TCR meta-clonotypes for biomarker discovery with tcrdist3: quantification of public, HLA-restricted TCR biomarkers of SARS-CoV-2 infection

Koshlan Mayer-Blackwell[1], Stefan Schattgen[2], Liel Cohen-Lavi[3,4], Jeremy Chase Crawford[2], Aisha Souquette[2], Jessica A. Gaevert[2,5], Tomer Hertz[6], Paul G. Thomas[2], Philip Bradley[7], Andrew Fiore-Gartland[1]

[1] Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, USA

[2] Immunology Department, St. Jude Children's Research Hospital, Memphis, USA

[3] Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Be'er-Sheva, Israel

[4] National Institute for Biotechnology in the Negev, Ben-Gurion University of the Negev, Be'er-Sheva, Israel

[5] St. Jude Graduate School of Biomedical Sciences, St. Jude Children's Research Hospital, Memphis, USA

[6] Shraga Segal Department of Microbiology and Immunology, Ben-Gurion University of the Negev, Be'er Sheva, Israel

[7] Public Health Science Division, Fred Hutchinson Cancer Research Center, Seattle, USA

ABSTRACT

As the mechanistic basis of adaptive cellular antigen recognition, T cell receptors (TCRs) encode clinically valuable information that reflects prior antigen exposure and potential future response. However, despite advances in deep repertoire sequencing, enormous TCR diversity complicates the use of TCR clonotypes as clinical biomarkers. We propose a new framework that leverages antigen-enriched repertoires to form meta-clonotypes – groups of biochemically similar TCRs – that can be used to robustly quantify functionally similar TCRs in bulk repertoires. We apply the framework to TCR data from COVID-19 patients, generating 1,915 public TCR meta-clonotypes from the 18 SARS-CoV-2 antigen-enriched repertoires with the strongest evidence of HLA-restriction. Applied to independent cohorts, meta-clonotypes targeting these specific epitopes were more frequently detected in bulk repertoires compared to exact amino acid matches, and 44% (845/1915) were significantly enriched among COVID-19 patients that expressed the putative restricting HLA allele, demonstrating the potential utility of meta-clonotypes as antigen-specific features for biomarker development. To enable further applications, we developed an open-source software package, *tcrdist3,* that implements this framework and facilitates workflows for distance-based TCR repertoire analysis.

INTRODUCTION

An individual's unique repertoire of T cell receptors (TCRs) is shaped by antigen exposure and is a critical component of immunological memory, contributing to recall responses against future infectious challenges (Emerson et al., 2017; Welsh and Selin, 2002). With the advancement of immune repertoire profiling, TCR repertoires are a largely untapped source of biomarkers that could potentially be used to predict immune responses to a wide range of exposures including microbial infections (Wolf et al., 2018), tumor neoantigens (Ahmadzadeh et al., 2019; Kato et al., 2018), or environmental allergens (Cao et al., 2020). The TCR repertoire is characterized by its extreme diversity, originating from the genomic V(D)J gene recombination of receptors in development. Between 10^9-10^{10} unique clonotypes - T cells with distinct nucleotide-encoded receptors - are maintained in an adult human TCR repertoire (Lythe et al., 2016). The diversity, both within and between individuals, presents major hurdles to biomarker development. This is further complicated by the breadth of potential TCRs able to recognize even a single antigen (Meysman et al., 2019), hampering detection of population-wide signatures of antigen exposure. Indeed, mathematical modeling suggests that only 10-15% of TCRs are public or shared frequently by multiple individuals (Elhanati et al., 2018), which is consistent with observations from extremely deeply sequenced human repertoires (Soto et al., 2019). Despite advances in high-throughput next-generation TCR amplicon sequencing, only a fraction of the repertoire can be assayed, making it difficult to reproducibly sample many relevant TCR clonotypes from an individual, let alone reliably detect public clonotypes in a population. In practice, the problem is exacerbated by unequal sampling depth. Thus, individual T cell clonotypes are currently sub-optimal and under-powered for population-level investigations of TCR specificity, which limits their application in the development of TCR-based clinical biomarkers**.**

In this study we explored the utility of defining meta-clonotypes to accelerate discovery of TCR biomarkers. We define meta-clonotypes as groups of TCRs with biochemically similar complementarity determining regions (CDRs) that are likely to share antigen recognition. By grouping similar TCRs together, repertoire analysis can more robustly identify and quantify "public" features shared across individuals. Such features could subsequently be leveraged for building population-level biomarkers of clinical outcomes that may depend on antigen-specific features of the TCR repertoire, such as disease severity in natural infection or the level of vaccine-induced protection. Shifting the focus of repertoire analysis from clonotypes to meta-clonotypes increases statistical power by reducing the inherent sparsity of finite repertoire samples and increasing the precision with which the frequencies of antigen-specific cells can be estimated. A number of existing tools already enable grouping TCRs by sequence similarity; for example, VDJtools (TCRNET) and ALICE evaluate networks of similar TCR β- or TCR α-chain CDR3s based on a maximum edit-distance of one amino acid substitution, insertion or deletion, while GLIPH2 groups similar TCRs based on shared amino acid k-mers in identical length CDR3s (Glanville et al., 2017; Huang et al., 2020; Pogorelyy et al., 2019; Pogorelyy and Shugay, 2019; Ritvo et al., 2018). Previously, we introduced TCRdist, a weighted multi-CDR, biochemically informed distance metric that enabled grouping of paired αβ TCRs by antigen specificity, based on their sequence similarity (Dash et al., 2017). Here, we describe a new application of TCRdist that guides formation of meta-clonotypes optimized for biomarker development. This application is made possible by a new open-source Python3 software package *tcrdist3* that brings new flexibility to distance-based repertoire analysis, allowing customization of the distance metric, analysis of γδ TCRs, and at-scale computation with sparse data representations and parallelized, byte-compiled code.

Here we first describe a novel analytical framework for identifying meta-clonotypes in antigen-enriched repertoires. The framework is then applied to a large publicly available dataset of putative SARS-CoV-2 antigen-associated TCRs with the objective of identifying meta-clonotypes that could be

used as features in further developing SARS-CoV-2 related biomarkers. The SARS-CoV-2 virus has caused a pandemic with more than 50 million recorded cases within one year of its initial discovery. One of the distinguishing characteristics of SARS-CoV-2 infection is the wide range of potential exposure outcomes, from transient, asymptomatic infection to severe disease requiring hospitalization and intensive care. While there are high quality biomarkers for detecting active SARS-CoV-2 infection via viral RNA qPCR (Nalla et al., 2020) and prior exposure via antibody ELISA (Espejo et al., 2020), additional biomarkers capable of predicting susceptibility to symptomatic infection or severe disease could help guide clinical care and public health policy. Several studies have begun to describe the cellular adaptive immune responses that are elicited by SARS-CoV-2 infection and how they correlate with disease severity (Le Bert et al., 2020; McMahan et al., 2020; Wang et al., 2020; Weiskopf et al., 2020). These and other studies have also established that 20-50% of unexposed individuals have T cell responses to SARS-CoV-2, raising the hypothesis that exposure to “common-cold” coronaviruses may shape the response to SARS-CoV-2 infection (Sette and Crotty, 2020; Welsh and Selin, 2002). T cells likely play an integral role in SARS-CoV-2 pathogenesis and may be an important target for biomarker development. For instance, a TCR biomarker of pre-existing SARS-CoV-2 responses could help predict the course of infection. A T cell-based biomarker might also play a role in vaccine development, for which immunological surrogates of vaccine-induced protection or response durability are highly valued. Most published studies have had limited ability to determine quantitative immunodominance hierarchies, relying on pooled peptide assays, due to the large size of the SARS-CoV-2 proteome and HLA diversity; direct repertoire measurement tied to identified epitopes is a complementary approach to resolve the associated T cell response.

One recent study to elucidate the role of cellular immune responses in acute SARS-CoV-2 infection examined the T cell receptor repertoires of patients diagnosed with COVID-19 disease. Researchers used an assay based on antigen stimulation and flow cytometric sorting of activated CD8+ T cells to sequence SARS-CoV-2 peptide-associated TCR β-chains; the assay is called “multiplex identification of T-cell receptor antigen specificity” or MIRA (Klinger et al., 2015). Data from these experiments were released publicly in July 2020 by Adaptive Biotechnologies and Microsoft as part of ‘immuneRACE’ and their efforts to stimulate science on COVID-19 (Nolan et al., 2020; Snyder et al., 2020). The MIRA antigen enrichment assays identified 269 sets of TCR β-chains associated with CD8+ T cells activated by exposure to SARS-CoV-2 peptides, with TCR sets ranging in size from 1 - 16,607 TCRs (Table S1). The deposited ImmuneRace datasets also included bulk TCR β-chain repertoires from 694 patients within 0-30 days of COVID-19 diagnosis. To demonstrate potential uses of our new analytical tools for TCR repertoire analysis and to accelerate understanding of the cellular responses to SARS-CoV-2 infection, we present analyses of these data with a focus on an integration of the peptide-associated MIRA TCR repertoires with bulk repertoires from four COVID-19 observational studies that enrolled patients with diversity in age and geography (Alabama, USA n = 374; Madrid, Spain, n=117; Pavia, Italy, n=125; Washington, USA, n=78).

FRAMEWORK

Experimental antigen-enrichment allows discovery of TCRs with biochemically similar neighbors

Searching for identical TCRs within a repertoire - arising either from clonal expansion or convergent nucleotide encoding of amino acids in the CDR3 - is a common strategy for identifying functionally important receptors. However, in the absence of experimental enrichment procedures, observing T cells with the same amino acid TCR sequence in a bulk sample is rare. For example, in 10,000 β-chain TCRs from an umbilical cord blood sample, less than 1% of TCR amino acid sequences

were observed more than once, inclusive of possible clonal expansions (Figure 1A). By contrast, a valuable feature of antigen-enriched repertoires is the presence of multiple T cells with identical or highly similar TCR amino acid sequences (Figure 1A). For instance, 45% of amino acid TCR sequences were observed more than once (excluding clonal expansions) in a set of influenza M1(GILGFVFTL)-A*02:01 peptide-MHC tetramer sorted sub-repertoires from 15 subjects (Dash et al., 2017). Enrichment was evident compared to cord blood for additional peptide-MHC tetramer sorted sub-repertoires obtained from VDJdb (Shugay et al., 2018), though the proportion of TCRs with an identical or similar TCR in each set was heterogeneous.

We investigated the degree to which the MIRA enrichment strategy employed by Nolan et al. (2020) identified TCRs with identical or similar amino acid sequences. In general, across multiple MIRA TCR β-chain antigen-enriched repertoires, the proportion of amino acid TCR sequences observed more than once was generally lower than in the tetramer-enriched repertoires and varied considerably across the sets; some MIRA sets resembled tetramer-sorted sub-repertoires (Figure 1B; see MIRA133), while others were more similar to unenriched repertoires (Figure 1B; see MIRA90). The increased diversity in MIRA-enriched TCR sets versus tetramer-enriched TCR sets may, in part, be explained by: (i) recruitment of lower affinity receptors, (ii) the dependence of MIRA on a diverse set of native host MHC presentation molecules, compared to a single peptide-MHC complex, or (iii) bystander activation in the MIRA stimulation assay.

TCR biochemical neighborhood density is heterogeneous in antigen-enriched repertoires

We next investigated the proportion of unique TCRs with at least one biochemically similar neighbor among TCRs with the same putative antigen specificity. We and others have shown that a single peptide-MHC epitope is often recognized by many distinct TCRs with closely related amino acid sequences (Dash et al., 2017); in fact, detection of such clusters in bulk-sequenced repertoires is the basis of several existing tools: GLIPH (Glanville et al., 2017; Huang et al., 2020), ALICE (Pogorelyy et al., 2019) and TCRNET (Ritvo et al., 2018). Therefore, to better understand new large-scale antigen-enriched datasets, like the SARS-CoV-2 MIRA data, we next evaluated the TCR biochemical neighborhoods, defined for each TCR as the set of similar TCRs whose sequence divergence is within a specified radius. We measured biochemical divergence using a position weighted, multi-CDR TCR distance metric (see Methods for details of *tcrdist3* re-implementation of TCRdist). As the radius about a TCR centroid expands, the number of TCRs it encompasses naturally increases. As a neighborhood radius extends, the number of proximal TCRs increases more rapidly in antigen-enriched repertoires compared to the unenriched repertoires.

To better understand the relationship between the TCR distance radius and the density of proximal TCRs, we constructed empirical cumulative distribution functions (ECDFs) for each individual TCR found within a repertoire (Figure 2). An ECDF was constructed for each centroid TCR (one line in Figure 2), and those with sparse neighborhoods appear as lines that remain flat and do not increase along the *y*-axis even as the search radius expands. Moreover, the proportion of TCRs with sparse or empty neighborhoods (ECDF proportion < 0.001) is indicated by the height of the gray area plotted below the ECDF (Figure 2); we observed the highest density neighborhoods within repertoires sorted based on peptide-MHC tetramer binding. For instance, with the influenza M1(GILGFVFTL)-A*02:01 tetramer-enriched repertoire from 15 subjects, we observed that many TCRs were concentrated in dense neighborhoods, which included as much as 30% of the other influenza M1-recognizing TCRs within a radius of 12 TCRdist units (tdus) (Figure 2A). Notably there were also many TCRs with empty or sparse neighborhoods using a radius of 12 tdus (11/247, 44%) or 24 tdus (83/247, 34%). Based on previous work (Dash et al., 2017), we assume that the majority of these tetramer-sorted CD8+ T cells without many

close proximity neighbors do indeed bind the influenza M1:A*02:01 tetramer. This suggests that TCRs within sparse neighborhoods represent less common modes of antigen recognition and highlights the broad heterogeneity of neighborhood densities even among TCRs recognizing a single pMHC.

Neighbor densities for individual TCRs within MIRA identified antigen-enriched repertoires were highly heterogeneous. Densities for an illustrative MIRA set are shown in Figure 2 (MIRA55:ORF1ab 4211:4252; peptide ALRKVPTDNYITTY). Within this antigen-enriched repertoire, at 24 tdus, 8.9% (44/497) of TCR neighborhoods included >10% of the other antigen-activated CD8+ TCRs (Figure 2B). As expected, TCR neighborhoods in the umbilical cord blood repertoire were sparser (Figure 2C); the densest neighborhood included only 0.13% of the repertoire at 24 tdus. We also noted that TCRs with empty neighborhoods tended to have longer CDR3 loops. This is consistent with mathematical modeling approaches that show that TCRs with shorter CDR3 loops have a higher generation probability (P_{gen}) during genomic recombination of the TCR locus (Marcou et al., 2018; Murugan et al., 2012; Sethna et al., 2019). Absent strong selection for antigen recognition, TCRs with a low generation probability are thus more likely to have a less dense biochemical neighborhood. Together, these observations suggest that biochemical neighborhood density is highly heterogeneous among TCRs and that it may depend on mechanisms of antigen-recognition as well as receptor V(D)J recombination biases (Thomas and Crawford, 2019).

