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

Thomas Dupic, Quentin Marcou, Aleksandra M. Walczak, Thierry Mora

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

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

**Thomas Dupic**[1,2], **Quentin Marcou**[2], **Aleksandra M. Walczak**[2☯]*, **Thierry Mora**[2☯]*

**1** Laboratoire de physique théorique et hautes énergies, CNRS and Sorbonne Université, 4 Place Jussieu, 75005 Paris, France, **2** Laboratoire de physique de l'ENS, CNRS, Sorbonne Université, and École normale supérieure (PSL), 24 rue Lhomond, 75005 Paris, France

☯ These authors contributed equally to this work.
* awalczak@lpt.ens.fr (AMW); tmora@lps.ens.fr (TM)

## Abstract

The T-cell (TCR) repertoire relies on the diversity of receptors composed of two chains, called $\alpha$ and $\beta$, 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 $\beta$ chain on the second chromosome upon failure or success on the first chromosome. Unlike $\beta$ chains, $\alpha$ 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$35% of cells express both $\alpha$ chains. 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 $10^{-12}$ and estimate the generation diversity and sharing properties of the $\alpha\beta$ TCR repertoire.

## Author summary

Receptors on the surface of T-cells recognize pathogens and initiate an immune response. Analyzing the sequences of human T-cell receptors we draw a detailed quantitative picture of the generation process of the two receptor chains allowing us to estimate the diversity of the complete repertoire. We discuss which elements of the receptor production processes are correlated and which are independent, proposing mechanistic models at the origin of the correlations. We discuss the implications of our findings for the functional role of each of these cells, and the diversity of the repertoire.



**Citation:** Dupic T, Marcou Q, Walczak AM, Mora T (2019) Genesis of the $\alpha\beta$ T-cell receptor. PLoS Comput Biol 15(3): e1006874. https://doi.org/10.1371/journal.pcbi.1006874

**Editor:** Benny Chain, UCL, UNITED KINGDOM

**Received:** August 25, 2018

**Accepted:** February 17, 2019

**Published:** March 4, 2019



**Data Availability Statement:** All relevant code and curated data used to produced the analyses is available at https://github.com/Thopic/TCR_pairings.

**Funding:** This work was supported by the European Research Council (erc.europa.eu) Consolidator Grant n. 724208 to AMW. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## Introduction

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, $\alpha$ and $\beta$, 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 $\alpha$ and $\beta$ 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 $\alpha$ and $\beta$ chains of thymocytes in the thymus. Each chain locus consists of a constant region (C), and multiple gene segments V (52 for the human $\beta$ chain and $\approx$ 70 for $\alpha$), 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. TCR$\beta$ 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 TCR$\alpha$ 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 $\beta$ chain on the second chromosome may be attempted if the initial recombination was unsuccessful. By contrast, the $\alpha$ 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 $\alpha$ and 2 $\beta$) 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 $\beta$-chains [19, 20], and %1 express both of them on the surface [21–23]. Allelic exclusion in the $\alpha$ 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 $\alpha$ 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

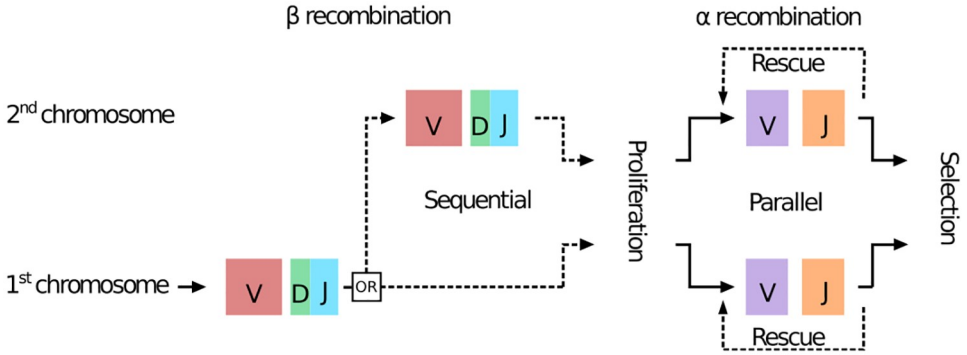


**Fig 1. Formation of a T-cell receptor.** The $\beta$ chain is rearranged before the $\alpha$ chain. The recombination on the two chromosomes is sequential for $\beta$, and parallel for $\alpha$. Dotted lines indicate optional events. Rescue events on the $\alpha$ chain correspond to successive recombinations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

https://doi.org/10.1371/journal.pcbi.1006874.g001

for a chain to pass selection, or the extent of allelic exclusion, have not been measured precisely. Here we re-analyse the data from [9] to link together each of the 4 $\alpha$ and $\beta$ chains of single clones, and study $\alpha$-$\alpha$ and $\beta$-$\beta$ pairs as well as $\alpha$-$\beta$ pairs. Using these pairings, we propose a mechanistic model of recombination of the two chains on the two chromosomes, inspired by [26], 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 $\alpha$ and $\beta$ 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 $10^{-10}$ [27], are not expected to be paired in this way.

We generalized the statistical method of [9] to associate $\alpha$-$\alpha$ and $\beta$-$\beta$ pairs present in the same clone. Along with $\alpha$-$\beta$ 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 (S1 Fig) to remove duplicates while minimizing the loss of real pairs. 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 $\alpha$-$\alpha$ and $\beta$-$\beta$ pairs than $\alpha$-$\beta$ pairs.

Table 1 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 $\alpha$-$\beta$ pairings, and also about the mechanisms and temporal organization of recombination on the two chromosomes in the case of $\alpha$-$\alpha$ and $\beta$-$\beta$ pairings. We computed the mutual information, a non-parametric measure of correlations (see Methods), between pairs of

**Table 1. Number of $\alpha$-$\beta$, $\alpha$-$\alpha$ and $\beta$-$\beta$ 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.

| Exp. | # cells | unique $\alpha$ | unique $\beta$ | pairs ($\alpha$, $\beta$) | pairs ($\alpha$, $\alpha$) | pairs ($\beta$, $\beta$) |
|---|---|---|---|---|---|---|
| 1 | $3.8 \times 10^5$ | $1.8 \times 10^6$ | $1.7 \times 10^6$ | 1098 | 336 | 30 |
| 2 | $1.5 \times 10^7$ | $2.7 \times 10^7$ | $3.3 \times 10^7$ | 79420 | 47665 | 7795 |
| 3 | $1.5 \times 10^7$ | $5.1 \times 10^7$ | $6.3 \times 10^7$ | 129757 | 89957 | 15361 |

https://doi.org/10.1371/journal.pcbi.1006874.t001

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 [28] 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, 28]. Correlations within single chains recapitulate previously reported results for the $\beta$ [12] and $\alpha$ [29] 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 inserted base pairs [30], 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 $\alpha$ and $\beta$ chains (Fig 2A and S2 Fig for an analysis of statistical significance), with a total sum of 0.36 bits, about 10 times lower than the total intra-gene correlations of the $\alpha$ chain. The largest correlation is between the choice of $V_\alpha$ and $V_\beta$ genes (0.036 bits) and $J_\alpha$ and $V_\beta$ 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 correlations 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. An analysis of the correlations between gene segments (S3 Fig) does not show any particular structure.

Pairs of $\beta$ chains show almost no correlations (Fig 2C and S2 Fig for an analysis of statistical significance). Looking in detail at the correlations between gene segments reveals a strongly negative correlation of TCRBV21-01 and TCRBV23-01 (both pseudogenes) with themselves (S4 Fig), which is expected because at least one of the two $\beta$ 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 $\beta$ 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 $\alpha$-$\alpha$ 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 $\alpha$ recombination events occur processively and simultaneously on the two chromosomes, as we analyse in more detail below.

