Supporting Information for

Measures of epitope binding degeneracy from T cell receptor repertoires

Andreas Mayer and Curtis G. Callan Jr

E-mail: andreas.mayer@ucl.ac.uk

This PDF file includes:

Supporting Information Text

1. Formal treatment of effect of selection on coincidence statistics

A. Coincidence changes are related to the cross-moments of the selection factors. We are interested in how the probability of coincidences changes as we modify a base measure $P(\sigma)$ with a weighting function $Q(\sigma)$ that represents the effect of selection according to

$$\tilde{P}(\sigma) = Q(\sigma)P(\sigma), \tag{1}$$

with $\langle Q(\sigma)\rangle_{P(\sigma)} = 1$ for normalization of $\tilde{P}$. Plugging in $\tilde{P}$ into the definition of the near-coincidence probability, we have

$$p_C[\tilde{P}](\Delta) = \sum_{\sigma,\sigma'} P(\sigma)P(\sigma')Q(\sigma)Q(\sigma')I_{d(\sigma,\sigma')=\Delta}. \tag{2}$$

This expression can be rewritten formally as an average over randomly picked pairs of sequences,

$$p_C[\tilde{P}](\Delta) = \langle Q(\sigma)Q(\sigma')I_{d(\sigma,\sigma')=\Delta}\rangle_{P(\sigma,\sigma')}, \tag{3}$$

where $P(\sigma,\sigma') = P(\sigma)P(\sigma')$. Only pairs that are at distance Δ contribute to the average, which suggests restricting the average to these pairs. When conditioning on pairs of sequences at distance Δ, the conditional probability of sequences σ,σ' reads

$$P(\sigma,\sigma'|d(\sigma,\sigma') = \Delta) = \frac{P(\sigma,\sigma')I_{d(\sigma,\sigma')=\Delta}}{P(d(\tilde{\sigma},\tilde{\sigma}') = \Delta)}, \tag{4}$$

where $P(d(\tilde{\sigma},\tilde{\sigma}') = \Delta) = \langle I_{d(\tilde{\sigma},\tilde{\sigma}')=\Delta}\rangle_{P(\tilde{\sigma},\tilde{\sigma}')} = p_C[P](\Delta)$ is the near-coincidence probability in the unselected set. The probability of coincidence thus changes upon selection according to

$$p_C[\tilde{P}](\Delta) = \langle Q(\sigma)Q(\sigma')\rangle_{P(\sigma,\sigma'|d(\sigma,\sigma')=\Delta)}\, p_C[P](\Delta), \tag{5}$$

which completes the derivation of Eqn. 3 in the main text.

B. Relation to the covariance of selection factors. To gain intuition into Eqn. 3 we can rewrite the first factor on the left hand side in terms of the covariance of selection factors. In general, $\mathrm{Cov}(X,Y) = \langle XY\rangle - \langle X\rangle\langle Y\rangle$. Thus

$$\langle Q(\sigma)Q(\sigma')\rangle = \mathrm{Cov}(Q(\sigma),Q(\sigma')) + \langle Q(\sigma)\rangle^2, \tag{6}$$

where the covariance is calculated across random pairs at distance Δ, $P(\sigma,\sigma'|d(\sigma,\sigma') = \Delta)$, and the average across the marginal distribution, $\sum_{\sigma'} P(\sigma,\sigma'|d(\sigma,\sigma') = \Delta)$. The normalization of $\tilde{P}$ implies $\langle Q(\sigma)\rangle_{P(\sigma)} = 1$, hence

$$p_C[\tilde{P}](\Delta) = \big(\mathrm{Cov}(Q(\sigma),Q(\sigma')) + 1\big)\, p_C[P](\Delta) \tag{7}$$

when the average of $Q(\sigma)$ over the marginal distribution can be approximated by its simple average over $P(\sigma)$.

To derive the condition for which Eqn. 7 is exact, we define the local neighbor density around a sequence σ,

$$n_\sigma(\Delta) = \sum_{\sigma'} P(\sigma')I_{d(\sigma,\sigma')=\Delta}. \tag{8}$$

Using this definition we can write

$$\langle Q(\sigma)\rangle_{P(\sigma,\sigma'|d(\sigma,\sigma')=\Delta)} = \left\langle Q(\sigma)\sum_{\sigma'} P(\sigma')I_{d(\sigma,\sigma')=\Delta}/p_C[P](\Delta)\right\rangle_{P(\sigma)} = \langle Q(\sigma)n_\sigma(\Delta)\rangle_{P(\sigma)} \,/\, p_c[P](\Delta). \tag{9}$$

When the probability of selecting a sequence is uncorrelated with its neighbor density, we have

$$\langle Q(\sigma)n_\sigma(\Delta)\rangle_{P(\sigma)} = \langle Q(\sigma)\rangle_{P(\sigma)}\langle n_\sigma(\Delta)\rangle_{P(\sigma)} = \langle Q(\sigma)\rangle_{P(\sigma)}p_C[P](\Delta), \tag{10}$$

and thus the averages are the same, $\langle Q(\sigma)\rangle_{P(\sigma,\sigma'|d(\sigma,\sigma')=\Delta)} = \langle Q(\sigma)\rangle_{P(\sigma)}$, and Eqn. 7 is exact.