Biochemical neighborhood radius can be tuned to balance a biomarker's sensitivity and specificity

The utility of a TCR-based biomarker depends on the antigen-specificity of the TCRs. Therefore, a key constraint on distance-based clustering is the presence of similar TCR sequences that may lack the ability to recognize the target antigen. To be useful, a biochemical neighborhood definition should be wide enough to capture multiple biochemically similar TCRs with shared antigen-recognition, but not excessively broad as to include a high number of sequences found in background repertoires that are antigen naive. Because the density of neighborhoods around TCRs are heterogeneous, we hypothesize that the optimal radius defining a meta-clonotype may differ for each TCR. To find an ideal radius we proposed comparing the relative density of a radius-defined target TCR neighborhood in the antigen-enriched sub repertoire (Figure 3A) to the density of the radius-defined neighborhood in an unenriched background repertoire (Figure 3B, 3C). This is similar to previous approaches taken by tools like ALICE and TCRNET, except that we employ a biochemically informed distance measure (TCRdist) and tune the radius around each TCR to balance the antigen-enriched and unenriched neighborhood densities. The radius around each TCR defines a meta-clonotype that can be used to search for and quantify the abundance of conformant sequences in bulk repertoires (Figure 4A, 4B). For each TCR, its radius-defined meta-clonotype tends to be more abundant within a repertoire and more prevalent in a population than the exact clonotype; for example, multiple TCR meta-clonotypes formed from the MIRA55:ORF1ab set were detected in 13 of 15 HLA-A*01 participants in the MIRA cohort, whereas the centroid TCRs from each of the meta-clonotypes were consistently less prevalent (Figure S1).

An ideal radius-defined meta-clonotype would include a high density of TCRs sharing antigen recognition, yet a low density of TCRs among an antigen-naive background. We demonstrate this approach for selecting an optimal radius for TCRs in the MIRA55:ORF1ab dataset, which includes TCRs from 15 COVID-19 diagnosed subjects (see Methods for details about MIRA and the immuneRACE dataset). First, an ECDF is constructed for each TCR showing the relationship between the meta-clonotype radius and its "sensitivity": its inclusion of similar antigen-recognizing TCRs, approximated by the proportion of TCRs in the antigen-enriched TCR set that are within the radius-defined neighborhood (Figure 3A). Next, an ECDF is constructed for each TCR showing the relationship between the meta-clonotype radius and its "specificity": its exclusion of TCRs with divergent antigen-recognition,

approximated by the proportion of TCRs in an unenriched background repertoire within the radius-defined neighborhood (Figure 3B). Generating an appropriate set of unenriched background TCRs is important; for each set of antigen-associated TCRs discovered by MIRA, we created a two part background. One part consisted of 100,000 synthetic TCRs whose TRBV- and TRBJ-gene frequencies matched those in the antigen-enriched repertoire; TCRs were generated using the software OLGA (Marcou et al., 2018; Sethna et al., 2019). The other part consisted of 100,000 umbilical cord blood TCRs sampled uniformly from 8 subjects (Britanova et al., 2017). This mix balanced dense sampling of sequences near the biochemical neighborhoods of interest with broad sampling of TCRs from an antigen-naive repertoire. Importantly, we adjusted for the biased sampling by using the TRBV- and TRBJ-gene frequencies observed in the cord blood data (see Methods for details about inverse probability weighting adjustment). Using this approach, we are able to estimate the abundance of TCRs similar to a centroid TCR in an unenriched background repertoire of effectively ~1,000,000 TCRs, using a comparatively modest background dataset of 200,000 TCRs. While this may underestimate the true specificity since some of the neighborhood TCRs in the unenriched background repertoire may in fact recognize the antigen of interest, this measure is useful for prioritizing neighborhoods and selecting a radius for each neighborhood that balances sensitivity and specificity.

We find that the neighborhoods around TCR centroids with higher probabilities of generation consistently span a higher proportion of unenriched background TCRs across a range of radii, suggesting that a smaller radius may be desirable for forming neighborhood meta-clonotypes from high P_{gen} TCRs. With a large neighborhood radius, all TCR centroids had high sensitivity and low specificity, indicated by the meta-clonotypes including both a high proportion of TCRs from the antigen-enriched and unenriched repertoires. Some TCRs had low sensitivity and specificity even at a radius of 24 tdus, indicative of a low P_{gen} or “snowflake” TCR: a seemingly unique TCR in both the antigen-enriched and unenriched repertoires. However, radius-defined neighborhoods around many TCRs in the MIRA55:ORF1ab repertoire included 1 - 10% of the antigen-enriched repertoire (5-50 clonotypes) with a radius that included fewer than 0.0001% of TCRs (equivalent to 1 out 10^6) in the unenriched background repertoire, demonstrating a level of sensitivity and specificity that would be favorable for development of a TCR biomarker (Figure 3C, one example meta-clonotype).

RESULTS

Engineering meta-clonotype features for SARS-CoV-2

The MIRA antigen enrichment assays identified 269 sets of TCR β-chains associated with recognition of a SARS-CoV-2 antigen using CD8+ T cell enriched PBMC samples from 68 COVID-19 diagnosed patients. Of these, 252 included at least 6 unique TCRs (unique TRBV-CDR3 amino acid sequences; referred to as MIRA1 - MIRA252; Table S1). All TCR clonotypes in the MIRA enriched repertoires, defined by identical TRBV gene and CDR3 at the amino acid level, were initially considered as candidate centroids; only 2.7% of the clonotypes were found in more than one MIRA participant. For each candidate TCR, a meta-clonotype was engineered by selecting the maximum distance radius that controlled the estimated number of neighboring TCRs in a bulk unenriched repertoire to less than 1 in 10^6, estimated using an inverse probability weighted antigen-naive background repertoire (see Methods). We then ranked the meta-clonotypes by their sensitivity approximated as the proportion of a centroid’s MIRA-enriched repertoire spanned by the search radius (diagrammed in Figure 4). Redundant, lower-ranked meta-clonotypes were eliminated if they were completely encompassed by a higher-ranked meta-clonotype. We further required that meta-clonotypes be public, including sequences from at least two subjects in the MIRA cohort. We found that 102 of the 252 MIRA sets (Table S6) had sufficiently similar

TCRs observed in multiple subjects allowing formation of public meta-clonotypes. From 100,135 TCR β-clones across these 102 MIRA sets, we engineered 6,478 public meta-clonotypes, which spanned 17% of the original TCR sequences (17,421 / 100,135). The proportion of MIRA-enriched TCRs spanned by the meta-clonotypes ranged widely from <1% with MIRA42 to 63% with MIRA7, reflecting broad heterogeneity in the diversity of TCRs activated by each peptide in the assay.

As an example, the MIRA repertoire MIRA55:ORF1ab 4211:4252 (TCRs associated with stimulation peptides ALRKVPTDNYITTY or KVPTDNYITTY) included 524 TCRs from 15 individuals. From the 524 potential centroids, we defined 46 public meta-clonotypes. Among these features, the radii ranged from 10-36 tdus (median 22 tdus), and the publicity - the number of unique subjects spanned by the meta-clonotype - ranged from 3 to 12 individuals (median 6). Meta-clonotype and meta-clonotype summary statistics for other enriched repertoires are provided in the Supplemental Materials (Table S6, S7, S8). The result was a set of non-redundant, public meta-clonotypes that could be used to search for and quantify putative SARS-CoV-2-specific TCRs in bulk repertoires (Table S7). In addition to the radius-defined meta-clonotypes (RADIUS), we also developed a modified approach that additionally enforced a motif-constraint (RADIUS + MOTIF). The constraint further limited sequence divergence in highly conserved positions of the CDR3, requiring that candidate bulk TCRs match specific amino acids found in the meta-clonotype CDR3s to be counted as part of the neighborhood (see Methods).

Evidence of HLA-restriction in SARS-CoV-2 antigen-enriched sub repertoires

Given the important role of HLA class I molecules in antigen presentation and given the role HLA genotype plays in shaping the TCR repertoire (DeWitt, 2018), we further focused on 18 of the 269 repertoires which showed strong evidence of HLA restriction based on two criteria: (i) computational prediction of HLA binding to the SARS-CoV-2 stimulation peptides, and (ii) HLA allele expression of MIRA participants contributing TCRs. With each set of the MIRA TCRs and the associated SARS-CoV-2 peptides we used HLA binding predictions (NetMHCpan4.0) to identify the class I HLA alleles that were predicted to bind with strong (IC50<50 nM) or weak (50 nm< IC50 <500 nM) affinity to any of the 8, 9, 10, or 11-mers derived from the stimulation peptides (Tables S2, S3). For instance, the peptides associated with MIRA55 TCRs (ORF1ab 4211:4252) are predicted to preferentially bind A*01 (IC50 21 nM), B*15 (IC50 120 nM), and B*35 (IC50 32 nM), and peptides associated with MIRA51 TCRs (nucleocapsid phosphoprotein 29348:29380) are predicted to bind A*03 (IC50 19 nM), A*11 (IC50 8 nM), and A*68 (IC50 9 nM).

Of the COVID-19 patients' samples screened using the MIRA assay, HLA genotype was available for 47 patients and only a subset of patients contributed TCRs to each of the MIRA sets. As a second indicator of HLA restriction, we tested whether the subgroup of patients contributing TCRs to each MIRA set was enriched with individuals expressing specific HLA class I alleles (2-digit resolution) (Table S4). We found that for 18 of the MIRA sets, the patients contributing TCRs were significantly enriched for at least one HLA allele (Fisher's exact test p<0.001). In one case, all 13 patients expressing an A*01 allele and only 2 of 34 patients not expressing A*01, contributed to the MIRA55 TCR set (p=1e-7); as noted above, A*01 was also strongly predicted by NetMHCpan4.0 to bind the MIRA55 peptides. Similar patterns of enrichment and predicted binding were seen with A*01 expressing individuals and recognition of MIRA1:ORF1ab 5171:5203 (HTTDPSFLGRY, p=1.9e-7) and MIRA45:ORF3a 25996:26037 (SYFTSDYYQ, p=1.9e-7). Notably, for all 18 MIRA sets, the enriched participant HLA allele was also predicted to bind the stimulating peptide (IC50 < 500 nM), which provided strong evidence of the HLA allele restricting the TCRs in the MIRA antigen-enriched sub repertoires (Table S5).

HLA-associated abundance of SARS-CoV-2 meta-clonotypes in bulk repertoires of COVID-19 patients

We focused confirmatory analyses on TCR meta-clonotypes derived from the 18 SARS-CoV-2 MIRA-identified TCR sets that showed strongest evidence of HLA restriction. We hypothesized that in an independent cohort of COVID-19 patients, the abundance of TCRs matching each meta-clonotype would be higher in patients expressing the restricting HLA allele. To test this hypothesis, we compared three TCR-based feature sets: (i) radius-defined meta-clonotypes, (RADIUS), (ii) radius and motif-defined meta-clonotypes (RADIUS+MOTIF) and (iii) centroid clonotypes alone, using TRBV-CDR3 amino acid (EXACT) matching (Tables S6, S7). Using the features in each set we screened TCRs from the bulk TCR β-chain repertoires of 694 COVID-19 patients whose repertoires were publicly released as part of the immuneRACE datasets (see Methods for details); these patients were not part of the smaller cohort that contributed samples to the MIRA experiments. Testing the HLA restriction hypothesis required having the HLA genotype of each individual, which was not provided in the dataset. To overcome this, we inferred each participant's HLA genotype with a classifier that was based on previously published HLA-associated TCR β-chain sequences (DeWitt et al., 2018) and their abundance in each patient's repertoire (see Methods for details). No MIRA TCRs were used to assign HLA-types to the 694 COVID-19 patients. We then used a beta-binomial counts regression model (Rytlewski et al., 2019) with each TCR feature to test for an association of feature abundance with presence of the restricting allele in the participant's HLA genotype, controlling for participant age, sex, and days since COVID-19 diagnosis.

The models revealed that there were radius-defined meta-clonotypes with a strong positive and statistically significant association (FDR < 0.001) for 10 of the 18 HLA-restricted-MIRA sets that were evaluated (Figure 5A, Table S7). Across all MIRA sets, a significant HLA-association was detected for 29% (657/1915) and 43% (845/1915) of the meta-clonotypes using the RADIUS or RADIUS+MOTIF definitions, respectively. In comparison, strong HLA-association was detected in fewer than 2% (27/1915) of exact clonotype features, largely because the specific TRBV gene and CDR3 sequences discovered in the MIRA experiments were infrequently observed in unenriched bulk samples (Figure 5B). When detectable, the abundance of exact TCR clonotypes in bulk repertoires tended to be positively associated with expression of the restricting HLA allele, as hypothesized. However, in most cases, the associated false discovery rate-adjusted q-value of these associations were orders of magnitude higher (i.e., less significant) than those obtained from using the engineered RADIUS or RADIUS+MOTIF feature with the same clonotype as a centroid (Figure 6B). The improved performance of meta-clonotypes as query features is particularly evident when testing for HLA-associated enrichment of TCRs recognizing immunodominant MIRA1:A*01 (Figure 5A, Figure 6A), MIRA48:A*02 , MIRA51:A*03, MIRA53:A*24, and MIRA55:A*01 (Figure 6B). Moreover, the regression models with meta-clonotypes also revealed possible negative associations between TCR abundance and participant age and positive associations with sample collection more than two days post COVID-19 diagnosis (Figure 6A).

DISCUSSION

Given the extent of TCR diversity across individuals, population-scale analysis of exact antigen-specific clonotype abundance is likely limited to public (i.e., higher P_{gen}) TCR features (Figure S4). To more fully understand the population-level dynamics of complex polyclonal T-cell responses across a gradient of generation probabilities, it is critical to develop methods for finding public TCR meta-clonotypes that capture otherwise private TCRs. We developed a novel framework, integrating antigen-enriched repertoires with efficiently sampled unenriched background repertoires, to engineer meta-clonotypes that balance the need for sufficiently public features with the need to maintain antigen

specificity. The output of the analysis framework (Figure 4A) is a set of meta-clonotypes (each represented by a (i) centroid, (ii) radius, and (iii) optionally, a motif-pattern) that can be used to rapidly search for and quantify similar TCRs – likely sharing antigen-recognition – in bulk repertoires. To demonstrate this analytical framework, we analyzed publicly available sets of antigen-enriched TCR β-chain sequences that putatively recognize SARS-CoV-2 peptides (Nolan et al., 2020). From these, we generated 6478 TCR radius-defined public meta-clonotypes that could be used to further investigate the CD8+ T cell response to SARS-CoV-2 (Tables S7, S8).

