remove all errors, most likely due to mutations that occur early during PCR [27]. Shugay *et al.* developed a highly effective secondary error-correction step based on read gain/loss filtering, in which clones are completely dropped from the analysis if after consensus building, these clones have lost reads mapping to their original, uncorrected clonal sequence [27].

An additional consideration is that UIDs should be designed to be diverse enough to prevent early saturation, which would lead to tagging different RNA or DNA molecules with the same barcode, a problem often described as the birthday paradox [47,48]. This design can be achieved either by adding sufficiently long, degenerate regions on the reverse primer [27,49], by performing a second-strand cDNA synthesis (which introduces an additional UID

[50]), and/or by also setting a similarity threshold for UID groups [48]. However, increasing the sequence space dedicated to the actual UID also proportionally increases the probability that UIDs will accumulate technical errors themselves [51]. Precise experimental control is also needed, such as limiting material inputs to ensure sufficient oversampling of UIDs in the NGS data. Additionally, the observed error across UIDs is even more dependent on the number amplification cycles during library preparation. To counteract this problem, Khan *et al.* recently developed a protocol known as molecular amplification fingerprinting (MAF), which tags single-molecule IgG cDNA with reverse UIDs and also adds forward UIDs via the forward primer set (specific for variable framework 1 regions) during multiplex PCR [49]. MAF uses both UIDs to enable error correction on the UID regions themselves before consensus building on the variable regions. Importantly, MAF also uses the ratio of forward to reverse UIDs to generate a normalization constant, which, when applied to Ig-seq data, resulted in nearly absolute correction of the bias introduced by multiplex PCR. This study also validated error and bias correction using synthetic spike-ins, providing a clear quantification of the degree of error and bias pre- and post-MAF bioinformatics processing. Synthetic spike-ins are becoming common tools for validation across various NGS methods [52,53] and it is recommended that they also become a standard in immune repertoire sequencing studies.

High-Throughput Pairing of Variable Regions

Initial studies in immune repertoire sequencing relied on bulk populations of lymphocytes. However, this approach is unable to recover the pairing of variable regions because they are expressed as unique transcripts from separate chromosomes. However, having additional functional information about variable region pairing is desirable because it can be used for a variety of applications, such as mAb discovery, vaccine profiling, and analysis of tumor-infiltrating lymphocytes. This kind of information has led to the development of several sophisticated methods that are able to achieve high-throughput sequencing of a paired repertoire. These pairing methods utilize the physical linkage of variable regions before sequencing or use a digital linkage that associates sequences via bioinformatic analysis (Figure 1, Key Figure). Additionally, when synthetic libraries are used (e.g., phage display), a physical linkage between chains typically exists in the display platform. In these cases, depending on the read length/coverage specifications, the CDR3s of both chains may be captured in a single PCR amplicon providing paired variable region data.

Primer Matrix Pairing

One of the initial steps in generating paired repertoire sequencing data is to devise a strategy that is able to generate libraries from genomic DNA or RNA of single cells. A straightforward way to achieve this is through flow-cytometric sorting of single cells into microwell plates, followed by single-cell library preparation [54–56] (Figure 1A). To increase throughput, each well can be used in conjunction with a specifically barcoded primer set [56] or, alternatively, in conjunction with a primer matrix, which utilizes row and column-specific forward and reverse index primers [54,55]. All of the wells from multiple plates are pooled together and, following sequencing, variable region pairing is identified by their barcode or index sequence. An advantage of this methodology is that sequencing is coupled to the flow cytometry data, thus providing greater resolution of the underlying phenotype of lymphocyte repertoires. So far, this technology has been able to process up to 4.6 $\times$ 10^4 cells with an efficiency of approximately 60% [54,55]. More recent technologies based on emulsion droplets have pushed the limit of cells that can be processed to even higher numbers (up to 10^7 cells, see below). However, these gains in throughput have so far come at the expense of collecting flow cytometry data.

