Genesis of the $\alpha\beta$ T-cell receptor

Thomas Dupic,[1] Quentin Marcou,[2] Thierry Mora,[3,4] and Aleksandra M. Walczak[2,4]

[1] *Laboratoire de physique théorique et hautes énergies, CNRS and Sorbonne Université, 4 Place Jussieu, 75005 Paris, France*
[2] *Laboratoire de physique théorique, CNRS, Sorbonne Université, and École normale supérieure (PSL), 24 rue Lhomond, 75005 Paris, France*
[3] *Laboratoire de physique statistique, CNRS, Sorbonne Université, Université Paris-Diderot, and École normale supérieure (PSL), 24 rue Lhomond, 75005 Paris, France*
[4] *These authors contributed equally. Please correspondance to* `tmora@lps.ens.fr`, `awalczak@lpt.ens.fr`

The T-cell (TCR) repertoire relies on the diversity of receptors composed of two chains, called α and β, to recognize pathogens. Using results of high throughput sequencing and computational chain-pairing experiments of human TCR repertoires, we quantitively characterize the $\alpha\beta$ generation process. We estimate the probabilities of a rescue recombination of the β chain on the second chromosome upon failure or success on the first chromosome. Unlike β chains, α chains recombine simultaneously on both chromosomes, resulting in correlated statistics of the two genes which we predict using a mechanistic model. We find that $\sim 28\%$ of cells express both α chains. We report that clones sharing the same β chain but different α chains are overrepresented, suggesting that they respond to common immune challenges. Altogether, our statistical analysis gives a complete quantitative mechanistic picture that results in the observed correlations in the generative process. We learn that the probability to generate any TCR$\alpha\beta$ is lower than 10^{-12} making it very unlikely for two people to share a full receptor by chance. We also estimate the generation diversity of the full TCR repertoire.

The adaptive immune system confers protection against many different pathogens using a diverse set of specialized receptors expressed on the surface of T-cells. The ensemble of the expressed receptors is called a repertoire and its diversity and composition encode the ability of the immune system to recognize antigens. T-cell receptors (TCR) are composed of two chains, α and β, that together bind antigenic peptides presented on the multihistocompatability complex (MHC). High-throughput immune sequencing experiments give us insight into the repertoire composition through lists of TCR, typically centered around the most diverse region, the Complimentary Determining Region 3 (CDR3) of these chains [1–5]. Until recently most experiments and analyses focused on only one of the two chains at a time, and studies of TCR with both chains were limited to low-throughput methods [6–8]. Recent technological and analytical breakthroughs now allow us to simultaneously determine the sequences of both α and β chains expressed on cells of the same clone in a high-throughput way [9] (see also analysis of unpublished data obtained by single-cell sequencing in [10]). These advances make it possible to study the repertoires of paired receptors, and to revisit the questions of the generation, distribution, diversity and overlap of TCR repertoires previously studied at the single-chain level [11–17], but also to gain insight into the mechanisms of T-cell recombination and maturation.

TCR receptor diversity arises from genetic recombination of the α and β chains of thymocytes in the thymus. Each chain locus consists of a constant region (C), and multiple gene segments V (52 for the human β chain and $\approx$ 70 for α), D (2 and 0) and J (13 and 61). Recombination proceeds by selecting one of each type of segment and joining them together, with additional deletions or insertions of base pairs at the junctions. TCRβ is first recombined and expressed along with the pre-T cell receptor alpha (a non-recombined template gene) on the surface of the cell to be checked for function. T cells then divide a few times before TCRα recombination begins, at which point the thymic selection process acts on the complete receptor. The recombination of each chain often result in non-productive genes (e.g. with frameshifts or stop codons). Subsequent rescue and selection mechanisms ensure that all mature T cells express at least one functional receptor. Recombination of the β chain on the second chromosome may be attempted if the initial recombination was unsuccessful. By contrast, the α chain is recombined on both chromosomes simultaneously [18], and proceeds through several recombination attempts that successively join increasingly distal V and J segments (Fig. 1). Taken together, recombination events can potentially produce up to 4 chains (2 α and 2 β) in each cell. In principle, allelic exclusion ensures that only one receptor may be expressed on the surface of the cell, but this process is leaky: 7% of T-cells have two productive β-chains [19, 20], and %1 express both of them on the surface [21–23]. Allelic exclusion in the α chain is less well quantified as it relies on different mechanisms [24, 25], with estimates ranging from 7% [8] to 30% [22] of cells with two functionally expressed α chains.

Despite the partial characterization of the various mechanisms underpinning the recombination, rescue and selection of the two TCR chains, a complete quantitative picture of these processes is still lacking. For instance, the probability of recombination rescue, the probability for a chain to pass selection, or the extend of allelic exclusion, have not been measured precisely. Here we reanalyse the data from [9] to link together each of the 4 α and β chains of single clones, and study α-α and β-β pairs as well as α-β pairs. Using these pairings, we

propose a mechanistic model of recombination of the two chains on the two chromosomes, and study the statistics of the resulting functional $\alpha\beta$ TCR.

RESULTS

Pairing multiple chains in the same clone.

We analysed previously published data on sequenced T-cell CDR3 regions obtained from two human subjects (PairSEQ), as described by Howie and collaborators [9]. In the original study, sequences of α and β chain pairs associated to the same clone were isolated using a combination of high-throughput sequencing and combinatorial statistics. Briefly, T cell samples were deposited into wells of a 96-well plate, their RNA extracted, reverse-transcribed into cDNA with the addition of a well-specific barcode, amplified by PCR, and sequenced. $\alpha\beta$ pairs appearing together in many wells were assumed to be associated with the same T-cell clone, and thus expressed together in the same cells. Because the method relies on the presence of cells of the same clone in many wells, the method can only capture large memory T cell clones present in multiple copies in the same blood sample. Naive clones which have a population size of around 10, or concentration of 10^{-10} [26], are not expected to be paired in this way.

We generalized the statistical method of [9] to associate α-α and β-β pairs present in the same clone. Along with α-β pairings, this allowed us to reconstruct the full TCR content of a cell. Two additional difficulties arise when trying to pair chains of the same type. First, truly distinct pairs of chains must be distinguished from reads associated with the same sequence but differing by a few nucleotides as a result of sequencing errors. We set a threshold of 11 nucleotide mismatches on the distribution of distances between paired chains (Fig. S1) to eliminate such duplicates. Second, because of allelic exclusions, one of the two chains of the same type is typically expressed in much smaller amounts than the other. As a result, we find much fewer α-α and β-β pairs than α-β pairs.