## Correlations between $\alpha$ chains can be explained by a rescue mechanism

We wondered whether the detailed structure of the observed correlations between the $\alpha$ chains on the two chromosomes could be explained by a simple model of recombination rescue. The correlations of the $V_\alpha$ segments on the two chromosomes and of the $J_\alpha$ 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 correlations between the $V_\alpha$ gene segment on the first chromosome and the $J_\alpha$ on the second chromosome also show a similar diagonal structure (S5 Fig).

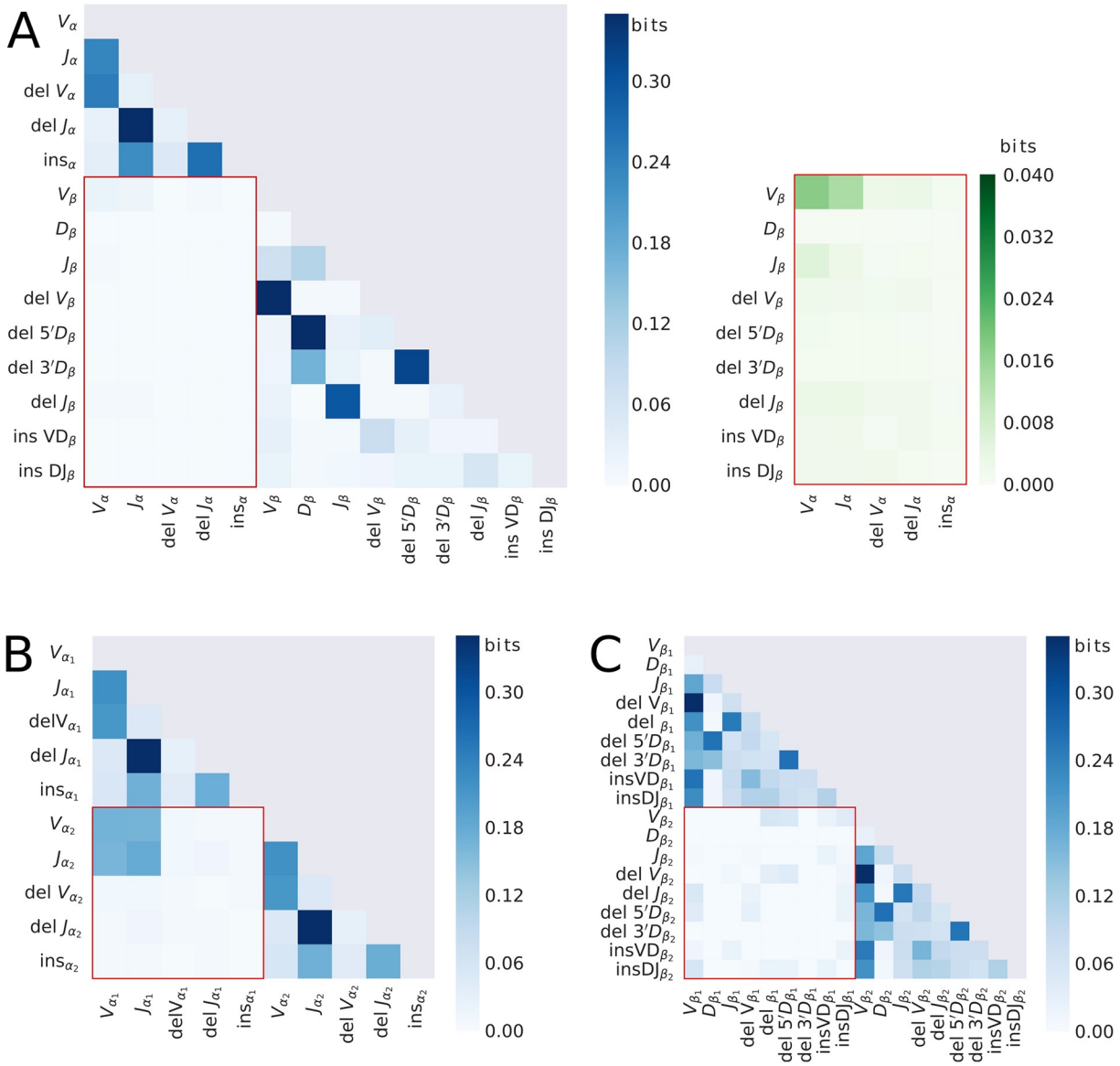


**Fig 2. Mutual information (a non-parametric measure of correlations) between the recombination events of the paired chains.** V, D, and J segment choice, numbers of bases deleted from the 3' end of the V-gene (delV), the 5′ end of the J-gene (delJ), and both ends of the D-gene for the $\beta$ chain (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 $\alpha$ chain, and between V and D (insVD) and between J and D (insDJ) segments for the $\beta$ chain. Mutual information for (**A**) $\alpha$-$\beta$ pairs (on the right in green: close-up of the inter-chain mutual information); (**B**) $\alpha$-$\alpha$ pairs; and (**C**) $\beta$-$\beta$ pairs. Inter-chain correlations are highlighted by red boxes. To remove systematic biases in mutual information estimation from finite data, the mutual information of shuffled data was subtracted (see Methods). For a statistical anlaysis of the significance of the reported mutual informations, see S2 Fig.

https://doi.org/10.1371/journal.pcbi.1006874.g002

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 recombination of a functional chain on either of the chromosomes immediately stops the process on both chromosomes. 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 V$\alpha$ and J$\alpha$ gene usages of the two chains.

