CellPress

Review

Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires

Victor Greiff,[1] Enkelejda Miho,[1] Ulrike Menzel,[1] and Sai T. Reddy[1,*]

High-throughput sequencing (HTS) of immune repertoires has enabled the quantitative analysis of adaptive immune responses and offers the potential to revolutionize research in lymphocyte biology, vaccine profiling, and monoclonal antibody engineering. Advances in sequencing technology coupled to an exponential decline in sequencing costs have fueled the recent overwhelming interest in immune repertoire sequencing. This, in turn, has sparked the development of numerous methods for bioinformatic and statistics-driven interpretation and visualization of immune repertoires. Here, we review the current literature on bioinformatic and statistical analysis of immune repertoire HTS data and discuss underlying assumptions, applicability, and scope. We further highlight important directions for future research, which could propel immune repertoire HTS to becoming a standard method for measuring adaptive immune responses.

Resolving the Complexity of Antigen Receptor Repertoires

B and T lymphocytes of the adaptive immune system have the ability to recognize foreign molecules via an immense array of antigen-binding receptors (B cell and T cell receptors, BCR/TCR) [1]. The diversity of lymphocyte repertoires (short: 'immune repertoires') is a result of genetic recombination and diversification mechanisms. Diversity is first created in the germline via recombination of variable V, (diversity D), and joining J gene segments [2], which form the antigen-binding variable region. Further diversification occurs through imprecise junction of these gene segments (addition of P- and N-nucleotides adjacent to the D segment), somatic hypermutation (SHM, in B cells), and combination of subunits (T cells) or heavy/light chains (B cells) [3].

Immune repertoire antigen-specificity and diversity is largely dominated by the junctional site of V(D)J recombination, which is known as the complementarity determining region 3 (CDR3) [4,5]. The CDR3 has thus served for a long time as a natural identifier of lymphocyte clonality: B and T cells with an identical CDR3 are classified as belonging to the same clone. Upon antigen challenge, B cells are activated and undergo clonal selection and expansion forming a clonal lineage [6].

Immune repertoires are an important target of immunological and clinical research because they harbor information on both past and ongoing immune responses [3]. HTS now enables the quantitative analysis of these highly diverse immune repertoires at an unprecedented depth [7–9] and has already shown tremendous promise for investigating immune repertoire changes during chronic viral infections (e.g., HIV-1) [10–12], autoimmune diseases [13–16], and with aging [17,18]. Continuous advances in sequencing technology have sparked the development of

Trends

High-throughput immune repertoire sequencing is becoming a core technology for advancing basic, applied, and clinical immunology.

Specialized bioinformatic and statistical methods for repertoire diversity and overlap analysis as well as for performing network and phylogenetic clustering enable the investigation of immune repertoire expansion, dynamics, architecture and evolution at an unprecedented level of detail.

There is a divergence of the underlying assumptions, applicability, and scope of bioinformatic and statistical methods, thus compromising the consistency of data analyses within and across studies that needs to (will) be addressed in the (near) future.

[1]Department of Biosystems Science and Engineering, Eidgenössische Technische Hochschule (ETH Zürich), Mattenstrasse 26, Basel 4058, Switzerland

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

 http://dx.doi.org/10.1016/j.it.2015.09.006
© 2015 Elsevier Ltd. All rights reserved.

CrossMark

numerous bioinformatic and statistical methods which aim to maximize information extraction from immune repertoire HTS data. We review the current literature on bioinformatic and statistical analysis of immune repertoires. Specifically, we discuss steps of the HTS (bioinformatic) workflow which can influence the biological conclusiveness of a study, such as representative repertoire sampling, data error correction, and sequence germline annotation, as well as statistical approaches for estimating diversity and visualization of repertoire selection, architecture, and evolution.

Sampling and Sequencing Depth: How Deep Is Deep Enough?

While HTS provides a tremendous amount of depth for analyzing immune repertoires, biologically meaningful information on repertoire diversity is substantially dependent on the comprehensive sampling of the cell population studied (**biological sampling**, Figure 1A, see Glossary, [19]) and on the comprehensive read coverage of DNA or RNA molecules encoding BCRs and TCRs (**technological sampling**, Figure 1A). The choices of the organism and cell population are key to achieving optimal repertoire coverage; in humans, the most common source for lymphocytes is the peripheral blood compartment, which contains only 2.5% ($\sim 10^9$ B or T cells) of the estimated total number of cells ($\approx 10^{11}$) [20–23]. In mice the coverage of immune repertoires is less problematic because all lymphoid organs are readily accessible, and the number of lymphocytes ($\sim 10^8$ B or T cells [24]) lies significantly closer to the current state of the art in sequencing depth (10^4 and 10^7 sequencing reads per sample [16,25]).

The consequences of insufficient biological sampling have been investigated previously by Warren and colleagues [26]: they showed that distinct 20 ml blood samples from the same individual captured only a portion of the TCR peripheral blood repertoire (biological undersampling). Furthermore, technological undersampling has been shown to compromise the detection of 'public' clones (clones shared across individuals), which are a common target in immune repertoire studies [27,28]. In fact, several studies indicated that there was a positive correlation between sequencing depth and the number of public clones detected [13,29,30]. Thus, the biological conclusiveness of a study benefits from the implementation of **biological replicates** (test for biological undersampling [26,31,32]) (Figure 1A) and **technical replicates** (test for technological undersampling [33–36]) (Figure 1A), which may be performed once for each cell population analyzed. It is important to note that biological undersampling can only be meaningfully addressed if sufficient technological sampling has been established [33]. Furthermore, **species accumulation and rarefaction analyses** may be performed to quantify the extent of (under)sampling [29,33,35,37] (Figure 1A).

Because the size of the cellular compartments studied differs widely by research question, universal guidelines for ensuring comprehensive sampling are challenging to implement. Nevertheless, two general rules to consider are: (i) the number of sequencing reads should at least exceed the clonal diversity of the sample if complete read coverage is unattainable, and (ii) the lower the frequency of a clone, the higher the sequencing depth must be for its accurate capture [34]. While knowing the exact clonal diversity of a lymphocyte population before HTS is not possible, basic knowledge of cell numbers and **clonal frequency distributions**, as well as mathematical modeling [34], facilitate the estimation of the required sequencing depth. For example, antigen-specific or clonally expanded populations (e.g., memory B and T cells, plasma cells) [34] will have a clone-to-cell ratio that is well below 1 [14,33,38–41], and thus less sequencing reads would be required to obtain a good snapshot of the clonal diversity. By contrast, clonal frequency distributions of naïve B and T cells have been shown to be more uniform [14,38–42] (i.e., higher clone-to-cell ratios than clonally expanded populations), requiring a substantially higher number of reads for clonal diversity description. In the future, as sequencing depth continues to increase, repertoire coverage may become more comprehensive.

