Identifying 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], H. Jabran Zahid[2], Rebecca Elyanow[1], Kathryn Doroschak[1], Matthew T. Noakes [1], Ruth Taniguchi[1], Zheng Yang[1], John R. Grino[1], Rachel Byron[1], Jamie Oaks[1], Anna Sherwood[1], Julia Greissl[2], Haiyin Chen-Harris[1], Bryan Howie[1], Harlan S. Robins[1]

1: Adaptive Biotechnologies - 1165 Eastlake Ave E, Seattle, WA 98109

2: Microsoft Research - 14820 NE 36th St, Redmond, WA 98052

ABSTRACT

Memory T cells are records of clonal expansion from prior immune exposures, such as infections, vaccines and chronic diseases like cancer. A subset of the receptors of these expanded T cells in a typical immune repertoire are highly public, i.e., present in many individuals exposed to the same exposure. For the most part, the exposures associated with these public T cells are unknown.

To identify public T-cell receptor signatures of immune exposures, we mined the immunosequencing repertoires of tens of thousands of donors to define clusters of co-occurring T cells. We first built co-occurrence clusters of T cells responding to antigens presented by the same Human Leukocyte Antigen (HLA) and then combined those clusters across HLAs. Each cross-HLA cluster putatively represents the public T-cell signature of a single prevalent exposure.

Using repertoires from donors with known serological status for 7 prevalent exposures (HSV-1, HSV-2, EBV, Parvovirus, *Toxoplasma gondii*, Cytomegalovirus and SARS-CoV-2), we identified a single T-cell cluster strongly associated with each exposure and used it to construct a highly sensitive and specific diagnostic model for the exposure.

These T-cell clusters constitute the public immune responses to prevalent exposures, 7 known and many others unknown. By learning the exposure associations for more T-cell clusters, this approach could be used to derive a ledger of a person's past and present immune exposures.

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 that is largely shaped by the combination of naive T-cell generation early in life and the exposure history of the individual (Goronzy & Weyand, 2017; Nikolich-Žugich, 2014; Qi et al., 2014).

T cells are activated when T-cell receptors (TCRs) recognize cognate antigens (Zinkernagel et al., 1978) presented by major histocompatibility complexes known as human leukocyte antigen (HLA) molecules in humans. TCRs are often observed to be specific to a combination of peptide antigen and restricting HLA (pHLA) (Babbitt et al., 1985; Brown et al., 1993; Fremont et al., 1996). HLA is the most polymorphic gene in the genome, and different HLAs present distinct and often complementary sets of antigens. T cells in subjects sharing HLAs and a common immune exposure will encounter at least some of the same pHLAs.

A large number ($\sim 10^6$) of TCRs in an individual may be measured via high-throughput sequencing of TCRs (Robins, 2013). An individual's immune history is encoded in the TCRs present in their T-cell repertoire (DeWitt et al., 2018) (here and onward, we define a TCR as a T cell's combination of TCRβ V gene, J gene and CDR3 amino acid sequence). However, given the generally unknown pHLA specificity of T cells, the high-dimensional nature of TCRs and the genetic diversity of individuals as encoded by their inherited HLAs, disentangling the many signals present in a repertoire is extremely challenging (Katayama et al., 2022; Liu & Wu, 2018; Pradier et al., 2023).

Subjects with overlapping HLAs and exposure histories will tend to share some TCRs responding to specific exposures. It has been previously shown that TCRs shared between individuals can be used to build diagnostic models of infectious diseases such as Cytomegalovirus (CMV) (Emerson et al., 2015a), SARS-CoV-2 (Snyder et al., 2020a), Lyme disease (Greissl et al., 2021) and herpes simplex virus 1 and 2 (HSV1/2) (Pradier et al., 2023). For each disease, the TCRs thus identified are specific to antigens derived from the exposure but may have various HLA restrictions.

Similarly, by identifying TCRs with higher prevalence in subjects expressing a particular HLA as compared to subjects not expressing that HLA, sets of TCRs may be associated to specific HLAs. Using TCRβ repertoires from 4,144 HLA genotyped subjects, Zahid et al. associate $\sim 10^6$ public TCRs (i.e., TCRs observed in multiple subjects) to hundreds of common HLAs (Zahid et al., in review). They show that these sets of TCRs are enriched for T cells with specific HLA restriction and build models to impute donor expression of

hundreds of HLAs with high sensitivity and specificity. These TCRs are associated with a single HLA but putatively respond to antigens derived from various prevalent exposures. We reason that, given the set of TCRs associated with the same HLA, each prevalent exposure is responsible for a different subset of these TCRs, and that those TCRs are more likely to be present in the repertoires of donors expressing the associated HLA who were exposed to the exposure.

Here, we introduce a method leveraging the co-occurrence patterns of HLA-restricted TCRs observed in a set of 30,674 T-cell repertoires to identify thousands of HLA-COclusters (HLA Co-Occurrence clusters), i.e., subsets of HLA-restricted TCRs that co-occur in subsets of repertoires expressing the HLA. We expect that each prevalent exposure is represented by a group of HLA-COclusters associated with different HLAs. Accordingly, we cluster the identified HLA-COclusters by their representation across all donors to derive ECOclusters (Exposure Co-Occurrence clusters). Each ECOcluster may contain TCRs associated with many different HLAs but is hypothesized to be enriched for TCRs associated with a specific prevalent exposure.

We validate our method using repertoires with serological labels for 7 common exposures with a wide range of prevalence. For each exposure, we identify a single ECOcluster that allows us to discriminate serological cases from controls in a holdout set of repertoires, thereby associating that ECOcluster with the exposure it responds to. By associating more ECOclusters with their exposures, we will decode more of the public T-cell repertoire.

RESULTS

We performed immunosequencing as previously described (Robins, 2013; Snyder et al., 2020a) to derive T-cell repertoires for 30,674 donors from our T-DETECT cohort (see Supplementary Figure 1 for donor demographics). These donors purchased Adaptive Biotechnologies' T-Detect COVID test for prior infection by SARS-CoV-2, and they consented to have their data used for research purposes. We then clustered TCRs (here defined as the combination of TCRβ V gene, J gene and CDR3 amino acid sequence) by their co-occurrence in those repertoires, first within HLA association and then across HLA associations. We determined the exposure association of 7 TCR clusters using repertoires from serologically labeled donors from other cohorts. Finally, we demonstrated the strong diagnostic performance of the exposure-associated clusters on holdout repertoires with serological labels for each exposure.

Public TCR Occurrence in Repertoires Is Determined by HLA Status and Exposures

Figure 1 illustrates the central idea that public TCRs tend to co-occur in individuals who share HLAs and common exposures. We consider a small subset of the TCRs associated with one of two Class II HLAs (DRB1*07:01 or DRB1*05:01) as well as with one of two exposures (CMV or SARS-CoV-2) using methods we will describe below. We visualize the occurrence of these TCRs within the repertoires of T-DETECT donors determined to have one or both HLAs using our HLA imputation models (Emerson et al., 2015a; Zahid et al., in review) and to have one or both exposures using our previously described diagnostic models (Emerson et al., 2015b; Snyder et al., 2020b).

