

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

3 Koshlan Mayer-Blackwell¹, Stefan Schattgen², Liel Cohen-Lavi^{3,4}, Jeremy Chase Crawford², Aisha
4 Souquette², Jessica A. Gaevert^{2,5}, Tomer Hertz⁶, Paul G. Thomas², Philip Bradley⁷, Andrew Fiore-
5 Gartland¹

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

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

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

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

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

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

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

17 **ABSTRACT**

18 As the mechanistic basis of adaptive cellular antigen recognition, T cell receptors (TCRs) encode
19 clinically valuable information that reflects prior antigen exposure and potential future response. However,
20 despite advances in deep repertoire sequencing, enormous TCR diversity complicates the use of TCR
21 clonotypes as clinical biomarkers. We propose a new framework that leverages antigen-enriched
22 repertoires to form meta-clonotypes – groups of biochemically similar TCRs – that can be used to robustly
23 quantify functionally similar TCRs in bulk repertoires. We apply the framework to TCR data from COVID-
24 19 patients, generating 1,915 public TCR meta-clonotypes from the 18 SARS-CoV-2 antigen-enriched
25 repertoires with the strongest evidence of HLA-restriction. Applied to independent cohorts, meta-
26 clonotypes targeting these specific epitopes were more frequently detected in bulk repertoires compared
27 to exact amino acid matches, and 44% (845/1915) were significantly enriched among COVID-19 patients
28 that expressed the putative restricting HLA allele, demonstrating the potential utility of meta-clonotypes as
29 antigen-specific features for biomarker development. To enable further applications, we developed an
30 open-source software package, *tcrdist3*, that implements this framework and facilitates workflows for
31 distance-based TCR repertoire analysis.

32 INTRODUCTION

33 An individual's unique repertoire of T cell receptors (TCRs) is shaped by antigen exposure and is
34 a critical component of immunological memory, contributing to recall responses against future infectious
35 challenges (Emerson et al., 2017; Welsh and Selin, 2002). With the advancement of immune repertoire
36 profiling, TCR repertoires are a largely untapped source of biomarkers that could potentially be used to
37 predict immune responses to a wide range of exposures including microbial infections (Wolf et al., 2018),
38 tumor neoantigens (Ahmadzadeh et al., 2019; Kato et al., 2018), or environmental allergens (Cao et al.,
39 2020). The TCR repertoire is characterized by its extreme diversity, originating from the genomic V(D)J
40 gene recombination of receptors in development. Between 10^9 - 10^{10} unique clonotypes - T cells with
41 distinct nucleotide-encoded receptors - are maintained in an adult human TCR repertoire (Lythe et al.,
42 2016). The diversity, both within and between individuals, presents major hurdles to biomarker
43 development. This is further complicated by the breadth of potential TCRs able to recognize even a single
44 antigen (Meysman et al., 2019), hampering detection of population-wide signatures of antigen exposure.
45 Indeed, mathematical modeling suggests that only 10-15% of TCRs are public or shared frequently by
46 multiple individuals (Elhanati et al., 2018), which is consistent with observations from extremely deeply
47 sequenced human repertoires (Soto et al., 2019). Despite advances in high-throughput next-generation
48 TCR amplicon sequencing, only a fraction of the repertoire can be assayed, making it difficult to
49 reproducibly sample many relevant TCR clonotypes from an individual, let alone reliably detect public
50 clonotypes in a population. In practice, the problem is exacerbated by unequal sampling depth. Thus,
51 individual T cell clonotypes are currently sub-optimal and under-powered for population-level
52 investigations of TCR specificity, which limits their application in the development of TCR-based clinical
53 biomarkers.

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

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

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

99 One recent study to elucidate the role of cellular immune responses in acute SARS-CoV-2
100 infection examined the T cell receptor repertoires of patients diagnosed with COVID-19 disease.
101 Researchers used an assay based on antigen stimulation and flow cytometric sorting of activated CD8+ T
102 cells to sequence SARS-CoV-2 peptide-associated TCR β -chains; the assay is called “multiplex
103 identification of T-cell receptor antigen specificity” or MIRA (Klinger et al., 2015). Data from these
104 experiments were released publicly in July 2020 by Adaptive Biotechnologies and Microsoft as part of
105 ‘immuneRACE’ and their efforts to stimulate science on COVID-19 (Nolan et al., 2020; Snyder et al.,
106 2020). The MIRA antigen enrichment assays identified 269 sets of TCR β -chains associated with CD8+ T
107 cells activated by exposure to SARS-CoV-2 peptides, with TCR sets ranging in size from 1 - 16,607 TCRs
108 (Table S1). The deposited ImmuneRace datasets also included bulk TCR β -chain repertoires from 694
109 patients within 0-30 days of COVID-19 diagnosis. To demonstrate potential uses of our new analytical
110 tools for TCR repertoire analysis and to accelerate understanding of the cellular responses to SARS-CoV-
111 2 infection, we present analyses of these data with a focus on an integration of the peptide-associated
112 MIRA TCR repertoires with bulk repertoires from four COVID-19 observational studies that enrolled
113 patients with diversity in age and geography (Alabama, USA n = 374; Madrid, Spain, n=117; Pavia, Italy,
114 n=125; Washington, USA, n=78).

115 116 **FRAMEWORK**

117
118 *Experimental antigen-enrichment allows discovery of TCRs with biochemically similar neighbors*
119

120 Searching for identical TCRs within a repertoire - arising either from clonal expansion or
121 convergent nucleotide encoding of amino acids in the CDR3 - is a common strategy for identifying
122 functionally important receptors. However, in the absence of experimental enrichment procedures,
123 observing T cells with the same amino acid TCR sequence in a bulk sample is rare. For example, in
124 10,000 β -chain TCRs from an umbilical cord blood sample, less than 1% of TCR amino acid sequences

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

133 We investigated the degree to which the MIRA enrichment strategy employed by Nolan et al.
134 (2020) identified TCRs with identical or similar amino acid sequences. In general, across multiple MIRA
135 TCR β -chain antigen-enriched repertoires, the proportion of amino acid TCR sequences observed more
136 than once was generally lower than in the tetramer-enriched repertoires and varied considerably across
137 the sets; some MIRA sets resembled tetramer-sorted sub-repertoires (Figure 1B; see MIRA133), while
138 others were more similar to unenriched repertoires (Figure 1B; see MIRA90). The increased diversity in
139 MIRA-enriched TCR sets versus tetramer-enriched TCR sets may, in part, be explained by: (i) recruitment
140 of lower affinity receptors, (ii) the dependence of MIRA on a diverse set of native host MHC presentation
141 molecules, compared to a single peptide-MHC complex, or (iii) bystander activation in the MIRA
142 stimulation assay.

143

144 *TCR biochemical neighborhood density is heterogeneous in antigen-enriched repertoires*

145

146 We next investigated the proportion of unique TCRs with at least one biochemically similar
147 neighbor among TCRs with the same putative antigen specificity. We and others have shown that a
148 single peptide-MHC epitope is often recognized by many distinct TCRs with closely related amino acid
149 sequences (Dash et al., 2017); in fact, detection of such clusters in bulk-sequenced repertoires is the
150 basis of several existing tools: GLIPH (Glanville et al., 2017; Huang et al., 2020), ALICE (Pogorelyy et al.,
151 2019) and TCRNET (Ritvo et al., 2018). Therefore, to better understand new large-scale antigen-enriched
152 datasets, like the SARS-CoV-2 MIRA data, we next evaluated the TCR biochemical neighborhoods,
153 defined for each TCR as the set of similar TCRs whose sequence divergence is within a specified radius.
154 We measured biochemical divergence using a position weighted, multi-CDR TCR distance metric (see
155 Methods for details of *tcrdist3* re-implementation of TCRdist). As the radius about a TCR centroid
156 expands, the number of TCRs it encompasses naturally increases. As a neighborhood radius extends,
157 the number of proximal TCRs increases more rapidly in antigen-enriched repertoires compared to the
158 unenriched repertoires.

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

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

175 Neighbor densities for individual TCRs within MIRA identified antigen-enriched repertoires were
176 highly heterogeneous. Densities for an illustrative MIRA set are shown in Figure 2 (MIRA55:ORF1ab
177 4211:4252; peptide ALRKVPTDNYITTY). Within this antigen-enriched repertoire, at 24 tdus, 8.9%
178 (44/497) of TCR neighborhoods included >10% of the other antigen-activated CD8+ TCRs (Figure 2B).
179 As expected, TCR neighborhoods in the umbilical cord blood repertoire were sparser (Figure 2C); the
180 densest neighborhood included only 0.13% of the repertoire at 24 tdus. We also noted that TCRs with
181 empty neighborhoods tended to have longer CDR3 loops. This is consistent with mathematical modeling
182 approaches that show that TCRs with shorter CDR3 loops have a higher generation probability (P_{gen})
183 during genomic recombination of the TCR locus (Marcou et al., 2018; Murugan et al., 2012; Sethna et al.,
184 2019). Absent strong selection for antigen recognition, TCRs with a low generation probability are thus
185 more likely to have a less dense biochemical neighborhood. Together, these observations suggest that
186 biochemical neighborhood density is highly heterogeneous among TCRs and that it may depend on
187 mechanisms of antigen-recognition as well as receptor V(D)J recombination biases (Thomas and
188 Crawford, 2019).