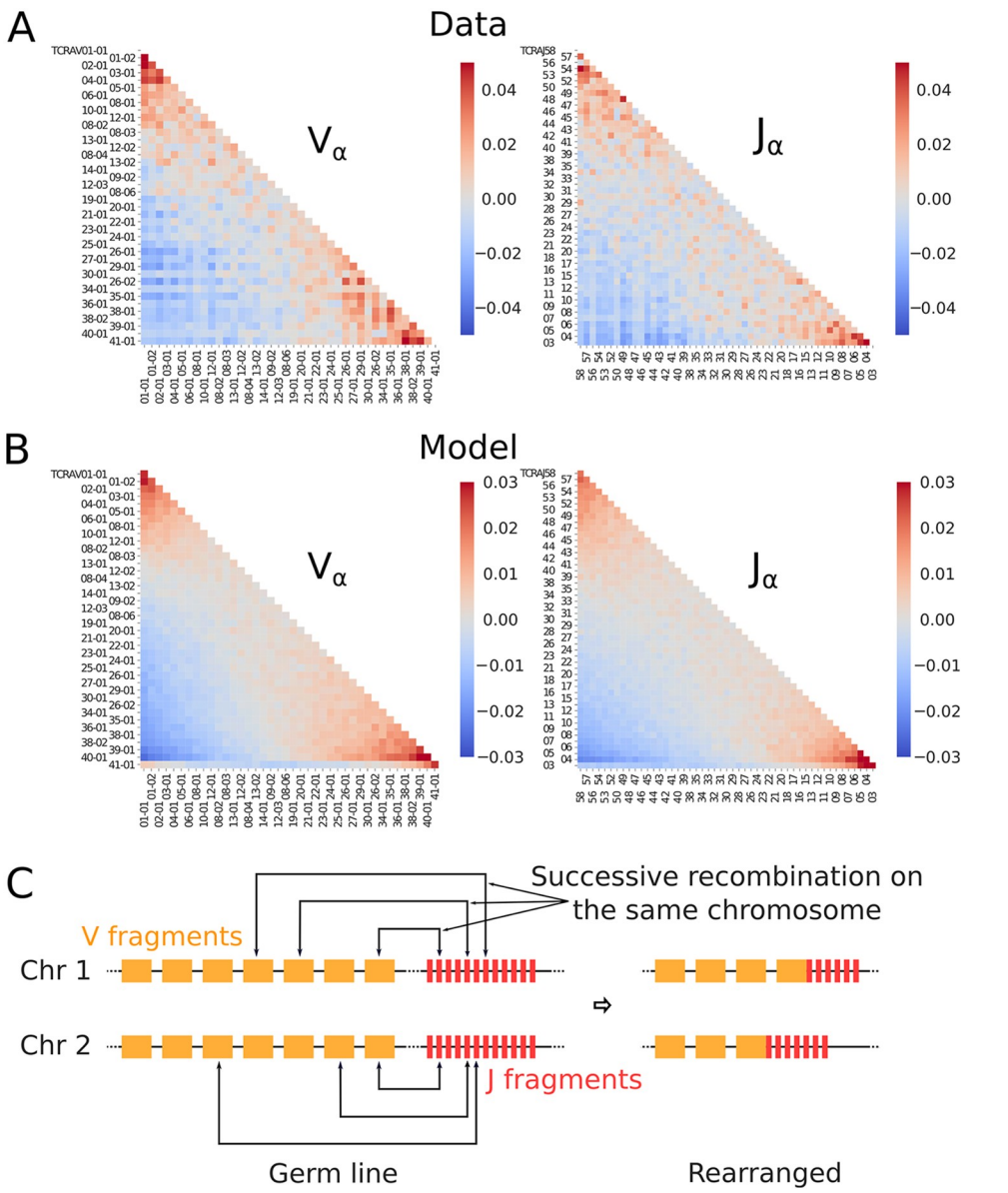


**Fig 3. Evidence of the rescue mechanism.** (A) Pearson correlation between V and J gene segment usage for TCR$\alpha$. The correlation is taken between the truth values of particular V and J gene choices (a value of 1 is assigned if a given segment is observed and 0 if it is not, see Methods for details). (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 rearrangement, recombination stops. In the end, the V and J gene segments selected on both chromosomes are close to each other in the germline ordering.

https://doi.org/10.1371/journal.pcbi.1006874.g003

To validate this hypothesis, we simulated a minimal model of the rescue process similar to [26] (Methods), in which the two chromosomes are recombined in parallel. If recombination happens to fail on both chromosomes, repeated "rescue" recombinations (which we limit to 5) take place between outward nearby segments (Fig 3C). The covariance matrices obtained from the simulations for both $V_\alpha$ and $J_\alpha$ (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. However, the actual distributions of V and J genes segments (see S6 Fig) are much more

heterogeneous than the slowly decaying distribution implied by our simple model: the question of gene usage is further complicated by other factors, such as gene accessibility and primer specificity.

## 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. While the authors of [9] also provided a genomic DNA (gDNA) dataset that does not have this issue, the number of sequences was too small to resolve statistically significant pairings. Nonetheless, we can derive strict bounds from the proportion of productive sequences found in this (unpaired) gDNA dataset. Following recombination, using IGoR we estimate $p^{\alpha}_{\mathrm{nc}} = 69.5\%$ of the $\alpha$ sequences, and $p^{\beta}_{\mathrm{nc}} = 73.5\%$ of $\beta$ 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 $p^{\alpha,\beta}_{\mathrm{c}} = 1 - p^{\alpha,\beta}_{\mathrm{nc}}$ of random rearrangements. We denote by $p^{\alpha}_{\mathrm{f}}$ and $p^{\beta}_{\mathrm{f}}$ the probability that a coding sequence can express a functional $\alpha$ or $\beta$ 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 $p'_{\mathrm{r}}$ when the first recombination succeeds. Then, the proportion $f_{\mathrm{nc}}$ of observed non-coding sequences can be written as (see tree in Fig 4 and Methods):

$$f_{\mathrm{nc}} = \frac{(p_{\mathrm{r}} + p'_{\mathrm{r}})p_{\mathrm{nc}}}{1 + p'_{\mathrm{r}} + 2(1 - p_{\mathrm{f}}p_{\mathrm{c}})p_{\mathrm{r}}}. \tag{1}$$

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 $\beta$ chain sequences calculated from Eq 1 with the values from gDNA data ($f^{\beta}_{\mathrm{nc}} = 18 \pm 1\%$ in [9] and 14% in [11]), allows us to constrain the values of $p^{\beta}_{\mathrm{r}}$ and $p^{\beta\prime}_{\mathrm{r}}$. The probability of a second recombination, even if the first recombination failed, is always lower than 65% (Fig 4A). By constrast, the observed fraction of non-coding sequences in the $\alpha$ chain, $f^{\alpha}_{\mathrm{nc}} = 40 \pm 1\%$, constrains the the rescue probabilities $p^{\alpha}_{\mathrm{r}}$ and $p^{\alpha\prime}_{\mathrm{r}}$ to be close to 100% (Fig 4B), in agreement with the fact that both chromosomes are believed to recombine independenly. Assuming strict independence, $p^{\alpha}_{\mathrm{r}} = p^{\alpha\prime}_{\mathrm{r}} = 1$ puts bounds on the probability that a random coding $\alpha$ sequence is functional, $70\% \leq p^{\alpha}_{\mathrm{f}} \leq 100\%$.

## Fraction of cells with two functional $\alpha$ 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 limited 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 beingfound and 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. We use both the second and the third experiments, which are more data-rich than the