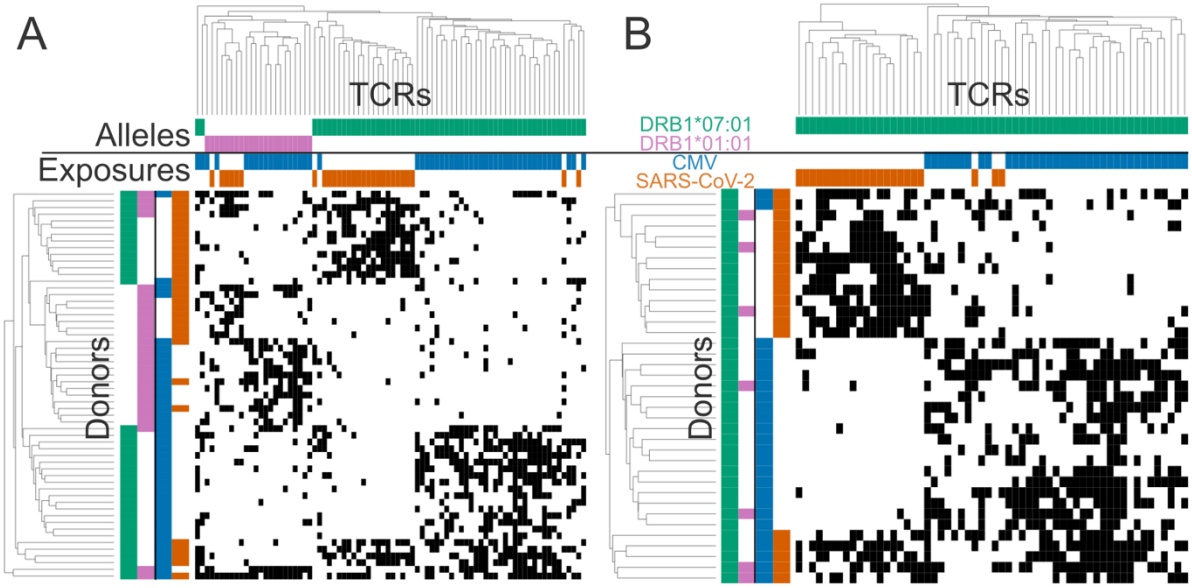

Figure 1: Donor HLA type and prior exposures determine TCR occurrence in repertoires. Heatmaps show presence (black) or absence (white) of 80 TCRs in 58 donor repertoires. TCRs are associated with the Class II HLA DRB1*07:01 (green, 58 TCRs) or with DRB1*01:01 (pink, 22 TCRs), and with exposure to CMV (blue, 52 TCRs) or SARS-CoV-2 (orange, 28 TCRs). Donors are assigned positive (corresponding color) or negative (white) labels for each HLA, and for each exposure, using previously described models. Dendrograms illustrate clustering of TCRs and repertoires by average linkage clustering. A. Considering donors with one or both HLAs and TCRs associated with one HLA or the other, donor repertoires cluster primarily by HLA status and secondarily by exposure status. B. As in Figure 1A, but considering only 37 donors with DRB1*07:01 and 58 DRB1*07:01-associated TCRs; donor repertoires cluster by exposure status.

We constructed a matrix of donors (rows) by TCRs (columns), with 1 representing TCR presence in the donor's repertoire and 0 representing absence. The full matrix is extremely sparse (which constitutes a central difficulty in TCR clustering), and so for this illustration we retained the 80 TCRs and 58 repertoires with maximum occurrence. For illustration, we use average-linkage agglomerative clustering (Sokal, 1958) to cluster the rows and the columns. Figure 1A demonstrates four characteristics of this clustering: donors cluster primarily by HLA type and secondarily by exposure status, and TCRs cluster primarily by HLA association and secondarily by exposure association.

Next, we restricted the matrix to the TCRs associated with DRB1*07:01, and the donors imputed to have that HLA, and performed the clustering again. In this HLA-restricted context, the donors cluster by their status with respect to the two exposures, and the TCRs cluster nearly perfectly by their associated exposures (Figure 1B). This restriction to TCRs and donors associated with / expressing the HLA in question is critical to isolating co-occurrence signatures of exposure.

As we will demonstrate, we can derive exposure-associated clusters of publicly HLA-associated TCRs without the *a priori* knowledge of TCR exposure association and the sparsity reduction used in this example.

Deriving clusters of co-occurring TCRs from Tens of Thousands of T-cell Repertoires

As we observed above, a public TCR's pattern of occurrence across a group of donor repertoires is influenced by its HLA association (i.e., the HLA presenting its cognate antigen in the context of a prevalent exposure) and the exposure it responds to, and by donor HLA expression and exposure history. Accordingly, to discover groups of public TCRs associated with exposures, we first developed the tools needed to associate millions of public TCRs with HLAs (Figure 2).

We used a "pseudolabeling" approach to expand our database of HLA-associated TCRs beyond the 2,904,747 TCRs previously described (Zahid et al., in review). We used our previously-described (Emerson et al., 2015b; Zahid et al., in review) HLA inference models to infer donor status with respect to 131 HLAs for 27,606 donors from our T-DETECT cohort. We then identified TCRs associated with each HLA using the imputed HLA types of the repertoires using the same approach with which we originally associated TCRs with HLAs using genotyped HLA status. This method yielded 3,805,455 TCRs associated with 131 HLAs.

