PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY
IMMUNOLOGY AND INFLAMMATION

Measures of epitope binding degeneracy from T cell receptor repertoires

Andreas Mayer[a,b,c] and Curtis G. Callan Jr.[d,e,1]

Contributed by Curtis G. Callan, Jr.; received August 2, 2022; accepted December 13, 2022; reviewed by Arup K. Chakraborty and Paul G. Thomas

Adaptive immunity is driven by specific binding of hypervariable 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 covaries 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.

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

Significance

Adaptive immunity relies on the binding of molecular targets by a few specific T cells in a highly diverse repertoire. Different T cell receptors can bind the same target, but a quantification of this recognition degeneracy is lacking. We develop a statistical approach that links distributions of sequence similarity among T cells of common specificity to how binding probability covaries 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 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 facilitate the development of repertoire sequencing-based immunodiagnostics.

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 histocompatibility 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 by counting sequence coincidences in epitope-specific TCR repertoires and comparing them with the rate at which they occur 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.

Author contributions: A.M. and C.G.C. designed research; performed research; and wrote the paper.

Reviewers: A.K.C., Massachusetts Institute of Technology; and P.G.T., St Jude Children's Research Hospital Department of Immunology.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: ccallan@princeton.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2213264120/-/DCSupplemental.

Published January 17, 2023.

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 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.

1. 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 are from Dash et al. (14) and were obtained using single-cell receptor sequencing of tetramer-sorted T cells binding the specific pMHC.

Fig. 1*A* shows a clustering by pairwise amino acid sequence 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 loops 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. 1*A* allows some direct conclusions about important features of the 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. 1*B* 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 that 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 nonoverlapping 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. Fig. 1 *C* and *D* show the histograms for α and β chains, respectively. Fig 1*E* 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 7, 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 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.

