Contrasting Effects of SARS-CoV-2 Vaccination vs. Infection on Antibody and TCR Repertoires

Jasper Braun[1], Elliot D. Hill[1], Elisa Contreras[1], Michie Yasuda[1], Alexandra Morgan[1], Sarah Ditelberg[1], Ethan Winter[1], Cody Callahan[1], Gabrielle Mazzoni[1], Andrea Kirmaier[1], Ghee Rye Lee[2*], Hamid Mirebrahim[3], Hosseinali Asgharian[3], Dilduz Telman[3], Ai-Ris Y. Collier[4,5,6], Dan H. Barouch[4,6,7,8], Stefan Riedel[1,4], Sanjucta Dutta[1], Florian Rubelt[3], and Ramy Arnaout[1,4,9,†]

[1]Division of Clinical Pathology, Department of Pathology, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA; [2]Department of Radiology, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA; [3]Roche Sequencing Solutions, Pleasanton, CA 94588, USA; [4]Harvard Medical School, Boston, MA 02215, USA; [5]Department of Obstetrics and Gynecology, Beth Israel Deaconess Medical Center, Boston, Massachusetts; [6]Center for Virology and Vaccine Research, Beth Israel Deaconess Medical Center, Boston, Massachusetts; [7]Ragon Institute of MGH, MIT, and Harvard, Cambridge, Massachusetts; [8]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts; [9]Division of Clinical Informatics, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA

*Current address: Ohio State University College of Medicine, Columbus, OH 43210

Contact Information

†Corresponding author: Ramy Arnaout, Beth Israel Deaconess Medical Center,330 Brookline Avenue, DA709, Boston, MA 02215, USA · rarnaout@bidmc.harvard.edu ·617-538-5681

Keywords

SARS-CoV-2, COVID-19, BCR repertoire, TCR repertoire, AIRRseq, ELISA, immunoPETE

Abstract

Antibodies and helper T cells play important roles in SARS-CoV-2 infection and vaccination. We sequenced B- and T-cell receptor repertoires (BCR/TCR) from the blood of 251 infectees, vaccinees, and controls to investigate whether features of these repertoires could predict subjects' SARS-CoV-2 neutralizing antibody titer (NAbs), as measured by enzyme-linked immunosorbent assay (ELISA). We sequenced recombined immunoglobulin heavy-chain (IGH), TCRβ (TRB), and TCRδ (TRD) genes in parallel from all subjects, including select B- and T-cell subsets in most cases, with a focus on their hypervariable CDR3 regions, and correlated this AIRRseq data with demographics and clinical findings from subjects' electronic health records. We found that age affected NAb levels in vaccinees but not infectees. Intriguingly, we found that vaccination and infection have an effect on non-productively recombined IGHs, suggesting an effect that precedes clonal selection. We found that repertoires' binding capacity to known SARS-CoV-2-specific CD4+ TRBs performs as well as the best hand-tuned approximate or "fuzzy" matching at predicting a protective level of NAbs, while also being more robust to repertoire sample size and not requiring hand-tuning. The overall conclusion from this large, unbiased, clinically well annotated dataset is that B- and T-cell adaptive responses to SARS-CoV-2 infection and vaccination are surprising, subtle, and diffuse. We discuss methodological and statistical challenges faced in attempting to define and quantify such strong-but-diffuse repertoire signatures and present tools and strategies for addressing these challenges.

41 ## Introduction

42 Since COVID-19's emergence, there has been great interest in identifying antibody (BCR) and T-
43 cell receptor (TCR) gene sequences specific to SARS-CoV-2. The pandemic presented a high-
44 profile opportunity to test the extent to which TCRs and BCR sequences against respiratory
45 viruses would be public, i.e. appearing across many different individuals, or private, present in
46 only one or a few individuals. New SARS-CoV-2-specific BCRs could form the basis for new
47 treatments. Understanding how to identify and characterize commonalities in such a setting
48 could help evaluate the viability of new diagnostics based on adaptive immune-receptor
49 repertoire sequencing (AIRRseq).[1] number of studies have succeeded in identifying public
50 immunoglobulin heavy-chain (IGH) and TCRβ (TRB) sequences. However, in part due to
51 exigencies and constraints imposed by the pandemic, and in part because it was impossible to
52 know a priori what study size would be adequate to identify public sequences comprehensively
53 in COVID-19, many of these studies involved relatively few individuals. Small sample sizes have
54 known limitations for AIRRseq studies. Because small samples may not be representative of
55 larger populations, results on small samples may not generalize. Small studies may be
56 insufficiently powered to detect subtle patterns. And the smaller the sample size, the more
57 likely it is that random fluctuations in the data—literally, the luck of the draw—will produce
58 results that appear to be statistically significant but do not reflect underlying relationships.
59 Moreover, most studies investigated only BCRs or only TCRs, but not both in the same cohort,
60 despite the importance of both antibody and T-cell responses in SARS-CoV-2 infection.[2,3] To our
61 knowledge only two previous COVID-19 studies[4,5] have sequenced TCR from γδ T cells, a little-
62 studied subset that may be important in mucosal antimicrobial immunity.[6,7] How SARS-CoV-2
63 virus or vaccine exposure affects different B- and T-cell subsets (IgM-positive vs. IgM-negative
64 B-cells, CD4-positive vs. CD8-positive T cells) has also been insufficiently explored.[2,3]

65 To our knowledge previous studies have identified 20 IGH V genes to be enriched in sequences
66 produced during various immune responses to SARS-CoV-2.[8–16] Given that human genomes
67 encode 54 IGH V genes,[17] collectively these studies implicate 37% of all V genes in the response

68 to this single viral exposure, indicating that the SARS-CoV-2 response is either quite broad
69 within individuals, quite heterogeneous among individuals, or both. There is no obvious reason
70 to think each V gene would contribute to only a single SARS-CoV-2-specific recombined IGH
71 sequence, much less a single IGH:immunoglobulin light chain (IGL) pair; therefore, collectively
72 these studies also suggest that the IGH response to SARS-CoV-2 might account for a significant
73 fraction of a given repertoire, a possibility that requires more comprehensive AIRRseq
74 investigation. Note that studies that compare only a single non-control cohort to a control
75 cohort cannot distinguish between features (clones, motifs, genes, CDR3 lengths, etc.) that
76 signify a disease-specific vs. a general immune response.

77 Regarding TCRs, one study[18] searched repertoires of 140 COVID-19 patients and another 140
78 pre-pandemic (and therefore unexposed) controls for the presence of each of 1,267 TRB
79 sequences that had independently been shown to recognize epitopes of the SARS-CoV-2 spike
80 protein. These authors showed that while the presence of some of the TRB sequences in almost
81 all of the repertoires suggested a public response to SARS-CoV-2 infection, the fraction of the
82 repertoire that matched the query sequences was similar between infectees and controls.
83 These authors also looked for SARS-CoV-2 specific TRBs in the brain tissue of COVID-19 patients,
84 since T-cell infiltration of the brain, an organ otherwise seldom infiltrated by T cells, is known to
85 occur during COVID-19 infection.[19–21] The 68 TRBs they identified were found in 40% of COVID-
86 19 repertoires vs. 17% of pre-pandemic controls. This suggests significant enrichment even as
87 the majority of individuals with COVID-19 lacked these sequences and a significant minority of
88 controls (1 in 6) had them despite these control samples having been collected before the
89 pandemic (which has been observed in other contexts,[22] perhaps indicating cross-reactivity
90 with previously circulating coronaviruses, which are very common in human populations[3]).

91 In addition, one study[23] identified a large database of SARS-CoV-2-specific TRB sequences as
92 being shared among infectees and unenriched among healthy controls. Subsequently, another
93 study[24] sequenced repertoires of individuals 0 and 4 weeks after vaccination with the Oxford-
94 AstraZeneca COVID-19 vaccine AZD1222 or the meningococcus vaccine MenACWY and matched

95 the database sequences to these repertoires. An increase of database sequences among
96 AZD1222-vaccinee repertoires but not MenACWY-vaccinee controls was seen between the two
97 timepoints.

98 Heterogeneity in clinical settings across studies complicates the interpretation of private vs.
99 public responses, for several reasons. First, there are important antigenic differences between
100 exposure and vaccination. This is especially true for the mRNA vaccines, which immunize
101 subjects with only spike protein, in contrast to the full complement of SARS-CoV-2 proteins to
102 which infectees are exposed. Second, demographics may play a role.

103 For example, it has long been recognized that individuals respond differently to vaccines by age,
104 with older individuals generally mounting less-robust and shorter-lasting responses as
105 measured by ELISA.[25–27] Other clinical features are also known to affect the adaptive immune
106 response to infection and vaccination, including immunosuppressive conditions such as organ
107 transplant or cancer therapy as well as metabolic disease.[28] Third, studies from different
108 periods of the pandemic likely measured responses to different strains. Fourth, the signal or
109 signature detected may differ depending on whether the controls were healthy, which might
110 result in detecting generalized responsiveness (e.g. bystander activation), or were instead
111 presenting with a non-COVID illness, making any signal/signature more likely to be specific for
112 COVID-19. Fifth, exposure, whether to replicating virus or to an inactivated or subcomponent
113 vaccine, may not be as clinically relevant as whether a substantial NAb response was mounted.
114 This is because NAbs are a marker of protection in SARS-CoV-2.[29–31] (T-cell-mediated immunity
115 may also play an important role[32]). And sixth, accessing clinical annotations from electronic
116 medical records can be challenging.[33] As a result, the effect of clinical heterogeneity in AIRRseq
117 studies in the setting of COVID-19 has been under-explored to date.

118 In all, the work above supports the view that there are commonalities in IGH and TRB at the
119 gene and sequence level in response to SARS-CoV-2 infection; however, the nature of the
120 signature is not well understood. One open question is to what extent infection affects

121 antibody and TCR repertoires as a whole vs. enriching specific clones within it. One can refer to
122 these ends of the continuum of possible effects as “diffuse” vs. “precise.” From previous work
123 on repertoires, “diffuse” features include CDR3 length, the frequency of usage of specific V or J
124 genes, and repertoire diversity as measured any of several ways (richness, Shannon entropy,
125 Simpson’s index, or their Hill-number equivalents).[34] At the other extreme, the most “precise”
126 features are the frequency of clones with specific sequences. Between these extremes is a
127 set of features that includes approximate or “fuzzy” matching of sequences[35] and other
128 clustering methods.[36,37] This middle ground has been less explored. Recently the concept of
129 binding capacity has been developed to measure the fraction of a repertoire that is “like” a
130 given query sequence in terms of target specificity (weighting the repertoire by the predicted
131 dissociation constants of its constituent antibodies or TCRs and by their sequence frequencies).
132 Whether or how binding capacity might be affected by SARS-CoV-2 infection and/or vaccination
133 is unknown.

134 Given this background, we sought to investigate the effects of SARS-CoV-2 infection and
135 vaccination on both antibody and TCR repertoires in a large clinical cohort, with attention to
136 major B- and T-cell subsets where possible, using NAbs via ELISA as a functional readout, with a
137 special focus on diffuse repertoire features and how they compare to both more traditional
138 features and to clinical correlates.

139 Results

140 NAbs vary with exposure, age, and immune status

141 Using immunoPETE (Roche; research use only),[5] we deep-sequenced IGH, TRB, and TRD from
142 the blood of 251 individuals: 36 vaccinees, all of whom had received an mRNA vaccine, 145
143 infectees, 53 healthy controls, and 20 with unknown SARS-CoV-2 exposure status. Three
144 individuals belonged to both the vaccinee and infectee groups. Forty-seven subjects were
145 considered immunosuppressed and the remaining 204 immunocompetent. Blood samples for
146 129/145 (89%) infectees were within 6 months of the most recent positive PCR test on record

147 and 121/145 (83%) were ≥7 days from the presumed most recent infection date (assuming a
148 mean of 4 days from exposure to testing). Fig. S1 presents a summary of the timeline and
149 sequencing yield. Tables S1 and S2 present demographics and relevant comorbidities for the
150 different cohorts. We measured plasma NAbs against SARS-CoV-2 spike for 237/251 subjects.
151 Fig. 1 presents a summary of the measured NAbs concentrations by cohort,
152 immunosuppression status, and age.

153 NAbs were undetectable in some infectees and one vaccinee (Fig. 1). The odds of producing
154 NAbs were significantly higher in immunocompetent subjects compared to immunosuppressed
155 subjects (OR=3.9, p_c=0.008—note, all p-values in this work have been corrected for multiple-
156 hypothesis testing; we write p_c to indicate this). Odds were also significantly higher in the
157 infectees (OR=5.8, p_c<0.001) and vaccinees (OR=4.7, p_c=0.034) compared to controls. Age was
158 not significantly associated with NAbs titer (p_c=0.801) and therefore age was excluded from
159 most of the subsequent models (below).