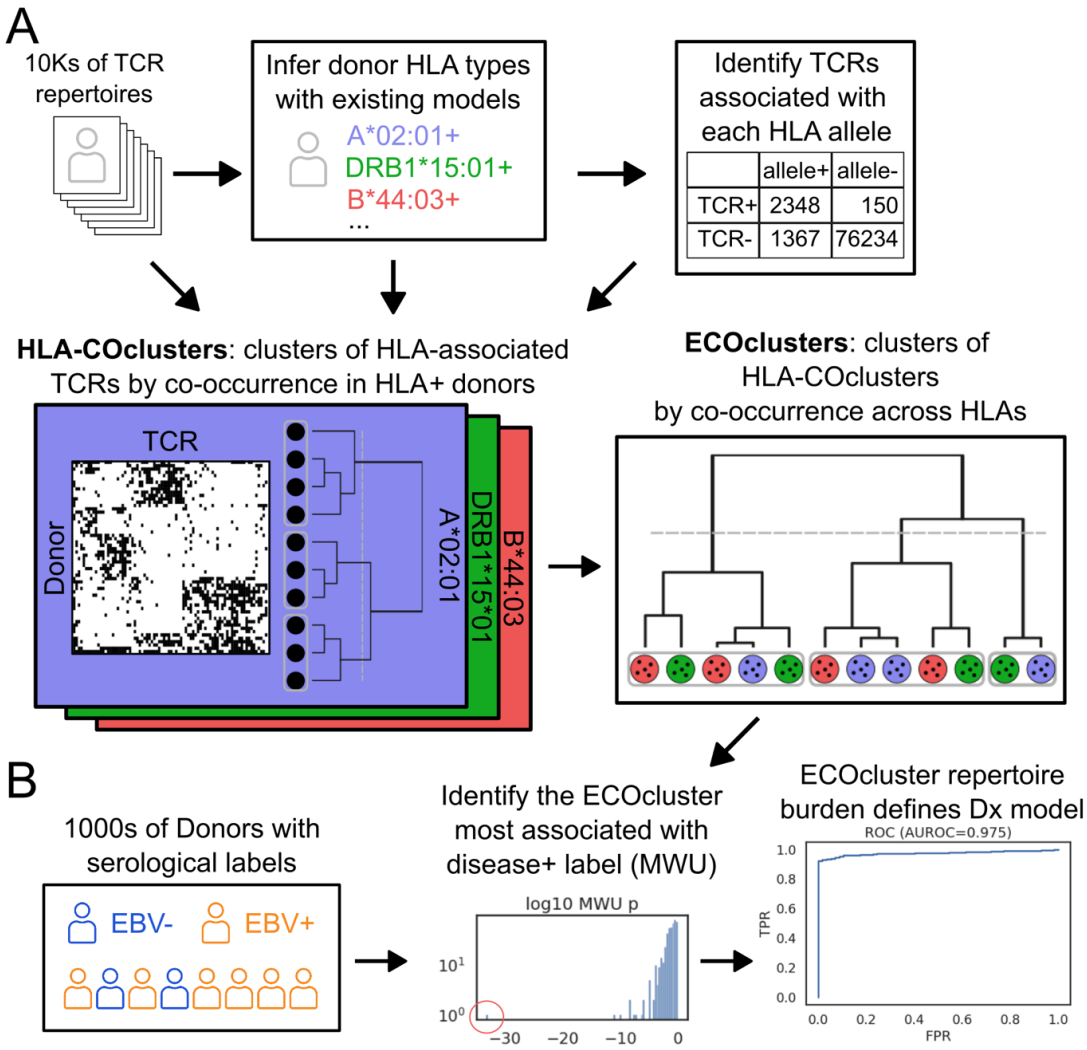

Figure 2: ECOcluster construction and disease modeling. A. Constructing ECOclusters. Starting with 10,000s of TCR repertoires, apply existing HLA inference models to infer all HLA types, then identify TCRs associated with inferred presence of each HLA using Fisher's Exact Test. Separately for each HLA, cluster HLA-associated TCRs by co-occurrence within HLA+ donors. Cluster these clusters using distance defined on donor occurrence correlation, considering only donors having the HLA(s) associated with both clusters. B. Deriving exposure biomarkers from ECOclusters. Serologically label thousands of donors for multiple disease labels. For each labeled donor repertoire, for each ECOcluster, calculate a measure of ECOcluster response (R_{EC}, see Methods). For each exposure, identify a single ECOcluster for which higher R_{EC} is most strongly associated with case label by Mann-Whitney U (MWU) test. Assess performance of disease-associated R_{EC} as a diagnostic classifier for exposure status.

Next, separately for each HLA, we constructed an occurrence matrix of TCRs associated with the HLA by donors inferred to express the HLA, of the kind shown in Figure 1B. We then performed density-based clustering of the TCRs by their co-occurrence in the donors (see Methods for details, clustering visualized in Supplementary Figure 2). This process yielded 43,643 HLA-COclusters (HLA-associated Co-Occurrence clusters) across all 131 HLA associations.

Finally, we clustered the HLA-COclusters by co-occurrence across all donors using HLA-masked Pearson correlation (see Methods), yielding 7,106 ECOclusters (Exposure Co-Occurrence Clusters, summarized in Supplementary Figure 3). Each ECOcluster comprises TCRs associated with one or more HLAs and putatively represents the public TCR signature, or a portion of the signature, of some unknown, prevalent exposure.

1,280 ECOclusters contained only a single HLA-COcluster, and 1,269 contained fewer than 50 TCRs. We suspect that many of these small ECOclusters represent HLA-bound partial exposure responses that failed to cluster across HLA associations due to insufficient donor HLA-sharing within the T-DETECT cohort. On the other end of the spectrum, 693 ECOclusters contained at least 10 HLA-COclusters, and 465 contained 500 or more TCRs. As a percentage of total sequenced TCRs in the repertoire, the TCRs that were members of any ECOcluster ranged from 0.01% to 6.05% (median: 0.91%) across the T-DETECT cohort.

Building Sensitive, Specific Diagnostic Models from Serological Labels

To identify the ECOcluster associated with a given exposure, we can collect many TCR repertoires from donors with known exposure status and identify the ECOcluster with the most significant difference in representation between exposed *vs*. unexposed donors. This approach is analogous to our previously described approach (Emerson et al., 2015b; Snyder et al., 2020b) to statistically associate individual TCRs with exposure for use in a diagnostic model. However, when considering ECOclusters as groups rather than TCRs individually, association with the positive label is greatly strengthened by combining the occurrence of hundreds or thousands of TCRs into a single test.

We restricted our analysis to the 465 “large” ECOclusters comprising 500 or more TCRs (Supplementary Figure 3D). For each ECOcluster, we can ask what proportion of donors are “HLA-matched” to the ECOcluster, i.e., have at least one imputed HLA among the HLAs with which ECOcluster-member TCRs are associated. All ECOclusters HLA-matched to at least 96% of donors were among the 465 “large” ECOclusters.

For each of seven different exposures, we collected repertoires from donors with positive and negative serological labels for the exposure. For Cytomegalovirus (CMV) and SARS-CoV-2, we used labeled repertoires described previously (Emerson et al., 2015b; Snyder et al., 2020b). For each of EBV, HSV-1, HSV-2, Parvovirus and *Toxoplasma gondii*, we derived new in-house serological labels on a shared set of donors (see Methods). We divided the labeled repertoires into training and holdout datasets (demographics in Supplementary Figure 4). Our case-control modeling approach was developed without any use of the holdout repertoires. For each of the in-house serological labels except *T. gondii*, positive and negative label counts roughly aligned with United States prevalence (CDC website, see Table 2). *T. gondii* positive labels represented only ~1% of our confident labels, suggesting our assay may be systematically failing to assign positive labels.

We calculated a "raw breadth" measure of the proportion of a repertoire's unique TCRs belonging to each ECOcluster (and also associated with an HLA the donor is inferred to express), termed B_{EC}. We then adjusted B_{EC} for each donor's inferred HLA type to derive a measure of each repertoire's response to each ECOcluster, termed R_{EC} (see Methods for precise formulations of B_{EC} and R_{EC}). We decided to adjust for donor HLA type after observing that donor expression or non-expression of the various HLAs represented in a given ECOcluster contributes significantly to donor B_{EC} for that ECOcluster.

For each exposure, we tested for higher R_{EC} among exposed than unexposed donors using a one-sided Mann-Whitney U test. For each exposure, a single ECOcluster had a highly significant *p* value (see Table 1) far lower than that of any other ECOcluster (Figure 3). As a percentage of total repertoire TCRs, the TCRs that were members of any of the 7 exposure-associated ECOclusters ranged from 0.001% to 2.83% (median: 0.09%) across the T-DETECT donor repertoires.

