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], Hamid Mirebrahim[2], Hosseinali Asgharian[2], Dilduz Telman[2], Ai-Ris Y. Collier[3,4,5], Dan H. Barouch[3,5,6,7], Stefan Riedel[1,3], Sanjucta Dutta[1], Florian Rubelt[2], and Ramy Arnaout[1,3,8,†]

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

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, but not infection, has a substantial effect on non-productively recombined IGHs, suggesting a vaccine 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 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.

Introduction

Since the emergence of COVID-19 there has been great interest in identifying antibody (BCR) and T-cell receptor (TCR) gene sequences that are specific to SARS-CoV-2. The pandemic presented perhaps the highest-profile opportunity to test the extent to which TCRs and BCRs against respiratory viruses would be public, i.e. with sequences appearing across many different individuals, or private, present in only one or a few individuals. New SARS-CoV-2-specific BCRs could form the basis for new treatments. Understanding how to identify and characterize

51 commonalities in such an important real-world setting could help evaluate the viability of new
52 diagnostics based on adaptive immune-receptor repertoire sequencing (AIRRseq).[1]

53 A number of studies have succeeded in identifying public immunoglobulin heavy-chain (IGH)
54 and TCRβ (TRB) sequences. However, in part due to exigencies and constraints imposed by the
55 pandemic, and in part because it was impossible to know a priori what study size would be
56 adequate to identify public sequences comprehensively in COVID-19, many of these studies
57 involved relatively small numbers of individuals. Small sample sizes have known limitations for
58 AIRRseq studies. Because small samples may not be representative of larger populations,
59 results on small samples may not generalize. Small studies may be insufficiently powered to
60 detect subtle patterns. And the smaller the sample size, the more likely it is that random
61 fluctuations in the data—literally, the luck of the draw—will produce results that appear to be
62 statistically significant but do not reflect underlying relationships. Moreover, most studies
63 investigated only BCRs or only TCRs, but not both in the same cohort, despite the importance of
64 both antibody and T-cell responses in SARS-CoV-2 infection.[2,3] To our knowledge only two
65 previous COVID-19 studies[4,5] have sequenced TCR from γδ T cells, a little-studied subset that
66 may be important in mucosal antimicrobial immunity.[6,7] How SARS-CoV-2 virus or vaccine
67 exposure affects different B- and T-cell subsets (IgM+ vs. non-IgM+ B-cells, CD4+ vs. CD8+ T
68 cells) has also been insufficiently explored.[2,3]

69 With these caveats in mind, to our knowledge previous studies have identified 20 IGH V genes
70 to be enriched in sequences produced during various immune responses to SARS-CoV-2.[8–16]
71 Given that human genomes encode 54 IGH V genes,[17] collectively these studies implicate 37%
72 of V genes in the response to this single viral exposure, indicating that the SARS-CoV-2 response
73 is either quite broad within individuals, quite heterogeneous among individuals, or both. There
74 is no obvious reason to think each V gene would contribute to only a single SARS-CoV-2-specific
75 recombined IGH sequence, much less a single IGH:immunoglobulin light chain (IGL) pair;
76 therefore, collectively these studies also suggest that the IGH response to SARS-CoV-2 might
77 account for a significant fraction of a given repertoire, a possibility that requires more

78 comprehensive sequence-level investigation such as AIRRseq can provide. Note that studies
79 that compare only a single non-control cohort to a control cohort cannot distinguish between
80 features (clones, motifs, genes, CDR3 lengths, etc.) that signify a disease-specific vs. a general
81 immune response.

82 Regarding TCRs, one study[18] searched repertoires of 140 COVID-19 patients and another 140
83 pre-pandemic (and therefore unexposed) controls for the presence of each of 1,267 TRB
84 sequences that had independently been shown to recognize epitopes of the SARS-CoV-2 spike
85 protein. These authors showed that while the presence of some of the TRB sequences in almost
86 all of the repertoires suggested a public response to SARS-CoV-2 infection, the fraction of the
87 repertoire that matched the query sequences was similar between infectees and controls.
88 These authors also looked for SARS-CoV-2 specific TRBs in the brain tissue of COVID-19 patients,
89 since T-cell infiltration of the brain, an organ otherwise seldom infiltrated by T cells, is known to
90 occur during COVID-19 infection.[19–21] The 68 TRBs they identified were found in 40% of COVID-
91 19 repertoires vs. 17% of pre-pandemic controls. This suggests significant enrichment even as
92 the majority of individuals with COVID-19 lacked these sequences and a significant minority of
93 controls (1 in 6) had them despite these control samples having been collected before the
94 pandemic (which has been observed in other contexts,[22] perhaps indicating cross-reactivity
95 with previously circulating coronaviruses, which are very common in human populations[3]).

96 In addition, one study[23] identified a large database of SARS-CoV-2-specific TRB sequences as
97 being shared among infectees and unenriched among healthy controls. Subsequently, another
98 study[24] sequenced repertoires of individuals 0 and 4 weeks after vaccination with the Oxford-
99 AstraZenica COVID-19 vaccine AZD1222 or the meningococcus vaccine MenACWY and matched
100 the database sequences to these repertoires. An increase of database sequences among
101 AZD1222-vaccinee repertoires but not MenACWY-vaccinee controls was seen between the two
102 timepoints.

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

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

123 In all, the work above supports the view that there are commonalities in IGH and TRB at the
124 gene and sequence level in response to SARS-CoV-2 infection; however the nature of the
125 signature is not well understood. One open question is to what extent infection affects
126 antibody and TCR repertoires as a whole vs. enriching specific clones within it. One can refer to
127 these ends of the continuum of possible effects as “diffuse” vs. “precise.” From previous work
128 on repertoires, “diffuse” features include CDR3 length, the frequency of usage of specific V or J
129 genes, and repertoire diversity as measured any of several ways (richness, Shannon entropy,

130 Simpson's index, or their Hill-number equivalents).[34] At the other extreme, the most "precise"
131 features are the frequency of clones with specific sequences. Between these extremes is a
132 set of features that includes fuzzy matching of sequences[35] and other clustering methods.[36,37]
133 This middle ground has been less explored. Recently the concept of binding capacity has been
134 developed to measure the fraction of a repertoire that is "like" a given query sequence in terms
135 of target specificity (weighting the repertoire by the predicted dissociation constants of its
136 constituent antibodies or TCRs and by their sequence frequencies). Whether or how binding
137 capacity might be affected by SARS-CoV-2 infection and/or vaccination is unknown.

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

143 Results

144 NAbs vary with exposure, age, and immune status

145 Using immunoPETE (Roche; research use only),[5] we deep-sequenced IGH, TRB, and TRD from
146 the blood of 251 individuals: 36 vaccinees, 145 infectees, 53 healthy controls, and 20 with
147 unknown SARS-CoV-2 exposure status. Three individuals belonged to both the vaccinee and
148 infectee groups. Forty-seven subjects were considered immunosuppressed and the remaining
149 204 immunocompetent. Blood samples for 129/145 (89%) infectees were within 6 months of
150 the most recent positive PCR test on record and 121/145 (83%) were ≥7 days from the
151 presumed most recent infection date (assuming a mean of 4 days from exposure to testing).
152 Fig. S1 presents a summary of the timeline and sequencing yield. Tables S1 and S2 present
153 demographics and relevant comorbidities for the different cohorts. We measured plasma NAbs
154 against SARS-CoV-2 spike for 237/251 subjects. Fig. 1 presents a summary of the measured
155 NAbs concentrations by cohort, immunosuppression status, and age.

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

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

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

171 The characteristic (e.g. mean or median) length of CDR3s is known to vary during development
172 and in response to various exposures, at least in productively recombined IGH genes, a.k.a.
173 "productive joins."[38] Because only productively recombined IGH genes can be expressed as
174 (BCR) proteins, such differences are generally considered evidence that the B cells that express
175 them are selected for having e.g. longer CDR3s. We found that vaccinees had shorter IGH
176 CDR3s than controls (p_c=0.024; Fig. 2a) or infectees (p_c=0.0046; Fig. 2b), indicating a repertoire-
177 wide difference in the B-cell response to vaccination vs. infection (Table S3, Figs. S2-S4).

178 The length of IGH CDR3s depends on the lengths of the constituent IGHV, IGD, and IGHJ genes,
179 as well as the number of N and P nucleotides inserted at the junctions between them.[39]
180 Annotating IGD and distinguishing mutated/truncated IGD sequence from N and P sequences is
181 challenging due to insertions, chewbacks, and somatic hypermutation. However, IGHV and IGHJ

182 can be annotated reliably, and so we tested whether the overall differences in CDR3 length
183 were attributable to differences in the use of longer vs. shorter IGHV and IGHJ genes.

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

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 and CD8 TRB
257 repertoires, the positive correlation was confined to CD4 repertoires. In contrast, no correlation
258 was seen for controls. Likewise, no correlation was seen for TRBs that were not specific for

259 SARS-CoV-2 in infectees, supporting the interpretation that this correlation is causal.
260 Nevertheless, this correlation alone performed poorly as a classifier of who had high enough
261 NAbs to be considered positive (per the ELISA test manufacturer), with an area under the
262 receiver-operator characteristic curve (AUROC) of 0.55 (95%CI, 0.46-0.63). Notably, there was
263 also a positive relationship between non-specific TRBs and NAbs in vaccinees, although the 95%
264 CI on the regression slope only narrowly missed including zero (Table S6).

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

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

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

283 We therefore tested a recently proposed alternative method for finding similar sequences:
284 measuring repertoires' binding capacity for the targets of reference sequences.[43] Binding

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

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

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

304 To do this, we calculated exact/fuzzy matches and binding capacities for SARS-CoV-2 specific
305 and non-specific sequences (from VDJDB) and measured their correlations to NAb titer. (Based
306 on the results above, we only used repertoires with at least ≥ 1, 000 sequences.) We used non-
307 specific sequences as a null model and kept only SARS-CoV-2-specific sequences with
308 correlations outside the middle 95% of the null model: specific sequences on the high end
309 correlated more with NAb titer than was expected by chance, while specific sequences on the
310 low end were correlated inversely with NAb titer to a larger degree than expected by chance
311 (Fig. 4a). Of 7,804 SARS-CoV-specific features with non-zero fractions or binding capacities, this

312 process filtered out all but 323. To reduce redundancy and further reduce the number of
313 features, we performed PCA on the results (the number of PCs to keep was tunable and fit by
314 each model). We did the same to reduce the number of V-gene features. To avoid data leakage,
315 we performed this dimensionality reduction procedure on training data only.