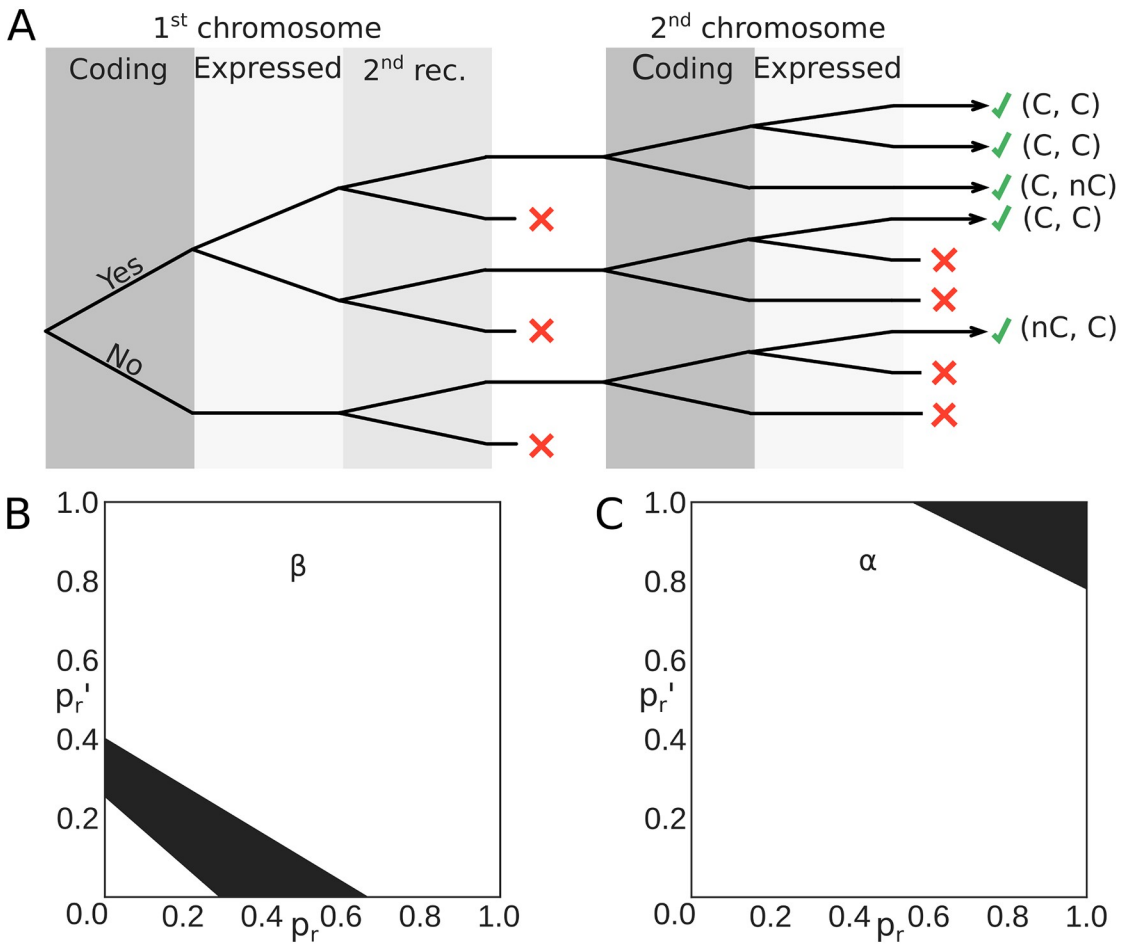


**Fig 4. Probability of recombination of the second chromosome.** (A) Decision tree of the recombination process for one chain ($\alpha$ or $\beta$). The first part shows the recombination of the first chromosome, the second part of the second chromosome. In each area a binary choice is made. Red crosses 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 $\beta$ 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 $\beta$ if the first was successful) and $p_r$ (probability of a second recombination for $\beta$ if the first was not successful). The bounds were obtained by imposing $0 < p_f^\beta < 1$ in Eq 1. (C) Bounds on the allowed values of rescue probabilities for the $\alpha$ chain. They are consistent with both chromosomes recombining simultaneously and independently, $p_r = p_r' = 1$.

https://doi.org/10.1371/journal.pcbi.1006874.g004

first one. The total read count of each sequence is obtained by summing its read count in individual wells.

Sequences of chains paired with a non-coding chain of the same type must be functional and expressed on the surface of the cell. These sequences are more expressed than non-coding ones, and their distributions of read counts are markedly different (S7A Fig).

Comparatively, coding sequences paired with a coding sequence of the same type 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 (S7B and S7E Fig) the latter being assumed to follow the same distribution as noncoding sequences. The best parameter fit of this mixture to the read counts of paired coding sequences (S7C and S7D Fig) yields the total proportion $p_e$ of functional sequences among the coding sequences coupled with another coding sequence. For $\alpha$ sequences, we

found $p^{\alpha}_{e,exp2} = 0.66 \pm 0.03$ for experiment 3 and $p^{\alpha}_{e,exp3} = 0.69 \pm 3$, meaning that around $2p^{\alpha}_{e} - 1 \sim 0.35 \pm 0.1$ of cells express two different $\alpha$ chains (see Methods). This number is consistent with older results [31], but higher than a recent estimate of 14% based on single-cell sequencing [8]. Another estimate from the same data [31], but taking into account material loss (see Methods), suggests that 24 ± 5% of cells have two functional and expressed $\alpha$ chains, more consistent with our own estimate. Of course, one of the major limitations of this method is that it only applies to relatively large clones, that can be paired by the PairSeq method, and it cannot be excluded that this ratio differs in naive cells for example. It should also be noted that the (fitted) mixture distribution and the original distribution do not coincide exactly (S7C and S7F Fig), this could be due to imperfect silencing or to a difference in the expression levels of non-coding and silenced sequences.

For $\beta$ chains, the fit is noisier, because non-coding sequences are much more suppressed and therefore scarcer than for the $\alpha$ 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 $\beta$ 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 2 we report the differences between the two ensembles in terms of CDR3 length (defined from the conserved cystein of V and the conserved phenylalanine or tryptophan of J, corresponding to IMGT positions 105 to 117) 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 2 and S8 Fig). 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 $\alpha$ or $\beta$ 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.

**Table 2. CDR3 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.** The error on the standard deviation of the length (estimated by bootstrap) is always lower than 0.2.

| chain | CDR3 length: mean ± st. deviation (nt) | | | Gene | KL divergence (bits) | |
|---|---|---|---|---|---|---|
| | functional | coding | non-coding | | functional | coding |
| $\alpha$ | 42.0 ± 5.00 | 39.12 ± 6.67 | 40.0 ± 7.00 | $V_\alpha$ | 0.66 ± 0.05 | 1.39 ± 0.01 |
| | | | | $J_\alpha$ | 0.110 ± 0.005 | 0.119 ± 0.004 |
| $\beta$ | 44.1 ± 5.03 | 43.17 ± 6.22 | 43.4 ± 7.82 | $V_\beta$ | 1.09 ± 0.06 | 1.03 ± 0.18 |
| | | | | $J_\beta$ | 0.12 ± 0.004 | 0.051 ± 0.008 |

https://doi.org/10.1371/journal.pcbi.1006874.t002

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 2), offering a more contrasted view. $V_\beta$ and $J_\alpha$ usages are similar in functional and coding sequences in terms of their divergence with non-coding sequences. For $J_\beta$ however, this divergence is higher in functional than in simply coding sequences, while the opposite is true for $V_\alpha$.

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

Ignoring small correlations between features of the $\alpha$ and $\beta$ 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 [28] for each paired chain in our datasets. The distribution of the pair generation probabilities obtained in this way (Fig 5A) 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 $\alpha$ and $\beta$ 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, only a negligible number of full TCR sequences will be shared in samples obtained from distinct individuals. To make that prediction more quantitative, we simulated a computational model of sequence generation followed by thymic selection. $\alpha$ and $\beta$ 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 $q_\alpha = 0.16$ (S9 Fig).

Using the model, we can make predictions about the expected number of TCR$\alpha\beta$ nucleotide sequences shared between any of 10 individuals (Fig 5B) for which a million unique

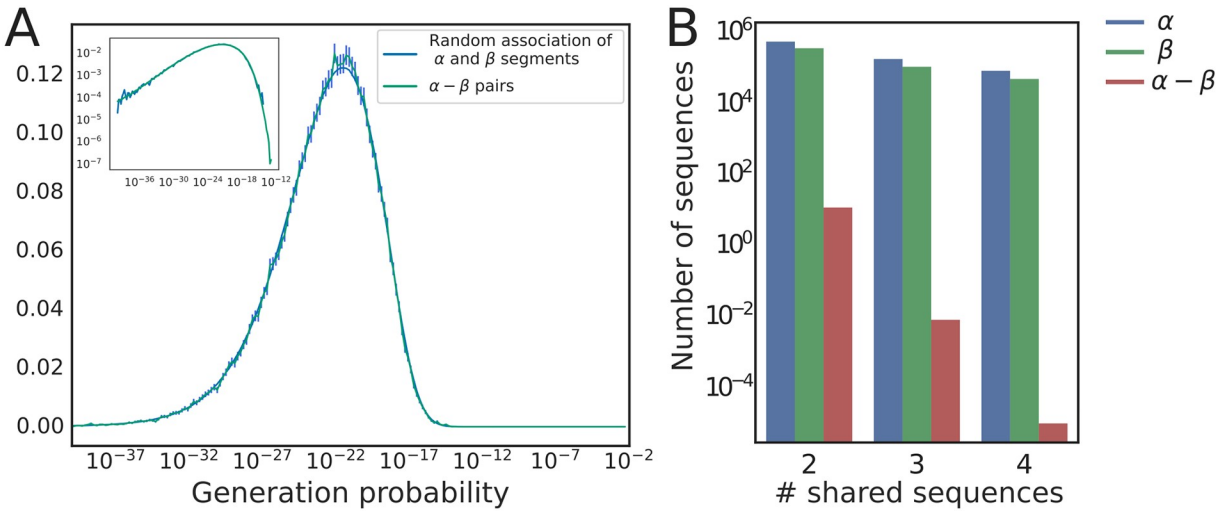


**Fig 5. Generation probability of a full $\alpha\beta$ TCR.** (A) Distribution of the generation probabilities of $\alpha\beta$ pairs, obtained by multiplying the generation probabilities of the $\alpha$ and $\beta$ sequences. The graph shows the distribution for paired sequences (blue) and random associations of $\alpha\beta$ pairs (green). The error bars represent three standard deviations, and the inset shows the same plot on a double logarithmic scale. (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.