Table I summarizes the numbers of pairs found in each experiment, with a significance threshold chosen to achieve a 1% false discovery rate (see Methods). This method can then be used to recreate the complete TCR content of a given clone, and set apart clones expressing multiple TCR receptors.

Correlations between chains of the same cell.

Correlations between the features of the recombination events of the chains present in the same cells are informative about the rules governing the formation of a mature $\alpha\beta$ TCR in the case of α-β pairings, and also about the mechanisms and temporal organization of recombination on the two chromosomes in the case of α-α

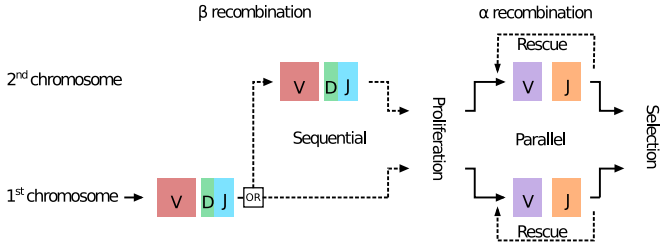

FIG. 1: Formation of a T-cell receptor. The β chain is rearranged before the α chain. The recombination on the two chromosomes is sequential for β, and parallel for α. Dotted lines indicate optional events. Rescue events on the α chain correspond to successive recombinations of the same locus (see also schematic in Fig. 3).

and β-β pairings. We computed the mutual information, a non-parametric measure of correlations (see Methods), between pairs of recombination features for each chain: V, D, and J segment choices, and the numbers of deletions and insertions at each junction (Fig. 2). Because recombination events cannot be assigned with certainty to a given sequence, we used the IGoR software [27] to associate recombination events to each sequence with a probabilistic weight reflecting the confidence we have in this assignment (see Methods). We have shown previously that this probabilistic correction removes spurious correlations between recombination events [12, 27]. Correlations within single chains recapitulate previously reported results for the β [12] and α [28] chains. Inter-chain correlations, highlighted by red boxes, are only accessible thanks to the chain pairings.

We find no correlation between the number of insertions in different chains across all pair types. Such a correlation could have been expected because Terminal deoxynucleotidyl transferase (TdT), the enzyme responsible for insertions, is believed to correlate with the number of insertions [29], and is expected to be constant across recombination events in each cell. The lack of correlation between different insertion events thus suggests that the there is no shared variability arising from differences in TdT concentration across cells.

We report generally weak correlations between the α and β chains (Fig. 2A), with a total sum of 0.36 bits, about 10 times lower than the total intra-gene correlations of the α chain. The largest correlation is between the choice of V_α and V_β genes (0.036 bits) and J_α and V_β genes (0.033 bits), in agreement with the analysis of [10] on unpublished single-cell data. These correlations probably do not arise from biases in the recombination process, because recombination of the two chains occurs on different loci (located on distinct chromosomes) and at different stages of T cell maturation. A more plausible explanation is that thymic selection preferentially selects some chain associations with higher folding stability or better peptide-MHC recognition properties. Distinguishing recombination- from selection-induced corre-

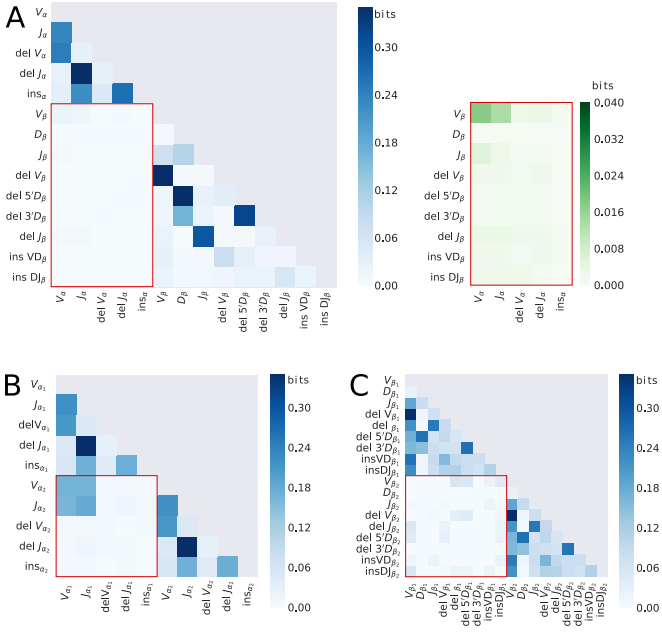

FIG. 2: Mutual information (a non-parametric measure of correlations) between the recombination events of the paired chains: V, D, and J segment choice, numbers of bases deleted from the 3' end of the V-gene (delV), the 5′ end of the J-gene (delJ), and both ends of the D-gene for the β chain (del5′D and del3′D for the 5′ and 3′ ends, respectively); number of insertions of random nucleotides between V and J segments (insVJ) for the α chain, and between V and D (insVD) and between J and D (insDJ) segments for the β chain. Mutual information for (**A**) α-β pairs (on the right in green: close-up of the inter-chain mutual information); (**B**) α-α pairs; and (**C**) β-β pairs. Inter-chain correlations are highlighted by red boxes.

lations would require analysing pairs of non-productive sequences, which are not subjected to selection, but the number of such pairs in the dataset was too small to extract statistically significant results.

Pairs of β chains show almost no correlations (Fig. 2C). Looking in detail at the correlations between gene fragments reveals a strongly negative correlation of TCRBV21-01 and TCRBV23-01 (both pseudogenes) with themselves (Fig. S2), which is expected because at least one of the two β chain must have a non-pseudogene V. More generally, correlations are likely to arise from selection effects, since the two recombination events of the two β chains are believed to happen sequentially and independently. The fact that at least one of the chains needs to be functional for the cell to survive breaks the independence between the two recombination events. By contrast, the α-α pairs have very strong correlations between the V and J usages of the two chromosomes, and none between any other pair of features (Fig. 2B). These correlations arise from the fact that the two α recombination events occur processively and simultaneously on the two chromosomes, as we analyse in more detail below.

Correlations between α chains can be explained by a rescue mechanism.