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

326 In sum, being infected and/or vaccinated—the gold-standard clinical model—lacked high
327 predictive power for Nab titer, although binding capacities, fuzzy matches with a tolerance of 2,
328 and TRJ frequencies on CD4 TRB sequences performed nearly as well and much better than
329 exact matches.

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

338 diversity measurements taken for different repertoire subsets might give insights about
339 exposure status in very different ways.

340 Discussion

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

351 As a result, statistical confidence requires large study sizes, which are challenging to obtain;
352 methods that can avoid spurious associations, which are common in high-dimensional systems;
353 appropriate controls, so that signatures are specific and not related to e.g. general immune
354 activation; and detailed clinical annotation, which we obtained from our electronic medical
355 record (as detailed in Materials and Methods). Even with these design safeguards in place, the
356 signature of exposure to a specific immunogen, such as SARS-CoV-2, may be broad or diffuse,
357 with different individuals' repertoires reacting in different ways. And factors and features
358 outside of repertoires may be important for determining exposure.

359 Given these considerations, our study was fairly large, with over 250 subjects, and involved
360 sequencing IGH and TRD as well as TRB, to a median depth of over 10^5 cells/subject, made
361 possible by ImmunoPETE's integrated library preparation.[5] To focus analysis on SARS-CoV-2-
362 specific signatures and patterns, controls in our study were not typical "healthy controls" but
363 rather patients presenting for care who had sufficient concern for SARS-CoV-2 infection, and

364 who were therefore tested, and were negative. At the time, hospital policy involved widespread
365 testing with very sensitive tests (limit of detection, 100 copies of viral mRNA/mL), so we
366 consider the probability of false negatives to be low. In addition, for infectees and controls, we
367 separately analyzed IGM- and non-IGM (predominantly IGG)-isotype antibodies and CD4 and
368 CD8 T cells. (Scheduling issues related to vaccine rollout prevented separate subset analysis for
369 vaccinees, a limitation of the study.) We also limited dimensionality, thereby increasing
370 statistical confidence, by filtering for features that correlate with the outcome measure of NAb
371 titer. And instead of simply combining all features into a single model, we compared models
372 with different feature sets to tease apart where signals might lie. Finally, we compared these to
373 the simplest model we could think of, made up of readily available clinical information: whether
374 or not a person was infected and/or vaccinated, to test how repertoire data compares (and
375 what, if anything, it could add). To our knowledge this is the largest such study, and possibly the
376 first. It led to several previously unreported patterns across multiple feature sets, for both IGH
377 and TRB, as well as in multiple subtypes of B and T cells, that merit discussion.

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

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

399 Second, binding capacity was shown to have essentially the same predictive power as the best-
400 performing version of fuzzy matching. Recall that both fuzzy matching and binding capacity
401 measure the size of groups of similar antibodies or TCRs. Here they were applied by taking a
402 reference sequence, for example a sequence previously reported in the literature to be
403 associated with SARS-CoV-2 (a "SARS-CoV-2-specific sequence") and ask what fraction of a
404 given subject's repertoire was similar to that index sequence. The methods differ in how they
405 view similarity. Fuzzy matching requires choice of tolerance: above a set number of amino-acid
406 mismatches, a query sequence is considered different to the index sequence. If the tolerance is
407 2, a query with 3 mismatches is considered just as different from the index sequence as a query
408 with 20 mismatches.

409 Binding capacity has neither problem. It is based on the measured relationship between
410 number of mismatches and change in dissociation constant (K_d), i.e. binding similarity (cite
411 Arora Arnaout 2023). This empirical data essentially substitutes for having to choose a
412 tolerance. In addition, binding capacity is continuous: a query with 3 mismatches is more similar
413 to the index than a query with 20 mismatches. Consequently, binding capacity can detect the
414 potential presence of a large group of sequences with low similarity, which collectively might
415 play as important a role as a small group of high-similarity sequences (or in the limit, the
416 presence of the index sequence as a high-frequency clone). The magnitude of the CDR3 length
417 effect supports the importance of being able to detect such diffuse/weak signals. We showed

418 that different tolerances had different ability to predict NAb titer. To us there is no obvious
419 reason that tolerance of 2 should outperform, e.g., a tolerance of 10. Possibly which tolerance
420 is best may differ by exposure. That binding capacity performs comparably to the best-
421 performing tolerance supports its utility for immune-repertoire analysis.

422 This study has several limitations. First, we were unable to sort vaccinee samples to obtain
423 separate IGM vs. IGG and CD4 vs. CD8 repertoires due to exigencies at the height of the
424 pandemic. Different subtypes may follow different (even opposite) trends, as did the
425 sensitivities and specificities of classifiers trained on IGG and IGM diversities. Any such patterns
426 in vaccinees were beyond our ability to measure. Second, we used concentrations of SARS-CoV-
427 2 anti-spike NAbs as our proxy of protection. Signals may be present that do not correlate with
428 antibodies binding this particular immunogen. For example, a signal might be seen in T cells or
429 antibodies that bind other SARS-CoV-2 proteins, which we are unable to evaluate given NAbs as
430 a readout. Third, although the sequence data in this study was quantitative, it contained only
431 single-chain, not paired-chain data. Fourth, the ability to define signatures is limited by
432 uncertainty about the specificity of reference sequences. Much effort is being put into methods
433 that predict receptor-antigen binding, but a unified, accepted, and feasible approach to
434 identifying all sequences that bind a given immunogen has yet to be established. Fifth, the
435 quality of binding capacity measurements is limited by the current measure of binding similarity
436 being based on mean behavior[43]; this is expected to improve with additional data and advances
437 in protein structure prediction.

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

445 ## Conflict of Interest Disclosures

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

448 ## Materials and Methods

449 **Study subjects**

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

454 **Institutional review board approval**

455 All work was carried out in accordance with BIDMC's Institutional Review Board protocols
456 2020P000634, 2021P000109 and 2020P000361.

457 **Cohort assignment**

458 All subjects from whom samples were obtained received RT-qPCR tests performed on two
459 Abbott Molecular platforms: m2000 and Alinity m (Abbott Molecular, Des Plaines, IL, U.S.A.).
460 These detect identical SARS-CoV-2 N and RdRp gene targets and are extremely sensitive for
461 SARS-CoV-2 infection, with limit of detection of 100 copies/mL.[45–47] Infectees had a positive
462 result at the time of sample acquisition. Controls were tested, but negative. COVID-19 test and
463 vaccination information were obtained using SQL queries from BIDMC's clinical data repository
464 and via a dedicated REDCap database set up to facilitate research involving vaccinees.[48]

465 Using these records, subjects were considered infectees if there was a record of a positive
466 COVID-19 test result dated before or on the sample collection date and non-infected otherwise.
467 If no medical record number was available for a subject, their infection status was considered
468 unknown. Subjects were considered vaccinees if vaccination prior to or on the day of sample

469 collection was indicated as the appropriate procedure code in the clinical data repository,
470 recorded in REDCap, or identified from Massachusetts' state Immunization Information System.
471 Subjects were considered non-vaccinated if the sample collection date preceded 12/15/2020
472 (the date of the first administered COVID-19 vaccine); if there was record of vaccination after
473 sample collection that was annotated as the first dose; if there were two vaccinations after
474 sample collection where the second was annotated as the second dose; or if there were two
475 vaccinations after sample collection within 42 days of each other (consistent with being the
476 primary series). Subjects that did not satisfy vaccinee or non-vaccinated criteria were
477 considered to have unknown vaccination status. Subjects were annotated as unexposed
478 controls if they were non-infected and non-vaccinated. Subjects whose vaccination status was
479 unknown or whose infection status was unknown and were neither vaccinees nor infectees
480 were considered to have an "unknown" SARS-CoV-2 exposure status.

481 **Clinical annotations**

482 *Immunosuppression*

483 Subjects were labelled either "immunosuppressed" or "immunocompetent." Subjects were
484 designated immunosuppressed if at least one of the following criteria was met:

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

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

497 *Demographics*

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

502 *Risk factors*

503 A semi-automated review of EHRs for ICD-10 diagnosis codes and related entries was used to
504 identify subjects having any of the medical conditions highlighted by the CDC as increasing risk
505 of severe illness from COVID-19.[49] Where feasible, the list of ICD-10 codes indicative of each
506 comorbidity was taken from the Elixhauser Comorbidity Software Refined for ICD-10-CM,[50]
507 version v2022.1, developed for the Healthcare Cost and Utilization Project (HCUP), which is
508 based on the work of Elixhauser et al.[51] In addition to these, another widely used set of
509 comorbidity measures is the Charlson Comorbidity Index.[52] For comorbidities not defined in the
510 HCUP software, the lists of ICD-10 codes defined by this study[53] were used where possible.
511 Comorbidities that were not codified in either resource were identified, where possible, using
512 ICD-10 codes or other automated chart queries, detailed as follows:

513 - Cancer: identified using ICD-10 codes in the HCUP software for "Leukemia,"
514 "Lymphoma," "Metastatic cancer," or "Solid tumor without metastasis, malignant."
515 - Chronic Kidney Disease: identified using ICD-10 codes in the HCUP software for "Renal
516 failure, moderate," and "Renal failure, severe."
517 - Chronic Liver Disease: identified using ICD-10 codes in the HCUP software for "Liver
518 disease, mild," and "Liver disease, moderate to severe."
519 - Chronic Lung Disease: The CDC website stipulates that asthma is of concern "if it's
520 moderate to severe," implying mild asthma is not of concern. The HCUP software
521 includes codes for all degrees of severity of asthma in the definition of "Chronic