Glossary

Biological replicates: HTS of different samples of the same underlying cell population [e.g., partitioning of PBMC (peripheral blood mononuclear cells)]. Biological replicates are used to assess biological sampling.

Biological sampling: the cell population sampled must be an approximate representation of the cellular compartment being investigated to allow meaningful conclusions to be drawn from the data.

Circos graph: a circular layout plot for the visualization of quantitative and qualitative relationships in complex and large datasets. In immune repertoire HTS data, it is used mainly to visualize frequencies of overlapping clones across time-resolved longitudinal data.

Clonal frequency distribution: for a given immune repertoire, the clonal frequency distribution describes the distribution of the number of sequencing reads (read abundance) that are allocated to each clonotype (commonly referred to as 'clone size'). The underlying power law of clonal frequency distributions is commonly visualized by plotting the logarithm of clonotype frequency as a function of the logarithm of clonal rank.

Clustering of sequences: clustering is the process of grouping a set of similar sequences (nt/aa sequences defined as strings of characters) in the same group based on a given sequence identity threshold. Hierarchical clustering is a connectivity algorithm that forms clusters of sequences based on their string distance.

Morisita–Horn overlap index: this is used to compare species (e.g., clone, germline genes) sequence and abundance overlap between any two immune repertoires. It is defined as the $MH = \frac{2\sum_{i=1}^{S} x_i y_i}{\sum_{i=1}^{S} x_i^2 + \sum_{i=1}^{S} y_i^2}$, where S is the number of unique species, and x and y denote the frequency of the ith species in either repertoire. The MH index ranges between 0 (no overlap) and 1 (complete species overlap and identical species frequencies).

Network: a network is a measurable pattern of relations among subunits. It represents a graph composed of a set of objects (vertices, nodes) and links (edges).

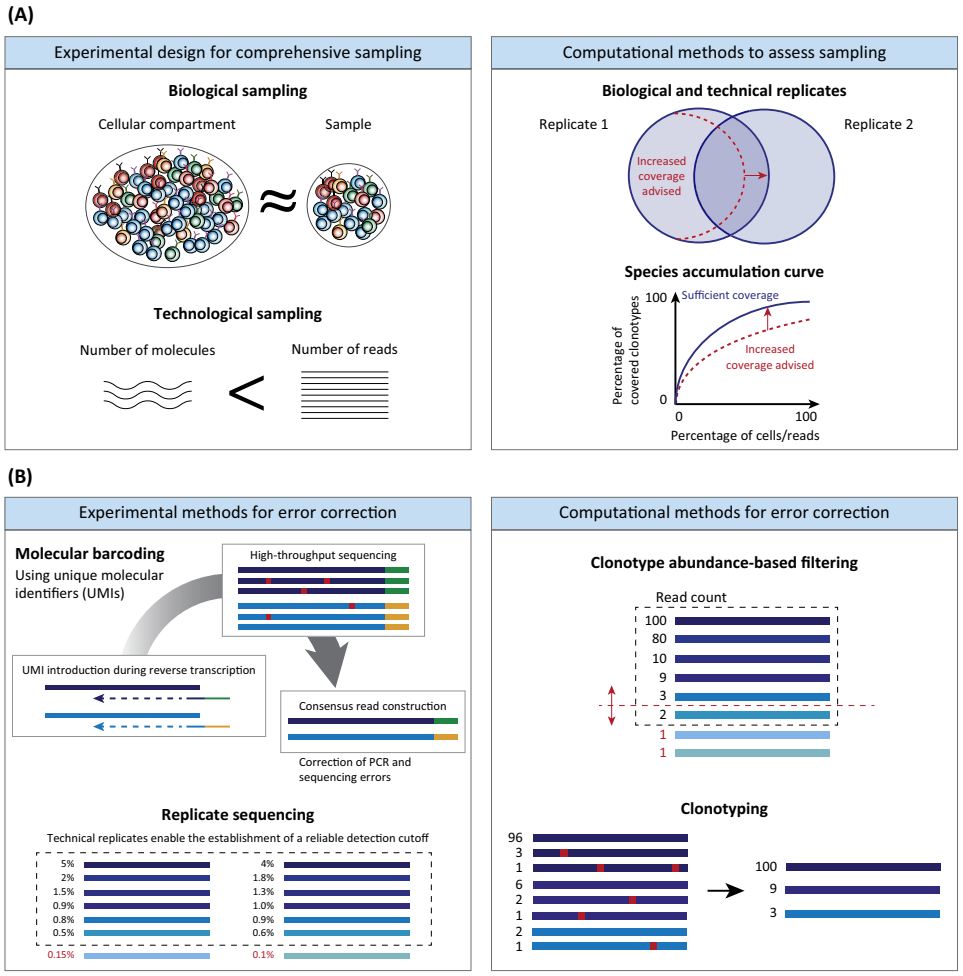

Figure 1. Considerations for the Experimental and Computational Design of an Immune Repertoire Study. (A) To comprehensively describe the population of interest, both biological and technological sampling deserve consideration. Reliable biological sampling ensures that the sampled population represents the diversity of the cellular compartment being investigated. For reliable HTS data it is equally important to calculate more reads than clones, or, if quantifiable, input molecules (DNA/RNA); this is referred to as technological sampling. The sufficiency of both biological and technological sampling can be assessed by clonal overlap analysis (e.g., Venn diagrams). Another typical means to assess sufficient sampling is by species accumulation/rarefaction curves; a curve that levels off indicates complete clonal coverage, and incomplete coverage is revealed by a curve with a positive slope (more cells/reads would reveal more unseen clones). (B) There exist different methods to account and correct for errors introduced by PCR and sequencing. Experimental methods comprise the addition of unique molecular identifiers (UMIs), which allow the construction of consensus reads. Replicate sequencing can be used to determine reliable detection cutoff. Errors can also be corrected computationally by heuristic abundance-based filtering of clones present with only a few reads or by clustering similar sequences based on a defined similarity threshold (clonotyping).

Bioinformatics Tools for Preprocessing of Immune Repertoire Data

Combining Experimental and Computational Approaches for Error Correction of Immune Repertoire HTS Data

Regardless of the sequencing platform, HTS has not yet reached the level of accuracy of Sanger sequencing because it suffers from errors introduced during library amplification (experimental) or sequencing (HTS, bridge amplification, platform-specific) [43]. Therefore, both experimental and computational strategies have been devised to attenuate the impact of errors on biological

Numbering schemes: complementarity determining regions and framework regions are identified as amino acid strings by different numbering schemes (i.e., IMGT, Kabat, Clothia). Numbering schemes define the start and ending positions of BCR and TCR regions.