189
190 *Biochemical neighborhood radius can be tuned to balance a biomarker's sensitivity and specificity*

191
192 The utility of a TCR-based biomarker depends on the antigen-specificity of the TCRs. Therefore,
193 a key constraint on distance-based clustering is the presence of similar TCR sequences that may lack the
194 ability to recognize the target antigen. To be useful, a biochemical neighborhood definition should be wide
195 enough to capture multiple biochemically similar TCRs with shared antigen-recognition, but not
196 excessively broad as to include a high number of sequences found in background repertoires that are
197 antigen naive. Because the density of neighborhoods around TCRs are heterogeneous, we hypothesize
198 that the optimal radius defining a meta-clonotype may differ for each TCR. To find an ideal radius we
199 proposed comparing the relative density of a radius-defined target TCR neighborhood in the antigen-
200 enriched sub repertoire (Figure 3A) to the density of the radius-defined neighborhood in an unenriched
201 background repertoire (Figure 3B, 3C). This is similar to previous approaches taken by tools like ALICE
202 and TCRNET, except that we employ a biochemically informed distance measure (TCRdist) and tune the
203 radius around each TCR to balance the antigen-enriched and unenriched neighborhood densities. The
204 radius around each TCR defines a meta-clonotype that can be used to search for and quantify the
205 abundance of conformant sequences in bulk repertoires (Figure 4A, 4B). For each TCR, its radius-defined
206 meta-clonotype tends to be more abundant within a repertoire and more prevalent in a population than
207 the exact clonotype; for example, multiple TCR meta-clonotypes formed from the MIRA55:ORF1ab set
208 were detected in 13 of 15 HLA-A*01 participants in the MIRA cohort, whereas the centroid TCRs from
209 each of the meta-clonotypes were consistently less prevalent (Figure S1).

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

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

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

247

248 RESULTS

249

250 *Engineering meta-clonotype features for SARS-CoV-2*

251

252 The MIRA antigen enrichment assays identified 269 sets of TCR β -chains associated with
253 recognition of a SARS-CoV-2 antigen using CD8+ T cell enriched PBMC samples from 68 COVID-19
254 diagnosed patients. Of these, 252 included at least 6 unique TCRs (unique TRBV-CDR3 amino acid
255 sequences; referred to as MIRA1 - MIRA252; Table S1). All TCR clonotypes in the MIRA enriched
256 repertoires, defined by identical TRBV gene and CDR3 at the amino acid level, were initially considered
257 as candidate centroids; only 2.7% of the clonotypes were found in more than one MIRA participant. For
258 each candidate TCR, a meta-clonotype was engineered by selecting the maximum distance radius that
259 controlled the estimated number of neighboring TCRs in a bulk unenriched repertoire to less than 1 in
260 10^6 , estimated using an inverse probability weighted antigen-naive background repertoire (see Methods).
261 We then ranked the meta-clonotypes by their sensitivity approximated as the proportion of a centroid's
262 MIRA-enriched repertoire spanned by the search radius (diagrammed in Figure 4). Redundant, lower-
263 ranked meta-clonotypes were eliminated if they were completely encompassed by a higher-ranked meta-
264 clonotype. We further required that meta-clonotypes be public, including sequences from at least two
265 subjects in the MIRA cohort. We found that 102 of the 252 MIRA sets (Table S6) had sufficiently similar

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

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

283

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

285

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

298 Of the COVID-19 patients' samples screened using the MIRA assay, HLA genotype was available
299 for 47 patients and only a subset of patients contributed TCRs to each of the MIRA sets. As a second
300 indicator of HLA restriction, we tested whether the subgroup of patients contributing TCRs to each MIRA
301 set was enriched with individuals expressing specific HLA class I alleles (2-digit resolution) (Table S4).
302 We found that for 18 of the MIRA sets, the patients contributing TCRs were significantly enriched for at
303 least one HLA allele (Fisher's exact test $p < 0.001$). In one case, all 13 patients expressing an A*01 allele
304 and only 2 of 34 patients not expressing A*01, contributed to the MIRA55 TCR set ($p = 1e-7$); as noted
305 above, A*01 was also strongly predicted by NetMHCpan4.0 to bind the MIRA55 peptides. Similar patterns
306 of enrichment and predicted binding were seen with A*01 expressing individuals and recognition of
307 MIRA1:ORF1ab 5171:5203 (HTTDP SFLGRY, $p = 1.9e-7$) and MIRA45:ORF3a 25996:26037
308 (SYFTSDYYQ, $p = 1.9e-7$). Notably, for all 18 MIRA sets, the enriched participant HLA allele was also
309 predicted to bind the stimulating peptide ($IC_{50} < 500$ nM), which provided strong evidence of the HLA
310 allele restricting the TCRs in the MIRA antigen-enriched sub repertoires (Table S5).

311

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

313

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

331

332 The models revealed that there were radius-defined meta-clonotypes with a strong positive and
333 statistically significant association (FDR < 0.001) for 10 of the 18 HLA-restricted-MIRA sets that were
334 evaluated (Figure 5A, Table S7). Across all MIRA sets, a significant HLA-association was detected for
335 29% (657/1915) and 43% (845/1915) of the meta-clonotypes using the RADIUS or RADIUS+MOTIF
336 definitions, respectively. In comparison, strong HLA-association was detected in fewer than 2% (27/1915)
337 of exact clonotype features, largely because the specific TRBV gene and CDR3 sequences discovered in
338 the MIRA experiments were infrequently observed in unenriched bulk samples (Figure 5B). When
339 detectable, the abundance of exact TCR clonotypes in bulk repertoires tended to be positively associated
340 with expression of the restricting HLA allele, as hypothesized. However, in most cases, the associated
341 false discovery rate-adjusted q-value of these associations were orders of magnitude higher (i.e., less
342 significant) than those obtained from using the engineered RADIUS or RADIUS+MOTIF feature with the
343 same clonotype as a centroid (Figure 6B). The improved performance of meta-clonotypes as query
344 features is particularly evident when testing for HLA-associated enrichment of TCRs recognizing
345 immunodominant MIRA1:A*01 (Figure 5A, Figure 6A), MIRA48:A*02, MIRA51:A*03, MIRA53:A*24, and
346 MIRA55:A*01 (Figure 6B). Moreover, the regression models with meta-clonotypes also revealed possible
347 negative associations between TCR abundance and participant age and positive associations with
348 sample collection more than two days post COVID-19 diagnosis (Figure 6A).

348

349 **DISCUSSION**

350

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

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

365 To evaluate the potential clinical relevance of radius-defined meta-clonotypes we focused on
366 those with the strongest evidence of HLA restriction (Table S7, $n = 1915$). We reasoned that we could
367 compare the abundance of these meta-clonotypes in COVID-19 patients with and without the restricting
368 HLA and that a significant positive association of abundance with expression of the restricting allele would
369 provide confirmatory evidence both of the SARS-CoV-2 specificity of the meta-clonotype and its HLA
370 restriction (Figure 4B). Overall, we found confirmation of HLA-restriction of meta-clonotype abundance for
371 a majority of the MIRA sets we analyzed (11/18) and nearly one-third of all engineered meta-clonotypes
372 (44% using the RADIUS+MOTIF approach). To demonstrate the possibility of employing other
373 complementary tools to generate public TCR features, we applied GLIPH2 to one of the HLA-restricted
374 MIRA sets (MIRA55:ORF1ab; see Methods for details). Some of the GLIPH2 k-mers enriched in MIRA55
375 TCRs showed evidence of HLA-restriction, supporting the general applicability of using antigen-enriched
376 repertoires to discover generalizable features of otherwise private antigen-recognizing TCRs (Figures S2
377 and S3). With *tcrdist3*, we also found meta-clonotypes with significantly higher abundance in samples that
378 were provided more than two days after COVID-19 diagnosis, which is consistent with expansions of
379 virus-specific TCRs that would be typical of responses to viral infection and have been shown
380 preliminarily for SARS-CoV-2 (Weiskopf et al., 2020). This further demonstrated the potential clinical
381 relevance of meta-clonotypes and suggests they could be used to study SARS-CoV-2 T cell response
382 kinetics longitudinally.