https://doi.org/10.1371/journal.pcbi.1006874.g005

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. A concomitant question concerns the total number of TCR sequences shared between two individuals. This number does not depend on selection or sample size, but rather on the total number of different clonotypes in an individual. While this last quantity is not precisely known, estimates range between $10^8$ and $10^{11}$ [13, 34]. Using the analytical formulas and numerical procedure described in [17] with these estimates of the repertoire size, we predict the proportion of shared clonotypes between two individuals to fall between 0.001% and 0.1% of their full repertoires (see Methods for details).

### Co-activation of cells sharing the same $\beta$ chain

To further investigate the effects of convergent selection, we quantified how often the same $\alpha$ chain was associated with distinct $\beta$ chains in different clones (S10A Fig), and *vice versa* (S10B Fig). While association of a $\beta$ with 2 distinct $\alpha$ chains could happen in the same cell because of the existence of two copies, we found a substantial fraction (3%) of all paired TCR$\beta$ that could be associated with three or more TCR$\alpha$.

Convergent recombination of $\beta$ can create clones that shares their $\beta$ but not their $\alpha$ chains. This effect can be quantified using the generation and thymic selection model introduced in the previous paragraph. Simulations with the same sample sizes as the data show that such convergent recombination is predicted to happen with a rate of 0.5%, and thus cannot explain the data. However, there is another effect at play: cells divide around 5 times between $\beta$ and $\alpha$ recombination, which leads to clones with the same $\beta$ chain but with up to $2^5 \sim 30$ distinct $\alpha$ chains. A simulation considering these two effects together (see Methods) predicts a sharing fraction of 3%, consistent with the fraction observed empirically.

## Discussion

Analysing computationally reconstructed pairs of TCR $\alpha$ and $\beta$ chains, as well as $\alpha$-$\alpha$ and $\beta$-$\beta$ 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 $\alpha$ 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 [26], but differs in its details and parameters, as the original model could not reproduce the correlation pattern of the data.

We estimated that $\sim 35 \pm 10\%$ of cells express two $\alpha$ chain, higher than a recent report of 14% using single-cell sequencing [8]. However, this fraction is very hard to assess directly, as material loss can lead to its underestimation—in fact, correcting for this loss with minimal assumptions gives a fraction $\sim 24\%$ instead of 14%. 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 $V_\beta$ and $(V_\alpha, J_\alpha)$ usage—is in agreement with recent observations using direct single-cell chain pairing [10]. While independence between the $\alpha$ and $\beta$ 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 $\alpha$ chain, and 38 bits for the $\beta$ 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 samples from different individuals. Our analysis predicts that sharing of $\alpha\beta$ pairs between two samples should be rare, and that sharing between more than two is exceptional. In a recent report [10], 26 TCR$\alpha\beta$ pairs were found to be shared between any 2 of 5 individuals. Our result indicate that such a high level of sharing cannot be explained by convergent recombination alone: by simulating samples of the same size as in [10], we estimated a total expected number of 0.001 sequences between all their pairs (see Methods). The much higher number of shared sequences reported in the original study may result from over-correcting for sequencing errors, or alternatively from strong convergent selection in all 5 donors. A clonotype expansion of $10^4$ (not unexpected in the context of an immune response, see e.g. [35]) would be sufficient to explain this result.

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 [28]. The IGoR software is able to learn, from out-of-frame receptor sequences, the statistics of a V(D)J recombination process. We don't use IGoR in its inference capacity here, but rather rely on the pre-inferred recombination model for TRA and TRB chains in humans supplied with IGoR, as the recombination process is widely shared between individuals [12]. Briefly, the probabilities of recombination of $\alpha$ and $\beta$ chains factorize as:

$$P^{\alpha}_{\text{recomb}} = P(V,J)P(\text{del}V|V)P(\text{del}J|J)P(\text{ins}VJ)\prod_{i}^{\text{ins}VJ} P_{VJ}(n_i|n_{i-1}), \tag{2}$$

$$P^{\beta}_{\text{recomb}} = P(V,D,J)P(\text{del}V|V)P(\text{ins}VJ)P(\text{del}D5'\text{del}D3'|D)P(\text{ins}DJ) \tag{3}$$

$$\times P(\text{del}J|J)\prod_{i}^{\text{ins}VD} P_{VD}(m_i|m_{i-1})\prod_{i}^{\text{ins}DJ} P_{DJ}(r_i|r_{i-1}), \tag{4}$$

where $(n_i)$, $(m_i)$, $(r_i)$ are the inserted nucleotides at the VJ, VD, and DJ junctions. IGoR infers these probabilities through an Expectation-Maximization algorithm as described previously.

We rely on IGoR for:

- The generation of synthetic sequences with the same statistic as V(D)J recombination, which we use to predict sharing between individuals.

- The computation of the probability of generation of a sequence $s$ by summing over all the scenarios that are compatible with it, $P_{\text{gen}}(s) = \sum_{\text{scenario}\to s} P_{\text{recomb}}(\text{scenario})$, which allows us to generate Fig 5. We also use this feature to predict sharing between very number large of sequences using Eq 7 (see [17] for details).

## 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 first select all the pairs under a given p-value threshold ($10^{-4}$). For $\alpha$–$\alpha$ and $\beta$–$\beta$ pairs, we apply a threshold on their Levenshtein distances in order to remove most of the false pairings (pairing of near identical sequences due to sequencing errors). Then for each pair of well occupation numbers ($w_1$, $w_2$), we set the p-value threshold so that false discovery rate (using the Benjamini—Hochberg procedure) is always less than 1%. Compared to the analysis of Ref. [9], where the discreteness of the p-value distribution was taken into account by using a permutation algorithm, our approach is more conservative, as we worried about the potential effect of fake pairings on the false discovery rate. Thus our reported number of pairs (Table 1) slightly differs from that reported in the original study.

## Information quantities

The mutual information (in bits) of two variables $X$, $Y$ with a joint distribution $p(x, y)$ is defined by: $I(X, Y) = \Sigma_{x,y}\, p(x, y)\log_2[p(x, y)/(p(x)p(y))]$. 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, a better correction can be obtained by computing the mutual information between shuffled sequences, where the two chains are assorted at random.

The Kullback-Leibler divergence between two distribution $p(x)$ and $q(x)$ of a variable $X$ is given by: $D_{\text{KL}}(\text{p}\|\text{q}) = \Sigma_x\, p(x)\log_2(p(x)/q(x))$.

## 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], after a success however, on any of the two chromosomes, it stops. This model is similar to that of [26], 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.