Phred score (Q): a measure for quality base calling. It is defined as $Q = -10 \log_{10} P$, where P is the base-calling error probability. For example, if $Q = 30$ for a given base, the probability that the base was called incorrectly is $P = 10^{-3}$.

R package: R is a statistical programming environment, and its package system enables the flexible and constant addition of newly developed statistical approaches.

Reliable clonal detection cutoff: clones in datasets of technical replicates are ranked in decreasing order of frequency and tested for simultaneous presence in all replicates to construct a list of reliably detected clones, which together are at least for example 90% (cutoff) present in all replicates. The reliable detection cutoff is valid for all HTS datasets prepared with experimental conditions identical to those of the technical replicates. Importantly, the meaningful application of reliable detection cutoffs depends on (near)-complete sample coverage.

Species accumulation and rarefaction analysis: species accumulation curves display the rate at which new clones are discovered with increasing number of sequencing reads. By contrast, rarefaction curves are used to estimate the number of clones at a particular level of sampling.

String distance: measures dissimilarity between any two sequences [e.g., germline reference sequence and sequencing reads for V(D)J annotation or two CDR3s for clonotyping] by counting the minimum number of operations required to transform one string into the other. Levenshtein or edit distance accounts for insertions, deletions and substitutions.

Technical replicates: replicate sequencing of the same immune repertoire library. A strict definition would be the resequencing of the same library, whereas a more lenient definition would consider also molecular replicates (separate library preparation of the same genetic material) adequate provided that

conclusions. A shared principle of all error-correction approaches listed below is that they rely (either explicitly or implicitly) on high read numbers. Thus, high sequencing depth not only ensures comprehensive sampling but can also increase the accuracy of repertoire measurements.

It is a well-known statistical principle that a given entity converges to its true ('expected') value (law of large numbers) if sampled sufficiently often. This principle is leveraged by an error-correction approach that is based on unique molecular identifiers [UMIs, also referred to as unique identifiers (UIDs) or barcodes] (Figure 1B), which are degenerate nucleotide sequences of high diversity that uniquely tag each DNA or RNA molecule [31,44–46]. Leveraging dedicated bioinformatic pipelines [45,47], UMI methods in immune repertoire sequencing have been shown to achieve up to a 100-fold error reduction [31,45], thus considerably reducing artificial repertoire diversity [48]. However, a study by Shugay and colleagues indicated that increased RNA input (increasing from ng to μg) required a considerable increase in sequencing depth (10^6 to 10^7 sequencing reads) and a switch in sequencing platform (Illumina MiSeq to HiSeq) to ensure consensus read construction (presence of multiple sequencing reads with identical UMIs) [45]. Therefore, to effectively use UMI approaches for error correction, technological over-sampling is needed, which may not always be feasible for highly-diverse and large cell populations (i.e., naïve B and T cells). In addition to error correction, UMI methods have also been applied to the problem of pairing TCR α and TCR β chains [49], and BCR heavy and light chains [50].

Another approach to experimental error correction is the use of technical replicates which can be used to establish **reliable clonal detection cutoffs** [32,33,36] (Figure 1B). While these cutoffs exploit the multiplicity of reads per clone as detection confidence [26], it has been indicated that hotspot PCR or sequencing errors are reproducible across technical replicates [45]. In these situations, the only approach for error correction would be UMI-based [45]. However, it should be noted that UMI-based approaches can still benefit by using technical replicates to establish sensitivity thresholds of error correction [51,52].

There is a vast array of approaches that could be considered as computational error correction. The simplest would be filtering HTS datasets (before any V(D)J annotation, Table 1) for low-quality reads (e.g., **Phred score**) using dedicated software packages such as pRESTO or FastQC [47,53]. Subsequently, several studies have employed heuristic clonal abundance cutoffs (e.g., removal of clones with only 1–5 reads, Figure 1B) [31–33,40] to decrease artificial diversity. Warren and colleagues showed that abundance filtering is superior to strict quality filtering in decreasing artificial diversity [26]. Furthermore, Bolotin and colleagues demonstrated that aggressive quality filtering can even lead to loss of a significant portion of the data [54]. In fact, lower-quality reads may be recovered from paired-end sequencing [55] (the inherently lower-quality 3′ ends of sequencing reads gain in confidence via an overlapping region in both forward and reverse reads) or by merging lower-quality reads with reads of higher quality and identical or very similar clonal identifiers (clonotyping, see below and Figure 1B) [45,54,56,57]. This ensures error correction and maximal data preservation while removing artificial diversity. Overall, however, it should be noted that clear guidelines for quality and abundance filtering do not yet exist [58,59].

Clonotyping: Defining Clonality from Error-Prone High-Throughput Sequencing Data

The investigation of the complexity of lymphocyte clonality in health and disease represents the core purpose of the majority of analytical approaches (Figure 2). While the definition of clonality in a biological sense is widely accepted (all lymphocytes having the same BCR or TCR belong to the same clone, see above), its translation to HTS data is challenging owing to the influence of PCR and sequencing errors, and of SHM.

A common approach is to cluster sequencing reads with high CDR3 homology (measured by edit **string distance**) as well as identical CDR3 length and V/J gene usage, and to refer to these

biological replicates have been performed to exclude biological undersampling. Technical replicates are used to assess technological sampling.

Technological sampling: ensuring that the number of sequencing reads exceeds the molecular diversity, or at least, the clonal diversity of the underlying sample.

Unique molecular identifiers (UMIs): pseudo-random sequences of several degenerate nucleotides (usually 8–12), which are added during library preparation by reverse transcription or ligation. Sequencing reads with identical UMIs are merged (consensus read construction) thus increasing the confidence in each base call, and consequently reducing the extent of PCR and sequencing error.

Table 1. Characterization of the Main Annotation Platforms

	IMGT/ High-V-Quest [62]	IgBlast [123]	iHMMune-align [124]	MIGEC [45]	MIXCR [56]
Analysis of TCR and BCR data	TCR and BCR	BCR	BCR	TCR and BCR	TCR and BCR
Prediction of germline sequences	Yes	Yes	Yes	No	Yes
Extraction of FR/ CDR/constant region (CR)	FR, CDR	For V region only (until V-part of CDR3)	No	CDR3	FR/CDR/CR
SHM extraction	Yes (but V region only)	Yes (entire V(D)J region)	Yes (entire V(D)J region)	No	Yes (entire V(D)J region)
Reference numbering scheme	IMGT	IMGT/Kabat/ NCBI	UNSWIg	IMGT	IMGT
Max number of sequences per analysis	≤500 000	~1000 (online) Unrestricted (standalone)	~2 Mb (Online), Unrestricted (standalone)	Unrestricted	Unrestricted
Processing of unique molecular identifiers	No	No	No	Yes	No
Consideration of sequencing quality information (Phred scores)	No	No	No	Yes	Yes
Speed (standard dataset of 1×10^6 reads)	Days	Hours	Hours	Minutes	Minutes
Supported input format	FASTA	FASTA	FASTA	FASTQ	FASTA, FASTQ
Platform	Online	Online/stand-alone	Online/stand-alone	Stand-alone	Stand-alone