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

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

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

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

430

431 **METHODS**

432

433 *TCR Data: immuneRACE datasets and MIRA assay*

434

435 The study utilized two primary sources of TCR data (Nolan et al. 2020; Snyder et al. 2020). The
436 first data source was a table of TCR β -chains amplified from CD8+ T cells activated after exposure to a
437 pool of SARS-CoV-2 peptides, using a Multiplex Identification of Receptor Antigen (MIRA) (Klinger et al.
438 2015). The samples used for the MIRA analysis included samples from 61 individuals diagnosed with
439 COVID-19, of whom 47 were HLA-genotyped. We analyzed the 252 MIRA sets with at least 6 unique
440 TCRs, referred to as M1-252 in rank order by their size (Table S1). Adaptive Biotechnologies also made
441 publicly available bulk unenriched TCR β -chain repertoires from COVID-19 patients participating in a
442 collaborative immuneRACE network of international clinical trials. We selected repertoires from
443 individuals where meta-data was available indicating that the sample was collected from 0 to 30 days
444 from the time of diagnosis. (COVID-19-DLS (Alabama, USA n = 374); COVID-19-HUniv12Oct (Madrid,
445 Spain n = 117); COVID-19-NIH/NIAID (Pavia, Italy n=125) + COVID-19-ISB (Washington, USA n = 78).
446 The sampling depth of these repertoires varied from 15,626-1,220,991 productive templates (median
447 208,709) and 10,395-1,038,012 productive rearrangements (median 113,716). We did not use bulk
448 samples from the COVID-19-ADAPTIVE dataset as the average age was lower than other immuneRACE
449 populations and some of the participants overlap with individuals in the Adaptive led MIRA-based antigen
450 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

545 data. We used an R package, *corncob*, which provides maximum likelihood methods for inference and
546 hypothesis testing with beta-binomial regression models (Martin et al., 2020). Due to the sparsity of some
547 meta-clonotypes, seven percent of coefficient estimates in regression models had p-values greater than
548 0.99 and unreliable high magnitude coefficient estimates. These values are not shown in the horizontal
549 range of the volcano plots.

550

551 *Creation of k-mer based TCR features with GLIPH2*

552

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

563

564 *tcrdist3: Software for TCR repertoire analysis*

565

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

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

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

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

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

606

607 SUPPLEMENTAL TABLES

608	Table S1	MIRA enriched repertoires MIRA0 - MIRA252
609	Table S2	HLA class I alleles capable of presenting the SARS-CoV-2 associated peptides MIRA 610 screen
611	Table S3	NetMHCpan4.0 peptide MHC class I binding affinity prediction
612	Table S4	Statistical associations between common HLA genotypes of COVID-19 exposed MIRA 613 participants and SARS-CoV-2 peptide-enriched TCR repertoires
614	Table S5	Peptide-associated repertoires by MIRA that showed both strong in silico evidence of 615 HLA-restriction and were highly associated with participants expressing a predicted MHC 616 Class 1 binder
617	Table S6	SARS-CoV-2 CD8+ meta clonotypes summarized by MIRA enriched repertoire
618	Table S7	Table S7 SARS-CoV-2 CD8+ meta clonotypes with strong evidence of HLA restriction (n = 619 1915)
620	Table S8	SARS-CoV-2 CD8+ meta clonotypes with less evidence of HLA restriction (n = 4562)
621	Table S9	Comparison of selected software tools for clustering TCRs

622

623 SUPPLEMENTAL FIGURES

624

625	Figure S1	Publicity analysis of CD8+ TCR β -chain features activated by SARS-CoV-2 peptide 626 ORF1ab 4211:4252.
627	Figure S2	Publicity and breadth analysis of CD8+ TCR β -chain features activated by 628 SARS-CoV-2 peptide ORF1ab 4211:4252 (MIRA55) using tcrdist3 and GLIPH2.
629	Figure S3	Associations of TCR features with participant age, days post diagnosis, HLA-type, and 630 sex in bulk TCR β -chain repertoires of COVID-19 patients (n=694).
631	Figure S4	Detectable HLA-association and CDR3 probability of generation.

632 DATA AVAILABILITY

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

635 **SOFTWARE AVAILABILITY**

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

640

641 **CONTRIBUTIONS**

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

646 **ACKNOWLEDGEMENTS**

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

648 **REFERENCES**

649

- 650 Ahmadzadeh M, Pasetto A, Jia L, Deniger DC, Stevanović S, Robbins PF, Rosenberg SA. 2019. Tumor-
651 infiltrating human CD4+ regulatory T cells display a distinct TCR repertoire and exhibit tumor and
652 neoantigen reactivity. *Sci Immunol* 4. doi:10.1126/sciimmunol.aao4310
- 653 Britanova OV, Shugay M, Merzlyak EM, Staroverov DB, Putintseva EV, Turchaninova MA, Mamedov IZ,
654 Pogorelyy MV, Bolotin DA, Izraelson M, Davydov AN, Egorov ES, Kasatskaya SA, Rebrikov DV,
655 Lukyanov S, Chudakov DM. 2016. Dynamics of individual T Cell repertoires: from cord blood to
656 centenarians. *The Journal of Immunology* 196:5005–5013.
- 657 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,
658 Fang M, Yue Y, Wang X, Ding Y, Zhou L, Zhao Q, Yang H, Wang J, He S, Liu X. 2020. T-cell
659 receptor repertoire data provides new evidence for hygiene hypothesis of allergic diseases.
660 *Allergy*. doi:10.1111/all.14014
- 661 Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, Crawford JC, Clemens EB,
662 Nguyen THO, Kedzierska K, La Gruta NL, Bradley P, Thomas PG. 2017. Quantifiable predictive
663 features define epitope-specific T cell receptor repertoires. *Nature* 547:89–93.
- 664 DeWitt WS 3rd, Smith A, Schoch G, Hansen JA, Matsen FA 4th, Bradley P. 2018. Human T cell receptor
665 occurrence patterns encode immune history, genetic background, and receptor specificity. *Elife* 7.
666 doi:10.7554/eLife.38358
- 667 Elhanati Y, Sethna Z, Callan CG Jr, Mora T, Walczak AM. 2018. Predicting the spectrum of TCR
668 repertoire sharing with a data-driven model of recombination. *Immunol Rev* 284:167–179.
- 669 Emerson RO, DeWitt WS, Vignali M, Gravley J, Hu JK, Osborne EJ, Desmarais C, Klinger M, Carlson
670 CS, Hansen JA, Rieder M, Robins HS. 2017. Immunosequencing identifies signatures of
671 cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet*
672 49:659–665.
- 673 Espejo AP, Akgun Y, Al Mana AF, Tjendra Y, Millan NC, Gomez-Fernandez C, Cray C. 2020. Review of
674 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 (TCR β) 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 $\gamma\delta$ 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

768 Welsh RM, Selin LK. 2002. No one is naive: the significance of heterologous T-cell immunity. *Nat Rev*
769 *Immunol* 2:417–426.
770 Wirasinha RC, Singh M, Archer SK, Chan A, Harrison PF, Goodnow CC, Daley SR. 2018. $\alpha\beta$ T-cell
771 receptors with a central CDR3 cysteine are enriched in CD8 $\alpha\alpha$ intraepithelial lymphocytes and
772 their thymic precursors. *Immunol Cell Biol* 96:553–561.
773 Wolf K, Hether T, Gilchuk P, Kumar A, Rajeh A, Schiebout C, Maybruck J, Buller RM, Ahn T-H, Joyce S,
774 DiPaolo RJ. 2018. Identifying and tracking low-frequency virus-specific TCR clonotypes using
775 high-throughput sequencing. *Cell Rep* 25:2369-2378.e4.
776
777

778 FIGURE CAPTIONS

779
780 **Figure 1. Experimental antigen-specific enrichment.** (A) Repertoire subsets obtained by single-cell
781 sorting with peptide-MHC tetramers (data from Dash et al. and Sewell et al. via VDJdb; greens), peptide
782 stimulation enrichment (MIRA M55, M48; purples), or random sub-sampling of umbilical cord blood (1,000
783 or 10,000 TCRs; blues). Biochemical distances were computed among all pairs of TCRs in each subset
784 using TCRdist. Neighborhoods were formed around each TCR using a variable radius (x-axis) and the
785 percent of TCRs in the set that were within the neighborhood was computed. A radius of zero indicates
786 the proportion of TCRs that have at least one TCR with an identical amino acid sequence (solid square).
787 (B) Analysis of 18 MIRA-enriched repertoires for which the volunteers contributing the TCRs were
788 significantly enriched with a specific class I HLA allele (Table S4).

789
790 **Figure 2. Heterogeneous TCR neighborhoods within experimentally antigen-enriched and**
791 **unenriched repertoire subsets.** TCRs from (A) a peptide-MHC tetramer-enriched sub-repertoire, (B) a
792 MIRA peptide stimulation-enriched sub-repertoire, or (C) an umbilical cord blood unenriched repertoire.
793 Within each sub-repertoire, an empirical cumulative distribution function (ECDF) was estimated for each
794 TCR (one line) acting as the centroid of a neighborhood over a range of distance radii (x-axis). Each
795 ECDF shows the proportion of TCRs within the sub-repertoire found within the indicated radius from the
796 centroid TCR. ECDF color corresponds to the number of amino acids in the TCR CDR3 loop. ECDF
797 curves were randomly shifted by <1 unit along the x-axis to reduce overplotting. Vertical ECDF lines at
798 10^{-6} indicate no similar TCRs at or below that radius. Percentage of TCRs with an ECDF proportion $< 10^{-3}$
799 (bottom panels), indicates the percentage of TCRs without, or with very few biochemically similar
800 neighbors at the given radius.

