Quantifying lymphocyte receptor diversity

Thierry Mora[1], Aleksandra M. Walczak[2]
[1] *Laboratoire de physique statistique, CNRS, UPMC and École normale supérieure, 24, rue Lhomond, Paris, France and*
[2] *Laboratoire de physique théorique, CNRS, UPMC and École normale supérieure, 24, rue Lhomond, Paris, France*

To recognize pathogens, B and T lymphocytes are endowed with a wide repertoire of receptors generated stochastically by V(D)J recombination. Measuring and estimating the diversity of these receptors is of great importance for understanding adaptive immunity. In this chapter we review recent modeling approaches for analyzing receptor diversity from high-throughput sequencing data. We first clarify the various existing notions of diversity, with its many competing mathematical indices, and the different biological levels at which it can be evaluated. We then describe inference methods for characterizing the statistical diversity of receptors at different stages of their history: generation, selection and somatic evolution. We discuss the intrinsic difficulty of estimating the diversity of receptors realized in a given individual from incomplete samples. Finally, we emphasize the limitations of diversity defined at the level of receptor sequences, and advocate the more relevant notion of functional diversity relative to the set of recognized antigens.

I. INTRODUCTION

To protect its host against pathogens, the adaptive immune system of jawed vertebrates expresses a large repertoire of distinct receptors on its B- and T lymphocytes. These receptors must recognize a wide range of pathogens to trigger the response of the adaptive immune system. Since each receptor is specialized in recognizing specific pathogens, a very diverse repertoire of receptors is required to cover all possible threats. While one can now sequence the repertoires of individuals with some depth, it remains unclear how to quantify or even define their diversity, and what aspects of this diversity are relevant for recognition. These fundamental questions are further obscured by the purely technical but important issue of reliably sampling immune repertoires.

The actual number of lymphocytes varies from species to species, but in all cases is large. Estimates of the number of T cells in humans are of the order of $3 \cdot 10^{11}$ cells [1]. Each cell expresses only one type of receptor. Cells proliferate and form clones, so that many distinct cells may share a common receptor. As we will discuss further, the number of unique distinct receptors is very hard to estimate. However, even a conservative lower bound of 10^6 unique receptors [2, 3] is much larger than the total number of genes in the human genome ($\sim 20,000$). This broad diversity of receptors is not hard-coded, but is instead generated by a unique gene rearrangement process that couples a combinatoric choice of genomic templates with additional randomness.

Each receptor is made up of two arms: B-cell receptors (BCR) have a light and a heavy chains, while T-cell receptors (TCR) have analogous α and β chains. Each chain is composed of three segments called V, D and J in the case of heavy or β chains, and two segments V and J in the case of light or α chains. These segments are combinatorically picked out of several genomic templates for each type, in a process called V(D)J recombination [4], as schematized in Fig. 1A. This recombination is achieved by looping DNA and excising the template genes that lie between the selected gene segments. In the case of heavy or β chains, the D-J junction is assembled first, followed by the V-D junction. The precise number of templates for each segment differs from species to species, but generally results in a combinatoric diversity of ~ 1000 for each chain. This combinatoric assortment is followed by stochastic nucleotide deletions and insertions at the junctions between the newly assorted V-D and D-J fragments (or V-J fragment for the shorter chain), forming what is termed junctional diversity. This stochastic step largely increases the repertoire diversity, as we will show in detail. As a result of this procedure the receptor DNA may be out-of-frame, or the encoded protein may not be functional or correctly folded. The newly assembled β chain sequences then are tested with a surrogate α chain for their binding and expression properties. If they pass this selection step, the second chain is assembled and the whole receptor undergoes a similar round of selection against proteins that are natural to the organism, or self proteins. Receptors that do not bind any self-protein or bind too strongly to self-proteins are discarded. If a receptor fails these tests, the cell may attempt to recombine its second chromosome.

The processes of recombination and selection are stochastic, and therefore are characterized by their own intrinsic diversity, which we may view as a statistical or potential diversity. It is distinct from the diversity realized in a given individual at a given time, with its finite number of recombined receptors, much like the potential diversity of the English language is distinct from – and much larger than – the diversity of texts found in a single library. While most previous discussions, with the expection of [5], have focused on the realized rather than potential diversity of receptors, in this chapter we will discuss both.

After generation and selection, B- and T cells feed the naive repertoire where they attempt to recognize foreign antigens (Fig. 1B). The dynamics of lymphocytes vary widely between B and T cells, as well as between species. However, a common feature is that cells whose receptors successfully bind to antigens proliferate, producing either

identical offspring (T-cells) or that differ by somatic point hypermutations (B-cells). A fraction of the cells that have undergone proliferation are kept in what is called the memory repertoire, while cells that have not received a proliferation signal stay in the naive repertoire. Cells that share a common receptor, or "clonotype," define a clone. The clonal structure of the lymphocyte repertoire is one of the characteristics of repertoire diversity.

The diversity of lymphocyte receptors can be studied with the help of repertoire high-throughput sequencing experiments [2, 6–8], which have been developing rapidly over the last few years [9–14]. These experiments focus on the region of the chain that encompasses the junctions between the recombined segments, allowing for the complete identification of the receptor chain. This region includes the Complementarity Determining Region 3 (CDR3), defined from roughly the end of the V segment to the beginning of the J segment, which is believed to play an important role in recognition. Because sequence reads can only cover one of the two chains making up the receptor, most studies have focused on the diversity of one chain at a time. However, new techniques make it possible to pair the two chains together [15–17], opening the way for the analysis of repertoires of complete receptors. In general, a tissue (blood, lymph node, thymus, germinal center, etc.) sample is taken and the mRNA or DNA of the lymphocytes of interest are sorted out. Different technologies have been developed for DNA and mRNA. Data are usually clustered and error-corrected for PCR and sequencing errors [18]. Many recent experiments use unique molecular barcodes associated to each initial mRNA molecule, which help correct for PCR amplification noise [19–21], and allow for the direct measurement of relative clone sizes using sequence counts. Unless an error occurred in the first round of PCR, barcodes can reliably pick up even very rare sequences, as long as they are present in the sample. These experiments result in a list of unique receptor chain sequences, and if the data was barcoded, of reliable counts for the corresponding number of RNA molecules in the initial sample. This information is the staring point for the analysis of repertoire diversity.

