

RESEARCH ARTICLE

TCR2HLA: Calibrated inference of HLA genotypes from TCR repertoires enables identification of immunologically relevant metaclonotypes

Koshlan Mayer-Blackwell^{1,2}, Anastasia Minervina³, Mikhail Pogorelyy³, Puneet Rawat⁴, Melanie R. Shapiro⁵, Leeana D. Peters⁵, Emily S. Ford^{1,6}, Amanda L. Posgai⁵, Kasi Vegesana³, Samuel Minot^{1,2}, David M. Koelle^{1,6,7,8,9}, Victor Greiff^{4,10}, Philip Bradley¹¹, Todd M. Brusko^{5,12,13}, Paul G. Thomas³, Andrew Fiore-Gartland^{1,2*}

1 Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, Seattle, Washington, United States of America, **2** Vaccine Immunology Statistical Center, Fred Hutchinson Cancer Center, Seattle, Washington, United States of America, **3** St. Jude Children's Research Hospital, Memphis, Tennessee, United States of America, **4** Department of Immunology, University of Oslo and Oslo University Hospital, Oslo, Norway, **5** Department of Pathology, Immunology, and Laboratory Medicine, Diabetes Institute, College of Medicine, University of Florida, Gainesville, Florida, United States of America, **6** Department of Medicine, University of Washington, Seattle, Washington, United States of America, **7** Department of Global Health, University of Washington, Seattle, Washington, United States of America, **8** Benaroya Research Institute, Seattle, Washington, United States of America, **9** Department of Laboratory Medicine and Pathology, University of Washington, Seattle, Washington, United States of America, **10** Imprint Labs, LLC., New York, New York, United States of America, **11** Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, Washington, United States of America, **12** Department of Pediatrics, Diabetes Institute, College of Medicine, University of Florida, Gainesville, Florida, United States of America, **13** Department of Biochemistry and Molecular Biology, College of Medicine, University of Florida, Gainesville, Florida, United States of America

* agartlan@fredhutch.org

Abstract

T cell receptors (TCRs) recognize peptides presented by polymorphic human leukocyte antigen (HLA) molecules, but HLA genotype data are often missing from TCR repertoire sequencing studies. To address this, we developed TCR2HLA, an open-source tool that infers HLA genotypes from TCR β repertoires. Expanding on work linking public TRBV-CDR3 sequences to HLA genotypes, we incorporated “quasi-public” metaclonotypes – composed of rarer TCR β sequences with shared amino acid features – enriched by HLA genotypes. Using four TCR β seq datasets from 3,150 individuals, we applied TRBV gene partitioning and locality-sensitive hashing to identify ~96,000 TCR β features strongly associated with specific HLA alleles from 71M input TCRs. Binary HLA classifiers built with these features achieved high balanced accuracy (>0.9) across common HLA-A (9/12), B (9/12), C (6/13), DRB1 (11/11) alleles and prevalent DPA1/DPB1 (6/10), DQA1/DQB1 (8/17) heterodimers. We also introduced a high-sensitivity calibration to support predictions in samples with as few as 5,000 unique clonotypes. Calibrated predictions with confidence filtering improved reliability. Beyond genotype imputation, TCR2HLA enables the discovery of novel HLA- and exposure-associated TCRs, as shown by the identification of SARS-CoV-2



OPEN ACCESS

Citation: Mayer-Blackwell K, Minervina A, Pogorelyy M, Rawat P, Shapiro MR, Peters LD, et al. (2026) TCR2HLA: Calibrated inference of HLA genotypes from TCR repertoires enables identification of immunologically relevant metaclonotypes. *PLoS Comput Biol* 22(1): e1013767. <https://doi.org/10.1371/journal.pcbi.1013767>

Editor: Brittany Rife Magalis, University of Louisville, UNITED STATES OF AMERICA

Received: August 8, 2025

Accepted: November 19, 2025

Published: January 16, 2026

Copyright: © 2026 Mayer-Blackwell et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: Data availability: TCR repertoires data used to identify HLA-associated features were analyzed from previous studies and are available (<https://>

clients.adaptivebiotech.com/pub/emerson-2017-natgen, <https://clients.adaptivebiotech.com/pub/elyanow-2022-jci>, <https://clients.adaptivebiotech.com/pub/musvosvi-2022-nm>). TCR repertoire data from Rawat et al. study (doi: [10.1101/2024.12.10.24318751](https://doi.org/10.1101/2024.12.10.24318751)) is deposited at the AIRR Data Commons using the iReceptor Gateway using Study ID IR-T1D-000004. HLA genotype labels are available from supporting information in primary manuscripts [doi: [10.7554/eLife.38358](https://doi.org/10.7554/eLife.38358), doi: [10.1101/2024.12.10.24318751](https://doi.org/10.1101/2024.12.10.24318751) and doi: [10.21417/mm2022nm](https://doi.org/10.21417/mm2022nm)]. HLA genotype data for participants in the Rawat et al. study are deposited in dbGaP under accession phs003979.v1.p1. HLA genotype data for participants in the Elyanow et al. study (doi: [10.1172/jci.insight.150070](https://doi.org/10.1172/jci.insight.150070)) are deposited in dbGaP under accession phs004406.v1.p1. Code Availability: Code used for feature discovery, model fitting, and calibration is available at github.com/kmayerb/TCR2HLA. A command-line python tool for running TCR2HLA with multiple cpus is available at github.com/kmayerb/TCR2HLA. An interactive serverless web-browser application is also provided: kmayerb.github.io/TCR2HLAi/. Tools for CPU or GPU-accelerated approximate TCRdist computation and network-based meta-clonotype motif discovery are available at github.com/kmayerb/tcrdistgpu.

Funding: KMB, AFG were supported by NIH grant number R01 AI136514 and by a grant from the Gates Foundation (INV-027499). The conclusions and opinions expressed in this work are those of the authors alone and shall not be attributed to the Foundation. Scientific Computing Infrastructure at Fred Hutchinson Cancer Research Center was funded by an ORIP grant S100D028685. Specimen collection included samples from persons enrolled in NCT04338360 or NCT04344977 funded by the US NIH, with data collection from these samples funded by the US NIH NIAID Contract 75N93019C00063 (DMK). Type 1 Diabetes datasets were supported by grants from The Leona M. and Harry B. Helmsley Charitable Trust (#2019PG-T1D011, to TMB and VG), the National Institutes of Health (P01 AI042288, to TMB; K99 DK140511, to MRS), and the American Diabetes Association (11-23-PDF-78, to LDP). VG was supported by a Norwegian Cancer Society Grant (#215817), Research Council of Norway grants (300740, 311341, 331890), and an ERC-CoG (101125630). PR was supported by the European Union's

related TCRs in a large COVID-19 dataset lacking HLA data. TCR2HLA provides a scalable framework for bridging the gap between TCRseq data and HLA genotype for biomarker discovery.

Author summary

T cells are a crucial component of the immune system. Each T cell has a unique T cell receptor (TCR) that can detect a potential infection by recognizing protein fragments presented on the surface of cells, bound by human leukocyte antigen (HLA) proteins. The HLA proteins are highly variable across individuals and determine what part of a pathogen each person's immune system can recognize. Thus, HLA genotype and pathogen exposure shape the composition of T cell memory over an individual's lifetime and are important for interpreting their T cell responses and TCR repertoire. However, HLA genotype data are often absent from TCR sequencing studies. To help, we developed TCR2HLA, an open-source tool that infers HLA genotype from the thousands of TCR sequences recoverable in blood samples. Statistical inference of HLA genotype probability enables grouping of data from participants with shared HLA gene variants, unlocking approaches to identify TCRs that could be biomarkers of infection, autoimmunity, or vaccine response.

Introduction

Highly diverse $\alpha\beta$ T cell receptors (TCRs) survey peptides presented by human leukocyte antigen (HLA) molecules. Although TCR repertoires are routinely sequenced from blood and tissues, identifying immunologically relevant features within diverse adaptive immune receptor repertoire (AIRR) data remains a challenging problem [1,2]. Within each individual, highly polymorphic HLA molecules restrict the range of peptides available to interact with TCRs. Sufficiently large cohorts can be leveraged to identify the combinations of immunological exposure and HLA genotypes [3–5] that increase the likelihood of observing specific TCR features above the prevalence modulated by V(D)J recombination and thymic selection [3]. The capacity to evaluate TCR feature enrichment following immunological exposures across datasets in an HLA-defined manner would be particularly helpful in identifying relevant peptide-HLA epitope interactions that underpin T cell-mediated immunological control or pathology.

HLA alleles are encoded by distinct genetic loci. The class I loci – HLA-A, HLA-B, and HLA-C – encode proteins that complex with β 2-microglobulin to present intracellular peptides to CD8⁺ T cells, while the class II loci – HLA-DRA1/DRB1, HLA-DQA1/DQB1, and HLA-DPA1/DPB1 – encode heterodimers that present peptides to CD4⁺ T cells [6–8]. To address the challenge that HLA genotypes are not universally collected in TCR repertoire sequencing (TCRseq) studies, recent work has explored the utility of using public TCR features to infer HLA genotypes from bulk TCR β repertoires

Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801133. AM, MP, and PGT were supported by NIH grants AI136514, AI144616, and AI165077, and American Lebanese Syrian Associated Charities (ALSAC) at St. Jude. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: The authors declare the following competing interests: V.G. declares advisory board positions in aiNET GmbH, Enpicom B.V, Absci, Fairjourney Biologics and Diagonal Therapeutics. V.G. is a consultant for Adaptyv Biosystems, Proteinea, and LabGenius. V.G. is an employee of Imprint LLC. PGT is on the Scientific Advisory Board of Immunoscope and Shennon Bio, has received research support and personal fees from Elevate Bio, and consulted for 10X Genomics, Illumina, Pfizer, Cytoagents, Sanofi, Merck, and JNJ. PGT, AAM, and MVP have patents related to TCR amplification, cloning, and/or applications thereof.

[9–11]. In this study, we exploit the fact that highly biochemically similar TCRs are often found across repertoires, such that searching for meta-clonotypes can increase feature publicity and reveal HLA-associated TCRs and motifs [11–13]. Due to biases in V(D)J recombination, many TCR sequences are uncommon and unlikely to be shared between individuals; however, rare non-identical receptors may undergo clonal expansion following antigen exposure [3,14–16]. Thus, we used near-exact neighborhood tabulation to capture “quasi-public” TCRs – rarer sequences [17] that, despite not being broadly shared themselves, occur within sequence neighborhoods that are public enough among persons with exposure to common pathogens to support statistical association with specific HLA alleles. For inference of HLA genotype directly from repertoire data, we then developed models that combine (i) exact public features with (ii) sparsely public near-exact TCRs to build a binary classifier for each allele.

This study contributes open-source software that permits the discovery of exact and near-exact (edit distance 1) HLA-associated TCR β features across thousands of input repertoires. The resulting high-confidence HLA genotype inferences allow for the identification of novel TCR:HLA associations in datasets lacking genotype labels. This facilitates the integration of labeled and unlabeled repertoires for discovering exposure- and HLA-associated TCRs that could serve as biomarkers for infectious disease, autoimmunity, cancer, and vaccine applications.

