bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Measures of epitope binding degeneracy from T cell receptor repertoires

Andreas Mayer[1,2] and Curtis G. Callan Jr[3,4]

[1] *Division of Infection and Immunity, University College London*
[2] *Lewis-Sigler Institute for Integrative Genomics, Princeton University*
[3] *Department of Physics, Princeton University*
[4] *Institute for Advanced Study, Princeton, NJ*

(Dated: November 15, 2022)

Adaptive immunity is driven by specific binding of hyper-variable receptors to diverse molecular targets. The sequence diversity of receptors and targets are both individually known but, because multiple receptors can recognize the same target, a measure of the effective 'functional' diversity of the human immune system has remained elusive. Here, we show that sequence near-coincidences within T cell receptors that bind specific epitopes provide a new window into this problem, and allow the quantification of how binding probability co-varies with sequence. We find that near-coincidence statistics within epitope-specific repertoires imply a measure of binding degeneracy to amino acid changes in receptor sequence that is consistent across disparate experiments. Paired data on both chains of the heterodimeric receptor are particularly revealing since simultaneous near-coincidences are rare and we show how they can be exploited to estimate the number of epitope responses that created the memory compartment. In addition, we find that paired-chain coincidences are strongly suppressed across donors with different human leukocyte antigens, evidence for a central role of antigen-driven selection in making paired chain receptors public. These results demonstrate the power of coincidence analysis to reveal the sequence determinants of epitope binding in receptor repertoires.

Keywords: T cells | receptor-ligand binding | repertoire sequencing | specificity

Which epitopes are recognized by an individual's T cells? The specificity of T cells is encoded genetically in the loci coding for the hypervariable loops of the T cell receptor (TCR) chains [1], and thus in principle reading out the immune repertoire by sequencing provides the information to answer this question [2–4]. Yet, deciphering the complex sequence 'code' for the many-to-many mapping between TCRs and peptide-major histocompability complexes (pMHCs) remains an open problem [5].

Aspects of this code are coming into view thanks to data from multiple experimental approaches. Structural studies have revealed the spatial arrangements in which TCRs bind pMHCs [6–11]. Mutagenesis experiments [12, 13] have provided early evidence that some amino acid substitutions in TCRs maintain or even increase binding affinity to a given epitope. Such local degeneracy of the binding code has been confirmed more recently by sequencing of epitope-specific groups of TCRs [14–21], and sequence patterns in these datasets are now used in machine learning approaches to predict further binders to the same epitope [22–26].

Direct experiment can, however, examine only a minute fraction of all the possible binding combinations, due to the enormous diversity of potential receptors and epitopes: more than 10^{12} different peptides [27] are presented on 1000s of human MHC alleles [28] to up to 10^{61} possible TCRs [29] generated by the recombination machinery. As a result, rules that generalize across epitopes would be of utmost utility, but TCR diversity has made it difficult to find such rules.

To address this problem, we here introduce a statistical framework that quantifies the sequence degeneracy of receptors that bind to a common target via the enhanced frequencies of sequence coincidences in epitope-specific TCR repertoires as compared to their frequencies in suitably chosen 'background' repertoires. The specific repertoires we study can be created in a controlled way in an experiment, or can arise organically, as when a memory compartment is formed in response to an infection. Generalizing the analysis to inexact coincidences (pairs of sequences with high sequence similarity), we find that they, too, are enhanced in epitope-specific repertoires. We demonstrate mathematically that the ratio of near-coincidence probabilities between data and background, as a function of sequence distance, is a direct measure for how specificity is correlated across sequence space.

Applying this framework to epitope-specific T cell repertoires that have been acquired in different ways [14–17] reveals a common coincidence enhancement signature of specific binding across disparate experiments. We relate this signature to the existence of a typical average local binding degeneracy, defined as the fraction of the available sequence neighbors of a specific T cell receptor (available in the sense of being present in a natural repertoire) that will also bind to the same pMHC. In addition, we see a weaker version of this signature in paired chain repertoires that have not been subjected to explicit *ex vivo* enrichment for epitope-specific T cells [30]. We exploit this observation in two ways: we provide clear evidence that this signature is associated with MHC presentation of antigen by demonstrating that coincidences between different donors are strongly affected by the overlap between their human leukocyte antigen (HLA) types; in addition, after some mathematical analysis, we use it to quantify the effective functional diversity of the memory repertoire, in the sense of an estimated number of

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

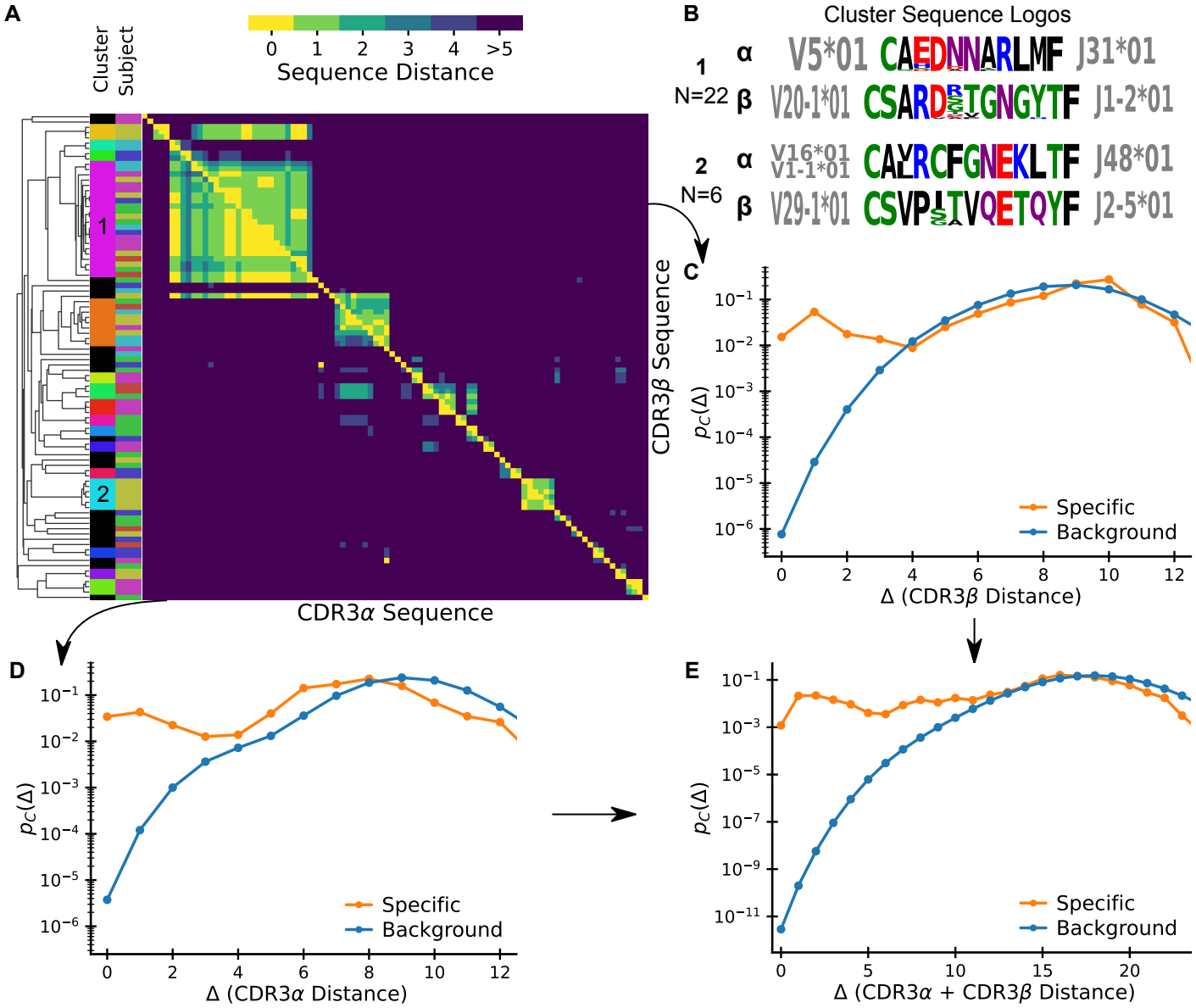

FIG. 1: **Patterns of sequence similarity within an epitope-specific repertoire.** (A) Sequence-similarity clustermap of TCRs binding to an Epstein-Barr Virus epitope as obtained by single cell TCR sequencing following tetramer sorting (Data: Dash et al. [14], antigen BMLF). Lower (upper) triangle shows pairwise distances of CDR3α (CDR3β) sequences. Sequences are ordered by average linkage hierarchical clustering based on summed $\alpha\beta$ distance. Columns on the left show subject of origin and cluster assignment; sequences not belonging to a cluster based on a cutoff distance of 6 are shown in black. (B) Sequence logos for two clusters of specific sequences. Amino acids are colored by their chemical properties, and V and J gene usage within the cluster is displayed alongside the logo. (C-E) Normalized histograms of pairwise distances between (C) CDR3β, (D) CDR3α, and (E) CDR3$\alpha\beta$ sequences specific to the epitope show vastly increased sequence similarity relative to background expectations.

epitope recognition events it records. Taken together, these results illustrate how coincidence analysis can help to quantitatively address immunological questions whose answers have so far remained elusive.

I. OVERVIEW OF ANALYSIS STRATEGY

We illustrate the broad strategy of our approach on a repertoire of CD8$^+$ T cells specific to an Epstein-Barr Virus peptide presented on HLA-A*02:01. The data is from Dash et al. [14] and was obtained using single cell receptor sequencing of tetramer-sorted T cells binding the specific pMHC.

Fig. 1A shows a clustering by pairwise amino acid se-

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

quence distance of all distinct nucleotide sequence clones. In this visualization, each position in the heatmap records the sequence distance Δ between the amino acid sequences of a pair of distinct T cell clones. TCRs are heterodimeric, and the heatmaps above (below) the diagonal record distances between the β (α) hypervariable complementary-determining region 3 (CDR3) loops of the sequence pair. Clustering is based on the sum of distances between α and β chains. Here, and throughout this work, we define sequence distances as the minimal number of edits (insertions, deletion, or substitutions) that change one sequence into another, known as the Levenshtein distance. We only consider sequence distances between CDR3 loop for simplicity, but the mathematical framework we develop is general and could also be used with distance measures that include other hypervariable receptor regions. By clones we mean lineages of cells that go back to the same ancestral recombination event, which we define in practice based on nucleotide sequence identity. A zero distance pair arises when due to convergent recombination two distinct nucleotide sequences have the same amino acid translation. We chose to ignore the number of times a given nucleotide sequence is sampled, as clone sizes also reflect TCR-independent lineage differences [31, 32]. Instead, we analyze convergent selection imposed on distinct clones with the same or similar TCR as a stringent measure of epitope-driven functional selection. In the experiments that we consider in this manuscript, TCRs are selected for binding to a specific pMHC ligand, and our analysis quantifies the imprint of this filtering funnel on TCR sequence statistics. We use the word "selection" to refer to this filtering process, which is distinct from, and not to be confused with, thymic selection.

Fig. 1A allows some direct conclusions about important features of TCR-pMHC binding code: First, it highlights the remarkable sequence similarity among specific TCRs and it shows that this similarity also holds for TCRs from different donors. Second, it shows that there are several clusters of sequences differing by a few substitutions from each other, plus a substantial number of isolated sequences that differ from all other sequences by many substitutions. Fig. 1B shows sequence logos for two prominent clusters. Interestingly, they are quite different from each other, even when accounting for chemical similarity of amino acids. This suggests clusters might represent broad structurally distinct binding solutions, each with local residue degeneracy. This view is supported by the V and J gene usage, which is highly restricted within each cluster but non-overlapping between them. Third, it demonstrates that chain-pairing is biased even among specific binders as similarity on one chain is often associated with similarity on the other chain.