160 Among the subjects who did produce detectable NAbs, NAb concentration was notably higher
161 in the vaccinated and infected groups compared to the control group (Fig. 1a). The relationship
162 between concentration and age was more complex and depended on the cohort (significant
163 age × cohort interaction). Age affected titer only in vaccinees, with NAbs being lower in older
164 individuals: above age 65, individuals had higher NAbs with infection than vaccination (Fig. 1a).
165 Meanwhile, age did not affect NAbs in the control or infected groups. We found no significant
166 effect of immunocompetence on NAbs in subjects who had non-zero NAbs.

167 **Vaccination is associated with shorter IGH CDR3s in productive joins**

168 The characteristic (e.g. mean or median) length of CDR3s is known to vary during development
169 and in response to various exposures, at least in productively recombined IGH genes, a.k.a.
170 "productive joins."[38] Because only productively recombined IGH genes can be expressed as
171 (BCR) proteins, such differences are generally considered evidence that the B cells that express
172 them are selected for having e.g. longer CDR3s. We found that vaccinees had shorter IGH

173 CDR3s than controls (p_c=0.024; Fig. 2a) or infectees (p_c=0.0046; Fig. 2b), indicating a repertoire-
174 wide difference in the B-cell response to vaccination vs. infection (Table S3, Figs. S2-S4).

175 The length of IGH CDR3s depends on the lengths of the constituent IGHV, IGHD, and IGHJ
176 genes, as well as the number of non-templated (N) and palindromic (P) nucleotides inserted at
177 the junctions between them.[39] Annotating IGHD and distinguishing mutated/truncated IGHD
178 sequence from N and P sequences is challenging due to insertions, chewbacks, and somatic
179 hypermutation. However, IGHV and IGHJ can be annotated reliably, and so we tested whether
180 the overall differences in CDR3 length were attributable to differences in the use of longer vs.
181 shorter IGHV and IGHJ genes.

182 We grouped IGHV genes by the number of amino acids that their germline contributes to
183 CDR3s, and similarly for IGHJ genes. The 54 IGHV genes hard-coded as part of the human
184 germline contribute either 3 or 4 amino acids to the CDR3, depending on the gene. We found
185 that vaccinees generally used more of the IGHV genes that contribute 3 residues (p_c=0.021 vs.
186 controls and p_c=0.0028 vs. infectees) and fewer of the IGHV genes that contribute 4 residues
187 (again p_c=0.021 vs. controls and p_c=0.0028 vs. infectees; Fig. 2c and Table S4). The same pattern
188 held for IgM-positive and IgM-negative (mainly IgG; hereafter "IgG") in controls and infectees.
189 Meanwhile, the six IGHJ germline genes contribute 5 (IGH J4), 6 (IGH J3 and J5), 7 (IGH J1 and
190 J2), or 10 (IGH J6) amino acids to the CDR3 (Fig. 2d). We found that vaccinees used more J4
191 (p_c=9.4e-5 vs. controls and p_c=1.9e-6 vs. infectees) and fewer J3 & J5 (p_c=0.0002 vs. controls and
192 p_c=0.0021 vs. infectees; Table S4). Thus, the preference of shorter IGH CDR3s after vaccination
193 can at least partially be explained by selection for V and J genes that contribute fewer residues
194 to the CDR3.

195 No such differences were observed in TCR CDR3s, which have a far narrower length
196 distribution.

197 **Vaccination is associated with longer IGH CDR3s in non-productive joins**

198 Next we sought to estimate the strength of selection for IGH CDR3s of different lengths in
199 vaccinees, infectees, and controls. This can be done by comparing the length distribution of
200 productive joins to the distribution in non-productive joins, i.e. those in which VDJ
201 recombination occurs out of frame or produces stop codons. Because non-productive joins do
202 not produce functional antibodies, the B cells that contain them cannot be selected for or
203 against based on them. Nevertheless, the lengths of the CDR3 regions in non-productive joins
204 can be measured. Thus, any differences in length between non-productive joins and productive
205 joins reflect selection on (some aspect of) the productive joins, for example by exposure to
206 SARS-CoV-2 (in infectees) or vaccine contents (vaccinees).

207 Our null hypothesis was that the lengths of non-productive joins would be similar for vaccinees,
208 infectees, and controls. Surprisingly, we found that CDR3s in non-productive joins differed
209 across these three cohorts. In fact, we observed reverse relationships from the ones we saw in
210 productive joins: CDR3s in non-productive joins were longer in vaccinees and infectees than in
211 controls (p_c=0.039 and 0.0021, respectively). Vaccinees' non-productive CDR3s used the
212 shortest J gene, J4, less often and the longest J, J6, more often than controls' (p_c=0.00011 and
213 0.022, respectively; Fig. S5). Thus, selection for shorter CDR3s in vaccinees is even stronger than
214 indicated from the comparison of productive joins in the previous section, because in
215 vaccinees, recombination, which precedes selection, is biased toward longer CDR3s. Again, no
216 such differences were observed in TCR CDR3s (Figs. S6 and S7).

217 **Vaccination affects at least one-sixth of the pre-selection IGH repertoire**

218 We next sought to better characterize this apparent effect of vaccine exposure on IGH
219 recombination. The results in the previous section were regarding differences in subjects' entire
220 IGH CDR3 repertoires. However, vaccine exposure is generally thought to affect only a portion
221 of the repertoire. The rest of the repertoire, the unaffected portion, should be the same as a
222 control's. Therefore, conceptually, each vaccinee's repertoire can be thought of as a weighted
223 sum of two parts: a vaccine-responsive part and a control part. We asked what the minimum
224 size of the vaccine-responsive part would have to be, in order to explain the difference in the

225 length distribution of non-productive joins between vaccinees and controls.

226 To do this, we analyzed the differences between the mean IGH CDR3 length-distribution curves
227 of vaccinees and controls. By calculating differences at each length, we generated the length
228 distribution that the putative vaccine-responsive part would have to have, in order for the
229 vaccinee curve to be a weighted sum of the control curve and the vaccine-responsive part, for a
230 given size of the vaccine-responsive part (Fig. 2e). Inevitably, there will an inverse relationship
231 between how different the length distribution of the vaccine-responsive part is, and its size.
232 This fact sets a floor on the size of the vaccine-responsive part: any smaller, and the vaccine-
233 responsive part would have to be so different that at least one of its lengths would have a
234 negative frequency.

235 For example, 20-amino-acid-long CDR3s constituted an average of 9% of non-productive joins in
236 controls but 8% in vaccinees. Considering just this length for the moment, if length-20 CDR3s
237 constituted 7% in the vaccine-responsive part, then the vaccine-responsive part would have to
238 be 50% of the repertoire, since 50%×9% + 50%×7% = 8%. If instead length-20 CDR3s constituted
239 3.5%, the vaccine-responsive part would only have to be 18%, since (100-18)%×9% + 18%×3.5%
240 = 8%. In this example, the vaccine-responsive part could never be as small as 1%, since in that
241 case length-20 CDR3s would have to have a negative frequency. By this approach, we found
242 that the size of the vaccine-responsive part could be no smaller than 16%, or one-sixth, of the
243 vaccinees' non-productive joins.

244 **Vaccinees and infectees with more SARS-CoV-2-specific TRBs have higher NAbs**

245 Next, we tested whether TRB and IGH CDR3s that had been previously found to be associated
246 with SARS-CoV-2 exposure, including by structural studies, were enriched among our vaccinee
247 and infectee cohorts (see Methods). We obtained SARS-CoV-2-specific TCRs from CD4 and CD8
248 T cells from Nolan et al.[23] and obtained non-CD-restricted SARS-CoV-2- and non-SARS-CoV-2-
249 specific TRBs and IGHs from CoV-AbDab, PDB, and VDJDb.[40–42] These comprised totals of
250 184,100 unique SARS-CoV-2-specific TRBs and 1,630 unique SARS-CoV-2-specific IGHs (Table

251 S5).

252 We found a much higher proportion of SARS-CoV-2-specific TRB sequences than IGH sequences
253 had exact matches in our samples: ≥ 12% vs. 0.1%, respectively, with the 0.1% representing just
254 a single sequence (Table S5). The fraction of each repertoire that matched SARS-CoV-2-specific
255 TRBs correlated positively with NAbs, as measured by ELISA titer, in infectees and vaccinees
256 (Fig. 3a-b, Table S6, and Fig. S9). In infectees, for whom we had separate CD4-positive and CD4-
257 negative (mainly CD8-positive; hereafter "CD8") TRB repertoires, the positive correlation was
258 confined to CD4 repertoires. In contrast, no correlation was seen for controls. Likewise, no
259 correlation was seen for TRBs that were not specific for SARS-CoV-2 in infectees, supporting the
260 interpretation that this correlation is causal. Nevertheless, this correlation alone performed
261 poorly as a classifier of who had high enough NAbs to be considered positive (per the ELISA test
262 manufacturer), with an area under the receiver-operator characteristic curve (AUROC) of 0.55
263 (95%CI, 0.46-0.63). Notably, there was also a positive relationship between non-specific TRBs
264 and NAbs in vaccinees, although the 95% CI on the regression slope only narrowly missed
265 including zero (Table S6).

266 **Binding capacity outperforms fuzzy matching for measuring similarity**

267 That subjects had almost no exact matches to SARS-CoV-2-specific IGH sequences did not
268 exclude the possibility that they have sequences that are functionally similar to these reference
269 sequences. The same possibility exists for TRBs. A standard method for finding similar
270 sequences is using the Levenshtein distance (i.e. edit distance). Sequences with a distance of
271 less than or equal to a tolerance t are considered similar (for example, sequences that differ by
272 no more than t=1 amino acid). This is known as "fuzzy matching" with tolerance t. (Note that
273 exact matches are just fuzzy matches with tolerance 0.) Unfortunately, there is no consensus on
274 what t should be chosen. Also, the fraction of a repertoire that fuzzy-matches a set of
275 references could depend on repertoire size because of the nature of sampling, potentially
276 complicating the use of fuzzy matching.

277 To test this possibility, we subsampled 30 subjects' repertoires (10 controls, 10 infectees, and
278 10 vaccinees) and measured the fraction of the repertoire that fuzzy-matched SARS-CoV-2-
279 specific CD4 TRBs at tolerances of 0, 2, 4, 6, 8, and 10 amino acids. We fit a linear mixed model
280 grouped by subject for all repertoires with at least 1,000 sequences. We found the fraction of
281 fuzzy matches depended strongly on repertoire size for all repertoire sizes measured (up to 1
282 million sequences), falling steeply and continuously throughout (Fig. 3c). Thus, fuzzy matching
283 was shown to not be a robust measure of repertoire content in this study.

284 We therefore tested a recently proposed alternative method for finding similar sequences:
285 measuring repertoires' binding capacity for the targets of reference sequences.[43] Binding
286 capacity is the average similarity of a repertoire to one or more reference sequences, with
287 similarity estimated according to a general model of the likelihood of a given sequence in the
288 repertoire to bind the same antigen as a reference sequence. In contrast to fuzzy matching, we
289 found the binding capacity remained robust for sample sizes above 1,000 sequences, with only
290 minimal dependence on repertoire size (Fig. 3d). Binding capacity was more robust to
291 repertoire size than fuzzy matching at all tolerances tested (Fig. 3e, Table S7; note that binding
292 capacity does not require a choice of tolerance; it is independent of and therefore robust to
293 tolerance; technically it is a nonlinear weighted mean across all tolerances). Thus, binding
294 capacity provides a robust way to measure the fraction of these TRB repertoires that is similar
295 to reference SARS-CoV-2-specific TRB sequences.

296 **Repertoire features predict levels of NAbs consistent with exposure comparably to clinical**
297 **data**

298 Finally, we compared how well above feature sets predicted exposure-level NAbs titers. To do
299 so, we trained machine-learning models that used each of these feature sets. Because there
300 were many reference SARS-CoV-2-specific TCR sequences to consider, each of which produces
301 one exact-matching fraction, several fuzzy-matching fractions (one for each chosen tolerance),
302 and one binding capacity measurement, there was a risk of overfitting (true whenever the
303 number of features exceeds the number of datapoints). Therefore we first filtered out

304 uninformative features.

305 To do this, we calculated exact/fuzzy matches and binding capacities for SARS-CoV-2 specific
306 and non-specific sequences (from VDJDB) and measured their correlations to NAb titer. (Based
307 on the results above, we only used repertoires with ≥ 1,000 sequences.) We used non-specific
308 sequences as a null model and kept only SARS-CoV-2-specific sequences with correlations
309 outside the middle 95% of the null model: specific sequences on the high end correlated more
310 with NAb titer than was expected by chance, while specific sequences on the low end were
311 correlated inversely with NAb titer to a larger degree than expected by chance (Fig. 4a). Of
312 7,804 SARS-CoV-specific features with non-zero fractions or binding capacities, this process
313 filtered out all but 323. To reduce redundancy and further reduce the number of features, we
314 performed Principal Component Analysis on the results (the number of Principal Components
315 to keep was tunable and fit by each model). We did the same to reduce the number of V-gene
316 features. To avoid data leakage, we performed this dimensionality reduction procedure on
317 training data only.