## Simple model of selection based on the V genes segments

We have shown that the pairs $V_\alpha - V_\beta$ and $J_\alpha - V_\beta$ were not independent (Fig 2). In this section we define the simplest model that can reproduce these correlations. The marginal distributions $p_{V_\alpha,J_\alpha}$ and $p_{V_\beta}$, coupled with the experimental pairing data can be used to obtain selection factors $q_{V_\alpha,J_\alpha V_\beta}$:

$$p(V_\alpha, J_\alpha, V_\beta) = p_{V_\alpha,J_\alpha}\ p_{V_\beta}\ q_{V_\alpha,J_\alpha,V_\beta} \tag{5}$$

By adding a tunable temperature, we can modify the level of selection we want to observe:

$$p(V_\alpha, J_\alpha, V_\beta) \propto p_{V_\alpha}\ p_{V_\beta}\ (q_{V_\alpha,V_\beta})^{1/T} \tag{6}$$

When $T \to 0$, the selection conserves only a few specific pairs of V, while for $T \to \infty$ there is no selection. This modifies the mutual information between $V_\alpha$ and $V_\beta$ in the same cell, but also, because $V$ and $J$ on the same chromosome are not independent, the mutual information between $V_\alpha$ and $J_\beta$. In S11 Fig, we show the evolution of the mutual information between $V_\alpha$, $J_\alpha$, $V_\beta$ and $J_\beta$ as a function of $T$. The model underestimates the mutual information between $V_\alpha$ and $J_\beta$ which hints that it may be necessary to also include $J_\beta$ in the selection model.

## Copy number distributions

In order to estimate the ration of expressed sequences in a set of coding chains, we fit the empirical distribution of reads per coding chain, $\rho_{\mathrm{c}}$, with a mixture of two distributions (S7 Fig): $\rho_{\mathrm{e}}$, corresponding to chain sequences that could be paired with a non-coding sequence of the same type and thus believed to be expressed; and $\rho_{\mathrm{nc}}$ corresponding to non-expressed sequences and learned from non-coding sequences. Each distribution is estimating by taking histograms with bin size chosen using the Freedman-Draconis rule. For a given parameter $p_{\mathrm{e}}$, the mixture distribution is obtained by sampling $Np_{\mathrm{e}}$ expressed chains and $N(1 - p_{\mathrm{e}})$ non-coding sequences, with $N$ large. The fit is done by minimizing the (two-sample) Kolmogorov-Smirnov (KS) distance between the two distributions, and the error bars are obtained through bootstrapping. The result of the fit is a parameter $p_{\mathrm{e}}$ corresponding to the proportion of expressed sequences among the chains. Applying this method to coding chains paired with another coding chain, we can infer $p_{2\alpha}$, the proportion of cells with two expressed $\alpha$. The relation between $p_{2\alpha}$ and $p_{\mathrm{e}}$ should be $p_{\mathrm{e}} = (2p_{2\alpha} + (1 - p_{2\alpha}))/2$, hence $p_{2\alpha} = 2p_e - 1$. A different approach to estimate $p_{\mathrm{e}}$ from the data consists in comparing the mean of the distributions. While this gives poor results with raw data due to the long tail of the distributions, it matches the distance-minimization result when the distributions are log-transformed. We find a value $p_e^\alpha = 28\% \pm 10\%$, not compatible with the value of 14% ± 3% obtained in [8] (19 out of 139 cells in which at least one productive sequence was found). But the authors of [8] make their estimate by sequencing cDNA, which can lead to different drop-out rates depending on the nature of the sequence. Silenced productive sequences or non-productive sequences are less expressed and their drop-out rates are higher. They find two TCRA (productive or not) in

only 58% of cells, while both TCRA are expected to recombine [39]. In this context the 14% rate can only be understood as a lower bound. Assuming that non-productive and silenced sequences are expressed in similar quantities, we obtain an estimate for $p_e^x$ of 24% ± 5% (19 out of the 80 cells which had two sequences, productive or not) from their data, which is consistent with our result.

## Sharing estimation

We follow the methods of [17]. A large number of productive $\alpha$ and $\beta$ chain pair sequences are generated through a stochastic model of recombination using IGoR [28]. 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 $q_\alpha$ and $q_\beta$ 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 (S5 Fig), using the analytical expressions given in [17]. The predictions for the number of shared TCR$\alpha\beta$ nucleotide sequences reported in Fig 5B, as well as the estimation of the sharing between the full repertoire of two individuals, are computed using the analytical expressions of [17]. If $N$ sequences are sampled in $m$ individuals, the expected number of sequences which will be found in exactly $k$ individuals is:

$$M_{k,m}(N) = \int_0^\infty \mathrm{d}p P(p) \binom{m}{k} e^{Np(m-k)} (1 - e^{-Np})^k \tag{7}$$

Without selection $P(p)$ is the probability density function for of sequences probabilities. We used this formula with $p = P_{\mathrm{gen}}/q$ for selected sequences, and $p = 0$ otherwise. The integral in Eq 7 is evaluated using a Monte Carlo simulation. Derivations and details about the Monte Carlo simulation can be found in [17]. We use this formula to estimate the proportion of full receptors shared between two individuals.

## $\beta$ sharing

The results of [17] can also be used to estimate the theoretical proportion of clonotypes sharing a $\beta$ in a sample of size $N$. This sharing is due to two phenomena: the possibility of generating twice the same $\beta$ sequence and the division stage between the recombination of $\beta$ and $\alpha$. To simulate the first mechanism we can, following [17], generate an important number of $\beta$ sequences (in-frame, no-stop codons) with IGoR, associate to each of them a hash between 0 and 1 and then only keep the sequences whose hash is lower than $q_\beta$ to simulate the selection. The cellular division between $\beta$ and $\alpha$ recombination creates 30 cells with the same $\beta$ and different $\alpha$. Some of these cells won't have a functional $\alpha$ receptors, while others will not pass selection, while there is no precise way to quantify how many cells survive, we can consider an estimate of roughly $n_d \approx 10$ cells. Because the probability $p(s)$ of generating a given sequence is so low, this increase in cell number multiplies $p(s)$ by $n_d$, hence corresponds to a change $q_\beta \to q_\beta/n_d$. Then, for $10^5$ sequences and $n_d = 10$, we find that $\approx 3\%$ of clonotypes are expected to share their $\beta$ sequence with another TCR.

# Supporting information

**S1 Fig. Hamming distance between two TCR$\beta$ 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 TCR$\alpha$ chains. A threshold of 11 was

chosen to exclude pairs from sequencing errors from the analysis, while retaining as many pairs as possible, including some with the same gene usage.
(TIF)

**S2 Fig. Comparison between the observed mutual information and the null for $\alpha$-$\beta$ (A) and $\beta$-$\beta$ pairs (B).** The null distribution is obtained by shuffling the pairs, the error-bar represents the standard deviation over multiple shuffling. We consider the raw mutual information, not corrected with the shuffled distribution, contrary to Fig 2. With a false discovery rate of 0.01 (using the Benjamini—Hochberg procedure) and assuming a Gaussian distribution for the mutual information of shuffled sequences, we find that, for $\beta - \beta$ pairings, the only pairs of features passing the test are (in order of significance) $V_1 - V_2$, $V_1 - \mathrm{Ins}DJ_2$ and $\mathrm{Del}3'D_1 - \mathrm{Ins}DV_2$. By contrast, for $\alpha - \beta$ pairing, with the same false discovery rate (0.01), 36 out of the 45 possible feature pairings are significant.
(TIF)

**S3 Fig. Pearson correlation coefficient between TCRA and TCRB genes.** $V_\alpha - V_\beta$ (**A**), $V_\alpha - J_\beta$ (**B**), $J_\alpha - V_\beta$ (**C**) and $J_\alpha - J_\beta$ (**D**). The correlation are generically small and do not show a particular structure.
(TIF)

