frontiers in Immunology

TYPE Original Research
PUBLISHED 18 September 2026
DOI 10.3389/fimmu.2026.1944643

Check for updates

OPEN ACCESS

EDITED BY
Pieter Meysman,
University of Antwerp, Belgium

REVIEWED BY
Xin Gao,
Chinese Academy of Medical Sciences and Peking Union Medical College, China
Yuepeng Jiang,
City University of Hong Kong, Hong Kong SAR, China

*CORRESPONDENCE
Damon H. May
damon.may@gmail.com

†These authors have contributed equally to this work

RECEIVED 21 July 2026
REVISED 29 August 2026
ACCEPTED 01 September 2026
PUBLISHED 18 September 2026

CITATION
May DH, Woodhouse S, Noakes MT, Doroschak K, Elyanov R, Moore AJ, Osborne EJ, Greenfield B, Hayes T, Taniguchi R, Yang Z, Grino JR, Byron R, Oaks J, Chen-Harris H, Sherwood A, Greissl J, Howie B, Klinger M and Robins HS (2026) Immune signatures of common exposures through co-occurrence of T-cell receptors in tens of thousands of donors. *Front. Immunol.* 17:1944643. doi: 10.3389/fimmu.2026.1944643

COPYRIGHT
© 2026 May, Woodhouse, Noakes, Doroschak, Elyanov, Moore, Osborne, Greenfield, Hayes, Taniguchi, Yang, Grino, Byron, Oaks, Chen-Harris, Sherwood, Greissl, Howie, Klinger and Robins. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Immune signatures of common exposures through co-occurrence of T-cell receptors in tens of thousands of donors

Damon H. May[1*†], Steven Woodhouse[1†], Matthew T. Noakes[1], Kathryn Doroschak[1], Rebecca Elyanow[1], Amanda J. Moore[1], Edward J. Osborne[1], Brad Greenfield[1], Tim Hayes[1], Ruth Taniguchi[1], Zheng Yang[1], John R. Grino[1], Rachel Byron[1], Jamie Oaks[1], Haiyin Chen-Harris[1], Anna Sherwood[1], Julia Greissl[2], Bryan Howie[1], Mark Klinger[1] and Harlan S. Robins[1]

[1]Adaptive Biotechnologies, Seattle, WA, United States, [2]Microsoft Research, Redmond, WA, United States

Background: Memory T cells are records of clonal expansion from prior immune exposures such as infections, vaccines and chronic diseases. Some of the receptors of these expanded T cell clones in a typical immune repertoire are highly public (present in many individuals) because they respond to the same peptide from a prevalent immune exposure, presented by the same Human Leukocyte Antigen (HLA) allele. Only a tiny fraction of public T-cell receptor β sequences (TCRs) have known associations with exposures or specific peptides.
Methods: We mined the TCR repertoires of tens of thousands of donors to define "ECOclusters" (Exposure Co-Occurrence clusters): clusters of public TCRs that tend to occur in the same donors. First, we built models to infer donor HLA type from the TCR repertoire, then associated public TCRs with HLA alleles. Next, we derived co-occurrence clusters of TCRs responding to antigens presented by the same HLA allele, then combined those clusters by co-occurrence across HLA alleles. Each such cross-HLA ECOcluster putatively represents a public TCR signature of a single exposure.
Results: We constructed sensitive, specific models to predict the presence of 220 HLA alleles from TCR repertoires and clustered 8,618,285 HLA allele-associated TCRs to define 11,058 ECOclusters. Using serologically labeled repertoires, we identified ECOclusters associated with HSV-1, HSV-2, EBV, Parvovirus, *Toxoplasma gondii*, Cytomegalovirus (CMV) and SARS-CoV-2, and constructed sensitive, specific classifiers of exposure. We explore the 50,217 TCRs in the CMV ECOcluster in detail, including diagnostic performance of each HLA-associated component and intersection with public TCR databases. ECOclusters represent a step toward deciphering the ledger of immune exposure history encoded by the T-cell repertoire.

KEYWORDS
Cytomegalovirus, machine learning, T cell receptor, T cells, TCR repertoire, TCR-seq

1 Introduction

The enormous diversity of T cells in any individual allows for immune system recognition of many foreign pathogenic exposures. A given individual's T-cell repertoire is a mix of naive and memory T cells, largely shaped by the combination of naive T-cell generation early in life and the exposure history of the individual (1–3).

T cells are activated when T-cell receptors (TCRs) recognize cognate antigens (4) presented by major histocompatibility complexes known as human leukocyte antigen (HLA) molecules in humans. HLA genes are the most polymorphic in the genome. Different HLA alleles (defined here with 4-digit precision, e.g., A*02:01) present complementary sets of antigens, and T-cell receptors are often observed to be specific to a combination of peptide antigen and restricting HLA allele (pHLA) (5–7). T cells in subjects sharing both an HLA allele and a common immune exposure will encounter some of the same pHLAs.

Since this work deals with bulk TCRβ sequencing, we define a TCR as a productive combination of TCRβ V gene, J gene and CDR3 amino acid sequence. High-throughput sequencing of a peripheral blood sample may measure ~10^6 TCRs (8). An individual's immune history is encoded in these repertoire TCRs (9). However, given the generally unknown HLA and peptide associations of TCRs, the high-dimensional nature of repertoires and the genetic diversity of individuals expressed in their inherited HLA types, disentangling the many signals present in a repertoire presents many challenges (10–12).

Subjects who share a prior immune exposure tend to share some TCRs responding to the exposure. It has been previously shown that these "public" TCRs can be associated with immune exposures and used to build diagnostic models for exposures such as Cytomegalovirus (CMV) (13), SARS-CoV-2 (14), Lyme disease (15), herpes simplex virus 1 and 2 (HSV-1/HSV-2) (12) and others. The exposure-associated TCRs bind antigens derived from the exposure, each within the context of an HLA allele restriction (Figure 1A).

Similarly, by comparing TCR repertoires from subjects expressing *vs.* not expressing a particular HLA allele, we and others have associated public TCRs with HLA alleles (16–18). Each such TCR is associated with a single HLA allele due to its response to an HLA-presented antigen from some prevalent exposure.

Here, expanding on ideas first articulated by DeWitt et al. (9), we introduce an approach to identify public TCRs responding to prevalent exposures using the co-occurrence patterns of TCRs observed in tens of thousands of TCR repertoires. We first associate public TCRs with the HLA alleles that present the antigens they bind and use those TCRs to build repertoire-based classifiers for HLA allele presence (Figure 1B). Next, we use those classifiers to infer donor HLA type from 34,793 repertoires and then build a much larger database of HLA allele-associated TCRs. We then construct clusters of co-occurring public TCRs, in two stages (Figure 1C): separately for each HLA allele, we identify HLA-COclusters (HLA Co-Occurrence clusters), clusters of TCRs associated with the same HLA allele that co-occur in repertoires from donors inferred to have the allele; then, we cluster the HLA-COclusters by co-occurrence across HLA allele associations to derive ECOclusters (Exposure Co-Occurrence clusters). Each ECOcluster putatively represents a public TCR response to a single prevalent exposure.

We validate this approach using repertoires with serological labels for seven prevalent exposures. For each exposure, we identify associated ECOclusters that discriminate serological cases from controls in a held-out set of labeled repertoires. We also "deorphanize" several ECOclusters through their high concentration of TCRs known to bind to antigens presented by the same exposure, decoding more of the public T-cell repertoire.

An earlier version of ECOclusters was described in two 2024 preprints (19, 20). Since that time, further published papers and preprints have confirmed the clinical relevance of ECOclusters to Cytomegalovirus (21), Primary Sclerosing Cholangitis (22) and Celiac disease (23). The work described here represents a significant expansion on the earlier ECOclusters preprints. Here, we describe ECOclusters built from repertoires from more-diverse donors, incorporating TCRs associated with many more HLA alleles and associating ECOclusters with more exposures.

2 Materials and methods

2.1 Sample acquisition and TCR repertoire sequencing

A total of 48,440 repertoires were used in this study, of which 42,799 have been described previously and 5,641 were newly acquired to increase racial and HLA diversity (Supplementary Figure 2, Supplementary Table 1). Newly acquired repertoires were from donors with self-reported Asian and other ancestry in the USA (control_bloodworks) and donors from Kuwait (covid_kuwait). We obtained HLA genotyping (Scisco Genetics, ScisGo-HLA-v6 methodology) for 1,810 of the newly sequenced donors (1,328 from control_bloodworks and 482 from covid_kuwait). Together with previously described HLA-typed repertoires (24), this brought our total to 5,954 HLA-typed repertoires (Supplementary Figure 1). We performed immunosequencing on these new samples as previously described (8, 14).

In some cohorts (control_bloodworks, covid_dls, cro_deep_v4, cro_deep_v4b, cro_ultradeep_v4, cro_ultradeep_v4b), donors were screened for known active cancer and FDA-required infectious diseases at the time of blood donation. No explicit exclusion criteria were consistently applied in the collection of the remaining cohorts, including the largest cohort, tdetect_covid.

Between previously sequenced and newly sequenced repertoires, we obtained TCR repertoires for 48,440 donors from 15 cohorts (summary in Supplementary Table 1). The 5th–95th percentile range of the total number of unique rearrangements sequenced per donor was 93,278–562,172, with a median of 305,253. For a subset of these 48,440 donors, we also obtained serological labels (described below).