318 We performed 700 replicate logistic-regression fits on each of the above feature sets and
319 measured performance by AUROC (Fig. 4b). As a comparator, we also fit 700 replicates on
320 subjects' infection and vaccination status, which we reasoned would approximate the
321 maximum possible performance that should be achievable on this dataset. As expected, this
322 comparator resulted in the highest median AUROC of all the feature sets tested, at 0.72 (inter-
323 quartile range across the replicates, 0.66-0.79; Fig. 4b). Strong performance was also seen
324 when training on fuzzy matches with tolerance 2 on all TRB sequences (AUROC 0.71; IQR, 0.64-
325 77) and on TRJ frequencies for CDR4 TRB sequences (0.70; 0.62-0.77). Binding capacities on all
326 TRB sequences showed similar performance to these two (0.68; 0.61-0.74), while exact matches
327 on the same sequences showed poor performance (0.59; 051-0.66).

328 In sum, being infected and/or vaccinated—the gold-standard clinical model—lacked high
329 predictive power for Nab titer, although binding capacities, fuzzy matches with a tolerance of 2,

330 and TRJ frequencies on CD4 TRB sequences performed nearly as well and much better than
331 exact matches.

332 To better understand the characteristics of different feature sets, we also calculated sensitivity,
333 specificity, and precision for all replicates (see Fig. 4c). The clinical feature set's performance
334 metrics are relatively well balanced. In contrast, for most binding-capacity and fuzzy-matching
335 feature sets, sensitivity and precision were low, making them less desirable for screening.
336 Interestingly, features based on IgG diversity had the highest sensitivities while IgM diversities
337 had the lowest sensitivities. The reverse was true for specificities, with IgM diversities having
338 among the highest specificities and IgG diversities among the lowest. While it should be noted
339 that repertoire diversity is not disease specific, these observations suggest that trends in
340 diversity measurements taken for different repertoire subsets might give insights about
341 exposure status in very different ways.

342 Discussion

343 Detecting and defining signatures in repertoire sequence is challenging in part due to the large
344 number of features that can contribute to a signature. These include high-level features such as
345 CDR3 length and repertoire diversity, mid-level features such as frequencies of V and J gene
346 usage and VJ combinations (D genes are harder to assign), and low-level features such as the
347 frequencies of specific reference sequences. Repertoire diversity itself is actually a set of
348 features, some of which incorporate sequence similarity, which furthermore can be defined in
349 multiple ways[38,43]. Ideally the features above should be measured in both antibody and TCR
350 repertoires, since they act cooperatively[44], and in cell subsets defined by isotype (for B cells) or
351 CD4 vs. CD8 expression (for T cells). Thus, overall, the total number of features that can be used
352 to detect and define signatures reaches into the hundreds of thousands.

353 As a result, statistical confidence requires large study sizes, which are challenging to obtain;
354 methods that can avoid spurious associations, which are common in high-dimensional systems;
355 appropriate controls, so that signatures are specific and not related to e.g. general immune

356 activation; and detailed clinical annotation, which we obtained from our electronic medical
357 record (as detailed in Materials and Methods). Even with these design safeguards in place, the
358 signature of exposure to a specific immunogen, such as SARS-CoV-2, may be broad or diffuse,
359 with different individuals' repertoires reacting in different ways. And factors and features
360 outside of repertoires may be important for determining exposure.

361 Given these considerations, our study was fairly large, with over 250 subjects, and involved
362 sequencing IGH and TRD as well as TRB, to a median depth of over 10^5 cells/subject, made
363 possible by ImmunoPETE's integrated library preparation.[5] To focus analysis on SARS-CoV-2-
364 specific signatures and patterns, controls in our study were not typical "healthy controls" but
365 rather patients presenting for care who had sufficient concern for SARS-CoV-2 infection, and
366 who were therefore tested, and were negative. At the time, hospital policy involved widespread
367 testing with very sensitive tests (limit of detection, 100 copies of viral mRNA/mL), so we
368 consider the probability of false negatives to be low. In addition, for infectees and controls, we
369 separately analyzed IGM- and non-IGM (predominantly IGG+)-isotype antibodies and CD4+ and
370 CD4- (i.e. mainly CD8+) T cells. Scheduling issues related to vaccine rollout prevented separate
371 subset analysis for vaccinees, a limitation of the study. We also limited dimensionality, thereby
372 increasing statistical confidence, by filtering for features that correlate with the outcome
373 measure of NAb titer. And instead of simply combining all features into a single model, we
374 compared models with different feature sets to tease apart where signals might lie. Finally, we
375 compared these to the simplest model we could think of, made up of readily available clinical
376 information: whether or not a person was infected and/or vaccinated, to test how repertoire
377 data compares (and what, if anything, it could add). To our knowledge this is the largest such
378 study, and possibly the first. It led to the discovery of several previously unreported patterns
379 across multiple feature sets, for both IGH and TRB, as well as in multiple subtypes of B and T
380 cells, that merit discussion.

381 First, the pattern in IGH CDR3 lengths in vaccinees was curious for several reasons. First, it
382 involved a change in non-productive joins (which in our reading of the literature are usually

383 treated as a baseline and not compared between cohorts, as we did). This was unexpected
384 because B cells are selected for survival based on expressed B-cell receptors, and non-
385 productive joins are not expressed. Our finding seems to indicate selection independent of
386 expression (non-productive joins are not expressed). Second, this is a much larger effect than
387 would be expected from an antigen-specific adaptive immune response. Immunogen-specific B
388 cells rarely exceed low-single-digit percentages of the repertoire. Yet the effect we found
389 appears to involve at least one-sixth (~17%) of the repertoire. Third, the direction of the length
390 change in non-productive joins is opposite that of productive joins: Non-productive CDR3s are
391 longer in vaccinees than in controls and infectees, but productive CDR3s are shorter. And
392 fourth, while other patterns we found were fairly similar between vaccinees and infectees, this
393 CDR3 length effect appears confined to vaccinees.

394 We conclude that vaccination may have some undescribed effect on the V-D-J recombination
395 machinery, biasing recombination toward use of IGHJ genes (and secondarily IGHV genes) that
396 result in longer CDR3s. This effect would have to be due to some difference between the
397 vaccine and natural infection, or else it would have been seen in infectees. If our interpretation
398 is correct, it would mean the effect of selection for shorter CDR3s in productive joins is quite
399 strong, because there are fewer short joins from which to select. In any event, both vaccination
400 (in nonproductive and productive joins) and infection (in productive joins) affect a larger
401 proportion of IGH repertoires than is typically considered "specific."

402 Second, binding capacity was shown to have essentially the same predictive power as the best-
403 performing version of fuzzy matching. Recall that both fuzzy matching and binding capacity
404 measure the size of groups of similar antibodies or TCRs. Here they were applied by taking a
405 reference sequence, for example a sequence previously reported in the literature to be
406 associated with SARS-CoV-2 (a "SARS-CoV-2-specific sequence") and ask what fraction of a
407 given subject's repertoire was similar to that index sequence. The methods differ in how they
408 view similarity. Fuzzy matching requires choice of tolerance: above a set number of amino-acid
409 mismatches, a query sequence is considered different to the index sequence. If the tolerance is

410 2, a query with 3 mismatches is considered just as different from the index sequence as a query
411 with 20 mismatches.

412 Binding capacity has neither problem. It is based on the measured relationship between
413 number of mismatches and change in dissociation constant (K_d), i.e. binding similarity (cite
414 Arora Arnaout 2023). This empirical data essentially substitutes for having to choose a
415 tolerance. In addition, binding capacity is continuous: a query with 3 mismatches is more similar
416 to the index than a query with 20 mismatches. Consequently, binding capacity can detect the
417 potential presence of a large group of sequences with low similarity, which collectively might
418 play as important a role as a small group of high-similarity sequences (or in the limit, the
419 presence of the index sequence as a high-frequency clone). The magnitude of the CDR3 length
420 effect supports the importance of being able to detect such diffuse/weak signals. We showed
421 that different tolerances had different ability to predict NAb titer. To us there is no obvious
422 reason that tolerance of 2 should outperform, e.g., a tolerance of 10. Possibly which tolerance
423 is best may differ by exposure. That binding capacity performs comparably to the best-
424 performing tolerance supports its utility for immune-repertoire analysis.

425 We note that in principle, diffuse effects could reflect superantigen-type activity.[45] However, to
426 our knowledge, clinical evidence for a superantigen in SARS-CoV-2 is scant and has focused on T
427 cells. Although proposed to explain the rare sequela of multisystem inflammatory syndrome in
428 children,[46] wet-lab investigation has not demonstrated superantigen activity in SARS-CoV-2.[47]

429 This study has several limitations. First, we were unable to sort vaccinee samples to obtain
430 separate IgM vs. IgG B-cell repertoires and CD4 vs. CD8 T-cell repertoires due to exigencies at
431 the height of the pandemic. Different subtypes may follow different (even opposite) trends, as
432 did the sensitivities and specificities of classifiers trained on IGG and IGM diversities. Any such
433 patterns in vaccinees were beyond our ability to measure. Second, we used concentrations of
434 SARS-CoV-2 anti-spike NAbs as our proxy of protection. Signals may be present that do not
435 correlate with antibodies binding this particular immunogen. For example, a signal might be

436 seen in T cells or antibodies that bind other SARS-CoV-2 proteins, which we are unable to
437 evaluate given NAbs as a readout. Third, although the sequence data in this study was
438 quantitative, it contained only single-chain, not paired-chain data. Fourth, the ability to define
439 signatures is limited by uncertainty about the specificity of reference sequences. Much effort is
440 being put into methods that predict receptor-antigen binding, but a unified, accepted, and
441 feasible approach to identifying all sequences that bind a given immunogen has yet to be
442 established. Fifth, the quality of binding capacity measurements is limited by the current
443 measure of binding similarity being based on mean behavior[43]; this is expected to improve with
444 additional data and advances in protein structure prediction. And sixth, we did not have the
445 benefit of HLA typing for these samples, which could be valuable because of its influence
446 especially on the TCR repertoire.

447 It will be valuable to see the methodology presented here, with its many steps taken to
448 maximize robustness and avoid statistical artifacts, applied to additional datasets. This will give
449 additional evidence of how well these results and this approach generalize for SARS-CoV-2 in
450 general, for immune responses to variants of the virus, and for other pathogens and
451 immunogens. A careful statistical approach applied to multiple, functional features, measured
452 on unbiased repertoire sequence from TCR and BCR subsets from large cohorts, is, in our
453 opinion, the best way to decipher the rich information that the adaptive immunome encodes.

Acknowledgements

This work was supported by the National Institutes of Health (NIH) National Institute for Allergy and Infectious Diseases (NIAID) R01AI148747; the Massachusetts Life Sciences Center Bits to Bytes program; the Extreme Science and Engineering Discovery Environment, which is supported by NSF Grant ACI-1548562; and the Comet supercomputer at the San Diego Supercomputer Center, which is supported by NSF Grant ACI-1341698.

Conflict of Interest Disclosures

HM, HA, DT, and FR are employees of Roche Molecular Systems Inc. JB, EDH, EC, MY, AM, SD, EW, CC, GM, AK, A-RYC, DHB, SR, SD, and RA have no conflicts of interest to declare.

Materials and Methods

Study subjects

The subjects in this study were patients seeking clinical care at the Beth Israel Deaconess Medical Center (BIDMC), a 743-bed tertiary care medical center in Boston, MA, USA. BIDMC serves a large and diverse population in and around eastern Massachusetts, USA, centered on the Boston metropolitan area.

Institutional review board approval

All work was carried out in accordance with BIDMC's Institutional Review Board protocols 2020P000634, 2021P000109 and 2020P000361.

Cohort assignment

All subjects from whom samples were obtained received RT-qPCR tests performed on two Abbott Molecular platforms: m2000 and Alinity m (Abbott Molecular, Des Plaines, IL, U.S.A.). These detect identical SARS-CoV-2 N and RdRp gene targets and are extremely sensitive for SARS-CoV-2 infection, with limit of detection of 100 copies/mL.[48–50] Infectees had a positive

477 result at the time of sample acquisition. Controls were tested, but negative. COVID-19 test and
478 vaccination information were obtained using SQL queries from BIDMC's clinical data repository
479 and via a dedicated REDCap database set up to facilitate research involving vaccinees.[51]

480 Using these records, subjects were considered infectees if there was a record of a positive
481 COVID-19 test result dated before or on the sample collection date and non-infected otherwise.
482 If no medical record number was available for a subject, their infection status was considered
483 unknown. Subjects were considered vaccinees if vaccination prior to or on the day of sample
484 collection was indicated as the appropriate procedure code in the clinical data repository,
485 recorded in REDCap, or identified from Massachusetts' state Immunization Information System.
486 The entire vaccinated cohort had received mRNA-based vaccines, with 80% receiving Pfizer
487 Comirnaty and the rest receiving Moderna Spikevax. Samples were collected from the vaccinee
488 cohort 4 to 84 days (mean = 44.3 days, standard deviation = 15.3 days) after administration of
489 the initial vaccine dose.