**S4 Fig. Normalized covariance between V (left) and J (right) gene usages of pairs of $\beta$ sequences found in the same clone.** The V21-01 and V23-01 genes are non-functional pseudogenes and are thus anticorrelated.
(TIF)

**S5 Fig. Pearson correlation between the $V_\alpha$ gene segment on the first chromosome and the $J_\alpha$ gene segment on the second chromosome.** The correlations observed in Fig 3A and 3B are also observed here.
(TIF)

**S6 Fig. Distribution of the V and J gene segments.** In both case, they are ordered along the germline, 5' to 3'.
(TIF)

**S7 Fig. Distribution of the number of reads of different types of TCR$\alpha$ RNA sequences.** (A) displays the distribution (normalized histogram and kernel density estimation) of the total number of read counts (all wells summed) of subsets of paired TCR sequences in experiments 2 and 3. The blue histograms look only at the sequences which are paired and non-coding, while the yellow ones focus on sequences paired with a non-coding sequence, hence expected to be expressed. The histograms are normalized so that the area under them is equal to one. The bin width is chosen using the Freedman-Draconis rule. (B) (resp. (E)) shows the distribution of the log-transformed read counts for experiment 3 (resp. 2). In blue, paired non-coding sequences and in yellow functional sequences again. The green histogram corresponds to coding sequences paired with another coding sequence (CC). This last type of sequences contains both expressed and silenced sequences, the distribution of its read counts should be a mixture of the two other distributions. The parameter $p_e^\alpha$ of this mixture can be related to the proportion of cells exhibiting two functional TCR$\alpha$ chains (see Methods). In plot (C) (exp. 3) and (F) (exp. 2), the mixture distribution, with parameter $p_e^\alpha$ minimizing the Kolmogorov-Smirnov (KS) distance between the two distributions, is represented in black, while the distribution (CC) is shown in green. Plots (D) and (G) show (for experiments 3 and 2 respectively), the KS distance between the mixture distribution and the (CC) distribution for different values of the parameter $p_e^\alpha$. The fit did not depend significantly on the bin width of the histograms. The

black vertical line corresponds to the value of $p_e^z$ giving the minimum distance, respectively 0.66 ± 0.03 and 0.69 ± 0.03 in Exp. 2 and Exp. 3.
(TIF)

**S8 Fig. CDR3 length distribution of expressed and out-of-frame TCR$\alpha$ sequences.** Expressed sequences have a narrowed distribution than unselected ones. All sequences used in these distributions were paired.
(TIF)

**S9 Fig. Number of unique amino-acid (translated) sequences as a function of the number of unique nucleotide sequences for (A) $\alpha$ and (B) $\beta$ chains.** Red crosses are experimental data, blue line comes from simulations of the recombination model with random selection. For $\alpha$ the value of $q$ is inferred by least-square minimisation to be $q_\alpha = 0.16$, while for $\beta$ we used the value of $q_\beta = 0.037$ reported in Elhanati et al., *Immunological Reviews*, in press (2018).
(TIF)

**S10 Fig. (A) Distribution of the number of distinct $\alpha$ sequences that could be paired with a given $\beta$ sequence. (B) Distribution of the number of distinct $\beta$ sequences that could be paired with a given $\alpha$ 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.
(TIF)

**S11 Fig. The full blue (resp. yellow, green) line represent the mutual information between $V_\alpha/V_\beta$ (resp. $V_\beta/J_\alpha$, $J_\beta/V_\alpha$), as a function of temperature $T$, as described in the Methods section.** The dot are the observed values in the dataset.
(TIF)

## Author Contributions

**Conceptualization:** Aleksandra M. Walczak, Thierry Mora.

**Data curation:** Thomas Dupic.

**Formal analysis:** Thomas Dupic, Quentin Marcou, Aleksandra M. Walczak, Thierry Mora.

**Investigation:** Thomas Dupic.

**Methodology:** Thomas Dupic, Aleksandra M. Walczak, Thierry Mora.

**Supervision:** Aleksandra M. Walczak, Thierry Mora.

**Visualization:** Thomas Dupic.

**Writing – original draft:** Thomas Dupic, Aleksandra M. Walczak, Thierry Mora.

**Writing – review & editing:** Aleksandra M. Walczak, Thierry Mora.

## References

1. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, et al. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood. 2009; 114(19):4099–4107. https://doi.org/10.1182/blood-2009-04-217604 PMID: 19706884
2. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel {VDJ} pyrosequencing. Sci Transl Med. 2009; 1(12):12ra23. https://doi.org/10.1126/scitranslmed.3000540 PMID: 20161664

3. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S. Rep-Seq: Uncovering the immunological repertoire through next-generation sequencing; 2012. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3311040&tool=pmcentrez&rendertype=abstract.
4. Robins H. Immunosequencing: applications of immune repertoire deep sequencing. Curr Opin Immunol. 2013; 25(5):646–652. https://doi.org/10.1016/j.coi.2013.09.017 PMID: 24140071
5. Six A, Mariotti-Ferrandiz ME, Chaara W, Magadan S, Pham HPP, Lefranc MPP, et al. The past, present and future of immune repertoire biology—the rise of next-generation repertoire analysis. Front Immunol. 2013; 4(413):413. https://doi.org/10.3389/fimmu.2013.00413 PMID: 24348479
6. Kim SM, Bhonsle L, Besgen P, Nickel J, Backes A, Held K, et al. Analysis of the paired TCR α- and β-chains of single human T cells. PLoS One. 2012; 7(5). https://doi.org/10.1371/journal.pone.0037338
7. Turchaninova Ma, Britanova OV, Bolotin Da, Shugay M, Putintseva EV, Staroverov DB, et al. Pairing of T-cell receptor chains via emulsion PCR. Eur J Immunol. 2013; 43(9):2507–2515. https://doi.org/10.1002/eji.201343453 PMID: 23696157
8. Han A, Glanville J, Hansmann L, Davis MM. Linking T-cell receptor sequence to functional phenotype at the single-cell level. Nat Biotechnol. 2014; 32(7):684–692. https://doi.org/10.1038/nbt.2938 PMID: 24952902
9. Howie B, Sherwood AM, Berkebile AD, Berka J, Emerson RO, Williamson DW, et al. High-Throughput Pairing of T Cell Receptor α and β Sequences. Science Translational Medicine. 2015; 7(301):301ra131–301ra131. https://doi.org/10.1126/scitranslmed.aac5624 PMID: 26290413
10. Grigaityte K, Carter JA, Goldfless SJ, Jeffery EW, Ronald J, Jiang Y, et al. Single-cell sequencing reveals αβ chain pairing shapes the T cell repertoire. 2017.
11. Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, et al. Overlap and effective size of the human CD8+ T cell receptor repertoire. Sci Transl Med. 2010; 2(47):47ra64. https://doi.org/10.1126/scitranslmed.3001442 PMID: 20811043
12. Murugan A, Mora T, Walczak AM, Callan CG. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. Proc Natl Acad Sci. 2012; 109(40):16161–16166. https://doi.org/10.1073/pnas.1212755109 PMID: 22988065
13. Qi Q, Liu Y, Cheng Y, Glanville J, Zhang D, Lee JY, et al. Diversity and clonal selection in the human T-cell repertoire. Proc Natl Acad Sci U S A. 2014; 111(36):13139–44. https://doi.org/10.1073/pnas.1409155111 PMID: 25157137
14. Elhanati Y, Murugan A, Callan CG, Mora T, Walczak AM. Quantifying selection in immune receptor repertoires. Proc Natl Acad Sci. 2014; 111(27):9875–9880. https://doi.org/10.1073/pnas.1409572111 PMID: 24941953
15. Mora T, Walczak A. Quantifying lymphocyte receptor diversity. In: Das JD, Jayaprakash C, editors. Syst. Immunol. CRC Press; 2018. p. 185–199. Available from: http://arxiv.org/abs/1604.00487.
16. Pogorelyy MV, Elhanati Y, Marcou Q, Sycheva AL, Komech EA, Nazarov VI, et al. Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. PLoS Comput Biol. 2017; 13(7):e1005572. https://doi.org/10.1371/journal.pcbi.1005572 PMID: 28683116
17. Elhanati Y, Sethna Z, Callan CG Jr, Mora T, Walczak AM. Predicting the Spectrum of TCR Repertoire Sharing with a Data-Driven Model of Recombination. Immunological Reviews. 2018;in press. https://doi.org/10.1111/imr.12665 PMID: 29944757
18. Petrie HT, Livak F, Schatz DG, Strasser A, Crispe IN, Shortman K. Multiple Rearrangements in T Cell Receptor Alpha Chain Genes Maximize the Production of Useful Thymocytes. Journal of Experimental Medicine. 1993; 178(2):615–622. https://doi.org/10.1084/jem.178.2.615 PMID: 8393478
19. Stubbington MJT, Lönnberg T, Proserpio V, Clare S, Speak AO, Dougan G, et al. T Cell Fate and Clonality Inference from Single-Cell Transcriptomes. Nature Methods. 2016; 13(4):329. https://doi.org/10.1038/nmeth.3800 PMID: 26950746
20. Eltahla AA, Rizzetto S, Pirozyan MR, Betz-Stablein BD, Venturi V, Kedzierska K, et al. Linking the T Cell Receptor to the Single Cell Transcriptome in Antigen-Specific Human T Cells. Immunology and Cell Biology. 2016; 94(6):604–611. https://doi.org/10.1038/icb.2016.16 PMID: 26860370
21. Davodeau F, Peyrat MA, Romagné F, Necker A, Hallet MM, Vié H, et al. Dual T Cell Receptor Beta Chain Expression on Human T Lymphocytes. Journal of Experimental Medicine. 1995; 181(4):1391–1398. https://doi.org/10.1084/jem.181.4.1391 PMID: 7699325
22. Padovan E, Giachino C, Cella M, Valitutti S, Acuto O, Lanzavecchia A. Normal T Lymphocytes Can Express Two Different T Cell Receptor Beta Chains: Implications for the Mechanism of Allelic Exclusion. Journal of Experimental Medicine. 1995; 181(4):1587–1591. https://doi.org/10.1084/jem.181.4.1587 PMID: 7699339