2. Decomposing coincidences in mixtures

Here we give the formal derivation of Eqn. 10 in the text which describes a situation where the underlying distribution on sequences is a mixture of distributions, each one describing the sequences selected by a particular antigen (or peptide-MHC complex). As discussed in the text, this is a way to describe the memory compartment of the immune system, or the set of TCRs selected by a peptide pool. We define the mixture distribution as

$$P(\sigma) = \sum_{\pi \in \Pi} P(\sigma|\pi)P(\pi), \tag{11}$$

where $P(\pi)$ are the mixtures weights with $\sum_\pi P(\pi) = 1$. To simplify notations, our derivation will use the exact coincidence definition introduced in Eqn. 2, which uses $P(\tau)$ the marginalized distribution over amino acid sequences τ, but we expect the results to hold more generally. We start our derivation by inserting the mixture distribution definition into the formula for the coincidence probability, and split the resulting expression into two terms:

$$p_C[P(\tau)] = \sum_\tau \left(\sum_\pi P(\tau|\pi)P(\pi) \right)^2, \tag{12}$$

$$= \sum_\pi P(\pi)^2 \sum_\tau P(\tau|\pi)^2 + \sum_{\pi_1 \neq \pi_2} P(\pi_1)P(\pi_2) \sum_\tau P(\tau|\pi_1)P(\tau|\pi_2). \tag{13}$$

We identify the last factor in the second term as a generalized coincidence probability for samples drawn from two probability distributions,

$$p_C[P_i(\tau), P_j(\tau)] = \sum_\tau P_i(\tau)P_j(\tau), \tag{14}$$

which allows us to rewrite the coincidence probability s

$$p_C[P(\tau)] = \sum_\pi P(\pi)^2 p_C[P(\tau|\pi)] + \sum_{\pi_1 \neq \pi_2} P(\pi_1)P(\pi_2) p_C[P(\tau|\pi_1), P(\tau|\pi_2)]. \tag{15}$$

We now further note that

$$P(\pi_1, \pi_2|\pi_1 \neq \pi_2) = \frac{P(\pi_1)P(\pi_2)}{1 - p_C[P(\pi)]}, \tag{16}$$

and finally that

$$P(\pi|\pi_1 = \pi_2 = \pi) = \frac{P(\pi)^2}{p_C[P(\pi)]}. \tag{17}$$

Putting it all together we obtain a rather simple decomposition of the coincidence probability in mixtures:

$$p_C[P(\tau)] = p_C[P(\pi)] \left\langle p_C[P(\tau|\pi)] \right\rangle_{P(\pi|\pi_1=\pi_2=\pi)} + (1 - p_C[P(\pi)]) \left\langle p_C[P(\tau|\pi_1), P(\tau|\pi_2)] \right\rangle_{P(\pi_1,\pi_2|\pi_1\neq\pi_2)} \tag{18}$$

This is equivalent to the equation presented in the main text, expressed there in terms of $P(\sigma)$. Let us finally note that we can also generalize Eqn. 14 to inexact coincidences, which defines the generalized cross-sample near-coincidence probabilities,

$$p_C[P_i(\sigma), P_j(\sigma)](\Delta) = \sum_{\sigma,\sigma'} P_i(\sigma)P_j(\sigma') I_{d(\sigma,\sigma')=\Delta}. \tag{19}$$

3. Decomposing selection on paired chain data

How does selection on the heterodimeric TCR protein restrict diversity on the two constituent chains? To answer this question we start by introducing some notation. The complete clone σ is defined by both its α chain sequence, which we denote σ_α, and its β chain sequence, which we denote σ_β. We define the set of all complete TCRs that bind a specific epitope $\mathcal{S}$, such that $\sigma \in \mathcal{S}$ implies the specificity of the receptor encoded by the nucleotide sequence σ. The overall selection factor is then equal to

$$Q(\sigma) = \frac{1}{p(\mathcal{S})} I_{\mathcal{S}}(\sigma), \tag{20}$$

and the exact coincidences in the paired chain data are enriched by a factor $\chi_{\alpha\beta}$,

$$\chi_{\alpha\beta} = \frac{p_C[Q(\sigma)P(\sigma)](0)}{p_C[P(\sigma)](0)} = \frac{1}{p(\mathcal{S})}. \tag{21}$$

In the following we will ask how this ratio relates to the same ratios calculated for the individuals chains, this is to

$$\chi_\alpha = \frac{p_C[Q(\sigma_\alpha)P(\sigma_\alpha)](0)}{p_C[P(\sigma_\alpha)](0)}, \quad \chi_\beta = \frac{p_C[Q(\sigma_\beta)P(\sigma_\beta)](0)}{p_C[P(\sigma_\beta)](0)}, \tag{22}$$

where we will assume independent chain pairing in the background $P(\sigma) = P(\sigma_\alpha)P(\sigma_\beta)$.

