

1 **Antigenic variation of SARS-CoV-2 in response to immune pressure**

2

3

4 Diego Forni^{1,*}, Rachele Cagliani¹, Chiara Pontremoli¹, Alessandra Mozzi¹, Uberto Pozzoli¹, Mario

5 Clerici^{2,3}, Manuela Sironi¹

6

7

8 ¹ Scientific Institute IRCCS E. MEDEA, Bioinformatics, Bosisio Parini, Italy;

9 ² Department of Physiopathology and Transplantation, University of Milan, Milan, Italy;

10 ³ Don C. Gnocchi Foundation ONLUS, IRCCS, Milan, Italy.

11 * Correspondence: diego.forni@lanostrafamiglia.it

12

13

14

15

16 **Summary**

17

18 The ongoing evolution of SARS-CoV-2 is expected to be at least partially driven by the selective
19 pressure imposed by the human immune system. We exploited the availability of a large number of
20 high-quality SARS-CoV-2 genomes, as well as of validated epitope predictions, to show that B cell
21 epitopes in the spike glycoprotein (S) and in the nucleocapsid protein (N) have higher diversity than
22 non-epitope positions. Similar results were obtained for other human coronaviruses. Conversely, in the
23 SARS-CoV-2 population, epitopes for CD4⁺ and CD8⁺ T cells were not more variable than non-epitope
24 positions. A significant reduction in epitope variability was instead observed for some of the most
25 immunogenic proteins (S, N, ORF8, and ORF3a). Analysis over longer evolutionary time-frames
26 indicated that this effect is not due to differential constraints. These data indicate that SARS-CoV-2 is
27 evolving to elude the host humoral immune response, whereas recognition by T cells might benefit the
28 virus.

29

30

31

32

33 **Keywords:** SARS-CoV-2; COVID-19; Human Coronavirus; Sarbecovirus; B cell epitope; T cell

34 epitope

35

36

37

38

39 **Introduction**

40

41 The COVID-19 pandemic is caused by a novel coronavirus named SARS-CoV-2 (Coronaviridae Study
42 Group of the International Committee on Taxonomy, of Viruses, 2020). Most likely, SARS-CoV-2
43 originated and evolved in bats, eventually spilling over to humans, either directly or through an
44 intermediate host (Killerby et al., 2020; Lam et al., 2020; Liu et al., 2020; Sironi et al., 2020; Wong et
45 al., 2020; Xiao et al., 2020; Zhou et al., 2020a). Sustained human-to-human transmission determined
46 the global spread of the virus, which has now resulted in an unprecedented global sanitary crisis. In
47 fact, whereas the majority of COVID-19 cases are relatively mild, a significant proportion of patients
48 develop a serious, often fatal illness, characterized by acute respiratory distress syndrome (Wu and
49 McGoogan, 2020). Both viral-induced lung pathology and overactive immune responses are thought to
50 contribute to disease severity (St John and Rathore, 2020; Vabret et al., 2020).

51 Ample evidence suggests that coronaviruses can easily cross species barriers and have high zoonotic
52 potential. Indeed, seven coronaviruses are known to infect humans and all of them originated in
53 animals (Cui et al., 2019; Forni et al., 2017; Ye et al., 2020). Among these, HCoV-OC43, HCoV-
54 HKU1, HCoV-NL63 and HCoV-229E have been circulating for decades in human populations and
55 usually cause limited disease (Bucknall et al., 1972; Forni et al., 2017; Woo et al., 2005). They are thus
56 referred to as “common cold” coronaviruses. Conversely, MERS-CoV and SARS-CoV, whose
57 emergence in the 2000s preceded that of SARS-CoV-2, can cause serious illness and respiratory
58 distress syndrome in a non-negligible proportion of infected individuals (Petrosillo et al., 2020). Like
59 all coronaviruses, these human-infecting viruses have positive-sense, single stranded RNA genomes.
60 Two thirds of the coronavirus genome are occupied by two large overlapping open reading frames
61 (ORF1a and ORF1b), that are translated into polyproteins. These latter are processed to generate 16
62 non-structural proteins (nsp1 to nsp16). The remaining portion of the genome includes ORFs for the

63 structural proteins (spike, envelope, membrane, and nucleocapsid) and a variable number of accessory
64 proteins (Cui et al., 2019; Forni et al., 2017).