Results

TCR2HLA identifies exact and near-exact HLA-associated TCR features across cohorts

TCR2HLA is a statistical tool for inferring HLA genotypes from the occurrence patterns of receptors in peripheral blood TCR β repertoires. Leveraging a training set of 3,125 repertoires from four independent cohorts with HLA-labeled data [18–21], we identified exact and near-exact TRBV-CDR3 β features associated with common HLA alleles (Fig 1A–C). Building on prior work identifying public TRBV-CDR3 sequences linked to HLA genotypes, we aimed to incorporate additional, “quasi-public” clonotypes with HLA-association evident based on detection of a TRBV-CDR3 AA feature or any single mutational variant. Notably, the probability of generation of TCRs with near-exact HLA-associated neighborhoods tended to be lower than that of exact TCR features (S1A Fig). Combining distinct sets of exact and near-exact TRBV-CDR3 β amino acid features (S2 Table), we fit allele-specific classifiers via L1-regularized logistic regression, retaining the subset of features with non-zero weights (S3 Table). Models used inverse statistical weighting to address class imbalance and post-estimation recalibration to standardize predicted probabilities across alleles and account for repertoire depth, helping to distinguish high-confidence positive or negative genotype calls from ambiguous ones (Fig 1D). We evaluated model performance in multiple external validation cohorts with varied TCRseq amplification protocols and differing levels of richness (S2 Fig and S4 Table) before we applied TCR2HLA models to 1,355 previously ungenotyped repertoires from SARS-CoV-2–exposed individuals, identifying novel exposure-associated meta-clonotypes through clustering of

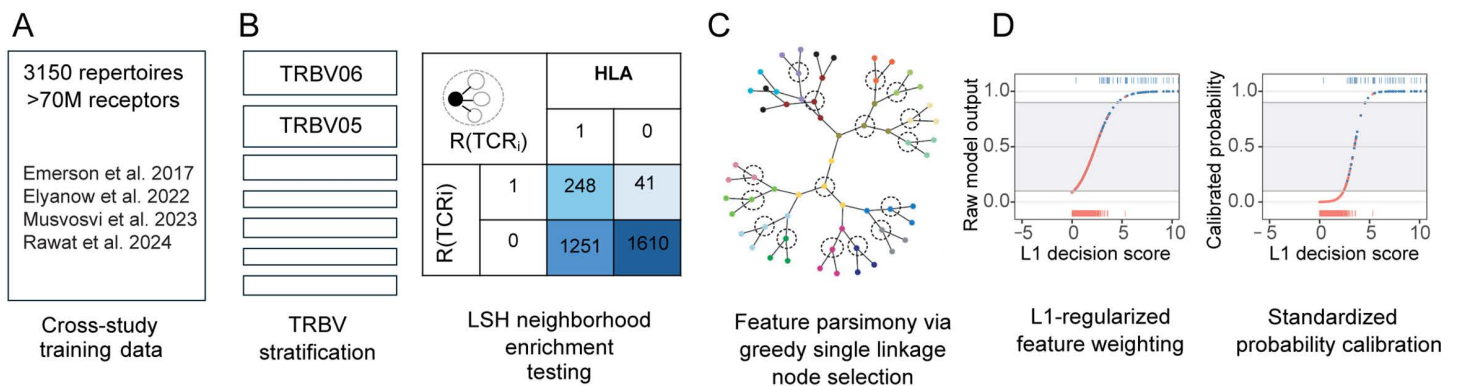


Fig 1. Schematic for generating calibrated HLA genotype predictions from TCR β repertoires. Schematic of statistical framework for discovering near-exact (edit distance 1) TCR features (defined by TRBV-family + CDR3 (amino acid)) in a multi-cohort wide association test. **(A)** 71 M receptors present in >1 copy were assembled across 3150 repertoires from 4 studies (S1 Table). **(B)** Use of TRBV-family stratification combined with locality-sensitive hashing facilitated discovery of exact and inexact HLA-associated TCR features. Fisher's exact test is applied to an indicator function defined over the set $\{R\}$ of neighboring TCRs to TCR_i, where neighbors are defined as those TCRs within a mutational distance of 1 (i.e., differing by a single substitution, insertion, or deletion in the CDR3 sequence) and detected at a frequency exceeding the threshold 2×10^{-6} . **(C)** A greedy single linkage algorithm was used to reduce feature redundancy and potential variable collinearity prior to learning feature weights in L1-regularized logistic regression. **(D)** Raw outputs from class-weighted L1-logistic regression optimized for AUROC were recalibrated to standardized probabilities, enabling standardized decision thresholding and confidence-based sample retention for downstream analysis.

<https://doi.org/10.1371/journal.pcbi.1013767.g001>

similar TCR sequences within HLA-associated sequence neighborhoods. More detailed descriptions of feature discovery, model training, and validation are provided in the Methods section.

Allelic prevalence is similar across training and validation data sets

The prevalence of common HLA class I alleles was similar among the training data and the U.S. based validation datasets (Fig 2A). By contrast, the prevalence of class II alleles differed between the training and validation sets. As expected, alleles associated with T1D risk, particularly DRB1*03:01, DQA1*03:02, DQA1*05:01, DQB1*03:02, and DQB1*02:01 and associated haplotypes, were enriched in the training data participants relative to those in the external validation datasets. Given strong geographical effects on the distribution of certain HLA alleles in human populations, we also examined the prevalence of HLA alleles from participants in a study conducted in South Africa (Fig 2A, Musvosvi dataset). Reflective of the diversity represented in this dataset, certain alleles – particularly A*01:01, A*02:01, A*03:01, B*08:01, B*44:02, C*07:02 – were underrepresented in the South African cohort (Fig 2A and S5 Table). Some alleles (e.g., HLA-A*02:05, B*58:02, B*42:01, DRB1*15:03) that were strongly enriched in the South African participants but were not considered in the development of the cross-study common allele model because they were rare in the training and primary validation datasets.

Common HLA genotypes can be inferred based on the presence of TCR features

To estimate model performance on healthy adult HLA-unlabeled TCR β repertoires, we evaluated each common-allele model using validation repertoires that were neither used for feature discovery nor weighting. The model's decision scores robustly separated samples by their HLA-status (AUC > 0.95) for most alleles across HLA-A, HLA-B, HLA-DR, and HLA-DQ loci (Fig 2B). We also assessed the models' balanced accuracies (BA) – measuring the ability to correctly predict positive and negative cases with a predefined decision threshold (Fig 2C). In the primary validation repertoires, models correctly predicted the HLA-status with high balanced accuracy (BA > 0.9) for some, but not all alleles, with performance differing by HLA-locus: HLA-A (9/12 alleles), B (9/12 alleles), C (6/13 alleles), DRB1 (11/11 alleles), DQA1 (6/9 alleles),

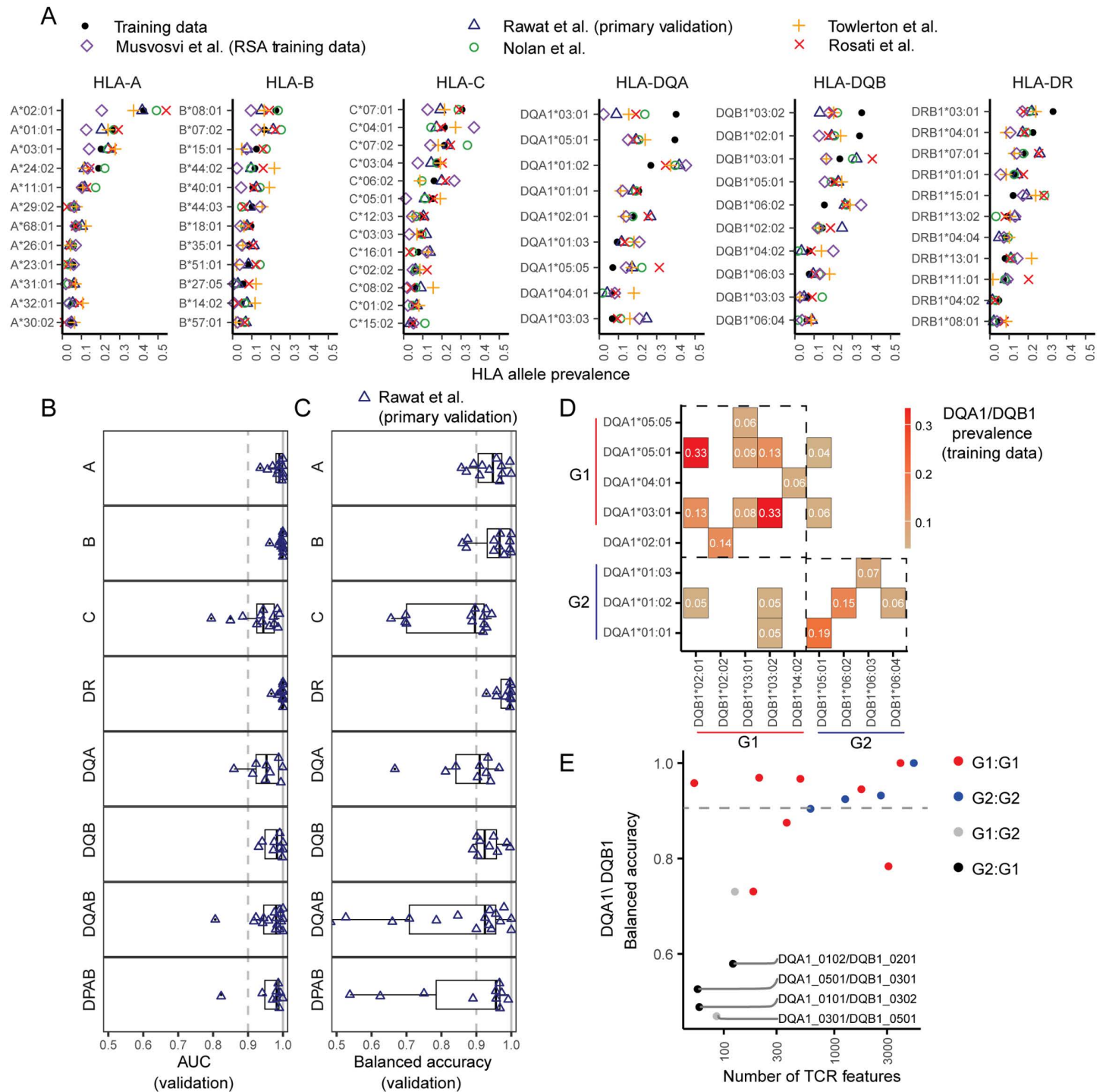


Fig 2. TCR2HLA classifier performance varies by locus on external validation datasets. (A) Prevalence of alleles in training data versus validation datasets (Rawat et al., cohort 3 (primary validation), Towler et al., Nolan et al. and Rosati et al.). The Musvosvi et al. dataset represents a subset of the training data that includes participants from the Republic of South Africa (RSA). The model was fit for alleles or allele heterodimer (DQ, DP) combinations present in greater than four percent of overall training data – HLA-A (alleles, $n=12$), HLA-B ($n=12$), HLA-C ($n=13$), HLA-DR ($n=11$), HLA-DQA ($n=10$), HLA-DQB ($n=10$), HLA-DPA ($n=4$), HLA-DPB ($n=6$), and HLA-DQAB ($n=17$) and HLA-DPAB ($n=9$). (Prevalence of HLA-DP, HLA-DPAB heterodimers, and HLA-DQAB heterodimers, [S5 Table](#)) (B) Receiver operator area under the curve (AUC) in the primary external validation cohort (adults

18-65, unique TCR > 25,000) by allele. (C) Balanced accuracy (average of sensitivity and specificity) in the primary external validation cohort (adults 18-65, unique TCR > 25,000) by allele. (D) Prevalence of DQA1/DQB1 group 1 (G1:G1), group 2 (G2:G2), and cross-group (G1:G2 or G2:G1) heterodimers in training data (E) Balanced accuracy and number of unique TCR model features for each DQA1/DQB1 heterodimer. Colors indicate within-group or cross-group pairings.

<https://doi.org/10.1371/journal.pcbi.1013767.g002>

and DQB1 (8/10 alleles), as well as for DPA1/DPB1 (6/10) and DQA1/DQB1 (9/17) heterodimeric pairings (performance by allele, [S6 Table](#)). Using the standard calibration (unadjusted for total repertoire diversity), models had high specificity on the primary validation data (i.e., median.99, IQR [0.95,0.99]) but a wider range of sensitivities across alleles (median 0.92, IQR [0.80,1]), with incorrect predictions across all models split between false positives (2.8%) and false negatives (1.5%) ([S7 Table](#)).