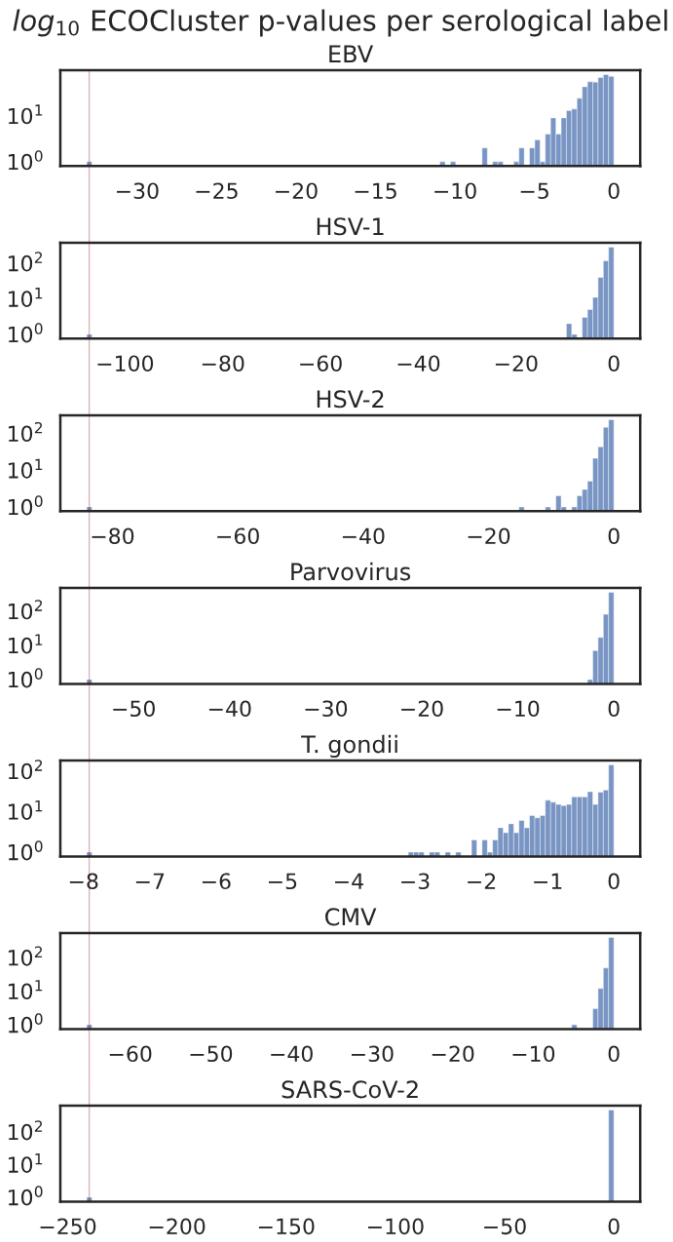

Figure 3: Histograms of $\log_{10}$ p-values, per label, for all large ECOclusters. For each exposure with serological labels, we tested association of each large (>500-TCR) ECOcluster with the positive label using a one-sided Mann-Whitney U test. The figure shows, for each label, the distribution of $\log_{10}$ *p*-values per ECOcluster (vertical axis on $\log_{10}$ scale for visibility). The red line at the left highlights the position of the single lowest p-value for each exposure, which for all exposures is much lower than the rest.

We used R_{EC} on the ECOcluster associated with each exposure as a metric for exposure classification. We evaluated each classifier on the held-out labeled repertoires (Table 1, Figure 4b). Models have AUROC in the range: 0.876-1.0. The CMV model has AUROC 0.96, compared with 0.93 reported in cross-validation using our previously described approach(Emerson et al., 2015a). 5 of the 7 models also have at least 80% sensitivity at 99% specificity. Performance of classifiers using B_{EC} instead of R_{EC} was notably inferior but still strong (Supplementary Figure 5).

Table 1: Serological label counts, ECOcluster statistical significance, and classifier performance.

Exposure	**+/- training labels***	**+/- holdout labels***	**ECOcluster *p*-value****	**Cluster TCRs**	**AUROC*****	**Sensitivity at 99% Specificity**
CMV	289/352	51/66	4.0e-66	26,139	.96 (.93-.98)	92.1%
EBV	1,046/57	365/11	2.2e-18	9,704	.99 (.97-1.0)	97.3%
HSV-1	521/414	167/153	1.0e-72	11,579	.88 (.84-.91)	7.8%
HSV-2	191/623	73/159	3.9e-55	938	.99 (.97-1.0)	86.3%
Parvovirus	652/176	172/37	7.4e-25	4,359	.93 (.87-.98)	18.0%
SARS-CoV-2	1,130/ 2,669	463/4,287	5.0e-242	16,472	.95 (.93-.97)	84.8%
T. gondii	14/1003	3/252	6.5e-7	1,058	1.0 (N/A)	100.0%

* +/- training (holdout) labels: the number of positively and negatively labeled samples in the training (holdout) set

** ECOcluster *p*-value: the one-sided Mann-Whitney U test *p*-value of the exposure-associated ECOcluster

*** AUROC: area under the receiver operating characteristic curve; values in parentheses indicate 95% confidence interval from 1,001 bootstrapping iterations (*T. gondii* had too few holdout samples for bootstrapping)

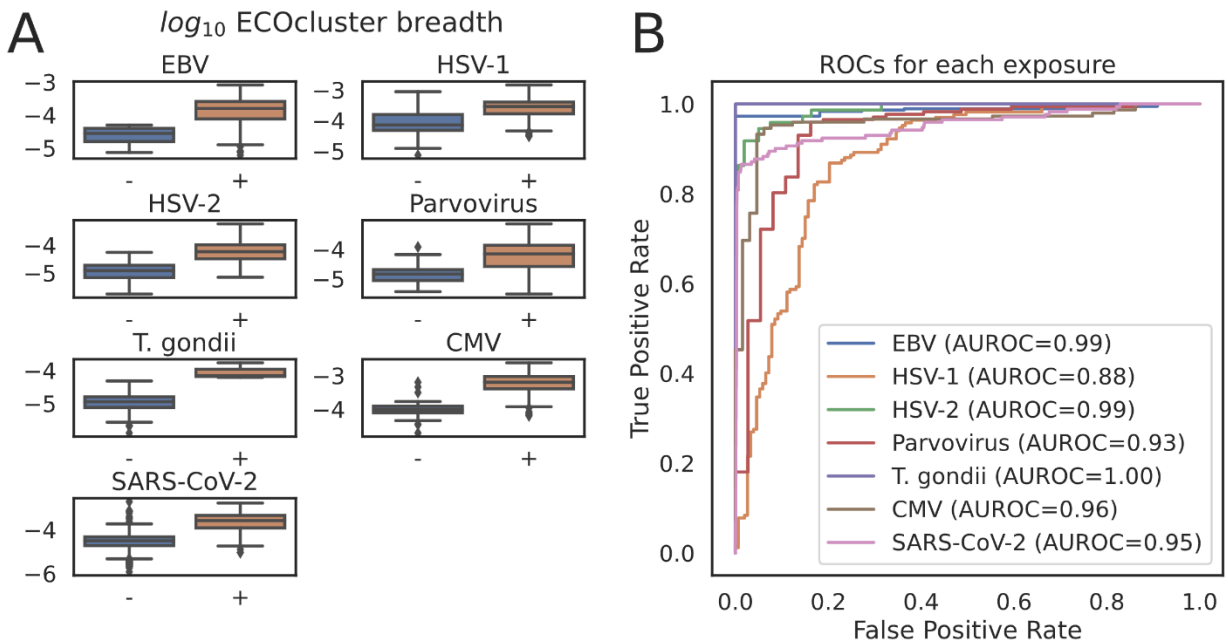

Figure 4: Exposure-associated ECOclusters. A) Box plots of B_{EC} (raw $\log_{10}$ HLA-aware breadth) on the ECOcluster TCRs associated with each exposure, for positive and negative serologically labeled holdout samples. B) Receiver Operating Characteristic curves describing the performance of R_{EC} (donor HLA-adjusted B_{EC}) calculated on the exposure-associated ECOcluster as an exposure classifier in each holdout set.