Given the full model the selection coefficients for single chains are marginalized averages over all possible choices for the second chain. To calculate these we can rearrange terms in the definition of the marginal single chain post-selection distributions:

$$\tilde{P}(\sigma_\alpha) = \sum_{\sigma_\beta} Q(\sigma)P(\sigma) \tag{23}$$

$$= P(\sigma_\alpha) \sum_{\sigma_\beta} \frac{I_{\mathcal{S}}(\sigma_\alpha, \sigma_\beta)}{P(\mathcal{S})} P(\sigma_\beta), \tag{24}$$

from which we read off

$$Q(\sigma_\alpha) = \frac{\langle I_{\mathcal{S}}(\sigma_\alpha, \sigma_\beta)\rangle_{P(\sigma_\beta)}}{P(\mathcal{S})}. \tag{25}$$

Importantly, we show in the following that $\chi_{\alpha\beta} = \chi_\alpha \chi_\beta$, when there are no chain pairing biases within the epitope-specific set of sequences. To start, note that the set of specific $\alpha\beta$ sequences is given by the cartesian product

$$\mathcal{S} = \mathcal{S}_\alpha \times \mathcal{S}_\beta, \tag{26}$$

of the single chain sets

$$\mathcal{S}_\alpha = \{\sigma_\alpha : \exists \sigma_\beta : (\sigma_\alpha, \sigma_\beta) \in \mathcal{S}\}, \quad \mathcal{S}_\beta = \{\sigma_\beta : \exists \sigma_\alpha : (\sigma_\alpha, \sigma_\beta) \in \mathcal{S}\}, \tag{27}$$

which are composed of all chains that when paired with any of the opposite chains are specific. From these definitions it follows that

$$P(\mathcal{S}) = P(\mathcal{S}_\alpha)P(\mathcal{S}_\beta). \tag{28}$$

Furthermore some algebra shows

$$Q(\sigma_\alpha) = \frac{I_{\mathcal{S}_\alpha}(\sigma_\alpha)}{P(\mathcal{S}_\alpha)}, \quad Q(\sigma_\beta) = \frac{I_{\mathcal{S}_\beta}(\sigma_\beta)}{P(\mathcal{S}_\beta)}. \tag{29}$$

and thus

$$\chi_\alpha = \frac{1}{P(\mathcal{S}_\alpha)}, \quad \chi_\beta = \frac{1}{P(\mathcal{S}_\beta)}. \tag{30}$$

Combining this result with Eqn. 28 we complete the derivation of the equality

$$\chi_{\alpha\beta} = \chi_\alpha \chi_\beta \tag{31}$$

In the general case, the set $\mathcal{S}$ of all specific sequences is a proper subset of the cartesian product $\mathcal{S}_\alpha \times \mathcal{S}_\beta$, and thus $P(\mathcal{S}) < P(\mathcal{S}_\alpha)P(\mathcal{S}_\beta)$. We thus expect the coincidence ratio for the paired chain receptors to increase. This motivates using the ratio

$$\frac{\chi_{\alpha\beta}}{\chi_\alpha \chi_\beta} \tag{32}$$

as a measure of pairing biases among the specific receptor chains.

4. Random models of sequence-correlated selection

In Sec. D of the main text we defined a simple random model of sequence-correlated selection on a background repertoire of recombination events (or clones) by a notional epitope as follows: choose at random a fraction (1% in the examples studied) of the CDR3 amino acid sequences that appear in this background and declare that any recombination event with a CDR3 amino acid sequence included in this list is 'selected'; in addition, let a random fraction (10% in the examples that follow) of all background events with CDR3 amino acid sequence lying at Levenshtein distance one from an element of the selected list be declared to be selected as well. This procedure is motivated by the observation that systematic studies of T cell selection by individual epitopes show a) the existence of multiple unrelated 'solutions' in sequence space to the problem of binding a specific epitope, and b) the existence of some degree of sequence degeneracy within individual binding solutions.

This protocol will bias selection toward sequences that have more than the average number of near neighbors in sequence space. While there is nothing wrong in principle with this, systematic studies of selection by multiple epitopes (such as (1)) show little if any such bias. This motivates us to modify the basic random selection procedure to make the probability of selecting a given background sequence independent of the number of its neighbors in sequence space. In what follows, we approximately solve a simplified version of this problem using the language of random graphs, and use that solution to propose a modification of the sequence-correlated random selection algorithm. We then show by concrete example that this modification achieves the desired result in the more demanding context of selection from realistic T cell sequence ensembles.