We wondered whether the detailed structure of the observed correlations between the α chains on the two chromosomes could be explained by a simple model of recombination rescue. The correlations of the V_α segments on the two chromosomes and of the J_α segments show a similar spatial structure as a function of their ordering on the chromosome (see Fig 3A): proximal genes are preferentially chosen together on the two chromosomes, as are distal genes.

The two chromosomes recombine simultaneously, and proceed by successive trials and rescues. If the first recombination attempt fails to produce a functional chain, another recombination event may happen on the same chromosome between the remaining distal V and J segments, excising the failed rearranged gene in the process. The process stops when a functional chain has been recombined. By the time this happens on one chromosome, a similar number of recombination attempts will have occurred on the other chromosome. We hypothesize that this synchrony is the main source of correlations between the Vα and Jα gene usages of the two chains.

To validate this hypothesis, we simulated a minimal model of the rescue process similar to [30] (Methods), in which the two chromosomes are recombined in parallel. If recombination happens to fail on both chromosomes, repeated "rescue" recombinations take place between outward nearby segments (Fig 3C). The covariance matrices obtained from the simulations for both V_α and J_α (Fig. 3B) show profiles that are very similar to the data, with positive correlations along the diagonal, in particular at the two ends of the sequence.

Probability of recombination of the second chromosome.

We wondered if the paired data could be used to estimate the percentage of cells with two recombined chains of the same type. However, since pairing was done based on mRNA transcripts through cDNA sequencing, silenced or suppressed genes are not expected to be among the identified pairs, leading to a systematic underestimation of double recombinations. Analysing sequence data derived from genomic DNA (gDNA) instead of cDNA can solve this problem, but the datasets were too small to get statistically significant pairings. However, we can derive strict bounds from the proportion of productive sequences found in the (unpaired) gDNA dataset of [9]. Following recombination, using IGoR we estimate $p_{\rm nc}^{\alpha} = 69.5\%$ of the α sequences, and $p_{\rm nc}^{\beta} = 73.5\%$ of β sequences are non-coding or contain a stop codon. We collectively refer to as "non-coding" sequences. The remaining sequences, called "coding", make up a fraction $p_{\rm c}^{\alpha,\beta} = 1 - p_{\rm nc}^{\alpha,\beta}$ of random rearrangements. We denote by $p_{\rm f}^{\alpha}$ and $p_{\rm f}^{\beta}$ the probability that a coding sequence can

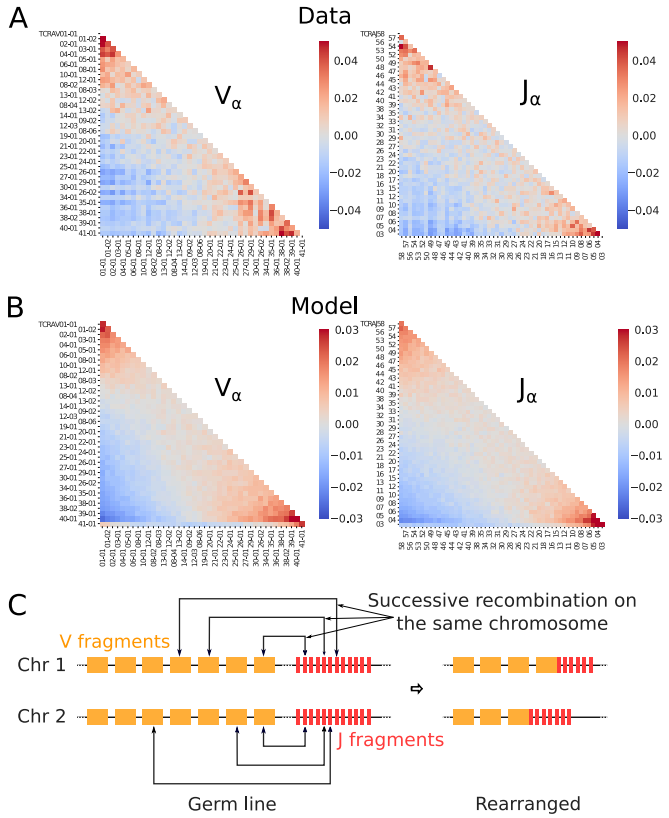

FIG. 3: **A.** Pearson correlation between V and J gene segment usage for TCRα. The correlation is taken between the truth values of particular V and J gene choices (a value of 1 is assigned if a given segment is observed and 0 if it is not, see Methods for details). **B.** Same Pearson correlation as in (A) calculated from simulations of the rescue mechanism model depicted in (C). **C.** Cartoon of the rescue mechanism. The rescue happens simultaneously on the two chromosomes. Once one of the re-arrangements results in a functional re-arrangement, recombination stops. In the end, the V and J gene segments selected on both chromosomes are close to each other in the germline ordering.

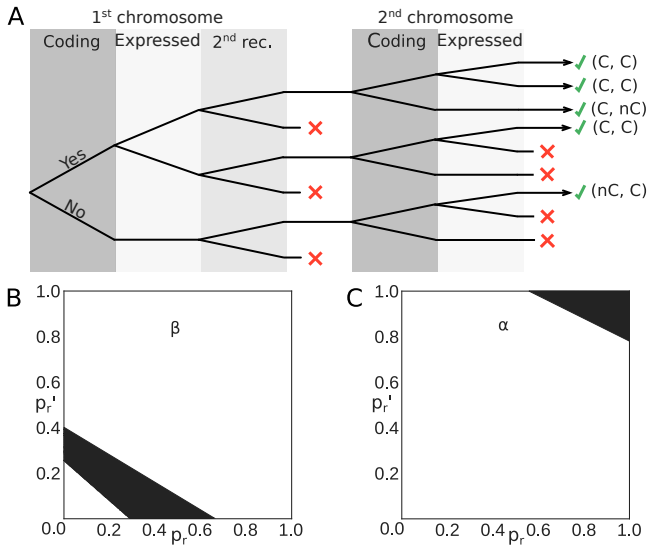

FIG. 4: **A.** Decision tree of the recombination process for one chain (α or β). The first part shows the recombination of the first chromosome, the second part of the second chromosome. In each area a binary choice is made. Red crosses indicate decision outcomes that lead to no observed sequence. Observable outcomes (with at least one coding sequence) are indicated at the end of the tree by green ticks. C stands for coding, nC for non-coding. **B.** Bounds on the allowed values of rescue probabilities for the β chain calculated from the decision tree in (A). The black part of the graph corresponds to the allowed values of p_r' (probability of a second recombination for β if the first was successful) and p_r (probability of a second recombination for β if the first was not successful). The bounds were obtained by imposing $0 < p_f^\beta < 1$ in Eq. 1. **B.** Bounds on the allowed values of rescue probabilities for the α chain are consistent with both chromosomes recombining simultaneously and independently, $p_r = p_r' = 1$.

express a functional α or β chain that can ensure its selection.