2.2 Constructing classifiers for HLA allele status

We built classification models for each of 220 HLA alleles (where an "HLA allele" can be an individual allele, a class II DP/

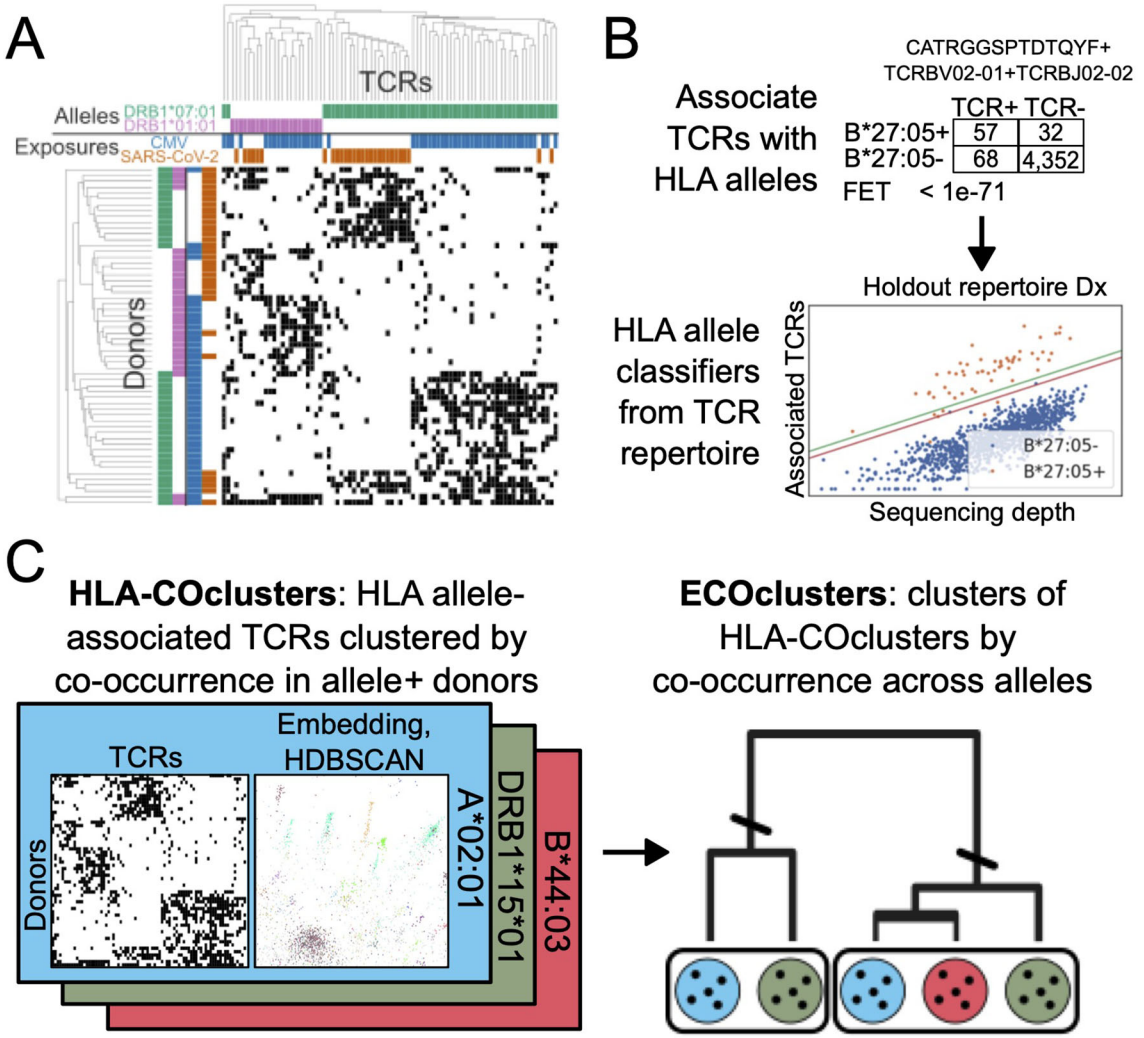


FIGURE 1
(A) Illustration of TCR clustering by repertoire HLA type and exposure history. Heatmap shows presence (black) or absence (white) of 80 TCRs in 58 donor repertoires. TCRs have known HLA allele association (green or pink) and association with exposure to CMV (blue) or SARS-CoV-2 (orange). Donors have known positive (color) or negative (white) HLA alleles and exposures. TCRs and repertoires cluster (here, average-linkage hierarchical clustering) primarily by HLA allele and secondarily by exposure status. **(B)** Building HLA classifiers. We associate TCRs with HLA alleles by occurrence in allele+ and allele– donors and train a predictive model predicting HLA allele presence (orange) or absence (blue) based on allele-associated TCR count and sequencing depth, defining positive (green) and negative (red) call thresholds. **(C)** Constructing ECOclusters. Separately for each HLA allele, cluster HLA allele-associated TCRs by co-occurrence within allele+ donors. Cluster these clusters by donor occurrence correlation, considering only donors having the allele(s) associated with both clusters.

DQ heterodimer, or a p-group—a set of HLA alleles with identical antigen-binding domain protein sequences (25)) with a modified version of the methodology previously described (24) (Supplementary Figure 1C, details in Supplementary Methods).

These 220 models differ from those described in Zahid et al. in including more training data from more diverse donors, modeling more HLA alleles, and introducing separate positive and negative call thresholds (and allowing models to abstain on some repertoires). We also modeled some sets of HLA alleles collectively as p-groups, after observing that models trained individually were essentially indistinguishable in our data or that high-precision models for alleles within a p-group could not be trained individually.

2.3 Associating public TCRs with HLA alleles

To construct a database of ~8.6 million HLA allele-associated public TCRs, we used our HLA classifiers to computationally HLA-type a cohort of 34,793 donors with respect to the 220 modeled HLA alleles. We then identified candidate TCRs overrepresented in predicted positives for each HLA allele with Fisher's Exact Test and estimated a False Discovery Rate (FDR, method described in Supplementary Methods) for each TCR. We retained TCRs with $p < 0.0001$ and FDR <0.001. For TCRs with multiple retained allele associations, we chose the allele with the lowest FDR.

2.4 Clustering TCRs by co-occurrence in repertoires

For each HLA allele *a*, we then constructed a matrix indicating the convergent recombination count of each *a*-associated TCR (the number of unique nucleotide rearrangements encoding the TCR) in each repertoire imputed to have *a*. Clustering on these matrices proved to be intractable due to their extreme sparsity. We therefore first transformed each matrix through a series of embedding and dimensionality-reduction steps.

For each such matrix, we embedded both TCRs and donors into a shared space of 150 dimensions using spectral co-clustering (26). The input bipartite graph is between TCRs and donors, and edges are weighted by convergent recombination count. This embedding reduces the dimensionality and sparsity of the data and relates the problem of clustering TCRs in terms of donors to the problem of clustering donors in terms of TCRs. Next, we applied UMAP (27) to further reduce the dimensionality to 15, which enabled application of the density-based clustering algorithm HDBSCAN (28) with a minimum cluster size of 10 TCRs and/or donors. The minimum number of samples for HDBSCAN was set to the same value as the minimum cluster size, and the Euclidean distance metric was used. A total of 15 UMAP dimensions were chosen by decreasing the dimensionality until HDBSCAN was able to reliably find clusters. If the dimensionality was too large, HDBSCAN reported a very small number of clusters, with most points classified as noise. HLA-COclusters are the TCR members of each derived cluster.

Next, we built cross-HLA ECOclusters by clustering HLA-COclusters across allele associations. Let X denote the matrix of donors by HLA-COclusters, with values as summed convergent recombination counts across the HLA-COcluster TCRs. We defined an HLA-masked correlation matrix C on all pairs of HLA-COclusters:

$$C_{i,j} = corr\ \left(X_{d,i}, X_{d,j}\right)$$

where corr denotes Pearson correlation and **d** denotes the donors imputed to have the HLA allele or alleles associated with both HLA-COclusters *i* and *j*.

We then performed agglomerative hierarchical clustering using $1 - C$ as a pairwise distance between HLA-COclusters and a custom linkage function *f*:

$$S_{I,J} = \{(i,j) \in I \times J\}$$

$$f(I,J) = \frac{1}{\left|S_{I,J}\right|} \sum_{(i,j)\in S_{I,J}} \left(1 - C_{i,j}\right)$$

f preserves the average HLA-masked correlation distances between sets of HLA-COclusters at every branch point with higher fidelity than standard linkage functions such as the UPGMA linkage, which is defined by repeated averaging.

Finally, we cut the resulting hierarchical clustering tree to define a discrete set of clusters, using natural cut points learned from the structure of the tree. We observed that most paths from leaf HLA-COclusters to the root of the tree exhibited sudden spikes in the value of f at which many HLA-COclusters come together. Accordingly, for each leaf HLA-COcluster, we located the branch point maximizing the second derivative of f, f'', along the path from root to leaf. Many HLA-COclusters share a maximum f'' point, converging to the same cluster assignment. However, along some branches we obtained nested clusters. In these cases, we retained the cluster closest to the root.

2.5 Quantifying ECOcluster breadth in repertoires

To quantitatively measure the TCR burden of an ECOcluster *e* within a repertoire *r*, we defined an HLA-invariant estimate of the "breadth" of *e* within *r*, B_{er}. An ECOcluster *e* comprises HLA-COclusters associated with a set of HLA alleles A_e. A repertoire *r* has a set of inferred HLA alleles A_r. Define A_{re} as the intersection of A_r and A_e. Each repertoire can have a different A_{re} with respect to the same ECOcluster *e*; therefore, B_{er} is undefined if A_{re} is the empty set, and it must account for A_{re} in order to be comparable across repertoires.

Conceptually, we build B_{er} on ECOcluster *e* within repertoire *r* as follows: calculate the simple breadth of each HLA-COcluster associated with an HLA allele in A_{re}; normalize the breadth on each HLA-COcluster *h* as its percentile within the breadths on *h* from a large reference distribution of repertoires; calculate B_{er} as the weighted mean of those breadths (using different weights in different contexts, see below). B_{er} is interpretable as an estimate of *r*'s percentile within a notional HLA-independent distribution of breadths on *e*. A formal definition follows.