After considering models for individual DQA1 or DQB1 alleles, we trained separate classifiers for DQA1/DQB1 heterodimers. While most models for common DQA1/DQB1 dimers had moderate to high balanced accuracy (>0.8-1.0 BA), the models predicting certain DQA1/DQB1 heterodimer pairings had low accuracy (BA < 0.6), despite the co-occurring alleles being prevalent in > 4% percent of training participants ([Fig 2D](#)). At the DQ loci, two main classes of heterodimers are predicted to form stable HLA-DQ $\alpha\beta$ molecules: Group 1 (DQA1*02/04/05/06 α + DQB1*02/03/04 β) and Group 2 (DQA1*01 α + DQB1*05/06 β) [[22,23](#)]. Model accuracy for most Group 1 (G1 α :G1 β) and Group 2 (G2 α :G2 β) heterodimers was high. However, the remarkably low model accuracy for some heterodimers (e.g., DQA1*03:01/DQB1*05:01, DQA1*01:01/DQB1*03:02, DQA1*05:01/DQB1*03:01, [Fig 2E](#)) is consistent with the hypothesized structural incompatibility of some cross-group pairings (i.e., G1 α :G2 β and G2 α :G1 β).

Model recalibration boosts balanced accuracy of genotype prediction in shallowly sequenced adult repertoires

The alpha-diversity of the primary validation repertoires ([Fig 3A](#)) resembled that of repertoires used to train the models ([Fig 3B](#)), however, other TCRseq datasets can have fewer unique TCR clonotypes per repertoire. To evaluate model performance in lower-diversity repertoires, we applied our HLA prediction model to data from Towler et al. [[24](#)], where many samples contained only 5,000–25,000 unique TCR β s – substantially fewer than in the primary validation set (25,000–450,000 unique TCR β s per sample; [Fig 3A](#) and [3C](#)). Despite high class I HLA areas under the receiver operator curve (AUROCs) ([Fig 3E](#)), balanced accuracy was lower using the standard threshold calibrated from training data ($P(\text{HLA} = 1) > 0.5$, [Fig 3F](#)), due to fewer detections of HLA-enriched features in low diversity samples. The lower AUC and predictive performance observed for class II alleles in the Towler et al. validation dataset may be a consequence of depleted CD4⁺ T cell diversity expected in repertoires from participants living with human immunodeficiency virus (HIV), with one third of repertoires in that study collected prior to initiation of antiretroviral therapy [[25](#)].

To improve HLA inference in lower diversity samples, we implemented a high-sensitivity (HS) recalibration, adding log₁₀-transformed repertoire size to the model calibration with downsampled synthetic data ([Fig 3D](#), see Methods). For the Towler validation repertoires, the HS calibration improved balanced accuracy compared with standard calibration (IQR [0.72,0.91] vs. IQR [0.60,0.82] across all alleles) and reduced the number of false negatives, particularly in shallow repertoires ([Fig 3G](#)) with a slight cost in specificity (IQR [.97-1] vs. IQR [.99-1] across all models). The HS calibration yielded more reliable probabilistic outputs, with fewer extreme false negative misclassifications. Overall, while recalibration markedly improved performance on shallow repertoires (Towler et al., median = 27,119 unique TCR β), balanced accuracy achieved on these repertoires remained lower than in the primary validation set repertoires, possibly reflecting fundamental limits of inference from shallow sequencing data or in persons with highly depleted T cell memory. Further validation in additional DNA- and RNA-based TCRseq datasets showed sensitivity consistently dropping for repertoires with fewer than 25,000 unique clones ([S2 Fig](#) and [S6 Table](#)).

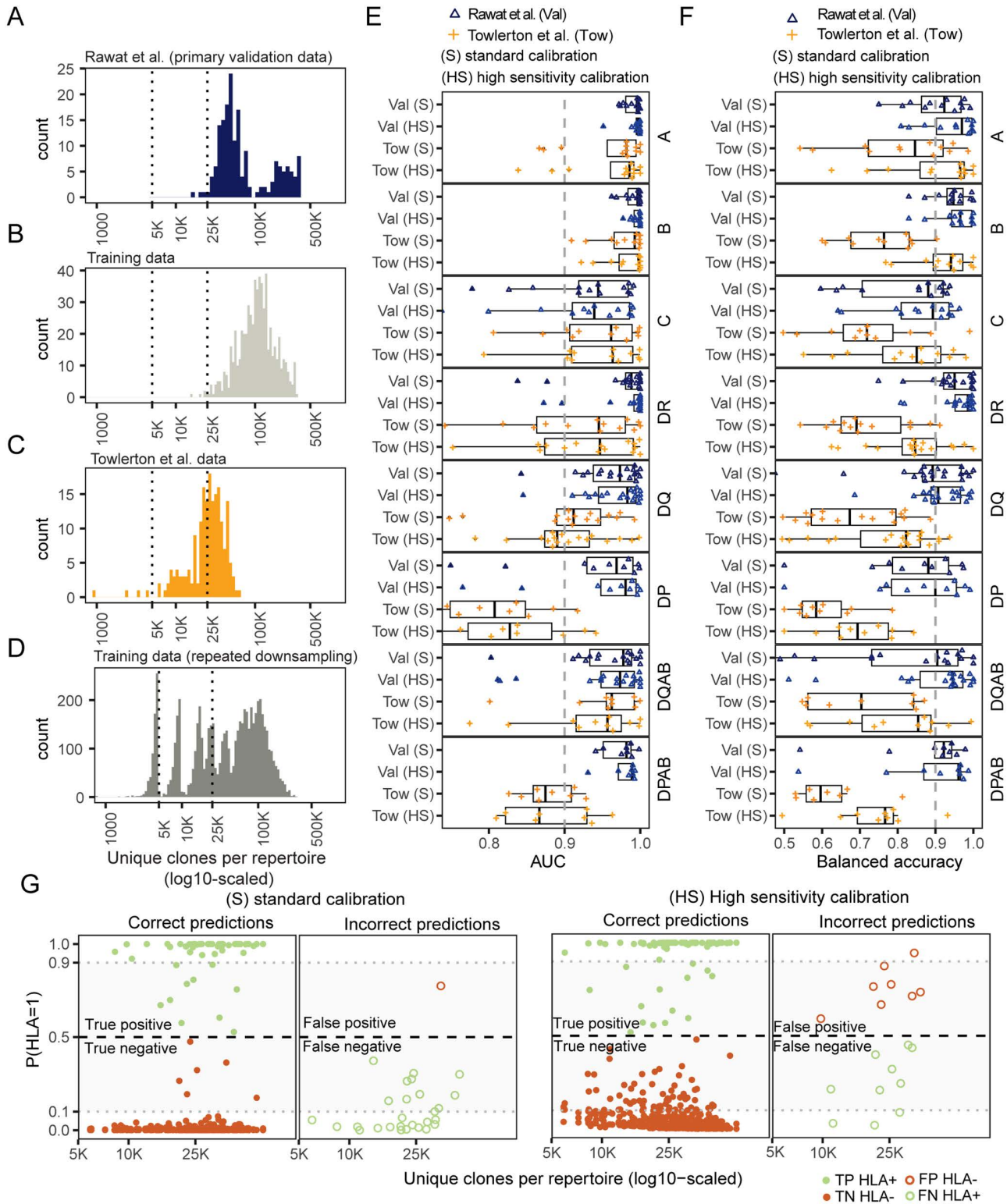


Fig 3. High sensitivity model calibration boosts sensitivity in lower diversity repertoires. Distribution of unique productive TCR β clonotypes sampled per repertoire in: (A) Rawat et al. primary validation (Val, unique TCR > 5,000) data (B) training data, (C) Towlerton et al. (Tow, unique TCR > 5,000) data with shallower depth of sampling (D) and repeated down sampled training data. (E) AUC in external validation datasets using either a standard (S) or high sensitivity (HS) calibration. (F) Balanced accuracy in external validation datasets using either a standard (S) or high sensitivity (HS) calibration. (G) HLA prediction probabilities. Left: (S) standard calibration. Right: (HS) High sensitivity calibration. Columns: Correct predictions, Incorrect predictions. Y-axis: $P(\text{HLA}=1)$ (0.0 to 1.0). X-axis: Unique clones per repertoire (log₁₀-scaled) (5K to 25K). Legend: TP HLA+ (green circle), FP HLA- (red circle), TN HLA- (orange circle), FN HLA+ (green circle).

or high sensitivity (HS) calibration. **(F)** Balanced accuracy in validation datasets using either the standard (S) or high sensitivity (HS) calibration. **(G)** Correct and incorrect predictions using the S or HS calibration. Prediction shown for A*03:01, A*02:01, A*01:01 alleles, where the colors of points show true HLA-status of the donor (green: HLA positive; orange: HLA negative), with closed dots indicating correct predictions (TP or TN) and open dots indicating incorrect predictions (FP or FN).

<https://doi.org/10.1371/journal.pcbi.1013767.g003>

Calibrated HLA genotype probabilities enable confidence-based label retention

While highly sensitive and accurate prediction was only possible for some HLA alleles, well-calibrated genotype probabilities can differentiate high from lower-confidence genotype inferences for those alleles that are more challenging to predict. To account for differences in model sensitivity across HLA loci and variability in repertoire sampling depth, we explored the use of calibrated HLA-genotype probabilities to implement confidence-based label retention. If probabilities are well-calibrated, masking of intermediate-confidence predictions should substantially increase balanced accuracy in retained data.

Thus, to assess the model calibration and feasibility of foregoing prediction in cases where the information content of a repertoire is low, we retained only high-confidence predictions ($P < 0.1$ or $P > 0.9$; Fig 4A–4B). This resulted in masking less than 10% of the predictions for most HLA-A, -B, and -DR alleles, and less than 20% of predictions for HLA-C and -DQ alleles (Fig 4C). For HLA-DP alleles, this procedure resulted in data loss due to more intermediate-confidence predictions. After masking, median balanced accuracy exceeded 0.95 across most loci. While median accuracy among HLA-C alleles was less than 0.9 after data filtering, high balanced accuracy was possible for some HLA-C alleles (i.e., C*03:04, *04:01, *05:01, *06:02, *07:01). These results show that confidence-based masking can yield high-fidelity HLA labels for many alleles while retaining allele-level predictions from $\geq 80\%$ of the repertoires.

HLA-inference facilitates the discovery of new HLA- and exposure-associated meta-clonotypes

Accurate HLA prediction is particularly valuable because it increases the possibility of discovering infection- or vaccination-associated TCRs in repertoires lacking explicit HLA genotype information. To this end, we applied TCR2HLA to 1,355 repertoires from individuals with acute or convalescent COVID-19 (Snyder et al.). We sought to use high-confidence genotype inferences ($P < 0.1$ and $P > 0.9$) to search the repertoires for novel HLA-associated receptors potentially associated with their common recent immunologic exposure. Using the inferred participant genotypes, we found numerous TCRs that were enriched in participants expressing specific alleles, but that did not match previously identified HLA-associated TCRs, including for representative allele DRB1*01:01 (Fig 5A). A subset of the TCRs statistically associated with DRB1*01:01 formed distinct public sequence clusters (Fig 5B). Because the reference datasets used to impute HLA labels were derived primarily from pre-2019 repertoires, we hypothesized that the unique TCR clusters discovered in the Snyder et al. cohort reflect expansions of SARS-CoV-2 reactive clonotypes.

To reduce redundancy and engineer features generalizable to closely related receptors, we organized novel enriched TCRs into meta-clonotypes defined by an anchor sequence and its first-degree neighbors based on TCRdist-based similarity networks. These clusters, represented by a V-gene family and CDR3 regular expression, capture quasi-public TCRs hypothesized to recognize the same HLA-presented epitope (Fig 5C). We screened the meta-clonotypes for their prevalence in individuals with and without the associated HLA allele in the Snyder et al. cohort and a pre-COVID cohort (Rawat et al. cohorts 1 and 2) [18,26]. When stratified by exposure and genotype, many of these novel HLA-associated meta-clonotypes showed specific enrichment in repertoires from SARS-CoV-2-infected individuals carrying the relevant restricting HLA allele (Fig 5D), supporting their potential application as antigen-specific biomarkers.