490 Subjects were considered non-vaccinated if the sample collection date preceded 12/15/2020
491 (the date of the first administered COVID-19 vaccine); if there was record of vaccination after
492 sample collection that was annotated as the first dose; if there were two vaccinations after
493 sample collection where the second was annotated as the second dose; or if there were two
494 vaccinations after sample collection within 42 days of each other (consistent with being the
495 primary series). Subjects that did not satisfy vaccinee or non-vaccinated criteria were
496 considered to have unknown vaccination status. Subjects were annotated as unexposed
497 controls if they were non-infected and non-vaccinated. Subjects whose vaccination status was
498 unknown or whose infection status was unknown and were neither vaccinees nor infectees
499 were considered to have an "unknown" SARS-CoV-2 exposure status.

500 **Clinical annotations**

501 *Immunosuppression*

502 Subjects were labelled either "immunosuppressed" or "immunocompetent." Subjects were

503 designated immunosuppressed if at least one of the following criteria was met:

504 - the most recent CD4+ cell count was less than 100 cells/µl;
505 - there was a diagnosis of lymphoma or leukemia associated with a healthcare encounter
506 (visit, admission, or phone call) either before or within 60 days after sample collection;
507 or
508 - the subject was prescribed any of the following medications on an ongoing basis prior to
509 sample collection and with enough refills to include up to 30 days **after**: abatacept,
510 adalimumab, anakinra, azathioprine, basiliximab, budesonide, certolizumab,
511 cyclosporine, daclizumab, dexamethasone, everolimus, etanercept, golimumab,
512 infliximab, ixekizumab, leflunomide, lenalidomide, methotrexate, mycophenolate,
513 natalizumab, pomalidomide, prednisone, rituximab, secukinumab, sirolimus, tacrolimus,
514 tocilizumab, tofacitinib, ustekinumab, and vedolizumab.

515 If none of these criteria were met, subjects were considered immunocompetent.

516 *Demographics*

517 If a subject had a COVID test, the sex and date of birth were read from the corresponding
518 record. Otherwise, we read sex and date of birth from other records of lab specimens, the
519 electronic health record (EHR), or the project's REDCap database (always in structured fields,
520 not using natural-language processing). Self-reported race was read from the EHR.

521 *Risk factors*

522 A semi-automated review of EHRs for ICD-10 diagnosis codes and related entries was used to
523 identify subjects having any of the medical conditions highlighted by the CDC as increasing risk
524 of severe illness from COVID-19.[52] Where feasible, the list of ICD-10 codes indicative of each
525 comorbidity was taken from the Elixhauser Comorbidity Software Refined for ICD-10-CM,[53]
526 version v2022.1, developed for the Healthcare Cost and Utilization Project (HCUP), which is
527 based on the work of Elixhauser et al.[54] In addition to these, another widely used set of

528 comorbidity measures is the Charlson Comorbidity Index.[55] For comorbidities not defined in the
529 HCUP software, the lists of ICD-10 codes defined by this study[56] were used where possible.
530 Comorbidities that were not codified in either resource were identified, where possible, using
531 ICD-10 codes or other automated chart queries, detailed as follows:

532 - Cancer: identified using ICD-10 codes in the HCUP software for "Leukemia,"
533 "Lymphoma," "Metastatic cancer," or "Solid tumor without metastasis, malignant."
534 - Chronic Kidney Disease: identified using ICD-10 codes in the HCUP software for "Renal
535 failure, moderate," and "Renal failure, severe."
536 - Chronic Liver Disease: identified using ICD-10 codes in the HCUP software for "Liver
537 disease, mild," and "Liver disease, moderate to severe."
538 - Chronic Lung Disease: The CDC website stipulates that asthma is of concern "if it's
539 moderate to severe," implying mild asthma is not of concern. The HCUP software
540 includes codes for all degrees of severity of asthma in the definition of "Chronic
541 pulmonary disease." Thus, chronic lunch disease was identified using ICD-10 codes in
542 the HCUP software for "Chronic pulmonary disease," excluding any ICD-10 codes
543 beginning with J452 or J453 (mild intermittent or mild persistent asthma, respectively).
544 - Cystic Fibrosis: Identified by any ICD-10 code beginning with E84.
545 - Dementia or other neurological condition: identified using ICD-10 codes in the HCUP
546 software for "Dementia," "Neurological disorders affecting movement," "Seizures and
547 epilepsy," and "Other neurological disorders."
548 - Diabetes: identified using ICD-10 codes in the HCUP software for "Diabetes with chronic
549 complications" and "Diabetes without chronic complications."
550 - Disabilities: identified using ICD-10 codes in the HCUP software for "Paralysis" plus any
551 ICD-10 code beginning with Q (birth defects and chromosomal abnormalities). Note that
552 this omits many, possibly most, forms of disabilities, including non-congenital blindness
553 and deafness, cognitive impairments not due to chromosomal abnormalities, autism
554 spectrum disorders of unknown etiology, etc., but these are of dubious connection to
555 COVID-19.

556 ● Heart conditions: identified using ICD-10 codes in the HCUP software for "Heart failure,"
557 the ICD-10 codes listed in the referenced study[50] for "Myocardial Infarction," and/or any
558 ICD-10 code starting with any of these prefixes: A1884, A3282, A3681, A381, A395,
559 A5203, B2682, B332, B376, B5881, C452, D8685, G130, G712, G713, G720, G721, G722,
560 G7249, G7281, G7289, G729, G737, I01, I02, I05, I06, I07, I08, I09, I11, I13, I20, I23, I24,
561 I25, I3, I4, I5, I70, I9713, J1082, J1182, O101, OO2912, O903, Q2, R570, S26, T82, and
562 Z95.
563 ● HIV: identified using ICD-10 codes in the HCUP software for "Acquired immune
564 deficiency syndrome."
565 ● Mental health conditions: identified using ICD-10 codes in the HCUP software for
566 "Depression" and "Psychoses." Note that this may omit many other forms of mental
567 illness, such as obsessive-compulsive disorder, post-traumatic stress syndrome,
568 borderline personality disorder, etc. Note that there is overlap between conditions
569 considered mental health conditions and those considered disabilities (such as autism
570 spectrum disorders) as well as between mental health conditions and other medical
571 conditions (such as substance abuse disorders).
572 ● Overweight or obese: Subjects were considered to be overweight or obese if their BMI
573 was ≥25. If multiple BMI or height-and-weight values were recorded in the database
574 over time for a given subject, the value(s) used were those closest in time to the date of
575 sample collection.
576 ● Pregnancy or recent pregnancy: Electronic medical records of all female subjects under
577 the age of 69 were searched for: ICD-10 codes starting with Z3A and records of hospital
578 admissions which include a baby delivery time. The timespans of the pregnancy and
579 puerperium periods were estimated from either type of record. In the case of ICD-10
580 codes starting with Z3A, the final digits of the ICD-10 code encode weeks of gestation at
581 the time of the encounter, from which a start and end date of the pregnancy can be
582 estimated. If only a delivery date is known, the pregnancy is estimated to have begun 40
583 weeks earlier, unless "PRETERM" is found in the free-text diagnosis. Subjects were
584 marked as "pregnancy or recent pregnancy" only if their COVID-19 test date fell

585 between the estimated start date of the pregnancy and 42 days after the estimated end
586 date (to allow for post-term pregnancy). Where there was no COVID test date, the date
587 of the blood sample collection was used.

588 - Sickle cell or Thalassemia: Identified by any ICD-10 code beginning with D56 or D57.
589 - Smoking, current or former: Electronic medical records were searched for any non-zero
590 "Tobacco pack years," and for a free-text description of their tobacco usage including
591 the text "current smoker," "former," "every day," "some days," "light," "heavy," "less
592 than 10," "10+," "yes," or "counseling provided."
593 - Solid organ or blood stem cell transplant: Identified by any ICD-10 code beginning with
594 Z94.
595 - Stroke or cerebrovascular disease: identified using ICD-10 codes in the HCUP software
596 for "Cerebrovascular disease," which includes ICD-10 codes for both CBVD POA and
597 CBVD SQLA.
598 - Substance abuse: identified using ICD-10 codes in the HCUP software for "Drug abuse"
599 and for "Alcohol abuse."
600 - Tuberculosis: Identified by any ICD-10 code beginning with A15.

601 **Sample collection, cell separation, and DNA extraction (Fig. S13-S14)**

602 2mL aliquots were taken from EDTA-anticoagulated venous blood collected in the course of
603 standard clinical care (via "purple-top" tubes; Becton, Dickinson, Franklin Park, NJ). Tubes were
604 stored at 4°C between collection and processing, never more than 12 hours. Each aliquot was
605 mixed 1:1 dilution in phosphate-buffered saline (PBS) and centrifuged over Ficoll-Paque-plus
606 (Cytiva, Marlborough) to obtain peripheral blood mononuclear cells (PBMCs). Plasma was
607 collected and stored at 80°C. PBMCs were washed with PBS and resuspended in a sorting buffer
608 of PBS, 1% bovine serum albumin (BSA), and 0.01% sodium azide.

609 Magnetically-labeled anti-CD4 and anti-IgM microbeads (Miltenyi, Bergisch Gladbach) were
610 used to label and column-separate for infectee and control samples; vaccinee samples cells
611 were not separated. This process divided the samples into CD4-positive T cells and IgM-positive

612 B cells in one fraction, and CD4-negative T cells and IgM-negative B cells (principally CD8-
613 positive T cells and IgG-positive B cells) in the other fraction. DNA was isolated for each fraction
614 using EZ1&2 DNA Blood 350µL kits (Qiagen, Hilden) and the EZ1 Advanced XL automated
615 system (Qiagen, Hilden). DNA concentration was assessed via Nanodrop (Thermo Fisher,
616 Waltham).

617 **Sequencing library preparation**

618 AIRRseq libraries were generated using the immunoPETE method as described.[5] ImmunoPETE is
619 a two-step primer extension based targeted gene enrichment assay designed to specifically
620 target and quantitatively amplify recombined human TRB, TRD, and IGH from genomic DNA
621 simultaneously. Briefly, V gene-based primers containing unique molecular identifiers (UMI) as
622 well as universal PCR amplification handles were annealed to the chromosomal VDJ rearranged
623 loci. The first primer extension products, spanning the VDJ rearrangement, were purified from
624 any remaining oligos by a combination of beads (KAPA HyperPure, Roche) and enzymatic
625 treatment with Thermolabile Exonuclease I (New England Biolabs). A second primer extension
626 and amplification master mix containing a pool of J-gene oligos and an Illumina i7 primer
627 generated VDJ amplicons after 10 cycles of target amplification. Illumina sequencing library
628 amplification was performed using the i7/i5 primer pairs with dual sample indexes. All primer
629 extensions and amplifications were performed using the KAPA Long Range HotStart Ready Mix
630 (Roche). The resulting libraries underwent purification using KAPA HyperPure beads (Roche),
631 followed by quantification with the Qubit dsDNA HS Assay kit (Thermo Fisher) and fragment
632 analysis (Agilent TapeStation). Individual sample libraries were pooled in equal mass. A final
633 round of quantification and fragment analysis was then performed. Finally, libraries were
634 sequenced using the Illumina NextSeq 500/550 High Output Kit v2.5 (300 cycles).

635 **Sequencing and bioinformatics**

636 ImmunoPETE sequencing libraries were analyzed using the Roche in-house bioinformatics
637 pipeline, Daedalus (https://github.com/bioinform/Daedalus). After quality filtering of reads and

638 trimming off primers, the pipeline identified V and J genes using a Smith-Waterman alignment
639 approach (https://github.com/pgngp/swift) against an in-house curated V and J gene database.
640 Original V and J gene data and sequences were sourced from HGNC
641 (https://www.genenames.org/) and ENSEMBL (https://ensemblgenomes.org/). CDR3
642 sequences were identified for all V-J pairs, capturing both functional (functional V/J gene AND
643 coding CDR3) and non-functional (annotated non-functional or pseudogene V/J gene in the
644 database OR stop codon/frameshift in CDR3) rearrangements. Sequences are deduplicated by
645 clustering UMI and CDR3 sequences to identify UMI families. Consensus sequences were
646 derived for the CDR3 and UMI segments of each UMI family, suppressing sequencing and PCR
647 errors, and identifying CDR3 rearrangements at single molecule resolution. High quality CDR3
648 rearrangements were further analyzed for cell counting, clonal diversity, and other calculations.
649 Terms used are listed alphabetically and defined as follows:

650 - Cell count: the total number of functional IGH, TRD, and TRB rearrangements in a
651 sample
652 - Cell type percentages: the total number of functional rearrangements from each heavy
653 chain divided by the total cell count × 100
654 - CDR3 clone: BCR or TCR sequences from the same individual with matching V gene,
655 CDR3 amino acid sequence (CDR3-AA), and J gene assignment arising from two or more
656 UMI families
657 - CDR3 clonal type: BCR or TCR sequences from multiple UMI families from multiple
658 individuals with matching V gene, CDR3-AA, and J gene assignment
659 - Clone count: total number of UMI families from the same individual with the same V
660 gene, CDR3-AA, and J gene
661 - UMI family: a set of reads that have been clustered together based on the similarities of
662 the 9-nt UMI sequence and the CDR3-nt region

663 Both UMI and CDR3 sequences are clustered based on a Levenshtein edit distance of 1,
664 capturing likely PCR and sequencing errors. A UMI family represents a single captured DNA

665 molecular fragment from the immunoPETE reaction.