Our model of HSV-1 exposure performs far more poorly than the rest of the models. HSV-1 model performance segregates clearly by HSV-2 status: the HSV-1 model has AUROC 0.69 on HSV-2 positive-label repertoires and 0.97 on HSV-2 negative-label repertoires (Supplementary Figure 6). This strongly suggests that poor HSV-1 classification is due to the roughly 80% genome homology between the two viruses (Greninger et al., 2018). Our previous work (Pradier et al., 2023) demonstrates a successful approach to disentangling the TCR signals of these two viruses by jointly modeling the two diseases and learning a low-dimensional compositional representation of TCR repertoires. By contrast, our current approach produced an ECOcluster highly specific to HSV-2, but no ECOcluster specific to HSV-1. This is likely because our approach forces each TCR to be a member of at most one ECOcluster: if a TCR responds to an antigen derived from both viruses, its occurrence in repertoires would be dominated by responses to the more-prevalent exposure.

ECOclusters representing other, cryptic, highly homologous pairs of exposures are very likely similarly entangled.

The second weakest model is Parovirus (AUROC 0.93). We propose that the ECOcluster breadth of acute exposures like Parvovirus may diminish with time since exposure. Analysis of model-estimated prevalence by age, below, supports this inference.

While our model of *T. gondii* exposure predicts holdout label status perfectly, our holdout repertoires contain only three positive labels, with the additional aforementioned caveat that our positive label may lack sensitivity. Therefore, while the ROC curve is suggestive of a *T. gondii*-specific response, the performance of the *T. gondii* model cannot be accurately assessed.

The remainder of our serological labels (HSV-2, EBV, CMV and SARS-CoV-2), for which we have developed very strong classifiers, are potentially chronic infections, except SARS-CoV-2. Notably, since our SARS-CoV-2 positively-labeled samples were all acquired before October 2021, the SARS-CoV-2-positive donors were necessarily exposed less than two years prior to sample collection.

Interpreting ECOclusters through intersection with public databases

To investigate the agreement between our results and publicly available TCR data, we intersected ECOcluster TCRs with three public databases of associations between TCRs and peptide antigens (VDJDB (Shugay et al., 2018), IEDB (Vita et al., 2019) and McPAS(Tickotsky et al., 2017)). We looked for ECOclusters that were significantly enriched for TCRs associated with antigens from a single taxon, using an approach inspired by gene set enrichment analysis (see Methods).

This approach lent further support to the association of some of our ECOclusters with exposures via serological labels. 310 of our EBV-associated ECOcluster's 9,704 TCRs were associated with EBV antigens in the public databases, a significant enrichment with hypergeometric test $p < 1e\text{-}15$. The SARS-CoV-2 ECOcluster association was similarly supported (1,005 of 16,429 ECOcluster TCRs found in public databases, $p < 1e\text{-}15$). Even though only 5 of the 938 HSV-2-associated ECOcluster's TCRs were associated with HSV-2 in the public databases, that association was still highly significant ($p < 1e\text{-}13$) because the public databases only contained a total of 59 HSV-2-associated TCRs. While we observed a large overlap of 249 TCRs between our CMV-associated ECOcluster and CMV-associated TCRs in public databases, this overlap was not statistically significant ($p = 0.21$) due to the large number of TCRs in both the ECOCluster (26,106) and the public databases (24,828).

One ECOcluster of previously unknown exposure association was significantly enriched for public database TCRs from influenza antigens: 375 of this ECOcluster's 4,746 TCRs were among the 10,379 public-database TCRs associated with influenza antigens ($p < 1e-15$). We will attempt to validate this association with further experiments.

Table 2: Summary of ECOcluster-associated exposure status across the T-DETECT cohort.

Exposure	Sensitivity	Specificity	Donors Labeled*	Estimated Prevalence**	U.S. Prevalence (CDC)	% Exposed Female
CMV	94%	93%	30,674	42%	>50%	57%
EBV	100%	97%	30,666	95%	90%	53%
HSV-1	82%	83%	30,673	46%	48.1% aged 14-49	56%
HSV-2	95%	95%	30,499	23%	12.1% aged 14-49	59%
Parvovirus	86%	84%	30,430	58%	40%-60%	52%
SARS-CoV-2	89%	90%	30,674	64%	N/A	52%
T. gondii	100%	100%	30,619	6%	11%	48%

* Donors with at least one HLA matching the exposure-associated ECOcluster were labeled as positive or negative using a classification threshold with the indicated sensitivity and specificity

**Estimated prevalences are broadly similar to estimates from the CDC website (SARS-CoV-2 prevalence is not relevant due to the timing of sample collection). Full T-DETECT cohort is 52% female.

These estimates of infection prevalence derived from T-DETECT donor TCR repertoires broadly aligned with expectations current Centers for Disease Control and Prevention (CDC) estimates for adults in the United States (CDC website). HSV-2, CMV and HSV-1 all showed strong bias toward female exposures (59%, 57% and 56%, respectively, compared with 52% female donors in the full T-DETECT cohort) among exposed individuals, also consistent with CDC estimates (HSV-1, HSV-2) and literature (CMV (Fowler et al., 2022)).

To further validate this approach, we examined estimated seroprevalence of each exposure as a function of age (Figure 5). EBV prevalence increases dramatically until roughly age 32 and flattens significantly afterward, consistent with expectations. In contrast, *T. gondii* exposure prevalence is around 1% at age 20 but increases steadily throughout the full range of observed ages. Parvovirus seropositivity appears to *decrease* in prevalence starting around age 40. The sensitivity of our Parvovirus model likely decreases with time since exposure because it is an acute infection that does not continue to stimulate an immune response. Acute exposures like Parvovirus present an opportunity to develop models that retain sensitivity for a longer time after exposure, perhaps by identifying subsets of the TCR response that tend to persist longer than others.