Bioinformatic Pairing

A simple way to extract pairing information bioinformatically is to link variable regions based on their relative frequency, assuming that naturally paired chains are found in a similar order when

Key Figure

Overview of Immune Repertoire Pairing and Sequencing Methods

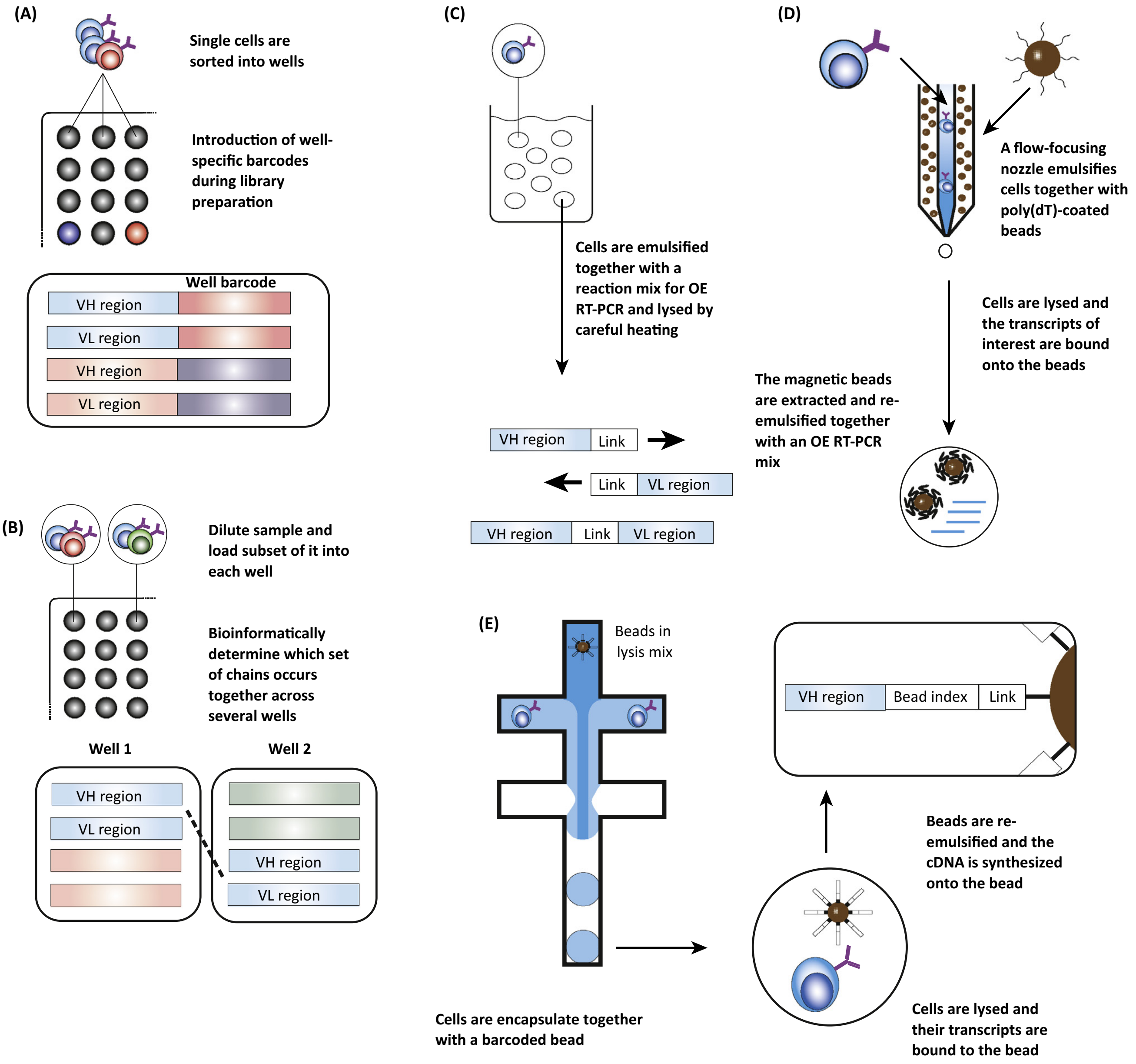
(A)
Single cells are sorted into wells
Introduction of well-specific barcodes during library preparation
Well barcode
VH region
VL region
VH region
VL region
(B)
Dilute sample and load subset of it into each well
Bioinformatically determine which set of chains occurs together across several wells
Well 1
Well 2
VH region
VL region
VH region
VL region
(C)
Cells are emulsified together with a reaction mix for OE RT-PCR and lysed by careful heating
VH region
Link
Link
VL region
VH region
Link
VL region
(D)
A flow-focusing nozzle emulsifies cells together with poly(dT)-coated beads
Cells are lysed and the transcripts of interest are bound onto the beads
The magnetic beads are extracted and re-emulsified together with an OE RT-PCR mix
(E)
Beads in lysis mix
VH region
Bead index
Link
Beads are re-emulsified and the cDNA is synthesized onto the bead
Cells are lysed and their transcripts are bound to the bead
Cells are encapsulate together with a barcoded bead
Trends in Biotechnology