In this chapter we discuss approaches for estimating repertoire diversity from the datasets generated by these new technologies. We first review and discuss the different definitions of diversity – species richness, entropy, and other diversity indices – and their relation to the distribution of clonotype frequencies. We also emphasize the need to distinguish the different levels at which diversity may be evaluated: recombination diversity, post-selection potential diversity, actual diversity realized in a particular individual, in a particular tissue, or with a particular phenotype, etc. We review recent efforts to calculate accurately the diversity of receptors generated by V(D)J recombination using high-throughput sequencing data. We discuss the challenges of estimating diversity when the clonal structure is scale-free, as is generically the case in many reported cases. We conclude by discussing the importance of sequence diversity and contrast it with more biologically relevant but elusive notion of functional diversity.

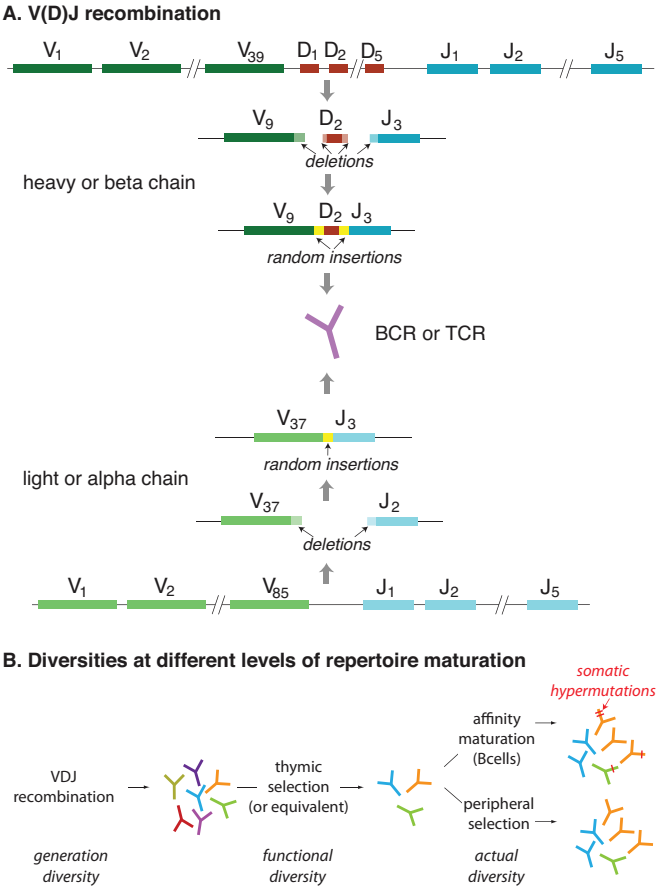

FIG. 1: A. V(D)J recombination of T- and B cell receptors (TCR and BCR). TCRs and BCRs are made of two chains, one shorter and one longer, called the α and β chains for TCRs, and the light and heavy chains for BCRs. Each chain is obtained by a gene rearrangement process called V(D)J recombination, by which two (for the shorter chain) or three (for the longer chain) segments are assembled together from palettes of templates encoded in the genome. At each of the junction between these segments, further diversity is added by stochastic deletions and insertions of random, non-templated nucleotides. B. Evolution of repertoires of TCR and BCR. After their generation by V(D)J recombination, receptors first pass a selection process, called thymic selection for TCRs, whereby nonfunctional and self-reactive receptors are discarded. They are then released into the periphery, where they may divide, die, proliferate and differentiate as a function of the signals they receive from antigens or other immune cells. In addition, BCRs are subject to somatic hypermutations as B cells mature in germinal centers following an infection.

II. A FAMILY OF DIVERSITY MEASURES

A number of different diversity measures have been proposed to quantify the vastness of lymphocyte reper-

toires [22–24]: the Shannon entropy [25], the Simpson index [26], and most commonly the total number of clonotypes or species richness [2, 3, 27–29]. These diversity measures are taken from ecology, where they are used to quantify the diversity of species. They are all related to a generalized family of diversity measures called the Rényi entropy [30], parametrized by β and defined as:

$$H_\beta = \frac{1}{1-\beta} \ln \left[\sum_s p(s)^\beta \right], \tag{1}$$

where $p(s)$ is the probability, frequency or abundance of a given receptor sequence or clonotype s. For $\beta \to 1$ we recover Shannon's entropy:

$$H_1 = -\sum_s p(s) \ln p(s). \tag{2}$$

The exponential of the Rényi entropy defines a generalized class of diversity indices called Hill diversities [31]:

$$D_\beta = \exp[H_\beta]. \tag{3}$$

This index can be interpreted as an effective number of clonotypes in the data. For $\beta = 1$, it is simply the exponential of Shannon's entropy, and we will refer to it as Shannon's diversity. For $\beta = 2$, it reduces to the inverse of Simpson's diversity index, $D_2 = 1/\sum_s p(s)^2$. The Simpson index gives the probability that two sequences drawn at random from the distribution are identical, and is related to a common measure of inequality, the Gini-Simpson index, defined as $1 - 1/D_2$. D_0 is the species richness, while $D_\infty = 1/\max_s p(s)$ is the inverse of the Berger-Parker index.

Each of these diversity indices is a summary statistics of the information contained in the distribution of clonotype frequencies, *i.e.* the distribution of values of $p(s)$ themselves. This frequency distribution may in fact be viewed as the most complete description of the diversity of the repertoire. Conversely, the whole spectrum of Rényi entropies H_β is sufficient to reconstruct the full clonotype frequency distribution. In other words, the functions H_β, D_β, and the distribution of frequencies carry the exact same information [32]. The choice of a single diversity measure D_β, rather than the full frequency distribution, is often useful to make comparisons between individuals, tissues, experiments, etc. When β is large enough, it may also be less sensitive to experimental noise than the frequency distribution.

It is possible to get a rough estimate of Hill diversities by simple inspection of the frequency distribution, represented as a rank-frequency graph with a double logarithmic scale [32]. A simple geometric construction, illustrated by Fig. 2, helps understand the meaning of the various indices, what properties of the underlying cumulative clone size distribution they are most likely to capture, and where one should stop trusting them because of insufficient sampling. The intersection of the the tangents of slope -1 and $-\beta^{-1}$ to the rank-frequency curve