Define N_r as the total number of unique rearrangements observed in *r*, and C_{hr} as the number of rearrangements in both repertoire *r* and HLA-COcluster *h*. The simple breadth of HLA-COcluster *h* in repertoire *r*, B_{hr}, is calculated as:

$$B_{hr} = \frac{C_{hr}}{N_r}$$

We wish to compare breadth values B_{hr} within the same ECOcluster between donors having different HLA types and therefore having non-missing B_{hr}values for different HLA-COclusters. However, HLA-COclusters differ in repertoire impact, due to, e.g., varying TCR count and cognate peptide immunogenicity, and B_{hr}values for different HLA-COclusters are not directly comparable. We therefore normalize B_{hr} by calculating $\tilde{B}_{hr}$, the percentile of B_{hr} within all the breadth values for *h* from a large database of repertoires (the repertoires used in the construction of ECOclusters) inferred to have the HLA allele *a* associated with *h*. That is:

$$\tilde{B}_{hr} = F_h(B_{hr})$$

where F_h is the empirical cumulative distribution function on $B_{hr}$for repertoires inferred to have *a*. $\tilde{B}_{hr}$, then, has the same interpretation for each HLA-COcluster within the same ECOcluster: a percentile-transformed measure of each donor's response to the exposure represented by the ECOcluster, among all HLA-matched donors.

Finally, some HLA-COclusters may be thought to be more salient or more accurately measured than others (see below), so

10.3389/fimmu.2026.1944643

their $\tilde{B}_{hr}$ values should be weighted more highly. Let w_h denote some nonnegative weight assigned to HLA-COcluster h. We define a measure of the breadth of ECOcluster e in repertoire r, B_{er}, as the weighted mean of the normalized breadths of the HLA-COclusters in e HLA-matched to r. More formally, for repertoire r with inferred HLA alleles A_r, define the set of HLA-COclusters $H_r = \{h{:}h$ is associated with an allele in $A_r\}$, and define H_{er} as $H_e \cap H_r$. Given weights w_h on each HLA-COcluster h, we calculate B_{er} as:

$$B_{er} = \frac{\sum_{h \in H_{er}} w_h \tilde{B}_{hr}}{\sum_{h \in H_{er}} w_h}$$

Intuitively and empirically, larger HLA-COclusters tend to have less dropout (repertoires containing no member TCRs) and therefore represent a more accurate measure of breadth. Therefore, as a general-purpose breadth measure, we define the weight of each HLA-COcluster h as n_h, the count of TCRs in h, to derive B_{er}^{count}:

$$B_{er}^{count} = \frac{\sum_{h \in H_{er}} n_h \tilde{B}_{hr}}{\sum_{h \in H_{er}} n_h}$$

An alternative weighting scheme specific to the context of groups of serologically labeled repertoires is described below.

2.6 Constructing sets of serologically labeled repertoires

For seven prevalent exposures, we assembled collections of repertoires from donors with positive or negative serolabels for each exposure. For CMV and SARS-CoV-2, we used previously acquired serologically labeled samples as described previously (13, 14). For EBV, Parvovirus, HSV-1, HSV-2 and *T. gondii*, we derived new serological labels on previously acquired samples using U-PLEX Development Pack from Meso Scale Discovery (see Supplementary Methods for details).

For each indication, labeled repertoires were divided into training and holdout sets. Where possible, this division separated meaningfully different cohorts. The holdout set samples for EBV and *T. gondii* were acquired with the explicit aim of increasing diversity with respect to HLA type and race/ethnicity (see above). All holdout repertoires from all indications were excluded from the construction of the HLA allele-associated TCR database and ECOclusters.

The samples labeled for HSV-2 were from a single cohort, but with a mix of self-reported Black and White donors. We split the labeled repertoires into a training set of White donors and a holdout set of all Black donors. Separately within the training and holdout sets, cases and controls were matched on age and sequencing depth. These training and holdout sets were provided to support AIRR-ML-25, a completed public machine learning competition (https://www.kaggle.com/competitions/adaptive-immune-profiling-challenge-2025) whose results will be published in Nature Methods.

2.7 Associating ECOclusters with serological labels

For each indication, we calculated repertoire breadth B_{er}^{count} as described above for every ECOcluster on all repertoires in the training set. We tested each ECOcluster for differential B_{er}^{count} in cases using a two-sided Mann–Whitney U test (MWU) on serologically positive *vs.* negative B_{er}^{count} for each ECOcluster, applying Bonferroni correction for multiple testing. For all ECOclusters with Bonferroni-corrected MWU $p < 0.001$, we calculated sensitivity at 98% specificity for B_{er}^{count} as a predictor of case status within the training set. We considered all ECOclusters with higher mean B_{er}^{count} in cases than controls, MWU $p < 0.001$ and >20% sensitivity at 98% specificity to be associated with the indication.

2.8 Constructing classifiers for prior immune exposure

For each indication, for each indication-associated ECOcluster e, for each HLA-COcluster h within e, we calculated the ROC curve relating B_{hr} to the positive serological label in the training set. For our classifier, we retained all HLA-COclusters with area under the receiver-operating characteristic curve (AUROC) >0.6, >20% sensitivity at 98% specificity, and at least five positive and five negative labels in HLA-matched repertoires (the 5-positive requirement was waived for *T. gondii* due to low prevalence in training samples).

Whether they came from a single ECOcluster or multiple, we treated all retained HLA-COclusters for an indication as though they were members of a single ECOcluster e and calculated an alternatively weighted B_{er} breadth measure for this task, as described above, with one difference: the weights for each HLA-COcluster h in this breadth measure were defined by the AUROC for h in the training set:

$$a_h = \left(2(AUROC_h - 0.5)\right)^2$$

To classify new repertoires with respect to the label, we used a weighted breadth B_{er}^{AUROC} calculated using these weights a_h:

$$B_{er}^{AUROC} = \frac{\sum_{h \in H_{er}} a_h \tilde{B}_{hr}}{\sum_{h \in H_{er}} a_h}$$

We chose the B_{er}^{AUROC} threshold for binary classification based on the training ROC curve, as the value maximizing Youden's J statistic (J = true positive rate − false positive rate).

Alternative models (e.g., gradient-boosted trees) were considered that more directly inferred each HLA-COcluster's contribution to the classification task. We chose this model because it outperformed those models in cross-validation and because B_{er}^{AUROC} is interpretable as an HLA-independent estimate of the donor's breadth on exposure-responding TCRs, expressed as a percentile in a large reference distribution.

2.9 Computing enrichment of ECOclusters for TCRs with known association

To associate ECOclusters with exposures for which we lacked labeled repertoires, we evaluated ECOclusters for statistically unlikely representation of TCRs with known association with the same taxon. We separately considered two sources for those associations: public databases and our MIRA assay (29), which identifies TCRs from T cells activated by specific peptides.

We combined three public databases of associations between TCRs and peptide antigens: VDJDB (score >1) (30), IEDB (31), and McPAS (32), downloaded from their respective websites, into a single dataset. In this analysis, because of differences in gene naming between our data and the databases, we matched TCRs by CDR3 amino acid sequence, V gene family, and J gene family. For the MIRA analysis, we used V gene, J gene, and CDR3 amino acid sequence.

Given each of these datasets of TCR-pHLA associations, we separately tested each ECOcluster for enrichment of TCRs associated with pHLAs from each represented taxon using a method analogous to gene set enrichment analysis. For each combination of ECOcluster and taxon, we calculated the number x of unique TCRs shared between the ECOcluster and the taxon-associated TCR list. We then computed the probability $p(x)$ of observing an intersection of x or more TCRs with an ECOcluster by chance, using the hypergeometric distribution:

$$p(x) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}$$

where:

x = count of intersecting TCRs.

m = count of TCRs in ECOcluster.

k = count of TCRs in TCR list.

n = (estimate of total number of public TCRs) − m.

For the total number of public TCRs $n + m$, smaller values are more conservative in the sense of leading to higher p-values for a given intersection. Therefore, we estimated this value conservatively as 8,618,285, the size of our HLA allele-associated TCR database and therefore the smallest possible estimate. We retained associations with $p < 1e{-}10$ and at least 15 intersecting TCRs associated with at least two different peptides.

3 Results

3.1 HLA-associated public TCRs from tens of thousands of repertoires with imputed HLA

Most public TCRs notionally respond to a cognate antigen from a prevalent immune exposure presented by the MHC encoded by an HLA allele or else have extremely high V(D)J generation probability (pGen, here computed by OLGA (33) and TCRsep (37)). As described above, a public TCR's pattern of occurrence across a group of donor repertoires tends to be defined by which donors have its corresponding HLA allele and which donors have encountered its associated exposure.

Accordingly, to discover groups of public TCRs associated with exposures, we first inferred donor HLA type with respect to 220 HLA alleles on 34,793 donors and associated 8,618,285 of their public TCRs with HLA alleles (see Supplementary Results for details). Ninety-two percent of TCRs were associated with class II HLA alleles and 8% with class I alleles. We held out from this process all repertoires later used to evaluate diagnostic model performance.

We note the striking imbalance between the number of TCRs associated with class I alleles and the number associated with class II alleles. This is primarily due to a much larger number of TCRs significantly associated with class II alleles by FET at a fixed p-value threshold, rather than a bias in the method we subsequently use to assign TCRs uniquely to an allele. There are several possible explanations for this, including greater diversity in the CD4 repertoire than the CD8 repertoire (as others have previously noted), or differences in LD structure between class I alleles and class II alleles that make class I-associated TCRs more difficult to discover statistically.

We expect the imputed HLA associations of individual TCRs to have lower accuracy than the HLA model predictions. Multiple allele associations frequently yield low FET p-values; we choose the allele with the lowest FDR as our best estimate. Because of strong linkage disequilibrium (LD) between many pairs of HLA alleles, it is fundamentally difficult to statistically distinguish between multiple possible allele associations for a TCR. However, we expect these errors to have little effect on the correctness of an ECOcluster's exposure association: if we incorrectly associate the TCRs responding to an exposure between two alleles in strong LD, these TCRs will nevertheless tend to occur in the same donors and be combined into the same ECOcluster because donors with one allele by definition very often have the other.

3.2 Clusters of co-occurring TCRs in tens of thousands of repertoires

We next clustered HLA allele-associated TCRs by their co-occurrence in donor repertoires.