2. 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

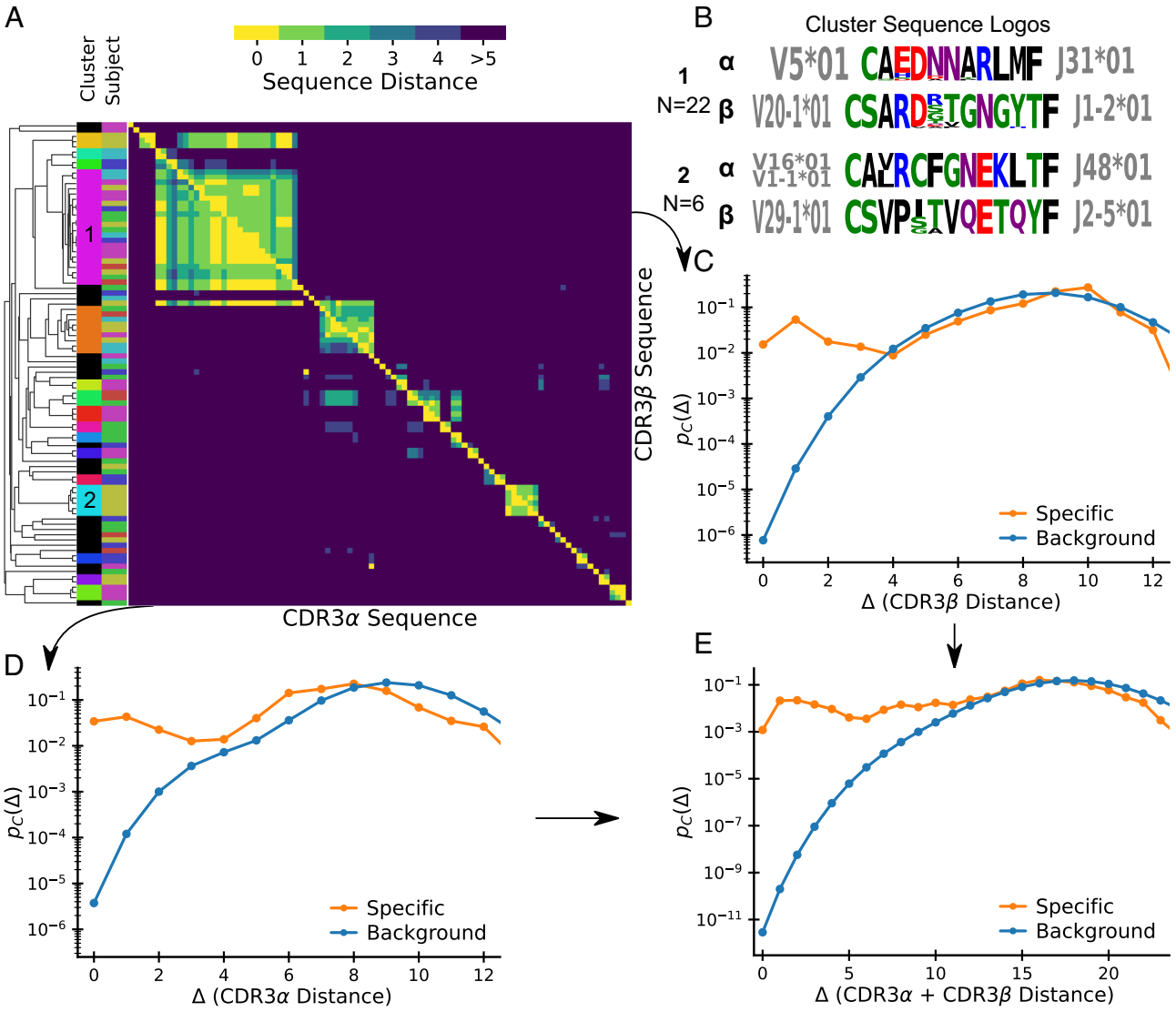

Fig. 1. Patterns of sequence similarity within an epitope-specific repertoire. (*A*) Sequence-similarity clustermap of TCRs binding to an Epstein-Barr Virus epitope as obtained by single-cell TCR sequencing following tetramer sorting (Data: Dash et al. (14), antigen BMLF). *Lower* (*Upper*) triangle 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 the 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.

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 empirical 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. 1 *C–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 Eq. 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 that 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, Eq. **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. 2*A* 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 preselection repertoire is shown in the left panel, where links were arbitrarily chosen such that $p_C = 0.02$. The middle and 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 *SI Appendix*, Appendix 1, the coincidence distributions of the two ensembles are related via the cross-moments of the selection factors,

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

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:

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

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 Eq. **4** into Eq. **3** we find that the coincidence enhancement ratio is proportional to the average of that fraction over the selected sequences $\langle f_\sigma(\Delta)\rangle_{\sigma \in \mathcal{S}} = \langle I_{\mathcal{S}}(\sigma')\rangle_{\sigma \overset{\Delta}{\sim} \sigma', \sigma \in \mathcal{S}}$:

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

Note that $\langle f_\sigma(\Delta = 0)\rangle_{\sigma \in \mathcal{S}} = 1$ 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

