Bioinformatics, 35(17), 2019, 2974–2981
doi: 10.1093/bioinformatics/btz035
Advance Access Publication Date: 18 January 2019
Original Paper

Sequence analysis

OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs

Zachary Sethna[1], Yuval Elhanati[1], Curtis G. Callan Jr[1,2], Aleksandra M. Walczak[2,*,†] and Thierry Mora[2,*,†]

[1]Joseph Henry Laboratories, Princeton University, Princeton, NJ 08544, USA and [2]Laboratoire de physique de l'Ecole normale supérieure (PSL University), Centre national de la recherche scientifique, Sorbonne University, University Paris-Diderot, 75005 Paris, France

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.

Associate Editor: Bonnie Berger

Received on July 12, 2018; revised on November 10, 2018; editorial decision on January 9, 2019; accepted on January 13, 2019

Abstract

Motivation: High-throughput sequencing of large immune repertoires has enabled the development of methods to predict the probability of generation by V(D)J recombination of T- and B-cell receptors of any specific nucleotide sequence. These generation probabilities are very nonhomogeneous, ranging over 20 orders of magnitude in real repertoires. Since the function of a receptor really depends on its protein sequence, it is important to be able to predict this probability of generation at the amino acid level. However, brute-force summation over all the nucleotide sequences with the correct amino acid translation is computationally intractable. The purpose of this paper is to present a solution to this problem.

Results: We use dynamic programming to construct an efficient and flexible algorithm, called OLGA (Optimized Likelihood estimate of immunoGlobulin Amino-acid sequences), for calculating the probability of generating a given CDR3 amino acid sequence or motif, with or without V/J restriction, as a result of V(D)J recombination in B or T cells. We apply it to databases of epitope-specific T-cell receptors to evaluate the probability that a typical human subject will possess T cells responsive to specific disease-associated epitopes. The model prediction shows an excellent agreement with published data. We suggest that OLGA may be a useful tool to guide vaccine design.

Availability and implementation: Source code is available at https://github.com/zsethna/OLGA.

Contact: awalczak@lpt.ens.fr or tmora@lps.ens.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The ability of the adaptive immune system to recognize foreign peptides, while avoiding self peptides, depends crucially on the specificity of receptor-antigen binding and the diversity of the receptor repertoire. Immune repertoire sequencing (Repseq) of B- and T-cell receptors (BCR and TCR) (Heather *et al.*, 2017; Lindau and Robins, 2017; Six *et al.*, 2013; Woodsworth *et al.*, 2013) offers an efficient experimental tool to probe the diversity of full repertoires in healthy individuals (Freeman *et al.*, 2009; Howie *et al.*, 2015; Mora and Walczak, 2018; Pogorelyy *et al.*, 2017; Robins *et al.*, 2009, 2010; Weinstein *et al.*, 2009), in cohorts with specific conditions (DeWitt *et al.*, 2018; Emerson *et al.*, 2017; Faham *et al.*, 2017; Horns *et al.*, 2017; Jiang *et al.*, 2013; Komech *et al.*, 2018; Pogorelyy *et al.*, 2018a; Vollmers *et al.*, 2013) and evaluate the response to specific fluorescent MHC-multimers (Dash *et al.*, 2017; Glanville *et al.*, 2017). Recent work has shown that responding clonotypes often form disjoint clusters of similar amino acid sequences, which has lead to the identification of responsive amino acid motifs (Dash *et al.*, 2017; Glanville *et al.*, 2017). In order for these techniques to have practical applications in therapy and vaccine design, one needs a fast and efficient algorithm to evaluate which specific amino acid sequences and sequence motifs are likely to be generated and found in repertoires. We present a solution to this problem in the form of an algorithm and computational tool, called OLGA, which implements an exact computation of the generation probability of any BCR or TCR sequence (nucleotide or amino acid), or motif.

BCR and TCR are stochastically generated by choosing a germline genetic template in each of several cassettes of alternates [V, (D), or J] and then splicing them together with random nucleotide deletions and insertions at the junctions. Given a generative model, one can define the generation probability of any nucleotide sequence as the sum of the probabilities of all the generative events that can produce that sequence (Elhanati *et al.*, 2015, 2016; Marcou *et al.*, 2018; Murugan *et al.*, 2012). However, computing the generation probability of amino acid sequences by summing over all consistent nucleotide sequences is impractical: because of codon degeneracy, the number of nucleotide sequences to be summed grows exponentially with sequence length. OLGA is powered by an efficient dynamic programming method to exactly sum over generative events and obtain net probabilities of amino acid sequences and motifs.

We validate our algorithm by comparing its results and performance to Monte-Carlo sampling estimates. We present results using publicly available data for both TCR α (TRA, Pogorelyy *et al.*, 2017) and β (TRB, Emerson *et al.*, 2017) chains and BCR heavy chains (IGH, DeWitt *et al.*, 2016) of humans, and TRB of mice (Sethna *et al.*, 2017). We applied OLGA to a TCR database that catalogs the different CDR3 amino acid sequences responding to a variety of different epitopes associated with disease (Shugay *et al.*, 2018). We computed the generation probability of particular CDR3 amino acid sequences, as well as the net generation probability of all the TCR that respond to a particular epitope. Finally, we discuss OLGA's applications in vaccine design and other therapeutic contexts.