To compare statistics of sequence similarity across epitope targets, we next compress the off-diagonal elements of the clustermap into a normalized pairwise distance histogram that we denote by $p_C(\Delta)$. We normalize coincidences by $N(N-1)/2$, the number of possible pairs (i.e. upper diagonal elements in the matrix), so that $p_C(\Delta)$ is a probability distribution on Δ. Figs. 1C,D show the histograms for α and β chains, respectively. Fig 1E shows the histogram for the complete $\alpha\beta$-TCR, with paired chain sequence distance defined as the sum of distances of both chains. These normalized pairwise distance distributions are the basic element of our analysis framework. We also plot the $p_C(\Delta)$ distributions derived from bulk sequencing of a "background" sample as a proxy for the expected distribution prior to selection. We use sequencing data from Minervina et al. [16] of total peripheral blood mononuclear cells (PBMCs) from a healthy individual for these background curves for α and β chains. For the paired chain background curve we currently lack sufficiently deeply sequenced datasets. Fortunately, previous studies have concluded that α and β chain gene usages are largely uncorrelated [30, 33, 34], so we use the convolution of the α chain and β chain distributions from Minervina et al. [16] as a plausible paired chain background prior to selection. In section VII we will present further evidence supporting the use of this assumption.

The central observation is that $p_C(\Delta)$ is orders of magnitude larger in epitope-specific repertoires than the corresponding background for small Δ. Exact coincidence frequencies are in excess by surprisingly large factors ($\sim 10^9$ and $\sim 10^4$ for paired and unpaired chains, respectively). This excess extends to near-coincidences, but interestingly, for large enough Δ, the selected and the background values of $p_C(\Delta)$ approach each other. The manner in which their ratio falls to unity will turn out to be roughly the same across different types of experiments, an intriguing fact that points to shared underlying biophysical rules of specific binding.

II. THEORY OF COINCIDENCE ANALYSIS

A. Definitions and statistical estimation

The T cell clones that enter the immune repertoire are drawn from a background distribution $P(\sigma)$ over all possible nucleotide sequences σ that code for the TCR hypervariable chains. This distribution summarizes the statistics of the recombination process by which the receptor coding genes are rearranged, and it is known that probabilities of individual sequences range over many orders of magnitude [35]. Experimentally, clones are identified by distinct nucleotide sequences, and coincidences (exact or near) are defined by the corresponding amino acid sequence (since that is what determines functional identity or similarity). Generation probabilities are such that it is unlikely that two separate T cell generation events will give the same nucleotide sequence, but it is less uncommon for them to give the same CDR3 amino acid sequence. Therefore the practical limitation of identifying clones by distinct nucleotide sequences instead of recombination events introduces only minimal bias. The normalized histogram of pairwise distances defined operationally in the previous section is then an empiri-

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

cal estimate of coincidence probabilities, more formally defined as

$$p_C[P](\Delta) = \sum_{\sigma,\sigma'} P(\sigma)P(\sigma')I_{d(\sigma,\sigma')=\Delta}, \qquad (1)$$

where I is the indicator function, the sum is over independent pairs of nucleotide sequences, and $d(\sigma,\sigma')$ is the sequence distance between the amino acid translations of the sequences.

Given the diversity of TCRs it is surprising that we are able to find any coincidences in small epitope-specific repertoires. The occurrence of coincidences at sample sizes much smaller than the space of all sequences is connected to the "birthday problem" in probability theory [36, 37]: In a sample of N distinct sequences there are $N(N-1)/2$ distinct pairs, and the expected number of pairs at distance Δ is thus $p_C(\Delta)N(N-1)/2$. This means that we can estimate normalized pair probabilities $p_C(\Delta) \sim 10^{-3}$ using repertoires of only $N \sim 10^2$ sequences. This is fortunate since it is precisely this combination of orders of magnitude that we encounter when we estimate $p_C(\Delta)$ from epitope-specific repertoires at small values of Δ (Fig. 1C-E).

B. Intuition for why coincidences increase in epitope-specific repertoires

To gain intuition, we define a probability distribution on amino acid sequences by marginalizing over nucleotide sequences, $P(\tau) = \sum_{\sigma \in T_\tau} P(\sigma)$, where T_τ is the set of sequences that translate to amino acid sequence τ. In this notation, we can give an alternative definition of the exact coincidence probability (the value of Eqn. 1 at $\Delta = 0$) as

$$p_C[P] = \sum_{\tau} P(\tau)^2. \qquad (2)$$

This expression is Simpson's diversity index from ecology [36]. Its inverse $1/p_C$ is known as a true diversity, an estimate of an effective number of species present in a population. Here, amino acid receptor sequences take the role of species, which means p_C is an index of the diversity of amino acid sequences coded for by the different clones in the repertoire. Only some receptors bind an epitope, thus we expect epitope-specific repertoires to have lower diversity. This provides an intuitive explanation why p_C, the inverse of true diversity, increases with selection. From this perspective, Eqn. 1 represents a generalization of Simpson's index to a similarity-weighted measure of diversity [38]. As epitope-specific repertoires consist of TCRs with similar sequences, we expect similarity-weighted diversity to also be restricted. This in turn helps rationalize why $p_C(\Delta)$ is increased in epitope-specific repertoires for some range of small Δ. A central point of this paper is that a great deal of information is contained in the generalization of Simpson's index to inexact coincidences.

To develop this intuition further, let us represent T cells with distinct nucleotide sequences as nodes in a graph, and connect pairs of clones with the same TCR amino acid sequence with a link. Fig. 2A displays such a graph representation for 100 notional background T cells, together with the result of selecting half of them according to two different protocols. The probability that a randomly chosen pair of nodes are linked is equal to $p_C = 2|E|/(|V|(|V|-1))$, where $|E|$ is the number of edges and $|V|$ is the number of vertices. The pre-selection repertoire is shown in the left panel, where links were arbitrarily chosen such that $p_C = 0.02$. The middle and

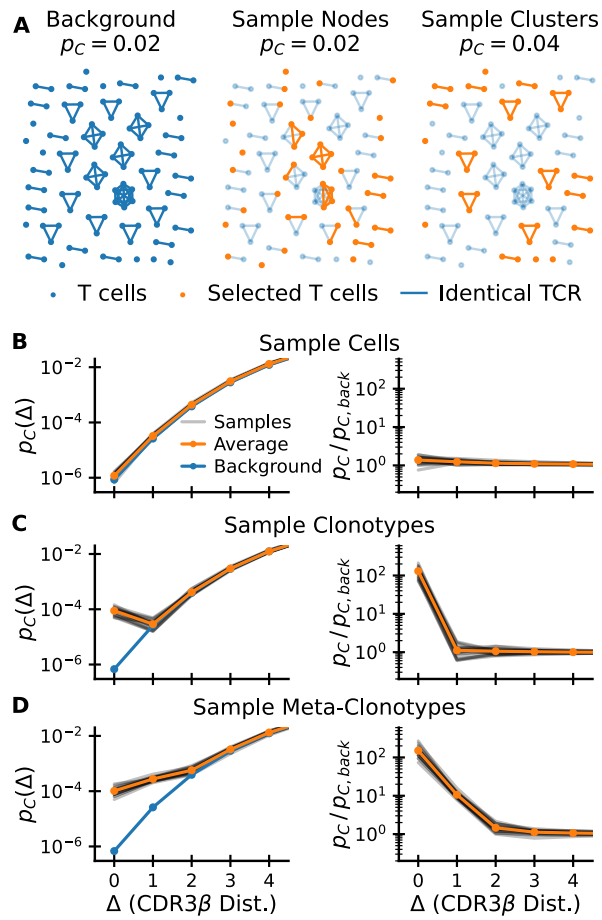

FIG. 2: **How selection increases coincidences.** (A) How different selection procedures change the graph of sequence neighbors. Cells (nodes) in a background graph (left) are connected by edges if they share an identical TCR. Random sampling of nodes (middle) does not change the coincidence probability. Random sampling of clusters (right) increases the coincidence probability. Selected nodes and links in orange; unselected background nodes in light blue. (B-D) Coincidence probabilities for synthetic data generated by selecting 1% of cells (B), 1% of amino acid clonotypes (C), and 1% of meta-clonotypes (generated by including 10% of neighbors of each selected sequence, see text) (D) at random. These random selection protocols act on a a background CDR3β repertoire (data from Minervina et al. [16]). The grey lines show estimates for 20 repetitions of the sampling procedure, and the orange line their average.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

right panels show the results of two selection protocols mimicking random subsampling and epitope-specific sorting, respectively: selecting nodes with probability 1/2, ignoring linkage (center), or selecting clusters of nodes with probability 1/2 (right). When selecting cells at random, the coincidence probability $p_C = 0.02$ is unchanged: the mean number of linked pairs decreases by a factor 4, but so does the total number of possible node pairs. Selecting clusters in contrast, implies that the number of edges decreases by only a factor 2. Normalizing by the total number of node pairs, the coincidence probability increases two-fold to $p_C = 0.04$. The selection of connected clusters mimics sorting by epitope-specificity, in the sense that cells belonging to the same clonotype, defined by identical amino acid sequence, all share the same specificity.

C. Formal analysis

We now mathematically derive how coincidence probabilities change when specific TCRs are identified within a larger pool. We analyze this as follows: let $Q(\sigma)$, normalized by $\langle Q(\sigma)\rangle_{P(\sigma)} = 1$, be a selection factor that characterizes whether sequence σ meets the chosen selection condition. The distribution of selected sequences is then $Q(\sigma)P(\sigma)$. As we derive in Appendix A the coincidence distributions of the two ensembles are related via the cross-moments of the selection factors,

$$\frac{p_C[QP](\Delta)}{p_C[P](\Delta)} = \langle Q(\sigma)Q(\sigma')\rangle_{\sigma\overset{\Delta}{\sim}\sigma'}, \tag{3}$$

where $\langle . \rangle_{\sigma\overset{\Delta}{\sim}\sigma'}$ indicates that the average is calculated over random pairs of sequences at distance Δ, i.e. over the distribution $P(\sigma, \sigma'|d(\sigma, \sigma') = \Delta)$.

To gain intuition we consider a simple class of selection functions of relevance to antigen-specific selection, where Q weights equally a specific subset $\mathcal{S}$ of sequences and gives zero weight to all others:

$$Q(\sigma) = I_{\mathcal{S}}(\sigma)/P(\mathcal{S}), \tag{4}$$

where $P(\mathcal{S}) = \sum_{\sigma\in\mathcal{S}} P(\sigma)$ is the fraction of all clones (i.e. distinct nucleotide sequences) that are specific to the epitope in question. Given the statistical process that created the background repertoire, any given background sequence has an expected number of 'neighbors' at sequence distance Δ; if the sequence in question is selected, we can ask what fraction $f_\sigma(\Delta)$ of its neighbors at distance Δ are also selected. Plugging Eqn. 4 into Eqn. 3 we find that the coincidence enhancement ratio is proportional to the average of that fraction over the selected sequences $\langle f_\sigma(\Delta)\rangle_{\sigma\in\mathcal{S}} = \langle I_{\mathcal{S}}(\sigma')\rangle_{\sigma\overset{\Delta}{\sim}\sigma',\sigma\in\mathcal{S}}$:

$$\frac{p_C[QP](\Delta)}{p_C[P](\Delta)} = \frac{\langle f_\sigma(\Delta)\rangle_{\sigma\in\mathcal{S}}}{P(\mathcal{S})}. \tag{5}$$

Note that $\langle f_\sigma(\Delta = 0)\rangle_{\sigma\in S} = 1$ because specific binding only depends on amino acid sequence, so that all exact coincidences with a selected sequence must also be selected. Thus, the increase in exact coincidence probability is inversely proportional to the selection fraction $P(\mathcal{S})$. If the selection fraction is small, the coincidence ratio is large, in line with the interpretation of this ratio as a measure of the strength of selection. What follows is a direct way to estimate the average number of specific neighbors:

$$\frac{p_C[QP](\Delta)}{p_C[P](\Delta)} = \frac{p_C[QP](0)}{p_C[P](0)}\,\langle f_\sigma(\Delta)\rangle_{\sigma\in S}. \tag{6}$$

How coincidence ratios decrease with distance Δ is thus a measure of the average sequence degeneracy of specific binding. Applying this equation to experimental data will allow us to estimate this fundamental quantity. In comparing with data the empirical coincidence distribution within an epitope-specific repertoire is our measure of $p_C[QP]$, and $p_C[P]$ is determined from a background set of sequences. To simplify notations we will thus refer to their ratio as $p_C/p_{C,back}$.

D. Simulation of selection on real data

To make the preceding formal analysis concrete we next turned to numerical simulation of selection of sequences from a realistic background T cell repertoire. To get intuition of the effect of selection by a generic pMHC complex at a gross statistical level, we filter sequences from a background dataset of approximately 10^5 CDR3β sequences taken from whole blood (data from [16]) according to different random sampling protocols.