666 **NAbs ELISA titers**

667 The SARS-CoV-2 Surrogate Virus Neutralization Test Kit (GenScript, L00847-A) is designed to
668 quantify the circulating antibodies capacity of blocking the interaction between the receptor
669 binding domain of SARS-CoV-2 spike glycoprotein and the ACE2 human cell surface receptor.
670 The kit was used according to the manufacturer's instructions as follows. A standard curve was
671 generated using a serial dilution of the standard (GenScript, A02087-20) with a dilution factor of
672 1:2. Each subject's serum sample was mixed with sample dilution buffer (1:10) and horseradish
673 peroxidase-conjugated recombinant SARS-CoV-2 receptor-binding domain (HRP-RBD). The
674 mixture was incubated at 37°C for 30 min to allow the circulating NAbs to bind to HRP-RBD. The
675 mixture was then added to an ACE2 protein-coated plate and incubated for an additional 15
676 min at 37°C. Unbound HRP-RBD and HRP-RBD bound to non-neutralizing antibodies were
677 bound to the plate while circulating neutralization antibody HRP-RBD complexes remained in
678 the supernatant for subsequent wash steps. After washing, tetramethylbenzidine solution was
679 added, followed by a stop solution to quench the reaction, turning wells yellow. The plate was
680 read immediately at 450 nm in a microtiter plate reader. Statistical analysis was performed with
681 GraphPad Prism using a 4PL model for linear regression. Results were reported by interpolating
682 the OD450 values to the standard curve values.

683 Results are reported in the WHO standard units (BAU/mL), which can be converted to the
684 manufacturer's U/mL by dividing numbers by a factor of 1.626 (per the manufacturer).

685 **pymmunomics**

686 Code used for the analyses was written up as a python package and made publicly available on
687 github (https://www.github.com/ArnaoutLab/pymmunomics). Reference is made in the
688 following sections wherever that is the case.

689 **Validation of pooled sequences from sorted subsets vs. unsorted sequences**

690 Additional test samples were split into two replicates. One replicate was sorted into IgM/CD4
691 and IgG/CD8 fractions as above. The other replicate was unsorted. Each of the IgM/CD4,
692 IgG/CD8, and unsorted fractions had libraries prepared as above and were sequenced as above.
693 V- and J-gene frequencies from the IgM/CD4 and IgG/CD8 fractions and from the unsorted
694 fraction were measured, and the frequencies from the IgM/CD4 and IgG/CD8 fractions were
695 averaged together. These averages were compared to the frequencies in the unsorted fraction,
696 and were shown to be indistinguishable by t-test (p-values ranging from 0.19 to 0.95 for
697 comparisons of the frequencies of each gene), demonstrating that IgM/CD4 + IgG/CD8 is
698 statistically indistinguishable from unsorted, supporting comparisons between infectees
699 (sorted) and vaccinees (unsorted) and between controls (sorted) and vaccinees. Dependence of
700 antibody concentrations on age, immunocompetence, and SARS-CoV-2 exposure

701 Univariate and bivariate exploratory plots suggested zero antibody concentration to be a
702 special category. Therefore, we first modeled the ability to produce zero vs. non-zero amounts
703 of antibody using logistic regression. We then performed linear regression to model the $\log_{10}$-
704 transformed concentration of the nonzero values on our set of covariates. In both cases, we
705 started with a full model incorporating age, immunocompetence status, cohort, and all of their
706 two-way and three-way interactions. Starting with the interaction terms and then proceeding
707 to the main effects, we sequentially eliminated covariates that were not significant at α=0.05.
708 This did not change the regression coefficients of any of the significant terms by >20% (i.e. were
709 not confounders). Finally, we confirmed that the best model had lower AIC (logistic regression)
710 or higher adjusted R^2 (linear regression) compared with the alternative models.

711 **CDR3 length analysis**

712 For calculating lengths, the CDR3 region was defined to include the preceding Cysteine anchor
713 residue and the terminating Tryptophan (for IGH) or Phenylalanine (for TCR) anchor residue.
714 CDR3 length frequencies for each available functional and non-functional pooled IGH, TRB, TRD,
715 and subtyped IGG, IGM, CD4 TRB/D, CD8 TCB/D repertoire of immunocompetent subjects were
716 calculated using the pymmunomics python package (above). Since vaccinee samples were not

717 sorted into subtypes, pooled repertoire CDR3 length frequency distributions were used to
718 compare vaccinees to controls and infectees. $CD4^+/IGM^+$ and CD8/IGG repertoire CDR3 length
719 frequency distributions were compared independently between controls to infectees.

720 To compare CDR3 length distributions between cohorts without simplifying them down to their
721 mean or median distribution, which ignores variance within groups, we chose a threshold CDR3
722 length ℓ and compared the cumulative frequencies of sequences on each side of that length
723 using a two-tailed Mann-Whitney-U test. The threshold length was determined by estimating
724 the difference of length frequencies between cohorts for each CDR3 length. These estimates
725 were calculated by taking the median difference in frequency between members of one cohort
726 and members of the other. The dividing line is then placed between the lengths ℓ and ℓ+1,
727 where ℓ is the CDR3 length that maximizes the magnitudes of the total areas under the curve of
728 estimated frequency differences to the left and right of the line, i.e. the best dividing line
729 between patterns:

730 $$\left|\sum_{\ell'<\ell} d_{\ell'}\right| + \left|\sum_{\ell'>\ell} d_{\ell'}\right|$$

731 Here $d_{\ell'}$ denotes the estimated difference of frequencies of CDR3s of length ℓ' between the two
732 cohorts. Note that the absolute values are taken after summing group differences on one side
733 of the dividing line (making positive and negative differences cancel each other out before
734 taking the absolute value), favoring a dividing line that splits the median differences into large
735 same-signed runs. P-values were corrected for multiple hypotheses via the Holm-Bonferroni
736 method (Table S3).

737 To identify trends among lengths of V and J genes, V and J genes (from IMGT) of the relevant
738 cell types (IGH for the functional pooled IGH comparisons and pooled IGH, IGG, and IGM for the
739 non-functional comparisons) which had a corrected p-value below 0.05 were grouped into the
740 number of residues that fall into the CDR3 region. Usage frequencies of V- and J-gene groups

741 were compared between cohorts using two-tailed Mann-Whitney-U and a second correction
742 round was conducted to correct all original p-values of the CDR3 length comparisons at the
743 same time as the p-values obtained from the follow-up tests.

744 **Sets of known SARS-CoV-2 binders and binders to other pathogens**

745 MIRA-identified SARS-CoV-2 specific T-cell receptor sequences[23] were downloaded from
746 https://clients.adaptivebiotech.com/pub/covid-2020 on April 19, 2021.

747 Query B-cell and T-cell receptor sequences (CDR3) of cells known to bind to SARS-CoV-2 were
748 downloaded from CoVAbDab, PDB, and VDJDB. The CovAbDab sequences were downloaded on
749 April 20, 2022 and consists of all SARS-CoV-2-WT-neutralizing human antibodies with CDRH3
750 sequence listed in the database at the time and added since May 04, 2020. PDB sequences
751 were download on May 03, 2022 searching for all structures of source organism Homo sapiens,
752 containing in the title one of “antibody” or “Fab,” and one of “CMV,” “cytomegalovirus,”
753 “DENV” (i.e. dengue virus), “dengue,” “EBV,” “Epstein-Barr,” “hepatitis,” “HIV,” “human
754 immunodeficiency virus,” “influenza,” “SARS-CoV-2,” or “tetanus.” The resulting entries were
755 filtered for sequences in which a CDRH3 sequence of length at least 6 and at most 40 could be
756 detected using in-house Python code. For each sequence, the name of the binding target was
757 extracted from the structure title. VDJDB sequences were also downloaded on April 20, 2022 to
758 obtain human TRB sequences with CDR3 and J-gene specified that bind to their listed target
759 with a non-zero score.

760 To conform with the gene database used for V- and J-gene assignment of repertoire sequences
761 (see Sequencing and bioinformatics), the same gene sequences were aligned (blastp and blastp-
762 short for V genes and J genes, respectively; BLAST+ v2.12.0) to the sequences from PDB and
763 CoVAbDab, setting the max target seqs parameter to 10,000—a number much larger than the
764 total number of genes in the query to avoid missing the best matching genes.[57] V-gene matches
765 with query coverage less than 30% or percent identity less than 40% and J-gene matches with
766 query coverage less than 50% or percent identity less than 40% were filtered out. From the

767 remainder, the best V- and J-gene matches according to percent identity and gene sequence
768 coverage (lexicographically) were assigned to each query sequence. Data downloaded from
769 VDJDB contained sequence only for the CDR3 region, so the V, and J-gene annotation provided
770 by the database was used (as opposed to using e.g. BLAST).

771 To calculate the fractions of query sequences sets matching subject repertoire sequences and
772 the fractions of subject TRB repertoires matching query TRB sequences sets, a pair of sequences
773 is considered to match if their V gene, J gene, and CDR3 sequence are identical.

774 **Binding-capacity measurements**

775 Binding capacities to the MIRA-identified HLA class II T-cell sequences were measured for all
776 subject pooled (CD4+CD8), and CD4 TRB repertoires, wherever possible. The binding capacity of
777 a repertoire R to a clone c is defined as:

778 $$\tau(c; R) = \sum_{c' \in R} p(c') \cdot s(c, c')$$

779 where $p(c')$ denotes the frequency of clone c' in repertoire R and s is the binding similarity
780 between sequences. Here, s as previously described,[43] which accounted only for the
781 relationship between Levenshtein distance of CDR3s and the predicted difference in strength of
782 their binding to the same target(s) (in terms of relative K_d), was constrained as follows to
783 require matching V and/or J genes:

$$s(c, c') = \begin{cases} 0.3^{Lev(c,c')} & \text{if V and J genes match} \\ 0 & \text{otherwise} \end{cases}$$

784 Here, $Lev(c, c')$ is the Levenshtein distance between the CDR3 amino acid sequences of
785 sequences c and c'. The pymmunomics Python package was used to calculate similarity
786 matrices and binding capacities.

787 **Fuzzy query sequence matching**

788 Fuzzy sequence matching measurements for each pooled CD4+CD8 and each CD4-only TRB
789 subject repertoire to the MIRA-identified HLA class II query sequences were tabulated from the
790 similarity matrices that are calculated as part of determining binding capacities. For each query
791 sequence and each subject repertoire, we measured the fraction of repertoire sequences for
792 with the same V and J genes as the query sequence, and whose CDR3 sequence was within
793 Levenshtein distances 0-10 of the query's CDR3. Note that exact matching is equivalent to fuzzy
794 matching with a Levenshtein distance of 0.

795 **Binding-capacity and fuzzy-matching robustness experiments**

796 To compare robustness to variations in repertoire size of binding capacity and fuzzy matching
797 features, we conducted subsampling experiments. We randomly chose 10 subjects from each
798 of the vaccinee, infectee, and control cohorts that had a pooled TRB repertoire size of at least
799 80,000 cells, i.e. 80,000 distinct corrected UMIs. (This size was chosen in order to guarantee at
800 least 10 subjects from the control cohort to choose from.) Each of these repertoires was
801 sampled down to 20 different subsample sizes chosen to be equidistantly spaced between 10
802 and 80,000 at log-scale. For each subsample, we calculated binding capacities as well as fraction
803 of fuzzy matches for fuzzy-match tolerances 0, 2, 4, 6, 8, and 10 amino acids to CD4 TRB
804 reference sequences from MIRA. The slopes and their surrounding 95% confidence intervals
805 were obtained by fitting a linear mixed model that groups the data by subject.

806 **Feature selection**

807 Preferring the use of domain knowledge over generic feature selection mechanisms for
808 selecting from the high-dimensional query sequence matching features (binding capacity and
809 fuzzy matching), a custom feature selection method is developed and implemented in the
810 python package pymmunomics. For this mechanism we use binding capacity and fuzzy
811 matching measurements to sequence specific to pathogens other than SARS-CoV-2 ("SARS-CoV-
812 2 non-specific sequences") as a null distribution to which to compare the measurements for

813 MIRA-identified SARS-CoV-2-specific sequences. We calculated the (Stuart-)Kendall Tau-c
814 correlation coefficient between each feature's measurement and NAb titer. For each feature
815 group (binding capacity, fuzzy matching with tolerances 0, 1, …, 10, etc.), the correlation
816 coefficients of measurements for non-SARS-CoV-2 specific sequences form the null distribution
817 and correlation coefficients of SARS-CoV-2 specific features below the 2.5^{th} and above the
818 97.5^{th} percentile are selected (cumulatively, the most correlated and anti-correlated 5%).