as 'clonotypes' [6,60]. Using publicly-available resources [45,56,61–64] and in-house developed software [65,66], **clustering** by CDR3 homology at the nucleotide level has been performed in the following ways: (i) inferring unmutated common ancestors [67,68]; (ii) absolute edit distance cutoffs in hierarchical clustering linkage trees [68], allowing a range of mismatches (one [54,69,70], three [71], or five [10]) in sequences within one clonotype; or (iii) clustering by using relative thresholds (90% [65,72], 95% [73], 97.25% [12], 100% [41]). At the amino acid level, clonotypes have been built based on 80–100% CDR3 homology [13,33,36,74,75].

Clonotyping reduces the influence of PCR and sequencing errors on clonal diversity estimations but also, in the case of B cells, serves to group clones that belong to the same clonal lineage. A robust clonotype definition is, therefore, a defining step in every immune repertoire HTS study because it has a large impact on biological conclusions drawn (especially in diversity analyses, Figure 2). Tipton *et al.* recently defined clonotypes by experimental validation as sequences with CDR3 (hamming) nucleotide identity of >85% using replicate sequencing [14].

Annotation of Immune Repertoire Data

Raw sequencing reads require annotation for downstream statistical analysis (Figure 2). Annotation tools vary widely with regard to the extent of output information, which can range from the sole identification of the CDR3 [45] to comprehensive information (e.g., germline gene usage, framework regions, CDRs, and the extent of SHM, Table 1). While annotation speeds differ

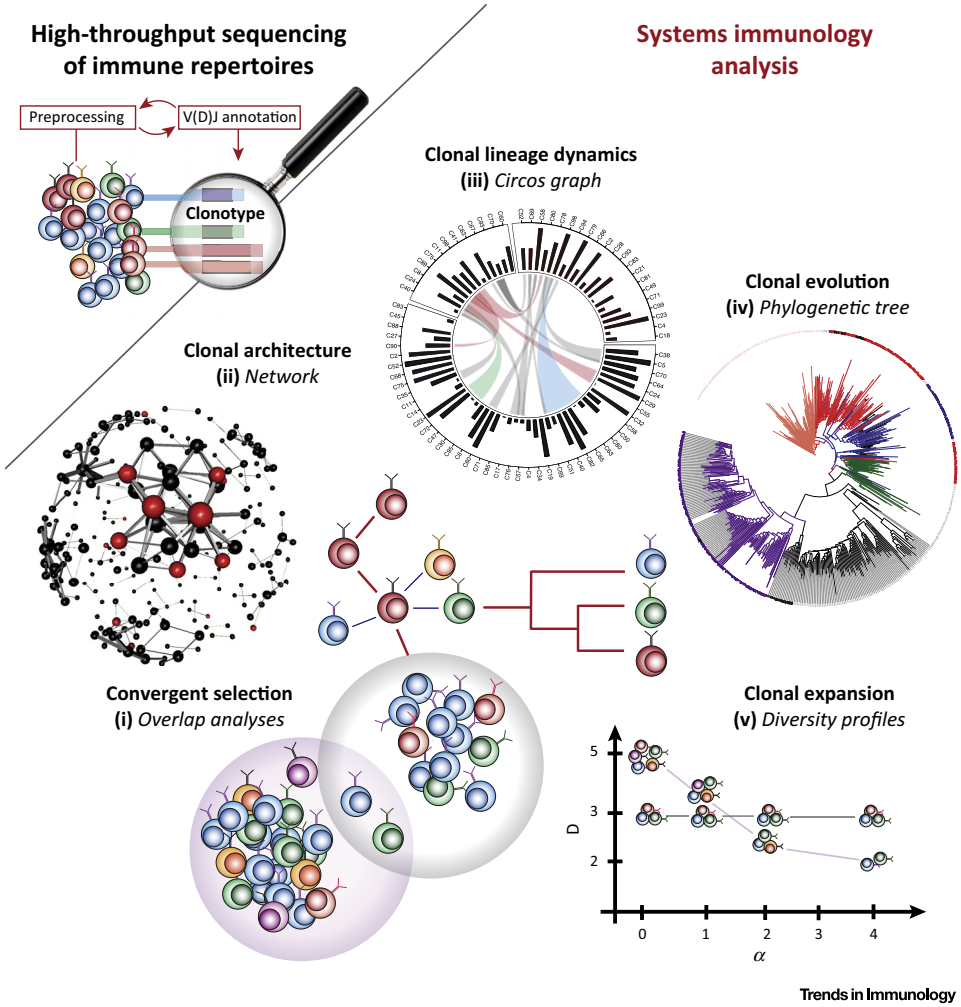

Figure 2. Statistical Analysis and Visualization of High-Throughput Immune Repertoires. The complexity of immune repertoire data necessitates sequence-dependent and sequence-derived analysis. Statistical analyses rely predominantly on clonotyped data and are therefore preceded by a workflow composed of raw data preprocessing (read filtering, error correction), germline annotation (Table 1), and clonotyping. Sequence-dependent approaches (i) visualize convergence of repertoires by quantifying clonal overlap [Venn diagrams; overlap indices such as Morisita–Horn (not shown)]; (ii) display the clonal architecture of repertoires (networks) highlighting denser (clonal expansion) or sparser regions of the repertoire (each vertex is a clone, the size of each vertex is proportional to its abundance, red color highlights selected clones); (iii) reveal dynamics of clones (Circos graphs) shared across samples (sections) by visualizing their change in frequency (bars); or (iv) retrace clonal evolution (phylogenetic trees) helping for instance the visualization of the phylogenetic relation of different clonal lineages (color-coded). Sequence-derived approaches consist mainly of (v) diversity (*D*) profiles (Box 1), which enable the comparison of repertoire diversity and clonal expansion (each line represents the diversity profile of one repertoire, that in purple being more clonally expanded). Legend: each color represents one clone.

across several orders of magnitude (minutes to potentially days, Table 1), IMGT (International Immunogenetics Information System) has become the germline and **numbering scheme** database of choice for the main annotation platforms (Table 1). Because annotation accuracy can vary widely across different platforms [56], the use of simulated V(D)J repertoires [56,76] may now offer the potential to help standardize annotation algorithms.