522 pulmonary disease." Thus, chronic lunch disease was identified using ICD-10 codes in
523 the HCUP software for "Chronic pulmonary disease," excluding any ICD-10 codes
524 beginning with J452 or J453 (mild intermittent or mild persistent asthma, respectively).
525 - Cystic Fibrosis: Identified by any ICD-10 code beginning with E84.
526 - Dementia or other neurological condition: identified using ICD-10 codes in the HCUP
527 software for "Dementia," "Neurological disorders affecting movement," "Seizures and
528 epilepsy," and "Other neurological disorders."
529 - Diabetes: identified using ICD-10 codes in the HCUP software for "Diabetes with chronic
530 complications" and "Diabetes without chronic complications."
531 - Disabilities: identified using ICD-10 codes in the HCUP software for "Paralysis" plus any
532 ICD-10 code beginning with Q (birth defects and chromosomal abnormalities). Note that
533 this omits many, possibly most, forms of disabilities, including non-congenital blindness
534 and deafness, cognitive impairments not due to chromosomal abnormalities, autism
535 spectrum disorders of unknown etiology, etc., but these are of dubious connection to
536 COVID-19.
537 - Heart conditions: identified using ICD-10 codes in the HCUP software for "Heart failure,"
538 the ICD-10 codes listed in the referenced study[50] for "Myocardial Infarction," and/or any
539 ICD-10 code starting with any of these prefixes: A1884, A3282, A3681, A381, A395,
540 A5203, B2682, B332, B376, B5881, C452, D8685, G130, G712, G713, G720, G721, G722,
541 G7249, G7281, G7289, G729, G737, I01, I02, I05, I06, I07, I08, I09, I11, I13, I20, I23, I24,
542 I25, I3, I4, I5, I70, I9713, J1082, J1182, O101, OO2912, O903, Q2, R570, S26, T82, and
543 Z95.
544 - HIV: identified using ICD-10 codes in the HCUP software for "Acquired immune
545 deficiency syndrome."
546 - Mental health conditions: identified using ICD-10 codes in the HCUP software for
547 "Depression" and "Psychoses." Note that this may omit many other forms of mental
548 illness, such as obsessive-compulsive disorder, post-traumatic stress syndrome,
549 borderline personality disorder, etc. Note that there is overlap between conditions
550 considered mental health conditions and those considered disabilities (such as autism

551 spectrum disorders) as well as between mental health conditions and other medical
552 conditions (such as substance abuse disorders).

553 - Overweight or obese: Subjects were considered to be overweight or obese if their BMI
554 was ≥25. If multiple BMI or height-and-weight values were recorded in the database
555 over time for a given subject, the value(s) used were those closest in time to the date of
556 sample collection.

557 - Pregnancy or recent pregnancy: Electronic medical records of all female subjects under
558 the age of 69 were searched for: ICD-10 codes starting with Z3A and records of hospital
559 admissions which include a baby delivery time. The timespans of the pregnancy and
560 puerperium periods were estimated from either type of record. In the case of ICD-10
561 codes starting with Z3A, the final digits of the ICD-10 code encode weeks of gestation at
562 the time of the encounter, from which a start and end date of the pregnancy can be
563 estimated. If only a delivery date is known, the pregnancy is estimated to have begun 40
564 weeks earlier, unless "PRETERM" is found in the free-text diagnosis. Subjects were
565 marked as "pregnancy or recent pregnancy" only if their COVID-19 test date fell
566 between the estimated start date of the pregnancy and 42 days after the estimated end
567 date (to allow for post-term pregnancy). Where there was no COVID test date, the date
568 of the blood sample collection was used.

569 - Sickle cell or Thalassemia: Identified by any ICD-10 code beginning with D56 or D57.

570 - Smoking, current or former: Electronic medical records were searched for any non-zero
571 "Tobacco pack years," and for a free-text description of their tobacco usage including
572 the text "current smoker," "former," "every day," "some days," "light," "heavy," "less
573 than 10," "10+," "yes," or "counseling provided."

574 - Solid organ or blood stem cell transplant: Identified by any ICD-10 code beginning with
575 Z94.

576 - Stroke or cerebrovascular disease: identified using ICD-10 codes in the HCUP software
577 for "Cerebrovascular disease," which includes ICD-10 codes for both CBVD POA and
578 CBVD SQLA.

579 - Substance abuse: identified using ICD-10 codes in the HCUP software for "Drug abuse"
580 and for "Alcohol abuse."
581 - Tuberculosis: Identified by any ICD-10 code beginning with A15.

582 **Sample collection, cell separation, and DNA extraction**

583 2mL aliquots were taken from EDTA-anticoagulated venous blood collected in the course of
584 standard clinical care (via "purple-top" tubes; BD). Tubes were stored at 4°C between collection
585 and processing, never more than 12 hours. Each aliquot was mixed 1:1 dilution in phosphate-
586 buffered saline (PBS) and centrifuged over Ficoll-Paque-plus (Cytiva, Marlborough) to obtain
587 peripheral blood mononuclear cells (PBMCs). Plasma was collected and stored at 80°C. PBMCs
588 were washed with PBS and resuspended in a sorting buffer of PBS, 1% bovine serum albumin
589 (BSA), and 0.01% sodium azide.

590 Magnetically-labeled anti-CD4 and anti-IgM microbeads (Miltenyi, Bergisch Gladbach) were
591 used to label and column-separate for infectee and control samples; vaccinee samples cells
592 were not separated. This process divided the samples into CD4+ T cells and IgM+ B cells in one
593 fraction and CD8+ T cells and non-IgM+ B cells (principally IgG+) in another fraction. DNA was
594 isolated for each fraction using EZ1&2 DNA Blood 350µL kits (Qiagen, Hilden) and the EZ1
595 Advanced XL automated system (Qiagen, Hilden). DNA concentration was assessed via
596 Nanodrop (Thermo Fisher, Waltham).

597 **Sequencing library preparation**

598 AIRRseq libraries were generated using the immunoPETE method as described.[5] ImmunoPETE is
599 a two-step primer extension based targeted gene enrichment assay designed to specifically
600 target and quantitatively amplify recombined human TRB, TRD, and IGH from genomic DNA
601 simultaneously. Briefly, V gene-based primers containing unique molecular identifiers (UMI) as
602 well as universal PCR amplification handles were annealed to the chromosomal VDJ rearranged
603 loci. The first primer extension products, spanning the VDJ rearrangement, were purified from
604 any remaining oligos by a combination of beads (KAPA HyperPure, Roche) and enzymatic

605 treatment with Thermolabile Exonuclease I (New England Biolabs). A second primer extension
606 and amplification master mix containing a pool of J-gene oligos and an Illumina i7 primer
607 generated VDJ amplicons after 10 cycles of target amplification. Illumina sequencing library
608 amplification was performed using the i7/i5 primer pairs with dual sample indexes. All primer
609 extensions and amplifications were performed using the KAPA Long Range HotStart Ready Mix
610 (Roche). The resulting libraries underwent purification using KAPA HyperPure beads (Roche),
611 followed by quantification with the Qubit dsDNA HS Assay kit (Thermo Fisher) and fragment
612 analysis (Agilent TapeStation). Individual sample libraries were pooled in equal mass. A final
613 round of quantification and fragment analysis was then performed. Finally, libraries were
614 sequenced using the Illumina NextSeq 500/550 High Output Kit v2.5 (300 cycles).

615 **Sequencing and bioinformatics**

616 ImmunoPETE sequencing libraries were analyzed using the Roche in-house bioinformatics
617 pipeline, Daedalus (https://github.com/bioinform/Daedalus). After quality filtering of reads and
618 trimming off primers, the pipeline identified V and J genes using a Smith-Waterman alignment
619 approach (https://github.com/pgngp/swift) against an in-house curated V and J gene database.
620 Original V and J gene data and sequences were sourced from HGNC
621 (https://www.genenames.org/) and ENSEMBL (https://ensemblgenomes.org/). CDR3
622 sequences were identified for all V-J pairs, capturing both functional (functional V/J gene AND
623 coding CDR3) and non-functional (annotated non-functional or pseudogene V/J gene in the
624 database OR stop codon/frameshift in CDR3) rearrangements. Sequences are deduplicated by
625 clustering UMI and CDR3 sequences to identify UMI families. Consensus sequences were
626 derived for the CDR3 and UMI segments of each UMI family, suppressing sequencing and PCR
627 errors, and identifying CDR3 rearrangements at single molecule resolution. High quality CDR3
628 rearrangements were further analyzed for cell counting, clonal diversity, and other calculations.
629 Terms used are listed alphabetically and defined as follows:

630 - Cell count: the total number of functional IGH, TRD, and TRB rearrangements in a
631 sample

632 - Cell type percentages: the total number of functional rearrangements from each heavy
633 chain divided by the total cell count × 100
634 - CDR3 clone: BCR or TCR sequences from the same individual with matching V gene,
635 CDR3 amino acid sequence (CDR3-AA), and J gene assignment arising from two or more
636 UMI families
637 - CDR3 clonal type: BCR or TCR sequences from multiple UMI families from multiple
638 individuals with matching V gene, CDR3-AA, and J gene assignment
639 - Clone count: total number of UMI families from the same individual with the same V
640 gene, CDR3-AA, and J gene
641 - UMI family: a set of reads that have been clustered together based on the similarities of
642 the 9-nt UMI sequence and the CDR3-nt region

643 Both UMI and CDR3 sequences are clustered based on a Levenshtein edit distance of 1,
644 capturing likely PCR and sequencing errors. A UMI family represents a single captured DNA
645 molecular fragment from the immunoPETE reaction.

646 **NAbs ELISA titers**

647 The SARS-CoV-2 Surrogate Virus Neutralization Test Kit (GenScript, L00847-A) was used
648 according to the manufacturer's instructions as follows. A standard curve was generated using a
649 serial dilution of the standard (GenScript, A02087-20) with a dilution factor of 1:2. Each
650 subject's serum sample was mixed with sample dilution buffer (1:10) and horseradish
651 peroxidase-conjugated recombinant SARS-CoV-2 receptor-binding domain (HRP-RBD). The
652 mixture was incubated at 37°C for 30 minutes to allow the circulating NAbs to bind to HRP-RBD.
653 The mixture was then added to an ACE2 protein-coated plate and incubated for an additional
654 15 minutes at 37°C. Unbound HRP-RBD and HRP-RBD bound to non-neutralizing antibodies
655 were bound to the plate while circulating neutralization antibody HRP-RBD complexes
656 remained in the supernatant for subsequent wash steps. After washing, tetramethylbenzidine
657 solution was added, followed by a stop solution to quench the reaction, turning wells yellow.
658 The plate was read immediately at 450nm in a microtiter plate reader. Statistical analysis was

659 performed with GraphPad Prism using a 4PL model for linear regression. Results were reported
660 by interpolating the OD450 values to the standard curve values.