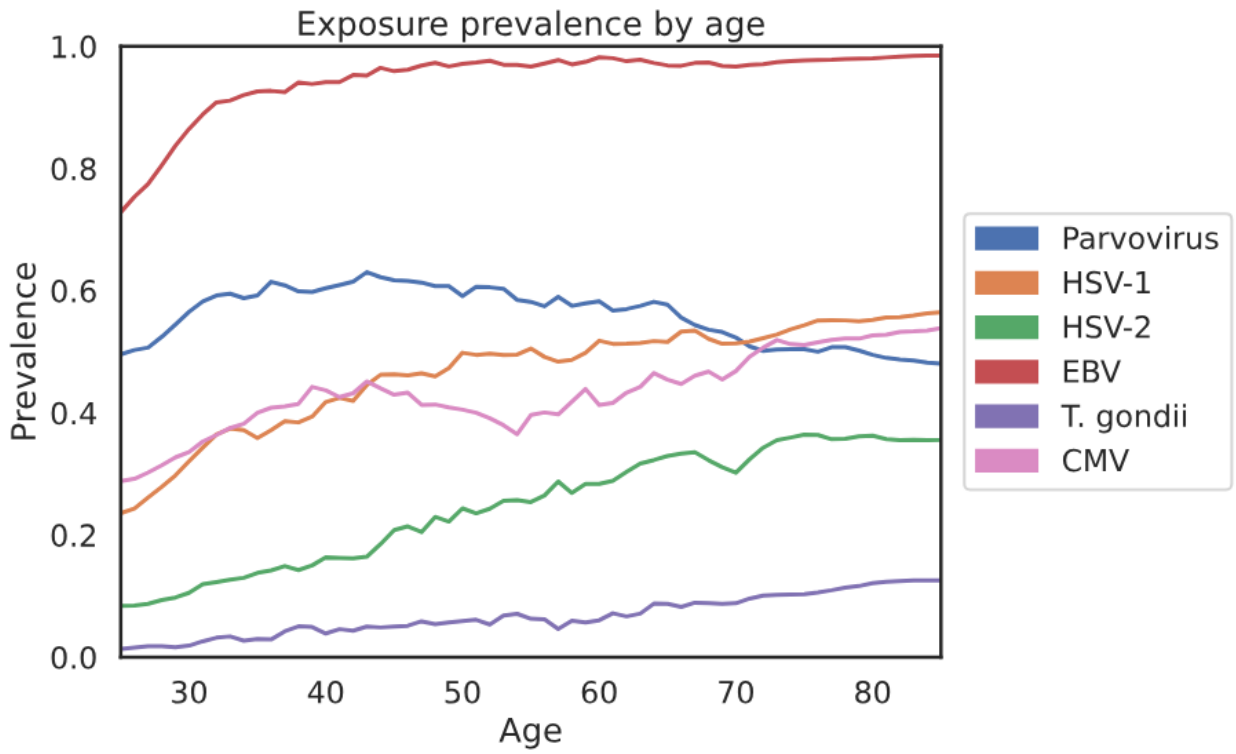

Figure 5: Prevalence of ECOcluster-associated exposures by age. Prevalence (vertical axis) calculated over rolling mean of 1,000 donors, by age (horizontal axis). Each line represents a different exposure, with known exposures indicated by color. SARS-CoV-2 is excluded because it is a novel virus; its prevalence by age doesn't have the same interpretation.

DISCUSSION

The TCR clusters we have derived from tens of thousands of human immune repertoires putatively represent the public T-cell responses to hundreds of prevalent exposures. Most of these exposures remain unknown. Each ECOcluster could potentially represent the immune signature of any kind of prevalent immune exposure, including acute or chronic viral or bacterial infections, vaccines and medications. ECOclusters represent exposures at least as rare as *T. gondii* (~11% prevalence in the U.S.) and at least as prevalent as EBV (~90%).

Unlike serological and PCR-based tests, ECOclusters enable the potential determination of exposure status for many different exposures with a single test. Toward that end, we intend to discover the exposures associated with many more ECOclusters. We will generate data associating TCRs with antigens from prevalent exposures, using our MIRA assay(Klinger et al., 2015), and apply these new data to identify the ECOclusters associated with more exposures.

Our hundreds of large ECOclusters may be clinically relevant, even without knowledge of their associated exposure, in diseases such as autoimmune disorders. Serology and PCR-based detection of common infections like EBV and coxsackie virus have made important links between these infections and the incidence of autoimmune diseases like MS and type I diabetes. In this work we have used ECOclusters to classify exposure status, a binary classification, but our measure of ECOcluster response (R_{EC}) is a quantitative measure. The relationship between prior viral immune responses and autoimmune disorders is complex (Shim et al., 2022). While bystander reactivation of viral responses may have a limited role in autoimmune disease pathogenesis, measures of the degree of such reactivation for many prevalent exposures may provide insight into autoimmune disease severity or treatment efficacy (Guan et al., 2019).

We intend to continue improving ECOclusters. As observed in Table 2, not all ECOclusters are HLA-matched to all donors in the T-DETECT cohort. Adding more repertoires to the clustering could help cluster more disease signatures across more HLA associations, as well as identify new TCR clusters for HLAs that are rare in that cohort. Further, the T-DETECT cohort used to derive ECOclusters represents a limited portion of worldwide HLA diversity. We are actively collecting repertoires with greater HLA diversity to increase applicability of the ECOclusters to all populations.

The ECOclusters are a powerful tool for discoveries about TCR-pHLA binding. Each HLA-COcluster is a group of T cells responding to a constrained set of antigens, presented by a known HLA. By applying some basic assumptions about TCR sequence similarity required

to achieve the same binding solution, we can derive a dataset comprising the TCRβ responses to tens or hundreds of thousands of (largely cryptic) antigens presented by known HLAs. By combining these data with TCRβ-TCRα pairing data such as public single-cell experiments, we can construct a trove of data to illuminate the relationship between TCR sequence and HLA-presented antigen specificity.

Sequencing and analyzing tens of thousands of TCRβ repertoires has led to the discovery of hundreds of public T-cell signatures of immune exposures. As these approaches are applied to more varied data at greater scale, they will help decode a greater portion of the public T-cell repertoire.

METHODS

Identifying HLA-associated TCRs

To construct a large database of TCRs publicly associated with HLAs, we first constructed diagnostic models for each of 131 Human Leukocyte Antigen (HLA) genes, using HLA-typed donor repertoires, as described previously(Emerson et al., 2015b; Zahid et al., in review). Next, we applied those HLA-imputation models to a much larger pool of donor repertoires of unknown HLA (27,606 of our T-DETECT donors) to infer those donors' HLA types. Finally, we used those imputed HLA types to identify 3,805,455 TCRs having strong statistical association with one or more HLAs (one-sided Fisher's Exact Test $p < 1e\text{-}4$) using the previously described L1LR method (Zahid et al., in review).

Constructing ECOclusters

We employed distinct methods for clustering TCRs into HLA-COclusters and for clustering HLA-COclusters into ECOclusters. For the latter, we opted to use agglomerative clustering on a correlation matrix, as this allowed us to explore an interpretable clustering threshold. For clustering TCRs into HLA-COclusters, direct computation of pairwise correlations performs very poorly due to the extreme sparsity of the TCR-by-donor matrix. We therefore first transformed the matrix through embedding and dimensionality reduction steps. These transformations lose the interpretability of the distance measure, and so we opted to use density-based clustering to define HLA-COclusters, rather than tuning HLA-specific clustering thresholds.

In more detail, to construct HLA-COclusters we used density-based clustering to identify, for each of 131 HLAs with at least 2,000 imputed donors, clusters of HLA-associated TCRs that tend to co-occur in a subset of donors inferred to have the HLA. We first used spectral co-clustering (Dhillon, 2001) to embed both TCRs and donors into a shared space of 150 dimensions, and to relate the problem of clustering TCRs to the problem of clustering

donors in terms of their TCRs. Next, we applied UMAP (McInnes et al., 2018) to reduce the dimensionality of this space to 15. Finally, we applied HDBSCAN (Campello et al., 2013) with a minimum cluster size of 10 TCRs and/or donors to define HLA-COclusters.