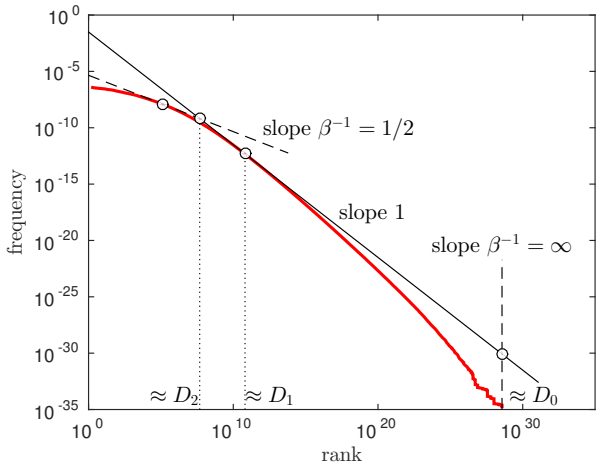

FIG. 2: Geometric construction of Hill diversities from a rank-frequency curve. The Hill diversity of order β, $D_\beta = [\sum_s p(s)^\beta]^{1/(1-\beta)}$, can be approximated from the intersection between the tangents of slope -1 and $-1/\beta$. D_0 is the total number of types or species richness; D_1 is the exponential of Shannon's entropy; and D_2 is the inverse of Simpson's diversity index.

gives the Hill diversity index D_β. This construction emphasizes the fact that different diversity measures focus on sequences of various frequencies: large values of β tend to favor very common clonotypes, while low values favor rare ones. Geometrically, tangents of small slopes (large β, e.g. Simpson's index or Shannon's entropy) osculate the rank-frequency curve at high frequencies, while large slopes do so at low frequencies. Thus, diversity indices D_β with a small β rely very strongly on correclty capturing the tail of rare clonotypes. This is particularly true for D_0, the species richness, which is very hard to estimate as it requires to estimate the number of unseen clonotypes. This observation warns us against the pitfalls of estimating diversity when dealing with incomplete samples. The larger the β, the more reliable the Hill index D_β should be. In general, estimates of the species richness D_0 should be taken with extreme caution, as we will further discuss in concrete examples.

III. QUANTIFYING V(D)J RECOMBINATION

The repertoire is a dynamic ensemble of receptors that evolves somatically. As the repertoire is shaped, its diversity changes significantly. Repertoires at different functional stages, from generation to memory, show different levels of potential and realized diversity. By analyzing unique receptors from high-throughput sequencing data, one can track these changes. We start by decribing the diversity of the initial stochastic recombination of receptors.

Each cell has two sets of chromosomes. If the first V(D)J rearrangement results in a non-functional recep-

tor, the second one recombines [33]. When this second rearrangement is successful, the cell expresses the functional receptor, but keeps the rearranged nonfunctional DNA. This nonfunctional receptor is expressed at a basal, leaky level despite allelic exclusion, especially for α chains, and may also be captured by genomic DNA sequencing. These out-of-frame receptors offer unique insight into the raw generation process, because they were never selected for, as they owe their survival to the gene expressed from the other chromosome. We can therefore use these sequences to gain insight into the generation process, and analyze the *potential* diversity of recombination, *i.e.* the statistics of unique receptors that can ever be formed as a result of V(D)J recombination. As already noted, this diversity of the generation process should not be confused with the actually realized diversity in a given individual, which is generically smaller.

As the numbers will show, the recombination probability of each generated sequence is so small that it is hopeless to sample their distribution by simply counting how often we observe them. Besides, this counting number is not expected to reflect the frequency of generation alone, because of lymphocyte population dynamics. As we pointed out, cell proliferation is independent of the identity of the out-of-frame sequence of interest, and in the limit of infinite data should not in principle affect such an estimate. However, for any dataset coming from a single individual, these heterogeneities in the clone size completely dominate the sequence counts. For this reason, it is suitable to count each unique sequence only once to remove these possible biases. Starting with a dataset of unique realizations of the recombination process, we need a model to describe their probability distribution. This model is based on what we know about the recombination process: choice of V(D)J segments, stochastic number of deletions of each gene segments, stochastic number and identities of inserted nucleotides at each junction. Thus, taking the simpler case of α or light chains, the probability of a given recombination scenario r can be written as:

$$P_{\rm rearr}(r) = P(V,J)P(\mathrm{del}V|V)P(\mathrm{del}J|J)P(\mathrm{ins}), \quad (4)$$

where del$V$ and del$J$ denote the number of deletions at the V and J ends, and "ins" is the list of inserted nucleotides. A very similar expression accounting for three genes and two junctions can be written for the β or heavy chains. The form of the model is motivated by biophysical considerations: the number of deletions of the J end does not depend on the choice of the V segment, the number and identities of insertions does not depend on the gene choice, and follows a Markov chain. These assumptions, however, should and can be checked consistently by verifying that no correlations in the data remains unaccounted for by the model [34].

The parameters of the generation model (4) cannot be directly read off the sequences, because it is impossible in general to assign with certainty a recombination scenario to a given sequence, as many distinct scenarios can lead to the same sequence through convergent recombination [35]. As we will quantify below, this effect is very significant and cannot be ignored. Importantly, it forces us to think of scenarios or sequence annotation in a probabilistic manner, rather than try to select the most probable one as is often done in annotation software [36–38]. The generation parameters can be inferred using a standard implementation of the Expectation-Maximization algorithm, an iterative procedure that maximizes the likelihood of the data. The algorithm works by collecting summary statistics about the elements of the recombination scenarios to build the model distribution (4). The recombination scenarios are themselves assigned probabilistically using the previous iteration of the model. The algorithm, which relies on the enumeration of all plausible scenarios giving rise to each sequence, is computationnally heavy, but can be significantly sped up after mapping the problem onto a hidden Markov model and using standard dynamic programming tools [39].

Once a recombination model such as Eq. 4 has been inferred, it can be used to generate and analyse sequences with the same statistical properties as the original data. It can also be used to quantity the various types of diversity indices discussed in the previous section. Note that, because of convergent recombination, the diversity of generated sequences is expected to be smaller than the diversity of the scenarios that produce them. The generation probability of a sequence s is given by the sum of the probabilities of all scenarios that could have given rise to this sequence:

$$P_{\rm gen}(s) = \sum_{r\to s} P_{\rm rearr}(r). \quad (5)$$

The diversity measures calculated from $P_{\rm gen}$ and $P_{\rm rearr}$ are therefore distinct.

Recombination models have been inferred for T cell β [34] and α [39] chains, as well as for B cell heavy chains [40]. In all these cases, the distributions inferred from different individuals were found to be surprisingly similar, with some variability in the gene segment usage, but very reproducible deletion and insertion profiles, consistent with a common biophysical mechanism of enzyme function. The entropy H_1 of sequences and recombination scenarios obtained from these models are reported in Fig. 3A. Because the distribution of scenarios (4) is a product of its various elements (gene choice, deletions, insertions), its entropy can also be broken up into their respective contributions. The entropy difference between recombination events (in purple) and sequences (in red), is the entropy of convergent recombination (in gray), which quantifies the diversity of scenarios resulting in the same sequence. For example, it is 5 bits for TCR β chains, corresponding to a fairly large Shannon diversity number, $D_1 \sim 30$. Note that the total number of possible scenarios for a given sequence, D_0 is much larger, but its precise definition depends on the cutoff we impose on the possible number of deletions and insertions.