A major strength of TCR2HLA and our cross-study reference set of broadly HLA-associated TCR clones is the potential to discover novel HLA-associated TCRs that are selectively enriched by exposure and HLA genotype. In this example, we show how TCR2HLA can be used to infer the HLA status of a large set of unlabeled repertoires and identify new clusters of TCRs with strongly HLA-associated neighborhoods. This is particularly valuable for the identification of

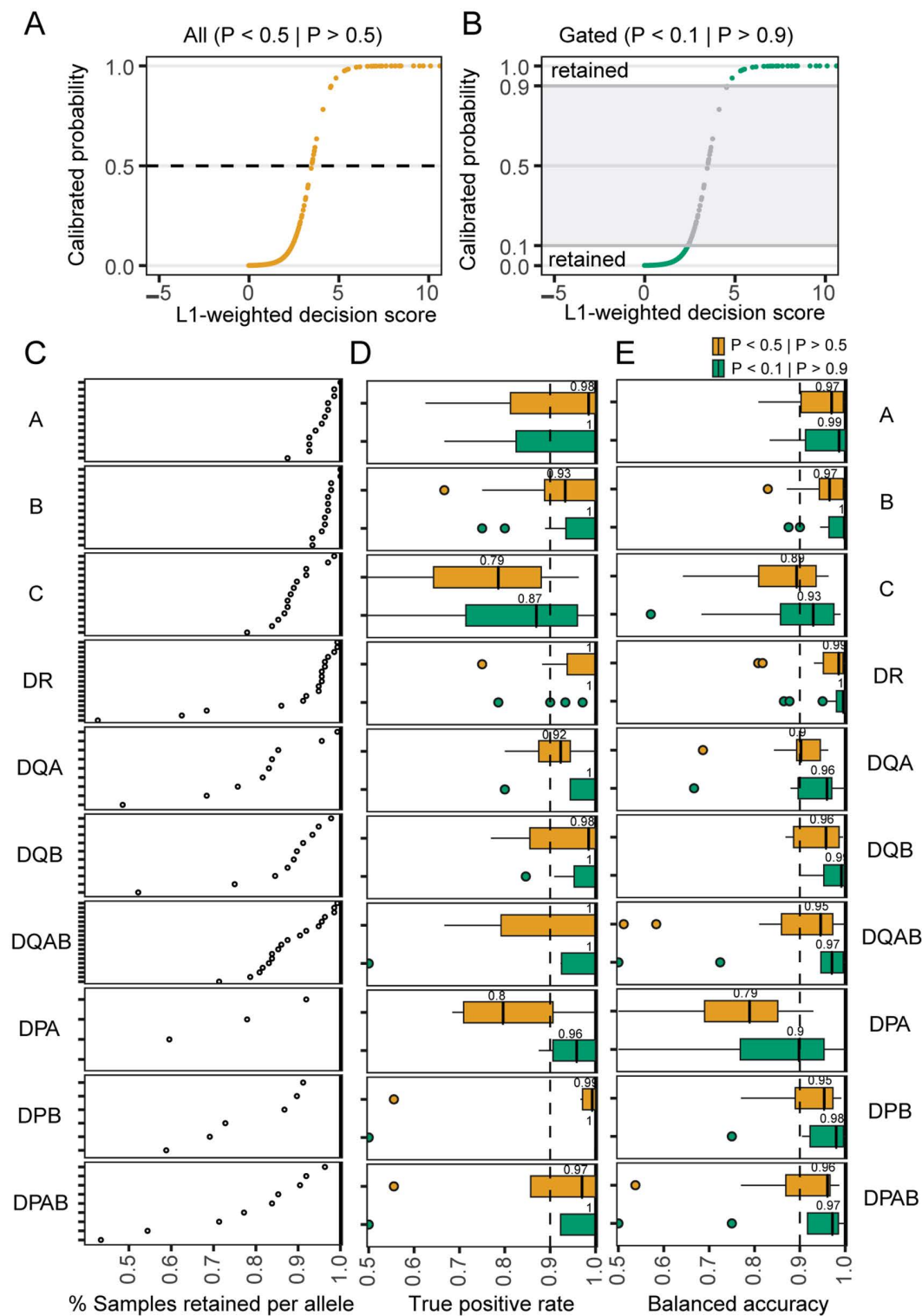


Fig 4. Confidence-based sample retention increases sensitivity and accuracy. (A) Representative score reliability curve, showing mapping of a classifier's raw decision scores to calibrated probabilities. When all data is considered (yellow) the prediction is based on a threshold probability of 0.5. (B) Curve with intermediate probability estimates masked ($0.1 > P < 0.9$) to retain high-confidence negative and positive predictions. (C) Fraction of samples retained per allele after confidence-based gating in the primary validation data (not used for feature discovery or feature weighting) from adults aged

18–65 with > 25,000 unique clones (n = 144). **(D)** True positive rate of all (yellow) and high-confidence (green) predictions. **(E)** Balanced accuracy for all (yellow) or high-confidence (green) predictions.

<https://doi.org/10.1371/journal.pcbi.1013767.g004>

exposure-associated CD4⁺ T cell receptors since, compared to CD8⁺ T cells, this population is less identifiable by longitudinal expansion analysis, tetramer staining, or activation-induced markers, with potential confoundment by bystander activation [27,28].

TCR2HLA models underperform when applied to TCR repertoires from children

The human TCR repertoire size and diversity change over the lifespan, in part, due to declining thymic output and clonal expansion of memory T cells [29,30], with diversity negatively associated with age and cytomegalovirus (CMV) exposure [31]. Thus, despite the utility of the TCR2HLA framework, we conducted sensitivity analyses to assess whether model performance varied by age. We hypothesized that predicting HLA status might be more difficult in children than adults due to higher repertoire diversity maintained by higher thymic output, fewer total exposures, and smaller T cell memory fractions in early life [32,33]. To test this, we evaluated classifier performance in childhood repertoires not used in training from the DAISY (Diabetes Autoimmunity Study in the Young) cohort described in studies by Mitchell et al. (2022, 2023), which included children aged 1–18, many diagnosed or at risk for type 1 diabetes [34,35]. Compared to observed performance in adults, balanced accuracy and AUCs were notably lower, even for common HLA alleles (representative results for HLA-B and HLA-DR, Fig 6). While models retained high specificity, low sensitivity led to significantly lower accuracy in children than adults (p-value < 0.0001), which was lowest in children under 5 years and increased with age. A recalibrated model, using the original TCR features and decision scores, but calibrated only with training data from participants under 13 years (U13 model), did not significantly improve childhood predictions relative to calibration with the full training dataset (p-value = 0.3). Adolescents (ages 13–18) showed intermediate accuracy compared to young children, suggesting key HLA-linked TCR features may arise due to antigen exposures during adolescence or early adulthood. Reduced overall TCR diversity in adult repertoires may also enhance feature detection, as memory clones become enriched relative to the naive T cell population. We also noted a trend of slightly reduced accuracy for HLA-B alleles in adults over 65 years, despite consistently high AUCs, potentially reflecting immunosenescence combined with potential calibration biases, as most of the training data came from individuals under age 65.

The relatively small number of repertoires from adults aged ≥65 years, limited our ability to assess performance in geriatric populations. While the overall predictive accuracy of our HLA-inference models was similar between adults and older adults, modest decreases in sensitivity were observed for some class I HLA alleles in geriatric repertoires. These differences may reflect both the limited representation of geriatric repertoires in the training and calibration data combined with biological changes in T-cell repertoire diversity associated with immune aging. The rate of age-related contraction of T-cell diversity could be uneven between CD4⁺ and CD8⁺ populations, which could partially explain the more pronounced decline in sensitivity we observed for some HLA-B (class I) relative to HLA-DR (class II) predictions in the small geriatric validation subset. Consistent with this interpretation, prior studies have shown that diverse CD4⁺ T cells may be maintained longer than CD8⁺ T cells, which undergo more significant age-related loss due chronic antigen exposure and more uneven homeostatic proliferation. [36,37]

Scope and accuracy of HLA genotype inference are limited by a lack of ethnic and geographic diversity

Because HLA allele frequencies and immunological exposures vary substantially by geographic region, we evaluated the limitations of using a primarily U.S.-based training dataset to predict HLA status of individuals from the Republic of South Africa (RSA), who differ partially in their HLA genotype diversity and environmental exposures (e.g., higher prevalence of HIV and tuberculosis [38]). To assess this, we removed all South African repertoires (n = 144) from feature discovery,

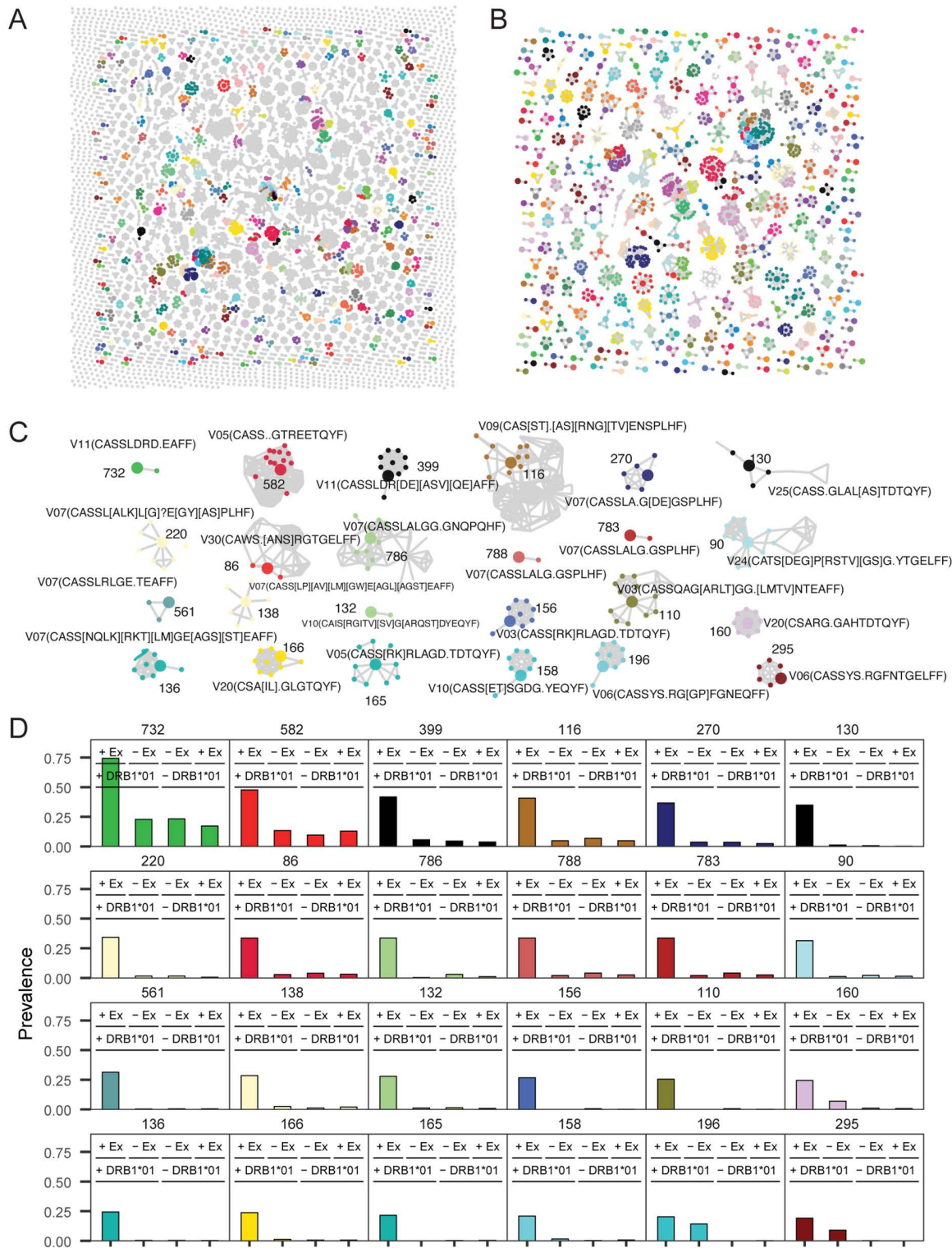


Fig 5. Identification of new HLA-associated TCRs in COVID-19-exposed repertoires. (A) Sequence similarity network of TRBV-CDR3 clonotypes with DRB1*01:01-associated edit-1 neighborhoods. Gray nodes represent clones within one mutation of a previously identified DRB1*01:01 reference clone used for HLA genotype inference with TCR2HLA. Colored nodes indicate potentially novel DRB1*01:01-associated TCRs that may reflect recent

SARS-CoV-2 exposure in this cohort (n = 1,355; Snyder et al.). **(B)** HLA-associated TCRs were clustered by sequence similarity ($\text{TCRdist} \leq 14$) and grouped into meta-clonotypes using a degree-based greedy graph-based algorithm. Large nodes represent anchor clones with the highest number of first-degree neighbors. **(C)** Twenty-four strongly HLA-associated meta-clonotypes, defined by CDR3 regular expression (regex) motifs and TRBV gene usage. Regex patterns were generated by aligning first-degree neighbors (smaller nodes) to their corresponding anchor TCRs (large nodes). **(D)** Relative prevalence of meta-clonotypes in exposed (+Ex, n = 1,355) and unexposed (-Ex, n = 1,977) repertoires, stratified by DRB1*01:01 genotype (+/- DRB1*01). Colors match the meta-clonotypes identified in previous panels.