We first compare selecting random cells (Fig. 2B) with selecting random clonotypes (Fig. 2C), in each instance selecting 1% of sequences. For the former, apart from statistical noise, $p_C(\Delta)$ is the same for the selected set as for the background. For the latter, the exact coincidence frequency increases hundredfold. This increase corresponds to the inverse of the selection fraction $P(\mathcal{S}) \sim 10^{-2}$, exactly as predicted by Eqn. 3. Such random selection of clonotypes was used successfully in Elhanati et al. [39] to predict TCR sharing numbers among a large number of human individuals. However, for $\Delta \neq 0$ coincidence frequencies do not differ from the background (in contrast to empirical data, such as Fig. 2C).

We thus next sought to incorporate sequence correlation in selection between similar amino acid sequences to model the local degeneracy in antigen recognition apparent in Fig. 1A. To this end, for each selected sequence σ, we also select a fraction p_{corr} of sequences that are within sequence distance Δ_{corr} from σ. The construction of such a sequence-correlated random selection model is somewhat subtle as a naive scheme oversamples sequences with many neighbors. We derived a corrected sampling scheme explained in Appendix D that overcomes this bias. The results of such a selection of metaclonotypes for $\Delta_{corr} = 1$ and $p_{corr} = 0.1$ are shown in Fig. 2D. As expected, sequence correlations lead to an enhancement of $p_C(\Delta)$ over background that extends to near coincidences. Also, the selection enhancement ratio changes by a factor

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

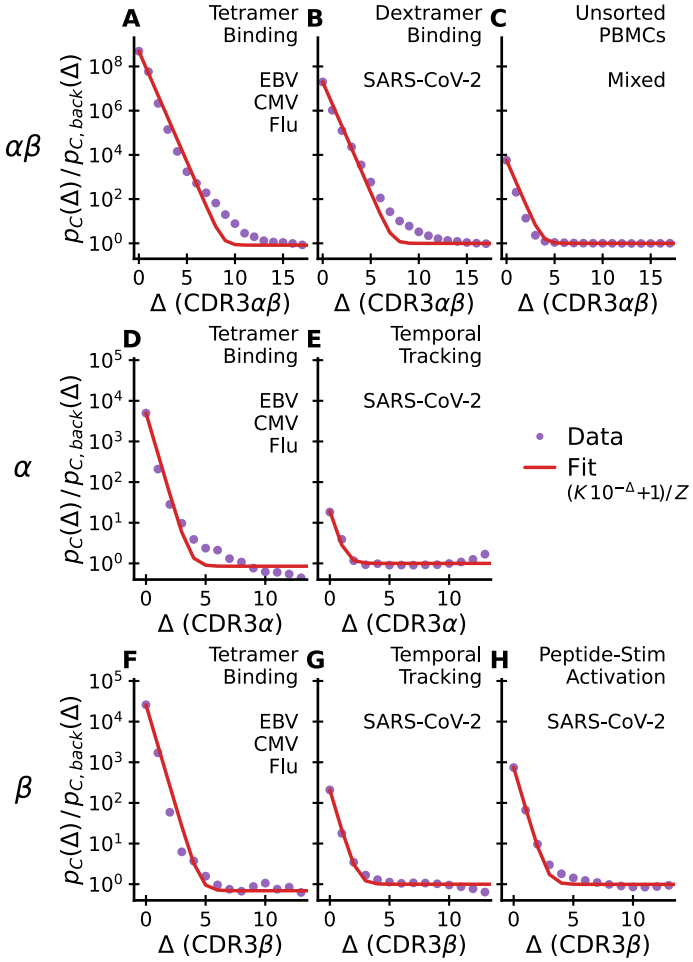

FIG. 3: **Excess coincidences follow a common functional form across experiments.** Sequence similarity of specific T cells for paired $\alpha\beta$-chain repertoires (top row), α-chain repertoires (middle row) and β-chain repertoires (lower row) compared with background expectations. In each panel the assay type used to enrich for epitope-specific T cells and the antigen source are noted in the upper right. Panel C is special as analyzed TCRs are from unsorted blood and have not been explicitly selected for binding to a specific epitope. A common reference curve is plotted for visual guidance. Its parameter K is set equal to the empirical value at $\Delta = 0$. Z is determined by normalization. Datasets: A,D,F - Dash et al. [14]; E,G - Minervina et al. [16]; H - Nolan et al. [15]; B - Minervina et al. [17]; C - Tanno et al. [30].

of ~ 0.1 (the value of p_{corr}) between $\Delta = 0$ and $\Delta = 1$, in accord with our expectation from Eqn. 6.

We note from these illustrations that the the enhancement ratio $p_C(\Delta)/p_{C,back}(\Delta)$ (plotted in the right-hand columns of Fig. 2B-D) gives a particularly direct diagnostic of the nature and strength of the selection that acts on the background. We will use it in the next sections to put a wide range of experimental data into a common framework.

III. COMMON FEATURES OF SELECTION ACROSS MULTIPLE DATASETS

We now use the lens of coincidence analysis to examine a broad set of experimental datasets that use different assays to select T cell repertoires specific to epitopes from different sources (details in Methods and Data) [14–17, 30]. Our analysis of these diverse datasets (Fig. 3) reveals striking similarities in the functional dependence of excess coincidences on sequence distance, together with wide variation in the magnitude of the enhancement of coincidence frequencies over background.

We first apply coincidence analysis to paired chain data from Dash et al. [14] (Fig. 3A), Minervina et al. [17] (Fig. 3B), and Tanno et al. [30] (Fig. 3C), taking the distance between two paired sequences to be the sum of distances between the two chains. Minervina et al. sequenced paired-chain $\alpha\beta$ TCRs that were determined by DNA-barcoded MHC dextramers to have specificity to chosen SARS-CoV-2 epitopes, while Tanno et al. provides a large dataset of paired-chain total T cell repertoires that have not been directly subjected to *ex vivo* selection. We compute the coincidence probability ratio $p_C(\Delta)/p_{C,back}(\Delta)$ against a synthetic background computationally constructed from single chain data under an independent pairing assumption, as described previously.

We next apply coincidence analysis to the single chain data from Nolan et al. [15] (Fig. 3H) and Minervina et al. [16] (Fig. 3E,G). Nolan et al. sequenced β-chain sequences of T cells selected (by passage through the MIRA pipeline [40]) for recognizing individual peptides in a broad panel of peptides from the SARS-Cov-2 genome, while Minervina et al. identified α and β chain sequences of T cells that responded dynamically during the SARS-Cov-2 infection of two human subjects using longitudinal sequencing. As a comparison we also analyze single chain sequences from the Dash et al. [14] paired chain dataset across the three studied viral epitopes, ignoring the chain pairing (Fig. 3D,F). For each repertoire, we compute the coincidence probability ratio $p_C(\Delta)/p_{C,back}(\Delta)$ against background bulk sequences of the same chain. To smooth out variability, we then average over epitopes or subjects, respectively.

Together, these analyses highlight major differences across chains and experiments (Fig. 3, rows and columns, respectively) in how much coincidence probabilities are increased relative to background, $p_C(\Delta)/p_{C,back}(\Delta)$ at small Δ. The fold increase for sequence identity ($\Delta = 0$) is highest in paired chain tetramer-sorted repertoires against immunodominant epitopes of common viruses (Fig. 3A), and decreases from this value when chains are considered separately (Fig. 3 2nd and 3rd row) or in sequence repertoires identified by other assays (Fig. 3 2nd and 3rd column). We will provide a potential mechanistic explanation for some of these differences in Section VI.

There are also some striking common features to note. First, the analyses show that, for small Δ and across experiments, the excess coincidence ratio declines from

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

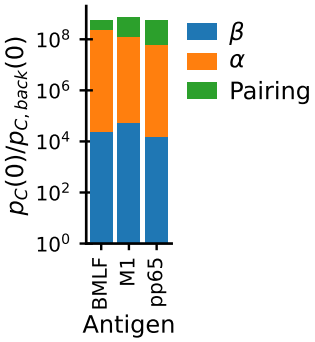

FIG. 4: **Epitope binding restricts diversity of both chains individually, and also restricts their pairing.** Bar chart shows the decomposition of paired chain exact coincidence probability ratios (Fig. 3A) for individual epitopes in the dataset from Dash et al. [14] into contributions from selection of α chains (Fig. 3D) and β (Fig. 3F) individually (blue,orange), plus a smaller contribution from restricting the pairing of the two chains (green).

its value at $\Delta = 0$ at a similar exponential rate; second, across all datasets, coincidence rates reduce to those of the background for distances substantially less than the mean distance between sequences in the background. In other words, the statistical differences between selected repertoires and the background are limited to small sequence distances Δ. The red curves plot a simple parametric function (specified in the legend) that captures the two key features: it interpolates between an initial exponential decrease by roughly one power of ten per unit increase in Δ and asymptotes to a constant. The parameter K is set to the value of excess coincidences at $\Delta = 0$, and the parameter Z is determined self-consistently by normalization. Without any additional fitting parameters the reference curve is in good agreement with the empirical data across all experiments, highlighting their similarity.

The exponential falloff for small Δ is a quantitative measure of binding degeneracy with respect to small sequence changes. According to Eqn. 6 the observed common falloff rate means that, on average, about one tenth of the $\Delta = 1$ sequence neighbors of a T cell that recognizes an epitope will also recognize the same epitope (and roughly one percent of the $\Delta = 2$ neighbors, etc). This degree of sequence degeneracy is observed both for α-chains (Fig. 3D,E) and β-chains (Fig. 3F-H). Note that this analysis relates to the fraction of available sequence neighbors, i.e. those present in the pool before sorting for specificity in accord with the TCR generation probabilities and sample size, and takes into account only the CDR3 region and not other hypervariable regions. The observation that this parameter agrees across experiments and chains is striking, and suggests it is a fundamental biophysical feature of TCR-pMHC binding interactions.

IV. DIVERSITY OF BOTH CHAINS AND THEIR PAIRING IS RESTRICTED IN SPECIFIC TCRS

Epitope-specific repertoires sequenced at the paired-chain level can be used to quantify the relative contribution to binding specificity of the two chains. Figs. 3D,F show that there is, on average, a strong diversity restriction (as measured by excess coincidences) for both chains individually due to epitope selection. If the selected chains could be freely paired without affecting specificity, then the overall excess coincidence factor for paired chains would be the product of the factors for the individual chains (as discussed in Appendix C). In fact, Fig. 3 A shows that paired chain coincidences are more frequent than this expectation by perhaps as much as a factor 10 (out of an overall increase by a factor of $\sim 10^9$). For further insight, we repeated the analysis separately for each individual epitope (Fig. 4): the paired chain selection factor is in each instance the product of two large factors due to selection of the β and α chains individually times a smaller factor that arises from restricting pairing among the selected sets of chains, and there is only limited variation in the contributions of the three terms across epitopes. These results show why paired chain information is essential for accurately predicting the specificity of a TCR. An important correlate of the strong restriction of diversity within epitope-specific repertoires is that when fixing one chain the other shows only very limited variation: As shown in Fig. 1 paired chain coincidences are nearly as frequent as coincidences on either chain alone. A related phenomenon was recently described comparing naive and memory antibodies [41], and termed chain coherence. Our analyses suggest that such coherence also occurs for TCRs.

V. THE SELECTION SIGNATURE CONSTRAINS THE BINDING LANDSCAPE

What are minimal features of a T cell-epitope binding landscape that can explain the coincidence enhancement signature? To explore this question we go beyond the random selection models considered in Fig. 2 and treat selection more realistically as due to sequence-dependent binding. This exercise could be carried out at many levels of sophistication [42, 43], but we will focus on a simple, schematic, and analytically tractable model for TCR-pMHC interactions. In what follows, we sketch the model and the conclusions we draw from it. Details are presented SI Sec. F.