65 Analysis of the bat viruses most closely related to SARS-CoV-2 indicated that, in analogy to SARS-
66 CoV, the virus most likely required limited adaptation to gain the ability to infect and spread in our
67 species (Boni et al., 2020; Cagliani et al., 2020). Nonetheless, since its introduction in human
68 populations SARS-CoV-2 must have been subject to the selective pressure imposed by the human
69 immune system. In fact, as with most other viruses, data from COVID-19, SARS, and MERS patients
70 indicate that both B and T lymphocytes play a role in controlling infection (Channappanavar et al.,
71 2014; St John and Rathore, 2020; Vabret et al., 2020).

72 Recent efforts predicted B cell and T cell epitopes in SARS-CoV-2 proteins (Grifoni et al., 2020a) and
73 validated such predictions using sera from convalescent COVID-19 patients (Grifoni et al., 2020b).

74 These works, as well as others (Farrera et al., 2020; Poh et al., 2020), revealed that the cell-mediated
75 responses against SARS-CoV-2 are not restricted to the nucleocapsid (N) and spike (S) proteins, but
76 rather target both structural and non-structural viral products. In parallel, analyses of B cell responses
77 in SARS-CoV-2 infected patients showed that the S and N proteins are the major target of the antibody
78 response and identified specific B cell epitopes in the S protein (Farrera et al., 2020; Jiang et al., 2020;
79 Poh et al., 2020). We exploited this growing wealth of information to investigate whether, after a few
80 months of sustained transmission, the selective pressure exerted by the human adaptive immune
81 response is already detectable in the SARS-CoV-2 population and to investigate how the virus is
82 evolving in response to such a pressure.

83

84 **Results**

85

86 **Antigenic variability of SARS-CoV-2 proteins**

87

88 To analyze B cell epitope diversity in SARS-CoV-2, we randomly selected 10,000 high-quality viral
89 genomes from those available in the GISAID database (as of June 5th, 2020) (Elbe and Buckland-
90 Merrett, 2017). Potential epitopes were predicted using Immune Epitope Database (IEDB) tools, as
91 previously described (Grifoni et al., 2020a). Specifically, because they are the major targets of the
92 humoral immune response (Channappanavar et al., 2014; St John and Rathore, 2020; Vabret et al.,
93 2020), we predicted both linear and conformational B epitopes for the S and N proteins, whereas only
94 linear epitopes were predicted for the other viral proteins (Table S1). A good correspondence was
95 observed between B cell epitope predictions for the S protein and epitopes identified in two works that
96 systematically mapped antibody responses in the sera of convalescent COVID-19 patients (Farrera et
97 al., 2020; Poh et al., 2020) (Figure 1).

98 Variability at each amino acid site of the proteins encoded by SARS-CoV-2 was quantified using
99 Shannon's entropy (H). Specifically, only predicted proteins longer than 60 amino acids were analyzed.
100 Because most positions in SARS-CoV-2 genomes are invariable across the sampled genomes, the
101 distribution of H is zero-inflated, making the use of conventional statistical tests inappropriate
102 (McElduff et al., 2010). We thus calculated statistical significance by permutations - i.e., by reshuffling
103 epitope positions as amino acid stretches of the same size as the predicted epitopes. This approach also
104 has the advantage of accounting for the possibility that, as a result of locally varying selective
105 constraints, H is not independent among continuous protein positions.

106 Using this methodology, we found that, for the N and nsp16 proteins, positions mapping to predicted B
107 cell linear epitopes are significantly more variable than those not mapping to these epitopes. Higher
108 diversity of B cell epitopes was also observed for S, although it did not reach statistical significance
109 (Figure 2). However, the H distribution for the spike protein includes a clear outlier represented by
110 position 614 (Figure 1). Recent works indicated that the D614G variant, which is now prevalent