One limitation of germline reference databases is their inherent incompleteness: recent studies have highlighted the uncertainty regarding the extent of germline polymorphism in humans [77]

and even among widely utilized mouse strains with defined genetic backgrounds (e.g., BALB/c, C57BL/6) [78].

Immune Repertoire Analysis and Visualization

Methods for Quantifying Clonal Convergent Selection, Dynamics, Architecture, and Evolution

The use of Venn diagrams is a classic approach to studying convergence (or overlap) of repertoires (e.g., quantification of shared antigen-specific [25,40,79] or evolutionarily conserved public clones [30]) (Figure 2). While Venn diagrams merely quantify the clonal sequence overlap, the **Morisita–Horn index** [80] (and other overlap indices [81]) quantifies the convergence of both clonal sequences and respective abundance across samples. The use of **Circos graphs** represents a recent advance in the visualization of overlap from complex large datasets [82]; for example, these plots have been used for studying the dynamics of B cell clonal expansion after influenza vaccination [25], and visualized the contribution of specific subsets of naïve B cells to the compartment of antibody-secreting cells in autoimmune disease patients [14]. Unfortunately, Venn diagrams and Circos plots do not scale well with increasing numbers of samples, thus rendering the visualization of the clonal overlap of more than 10 datasets virtually impossible.

Although clonotyping and clonal lineage reconstruction are widespread in the literature, the quantification of overlap of entire clonotypes/clonal lineages across samples remains an unresolved issue of great importance. This is due to mathematical challenges in determining the overlap of sets (repertoire of clonotypes, clonal lineages), which are themselves composed of sets (sequences within clonotypes, clonal lineages). This problem has been partly circumvented by considering either core clonotypes (most-reliable and abundant sequences within clonotypes) [54] or by considering partial overlap of clonal lineage members [25].

Large-scale connectivity analysis between and within repertoires on both non-temporal and temporal scales has been attempted using **network** [69,70,83] and phylogenetic analyses [12,14,84] (Figure 2). Network analysis was used for the visualization of differences in repertoire architecture of individuals of differential immunological status (e.g., healthy and cancer or HIV-attained individuals) by highlighting dense (highly connected clonally expanded clones) and sparse repertoire regions [38,69,70] (Figure 2). These networks are usually constructed by drawing edges between clones (termed vertices or nodes) which differ by a given number of amino acid/nucleotide changes. The size of the vertices may be drawn relative to the abundance of a clone within the repertoire. This strategy enables one to relate clonal sequence architecture to clonal frequency distributions, thus further highlighting regions of the repertoire that have undergone potential disease-specific clonal expansions (Figure 2). Immune repertoire networks have been visualized through the use of software packages such as igraph [69,70,83] and Gephi [38,85].

In contrast to networks, phylogenetic analysis allows the reconstruction of clonal lineage evolution [12,14,84,86] (Figure 2). They have recently been applied for tracing a lineage of HIV-1 broadly-neutralizing antibodies over the timecourse of 15 years [12]. In addition, phylogenetic clustering was used to determine the pairing of antibody heavy and light chains from HIV-1 repertoires to discover novel neutralizing antibody variants [65]. The foundation of all phylogenetic tree construction algorithms represents a sequence alignment [63], which is fed into phylogeny (clonal tree) inferral algorithms such as IgTree [14,84] or Phylip, the latter of which had been originally developed for applications in ecology and macroevolution [87–89]. Visualization of trees is often performed using Dendroscope [12,90]. To date, neither network nor phylogenetic methods are well adapted to the complexity of immune repertoire HTS data, and additional work will be necessary to fully exploit these analytical techniques.

Box 1. Repertoire Diversity Analysis

The diversity (${}^{\alpha}\mathrm{D}$) of a repertoire of S clones is usually calculated as follows: ${}^{\alpha}D = \left(\sum_{i=1}^{S} f_i^{\alpha}\right)^{\frac{1}{1-\alpha}}$ (Hill diversity) [125], where f_i is the frequency of the ith clone weighted by the parameter α. Special cases of this Diversity function correspond to popular diversity indices in the immune repertoire field: species richness ($\alpha = 0$), the exponential Shannon–Weiner ($\alpha \to 1$), the inverse of the Simpson index ($\alpha \to 2$), and the Berger–Parker index ($\alpha \to \infty$). The higher the value of $\propto$, the higher becomes the influence of the higher-abundance clones on the diversity. Owing to the mathematical properties of the diversity function (Schur concavity [104]), two repertoires may yield qualitatively different ${}^{\alpha}D$ values depending on the diversity index used (see Figure 1 in Greiff *et al.* [42]). Diversity profiles, which are vectors of several diversity indices, have, therefore, been suggested to be superior to single diversity indices [42] and are increasingly used in repertoire analyses [42,94,107]. Figure 2 shows two diversity profiles of two immune repertoires of differential clonal expansion. Of note, Chao *et al.* published recently a rarefaction framework for the Hill diversity formula [126], and this will enable the estimation of diversity profiles in the case of undersampled data.

To quantify clonal expansion, diversity can be divided into evenness (${}^{\alpha}\mathrm{e}/{}_{0}\mathrm{D}$) and species richness (${}^{0}\mathrm{D}$) [42,127]. Evenness ranges between 1 (uniform clonal population, every clone occurring in the frequency of ${}^{1}/_{0}\mathrm{D}$) and $\approx {}^{1}/_{0}\mathrm{D}$, in which case one clone completely dominates the immune repertoire.

Methods for Quantification of Clonal Diversity, Clonal Expansion, and SHM

Recent studies suggest that, in general, lymphocyte repertoires are quasi-distinct in clonal composition (see discussion of this phenomenon and associated references in Greiff *et al.* [42]). This restricts sequence-dependent comparisons of immune repertoires across individuals to the comparably small number of public clones, thus disregarding the wealth of information present in entire immune repertoires. However, the comparisons of sequence-derived characteristics, such as diversity, clonal expansion, and SHM count, can be performed at the whole-repertoire level, thereby complementing sequence-dependent analyses [42].

The quantification of clonal expansion and repertoire diversity (Box 1) represents a major goal in HTS repertoire studies because it yields information on the current immunological status of a host [25,42,69,91], which is particularly important for disease and vaccine profiling. The mathematical foundations of biological diversity assessment were developed decades ago for ecological research [92]. Several dedicated **R packages** already exist for diversity index calculations [93–96].

Diversity indices are highly dependent on comprehensive and accurate sequencing [81]. While error correction approaches function to limit overestimation of repertoire diversity (see above), comprehensive sampling of repertoires is challenging to achieve owing to their heavy-tailed clonal frequency distributions – that is, few highly-abundant clones and many low-abundance clones [42,97–99]. The precision of diversity calculation in case of insufficient sampling can be increased using diversity index estimators [81,100,101]. Furthermore, Laydon and colleagues recently published a novel rarefaction-based method for estimating total repertoire size [101,102], which offers advantages to commonly used estimators of species richness such as Chao1 [100,102] and Good–Turing [81,103].