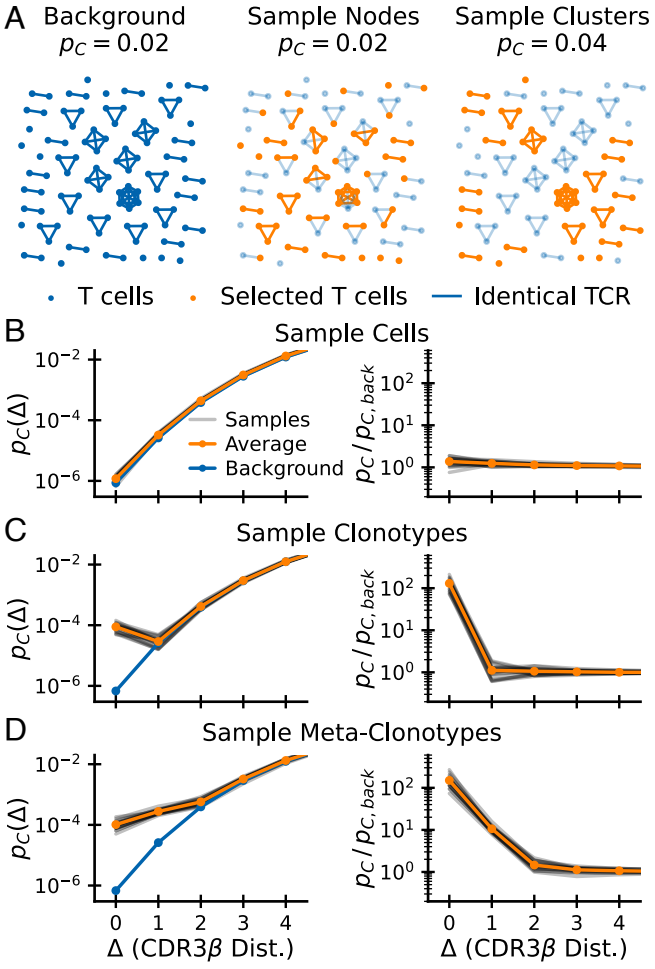

Fig. 2. How selection increases coincidences. (*A*) How different selection procedures change the graph of sequence neighbors. Cells (nodes) in a background graph (*Left*) are connected by edges if they share an identical TCR. Random sampling of nodes (*Middle*) does not change the coincidence probability. 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). (*D*) at random. These random selection protocols act on a background CDR3β repertoire (data from ref. 16). The gray lines show estimates for 20 repetitions of the sampling procedure, and the orange line shows their average.

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:

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

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 CDR3 β sequences taken from whole blood (data from ref. 16) according to different random sampling protocols.

We first compare selecting random cells (Fig. 2*B*) with selecting random clonotypes (Fig. 2*C*), 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 Eq. **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. 2*C*).

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. 1*A*. 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 *SI Appendix*, Appendix 4 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. 2*D*. 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 of ~0.1 (the value of p_{corr}) between $\Delta = 0$ and $\Delta = 1$, in accord with our expectation from Eq. **6**.

We note from these illustrations that the enhancement ratio $p_C(\Delta)/p_{C,back}(\Delta)$ (plotted in the right-hand columns of Fig. 2 *B*–*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.

3. Common Features of Selection Across 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 *Material and Methods*) (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. 3*A*), Minervina et al. (17) (Fig. 3*B*), and Tanno et al. (30) (Fig. 3*C*), 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. 3*H*) and Minervina et al. (16)

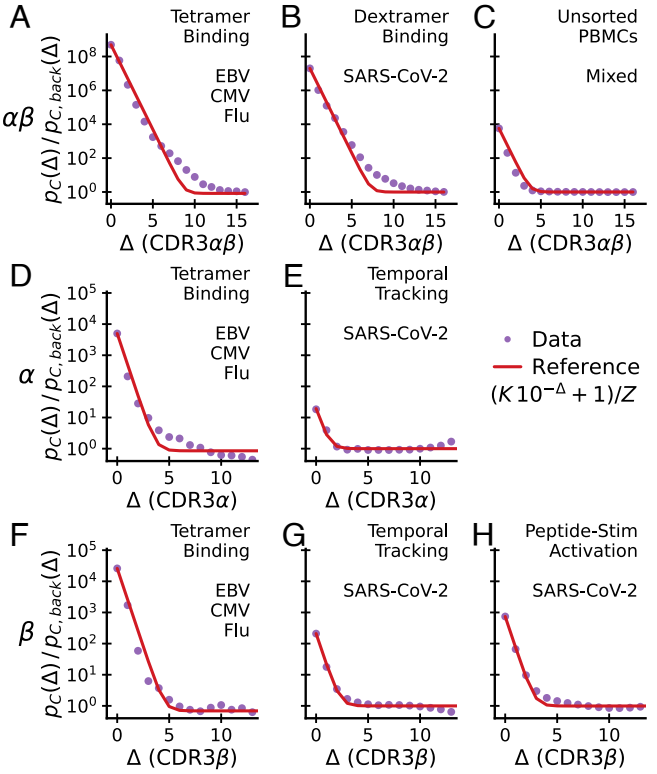

Fig. 3. Excess coincidences follow a common functional form across experiments. Sequence similarity of specific T cells for paired $\alpha\beta$-chain repertoires (*Top*), α-chain repertoires (*Middle*) and β-chain repertoires (*Lower*) compared with background expectations. In each panel, the assay 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*, and *F*–(14); *E* and *G*–(16); *H*–(15); *B*–(17); *C*–(30).