819 Following the same idea, V-gene frequencies were also selected from among the 54 total
820 possibilities (one for each V gene). Here, V-gene frequencies in non-functional repertoires were
821 taken as the null distribution against which to compare functional repertoires' V-gene
822 frequencies, since non-functional sequences do not undergo SARS-CoV-2 specific clonal
823 expansion. Since the functional and non-functional frequencies can be viewed as paired
824 measurements, the distribution of differences between their correlation coefficients was
825 calculated, and the most correlated and anti-correlated 5% (as defined above) were selected as
826 features.

827 **Machine learning to classify subjects with a protective NAb titer**

828 Machine learning classifiers of high or low neutralizing antibody concentration were fit to
829 various feature groups and for various cell types. For the CD4 and pooled TRB receptor
830 repertoires, binding capacities as well as fuzzy matching features with tolerances 0, 1, …, 10 to
831 the MIRA-identified CD4 clones from Nolan et al.[23] were used. Another set of models was
832 derived from these by adding a mechanism at the end of feature selection that aggregates the
833 selected features into their sums. For the pooled IGH, TRB, and TRD as well as the IGM, non-
834 IGM (predominately IGG), CD4 TRB and CD8 TRB repertoires models are fit on the following
835 feature sets:

836 - CDR3 length frequencies, summarized by 3 features: mean, variance and skewness;
837 - diversity, with Recon[58] (https://github.com/ArnaoutLab/Recon) being used to correct
838 Hill D_q numbers for q=0, 1, …, ∞ to correct for missing species;

839 - J-gene frequencies (with only 6 J genes, no further feature selection was required);
840 - V-gene frequencies for select V genes as described above;
841 - Baseline/clinical features: age, sex, days since infection (runs of positive COVID-19 PCR
842 tests successively within 28 days of each other and not interrupted by negative tests are
843 considered infected periods; to account for incubation of the virus prior to taking the
844 test, the start date of an infection is predicted as 4 days before the first positive test in
845 the corresponding run of tests; when a negative test was performed within those 4
846 days, that test's date is considered the infection start date; for the model, the predicted
847 start date of the most recent infection before sample collection was used, or 0 if the
848 subject was not infected), and days since vaccination (the number of days between
849 sample collection and most recent vaccination on record).

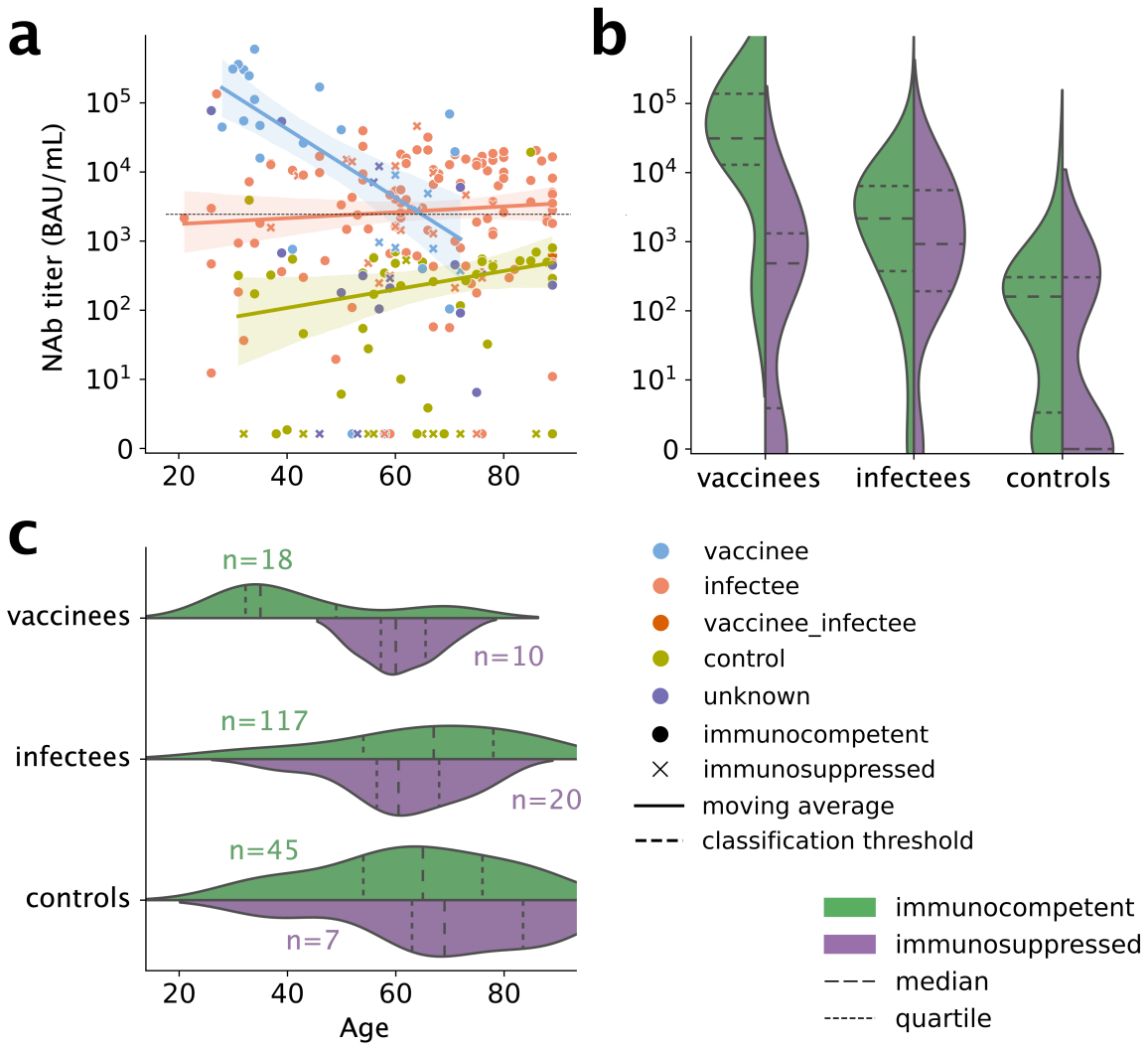
850 The machine learning framework was set up as follows. For each feature group, 700 replicate
851 performances were measured via repetition of 7-fold cross-validation 100 times, each time
852 choosing a different split of the data into 7 folds at random. For each replicate, 10-fold cross-
853 validation was used to tune hyperparameters via Bayesian optimization. For each model fit, the
854 training data was standardized, then underwent principal component analysis, and finally was
855 used to train an L2-regularized regression. There were two tuned hyperparameters:
856 regularization strength (with a log-uniform search space distribution between 10^{-8} and 10^{-2})
857 and the amount of variance to be explained by chosen principal components (with uniform
858 search space distribution between 0.50 and 0.99; e.g. if the value was 0.75 and the first four
859 PCs account for 75% of variance, these four PCs would be chosen). For feature sets relating to
860 similarity—binding capacities, fuzzy-matching features at various tolerances, and their
861 aggregated versions—and for V-gene features, feature selection was performed on the training
862 data before standardization for each model fit. To facilitate avoidance of train-test leakage, the
863 mechanisms are implemented in the pymmunomics python package to fit into the popular
864 scikit-lean API framework.

865 **Data availability**

All raw data for this study can be found at: https://data.entrop.ai/cdr3/data/index.html

867 **Figures and Figure Legends**

868 **Figure 1: Anti-SARS-CoV-2 ELISA trends and distributions by age for immunocompetent and**
869 **immunosuppressed vaccinees, infectees, and controls**

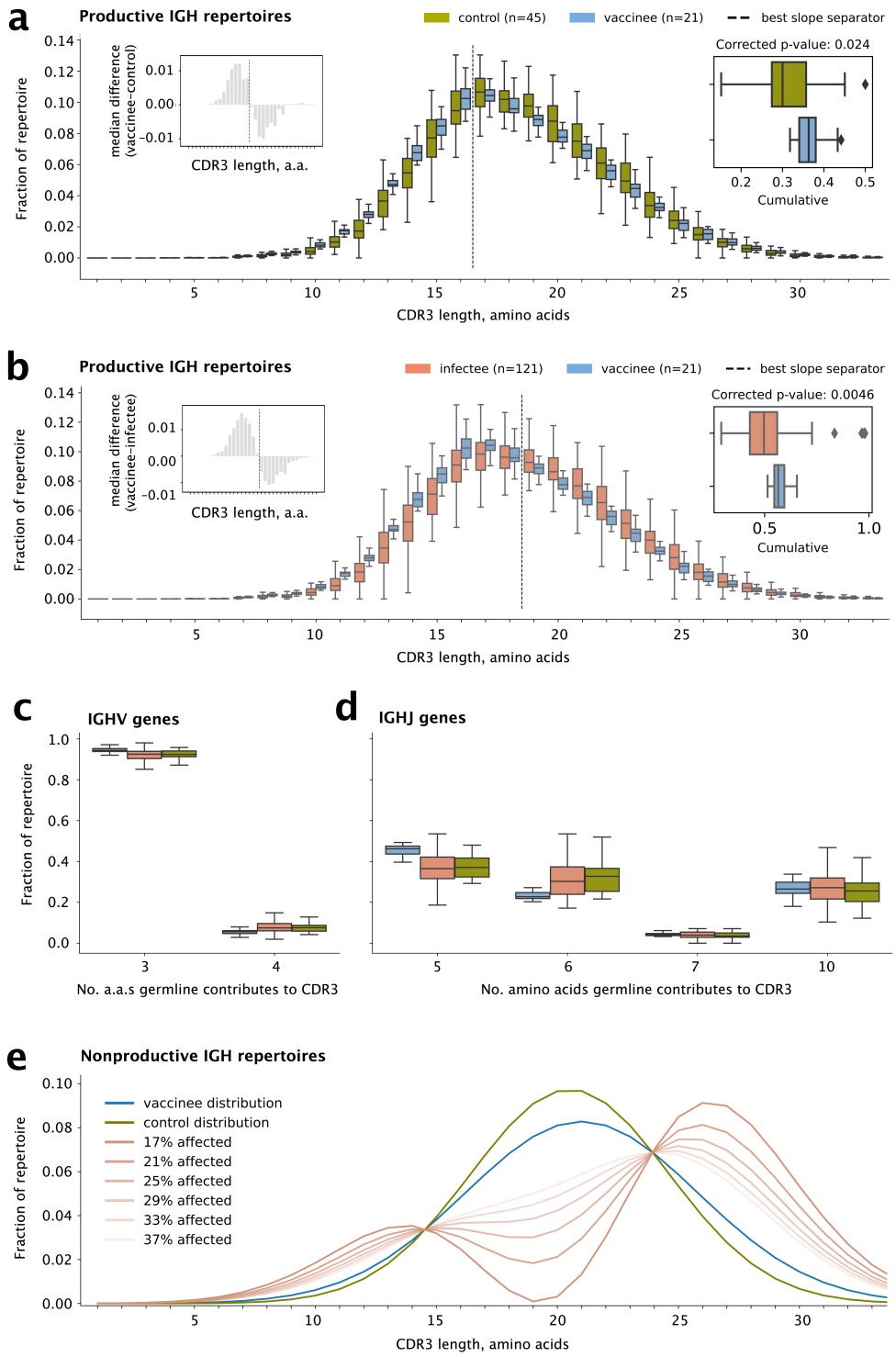


870

871 NAbs target SARS-CoV-2 spike protein. (a) ELISA titers for each subject. Solid lines indicate
872 regression fits; shaded areas indicate 95% confidence intervals. Dotted black line at ~10^3
873 indicates manufacturer's cutoff for positive vs. negative. Note strong negative trend with age in
874 vaccinees (blue) but not infectees (salmon). Note mild positive trend with age in controls
875 (olive), even as titers in this cohort remain below the cutoff for almost all individuals. (b)

Distribution of titers in the three cohorts, split by immune status. (c) Distribution of ages in these cohorts, again split by immune status, with numbers of subjects in each sub-cohort.