Adding to the problems in repertoire diversity analysis, it has been found that single diversity indices might lead to contradicting qualitative outcomes depending on the diversity measure used ([104] and Box 1). This could yield qualitatively different conclusions regarding the clonal expansion status of a given repertoire [42], and would be especially problematic in the example of clinical lymphoma and immune disorder monitoring [42,69,105,106]. Several groups [77,107] including ourselves [42] have recently published a potential solution to this problem in the form of diversity profiles (Box 1).

While diversity profiles are suitable for comparative analyses of clonal diversity and expansion, estimates of total repertoire size remain an unresolved issue. However, estimates of total

repertoire size remain an unresolved issue [26,102,108]. Reasons for this are: (i) so far it has remained a challenge to cover immune repertoires in their entirety (except for smaller ones such as that of zebrafish [35]); (ii) the discrimination between rare clones and sequencing errors remains a challenge [26]; and (iii) the absence of a validated framework for the diversity profile estimation of undersampled immune repertoire data (Box 1).

SHM is a defining step of clonal selection and expansion during the generation of the B cell response [2]. Annotation methods in Table 1 determine SHM counts independently of clonal-relatedness [62] by alignment with germline reference databases, while other approaches assess SHM phylogenetically [84]. For fundamental immunology, the elucidation of SHM patterns is of high importance for the understanding of how activation-induced deaminase (AID) targets V(D)J regions [109]. In HIV, high SHM counts are a hallmark of HIV-specific broadly-neutralizing antibodies [110], an outstanding feature that could be exploited for discovery of novel therapeutic candidates. The inability, however, to unequivocally separate SHM from PCR and sequencing error [6], as well as the incompleteness of reference germline databases [77], remain fundamental problems in obtaining true absolute SHM counts. Therefore, performing relative SHM analyses, which quantify the differences of SHM counts between cell populations of interest, together with appropriate controls (e.g., naïve B cells, or synthetic spike-ins), may be preferable to interpretations based on absolute SHM counts.

Concluding Remarks

The increased application of immune repertoire HTS to research in immunodiagnostics [42,111], immune response profiling [25,86,112,113], antibody engineering [12,114,115], and lymphocyte development [116,117] continues to expand the rapidly developing field of systems immunology. However, a lack of standardization in bioinformatic and statistical analysis renders the comparison of results across studies challenging. The sensitivity of immune repertoire data analyses would dramatically benefit from the establishment of standards for HTS data pre-processing (e.g., quality filtering, clonotype definition, etc.). Moreover, further development of visualization methods for these high-dimensional, highly diverse, and interconnected data will improve the knowledge gained from immune repertoires (see Outstanding Questions).

Systems immunology-driven studies hold the promise of resolving some longstanding questions in adaptive immunity: (i) what principles drive immune repertoire construction; (ii) what is the size and extent of variation of the expressed immune repertoire; and (iii) are immune repertoires complete in the sense that they could recognize any antigen [108,118]? Answering these questions will require detailed knowledge of interindividual germline variance [77], statistical models of repertoire generation [119–122], and continuous technological advances in DNA sequencing technology and computational biology.

Outstanding Questions

How to standardize HTS and the analysis of immune repertoires? An experimental framework mimicking the large diversity of immune repertoires for the unbiased validation of HTS library preparation methods (PCR, primer bias, and error correction) is missing. Similarly, a standardized repertoire simulation framework for validating bioinformatics processing and analysis pipelines remains to be developed.

How to visualize the connectivity of immune repertoires? Network and phylogenetic analyses are currently visually and computationally unsuitable for large datasets (>10 000 sequences; one small HTS dataset usually has >100 000 sequences). In addition, both phylogenetic and network results can vary substantially in the parameter values used (clonotyping parameters, alignment method, molecular clock model, substitution model).

How to analyze antibody repertoire evolution across time-course (longitudinal) samples? The establishment of a mathematical framework for clonotype/lineage overlap and phylodynamic network analyses on complex datasets will be necessary to investigate antibody (immune) repertoire evolution on a large scale.

References

1. Glanville, J. *et al.* (2009) Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci. U.S.A.* 106, 20216–20221
2. Tonegawa, S. (1983) Somatic generation of antibody diversity. *Nature* 302, 575–581
3. Janeway, C. *et al.* (2004) *Immunobiology* (6th edn), Garland Science
4. Xu, J.L. and Davis, M.M. (2000) Diversity in the CDR3 region of VH is sufficient for most antibody specificities. *Immunity* 13, 37–45
5. Rudolph, M.G. *et al.* (2006) How TCRs bind MHCs, peptides, and coreceptors. *Annu. Rev. Immunol.* 24, 419–466
6. Hershberg, U. and Prak, E.T.L. (2015) The analysis of clonal expansions in normal and autoimmune B cell repertoires. *Philos. Trans. R. Soc. B* 370, 20140239
7. Calis, J.J.A. and Rosenberg, B.R. (2014) Characterizing immune repertoires by high throughput sequencing: strategies and applications. *Trends Immunol.* 35, 581–590
8. Georgiou, G. *et al.* (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* 32, 158–168
9. Robinson, W.H. (2014) Sequencing the functional antibody repertoire – diagnostic and therapeutic discovery. *Nat. Rev. Rheumatol.* 11, 171–182
10. Wu, X. *et al.* (2011) Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* 333, 1593–1602
11. Wang, C. *et al.* (2014) Effects of aging, cytomegalovirus infection, and EBV infection on human B Cell repertoires. *J. Immunol.* 192, 603–611
12. Wu, X. *et al.* (2015) Maturation and diversity of the VRC01-antibody lineage over 15 years of chronic HIV-1 infection. *Cell* 161, 470–485
13. Madi, A. *et al.* (2014) T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Res.* 24, 1603–1612