We model TCRs as random amino acid strings of fixed length k=6 (corresponding to the mean number of hypervariable residues within a typical CDR3 loop). Background TCR sequences are generated by drawing 6 amino acids independently at random from the $q = 20$ amino acids. The set of TCRs binding to a specific pMHC is specified by a sequence logo, or motif, condition: at each of the k variable positions, we require that the residue lie

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

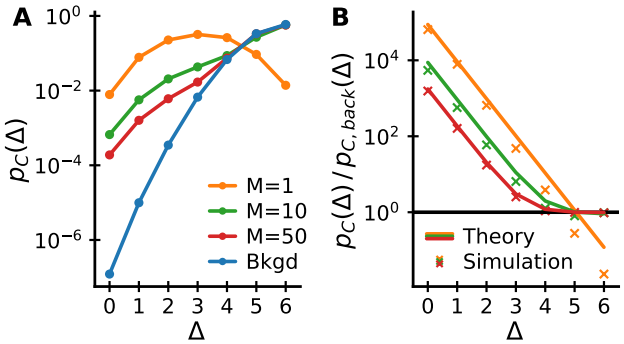

FIG. 5: **Coincidences in a mixture of motifs model.** (A) Coincidence probabilities and (B) coincidence probability ratios to background for simulated data generated from a mixture of motifs model with different numbers of motifs M and $c = 3$. (B) also shows analytical expectations from Eqn. 8 (lines), which agree well with the numerical results (crosses). The model reproduces key features of the empirical data: $p_C/p_{C,back}$ decays exponentially for small Δ and asymptotes to a constant for large Δ at sufficiently large M.

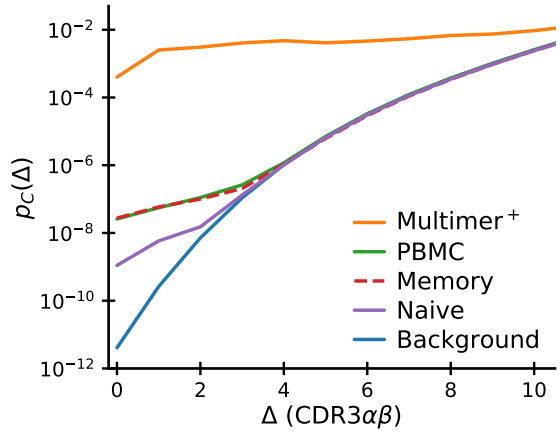

FIG. 6: **Comparison of near-coincidence probabilities across paired-chain datasets.** The highest values come from TCR repertoires specific to individual epitopes (solid orange curve: average over epitopes studied in Dash et al. [14] and Minervina et al. [17]). Paired-chain sequencing of whole blood (green), sorted CD4$^+$ memory (dashed red) and CD4$^+$ naive (purple) repertoires (data averaged over subjects from Tanno et al. [30]) give much smaller values. Background coincidence probabilities (calculated assuming independent chain pairing) are shown in blue. See text for discussion of the large difference in coincidence probabilities between repertoires.

in a randomly chosen subset of size $c \leq q$ of the amino acids (a different subset at each position).

Calculating the coincidence enhancement factor for a particular epitope and binding motif reduces to a combinatorial excercise in this model, with the result:

$$\frac{p_C(\Delta)}{p_{C,back}(\Delta)} = \left(\frac{c-1}{q-1}\right)^{\Delta} \left(\frac{q}{c}\right)^{k}. \tag{7}$$

This expression reproduces the exponential falloff of excess near-coincidences with Δ that is seen in real data. The falloff rate depends on the number of allowed amino acids c at each position, with $c \sim 3$ amino acids per position reproducing the empirical rate.

However, this expression does not capture the second observation in the empirical data, namely, that beyond a certain sequence distance Δ, the enhancement ratio asymptotes to a roughly constant value. To address this, we recall that Fig. 1 strongly suggests that there are multiple "solutions" to the problem of recognizing a given epitope. Sequence similarity between TCRs binding in different manners is expected to be low, thus the existence of multiple solutions might explain the flattening of the coincidence probabilities for large Δ. We thus extend our binding model to incorporate this idea: For each epitope, let there be M different randomly chosen motifs and declare that a T cell recognizes the epitope if any of the motifs are satisfied. T cells selected by this model are a mixture of those selected by the individual motifs. Applying results for coincidences in mixture distributions (derived in Appendix B), we obtain an analytical prediction for excess coincidences:

$$\frac{p_C(\Delta)}{p_{C,back}(\Delta)} \approx \frac{1}{M}\left(\frac{c-1}{q-1}\right)^{\Delta} \left(\frac{q}{c}\right)^{k} + 1 - \frac{1}{M}. \tag{8}$$

Fig. 5 displays this analytical result for different values of M. In addition it shows the almost identical results of numerical simulations of the model with a more realistic non-uniform amino acid usage (drawn according to the amino acid usage in CDR3α hypervariable chains reported in [16]). The key observation is that, for multiple motifs, the ratio $p_C/p_{C,back}$ both shows exponential decay for small Δ and asymptotes to a constant (close to unity) as Δ approaches the maximum possible value in this setup, $\Delta = 6$.

VI. FUNCTIONAL DIVERSITY LINKS COINCIDENCES ACROSS SCALES

We now revisit the intriguing observation of a selection-like signature in paired chain sequencing data from whole blood (specifically, the coincidence enhancement displayed in Fig. 3C). In Fig. 6 we compare coincidence frequencies obtained from direct paired chain sequencing of blood samples with coincidence frequencies among multimer-sorted T cells that recognize individual epitopes. We note that coincidences within multimer-sorted repertoires exceed those in blood samples by four orders of magnitude. Also, the comparison with sorted memory and naive repertoires shows that coincidences in the total repertoire are primarily driven by memory cells. Bearing in mind that the whole blood coincidence analysis compares sequences

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

within and between all the memory sub-compartments created by past infections, we hypothesize that the coincidences in whole blood reflect high-levels of sequence similarity among groups of memory cells selected in response to specific epitopes encountered in the past. Intuitively, we then expect coincidences in whole blood to depend on the diversity of the memory repertoire, i.e. on how many different epitope exposures the immune system is remembering. To make this intuition quantitative we develop a mathematical formalism to predict coincidences in mixture distributions.

We propose to model TCRs in an individual's memory compartment as a mixture distribution over the set Π of peptide-MHC complexes (pMHCs) that have driven past immune responses in that individual. For each $\pi \in \Pi$, there is a distribution of T cell sequences $P(\sigma|\pi)$ that target π. The distribution of TCRs in the memory compartment will then be the mixture distribution

$$P(\sigma) = \sum_{\pi \in \Pi} P(\sigma|\pi) P(\pi), \tag{9}$$

where $P(\pi)$ is the proportion of all TCRs selected for binding to pMHC π. The coincidence probability for mixtures can be calculated using the following mixture decomposition theorem, which we derive in Appendix B:

$$\begin{aligned} p_C[P(\sigma)] = & \, p_C[P(\pi)] \, \langle p_C[P(\sigma|\pi)] \rangle \\ & + (1 - p_C[P(\pi)]) \, \langle p_C[P(\sigma|\pi_1), P(\sigma|\pi_2)] \rangle, \end{aligned} \tag{10}$$

where the averages are over $P(\pi|\pi_1 = \pi_2 = \pi)$ and $P(\pi_1, \pi_2|\pi_1 \neq \pi_2)$, respectively. It is noteworthy that such an exact decomposition of coincidence probabilities in mixtures exists. For example, no equivalent formula exists for Shannon entropy, an alternative measure of diversity, which has led to long-running debates within ecology about the decomposition of diversity in pooled communities [44–46].

Eqn. 10 is a sum of two non-negative terms, each of which can be given an intuitive interpretation. We recall that the probability of exact coincidence is the probability with which two randomly chosen sequences σ_1 and σ_2 are coding for the same TCR, $p_C[P(\sigma)] = \sum_{\sigma_1,\sigma_2} P(\sigma_1) P(\sigma_2) I_{d(\sigma_1,\sigma_2)=0}$. The decomposition formula then represents a conditioning on the mixture identity for σ_1 and σ_2: The overall probability of coincidence is a weighted mean of average within-group coincidence probabilities (first term) and of average between-group coincidence probabilities (second term). The relative weight given to within group comparisons is given by the probability with which two randomly chosen elements come from the same group, i.e. the coincidence probability of the group assignments $p_C[P(\pi)]$ (defined in the sense of Eqn. 2).

Multimer sorting followed by sequencing gives draws from $P(\sigma|\pi)$ for specific pMHCs π [14, 17], and this data can be used to estimate the average within-epitope-group coincidence probability $\langle p_C[P(\sigma|\pi)] \rangle$. Absent better information, we shall assume that the average value $p_C[P(\sigma|\pi)] \sim 10^{-4}$ in these experiments is the typical order of magnitude for all epitopes. We further assume that the between-epitope-group term in Eqn. 10 is negligible. Then the only remaining quantity is $p_C[P(\pi)]$, the Simpson diversity of the set of epitope-specific groups within the repertoire. Putting the numbers together we obtain an effective diversity $1/p_C[P(\pi)] \sim 10^4$, a not implausible value for the pMHC diversity of a memory compartment.

In other words, the large ratio between coincidence frequencies in a repertoire selected *ex vivo* by an individual pMHC complex and the coincidence frequencies in the memory compartment as a whole is informative about the number of epitope recognition events that created the memory compartment in the first place. While the precise numbers are likely to change as more comprehensive data becomes available, the calculation above gives a clear recipe to settle the question of how functionally diverse our immune repertoire is. More broadly, mixture averaging also likely explains why coincidence probabilities among longitudinally identified TCRs (presumably specific to multiple immunodominant epitopes) are lower than among TCRs specific to individual epitopes (Fig. 3E vs. 3D and Fig. 3G vs. 3F,H).

VII. HLA OVERLAP DETERMINES COINCIDENCES BETWEEN DONORS

How many TCRs are shared between donors? In previous studies of T cell repertoires, there has been much interest in such shared sequences, on the grounds that such 'public' sequences may point towards common pathogen exposures [39, 47]. Since in order to mount a common response to a pathogen epitope, two subjects must not only share (up to near-coincidence) T cells with the same TCR, but must also share an MHC molecule on which the epitope can be presented, we expect more T cell sharing between donors that share HLA alleles. In line with this expectation, Tanno et al. [30] observed an association between exact sharing of paired $\alpha\beta$ TCRs and the number of shared HLA alleles. By our logic, it makes sense to broaden the definition of public T cells to those that are nearly coincident across donors, and present at rates well above an appropriately estimated background. We will thus revisit the analysis by Tanno and co-workers by applying our coincidence analysis framework to their dataset. Specifically, we calculate the histogram of sequence distances between TCRs drawn from pairs of repertoires, and ask how the strength of any selection signal depends on the similarity of HLA type between the two repertoires.