Diversity in the heavy chain of B-cells is larger than

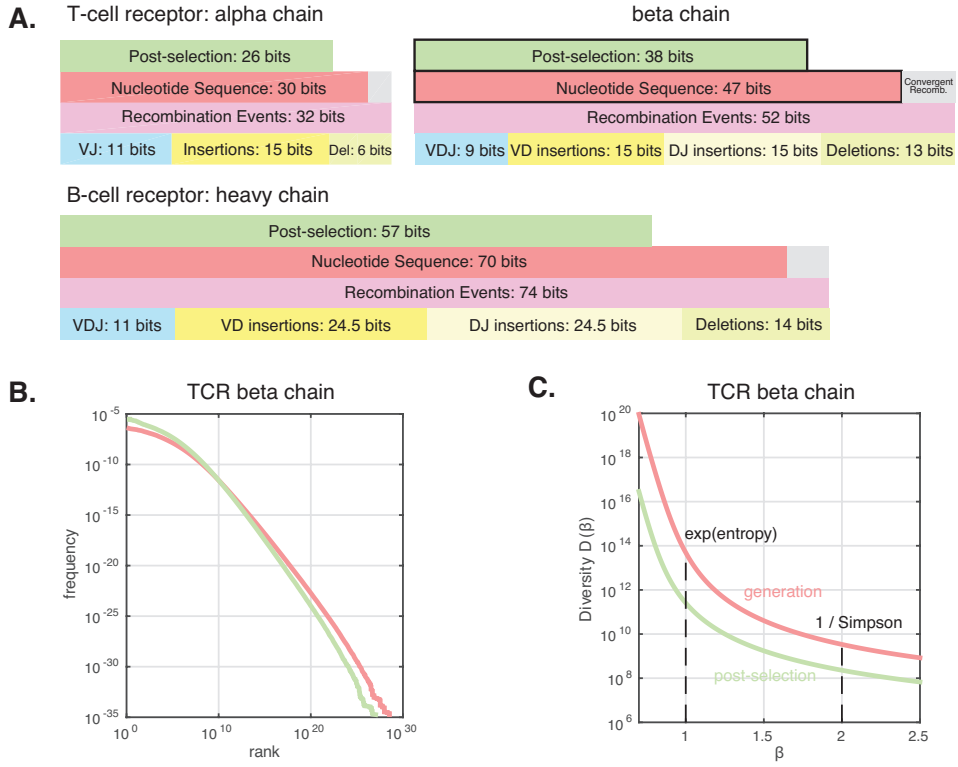

FIG. 3: Entropies and diversity indices of the receptor generation and selection process. A. Entropy of the V(D)J recombination process in TCR α and β chains, and in BCR heavy chains. The entropy of recombination events (purple) can be decomposed into contributions for the choice of the V(D)J genes (blue), the number and identity of insertions (yellow), and deletions (light green). The sequence entropy (red) is slightly smaller than the recombination entropy because several recombination events can lead to the same sequence (convergent recombination, in gray). Following thymic selection, or the B-cell counterpart, the entropy is further reduced (green). B. Rank-frequency curves of TCR β chain sequences, upon generation (red), and following thymic selection (green). C. Hill diversities for the same statistical ensembles. The Shannon diversity $D(1)$ is the exponential of the entropies shown in black boxes in A.

that of T-cells. This difference can be attributed to longer CDR3 regions due to many more insertions at the junctions between the genes. The receptor generation process is characterized by an entropy of ~ 70 bits for BCR heavy chains and ~ 43 bits for TCR β chains. These numbers correspond to a Shannon diversity index $D_1 \sim 10^{21}$ and $\sim 10^{14}$, respectively.

Although most studies have focused on the Shannon diversity index D_1, the full diversity spectrum of the generation process can be calculated. In Fig. 3B we show the rank-frequency curve of human TCR β chains, taken from Ref. [32] based on the model of Ref. [34]. As explained in the previous section, the full range of diversity indices D_β can be calculated from that curve, and are shown in Fig. 3C. In addition to the Shannon diversity D_1 already discussed, of special interest is the inverse of the Simpson index, D_2. The Simpson index corresponds to the probability that the same nucleotide sequence is obtained from two independent draws. It gives the expected number of shared sequences between two individuals, normalized by the product of their repertoire sizes, assuming that their receptor sequences were generated independently from the same source. Thus, it is deeply linked to the notion of "public" sequences found in several individuals, and making up the public repertoire [26, 35, 41]. This number, estimated to be $1/D_2 \sim 3 \cdot 10^{-10}$ for human TCR β chains from the model, is in fact very close to that measured in the data for out-of-frame sequences [34].

It is important to stress that, however large, these numbers are *not* the total number of possible receptor sequences, D_0, which is much larger. As we can see from the rank-frequency plot of generated TCR β chain sequences (Fig. 3B, red), generation probabilities span over 20 orders of magnitude. The largest rank of $\sim 10^{30}$ is in fact a lower bound to D_0 limited by the finite sampling of sequences by the model. To better estimate D_0, one may count the total number of possible deletion profiles reported for each gene, and multiply that number by the total number of possible insertion profiles of at most $L_{\max}$ nucleotides, $(4^{L_{\max}-1})/3$, for each of the two junctions. Doing so with $L_{\max} = 26$, the largest number of insertions reported in [34], yields an upper bound of $D_0 \sim 2 \cdot 10^{39}$ for the TCR β chain alone. However, because this estimate is very sensitive to the value of $L_{\max}$, which is not precisely known and may depend on the

sample size, it must be taken with some caution.

The above estimates only include heavy or β chains. Coupling this chain with the light or α chain adds further diversity. Since the shorter (α and light) chains have only one junctional region between the V and J genes, their diversity is much lower. For example, TCR α chains were estimated to have a generation Shannon entropy of $H_1 = 30$ bits, or $D_1 \sim 10^9$ [39]. The part of the entropy that is attributable to the gene choice is similar to that reported for the β chain, of the order of 10 bits. While that contribution was only a small fraction of the overall diversity for the β chain, it is comparable to that of insertions for the α chain. The number of possible α chain sequences can be estimated similarly to the β chain, yielding $D_0 \sim 5 \cdot 10^{21}$.