<https://doi.org/10.1371/journal.pcbi.1013767.g005>

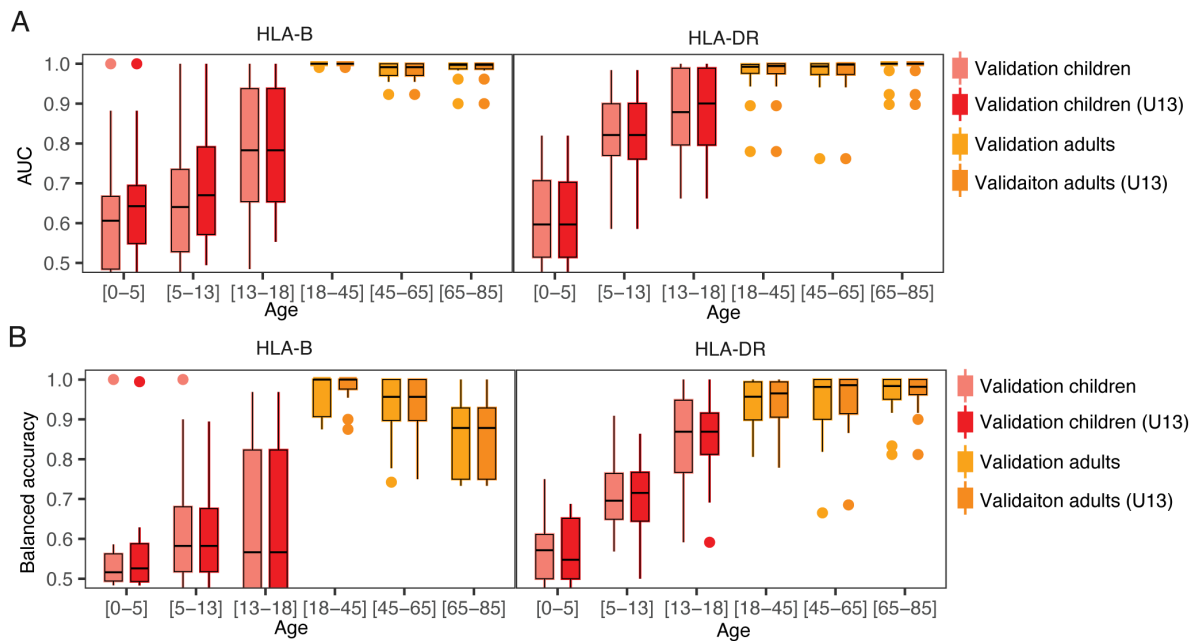


Fig 6. Sensitivity of HLA inference differs by age group. For HLA-B and HLA-DR alleles, **(A)** model AUROC performance separating HLA-genotype status of samples and **(B)** balanced accuracy predicting HLA-genotype. Probabilities were calibrated using either the full training data or only data from participants under 13 years of age (U13). External validation dataset from children longitudinally sampled (Mitchell et al.) and standard validation dataset (n = 142, ages 18-65 years, and n = 32, >65 years). Age category [0-5], [5-13], and [13-18] includes children under the age of 5, 13, and 18, respectively.

<https://doi.org/10.1371/journal.pcbi.1013767.g006>

selection, model weighting, and calibration steps, generating a U.S.-only model (xRSA). We then compared the predictions (S3A Fig) and predictive accuracy (S3B–S3C Fig) of the full model (trained on all data) and the xRSA model on the RSA repertoires. We observed that for several alleles common to both U.S. and RSA cohorts (e.g., A*01:01 and A*02:01), the xRSA model retained similar accuracy. While this comparison is imperfect, as models generally perform better against data used in training, we did observe that prediction accuracy and AUCs were reduced after removal of RSA training data for several common alleles, including A*11:01, A*24:02, B*44:03, C*07:01, C*07:02, C*16:01, DQA1*05:05, and DQB1*04:02 (S3B–S3C Fig). However, the AUROC values of the xRSA models indicate that raw decision scores still robustly separate HLA genotype-positive and -negative samples among RSA participants, suggesting that many of the selected TCR features are shared across geographic contexts. The reliability of the xRSA model declined relative to the full model, with more HLA-positive samples assigned moderate confidence scores ($0.4 < P(\text{HLA} = 1) < 0.9$). In some models, such as DQA1*05:01, excluding RSA repertoires from model training led to an increased number of false positives, possibly because the xRSA model failed to downweight TCRs that, in the RSA context, are associated with a different HLA allele not as commonly found in the U.S. population.

Despite reasonable performance for many common shared alleles, it is important to recognize that many HLA alleles common among the RSA participants were extremely rare or altogether absent in the U.S. training data (e.g., representative HLA-B alleles in [S3D Fig](#); HLA-DR in [S3E Fig](#)). For alleles present in 1–4% of the combined U.S.–R.S.A. dataset, we developed a rare-alleles model ([S2 Table](#)), but we excluded from consideration models for alleles with <1% prevalence. Models trained primarily on U.S. data provided less coverage of the full allelic diversity among South African participants. For example, the common-allele model set developed from the training data in this study included at least one HLA-A, -B, and -C allele for 95%, 87%, and 96% of U.S. participants, respectively – but only 76%, 54%, and 42% of R.S.A. participants ([Fig 7A](#)). Including rare allele models (which could not be externally validated) expanded this coverage (at least 1 allele per loci) to > 89% of participants in both U.S. and R.S.A. datasets across considered HLA-class I and HLA-class II loci ([Fig 7B](#)).

Discussion

In this study, we developed open-source software for the large-scale discovery and selection of HLA-associated TCR features. Our work builds on prior efforts to infer HLA genotypes from unlabeled TCR repertoire data. For example, Mayer-Blackwell et al. (2021) employed public HLA-associated TCR β sequences previously identified by DeWitt et al. (2018) to predict expression of common HLA alleles in unlabeled repertoires. The inferred HLA status of each repertoire was used to evaluate TCR meta-clonotypes and GLIPH motifs – learned from MIRA assays – for evidence of HLA restriction after SARS-CoV-2 exposure [[4,11,15,39](#)]. Ortega et al. (2025) used L1-regularized models trained on TCR α and TCR β features derived from Rosati et al. (2022), Russell (2022), and Emerson et al. (2017) datasets to predict participant HLA status [[9,19,40,41](#)]. Zahid et al. (2025) identified ~1 million HLA-associated public TCR clones from 4,144 genotyped individuals but have not released those features or an open-source predictive tool [[42](#)]. Recently, ElAbd et al. (2025,

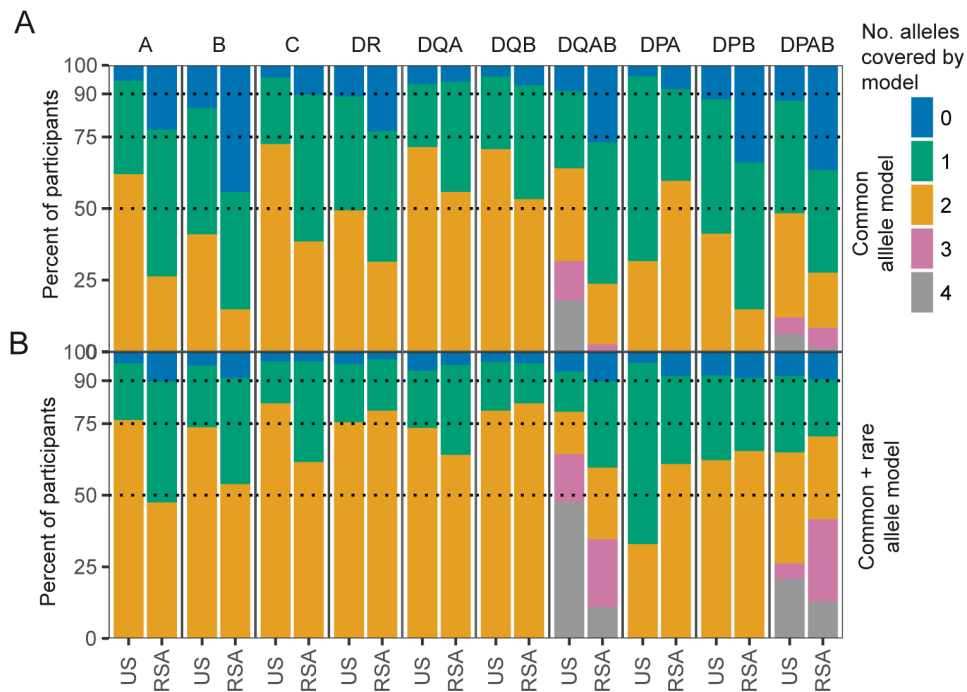


Fig 7. Model coverage of HLA-diversity differs by geographic region. (A) Common allele model coverage of alleles by locus in training data from the United States (US) and the Republic of South Africa (RSA). **(B)** Enhanced coverage by combined common and rare allele models.

<https://doi.org/10.1371/journal.pcbi.1013767.g007>

preprint) developed an HLA imputation framework using public HLA-associated features from 5,554 TCR β repertoires across cohorts from Germany, the U.S., and Norway, primarily involving individuals with Crohn's Disease and Ulcerative Colitis [10]. Pan et al. (2025) applied the GIANNA clustering algorithm [43,44] to publicly available HLA-genotyped datasets, identifying CDR3s from GIANNA clusters predictive of HLA genotype. While these efforts all make unique contributions, our study offers a scalable method for discovering HLA-association among related sets of similar TCRs across studies and demonstrates how well-calibrated HLA-genotype probabilities estimated from the occurrence of non-exact TCR features can be leveraged to discover new HLA-restricted and exposure-associated receptors.

In this study, we integrated more than 1,900 TCR β repertoires with HLA genotypes from a cohort comprising individuals with T1D, their close relatives without diabetes, and unrelated controls without diabetes [18] with 1,250 TCR β repertoires from three other previously available TCR β datasets to develop an open-source software for calibrated HLA genotype predictions. Expanding on Ortega et al.'s approach [9], we evaluated public feature tabulation strategies that use edit-1 neighborhoods instead of exact sequence matches. We first observed that “fuzzy” edit-1 tabulation reduced HLA specificity for many individual public HLA-enriched features with high probability of generation (TCR β formed from commonly used TRBV- and TRBJ-genes with CDR3 junctions more easily formed by the V(D)J recombination mechanism). Ortega et al.'s use of L1-penalized regression likely counteracted this through strong L1-penalized regression, effectively dropping fuzzy features lacking an HLA-specific signal. In contrast, in the present study, we fundamentally changed the way HLA-association testing was performed: rather than limiting enrichment tests to the neighborhoods of known public features, we extended *edit-1* tabulation into the feature discovery phase itself, testing millions of candidate neighborhoods across the multi-study cohorts. This allowed us to identify quasi-public TCRs with lower generation probabilities and robust HLA-specific enrichment that would be undetectable using exact-match methods alone.