879 **Figure 2: IGH CDR3 length distributions**



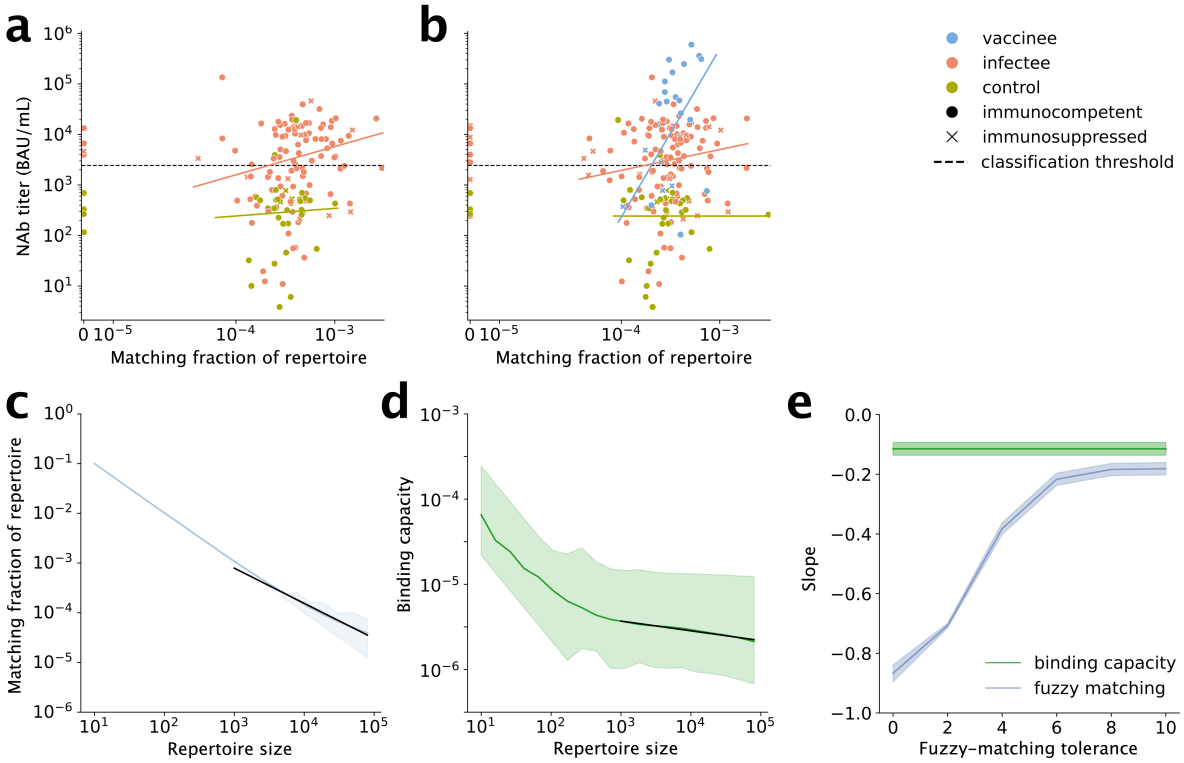
880

881 **(a)** CDR3 length comparison plots for productive IGH repertoires of vaccinees vs. controls. Left

882 inset: the median differences of frequencies at each length, showing that CDR3s of length 16 or
883 shorter are more frequent in vaccinees, whereas CDR3s of length 17 or longer are less frequent.
884 The pattern reverses at the dividing line between 16 and 17 amino acids (vertical dotted line).
885 Right inset: total fraction of the repertoire up to the dividing line. The p-value is obtained by
886 applying Mann-Whitney U to the cumulatives followed by Holm-Bonferroni multiple-hypothesis
887 correction. **(b)** The same for vaccinees vs. infectees, showing the same pattern but with a
888 dividing line between 18 and 19 amino acids. **(c)**-**(d)** Frequencies of V and J genes grouped by
889 the number of residues each gene contributes to the CDR3 according to germline (Fig. S8). Note
890 the only J gene that contributes 5 residues is IGHJ4. **(e)** Assuming the nonproductive IGH
891 vaccinee repertoire (blue) is made up of a part that is unaffected by vaccination and therefore
892 looks like the control repertoire (green) and a part that is affected by vaccination (salmon), this
893 plot shows what the distribution of the affected part would have to look like so the two parts
894 add up correctly, for different fractions affected (dark to light salmon lines). Estimated means
895 for vaccinee and control distributions are shown. The smaller the affected portion, the more
896 extreme the effect must be. The minimum possible effect size is that for which a CDR3 length
897 for the affected portion is zero; any smaller, and a negative frequency at that CDR3 length
898 would be required (negative frequencies are not possible). Boxes show the 25th and 75th
899 percentiles. Horizontal lines show medians. Whiskers show the min and max unless there are
900 outliers (diamonds).

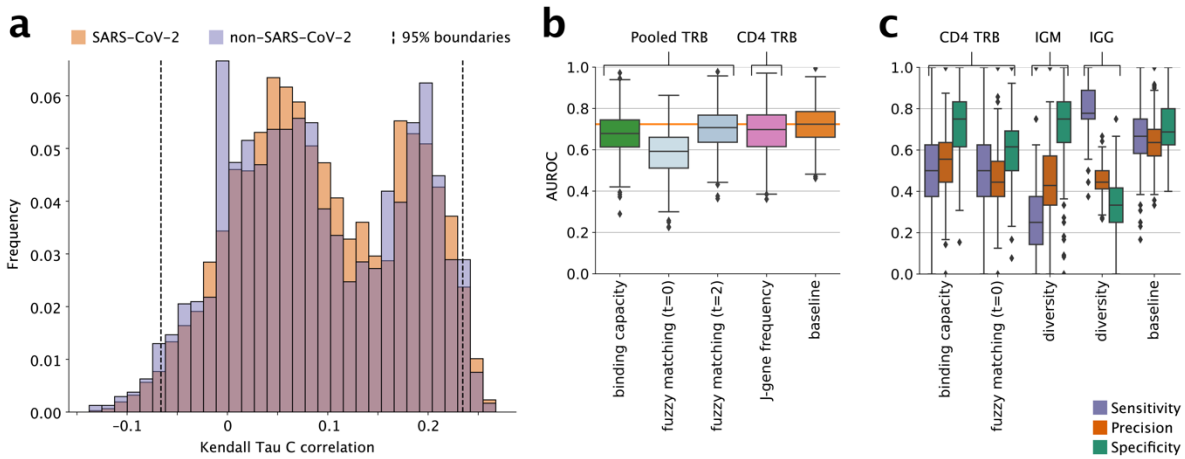
901

Figure 3: SARS-CoV-2-specific TRBs vs. NAb titers



(a)-(b) Fraction of TRB repertoires matching the SARS-CoV-2-specific CD4 TRB sequences obtained from Nolan et al.[23] against SARS-CoV-2 NAb titer. Panel (a) shows repertoires from CD4+ T cells, which were available for infectees and controls but not vaccinees, while panel (b) shows repertoires from all T cells, which were available for all three cohorts. Theil-Sen regression fits (solid lines) show positive relationships for infectees and vaccinees but not controls. **(c)** The fraction of a repertoire that matches reference TRBs within a chosen tolerance (here, 2 amino-acid differences) depends strongly on the number of cells in the repertoire (i.e., repertoire size; fig. S10). **(d)** In contrast, binding capacity is much more robust. The slope of the dependency on size for repertoires above 1,000 cells are shown as black lines. **(e)** Slope as a function of fuzzy-binding tolerance, demonstrating binding capacity is more robust regardless of tolerance.

Figure 4: Predicting positive NAbs



(a) Feature selection mechanism used for binding capacity and fuzzy matching features on the binding capacity measurements of all TCR repertoires of size at least 1,000 using the SARS-CoV-2-specific CD4 TCR sequences and non-SARS-CoV-2-specific TCR sequences obtained from VDJDB. Of the 7,804 SARS-CoV-2-specific features' correlations, 323 fall outside the selection boundaries set by the 95% boundaries of the correlations of non-SARS-CoV-2-specific features with NAb titer. **(b)-(c)** Machine learning performance results for a selected group of feature sets and cell types across all 700 replicates (100 repeats of 7-fold cross-validation). (a) shows areas under receiver operating curves and (b) breaks down the performances into sensitivity, precision, and specificity. The same plots for all feature sets and cell types can be found in Fig. S11 and S12. Median values and interquartile ranges for all metrics are reported in Table S8. Boxes and whiskers as in Fig. 2.

929 **References**

930 1. Arnaout, R.A., Prak, E.T.L., Schwab, N., Rubelt, F., and the Adaptive Immune Receptor
931 Repertoire Community (2021). The Future of Blood Testing Is the Immunome. Front.
932 Immunol. *12*, 626793. 10.3389/fimmu.2021.626793.

933 2. Primorac, D., Vrdoljak, K., Brlek, P., Pavelić, E., Molnar, V., Matišić, V., Erceg Ivkošić, I., and
934 Parčina, M. (2022). Adaptive Immune Responses and Immunity to SARS-CoV-2. Frontiers in
935 Immunology *13*.

936 3. Shen, J., Fan, J., Zhao, Y., Jiang, D., Niu, Z., Zhang, Z., and Cao, G. (2023). Innate and adaptive
937 immunity to SARS-CoV-2 and predisposing factors. Front Immunol *14*, 1159326.
938 10.3389/fimmu.2023.1159326.

939 4. Joseph, M., Wu, Y., Dannebaum, R., Rubelt, F., Zlatareva, I., Lorenc, A., Du, Z.G., Davies, D.,
940 Kyle-Cezar, F., Das, A., et al. (2022). Global patterns of antigen receptor repertoire
941 disruption across adaptive immune compartments in COVID-19. Proc Natl Acad Sci U S A
942 *119*, e2201541119. 10.1073/pnas.2201541119.

943 5. Dannebaum, R., Suwalski, P., Asgharian, H., Du Zhipei, G., Lin, H., Weiner, J., Holtgrewe, M.,
944 Thibeault, C., Müller, M., Wang, X., et al. (2022). Highly multiplexed immune repertoire
945 sequencing links multiple lymphocyte classes with severity of response to COVID-19.
946 EClinicalMedicine *48*, 101438. 10.1016/j.eclinm.2022.101438.

947 6. Fujihashi, K., McGhee, J., Yamamoto, M., Hiroi, T., and Kiyono, H. (1996). Role of gamma
948 delta T cells in the regulation of mucosal IgA response and oral tolerance. Ann N Y Acad Sci
949 *778*, 55–63. 10.1111/j.1749-6632.1996.tb21114.x.

950 7. Rezende, R.M., Cox, L.M., Moreira, T.G., Liu, S., Boulenouar, S., Dhang, F., LeServe, D.S.,
951 Nakagaki, B.N., Lopes, J.R., Tatematsu, B.K., et al. (2023). Gamma-delta T cells modulate the
952 microbiota and fecal micro-RNAs to maintain mucosal tolerance. Microbiome *11*, 32.
953 10.1186/s40168-023-01478-1.

954 8. Brouwer, P.J.M., Caniels, T.G., van der Straten, K., Snitselaar, J.L., Aldon, Y., Bangaru, S.,
955 Torres, J.L., Okba, N.M.A., Claireaux, M., Kerster, G., et al. (2020). Potent neutralizing
956 antibodies from COVID-19 patients define multiple targets of vulnerability. Science *369*,
957 643–650. 10.1126/science.abc5902.

958 9. Cao, Y., Su, B., Guo, X., Sun, W., Deng, Y., Bao, L., Zhu, Q., Zhang, X., Zheng, Y., Geng, C., et
959 al. (2020). Potent Neutralizing Antibodies against SARS-CoV-2 Identified by High-
960 Throughput Single-Cell Sequencing of Convalescent Patients' B Cells. Cell *182*, 73-84.e16.
961 10.1016/j.cell.2020.05.025.

962 10. Chen, E.C., Gilchuk, P., Zost, S.J., Suryadevara, N., Winkler, E.S., Cabel, C.R., Binshtein, E.,
963 Chen, R.E., Sutton, R.E., Rodriguez, J., et al. (2021). Convergent antibody responses to the
964 SARS-CoV-2 spike protein in convalescent and vaccinated individuals. Cell Reports *36*,
965 109604. 10.1016/j.celrep.2021.109604.

966 11. Cerutti, G., Guo, Y., Zhou, T., Gorman, J., Lee, M., Rapp, M., Reddem, E.R., Yu, J., Bahna, F.,
967 Bimela, J., et al. (2021). Potent SARS-CoV-2 neutralizing antibodies directed against spike N-
968 terminal domain target a single supersite. Cell Host & Microbe *29*, 819-833.e7.
969 10.1016/j.chom.2021.03.005.

970 12. Galson, J.D., Schaetzle, S., Bashford-Rogers, R.J.M., Raybould, M.I.J., Kovaltsuk, A.,
971 Kilpatrick, G.J., Minter, R., Finch, D.K., Dias, J., James, L., et al. (2020). Deep sequencing of B
972 cell receptor repertoires from COVID-19 patients reveals strong convergent immune
973 signatures. bioRxiv, 2020.05.20.106294. 10.1101/2020.05.20.106294.

974 13. Lima, N.S., Mukhamedova, M., Johnston, T.S., Wagner, D.A., Henry, A.R., Wang, L., Yang,
975 E.S., Zhang, Y., Birungi, K., Black, W.P., et al. (2022). Convergent epitope specificities, V gene
976 usage and public clones elicited by primary exposure to SARS-CoV-2 variants. bioRxiv,
977 2022.03.28.486152. 10.1101/2022.03.28.486152.

978 14. McCallum, M., De Marco, A., Lempp, F.A., Tortorici, M.A., Pinto, D., Walls, A.C., Beltramello,

979 M., Chen, A., Liu, Z., Zatta, F., et al. (2021). N-terminal domain antigenic mapping reveals a
980 site of vulnerability for SARS-CoV-2. Cell *184*, 2332-2347.e16. 10.1016/j.cell.2021.03.028.

981 15. Rapp, M., Guo, Y., Reddem, E.R., Yu, J., Liu, L., Wang, P., Cerutti, G., Katsamba, P., Bimela,
982 J.S., Bahna, F.A., et al. (2021). Modular basis for potent SARS-CoV-2 neutralization by a
983 prevalent VH1-2-derived antibody class. Cell Reports *35*, 108950.
984 10.1016/j.celrep.2021.108950.