Assuming that the two chain rearrangements are independent, the overall diversity of the pool from which TCRs are generated is about $H_1 \sim 75$ bits, or $D_1 \sim 10^{23}$, and a total potential repertoire of size $D_0 \sim 10^{61}$. Note that this last estimate is much larger than the classically quoted number of 10^{15} from [42], which assumed a much more restricted junctional diversity. Analysis of recently published α-β sequence pairings should allow for more precise estimates of these diversity numbers for TCRs [17] and BCRs [15].

All these diversity numbers are very large. Clearly, a single individual is only able to sample a tiny fraction of the potential pool of receptor sequences, with a total T-cells count of $\sim 3 \cdot 10^{11}$ in humans [1].

IV. THYMIC SELECTION AND HYPERMUTATIONS

After sequences have been generated by V(D)J recombination, they undergo an initial selection process. For T-cells, this takes place in the thymus and is called thymic selection. An analogous process occurs for B-cells. Sequences that bind too strongly to the host's own self-proteins, as well as those that bind too weakly to them, are discarded. By analyzing the in-frame naive receptor repertoire, one can study how the diversity of the repertoire is affected by this initial selection process. While the recombination diversity, $P_{\rm gen}(s)$, described the potential variability from the gene rearrangement process, this post-selection naive diversity, $P_{\rm sel}(s)$, describes the statistics of sequences actually found in the naive repertoire. It is still a potential diversity, as it refers to a statistical ensemble of receptors, rather than a finite set of receptors found in a given individual.

One can define a sequence-dependent selection factor $Q(s) = P_{\rm sel}(s)/P_{\rm gen}(s)$ quantifying how the distribution of sequences is affected by thymic selection. As before, sampling from $P_{\rm sel}(s)$ is impossible in practice because of the too large number of sequences, and models of the selection factor $Q(s)$ are needed. For example, it may take the factorized form

$$Q(s) = \prod_{i=1}^{L} q_{i;L}(a_i), \tag{6}$$

where $(a_1, a_2, \ldots, a_L)$ is the amino-acid sequence of the CDR3 region of length L, and the single-position factors $q_{i;L}(a)$ are inferred from the data using maximum likelihood. This model describes very well the statistics of naive and memory TCR β-chain sequences [43], α-chain sequences [44], and naive BCR heavy chain sequences [40]. The selection factors $Q(s)$ were shown to depend only on the amino-acid rather than nucleotide sequence, consistent with our hypothesis that selection acts on the protein product and its functional properties (folding, stability, binding, etc.). Although selection factors may vary significantly from individual to individual in the statistical sense, these differences are relatively small. In addition, models inferred from the memory and naive sequence repertoires were found to be similar, suggesting that the selection factors $Q(s)$ capture universal functional properties of the receptor proteins.

Diversity numbers can be estimated from the model of Eq. 6. The entropy of the post-selection distributions of receptor sequences, $P_{\rm sel}(s) = Q(s)P_{\rm gen}(s)$ are shown in green in Fig. 3A. The rank-frequency distribution and Hill diversities D_β of the post-selection ensemble of TCR β chain sequences are shown in green in Fig. 3B and C.

Diversity is reduced by selection from 47 to 38 bits for TCR β chains, from 30 to 26 bits for α chains, and from 70 to 58 bits for BCR heavy chains, corresponding to $D_1 \sim 3 \cdot 10^{11}$ for β chains, $D_1 \sim 7 \cdot 10^7$ for α chains (or a combined TCR diversity of $2 \cdot 10^{19}$ assuming independence between the two chains), and $D_1 \sim 3 \cdot 10^{17}$ for heavy chains. About 2 bits of this reduction are due to the removal of visibly nonfunctional sequences (out-of-frame or having stop codons). However, most of the diversity loss is caused by negative selection against sequences that were unlikely to be produced in the first place. Frequent sequences are enriched by the selection process, while rare ones are more likely to be removed. This enhancement of inequalities between sequences is the main source of entropy reduction by selection.

It should be noted that these estimate rely on an effective model (6), which may miss many important aspects of the selection process. In particular, negative selection, which prunes the repertoire of specific sequences that bind to self-antigens, is likely not accounted for by the model. This further diversity loss would be specific to each individual and its set of self-antigens, which depends on its HLA types. To assess whether all the aspects of selection that are not individual specific are well captured by Eq. 6, one can ask whether the Simpson index calculated with the model, $1/D_2$, is consistent with the observed repertoire overlap between distinct individuals, as it should if the two repertoires were drawn independently from the same distribution $P_{\rm sel}(s)$. Indeed the model and data showed good agreement [43], confirm-

ing that the model describes the statistics of sequences accurately.

Following their release into the periphery, cells undergo a somatic evolution process by which they divide, die or proliferate depending on the signals they receive. In the case of T cells, it is not clear how this evolution affects the potential naive diversity, as TCR β-chain sequences expressed by memory cells are statistically indistinguishable from naive ones [43]. In contrast, BCRs experience somatic hypermutations as B cells proliferate upon antigen recognition, during the process of affinity maturation. These hypermutations are stochastic but do not occur uniformly across the receptor, favoring instead sequence context dependent 'hotspots' [45, 46]. High-throughput repertoire sequencing now makes it possible to build predictive statistical models of hypermutations, by disentangling mutation from substitution rates using either synonymous mutants [47] or out-of-frame sequences [40, 48]. Out-of-frame sequences have a raw mutation rate ranging from a 5% to 10%, implying an additional 0.4 bits per nucleotide. This additional diversity is a huge boost if this estimate holds for the whole length of the receptor sequence. However, the increase in diversity due to hypermutations should depend on how long cells have been allowed to evolve. As affinity maturation consists of alternating cycles of mutation and selection, the effects of hypermutations on diversity cannot entirely be decoupled from selective pressures. The inference of selection during affinity maturation using repertoire sequencing is currently a very active field of study [23, 49–55].

V. REALIZED DIVERSITY

Thus far we have focused on the potential diversity of lymphocyte receptors. Its object is the probability that each receptor sequence has been generated, selected and, in the case of BCR, hypermutated into its final form. One can also study the realized diversity of receptor clonotypes actually present in a given individual at a given time. The relative frequency of clonotypes in an individual can vary greatly depending on the history of cell divisions and deaths, and is in general distinct from the probabilities $P_{\rm gen}$ and $P_{\rm sel}$ discussed so far. Measuring accurate clonotype frequencies relies on trustworthy counts made possible by unique molecular barcodes associated to original mRNA molecule [19–21] (with the caveat that cells may express variable amounts of mRNA molecules). One can build the rank-frequency relation as before, by ranking clonotypes in a given individual from most common to rarest. This relation can be measured for different phenotypes (naive or memory, CD4 or CD8), in different tissues or organs, or at different ages, to study the organisation and evolution of diversity.