Figure 1. The common core of many pairing technologies is to first separate the bulk cell population into single cells and then to pair the sequences of interest [depicted here are variable heavy (V_H) and light (V_L) regions of the BCR] either via physical or digital linkage. (A) One straightforward approach is to sort single cells into wells on a well plate and introduce well-specific barcodes during library preparation (which is carried out within each of the wells). The resulting sequencing reads can then be linked via this unique molecular identifier (UID) [54]. (B) Instead of sorting single cells into each well, aliquots of a larger cell population can be distributed across the plate, and well-specific barcodes can be introduced. Receptor pairing is then performed bioinformatically by determining which set of sequences occurs together across several wells [58]. (C) Another way to separate single cells is to disperse them within water-in-oil emulsion (OE) droplets and then lyse the cells within each droplet by heat shock. The resulting sequences are then linked physically by overlap-extension reverse transcription (RT)-PCR [59]. (D) A more sophisticated approach divides the cell lysis and the RT-PCR step into different compartments. Cells are first emulsified together with poly(dT)-beads and lysis buffer. Transcripts are captured by the beads, which are

(Figure legend continued on the bottom of the next page.)

ranked by frequency. However, immune repertoires tend to follow power-law distributions, which implies that most cells in a repertoire at a given time comprise a few dominant clones, whereas most clones are present at low frequencies [57]. Given this phenomena, the above-mentioned pairing approach is generally only feasible for a small subset of highly abundant clones [10]. A novel approach, which includes both experimental and bioinformatic elements, has recently been described for high-throughput TCR pairing [58]. In this method, small subsets of T cells are separated into microwells, from which sequencing libraries indexed by wells are then bulk prepared, similar to the above-mentioned protocols (Figure 1B). After TCR-seq, a list with all possible V∝ and Vβ combinations within each well is generated. Assuming that the initial cell dilution across all wells was sufficiently high enough, cross-referencing all combinatorial indices elucidates pairings that only occur in combination in the same wells (i.e., providing V∝ and Vβ pairing). Compared with single-cell sequencing, the number of wells from which sequencing libraries have to be prepared is substantially reduced.