First, we built *HLA-COclusters*: TCRs that co-occur in donors within the context of an HLA allele. Separately for each HLA h, we embedded donors inferred positive for h and TCRs associated with h into a shared space using an embedding adopted from the spectral co-clustering algorithm (26) and then reduced this space to 15 dimensions by applying UMAP (27). We note that any method for dimensionality reduction can lose important features of the correlation structure of a dataset, but we found these steps necessary to deal with the extreme sparsity and high dimensionality of the data (the "curse of dimensionality" in machine learning (34)). We used density-based clustering (28) to derive TCR clusters, obtaining 44,010 HLA-COclusters across 196 HLA alleles.

To bridge across HLA contexts, we built *ECOclusters*: clusters of HLA-COclusters that span one or more HLA alleles. To do this, we computed an HLA-masked correlation matrix between all pairs of HLA-COclusters whose entries are the Pearson correlation coefficient between pairs of HLA-COclusters, in only the donors that are positive for the HLA allele(s) associated with the pair. We then applied hierarchical clustering and cut the clustering tree at naturally defined "elbows" to derive 11,058 ECOclusters. We used a custom linkage function that computes the mean HLA-masked correlation distance across all pairs of HLA-COcluster members drawn from the two clusters being merged, preserving HLA-masking throughout the clustering tree. With a standard linkage function such as UPGMA, HLA-masking would be applied only when computing the initial pairwise distances between leaves and

10.3389/fimmu.2026.1944643

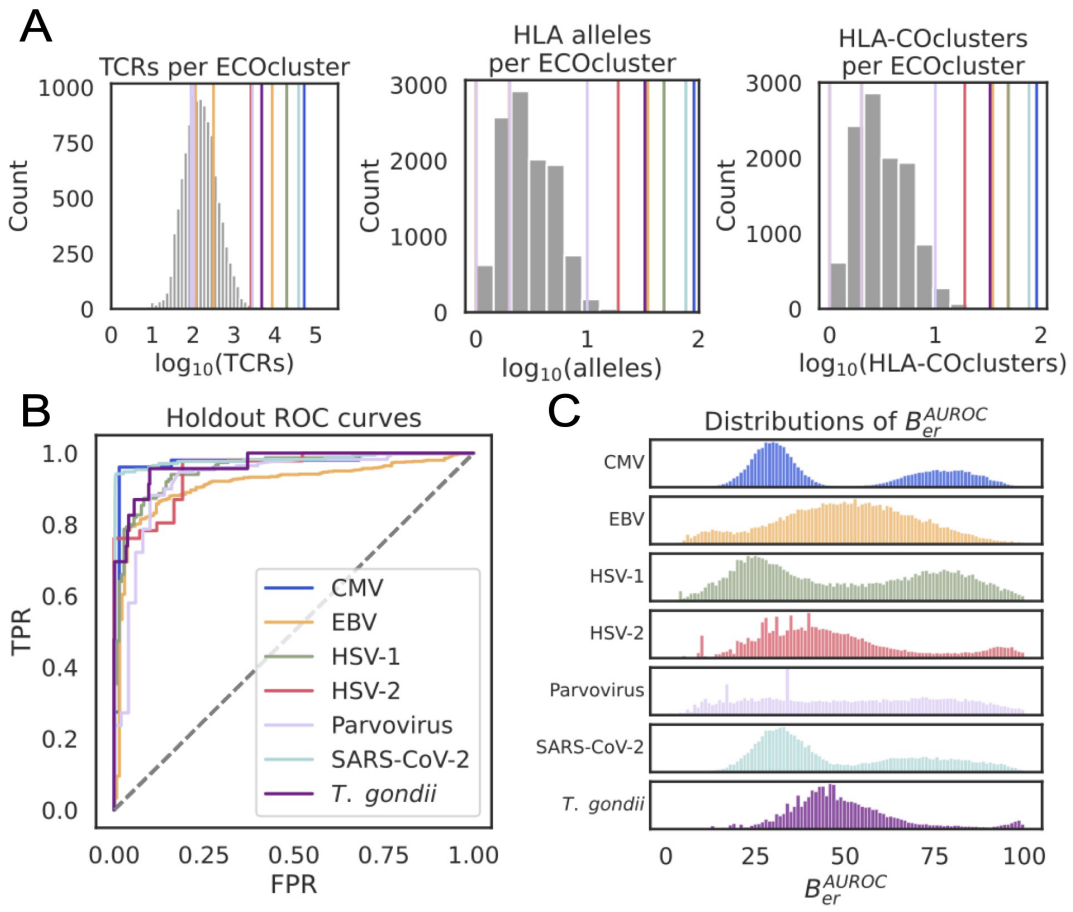


FIGURE 2
(A) Histograms of the number of TCRs, HLA-COclusters, and unique HLA allele associations represented by each ECOcluster. ECOclusters used in diagnostic models indicated with colored lines. **(B)** ROC curves describing the performance of B_{er}^{AUROC} (estimated HLA-independent breadth) on case-associated ECOclusters as a classifier for seven prevalent exposures on held-out labeled repertoires. **(C)** Distributions of B_{er}^{AUROC} for ECOclusters associated with seven prevalent exposures within all donors from the 32,276-repertoire T-DETECT cohort inferred to have at least one HLA allele contributing to B_{er}^{AUROC}. Distributions exhibit bimodality ranging from weak (Parvovirus) to nearly complete (CMV) separation between high (presumed exposed) and low (presumed unexposed) modes.

would not be preserved when calculating distances between subsequently merged clusters. ECOclusters contain between six and 207,512 TCRs, associated with between one and 81 HLA alleles (Figure 2A).

3.3 ECOclusters associated with specific exposures

3.3.1 Diagnostic models of infection from serologically labeled repertoires

For each indication (CMV, EBV, HSV-1, HSV-2, Parvovirus, SARS-CoV-2, and *T. gondii*), we acquired repertoires from serologically labeled samples as described in *Materials and methods*. Depending on labeled sample availability for each indication, we separated repertoires into train and holdout sets by cohort (samples acquired at different times and places), by self-reported race, or, when biologically meaningful separation was unavailable, by sequencing batch. All labeled samples except SARS-CoV-2 were first acquired and subsequently serologically labeled, and so positive label frequency approximates exposure prevalence among the groups sampled.

For EBV and *T. gondii*, holdout samples were acquired specifically to represent greater racial and HLA diversity than the training sets; EBV holdout samples were propensity matched on age, and *T. gondii* on unique productive rearrangements, due to observed imbalances. For HSV-1, HSV-2, and Parvovirus, training and holdout samples were acquired as part of the same cohort and then serologically labeled, minimizing confounding variables that differentiate positive and negative labels. Table 1 gives counts of training and holdout positive- and negative-labeled repertoires per indication (more details in Supplementary Table 1, Supplementary Figures 2, 3). All holdout labeled repertoires for all indications were excluded when constructing our HLA allele-associated TCR database and ECOclusters.

In the training samples for each indication, for each repertoire r, for each ECOcluster e, we calculated B_{er}^{count}, a general-purpose estimate of the HLA-independent contribution of e to r as a percentile of a reference distribution with respect to e (see *Materials and methods*). This metric is useful in quantitatively comparing groups of repertoires with respect to an ECOcluster without knowing which exposure is driving the ECOcluster. As such, it is a good metric for identifying ECOclusters associated with

TABLE 1 Classifier summary.

Indication	Train-holdout separation	+/− Training labels	+/− Holdout labels	ECO-clusters	HLA-COclusters	TCRs	Percent abstention	AUROC	AUPR	Sensitivity at 98% specificity
CMV	Cohort	289/352	51/69	1	79	50,217	0%	0.97	0.96	96%
EBV	cohort, race	1,434/72	750/150	4	28	13,437	7%	0.91	0.98	53%
HSV-1	batch	521/414	167/155	1	25	9,914	8%	0.95	0.96	64%
HSV-2	race	63/315	50/50	1	14	2,133	12%	0.95	0.96	76%
Parvovirus	batch	652/176	199/63	3	12	2,933	16%	0.93	0.97	27%
SARS-CoV-2	cohort	1,130/2,669	172/3,845	1	65	37,704	0%	0.98	0.87	94%
T. gondii	cohort, race	17/1,345	25/750	1	17	3,644	5%	0.97	0.77	69%

For each indication, indicates how training and holdout data were separated; the number of positive and negative labels used in training and holdout; the number of ECOclusters, HLA-COclusters and TCR used; the percent of holdout samples on which model abstained; and the AUROC and sensitivity at 98% specificity of the classifier on holdout labeled repertoires.

positive vs. negative labels of exposure. We identified between one and four ECOclusters specifically associated with the positive label by B_{er}^{count} (four for EBV, three for Parvovirus, and one each for the remaining labels), comprising between 12 and 81 HLA-COclusters and containing between 2,133 and 50,457 TCRs.

For each label, we used the training set to learn positive-label-association weights for each associated HLA-COcluster and calculate B_{er}^{AUROC}. This metric is an exposure-specific, quantitative estimate of the HLA-independent contribution of the ECOcluster(s) associated with an exposure to each repertoire that takes into account HLA-COcluster sensitivity and specificity with respect to the exposure. We assessed B_{er}^{AUROC} with respect to the training labels for each exposure to predict exposure and set a classification threshold, defining a classifier (see Materials and methods for details). B_{er}^{AUROC} is a more complex statistic than simple TCR breadth or HLA-aware breadth (i.e., matching TCRs in repertoires only if the donor is inferred to have the associated HLA allele), or even B_{er}^{count}. B_{er}^{AUROC} learns the diagnostic strength of each HLA-COcluster and then uses all HLA-COclusters that are matched to each repertoire, weighted appropriately. We found this additional complexity instrumental to extracting the highest-specificity diagnostic signal, particularly in indications that were more difficult to classify (Supplementary Figure 4 compares different breadth measures). As described previously (12), HSV-1 and HSV-2 share many common responding TCRs due to high protein homology, but the weights learned in constructing B_{er}^{AUROC} allowed those models to use the most indication-specific HLA-COclusters available to each donor.

Due to the HLA-bound nature of each component of diagnostic signal (HLA-COcluster), these models abstain from prediction on any repertoire not inferred to have any alleles with label-associated HLA-COclusters. Practically speaking, the models are more likely to abstain from prediction on repertoires from donors with few HLA alleles present in the repertoires used to construct ECOclusters.