The number of observed non-coding sequences depends on whether the second chromosome attempts to recombine following the recombination of the first one. We call p_r the probability that a second recombination happens when the first recombination fails to produce a functional chain, and p_r' when the first recombination succeeds. Then, the proportion f_c of observed non-coding sequences can be written as (see tree in Fig. 4 and Methods):

$$f_c = \frac{(p_r + p_r')p_{nc}}{1 + p_r' + 2(1 - p_f p_c)p_r}. \quad (1)$$

Note that this formula assumes that the presence of more than one functional chain does not affect its selection probability. Comparing the proportion of observed non-coding β chain sequences calculated from Eq. 1 with the values from gDNA data ($f_c^\beta = 18 \pm 1\%$ in [9] and 14% in [11]), allows us to constrain the values of p_r^β and $p_r^{\beta\prime}$. The probability of a second recombination, even if the first recombination failed, is always lower than 65% (Figure 4A). By constrast, the observed fraction of non-coding sequences in the α chain, $f_c^\alpha = 40 \pm 1\%$, constrains the the rescue probabilities p_r^α and $p_r^{\alpha\prime}$ to be close to 100% (Figure 4B), in agreement with the fact that both chromosomes are believed to recombine independenly. Assuming strict independence, $p_r^\alpha = p_r^{\alpha\prime} = 1$ puts bounds on the probability that a random coding α sequence is functional, $70\% \leq p_f^\alpha \leq 100\%$.

Fraction of cells with two functional α chains.

Can we learn from pairing data what fraction of cells expressed two chains of the same type? gDNA pairings do not allow us to do that, because they are severely lim-

ited by sequencing depth: most chains cannot be paired because of material losses, and estimating the fraction of cells with several chains is impossible. While cDNA pairings are in principle less susceptible to material loss, non-functional sequences are much less expressed than functional ones [23, 25], lowering their probability of being paired and introducing uncontrolled biases in the estimate of fractions of cells with different chain compositions. However, we can use this difference in expression patterns by examining the distribution of read counts for each type of chain. Sequences of chains paired with a non-coding chain of the same type must be functional and expressed on the surface of the cell. Those sequences have a markedly different distribution of read counts than non-coding sequences (Fig. S3A and B). Coding sequences that are coupled with another coding sequence can be either expressed or silenced, depending on their own functionality and the status of the other chain. Thus, their read count should follow a mixture distribution of both expressed and silenced sequences, the latter being assumed to follow the same distribution as non-coding sequences. Fitting the parameters of this mixture to the read counts of paired coding sequences (Fig. S3C) yields the total proportion $p_{\mathrm{e}} \leq p_{\mathrm{f}}$ of functional and expressed sequences, among all functional coding sequences p_{f}.

For α sequences, we found $p_{\mathrm{e}}^{\alpha} = 64 \pm 5\%$, meaning that $2p_{\mathrm{e}}^{\alpha} - 1 = 28\% \pm 10\%$ of cells express two different α chains (see Methods). This number is consistent with older results [31], but slightly higher than a recent estimate of 14% based on single-cell sequencing [8], although that estimate may be affected by material loss and should be viewed as a lower bound.

For β chains, the fit is noisier, because non-coding sequences are much more suppressed and therefore scarcer than for the α chain (only 4.5% of sequences are non-coding). We estimate that there are 8-10 times more silenced coding sequences than non-coding sequences, but the fit does not allow us to estimate the fraction of cells with two expressed β chains, although this number is consistent with 0 according to the data.

Functional sequences are more restricted than 'just coding' sequences.

It is often assumed that all coding sequences must be functional, and previous studies have used the difference between coding and non-coding sequences to quantify the effects of selection [14, 32, 33]. However, some fraction of coding sequences may actually be disfunctional, silenced, or not properly expressed on the cell surface. By contrast, sequences that can be paired with a non-coding sequence of the same type must be functional and expressed on the cell surface, lest the cell that carries them dies. These sequences represent a non-biased sample of all functional sequences, and their statistics may differ from those of 'just coding' sequences. In Table II we report the differences between the two ensembles in terms of sequence length and gene usage. All comparisons are with sequences that could be paired with another one to remove possible biases from the pairing process.

We find that functional sequences are on average slightly larger (by 1-2 nucleotides) than coding and non-coding sequences (Table II and Fig. S4). More markedly, the variance of their length is smaller, implying stronger selection towards a prefered length in the functional ensemble than in the coding and non-coding ensembles. These observations, which hold for both the α or β chains, indicate that the functional ensemble (as defined here using pairing information) is more restricted than 'just coding' sequences, and gives a more precise picture of the selected repertoire.

The impact of selection can also be measured by how much gene usage departs from the unselected ensemble using the Kullback-Leibler divergence (see Methods and Table II), offering a more contrasted view. V_{β} and J_{α} usages are similar in functional and coding sequences in terms of their divergence with non-coding sequences. For J_{β} however, this divergence is higher in functional than in simply coding sequences, while the opposite is true for V_{α}.

Model predicts very rare $\alpha\beta$ TCR sharing

Ignoring small correlations between features of the α and β chains reported in Fig. 2, we can assume that the probability of generating a $\alpha\beta$ pair is given by the product of the probabilities of generating each chain independently. These probabilities can be calculated using the IGoR software [27] for each paired chain in our datasets. The distribution of the pair generation probabilities obtained in this way (Fig. 5 A) shows an enormous breadth, spanning more than 20 orders of magnitude. We self-consistently validated the assumption of independence by showing that random assortments of α and β chains yielded an identical distribution of generation probabilities (green curve).