111 worldwide, enhances viral infectivity without affecting neutralization by convalescent patient plasma
112 (Korber et al., 2020; Yurkovetskiy et al., 2020; Zhang et al., 2020b). Hence, the frequency increase of
113 this variant is unlikely to be related to immune evasion. We thus repeated the analyses after excluding
114 position 614 and we observed that predicted B cell linear epitopes in the spike protein are significantly
115 more variable than non-epitope positions (Figure 2). The same analysis for B cell conformational
116 epitopes in the N and S proteins indicated a similar trend, although statistical significance was not
117 reached (not shown). This is most likely due to the small number of positions in these epitopes.
118 Overall, these data fit very well with the observation that most humoral immune responses against
119 SARS-CoV-2 and other human coronaviruses are directed against the S and N proteins (Farrera et al.,
120 2020; Jiang et al., 2020; Poh et al., 2020). These results also support the notion that the selective
121 pressure exerted by the human antibody response is already detectable in the SARS-CoV-2 population.
122 We next assessed whether epitopes for cell-mediated immune responses are also more variable than
123 non-epitope positions. We thus retrieved predicted CD4⁺ and CD8⁺ T cell epitopes from a previous
124 work (Grifoni et al., 2020a). These epitope predictions were shown to be reliable, as they capture a
125 significant proportion of T cell responses in the sera of convalescent COVID-19 patients (Grifoni et al.,
126 2020b). Analysis of entropy values indicated that CD4⁺ T cell epitopes are significantly less variable
127 than non-epitope positions for the N and nsp16 proteins (Figure 2). A similar trend was observed for
128 ORF8, E, and S, although significance was not reached. Reduction of variability was also observed for
129 CD8⁺ T cell epitopes for the N protein, as well as for ORF3a. Higher variability in epitope positions
130 was observed for nsp8 and nsp14 for CD4⁺ T cells alone (Figure 2). Because several epitopes for T
131 cells co-map with B cell epitopes, which tend to show higher diversity, we compared positions within
132 CD4⁺ or CD8⁺ T cell epitopes only (not overlapping with B cell epitopes) with positions not mapping
133 to any of these epitopes. A significant reduction of variability was observed for S, N, ORF8, nsp15, and
134 nsp16, whereas higher diversity was still evident for nsp8 (Figure 2).

135 Overall, these data indicate that T cell epitopes in the most immunogenic SARS-CoV-2 proteins (S, N,
136 ORF3a, and ORF8) (Grifoni et al., 2020b; Peng et al., 2020b) tend to be more conserved than non-
137 epitopes. However, this was not the case for other proteins targeted by T cell responses, namely M,
138 ORF7a, nsp3, nsp4, and nsp6.

139 Clearly, protein sequence variability is strongly influenced by functional and structural constraints. We
140 thus reasoned that if the observations reported above were secondary to the incidental co-localization of
141 T cell epitopes with more constrained regions, a similar pattern should be observed for H values
142 calculated on an alignment of proteins from other sarbecoviruses. In fact, all these viruses, with the
143 exclusion of SARS-CoV, were sampled from bats. Thus, whereas structural/functional constraints are
144 expected to be maintained across long evolutionary time frames, the pressure exerted by the human
145 cell-mediated immune response is not, as, in different species, antigen processing within host cells
146 results in the preferential presentation of diverse viral epitopes to T lymphocytes depending on the
147 MHC gene repertoire and on distinct preferences of the antigen processing pathway (Abduriyim et al.,
148 2019; Burgevin et al., 2008; Hammer et al., 2007; Lu, Dan AND Liu, Kefang AND Zhang, Di AND
149 Yue, Can AND Lu, Qiong AND Cheng, Hao AND Wang, Liang AND Chai, Yan AND Qi, Jianxun
150 AND Wang, Lin-Fa AND Gao, George F. AND Liu, William J., 2019; Wynne et al., 2016). Conversely,
151 epitopes for antibodies tend to be conserved across species (Tse et al., 2017; Wiehe et al., 2014) and
152 consequently the selective pressure acting on these positions is expected to be constant across time and
153 hosts.

154 We thus aligned the SARS-CoV-2 reference sequences of proteins showing decreased or increased
155 variability in T cell epitopes with those of 45 sarbecoviruses (Table S2). Calculation of H indicated a
156 significant difference only for CD4⁺ T cell epitopes in the N protein. Conversely, B cell epitopes were
157 more variable than non-epitope positions for the S, N and nsp16 proteins (Figure 3). Overall, these
158 results indicate that the variability within SARS-CoV-2 T cell epitopes is not primarily driven by

159 functional/structural constraints, but most likely results from the interaction with the human adaptive
160 immune response.

161

162 **Comparison with other human coronaviruses**

163

164 Given the results above we set out to determine whether the other human coronaviruses show the same
165 tendency of reduced and increased variability at T cell and B cell epitopes, respectively. For these
166 viruses, analyses were restricted to the N and S proteins, as they are the most antigenic proteins and
167 because the number of complete viral genomes is relatively limited (Table S3).