To construct ECOclusters, we clustered the HLA-COclusters: we constructed the matrix X of 30,674 donors by 43,673 HLA-COclusters, with values indicating the count of the TCR members of each HLA-COcluster occurring in each donor. We then computed the HLA-masked Pearson correlation matrix P between all pairs of HLA-COclusters, where each entry $P_{i,j}$ is equal to the Pearson correlation of $X_{\boldsymbol{d},i}$ and $X_{\boldsymbol{d},j}$, where $\boldsymbol{d}$ are the donors imputed to have the HLA or HLAs associated with both HLA-COclusters i and j. We defined a distance metric between all pairs of HLA-COclusters, $D = 1 - P$, which ranged from 0.0 to 1.0. We performed average-linkage (UPGMA) agglomerative clustering with a D threshold of 0.8 (corresponding to a Pearson correlation coefficient of 0.2). The 7,106 clusters thus defined each comprised between 5 and 287,393 TCRs and combining between 1 and 1,129 HLA-COclusters.

Serological labeling of exposure status

For CMV and SARS-CoV-2 labels, we used previously acquired serologically labeled samples as described previously (Emerson et al., 2015b; Snyder et al., 2020a).

For EBV, Parvovirus, HSV-1, HSV-2 and *T. gondii*, we derived new serological labels on previously acquired samples. A multiplexed serological testing method was developed in house using U-PLEX Development Pack from Meso Scale Discovery (MSD). Purified antigens (recombinant VCA p18 and EBNA-1 proteins for EBV, recombinant HSV-1 gG protein, recombinant HSV-2 gG protein and *T. gondii* antigen were purchased from Meridian Life Science. Parvovirus B19 VLP/VP1/VP2 Co-Capsid Recombinant protein was purchased from Raybiotech) were biotinylated at optimized biotin-to-protein ratios that generated biotinylated proteins with 1-3 biotin(s) per molecule.

Biotinylated antigens were coated on the plate simultaneously at optimized concentrations onto different spots via linker provided by MSD. After washing off the unbound antigens, sera samples diluted to optimized concentration with assay diluent were applied to the plate. Antibodies in the serum that recognize the plate bound antigens were detected by a sulfo-tag labeled anti-human IgG antibody. The signal level of each spot is in direct correlation with the amount of antigen-specific antibodies in the serum sample. A positive control that contains antibodies against all the antigens in the panel, a negative control that does not have detectable antibodies against any of the antigens in the panel and two cutoff samples that contain threshold level of antibodies against the antigens in the panel were run on each plate. The multiplexed serological testing method was validated using clinically labeled serum samples and using commercially available ELISA kits.

We normalized MSD signal in two steps to remove variation in background signal among (1) wells and samples, and (2) MSD plates. First, within each well of an MSD plate, we used the mean signal of spots without antigens as a measure of background signal and subtracted it from the signal of spots with bound disease antigens. We ran each sample in three wells and took the mean of the three background-adjusted signal values for each disease. Second, to remove variation among MSD plates, we included a cutoff sample in three wells of every plate and calculated the normalized signal, *S*, for each disease for each sample; for each antigen, we divided the mean background-adjusted signal of every sample on the plate by the mean background-adjusted signal of the cutoff sample. The signal of the cutoff sample was always greater than the background signal, so the denominator of *S* was always positive, but for some samples with low signal for a disease, the numerator (and *S*) was negative.

Statistical methods for identifying high-confidence serological labels

We observed a bimodal distribution of log *S* for each disease, so we modeled log *S* as a mixture of two univariate Gaussian distributions, assuming the component distributions with lower and higher signal represented controls and cases, respectively. When fitting the mixture model, we ignored all samples with negative *S;* this ranged from 1-6% of the samples among the diseases. After fitting the means, variances, and mixture proportion of the mixture model using all samples with positive *S*, we used Bayes' rule to calculate the probability each sample was a case given its value of log *S*. When calculating this probability for samples with negative *S*, we used the smallest positive *S* among the samples for the disease. For model training and evaluation, we considered samples with probability less than 0.01 and greater than 0.99 as high-confidence controls and cases, respectively.

For the EBV labels, as described above, we had labels and confidence estimates for two antigens, VCA and EBNA1. We used VCA as our primary indicator of donor EBV status. However, a small number of donors had a confident negative label for VCA but also had a label for EBNA1 that was insufficiently confidently negative (>0.1 posterior probability). We removed those labels from consideration.

Measuring each donor's response to each ECOcluster

For each ECOcluster, we define each donor's "ECOcluster count" C_{EC} as the number of unique ECOcluster-member TCRs in the donor's repertoire with HLA associations matching the donor's imputed HLA type. We define each donor's *UPR*, or Unique Productive Rearrangement count, as the total number of unique TCRs observed in the donor's repertoire. We then calculate "raw ECOcluster breadth" B_{EC} as $log_{10}(C_{EC} / UPR)$.

Next, we developed a measure of ECOcluster response adjusted for the donor's HLA type. For each ECOcluster, we constructed a linear regression model to predict a donor's B_{EC} (denoted $Bpred_{EC}$) from presence (encoded as 1) or absence (encoded as 0) of each ECOcluster-associated HLA according to their imputed HLA type. We calculated an "ECOcluster response" R_{EC}, adjusted for each donor's HLA type, as B_{EC} - $Bpred_{EC}$.

Building classifiers for disease labels

For each exposure, we divided the labeled repertoires into training and holdout sets, with labeled repertoire counts described in Table 1. All model development and selection, including the definition of the ECOclusters, B_{EC} and R_{EC}, was performed without any use of the holdout set. Within the training set, we tested each ECOcluster for association with case status, using a one-sided Mann-Whitney U test (MWU) on serologically positive *vs.* negative R_{EC} for each ECOcluster. We declared the ECOcluster with the lowest MWU *p* value to be the single ECOcluster associated with the exposure.

Computing enrichment of ECOclusters for TCRs with known association

Given the dataset of TCR-pHLA associations from public databases, we tested each ECOcluster for enrichment of TCRs associated with pHLAs from each different taxon. In this analysis, because of differences in V gene naming between our data and the databases, we matched TCRs by CDR3 amino acid sequence, V gene family and J gene family. 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. *p(x)* is computed using the hypergeometric distribution as follows:

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

where:

x = count of intersecting TCRs

m = count of TCRs in ECOcluster

k = count of TCRs in TCR list

n = estimate for total number of public TCRs (5 million) - *m*

BIBLIOGRAPHY

Babbitt, B. P., Allen, P. M., Matsueda, G., Haber, E., & Unanue, E. R. (1985). Binding of immunogenic peptides to Ia histocompatibility molecules. *Nature*, *317*(6035), 359–361.

Brown, J. H., Jardetzky, T. S., Gorga, J. C., Stern, L. J., Urban, R. G., Strominger, J. L., & Wiley, D. C. (1993). Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature*, *364*(6432), 33–39.

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7819 LNAI*(PART 2). https://doi.org/10.1007/978-3-642-37456-2_14

DeWitt, W. S., Smith, A., Schoch, G., Hansen, J. A., Matsen, F. A., & Bradley, P. (2018). Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *ELife*, *7*. https://doi.org/10.7554/eLife.38358

Dhillon, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/502512.502550

Emerson, R., DeWitt, W., Vignali, M., Gravley, J., Desmarais, C., Carlson, C., Hansen, J., Rieder, M., & Robins, H. (2015a). Immunosequencing reveals diagnostic signatures of chronic viral infection in T cell memory. *BioRxiv*, *January*.