Emulsion-Based Pairing

Droplet emulsions are a proven method for capturing and performing RT followed by PCR reactions on single cells. Therefore, using emulsions for library preparation in immune repertoire sequencing offers a strategy for high-throughput variable region pairing. A major advantage of emulsions is that they reduce the overall reaction volume and offer straightforward scalability. A simple and cost-efficient approach has been established by Turchaninova *et al.*; this protocol used water-in-oil droplets to encapsulate T cells, whereby lysis and library preparation were both performed within the droplet [59]. Together with the cell, each droplet also contains a reaction mixture suitable for the RT and PCR steps. The cells are lysed by heat shock, and the variable regions are linked by overlap-extension RT-PCR (Figure 1C). A challenge of this methodology is that the heat-shock step may reduce the efficiency of the RT enzyme and the subsequent cDNA synthesis. One way to overcome this problem is to use a methodology that separates the lysis and RT steps, as described by DeKosky *et al.* [60]. In this protocol, B cells are randomly distributed into high-density microtiter chips (60 000 wells/plate), comprising picoliter wells patterned across a PDMS slide. Additionally, each well is loaded with beads covered with oligo (dT)-primers, which base-pair with the 3′ poly(A) tail of (and, thus, specifically enrich for) mRNA. After the final addition of a lysis buffer, the chip is sealed off. Following lysis, the beads are recovered in batch and then encapsulated into single-bead emulsion droplets (Figure 1D). Within each emulsion, cDNA is synthesized and the variable regions are physically linked by an overlap-extension PCR step. This pairing approach was used to identify, with high fidelity, mAbs from the plasmablast compartment of a recently vaccinated human subject [60]. However, the initial encapsulation of cells in the high-density microtiter chip limits the number of cells that can be screened. In a further improvement, a flow-focusing device was used to first encapsulate cells and magnetic poly(dT) beads in a primary emulsion droplet [61]. Cells were then lysed within the droplets and transcripts were again bound to the poly(dT) beads, which were then extracted via magnetic force. Afterwards, the beads were re-emulsified and underwent the same process as in the earlier version of this approach. Given that the updated version does not rely on PDMS slides, the number of cells that can be sequenced increases substantially, up to 10^6–10^7 cells [62].

Combining Immune Repertoires with Single-Cell Transcriptomics

Excitingly, several studies already highlight the expansion of the tools that comprise advanced repertoire sequencing. In one example by Han *et al.*, a primer matrix approach for single-cell TCR-seq was combined with targeted RNA-seq of T cells [63]. In another example, a newly

re-emulsified in a second step. Afterwards, the sequences of interest are again linked by RT-PCR [61]. (E) Sequencing by droplet microfluidics (Drop-Seq) utilizes digital linking of not only receptor sequences, but also entire transcriptomes. Here, cells and beads linked to poly(dT) primers (or gene specific primers), which carry a bead-specific UID, are encapsulated together in a mix of lysis buffer and RT-reaction buffer. Cells are lysed, and their transcriptome is subsequently reversed transcribed onto the uniquely barcoded bead. Sequencing libraries, which incorporate each barcode, are then built directly from each bead [67].

Table 2. A Brief Overview of Single-Cell Sequencing Technologies

Application	Technology	Company	Sequencing targets	Primary purpose	Refs
Immune repertoire sequencing technologies	Plate based	Single Cell Technology	B cell receptors (BCRs)	Antibody discovery	[71]
	Droplet-based microfluidics	Juno (formerly AbVitro)	BCRs + TCRs	CAR T cell therapy	[72]
	Plate-based and/or combinatorial pairing	Adaptive Biotechnologies	BCRs + TCRs	Clinical diagnostics	[58]
	Droplet-based microfluidics	Atreca	BCRs + TCRs	Antibody discovery	[73]
General single-cell sequencing technologies	Continuous-flow microfluidics	Fluidigm	Transcriptome, genome, and/or epigenome	Provision of technology to end user	[74]
	Plate based	Cellular Research/BD Biosciences	Transcriptome	Provision of technology to end user	[65]
	Droplet-based microfluidics	10X Genomics	Transcriptome	Provision of technology to end user	[75]