985 16. Wec, A.Z., Wrapp, D., Herbert, A.S., Maurer, D.P., Haslwanter, D., Sakharkar, M., Jangra,
986 R.K., Dieterle, M.E., Lilov, A., Huang, D., et al. (2020). Broad neutralization of SARS-related
987 viruses by human monoclonal antibodies. Science *369*, 731–736. 10.1126/science.abc7424.

988 17. Manso, T., Folch, G., Giudicelli, V., Jabado-Michaloud, J., Kushwaha, A., Nguefack Ngoune,
989 V., Georga, M., Papadaki, A., Debbagh, C., Pégorier, P., et al. (2022). IMGT® databases,
990 related tools and web resources through three main axes of research and development.
991 Nucleic Acids Research *50*, D1262–D1272. 10.1093/nar/gkab1136.

992 18. Simnica, D., Schultheiß, C., Mohme, M., Paschold, L., Willscher, E., Fitzek, A., Püschel, K.,
993 Matschke, J., Ciesek, S., Sedding, D.G., et al. (2021). Landscape of T-cell repertoires with
994 public COVID-19-associated T-cell receptors in pre-pandemic risk cohorts. Clinical &
995 Translational Immunology *10*, e1340. 10.1002/cti2.1340.

996 19. Ellwardt, E., Walsh, J.T., Kipnis, J., and Zipp, F. (2016). Understanding the Role of T Cells in
997 CNS Homeostasis. Trends in Immunology *37*, 154–165. 10.1016/j.it.2015.12.008.

998 20. Matschke, J., Lütgehetmann, M., Hagel, C., Sperhake, J.P., Schröder, A.S., Edler, C.,
999 Mushumba, H., Fitzek, A., Allweiss, L., Dandri, M., et al. (2020). Neuropathology of patients
1000 with COVID-19 in Germany: a post-mortem case series. The Lancet Neurology *19*, 919–929.
1001 10.1016/S1474-4422(20)30308-2.

1002 21. Schwabenland, M., Salié, H., Tanevski, J., Killmer, S., Lago, M.S., Schlaak, A.E., Mayer, L.,
1003 Matschke, J., Püschel, K., Fitzek, A., et al. (2021). Deep spatial profiling of human COVID-19

1004 brains reveals neuroinflammation with distinct microanatomical microglia-T-cell
1005 interactions. Immunity *54*, 1594-1610.e11. 10.1016/j.immuni.2021.06.002.

1006 22. Arora, R., and Arnaout, R. (2020). Private Antibody Repertoires Are Public.
1007 10.1101/2020.06.18.159699.

1008 23. Nolan, S., Vignali, M., Klinger, M., Dines, J., Kaplan, I., Svejnoha, E., Craft, T., Boland, K.,
1009 Pesesky, M., Gittelman, R., et al. (2020). A large-scale database of T-cell receptor beta
1010 (TCRβ) sequences and binding associations from natural and synthetic exposure to SARS-
1011 CoV-2. 10.21203/rs.3.rs-51964/v1.

1012 24. Swanson, P.A., Padilla, M., Hoyland, W., McGlinchey, K., Fields, P.A., Bibi, S., Faust, S.N.,
1013 McDermott, A.B., Lambe, T., Pollard, A.J., et al. (2021). AZD1222/ChAdOx1 nCoV-19
1014 vaccination induces a polyfunctional spike protein–specific T_H 1 response with a diverse
1015 TCR repertoire. Sci. Transl. Med. *13*, eabj7211. 10.1126/scitranslmed.abj7211.

1016 25. Bates, T.A., Leier, H.C., Lyski, Z.L., Goodman, J.R., Curlin, M.E., Messer, W.B., and Tafesse,
1017 F.G. (2021). Age-Dependent Neutralization of SARS-CoV-2 and P.1 Variant by Vaccine
1018 Immune Serum Samples. JAMA *326*, 868–869. 10.1001/jama.2021.11656.

1019 26. Bonfante, F., Costenaro, P., Cantarutti, A., Di Chiara, C., Bortolami, A., Petrara, M.R.,
1020 Carmona, F., Pagliari, M., Cosma, C., Cozzani, S., et al. (2021). Mild SARS-CoV-2 Infections
1021 and Neutralizing Antibody Titers. Pediatrics *148*, e2021052173. 10.1542/peds.2021-052173.

1022 27. Dyer, A.H., Noonan, C., McElheron, M., Batten, I., Reddy, C., Connolly, E., Pierpoint, R.,
1023 Murray, C., Leonard, A., Higgins, C., et al. (2022). Previous SARS-CoV-2 Infection, Age, and
1024 Frailty Are Associated With 6-Month Vaccine-Induced Anti-Spike Antibody Titer in Nursing
1025 Home Residents. Journal of the American Medical Directors Association *23*, 434–439.
1026 10.1016/j.jamda.2021.12.001.

1027 28. Sadighi Akha, A.A. (2018). Aging and the immune system: An overview. Journal of
1028 Immunological Methods *463*, 21–26. 10.1016/j.jim.2018.08.005.

1029 29. Haveri, A., Ekström, N., Solastie, A., Virta, C., Österlund, P., Isosaari, E., Nohynek, H., Palmu,
1030 A.A., and Melin, M. (2021). Persistence of neutralizing antibodies a year after SARS-CoV-2
1031 infection in humans. European Journal of Immunology *51*, 3202–3213.
1032 10.1002/eji.202149535.

1033 30. Korobova, Z.R., Zueva, E.V., Arsentieva, N.A., Batsunov, O.K., Liubimova, N.E., Khamitova,
1034 I.V., Kuznetsova, R.N., Rubinstein, A.A., Savin, T.V., Stanevich, O.V., et al. (2022). Changes in
1035 Anti-SARS-CoV-2 IgG Subclasses over Time and in Association with Disease Severity. Viruses
1036 *14*, 941. 10.3390/v14050941.

1037 31. Luo, Y.R., Chakraborty, I., Yun, C., Wu, A.H.B., and Lynch, K.L. (2021). Kinetics of Severe
1038 Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Antibody Avidity Maturation and
1039 Association with Disease Severity. Clinical Infectious Diseases *73*, e3095–e3097.
1040 10.1093/cid/ciaa1389.

1041 32. Moss, P. (2022). The T cell immune response against SARS-CoV-2. Nat Immunol *23*, 186–
1042 193. 10.1038/s41590-021-01122-w.

1043 33. Morgan, A., Contreras, E., Yasuda, M., Dutta, S., Hamel, D., Shankar, T., Balallo, D., Riedel,
1044 S., Kirby, J.E., Kanki, P.J., et al. (2023). The Coviral Portal: Multi-Cohort Viral Loads and
1045 Antigen-Test Virtual Trials for COVID-19. 10.1101/2023.05.05.23289582.

1046 34. Brown, A.J., Snapkov, I., Akbar, R., Pavlović, M., Miho, E., Sandve, G.K., and Greiff, V. (2019).
1047 Augmenting adaptive immunity: progress and challenges in the quantitative engineering
1048 and analysis of adaptive immune receptor repertoires. Mol. Syst. Des. Eng. *4*, 701–736.
1049 10.1039/C9ME00071B.

1050 35. Brown, S.D., Raeburn, L.A., and Holt, R.A. (2015). Profiling tissue-resident T cell repertoires
1051 by RNA sequencing. Genome Med *7*, 125. 10.1186/s13073-015-0248-x.

1052 36. Huang, H., Wang, C., Rubelt, F., Scriba, T.J., and Davis, M.M. (2020). Analyzing the
1053 Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2

1054 and genome-wide antigen screening. Nat Biotechnol *38*, 1194–1202. 10.1038/s41587-020-
1055 0505-4.

1056 37. Mayer, A. and Curtis G. Callan (2023). Measures of epitope binding degeneracy from T cell
1057 receptor repertoires. Proceedings of the National Academy of Sciences *120*, e2213264120.
1058 10.1073/pnas.2213264120.

1059 38. Kaplinsky, J., Li, A., Sun, A., Coffre, M., Koralov, S.B., and Arnaout, R. (2014). Antibody
1060 repertoire deep sequencing reveals antigen-independent selection in maturing B cells. Proc
1061 Natl Acad Sci U S A *111*, E2622-9. 10.1073/pnas.1403278111.

1062 39. Murphy, K., and Weaver, C. (2016). Janeway's Immunobiology (Garland Science).

1063 40. Raybould, M.I.J., Kovaltsuk, A., Marks, C., and Deane, C.M. (2020). CoV-AbDab: the
1064 Coronavirus Antibody Database. bioRxiv, 2020.05.15.077313. 10.1101/2020.05.15.077313.

1065 41. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N.,
1066 and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

1067 42. Goncharov, M., Bagaev, D., Shcherbinin, D., Zvyagin, I., Bolotin, D., Thomas, P.G., Minervina,
1068 A.A., Pogorelyy, M.V., Ladell, K., McLaren, J.E., et al. (2022). VDJdb in the pandemic era: a
1069 compendium of T cell receptors specific for SARS-CoV-2. Nature Methods *19*, 1017–1019.
1070 10.1038/s41592-022-01578-0.

1071 43. Arora, R., and Arnaout, R. (2022). Repertoire-scale measures of antigen binding. Proc. Natl.
1072 Acad. Sci. U.S.A. *119*, e2203505119. 10.1073/pnas.2203505119.

1073 44. Arnaout, R.A., and Nowak, M.A. (2000). Competitive coexistence in antiviral immunity. J
1074 Theor Biol *204*, 431–441. 10.1006/jtbi.2000.2027.

1075 45. Arnaout, R., Lee, R.A., Lee, G.R., Callahan, C., Cheng, A., Yen, C.F., Smith, K.P., Arora, R., and
1076 Kirby, J.E. (2021). The Limit of Detection Matters: The Case for Benchmarking Severe Acute
1077 Respiratory Syndrome Coronavirus 2 Testing. Clinical Infectious Diseases *73*, e3042–e3046.

1078 10.1093/cid/ciaa1382.

1079 46. Callahan, C., Lee, R.A., Lee, G.R., Zulauf, K., Kirby, J.E., and Arnaout, R. (In press). Nasal Swab
1080 Performance by Collection Timing, Procedure, and Method of Transport for Patients with
1081 SARS-CoV-2. J Clin Microbiol.

1082 47. Callahan, C., Ditelberg, S., Dutta, S., Littlehale, N., Cheng, A., Kupczewski, K., McVay, D.,
1083 Riedel, S., Kirby, J.E., and Arnaout, R. (2021). Saliva is Comparable to Nasopharyngeal Swabs
1084 for Molecular Detection of SARS-CoV-2. Microbiol Spectr *9*, e0016221.
1085 10.1128/Spectrum.00162-21.

1086 48. Liu, J., Chandrashekar, A., Sellers, D., Barrett, J., Jacob-Dolan, C., Lifton, M., McMahan, K.,
1087 Sciacca, M., VanWyk, H., Wu, C., et al. (2022). Vaccines elicit highly conserved cellular
1088 immunity to SARS-CoV-2 Omicron. Nature *603*, 493–496. 10.1038/s41586-022-04465-y.

1089 49. People with Certain Medical Conditions (2023). Centers for Disease Control and Prevention.
1090 https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-
1091 medical-conditions.html.

1092 50. ELIXHAUSER COMORBIDITY SOFTWARE REFINED FOR ICD-10-CM Agency for Healthcare
1093 Research and Quality. https://hcup-
1094 us.ahrq.gov/toolssoftware/comorbidityicd10/comorbidity_icd10.jsp.

1095 51. Elixhauser, A., Steiner, C., Harris, D.R., and Coffey, R.M. (1998). Comorbidity measures for
1096 use with administrative data. Med Care *36*, 8–27. 10.1097/00005650-199801000-00004.

1097 52. Charlson, M.E., Pompei, P., Ales, K.L., and MacKenzie, C.R. (1987). A new method of
1098 classifying prognostic comorbidity in longitudinal studies: development and validation. J
1099 Chronic Dis *40*, 373–383. 10.1016/0021-9681(87)90171-8.

1100 53. Garry, E.M., Weckstein, A.R., Quinto, K., Bradley, M.C., Lasky, T., Chakravarty, A., Leonard,
1101 S., Vititoe, S.E., Easthausen, I.J., Rassen, J.A., et al. (2022). Categorization of COVID-19

1102 severity to determine mortality risk. Pharmacoepidemiol Drug Saf *31*, 721–728.
1103 10.1002/pds.5436.

1104 54. Shah, N., Nute, M.G., Warnow, T., and Pop, M. (2019). Misunderstood parameter of NCBI
1105 BLAST impacts the correctness of bioinformatics workflows. Bioinformatics *35*, 1613–1614.
1106 10.1093/bioinformatics/bty833.

1107 55. Kaplinsky, J., and Arnaout, R. (2016). Robust estimates of overall immune-repertoire
1108 diversity from high-throughput measurements on samples. Nat Commun *7*, 11881.
1109 10.1038/ncomms11881.

1110