2 Materials and methods

2.1 Stochastic model of VDJ recombination

V(D)J recombination is a stochastic process involving several events (gene template selection, terminal deletions from the templates, random insertions at the junctions), each of which has a set of possible outcomes chosen according to a discrete probability distribution. The probability $P^{\text{rec}}_{\text{gen}}(E)$ of any generation event E, defined as a combination of the above-mentioned processes is, for the TRB locus:

$$\begin{aligned} P^{\text{rec}}_{\text{gen}}(E) &= P_{\text{V}}(V)P_{\text{DJ}}(D,J)P_{\text{delV}}(d_V|V)P_{\text{delJ}}(d_J|J) \\ &\times P_{\text{delD}}(d_D,d'_D|D)P_{\text{insVD}}(\ell_{\text{VD}})p_0(m_1)\left[\prod_{i=2}^{\ell_{VD}} S_{\text{VD}}(m_i|m_{i-1})\right] \\ &\times P_{\text{insDJ}}(\ell_{\text{DJ}})q_0(n_{\ell_{DJ}})\left[\prod_{i=1}^{\ell_{DJ}-1} S_{\text{DJ}}(n_i|n_{i+1})\right], \end{aligned} \tag{1}$$

where (V, D, J) identify the choices of gene templates, (d_V, d_D, d'_D, d_J) are the numbers of deletions at each end of the segments, and $(m_1,\ldots,m_{\ell_{VD}})$ and $(n_1,\ldots,n_{\ell_{DJ}})$ are the untemplated inserted nucleotide sequences at the VD and DJ junctions. These variables specify the recombination event E, and are drawn according to the probability distributions (P_{V}, P_{DJ}, P_{delV}, P_{delD}, P_{delJ}, P_{insVD}, P_{insDJ}, p_0, q_0, S_{VD}, S_{DJ}). The inserted segments are drawn according to a Markov process starting with the nucleotide distribution p_0 and with the transition matrix R, and running from the $5'$ side (left to right) for the VD segment, and from the $3'$ side (right to left) from the DJ segment. Similar models can be defined for the α chain or for BCR chains. Although here we describe the method for TRB only, it is also implemented for other chains in the software.

Since the same nucleotide sequence can be created by more than one specific recombination event, the generation probability of a nucleotide sequence is the sum of the probabilities of all possible events that generate the sequence: $P^{\text{nt}}_{\text{gen}}(\boldsymbol{\sigma}) = \sum_{\text{E}\to\boldsymbol{\sigma}\,\text{P}^{\text{rec}}_{\text{gen}}(\text{E}),}$ where the sum is over all recombination events E that produce the sequence $\boldsymbol{\sigma} = (\sigma_1,\ldots,\sigma_{\text{n}})$. The probability of generation of an amino acid sequence, $\mathbf{a} = (a_1,\ldots,a_L)$ is the sum of the probabilities of all nucleotide sequences that translate into the amino acid sequence:

$$P^{\text{aa}}_{\text{gen}}(a_1,\ldots,a_L) = \sum_{\boldsymbol{\sigma}\sim\mathbf{a}} P^{\text{nt}}_{\text{gen}}(\sigma_1,.,\sigma_{3L}) = \sum_{E\to\boldsymbol{\sigma}\sim\mathbf{a}} P^{\text{rec}}_{\text{gen}}(E), \tag{2}$$

where the $\sim$ sign indicates that $\boldsymbol{\sigma}$ translates into $\mathbf{a}$. We can generalize this approach to any scheme that groups nucleotide triplets, or codons, into arbitrary classes, which we still denote by $\boldsymbol{\sigma} \sim \mathbf{a}$. In the formulation above, these classes simply group together codons with the same translation according to the standard genetic code. In an example of generalization, all codons that code for amino acids with a common chemical property, e.g. hydrophobicity or charge, could be grouped into a single class. In that formulation, $(a_1,\ldots,a_L)$ would correspond to a sequence of symbols denoting that property. More generally, any grouping of amino acids can be chosen (including one where any amino acid is acceptable), and the partition can be position dependent. Thus, the generation probability of arbitrary 'motifs' can be queried. In the following, for ease of exposition, we restrict our attention to the case where $\mathbf{a}$ is an amino acid sequence.

2.2 Dynamic programming computation of the generation probability of amino acid sequences

We now give an overview of how OLGA computes Eq. 2 without performing the sum explicitly, using dynamic programming. Supplementary Figures S1 and S2 give a graphical overview of the method, and details of the method implementation can be found in Supplementary Sections I and II and in the code manual. Given the genomic nucleotide sequences of the possible gene templates, together with a specific model of the type described in Eq. 1, the algorithm computes the net probability of generating a recombined gene with a given CDR3 amino acid sequence under a given set of V and J gene choices.

Each recombination event implies an annotation of the CDR3 sequence, assigning a different origin to each nucleotide (V, N1, D, N2, or J, where N1 and N2 are the VD and DJ insertion segments, respectively) that parses the sequence into five contiguous segments (see schematic in Fig. 1). The principle of the method is to sum over the probabilities of all choices of nucleotides consistent with the known amino acid sequence, over the possible locations of the four boundaries (x_1, x_2, x_3 and x_4) between the five segments, and over the possible V, D and J genomic templates (Fig. 1). We do this in a recursive way using matrix operations by defining weights that accumulate the probabilities of events from the left of a position x (i.e. up to x), and weights that accumulate events from the right of x

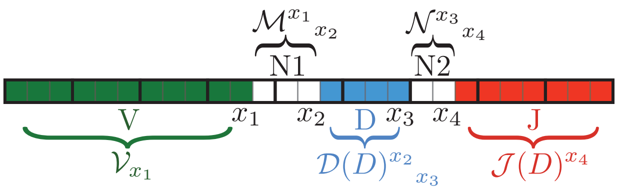
$\mathcal{M}^{x_1}{}_{x_2}$
$\mathcal{N}^{x_3}{}_{x_4}$
N1
N2
V
x_1
x_2
D
x_3
x_4
J
$\mathcal{V}_{x_1}$
$\mathcal{D}(D)^{x_2}{}_{x_3}$
$\mathcal{J}(D)^{x_4}$

Fig. 1. Partitioning a CDR3 sequence: boxes correspond to nucleotides and are indexed by integers. Each group of three boxes (identified by heavier boundary lines) corresponds to an amino acid. The nucleotide positions $x_1, \ldots, x_4$ identify the boundaries between different elements of the partition. The $\mathcal{V}$, $\mathcal{M}$, $\mathcal{D}(D)$, $\mathcal{N}$ and $\mathcal{J}(D)$ matrices define cumulated weights corresponding to each of the five elements