3.3.2 Performance of diagnostic models

In holdout labeled repertoires, our B_{er}^{AUROC}-based classifiers achieved AUROCs ranging from 0.93 (Parvovirus) to 0.97 (CMV and SARS-CoV-2), areas under the precision-recall curve (AUPRs) ranging from 0.77 to 0.98, and sensitivity at 98% specificity ranging from 27% (Parvovirus) to 96% (CMV). Abstention rates ranged per indication from 0% (CMV and SARS-CoV-2) to 16% (Parvovirus). See Figure 2B, Table 1, and Supplementary Figure 5 for details.

The CMV and SARS-CoV-2 classifiers had the strongest performance. In SARS-CoV-2, samples were acquired before October 2021 and thus were likely all either exposed relatively recently or never exposed/vaccinated, a bygone milieu which necessarily contributes to the very strong performance. CMV dramatically reshapes the T-cell repertoire, and we found 50,217 TCRs in 79 HLA-COclusters specifically associated with CMV. The CMV training and holdout repertoires used here have been used by other groups previously; the model described here achieves higher performance than MotifBoost (36) (AUROC = 0.79), TCRsep (37) (AUROC = 0.90), or an updated version of the Emerson classifier (13) (AUROC = 0.96). Notably, while all four methods give higher average scores to 33 CMV+ self-identified Caucasian donors than to

CMV+ self-identified Asian or Pacific Islander donors, the difference is statistically insignificant for the ECOcluster, MotifBoost, and TCRsep classifiers (all with two-sided MWU p >0.3) but highly significant for the Emerson classifier (p = 0.007). This suggests that the Emerson classifier's diagnostic signal is particularly concentrated in HLA alleles with high prevalence among Caucasians (see Supplementary Figure 6).

Caution is always required when applying diagnostic models for a low-prevalence condition. A classifier that performs well on individual samples may nevertheless call many false positives when applied to samples that are overwhelmingly negative for the condition. Since the labels for all indications except SARS-CoV-2 were determined post-collection, positive label frequency approximates indication prevalence in the population sampled. The AUROC of the *T. gondii* classifier is remarkably strong given the low prevalence of infection in both the population and our training and holdout donors, possibly indicating a large impact of *T. gondii* on the T-cell repertoire. However, due to this very low prevalence, the AUPR (which, unlike AUROC, is sensitive to class prevalence and so reflects the precision achievable at the prevalence captured by evaluation labels) is a relatively low 0.77. Conversely, the high prevalence of EBV leads to a low AUPR (0.59) when predicting EBV seronegativity rather than seropositivity. These statistics provide a better description than AUROC of practical diagnostic performance in low- and high-prevalence regimes.

The Parvovirus classifier may be weaker because Parvovirus is neither a chronic infection, like the herpesviruses (which continually present antigens after initial infection), nor recently acquired in all positive samples, like SARS-CoV-2. But other factors may be involved, e.g., Parvovirus may evoke a weaker T-cell response, or the most immunogenic Parvovirus antigens may be presented by rarer HLA alleles not reflected in our models.

In HSV-2, we separated our training and holdout sets of repertoires by self-reported race, with White donors used for training and Black donors held out for evaluation. This created a meaningful biological difference between the training and holdout sets due to the greater HLA diversity among Black donors. We made this same HSV-2-labeled dataset available for AIRR-ML-25, a public machine learning competition whose results will be published in Nature Methods (35). The strongest competitor's model achieved an AUROC of 0.67 (with no abstention). When comparing the model described here with the competitors' models, we acknowledge two critical caveats: we had prior access to all training and holdout labeled repertoires, and we had much more time to build our models. Since our model abstained on 12 holdout repertoires (12%), we performed 1,000 assignments of scores randomly chosen from other repertoires to the 12 abstained repertoires. The median AUROC among those 1,000 predictions on all 100 samples was 0.90. With the aforementioned caveats firmly in mind, our model's overall performance in a difficult setting demonstrates the power of the ECOclusters approach to exposure classification.

3.3.3 Model score bimodality in tens of thousands of repertoires

To assess the separation between cases and controls in a broader context, we calculated B_{er}^{AUROC} values for each indication on 31,611 repertoires from the previously described T-DETECT cohort (36), which was used in the creation of ECOclusters. For all indications, we observed a bimodal distribution of scores, ranging from weakly (Parvovirus) to starkly (CMV) bimodal (Figure 2C). Given the strong ROC curves for each indication in our labeled samples, we hypothesize that the high and low modes represent exposed and unexposed donors, respectively, with varying separation by indication due to varying discriminatory signal.

For both exposed and unexposed donors, we can ask whether there are variables that influence the position each repertoire has within the distribution of scores with the same label. Among the EBV+ holdout labeled samples, 215 self-reported Caucasian donors have higher average B_{er}^{AUROC} than 263 self-reported Asian donors (two-sided MWU p = 0.0006). Among holdout labeled EBV+ donors, older donors have higher B_{er}^{AUROC} (r = 0.15, Spearman r = 0.0001), and female donors have higher B_{er}^{AUROC} than male donors (MWU p = 0.037). In a linear model predicting B_{er}^{AUROC} among the 478 holdout EBV+ White and Asian donors using all three variables, all three coefficients retained the same directionality as in the individual comparisons, and all were significant (race: p = 0.001, age: p = 0.0003, sex: 0.015). Since most donors used to build the HLA models and ECOclusters were Caucasian, HLA alleles more prevalent among Caucasian donors are expected to be better represented in the ECOclusters. Higher EBV repertoire breadth in older EBV+ donors is to be expected due to increased memory fraction with age, and higher B_{er}^{AUROC} among female donors may represent the established phenomenon of more robust antiviral T-cell responses in women (38).

3.3.4 Diagnostic models using multiple ECOclusters

Two of these classifiers, EBV and Parvovirus, make use of multiple ECOclusters, each of which has HLA-COclusters whose breadth is a sensitive and specific predictor of positive label. To investigate why, we assembled a database of TCR-pHLA pairs derived from the public databases VDJDB (30), IEDB (31), and McPAS (32), and intersected the TCRs with ECOclusters.

Three of the four EBV-associated ECOclusters have HLA-COclusters associated with HLA allele A*02:01 and containing TCRs associated with EBV antigens in public databases. In the first, the majority of database-intersecting TCRs are associated with peptide GLCTLVAML from BMLF1 (an early lytic-phase protein involved in mRNA export and processing). In the second, the majority peptide is YVLDHLIVV from BRLF1 (an immediate-early transcription factor that initiates lytic replication), and in the third, it is FLYALALLL from LMP2 (a latent membrane protein expressed during persistent infection). Hypothetically, these three EBV-associated ECOclusters might represent immune responses to distinct stages of the viral life cycle that occur at different times in different people.

3.3.5 ECOclusters enriched for TCRs responding to peptides from the same taxon

Following methods used in gene set enrichment analysis, we used a hypergeometric test to identify ECOclusters statistically

enriched for TCRs known to respond to peptides from the same taxon (see *Materials and methods*). We used this approach separately with our database of TCR-pHLA pairs derived from public databases and a database derived from experiments using our MIRA assay.

Two ECOclusters used by the EBV classifier were also associated with EBV by intersection with MIRA and a third via intersection with public databases. Two ECOclusters used by the Parvovirus classifier also gained Parvovirus associations from MIRA. One ECOcluster was associated with Influenza using both databases and additionally with SARS-CoV-2 using the public database. Public databases associated two more ECOclusters with *M. tuberculosis*, while MIRA associated additional ECOclusters with Adenovirus D, Influenza, Coxsackievirus, Poliovirus (potentially driven by vaccination), and HHV-6B. Until these observations are confirmed with direct experimental evidence that these ECOclusters represent sensitive, specific biomarkers for exposure, we consider these associations to be putative.

The ECOcluster used by the CMV classifier ("the CMV ECOcluster") also has a large intersection with CMV MIRA TCRs (656) and public database CMV TCRs (239). However, due to the large size of the CMV ECOcluster (50,217 TCRs) and the large amount of MIRA and public CMV data, these large intersections failed to achieve significance via our test.

3.3.6 The CMV ECOcluster in detail

The CMV ECOcluster comprises 50,217 TCRs: 5,323 TCRs from 33 HLA-COclusters associated with class I HLA alleles and 47,124 TCRs from 58 HLA-COclusters associated with class II alleles (see Supplementary Data).

Considering B_{hr} (HLA-COcluster breadth) calculated on *all* labeled repertoires in the CMV holdout set, without regard to donor HLA, AUROCs for the 91 HLA-COclusters ranged from 0.45 to 0.80 (median: 0.55). A total of 45 of 91 HLA-COclusters had at least three positive-labeled and three negative-labeled repertoires inferred to have the HLA allele with which they are associated. AUROCs on those HLA-restricted repertoires ranged from 0.52 to 1.0 (median: 0.92). All but three HLA-COclusters had dramatically higher AUROC within the HLA-restricted repertoires, illustrating the necessity of accounting for HLA type when comparing TCRs between repertoires (Figure 3A, backing data in Supplementary Data).

CMV is well-studied, and our compilation of public TCR databases contains 23,318 TCRs annotated as responding to CMV antigens. For CMV TCRs from the CMV ECOcluster and the public databases, we used the TCRsep CMV model to calculate P_{gen} and P_{post}, estimates of the probability of occurrence in a pre-selection and post-selection repertoire, respectively, and q, a factor that quantifies the degree of functional selection on the TCR. Mean P_{gen}, P_{post}, and q are all significantly higher in the CMV ECOcluster TCRs than in the public database TCRs (two-sided MWU p = 1.3e−45, 1.3e−61, and 1.0e−22, respectively, Figure 3B). These observations have very different interpretations. For P_{gen}, the experimental methods typically used to annotate TCRs in public databases impose no practical constraint on P_{gen}, but the CMV ECOcluster TCRs must be generated easily enough to find their cognate antigens in many repertoires. For P_{post}, the CMV ECOcluster TCRs as a group are demonstrably very often part of a typical CMV response and thus more likely to be observed in a typical CMV+ post-selection repertoire (higher q indicates they are over-represented relative to their P_{gen}), while TCRs in the public databases are only asserted to bind a presented CMV peptide, not to be part of an active response to CMV.

pGen is less of a factor in the clustering of TCRs into HLA-COclusters than might be expected. Altogether, the CMV ECOcluster TCRs have log10(pGen) spanning a range of 3.06 (5th to 95th percentile). On average, an HLA-COcluster in the CMV ECOcluster has TCRs spanning a similarly broad range of 2.80 (Supplementary Figure 7). Some HLA alleles have multiple associated CMV HLA-COclusters, but the differences in mean log10(pGen) between same-allele HLA-COclusters are no greater than between different-allele HLA-COclusters (two-sided MWU p = 0.81). This suggests that pGen, while influencing which TCRs are able to be clustered, does not dominate the clustering behavior.