168 SARS-CoV, the human coronavirus most similar to SARS-CoV-2, caused the first human outbreak in
169 2002/2003 after a spill-over from palm civets, followed by human-to-human transmission chains (Shi
170 and Wang, 2017). A second zoonotic transmission occurred in December 2003 and caused a limited
171 number of cases (Shi and Wang, 2017; Wang et al., 2005). Viral genomes sampled during the second

172 outbreak were not included in the analyses because their evolution occurred in the civet reservoir
173 (Table S4). Four other human coronaviruses – HCoV-OC43, HCoV-HKU1 (members of the

174 *Embecovirus* subgenus), HCoV-229E (*Duvinavirus* subgenus), and HCoV-NL63 (*Setracovirus*
175 subgenus) - have been transmitting within human populations for at least 70 years (Forni et al., 2017).

176 Thus, all available S and N sequences were included in the analyses (Table S4). Conversely, MERS-
177 CoV displays limited ability of human-to-human transmission and outbreaks were caused by repeated
178 spill-over events from the camel host (Cui et al., 2019). For this reason MERS-CoV was excluded from
179 the analyses.

180 Quantification of sequence variability by H calculation indicated that B cell epitopes in the S protein
181 are significantly more variable than non-epitopes for SARS-CoV, HCoV-OC43 and HCoV-HKU1
182 (Figure 4). Analysis of CD4⁺ and CD8⁺ T cell epitopes in these viruses indicated no increased diversity

183 for epitope compared to non-epitope positions, with the exclusion of the S protein of SARS-CoV for
184 CD4⁺ T cells. However, when positions within B cell epitopes were excluded from the analysis, this
185 difference disappeared and T cell epitopes were found to be significantly less variable than non-
186 epitopes for the spike proteins of HCoV-HKU1 and HCoV-OC43, as well as for the N protein of
187 HCoV-229E (figure 4). Thus, the lack of antigenic diversity at T cell epitopes is a common feature of
188 human coronaviruses, which instead tend to maintain sequence conservation of such epitopes.

189

190 **Discussion**

191 The origin of SARS-CoV-2 is still uncertain and it is presently unknown whether the virus spilled over
192 from a bat or another intermediate host. The hypothesis of a zoonotic origin is strongly supported by
193 multiple lines of evidence, although it cannot be excluded that SARS-CoV-2 transmitted cryptically in
194 humans before gaining the ability of spreading efficiently among people (Andersen et al., 2020; Sironi
195 et al., 2020). Whatever the initial events associated with the early phases of the pandemic, it is clear
196 that circulating SARS-CoV-2 viruses shared a common ancestor at the end of 2019 (Li et al., 2020; van
197 Dorp et al., 2020). Due to its recent origin, the genetic diversity of the SARS-CoV-2 population is still
198 limited. This is also the result of the relatively low mutation rate of coronaviruses (as compared to other
199 RNA viruses), which encode enzymes with some proofreading ability (Denison et al., 2011; Forni et
200 al., 2017). Nonetheless, the huge number of transmissions worldwide have allowed thousands of
201 mutations to appear in the viral population and, thanks to enormous international sequencing efforts,
202 more than 14,000 amino acid replacements have currently been reported (<http://cov-glue.cvr.gla.ac.uk>).
203 Irrespective of the host, most variants are expected to be deleterious for viral fitness, or to have no
204 consequences (Cagliani et al., 2020; Grubaugh et al., 2020; van Dorp et al., 2020). However, a portion
205 of replacements may favor the virus and some of these may contribute to adaptation to the human host.
206 In particular, the recent and ongoing evolution of SARS-CoV-2 is expected to be at least partially

207 driven by the selective pressure imposed by the human immune system. Indeed, antigenic drift or
208 immune evasion mutations have been reported for other zoonotic viruses such as Lassa virus (Andersen
209 et al., 2015) and Influenza A virus (Su et al., 2015). The emergence of immune evasion variants was
210 also observed during an outbreak of MERS-CoV in South Korea, when mutations in the spike proteins
211 were positively selected as they facilitated viral escape from neutralizing antibodies, even though the
212 same variants decreased binding to the cellular receptor (Kim et al., 2019; Kim et al., 2019; Kleine-
213 Weber et al., 2019; Rockx et al., 2010). This exemplifies a phenomenon often observed in other
214 viruses, most notably HIV-1 (Liu et al., 2007; Martinez-Picado et al., 2006; Schneidewind et al., 2007;
215 Schneidewind et al., 2008), whereby the virus trades off immune evasion with a fitness cost. As a
216 consequence, immune evasion mutations may be only transiently maintained in viral populations. For
217 this reason we decided to quantify epitope variability in terms of entropy, rather than relying on
218 measures based on substitution rates (dN/dS), which were developed for application to variants that go
219 to fixation in different lineages over time (Kryazhimskiy and Plotkin, 2008).