developed computational tool, known as TraCeR, was designed to assemble transcripts obtained from single-cell RNA-seq experiments and reconstruct the complete transcriptome of single T cells, also including their TCR variable region sequences [64]. The importance of these methods will only increase as the price and throughput of commercially available single-cell technologies continues to fall (Table 2). Fan *et al.* recently established a technology termed 'CytoSeq', which was able to perform high-throughput gene expression profiling on single cells [65]. CytoSeq couples microscale plates (patterned with up to 100 000 wells) with magnetic beads that are covalently bound to oligo(dT) primers that carry a particular UID for each bead. Such uniquely barcoded beads are prepared by **split-pool synthesis**, a concept borrowed from combinatorial chemistry. These microbeads are then separately encapsulated with single cells in each well, while also extracting the transcriptome cells based on the bead UID. So far, CytoSeq has enabled the simultaneous interrogation of transcriptomes from up to 4×10^5 single cells [65]. Two recent studies established a technology that combines droplet microfluidics with transcriptome sequencing (Drop-seq) [66,67] (Figure 1E). Similar to CytoSeq, Drop-seq uses specifically barcoded microspheres to group sequencing reads coming from the same cell together. To date, these approaches have yet to be applied to immune repertoire analysis, but we anticipate this integration in the future, as the throughput and overall wealth of information provided will lead to substantially greater insights.

Concluding Remarks and Future Directions

The advanced methods in immune repertoire sequencing described above demonstrate the capability to obtain accurate (error and bias-corrected) and **variable region paired** data. These capabilities have set a new standard for the field of immune repertoire sequencing. While the technologies established for error and bias correction were largely focused on *ex vivo* biological repertoires, similar approaches may be adapted to synthetic repertoires (e.g., phage, bacteria, and yeast display [68,69]). Bias is less of a concern in synthetic libraries, since they typically use a single forward and reverse primer. However, the application of error correction (e.g., the use of UIDs) is still relevant to synthetic repertoires. Using error-corrected data for synthetic libraries reliably tracks clonal enrichment, accurately identifies conserved versus nonconserved amino acids, and decreases the likelihood of overestimating library diversity. Now that the faults and

Outstanding Questions

Experiments comparing the reviewed technologies are still missing: how do these methods compare with each other in terms of efficiency, error rates, and reproducibility? Can spike-in data sets be used to benchmark different methods?

To what extent is variable region pairing stochastic? Are there preferential germline partners for pairing?

Do linked V_H and V_L chains exhibit the same or similar levels of somatic hypermutation?

Is it possible to predict the functionality of a receptor based on transcriptomic signatures?