A. Mathematical analysis of the two-step selection algorithm. Consider the following random graph problem: we have a set of N nodes, with links between the nodes defined by an adjacency matrix A_{ij} where $A_{ij} = 1$ if node i and j are connected by an edge, and $A_{ij} = 0$ otherwise. The nodes represent amino acid sequences in a background repertoire, and links connect sequence distance one sequence neighbors. We define a 'selected' subgraph by assigning Ising variables s_i to each node to indicate whether the node is selected ($s_i = 1$) or not ($s_i = 0$). Our goal is to find a way of choosing the s_i such that a specified fraction of the nodes have $s_i = 1$ and the probability that a particular node i is selected is independent of the number of links that node participates in. Thus the selected subgraph inherits a subset of the nodes i of the original graph and its adjacency matrix is the projection of A_{ij} on the surviving nodes (no new nodes are created).

We define a two step selection process: We first select a fraction $q_1 \ll 1$ of the nodes and we then select a fraction q_2 (typically $q_2 \gg q_1$) of the links that connect a selected node to one that was not selected; when such a link is selected, we add the originally unselected node to which it connects to the list of selected nodes (i.e. we capture some of the neighbors of a selected node). We take η_i (θ_{ij}) as independent binary random variables (0 or 1) that describe which nodes (links) are picked during the first (second) step of the selection procedure, respectively and have averages q_1 (q_2) respectively.

The variable s_i that indicates whether a given node was picked in either of the two steps can be written as

$$s_i = \eta_i + (1 - \eta_i) \min \left[\sum_{j \neq i} A_{ij} \theta_{ij} \eta_j, 1 \right], \qquad [33]$$

The min function in the second term accounts for the possibility of multiple selection of the unselected node i in the link selection step. In practice, this occurs very rarely for the values of the parameters $q_{1,2}$ that are of interest to us. We will thus ignore this non-linearity in what follows for tractability. We can calculate various marginals and correlations involving the s_i by averaging over the independent, uncorrelated, binary variables η_i and θ_{ij}. Doing so, we find

$$\langle s_i \rangle = \langle \eta_i \rangle + (1 - \langle \eta_i \rangle) \sum_{j \neq i} A_{ij} \langle \theta_{ij} \rangle \langle \eta_j \rangle, \qquad [34]$$

$$\begin{aligned} \langle s_i s_j \rangle =& \langle \eta_i \rangle \langle \eta_j \rangle + (1 - \langle \eta_i \rangle) \sum_{k \neq i} A_{ik} \langle \theta_{ik} \rangle \langle \eta_k \eta_j \rangle + (1 - \langle \eta_j \rangle) \sum_{l \neq j} A_{jl} \langle \theta_{jl} \rangle \langle \eta_l \eta_i \rangle \\ &+ (1 - \langle \eta_i \rangle)(1 - \langle \eta_j \rangle) \sum_{k,l \neq i,j} A_{ik} A_{jl} \langle \theta_{ik} \rangle \langle \theta_{jl} \rangle \langle \eta_k \eta_l \rangle. \end{aligned} \qquad [35]$$

For $i \neq j$, η_i and η_j are statistically independent by assumption and, because η_i is binary, $\eta_i^2 = \eta_i$. This yields the identity

$$\langle \eta_i \eta_j \rangle = (1 - \delta_{ij}) \langle \eta_i \rangle \langle \eta_j \rangle + \delta_{ij} \langle \eta_i \rangle, \qquad [36]$$

which we can use to simplify the equation for the joint selection probability of a pair of nodes as follows:

$$\begin{aligned}\langle s_i s_j\rangle =&\langle\eta_i\rangle\langle\eta_j\rangle\\ &+(1-\langle\eta_i\rangle)\langle\eta_j\rangle A_{ij}\langle\theta_{ij}\rangle\\ &+(1-\langle\eta_i\rangle)\langle\eta_j\rangle\sum_{k\neq i,j}A_{ik}\langle\theta_{ik}\rangle\langle\eta_k\rangle\\ &+(1-\langle\eta_j\rangle)\langle\eta_i\rangle A_{ji}\langle\theta_{ji}\rangle\\ &+(1-\langle\eta_j\rangle)\langle\eta_i\rangle\sum_{l\neq i,j}A_{jl}\langle\theta_{jl}\rangle\langle\eta_l\rangle\\ &+(1-\langle\eta_i\rangle)(1-\langle\eta_j\rangle)\sum_{k\neq i,j}A_{ik}A_{jk}\langle\theta_{ik}\rangle\langle\theta_{jk}\rangle\langle\eta_k\rangle\\ &+(1-\langle\eta_i\rangle)(1-\langle\eta_j\rangle)\sum_{k\neq l\neq i,j}A_{ik}A_{jl}\langle\theta_{ik}\rangle\langle\theta_{jl}\rangle\langle\eta_k\rangle\langle\eta_l\rangle.\end{aligned}\quad[37]$$