661 **pymmunomics**

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

665 **Dependence of antibody concentrations on age, immunocompetence, and SARS-CoV-2**
666 **exposure**

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

677 **CDR3 length analysis**

678 CDR3 length frequencies for each available functional and non-functional pooled IGH, TRB, TRD,
679 and subtyped IGG, IGM, CD4 TRB/D, CD8 TCB/D repertoire of immunocompetent subjects were
680 calculated using the pymmunomics python package (above). Since vaccinee samples were not
681 sorted into subtypes, pooled repertoire CDR3 length frequency distributions were used to
682 compare vaccinees to controls and infectees. CD4/IGM and CD8/IGG repertoire CDR3 length
683 frequency distributions were compared independently between controls to infectees.

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

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

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

To identify trends among lengths of V and J genes, V and J genes (from IMGT) of the relevant cell types (IGH for the functional pooled IGH comparisons and pooled IGH, IGG, and IGM for the non-functional comparisons) which had a corrected p-value below 0.05 were grouped into the number of residues that fall into the CDR3 region. Usage frequencies of V- and J-gene groups were compared between cohorts using two-tailed Mann-Whitney-U and a second correction round was conducted to correct all original p-values of the CDR3 length comparisons at the same time as the p-values obtained from the follow-up tests.

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

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

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

723 To conform with the gene database used for V- and J-gene assignment of repertoire sequences
724 (see Sequencing and bioinformatics), the same gene sequences were aligned (blastp and blastp-
725 short for V genes and J genes, respectively; BLAST+ v2.12.0) to the sequences from PDB and
726 CoVAbDab, setting the max target seqs parameter to 10,000—a number much larger than the
727 total number of genes in the query to avoid missing the best matching genes.[54] V-gene matches
728 with query coverage less than 30% or percent identity less than 40% and J-gene matches with
729 query coverage less than 50% or percent identity less than 40% were filtered out. From the
730 remainder, the best V- and J-gene matches according to percent identity and gene sequence
731 coverage (lexicographically) were assigned to each query sequence. Data downloaded from
732 VDJDB contained sequence only for the CDR3 region, so the V, and J-gene annotation provided
733 by the database was used (as opposed to using e.g. BLAST).

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

737 **Binding-capacity measurements**

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

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

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

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

746 Here, $Lev(c, c')$ is the Levenshtein distance between the CDR3 amino acid sequences of
747 sequences c and c'. The pymmunolib Python package was used to calculate similarity matrices
748 and binding capacities.

749 **Fuzzy query sequence matching**

750 Fuzzy sequence matching measurements for each pooled CD4+CD8 and each CD4-only TRB
751 subject repertoire to the MIRA-identified HLA class II query sequences were tabulated from the
752 similarity matrices that are calculated as part of determining binding capacities. For each query

753 sequence and each subject repertoire, we measured the fraction of repertoire sequences for
754 with the same V and J genes as the query sequence, and whose CDR3 sequence was within
755 Levenshtein distances 0-10 of the query's CDR3. Note that exact matching is equivalent to fuzzy
756 matching with a Levenshtein distance of 0.

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

758 To compare robustness to variations in repertoire size of binding capacity and fuzzy matching
759 features, we conducted subsampling experiments. We randomly chose 10 subjects from each
760 of the vaccinee, infectee, and control cohorts that had a pooled TRB repertoire size of at least
761 80,000 cells, i.e. 80,000 distinct corrected UMIs. (This size was chosen in order to guarantee at
762 least 10 subjects from the control cohort to choose from.) Each of these repertoires was
763 sampled down to 20 different subsample sizes chosen to be equidistantly spaced between 10
764 and 80,000 at log-scale. For each subsample, we calculated binding capacities as well as fraction
765 of fuzzy matches for fuzzy-match tolerances 0, 2, 4, 6, 8, and 10 amino acids to CD4 TRB
766 reference sequences from MIRA. The slopes and their surrounding 95% confidence intervals
767 were obtained by fitting a linear mixed model that groups the data by subject.

768 **Feature selection**

769 Preferring the use of domain knowledge over generic feature selection mechanisms for
770 selecting from the high-dimensional query sequence matching features (binding capacity and
771 fuzzy matching), a custom feature selection method is developed and implemented in the
772 python package pymmunomics. For this mechanism we use binding capacity and fuzzy
773 matching measurements to sequence specific to pathogens other than SARS-CoV-2 ("SARS-CoV-
774 2 non-specific sequences") as a null distribution to which to compare the measurements for
775 MIRA-identified SARS-CoV-2-specific sequences. We calculated the (Stuart-)Kendall Tau-c
776 correlation coefficient between each feature's measurement and NAb titer. For each feature
777 group (binding capacity, fuzzy matching with tolerances 0, 1, …, 10, etc.), the correlation
778 coefficients of measurements for non-SARS-CoV-2 specific sequences form the null distribution

779 and correlation coefficients of SARS-CoV-2 specific features below the 2.5th and above the
780 97.5th percentile are selected (cumulatively, the most correlated and anti-correlated 5%).

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

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

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

798 - CDR3 length frequencies, summarized by 3 features: mean, variance and skewness;
799 - diversity, with Recon[55] (https://github.com/ArnaoutLab/Recon) being used to correct
800 Hill D_q numbers for q=0, 1, …, ∞ to correct for missing species;
801 - J-gene frequencies (with only 6 J genes, no further feature selection was required);
802 - V-gene frequencies for select V genes as described above;
803 - Baseline/clinical features: age, sex, days since infection (runs of positive COVID-19 PCR
804 tests successively within 28 days of each other and not interrupted by negative tests are

considered infected periods; to account for incubation of the virus prior to taking the test, the start date of an infection is predicted as 4 days before the first positive test in the corresponding run of tests; when a negative test was performed within those 4 days, that test's date is considered the infection start date; for the model, the predicted start date of the most recent infection before sample collection was used, or 0 if the subject was not infected), and days since vaccination (the number of days between sample collection and most recent vaccination on record).

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

828 ## Figure Legends

829 **Figure 1: Anti-SARS-CoV-2 ELISA trends and distributions by age for immunocompetent and**
830 **immunosuppressed vaccinees, infectees, and controls.** NAbs are to SARS-CoV-2 spike protein.
831 **(a)** ELISA titers for each subject. Solid lines indicate regression fits; shaded areas indicate 95%
832 confidence intervals. Dotted black line at ~10^3 indicates manufacturer's cutoff for positive vs.
833 negative. Note strong negative trend with age in vaccinees (blue) but not infectees (salmon).
834 Note mild positive trend with age in controls (olive), even as titers in this cohort remain below
835 the cutoff for almost all individuals. **(b)** Distribution of titers in the three cohorts, split by
836 immune status. **(c)** Distribution of ages in these cohorts, again split by immune status, with
837 numbers of subjects in each sub-cohort.

838 **Figure 2: IGH CDR3 length distributions. (a)** CDR3 length comparison plots for productive IGH
839 repertoires of vaccinees vs. controls. Left inset: the median differences of frequencies at each
840 length, showing that CDR3s of length 16 or shorter are more frequent in vaccinees, whereas
841 CDR3s of length 17 or longer are less frequent. The pattern reverses at the dividing line
842 between 16 and 17 amino acids (vertical dotted line). Right inset: total fraction of the
843 repertoire up to the dividing line. The p-value is obtained by applying Mann-Whitney U to the
844 cumulatives followed by Holm-Bonferroni multiple-hypothesis correction. **(b)** The same for
845 vaccinees vs. infectees, showing the same pattern but with a dividing line between 18 and 19
846 amino acids. **(c)**-**(d)** Frequencies of V and J genes grouped by the number of residues each gene
847 contributes to the CDR3 according to germline. Note the only J gene that contributes 5 residues
848 is IGHJ4. **(e)** Assuming the nonproductive IGH vaccinee repertoire (blue) is made up of a part
849 that is unaffected by vaccination and therefore looks like the control repertoire (green) and a
850 part that is affected by vaccination (salmon), this plot shows what the distribution of the
851 affected part would have to look like so the two parts add up correctly, for different fractions
852 affected (dark to light salmon lines). Estimated means for vaccinee and control distributions are
853 shown. The smaller the affected portion, the more extreme the effect must be. The minimum
854 possible effect size is that for which a CDR3 length for the affected portion is zero; any smaller,

855 and a negative frequency at that CDR3 length would be required (negative frequencies are not
856 possible).

857 **Figure 3: SARS-CoV-2-specific TRBs vs. NAb titers. (a)-(b)** Fraction of TRB repertoires matching
858 the SARS-CoV-2-specific CD4 TRB sequences obtained from Nolan et al.[23] against SARS-CoV-2
859 NAb titer. Panel (a) shows repertoires from CD4+ T cells, which were available for infectees and
860 controls but not vaccinees, while panel (b) shows repertoires from all T cells, which were
861 available for all three cohorts. Theil-Sen regression fits (solid lines) show positive relationships
862 for infectees and vaccinees but not controls. **(c)** The fraction of a repertoire that matches
863 reference TRBs within a chosen tolerance (here, 2 amino-acid differences) depends strongly on
864 the number of cells in the repertoire (i.e., repertoire size). **(d)** In contrast, binding capacity is
865 much more robust. The slope of the dependency on size for repertoires above 1,000 cells are
866 shown as black lines. **(e)** Slope as a function of fuzzy-binding tolerance, demonstrating binding
867 capacity is more robust regardless of tolerance.

868 **Figure 4: Predicting positive NAbs. (a)** Feature selection mechanism used for binding capacity
869 and fuzzy matching features on the binding capacity measurements of all TCR repertoires of
870 size at least 1,000 using the SARS-CoV-2-specific CD4 TCR sequences and non-SARS-CoV-2-
871 specific TCR sequences obtained from VDJDB. Of the 7,804 SARS-CoV-2-specific features'
872 correlations, 323 fall outside the selection boundaries set by the 95% boundaries of the
873 correlations of non-SARS-CoV-2-specific features with NAb titer. **(b)-(c)** Machine learning
874 performance results for a selected group of feature sets and cell types across all 700 replicates
875 (100 repeats of 7-fold cross-validation). a shows areas under receiver operating curves and b
876 breaks down the performances into sensitivity, precision, and specificity. The same plots for all
877 feature sets and cell types can be found in Fig. S11 and S12. Median values and interquartile
878 ranges for all metrics are reported in Table S8.