In Fig. 4 we plot the rank-frequency relation for the unpartitioned TCR β-chain repertoires sampled from the blood of six individuals [44] and sequenced using unique molecular barcodes. A striking feature of these relations is that they seem to follow a power law, $f \propto 1/r^\alpha$, where f and r denote the clonotype frequency and rank, with exponent α ranging from 0.65 to 1, with a mean of 0.78. This observation is consistent with previous reports on zebrafish BCR [6, 25] or mouse TCR repertoires [5]. These power laws cannot be explained by a neutral model in which cells divide and die stochastically at a constant rate. Instead, they are consistent with models where each clone evolves under a fluctuating fitness shaped by its changing antigenic environment [56].

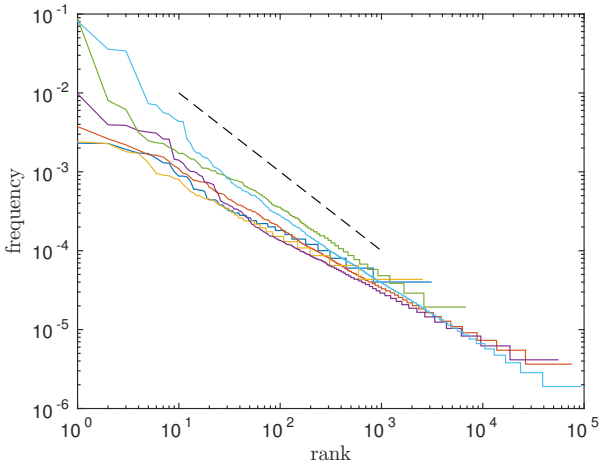

FIG. 4: Clonotype frequency vs. rank in the sequenced unpartitioned repertoires of six individuals from [44]. These relations are close to a power law with exponents ranging from -0.65 to -1. The dashed line shows a slope of -1.

Power-law frequency distributions make it challenging to estimate diversity measures D_β [32]. This difficulty can be understood by considering the geometric construction of diversities of Fig. 2: examining the rank-frequency curve of Fig. 4, no tangent of slope -1 can be easily defined. Mathematically, the normalization of the distribution strongly depends on the maximal rank, as $\sum_r 1/r^\alpha$ is a diverging series, meaning that the distribution is dominated by a very large number of very small clonotypes. This is particularly problematic as these rare clonotypes are not well captured by incomplete sampling.

Most past studies of repertoire diversity have actually focused on the hardest diversity measure to estimate in the face of these sampling issues, namely the species richness index D_0. By sequencing a subset of the repertoire with low-throughput techniques and extrapolating to the entire repertoire, Arstila and collaborators found a lower bound to the total size of the TCR repertoire of 10^6 distinct β chains, each pairing to 25 distinct α chains, *i.e.* $2.5{\cdot}10^7$ distinct TCRs [27]. This bound has since been revisited using high-throughput sequencing data, yielding the same order of magnitude of a few millions [2, 3].

In practice, most experiments are performed on samples of blood or tissues and do not sequence every single

cell. Even experiments using a whole tissue are subject to losses. The problem of species richness estimation from incomplete samples is not specific to lymphocyte repertoires and has been extensively discussed in ecology. A number of estimators of D_0, such as Chao1 [57], the abundance-based coverage estimator [58], or more recently DivE proposed in the context of TCRs [29], have been developed to address this issue. Another estimator using multiple samples, Chao2 [59], has recently been used to yield a lower bound of 10^8 distinct TCR β chains in humans [28]. All these estimators implictly assume that the distribution of frequencies is reasonably peaked, and may not be appropriate for broad distributions such as power laws.

To illustrate the inadequacy of most estimators to capture the true species richness of power-law distributed clone sizes, we numerically generated $D_0 = 10^7$ distinct clonotypes, and fixed their abundance to

$$C_r = (D_0/r)^\alpha, \tag{7}$$

where $r = 1, \ldots, D_0$ is the rank of the clonotype ordered by abundance, and $\alpha = 0.8$ to mimick the data of Fig. 4. We simulated a sample comprising 1% of the entire dataset, by drawing S_r, the size of clonotype of rank r in the sample, from a Poisson distribution of mean $C_r/100$. We calculated Chao1,

$$D_0 \approx D_0^{\text{raw}} + \frac{n_1^2}{2n_2}, \tag{8}$$

where D_0^{raw} is the number of sampled clonotypes ($S_r > 0$, n_1 is the number of singletons ($S_r = 1$), and n_2 the number of doubletons ($S_2 = 2$). This estimate gave $D_0 = 3 \cdot 10^6$ instead of the true value of 10^7. Dividing the dataset into 5 subsamples as in [28], and calculating Chao2 yields a similar estimate, $3.2 \cdot 10^6$. The reason for this underestimation is deep and does not depend much on the details of the estimator. When downsampling, one loses information about the rare clones, which dominate the species richness. Extrapolating their number from larger clones must rely on implicit or explicit assumptions about the clonal distribution, which are likely not satisfied by fat-tailed distributions such as power laws. It is therefore likely that most current estimates from high-throughput sequencing data are only lower bounds to the true species richness.

In fact, simple theoretical arguments based on thymic output estimates and neutral models of clonal evolution give upper bounds of 10^{10}-10^{11}[60, 61]. However, since we have argued that the power-law in the rank-frequency curve did not support the hypothesis of neutrality, it is legitimate to ask what species richness would be predicted from a power-law distribution of clone sizes. Assuming that the rank-size relation is given by Eq. 7, the average clonotype size reads:

$$\langle C \rangle = \frac{1}{D_0} \sum_{r=1}^{D_0} \left(\frac{D_0}{r} \right)^\alpha \approx D_0^{1-\alpha} \int_1^{D_0} \frac{1}{r^\alpha} = \frac{1}{1-\alpha}, \tag{9}$$

where we have approximated the sum by an integral, which is valid for large D_0. Plugging $\alpha = 0.8$ gives an average clone size of 5 cells, and hence a species richness $D_0 = 3 \cdot 10^{11}/5 \sim 10^{11}$ of the same order of magnitude as total number of T cells. Note however that this estimate is very sensitive to the value of α, as the average clone size becomes $\sim \ln(D_0)$ for $\alpha = 1$, and $\sim \zeta(\alpha) D_0^{\alpha-1}$ for $\alpha > 1$, where $\zeta(\alpha)$ is the Riemann zeta function.