Let us use these expressions to analyze the naive algorithm for sequence-correlated random selection that was described at the beginning of this appendix. In this scheme $\forall i:\langle\eta_i\rangle=q_1\ll 1$ and $\forall i,j:\langle\theta_{ij}\rangle=q_2\ll 1$. Importantly, for nodes i,j connected by an edge (so that $A_{ij}=1$), the leading order terms in the above expression for $\langle s_i s_j\rangle$ give, for $q_1\ll q_2$ (the case of interest):

$$\langle s_i s_j\rangle\approx\langle\eta_j\rangle\langle\theta_{ij}\rangle+\langle\eta_i\rangle\langle\theta_{ji}\rangle=2q_1q_2.\quad[38]$$

In other words, the procedure allows us to enhance the probability of picking adjacent vertices beyond the independent probability q_1^2 of picking two isolated vertices.

However, as mentioned above, this procedure introduces a bias towards sampling highly connected sequences. More precisely, according to Eqn. 34

$$\langle s_i\rangle=q_1+(1-q_1)q_1q_2N_i\quad\text{with } N_i=\sum_{j\neq i}A_{ij}\quad[39]$$

where N_i is the number of edges connecting to the i-th vertex: the more connected nodes i will have higher probability of being selected. The idea of the corrected procedure is to make the selection probability in the first step dependent on the neighbor number: $\langle\eta_i\rangle=q_1(N_i)$ where $q_1(N)$ is a function to be determined. The probability with which node i is selected in this procedure can be expressed as

$$\langle s_i\rangle=q_1(N_i)+(1-q_1(N_i))q_2N_i\left(\frac{1}{N_i}\sum_{j\neq i}A_{ij}q_1(N_j)\right),\quad[40]$$

where the term within brackets is the average first step selection probability of all neighboring sequences connected to i by a link. To derive an appropriate functional form for $q_1(N_i)$ we make the further assumption that neighboring nodes have the same number of edges on average, so that $\frac{1}{N_i}\sum_{j\neq i}A_{ij}q_1(N_j)\approx q_1(N_i)$. One can then show that the choice of

$$q_1(N_i)=\frac{q_1}{1+q_2N_i},\quad[41]$$

leads to an approximately constant probability of selection, $\langle s_i\rangle\approx q_1$, independent of i, if $q_1\ll 1$.

B. Numerical tests of the corrected algorithm. How well does proposed neighbor-number corrected selection algorithm work on realistic T cell sequence data? We have applied the algorithm described in the previous paragraphs to background repertoires of a sample of 10^5 CDR3β sequences of PBMC T cells obtained from a human subject (data from (3)). We generated 50 selected repertoires of $\sim 10^3$ sequences from this background using the neighbor number corrected algorithm. In this background repertoire, the variance of the number of neighbors is large, with a coefficent of variation of ~ 2.5. This analysis is thus a good test for how well the corrected selection algorithm succeeds in equalizing selection probability for nodes with different numbers of neighbors.

In Fig. S1 we plot the probability of selection of a recombination event from the background repertoire, averaged over 50 realizations of the algorithm. We use a fixed selection parameter $q_1=.01$ and neighbor selection parameters q_2 ranging from 0.0 to 0.1. We plot the selection fraction as a function of the number of distance one neighbors the selected node has in the background repertoire. The net selection probability except for fluctuations is nearly constant out to neighbor number of ~ 10, a value that captures more than 90% of the nodes in the background.

It is instructive to display the near-coincidence frequency distributions that result from the corrected selection algorithm for different values of q_2 (Fig. S2). As expected from Eqn. 6 the rate of falloff of the coincidence frequency enhancement over background at small sequence distance Δ depends on the parameter q_2 that governs the fraction of the neighbors of a selected sequence that will also be selected.