14. Tipton, C.M. *et al.* (2015) Diversity, cellular origin and autoreactivity of antibody-secreting cell population expansions in acute systemic lupus erythematosus. *Nat. Immunol.* 16, 755–765
15. Hehle, V. *et al.* (2015) Immunoglobulin kappa variable region gene selection during early human B cell development in health and systemic lupus erythematosus. *Mol. Immunol.* 65, 215–223
16. Muraro, P.A. *et al.* (2014) T cell repertoire following autologous stem cell transplantation for multiple sclerosis. *J. Clin. Invest.* 124, 1168–1172
17. Dunn-Walters, D.K. and Ademokun, A.A. (2010) B cell repertoire and ageing. *Curr. Opin. Immunol.* 22, 514–520
18. Britanova, O.V. *et al.* (2014) Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *J. Immunol.* 1950, 2689–2698
19. Benichou, J. *et al.* (2012) Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* 135, 183–191
20. Morbach, H. *et al.* (2010) Reference values for B cell subpopulations from infancy to adulthood. *Clin. Exp. Immunol.* 162, 271–279
21. Ganusov, V.V. and De Boer, R.J. (2007) Do most lymphocytes in humans really reside in the gut? *Trends Immunol.* 28, 514–518
22. Farber, D.L. *et al.* (2014) Human memory T cells: generation, compartmentalization and homeostasis. *Nat. Rev. Immunol.* 14, 24–35
23. Trepel, F. (1974) Number and distribution of lymphocytes in man. A critical analysis. *J. Mol. Med.* 52, 511–515
24. Johnson, K.M. *et al.* (2002) Aging and developmental transitions in the B cell lineage. *Int. Immunol.* 14, 1313–1323
25. Jackson, K.J.L. *et al.* (2014) Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. *Cell Host Microbe* 16, 105–114
26. 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
27. Jackson, K.J.L. *et al.* (2013) The shape of the lymphocyte receptor repertoire: lessons from the B cell receptor. *Front. B: Cell Biol.* 4, 263
28. Venturi, V. *et al.* (2013) Specificity, promiscuity, and precursor frequency in immunoreceptors. *Curr. Opin. Immunol.* 25, 639–645
29. Shugay, M. *et al.* (2013) Huge overlap of individual TCR beta repertoires. *T Cell Biol.* 4, 466
30. Robins, H.S. *et al.* (2010) Overlap and effective size of the human $CD8^+$ T-cell receptor repertoire. *Sci. Transl. Med.* 2, 47ra64
31. 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
32. Becattini, S. *et al.* (2014) Functional heterogeneity of human memory $CD4^+$ T cell clones primed by pathogens or vaccines. *Science* 347, 400–406
33. 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
34. Bashford-Rogers, R.J. *et al.* (2014) Capturing needles in haystacks: a comparison of B-cell receptor sequencing methods. *BMC Immunol.* 15, 29
35. Weinstein, J.A. *et al.* (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science* 324, 807–810
36. Menzel, U. *et al.* (2014) Comprehensive evaluation and optimization of amplicon library preparation methods for high-throughput antibody sequencing. *PLoS ONE* 9, e96727
37. Gotelli, N.J. and Colwell, R.K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* 4, 379–391
38. Lindner, C. *et al.* (2015) Diversification of memory B cells drives the continuous adaptation of secretory antibodies to gut microbiota. *Nat. Immunol.* 16, 880–888
39. Nair, N. *et al.* (2015) High-dimensional immune profiling of total and rotavirus VP6-specific intestinal and circulating B cells by mass cytometry. *Mucosal Immunol.* Published online April 22, 2015. http://dx.doi.org/10.1038/mi.2015.36
40. Estorninho, M. *et al.* (2013) A novel approach to tracking antigen-experienced CD4 T cells into functional compartments via tandem deep and shallow TCR clonotyping. *J. Immunol.* 191, 5430–5440
41. Venturi, V. *et al.* (2011) A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *J. Immunol.* 186, 4285–4294
42. Greiff, V. *et al.* (2015) A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med.* 7, 49
43. Loman, N.J. *et al.* (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* 30, 434–439
44. 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
45. Shugay, M. *et al.* (2014) Towards error-free profiling of immune repertoires. *Nat. Methods* 11, 653–655
46. 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
47. Heiden, J.A.V. *et al.* (2014) pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* 30, 1930–1932
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. Howie, B. *et al.* (2015) High-throughput pairing of T cell receptor α and β sequences. *Sci. Transl. Med.* 7, 301ra131
50. 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
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. Nguyen, L.V. *et al.* (2014) Clonal analysis via barcoding reveals diverse growth and differentiation of transplanted mouse and human mammary stem cells. *Cell Stem Cell* 14, 253–263
53. Andrews, S. (2010) FastQC: a quality control tool for high throughput sequence data
54. 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
55. Masella, A.P. *et al.* (2012) PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* 13, 31
56. Bolotin, D.A. *et al.* (2015) MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* 12, 380–381
57. Bolotin, D.A. *et al.* (2013) MiTCR: software for T-cell receptor sequencing data analysis. *Nat. Methods* 10, 813–814
58. Li, S. *et al.* (2014) Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.* 32, 888–895
59. Seqc/Maqc-Iii Consortium (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* 32, 903–914
60. Yassai, M.B. *et al.* (2009) A clonotype nomenclature for T cell receptors. *Immunogenetics* 61, 493–502
61. Chen, Z. *et al.* (2010) Clustering-based identification of clonally-related immunoglobulin gene sequence sets. *Immunome Res.* 6, S4
62. 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
63. Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461
64. Fu, L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152
65. 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