To evaluate the potential clinical relevance of radius-defined meta-clonotypes we focused on those with the strongest evidence of HLA restriction (Table S7, n = 1915). We reasoned that we could compare the abundance of these meta-clonotypes in COVID-19 patients with and without the restricting HLA and that a significant positive association of abundance with expression of the restricting allele would provide confirmatory evidence both of the SARS-CoV-2 specificity of the meta-clonotype and its HLA restriction (Figure 4B). Overall, we found confirmation of HLA-restriction of meta-clonotype abundance for a majority of the MIRA sets we analyzed (11/18) and nearly one-third of all engineered meta-clonotypes (44% using the RADIUS+MOTIF approach). To demonstrate the possibility of employing other complementary tools to generate public TCR features, we applied GLIPH2 to one of the HLA-restricted MIRA sets (MIRA55:ORF1ab; see Methods for details). Some of the GLIPH2 k-mers enriched in MIRA55 TCRs showed evidence of HLA-restriction, supporting the general applicability of using antigen-enriched repertoires to discover generalizable features of otherwise private antigen-recognizing TCRs (Figures S2 and S3). With *tcrdist3,* we also found meta-clonotypes with significantly higher abundance in samples that were provided more than two days after COVID-19 diagnosis, which is consistent with expansions of virus-specific TCRs that would be typical of responses to viral infection and have been shown preliminarily for SARS-CoV-2 (Weiskopf et al., 2020). This further demonstrated the potential clinical relevance of meta-clonotypes and suggests they could be used to study SARS-CoV-2 T cell response kinetics longitudinally.

Recently, Snyder et al. (2020) analyzed 1,521 bulk TCR β-chain repertoires from COVID-19 patients in the immuneRACE dataset and an additional 3,500 (non-publicly available) repertoires from healthy controls to identify public TCR β-chains that could be used to identify SARS-COV-2 infected individuals with high sensitivity and specificity. Their results show that with sufficient data it is possible to engineer highly performant TCR biomarkers of antigen exposure from exact clonotypes. We show that by leveraging antigen-enriched TCR repertoires it is possible to engineer radius-defined TCR meta-clonotypes from a relatively small group of COVID19 diagnosed individuals (n=61; HLA-typed n=47) that are frequently detected in much larger independent cohorts. We propose that meta-clonotypes constitute a set of potential features that could be leveraged in developing TCR-based clinical biomarkers that go beyond detection of infection or exposure. For example, biomarkers predictive of infection, disease severity or vaccine protection may all require different TCR features. Statistical and machine learning tools can be employed to identify the meta-clonotypes and meta-clonotype combinations that carry the desired clinical signal. Much like any biomarker study, to establish a TCR-based predictor of a particular outcome, the features must be measured among a sufficiently large cohort of individuals, with a sufficient mix of outcomes.

Though demonstrating HLA restriction of the SARS-CoV-2 meta-clonotypes helped establish their potential utility, it also highlighted how HLA diversity could be a major hurdle to biomarker development. The sensitivity of a TCR-based biomarker in a diverse population may depend on combining meta-clonotypes with diverse HLA restrictions since individuals with different HLA genotypes often target different epitopes using divergent TCRs. Our analysis shows that having HLA genotype information for TCR repertoire analysis can be critical to interpreting results. The simple HLA classifier we developed suggests that in the near future it may be possible to infer high-resolution HLA genotype from bulk TCR

repertoires, but until then it is valuable to have sequenced-based HLA genotyping. In the absence of HLA genotype information, it may still be feasible to generate informative TCR meta-clonotypes. For example, a polyantigenic TCR-enrichment strategy (i.e., peptide pools or whole-proteins) could help generate meta-clonotypes that broadly cover HLA diversity if the samples are racially, ethnically and geographically representative of the ultimate target population. For these reasons, donor unrestricted T cells and their receptors (e.g., MAITs, γδT cells) may also be good targets for TCR biomarker development.

To enable TCR biomarker development and innovative extensions of distance-based immune repertoire analysis, we developed *tcrdist3*, which provides Python3, open-source (https://github.com/kmayerb/tcrdist3), well-documented (https://tcrdist3.readthedocs.io) computational building blocks for a wide array of TCR repertoire workflows. The software is highly flexible, allowing for: (i) customization of the distance metric with position-specific weights or amino acid substitution matrices, (ii) inclusion of CDRs beyond the CDR3, (iii) clustering based on single-chain or paired-chain data for α/β or γ/δ TCRs, and (iv) use of default as well as user-provided TCR repertoires as background for controlling meta-clonotype specificity (e.g., users may want to use strain-specific, HLA-genotyped, or age-matched backgrounds). *tcrdist3* makes efficient use of available CPU and memory resources; as a reference, application of the biomarker analysis framework to the MIRA55:ORF1ab (n = 525 TCRs) dataset can be completed in less than 2 hours using 1 CPU and < 6 GB or memory, including distance computation, radius optimization, and quantification of meta-clonotypes (n=46) in 694 bulk TCR β-chain repertoires, ranging in size from 10,395 to 1,038,012 in-frame clones (~5 billion total pairwise comparisons). The package also can generate multiple types of publication-ready figures (e.g., background-adjusted CDR3 sequence logos, V/J-gene usage chord diagrams, and annotated TCR dendrograms). The continued maturation of multiple adaptive immune receptor repertoire sequencing technologies will open possibilities for basic immunology and clinical applications, and *tcrdist3* will remain a flexible tool that researchers can use to integrate the data sources needed to detect and quantify antigen specific TCR features.

METHODS

TCR Data: immuneRACE datasets and MIRA assay

The study utilized two primary sources of TCR data (Nolan et al. 2020; Snyder et al. 2020). The first data source was a table of TCR β-chains amplified from CD8+ T cells activated after exposure to a pool of SARS-CoV-2 peptides, using a Multiplex Identification of Receptor Antigen (MIRA) (Klinger et al. 2015). The samples used for the MIRA analysis included samples from 61 individuals diagnosed with COVID-19, of whom 47 were HLA-genotyped. We analyzed the 252 MIRA sets with at least 6 unique TCRs, referred to as M1-252 in rank order by their size (Table S1). Adaptive Biotechnologies also made publicly available bulk unenriched TCR β-chain repertoires from COVID-19 patients participating in a collaborative immuneRACE network of international clinical trials. We selected repertories from individuals where meta-data was available indicating that the sample was collected from 0 to 30 days from the time of diagnosis. (COVID-19-DLS (Alabama, USA n = 374); COVID-19-HUniv12Oct (Madrid, Spain n = 117); COVID-19-NIH/NIAID (Pavia, Italy n=125) + COVID-19-ISB (Washington, USA n = 78). The sampling depth of these repertoires varied from 15,626-1,220,991 productive templates (median 208,709) and 10,395-1,038,012 productive rearrangements (median 113,716). We did not use bulk samples from the COVID-19-ADAPTIVE dataset as the average age was lower than other immuneRACE populations and some of the participants overlap with individuals in the Adaptive led MIRA-based antigen mapping study that we used to identify antigen-specific meta-clonotypes.

451
452 *HLA genotypes and HLA genotype inferences*
453
454 No publicly available HLA genotyping was available for the 694 bulk unenriched immuneRACE T
455 cell repertoires (Nolan et al. 2020). Before considering SARS-CoV-2 specific features, we inferred the
456 HLA likely expressed by these participants based on their TCR repertoires. Predictions were based on
457 previously published HLA-associated TCR β-chain sequences (DeWitt et al., 2018) and their abundance
458 in each volunteer's repertoire. Briefly, a weight-of-evidence classifier for each HLA loci was computed as
459 follows. For each sample and for each common allele, the number of detected HLA-diagnostic TCR β-
460 chains was divided by the total possible number of HLA-diagnostic TCR β-chains. The weights were
461 normalized as a probability vector and the two highest HLA-allele probabilities (if the probability was
462 greater than 0.2) were assigned to each sample. The sensitivity and specificity of this simple classifier for
463 each allele prediction were assessed using 550 HLA-typed bulk repertoires (Emerson et al., 2017).
464 Sensitivities for common HLA-A alleles A*01:01, A*02:01, A*03:01, A*24:02, and A*11:01 were 0.96,
465 0.91, 0.90. 0.84, 0.94, respectively. Importantly, specificity for major HLA-A alleles was between 0.97-1.0.
466 With such a low false positive rate, inference of the HLA genotype of most participants was deemed
467 sufficient in the absence of available direct HLA genotyping.
468
469 *Peptide-HLA binding prediction*
470
471 HLA binding affinities of peptides used in the MIRA stimulation assay were computationally
472 predicted using NetMHCpan4.0 (Jurtz et al., 2017). Specifically, the affinities of all 8, 9, 10 and 11mer
473 peptides derived from the stimulation peptides were computed with each of the class I HLA alleles
474 expressed by participants in the MIRA cohort (n=47). From this data we derived 2-digit HLA binding
475 predictions (e.g., A*02) for each MIRA set by pooling the predictions for all the 4-digit HLA variants (e.g.
476 A*02:01, A*02:02) across all the derivative peptides and selecting the lowest IC50 (strongest affinity).
477 Predictions with IC50 < 50 nM were considered strong binders and IC50 < 500 nM were considered weak
478 binders.
479
480 *TCR distances*
481
482 Weighted multi-CDR distances between TCRs were computed in a *tcrdist3*, a open-source
483 Python3 package for TCR repertoire analysis and visualization, using the procedure first described in
484 (Dash et al., 2017). The package has been expanded to include gamma-delta TCRs; it has also been re-
485 coded to increase CPU efficiency using *numba*, a high-performance just-in-time compiler. A numba-
486 coded edit/Levenshtein distance is also included for comparison, with the flexibility to accommodate novel
487 TCR metrics as they are developed.
488 Briefly, the distance metric in this study is based on comparing TCR β-chain sequences. The
489 *tcrdist3* default settings compare TCRs at the CDR1, CDR2, and CDR2.5 and CDR3 positions. By default,
490 IMGT aligned CDR1, CDR2, and CDR2.5 amino acids are inferred from TRVB gene names, using the *01
491 allele sequences when allele level information is not available. The CDR3 junction sequences are
492 trimmed 3 amino acids on the N-terminal side and 2 amino acids on the C-terminus, positions that are
493 highly conserved and less crucial for mediation antigen specific recognition. Trimmed CDR3 sequences
494 are aligned with a single gap, positioned to minimize alignment penalties incurred by a BLOSUM62
495 substitution matrix. Distances are then the weighted sum of substitution penalties across all CDRs, with
496 the CDR3 penalty weighted 3 times greater than other CDRs.
497

498 *Optimized TCR-specific radius*
499
500 To find biochemically similar TCRs while maintaining a high level of specificity, we used the
501 packages *tcrdist3* and *tcrsampler* to generate an appropriate set of unenriched antigen-naive background
502 TCRs. A background repertoire was created for each MIRA set, each consisting of two parts. First, we
503 combine a set of 100,000 synthetic TCRs generated using the software OLGA (Marcou et al., 2018;
504 Sethna et al., 2019), whose TRBV- and TRBJ-gene frequencies match those in the antigen-enriched
505 repertoire. Second we used 100,000 umbilical cord blood TCRs sampled evenly from 8 subjects
506 (Britanova et al., 2016). This mix balances dense sampling of background sequences near the
507 biochemical neighborhoods of interest with broad sampling of common TCR representative of antigen-
508 naive repertoire. We then adjust for the biased sampling by using the TRBV- and TRBJ-gene frequencies
509 observed in the cord blood data**.** The adjustment is a weighting based on the inverse of each TCR's
510 sampling probability. Because we oversampled regions of the "TCR space" near the candidate centroids
511 we were able to estimate the density of the meta-clonotype neighborhoods well below 1 in 200,00. This
512 is important because ideal meta-clonotypes would be highly specific even in repertoires larger than
513 200,000 sequences. With each candidate centroid, a meta-clonotype was engineered by selecting the
514 maximum distance radius that still controlled the number of neighboring TCRs in the weighted unenriched
515 background to 1 in 10^6.
516
517 *TCR meta-clonotype MOTIF constraint*
518
519 We leveraged the resulting clustering of antigen-enriched TCR sequences within a stringent
520 TCRdist radius to discover key conserved residues most likely to determine antigen specificity. To this
521 end, we developed a "motif" constraint as an optional part of each meta-clonotype definition that limited
522 allowable amino-substitutions in highly conserved positions of the CDR3 among the known antigen-
523 enriched TCRs. The motif constraint for each radius-defined meta-clonotype was defined by aligning all of
524 the CDR3 amino acid sequences within the allowable radius of the meta-clonotype centroid. Alignment
525 positions with five or fewer distinct amino acids were considered conserved and added to the motif. The
526 motif constraint is permissive of substitutions in select positions relative to the centroid, however these
527 substitutions are penalized by the radius constraint. Where a gap existed in the alignment of antigen-
528 specific MIRA-derived TCRs, that position was made optional in the motif constraint. The motif constraint
529 was encoded as a regular-expression, with the "." character indicating non-conserved positions and
530 specified degenerate amino acid indicated by the set of allowable residues in brackets (e.g.,
531 "SL[RK]?[ND]YEQ"). Since the motif constraints form regular expressions, they can be used to rapidly
532 scan large repertoires for matching TCRs or validate positional similarity of key residues between a
533 centroid and the set of TCRs found within its specified TCRdist radius. When applied to bulk repertoires,
534 the motif constraint eliminates CDR3s that didn't match key conserved residues.
535
536 *TCR abundance regression modeling*
537
538 Similar to bulk RNA sequencing data, TCR frequencies are count data drawn from samples of
539 heterogeneous size. Thus we initially attempted to fit a negative binomial model to the data (e.g. DESEQ2
540 (Love et al., 2013)). We found that the negative binomial model did not adequately fit TCR counts, which
541 compared to transcriptomic data, were characterized by more technical zeros, due to inevitable under
542 sampling, and even greater over-dispersion, which could be due to clonal expansions and HLA genotype
543 diversity. Instead we found that the beta-binomial distribution, which was recently used for TCR
544 abundance modeling (Rytlewski et al., 2019), provided the flexibility needed to adequately fit the TCR

data. We used an R package, *corncob*, which provides maximum likelihood methods for inference and hypothesis testing with beta-binomial regression models (Martin et al., 2020). Due to the sparsity of some meta-clonotypes, seven percent of coefficient estimates in regression models had p-values greater than 0.99 and unreliable high magnitude coefficient estimates. These values are not shown in the horizontal range of the volcano plots.