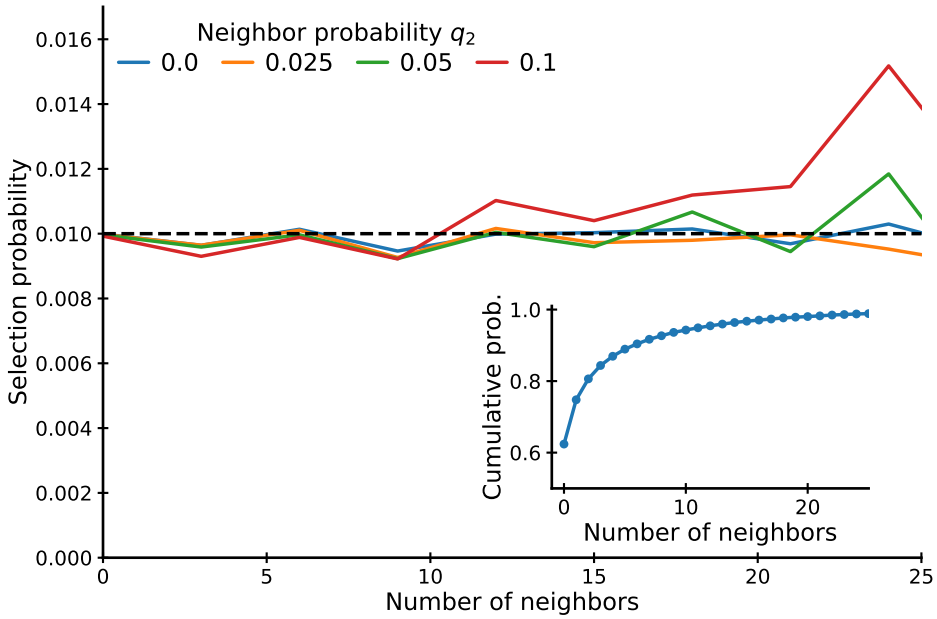

Fig. S1. Assessment of the uniformity of random selection from background sequence repertoires across groups of sequences classified by different numbers of Levenshtein distance one neighbors. The plot shows the fraction of each class in the background repertoire that appears in the selected repertoire, averaged over 50 realizations of the selection algorithm.

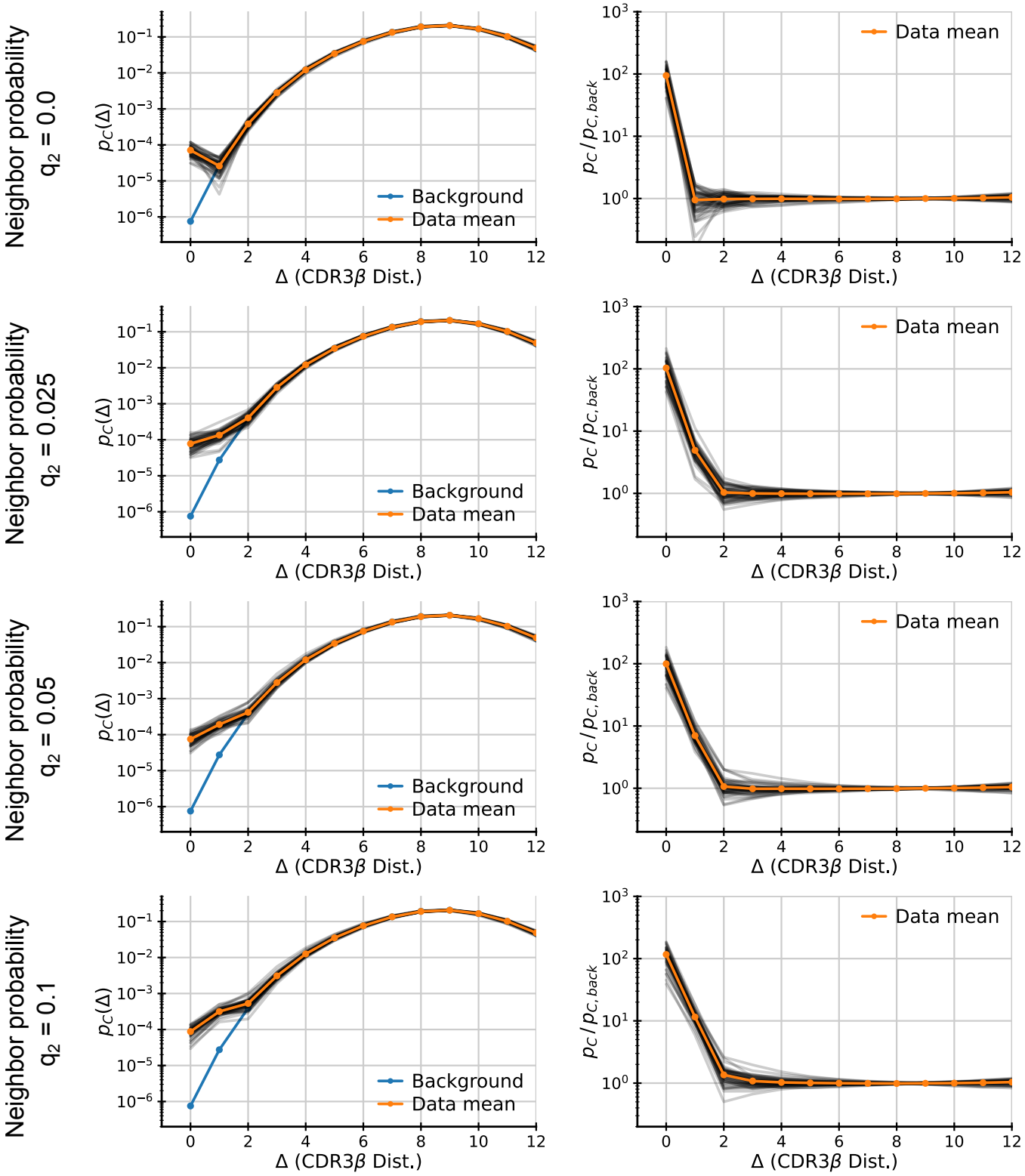

Fig. S2. Coincidence frequency distributions for different realizations of the neighbor-corrected random selection algorithm. The black lines give the results of individual random realizations of selection, while the orange curves give averages over the full set of 100 realizations. The key feature to note is the inverse correlation between q_2, the parameter governing the sequence neighbor selection probability, and the initial slope of the log ratio of near-coincidence frequencies to background.