(Fig. 3 *E* and *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. 3 *D* and *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. 3*A*) 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 6.

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 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 Eq. **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. 3 *D* and *E*) and β-chains (Fig. 3 *F*–*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 that it is a fundamental biophysical feature of TCR-pMHC binding interactions.

4. 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. Fig. 3 *D* and *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 *SI Appendix*, Appendix 3). 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.

5. 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 Appendix*, Appendix 6.

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 six 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

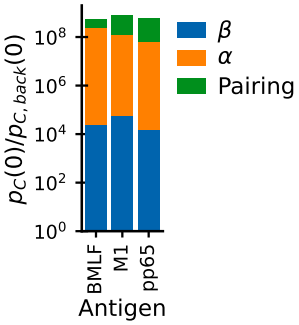

Fig. 4. Epitope binding restricts diversity of both chains individually and also restricts their pairing. Bar chart shows the decomposition of paired chain exact coincidence probability ratios (Fig. 3*A*) for individual epitopes in the dataset from Dash et al. (14) into contributions from selection of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

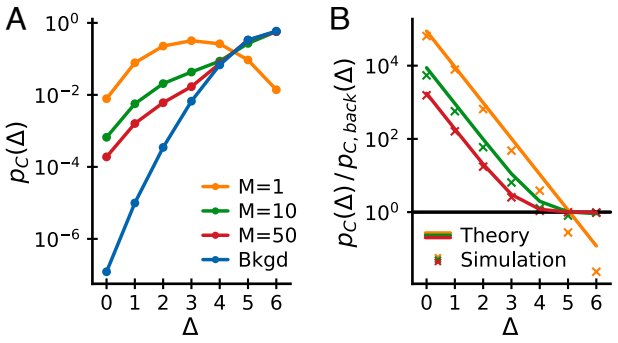

Fig. 5. Coincidences in a mixture of motifs model. (*A*) Coincidence probabilities and (*B*) coincidence probability ratios to background for simulated data generated from a mixture of motifs model with different numbers of motifs M and $c = 3$. (*B*) also shows analytical expectations from Eq. **8** (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.

motif, condition: at each of the k variable positions, we require that the residue lie 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 exercise in this model, with the result:

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

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 *SI Appendix*, Appendix 2), we obtain an analytical prediction for excess coincidences:

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

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 nonuniform amino acid usage (drawn according to the amino acid usage in CDR3α hypervariable chains reported in ref. 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, $\Delta = 6$.

6. 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. 3*C*). 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 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

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

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 *SI Appendix*, Appendix 2:

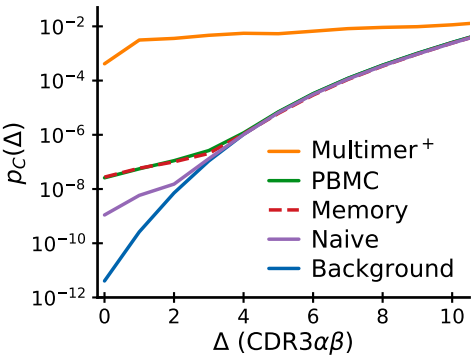

Fig. 6. Comparison of near-coincidence probabilities across paired-chain datasets. The highest values come from TCR repertoires specific to individual epitopes (solid orange curve: average over epitopes studied in Dash et al. (14) and Minervina et al. (17)). Paired-chain sequencing of whole blood (green), 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 a discussion of the large difference in coincidence probabilities between repertoires.

$$p_C[P(\sigma)] = p_C[P(\pi)]\,\langle p_C[P(\sigma|\pi)]\rangle + (1 - p_C[P(\pi)])\,\langle p_C[P(\sigma|\pi_1), P(\sigma|\pi_2)]\rangle, \quad [10]$$

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).

Eq. **10** is a sum of two nonnegative 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 Eq. **2**).