(i.e. from $x+1$ on). Specifically, we define the following index notation: $\mathcal{X}_x$ with a subscript called left index, accumulates weights from the left of x; $\mathcal{Y}^x$, with a superscript called right index, accumulates weights from the right of x; a matrix $\mathcal{X}^x{}_y$ corresponds to accumulated weights from position $x+1$ to y (as will be explained shortly, these objects may have suppressed nucleotide indices as well). $P^{\rm aa}_{\rm gen}$ is calculated recursively by matrix-like multiplications as:

$$P^{\rm aa}_{\rm gen}(\mathbf{a}) = \sum_{x_1,x_2,x_3,x_4} \mathcal{V}_{x_1}\mathcal{M}^{x_1}{}_{x_2}\sum_D \left[\mathcal{D}(D)^{x_2}{}_{x_3}\mathcal{N}^{x_3}{}_{x_4}\mathcal{J}(D)^{x_4}\right]. \qquad (3)$$

The vector $\mathcal{V}_x$ corresponds to a cumulated probability of the V segment finishing at position x; $\mathcal{M}^x{}_y$ is the probability of the VD insertion extending from $x+1$ to y; $\mathcal{N}^x{}_y$ is the same for DJ insertions; $\mathcal{D}^x{}_y$ corresponds to weights of the D segment extending from $x+1$ to y, conditioned on the D germline choice being D; $\mathcal{J}^x(D)$ gives the weight of J segments starting at position $x+1$ conditioned on the D germline being D. This D dependency is necessary to account for the dependence between the D and J germline segment choices (Murugan *et al.*, 2012). All the defined vectors and matrices depend implicitly on the amino acid sequence $(a_1, \ldots, a_L)$, but we leave this dependency implicit to avoid making the notation too cumbersome.

Because we are dealing with amino acid sequences encoded by triplet nucleotide codons, we need to keep track of the identity of the nucleotide at the beginning or the end of a codon. Depending on the position of the index x in the codon, the objects defined above may be vectors of size four (or 4×4 matrices) in the suppressed nucleotide index. We use conventions that depend on whether we are considering left or right indices, as follows.

If x is a multiple of three, i.e. $x = 0\ (\mathrm{mod}\ 3)$, then we do not keep nucleotide information and both $\mathcal{X}_x$ and $\mathcal{Y}^x$ are scalars (whether x is a left or a right index). If $x = 1\ (\mathrm{mod}\ 3)$, then $\mathcal{X}_x$ must be interpreted as a row vector of four numbers, $\mathcal{X}_x(\sigma)$, $\sigma = A, T, G, C$, corresponding to the cumulated probability weight that the nucleotide at position x (first position of the codon) takes value σ. If $x = 2\ (\mathrm{mod}\ 3)$, then $\mathcal{X}_x$ is also a row vector of four numbers, $\mathcal{X}_x(\sigma)$, but with a different interpretation: it corresponds to the cumulated probability up to position x, with the additional constraint that the nucleotide at position $x+1$ (the last position in the codon) *can* take value σ (the value is 0 otherwise). For right indices, the interpretation is reversed and the entries are column vectors: when $x = 1\ (\mathrm{mod}\ 3)$ the $\mathcal{Y}^x$ is a column vector containing the cumulated weights from $x+1$ onwards, with the constraint that the nucleotide at x *can* be σ, and when $x = 2\ (\mathrm{mod}\ 3)$, it is the probability weight that the nucleotide at position $x+1$ *is* σ. Generalizing to matrices, $\mathcal{X}^x{}_y$ is a 4×4, 4×1, 1×4, or 1×1 matrix depending on whether the x and y positions are multiples of 3 or not, with the same rules as for vectors for each type of index.

Entries with left indices are interpreted as row vectors, and entries with right indices as column vectors. Thus, in Eq. 3 contractions between left and right indices correspond to dot products over the four nucleotides when the index is not a multiple of three, and simply a product of scalars when it is.

The entries of the matrices corresponding to the germline segments, $\mathcal{V}$, $\mathcal{D}(D)$ and $\mathcal{J}(D)$, can be calculated by simply summing over the probabilities of different germline nucleotide segments compatible with the amino acid sequence $(a_1, \ldots, a_L)$ with conditions on deletions to achieve the required segment length. For instance, the $\mathcal{V}$ matrix elements are given by:

$$\begin{aligned}\mathcal{V}_x(\sigma) &= \sum_V P_{\rm V}(V)P_{\rm delV}(l_V - x)\mathbb{I}(s^V_x = \sigma)\mathbb{I}(\mathbf{s}^V_{1:x} \sim \mathbf{a}_{1:i}) \text{ if } u = 1\\ \mathcal{V}_x(\sigma) &= \sum_V P_{\rm V}(V)P_{\rm delV}(l_V - x)\mathbb{I}((\mathbf{s}^V_{1:x}, \sigma) \sim \mathbf{a}_{1:i}) \text{ if } u = 2,\\ \mathcal{V}_x &= \sum_V P_{\rm V}(V)P_{\rm delV}(l_V - x)\mathbb{I}(\mathbf{s}^V_{1:x} \sim \mathbf{a}_{1:i}) \text{ if } u = 3,\end{aligned} \qquad (4)$$

where $x = 3(i-1)+u$, i.e. x is the uth nucleotide of the ith codon, $\mathbf{s}^V$ the sequence of the V germline gene, and $\mathbb{I}$ the indicator function. The $\sim$ sign is generalized to incomplete codons so that it returns a true value if there exists a codon completion that agrees with the motif $\mathbf{a}$. Detailed formulas for the other segments are derived using the same principles and are given in the SI Appendix. The sums in Eq. 4 (and equivalent expressions for J) can be restricted to particular germline genes to compute the generation probability of particular VJ-CDR3 combinations.

The entries of the insertion segment N1 are calculated using the following formula:

$$\mathcal{M}^x{}_{\rm y} = P_{\rm insVD}(y-x)L^u_{a_i}T_{a_{i+1}}\ldots T_{a_{j-1}}R^v_{a_j}, \qquad (5)$$

with $y = 3(j-1)+v$ (and $x = 3(i-1)+u$ as in Eq. 4). The transfer matrix

$$T_a(\tau,\sigma) = \sum_{(n_1,n_2,\sigma)\sim a} S_{\rm VD}(\sigma|n_2)S_{\rm VD}(n_2|n_1)S_{\rm VD}(n_1|\tau) \qquad (6)$$

corresponds to the probability of inserting a codon coding for a and ending with nucleotide σ, knowing that the previous codon ended with nucleotide τ. L^u_a and R^v_a are vectors or matrices with different definitions depending on the values of x and y modulo 3, corresponding to the probabilities of inserting incomplete codons on the left and right ends of the insertion segment. Eq. 5 is only valid for $j > i$, but similar formulas describe the case $i = j$. The precise definitions of L and R, the $i = j$ case, and the formulas for $\mathcal{N}$ and the N2 insertion segment, which is exactly equivalent, are all given in detail in the SI Appendix.