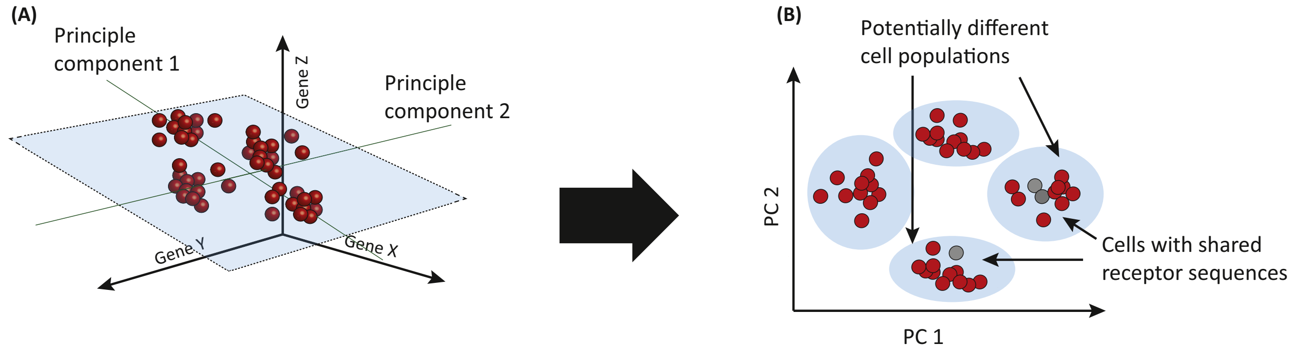

Figure 2. Example Highlighting Linking Information about Specific Immune Receptors and Cellular Transcriptional Profiles. Each single cell exhibits unique levels of transcriptional activity (A). In reality, one sequencing run can measure the expression level of many thousand genes. Dimensionality reduction techniques such as principle component analysis or t-SNE can aid the visual inspection of these data sets. Here, principal component analysis orthogonally projects the data points onto a lower dimension, which minimizes the (squared) distances between the original points and their projection (B). This projection reveals cells of similar transcriptional activity and subsets of potentially different cell populations, while the information about the immune receptor of each cell (here pictured in gray) illuminates the developmental lineages of lymphocytes across these compartments.

limitations of TCR- and Ig-seq are well established, researchers can ensure high-quality results and interpretations using these methods. However, gaps still remain (see Outstanding Questions) and, to gain further insight into lymphocyte function, it may be beneficial to link immune repertoires with transcriptional or epigenomic data [64] (Figure 2). This high-dimensional information may lead to the ability to predict TCR or antibody specificity and functionality based on transcriptional profiles and refined phenotypic markers.

As more advanced methods continue to be established, increased reliability will lead to widespread usage of immune repertoire sequencing for a variety of healthcare applications. These applications are already being commercialized for characterizing tumor-infiltrating lymphocytes, tracking of residual disease in blood-borne cancer, and mAb engineering. The ability to interrogate millions of sequences enables not only the selection of mAbs based on binding, but also selection of candidates that are more developable (e.g. removing candidates with undesirable T cell epitopes or chemical degradation hotspots). Overall, it is expected that advanced immune repertoire sequencing will lead to more efficient and faster development of immunological therapeutics and diagnostics.

References

1. Benichou, J. *et al.* (2012) Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* 135, 183–191
2. Georgiou, G. *et al.* (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* 32, 158–168
3. Robins, H.S. *et al.* (2009) Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* 114, 4099–4107
4. Warren, R.L. *et al.* (2011) Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res.* 21, 790–797
5. Robins, H.S. *et al.* (2010) Overlap and effective size of the human CD8+ T cell receptor repertoire. *Sci. Transl. Med.* 2, 47ra64
6. Weinstein, J.A. *et al.* (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science* 324, 807–810
7. Wu, D. *et al.* (2012) High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. *Sci. Transl. Med.* 4, 134ra63
8. Kurtz, D.M. *et al.* (2015) Noninvasive monitoring of diffuse large B-cell lymphoma by immunoglobulin high-throughput sequencing. *Blood* 125, 3679–3687
9. Faham, M. *et al.* (2012) Deep-sequencing approach for minimal residual disease detection in acute lymphoblastic leukemia. *Blood* 120, 5173–5180
10. Reddy, S.T. *et al.* (2010) Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat. Biotechnol.* 28, 965–969
11. Zhu, J. *et al.* (2013) Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proc. Natl. Acad. Sci. U.S.A.* 110, 6470–6475
12. Zhu, J. *et al.* (2013) De novo identification of VRC01 class HIV-1-neutralizing antibodies by next-generation sequencing of B-cell transcripts. *Proc. Natl. Acad. Sci. U.S.A.* 110, E4088–E4097
13. Cheung, W.C. *et al.* (2012) A proteomics approach for the identification and cloning of monoclonal antibodies from serum. *Nat. Biotechnol.* 30, 447–452