Though HLA-COclusters are defined on TCR co-occurrence, not TCR sequence, we very often observe clusters of sequence-similar TCRs within the same HLA-COcluster. Using a simple definition of a TCR sequence cluster (a connected component of the graph defined by pairs of TCRs with the same V and J gene whose CDR3 sequences differ by a single amino acid substitution), 311 of the 540 TCRs in the CMV HLA-COcluster associated with A*02:01 are members of 79 sequence clusters ranging in cardinality from two to 26 TCRs (median: 3). Each sequence cluster notionally represents a TCR binding solution to a single pHLA (Figure 3C).

A total of 732 TCRs in the CMV HLA-COcluster are present in IEDB, VDJDB, and/or McPAS. Of those TCRs, 259 are associated with CMV peptides (hypergeometric test p = 1.8e−14, Bonferroni-corrected for multiple species tests). Of the 540 TCRs in the CMV HLA-COcluster associated with A*02:01, 174 are present in the public databases, 60 associated with CMV antigens (Bonferroni-corrected p = 2.4e−49). Of these, 58 are associated with CMV peptide NLVPMVATV and two with VLEETSVML. Forty-six NLVPMVATV TCRs are members of 21 sequence clusters, 16 of which also contain 50 TCRs not in public databases. Known peptide associations can potentially be propagated within sequence clusters, though confident propagation would require an algorithm proven to cluster TCRs that share pHLA binding.

CMV TCRs are highly enriched in CMV+ donors with their associated HLA allele compared to CMV+ donors without the allele. There are 25 class I HLA alleles with associated HLA-COclusters in the CMV ECOcluster, and class I HLA typing is available for donors in the CMV training set. Among the 274 CMV+ donors in the training set, 12 alleles have at least 10 donors with the allele and are therefore suitable for evaluation. For each allele, we assessed raw breadth on the allele-associated HLA-COcluster (fraction of TCRs in the repertoire that are members of the HLA-COcluster) as a predictor of allele positivity in these CMV+ donors. Per-allele AUROC ranged from 0.78 to 0.99 (median 0.9), and sensitivity at 98% specificity ranged from 0.26 to 0.89 (median 0.69). These associations provide evidence of the correctness of the HLA allele associations of these CMV ECOcluster TCRs. They also reinforce the point that allele context is critical when using TCR-based biomarkers (Figure 3D).

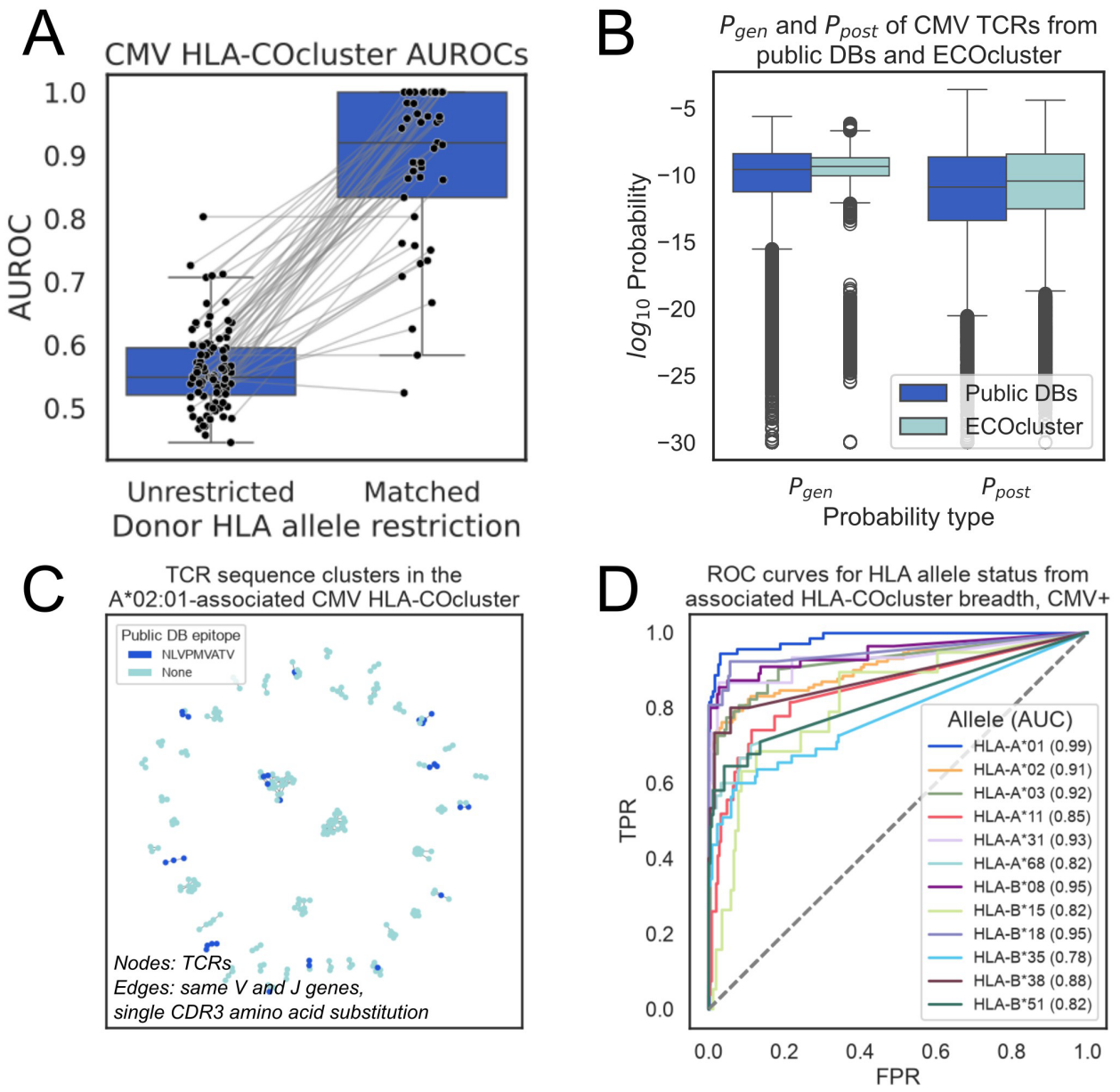


FIGURE 3
The CMV ECOcluster. **(A)** AUROCs describing B_{hr} on HLA-COclusters in the CMV ECOcluster as a classifier of CMV status. All but three HLA-COclusters have dramatically higher AUROC when considering only donors inferred to have the associated HLA allele ("Matched," 45 HLA-COclusters) than when considering all donors ("Unrestricted," 91). **(B)** TCRs in the CMV ECOcluster have higher P_{gen}, P_{post}, and q (q not shown) than CMV-associated TCRs in public databases. **(C)** 311 of the 540 CMV ECOcluster TCRs associated with A*02:01 have at least one sequence neighbor with the same V and J genes and a single CDR3 amino acid substitution (edges). Dark blue dots: TCRs associated with CMV peptide NLVPMVATV in public databases. Edges: sequence neighbors. **(D)** ROC curves describing performance of breadth on allele-associated CMV HLA-COclusters as a predictor of allele presence among CMV+ holdout donors, for 12 class I HLA alleles.

3.4 Typical repertoires contain many ECOcluster TCRs

To investigate the proportion of a typical TCR repertoire annotated by ECOclusters, we assembled 2,000 repertoires from the tdetect_covid, cro_deep_v4b, and cro_ultradeep_v4b cohorts that were held out from the construction of our HLA allele-associated TCR database and ECOclusters. Repertoires had between 53,008 and 645,051 unique productive rearrangements (median: 311,524; 5th–95th percentile: 121,495.9–545,447.9). We inferred HLA type as described for each repertoire and intersected TCRs from every HLA-COcluster with each repertoire inferred to have its associated HLA allele.

These 2,000 repertoires had between 415 and 16,697 such HLA-matched ECOcluster TCRs (median: 3,327.0; 5th–95th percentile: 1,950.8–11,816.1), representing breadths (proportions of repertoire TCRs in any ECOcluster) ranging from 0.001 to 0.041 (median: 0.020; 5th–95th percentile: 0.010–0.032) and depths (sums of TCR frequencies) ranging from 0.003 to 0.250 (median: 0.030; 5th–95th percentile: 0.014–0.074). Thus, ECOcluster TCRs represent a median of 2.0% of the TCRs in these repertoires and 3.0% of the templates (Figure 4A).