Creation of k-mer based TCR features with GLIPH2

GLIPH2 (Huang et al., 2020) was applied to the MIRA55:ORF1ab antigen-enriched sub-repertoire of TCRs (n=524 TCRs) to demonstrate how a kmer-based tool might also be used to cluster biochemically similar antigen-specific TCRs to discover potential TCR biomarker features. Similar to *tcrdist3*, GLIPH2 attempts to find enriched features via comparisons to a background repertoire of TCRs. The MIRA55 set was chosen because it is comprised of CD8+ TCR β-chains activated by a peptide with strong evidence of HLA-restriction, primarily HLA-A*01 (see Table S4). GLIPH2 returned 147 CDR3 patterns enriched relative its default CD8+ TCR background (Fisher's $P < 0.001$). The GLIPH2 features and TRBV gene usages were then used to search for conforming TCRs in the 694 bulk sequenced COVID-19 repertoires, allowing comparison to the TCR clonotype (EXACT) and meta-clonotype features (Figure S3).

tcrdist3: Software for TCR repertoire analysis

tcrdist3 is an open-source Python3 package for TCR repertoire analysis and visualization. The core of the package is the TCRdist, a distance metric for relating two TCRs, which has been expanded beyond what was previously published (Dash et al., 2017) to include γδTCRs. It has also been re-coded to increase CPU efficiency using *numba*, a high-performance just-in-time compiler. A numba-coded edit/Levenshtein distance is also included for comparison, with the flexibility to accommodate novel TCR metrics as they are developed. The package can accommodate data in standardized format including AIRR, vdjdb exports, MIXCR output, 10x Cell Ranger output or Adaptive Biotechnologies immunoSeq output. The package is well documented including examples and tutorials, with source code available on github.com under an MIT license (http://github.com/kmayerbl/tcrdist3). *tcrdist3* imports modules from several other open-source, pip installable packages by the same authors that support the functionality of tcrdist, while also providing more general utility. Briefly, the novel features of these packages and their relevance for TCR repertoire analysis is described here:

pwseqdist enables fast and flexible computation of pairwise sequence-based distances using either numba-enabled tcrdist and edit distances or any user-coded Python3 metric to relate TCRs; it can also accommodate computation of "rectangular" pairwise matrices: distances between a relatively small set of TCRs with all TCRs in a much larger (e.g., bulk) repertoire. On a modern laptop distances can be computed at a rate of ~70M per minute, per CPU.

tcrsampler is a tool for sub-sampling large bulk datasets to estimate the frequency of TCRs and TCR neighborhoods in non-antigen-enriched background repertoires. The module comes with large, bulk sequenced, default databases for human TCR α, β, γ and δ and mouse TCR β (Britanova et al., 2016; Ravens et al., 2018; Wirasinha et al., 2018). Datasets were selected because they represented the largest pre-antigen exposure TCR repertoires available; users can optionally supply their own background repertoires when applicable. An important feature of tcrsampler is the ability to specify sampling strata; for example, sampling is stratified on individual by default so that results are not biased by on individual with deeper sequencing. Sampling can also be stratified on V and/or J-gene usage to over-sample TCRs that are somewhat similar to the TCR neighborhood of interest. This greatly improves sampling efficiency,

since comparing a TCR neighborhood to a background set of completely unrelated TCRs is computationally inefficient; however, we note that it is important to adjust for biased sampling approaches via inverse probability weighting to estimate the frequency of oversampled TCRs in a bulk-sequenced repertoire.

palmotif is a collection of functions for computing symbol heights for sequence logo plots and rendering them as SVG graphics for integration with hierdiff interactive trees or print publication. Much of the computation is based on existing methods that use either KL-divergence/entropy or odds-ratio based approaches to calculate symbol heights. We contribute a novel method for creating a logo from CDR3s with varying lengths. The target sequences are first globally aligned (parasail C++ implementation of Needleman-Wunsch) to a pre-selected centroid sequence. For logos expressing relative symbol frequency, background sequences are also aligned to the centroid. Logo computation then proceeds as usual, estimating the relative entropy between target and background sequences at each position in the alignment and the contribution of each symbol. Gaps introduced in the centroid sequence are ignored, while gap symbols in the aligned sequences are treated as an additional symbol.

SUPPLEMENTAL TABLES

Table S1	MIRA enriched repertoires MIRA0 - MIRA252
Table S2	HLA class I alleles capable of presenting the SARS-CoV-2 associated peptides MIRA screen
Table S3	NetMHCpan4.0 peptide MHC class I binding affinity prediction
Table S4	Statistical associations between common HLA genotypes of COVID-19 exposed MIRA participants and SARS-CoV-2 peptide-enriched TCR repertoires
Table S5	Peptide-associated repertoires by MIRA that showed both strong in silico evidence of HLA-restriction and were highly associated with participants expressing a predicted MHC Class 1 binder
Table S6	SARS-CoV-2 CD8+ meta clonotypes summarized by MIRA enriched repertoire
Table S7	Table S7 SARS-CoV-2 CD8+ meta clonotypes with strong evidence of HLA restriction (n = 1915)
Table S8	SARS-CoV-2 CD8+ meta clonotypes with less evidence of HLA restriction (n = 4562)
Table S9	Comparison of selected software tools for clustering TCRs

SUPPLEMENTAL FIGURES

Figure S1	Publicity analysis of CD8+ TCR β-chain features activated by SARS-CoV-2 peptide ORF1ab 4211:4252.
Figure S2	Publicity and breadth analysis of CD8+ TCR β-chain features activated by SARS-CoV-2 peptide ORF1ab 4211:4252 (MIRA55) using tcrdist3 and GLIPH2.
Figure S3	Associations of TCR features with participant age, days post diagnosis, HLA-type, and sex in bulk TCR β-chain repertoires of COVID-19 patients (n=694).
Figure S4	Detectable HLA-association and CDR3 probability of generation.

DATA AVAILABILITY

ImmuneRace data is publicly available: https://immunerace.adaptivebiotech.com/data/. All other TCR data is publicly available from VDJdb (https://vdjdb.cdr3.net/) or the cited research.

SOFTWARE AVAILABILITY

The *tcrdist3* code base used in this analysis is freely available at https://github.com/kmayerb/tcrdist3/ with documented examples at http://tcrdist3.readthedocs.io. *tcrdist3* relies on the Python package *pwseqdist* - freely available at https://github.com/agartland/pwseqdist - for numba-optimized just-in-time compiled versions of the TCRdist measure.

CONTRIBUTIONS

Conceptualization: KM, SS, LC, JC, AS, JG, TH, PT, PB, AF; Methodology; Software: KM, AF; Validation; Formal analysis; Investigation: KM, AF; Data Curation; Writing – original draft preparation: KM, AF; Writing – review & editing: KM, SS, LC, JC, AS, JG, TH, PT, PB, AF; Supervision: TH, PT, PB, AF; Funding acquisition: TH, PT, PB, AF

ACKNOWLEDGEMENTS

This work was funded by NIH NIAID R01 AI136514-03 (PI Thomas) and ALSAC at St. Jude.

REFERENCES

Ahmadzadeh M, Pasetto A, Jia L, Deniger DC, Stevanović S, Robbins PF, Rosenberg SA. 2019. Tumor-infiltrating human CD4+ regulatory T cells display a distinct TCR repertoire and exhibit tumor and neoantigen reactivity. *Sci Immunol* 4. doi:10.1126/sciimmunol.aao4310

Britanova OV, Shugay M, Merzlyak EM, Staroverov DB, Putintseva EV, Turchaninova MA, Mamedov IZ, Pogorelyy MV, Bolotin DA, Izraelson M, Davydov AN, Egorov ES, Kasatskaya SA, Rebrikov DV, Lukyanov S, Chudakov DM. 2016. Dynamics of individual T Cell repertoires: from cord blood to centenarians. *The Journal of Immunology* 196:5005–5013.

Cao K, Wu J, Li Xuemei, Xie H, Tang C, Zhao X, Wang S, Chen L, Zhang W, An Y, Li Xin, Lin L, Chai R, Fang M, Yue Y, Wang X, Ding Y, Zhou L, Zhao Q, Yang H, Wang J, He S, Liu X. 2020. T-cell receptor repertoire data provides new evidence for hygiene hypothesis of allergic diseases. *Allergy*. doi:10.1111/all.14014

Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, Crawford JC, Clemens EB, Nguyen THO, Kedzierska K, La Gruta NL, Bradley P, Thomas PG. 2017. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* 547:89–93.

DeWitt WS 3rd, Smith A, Schoch G, Hansen JA, Matsen FA 4th, Bradley P. 2018. Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *Elife* 7. doi:10.7554/eLife.38358

Elhanati Y, Sethna Z, Callan CG Jr, Mora T, Walczak AM. 2018. Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunol Rev* 284:167–179.

Emerson RO, DeWitt WS, Vignali M, Gravley J, Hu JK, Osborne EJ, Desmarais C, Klinger M, Carlson CS, Hansen JA, Rieder M, Robins HS. 2017. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet* 49:659–665.

Espejo AP, Akgun Y, Al Mana AF, Tjendra Y, Millan NC, Gomez-Fernandez C, Cray C. 2020. Review of current advances in serologic testing for COVID-19. *Am J Clin Pathol*.

675 Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, Ji X, Han A, Krams SM, Pettus C, Haas N,
676 Arlehamn CSL, Sette A, Boyd SD, Scriba TJ, Martinez OM, Davis MM. 2017. Identifying
677 specificity groups in the T cell receptor repertoire. *Nature* 547:94–98.
678 Huang H, Wang C, Rubelt F, Scriba TJ, Davis MM. 2020. Analyzing the Mycobacterium tuberculosis
679 immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening.
680 *Nat Biotechnol*. doi:10.1038/s41587-020-0505-4
681 Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. 2017. NetMHCpan-4.0: Improved peptide–
682 MHC Class I interaction predictions integrating eluted ligand and peptide binding affinity data. *The*
683 *Journal of Immunology* 199:3360–3368.
684 Kato T, Matsuda T, Ikeda Y, Park J-H, Leisegang M, Yoshimura S, Hikichi T, Harada M, Zewde M, Sato
685 S, Hasegawa K, Kiyotani K, Nakamura Y. 2018. Effective screening of T cells recognizing
686 neoantigens and construction of T-cell receptor-engineered T cells. *Oncotarget* 9:11009–11019.
687 Klinger M, Pepin F, Wilkins J, Asbury T, Wittkop T, Zheng J, Moorhead M, Faham M. 2015. Multiplex
688 identification of antigen-specific t cell receptors using a combination of immune assays and
689 immune receptor sequencing. *PLoS One* 10:e0141561.
690 Le Bert N, Tan AT, Kunasegaran K, Tham CYL, Hafezi M, Chia A, Chng MHY, Lin M, Tan N, Linster M,
691 Chia WN, Chen MI-C, Wang L-F, Ooi EE, Kalimuddin S, Tambyah PA, Low JG-H, Tan Y-J,
692 Bertoletti A. 2020. SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and
693 uninfected controls. *Nature* 584:457–462.
694 Love M, Anders S, Huber W. 2013. Differential analysis of RNA-Seq data at the gene level using the
695 DESeq2 package. *Heidelberg: European Molecular Biology Laboratory (EMBL)*.
696 Lythe G, Callard RE, Hoare RL, Molina-París C. 2016. How many TCR clonotypes does a body maintain?
697 *J Theor Biol* 389:214–224.
698 Marcou Q, Mora T, Walczak AM. 2018. High-throughput immune repertoire analysis with IGoR. *Nat*
699 *Commun* 9:561.
700 Martin BD, Witten D, Willis AD. 2020. Modeling microbial abundances and dysbiosis with beta-binomial
701 regression. *Ann Appl Stat* 14:94–115.
702 McMahan K, Yu J, Mercado NB, Loos C, Tostanoski LH, Chandrashekar A, Liu J, Peter L, Atyeo C, Zhu
703 A, Bondzie EA, Dagotto G, Gebre MS, Jacob-Dolan C, Li Z, Nampanya F, Patel S, Pessaint L,
704 Van Ry A, Blade K, Yalley-Ogunro J, Cabus M, Brown R, Cook A, Teow E, Andersen H, Lewis
705 MG, Lauffenburger DA, Alter G, Barouch DH. 2020. Correlates of protection against SARS-CoV-2
706 in rhesus macaques. *Nature*. doi:10.1038/s41586-020-03041-6
707 Meysman P, De Neuter N, Gielis S, Bui Thi D, Ogunjimi B, Laukens K. 2019. On the viability of
708 unsupervised T-cell receptor sequence clustering for epitope preference. *Bioinformatics* 35:1461–
709 1468.
710 Murugan A, Mora T, Walczak AM, Callan CG Jr. 2012. Statistical inference of the generation probability of
711 T-cell receptors from sequence repertoires. *Proc Natl Acad Sci U S A* 109:16161–16166.
712 Nalla AK, Casto AM, Huang M-LW, Perchetti GA, Sampoleo R, Shrestha L, Wei Y, Zhu H, Jerome KR,
713 Greninger AL. 2020. Comparative Performance of SARS-CoV-2 Detection Assays Using Seven
714 Different Primer-Probe Sets and One Assay Kit. *J Clin Microbiol* 58. doi:10.1128/JCM.00557-20
715 Nolan S, Vignali M, Klinger M, Dines JN, Kaplan IM, Svejnoha E, Craft T, Boland K, Pesesky M,
716 Gittelman RM, Snyder TM, Gooley CJ, Semprini S, Cerchione C, Mazza M, Delmonte OM, Dobbs
717 K, Carreño-Tarragona G, Barrio S, Sambri V, Martinelli G, Goldman JD, Heath JR, Notarangelo
718 LD, Carlson JM, Martinez-Lopez J, Robins HS. 2020. A large-scale database of T-cell receptor
719 beta (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-
720 CoV-2. *Res Sq*. doi:10.21203/rs.3.rs-51964/v1