14. Lavinder, J.J. *et al.* (2014) Identification and characterization of the constituent human serum antibodies elicited by vaccination. *Proc. Natl. Acad. Sci. U.S.A.* 111, 2259–2264
15. 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
16. 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
17. Jiang, N. *et al.* (2013) Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci. Transl. Med.* 5, 171ra19
18. Laserson, U. *et al.* (2014) High-resolution antibody dynamics of vaccine-induced immune responses. *Proc. Natl. Acad. Sci. U.S.A.* 111, 4928–4933
19. Best, K. *et al.* (2015) Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. *Sci. Rep.* 5, 14629
20. Glanville, J. *et al.* (2015) Deep sequencing in library selection projects: what insight does it bring? *Curr. Opin. Struct. Biol.* 33, 146–160
21. Robasky, K. *et al.* (2014) The role of replicates for error mitigation in next-generation sequencing. *Nat. Rev. Genet.* 15, 56–62
22. Ellefson, J.W. *et al.* (2016) Synthetic evolutionary origin of a proofreading reverse transcriptase. *Science* 352, 1590–1593
23. Malboeuf, C.M. *et al.* (2001) Thermal effects on reverse transcription: improvement of accuracy and processivity in cDNA synthesis. *Biotechniques* 30, 1074–1078
24. Eckert, K.A. and Kunkel, T.A. (1991) DNA polymerase fidelity and the polymerase chain reaction. *PCR Methods Appl.* 1, 17–24
25. Menzel, U. *et al.* (2014) Comprehensive evaluation and optimization of amplicon library preparation methods for high-throughput antibody sequencing. *PLoS ONE* 9, e96727
26. Carlson, C.S. *et al.* (2013) Using synthetic templates to design an unbiased multiplex PCR assay. *Nat. Commun.* 4, 2680
27. Shugay, M. *et al.* (2014) Towards error-free profiling of immune repertoires. *Nat. Methods* 11, 653–655
28. Quigley, M.F. *et al.* (2011) Unbiased molecular analysis of T cell receptor expression using template-switch anchored RT-PCR. *Curr. Protoc. Immunol.* 10, 10.33
29. Li, S. *et al.* (2013) IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nat. Commun.* 4, 2333
30. Wang, B. *et al.* (2015) Facile discovery of a diverse panel of anti-ebola virus antibodies by immune repertoire mining. *Sci. Rep.* 5, 13926
31. Bokulich, N.A. *et al.* (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* 10, 57–59
32. Schirmer, M. *et al.* (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* 43, e37
33. Greiff, V. *et al.* (2015) Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends Immunol.* 36, 738–749
34. Greiff, V. *et al.* (2014) Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. *BMC Immunol.* 15, 40
35. Brodin, J. *et al.* (2015) Challenges with using primer IDs to improve accuracy of next generation sequencing. *PLoS ONE* 10, e0119123
36. Quail, M.A. *et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13, 341
37. Kuchenbecker, L. *et al.* (2015) IMSEQ–a fast and error aware approach to immunogenetic sequence analysis. *Bioinformatics* 31, 2963–2971
38. Safonova, Y. *et al.* (2015) IgRepertoireConstructor: a novel algorithm for antibody repertoire construction and immunoproteogenomics analysis. *Bioinformatics* 31, i53–i61
39. Bolotin, D.A. *et al.* (2012) Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. *Eur. J. Immunol.* 42, 3073–3083
40. Bolotin, D.A. *et al.* (2015) MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* 12, 380–381
41. Jabara, C.B. *et al.* (2011) Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc. Natl. Acad. Sci. U.S.A.* 108, 20166–20171
42. Kinde, I. *et al.* (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9530–9535
43. Stern, J.N. *et al.* (2014) B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci. Transl. Med.* 6, 248ra107
44. 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
45. Kivioja, T. *et al.* (2012) Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* 9, 72–74
46. Cole, C. *et al.* (2016) Highly accurate sequencing of full-length immune repertoire amplicons using Tn5-enabled and molecular identifier-guided amplicon assembly. *J. Immunol.* 196, 2902–2907
47. Sheward, D.J. *et al.* (2012) Degenerate primer IDs and the birthday problem. *Proc. Natl. Acad. Sci. U.S.A.* 109, E1330 author reply E1331
48. Egorov, E.S. *et al.* (2015) Quantitative profiling of immune repertoires for minor lymphocyte counts using unique molecular identifiers. *J. Immunol.* 194, 6155–6163
49. Khan, T.A. *et al.* (2016) Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci. Adv.* 2, e1501371
50. 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
51. Deakin, C.T. *et al.* (2014) Impact of next-generation sequencing error on analysis of barcoded plasmid libraries of known complexity and sequence. *Nucleic Acids Res.* 42, e129
52. Deveson, I.W. *et al.* (2016) Representing genetic variation with synthetic DNA standards. *Nat. Methods* 13, 784–791
53. Hardwick, S.A. *et al.* (2016) Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat. Methods* 13, 792–798
54. 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
55. 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
56. Tan, Y.C. *et al.* (2014) High-throughput sequencing of natively paired antibody chains provides evidence for original antigenic sin shaping the antibody response to influenza vaccination. *Clin. Immunol.* 151, 55–65
57. Greiff, V. *et al.* (2015) A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med.* 7, 49
58. Howie, B. *et al.* (2015) High-throughput pairing of T cell receptor alpha and beta sequences. *Sci. Transl. Med.* 7, 301ra131
59. Turchaninova, M.A. *et al.* (2013) Pairing of T-cell receptor chains via emulsion PCR. *Eur. J. Immunol.* 43, 2507–2515
60. DeKosky, B.J. *et al.* (2013) High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.* 31, 166–169
61. DeKosky, B.J. *et al.* (2015) In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat. Med.* 21, 86–91
62. McDaniel, J.R. *et al.* (2016) Ultra-high-throughput sequencing of the immune receptor repertoire from millions of lymphocytes. *Nat. Protoc.* 11, 429–442
63. Han, A. *et al.* (2014) Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nat. Biotechnol.* 32, 684–692
64. Stubbington, M.J. *et al.* (2016) T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* 13, 329–332