The maximum TCR generation probability is $< 10^{-12}$, meaning that generating the same pair twice independently is extremely unlikely. This suggests that, without strong antigenic selection, distinct individuals should share only a negligible number of TCR sequences. To make that prediction more quantitative, we simulated a computational model of sequence generation followed by thymic selection. α and β chains were generated by IGoR, and then each TCR$\alpha\beta$ amino-acid sequence was kept with probability q to mimick thymic selection [17]. We further assume that selection acts on each chain independently, so that the ratio q is given by $q_{\alpha}q_{\beta}$, where $q_{\alpha,\beta}$ are the selection probabilities infered from the analysis of single chains. These selection factors can be obtained by fitting the curve giving the number of unique amino-acid sequences as a function of unique nucleotide sequences [17], yielding $q_{\beta} = 0.037$ and $q_{\alpha} = 0.16$ (Fig. S5).

Using the model, we can make predictions about the expected number of TCR$\alpha\beta$ nucleotide sequences shared between any of 10 individuals (Fig. 5 B) for which a million unique synthetic TCR$\alpha\beta$ were obtained. We find that, while a substantial fraction of sequences of each chain are expected to be shared by several individuals, sharing the full TCR$\alpha\beta$ is very unlikely, and drops well below 1 for more than 2 individuals. This suggests that the existence in real data of any TCR$\alpha\beta$ shared between several individuals should be interpreted as resulting from strong common selection processes, probably associated with antigen-specific proliferation, leading to convergent selection of the shared sequences.

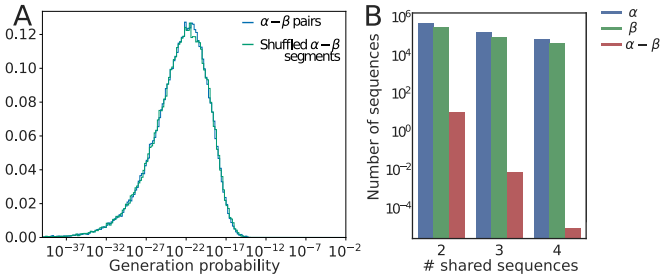

FIG. 5: Generation probability of a full $\alpha\beta$ TCR. **A.** Distribution of the generation probabilities of $\alpha\beta$ pairs, obtained by multiplying the generation probabilities of the α and β sequences. The graph shows the distribution for paired sequences (blue) and random associations of $\alpha\beta$ pairs (green). **B.** Number of CDR3 nucleotide sequences found in n among 10 individuals with a sample depth of $N = 10^6$ unique $\alpha\beta$ TCR per individual. The probability of more than two people sharing the same TCR receptor is extremely small.

Co-activation of cells sharing the same β chain.

To further investigate the effects of convergent selection, we quantified how often the same α chain was associated with distinct β chains in different clones (Fig. S6A), and *vice versa* (Fig. S6B). While association of a β with 2 distinct α chains could happen in the same cell because of the existence of two copies, we found a substantial fraction (3%) of all paired TCRβ that could be associated with three or more TCRα. There can be several explanations for this observation. First, convergent recombination of β can create clones sharing β but not α. This effect can be quantified using the generation and thymic selection model introduced in the previous paragraph. Simulations show that convergent recombination happens with a rate of 0.5%, and thus cannot explain the data. Second, cells divide around 5 times between the β and α recombinations, naturally leading to clones with the same β but distinct α. However, because the sampled repertoire ($\sim 10^5$ clones) only covers a minute fraction of all clones (estimated to be $> 10^8$ for the β chain only [13]), this effect can only explain $\sim 10^5/10^8 = 0.1\%$ of pairing multiplicity.

Thus, the data suggests that cells with identical β chains are convergently selected in the periphery as a result of antigen exposure, even if their α sequences differ. This agrees with the observation that the β chain is a strong indicator of epitope specificity [34].

DISCUSSION

Analysing computationally reconstructed pairs of TCR α and β chains, as well as α-α and β-β pairs, allowed us to quantify the various steps of sequence generation, rescue mechanisms, convergent selection, and sharing that were not accessible from just single-chain data.

Pairing α chains in single cells revealed correlations that were suggestive of a parallel and processive mechanism of VJ recombination in the two chromosomes. These signatures were well recapitulated by a simple computational model of successive rescue recombinations. Our model is similar to that of [30], but differs in its details and parameters, as the original model could not reproduce the correlation pattern of the data.

We estimated that $\sim 28\%$ of cells express two α chain, higher than a recent report of 14% using single-cell sequencing [8]. However, this fraction is very hard to assess experimentally from high-throughput sequencing, as material loss can lead to its underestimation. While our estimate is indirect, we expected it to be more robust to such loss.

Our finding that the statistics of the two chains are largely independent of each other — with only a weak correlation between V_β and (V_α, J_α) usage — is in agreement with recent observations using direct single-cell chain pairing [10]. While independence between the α and β recombination processes is perhaps expected because they occur at different stages of T-cell development, it is worth emphasizing that the absence of correlations reported here involves coding TCR$\alpha\beta$ sequences, which are believed to be largely restricted by thymic selection. This restriction can introduce correlations, notably through negative selection which could forbid certain $\alpha\beta$ combinations. Our results do not exclude such joint selection, but suggests that it does not introduce observable biases. The independence between the two chains implies that the entropies of the two generation processes can be simply summed to obtain the entropy of the full TCR$\alpha\beta$. Taking the values previously reported in [15] of 26 bits for the α chain, and 38 bits for the β chain, yields 64 bits for the TCR$\alpha\beta$, *i.e.* a diversity number of $2^{64} \approx 2 \cdot 10^{19}$.

The independence between the chains also allowed us to make predictions about the amount of TCR repertoire overlap one should expect between individuals. Our analysis predicts that sharing of $\alpha\beta$ pairs between more than two individuals should be extremely rare. In a recent report [10], 26 $\alpha\beta$ pairs were found to be shared between 5 individuals. Our result indicate that such a high level of sharing cannot be explained by convergent recombination

alone, but may result from over-correcting for sequencing errors, or alternatively from strong convergent selection in all 5 donors.

Future studies collecting the $\alpha\beta$ repertoires of more individuals, as promised by the rapid development of single-cell sequencing techniques, will help us get a more detailed picture of the diversity and sharing properties of the TCR$\alpha\beta$ repertoires. Our analysis provides a useful baseline against which to compare and assess the results of these future works.