721 Pogorelyy MV, Minervina AA, Shugay M, Chudakov DM, Lebedev YB, Mora T, Walczak AM. 2019.
722 Detecting T cell receptors involved in immune responses from single repertoire snapshots. *PLOS*
723 *Biology*. doi:10.1371/journal.pbio.3000314
724 Pogorelyy MV, Shugay M. 2019. A framework for annotation of antigen specificities in high-throughput T-
725 Cell repertoire sequencing studies. *Front Immunol* 10:2159.
726 Ravens S, Schultze-Florey C, Raha S, Sandrock I, Drenker M, Oberdörfer L, Reinhardt A, Ravens I, Beck
727 M, Geffers R, von Kaisenberg C, Heuser M, Thol F, Ganser A, Förster R, Koenecke C, Prinz I.
728 2018. Publisher Correction: Human γδ T cells are quickly reconstituted after stem-cell
729 transplantation and show adaptive clonal expansion in response to viral infection. *Nature*
730 *Immunology*. doi:10.1038/s41590-018-0054-x
731 Ritvo P-G, Saadawi A, Barennes P, Quiniou V, Chaara W, El Soufi K, Bonnet B, Six A, Shugay M,
732 Mariotti-Ferrandiz E, Klatzmann D. 2018. High-resolution repertoire analysis reveals a major
733 bystander activation of Tfh and Tfr cells. *Proc Natl Acad Sci U S A* 115:9604–9609.
734 Rytlewski J, Deng S, Xie T, Davis C, Robins H, Yusko E, Bienkowska J. 2019. Model to improve
735 specificity for identification of clinically-relevant expanded T cells in peripheral blood. *PLoS One*
736 14:e0213684.
737 Sethna Z, Elhanati Y, Callan CG, Walczak AM, Mora T. 2019. OLGA: fast computation of generation
738 probabilities of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics* 35:2974–
739 2981.
740 Sette A, Crotty S. 2020. Pre-existing immunity to SARS-CoV-2: the knowns and unknowns. *Nat Rev*
741 *Immunol* 20:457–458.
742 Shugay M, Bagaev DV, Zvyagin IV, Vroomans RM, Crawford JC, Dolton G, Komech EA, Sycheva AL,
743 Koneva AE, Egorov ES, Eliseev AV, Van Dyk E, Dash P, Attaf M, Rius C, Ladell K, McLaren JE,
744 Matthews KK, Clemens EB, Douek DC, Luciani F, van Baarle D, Kedzierska K, Kesmir C,
745 Thomas PG, Price DA, Sewell AK, Chudakov DM. 2018. VDJdb: a curated database of T-cell
746 receptor sequences with known antigen specificity. *Nucleic Acids Res* 46:D419–D427.
747 Snyder TM, Gittelman RM, Klinger M, May DH, Osborne EJ, Taniguchi R, Zahid HJ, Kaplan IM, Dines JN,
748 Noakes MN, Pandya R, Chen X, Elasady S, Svejnoha E, Ebert P, Pesesky MW, De Almeida P,
749 O'Donnell H, DeGottardi Q, Keitany G, Lu J, Vong A, Elyanow R, Fields P, Greissl J, Baldo L,
750 Semprini S, Cerchione C, Mazza M, Delmonte OM, Dobbs K, Carreño-Tarragona G, Barrio S,
751 Imberti L, Sottini A, Quiros-Roldan E, Rossi C, Biondi A, Bettini LR, D'Angio M, Bonfanti P,
752 Tompkins MF, Alba C, Dalgard C, Sambri V, Martinelli G, Goldman JD, Heath JR, Su HC,
753 Notarangelo LD, Martinez-Lopez J, Carlson JM, Robins HS. 2020. Magnitude and dynamics of
754 the T-Cell response to SARS-CoV-2 infection at both individual and population levels. *medRxiv*.
755 doi:10.1101/2020.07.31.20165647
756 Soto C, Bombardi RG, Branchizio A, Kose N, Matta P, Sevy AM, Sinkovits RS, Gilchuk P, Finn JA, Crowe
757 JE Jr. 2019. High frequency of shared clonotypes in human B cell receptor repertoires. *Nature*
758 566:398–402.
759 Thomas PG, Crawford JC. 2019. Selected before selection: A case for inherent antigen bias in the T cell
760 receptor repertoire. *Curr Opin Syst Biol* 18:36–43.
761 Wang Z, Yang X, Zhou Y, Sun J, Liu X, Zhang J, Mei X, Zhong J, Zhao J, Ran P. 2020. COVID-19
762 severity correlates with weaker T-Cell immunity, hypercytokinemia, and lung epithelium injury. *Am*
763 *J Respir Crit Care Med* 202:606–610.
764 Weiskopf D, Schmitz KS, Raadsen MP, Grifoni A, Okba NMA, Endeman H, van den Akker JPC,
765 Molenkamp R, Koopmans MPG, van Gorp ECM, Haagmans BL, de Swart RL, Sette A, de Vries
766 RD. 2020. Phenotype and kinetics of SARS-CoV-2-specific T cells in COVID-19 patients with
767 acute respiratory distress syndrome. *Sci Immunol* 5. doi:10.1126/sciimmunol.abd2071

Welsh RM, Selin LK. 2002. No one is naive: the significance of heterologous T-cell immunity. *Nat Rev Immunol* 2:417–426.

Wirasinha RC, Singh M, Archer SK, Chan A, Harrison PF, Goodnow CC, Daley SR. 2018. αβ T-cell receptors with a central CDR3 cysteine are enriched in CD8αα intraepithelial lymphocytes and their thymic precursors. *Immunol Cell Biol* 96:553–561.

Wolf K, Hether T, Gilchuk P, Kumar A, Rajeh A, Schiebout C, Maybruck J, Buller RM, Ahn T-H, Joyce S, DiPaolo RJ. 2018. Identifying and tracking low-frequency virus-specific TCR clonotypes using high-throughput sequencing. *Cell Rep* 25:2369-2378.e4.

FIGURE CAPTIONS

Figure 1. Experimental antigen-specific enrichment. (A) Repertoire subsets obtained by single-cell sorting with peptide-MHC tetramers (data from Dash et al. and Sewell et al. via VDJdb; greens), peptide stimulation enrichment (MIRA M55, M48; purples), or random sub-sampling of umbilical cord blood (1,000 or 10,000 TCRs; blues). Biochemical distances were computed among all pairs of TCRs in each subset using TCRdist. Neighborhoods were formed around each TCR using a variable radius (x-axis) and the percent of TCRs in the set that were within the neighborhood was computed. A radius of zero indicates the proportion of TCRs that have at least one TCR with an identical amino acid sequence (solid square). (B) Analysis of 18 MIRA-enriched repertoires for which the volunteers contributing the TCRs were significantly enriched with a specific class I HLA allele (Table S4).

Figure 2. Heterogeneous TCR neighborhoods within experimentally antigen-enriched and unenriched repertoire subsets. TCRs from (A) a peptide-MHC tetramer-enriched sub-repertoire, (B) a MIRA peptide stimulation-enriched sub-repertoire, or (C) an umbilical cord blood unenriched repertoire. Within each sub-repertoire, an empirical cumulative distribution function (ECDF) was estimated for each TCR (one line) acting as the centroid of a neighborhood over a range of distance radii (x-axis). Each ECDF shows the proportion of TCRs within the sub-repertoire found within the indicated radius from the centroid TCR. ECDF color corresponds to the number of amino acids in the TCR CDR3 loop. ECDF curves were randomly shifted by <1 unit along the x-axis to reduce overplotting. Vertical ECDF lines at 10-6 indicate no similar TCRs at or below that radius. Percentage of TCRs with an ECDF proportion < 10-3 (bottom panels), indicates the percentage of TCRs without, or with very few biochemically similar neighbors at the given radius.

Figure 3. Radius-defined neighborhood densities within an antigen-enriched and a synthetic background repertoire. (A) Each TCR in the MIRA55 antigen-enriched sub-repertoire (one line) acts as the centroid of a neighborhood and an empirical cumulative distribution function (ECDF) is estimated over a range of distance radii (x-axis). Each ECDF shows the proportion of TCRs within the sub-repertoire having a distance to the centroid less than the indicated radius. The ECDF line color corresponds to the TCR generation probability estimated using OLGA (Pgen). The ECDF curves are randomly shifted by <1 unit along the x-axis to reduce overplotting. The bottom panel shows the percentage of TCRs with an ECDF proportion < 10-3. (B) Estimated ECDF for each MIRA55 TCR based on the proportion of TCRs in a synthetic background repertoire that is within the indicated radius (x-axis). The synthetic background was generated using 100,000 OLGA-generated TCRs and 100,000 TCRs sub-sampled from umbilical cord blood; sampling was matched to the VJ-gene frequency in the MIRA55 sub-repertoire, with inverse probability weighting to account for the sampling bias (see Methods for details). (C) Antigen-enriched ECDF (y-axis) of one example TCR's neighborhood (red line) plotted against ECDF within the synthetic

background (x-axis). Example TCR neighborhood is the same indicated by the red line in (A) and (B). The dashed line indicates neighborhoods that are equally dense with TCRs from the antigen-enriched and unenriched background sub-repertoires.

Figure 4. TCR meta-clonotype framework and application. (A) Framework: sets of CD8+ TCRs activated by SARS-CoV-2 peptides were previously discovered in 61 individuals diagnosed with COVID-19 using a Multiplex Identification of Antigen-Specific T Cell Receptors Assay (MIRA) by Nolan et al. (2020). For each TCR clone activated by a given peptide, we used tcrdist3 to evaluate the repertoire fraction spanned at different TCRdist radii within (i) its antigen enriched repertoire (black) and (ii) a control V- and J-gene matched, inverse probability weighted background repertoire (purple). Within the radius (R) estimated to control a specified background TCR discovery frequency θ, we construct the set of all antigen enriched TCRs neighboring a given centroid. The diagram shows, for instance, the centroid TRBV28*01+CASSLKTNSYEQYF with θ set to 1e-6. The resulting logo plot depicts the conserved CDR3 β-chain motif formed from the other SARS-CoV-2 ORF1ab 4211:4252-activated TCRs within a TCRdist radius 16 of this centroid. (B) Application: we then applied public meta-clonotypes derived from antigen enriched repertoires to quantify the enrichment of public SARS-CoV-2 antigen-specific TCRs in a large diverse cohort, from whom unenriched bulk TCR repertoires were collected 0-30 days from COVID-19 diagnosis (n = 694). In almost all cases, the evidence of predicted HLA-mediated enrichment is stronger for public meta-clonotypes compared to exact TCR β-chain amino acid sequences. This is indicated by more statistically significant HLA coefficients in beta-binomial count regressions controlling for sequencing depth, subject age, sex, and days from diagnosis.

Figure 5. HLA restriction of TCR clonotypes and meta-clonotypes in bulk sequenced TCRβ repertoires of COVID-19 patients. (A) Percentage of TCR features with a statistically significant (FDR < 0.001) positive association of its abundance in COVID-19 repertoires and patients' expression of the restricting HLA allele. We tested for associations using beta-binomial regression controlling for sequencing depth, age, sex, and days since COVID-19 diagnosis. (B) For each clonotype/meta-clonotype, the percent of bulk repertoires from COVID-19 patients (n=694) containing TCRs meeting the criteria defined by (1) EXACT (TCRs matching the centroid TRBV gene and CDR3 amino acids), (2) RADIUS (TCR centroid with inclusion criteria defined by an optimized TCRdist radius, with θ set to 1e-6), or (3) RADIUS + MOTIF (inclusion criteria defined by TCR centroid, optimized radius and matching of CDR3 amino-acid residues at conserved positions). See Methods for details.

Figure 6. Associations of TCR features with participant age, days post diagnosis, HLA-genotype, and sex in TCR β-chain repertoires of COVID-19 patients (n=694). (A) Beta-binomial regression coefficient estimates (x-axis) and negative log10 false discovery rates (y-axis) for features developed from CD8+ TCRs activated by SARS-CoV-2 MIRA55: ORF1ab 5171:5203. The abundances of TCR meta-clonotypes are more robustly associated with predicted HLA type than exact clonotypes. (B) Signal strength of enrichment by participant HLA-type (2-digit) of TCR β-chain clonotypes (EXACT) and meta-clonotypes (RADIUS or RADIUS+MOTIF) predicted to recognize additional HLA-restricted SARS-CoV-2 peptides: (1) MIRA48: surface glycoprotein 22355:22393 (2) MIRA51: nucleocapsid phosphoprotein 29348:29380), (3) MIRA53: surface glycoprotein 22904:22936 (4) MIRA55: ORF1ab 4211:4252 (5) MIRA110: nucleocapsid phosphoprotein 29138:29176 (6) MIRA11: ORF1ab, ORF3a 3875:25593. Models were estimated with counts of TCR matching clonotypes (EXACT) or meta-clonotypes (RADIUS or RADIUS+MOTIF) with the following definitions: (1) EXACT (inclusion of TCRs matching the centroid TRBV gene and CDR3 amino acids), (2) RADIUS (TCR centroid with inclusion criteria defined by an optimized TCRdist radius), (3) RADIUS + MOTIF (inclusion criteria defined by TCR centroid, optimized

radius and matching of CDR3 amino acid residues at conserved positions within the meta-clonotype CDR3s). See methods for details.

Figure S1: Publicity analysis in MIRA participants of CD8+ TCR β-chain features activated by SARS-CoV-2 peptide ORF1ab 4211:4252 (MIRA55) predicted to bind HLA-A*01. TCR features publicity across individual was determined using two methods for clustering similar TCR sequences: (A) tcrdist3 meta-clonotyopes defined by a centroid TCR and all TCRs within an optimized radius chosen to span 1e-6 TCRs in a bulk unenriched TRBV-TRBJ matched background data, and (B) public clonotypes defined by identical TRBV gene usage and identical CDR3, at the amino acid level.

Figure S2: Publicity and breadth analysis of CD8+ TCR β-chain features activated by SARS-CoV-2 peptide ORF1ab 4211:4252 (MIRA55) using tcrdist3 and GLIPH2. TCR feature publicity was determined using two methods for clustering similar TCR sequences: (A) tcrdist3 meta-clonotypes (with radii based on θ=1Ee-6) and (B) GLIPH2-groups, sets of TCRs with a shared CDR3 k-mer pattern uncommon in the default background CD8+ receptor data (using default Fisher's p-value < 0.001). Grid fill color shows the breadth of clones clustered. See Methods for details.

Figure S3: Associations of TCR features with participant age, days-post diagnosis, HLA-type, and sex in bulk TCR β-chain repertoires of COVID-19 patients (n=694). TCR features shown here were identified from publicity analysis of CD8+ TCRs activated by an illustrative SARS-CoV-2 peptide ORF1ab 4211:4252 (MIRA55), which is predicted to bind HLA-A*01. Using beta-binomial regression models estimated for each feature, volcano plots show associations between participant characteristics and abundance of TCRs matching either: (A) EXACT clonotypes (TRBV+CDR3), (B) GLIPH2 patterns (TRBV+CDR3 k-mer), or (C) tcrdist3 based RADIUS+MOTIF meta-clonotype. See Methods for details.

Figure S4. Detectable HLA-association and CDR3 probability of generation. We evaluated meta-clonotypes from 18 MIRA sets in a cohort of 694 COVID-19 patients for their association with predicted HLA-restricting alleles. Evidence of the HLA association for each meta-clonotype (RADIUS or RADIUS+MOTIF) and the centroid alone (EXACT) is indicated by the associated false discovery rate adjusted p-value (FDR; y-axis) in beta-binomial regressions (see Methods for model details). The probability of generation (Pgen) of each centroid's CDR3-β was estimated using the software OLGA (x-axis). Given the extent of TCR diversity across individuals, population-scale analysis of exact antigen-specific clonotype abundance is likely limited to public (i.e., higher Pgen) TCR features. Using meta-clonotypes, *tcrdist3* revealed strong evidence of HLA-restriction for TCRs with both high and low Pgen.