The matrix product of Eq. 5 can be calculated recursively, requiring only 4×4 matrix multiplications. Thus, all $\mathcal{M}^x{}_{\rm y}$ elements can be calculated in $\mathcal{O}(L^2)$ operations, instead of the exponential time that would be required using brute-force summation over nucleotides in degenerate codons. Finally, since the sums of Eq. 3 can also be done recursively through $L\times L$ matrix operations, the whole procedure has $\mathcal{O}(L^2)$ computational complexity.

3 Results

3.1 Method validation

To verify the correctness of the OLGA code, we compared its predictions for generation probabilities to those estimated by Monte Carlo (MC) sequence generation (Pogorelyy *et al.*, 2018a). MC estimation is done by drawing events from a given generative model, binning according to the resulting CDR3 amino acid sequence, and normalizing by the total number of recombination events. The scatter plot of the estimated generation probabilities for these sequences against the values predicted by OLGA gives a direct test of the algorithm. As MC estimation is susceptible to Poisson sampling noise, it is important to ensure that enough events are drawn to accurately assess the generative probabilities of individual CDR3 sequences. For this reason, we made the comparison using a generative model inferred from a mouse, rather than human, T cell repertoire, because of the significantly lower entropy of mouse repertoires (Sethna *et al.*, 2017). The specific model was inferred by IGoR (Marcou *et al.*, 2018) using ~70 000 out-of-frame TRB sequences from a mature mouse thymus. MC estimation was done by generating 5×10^{11} recombination events, from which the first 10^6 unique CDR3 amino acid sequences are counted to serve as a sample for the comparison. This procedure provided good sequence coverage, with $> 98\%$ of sequences generated at least twice and $> 95\%$ of sequences generated at least 10 times. As Figure 2 shows for mouse TRB (see Supplementary Fig. S3 for human TRA), MC estimation and OLGA calculation are in agreement (up to Poisson noise in the MC estimate). The Kullback-Leibler divergence between the two distributions, a formal measure of their agreement, is a mere 4.82×10^{-7} bits.

3.2 Comparison of performance with existing methods

We compared the performance of OLGA to other methods. Direct calculation of amino acid sequence generation probability using OLGA is orders of magnitude faster than the two possible alternative methods: MC estimation (as described above), or exhaustive enumeration of the generative events giving rise to a given amino acid sequence. OLGA took six CPU hrs to compute the generation probabilities of the 10^6 amino acid sequences, i.e. 47 seqs/CPU/sec for mouse TRB (see Supplementary Section III and Supplementary Table S1 for runtimes of other loci). By comparison, MC estimation required 4313 CPU hrs. The scaling for the MC estimation does not depend on the number of queried sequences, but instead is determined by the number of recombinations needed to control the Poisson noise, which scales inversely with generation probability. In practice, to determine the generation probability of a typical sequence (which can be as low 10^{-20}, see Fig. 3 and below), one needs to generate very large datasets, and thus the generation probability of many sequences cannot be calculated by the MC method.

Alternatively, one could list all possible nucleotide sequences that translate to a particular amino acid CDR3 and sum the generation probabilities of each nucleotide sequence, using the IGoR algorithm (Marcou *et al.*, 2018). Each amino acid sequence in the mouse validation sample is, on average, coded for by 1.84 billion nucleotide sequences (and much more for human TRB). Since IGoR computes generation probabilities of nucleotide sequences at the rate of ~60 seqs/CPU/sec, it would take ~8500 CPU hrs to compute the generation probability of a *single* amino acid sequence. A systematic comparison of OLGA with IGoR (Supplementary Fig. S4) and MC estimation (Supplementary Figs S4 and S5) as a function of the number of analyzed sequences and their CDR3 lengths shows that OLGA is faster than both other methods for all practical purposes (see Supplementary Section IV for details).

3.3 Distribution of generation probabilities and diversity

V(D)J recombination produces very diverse repertoires of nucleotide sequences, with a very broad distribution of generation probabilities spanning up to 20 orders of magnitude (Elhanati *et al.*, 2015; Murugan *et al.*, 2012). This distribution gives a comprehensive picture of the diversity of the process, and can be used to recapitulate many classical diversity measures (Mora and Walczak, 2018), and to predict the overlap between the repertoires of different individuals (Elhanati *et al.*, 2018). In particular, the opposite of the mean logarithm of the generation probability, $-\langle \log_2 P_{\rm gen} \rangle$, is equal to the entropy of the process. While previous work focused on nucleotide sequence generation, OLGA allows us to compute this distribution for amino acid sequences.

Figure 3 shows the distribution of $P^{\rm aa}_{\rm gen}$ for four loci: human and mouse TRB, human TRA and human IGH, and compares it to the distributions of nucleotide sequence generation probabilities, $P^{\rm nt}_{\rm gen}$, and recombination event probabilities, $P^{\rm rec}_{\rm gen}$. While all these datasets are based on DNA RepSeq, we checked that the generation probability distribution was robust to the choice of protocol by computing the TRB distribution for independent datasets generated by

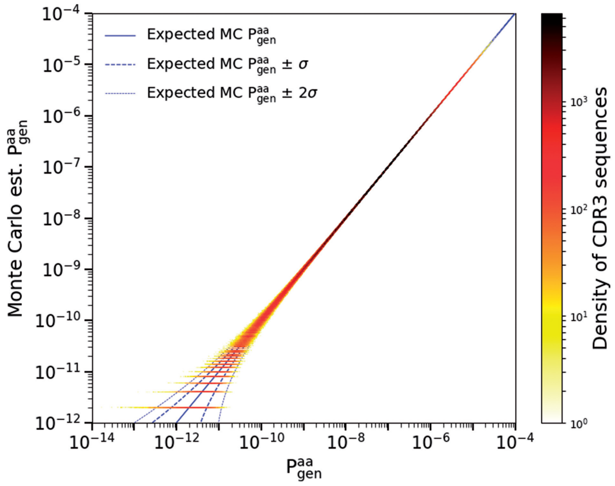

Fig. 2. Monte Carlo estimate of the generation probability of amino acid CDR3 sequences, $P^{\rm aa}_{\rm gen}$, versus OLGA's predictions (mouse TRB). The horizontal lines at the lower left of the plot represent CDR3s that were generated once, twice, etc., in the MC sample. The one- and two-sigma curves display the deviations from exact equality between simulated and computed P_{gen} to be expected on the basis of Poisson statistics