5. Analysis of paired chain sequencing data from Tanno et al.

In the following we describe some observations about the nature of sequence coincidences in the paired chain sequencing data from Tanno et al. (2), that have prompted us to perform further filtering steps beyond the analysis pipeline proposed in the previous work.

The experimental protocol uses overlap extension polymerase chain reaction to link both α and β hypervariable chains before sequencing. Linking predominantly occurs between mRNA from the same single cell as those are captured on the same bead. However, cross-contamination can lead to erroneous pairing of mRNA from different cells. To reduce cross-contamination the authors of (2) clustered CDR3β on 95% sequence identity (for typical sequence lengths this is ≤ 2 nts), and kept only the most frequent CDR3α sequence from each cluster. No such clustering had been performed for CDR3α, and we thus asked to what extent contamination accounts for CDR3α sequences associated with multiple CDR3β sequences. To assess this we calculated the ratio between CDR3α coincidence probabilities on the nucleotide and amino acid level (Fig. S3A). As a comparison we calculated the same ratio for sequences generated using a probabilistic model of VDJ recombination (4) and found that ratios greatly exceed expectations. This finding suggests substantial cross-contamination that might impact downstream analysis. We thus implemented a filtering step clustering CDR3α sequences with exact sequencing identity, and kept only the most frequent CDR3β sequence from each cluster. Such filtering substantially reduces coincidence probabilities bringing them closer in line with those observed in a bulk-sequenced α chain repertoire dataset (Fig. S3B). Variability across subjects is also reduced, suggesting different levels of cross-contamination across samples.

We next asked whether there was anything special about coincidences observed among pairs of subjects with the lowest HLA overlap. On examining the specific sequences responsible for exact coincidences in these pairs, we found that certain VJ combinations were heavily overrepresented among their α chains. These VJ combinations map to known signatures of noncanonical T cells with semi-invariant receptors, so called MAIT and iNKT cells (5). These T cells have semi-invariant α-chains and restricted β chain diversity, which explains why they contribute heavily to near-coincidences. Both T cell subsets bind to non-peptide ligands, which are not presented on MHC molecules. As we are interested in identifying signatures of selection driven by pMHC binding, we exclude the VJα combinations coding for these invariant T cells (TRAV1-2 paired with TRAJ12/TRAJ20/TRAJ33 for MAITs, and TRAV10 paired with TRAJ18 for iNKTs).

While these additional filtering steps have a numerically small effect on the total numbers of near-coincidences, we believe that this careful approach is warranted as some of the most interesting comparisons rest on the small numbers of exact or nearly exact coincidences.

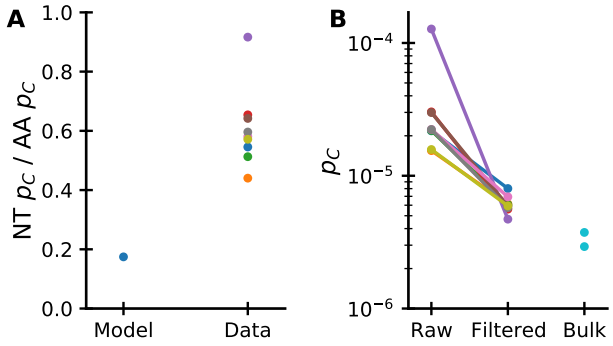

Fig. S3. Rationale for collapsing α chains paired with multiple β chains in the paired chain data from Tanno et al. (2). (A) The ratio between nucleotide and amino acid coincidence probabilities for α chains is calculated for each subject and for data generated from SONIA, a model of VDJ recombination and thymic selection. The ratio is at least two-fold higher in the data than expected suggesting substantial cross-contamination. (B) Collapsing redundant alpha chains reduces variation in amino acid coincidence probabilities across samples and makes them more comparable to those found in bulk single chain datasets (donors M and W at baseline from Minervina et al. (3)).

6. Detailed analysis of motif mixture model

In Sec. 5 we sketched a schematic model of TCR-pMHC binding as a string-matching problem (6, 7). In this appendix we present a fuller account of the formulation and analysis of this model.