Figure 1

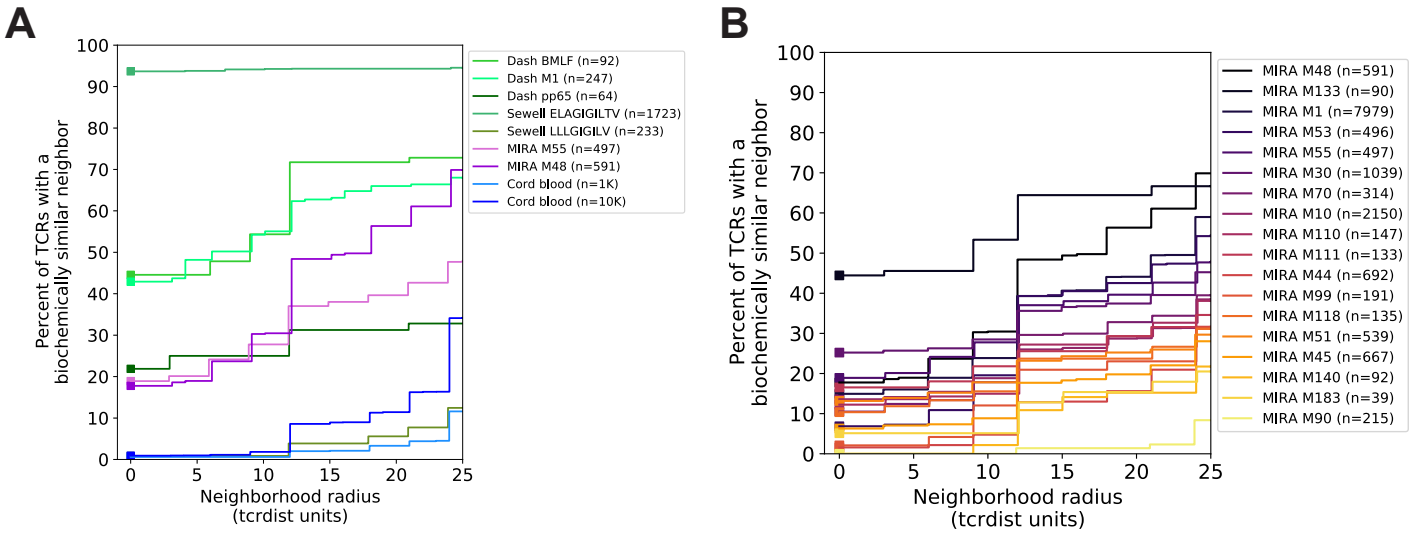


Figure 1. Experimental antigen-specific enrichment. (A) Repertoire subsets obtained by single-cell sorting with peptide-MHC tetramers (data from Dash et al. and Sewell et al. via VDJdb; greens), peptide stimulation enrichment (MIRA M55, M48; purples), or random sub-sampling of umbilical cord blood (1,000 or 10,000 TCRs; blues). Biochemical distances were computed among all pairs of TCRs in each subset using TCRdist. Neighborhoods were formed around each TCR using a variable radius (x-axis) and the percent of TCRs in the set that were within the neighborhood was computed. A radius of zero indicates the proportion of TCRs that have at least one TCR with an identical amino-acid sequence (solid square). (B) Analysis of 18 MIRA-enriched repertoires for which the volunteers contributing the TCRs were significantly enriched with a specific class I HLA allele (Table S4).

Figure 2

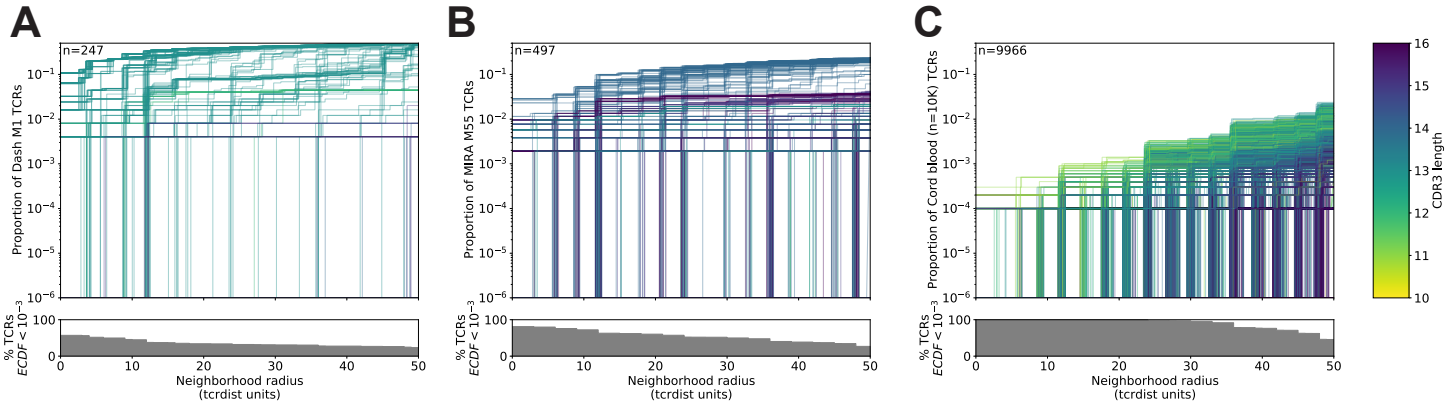


Figure 2. Heterogeneous TCR neighborhoods within experimentally antigen-enriched and unenriched repertoire subsets. TCRs from (A) a peptide-MHC tetramer-enriched sub-repertoire, (B) a MIRA peptide stimulation-enriched sub-repertoire, and (C) an umbilical cord-blood unenriched repertoire. Within each sub-repertoire, an empirical cumulative distribution function (ECDF) was estimated for each TCR (one line) acting as the centroid of a neighborhood over a range of distance radii (x-axis). Each ECDF shows the proportion of TCRs within the sub-repertoire found within the indicated radius from the centroid TCR. ECDF color corresponds to the number of amino acids in the TCR CDR3 loop. ECDF curves were randomly shifted by <1 unit along the x-axis to reduce overplotting. Vertical ECDF lines at 10-6 indicate no similar TCRs at or below that radius. Percentage of TCRs with an ECDF proportion < 10-3 (bottom panels), indicates the percentage of TCRs without, or with very few biochemically similar neighbors at the given radius.

Figure 3

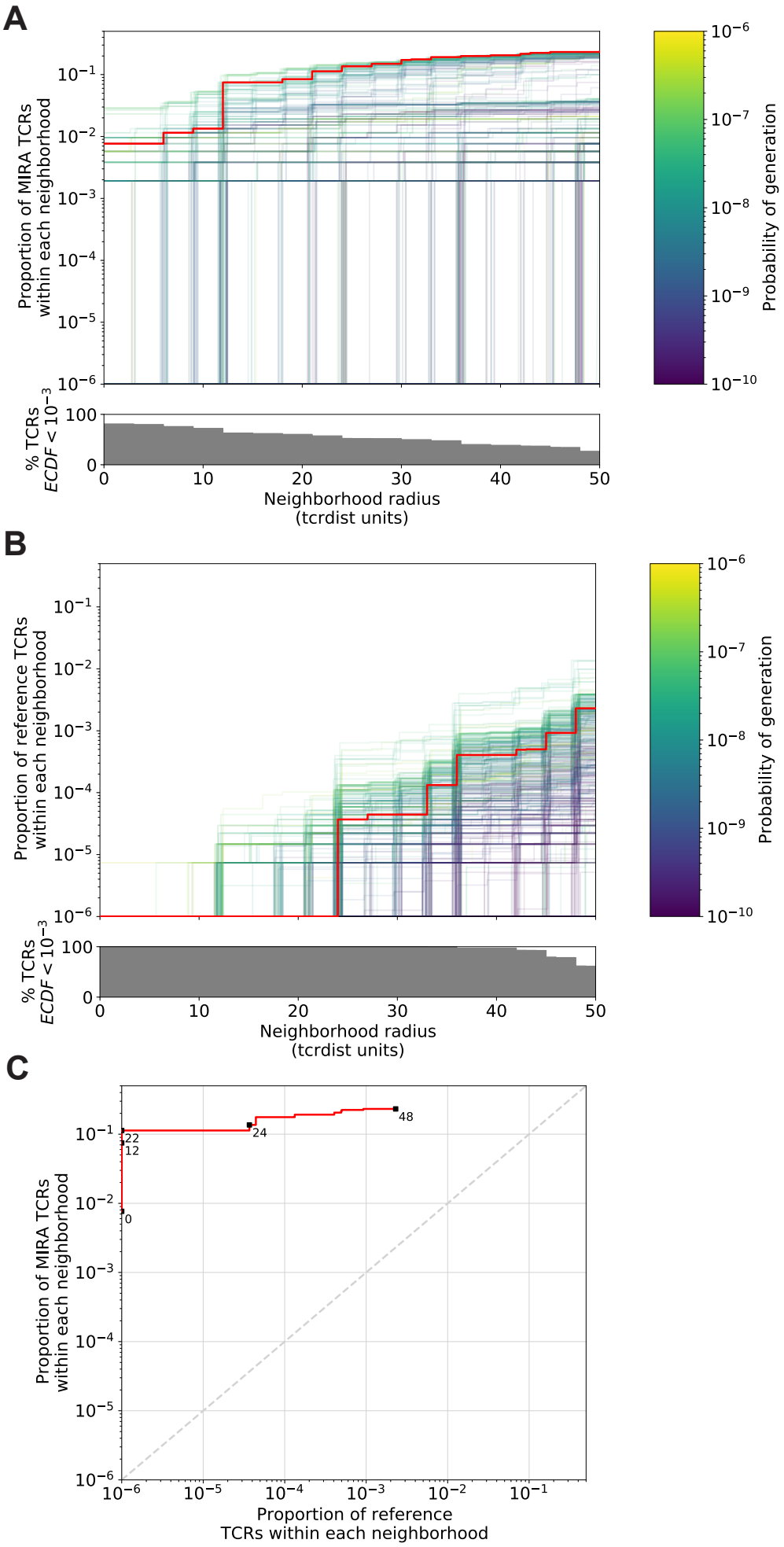
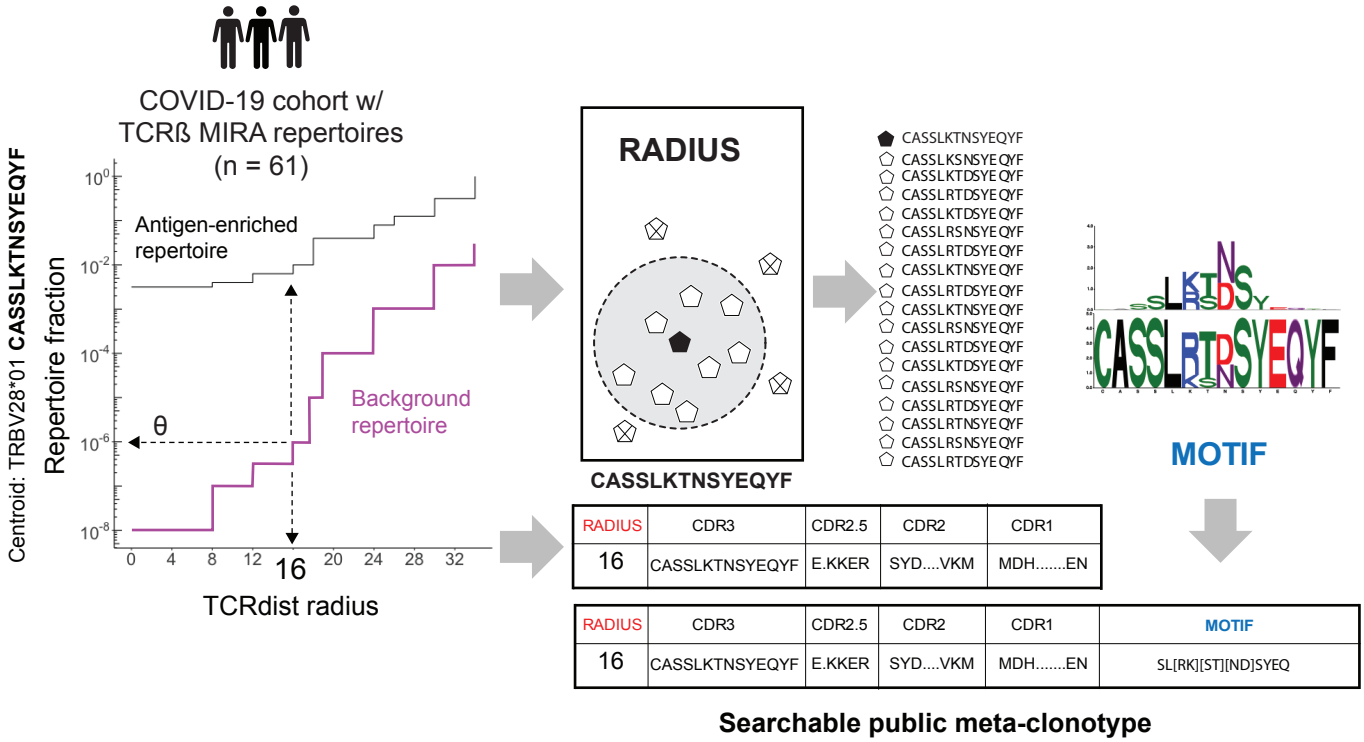


Figure 3. Radius-defined neighborhood densities within an antigen-enriched and a synthetic background repertoire. (A) Each TCR in the MIRA55 antigen-enriched sub-repertoire (one line) acts as the centroid of a neighborhood and an empirical cumulative distribution function (ECDF) is estimated over a range of distance radii (x-axis). Each ECDF shows the proportion of TCRs within the sub-repertoire having a distance to the centroid less than the indicated radius. The ECDF line color corresponds to the TCR generation probability estimated using OLGA (Pgen). The ECDF curves are randomly shifted by <1 unit along the x-axis to reduce overplotting. The bottom panel shows the percentage of TCRs with an ECDF proportion < 10-3. (B) Estimated ECDF for each MIRA55 TCR based on the proportion of TCRs in a synthetic background repertoire that is within the indicated radius (x-axis). The synthetic background was generated using 100,000 OLGA-generated TCRs and 100,000 TCRs sub-sampled from umbilical cord-blood; sampling was matched to the VJ-gene frequency in the MIRA55 sub-repertoire, with inverse probability weighting to account for the sampling bias (see Methods for details). (C) Antigen-enriched ECDF (y-axis) of one example TCR's neighborhood (red line) plotted against ECDF within the synthetic background (x-axis). Example TCR neighborhood is the same indicated by the red line in (A) and (B). The dashed line indicates neighborhoods that are equally dense with TCRs from the antigen-enriched and unenriched background sub-repertoires.