66. 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
67. Liao, H-X. *et al.* (2013) Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* 496, 469–476
68. Sok, D. *et al.* (2013) The effects of somatic hypermutation on neutralization and binding in the PGT121 family of broadly neutralizing hiv antibodies. *PLoS Pathog.* 9, e1003754
69. Bashford-Rogers, R.J.M. *et al.* (2013) Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res.* 23, 1874–1884
70. Hoehn, K.B. *et al.* (2015) Dynamics of immunoglobulin sequence diversity in HIV-1 infected individuals. *Philos. Trans. R. Soc. B* 370, 20140241
71. Laserson, U. *et al.* (2014) High-resolution antibody dynamics of vaccine-induced immune responses. *Proc. Natl. Acad. Sci. U.S.A.* 111, 4928–4933
72. Sundling, C. *et al.* (2014) Single-cell and deep sequencing of IgG-switched macaque B cells reveal a diverse Ig repertoire following immunization. *J. Immunol.* 192, 3637–3644
73. Zhu, J. *et al.* (2012) Somatic populations of PGT135-137 HIV-1-neutralizing antibodies identified by 454 pyrosequencing and bioinformatics. *Front. Microbiol.* 3, 315
74. Wang, C. *et al.* (2015) B-cell repertoire responses to varicella-zoster vaccination in human identical twins. *Proc. Natl. Acad. Sci. U.S.A.* 112, 500–505
75. Francica, J.R. *et al.* (2015) Analysis of immunoglobulin transcripts and hypermutation following SHIVAD8 infection and protein-plus-adjuvant immunization. *Nat. Commun.* 6, 1–14
76. Safonova, Y. *et al.* (2015) IgSimulator: a versatile immunosequencing simulator. *Bioinformatics* 31, 3213–3215
77. 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
78. Collins, A.M. *et al.* (2015) The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. *Philos. Trans. R. Soc. B* 370, 20140236
79. 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
80. Horn, H.S. (1966) Measurement of 'overlap' in comparative ecological studies. *Am. Nat.* 100, 419–424
81. Rempala, G.A. and Seweryn, M. (2013) Methods for diversity and overlap analysis in T-cell receptor populations. *J. Math. Biol.* 67, 1–30
82. Krzywinski, M. *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645
83. Csardi, G. and Nepusz, T. (2006) The igraph software package for complex network research. *Int. J. Complex Syst.* 1695
84. Barak, M. *et al.* (2008) IgTree©: creating Immunoglobulin variable region gene lineage trees. *J. Immunol. Methods* 338, 67–74
85. Bastian, M. *et al.* (2009) Gephi: an open source software for exploring and manipulating networks. *ICWSM* 8, 361–362
86. Jiang, N. *et al.* (2013) Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci. Transl. Med.* 5, 171ra19
87. Di Niro, R. *et al.* (2015) *Salmonella* infection drives promiscuous B cell activation followed by extrafollicular affinity maturation. *Immunity* 43, 120–131
88. Felsenstein, J. (1989) PHYLIP – phylogeny inference package (version 3.2). *Cladistics* 5, 164–166
89. Revell, L.J. and Chamberlain, S.A. (2014) Rphylip: an R interface for PHYLIP. *Methods Ecol. Evol.* 5, 976–981
90. Huson, D.H. *et al.* (2007) Dendroscope: an interactive viewer for large phylogenetic trees. *BMC Bioinformatics* 8, 460
91. Attaf, M. *et al.* (2015) αβ T cell receptors as predictors of health and disease. *Cell. Mol. Immunol.* 12, 391–399
92. Magurran, A.E. (1988) *Ecological Diversity and Its Measurement,* Princeton University Press
93. Cortina-Ceballos, B. *et al.* (2015) Reconstructing and mining the B cell repertoire with ImmunediveRsity. *MAbs* 7, 516–524
94. Gupta, N.T. *et al.* (2015) Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* Published online June 10, 2015. http://dx.doi.org/10.1093/bioinformatics/btv359
95. Oksanen, J. *et al.* (2015) *Vegan: Community Ecology Package*. Published online September 25, 2015. https://cran.r-project.org/web/packages/vegan/vegan.pdf
96. Nazarov, V.I. *et al.* (2015) tcR: an R package for T cell receptor repertoire advanced data analysis. *BMC Bioinformatics* 16, 175
97. Mora, T. *et al.* (2010) Maximum entropy models for antibody diversity. *Proc. Natl. Acad. Sci. U.S.A.* 107, 5405–5410
98. Schwab, D.J. *et al.* (2014) Zipf's law and criticality in multivariate data without fine-tuning. *Physical. Rev. Lett.* 113, 068102
99. Bolkhovskaya, O.V. *et al.* (2014) Assessing T cell clonal size distribution: a non-parametric approach. *PLoS ONE* 9, e108658
100. Chao, A. and Shen, T-J. (2003) Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environ. Ecol. Stat.* 10, 429–443
101. Laydon, D.J. *et al.* (2014) Quantification of HTLV-1 clonality and TCR diversity. *PLoS Comput. Biol.* 10, e1003646
102. Laydon, D.J. *et al.* (2015) Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. *Philos. Trans. R. Soc. B* 370, 20140291
103. Good, I.J. (1953) The population frequencies of species and the estimation of population parameters. *Biometrika* 40, 237–264
104. Solomon, D.L. *et al.* (1975) A comparative approach to species diversity *Biometrics Unit Technical Reports* BU-573-M. http://ecommons.library.cornell.edu/handle/1813/32672
105. Boyd, S.D. *et al.* (2009) Measurement and clinical monitoring of human lymphocyte clonality by massively parallel V-D-J pyrosequencing. *Sci. Transl. Med.* 1, 12ra23
106. Roskin, K.M. *et al.* (2015) lgH sequences in common variable immune deficiency reveal altered B cell development and selection. *Sci. Transl. Med.* 7, 302ra135
107. Snir, O. *et al.* (2015) Analysis of celiac disease autoreactive gut plasma cells and their corresponding memory compartment in peripheral blood using high-throughput sequencing. *J. Immunol.* 194, 5703–5712
108. Zarnitsyna, V. *et al.* (2013) Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. *T Cell Biol.* 4, 485
109. Yaari, G. *et al.* (2013) Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput Immunoglobulin sequencing data. *Front. B: Cell Biol.* 4, 358
110. Klein, F. *et al.* (2013) Somatic mutations of the immunoglobulin framework are generally required for broad and potent HIV-1 neutralization. *Cell* 153, 126–138
111. Chaussabel, D. (2015) Assessment of immune status using blood transcriptomics and potential implications for global health. *Semin. Immunol.* 27, 58–66
112. Parameswaran, P. *et al.* (2013) Convergent antibody signatures in human dengue. *Cell Host Microbe* 13, 691–700
113. Luciani, F. *et al.* (2012) Next generation deep sequencing and vaccine design: today and tomorrow. *Trends Biotechnol.* 30, 443–452
114. 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
115. Cheung, W.C. *et al.* (2012) A proteomics approach for the identification and cloning of monoclonal antibodies from serum. *Nat. Biotechnol.* 30, 447–452
116. Kaplinsky, J. *et al.* (2014) Antibody repertoire deep sequencing reveals antigen-independent selection in maturing B cells. *Proc. Natl. Acad. Sci. U.S.A.* 111, E2622–E2629
117. Shi, W. *et al.* (2015) Transcriptional profiling of mouse B cell terminal differentiation defines a signature for antibody-secreting plasma cells. *Nat. Immunol.* 16, 663–673

118. Perelson, A.S. and Oster, G.F. (1979) Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self-non-self discrimination. *J. Theor. Biol.* 81, 645–670

119. Murugan, A. *et al.* (2012) Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl. Acad. Sci. U.S.A.* 109, 16161–16166

120. Elhanati, Y. *et al.* (2015) Inferring processes underlying B-cell repertoire diversity. *Philos. Trans. R. Soc. B* 370, 20140243

121. Zvyagin, I.V. *et al.* (2014) Distinctive properties of identical twins' TCR repertoires revealed by high-throughput sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 111, 5980–5985

122. Elhanati, Y. *et al.* (2014) Quantifying selection in immune receptor repertoires. *Proc. Natl. Acad. Sci. U.S.A.* 111, 9875–9880

123. Ye, J. *et al.* (2013) IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* 41, W34–W40

124. 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

125. Hill, M.O. (1973) Diversity and evenness: a unifying notation and its consequences. *Ecology* 54, 427–432

126. Chao, A. *et al.* (2013) Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecol. Monogr.* 84, 45–67

127. Jost, L. (2010) The relation between evenness and diversity. *Diversity* 2, 207–232