Multimer sorting followed by sequencing gives draws from $P(\sigma|\pi)$ for specific pMHCs π (14, 17), and these data can be used to estimate the average within-epitope-group coincidence probability $\langle p_C[P(\sigma|\pi)]\rangle$. In the absence of better information, we shall assume that the average value $p_C[P(\sigma|\pi)] \sim 10^{-4}$ found in these experiments is the typical order of magnitude for all epitopes. We further assume that the between-epitope-group term in Eq. **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. 3 *E* vs. *D* and *G* vs. *F* and *H*).

7. HLA Overlap Determines Cross-Donor Coincidences

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 toward 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 et al. 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 (*SI Appendix*, Appendix 5). To mimic the filtering applied to intrasample analyses of the data from Tanno et al. (30), we did not count coincidences 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 noncanonical T cell subsets, mucosal associated invariant T cells (MAITs), and invariant natural killer T cells (iNKTs), that recognize nonpeptide 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. 7*A*), and the same trend holds in sorted $CD4^+$ memory (Fig. 7*B*) and $CD4^+$ naive cells (Fig. 7*C*). 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. 7*C* shows no significant deviations from pairing independence (dashed lines in the figure) across naive cells from nontwin 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. 7*C* also shows a clear signal of correlated chain pairing in naive cells both intrasample (black line) and across twin pairs (blue line). This strongly

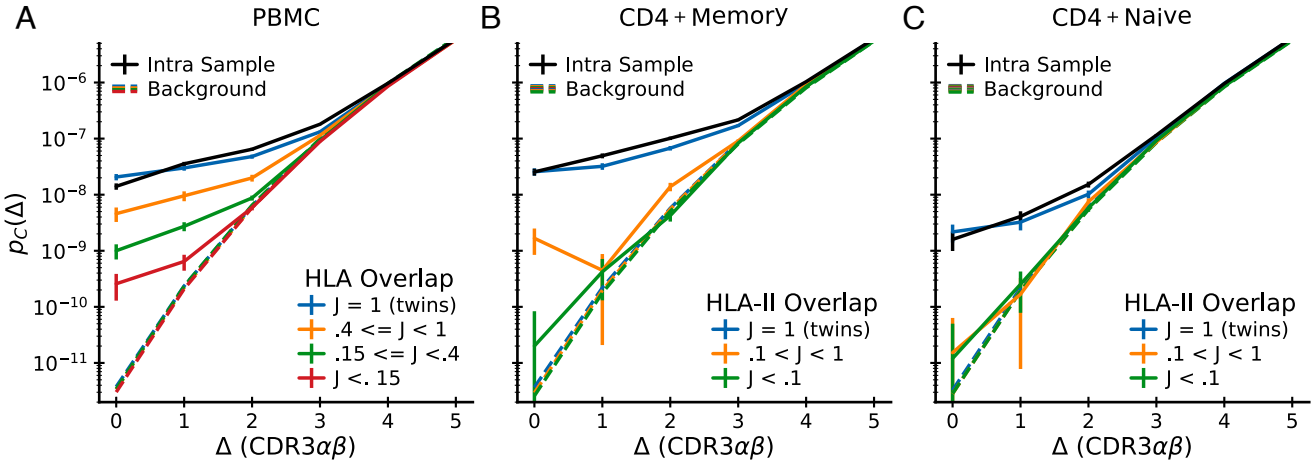

Fig. 7. Intersubject coincidences depend on HLA overlap. Pairwise interdataset coincidence frequency analysis for the 15 paired-seq datasets from Tanno et al. (30) grouped by pairwise HLA overlap. *A*: pairs of unsorted PBMC repertoires; *B*: pairs of $CD4^+$ memory repertoires; *C*: pairs of $CD4^+$ 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 intradataset 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.

suggests that thymic positive and negative selection substantially contribute to the pairing biases. Third, Fig. 7*B* shows that within the memory repertoire coincidences between twins occur at remarkably similar rates to the intrasample 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 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.

8. 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 weighing 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 why structural modeling might be needed to resolve the complex sequence determinants of the different binding solutions (65).

Materials and Methods

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 eight 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 the 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 d and a postinfection time point at 85 d. 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 7,400 and maximum of 69,000). 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 *SI Appendix*, Appendix 5. 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 background coincidence probability distributions, we used unpaired PBMC α and β chain data from ref. 16 (sample F1 from a pre-COVID baseline sample in 2018 from donor "W"). 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.