Although the validity of the power law across the entire spectrum of clone sizes is a matter of debate, this example emphasizes the need for models to extrapolate the size distribution to the very rare clonotypes, the knowledge of which is essential for evaluating species richness.

VI. TOWARDS A FUNCTIONAL DIVERSITY

All the diversities discussed in this chapter apply to nucleotide sequences. These estimates demonstrate the potential of the adaptive immune system to generate a huge diversity of sequences, while identifying the biases of their generation and selection. However, they do not directly inform us about the functional diversity of the repertoire, defined as its capacity to recognize a wide variety of antigens. First of all, the binding properties of receptors are determined by their amino-acid sequences, the diversity of which is smaller due to the degeneracy of the genetic code. But more fundamentally, a given antigen can be recognized by many receptors — a phenomenon termed cross-reactivity or polyspecificity. Mason [62] argued that if not for cross-reactivity, an individual would need a repertoire as large as the number of antigens it can encounter, or $\sim 10^{15}$ for TCRs, which is well beyond the number of lymphocytes a human or a mouse can afford. Simple models can help estimate the minimal size of the functional repertoire [5, 63, 64]. Theoretical arguments also suggests that cross-reactivity gives a certain freedom in the identity and binding properties of the receptors, implying that two individuals experiencing similar antigenic environments need not share common receptors through the convergent evolution of their repertoires [65].

Quantifying the functional diversity of the repertoire is arduous because it requires to precisely characterize cross-reactivity by mapping the sequence of receptors to their binding properties. The identification of TCRs that bind to specific antigens using tetramer experiments in mouse [66] shows that a single antigen is bound by 20-200 out of $4 \cdot 10^7$ CD4+ T cells, *i.e* a fraction $5 \cdot 10^{-7}$-$5 \cdot 10^{-6}$ of the total population. Conversely, a single TCR can recognize many antigens. A lower bound of 10^6 has been reported for an autoimmune TCR from a human patient [67], but that number must be much larger ($> 5 \cdot 10^{-7} \times 10^{15} = 5 \cdot 10^8$) so that the TCR repertoire may cover the entire set of possible peptides.

Assessing cross-reactivity in a more quantitative and systematic way requires to massively measure the binding properties of a huge numbers of receptor-antigens pairs.

High-throughput mutational scans combining binding assays with next-generation sequencing technologies now make it possible to measure the binding properties of a single receptor against many peptides [68], or of many mutagenized receptors againt a single antigen [69]. Integrating these measurements into predicitve models of receptor-antigen binding would provide powerful tools for analysing lymphocyte repertoires. The diversity of receptor sequences could then be augmented by the more relevant diversity of antigens that can be recognized by them, with varying potencies and frequencies.

This work was supported in part by grant ERCStG n. 306312, and by the National Science Foundation under Grant No. NSF PHY11-25915 through the KITP where part of the work was done.

[1] Jenkins MK, Chu HH, McLachlan JB, Moon JJ (2009) On the composition of the preimmune repertoire of T cells specific for Peptide-major histocompatibility complex ligands. *Annu. Rev. Immunol.* 28:275–294.

[2] Robins HS, et al. (2009) Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* 114:4099–4107.

[3] Warren RL, 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.

[4] Hozumi N, Tonegawa S (1976) Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proc. Natl. Acad. Sci.* 73:3628–3632.

[5] Zarnitsyna VI, Evavold BD, Schoettle LN, Blattman JN, Antia R (2013) Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. *Front. Immunol.* 4:485.

[6] Weinstein Ja, Jiang N, White Ra, Fisher DS, Quake SR (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science* 324:807–810.

[7] Freeman JD, Warren RL, Webb JR, Nelson BH, Holt Ra (2009) Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res.* 19:1817–1824.

[8] Robins HS, et al. (2010) Overlap and effective size of the human CD8+ T cell receptor repertoire. *Sci. Transl. Med.* 2:47ra64.

[9] Benichou J, Ben-Hamo R, Louzoun Y, Efroni S (2012) Rep-Seq: Uncovering the immunological repertoire through next-generation sequencing. *Immunology* 135:183–191.

[10] Warren EH, Matsen Fa, Chou J (2013) High-throughput sequencing of B- and T-lymphocyte antigen receptors in hematology. *Blood* 122:19–22.

[11] Six A, et al. (2013) The past, present and future of immune repertoire biology - the rise of next-generation repertoire analysis. *Front. Immunol.* 4:413.

[12] Woodsworth DJ, Castellarin M, Holt Ra (2013) Sequence analysis of T-cell repertoires in health and disease. *Genome Med.* 5:98.

[13] Georgiou G, et al. (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* 32:158–68.

[14] Calis JJ, Rosenberg BR (2014) Characterizing immune repertoires by high throughput sequencing: strategies and applications. *Trends Immunol.* pp 1–10.

[15] Dekosky BJ, et al. (2014) In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat. Med.* 21:1–8.

[16] Turchaninova Ma, et al. (2013) Pairing of T-cell receptor chains via emulsion PCR. *Eur. J. Immunol.* 43:2507–2515.

[17] Howie B, et al. (2015) High-throughput pairing of T cell receptor a and b sequences. *Sci. Transl. Med.* 7:301ra131.

[18] Shugay M, et al. (2014) Towards error-free profiling of immune repertoires. *Nat. Methods* 11:653–5.

[19] Vollmers C, Sit RV, Weinstein Ja, Dekker CL, Quake SR (2013) Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc. Natl. Acad. Sci.* 110:13463–8.

[20] Egorov ES, et al. (2015) Quantitative Profiling of Immune Repertoires for Minor Lymphocyte Counts Using Unique Molecular Identifiers. *J. Immunol.* 194:6155–63.

[21] Best K, Oakes T, Heather JM, Shawe-Taylor J, Chain B (2015) Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. *Sci. Rep.* 5:14629.

[22] Greiff V, Miho E, Menzel U, Reddy ST (2015) Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires. *Trends Immunol.* 36:738–749.

[23] Yaari G, Kleinstein SH (2015) Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med.* 7:121.

[24] Greiff V, et al. (2015) A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med.* 7:49.

[25] Mora T, Walczak AM, Bialek W, Callan CG (2010) Maximum entropy models for antibody diversity. *Proc. Natl. Acad. Sci.* 107:5405–5410.

[26] Venturi V, Kedzierska K, Turner SJ, Doherty PC, Davenport MP (2007) Methods for comparing the diversity of samples of the T cell receptor repertoire. *J. Immunol. Methods* 321:182–195.

[27] Arstila TP, et al. (1999) A direct estimate of the human alphabeta T cell receptor diversity. *Science* 286:958–961.