220 The MERS-CoV mutants responsible for the outbreak in South Korea also testify the relevance of the
221 antibody response in coronavirus control and the selective pressure imposed by humoral immunity on
222 the virus (Kim et al., 2019; Kleine-Weber et al., 2019; Rockx et al., 2010). This is most likely the case
223 for SARS-CoV-2, as well, as recent report indicated that the sera of most COVID-19 convalescent
224 patients have virus-neutralization activities and antibody titers negatively correlate with viral load
225 (OKBA et al., 2020; Vabret et al., 2020; Wu et al., 2020; Zhou et al., 2020b). Nonetheless, studies on
226 relatively large COVID-19 patient cohorts reported that patients with severe disease display stronger
227 IgG responses than milder cases and a negative correlation between anti-S antibody titers and
228 lymphocyte counts was reported (Jiang et al., 2020; Vabret et al., 2020; Wu et al., 2020; Zhang et al.,
229 2020a; Zhao et al., 2020). Consistently, asymptomatic SARS-CoV-2 infected individuals were recently
230 reported to have lower virus-specific IgG levels than COVID-19 patients (Long et al., 2020). These

231 observations raised concerns that humoral responses might not necessarily be protective, but rather
232 pathogenic, either via antibody-dependent enhancement (ADE) or other mechanisms (Cao, 2020;
233 Iwasaki and Yang, 2020; Wu et al., 2020).

234 Clearly, gaining insight into the dynamic interaction between SARS-CoV-2 and the human immune
235 system is of fundamental importance not only to understand COVID-19 immunopathogenesis, but also
236 to inform therapeutic and preventive viral control strategies. We thus exploited the availability of a
237 large number of fully sequenced high-quality SARS-CoV-2 genomes, as well as validated predictions
238 of B cell and T cell epitopes, to investigate whether the selective pressure exerted by the adaptive
239 immune response is detectable in global SARS-CoV-2 population, and if the virus is evolving to evade
240 it. Results indicated that B cell epitopes in the N and S proteins, which represent the major targets of
241 the antibody response, have higher diversity than non-epitope positions. The same was observed for the
242 spike proteins of HCoV-HKU1, HCoV-OC43 and SARS-CoV, although data on the latter virus should
243 be taken with caution as they derive from a relatively small number of sequences sampled over a short
244 time frame. Conversely, no evidence of antibody-mediated selective pressure was evident for HCoV-
245 229E and HCoV-NL63. The reasons underlying these differences are unclear, but recent data on a
246 relatively small population of patients with respiratory disease indicated that the titers of neutralizing
247 antibodies against HCoV-OC43 tend to be higher compared to those against HCoV-229E and HCoV-
248 NL63 (HCoV-HKU1 was not evaluated), suggesting the two latter viruses elicit mainly non-
249 neutralizing responses (Gorse et al., 2020).

250 B cell epitopes within nsp16 were also found to be variable, although this protein was not reported to
251 be immunogenic (Grifoni et al., 2020b). However, the antibody response to SARS-CoV-2 has presently
252 been systematically analyzed in a relatively small number of patients and most studies focused on
253 structural proteins. It is thus possible that, during infection, antibodies against nsp16 are raised but they
254 have not been detected yet. An alternative possibility is that B cell epitopes in nsp16, which is highly

255 conserved in SARS-CoV-2 strains (Cagliani et al., 2020), coincide with regions of relatively weaker
256 constraint. This hypothesis is partially supported by the observation that these same positions also
257 display higher diversity when entropy is calculated on an alignment of sarbecovirus nsp16 proteins.
258 More intriguingly, this result may indicate that nsp16, together with S and N, is a target of B cell
259 responses in the bat reservoirs. In fact, as mentioned above, antibody binding sites tend to be conserved
260 across species (Tse et al., 2017; Wiehe et al., 2014) and thus the selective pressure exerted on B cell
261 epitopes is likely to be constant across hosts. Whereas the immunogenicity of nsp16 remains to be
262 evaluated, these data suggest that SARS-CoV-2 is evolving to elude the host humoral immune
263 response. We however note that this observation does not necessarily imply that antibodies against
264 SARS-CoV-2 are protective and it does not rule out the possibility that humoral responses contribute to
265 COVID-19 pathogenesis.