Data, Materials, and Software Availability. To facilitate adoption of the methodology presented in this paper by the field, we alongside this paper release a Python package for immune repertoire analysis called *Pyrepseq*, available at https://github.com/andim/pyrepseq. This package implements key algorithms for coincidence analysis in a modular, easy-to-reuse manner. Detailed source code reproducing the results reported in this manuscript is available online at https://github.com/andim/paper_coincidences. All the data used in our analyses are publicly available and scripts for downloading it from the experimental data repositories are included in our software repository.

ACKNOWLEDGMENTS. 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. The work of A.M. was supported by a Lewis-Sigler fellowship; the work of C.G.C. was supported in part by NSF grants PHY-1607612 and PHY-1734030. C.G.C. is grateful to the Lustgarten Foundation for support for an extended visit to the Institute for Advanced Study, where some of this work was performed. C.G.C. is also grateful to the Ecole Normale Superieure, Paris, for hospitality during parts of the research reported here.

Author affiliations: [a]Division of Infection and Immunity, University College London, London WC1E 6BT, UK; [b]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton 08544, NJ; [c]Institute for the Physics of Living Systems, University College London, London WC1E 6BT, UK; [d]Department of Physics, Princeton University, Princeton 08544, NJ; and [e]Institute for Advanced Study, Princeton 08540, NJ