[28] Qi Q, et al. (2014) Diversity and clonal selection in the human T-cell repertoire. *Proc. Natl. Acad. Sci.*

[29] Laydon DJ, et al. (2014) Quantification of HTLV-1 Clonality and TCR Diversity. *PLoS Comput. Biol.* 10:1–13.

[30] Rényi A (1961) On measures of entropy and information. *Entropy* 547:547–561.

[31] Hill AMO (1973) Diversity and Evenness : A Unifying Notation and Its Consequences. *Ecology* 54:427–432.

[32] Mora T, Walczak AM (2016) Renyi entropy, abundance distribution and the equivalence of ensembles. *arXiv* qbio:1603.05458.

[33] Janeway C, Murphy KP, Travers P, Walport M (2008) *Janeway's immunobiology* (Garland Science).

[34] Murugan A, Mora T, Walczak AM, Callan CG (2012)

Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl. Acad. Sci.* 109:16161–16166.

[35] Venturi V, et al. (2006) Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proc. Natl. Acad. Sci.* 103:18691–18696.

[36] Volpe JM, Cowell LG, Kepler TB (2006) SoDA: implementation of a 3D alignment algorithm for inference of antigen receptor recombinations. *Bioinformatics* 22:438–44.

[37] Gaëta BA, et al. (2007) iHMMune-align: Hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics* 23:1580–1587.

[38] Munshaw S, Kepler TB (2010) SoDA2: a Hidden Markov Model approach for identification of immunoglobulin rearrangements. *Bioinformatics* 26:867–72.

[39] Elhanati Y, Marcou Q, Mora T, Walczak AM (2016) repgenHMM: a dynamic programming tool to infer the rules of immune receptor generation from sequence data. *Bioinformatics* In press.

[40] Elhanati Y, et al. (2015) Inferring processes underlying B-cell repertoire diversity. *Philos Trans R Soc Lond, B, Biol Sci* 370:20140243.

[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] Davis MM, Bjorkman PJ (1988) T-cell antigen receptor genes and T-cell recognition. *Nature* 334:395–402.

[43] Elhanati Y, Murugan A, Callan CG, Mora T, Walczak AM (2014) Quantifying selection in immune receptor repertoires. *Proc. Natl. Acad. Sci.* 111:9875–9880.

[44] Pogorelyy MV, et al. (2016) Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. *arXiv* qbio:1–21.

[45] Shapiro GS, Aviszus K, Ikle D, Wysocki LJ (1999) Predicting regional mutability in antibody V genes based solely on di- and trinucleotide sequence composition. *J. Immunol.* 163:259–68.

[46] Cowell LG, Kepler TB (2000) The Nucleotide-Replacement Spectrum Under Somatic Hypermutation Exhibits Microsequence Dependence That Is Strand-Symmetric and Distinct from That Under Germline Mutation. *J. Immunol.* 164:1971–1976.

[47] Yaari G, et al. (2013) Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front. Immunol.* 4:358.

[48] Dunn-Walters DK, Dogan A, Boursier L, MacDonald CM, Spencer J (1998) Base-Specific Sequences That Bias Somatic Hypermutation Deduced by Analysis of Out-of-Frame Human IgVH Genes. *J. Immunol.* 160:2360–2364.

[49] Uduman M, et al. (2011) Detecting selection in immunoglobulin sequences. *Nucleic Acids Res.* 39:W499–W504.

[50] Yaari G, Uduman M, Kleinstein SH (2012) Quantifying selection in high-throughput Immunoglobulin sequencing data sets. *Nucleic Acids Res.* 40:e134.

[51] Kepler TB, et al. (2014) Reconstructing a B-cell Clonal Lineage. II. Mutation, Selection, and Affinity Maturation. *Front. Immunol.* 5.

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

[53] Uduman M, Shlomchik MJ, Vigneault F, Church GM, Kleinstein SH (2014) Integrating B cell lineage information into statistical tests for detecting selection in Ig sequences. *J. Immunol.* 192:867–74.

[54] Mccoy CO, et al. (2015) Quantifying evolutionary constraints on B-cell affinity maturation. *Philos Trans R Soc Lond, B, Biol Sci* 370:20140244.

[55] Yaari G, Benichou JIC, Heiden JAV, Kleinstein SH, Louzoun Y (2015) The mutation patterns in B-cell immunoglobulin receptors reflect the influence of selection acting at multiple time-scales. *Philos Trans R Soc Lond, B, Biol Sci* 370:20140242.

[56] Desponds J, Mora T, Walczak AM (2016) Fluctuating fitness shapes the clone-size distribution of immune repertoires. *Proc. Natl. Acad. Sci.* 113:274–9.

[57] Chao A (1984) Nonparametric estimation of the number of classes in a population. *Scand J Stat.* 11:265–270.

[58] Chao A, Lee SM (1992) Estimating the Number of Classes via Sanple Coverage. *J. Am. Stat. Assoc.* 87:210–217.

[59] Chao A, Bunge J (2002) Estimating the number of species in a stochastic abundance model. *Biometrics* 58:531–539.

[60] Kesmir C, Borghans J, de Boer RJ (2000) Diversity of Human T Cell Receptors. *Science* 288:1135.

[61] Lythe G, Callard RE, Hoare R, Molina-París C (2015) How many TCR clonotypes does a body maintain? *J. Theor. Biol.* 389:214–224.

[62] Mason D (1998) A very high level of crossreactivity is an essential feature of the T- cell receptor. *Immunol. Today* 19:395–404.

[63] Perelson AS, Oster GF (1979) Theoretical studies of clonal selection minimal antibody repertoire size and reliability of self non self discrimination. *J. Theor. Biol.* 81:645–670.

[64] de Boer RJ, Perelson AS (1993) How diverse should the immune system be? *Proc R Soc Lond, B, Biol Sci* 252:171.

[65] Mayer A, Balasubramanian V, Mora T, Walczak AM (2015) How a well-adapted immune system is organized. *Proc. Natl. Acad. Sci.* 112:5950–5955.

[66] Moon JJ, et al. (2007) Naive CD4+ T Cell Frequency Varies for Different Epitopes and Predicts Repertoire Diversity and Response Magnitude. *Immunity* 27:203–213.

[67] Wooldridge L, et al. (2012) A single autoimmune T cell receptor recognizes more than a million different peptides. *J. Biol. Chem.* 287:1168–1177.

[68] Birnbaum ME, et al. (2014) Deconstructing the peptide-MHC specificity of t cell recognition. *Cell* 157:1073–1087.

[69] Adams RM, Kinney JB, Mora T, Walczak AM (2016) Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *arXiv* qbio:1601.02160.