Emerson, R., DeWitt, W., Vignali, M., Gravley, J., Desmarais, C., Carlson, C., Hansen, J., Rieder, M., & Robins, H. (2015b). Immunosequencing reveals diagnostic signatures of chronic viral infection in T cell memory. *BioRxiv*, *January*.

Fowler, K., Mucha, J., Neumann, M., Lewandowski, W., Kaczanowska, M., Grys, M., Schmidt, E., Natenshon, A., Talarico, C., Buck, P. O., & Diaz-Decaro, J. (2022). A systematic literature review of the global seroprevalence of cytomegalovirus: possible implications for treatment, screening, and vaccine development. *BMC Public Health*, *22*(1). https://doi.org/10.1186/s12889-022-13971-7

Fremont, D. H., Hendrickson, W. A., Marrack, P., & Kappler, J. (1996). Structures of an MHC class II molecule with covalently bound single peptides. *Science*, *272*(5264), 1001–1004.

Goronzy, J. J., & Weyand, C. M. (2017). Successful and maladaptive T cell aging. *Immunity*, *46*(3), 364–378.

Greissl, J., Pesesky, M., Dalai, S. C., Rebman, A. W., Soloski, M. J., Horn, E. J., Dines, J. N., Gittelman, R. M., Snyder, T. M., Emerson, R. O., Meeds, E., Manley, T., Kaplan, I. M., Baldo, L., Carlson, J. M., Robins, H. S., & Aucott, J. N. (2021). Immunosequencing of

the T-cell receptor repertoire reveals signatures specific for diagnosis and characterization of early Lyme disease. *MedRxiv*.

Greninger, A. L., Roychoudhury, P., Xie, H., Casto, A., Cent, A., Pepper, G., Koelle, D. M., Huang, M.-L., Wald, A., Johnston, C., & Jerome, K. R. (2018). Ultrasensitive Capture of Human Herpes Simplex Virus Genomes Directly from Clinical Samples Reveals Extraordinarily Limited Evolution in Cell Culture. *MSphere*, *3*(3). https://doi.org/10.1128/mSphereDirect.00283-18

Guan, Y., Jakimovski, D., Ramanathan, M., Weinstock-Guttman, B., & Zivadinov, R. (2019). The role of Epstein-Barr virus in multiple sclerosis: From molecular pathophysiology to in vivo imaging. In *Neural Regeneration Research* (Vol. 14, Issue 3). https://doi.org/10.4103/1673-5374.245462

Katayama, Y., Yokota, R., Akiyama, T., & Kobayashi, T. J. (2022). Machine learning approaches to TCR repertoire analysis. *Frontiers in Immunology*, *13*, 858057.

Klinger, M., Pepin, F., Wilkins, J., Asbury, T., Wittkop, T., Zheng, J., Moorhead, M., & Faham, M. (2015). Multiplex identification of antigen-specific T cell receptors using a combination of immune assays and immune receptor sequencing. *PLoS ONE*, *10*(10). https://doi.org/10.1371/journal.pone.0141561

Kumar, B. V, Connors, T. J., & Farber, D. L. (2018). Human T cell development, localization, and function throughout life. *Immunity*, *48*(2), 202–213.

Liu, X., & Wu, J. (2018). History, applications, and challenges of immune repertoire research. *Cell Biology and Toxicology*, *34*, 441–457.

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, *3*(29). https://doi.org/10.21105/joss.00861

Nikolich-Žugich, J. (2014). Aging of the T cell compartment in mice and humans: from no naive expectations to foggy memories. *The Journal of Immunology*, *193*(6), 2622–2629.

Pradier, M. F., Prasad, N., Chapfuwa, P., Ghalebikesabi, S., Ilse, M., Woodhouse, S., Elyanow, R., Zazo, J., Gonzalez Hernandez, J., Greissl, J., & Meeds, E. (2023). AIRIVA: A Deep Generative Model of Adaptive Immune Repertoires. In *Proceedings of Machine Learning Research* (Vol. 219).

Pradier, M. F., Prasad, N., Chapfuwa, P., Ghalebikesabi, S., Ilse, M., Woodhouse, S., Elyanow, R., Zazo, J., Hernandez, J. G., Greissl, J., & others. (2023). AIRIVA: a deep generative model of adaptive immune repertoires. *Machine Learning for Healthcare Conference*, 588–611.

Qi, Q., Liu, Y., Cheng, Y., Glanville, J., Zhang, D., Lee, J.-Y., Olshen, R. A., Weyand, C. M., Boyd, S. D., & Goronzy, J. J. (2014). Diversity and clonal selection in the human T-cell repertoire. *Proceedings of the National Academy of Sciences*, *111*(36), 13139–13144.

Robins, H. (2013). Immunosequencing: applications of immune repertoire deep sequencing. *Current Opinion in Immunology*, *25*(5), 646–652.

Shim, C. H., Cho, S., Shin, Y. M., & Choi, J. M. (2022). Emerging role of bystander T cell activation in autoimmune diseases. *BMB Reports*, *55*(2). https://doi.org/10.5483/BMBRep.2022.55.2.183

Shugay, M., Bagaev, D. V., Zvyagin, I. V., Vroomans, R. M., Crawford, J. C., Dolton, G., Komech, E. A., Sycheva, A. L., Koneva, A. E., Egorov, E. S., Eliseev, A. V., Van Dyk, E., Dash, P., Attaf, M., Rius, C., Ladell, K., McLaren, J. E., Matthews, K. K., Clemens, E. B., … Chudakov, D. M. (2018). VDJdb: A curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Research*, *46*(D1). https://doi.org/10.1093/nar/gkx760

Snyder, T. M., Gittelman, R. M., Klinger, M., May, D. H., Osborne, E. J., Taniguchi, R., Zahid, H. J., Kaplan, I. M., Dines, J. N., Noakes, M. N., Pandya, R., Chen, X., Elasady, S., Svejnoha, E., Ebert, P., Pesesky, M. W., De Almeida, P., O'Donnell, H., DeGottardi, Q., … Robins, H. S. (2020a). Magnitude and Dynamics of the T-Cell Response to SARS-CoV-2 Infection at Both Individual and Population Levels. *MedRxiv : The Preprint Server for Health Sciences*. https://doi.org/10.1101/2020.07.31.20165647

Snyder, T. M., Gittelman, R. M., Klinger, M., May, D. H., Osborne, E. J., Taniguchi, R., Zahid, H. J., Kaplan, I. M., Dines, J. N., Noakes, M. N., Pandya, R., Chen, X., Elasady, S., Svejnoha, E., Ebert, P., Pesesky, M. W., De Almeida, P., O'Donnell, H., DeGottardi, Q., … Robins, H. S. (2020b). Magnitude and Dynamics of the T-Cell Response to SARS-CoV-2 Infection at Both Individual and Population Levels. *MedRxiv : The Preprint Server for Health Sciences*. https://doi.org/10.1101/2020.07.31.20165647

Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull*, *38*.

Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E., & Friedman, N. (2017). McPAS-TCR: A manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics*, *33*(18). https://doi.org/10.1093/bioinformatics/btx286

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., & Peters, B. (2019). The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research*, *47*(D1). https://doi.org/10.1093/nar/gky1006

Zinkernagel, R. M., Callahan, G. N., Klein, J. A. N., & Dennert, G. (1978). Cytotoxic T cells learn specificity for self H–2 during differentiation in the thymus. *Nature*, *271*(5642), 251–253.