We grouped subject pairs by HLA overlap defined as $J = |A \cap B| / \max(|A|, |B|)$, where A and B are the sets of HLA alleles in the two subjects. The overlap ranges between $J = 1$ for identical twins, to $J = 0$ if there is no common HLA allele. We also applied additional filtering steps to control for confounding factors (Appendix E). To mimic the filtering applied to intra-sample analyses of the data from Tanno et al. [30], we did not count coinci-

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

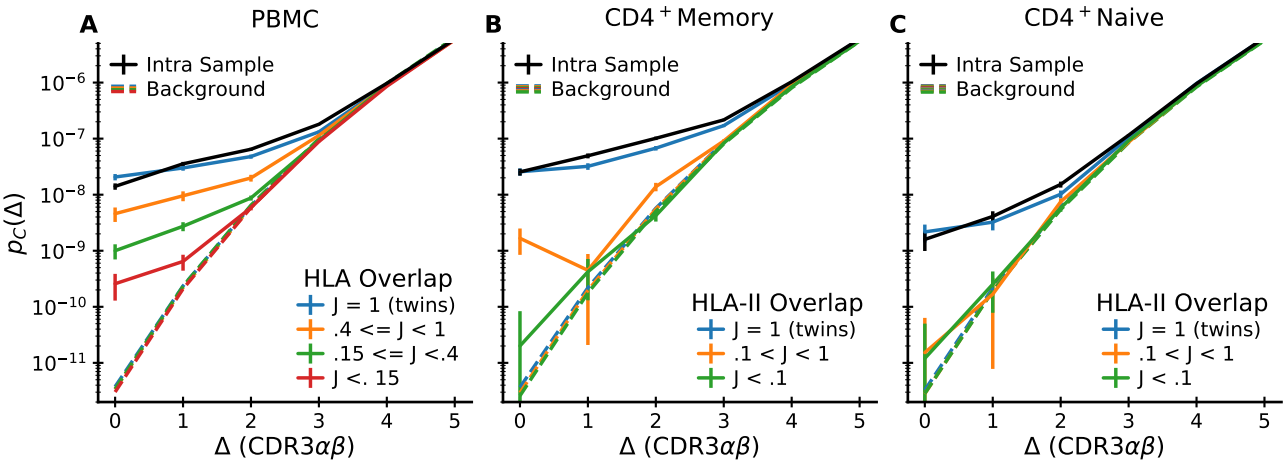

FIG. 7: **Inter-subject coincidences depend on HLA overlap.** Pairwise inter-dataset coincidence frequency analysis for the 15 paired-seq datasets from Tanno et al. [30] grouped by pairwise HLA overlap. A: pairs of unsorted PBMC repertoires; B: pairs of CD4$^+$ memory repertoires; C: pairs of CD4$^+$ naive repertoires. Each plot shows means over pairs whose HLA overlap lies within the indicated ranges together with estimated standard errors assuming Poisson sampling. For comparison, the mean intra-dataset coincidence distribution is shown in black. Background distributions constructed by scrambling the α and β chain associations within individuals are shown as dashed curves (colored according to the same HLA overlap code). These curves show no near-coincidence enhancement signal and very weak dependence on HLA overlap class.

dences where either chain had exact nucleotide identity. This filtering also allowed us to exclude exact nucleotide coincidences when comparing repertoires of twins. Exact nucleotide-level sharing of full $\alpha\beta$ TCRs between twins can represent long-lived clones shared via the blood supply during fetal development [48, 49], and is thus not necessarily evidence of convergent selection on the TCRs. Additionally, we removed sequences whose α-chain V and J genes match those of two non-canonical T cell subsets, mucosal associated invariant T cells (MAITs) and invariant natural killer T cells (iNKTs), that recognize non-peptide ligands not presented on classical MHC [50].

The results of the analysis are shown in Fig. 7: Near-coincidence probabilities between whole blood repertoires decrease systematically with decreasing HLA overlap (Fig. 7A), and the same trend holds in sorted CD4$^+$ memory (Fig. 7B) and CD4$^+$ naive cells (Fig. 7C). These HLA-dependent effects are large: exact coincidence probabilities range over two orders of magnitude as HLA overlap varies. This contrasts with prior studies that have found only a small influence of HLA type in single chain repertoires [51]. The interpretation suggested by our earlier analysis (Fig. 4) is that HLA binding requires specific pairs of α and β chains. To confirm that our observed large effect sizes are compatible with weak signals in single chain repertoires, we constructed synthetic distributions for randomized $\alpha\beta$ pairings by convolving the single chain distance distributions within HLA overlap groups. The results are shown as dashed lines in Fig. 7 (using the same color coding for the HLA overlap groups as for the real data). They reveal that single chain coincidences are almost independent of HLA overlap, even though this procedure retains the correlations between individual chains and HLA type.

The comparison of coincidence probabilities between these different ways of filtering and segregating the data is informative about how different mechanisms might contribute to chain pairing biases. First, Fig. 7C shows no significant deviations from pairing independence (dashed lines in the figure) across naive cells from non-twin donors. This limits the strength of chain pairing correlations that might arise through pMHC-independent processes, such as VDJ recombination, or from steric and biophysical constraints between chains for protein folding [33, 34]. We note that this finding validates the use of the independent chain pairing assumption for generating background distributions representative of repertoire statistics before selection has acted. Second, Fig. 7C also shows a clear signal of correlated chain pairing in naive cells both intra-sample (black line) and across twin pairs (blue line). This strongly suggests that thymic positive and negative selection substantially contribute to the pairing biases. Third, Fig. 7B shows that within the memory repertoire coincidences between twins occur at remarkably similar rates to the intra-sample coincidence rates, which suggests that memory selection is driven by prevalent pMHCs encountered by both donors (herpesviruses are one potential source of such pMHCs [52]). Alternatively, sequences binding a certain HLA might generally show substantially restricted pairing independently of which peptide is presented [53, 54] – something we will soon be able to test as more epitope-specific repertoires for different peptides presented on the same MHC are characterized. In summary, HLA-dependent selection leads to major biases in

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the pairing of TCR α and β chains at the repertoire level, the outcome of a combination of thymic and peripheral selection pressures. As dataset sizes continue to increase, the strategy we have described here provides a strategy for untangling these pressures in detail.

VIII. DISCUSSION

In this work we have introduced a versatile statistical framework for measuring selection in T cell receptor repertoires. Simply put, we have evidence of selection if the number of exactly (and nearly exactly) coincident receptor sequence pairs in a repertoire is substantially larger than the number that one would find in a reference repertoire. Importantly, we showed that this intuitive notion can be developed into a mathematical theory relating the number of excess coincidences to quantities of direct immunological interest, such as the extent of sequence degeneracy of T cell binding to particular epitopes, or the functional diversity of an individual's memory repertoire.

We take a probabilistic approach to selection, where each target epitope defines a probability distribution on the unselected, or naive, T cells that make up the immune repertoire. Experiments that query blood samples for binding to a specific pMHC represent a draw from this probability distribution, and experiments that capture T cell responses to multiple targets sample a mixture of distributions over targets. Certain global quantities of immunological interest are averages over these distributions and, in our approach, the experimental data serve to give empirical estimates of these averages. We highlight two salient examples:

First, we quantify the fraction of sequence neighbors of a typical specific sequence that share the same specificity. Our analysis predicts that when varying single amino acids in the hypervariable regions in accord with the TCR generative statistics, roughly one out of ten such changes lead to a receptor that still binds the same target. Across disparate datasets this measure of local recognition degeneracy shows remarkable consistency. We envisage that it can be used to guide bioinformatic clustering methods for finding groups of T cells with common specificity [14, 19, 55], for instance to put data-driven constraints on threshold choices. Importantly, the predicted level of local degeneracy is in rough accord with measured distributions of binding affinity changes between point-mutated TCRs [56, 57], and results from systematic mutational scans of specific binding upon changes in TCR hypervariable regions [58]. To quantitatively compare our results with such scans, it will be necessary to develop a framework for appropriately weighting the exhaustive mutational scanning data by the probability with which mutated TCRs occur in natural repertoires. With the rapid increase in the number of assayed epitopes, another area for future work will be to characterize in detail variation around the average selection strength and binding degeneracy, including for example between TCRs binding MHC-I or MHC-II (most data analyzed in the current study relates to MHC-I binding).

Second, we provided a recipe to quantify the functional diversity of a T cell compartment, as measured by the number of different epitopes that have selected the T cells comprising the compartment. From paired-chain sequencing data on human blood samples [30] we derived a rough estimate of the functional diversity of a typical memory compartment. This coarse-grained functional diversity is orders of magnitude smaller than TCR sequence diversity, which is consistent with the relatively small number of immunodominant epitopes typically targeted in response to individual pathogen infections [59] and theoretical predictions that adaptive immunity learns sparse features of the epitope distribution [60]. Additionally, cumulative coincidence probabilities at different sequence distances should provide a useful measure of repertoire diversity weighted by sequence similarity, a subject of recent interest in the field [38, 57, 61].

Beyond the quantification of functional diversity our analysis of deeply sequenced paired chain repertoires across individuals suggests additional research directions. We identify a substantial number of TCR specificity groups that the data suggest are in large part driven by common epitopes across individuals. Guided by such TCR groups it would be interesting to generalize the recently proposed reverse epitope discovery approach [62, 63] to the repertoire scale: cross-referencing coincident TCRs with other data, such as TCR-epitope databases [20, 64] and computationally predicted HLA binding of putative peptides, might guide the identification of the targets of these groups of T cells. More broadly, as dataset sizes increase an analysis of the dependence of cross-donor coincidence probabilities on which HLAs the two donors share could allow an unbiased apportionment of the immune repertoire selected by different HLA types.

In summary, our results reveal both complexity and predictability in the immune receptor code. The emerging picture is captured schematically in the mixture of motifs model that we have introduced: Epitope-specific repertoires are characterized by globally diverse binding solutions that sometimes share surprisingly little sequence similarity, but also display remarkably consistent signatures of local degeneracy. This picture, if further confirmed in structural studies [8, 9], can help focus future machine learning efforts in this area. The consistent signal of local degeneracy suggests that a promising direction will be to use machine learning to refine metrics, such as TCRdist [14, 55], that can group TCRs specific to a common target within large mixtures. Our framework should be of use in such efforts, as it can readily turn any definition of TCR similarity, not just the simple edit distance we have considered here, into probabilities of shared specificity. The existence of multiple binding solutions, on the other hand, might explain why purely sequence-based models for computationally predicting binding partners of epitopes (i.e. in the absence of any experimentally determined binders) have had limited success [23], and

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

why structural modeling might be needed to resolve the complex sequence determinants of the different binding solutions [65].

Methods and Data

In this paper, we analyze datasets that represent significantly different approaches, both conceptual and experimental, to creating functionally selected T cell repertoires. They are succinctly described as follows:

The Dash dataset [14] is based on tetramer sorting of $CD8^+$ T cells from blood, using three well-studied standard viral epitopes (HLA-A*02:01-BMLF1$_{280}$ (BMLF), HLA-B*07:02-pp65$_{495}$ (pp65), and HLA-A*02:01-M1$_{58}$ (M1)), followed by single-cell TCR sequencing to obtain paired TCRα and TCRβ reads of the captured cells. This protocol was repeated for 32 donors, resulting in a list of 415 paired $\alpha\beta$ TCRs associated with the three epitopes.

The Minervina 2022 dataset [17] uses DNA-barcoded MHC dextramers to identify T cells specific to 19 SARS-CoV-2 epitopes by sequencing. T cells were identified across a cohort of donors with a varied history of SARS-CoV-2 exposure and vaccination. We focused our analysis on the 8 epitopes for which there are at least 150 characterized $\alpha\beta$ TCRs each.

The Nolan dataset [15] is obtained by sorting about 3×10^7 T cells from a subject blood sample, then incubating the sorted cells with a cocktail of several hundred SARS-CoV-2 epitopes (chosen for their broad MHC presentability) to uniformly expand clones that recognize any of these epitopes. In a next step, aliquots of the expansion product are incubated with individual epitopes from the cocktail, followed by TCRβ sequencing to identify T cells that have expanded in this second step in response to individual epitopes. This yields a list of TCRβ clonotypes that recognize the epitope. This protocol is repeated for blood samples from about a hundred subjects, about a third of whom have had no known exposure to SARS-CoV-2 ("healthy" subjects). Summing over subject samples for each epitope, we get a list of a few tens to a few thousand clonotypes that recognize a given epitope. All told, the dataset is a list of some 10^5 TCRβ recombination events that respond to individual SARS-CoV-2 epitopes. We note that the α chains associated with each β chain are not known, and also that a given epitope may be presented on different MHC molecules in different individuals. To have adequate statistical power, we consider only epitopes from Nolan et al. [15] which are recognized by at least 150 distinct clones and we restrict our analysis to MHC-I epitopes.

The Minervina 2021 dataset [16] is based on a longitudinal study of TCRβ sequences in the blood of two unrelated subjects who contracted mild COVID-19. Analysis of time-separated samples allowed the identification of T cell clones, whose clone sizes changed significantly in response to infection. We focus on the several hundred $CD8^+$ clones, whose size decreased between the peak immune response at 15 days and a post-infection time point at 85 days. The specific epitopes to which these T cells respond are unknown, but they are presumably a subset of the SARS-CoV-2 viral epitopes that provoke the strongest immune response, and therefore constitute an interesting "selected" subset of the T cell repertoire.

The Tanno dataset [30] consists of paired-chain TCRs from a total of fifteen donors, including six pairs of twins. The mean number of reads is about 31,000 (minimum of 7400 and maximum of 69000). For three pairs of twins and three unrelated donors total PBMC samples were sequenced. Sorted $CD4^+$ naive ($CD45RA^+$, $CCR7^+$) and memory ($CD45RA^-$) cells were sequenced for three additional twin pairs. All fifteen subjects were HLA typed on the allele level. We used processed data as described in the original study but applied additional filtering steps, the rationale for which is described in Appendix E. For the naive repertoire we also removed any overlap with clonotypes that were also found within the memory repertoire from the same individual. To compare coincidence frequencies across repertoires from different individuals (Fig. 7) we sum the number of coincidences across all comparisons within an HLA overlap bin. We add a pseudocount of 0.1 to the summed counts for visualization purposes, and we display Poisson errorbars as $\sqrt{c}/c_{tot}$ where c is the count at a specific distance and c_{tot} the sum of all counts across distances. These errorbars represent lower bounds, as in addition to counting error there is heterogeneity between individuals.