To efficiently query neighborhoods among the ~70 million TCRs in our input datasets, we employed TRBV-family data partitioning and locality-sensitive hashing [45] to rapidly detect highly similar TRBV-CDR3 β sequences across thousands of repertoires. This approach parallels the GIANNA clustering strategy used by Pan et al. 2025 (preprint) [44], where vectorized CDR3 sequence embeddings were utilized for TCR clustering and HLA-enrichment testing. However, while the GIANNA algorithm initially clusters receptors using CDR3 information, the authors report that incorporating TRBV-gene information post-hoc boosted the HLA-signal [44]. In our study, we opted to define features at the TRBV-gene family level during the feature discovery phase to minimize batch artifacts that can arise at the TRBV-gene or allele level when combining datasets generated over time due to differences in amplification chemistries and advances in gene-calling pipelines.

A key aspect of our study was the evaluation of generalizability across heterogeneous datasets not used in feature discovery or model training, including repertoires varying from 5,000–450,000 unique clones. By calibrating prediction to stratify high- vs. low-confidence HLA genotype calls, we enabled efficient use of variable quality repertoires for downstream discovery of disease- or vaccine-specific HLA:TCR associations. Notably, the reduced sensitivity of *TCR2HLA* (high false negative rate) we observed in pediatric samples supports the hypothesis that enrichment of many public TCR-CDR3 features and neighborhoods results from common viral exposures and subsequent clonal expansion, rather than from imprinting by HLA proteins during thymic selection. This dependence on robust, exposure-driven response implies that *TCR2HLA*'s performance may decline in immunocompromised individuals or those with markedly different exposure histories or late-stage immunosenescence. Given that 96% of our training and calibration data came from younger individuals, caution should be exercised in interpretation of HLA inferences from geriatric (≥ 65 years old) repertoires.

We focused on inference of HLA genotypes from TCR β repertoires, because TCR β is the most commonly collected data type; however, the same approaches can be adapted to include TCR α occurrence alongside TCR β data as was shown by Ortega and colleagues [9]. The advent of new technologies for generating deep snapshots of TCR $\alpha\beta$ pairings [46] could be combined with single-chain HLA inference models to rapidly identify candidate TCR $\alpha\beta$ sequences' cognate HLA alleles to clone into cell lines for direct experimental validation. Future work might also take advantage of more sophisticated distance metrics to define a robust library of paired chain anchor features that could enhance HLA genotype

inference beyond what is possible with single chain data alone or reveal TCR β features that are conditionally linked to distinct HLA-alleles depending on α : β pairings.

As with other methods, *TCR2HLA* showed variable sensitivity across loci and by allele, motivating us to differentiate between high- and low-confidence predictions. Even when limiting to high-confidence imputations, a core limitation of the current *TCR2HLA* tool is that it only currently predicts presence and absence of HLA alleles with > 4% frequency in the U.S.-based training data. As such, *TCR2HLA* makes predictions for only a fraction of global HLA allelic variation. In sensitivity analyses, model performance on South African repertoires decreased for some alleles when these repertoires were withheld from the feature discovery and weighting steps, underscoring potential limitations to transporting models to ethnically divergent cohorts. For some HLA alleles common in both populations (e.g., A*02:01, B*07:02, DRB1*07:01), removal of South African repertoires did not substantially diminish predictive performance, most likely because many common human viruses circulate globally with associated TCR responses shared across geographically diverse populations.

Although Epstein-Barr virus (EBV)-, CMV-, and influenza-related TCRs appear in geographically distributed populations, HLA diversity could theoretically impact immunodominance hierarchies and alter population-specific TCR occurrence patterns even for globally common alleles. As with peptide-HLA binding prediction tools [47,48], limited allele representation for low- and medium-income countries in training data remains a barrier to model generalization [49]. Given the diagnostic value of HLA genotyping for maximizing insight from TCRseq data, expanding the geographical and genetic diversity of reference HLA-labeled TCR datasets is essential. The framework and software developed in this paper provide a scalable path forward for extending HLA genotype inference and feature discovery to new global datasets as they become available, as well as identifying novel exposure-elicited HLA-associated TCR features that could be leveraged as infectious disease, cancer, or autoimmune biomarkers.

Methods

HLA-labeled training data

To develop predictive models of HLA genotypes from peripheral blood TCR β repertoires, we assembled a training dataset of 3,125 repertoires from four studies (S1 Table) and constructed a bioinformatic pipeline to identify exact and near-exact (edit-distance 1) TCR features strongly associated with HLA alleles (Fig 1A-1C and S2 Table). The training data primarily included participants living in the United States from prior Emerson et al. (n=555), Elyanow et al. (n=276), and Rawat et al. (n=1,977) studies [18,19,21]. We further included a study with participants recruited in South Africa (n=144, repertoires=317) [20]. The largest study in the training set was conducted to study risk factors associated with type 1 diabetes (T1D; n=1,977); it includes individuals diagnosed with T1D as well as first-degree and second-degree relatives without T1D, and unrelated controls without T1D. Because certain class II alleles are strongly associated with the risk of developing T1D [50–52], the allelic prevalence in the training data is partially skewed from that of the general population of the United States and from other cohorts in the training data [53].

HLA alleles were inferred from genetic data at the 4-digit resolution. The methods for HLA-genotyping varied and are described in detail by study [4,18–21]. Briefly, participants in Rawat et al. were genotyped using the UFD1chip custom microarray, including more than 9000 markers covering the HLA genomic loci; genotypes were imputed from single nucleotide polymorphisms (SNPs) using Axiom HLA Analysis software (v1.2.0.38) [18]. Genotype data from Emerson et al. involved assembling genotypes from various methods as described by DeWitt et al. [4]. From the Emerson dataset, we restricted our analysis to 555 repertoires from participants with genotyping coverage across common class I and class II HLA alleles. High resolution HLA genotyping for HLA class I and class II loci for participants in Elyanow et al. and Towlerton et al. studies was performed using ScisGO HLA Typing kit v6 – designed for next-generation sequencing (NGS) [21,24,54]. Four-digit allele imputation accuracy of HLA-alleles from microarray-based SNP data varies by allele [55]; thus, we anticipated integrating this data with high-resolution NGS-derived HLA-genotype labels would improve the robustness of TCR β feature discovery.

Exact and near-exact HLA-associated feature discovery

Combining data across studies and HLA-genotyping methods, we attempted to identify robust cross-study HLA-associated TCR features, defined by TRBV family and CDR3 β amino acid (AA) sequence (TRBV-CDR3), based on statistical enrichment in the repertoires of participants with specific HLA alleles (4-digit resolution). HLA inference was framed as a series of separate allele-level binary classification tasks (e.g., HLA-A*02:01 positive vs. negative). To limit the feature space, we enriched for expanded T cells, using only TRBV-CDR3AA features observed in >1 template per repertoire. Sequences were partitioned by IMGT V-gene family. Edit-distance-1 (one substitution, insertion, or deletion) neighbors within V-family partitions were identified via locality-sensitive hashing – a computationally efficient method of sequence similarity search relying on an index of position-masked strings [45] related to other index-based methods for identifying neighboring CDR3 sequences [15,56]. TCR features found in ≥ 5 training set repertoires were tested for association with each common HLA-allele by Fisher’s Exact Test (*fishersapi*), with strong associations defined by (edit-1, odds ratio > 1, $p < 10^{-8}$, and absent in >90% HLA-mismatched repertoires).

For TRBV-CDR3 features with a strong HLA-association based on edit-1 detections, we constructed sequence similarity graphs with edges connecting observed TCRs within $\text{TCRdist} < 14$ (*tcrdist3* and *networkx*) [11,14]. To remove potentially spurious features (e.g., TCRs with HLA-enriched neighborhoods, but no individual HLA association), we pruned nodes that lacked any enrichment on an exact sequence basis (odds ratio < 1). To further reduce redundancy among remaining HLA-associated features, we applied a greedy algorithm to select “anchor” TCRs to retain for downstream prediction tasks: within each sequence similarity graph, the iterative algorithm selected the highest-degree node as a TCR anchor, then removed it and its neighbors from the graph, before iteratively selecting the next most densely connected anchor, until no nodes of degree > 1 remained. We also identified exact TRBV-family+CDR3 features: public HLA-associated features found across the four studies ($p < 10^{-8}$, odds ratio > 1) as well as those previously identified from the Emerson et al. cohort ($p < 10^{-6}$) as previously reported by Dewitt et al. [4].

Feature selection weighting

Both highly public TCRs and sparsely public TCRs (with sufficiently public sequence neighborhoods to power an association) could be combined to build a binary classifier for each allele (S1A Fig). We trained binary classifiers combining exact and edit-1 anchor features, estimating coefficients with L1-regularized logistic regression with inverse probability weighting. Five-fold cross-validation was used to tune regularization to achieve maximal area under the receiver operating characteristic curve (AUROC). Across all alleles, both edit-0 and edit-1 features were retained and assigned feature weights (S1 Fig and S3 Table). There was a modest positive correlation between the strength of the statistical HLA-association in the feature discovery phase and the logistic regression-derived model weights (S1C Fig). To address severe class imbalance, we used inverse class probability weighting in the model fitting stage, and post-estimation calibration by Platt’s method to refine predicted probabilities (Fig 1D). For application to samples with low numbers of unique clonotypes (5-25K), we implemented a high-sensitivity (HS) recalibration, using a second-degree polynomial adjustment based on the number of unique clones (\log_{10} -transformed) with down sampled synthetic data.

For each allele, after selecting highly HLA-enriched TCRs across four cohorts, we trained models one allele at a time, based on a set of n informative exact TCR β and a set of m near-exact anchor TCR β features. Each sample is thus represented as a vector $\mathbf{x} \in \{0, 1\}^{n+m}$ composed of two binary subvectors:

- $\mathbf{x}_0 \in \{0, 1\}^n$: indicates if a sample contains at least one clonotype at productive frequency $> 2E^{-6}$ matching that HLA-associated exact feature.
- $\mathbf{x}_1 \in \{0, 1\}^m$: indicates if a sample contains at least one clonotype at productive frequency $> 2E^{-6}$ within one mutation (substitution, insertion, or deletion) of an HLA-associated near-exact anchor feature.

Such that:

$$\mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

The probability that a repertoire comes from a person expressing an allelic variant (y) is expressed as:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-(\alpha_0 + \mathbf{x}_0^\top \beta_0 + \mathbf{x}_1^\top \beta_1))}$$

We used regularization to estimate a parsimonious set of parameters that maximizes the L1-penalized log-likelihood.

$$l_{\text{reg}}(\beta) = \sum_{i=1}^n [y_i \log \sigma(\mathbf{x}_i^\top \beta) + (1 - y_i) \log (1 - \sigma(\mathbf{x}_i^\top \beta))] - \lambda \|\beta\|_1$$

To account for extreme class imbalance, important for predicting the presence of rarer alleles, inverse probability weighting (IPW) was applied during model fitting.

$$w_i = \begin{cases} \frac{1}{n_1} & \text{if } y_i = 1 \\ \frac{1}{n_0} & \text{if } y_i = 0 \end{cases}$$

Where n_1 and n_0 are the number of allele-positive and allele-negative repertoires in the training data, respectively. The IP-weighted L1-regularized log-likelihood is:

$$l_{\text{weighted}}(\beta) = \sum_{i=1}^n w_i [y_i \log \sigma(\mathbf{x}_i^\top \beta) + (1 - y_i) \log (1 - \sigma(\mathbf{x}_i^\top \beta))] - \lambda \|\beta\|_1$$

This objective function ensures that both classes contribute during training, even under severe class imbalance. The L1-regularized logistic regression with feature weights β and regularization parameter λ learned by maximizing AUROC across 5-fold cross-validation, yields a decision score.

$$f(\mathbf{x}) = \mathbf{x}^\top \beta$$

However, after fitting with inverse probability weighting and regularization, the decision score may not be well-calibrated as a probability. Thus, Platt scaling fits a sigmoid to the raw decision function:

$$\hat{P}(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-(A f(\mathbf{x}) + B))}$$

For this standard calibration, parameters A and B were then fit on the full training set. For the high-sensitivity calibration – to account for low sample diversity < 25,000 unique clones – we include an additional adjustment using the covariate $u_i = \log_{10}(U_i)$, where U_i is the number of unique clonotypes (with frequency > $2E^{-6}$) in sample i . The calibrated probability is thus derived by adjusting the decision score as follows:

$$\hat{P}(y = 1|\mathbf{x}, u) = \frac{1}{1 + \exp(-(A f(\mathbf{x}) + B + C_1 u + C_2 u^2))}$$

Where the parameters C_1 and C_2 are fit simultaneously with A and B on the repeatedly down sampled draws from the training set, with the decision function extended to include the first- and second-degree polynomial terms. The synthetic training data was generated from 650 training data repertoires repeatedly sampled with replacement for 1, 5, 10, 25, 50, 100, and 500 thousand draws.