METHODS

Generation model. The generation model was obtained and used through the IGoR software [35]. It can be used to compute the probability of generation of both the α and β chains from generic, individual-independent models of recombination, by summing over the possible recombination scenarios leading to it. It can also be used to generate sequences *in silico* with the same statistics.

Pairing of sequences. We use the data and method of [9] to infer pairing from sequencing data of cells partitioned in $W = 95$ wells (instead of 96 as erroneously reported in the original paper, as one of the wells did not provide any results). We calculate the p-value that two sequences each present in w_1 and w_2 well are found together in w_{12} wells, under the null model that they are distributed randomly and independently: $p(w_{1,2}, w_1, w_2, W) = \sum_{u \geq w_{12}} \binom{w_1}{u}\binom{W-w_1}{w_2-u}/\binom{W}{w_2}$. We set the p-value threshold so that, for each well occupation number, the false discovery rate of 1% is respected.

Information quantities. The mutual information (in bits) of two variables X, Y with a joint distribution $p(x, y)$ is defined by: $I(X, Y) = \sum_{x,y} p(x, y) \log_2[p(x, y)/(p(x)p(y))]$. We estimated it from the empirical histogram of (x, y) using a finite size correction [36], $(n_X n_Y - n_X - n_Y + 1)/2N \log(2)$, where N is the sample size, and n_A is the number of different values the variable A can take. In the specific case of sequences in paired cells, an exact correction can be found by computing the mutual information between non-paired sequences. The Kullback-Leibler divergence between two distribution $p(x)$ and $q(x)$ of a variable X is given by: $D_{\mathrm{KL}}(p\|q) = \sum_x p(x) \log_2(p(x)/q(x))$.

Simulation of the rescue process. The V and J genes are indexed by i and j from most proximal to most distal along the chromosome: V_i, $i = 1, \ldots L_V$ and J_j, $j = 1, \ldots L_J$. In the first recombination attempt of the first chromosome, the model picks the V and J gene indices i_1 and j_1 from a truncated geometric distribution, $P(i_1 = i) \propto (1-p)^{i-1}$ (and likewise for j_1), with $p = 0.05$. The same process is simulated for the second chromosome. With probability 2/3 for each chromosome, the recombination fails. If both chromosome fail, a second recombination takes place on each between more distal genes indexed by $i_2 > i_1$ and $j_2 > j_1$, distributed as $P(i_2 = i) \propto (1-p)^{i_2-i_1-1}$ (and likewise for j_2), to reflect observations that successive recombination occur on nearby genes in the germline [37]. If recombination repeatedly fail on both chromosomes, the process is repeated up to 5 times [38]. This model is similar to that of [30], where a uniform instead of a geometric distribution was used.

Bounds on rescue probabilities. Non coding sequences can only appear in the TCR repertoire if they share a cell with a functional sequence. The probability of such a cell to appear in the selection process is $A = p_{\mathrm{nc}}(p_{\mathrm{r}} + p'_{\mathrm{r}})p_{\mathrm{c}}p_{\mathrm{f}}$. The probability for a cell to possess only one functional receptor is $B = p_{\mathrm{c}}p_{\mathrm{f}}(1 - p'_{\mathrm{r}})$, while the probability to possess two receptors and at least one functional one can be written as $C = p_{\mathrm{c}}p_{\mathrm{f}}\,[p_{\mathrm{r}}\,(1 - p_{\mathrm{c}}p_{\mathrm{f}}) + p'_{\mathrm{r}}]$. The proportion of non-coding reads is thus $A/(B + 2C)$, which gives Eq. 1.

Copy number distributions. We fit the empirical distribution of reads per coding chain, ρ_{c}, with a mixture of two distributions (Fig. S3): ρ_{e}, corresponding to chain sequences that could be paired with a non-coding sequence of the same type and thus believed to be expressed; and ρ_{nc} corresponding to non-expressed sequences and learned from non-coding sequences. The fit is done by mean square error minimization: $\int \mathrm{d}x(\rho_{\mathrm{c}}(x) - \lambda_1\rho_{\mathrm{nc}}(x) - \lambda_2\rho_{\mathrm{e}}(x))^2$. The fraction of expressed chains among coding ones is then given by $p_{\mathrm{e}} = \lambda_2/(\lambda_1 + \lambda_2)$. Calling $p_{2\alpha}$ the proportion of cells with two expressed α, the resulting fraction p_{e} of α sequences that are expressed should be $p_{\mathrm{e}} = (2p_{2\alpha} + (1 - p_{2\alpha}))/2$, hence $p_{2\alpha} = 2p_e - 1$.

Sharing estimation. We follow the methods of [17]. A large number of productive α and β chain pair sequences are generated through a stochastic model of recombination using IGoR [35]. Each TCR$\alpha\beta$ amino-acid sequence is then kept if its normalized hash (a hash is a deterministic but maximally disordered function) is $\leq q = q_\alpha q_\beta$, so that a random fraction q of sequences passes selection. The values of q_α and q_β are learned from rarefaction curves showing the number of unique amino-acid sequences of each chain as a function of the number of unique nucleotide sequences (Fig. S5), using the analytical expressions given in [17].

The predictions for the number of shared TCR$\alpha\beta$ nucleotide sequences reported in Fig. 5B are computed using the analytical expressions of [17].

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

[2] Boyd SD, et al. (2009) Measurement and clinical monitoring of human lymphocyte clonality by massively parallel {VDJ} pyrosequencing. *Sci Transl Med* 1:12ra23.

[3] Benichou J, Ben-Hamo R, Louzoun Y, Efroni S (2012) Rep-Seq: Uncovering the immunological repertoire through next-generation sequencing.

[4] Robins H (2013) Immunosequencing: applications of immune repertoire deep sequencing. *Curr. Opin. Immunol.* 25:646–652.

[5] 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.

[6] Kim SM, et al. (2012) Analysis of the paired TCR α- and β-chains of single human T cells. *PLoS One* 7.

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

[8] Han A, Glanville J, Hansmann L, Davis MM (2014) Linking T-cell receptor sequence to functional phenotype at

the single-cell level. *Nat. Biotechnol.* 32:684–692.

[9] Howie B, et al. (2015) High-throughput pairing of T cell receptor α and β sequences. *Science Translational Medicine* 7:301ra131–301ra131.