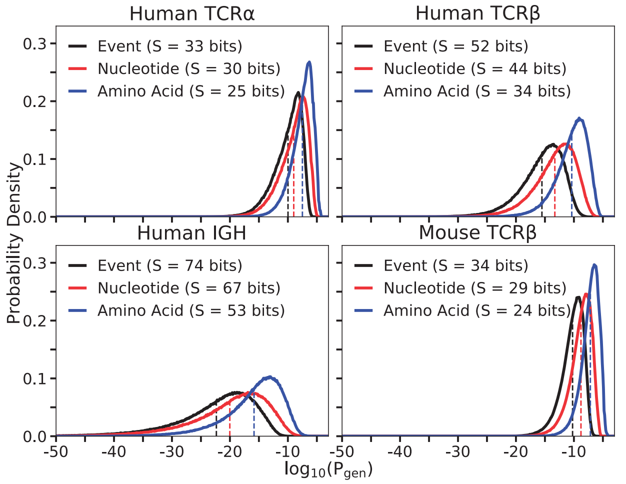

Fig. 3. Distributions of probabilities of recombination events ($P^{\rm rec}_{\rm gen}$), nucleotide CDR3 sequences ($P^{\rm nt}_{\rm gen}$) and CDR3 amino acid sequences ($P^{\rm aa}_{\rm gen}$) in different contexts. Each curve is determined by Monte Carlo sampling of 10^6 productive sequences for the indicated locus, and computing its generation probabilities at the three different levels. Entropies in bits (S) are, up to a ln(2)/ln(10) factor, the negative of the mean of each distributions, indicated by dotted lines

RNA RepSeq (Sims *et al.*, 2016; Wang *et al.*, 2010; Wu *et al.*, 2018) (Supplementary Figs S6 and S7, and Supplementary Section V). The generation models used here and elsewhere in this paper were taken from Marcou *et al.* (2018), except for the human TRB model which was relearned using IGoR from one individual in Emerson *et al.* (2017) as a check. Going from recombination events to nucleotide sequences to amino acid sequences leads to substantial shifts in the distribution, and corresponding drops in entropies, as the distribution is progressively coarse-grained. Higher generation probability of a given receptor sequence leads to higher chance of finding it in any given individual. Generation probabilities may be constrasted to the scale set by the inverse of the number of independent recombination events [estimated between 10^8 (Qi *et al.*, 2014) and 10^{10} (Lythe *et al.*, 2016) for human TCR]. Generation probabilities above this limit (10^{-10} to 10^{-8} for human TCR) can be considered 'large' as the corresponding receptor will almost surely exist in each individual (Elhanati *et al.*, 2018). Another relevant scale to distinguish small from large generation probabilities is given by their geometric mean (dashed lines in Fig. 3).

3.4 Cross-species generation probabilities

While distinct species differ in their generation mechanisms, they may yet be able to generate the same CDR3s. Using OLGA, we computed the probabilities of producing human TRB CDR3s by the mouse recombination model, and vice versa (details in Supplementary Section VI). An impressive 72.6% of human CDR3s can theoretically be produced by mice, and 100% of mouse CDR3s can be produced by humans. While cross-species generation probabilities are lower than intra-species ones (Supplementary Fig. S8), they are correlated (Supplementary Fig. S9). These results suggest that CDR3s observed in the repertoires of humanized mouse models of human diseases could be relevant for predicting their presence in human repertoires as well. OLGA allows for evaluating this potential, and could be used to inform clinical trials.

3.5 Generation probability of specific TCR

We can use OLGA to assess the total fraction of the generated repertoire that is specific to any given epitope, simply by summing the generation probabilities of all TRB sequences known to bind specifically to that epitope:

$$P_{\rm gen}^{\rm func}({\rm epitope}) = \sum_{\mathbf{a}|{\rm epitope}} P_{\rm gen}^{\rm aa}(\mathbf{a}), \qquad (7)$$

where '**a**|epitope' means that the amino acid sequence **a** recognizes the epitope. Many experiments, based e.g. on multimer sorting assays (Dash *et al.*, 2017; Glanville *et al.*, 2017) or T-cell culture assays, have established lists of epitope-specific TCR sequences for a number of disease-related epitopes. We used the VDJdb database (Shugay *et al.*, 2018), which aggregates such experiments, to compute $P_{\rm gen}^{\rm func}$ of all TRB known to be reactive against several epitopes. In Figure 4 we show results for four epitopes associated with Hepatitis C, and five epitopes associated with Influenza A. The net fraction of the repertoire specific to these epitopes (10^{-7} to 10^{-4}) is large in the sense defined above, meaning that any individual is likely to have many copies of reactive T cells in their naive repertoire.

The presence of any specific TCR in the repertoire will be affected by the recombination probability of both its α and β chains, and also by function-dependent selective pressures. Assessing accurately the fraction of reactive TCRs in the blood is beyond the scope of this method. However, it is still interesting to ask whether epitope-specific TRB sequences had higher generation probabilities than regular sequences, either because of observational biases, or because the immune system might have evolved to make them more likely to be produced. To answer that question, we display in Figure 5 the $P_{\rm gen}^{\rm aa}$ distribution of the sequences listed in VDJdb that are specific to any epitope of each of six commonly studied viruses. For comparison we plot the $P_{\rm gen}^{\rm aa}$ distribution of the full TRB sequence repertoire of a healthy donor [data taken from Emerson *et al.* (2017)].