879 **References**

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

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

886 3. Shen, J., Fan, J., Zhao, Y., Jiang, D., Niu, Z., Zhang, Z., and Cao, G. (2023). Innate and adaptive
887 immunity to SARS-CoV-2 and predisposing factors. Front Immunol *14*, 1159326.
888 10.3389/fimmu.2023.1159326.

889 4. Joseph, M., Wu, Y., Dannebaum, R., Rubelt, F., Zlatareva, I., Lorenc, A., Du, Z.G., Davies, D.,
890 Kyle-Cezar, F., Das, A., et al. (2022). Global patterns of antigen receptor repertoire
891 disruption across adaptive immune compartments in COVID-19. Proc Natl Acad Sci U S A
892 *119*, e2201541119. 10.1073/pnas.2201541119.

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

897 6. Fujihashi, K., McGhee, J., Yamamoto, M., Hiroi, T., and Kiyono, H. (1996). Role of gamma
898 delta T cells in the regulation of mucosal IgA response and oral tolerance. Ann N Y Acad Sci
899 *778*, 55–63. 10.1111/j.1749-6632.1996.tb21114.x.

900 7. Rezende, R.M., Cox, L.M., Moreira, T.G., Liu, S., Boulenouar, S., Dhang, F., LeServe, D.S.,
901 Nakagaki, B.N., Lopes, J.R., Tatematsu, B.K., et al. (2023). Gamma-delta T cells modulate the
902 microbiota and fecal micro-RNAs to maintain mucosal tolerance. Microbiome *11*, 32.
903 10.1186/s40168-023-01478-1.

904 8. Brouwer, P.J.M., Caniels, T.G., van der Straten, K., Snitselaar, J.L., Aldon, Y., Bangaru, S.,
905 Torres, J.L., Okba, N.M.A., Claireaux, M., Kerster, G., et al. (2020). Potent neutralizing
906 antibodies from COVID-19 patients define multiple targets of vulnerability. Science *369*,
907 643–650. 10.1126/science.abc5902.

908 9. Cao, Y., Su, B., Guo, X., Sun, W., Deng, Y., Bao, L., Zhu, Q., Zhang, X., Zheng, Y., Geng, C., et
909 al. (2020). Potent Neutralizing Antibodies against SARS-CoV-2 Identified by High-
910 Throughput Single-Cell Sequencing of Convalescent Patients' B Cells. Cell *182*, 73-84.e16.
911 10.1016/j.cell.2020.05.025.

912 10. Chen, E.C., Gilchuk, P., Zost, S.J., Suryadevara, N., Winkler, E.S., Cabel, C.R., Binshtein, E.,
913 Chen, R.E., Sutton, R.E., Rodriguez, J., et al. (2021). Convergent antibody responses to the
914 SARS-CoV-2 spike protein in convalescent and vaccinated individuals. Cell Reports *36*,
915 109604. 10.1016/j.celrep.2021.109604.

916 11. Cerutti, G., Guo, Y., Zhou, T., Gorman, J., Lee, M., Rapp, M., Reddem, E.R., Yu, J., Bahna, F.,
917 Bimela, J., et al. (2021). Potent SARS-CoV-2 neutralizing antibodies directed against spike N-
918 terminal domain target a single supersite. Cell Host & Microbe *29*, 819-833.e7.
919 10.1016/j.chom.2021.03.005.

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

924 13. Lima, N.S., Mukhamedova, M., Johnston, T.S., Wagner, D.A., Henry, A.R., Wang, L., Yang,
925 E.S., Zhang, Y., Birungi, K., Black, W.P., et al. (2022). Convergent epitope specificities, V gene
926 usage and public clones elicited by primary exposure to SARS-CoV-2 variants. bioRxiv,
927 2022.03.28.486152. 10.1101/2022.03.28.486152.

928 14. McCallum, M., De Marco, A., Lempp, F.A., Tortorici, M.A., Pinto, D., Walls, A.C., Beltramello,
929 M., Chen, A., Liu, Z., Zatta, F., et al. (2021). N-terminal domain antigenic mapping reveals a
930 site of vulnerability for SARS-CoV-2. Cell *184*, 2332-2347.e16. 10.1016/j.cell.2021.03.028.

931 15. Rapp, M., Guo, Y., Reddem, E.R., Yu, J., Liu, L., Wang, P., Cerutti, G., Katsamba, P., Bimela,
932 J.S., Bahna, F.A., et al. (2021). Modular basis for potent SARS-CoV-2 neutralization by a
933 prevalent VH1-2-derived antibody class. Cell Reports *35*, 108950.
934 10.1016/j.celrep.2021.108950.

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

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

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

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

948 20. Matschke, J., Lütgehetmann, M., Hagel, C., Sperhake, J.P., Schröder, A.S., Edler, C.,
949 Mushumba, H., Fitzek, A., Allweiss, L., Dandri, M., et al. (2020). Neuropathology of patients
950 with COVID-19 in Germany: a post-mortem case series. The Lancet Neurology *19*, 919–929.
951 10.1016/S1474-4422(20)30308-2.

952 21. Schwabenland, M., Salié, H., Tanevski, J., Killmer, S., Lago, M.S., Schlaak, A.E., Mayer, L.,
953 Matschke, J., Püschel, K., Fitzek, A., et al. (2021). Deep spatial profiling of human COVID-19
954 brains reveals neuroinflammation with distinct microanatomical microglia-T-cell
955 interactions. Immunity *54*, 1594-1610.e11. 10.1016/j.immuni.2021.06.002.

956 22. Arora, R., and Arnaout, R. (2020). Private Antibody Repertoires Are Public.
957 10.1101/2020.06.18.159699.

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

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

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

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

972 27. Dyer, A.H., Noonan, C., McElheron, M., Batten, I., Reddy, C., Connolly, E., Pierpoint, R.,
973 Murray, C., Leonard, A., Higgins, C., et al. (2022). Previous SARS-CoV-2 Infection, Age, and
974 Frailty Are Associated With 6-Month Vaccine-Induced Anti-Spike Antibody Titer in Nursing
975 Home Residents. Journal of the American Medical Directors Association *23*, 434–439.
976 10.1016/j.jamda.2021.12.001.

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

979 29. Haveri, A., Ekström, N., Solastie, A., Virta, C., Österlund, P., Isosaari, E., Nohynek, H., Palmu,
980 A.A., and Melin, M. (2021). Persistence of neutralizing antibodies a year after SARS-CoV-2
981 infection in humans. European Journal of Immunology *51*, 3202–3213.
982 10.1002/eji.202149535.

983 30. Korobova, Z.R., Zueva, E.V., Arsentieva, N.A., Batsunov, O.K., Liubimova, N.E., Khamitova,
984 I.V., Kuznetsova, R.N., Rubinstein, A.A., Savin, T.V., Stanevich, O.V., et al. (2022). Changes in
985 Anti-SARS-CoV-2 IgG Subclasses over Time and in Association with Disease Severity. Viruses
986 *14*, 941. 10.3390/v14050941.

987 31. Luo, Y.R., Chakraborty, I., Yun, C., Wu, A.H.B., and Lynch, K.L. (2021). Kinetics of Severe
988 Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Antibody Avidity Maturation and
989 Association with Disease Severity. Clinical Infectious Diseases *73*, e3095–e3097.
990 10.1093/cid/ciaa1389.

991 32. Moss, P. (2022). The T cell immune response against SARS-CoV-2. Nat Immunol *23*, 186–
992 193. 10.1038/s41590-021-01122-w.

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

996 34. Brown, A.J., Snapkov, I., Akbar, R., Pavlović, M., Miho, E., Sandve, G.K., and Greiff, V. (2019).
997 Augmenting adaptive immunity: progress and challenges in the quantitative engineering
998 and analysis of adaptive immune receptor repertoires. Mol. Syst. Des. Eng. *4*, 701–736.
999 10.1039/C9ME00071B.

1000 35. Brown, S.D., Raeburn, L.A., and Holt, R.A. (2015). Profiling tissue-resident T cell repertoires
1001 by RNA sequencing. Genome Med *7*, 125. 10.1186/s13073-015-0248-x.

1002 36. Huang, H., Wang, C., Rubelt, F., Scriba, T.J., and Davis, M.M. (2020). Analyzing the
1003 Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2
1004 and genome-wide antigen screening. Nat Biotechnol *38*, 1194–1202. 10.1038/s41587-020-
1005 0505-4.

1006 37. Mayer, A. and Curtis G. Callan (2023). Measures of epitope binding degeneracy from T cell
1007 receptor repertoires. Proceedings of the National Academy of Sciences *120*, e2213264120.
1008 10.1073/pnas.2213264120.

1009 38. Kaplinsky, J., Li, A., Sun, A., Coffre, M., Koralov, S.B., and Arnaout, R. (2014). Antibody
1010 repertoire deep sequencing reveals antigen-independent selection in maturing B cells. Proc
1011 Natl Acad Sci U S A *111*, E2622-9. 10.1073/pnas.1403278111.

1012 39. Murphy, K., and Weaver, C. (2016). Janeway's Immunobiology (Garland Science).

1013 40. Raybould, M.I.J., Kovaltsuk, A., Marks, C., and Deane, C.M. (2020). CoV-AbDab: the
1014 Coronavirus Antibody Database. bioRxiv, 2020.05.15.077313. 10.1101/2020.05.15.077313.

1015 41. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N.,
1016 and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

1017 42. Goncharov, M., Bagaev, D., Shcherbinin, D., Zvyagin, I., Bolotin, D., Thomas, P.G., Minervina,
1018 A.A., Pogorelyy, M.V., Ladell, K., McLaren, J.E., et al. (2022). VDJdb in the pandemic era: a
1019 compendium of T cell receptors specific for SARS-CoV-2. Nature Methods *19*, 1017–1019.
1020 10.1038/s41592-022-01578-0.

1021 43. Arora, R., and Arnaout, R. (2022). Repertoire-scale measures of antigen binding. Proc. Natl.
1022 Acad. Sci. U.S.A. *119*, e2203505119. 10.1073/pnas.2203505119.

1023 44. Arnaout, R.A., and Nowak, M.A. (2000). Competitive coexistence in antiviral immunity. J
1024 Theor Biol *204*, 431–441. 10.1006/jtbi.2000.2027.