For all datasets, we filtered out clones whose CDR3 amino acid sequence did not start with the conserved cysteine (C), or end on phenylalanine (F), tryptophan (W), or cysteine (C).

To calculate paired chain background coincidence probability distributions, we randomly associate chains from bulk single chain datasets. For efficient numerical calculation we exploit the fact that such independent pairing leads to coincidence probability distributions for paired chain TCRs that are a convolution of the single chain distributions, $p_{C,\alpha\beta}(\Delta) = \sum_{\delta=0}^{\Delta} p_{C,\alpha}(\delta) p_{C,\beta}(\Delta - \delta)$.

To generate the sequence logos displayed in Fig. 1 we built on the Python logomaker package [66], adding the ability to also display V and J gene usage. We colored amino acids by their chemical properties using the "chemistry" color scheme.

Acknowledgements. We are grateful to Yuval Elhanati and Léo Régnier for previous collaboration on the sequence space structure of T cell repertoires, unpublished work that was essential to the development of the ideas underlying the current work. We thank Hidetaka Tanno for providing processed data files and Giulio Isacchini for helpful discussions.

Funding information. The work of AM was supported by a Lewis-Sigler fellowship; the work of CC was supported in part by NSF grants PHY-1607612 and PHY-1734030. CC is grateful to the Lustgarten Foundation for support for an extended visit to the Institute for Advanced

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Study, where some of this work was performed. CC is also grateful to the Ecole Normale Superieure, Paris, for hospitality during parts of the research reported here.

Author contributions and competing interests. CC and AM conceived research, performed analysis, and wrote the paper. CC and AM declare no competing interests.

Data and software availability. All the data used in this work has been published and is available online (see the data publications for detailed instructions). We have made an implementation of coincidence analysis available as an open source Python package `https://github.com/andim/pyrepseq`.

Significance statement. Adaptive immunity relies on the binding of molecular targets by a few specific T cells out of a highly diverse repertoire. Different T cell receptors can bind the same target, but a quantification of this recognition degeneracy is lacking. Here we develop a statistical approach that links distributions of sequence similarity among T cells of common specificity to how binding probability co-varies with sequence. Applying our method to experimental data we determine the fraction of sequence neighbors of a specific T cell that also bind its target and we estimate how many response groups make up a memory compartment. Our study provides a quantitative framework for identifying the sequence determinants of specific binding, and will thus facilitate the development of repertoire sequencing-based immunodiagnostics.

[1] M. M. Davis and P. J. Bjorkman, Nature **334**, 395 (1988).

[2] H. S. Robins, P. V. Campregher, S. K. Srivastava, A. Wacher, C. J. Turtle, O. Kahsai, S. R. Riddell, E. H. Warren, and C. S. Carlson, Blood **114**, 4099 (2009).

[3] R. O. Emerson, W. S. DeWitt, M. Vignali, J. Gravley, J. K. Hu, E. J. Osborne, C. Desmarais, M. Klinger, C. S. Carlson, J. A. Hansen, et al., Nature Genetics **49**, 659 (2017).

[4] J. M. Heather, M. Ismail, T. Oakes, and B. Chain, Briefings in Bioinformatics **19**, 554 (2017).

[5] P. Bradley and P. G. Thomas, Annual Review of Immunology **37**, 547 (2019).

[6] K. C. Garcia, M. Degano, R. L. Stanfield, A. Brunmark, M. R. Jackson, P. A. Peterson, L. Teyton, and I. A. Wilson, Science **274**, 209 (1996).

[7] M. G. Rudolph, R. L. Stanfield, and I. A. Wilson, Annual Review Of Immunology **24**, 419 (2006).

[8] J. Rossjohn, S. Gras, J. J. Miles, S. J. Turner, D. I. Godfrey, and J. McCluskey, Annual Review of Immunology **33**, 169 (2015).

[9] I. Song, A. Gil, R. Mishra, D. Ghersi, L. K. Selin, and L. J. Stern, Nature Structural & Molecular Biology **24**, 395 (2017).

[10] C. H. Coles, R. M. Mulvaney, S. Malla, A. Walker, K. J. Smith, A. Lloyd, K. L. Lowe, M. L. McCully, R. Martinez Hague, M. Aleksic, et al., The Journal of Immunology **204**, 1943 (2020).

[11] P. Zareie, C. Szeto, C. Farenc, S. D. Gunasinghe, E. M. Kolawole, A. Nguyen, C. Blyth, X. Y. Sng, J. Li, C. M. Jones, et al., Science **372**, eabe9124 (2021).

[12] P. D. Holler, P. O. Holman, E. V. Shusta, S. O. Herrin, K. D. Wittrup, and D. M. Kranz, PNAS **97**, 5387 (2000).

[13] Y. Li, R. Moysey, P. E. Molloy, A.-l. Vuidepot, T. Mahon, E. Baston, S. Dunn, N. Liddy, J. Jacob, B. K. Jakobsen, et al., Nature Biotechnology **23**, 349 (2005).

[14] P. Dash, A. J. Fiore-Gartland, T. Hertz, G. C. Wang, S. Sharma, A. Souquette, J. C. Crawford, E. B. Clemens, T. H. Nguyen, K. Kedzierska, et al., Nature **547**, 89 (2017).

[15] S. Nolan, M. Vignali, M. Klinger, J. N. Dines, I. M. Kaplan, T. Craft, K. Boland, M. Mazza, and K. Dobbs, Research Square Preprint (2020).

[16] A. A. Minervina, E. A. Komech, A. Titov, M. B. Koraichi, E. Rosati, I. Z. Mamedov, A. Franke, G. A. Efimov, D. M. Chudakov, T. Mora, et al., eLife **10**, e63502 (2021).

[17] A. A. Minervina, M. V. Pogorelyy, A. M. Kirk, J. C. Crawford, E. K. Allen, C. H. Chou, R. C. Mettelman, K. J. Allison, C. Y. Lin, D. C. Brice, et al., Nature Immunology **23**, 781 (2022).

[18] P. Bacher, E. Rosati, D. Esser, P. Koehler, A. Franke, and A. Scheffold, Immunity **53**, 1258 (2020).

[19] J. Glanville, H. Huang, A. Nau, O. Hatton, L. E. Wagar, F. Rubelt, X. Ji, A. Han, S. M. Krams, C. Pettus, et al., Nature **547**, 94 (2017).

[20] M. Shugay, D. V. Bagaev, I. V. Zvyagin, R. M. Vroomans, J. C. Crawford, G. Dolton, E. A. Komech, A. L. Sycheva, A. E. Koneva, E. S. Egorov, et al., Nucleic Acids Research **46**, D419 (2017).

[21] H. Huang, C. Wang, F. Rubelt, T. J. Scriba, and M. M. Davis, Nature Biotechnology **38**, 1194 (2020).

[22] S. Gielis, P. Moris, W. Bittremieux, N. De Neuter, B. Ogunjimi, K. Laukens, and P. Meysman, Frontiers in Immunology **10**, 2820 (2019).

[23] D. S. Fischer, Y. Wu, B. Schubert, and F. J. Theis, Molecular Systems Biology **16**, e9416 (2020).

[24] W. Zhang, P. G. Hawkins, J. He, N. T. Gupta, J. Liu, G. Choonoo, S. W. Jeong, C. R. Chen, A. Dhanik, M. Dillon, et al., Science Advances **7**, eabf5835 (2021).

[25] A. Montemurro, V. Schuster, H. R. Povlsen, A. K. Bentzen, V. Jurtz, W. D. Chronister, A. Crinklaw, S. R. Hadrup, O. Winther, B. Peters, et al., Communications Biology **4**, 1060 (2021).

[26] B. Bravi, V. P. Balachandran, B. D. Greenbaum, A. M. Walczak, T. Mora, R. Monasson, and S. Cocco, PLoS Computational Biology **17**, e1009297 (2021).

[27] A. K. Sewell, Nature Reviews Immunology **12**, 669 (2012).

[28] J. Robinson, J. A. Halliwell, J. D. Hayhurst, P. Flicek, P. Parham, and S. G. Marsh, Nucleic Acids Research **43**, D423 (2015).

[29] T. Mora and A. Walczak, in *Systems Immunology*, edited by J. Das and C. Jayaprakash (CRC Press, 2019), chap. 11.

[30] H. Tanno, T. M. Gould, J. R. Mcdaniel, W. Cao, Y. Tanno, and R. E. Durrett, Proceedings of the National Academy