[10] Grigaityte K, et al. (2017) Single-cell sequencing reveals $\alpha\beta$ chain pairing shapes the T cell repertoire.

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

[12] 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.

[13] Qi Q, et al. (2014) Diversity and clonal selection in the human T-cell repertoire. *Proc. Natl. Acad. Sci. U. S. A.* 111:13139–44.

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

[15] Mora T, Walczak A (2018) in *Syst. Immunol.*, eds Das JD, Jayaprakash C (CRC Press), pp 185–199.

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

[17] Elhanati Y, Sethna Z, Callan Jr. CG, Mora T, Walczak AM (2018) Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunological Reviews* in press.

[18] Petrie HT, et al. (1993) Multiple rearrangements in T cell receptor alpha chain genes maximize the production of useful thymocytes. *Journal of Experimental Medicine* 178:615–622.

[19] Stubbington MJT, et al. (2016) T cell fate and clonality inference from single-cell transcriptomes. *Nature Methods* 13:329.

[20] Eltahla AA, et al. (2016) Linking the T cell receptor to the single cell transcriptome in antigen-specific human T cells. *Immunology and Cell Biology* 94:604–611.

[21] Davodeau F, et al. (1995) Dual T cell receptor beta chain expression on human T lymphocytes. *Journal of Experimental Medicine* 181:1391–1398.

[22] Padovan E, et al. (1995) Normal T lymphocytes can express two different T cell receptor beta chains: Implications for the mechanism of allelic exclusion. *Journal of Experimental Medicine* 181:1587–1591.

[23] Steinel N, Brady BL, Carpenter AC, Yang-Iott KS, Bassing CH (2010) Post-Transcriptional Silencing of VβDJβCβ Genes Contributes to TCRβ Allelic Exclusion in Mammalian Lymphocytes. *Journal of immunology (Baltimore, Md. : 1950)* 185:1055–1062.

[24] Rybakin V, et al. (2014) Allelic Exclusion of TCR α-Chains upon Severe Restriction of Vα Repertoire. *PLoS ONE* 9.

[25] Niederberger N, et al. (2003) Allelic Exclusion of the TCR α-Chain Is an Active Process Requiring TCR-Mediated Signaling and c-Cbl. *The Journal of Immunology* 170:4557–4563.

[26] Casrouge A, et al. (2000) Size Estimate of the $a\beta$ TCR Repertoire of Naive Mouse Splenocytes. *The Journal of Immunology* 164:5782–5787.

[27] Marcou Q, Mora T, Walczak AM (2018) High-throughput immune repertoire analysis with IGoR. *Nature Communications* 9:561.

[28] 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* 32:1943–1951.

[29] Motea EA, Berdis AJ (2010) Terminal Deoxynucleotidyl Transferase: The Story of a Misguided DNA Polymerase. *Biochimica et biophysica acta* 1804:1151–1166.

[30] Warmflash A, Dinner AR (2006) A Model for TCR Gene Segment Use. *The Journal of Immunology* 177:3857–3864.

[31] Padovan E, et al. (1993) Expression of two T cell receptor alpha chains: Dual receptor T cells. *Science* 262:422–424.

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

[33] Toledano A, et al. (2018) Evidence for shaping of l chain repertoire by structural selection. *Front. Immunol.* in press.

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

[35] Marcou Q, Mora T, Walczak AM (2017) IGoR: A tool for high-throughput immune repertoire analysis. *arXiv:1705.08246 [q-bio]*.

[36] Steuer R, Kurths J, Daub CO, Weise J, Selbig J (2002) The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics* 18:S231–S240.

[37] Pasqual N, et al. (2002) Quantitative and Qualitative Changes in V-J α Rearrangements During Mouse Thymocytes Differentiation. *The Journal of Experimental Medicine* 196:1163–1174.

[38] Murphy K, Weaver C (2016) *Janeway's Immunobiology, 9th Edition* (Garland Science).

TABLE I: Number of α-β, α-α and β-β statistically significant pairs in each of the three experiments from [9]. Samples were obtained from two human subjects X and Y and divided in three experiments (experiment 1, 2, and 3), with different sequencing depths and different subjects: experiment 3 contains only sequences from X, while experiments 1 and 2 contain sequences from both subjects.

Experiment	pairs (α, β)	pairs (α, α)	pairs (β, β)
1	1098	336	30
2	79420	47665	7795
3	129757	89957	15361

TABLE II: Length distribution and Kullback-Leibler (KL) divergence from the unselected (non-coding) ensemble for different types of sequences: functional (and expressed), coding, and non-coding.

chain	length: mean $\pm$ st. deviation (nt)			Gene	KL divergence (bits)	
	functional	coding	non-coding		functional	coding
α	42.0 ± 5.00	39.12 ± 6.67	40.0 ± 7.00	V_α	0.66 ± 0.05	1.39 ± 0.01
				J_α	0.110 ± 0.005	0.119 ± 0.004
β	44.1 ± 5.03	43.17 ± 6.22	43.4 ± 7.82	V_β	1.09 ± 0.06	1.03 ± 0.18
				J_β	0.12 ± 0.004	0.051 ± 0.008

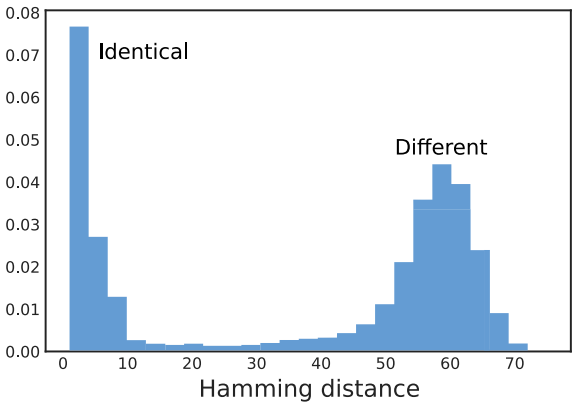

FIG. S1: Hamming distance between two TCRβ sequences identified as paired. Near-identical paired sequences are in their vast majority due to sequencing error. The Hamming distance permits to separate effectively these sequences from actually different sequences extracted from the same clone. A similar behaviour is observed for TCRα chains.