Assessing generalizable model performance via external validation

We tested model generalizability using several validation cohorts that were not used for feature discovery or model calibration. The primary validation, cohort 3 in Rawat et al., included 142 adults (18–65 years) unrelated to participants in the T1D cohorts 1,2 with >25,000 unique TCRs per sample. [18] We performed additional validation using a smaller COVID-19 cohort (Nolan et al., 63 repertoires), a dataset including low-diversity repertoires in HIV positive individuals (Towlerton et al., 188 repertoires), and an mRNA-based 5'-race derived TCRseq (Rosati et al., 112 repertoires) [18,24,26,40].

Models were built for common alleles (>4% prevalence in training data; Fig 2A and S2 Table, model v4) to ensure sufficient training and validation data were available to robustly examine allele and locus-specific predictive performance: HLA-A (n = 12 alleles), HLA-B (n = 12), HLA-C (n = 13), HLA-DR (n = 11), HLA-DQA (n = 10), HLA-DQB (n = 10), HLA-DPA (n = 4), HLA-DPB (n = 6). Separately, we also trained models for rare alleles (1–4% prevalence, S2 Table, model v7), though evaluation of these models' performance was limited due to insufficient validation data: HLA-A (n = 7 alleles), HLA-B (n = 21), HLA-C (n = 6), HLA-DR (n = 19), HLA-DQA (n = 2), HLA-DQB (n = 5), HLA-DPA (n = 1), HLA-DPB (n = 11). Due to inconsistent occurrence and genotype availability, we did not focus on TCRs related to DRB3, DRB4, or DRB5 loci.

Meta-clonotype discovery from unlabeled repertoires

We applied TCR2HLA to infer HLA genotypes for 1,355 previously unlabeled TCR repertoires from individuals with acute and convalescent COVID-19 or high-risk exposure [26]. For common HLA alleles, genotype predictions were retained when confidence was high (calibrated probability > 0.9 for presence or < 0.1 for absence). We then used TRBV-family partitioning and locality-sensitive hashing to identify TCRs that shared HLA-associated edit-1 neighborhoods in the COVID-19 cohort. To identify potentially novel, cohort-specific TCRs, we excluded receptors highly similar to the reference TCRs used to make HLA inferences, defined as sharing the same TRBV family and differing by no more than one AA substitution, insertion, or deletion in the CDR3 region.

To reduce feature redundancy, we grouped newly discovered HLA-associated TCRs into meta-clonotypes based on strong statistical associations with HLA alleles ($OR > 1$, $p < 1e-8$), after excluding TCRs from individuals lacking the relevant HLA genotype. For each allele, we constructed sequence similarity networks by connecting TCR β sequences with pairwise $TCRdist \leq 14$ (typically a strict similarity arising from matching TRBV gene-usage and 1–2 CDR3 amino acid substitutions). A degree-based greedy algorithm was used to identify anchor TCRs – clones with the most first-degree neighbors – and define meta-clonotypes as anchor-neighbor clusters. Each meta-clonotype was defined by a regular expression (regex) motif derived from aligned CDR3 sequences [11,57] and annotated with the TRBV family of the anchor clone. We then quantified the occurrence of TCRs matching each meta-clonotype definition across the Rawat et al. (pre-COVID) and Snyder et al. (COVID-exposed) cohorts to evaluate prevalence by HLA genotype and exposure status [18,26].

Supporting information

S1 Table. Validation and training datasets.

(XLSX)

S2 Table. Strongly HLA-associated features used to train binary classifiers.

(XLSX)

S3 Table. Feature count and sum of positive and negative weights by allele after L1 regression.

(XLSX)

S4 Table. Age and TCR diversity by validation and training dataset.

(XLSX)

S5 Table. Allele prevalence by dataset.

(XLSX)

S6 Table. External validation performance by allele and dataset.

(XLSX)

S7 Table. Prediction metrics across alleles for primary validation and Towlerston datasets.

(XLSX)

S1 Fig. Probability of generation estimates and L1-regularized feature weights by TCR feature type.

(PDF)

S2 Fig. Balanced accuracy on different validation datasets.

(PDF)

S3 Fig. HLA-inference in a geographically distinct population.

(PDF)

Acknowledgments

We thank Dr. Daniel Geraghty for providing high-resolution HLA-genotyping participants in the Elyanow et al. study, Ying Lei for helpful discussion about a memory-efficient algorithm for finding mutational variants, and Dr. Liel Cohen-Lavi for helpful early discussion on HLA feature selection. The authors used an LLM (Microsoft Copilot and ChatGPT-4) to assist with proofreading and reduction in word length for sections of the manuscript text. Proofread versions of the manuscript were compared with the authors' original text using Microsoft Word's compare documents tool. All proposed grammatical changes were manually reviewed and accepted or rejected. All scientific content, citations, and interpretations are the authors'.

Author contributions

Conceptualization: Koshlan Mayer-Blackwell, Victor Greiff, Philip Bradley, Paul G. Thomas, Andrew Fiore-Gartland.

Data curation: Koshlan Mayer-Blackwell, Puneet Rawat, Melanie R. Shapiro, Leeana D. Peters, Emily S. Ford, David M. Koelle.

Formal analysis: Koshlan Mayer-Blackwell, Anastasia Minervina, Mikhail Pogorelyy, Andrew Fiore-Gartland.

Funding acquisition: David M. Koelle, Victor Greiff, Philip Bradley, Todd M. Brusko, Paul G. Thomas.

Investigation: Koshlan Mayer-Blackwell.

Methodology: Koshlan Mayer-Blackwell, Anastasia Minervina, Mikhail Pogorelyy, Andrew Fiore-Gartland.

Software: Koshlan Mayer-Blackwell, Samuel Minot.

Supervision: Victor Greiff, Philip Bradley, Todd M. Brusko, Paul G. Thomas, Andrew Fiore-Gartland.

Validation: Koshlan Mayer-Blackwell, Kasi Vegesana.

Visualization: Koshlan Mayer-Blackwell.

Writing – original draft: Koshlan Mayer-Blackwell, Anastasia Minervina, Andrew Fiore-Gartland.

Writing – review & editing: Koshlan Mayer-Blackwell, Anastasia Minervina, Mikhail Pogorelyy, Puneet Rawat, Melanie R. Shapiro, Leeana D. Peters, Emily S. Ford, Amanda L. Posgai, Kasi Vegesana, Samuel Minot, David M. Koelle, Victor Greiff, Philip Bradley, Todd M. Brusko, Paul G. Thomas, Andrew Fiore-Gartland.

References

- O'Donnell TJ, Kanduri C, Isacchini G, Limenitakis JP, Brachman RA, Alvarez RA, et al. Reading the repertoire: Progress in adaptive immune receptor analysis using machine learning. *Cell Syst.* 2024;15(12):1168–89. <https://doi.org/10.1016/j.cels.2024.11.006> PMID: [39701034](https://pubmed.ncbi.nlm.nih.gov/39701034/)
- Mhanna V, Bashour H, Lê Quý K, Barennes P, Rawat P, Greiff V, et al. Adaptive immune receptor repertoire analysis. *Nat Rev Methods Primers.* 2024;4(1). <https://doi.org/10.1038/s43586-023-00284-1>
- Elhanati Y, Sethna Z, Callan CG Jr, Mora T, Walczak AM. Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunol Rev.* 2018;284(1):167–79. <https://doi.org/10.1111/imr.12665> PMID: [29944757](https://pubmed.ncbi.nlm.nih.gov/29944757/)
- DeWitt WS 3rd, Smith A, Schoch G, Hansen JA, Matsen FA 4th, Bradley P. Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *Elife.* 2018;7:e38358. <https://doi.org/10.7554/eLife.38358> PMID: [30152754](https://pubmed.ncbi.nlm.nih.gov/30152754/)
- May DH, Woodhouse S, Howie B, Robins HS. A Catalog of the Public T-cell Response to Cytomegalovirus. Cold Spring Harbor Laboratory. 2024. <https://doi.org/10.1101/2024.05.08.593237>
- Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC. Structure of the human class I histocompatibility antigen, HLA-A2. *Nature.* 1987;329(6139):506–12. <https://doi.org/10.1038/329506a0> PMID: [3309677](https://pubmed.ncbi.nlm.nih.gov/3309677/)
- 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(6432):33–9. <https://doi.org/10.1038/364033a0> PMID: [8316295](https://pubmed.ncbi.nlm.nih.gov/8316295/)
- Neefjes J, Jongsma MLM, Paul P, Bakke O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat Rev Immunol.* 2011;11(12):823–36. <https://doi.org/10.1038/nri3084> PMID: [22076556](https://pubmed.ncbi.nlm.nih.gov/22076556/)
- Ruiz Ortega M, 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(1):e1012724. <https://doi.org/10.1371/journal.pcbi.1012724> PMID: [39761303](https://pubmed.ncbi.nlm.nih.gov/39761303/)
- EIAbd H, Mahdy AKH, Wacker EM, Gretsova M, Ellinghaus D, Franke A. Decoding the restriction of T cell receptors to human leukocyte antigen alleles using statistical learning. Cold Spring Harbor Laboratory. 2025. <https://doi.org/10.1101/2025.02.06.636910>
- Mayer-Blackwell K, Schattgen S, Cohen-Lavi L, Crawford JC, Souquette A, Gaevart JA, et al. TCR meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public, HLA-restricted clusters of SARS-CoV-2 TCRs. *Elife.* 2021;10:e68605. <https://doi.org/10.7554/eLife.68605> PMID: [34845983](https://pubmed.ncbi.nlm.nih.gov/34845983/)
- Huang H, Wang C, Rubelt F, Scriba TJ, Davis MM. Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. *Nat Biotechnol.* 2020;38(10):1194–202. <https://doi.org/10.1038/s41587-020-0505-4> PMID: [32341563](https://pubmed.ncbi.nlm.nih.gov/32341563/)
- Mayer A, Callan CG Jr. Measures of epitope binding degeneracy from T cell receptor repertoires. *Proc Natl Acad Sci U S A.* 2023;120(4):e2213264120. <https://doi.org/10.1073/pnas.2213264120> PMID: [36649423](https://pubmed.ncbi.nlm.nih.gov/36649423/)
- Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature.* 2017;547(7661):89–93. <https://doi.org/10.1038/nature22383> PMID: [28636592](https://pubmed.ncbi.nlm.nih.gov/28636592/)
- Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, et al. Identifying specificity groups in the T cell receptor repertoire. *Nature.* 2017;547(7661):94–8. <https://doi.org/10.1038/nature22976>
- Mark M, Reich-Zeliger S, Greenstein E, Biram A, Chain B, Friedman N, et al. Viral infection reveals hidden sharing of TCR CDR3 sequences between individuals. *Front Immunol.* 2023;14:1199064. <https://doi.org/10.3389/fimmu.2023.1199064> PMID: [37325645](https://pubmed.ncbi.nlm.nih.gov/37325645/)
- Marcou Q, Mora T, Walczak AM. High-throughput immune repertoire analysis with IGoR. *Nat Commun.* 2018;9(1). <https://doi.org/10.1038/s41467-018-02832-w>
- Rawat P, Shapiro MR, Peters LD, Widrich M, Mayer-Blackwell K, Motwani K, et al. Identification of a type 1 diabetes-associated T cell receptor repertoire signature from the human peripheral blood. Cold Spring Harbor Laboratory. 2024. <https://doi.org/10.1101/2024.12.10.24318751>
- Emerson RO, DeWitt WS, Vignali M, Gravley J, Hu JK, Osborne EJ, et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet.* 2017;49(5):659–65. <https://doi.org/10.1038/ng.3822>
- Musvosvi M, Huang H, Wang C, Xia Q, Rozot V, Krishnan A, et al. T cell receptor repertoires associated with control and disease progression following Mycobacterium tuberculosis infection. *immuneACCESS. Adaptive Biotechnologies.* <https://doi.org/10.21417/mm2022nm>
- Elyanow R, Snyder TM, Dalai SC, Gittelman RM, Boonyaratankornkit J, Wald A, et al. T cell receptor sequencing identifies prior SARS-CoV-2 infection and correlates with neutralizing antibodies and disease severity. *JCI Insight.* 2022;7(10):e150070. <https://doi.org/10.1172/jci.insight.150070> PMID: [35439166](https://pubmed.ncbi.nlm.nih.gov/35439166/)
- Petersdorf EW. HLA structure and function in hematopoietic-cell transplantation. *Best Pract Res Clin Haematol.* 2024;37(3):101564. <https://doi.org/10.1016/j.beha.2024.101564> PMID: [39396254](https://pubmed.ncbi.nlm.nih.gov/39396254/)