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of Sciences **117**, 532 (2020).
[31] V. R. Buchholz, M. Flossdorf, I. Hensel, L. Kretschmer, B. Weissbrich, P. Gräf, A. Verschoor, M. Schiemann, T. Höfer, and D. H. Busch, Science **340**, 630 (2013).
[32] C. Gerlach, J. C. Rohr, L. Perié, N. Van Rooij, J. W. Van Heijst, A. Velds, J. Urbanus, S. H. Naik, H. Jacobs, J. B. Beltman, et al., Science **340**, 635 (2013).
[33] T. Dupic, Q. Marcou, A. M. Walczak, and T. Mora, PLoS Computational Biology **15**, e1006874 (2019).
[34] D. S. Shcherbinin, V. A. Belousov, and M. Shugay, PLoS Computational Biology **16**, e1007714 (2020).
[35] A. Murugan, T. Mora, A. M. Walczak, and C. G. Callan, Proceedings of the National Academy of Sciences **109**, 16161 (2012).
[36] E. H. Simpson, Nature **163**, 688 (1949).
[37] T. S. Nunnikhoven, American Statistician **46**, 270 (1992).
[38] T. Leinster and C. A. Cobbold, Ecology **93**, 477 (2012).
[39] Y. Elhanati, Z. Sethna, C. G. Callan, T. Mora, and A. M. Walczak, Immunological Reviews **284**, 167 (2018).
[40] M. Klinger, F. Pepin, J. Wilkins, T. Asbury, T. Wittkop, J. Zheng, M. Moorhead, and M. Faham, Plos One **10**, e0141561 (2015).
[41] D. B. Jaffe, P. Shahi, B. A. Adams, A. M. Chrisman, P. M. Finnegan, N. Raman, A. E. Royall, F. Tsai, T. Vollbrecht, D. S. Reyes, et al., Nature **611**, 352 (2022).
[42] A. Kosmrlj, A. K. Jha, E. S. Huseby, M. Kardar, and A. K. Chakraborty, Proceedings of the National Academy of Sciences **105**, 16671 (2008).
[43] J. T. George, D. A. Kessler, and H. Levine, Proceedings of the National Academy of Sciences p. 201708573 (2017).
[44] R. Lande, Oikos **76**, 5 (1996).
[45] L. Jost, *Entropy and diversity* (2006).
[46] A. Chao, N. J. Gotelli, T. C. Hsieh, E. L. Sander, K. H. Ma, R. K. Colwell, and A. M. Ellison, Ecological Monographs **84**, 45 (2014).
[47] V. Venturi, D. a. Price, D. C. Douek, and M. P. Davenport, Nature Reviews Immunology **8**, 231 (2008).
[48] M. V. Pogorelyy, Y. Elhanati, Q. Marcou, A. L. Sycheva, E. A. Komech, V. I. Nazarov, O. V. Britanova, D. M. Chudakov, I. Z. Mamedov, Y. B. Lebedev, et al., PLoS Computational Biology **13**, e1005572 (2017).
[49] M. U. Gaimann, M. Nguyen, J. Desponds, and A. Mayer, eLife **9**, e61639 (2020).
[50] D. I. Godfrey, A. P. Uldrich, J. Mccluskey, J. Rossjohn, and D. B. Moody, Nature Immunology **16** (2015).
[51] S. A. Johnson, S. L. Seale, R. M. Gittelman, J. A. Rytlewski, H. S. Robins, and P. A. Fields, PLoS ONE **16**, 1 (2021).
[52] G. J. Xu, T. Kula, Q. Xu, M. Z. Li, S. D. Vernon, T. Ndung'u, K. Ruxrungtham, J. Sanchez, C. Brander, R. T. Chung, et al., Science **348**, aaa0698 (2015).
[53] N. L. L. Gruta, S. Gras, S. R. Daley, P. G. Thomas, and J. Rossjohn, Nature Reviews Immunology 2018 p. 1 (2018).
[54] J. A. Carter, J. B. Preall, K. Grigaityte, S. J. Goldfless, E. Jeffery, A. W. Briggs, F. Vigneault, and G. S. Atwal, Frontiers in Immunology **10**, 1 (2019).
[55] K. Mayer-Blackwell, S. Schattgen, L. Cohen, J. C. Crawford, A. Souquette, J. A. Gaevert, T. Hertz, P. G. Thomas, P. Bradley, and A. Fiore, eLife **10**, e68605 (2021).
[56] J. Jankauskaite, B. Jiménez-García, J. Dapkunas, J. Fernández-Recio, and I. H. Moal, Bioinformatics **35**, 462 (2019).
[57] R. Arora, H. M. Burke, and R. Arnaout, bioRxiv preprint 483131 (2018).
[58] R. Vazquez-Lombardi, J. S. Jung, F. S. Schlatter, A. Mei, N. R. Mantuano, F. Bieberich, K.-L. Hong, J. Kucharczyk, E. Kapetanovic, E. Aznauryan, et al., Immunity (2022), ISSN 1074-7613.
[59] A. Cassotta, P. Paparoditis, R. Geiger, R. R. Mettu, S. J. Landry, A. Donati, M. Benevento, M. Foglierini, D. J. Lewis, A. Lanzavecchia, et al., Journal of Experimental Medicine **217**, e20200206 (2020).
[60] A. Mayer, V. Balasubramanian, A. M. Walczak, and T. Mora, Proceedings of the National Academy of Sciences **116**, 8815 (2019).
[61] M. Vujović, P. Marcatili, B. Chain, J. Kaplinsky, and T. Andresen, bioRxiv Preprint 2021.01.11.417444v2 (2021).
[62] P. A. Mudd, A. A. Minervina, M. V. Pogorelyy, J. S. Turner, W. Kim, E. Kalaidina, J. Petersen, A. J. Schmitz, T. Lei, A. Haile, et al., Cell **185**, 603 (2022).
[63] M. V. Pogorelyy, E. Rosati, A. A. Minervina, R. C. Mettelman, A. Scheffold, A. Franke, P. Bacher, and P. G. Thomas, Cell Reports Medicine p. 100697 (2022).
[64] N. Tickotsky, T. Sagiv, J. Prilusky, E. Shifrut, and N. Friedman, Bioinformatics **33**, 2924 (2017).
[65] K. K. Jensen, V. Rantos, E. C. Jappe, T. H. Olsen, M. C. Jespersen, V. Jurtz, L. E. Jessen, E. Lanzarotti, S. Mahajan, B. Peters, et al., Scientific Reports **9**, 14530 (2019).
[66] A. Tareen and J. B. Kinney, Bioinformatics **36**, 2272 (2020).
[67] Z. Sethna, G. Isacchini, T. Dupic, T. Mora, M. Walczak, and Y. Elhanati, PLoS Computational Biology **16**, e1008394 (2020).
[68] J. D. Farmer, N. H. Packard, and A. S. Perelson, Physica D: Nonlinear Phenomena **22**, 187 (1986).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Appendix A: Formal treatment of effect of selection on coincidence statistics

1. 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

$$\tilde{P}(\sigma) = Q(\sigma)P(\sigma), \tag{A1}$$

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

$$p_C[\tilde{P}](\Delta) = \sum_{\sigma,\sigma'} P(\sigma)P(\sigma')Q(\sigma)Q(\sigma')I_{d(\sigma,\sigma')=\Delta}. \tag{A2}$$

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

$$p_C[\tilde{P}](\Delta) = \langle Q(\sigma)Q(\sigma')I_{d(\sigma,\sigma')=\Delta}\rangle_{P(\sigma,\sigma')}, \tag{A3}$$

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

$$P(\sigma,\sigma'|d(\sigma,\sigma') = \Delta) = \frac{P(\sigma,\sigma')I_{d(\sigma,\sigma')=\Delta}}{P(d(\tilde{\sigma},\tilde{\sigma}') = \Delta)}, \tag{A4}$$

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

$$p_C[\tilde{P}](\Delta) = \langle Q(\sigma)Q(\sigma')\rangle_{P(\sigma,\sigma'|d(\sigma,\sigma')=\Delta)}\, p_C[P](\Delta), \tag{A5}$$

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

2. 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

$$\langle Q(\sigma)Q(\sigma')\rangle = \mathrm{Cov}(Q(\sigma),Q(\sigma')) + \langle Q(\sigma)\rangle^2, \tag{A6}$$

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

$$p_C[\tilde{P}](\Delta) = (\mathrm{Cov}(Q(\sigma),Q(\sigma')) + 1)\, p_C[P](\Delta) \tag{A7}$$

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. A7 is exact, we define the local neighbor density around a sequence σ,

$$n_\sigma(\Delta) = \sum_{\sigma'} P(\sigma')I_{d(\sigma,\sigma')=\Delta}. \tag{A8}$$

Using this definition we can write

$$\langle Q(\sigma)\rangle_{P(\sigma,\sigma'|d(\sigma,\sigma')=\Delta)} = \left\langle Q(\sigma)\sum_{\sigma'} P(\sigma')I_{d(\sigma,\sigma')=\Delta}/p_C[P](\Delta)\right\rangle_{P(\sigma)} = \langle Q(\sigma)n_\sigma(\Delta)\rangle_{P(\sigma)}\,/\,p_c[P](\Delta). \tag{A9}$$

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

$$\langle Q(\sigma)n_\sigma(\Delta)\rangle_{P(\sigma)} = \langle Q(\sigma)\rangle_{P(\sigma)}\langle n_\sigma(\Delta)\rangle_{P(\sigma)} = \langle Q(\sigma)\rangle_{P(\sigma)}p_C[P](\Delta), \tag{A10}$$

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. A7 is exact.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Appendix B: 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

$$P(\sigma) = \sum_{\pi \in \Pi} P(\sigma|\pi)P(\pi), \tag{B1}$$

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:

$$p_C[P(\tau)] = \sum_\tau \left(\sum_\pi P(\tau|\pi)P(\pi) \right)^2, \tag{B2}$$

$$= \sum_\pi P(\pi)^2 \sum_\tau P(\tau|\pi)^2 + \sum_{\pi_1 \neq \pi_2} P(\pi_1)P(\pi_2) \sum_\tau P(\tau|\pi_1)P(\tau|\pi_2). \tag{B3}$$

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

$$p_C[P_i(\tau), P_j(\tau)] = \sum_\tau P_i(\tau)P_j(\tau), \tag{B4}$$

which allows us to rewrite the coincidence probability s

$$p_C[P(\tau)] = \sum_\pi P(\pi)^2 p_C[P(\tau|\pi)] + \sum_{\pi_1 \neq \pi_2} P(\pi_1)P(\pi_2) p_C[P(\tau|\pi_1), P(\tau|\pi_2)]. \tag{B5}$$

We now further note that

$$P(\pi_1, \pi_2 | \pi_1 \neq \pi_2) = \frac{P(\pi_1)P(\pi_2)}{1 - p_C[P(\pi)]}, \tag{B6}$$

and finally that

$$P(\pi|\pi_1 = \pi_2 = \pi) = \frac{P(\pi)^2}{p_C[P(\pi)]}. \tag{B7}$$

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

$$p_C[P(\tau)] = p_C[P(\pi)] \left\langle p_C[P(\tau|\pi)] \right\rangle_{P(\pi|\pi_1=\pi_2=\pi)} + (1 - p_C[P(\pi)]) \left\langle p_C[P(\tau|\pi_1), P(\tau|\pi_2)] \right\rangle_{P(\pi_1,\pi_2|\pi_1\neq\pi_2)} \tag{B8}$$

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. B4 to inexact coincidences, which defines the generalized cross-sample near-coincidence probabilities,

$$p_C[P_i(\sigma), P_j(\sigma)](\Delta) = \sum_{\sigma,\sigma'} P_i(\sigma)P_j(\sigma') I_{d(\sigma,\sigma')=\Delta}. \tag{B9}$$

Appendix C: 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

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

$$Q(\sigma) = \frac{1}{p(\mathcal{S})} I_{\mathcal{S}}(\sigma), \tag{C1}$$

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

$$\chi_{\alpha\beta} = \frac{p_C[Q(\sigma)P(\sigma)](0)}{p_C[P(\sigma)](0)} = \frac{1}{p(\mathcal{S})}. \tag{C2}$$

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

$$\chi_\alpha = \frac{p_C[Q(\sigma_\alpha)P(\sigma_\alpha)](0)}{p_C[P(\sigma_\alpha)](0)}, \quad \chi_\beta = \frac{p_C[Q(\sigma_\beta)P(\sigma_\beta)](0)}{p_C[P(\sigma_\beta)](0)}, \tag{C3}$$

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:

$$\tilde{P}(\sigma_\alpha) = \sum_{\sigma_\beta} Q(\sigma)P(\sigma) \tag{C4}$$

$$= P(\sigma_\alpha) \sum_{\sigma_\beta} \frac{I_{\mathcal{S}}(\sigma_\alpha, \sigma_\beta)}{P(\mathcal{S})} P(\sigma_\beta), \tag{C5}$$

from which we read off

$$Q(\sigma_\alpha) = \frac{\langle I_{\mathcal{S}}(\sigma_\alpha, \sigma_\beta)\rangle_{P(\sigma_\beta)}}{P(\mathcal{S})}. \tag{C6}$$

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

$$\mathcal{S} = \mathcal{S}_\alpha \times \mathcal{S}_\beta, \tag{C7}$$

of the single chain sets

$$\mathcal{S}_\alpha = \{\sigma_\alpha : \exists \sigma_\beta : (\sigma_\alpha, \sigma_\beta) \in \mathcal{S}\}, \quad \mathcal{S}_\beta = \{\sigma_\beta : \exists \sigma_\alpha : (\sigma_\alpha, \sigma_\beta) \in \mathcal{S}\}, \tag{C8}$$

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

$$P(\mathcal{S}) = P(\mathcal{S}_\alpha)P(\mathcal{S}_\beta). \tag{C9}$$

Furthermore some algebra shows

$$Q(\sigma_\alpha) = \frac{I_{\mathcal{S}_\alpha}(\sigma_\alpha)}{P(\mathcal{S}_\alpha)}, \quad Q(\sigma_\beta) = \frac{I_{\mathcal{S}_\beta}(\sigma_\beta)}{P(\mathcal{S}_\beta)}. \tag{C10}$$

and thus

$$\chi_\alpha = \frac{1}{P(\mathcal{S}_\alpha)}, \quad \chi_\beta = \frac{1}{P(\mathcal{S}_\beta)}. \tag{C11}$$

Combining this result with Eqn. C9 we complete the derivation of the equality

$$\chi_{\alpha\beta} = \chi_\alpha \chi_\beta \tag{C12}$$

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

$$\frac{\chi_{\alpha\beta}}{\chi_\alpha \chi_\beta} \tag{C13}$$

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

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Appendix D: Random models of sequence-correlated selection

In Sec. II 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 [15]) 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.