1025 45. Arnaout, R., Lee, R.A., Lee, G.R., Callahan, C., Cheng, A., Yen, C.F., Smith, K.P., Arora, R., and
1026 Kirby, J.E. (2021). The Limit of Detection Matters: The Case for Benchmarking Severe Acute
1027 Respiratory Syndrome Coronavirus 2 Testing. Clinical Infectious Diseases *73*, e3042–e3046.
1028 10.1093/cid/ciaa1382.

1029 46. Callahan, C., Lee, R.A., Lee, G.R., Zulauf, K., Kirby, J.E., and Arnaout, R. (In press). Nasal Swab
1030 Performance by Collection Timing, Procedure, and Method of Transport for Patients with
1031 SARS-CoV-2. J Clin Microbiol.

1032 47. Callahan, C., Ditelberg, S., Dutta, S., Littlehale, N., Cheng, A., Kupczewski, K., McVay, D.,
1033 Riedel, S., Kirby, J.E., and Arnaout, R. (2021). Saliva is Comparable to Nasopharyngeal Swabs
1034 for Molecular Detection of SARS-CoV-2. Microbiol Spectr *9*, e0016221.
1035 10.1128/Spectrum.00162-21.

1036 48. Liu, J., Chandrashekar, A., Sellers, D., Barrett, J., Jacob-Dolan, C., Lifton, M., McMahan, K.,
1037 Sciacca, M., VanWyk, H., Wu, C., et al. (2022). Vaccines elicit highly conserved cellular
1038 immunity to SARS-CoV-2 Omicron. Nature *603*, 493–496. 10.1038/s41586-022-04465-y.

1039 49. People with Certain Medical Conditions (2023). Centers for Disease Control and Prevention.
1040 https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-
1041 medical-conditions.html.

1042 50. ELIXHAUSER COMORBIDITY SOFTWARE REFINED FOR ICD-10-CM Agency for Healthcare
1043 Research and Quality. https://hcup-
1044 us.ahrq.gov/toolssoftware/comorbidityicd10/comorbidity_icd10.jsp.

1045 51. Elixhauser, A., Steiner, C., Harris, D.R., and Coffey, R.M. (1998). Comorbidity measures for
1046 use with administrative data. Med Care *36*, 8–27. 10.1097/00005650-199801000-00004.

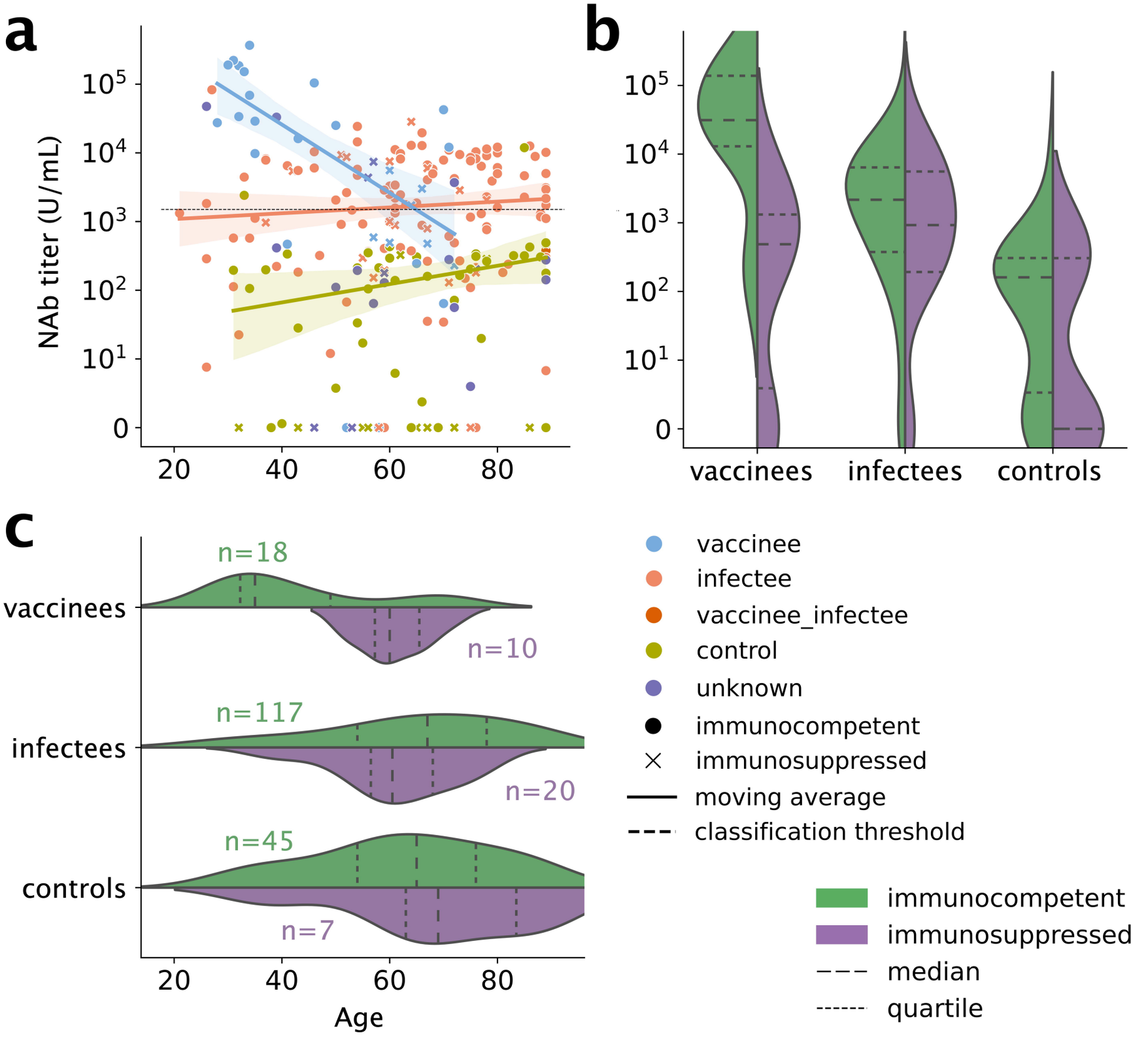
1047 52. Charlson, M.E., Pompei, P., Ales, K.L., and MacKenzie, C.R. (1987). A new method of
1048 classifying prognostic comorbidity in longitudinal studies: development and validation. J
1049 Chronic Dis *40*, 373–383. 10.1016/0021-9681(87)90171-8.

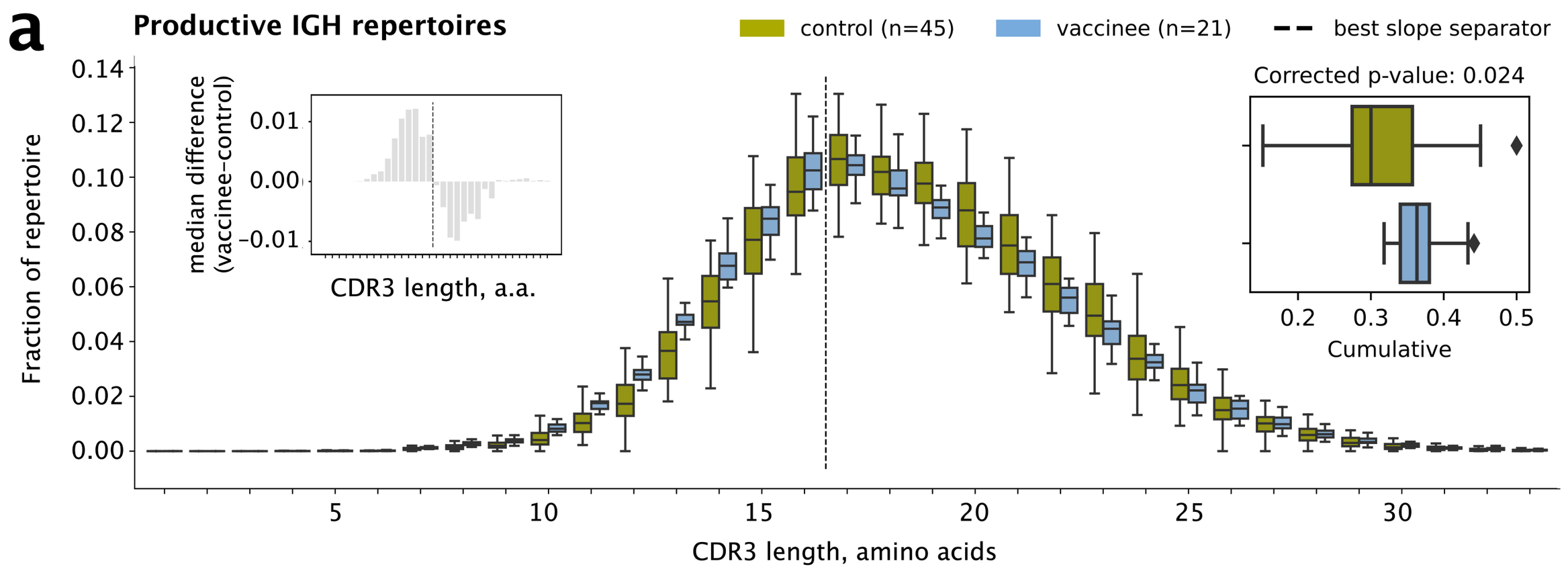
1050 53. Garry, E.M., Weckstein, A.R., Quinto, K., Bradley, M.C., Lasky, T., Chakravarty, A., Leonard,
1051 S., Vititoe, S.E., Easthausen, I.J., Rassen, J.A., et al. (2022). Categorization of COVID-19
1052 severity to determine mortality risk. Pharmacoepidemiol Drug Saf *31*, 721–728.
1053 10.1002/pds.5436.