10.3389/fimmu.2026.1944643

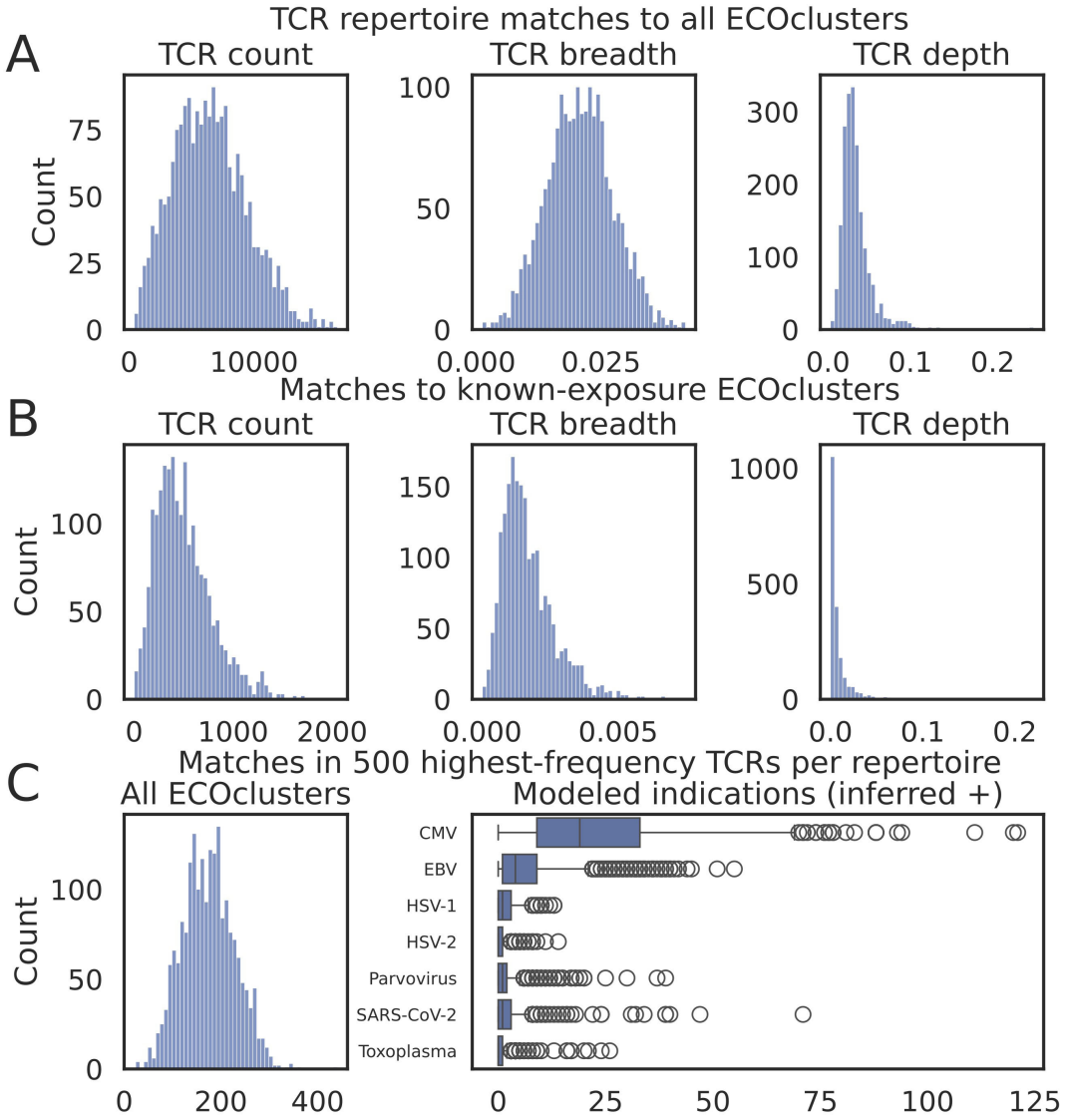


FIGURE 4
ECOcluster TCR representation in 2,000 repertoires not used to construct ECOclusters. **(A)** Distributions of per-repertoire counts, breadths, and depths of HLA-aware matches between repertoire TCRs and ECOclusters. **(B)** As **(A)**, but considering only ECOclusters associated with known exposures. **(C)** Representation of ECOcluster TCRs among the highest-frequency 500 TCRs per repertoire. Left: counts of HLA-aware ECOcluster-matches; right: median exposure-associated ECOcluster highest-frequency TCRs per inferred exposure-positive repertoire.

Considering only ECOclusters with a known exposure association (using all methods described above, Figure 4B), repertoires had between two and 2,053 total HLA-matched TCRs (median: 455.0; 5th–95th percentile: 137.0–1,041.2), representing breadths ranging from 0.000 to 0.007 (median: 0.001; 5th–95th percentile: 0.001–0.004) and depths ranging from 0.000 to 0.220 (median: 0.004; 5th–95th percentile: 0.001–0.037).

ECOcluster TCRs tend to be well represented among the highest-frequency TCRs in a repertoire (Figure 4C). HLA-matched ECOcluster TCRs represented between 24 and 441 of the 500 highest-frequency TCRs in these repertoires (median: 174.0; 5th–95th percentile: 89.0–266.0). The CMV ECOcluster was a major contributor to repertoires inferred to have CMV infection. A median of 19.0 of the highest-frequency TCRs per predicted CMV+ repertoire were members of the CMV-associated ECOcluster, eclipsing the other exposures with diagnostic models (EBV: 4.0; SARS-CoV-2, Parvovirus, HSV-1: 1.0; HSV-2, *T. gondii*: 0.0). While median breadth on the CMV ECOcluster among inferred CMV+ donors was just 0.0009 (5th–95th percentile: 0.0003–0.0022), median depth was 0.006 (5th–95th percentile: 0.0005–0.055), with the highest depth being 0.219: in one repertoire, 21.9% of the TCR templates were CMV ECOcluster members, illustrating the radical expansion of TCR clones responding to CMV.

4 Discussion

ECOclusters (TCR clusters derived from co-occurrence in tens of thousands of human immune repertoires) putatively represent the public T-cell responses to many prevalent immune exposures, including

viral and bacterial infections, vaccines, and medications. ECOclusters represent responses to exposures at least as rare as *T. gondii* (~11% prevalence in the U.S.) and at least as prevalent as EBV (~90%), but the exposures driving most ECOclusters remain unknown.

Our co-clustering approach has advantages over other means of annotating TCR specificity. TCRs (TCR β chains) experimentally determined to bind a pHLA could be part of a TCR α-β pair that binds a different pHLA presented by another prevalent immune exposure. By contrast, TCRs in an exposure-associated ECOcluster are unlikely to bind antigens presented during other prevalent exposures: if they did, they would not cluster by occurrence with other exposure-specific TCRs. TCR annotations derived from ECOclusters can therefore be made with greater confidence than annotations derived only from TCR-pHLA binding experiments.

An earlier version of the CMV-associated ECOcluster than the one described in this work, made publicly available in our earlier preprint, has already proven useful to researchers. Grabauskas et al. first validated that CMV ECOcluster's sensitivity and specificity as biomarkers of the CMV T-cell response. Then, by intersecting the CMV ECOcluster TCRs with single-cell T cell RNA data, they demonstrated that the well-known expansion of $CD8^+$ and $CD4^+$ TEMRA cells with CMV infection is reflected in T cells with CMV-specific TCRs. They also reported the novel finding that CMV-specific clones are enriched for $GZMK^+$ $CD8^+$ and $CD4^+$ Th1 T cells (21).

Though ECOclusters necessarily represent responses to prevalent exposures, multiple researchers have applied them to studies of autoimmune disorders, using an earlier version of ECOclusters encompassing fewer HLA alleles and built from fewer TCR repertoires. El Abd et al. demonstrated that donors with Primary Sclerosing Cholangitis (PSC) have more EBV-responding ECOcluster TCRs and that individual TCRs elevated in PSC are part of the public response to EBV (22). We previously used CMV- and EBV-associated ECOclusters to show that CMV and EBV TCRs were significantly elevated in Celiac+ donors vs. controls, but only among donors who had had the respective viral exposures (23). An earlier version of the CMV ECOcluster was used to establish a strong relationship between CMV and definite Susac's Syndrome (publication forthcoming).

The knowledge that a TCR is associated with a prevalent immune exposure can shed light on that TCR's role in an immune repertoire. We observe many ECOcluster-member TCRs (including CMV) in publicly available tumor single-cell RNA experiments (uTILity, https://github.com/ncborcherding/utility). Those CMV-associated TCRs could be present in the tumor due to their binding a tumor antigen, due to their high frequency in the repertoires of CMV+ donors, or both. While possibly binding a tumor-related antigen, those TCRs are more parsimoniously understood as bystanders. The fraction of tumor-infiltrating lymphocytes (TILs) that are tumor-responding can be an important prognostic indicator (37), and so understanding the composition of TILs can have implications for clinical outcomes.

The T-cell repertoire is a ledger of prior exposures that has remained largely uninterpretable. ECOclusters represent a step toward deciphering that ledger, with applications across many areas of medicine.

Data availability statement

The TCRs in the CMV ECOcluster, with their HLA-COcluster and HLA associations, are provided as a supplementary zip-compressed tab-separated-value file whose columns are described in Supplementary Materials. The code used to construct HLA-COclusters and ECOclusters is available in a public GitHub repository: https://github.com/swoodhouse/ECOclusters.

Ethics statement

This study used human blood samples obtained through two channels. Samples collected directly from participants were approved by the Comitato Etico Provinciale, Comité Ético del Hospital 12 de Octubre (Madrid IC), Providence St. Joseph's Health System IRB, and CEROM. Samples obtained from commercial specimen providers (Discovery Life Sciences, Protocol DLS13); Bloodworks, Research Donor Collection Protocol BT001) were previously collected and de-identified, with donors providing informed consent for research use through the providers, who contractually warranted this to Adaptive Biotechnologies. All studies were conducted in accordance with local legislation and institutional requirements, and all participants provided written informed consent.

Author contributions

DM: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. SW: Data curation, Formal Analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. MN: Conceptualization, Writing – review & editing. KD: Writing – review & editing, Data curation, Formal Analysis, Methodology, Writing – original draft. RE: Methodology, Writing – review & editing. AM: Methodology, Writing – review & editing, Formal Analysis. EO: Writing – review & editing, Investigation. BG: Writing – review & editing, Supervision. TH: Writing – review & editing, Formal Analysis. RT: Formal Analysis, Writing – review & editing, Methodology, Supervision, Writing – original draft. ZY: Formal Analysis, Writing – original draft, Writing – review & editing. JRG: Formal Analysis, Writing – original draft, Writing – review & editing. RB: Formal Analysis, Writing – original draft, Writing – review & editing. JO: Formal Analysis, Writing – original draft, Writing – review & editing, Methodology. HC-H: Writing – original draft, Writing – review & editing, Supervision. AS: Supervision, Writing – review & editing. JG: Supervision, Writing – review & editing, Formal Analysis, Methodology, Writing – original draft. BH: Supervision, Writing – review & editing. MK: Supervision, Writing – review & editing, Methodology. HR: Supervision, Writing – review & editing.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Acknowledgments

We thank Monera Alrukhayes at Kuwait University, who collected samples from which sequencing data was used for this study. We thank H. Jabran Zahid for initial development of models predicting HLA allele status from TCR repertoires and for contributions to the ideas embodied in this work.