1. 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

$$s_i = \eta_i + (1 - \eta_i) \min \left[\sum_{j \neq i} A_{ij} \theta_{ij} \eta_j, 1 \right], \tag{D1}$$

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

$$\langle s_i \rangle = \langle \eta_i \rangle + (1 - \langle \eta_i \rangle) \sum_{j \neq i} A_{ij} \langle \theta_{ij} \rangle \langle \eta_j \rangle, \tag{D2}$$

$$\begin{aligned} \langle s_i s_j \rangle =& \langle \eta_i \rangle \langle \eta_j \rangle + (1 - \langle \eta_i \rangle) \sum_{k \neq i} A_{ik} \langle \theta_{ik} \rangle \langle \eta_k \eta_j \rangle + (1 - \langle \eta_j \rangle) \sum_{l \neq j} A_{jl} \langle \theta_{jl} \rangle \langle \eta_l \eta_i \rangle \\ &+ (1 - \langle \eta_i \rangle)(1 - \langle \eta_j \rangle) \sum_{k,l \neq i,j} A_{ik} A_{jl} \langle \theta_{ik} \rangle \langle \theta_{jl} \rangle \langle \eta_k \eta_l \rangle. \end{aligned} \tag{D3}$$

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

$$\langle \eta_i \eta_j \rangle = (1 - \delta_{ij}) \langle \eta_i \rangle \langle \eta_j \rangle + \delta_{ij} \langle \eta_i \rangle, \tag{D4}$$

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

$$\begin{aligned}
\langle s_i s_j \rangle =& \langle \eta_i \rangle \langle \eta_j \rangle \\
&+ (1 - \langle \eta_i \rangle) \langle \eta_j \rangle A_{ij} \langle \theta_{ij} \rangle \\
&+ (1 - \langle \eta_i \rangle) \langle \eta_j \rangle \sum_{k \neq i,j} A_{ik} \langle \theta_{ik} \rangle \langle \eta_k \rangle \\
&+ (1 - \langle \eta_j \rangle) \langle \eta_i \rangle A_{ji} \langle \theta_{ji} \rangle \\
&+ (1 - \langle \eta_j \rangle) \langle \eta_i \rangle \sum_{l \neq i,j} A_{jl} \langle \theta_{jl} \rangle \langle \eta_l \rangle \\
&+ (1 - \langle \eta_i \rangle)(1 - \langle \eta_j \rangle) \sum_{k \neq i,j} A_{ik} A_{jk} \langle \theta_{ik} \rangle \langle \theta_{jk} \rangle \langle \eta_k \rangle \\
&+ (1 - \langle \eta_i \rangle)(1 - \langle \eta_j \rangle) \sum_{k \neq l \neq i,j} A_{ik} A_{jl} \langle \theta_{ik} \rangle \langle \theta_{jl} \rangle \langle \eta_k \rangle \langle \eta_l \rangle.
\end{aligned} \tag{D5}$$

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):

$$\langle s_i s_j \rangle \approx \langle \eta_j \rangle \langle \theta_{ij} \rangle + \langle \eta_i \rangle \langle \theta_{ji} \rangle = 2 q_1 q_2. \tag{D6}$$

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. D2

$$\langle s_i \rangle = q_1 + (1 - q_1) q_1 q_2 N_i \quad \text{with } N_i = \sum_{j \neq i} A_{ij} \tag{D7}$$

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

$$\langle s_i \rangle = q_1(N_i) + (1 - q_1(N_i)) q_2 N_i \left(\frac{1}{N_i} \sum_{j \neq i} A_{ij} q_1(N_j) \right), \tag{D8}$$

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

$$q_1(N_i) = \frac{q_1}{1 + q_2 N_i}, \tag{D9}$$

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

2. 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 approximately 10^5 sequences of naive CD8$^+$ T cells obtained from a human subject (part of the immuneCODE dataset [15]). We

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

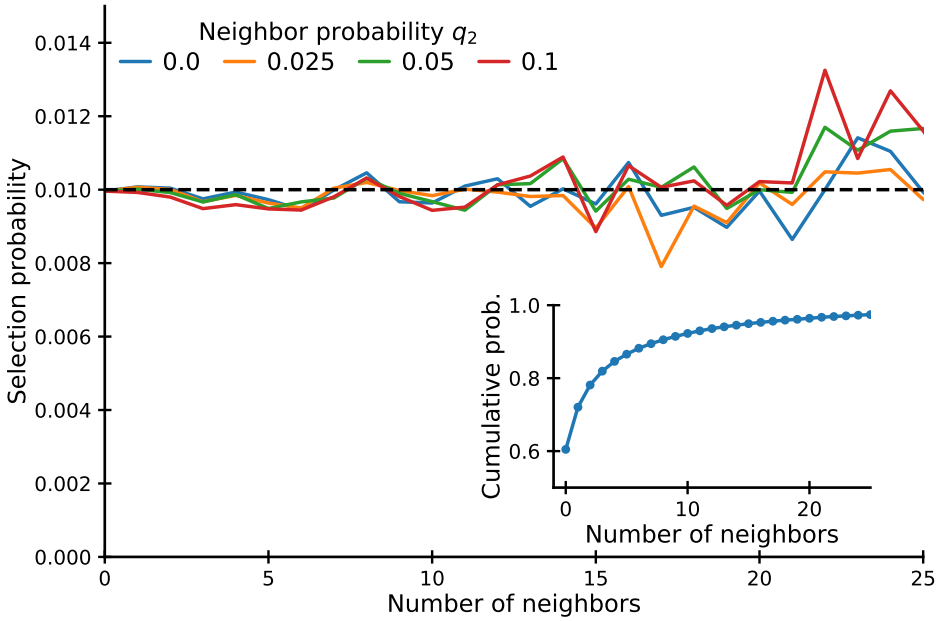

FIG. 8: Assessment of the uniformity of random selection from background sequence repertoires across groups of sequences classified by different numbers of Levenshtein distance one neighbors. The plot shows the fraction of each class in the background repertoire that appears in the selected repertoire, averaged over 100 realizations of the selection algorithm.

generated 100 selected repertoires of $\sim 10^3$ sequences from this background using the neighbor number corrected algorithm. In this background repertoire, the average number of distance-one neighbors of a sequence is close to three and the variance is very large. 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. 8 we plot the probability of selection of a recombination event from the background repertoire, averaged over 100 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 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. 9). 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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

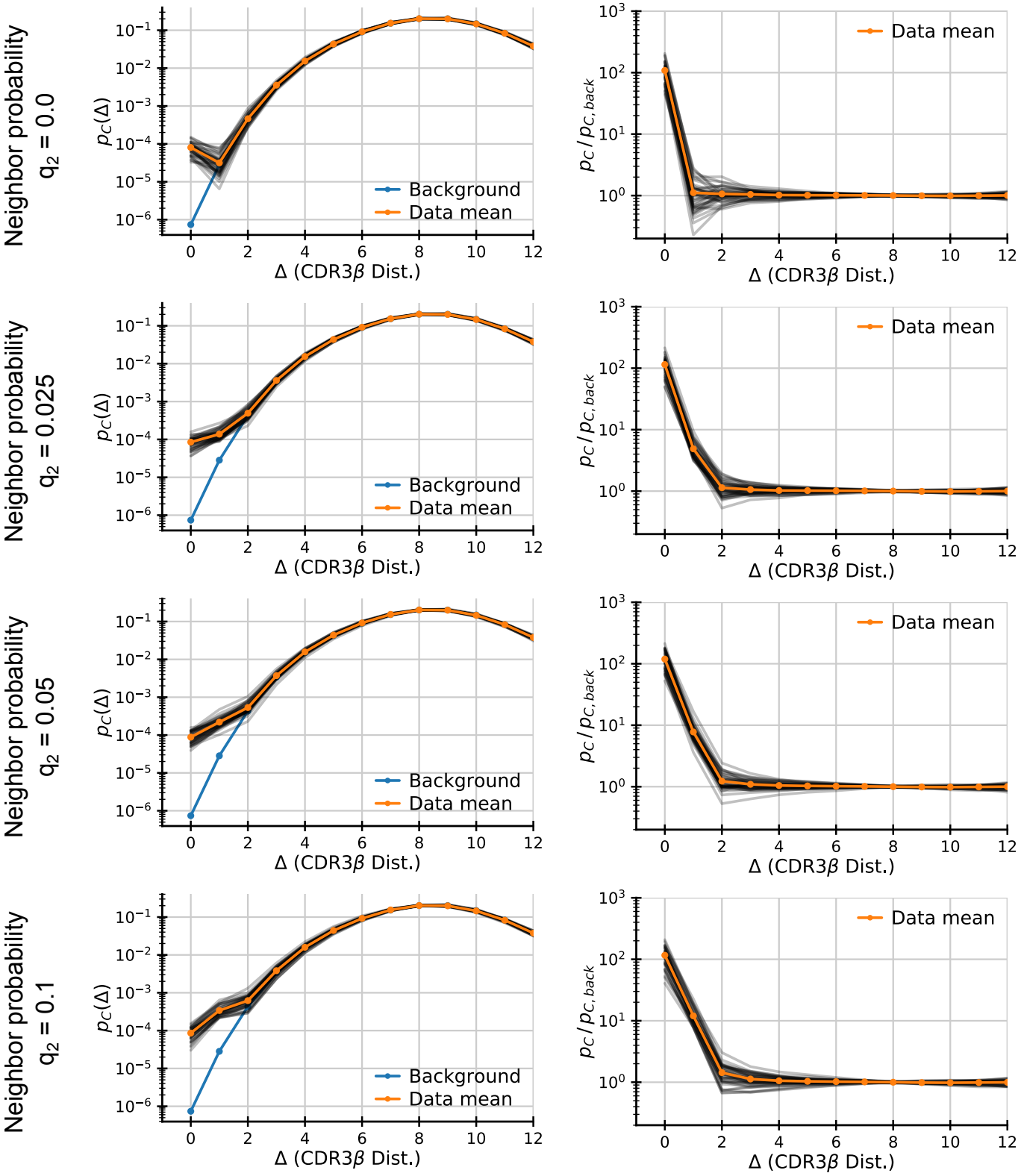

FIG. 9: Coincidence frequency distributions for different realizations of the neighbor-corrected random selection algorithm. The black lines give the results of individual random realizations of selection, while the orange curves give averages over the full set of 100 realizations. The key feature to 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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

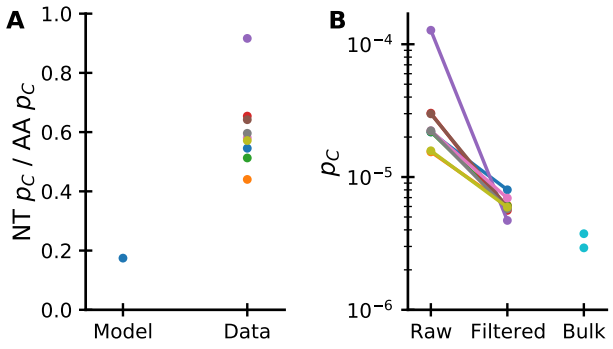

FIG. 10: **Rationale for collapsing α chains paired with multiple β chains in the paired chain data from Tanno et al. [30].** (A) The ratio between nucleotide and amino acid coincidence probabilities for α chains is calculated for each subject and for data generated from SONIA, a model of VDJ recombination 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. [16]).

Appendix E: 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. [30], 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 [30] 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. 10A). As a comparison we calculated the same ratio for sequences generated using a probabilistic model of VDJ recombination [67] 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. 10B). 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 [50]. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.25.501373; this version posted November 15, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Appendix F: Detailed analysis of motif mixture model

In Sec. V we sketched a schematic model of TCR-pMHC binding as a string-matching problem [42, 68]. 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:

$$E_p(\sigma) = \sum_{i=1}^{k} \epsilon_p(i, \sigma_i), \quad \epsilon_p(i, \sigma_i) = \begin{cases} -1 & \text{for } \sigma_i \in \mathcal{S}_i^p \\ 0, & \text{otherwise.} \end{cases} \tag{F1}$$

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

$$\begin{aligned} p_C(\Delta) &= \frac{1}{c^k}\binom{k}{\Delta}(c-1)^\Delta \quad \text{for} \quad \Delta = 0 \dots k \\ p_{C,back}(\Delta) &= \frac{1}{q^k}\binom{k}{\Delta}(q-1)^\Delta \quad \text{for} \quad \Delta = 0 \dots k. \end{aligned} \tag{F2}$$

The near-coincidence enhancement factor in this model is therefore

$$\frac{p_C(\Delta)}{p_{C,back}(\Delta)} = \left(\frac{c-1}{q-1}\right)^\Delta \left(\frac{q}{c}\right)^k. \tag{F3}$$

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. F1 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 B), we obtain an analytical prediction for excess coincidences

$$\frac{p_C(\Delta)}{p_{C,back}(\Delta)} \approx \frac{1}{M}\left(\frac{c-1}{q-1}\right)^\Delta \left(\frac{q}{c}\right)^k + 1 - \frac{1}{M}. \tag{F4}$$

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 [16]. 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.