801
802 **Figure 3. Radius-defined neighborhood densities within an antigen-enriched and a synthetic**
803 **background repertoire.** (A) Each TCR in the MIRA55 antigen-enriched sub-repertoire (one line) acts as
804 the centroid of a neighborhood and an empirical cumulative distribution function (ECDF) is estimated over
805 a range of distance radii (x-axis). Each ECDF shows the proportion of TCRs within the sub-repertoire
806 having a distance to the centroid less than the indicated radius. The ECDF line color corresponds to the
807 TCR generation probability estimated using OLGA (Pgen). The ECDF curves are randomly shifted by <1
808 unit along the x-axis to reduce overplotting. The bottom panel shows the percentage of TCRs with an
809 ECDF proportion $< 10^{-3}$. (B) Estimated ECDF for each MIRA55 TCR based on the proportion of TCRs in
810 a synthetic background repertoire that is within the indicated radius (x-axis). The synthetic background
811 was generated using 100,000 OLGA-generated TCRs and 100,000 TCRs sub-sampled from umbilical
812 cord blood; sampling was matched to the VJ-gene frequency in the MIRA55 sub-repertoire, with inverse
813 probability weighting to account for the sampling bias (see Methods for details). (C) Antigen-enriched
814 ECDF (y-axis) of one example TCR's neighborhood (red line) plotted against ECDF within the synthetic

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

818

819 **Figure 4. TCR meta-clonotype framework and application.** (A) Framework: sets of CD8+ TCRs
820 activated by SARS-CoV-2 peptides were previously discovered in 61 individuals diagnosed with COVID-
821 19 using a Multiplex Identification of Antigen-Specific T Cell Receptors Assay (MIRA) by Nolan et al.
822 (2020). For each TCR clone activated by a given peptide, we used tcrdist3 to evaluate the repertoire
823 fraction spanned at different TCRdist radii within (i) its antigen enriched repertoire (black) and (ii) a control
824 V- and J-gene matched, inverse probability weighted background repertoire (purple). Within the radius
825 (R) estimated to control a specified background TCR discovery frequency θ , we construct the set of all
826 antigen enriched TCRs neighboring a given centroid. The diagram shows, for instance, the centroid
827 TRBV28*01+CASSLKTNSYEQYF with θ set to $1e-6$. The resulting logo plot depicts the conserved CDR3
828 β -chain motif formed from the other SARS-CoV-2 ORF1ab 4211:4252-activated TCRs within a TCRdist
829 radius 16 of this centroid. (B) Application: we then applied public meta-clonotypes derived from antigen
830 enriched repertoires to quantify the enrichment of public SARS-CoV-2 antigen-specific TCRs in a large
831 diverse cohort, from whom unenriched bulk TCR repertoires were collected 0-30 days from COVID-19
832 diagnosis (n = 694). In almost all cases, the evidence of predicted HLA-mediated enrichment is stronger
833 for public meta-clonotypes compared to exact TCR β -chain amino acid sequences. This is indicated by
834 more statistically significant HLA coefficients in beta-binomial count regressions controlling for
835 sequencing depth, subject age, sex, and days from diagnosis.

836

837 **Figure 5. HLA restriction of TCR clonotypes and meta-clonotypes in bulk sequenced TCR β**
838 **repertoires of COVID-19 patients.** (A) Percentage of TCR features with a statistically significant (FDR <
839 0.001) positive association of its abundance in COVID-19 repertoires and patients' expression of the
840 restricting HLA allele. We tested for associations using beta-binomial regression controlling for
841 sequencing depth, age, sex, and days since COVID-19 diagnosis. (B) For each clonotype/meta-
842 clonotype, the percent of bulk repertoires from COVID-19 patients (n=694) containing TCRs meeting the
843 criteria defined by (1) EXACT (TCRs matching the centroid TRBV gene and CDR3 amino acids), (2)
844 RADIUS (TCR centroid with inclusion criteria defined by an optimized TCRdist radius, with θ set to $1e-6$),
845 or (3) RADIUS + MOTIF (inclusion criteria defined by TCR centroid, optimized radius and matching of
846 CDR3 amino-acid residues at conserved positions). See Methods for details.

847

848 **Figure 6. Associations of TCR features with participant age, days post diagnosis, HLA-genotype,**
849 **and sex in TCR β -chain repertoires of COVID-19 patients (n=694).** (A) Beta-binomial regression
850 coefficient estimates (x-axis) and negative log₁₀ false discovery rates (y-axis) for features developed from
851 CD8+ TCRs activated by SARS-CoV-2 MIRA55: ORF1ab 5171:5203. The abundances of TCR meta-
852 clonotypes are more robustly associated with predicted HLA type than exact clonotypes. (B) Signal
853 strength of enrichment by participant HLA-type (2-digit) of TCR β -chain clonotypes (EXACT) and meta-
854 clonotypes (RADIUS or RADIUS+MOTIF) predicted to recognize additional HLA-restricted SARS-CoV-2
855 peptides: (1) MIRA48: surface glycoprotein 22355:22393 (2) MIRA51: nucleocapsid phosphoprotein
856 29348:29380), (3) MIRA53: surface glycoprotein 22904:22936 (4) MIRA55: ORF1ab 4211:4252 (5)
857 MIRA110: nucleocapsid phosphoprotein 29138:29176 (6) MIRA11: ORF1ab, ORF3a 3875:25593. Models
858 were estimated with counts of TCR matching clonotypes (EXACT) or meta-clonotypes (RADIUS or
859 RADIUS+MOTIF) with the following definitions: (1) EXACT (inclusion of TCRs matching the centroid
860 TRBV gene and CDR3 amino acids), (2) RADIUS (TCR centroid with inclusion criteria defined by an
861 optimized TCRdist radius), (3) RADIUS + MOTIF (inclusion criteria defined by TCR centroid, optimized

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

864

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

871

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

878

879 **Figure S3: Associations of TCR features with participant age, days-post diagnosis, HLA-type, and**
880 **sex in bulk TCR β -chain repertoires of COVID-19 patients (n=694).** TCR features shown here were
881 identified from publicity analysis of CD8+ TCRs activated by an illustrative SARS-CoV-2 peptide ORF1ab
882 4211:4252 (MIRA55), which is predicted to bind HLA-A*01. Using beta-binomial regression models
883 estimated for each feature, volcano plots show associations between participant characteristics and
884 abundance of TCRs matching either: (A) EXACT clonotypes (TRBV+CDR3), (B) GLIPH2 patterns
885 (TRBV+CDR3 k-mer), or (C) *tcrdist3* based RADIUS+MOTIF meta-clonotype. See Methods for details.

886

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

Figure 1

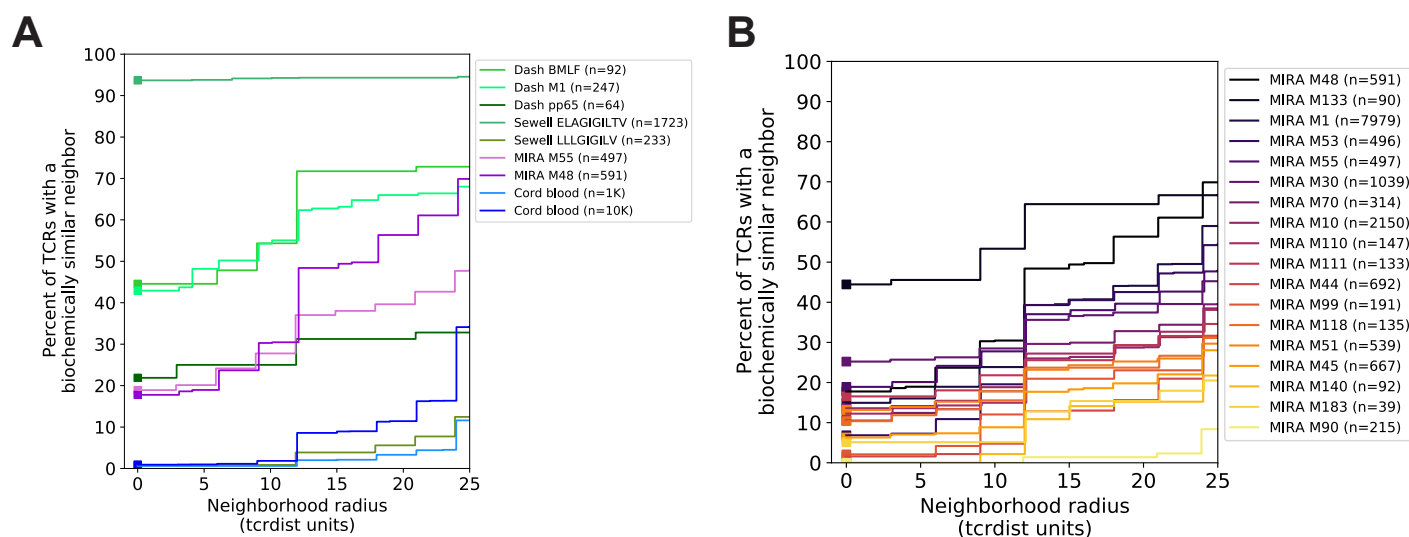


Figure 1. Experimental antigen-specific enrichment. (A) Repertoire subsets obtained by single-cell sorting with peptide-MHC tetramers (data from Dash et al. and Sewell et al. via VDJdb; greens), peptide stimulation enrichment (MIRA M55, M48; purples), or random sub-sampling of umbilical cord blood (1,000 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