23. Petersdorf EW, Bengtsson M, Horowitz M, McKallor C, Spellman SR, Spierings E, et al. HLA-DQ heterodimers in hematopoietic cell transplantation. *Blood*. 2022;139(20):3009–17. <https://doi.org/10.1182/blood.2022015860>
24. Towler AMH, Ravishankar S, Coffey DG, Puroton CE, Warren EH. Serial Analysis of the T-Cell Receptor β -Chain Repertoire in People Living With HIV Reveals Incomplete Recovery After Long-Term Antiretroviral Therapy. *Front Immunol*. 2022;13:879190. <https://doi.org/10.3389/fimmu.2022.879190> PMID: 35585986
25. Okoye AA, Picker LJ. CD4(+) T-cell depletion in HIV infection: mechanisms of immunological failure. *Immunol Rev*. 2013;254(1):54–64. <https://doi.org/10.1111/imr.12066> PMID: 23772614
26. 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. Cold Spring Harbor Laboratory. 2020. <https://doi.org/10.1101/2020.07.31.20165647>
27. Ford ES, Mayer-Blackwell K, Jing L, Laing KJ, Sholukh AM, St Germain R, et al. Repeated mRNA vaccination sequentially boosts SARS-CoV-2-specific CD8+ T cells in persons with previous COVID-19. *Nat Immunol*. 2024;25(1):166–77. <https://doi.org/10.1038/s41590-023-01692-x> PMID: 38057617
28. Zheng MZM, Burmas L, Tan H-X, Trieu M-C, Lee HJ, Rawlinson D, et al. Deconvoluting TCR-dependent & -independent activation is vital for reliable Ag-specific CD4+T cell characterization by AIM assay. Cold Spring Harbor Laboratory. 2024. <https://doi.org/10.1101/2024.12.10.627643>
29. Britanova OV, Shugay M, Merzlyak EM, Staroverov DB, Putintseva EV, Turchaninova MA, et al. Dynamics of Individual T Cell Repertoires: From Cord Blood to Centenarians. *J Immunol*. 2016;196(12):5005–13. <https://doi.org/10.4049/jimmunol.1600005> PMID: 27183615
30. Kumar BV, Connors TJ, Farber DL. Human T Cell Development, Localization, and Function throughout Life. *Immunity*. 2018;48(2):202–13. <https://doi.org/10.1016/j.immuni.2018.01.007> PMID: 29466753
31. Krishna C, Chowell D, Gönen M, Elhanati Y, Chan TA. Genetic and environmental determinants of human TCR repertoire diversity. *Immun Ageing*. 2020;17:26. <https://doi.org/10.1186/s12979-020-00195-9> PMID: 32944053
32. Naylor K, Li G, Vallejo AN, Lee W-W, Koetz K, Bryl E, et al. The influence of age on T cell generation and TCR diversity. *J Immunol*. 2005;174(11):7446–52. <https://doi.org/10.4049/jimmunol.174.11.7446> PMID: 15905594
33. Douek DC, McFarland RD, Keiser PH, Gage EA, Massey JM, Haynes BF, et al. Changes in thymic function with age and during the treatment of HIV infection. *Nature*. 1998;396(6712):690–5. <https://doi.org/10.1038/25374> PMID: 9872319
34. Mitchell AM, Alkanani AA, McDaniel KA, Pyle L, Waugh K, Steck AK, et al. T-cell responses to hybrid insulin peptides prior to type 1 diabetes development. *Proc Natl Acad Sci U S A*. 2021;118(6):e2019129118. <https://doi.org/10.1073/pnas.2019129118> PMID: 33542101
35. Mitchell AM, Baschal EE, McDaniel KA, Simmons KM, Pyle L, Waugh K, et al. Temporal development of T cell receptor repertoires during childhood in health and disease. *JCI Insight*. 2022;7(18):e161885. <https://doi.org/10.1172/jci.insight.161885> PMID: 35998036
36. Liu Z, Liang Q, Ren Y, Guo C, Ge X, Wang L, et al. Immunosenescence: molecular mechanisms and diseases. *Signal Transduct Target Ther*. 2023;8(1):200. <https://doi.org/10.1038/s41392-023-01451-2> PMID: 37179335
37. Czesnikiewicz-Guzik M, Lee W-W, Cui D, Hiruma Y, Lamar DL, Yang Z-Z, et al. T cell subset-specific susceptibility to aging. *Clinical Immunology*. 2008;127(1):107–18. <https://doi.org/10.1016/j.clim.2007.12.002>
38. Abdool Karim SS, Churchyard GJ, Karim QA, Lawn SD. HIV infection and tuberculosis in South Africa: an urgent need to escalate the public health response. *Lancet*. 2009;374(9693):921–33. [https://doi.org/10.1016/S0140-6736\(09\)60916-8](https://doi.org/10.1016/S0140-6736(09)60916-8) PMID: 19709731
39. Nolan S, Vignali M, Klinger M, Dines JN, Kaplan IM, Svejnova E, et al. A large-scale database of T-cell receptor beta sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. *Front Immunol*. 2025;16:1488851. <https://doi.org/10.3389/fimmu.2025.1488851> PMID: 40034696
40. Rosati E, Rios Martini G, Pogorelyy MV, Minervina AA, Degenhardt F, Wendorff M, et al. A novel unconventional T cell population enriched in Crohn's disease. *Gut*. 2022;71(11):2194–204. <https://doi.org/10.1136/gutjnl-2021-325373> PMID: 35264446
41. Russell ML, Souquette A, Levine DM, Schattgen SA, Allen EK, Kuan G, et al. Combining genotypes and T cell receptor distributions to infer genetic loci determining V(D)J recombination probabilities. *Elife*. 2022;11:e73475. <https://doi.org/10.7554/eLife.73475> PMID: 35315770
42. Zahid HJ, Taniguchi R, Ebert P, Chow I-T, Gooley C, Lv J, et al. Large-scale statistical mapping of T-cell receptor β sequences to human leukocyte antigens. *Front Immunol*. 2025;16:1603730. <https://doi.org/10.3389/fimmu.2025.1603730> PMID: 40936899
43. Zhang H, Zhan X, Li B. GIANA allows computationally-efficient TCR clustering and multi-disease repertoire classification by isometric transformation. *Nat Commun*. 2021;12(1):4699. <https://doi.org/10.1038/s41467-021-25006-7> PMID: 34349111
44. Pan M, Tan Y, Wang YT, Hu J, Fleming J, Hu H, et al. Quantitative and large-scale investigation of human TCR-HLA cross-reactivity. *Sci Adv*. 2025;11(43):eadx1751. <https://doi.org/10.1126/sciadv.adx1751> PMID: 41134880
45. Har-Peled S, Indyk P, Motwani R. *Theory of Comput*. 2012;8(1):321–50. <https://doi.org/10.4086/toc.2012.v008a014>
46. Pogorelyy MV, Kirk AM, Adhikari S, Minervina AA, Sundararaman B, Vegesana K, et al. TIRTL-seq: Deep, quantitative, and affordable paired TCR repertoire sequencing. *bioRxiv*. 2024;:2024.09.16.613345. <https://doi.org/10.1101/2024.09.16.613345> PMID: 39345544
47. Conev A, Fasoulis R, Hall-Swan S, Ferreira R, Kaviraki LE. HLAequity: Examining biases in pan-allele peptide-HLA binding predictors. *iScience*. 2023;27(1):108613. <https://doi.org/10.1016/j.isci.2023.108613> PMID: 38188519
48. Glynn E, Ghersi D, Singh M. Toward equitable major histocompatibility complex binding predictions. *Proc Natl Acad Sci U S A*. 2025;122(8):e2405106122. <https://doi.org/10.1073/pnas.2405106122> PMID: 39964728

49. Peng K, Safonova Y, Shugay M, Popejoy AB, Rodriguez OL, Breden F, et al. Diversity in immunogenomics: the value and the challenge. *Nat Methods*. 2021;18(6):588–91. <https://doi.org/10.1038/s41592-021-01169-5> PMID: [34002093](https://pubmed.ncbi.nlm.nih.gov/34002093/)
50. Todd JA, Bell JI, McDevitt HO. HLA-DQ beta gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. *Nature*. 1987;329(6140):599–604. <https://doi.org/10.1038/329599a0> PMID: [3309680](https://pubmed.ncbi.nlm.nih.gov/3309680/)
51. Herold KC, DeLong T, Perdigo AL, Biru N, Brusko TM, Walker LSK. The immunology of type 1 diabetes. *Nat Rev Immunol*. 2024;24(6):435–51. <https://doi.org/10.1038/s41577-023-00985-4> PMID: [38308004](https://pubmed.ncbi.nlm.nih.gov/38308004/)
52. Noble JA. Fifty years of HLA-associated type 1 diabetes risk: history, current knowledge, and future directions. *Front Immunol*. 2024;15:1457213. <https://doi.org/10.3389/fimmu.2024.1457213> PMID: [39328411](https://pubmed.ncbi.nlm.nih.gov/39328411/)
53. Erlich H, Valdes AM, Noble J, Carlson JA, Varney M, Concannon P, et al. HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk: analysis of the type 1 diabetes genetics consortium families. *Diabetes*. 2008;57(4):1084–92. <https://doi.org/10.2337/db07-1331> PMID: [18252895](https://pubmed.ncbi.nlm.nih.gov/18252895/)
54. Nelson WC, Pyo C-W, Vogan D, Wang R, Pyon Y-S, Hennessey C, et al. An integrated genotyping approach for HLA and other complex genetic systems. *Human Immunology*. 2015;76(12):928–38. <https://doi.org/10.1016/j.humimm.2015.05.001>
55. Perry DJ, Shapiro MR, Chamberlain SW, Kusmartseva I, Chamala S, Balzano-Nogueira L, et al. A genomic data archive from the Network for Pancreatic Organ donors with Diabetes. *Sci Data*. 2023;10(1):323. <https://doi.org/10.1038/s41597-023-02244-6> PMID: [37237059](https://pubmed.ncbi.nlm.nih.gov/37237059/)
56. Chotisorayuth T, Tiffeau-Mayer A. Lightning-fast adaptive immune receptor similarity search by symmetric deletion lookup. *arXiv*. 2024. <https://doi.org/10.48550/arXiv.2403.09010>
57. Turner CT, Tiffeau-Mayer A, Rosenheim J, Chandran A, Saxena R, Zhang P, et al. Evolution of T cell responses in the tuberculin skin test reveals generalisable Mtb-reactive T cell metaclones. *Cold Spring Harbor Laboratory*. 2025. <https://doi.org/10.1101/2025.04.12.648537>