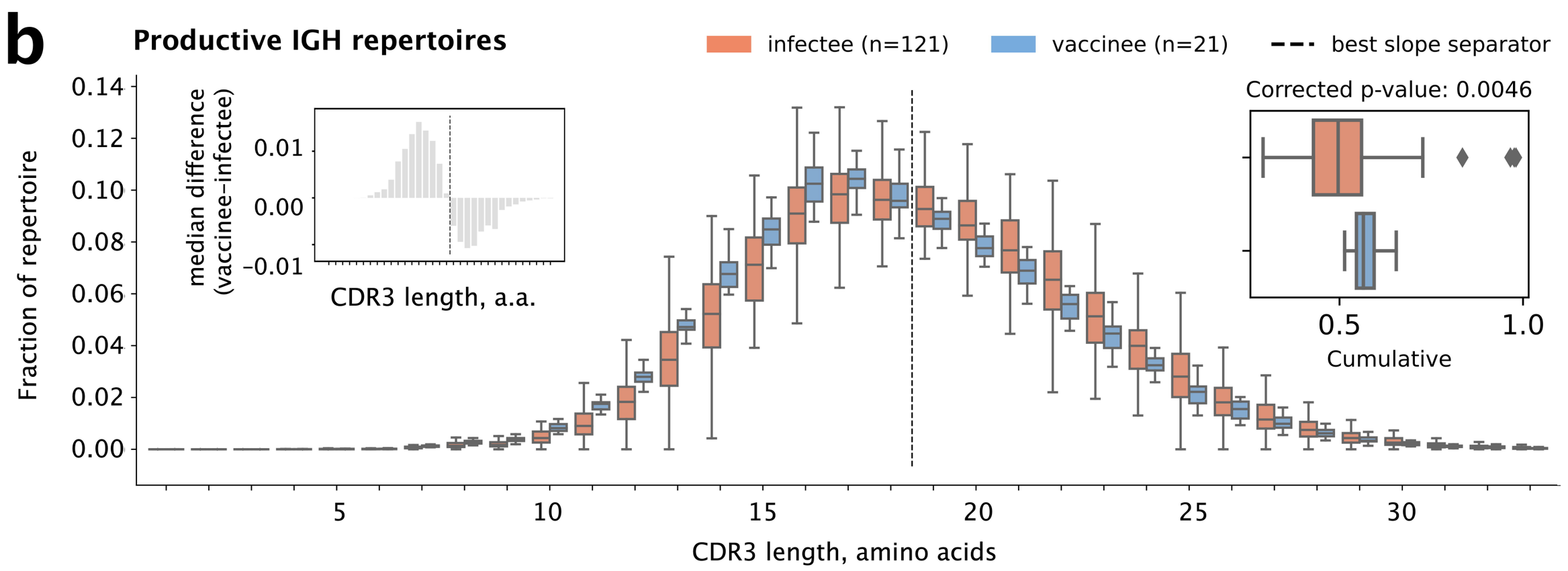
1054 54. Shah, N., Nute, M.G., Warnow, T., and Pop, M. (2019). Misunderstood parameter of NCBI
1055 BLAST impacts the correctness of bioinformatics workflows. Bioinformatics *35*, 1613–1614.
1056 10.1093/bioinformatics/bty833.

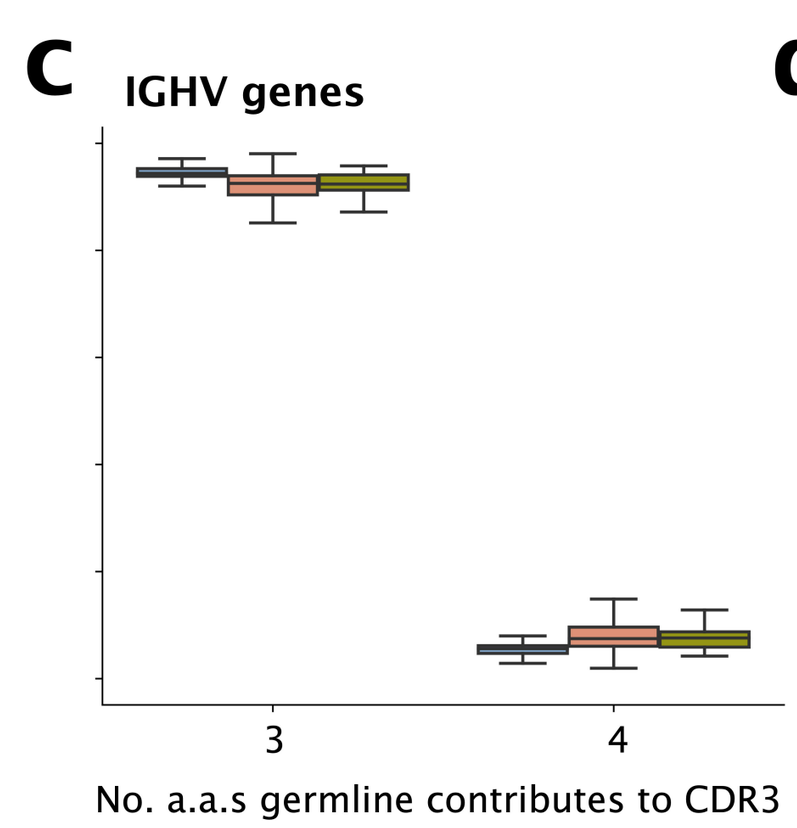
1057 55. Kaplinsky, J., and Arnaout, R. (2016). Robust estimates of overall immune-repertoire
1058 diversity from high-throughput measurements on samples. Nat Commun *7*, 11881.
1059 10.1038/ncomms11881.