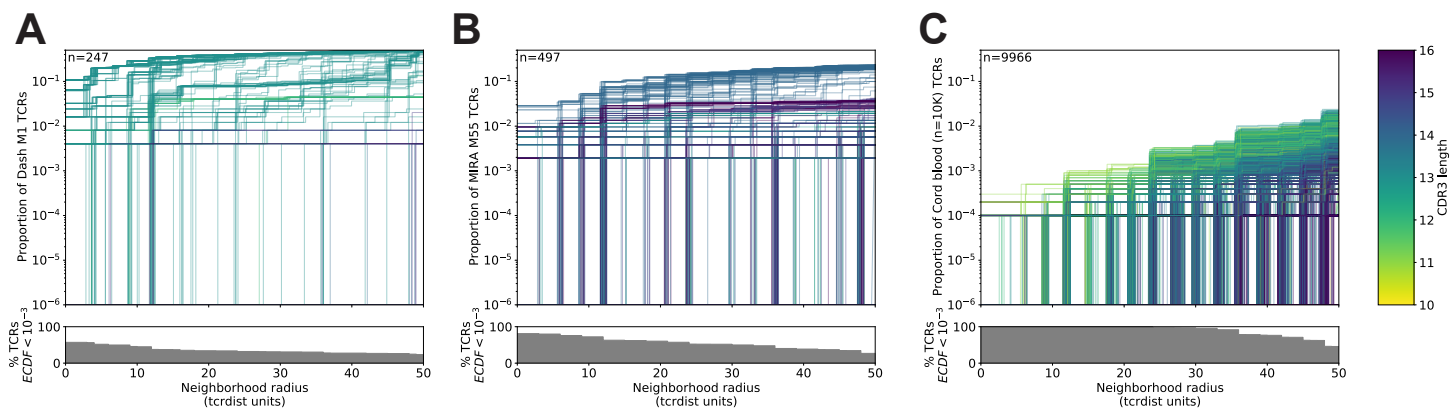


Figure 2. Heterogeneous TCR neighborhoods within experimentally antigen-enriched and unenriched repertoire subsets. TCRs from (A) a peptide-MHC tetramer-enriched sub-repertoire, (B) a MIRA peptide stimulation-enriched sub-repertoire, and (C) an umbilical cord-blood unenriched repertoire. Within each 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.

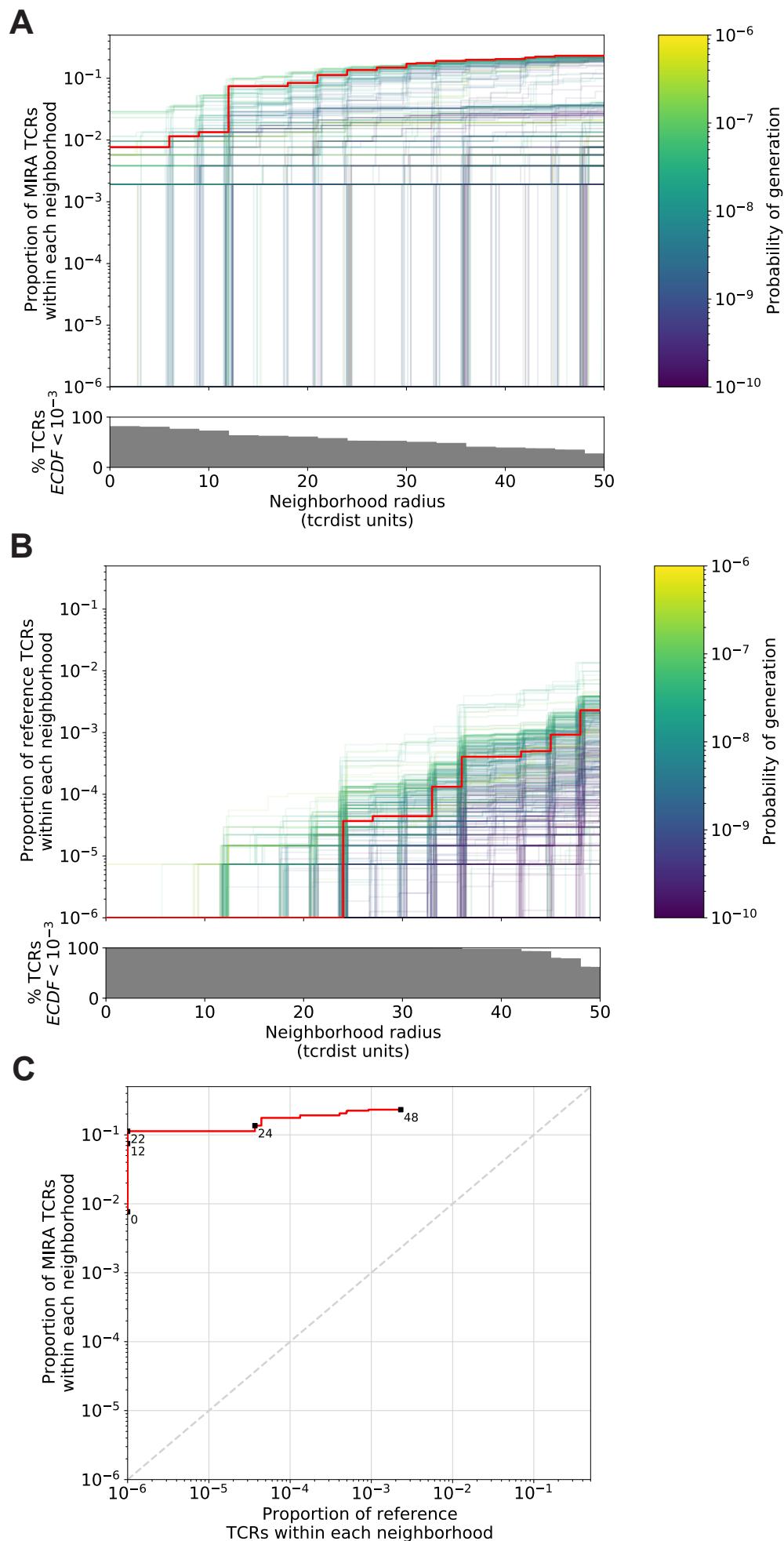
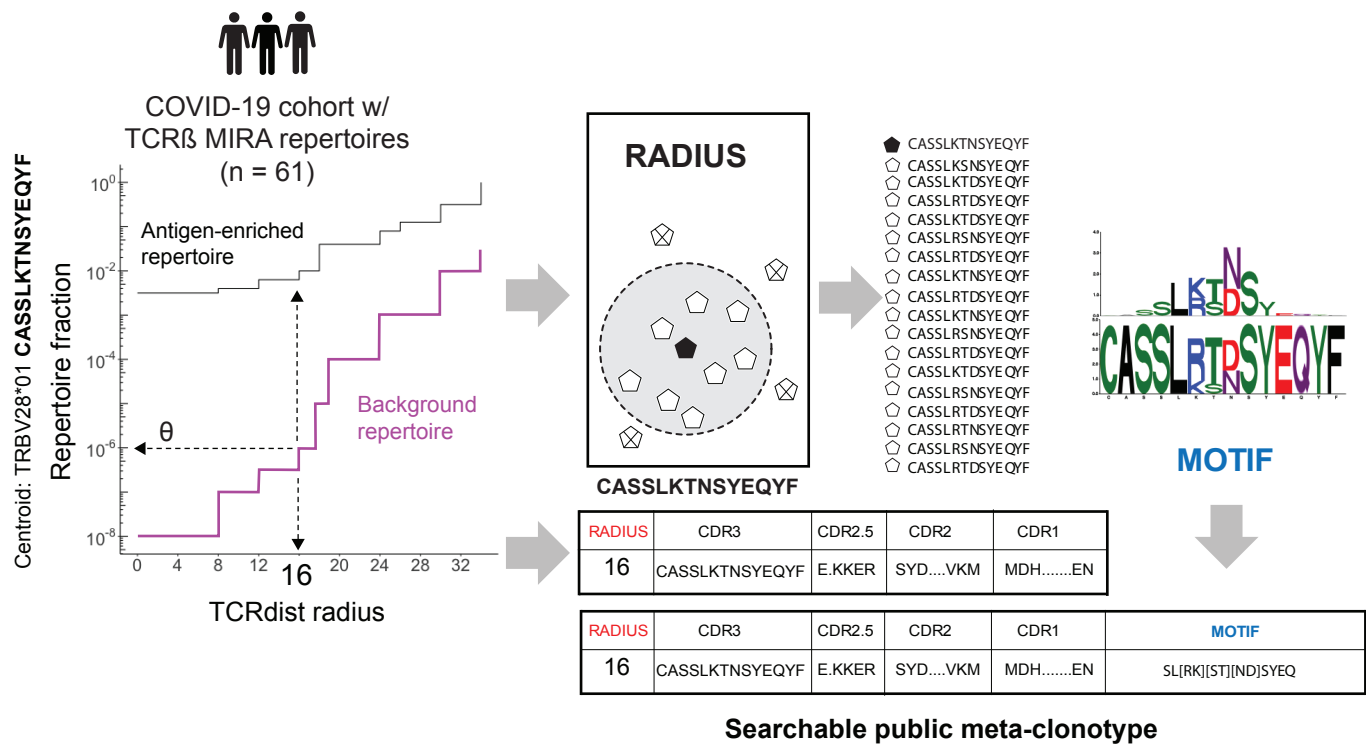