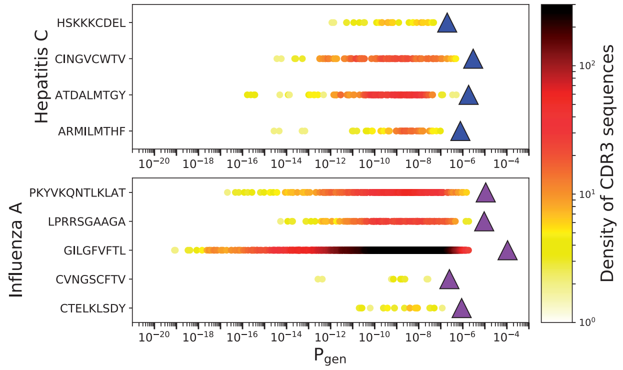

Fig. 4. Generation probabilities of human CDR3s that respond to hepatitis C and influenza A epitopes. $P_{\rm gen}^{\rm aa}$ of sequences that respond to an epitope are plotted as circles (color encodes density of the points). The fraction of the repertoire specific to each epitope ($P_{\rm gen}^{\rm func}$ as defined in Eq. 7) is obtained as the sum of the $P_{\rm gen}^{\rm aa}$ for each of the corresponding sequences (values plotted as triangles) (Color version of this figure is available at *Bioinformatics* online.)

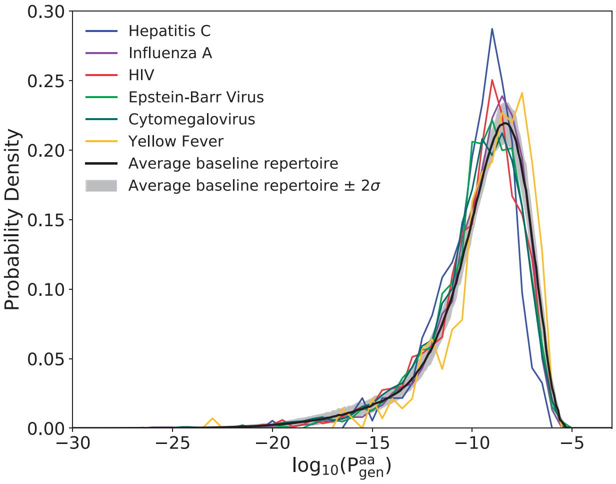

Fig. 5. Distributions of TRB generation probabilities $P_{\rm gen}^{\rm aa}$ for sequences in the VDJdb database that bind to any epitopes of six different viruses (colored curves). For comparison, we plot (black curve) the same distribution for the unsorted TRB repertoire of a typical healthy subject; the 2σ variance represents biological variability across multiple individuals [data from Emerson *et al.* (2017)] (Color version of this figure is available at *Bioinformatics* online.)

The viral distributions are very similar to each other, and also to the healthy repertoire background, meaning that the ability of a CDR3 to respond to a particular disease epitope is not strongly correlated with its generation probability. To see whether this result was confirmed in the case of a real infection, we repeated the same analysis on TRB RepSeq data from T-cells responding to three different types of pathogens (fungus, bacteria and toxin) (Becattini *et al.*, 2015). Consistently, we found that their distribution of generation probability was identical to that of naive sequences (Supplementary Fig. S10 and Supplementary Section VII).

3.6 Model accurately predicts the frequencies of sequences and of groups of specific sequences

To compare OLGA's predictions with sequence occurrence frequencies in real data, we used the aggregated TRB repertoire of 658 human subjects described in Emerson *et al.* (2017) as a test resource. More specifically, we measured the frequencies in this large dataset of the specific CDR3 sequences contained in the VDJdb database (Shugay *et al.*, 2018), and compared them to the values assigned by OLGA. When measuring frequencies we discarded read count information, recording only the presence or absence of nucleotide sequences in each individual in order to eliminate effects of clonal expansion and PCR amplification bias, averaging over the 648 individuals in the Emerson *et al.* (2017) dataset to get reliable estimates of frequencies. Each sequence in the VDJdb database is displayed as a dot in Figure 6, and the resulting distribution shows a strong correspondence between mean frequency in the large dataset and the predicted $P_{\mathrm{gen}}^{\mathrm{aa}}$ of that sequence.

We then measured the fraction of CDR3s in the aggregated repertoire that is specific to epitopes associated with six viruses (using lists of specific sequences in VDJdb), and compared it to OLGA's prediction, $P_{\mathrm{gen}}^{\mathrm{func}}$. The agreement was again excellent (triangles in Fig. 6). Again we observe that most epitope-specific sequence groups have large enough frequencies to be found in any individual. Thus, the model can be used to predict the size of repertoire subsets specific to any epitope, as long as specificity data are available for this epitope.

3.7 Generation probability of sequence motifs

OLGA can also compute the generation probability of any sequence motif, encoded by a string of multiple choices of amino acids. We apply this feature to calculate the net frequency of epitope-specific motifs, and of motifs that define the TRA sequence of invariant T-cells.

T-cell sequences that can bind a given epitope are often closely related to each other, and this similarity can sometimes be partially captured by sequence motifs. We evaluated the probabilities of motifs derived from a recent study of CDR3 sequence specificity to a variety of epitopes (Dash *et al.*, 2017). We took two motifs corresponding to TRA and TRB VJ-CDR3 combinations of TCRs that are known to bind the Epstein-Barr virus HLA-A*0201-BMLF$_{1280}$ (BMLF) and the influenza virus HLA-A*0201-M$_{158}$ (M1) epitopes. The motifs and generation probabilities are reported in Table 1.

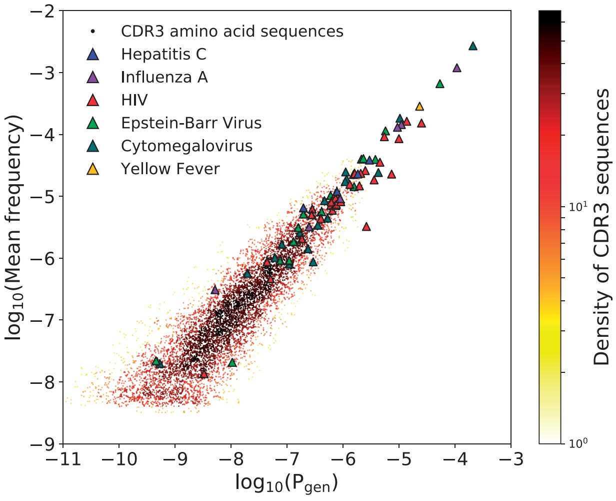

Fig. 6. Mean occurrence frequencies across a collection of 658 human samples of all CDR3 sequences in the VDJdb database, plotted against their computed $P_{\mathrm{gen}}^{\mathrm{aa}}$ (dots, colored by their density in the plot). Also, the net occurrence frequency in the VDJdb database of epitope-related collections of sequences, plotted against their computed $P_{\mathrm{gen}}^{\mathrm{func}}$ (triangles, colored to identify the virus the epitope belongs to) (Color version of this figure is available at *Bioinformatics* online.)