1. M. M. Davis, P. J. Bjorkman, The T cell receptor genes and T cell recognition. *Nature* **334**, 395 (1988).
2. H. S. Robins *et al*., Comprehensive assessment of T-cell receptor β-chain diversity in $\alpha\beta$ T cells. *Blood* **114**, 4099–4107 (2009).
3. R. O. Emerson *et al*., Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.* **49**, 659–665 (2017).
4. J. M. Heather, M. Ismail, T. Oakes, B. Chain, High-throughput sequencing of the T-cell receptor repertoire: Pitfalls and opportunities. *Briefings Bioinf.* **19**, 554–565 (2017).
5. P. Bradley, P. G. Thomas, Using T cell receptor repertoires to understand the principles of adaptive immune recognition. *Ann. Rev. Immunol.* **37**, 547–70 (2019).
6. K. C. Garcia *et al*., An αβ T cell receptor structure at 2.5A and its orientation in the TCR-MHC complex. *Science* **274**, 209–219 (1996).
7. M. G. Rudolph, R. L. Stanfield, I. A. Wilson, How TCRs bind MHCs, peptides, and coreceptors. *Ann. Rev. Immunol.* **24**, 419–466 (2006).
8. J. Rossjohn *et al*., T cell antigen receptor recognition of antigen-presenting molecules. *Ann. Rev. Immunol.* **33**, 169–200 (2015).
9. I. Song *et al*., Broad TCR repertoire and diverse structural solutions for recognition of an immunodominant CD8+ T cell epitope. *Nat. Struc. Mol. Biol.* **24**, 395–406 (2017).
10. C. H. Coles *et al*., TCRs with distinct specificity profiles use different binding modes to engage an identical peptide-HLA complex. *J. Immunol.* **204**, 1943–1953 (2020).
11. P. Zareie *et al*., Canonical T cell receptor docking on peptide-MHC is essential for T cell signaling. *Science* **372**, eabe9124 (2021).
12. P. D. Holler *et al*., In vitro evolution of a T cell receptor with high affinity for peptide/MHC. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5387–5392 (2000).
13. Y. Li *et al*., Directed evolution of human T-cell receptors with picomolar affinities by phage display. *Nat. Biotechnol.* **23**, 349–354 (2005).
14. P. Dash *et al*., Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89–93 (2017).
15. S. Nolan *et al*., A large-scale database of T-cell receptor beta (TCR β) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. Research Square Preprint (2020).
16. A. A. Minervina *et al*., Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T-cell memory formation after mild COVID-19 infection. *eLife* **10**, e63502 (2021).
17. A. A. Minervina *et al*., SARS-CoV-2 antigen exposure history shapes phenotypes and specificity of memory CD8+ T cells. *Nat. Immunol.* **23**, 781–790 (2022).
18. P. Bacher *et al*., Low-avidity CD4 + T cell responses to SARS-CoV-2 in unexposed individuals and humans with severe COVID-19. *Immunity* **53**, 1258–1271 (2020).
19. J. Glanville *et al*., Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94–98 (2017).
20. M. Shugay *et al*., VDJdb: A curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res.* **46**, D419–D427 (2017).
21. H. Huang, C. Wang, F. Rubelt, T. J. Scriba, M. M. Davis, Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. *Nat. Biotechnol.* **38**, 1194–1202 (2020).
22. S. Gielis *et al*., Detection of enriched T cell epitope specificity in full T cell receptor sequence repertoires. *Front. Immunol.* **10**, 2820 (2019).
23. D. S. Fischer, Y. Wu, B. Schubert, F. J. Theis, Predicting antigen specificity of single T cells based on TCR CDR 3 regions. *Mol. Syst. Biol.* **16**, e9416 (2020).
24. W. Zhang *et al*., A framework for highly multiplexed dextramer mapping and prediction of T cell receptor sequences to antigen specificity. *Sci. Adv.* **7**, eabf5835 (2021).
25. A. Montemurro *et al*., NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data. *Commun. Biol.* **4**, 1060 (2021).
26. B. Bravi *et al*., Probing T-cell response by sequence-based probabilistic modeling. *PLoS Comput. Biol.* **17**, e1009297 (2021).
27. A. K. Sewell, Why must T cells be cross-reactive? *Nat. Rev. Immunol.* **12**, 669–677 (2012).
28. J. Robinson *et al*., The IPD and IMGT/HLA database: Allele variant databases. *Nucleic Acids Res.* **43**, D423–D431 (2015).
29. T. Mora, A .Walczak, "Quantifying lymphocyte receptor diversity", in *Systems Immunology*, J. Das, C. Jayaprakash, Eds. (CRC Press, 2019).
30. H. Tanno *et al*., Determinants governing T cell receptor α / β -chain pairing in repertoire formation of identical twins. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 532–540 (2020).
31. V. R. Buchholz *et al*., Disparate individual fates compose robust CD8+ T cell immunity. *Science* **340**, 630–635 (2013).
32. C. Gerlach *et al*., Heterogeneous differentiation patterns of individual CD8+ T cells. *Science* **340**, 635–639 (2013).
33. T. Dupic, Q. Marcou, A. M. Walczak, T. Mora, Genesis of the $\alpha\beta$ T-cell receptor. *PLoS Comput. Biol.* **15**, e1006874 (2019).
34. D. S. Shcherbinin, V. A. Belousov, M. Shugay, Comprehensive analysis of structural and sequencing data reveals almost unconstrained chain pairing in TCR$\alpha\beta$ complex. *PLoS Comput. Biol.* **16**, e1007714 (2020).