Figure 4

A TCR META-CLONOTYPE FRAMEWORK



B APPLICATION

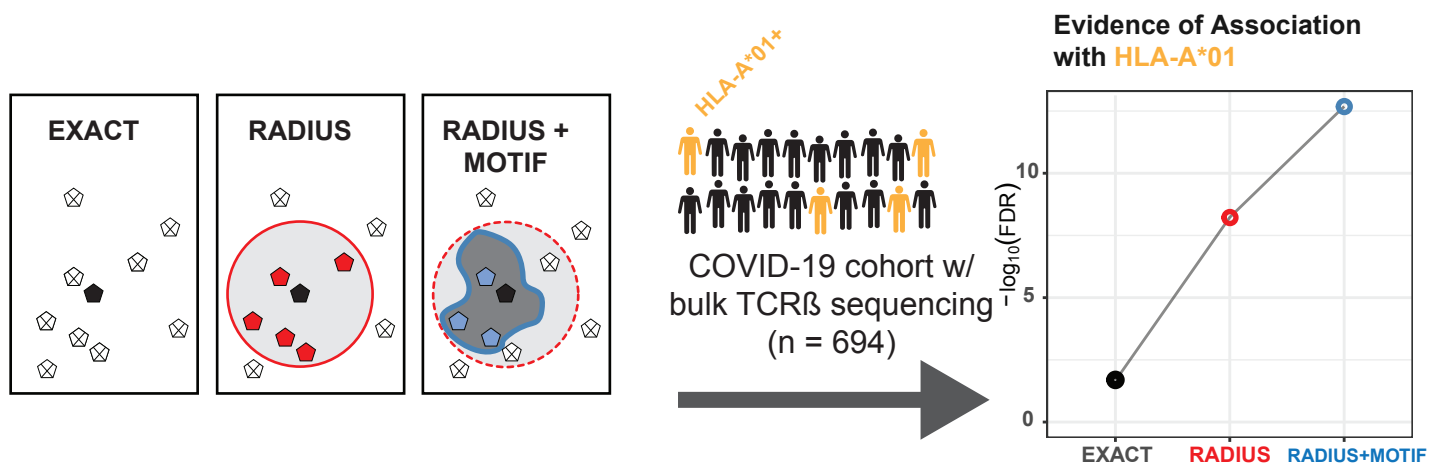


Figure 4. TCR meta-clonotype framework and application. (A) Framework: sets of CD8+ TCRs activated by SARS-CoV-2 peptides were previously discovered in 61 individuals diagnosed with COVID-19 using a Multiplex Identification of Antigen-Specific T Cell Receptors Assay (MIRA) by Nolan et al. (2020). For 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+CASSLK-TNSYEQYF 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

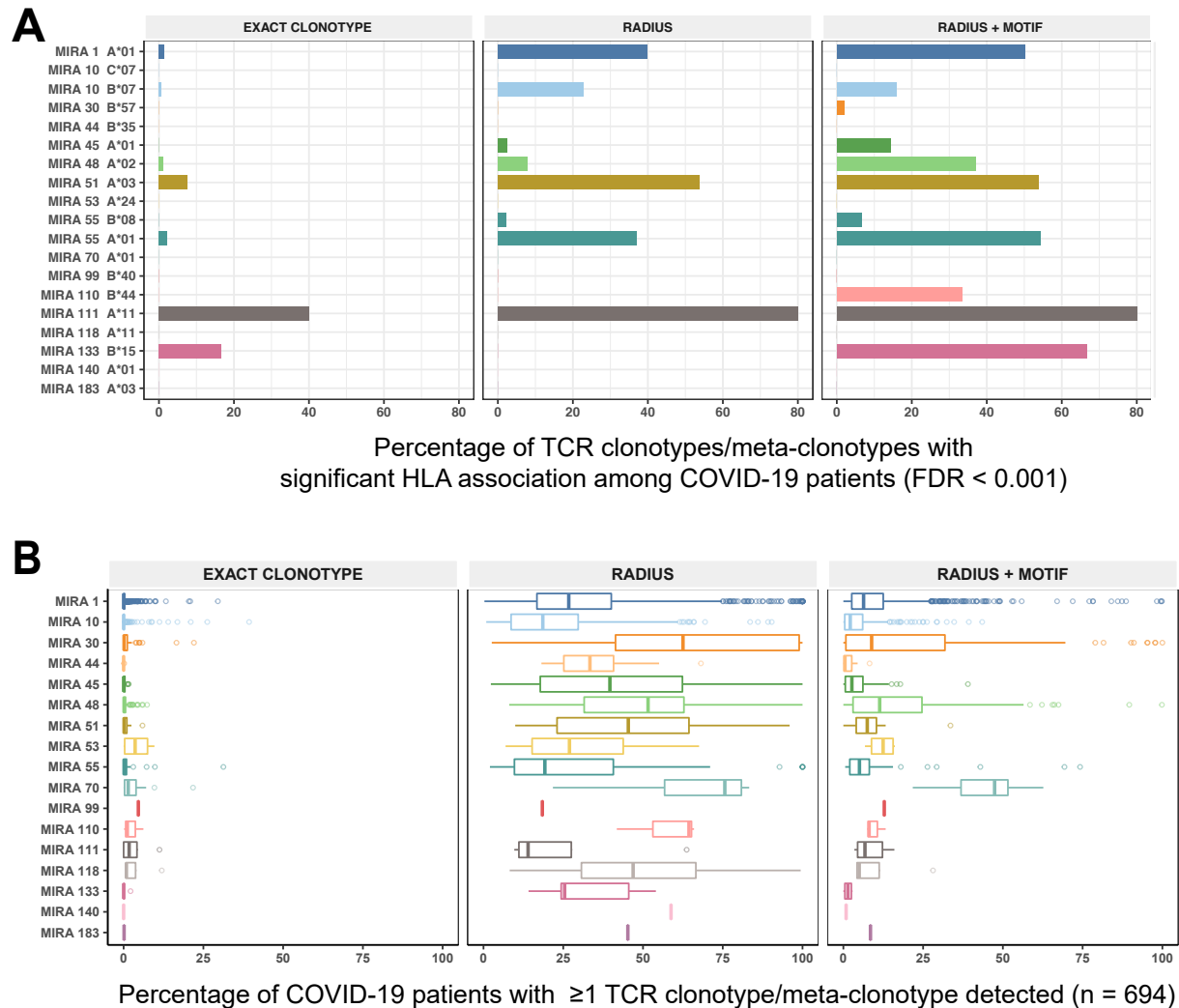


Figure 5. HLA restriction of TCR clonotypes and meta-clonotypes in bulk sequenced TCR β repertoires of COVID-19 patients. (A) Percentage of TCR features with a statistically significant (FDR < 0.001) positive association of its abundance in COVID-19 repertoires and patients' expression of the restricting 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.