Conflict of interest

Authors MTN, KD, RE, AJM, EJO, BG, TH, JRG, BH, MK and HSR have employment with and equity in Adaptive Biotechnologies. JG has employment with and equity in Microsoft.

The remaining author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor PM declared a past co-authorship with the author DHM.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2026.1944643/full#supplementary-material

References

1. Goronzy JJ, Weyand CM. Successful and maladaptive T cell aging. *Immunity*. (2017) 46:364–78. doi: 10.1016/j.immuni.2017.03.010

2. Nikolich-Žugich J. Aging of the T cell compartment in mice and humans: from no naive expectations to foggy memories. *J Immunol*. (2014) 193:2622–9. doi: 10.4049/jimmunol.1401174

3. Qi Q, Liu Y, Cheng Y, Glanville J, Zhang D, Lee JY, et al. Diversity and clonal selection in the human T-cell repertoire. *Proc Natl Acad Sci*. (2014) 111:13139–44. doi: 10.1073/pnas.1409155111

4. Zinkernagel RM, Callahan GN, Klein J, Dennert G. Cytotoxic T cells learn specificity for self H–2 during differentiation in the thymus. *Nature*. (1978) 271:251–3. doi: 10.1038/271251a0

5. Babbitt BP, Allen PM, Matsueda G, Haber E, Unanue ER. Binding of immunogenic peptides to Ia histocompatibility molecules. *Nature*. (1985) 317:359–61. doi: 10.1038/317359a0

6. Brown JH, Jardetzky TS, Gorga JC, Stern LJ, Urban RG, Strominger JL, et al. Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature*. (1993) 364:33–9. doi: 10.1038/364033a0

7. Fremont DH, Hendrickson WA, Marrack P, Kappler J. Structures of an MHC class II molecule with covalently bound single peptides. *Sci (1979)*. (1996) 272:1001–4. doi: 10.1126/science.272.5264.1001

8. Robins H. Immunosequencing: applications of immune repertoire deep sequencing. *Curr Opin Immunol*. (2013) 25:646–52. doi: 10.1016/j.coi.2013.09.017

9. DeWitt WS, Smith A, Schoch G, Hansen JA, Matsen FA, Bradley P. Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *Elife*. (2018) 7. doi: 10.7554/eLife.38358

10. Katayama Y, Yokota R, Akiyama T, Kobayashi TJ. Machine learning approaches to TCR repertoire analysis. *Front Immunol*. (2022) 13:858057. doi: 10.3389/fimmu.2022.858057

11. Liu X, Wu J. History, applications, and challenges of immune repertoire research. *Cell Biol Toxicol*. (2018) 34:441–57. doi: 10.1007/s10565-018-9426-0

12. Pradier MF, Prasad N, Chapfuwa P, Ghalebikesabi S, Ilse M, Woodhouse S, et al. (2023). "AIRIVA: a deep generative model of adaptive immune repertoires", in: *Machine Learning for Healthcare Conference*, 588–611.

13. Emerson R, DeWitt W, Vignali M, Gravley J, Desmarais C, Carlson C, et al. Immunosequencing reveals diagnostic signatures of chronic viral infection in T cell memory. *bioRxiv*. (2015) 49:659–665. doi: 10.1038/ng.3822.

14. Snyder TM, Gittelman RM, Klinger M, May DH, Osborne EJ, Taniguchi R, et al. Magnitude and dynamics of the T-cell response to SARS-CoV-2 infection at both individual and population levels. *Front Immunol*. (2024) 15. doi: 10.3389/fimmu.2024.1488860

15. Dalai SC, Greissl J, Pesesky M, Rebman AW, Soloski MJ, Horn EJ, et al. LB17. Immunosequencing of the T-cell receptor repertoire reveals signatures specific for diagnosis and characterization of early lyme disease. *Open Forum Infect Dis*. (2021) 8: S813–S814. doi: 10.1093/ofid/ofab466.1653

16. Zahid HJ, Taniguchi R, Ebert P, Chow IT, Gooley C, Lv J, et al. Large-scale statistical mapping of T-cell receptor β sequences to Human Leukocyte Antigens. *bioRxiv*. (2024), 2024.04.01.587617. doi: 10.1101/2024.04.01.587617

17. Mayer-Blackwell K, Minervina A, Pogorelyy M, Rawat P, Shapiro MR, Peters LD, et al. TCR2HLA: Calibrated inference of HLA genotypes from TCR repertoires enables identification of immunologically relevant metaclonotypes. *PloS Comput Biol*. (2026) 22:e1013767. doi: 10.1371/journal.pcbi.1013767

18. Ortega MR, Pogorelyy MV, Minervina AA, Thomas PG, Mora T, Walczak AM. Learning predictive signatures of HLA type from T-cell repertoires. *PloS Comput Biol*. (2025) 21. doi: 10.1371/journal.pcbi.1012724

19. May DH, Woodhouse S, Zahid HJ, Elyanow R, Doroschak K, Noakes MT, et al. Identifying immune signatures of common exposures through co-occurrence of T-cell receptors in tens of thousands of donors. *bioRxiv*. (2024), 2024.03.26.583354. doi: 10.1101/2024.03.26.583354

20. May DH, Woodhouse S, Howie B, Robins HS. A catalog of the public T-cell response to cytomegalovirus. *bioRxiv*. (2024), 2024.05.08.593237. doi: 10.1101/2024.05.08.593237

21. Grabauskas T, Trinity L, Verschoor CP. CMV-specific clonal expansion in human Th1 and GZMK$^+$ CD8$^+$ T cells. *Genome Biol*. (2026). doi: 10.1186/s13059-026-04198-5

22. ElAbd H, Pesesky M, Innocenti G, Chung BK, Mahdy AKH, Kriukova V, et al. T and B cell responses against Epstein–Barr virus in primary sclerosing cholangitis. *Nat Med*. (2025) 31:2306–16. doi: 10.1038/s41591-025-03692-w

23. Elyanow R, Choung RS, Marietta EV, Bharanikumar R, Zhou W, Chen-Harris H, et al. Shared TCRs in peripheral blood offer robust celiac disease classification independent of gluten Intake. *bioRxiv*. (2025), 2025.12.26.696039. doi: 10.64898/2025.12.26.696039

24. Zahid HJ, Taniguchi R, Ebert P, Chow IT, Gooley C, Lv J, et al. Large-scale statistical mapping of T-cell receptor β sequences to human leukocyte antigens. *Front Immunol*. (2025) 16. doi: 10.3389/fimmu.2025.1603730

25. Marsh SGE, Osoegawa K, Bodmer WF, Bontrop RE, Carrington MN, Erlich HA, et al. Nomenclature for factors of the HLA system, 2026. In: *Hla*. John Wiley and Sons Inc (2026). doi: 10.1111/tan.70595

26. Dhillon IS. (2001). "Co-clustering documents and words using Bipartite Spectral Graph Partitioning", in: *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. San Francisco, CA, 269–74. doi: 10.1145/502512

27. McInnes L, Healy J, Saul N, Großberger L. UMAP: uniform manifold approximation and projection. *J Open Source Softw*. (2018) 3:861. doi: 10.21105/joss.00861

28. Campello RJGB, Moulavi D, Sander J. (2013). "Density-based clustering based on hierarchical density estimates", in: *Advances in Knowledge Discovery and Data Mining*. Berlin, Heidelberg: Springer, 160–72.

29. Shugay M, Bagaev DV, Zvyagin IV, Vroomans RM, Crawford JC, Dolton G, et al. VDJdb: A curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res*. (2018) 46:D419–D427. doi: 10.1093/nar/gkx760

30. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res*. (2019) 47:D339–D343. doi: 10.1093/nar/gky1006

31. Tickotsky N, Sagiv T, Prilusky J, Shifrut E, Friedman N. McPAS-TCR: A manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics*. (2017) 33:2924–2929. doi: 10.1093/bioinformatics/btx286

32. Sethna Z, Elhanati Y, Callan CG, Walczak AM, Mora T. OLGA: fast computation of generation probabilities of B-and T-cell receptor amino acid sequences and motifs. *Bioinformatics*. (2019) 35:2974–81. doi: 10.1093/bioinformatics/btz035

33. Beyer K, Goldstein J, Ramakrishnan R, Shaft U. (1999). "When is "Nearest neighbor" Meaningful?", in: *Database Theory — ICDT'99*. Berlin, Heidelberg: Springer, 217–35.

34. Kanduri C, Konstantinovsky T, Rawat P, Pavlovic M, May DH, Elyanow R, et al. A community challenge to benchmark machine learning methods for adaptive immune profiling. (2026). doi: 10.6084/m9.figshare.31680751

35. Zahid HJ, May D, Robins H, Greissl J. A fundamental relationship between TCR diversity, repertoire size and systemic clonal expansion: insights from 30,000 TCRβ repertoires. *Front Immunol*. (2026) 16. doi: 10.3389/fimmu.2025.1707727

36. Katayama Y, Kobayashi TJ. Comparative study of repertoire classification methods reveals data efficiency of k-mer feature extraction. *Front Immunol*. (2022) 13:797640. doi: 10.3389/fimmu.2022.797640

37. Jiang Y, Zhang P, Huo M, Li SC. Deep learning-based prediction of the selection factors for quantifying selection in immune receptor repertoires. *Nat Mach Intell*. (2025) 7:1331–45. doi: 10.1038/s42256-025-01085-9

38. Sharma S, Gibbons A, Saphire EO. Sex differences in tissue-specific immunity and immunology. *Science*. (2025) 389:599–603. doi: 10.1126/science.adx4381