Figure 4

A TCR META-CLONOTYPE FRAMEWORK



B APPLICATION

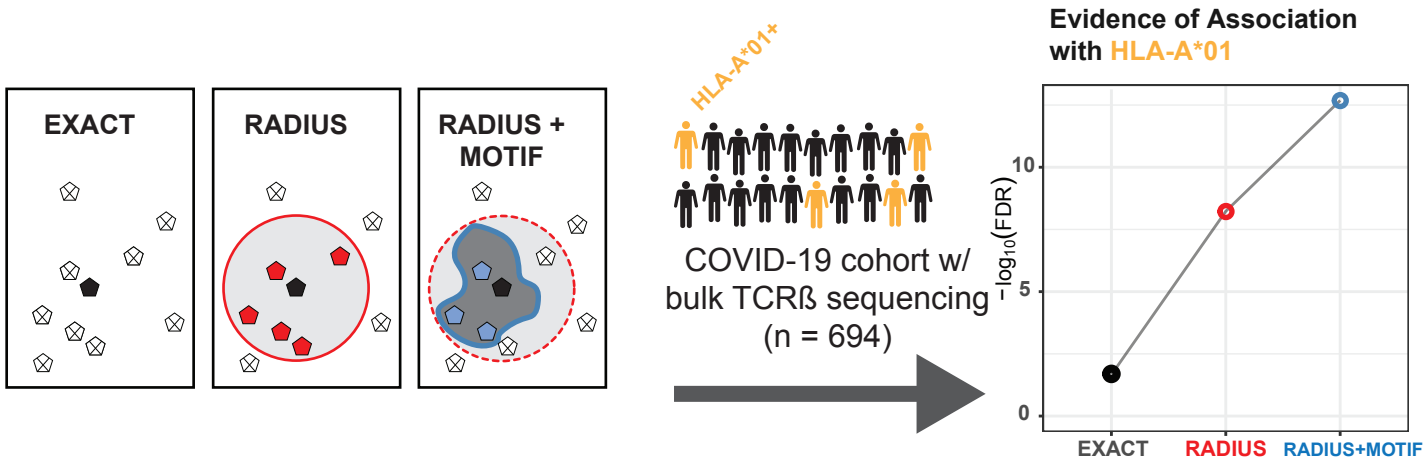


Figure 4. TCR meta-clonotype framework and application. (A) Framework: sets of CD8+ TCRs activated by SARS-CoV-2 peptides were previously discovered in 61 individuals diagnosed with COVID-19 using a Multiplex Identification of Antigen-Specific T Cell Receptors Assay (MIRA) by Nolan et al. (2020). For each TCR clone activated by a given peptide, we used *tcrdist3* to evaluate the repertoire fraction spanned at different TCRdist radii within (i) its antigen enriched repertoire (black) and (ii) a control V- and J-gene matched, inverse probability weighted background repertoire (purple). Within the radius (R) estimated to control a specified background TCR discovery frequency θ, we construct the set of all antigen enriched TCRs neighboring a given centroid. The diagram shows, for instance, the centroid TRBV28*01+CASSLKTNSYEQYF with θ set to 1e-6. The resulting logo plot depicts the conserved CDR3 β-chain motif formed from the other SARS-CoV-2 ORF1ab 4211:4252-activated TCRs within a TCRdist radius 16 of this centroid. (B) Application: we then applied public meta-clonotypes derived from antigen enriched repertoires to quantify the enrichment of public SARS-CoV-2 antigen-specific TCRs in a large diverse cohort, from whom unenriched bulk TCR repertoires were collected 0-30 days from COVID-19 diagnosis (n = 694). In almost all cases, the evidence of predicted HLA-mediated enrichment is stronger for public meta-clonotypes compared to exact TCR β-chain amino acid sequences. This is indicated by more statistically significant HLA coefficients in beta-binomial count regressions controlling for sequencing depth, subject age, sex, and days from diagnosis.

Figure 5

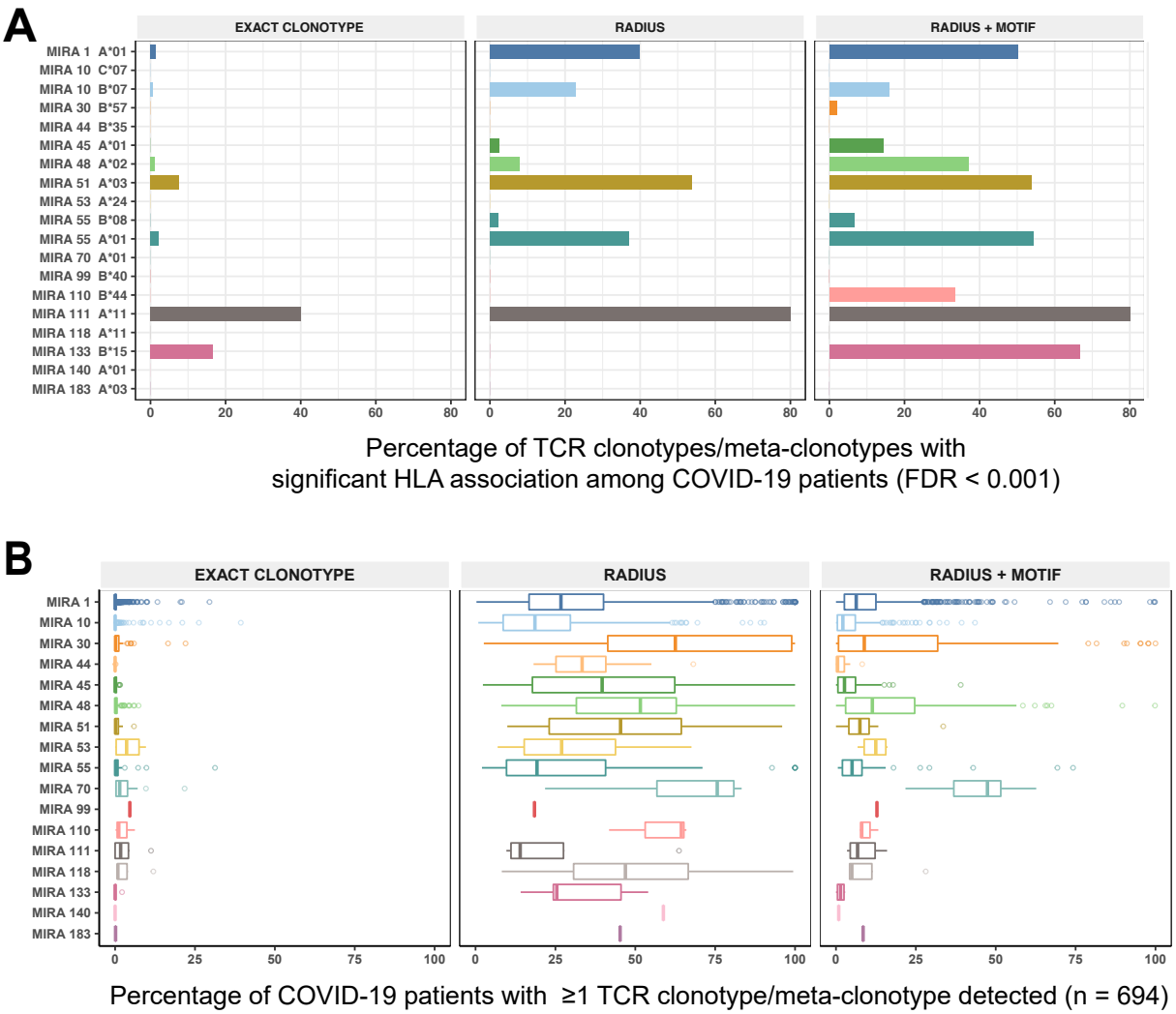


Figure 5. HLA restriction of TCR clonotypes and meta-clonotypes in bulk sequenced TCRβ repertoires of COVID-19 patients. (A) Percentage of TCR features with a statistically significant (FDR < 0.001) positive association of its abundance in COVID-19 repertoires and patients' expression of the restricting HLA allele. We tested for associations using beta-binomial regression controlling for sequencing depth, age, sex, and days since COVID-19 diagnosis. (B) For each clonotype/meta-clonotype, the percent of bulk repertoires from COVID-19 patients (n=694) containing TCRs meeting the criteria defined by (1) EXACT (TCRs matching the centroid TRBV gene and CDR3 amino acids), (2) RADIUS (TCR centroid with inclusion criteria defined by an optimized TCRdist radius, with θ set to 1e-6), or (3) RADIUS + MOTIF (inclusion criteria defined by TCR centroid, optimized radius and matching of CDR3 amino-acid residues at conserved positions). See Methods for details.

Figure 6

Figure 6. Associations of TCR features with participant age, days post diagnosis, HLA-genotype, and sex in TCR β-chain repertoires of COVID-19 patients (n=694). (A) Beta-binomial regression coefficient estimates (x-axis) and negative log10 false discovery rates (y-axis) for features developed from CD8+ TCRs activated by SARS-CoV-2 MIRA55: ORF1ab 5171:5203. The abundances of TCR meta-clonotypes are more robustly associated with predicted HLA type than exact clonotypes. (B) Signal strength of enrichment by participant HLA-type (2-digit) of TCR β-chain clonotypes (EXACT) and meta-clonotypes (RADIUS or RADIUS+MOTIF) predicted to recognize additional HLA-restricted SARS-CoV-2 peptides: (1) MIRA48: surface glycoprotein 22355:22393 (2) MIRA51: nucleocapsid phosphoprotein 29348:29380), (3) MIRA53: surface glycoprotein 22904:22936 (4) MIRA55: ORF1ab 4211:4252 (5) MIRA110: nucleocapsid phosphoprotein 29138:29176 (6) MIRA11: ORF1ab,ORF3a 3875:25593. Models were estimated with counts of TCR matching clonotypes (EXACT) or meta-clonotypes (RADIUS or RADIUS+MOTIF) with the following definitions: (1) EXACT (inclusion of TCRs matching the centroid TRBV gene and CDR3 amino acids), (2) RADIUS (TCR centroid with inclusion criteria defined by an optimized TCRdist radius), (3) RADIUS + MOTIF (inclusion criteria defined by TCR centroid, optimized radius and matching of CDR3 amino-acid residues at conserved positions within the meta-clonotype CDR3s) See methods for details.

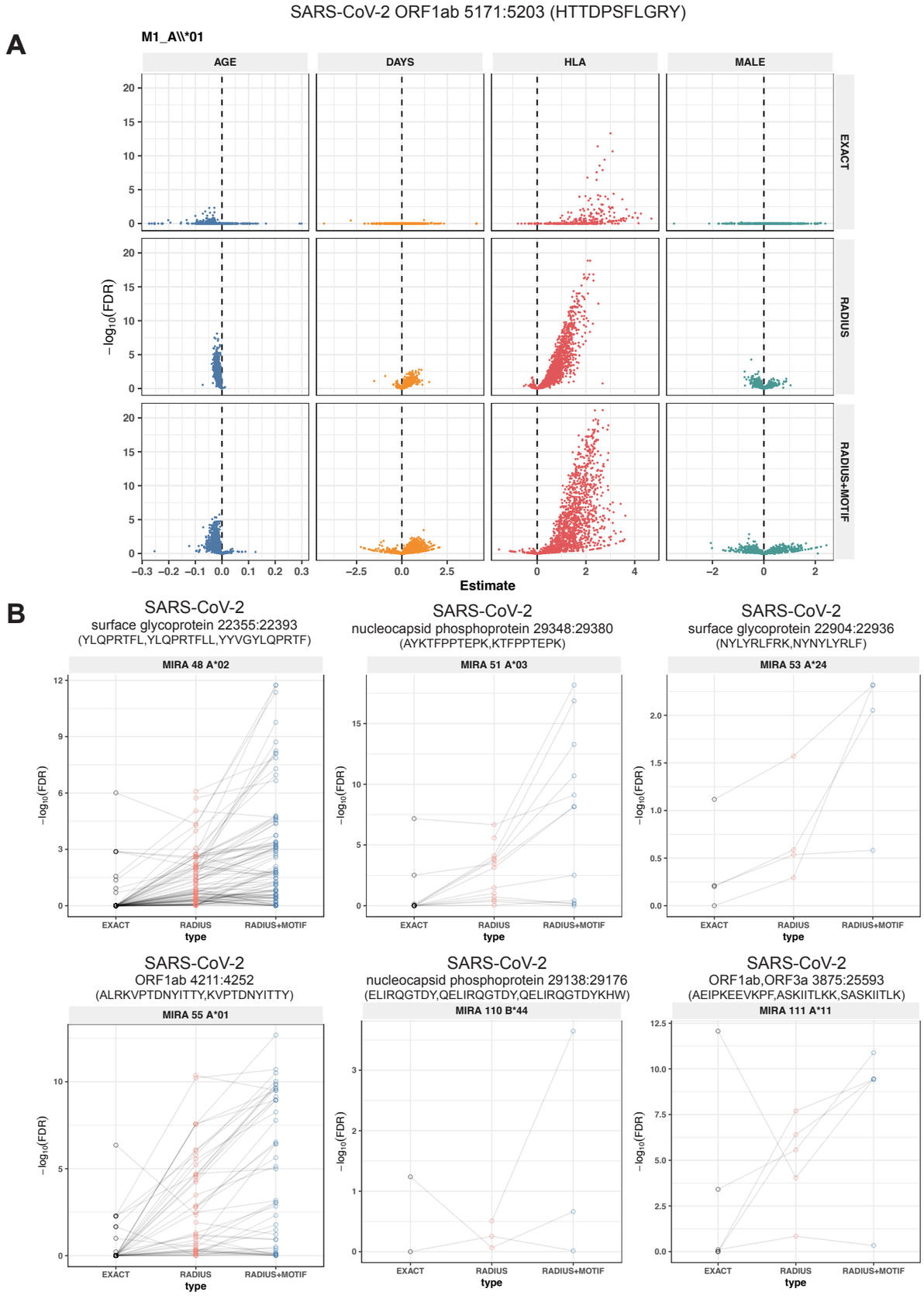
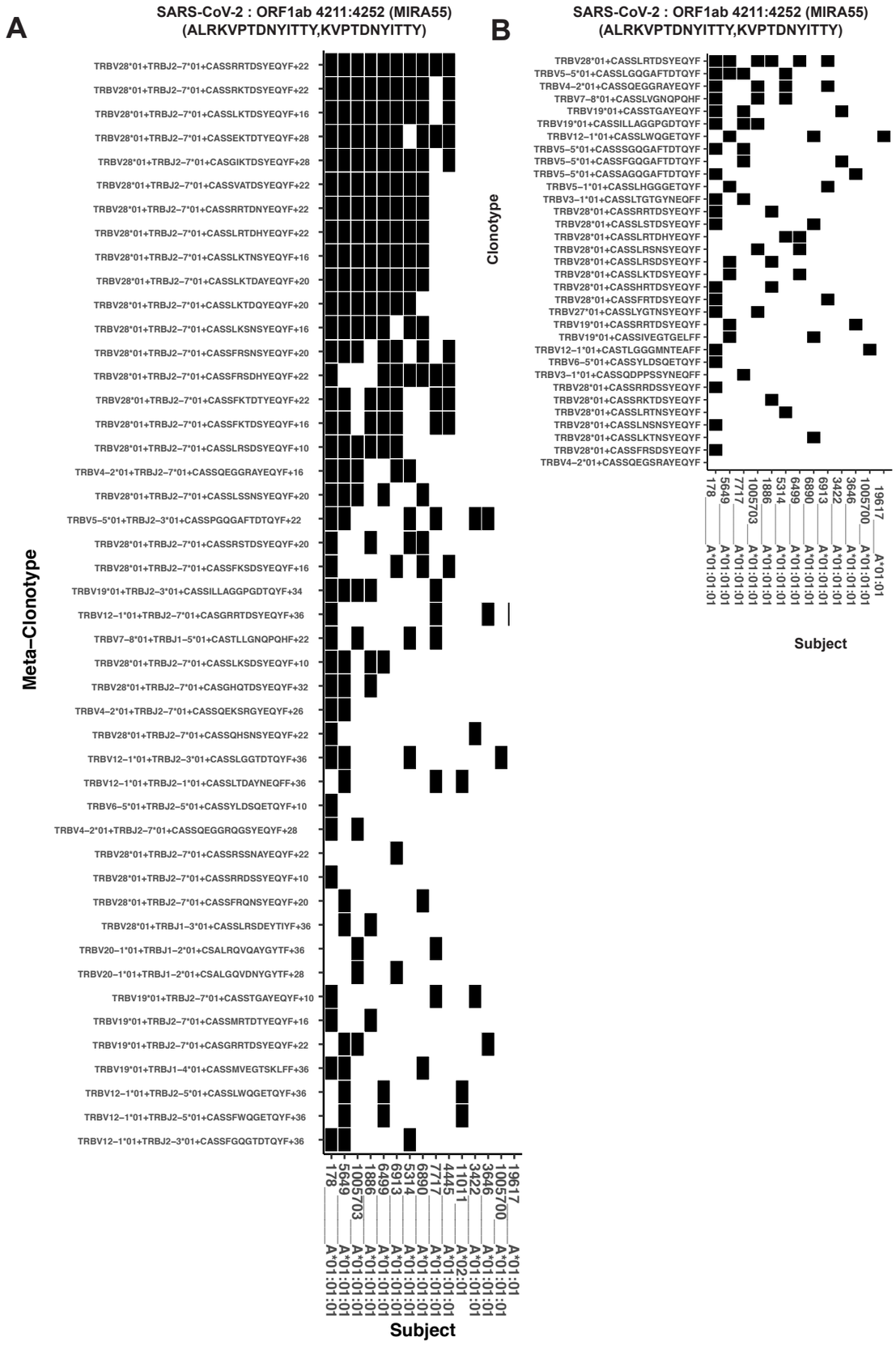