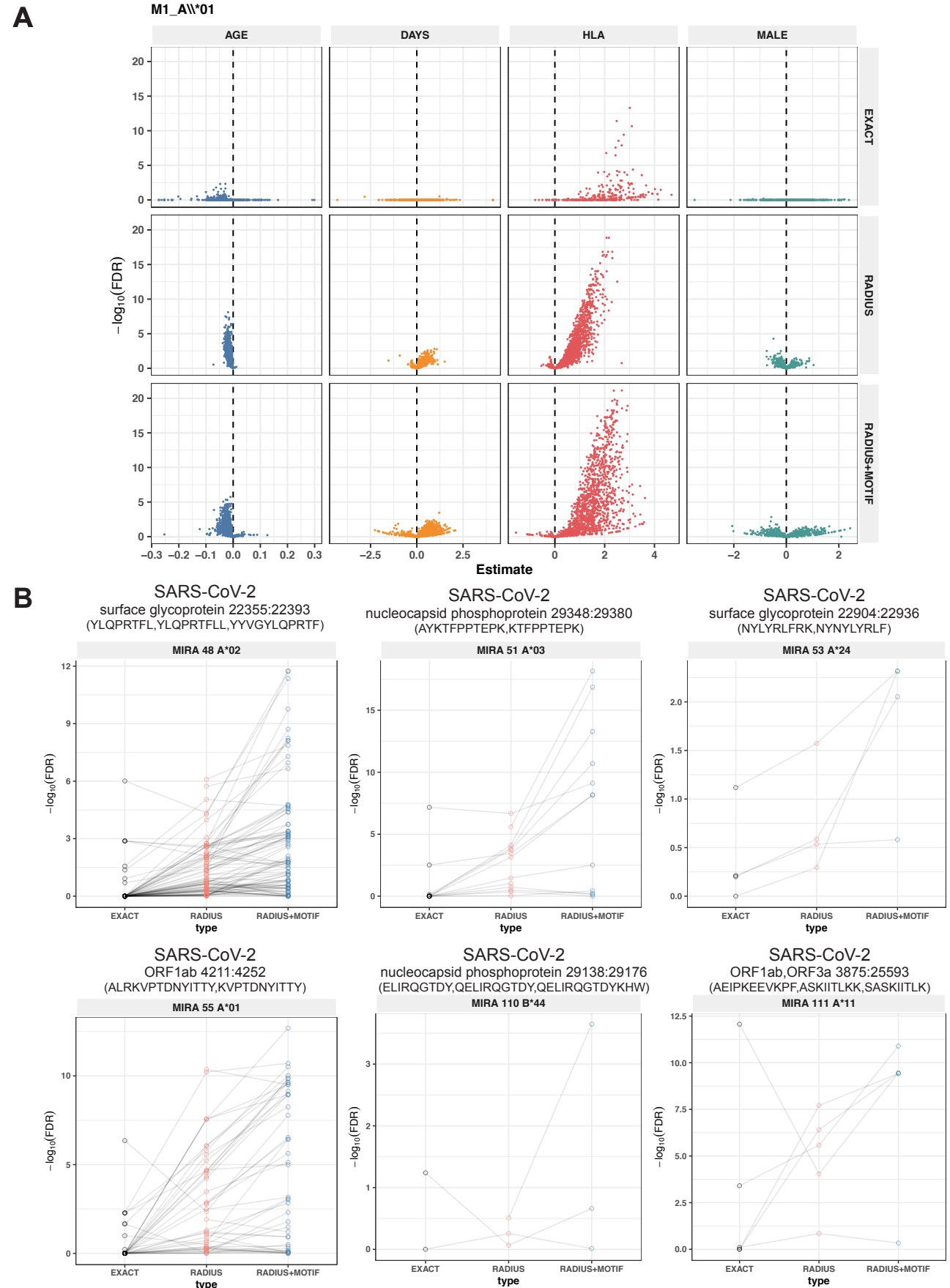


Figure S1

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-clonotypes 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.

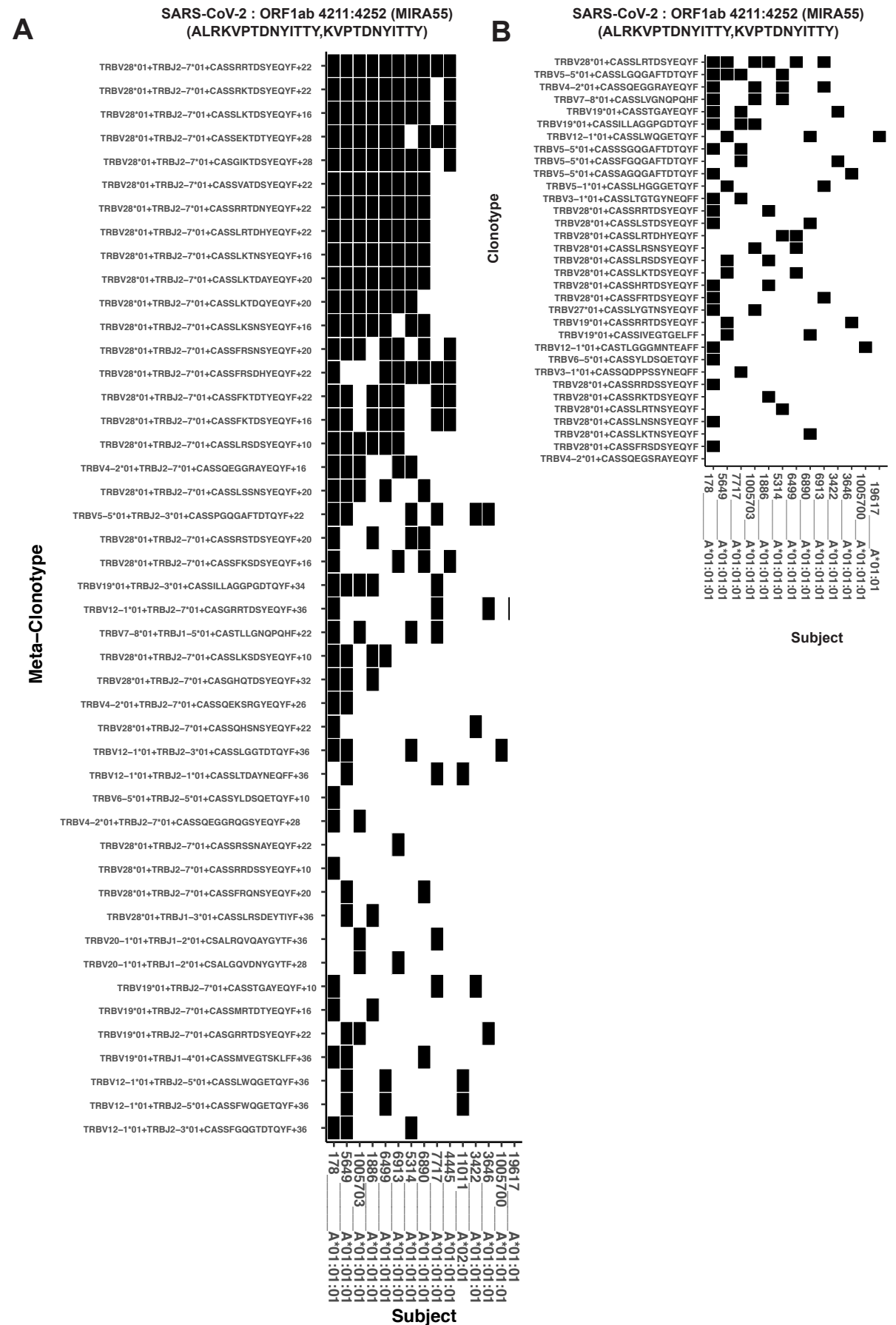


Figure S2

SARS-CoV2 : ORF1ab 4211:4252 (M55) (ALRKVPTDNYITTY,KVPTDNYITTY)

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 $\theta = 1E-6$) 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.

A

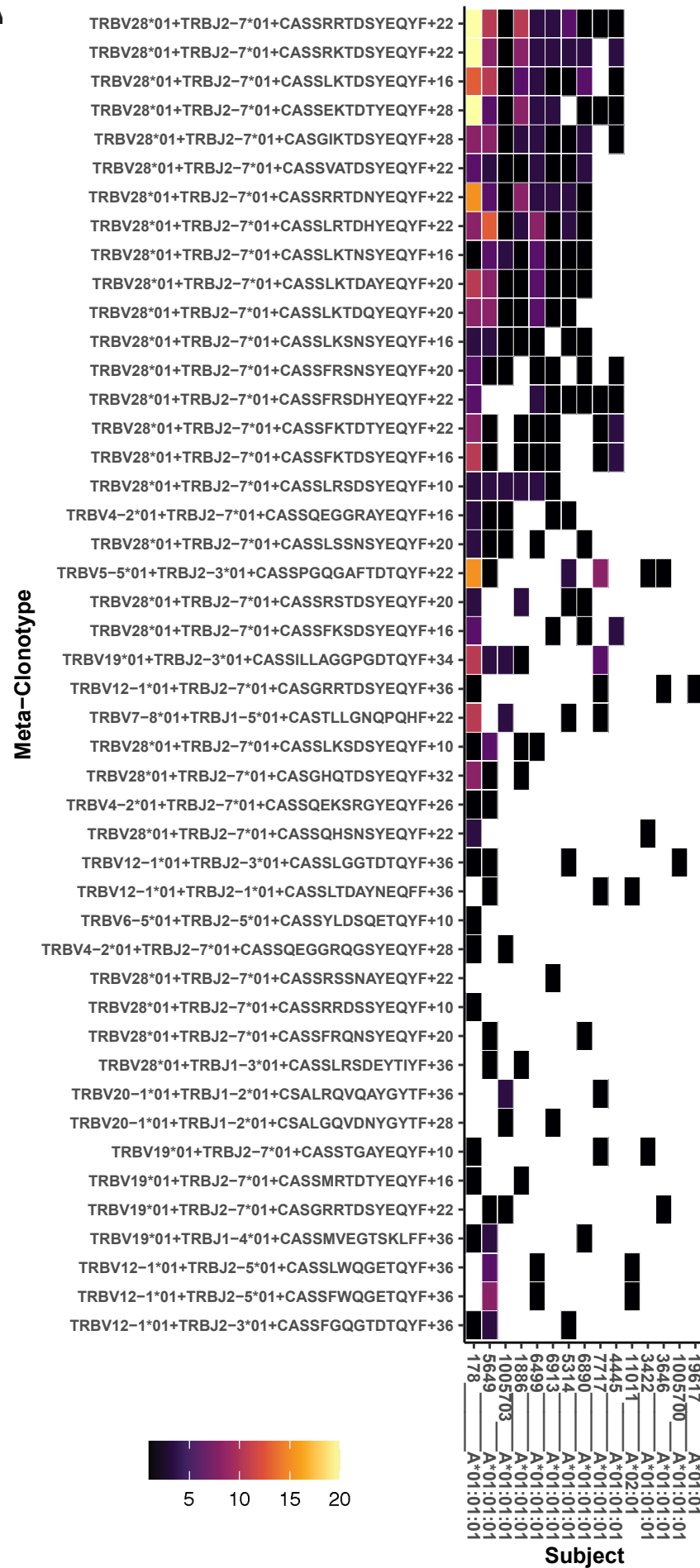
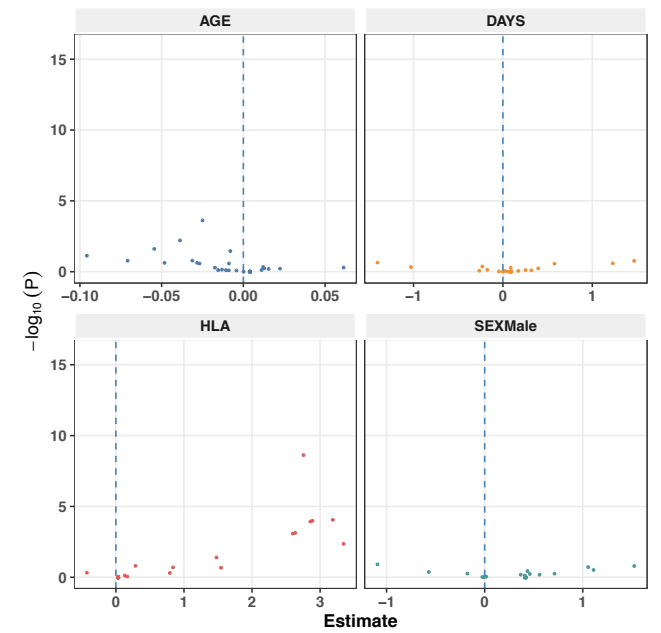
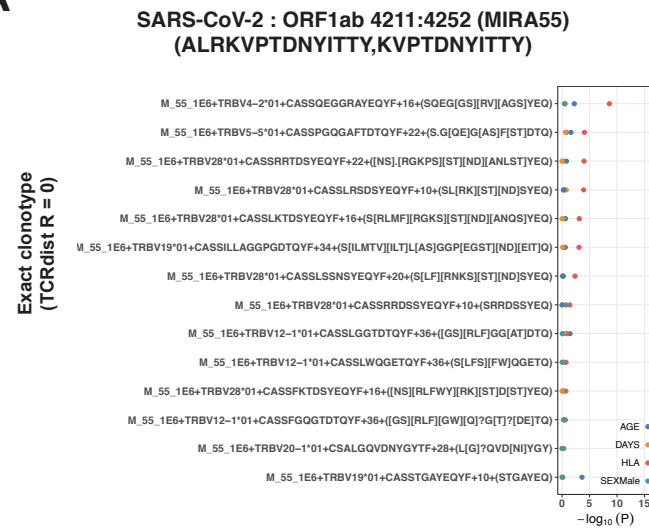