As a second application, we estimated the probabilities of generating a TRA chain corresponding to one of the motifs associated with Mucosal associated invariant T cells (MAIT) and invariant natural killer T cells (iNKT). The motifs, which were collected from Gherardin *et al.* (2016), and their probabilities are shown in Table 2. The relatively high values for these motifs imply that these invariant chains are generated with high frequency in the primary repertoire and shared by all individuals, confirming the conclusions of Venturi *et al.* (2013).

4 Discussion

Because the composition of the immune repertoire results from a stochastic process, the frequency with which distinct T- and B-cell receptors are generated is a quantity of primary interest. This frequency is computationally difficult to evaluate because each amino acid sequence can be created by a very large number of recombination events. Our tool overcomes that challenge with dynamic programming, allowing it to process ~50 sequences per second on a single CPU. In its current state OLGA can compute the probabilities of CDR3 sequences and motifs, with or without V/J restriction, of four chain loci (human and mouse TRB, human TRA and human IGH), but the list can readily be expanded by learning recombination models for other loci and species using IGoR (Marcou *et al.*, 2018) which shares the same model format. Obvious additions include the light chains of BCR (Toledano *et al.*, 2018), and more mouse models. While the algorithm evaluates the probability of single chains, recent analyses show that chain pairing in TCR is close to independent (Dupic *et al.*, 2018; Grigaityte *et al.*, 2017). The probability of generating a whole TCR receptor can thus be computed by taking the product over the two chains.

Table 1. Epitope-specific TCR motifs for the Epstein-Barr virus HLA-A*0201-BMLF$_{1280}$ (BMLF) and influenza virus HLA-A*0201-M$_{158}$ (M1) epitopes from Dash *et al.* (2017), and their generation probabilities

epitope: chain: V/J	CDR3 motif	P_{gen}
BMLF: α: 5/31	CAXD[NSDA]NARLMF	$1.8\cdot10^{-7}$
BMLF: β: 20-1/1-2, 1-3	CSARDX[TV]GNX{0,}	$5.1\cdot10^{-7}$
M1: α: 27/42	CAXGGSQGNLIF	$2.2\cdot10^{-5}$
M1: β: 19/all	CASSXR[SA][STAG]X[ET]Q[YF]F	$1.7\cdot10^{-6}$

Note: Each motif was associated with specific V/J gene choices. In the motifs we use the conventions: X, any one amino acid; [A.B], any one of the listed amino acids; X{0,}, arbitrary amino acid string.

Table 2. Generation probabilities of motifs corresponding to invariant T cell (iNKT and MAIT cells) TRA chain, assembled from sequence in Gherardin *et al.* (2016)

Type	V/J	CDR3 motif	P_{gen}
iNKT	10/18	CVVSDRGSTLGRLYF	$1.26\cdot10^{-6}$
MAIT	1-2/33	CAV[KSM]DSNYQLI[WF]	$1.79\cdot10^{-5}$
MAIT	1-2/12	CAVMDSSYKLIF	$4.71\cdot10^{-6}$
MAIT	1-2/20	CAVSDNDYKLSF	$3.11\cdot10^{-7}$

OLGA can be used to compute baseline receptor frequencies and to identify outlying sequences in repertoire sequencing datasets. In Elhanati *et al.* (2018) we used it to shed light on the question of public repertoires—composed of sequences shared by many individuals—and predict quantitatively its origin by convergent recombination (Madi *et al.*, 2014, 2017; Venturi *et al.*, 2008). Deviations from the baseline expectancy have been used to identify disease-associated TCR from cohorts of patients (Emerson *et al.*, 2017; Faham *et al.*, 2017; Fuchs *et al.*, 2017; Seay *et al.*, 2016; Zhao *et al.*, 2016), and to identify clusters of reactive TCRs from tetramer experiments (Glanville *et al.*, 2017) and vaccination studies (Pogorelyy *et al.*, 2018b). Such estimates could be made faster and more reliable by OLGA, especially for rare sequences, and without the need for a negative control cohort (Pogorelyy *et al.*, 2018a). In the future, OLGA could be useful in vaccine and therapy design by focusing attention on clonotypes that are likely to be present in every individual.

We applied OLGA to an experimental database of TCR responding to a variety of disease-associated epitopes. These selected TCR do not differ in their generation probabilities from those of random TCR found in the blood of healthy donors. However, some viral epitopes bind a much larger fraction of the repertoire than others. This observation has potentially important consequences for vaccine design. Since vaccine epitopes stimulate TCR in a pre-existing repertoire, epitopes targeting receptor sequences that are more likely to be generated will have a higher success rate in a wider range of individuals. OLGA can be used to identify such epitopes by computing their specific repertoire fractions, $P_{\text{gen}}^{\text{func}}$. While our examples are restricted to TCR, OLGA can also handle BCR and could be used to compute the generation probabilities of BCR precursors of highly reactive or broadly neutralizing antibodies, and thus guide vaccine design in that case as well. The algorithm does not yet handle hypermutations, and extending it to include them would be a useful development.

Funding

The work of TM and AMW was supported in part by grant European Research Council COG n. 724208. The work of ZS and CC was supported in part by NSF grant PHY-1607612. The work of CC was also supported in part by NSF grant PHY-1734030. The work of YE was supported by a fellowship from the V Foundation. The authors declare no conflicts of interest.

Conflict of Interest: none declared.

References

Becattini,S. *et al.* (2015) Functional heterogeneity of human memory cd4+ t cell clones primed by pathogens or vaccines. *Science*, **347**, 400–406.