23. Steinel N, Brady BL, Carpenter AC, Yang-Iott KS, Bassing CH. Post-Transcriptional Silencing of VβDJβCβ Genes Contributes to TCRβ Allelic Exclusion in Mammalian Lymphocytes. Journal of immunology (Baltimore, Md: 1950). 2010; 185(2):1055–1062. https://doi.org/10.4049/jimmunol.0903099

24. Rybakin V, Westernberg L, Fu G, Kim HO, Ampudia J, Sauer K, et al. Allelic Exclusion of TCR α-Chains upon Severe Restriction of Vα Repertoire. PLoS ONE. 2014; 9(12). https://doi.org/10.1371/journal.pone.0114320 PMID: 25500569

25. Niederberger N, Holmberg K, Alam SM, Sakati W, Naramura M, Gu H, et al. Allelic Exclusion of the TCR α-Chain Is an Active Process Requiring TCR-Mediated Signaling and c-Cbl. The Journal of Immunology. 2003; 170(9):4557–4563. https://doi.org/10.4049/jimmunol.170.9.4557 PMID: 12707333

26. Warmflash A, Dinner AR. A Model for TCR Gene Segment Use. The Journal of Immunology. 2006; 177 (6):3857–3864. https://doi.org/10.4049/jimmunol.177.6.3857 PMID: 16951348

27. Casrouge A, Beaudoing E, Dalle S, Pannetier C, Kanellopoulos J, Kourilsky P. Size Estimate of the *Aβ* TCR Repertoire of Naive Mouse Splenocytes. The Journal of Immunology. 2000; 164(11):5782–5787. https://doi.org/10.4049/jimmunol.164.11.5782 PMID: 10820256

28. Marcou Q, Mora T, Walczak AM. High-Throughput Immune Repertoire Analysis with IGoR. Nature Communications. 2018; 9(1):561. https://doi.org/10.1038/s41467-018-02832-w PMID: 29422654

29. Elhanati Y, Marcou Q, Mora T, Walczak AM. RepgenHMM: A dynamic programming tool to infer the rules of immune receptor generation from sequence data. Bioinformatics. 2016; 32(13):1943–1951. https://doi.org/10.1093/bioinformatics/btw112 PMID: 27153709

30. Motea EA, Berdis AJ. Terminal Deoxynucleotidyl Transferase: The Story of a Misguided DNA Polymerase. Biochimica et biophysica acta. 2010; 1804(5):1151–1166. https://doi.org/10.1016/j.bbapap.2009.06.030 PMID: 19596089

31. Padovan E, Casorati G, Dellabona P, Meyer S, Brockhaus M, Lanzavecchia A. Expression of Two T Cell Receptor Alpha Chains: Dual Receptor T Cells. Science. 1993; 262(5132):422–424. https://doi.org/10.1126/science.8211163 PMID: 8211163

32. Elhanati Y, Sethna Z, Marcou Q, Callan CGJ, Mora T, Walczak AM. Inferring processes underlying B-cell repertoire diversity. Philos Trans R Soc Lond, B, Biol Sci. 2015; 370:20140243. https://doi.org/10.1098/rstb.2014.0243 PMID: 26194757

33. Toledano A, Elhanati Y, Benichou J, M WA, Mora T, Louzoun Y. Evidence for shaping of L chain repertoire by structural selection. Front Immunol. 2018; 9:1307. https://doi.org/10.3389/fimmu.2018.01307 PMID: 29988361

34. Lythe G, Callard RE, Hoare R, Molina-París C. How many TCR clonotypes does a body maintain? Journal of Theoretical Biology 2015; 389:214–224. https://doi.org/10.1016/j.jtbi.2015.10.016 PMID: 26546971

35. Pogorelyy MV., Minervina AA., Touzel MP, Sycheva AL., Komech EA., Kovalenko EI., Karganova GG., et al. Precise tracking of vaccine-responding T cell clones reveals convergent and personalized response in identical twins. Proceedings of the National Academy of Sciences. 2018;201809642. https://doi.org/10.1073/pnas.1809642115

36. Steuer R, Kurths J, Daub CO, Weise J, Selbig J. The Mutual Information: Detecting and Evaluating Dependencies between Variables. Bioinformatics. 2002; 18(suppl_2):S231–S240. https://doi.org/10.1093/bioinformatics/18.suppl_2.S231 PMID: 12386007

37. Pasqual N, Gallagher M, Aude-Garcia C, Loiodice M, Thuderoz F, Demongeot J, et al. Quantitative and Qualitative Changes in V-J α Rearrangements During Mouse Thymocytes Differentiation. The Journal of Experimental Medicine. 2002; 196(9):1163–1174.

38. Murphy K, Weaver C. Janeway's Immunobiology, 9th Edition. Garland Science; 2016.

39. Niederberger N, Holmberg K, Alam SM, Sakati W, Naramura M, Gu H, Gascoigne N R J. Allelic Exclusion of the TCR α-Chain Is an Active Process Requiring TCR-Mediated Signaling and c-Cbl. The Journal of Immunology. 2003; 170:4557–4563. https://doi.org/10.4049/jimmunol.170.9.4557 PMID: 12707333