35. A. Murugan, T. Mora, A. M. Walczak, C. G. Callan, Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 16161–16166 (2012).
36. E. H. Simpson, Measurement of diversity. *Nature* **163**, 688 (1949).
37. T. S. Nunnikhoven, A birthday problem solution for nonuniform birth frequencies. *Am. Stat.* **46**, 270–274 (1992).
38. T. Leinster, C. A. Cobbold, Measuring diversity: The importance of species similarity. *Ecology* **93**, 477–489 (2012).
39. Y. Elhanati, Z. Sethna, C. G. Callan, T. Mora, A. M. Walczak, Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunol. Rev.* **284**, 167–179 (2018).
40. M. Klinger *et al*., Multiplex identification of antigen-specific T cell receptors using a combination of immune assays and immune receptor sequencing. *Plos One* **10**, e0141561 (2015).
41. D. B. Jaffe *et al*., Functional antibodies exhibit light chain coherence. *Nature* **611**, 352–357 (2022).
42. A. Kosmrlj, A. K. Jha, E. S. Huseby, M. Kardar, A. K. Chakraborty, How the thymus designs antigen-specific and self-tolerant T cell receptor sequences. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 16671–6 (2008).
43. J. T. George, D. A. Kessler, H. Levine, Effects of thymic selection on T cell recognition of foreign and tumor antigenic peptides. *Proc. Natl. Acad. Sci. U.S.A.* **114**, e7875–e7881 (2017).
44. R. Lande, Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos* **76**, 5–13 (1996).
45. L. Jost, Entropy and Diversity. *Oikos* **113**, 363–375 (2006).
46. A. Chao *et al*., Rarefaction and extrapolation with Hill numbers: A framework for sampling and estimation in species diversity studies. *Ecol. Monogr.* **84**, 45–67 (2014).
47. V. Venturi, Da. Price, D. C. Douek, M. P. Davenport, The molecular basis for public T-cell responses? *Nat. Rev. Immunol.* **8**, 231–238 (2008).
48. M. V. Pogorelyy *et al*., Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. *PLoS Comput. Biol.* **13**, e1005572 (2017).
49. M. U. Gaimann, M. Nguyen, J. Desponds, A. Mayer, Early life imprints the hierarchy of T cell clone sizes. *eLife* **9**, e61639 (2020).
50. D. I. Godfrey, A. P. Uldrich, J. Mccluskey, J. Rossjohn, D. B. Moody, The burgeoning family of unconventional T cells. *Nat. Immunol.* **16**, 1114–1123 (2015).
51. S. A. Johnson *et al*., Impact of HLA type, age and chronic viral infection on peripheral T-cell receptor sharing between unrelated individuals. *PLoS One* **16**, 1–18 (2021).
52. G. J. Xu *et al*., Comprehensive serological profiling of human populations using a synthetic human virome. *Science* **348**, aaa0698 (2015).
53. N. L. L. Gruta, S. Gras, S. R. Daley, P. G. Thomas, J. Rossjohn, Understanding the drivers of MHC restriction of T cell receptors. *Nat. Rev. Immunol.* **2018**, 1 (2018).
54. J. A. Carter *et al*., Single T cell sequencing demonstrates the functional role of $\alpha\beta$ TCR pairing in cell lineage and antigen specificity. *Front. Immunol.* **10**, 1–13 (2019).
55. K. Mayer-Blackwell *et al*., TCR meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public, HLA-restricted clusters of SARS-CoV-2 TCRs. *eLife* **10**, e68605 (2021).
56. J. Jankauskaite, B. Jiménez-García, J. Dapkunas, J. Fernández-Recio, I. H. Moal, SKEMPI 2.0: An updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics* **35**, 462–469 (2019).
57. R. Arora, R. Arnaout, Repertoire-scale measures of antigen binding. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2203505119 (2022).
58. R. Vazquez-Lombardi *et al*., High-throughput T cell receptor engineering by functional screening identifies candidates with enhanced potency and specificity. *Immunity* **55**, 1953–1966 (2022).
59. A. Cassotta *et al*., Deciphering and predicting CD4+ T cell immunodominance of influenza virus hemagglutinin. *J. Exp. Med.* **217**, e20200206 (2020).
60. A. Mayer, V. Balasubramanian, A. M. Walczak, T. Mora, How a well-adapting immune system remembers. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 8815–8823 (2019).
61. M. Vujović, P. Marcatili, B. Chain, J. Kaplinsky, T. Andresen, T-cell receptor diversity estimates for repertoires (TCRDivER) uses sequence similarity to find signatures of immune response. bioRxiv [Preprint] (2021). 2021.01.11.417444v2.
62. P. A. Mudd *et al*., SARS-CoV-2 mRNA vaccination elicits a robust and persistent T follicular helper cell response in humans. *Cell* **185**, 603–613.e15 (2022).
63. M. V. Pogorelyy *et al*., Resolving SARS-CoV-2 CD4+ T cell specificity via reverse epitope discovery. *Cell Rep. Med.* **3**, 100697 (2022).
64. N. Tickotsky, T. Sagiv, J. Prilusky, E. Shifrut, N. Friedman, McPAS-TCR: A manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* **33**, 2924–2929 (2017).
65. K. K. Jensen *et al*., TCRpMHCmodels: Structural modelling of TCR-pMHC class I complexes. *Sci. Rep.* **9**, 14530 (2019).
66. A. Tareen, J. B. Kinney, Logomaker: Beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2020).