65. Fan, H.C. *et al.* (2015) Expression profiling. Combinatorial labeling of single cells for gene expression cytometry. *Science* 347, 1258367

66. Klein, A.M. *et al.* (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161, 1187–1201

67. Macosko, E.Z. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214

68. D'Angelo, S. *et al.* (2014) From deep sequencing to actual clones. *Protein Eng. Des. Sel.* 27, 301–307

69. Spiliotopoulos, A. *et al.* (2015) Sensitive recovery of recombinant antibody clones after their in silico identification within NGS datasets. *J. Immunol. Methods* 420, 50–55

70. Seifert, D. *et al.* (2016) A comprehensive analysis of primer IDs to study heterogeneous HIV-1 populations. *J. Mol. Biol.* 428, 238–250

71. Chen, C. *et al.* Single Cell Technology. Method of obtaining antibodies of interest and nucleotides encoding same, US9328172 B2

72. Vigneault, F. *et al.* Abvitro. Single cell bar-coding for antibody discovery, WO2014144495 A1

73. Tan, Y.C *et al.* Atreca. Analysis of nucleic acids associated with single cells using nucleic acid barcodes, WO2015103339 A1

74. Kolodziejczyk, A.A. *et al.* (2015) The technology and biology of single-cell RNA sequencing. *Mol. Cell* 58, 610–620

75. Zheng, G.X. *et al.* (2016) Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* 34, 303–311

76. Murphy, K.M. *et al.* (2016) *Janeway's Immunobiology.* (9th), Garland Science

77. Haas, R.J. *et al.* (1972) Cytokinetic studies on slowly-renewing bone marrow and spleen lymphocytes in rats during a primary and a secondary immune response. *Acta Haematol.* 48, 39–48

78. Westermann, J. and Pabst, R. (1992) Distribution of lymphocyte subsets and natural killer cells in the human body. *Clin. Investig.* 70, 539–544

79. Deamer, D. *et al.* (2016) Three decades of nanopore sequencing. *Nat. Biotechnol.* 34, 518–524

80. Tan, J. *et al.* (2016) A LAIR1 insertion generates broadly reactive antibodies against malaria variant antigens. *Nature* 529, 105–109