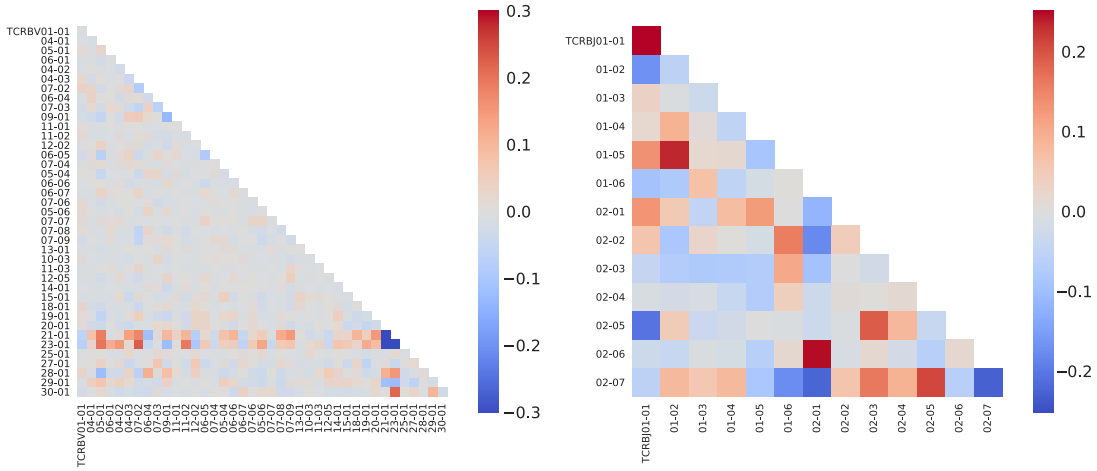

FIG. S2: Normalized covariance between V (left) and J (right) gene usages of pairs of β sequences found in the same clone. The V21-01 and V23-01 genes are non-functional pseudogenes and are thus anticorrelated.

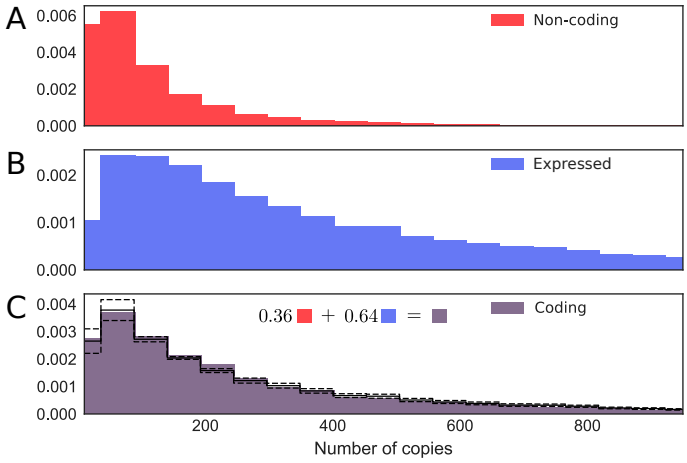

FIG. S3: Distribution of the number of reads of different types of TCRα RNA sequences: **(A)** non-coding; **(B)** functional and expressed (i.e. paired with a non-coding sequence); **(C)** 'just coding' sequences. In panel **(C)**, the full line represents the best fit for a mixture of 36% of non-coding and 64% of functional, expressed sequences. Dashed lines show 10% error intervals. To avoid biases in the comparisons, all sequences used in these distributions were paired.

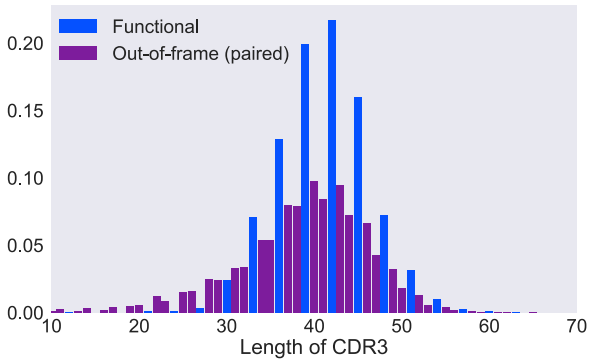

FIG. S4: CDR3 length distribution of expressed and out-of-frame TCRα sequences. Expressed sequences have a narrowed distribution than unselected ones. All sequences used in these distributions were paired.

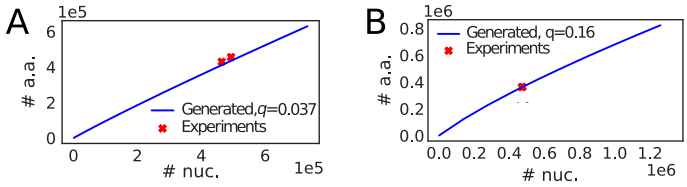

FIG. S5: Number of unique amino-acid (translated) sequences as a function of the number of unique nucleotide sequences for **(A)** α and **(B)** β chains. Red crosses are experimental data, blue line comes from simulations of the recombination model with random selection. For α the value of q is inferred by least-square minimisation to be $q_\alpha = 0.16$, while for β we used the value of $q_\beta = 0.037$ reported in Elhanati et al., *Immunological Reviews*, in press (2018).

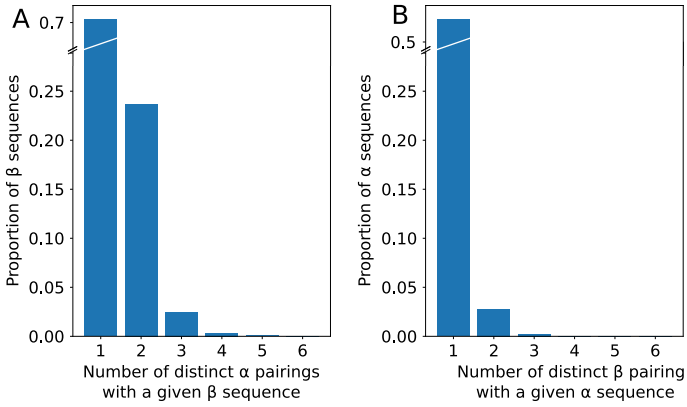

FIG. S6: **A.** Distribution of the number of distinct α sequences that could be paired with a given β sequence. **B.** Distribution of the number of distinct β sequences that could be paired with a given α sequence. Only sequences that appear in at least a pairing are considered. Since sequences may be paired with 2 chains of the other type in a single cell, only chains with 3 or more associations unambiguously correspond to the convergent selection of that chain in different clones.