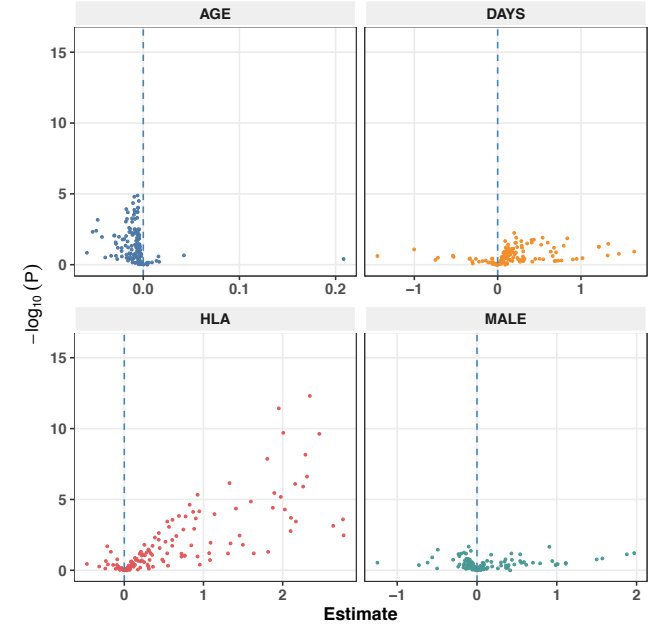
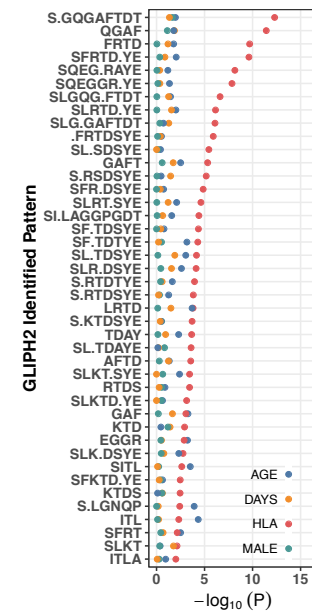
Figure S3

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) tcridist3 based RADIUS+MOTIF meta-clonotype. See Methods for details.

A



B



C

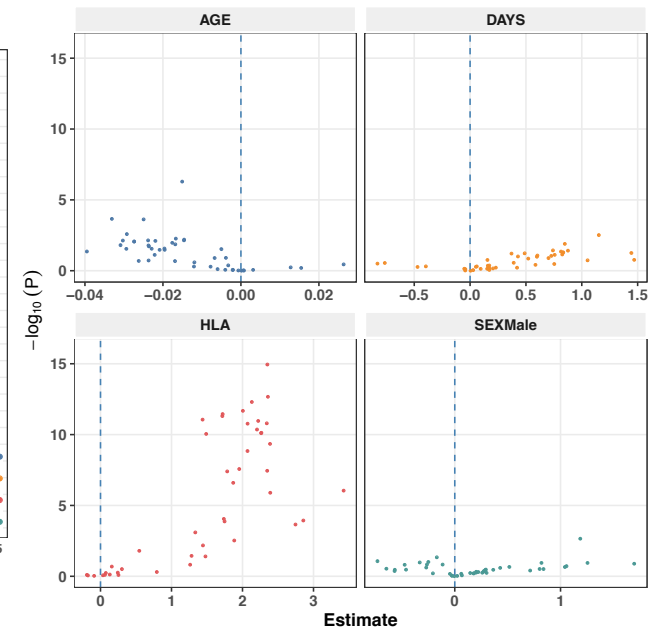


Figure S4

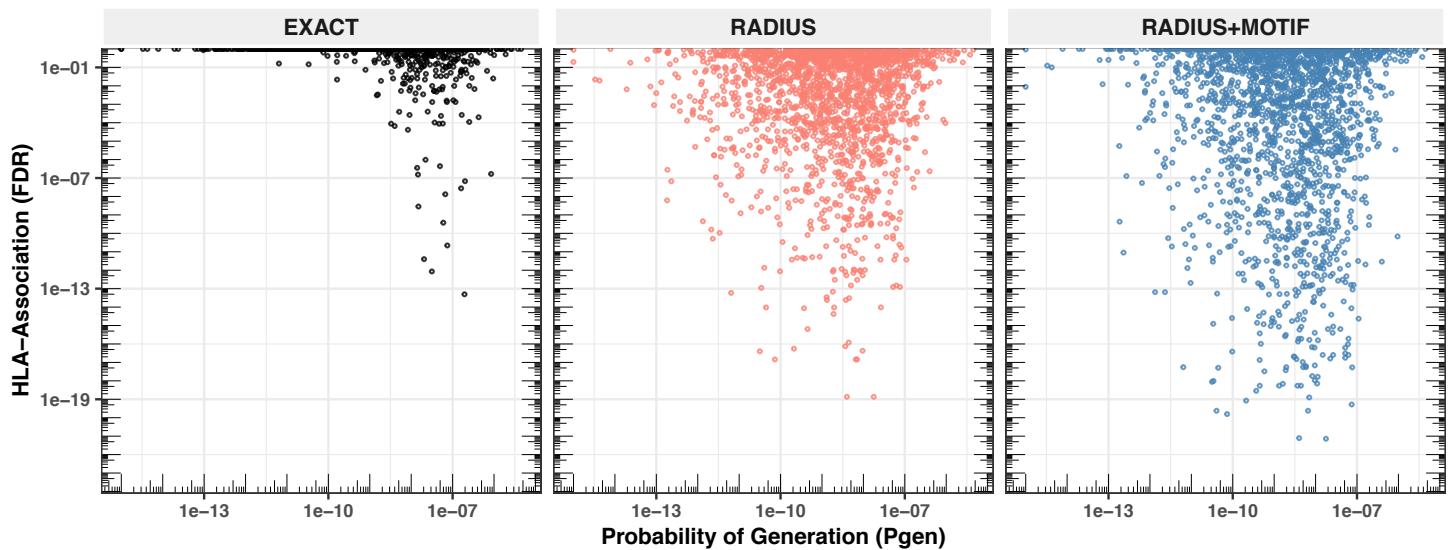


Figure S4. Detectable HLA-association and CDR3 probability of generation. We evaluated meta-clonotypes from 18 MIRA sets in a cohort of 694 COVID-19 patients for their association with predicted HLA-restricting alleles. Evidence of the HLA association for each meta-clonotype (RADIUS or RADIUS+MOTIF) 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.