Figure S1

Figure S1: Publicity analysis in MIRA participants of CD8+ TCR β-chain features activated by SARS-CoV-2 peptide ORF1ab 4211:4252 (MIRA55) predicted to bind HLA-A*01. TCR features publicity across individual was determined using two methods for clustering similar TCR sequences: **(A)** tcrdist3 meta-clonotyopes defined by a centroid TCR and all TCRs within an optimized radius chosen to span 1e-6 TCRs in a bulk unenriched TRBV-TRBJ matched background data, and **(B)** public clonotypes defined by identical TRBV gene usage and identical CDR3, at the amino acid level.



t3/SI_F2.R

Figure S2

Figure S2: Publicity and breadth analysis of CD8+ TCR β-chain features activated by SARS-CoV-2 peptide ORF1ab 4211:4252 (MIRA55) using tcrdist3 and GLIPH2. TCR feature publicity was determined using two methods for clustering similar TCR sequences: (A) tcrdist3 meta-clonotypes (with radii based on θ =1Ee-6) and (B) GLIPH2-groups, sets of TCRs with a shared CDR3 k-mer pattern uncommon in the default background CD8+ receptor data (using default Fisher's p-value < 0.001). Grid fill color shows the breadth of clones clustered. See Methods for details.

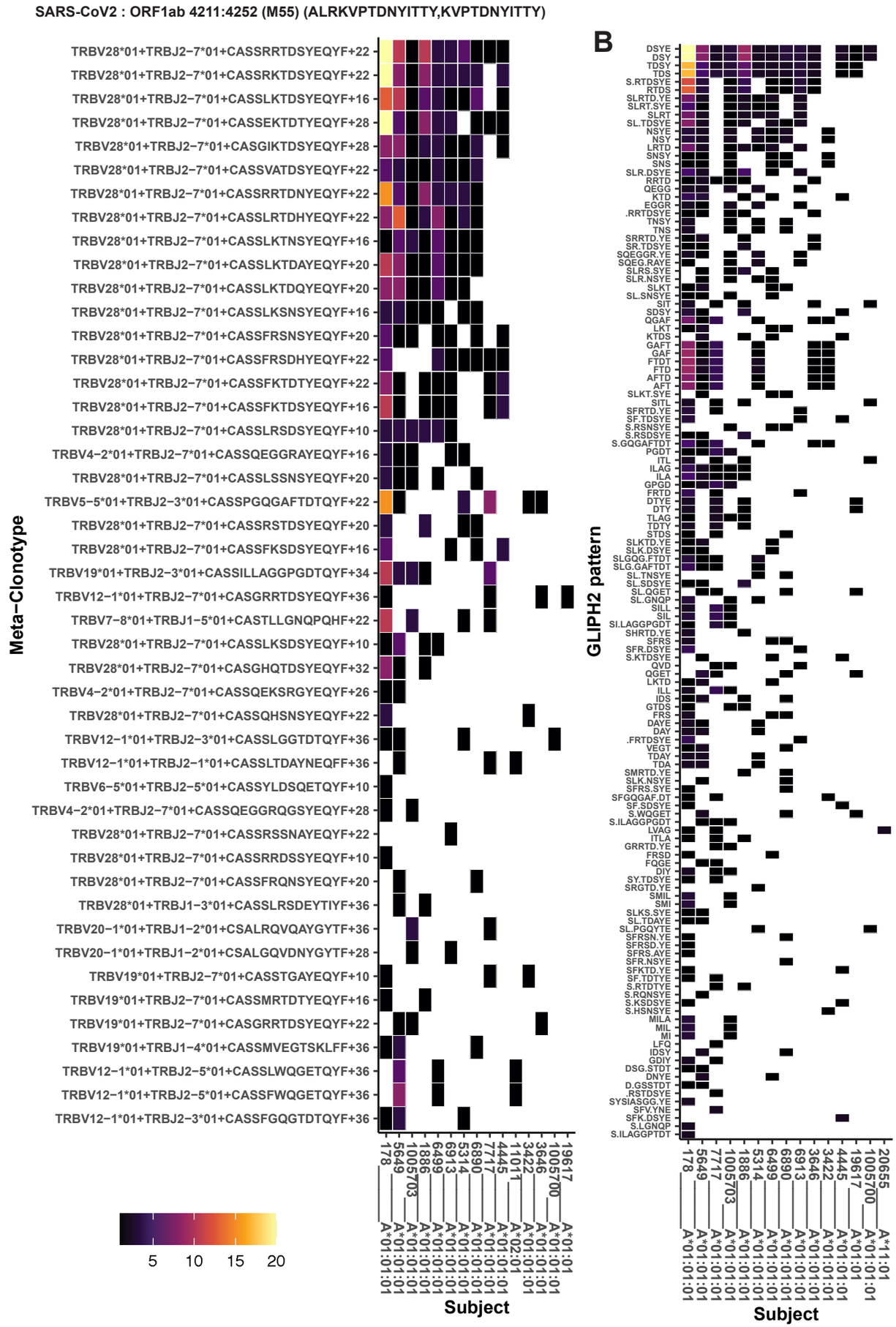


Figure S3

Figure S3 : Associations of TCR features with participant age, days-post diagnosis, HLA-type, and sex in bulk TCR β-chain repertoires of COVID-19 patients (n=694). TCR features shown here were identified from publicity analysis of CD8+ TCRs activated by an illustrative SARS-CoV-2 peptide ORF1ab 4211:4252 (MIRA55), which is predicted to bind HLA-A*01. Using beta-binomial regression models estimated for each feature, volcano plots show associations between participant characteristics and abundance of TCRs matching either: (A) EXACT clonotypes (TRBV+CDR3), (B) GLIPH2 patterns (TRBV+CDR3 k-mer), or (C) tcrdist3 based RADIUS+MOTIF meta-clonotype. See Methods for details.

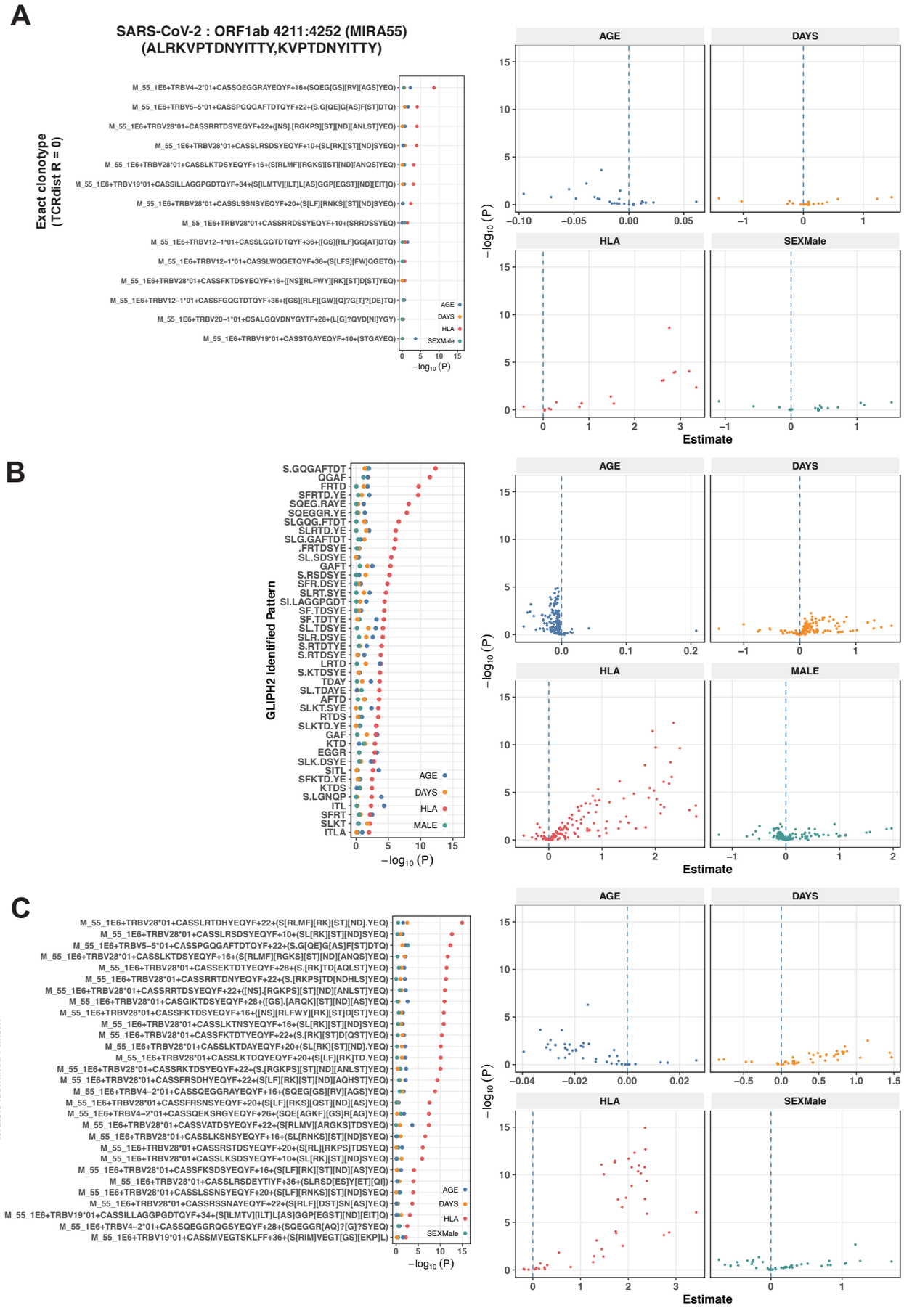


Figure S4

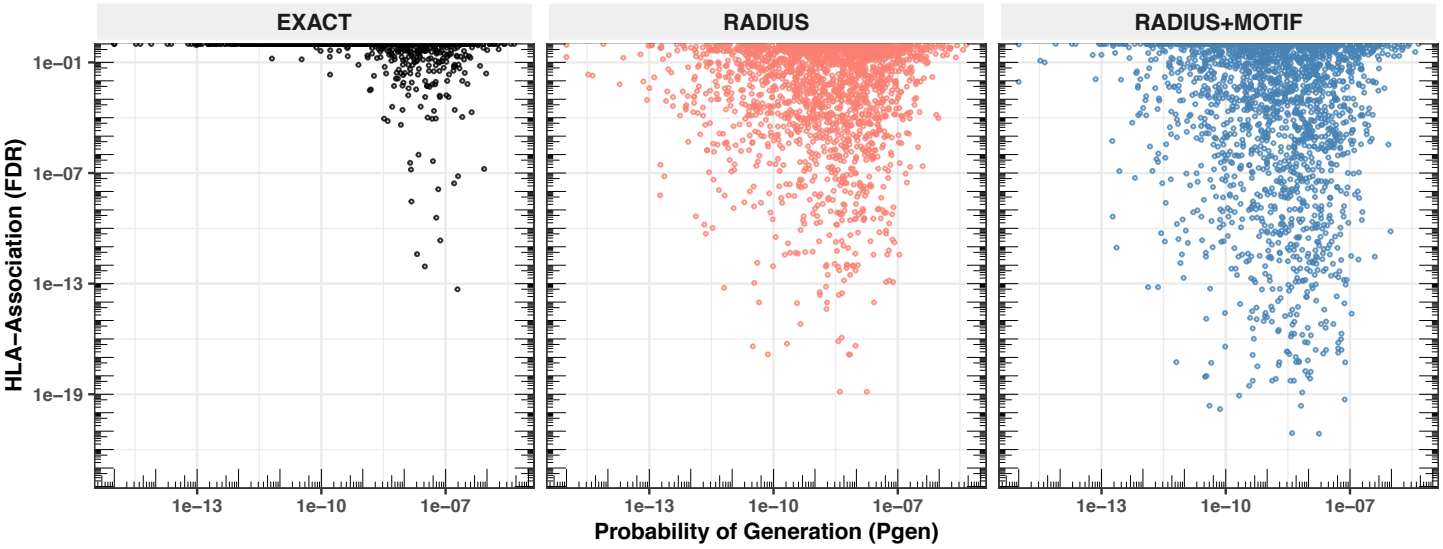


Figure S4. Detectable HLA-association and CDR3 probability of generation. We evaluated meta-clonotypes from 18 MIRA sets in a cohort of 694 COVID-19 patients for their association with predicted HLA-restricting alleles. Evidence of the HLA association for each meta-clonotype (RADIUS or RADIUS+MOTIF) and the centroid alone (EXACT) is indicated by the associated false discovery rate adjusted p-value (FDR; y-axis) in beta-binomial regressions (see Methods for model details). The probability of generation (Pgen) of each centroid's CDR3-β was estimated using the software OLGA (x-axis). Given the extent of TCR diversity across individuals, population-scale analysis of exact antigen-specific clonotype abundance is likely limited to public (i.e., higher Pgen) TCR features. Using meta-clonotypes, *tcrdist3* revealed strong evidence of HLA-restriction for TCRs with both high and low Pgen.