Dash,P. *et al.* (2017) Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, **547**, 89–93.

DeWitt,W.S. *et al.* (2016) A public database of memory and naive B-cell receptor sequences. *PLoS One*, **11**, e0160853.

DeWitt,W.S. *et al.* (2018) Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *eLife*, **7**, e38358.

Dupic,T. *et al.* (2018) Genesis of the $\alpha\beta$ T-cell receptor. *arXiv*: 1806.11030.

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

Elhanati,Y. *et al.* (2016) repgenhmm: a dynamic programming tool to infer the rules of immune receptor generation from sequence data. *Bioinformatics*, **32**, 1943–1951.

Elhanati,Y. *et al.* (2018) Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunol. Rev.*, **284**, 167–179.

Emerson,R.O. *et al.* (2017) Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.*, **49**, 659–665.

Faham,M. *et al.* (2017) Discovery of T cell receptor β motifs specific to HLA-B27-positive ankylosing spondylitis by deep repertoire sequence analysis. *Arthritis Rheumatol.*, **69**, 774–784.

Freeman,J.D. *et al.* (2009) Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res.*, **19**, 1817–1824.

Fuchs,Y.F. *et al.* (2017) CD8+T cells specific for the islet autoantigen IGRP are restricted in their T cell receptor chain usage. *Sci. Rep.*, **7**, 1–10.

Gherardin,N.A. *et al.* (2016) Diversity of T cells restricted by the MHC class I-related molecule MR1 facilitates differential antigen recognition. *Immunity*, **44**, 32–45.

Glanville,J. *et al.* (2017) Identifying specificity groups in the T cell receptor repertoire. *Nature*, **547**, 94–98.

Grigaityte,K. *et al.* (2017) Single-cell sequencing reveals $\alpha\beta$ chain pairing shapes the T cell repertoire. *bioRxiv: 213462*.

Heather,J.M. *et al.* (2017) High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. *Brief. Bioinform.*, **19**, 554–565.

Horns,F. *et al.* (2017) Signatures of selection in the human antibody repertoire: selective sweeps, competing subclones, and neutral drift. *bioRxiv*, doi.org/10.1101/145052.

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

Jiang,N. *et al.* (2013) Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci. Transl. Med.*, **5**, 171ra19.

Komech,E. *et al.* (2018) CD8+ T cells with characteristic TCR beta motif are detected in blood and expanded in synovial fluid of ankylosing spondylitis patients. *Rheumatology (Oxford, England)*, **57**, 1097–1104.

Lindau,P. and Robins,H.S. (2017) Advances and applications of immune receptor sequencing in systems immunology. *Curr. Opin. Syst. Biol.* **1**, 62–68.

Lythe,G. *et al.* (2016) How many TCR clonotypes does a body maintain? *J. Theor. Biol.*, **389**, 214–224.

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.

Madi,A. *et al.* (2017) T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences. *eLife*, **6**, e22057.

Marcou,Q. *et al.* (2018) High-throughput immune repertoire analysis with IGoR. *Nat. Commun.*, **9**, 561.

Mora,T. and Walczak,A. (2018) Quantifying lymphocyte receptor diversity. In: Das,J.D. and Jayaprakash,C. (eds.) *Systems Immunology: An Introduction to Modeling Methods for Scientists*. CRC Press, pp. 185–199. https://www.crcpress.com/Systems-Immunology-An-Introduction-to-Modeling-Methods-for-Scientists/Das-Jayaprakash/p/book/9781498717403.

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

Pogorelyy,M.V. *et al.* (2017) Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. *PLoS Comput. Biol.*, **13**, e1005572.

Pogorelyy,M.V. *et al.* (2018a) Method for identification of condition-associated public antigen receptor sequences. *Elife*, **7**, 1–13.

Pogorelyy,M.V. *et al.* (2018b) Precise tracking of vaccine-responding T-cell clones reveals convergent and personalized response in identical twins. *Proc. Natl Acad. Sci.*, **115**, 12704–12709.

Qi,Q. *et al.* (2014) Diversity and clonal selection in the human T-cell repertoire. *Proc. Natl. Acad. Sci. USA*, **111**, 13139–13144.

Robins,H.S. *et al.* (2009) Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood*, **114**, 4099–4107.

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

Seay,H.R. *et al.* (2016) Tissue distribution and clonal diversity of the T and B cell repertoire in type 1 diabetes. *JCI Insight*, **1**, 1–19.

Sethna,Z. *et al.* (2017) Insights into immune system development and function from mouse T-cell repertoires. *Proc. Natl. Acad. Sci. USA*, **114**, 2253–2258.

Shugay,M. *et al.* (2018) VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res.*, **46**, D419–D427.

Sims,J.S. *et al.* (2016) Diversity and divergence of the glioma-infiltrating t-cell receptor repertoire. *Proc. Natl. Acad. Sci. USA*, **113**, E3529–E3537.

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.

Toledano,A. *et al.* (2018) Evidence for shaping of light chain repertoire by structural selection. *Front. Immunol.*, **9**, 1307.

Venturi,V. *et al.* (2008) The role of production frequency in the sharing of simian immunodeficiency virus-specific CD8+ TCRs between macaques. *J. Immunol.*, **181**, 2597–2609.

Venturi,V. *et al.* (2013) Specificity, promiscuity, and precursor frequency in immunoreceptors. *Curr. Opin. Immunol.*, **25**, 639–645.

Vollmers,C. *et al.* (2013) Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc. Natl. Acad. Sci. USA*, **110**, 13463–13468.

Wang,C. *et al.* (2010) High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proc. Natl. Acad. Sci. USA*, **107**, 1518–1523.

Weinstein,J.A. *et al.* (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science* , **324**, 807–810.

Woodsworth,D.J. *et al.* (2013) Sequence analysis of T-cell repertoires in health and disease. *Genome Med.*, **5**, 98.

Wu,J. *et al.* (2018) Expanded tcrßcdr3 clonotypes distinguish Crohn's disease and ulcerative colitis patients. *Mucosal Immunol.*, **11**, 1487–1495.

Zhao,Y. *et al.* (2016) Preferential use of public TCR during autoimmune encephalomyelitis. *J. Immunol.*, **196**, 4905–4914.