For simplicity we consider synthetic TCR sequences σ of a fixed length $k = 6$, corresponding to the number of hypervariable residues within a typical CDR3 loop. For a particular pMHC complex p, we assume that the binding energy depends on the residues within the TCR additively, and that each site makes a binary contribution to the energy:

$$E_p(\sigma) = \sum_{i=1}^{k} \epsilon_p(i, \sigma_i), \quad \epsilon_p(i, \sigma_i) = \begin{cases} -1 & \text{for } \sigma_i \in \mathcal{S}_i^p \\ 0, & \text{otherwise.} \end{cases} \tag{42}$$

Here, for each site, $\mathcal{S}_i^p$ is a set of "good" amino acids that contribute to the binding of sequence σ to epitope p. A sequence is taken to bind only if $E_p(\sigma) = -k$, that is to say that the amino acid at each site i is in the allowed set $\mathcal{S}_i^p$. This definition of binding energy describes a random motif model, where a motif is defined as any combination of allowed amino acids at the different sites. The motif of course will vary from one epitope to another.

As in our model TCRs have fixed length we consider the simpler Hamming distance instead of edit distance, and we make two further assumptions to simplify analytical calculations: First, at each site there are an equal number c of allowed amino acids, $\forall i : |\mathcal{S}_i| = c$. Second, background sequences are drawn from the flat distribution over all k-mers: at each site we draw independently and uniformly at random one out of the $q = 20$ amino acids. Calculating the near-coincidence histograms within the background set and within the specific set for a particular epitope then reduces to purely combinatorial exercise with the analytical results

$$\begin{aligned} p_C(\Delta) &= \frac{1}{c^k} \binom{k}{\Delta} (c-1)^\Delta \quad \text{for} \quad \Delta = 0 \dots k \\ p_{C,back}(\Delta) &= \frac{1}{q^k} \binom{k}{\Delta} (q-1)^\Delta \quad \text{for} \quad \Delta = 0 \dots k. \end{aligned} \tag{43}$$

The near-coincidence enhancement factor in this model is therefore

$$\frac{p_C(\Delta)}{p_{C,back}(\Delta)} = \left(\frac{c-1}{q-1}\right)^\Delta \left(\frac{q}{c}\right)^k. \tag{44}$$

This expression reproduces the exponential falloff with Δ that is seen in real data, with the falloff rate dependent on the number of allowed amino acids c. To make the rate compatible with the observed factors of ten, requires that at each site there are on average $c = 3$ possible amino acids, such that the fraction of specific neighboring sequences is equal to $(c-1)/(q-1) = 2/19 \approx 0.1$.

We next considered a mixture of motif models, where all TCRs that conform to any of M randomly chosen motifs are specific. Each motif defines a different binding energy function as per Eqn. 42 with independently drawn sets of allowed amino acids $\mathcal{S}_i^p$. The binding energy of a TCR sequence is then taken to be the minimum over these motifs. The distribution of T cells selected by this binding energy can be approximated as a mixture of the distributions selected by the individual motifs. Applying results for coincidences in mixture distributions (derived in Appendix 2), we obtain an analytical prediction for excess coincidences

$$\frac{p_C(\Delta)}{p_{C,back}(\Delta)} \approx \frac{1}{M} \left(\frac{c-1}{q-1}\right)^\Delta \left(\frac{q}{c}\right)^k + 1 - \frac{1}{M}. \tag{45}$$

Numerical simulations of the model were performed as follows: We first draw a background set of $2 \cdot 10^7$ TCRs of length $k = 6$. Each TCR is drawn from an independent site model, where the probability of drawing a specific amino acid is set equal to the usage frequency of amino acids found within the CDR3α hypervariable chains in a human blood sample from (3). Next, we draw M different binding motifs, each defined by $c = 3$ amino acids drawn independently and evenly drawn from all possible amino acids. We then filter out all sequences from the background set that match the definition of any of the motifs. Finally, we calculate the coincidence probability for the specific sequences retained from the background.

References

1. S Nolan, et al., A large-scale database of T-cell receptor beta (TCR β) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2 . *Res. Sq. Prepr.* (2020).
2. H Tanno, et al., Determinants governing T cell receptor α / β -chain pairing in repertoire formation of identical twins. *Proc. Natl. Acad. Sci.* **117**, 532–540 (2020).
3. AA Minervina, et al., Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T-cell memory formation after mild COVID-19 infection. *eLife* **10**, e63502 (2021).
4. Z Sethna, et al., Population variability in the generation and selection of T-cell repertoires. *PLoS Comput. Biol.* **16**, e1008394 (2020).
5. DI Godfrey, AP Uldrich, J Mccluskey, J Rossjohn, DB Moody, The burgeoning family of unconventional T cells. *Nat. Immunol.* **16** (2015).

6. JD Farmer, NH Packard, AS Perelson, The immune system, adaptation, and machine learning. *Phys. D: Nonlinear Phenom.* **22**, 187–204 (1986).
7. A Kosmrlj, AK Jha, ES Huseby, M Kardar, AK Chakraborty, How the thymus designs antigen-specific and self-tolerant T cell receptor sequences. *Proc. Natl. Acad. Sci.* **105**, 16671–6 (2008).