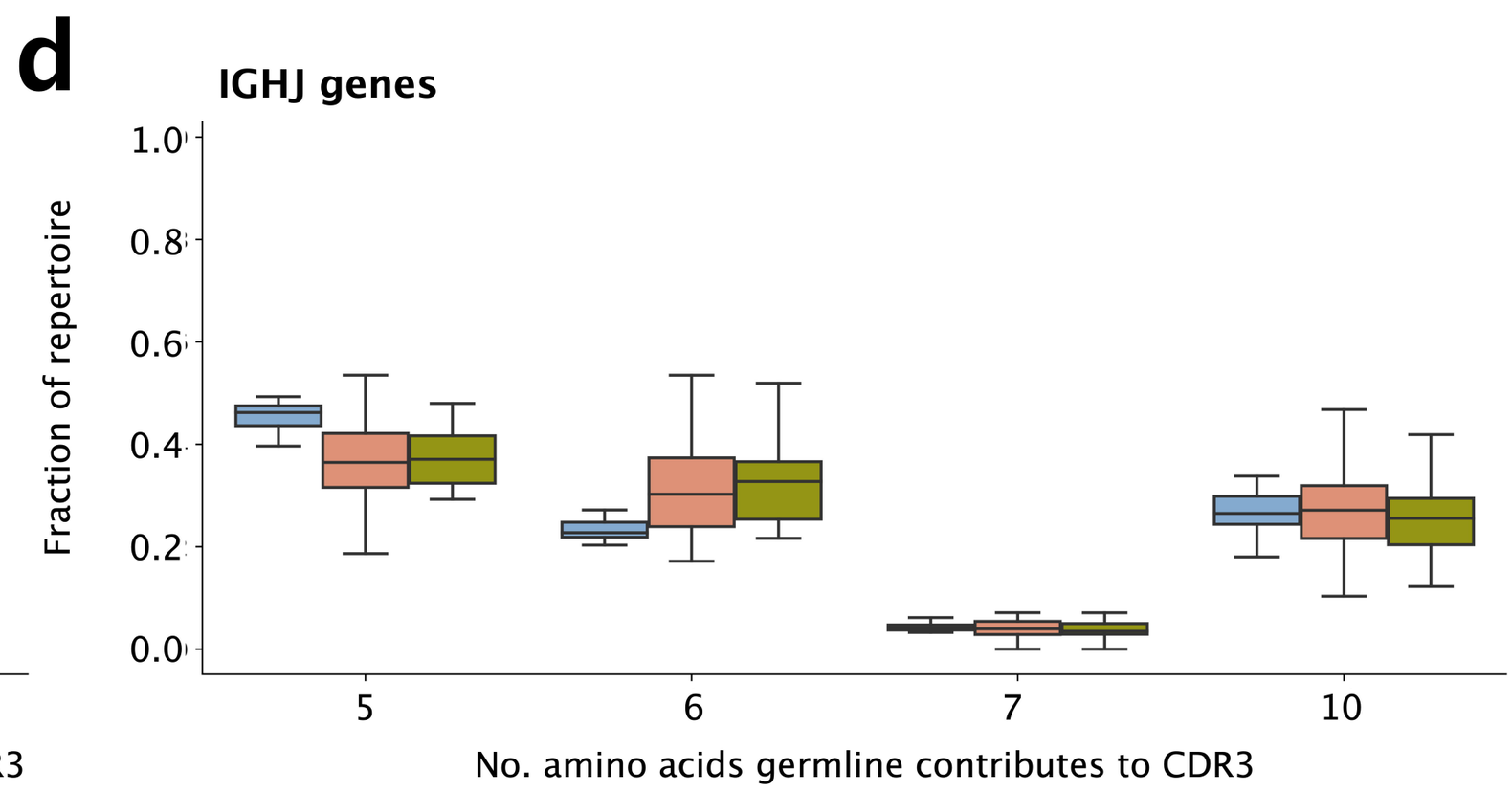
1060

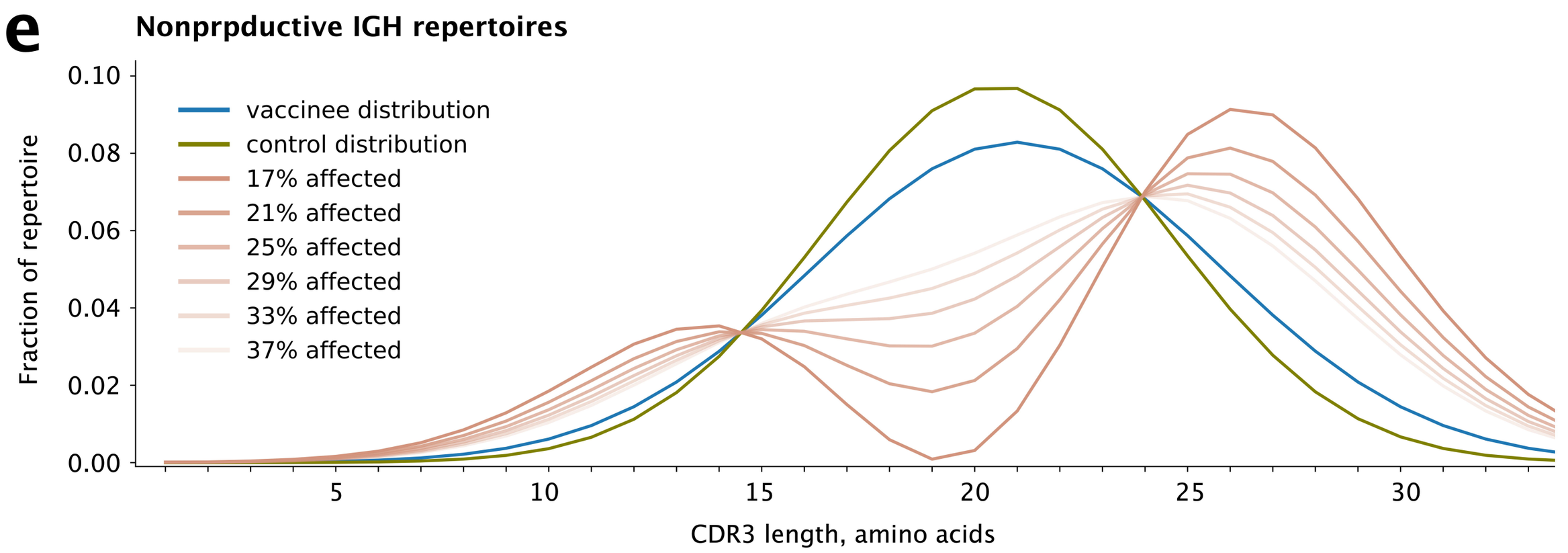

a
NAb titer (U/mL)
10^5
10^4
10^3
10^2
10^1
0
20
40
60
80
b
vaccinees
infectees
controls
c
vaccinees
infectees
controls
n=18
n=10
n=117
n=20
n=45
n=7
Age
vaccinee
infectee
vaccinee_infectee
control
unknown
immunocompetent
immunosuppressed
moving average
classification threshold
immunocompetent
immunosuppressed
median
quartile

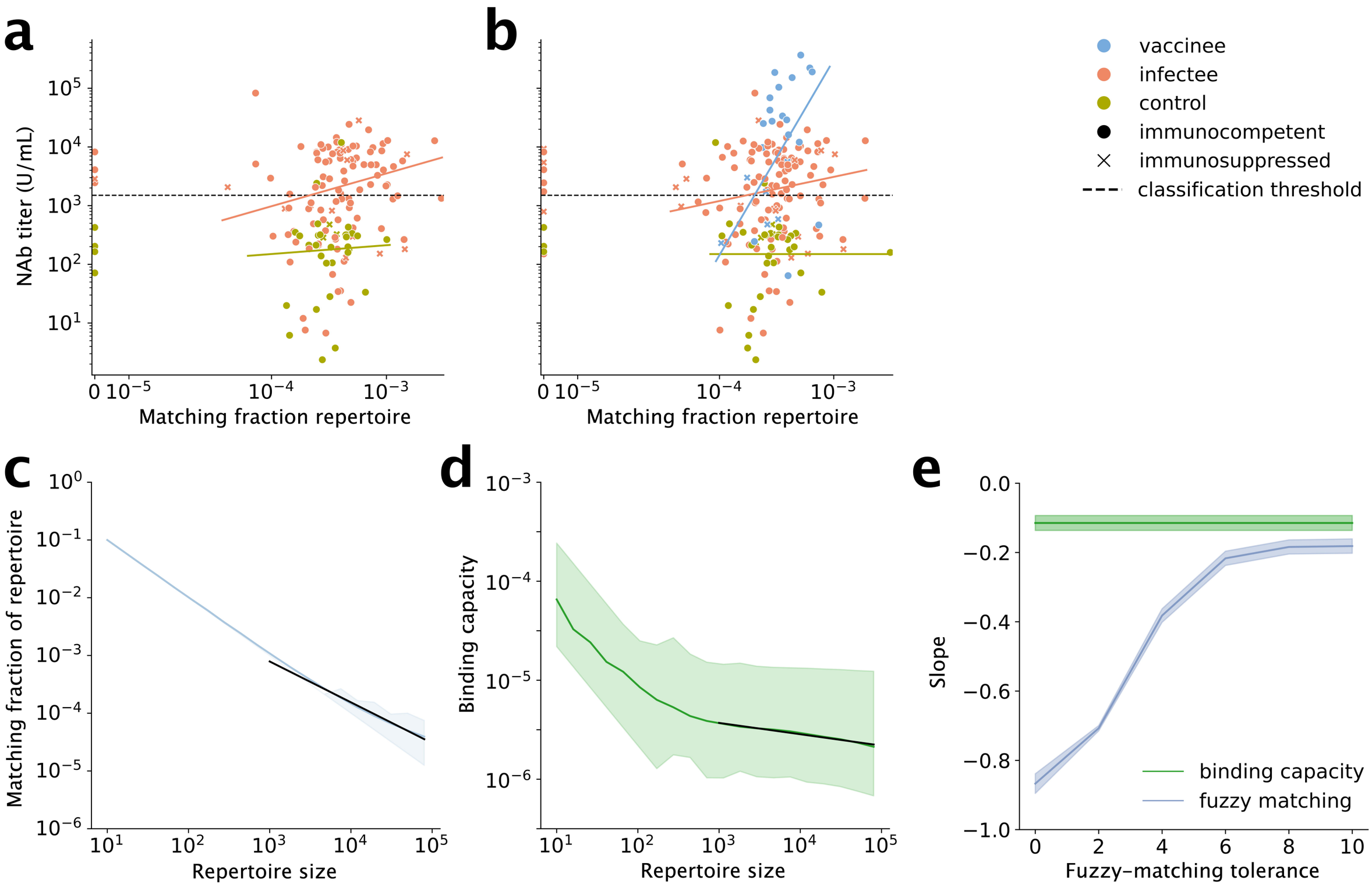

a
Productive IGH repertoires
control (n=45)
vaccinee (n=21)
best slope separator
Fraction of repertoire
median difference (vaccinee−control)
CDR3 length, a.a.
Corrected p-value: 0.024
Cumulative
CDR3 length, amino acids

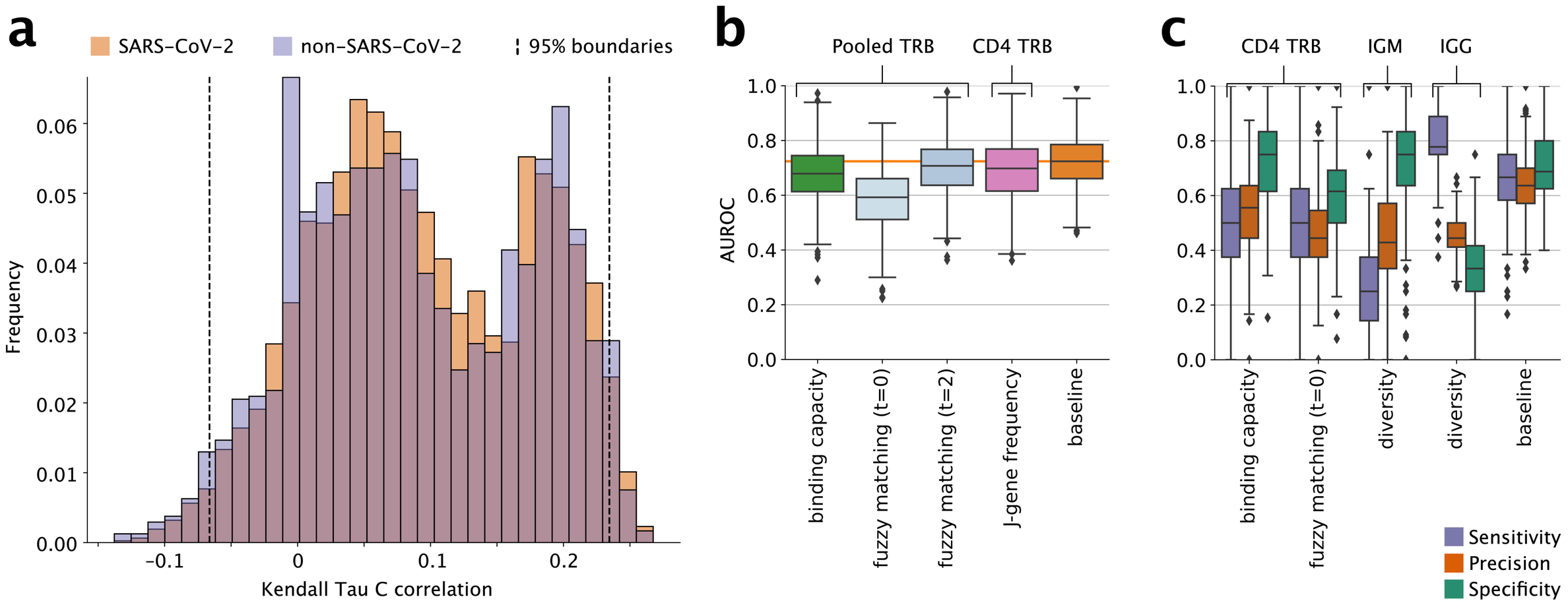

b
Productive IGH repertoires
infectee (n=121)
vaccinee (n=21)
best slope separator
Fraction of repertoire
median difference (vaccinee−infectee)
CDR3 length, a.a.
Corrected p-value: 0.0046
Cumulative
CDR3 length, amino acids


c
IGHV genes
No. a.a.s germline contributes to CDR3


d
IGHJ genes
Fraction of repertoire
No. amino acids germline contributes to CDR3


e
Nonprpductive IGH repertoires
vaccinee distribution
control distribution
17% affected
21% affected
25% affected
29% affected
33% affected
37% affected
Fraction of repertoire
CDR3 length, amino acids


a
b
c
d
e
NAb titer (U/mL)
Matching fraction repertoire
Matching fraction of repertoire
Repertoire size
Binding capacity
Slope
Fuzzy-matching tolerance
vaccinee
infectee
control
immunocompetent
immunosuppressed
classification threshold
binding capacity
fuzzy matching


a
SARS-CoV-2
non-SARS-CoV-2
95% boundaries
Frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
-0.1
0
0.1
0.2
Kendall Tau C correlation
b
Pooled TRB
CD4 TRB
AUROC
1.0
0.8
0.6
0.4
0.2
0.0
binding capacity
fuzzy matching (t=0)
fuzzy matching (t=2)
J-gene frequency
baseline
c
CD4 TRB
IGM
IGG
1.0
0.8
0.6
0.4
0.2
0.0
binding capacity
fuzzy matching (t=0)
diversity
diversity
baseline
Sensitivity
Precision
Specificity