266 In COVID-19 patients, antibody titers were found to correlate with the strength of virus-specific T cell
267 responses (Ni et al., 2020). Surprisingly, we found that, in the SARS-CoV-2 population, epitopes for
268 CD4⁺ and CD8⁺ T cells are not more variable than non-epitope positions. Conversely, a significant
269 reduction in epitope variability was observed for a subset of viral proteins, in particular for some of the
270 most immunogenic ones (S, N, ORF8, and ORF3a) (Grifoni et al., 2020b; Peng et al., 2020a). To check
271 that the result was not due to stronger structural/functional constraints acting on epitope positions, we
272 again used H values calculated on an alignment of sarbecovirus genomes, all of which, with the
273 exclusion of SARS-CoV, were sampled in bats. T cell responses are initiated by the presentation of
274 antigenic epitopes by MHC (major histocompatibility complex) class I and class II molecules. Different
275 mammals have diverse MHC gene repertoires and thus present distinct antigens. In particular, recent
276 data from various bat species indicated that many MHC class I molecules have a 3- or 5-amino acid
277 insertion in the peptide binding pocket, resulting in very different presented peptide repertoires
278 compared to the MHC class I molecules of other mammals (Abduriyim et al., 2019; Lu, Dan AND Liu,

279 Kefang AND Zhang, Di AND Yue, Can AND Lu, Qiong AND Cheng, Hao AND Wang, Liang AND
280 Chai, Yan AND Qi, Jianxun AND Wang, Lin-Fa AND Gao, George F. AND Liu, William J., 2019; Ng et
281 al., 2016; Papenfuss et al., 2012; Wynne et al., 2016) . Thus, the selective pressure acting on T cell
282 epitopes is most likely volatile and not conserved in humans and bats. Analysis of sarbecovirus proteins
283 indicated that, apart from CD4⁺ T cell epitopes in the N protein, the T cell epitopes predicted in SARS-
284 CoV-2 proteins are not less diverse than non-epitope positions, suggesting that epitope conservation is
285 not simply secondary to structural or functional constraints, but may result from interaction with human
286 T cell responses. Of course, another possible explanation for this finding is that the prediction tools
287 failed to identify real epitopes. However, we retrieved epitopes from a previous work and the authors
288 validated their predictions using the sera of 20 patients who recovered from COVID-19 (Grifoni et al.,
289 2020a; Grifoni et al., 2020b). Moreover, if a general artifact linked to epitope prediction was
290 introduced, we would not expect to observe significant differences and not specifically in the proteins
291 that represent the major targets of T cell responses.

292 Unexpected conservation of T cell epitopes was previously observed for HIV-1 and *Mycobacterium*
293 *tuberculosis* (MTB), both of which cause chronic infections in humans (Comas et al., 2010; Coscolla et
294 al., 2015; Lindestam Arlehamn et al., 2015; Sanjuán, Rafael AND Nebot, Miguel R. AND Peris, Joan
295 B. AND Alcamí, José, 2013). In the case of HIV-1, immune activation most likely favors the virus by
296 increasing the rate of CD4⁺ T cell trans-infection (Sanjuán, Rafael AND Nebot, Miguel R. AND Peris,
297 Joan B. AND Alcamí, José, 2013). Conversely, the mechanisms underlying MTB epitope conservation
298 are not fully elucidated. A possible explanation is that conserved epitopes generate a decoy immune
299 response and advantage the bacterium. An alternative possibility is that T cell activation results in lung
300 tissue inflammation and damage (cavitary tuberculosis), which favors MTB transmission by aerosol
301 (Coscolla et al., 2015; Lindestam Arlehamn et al., 2015). Although these mechanisms are unlikely to be
302 at play in the case of SARS-CoV-2, a deregulated immune response has been associated with COVID-

303 19 pathogenesis (Hannan et al., 2020). Specifically, recent data indicated that patients recovering from
304 severe COVID-19 have broader and stronger T cell responses compared to mild cases (Peng et al.,
305 2020b). This was particularly evident for responses against the S, membrane (M), ORF3a, and ORF8
306 proteins (Peng et al., 2020b). Although this observation might simply reflect higher viral loads in
307 severe cases, the possibility that the T cell response itself is deleterious cannot be excluded. Moreover,
308 the same authors reported that CD8⁺ T cells targeting different virus proteins have distinct cytokine
309 profiles, suggesting that the virus can modulate the host immune response to its benefit (Peng et al.,
310 2020b). Additionally, a post-mortem study on six patients who died from COVID-19 indicated that
311 infection of macrophages can lead to activation-induced T cell death, which may eventually be
312 responsible for lymphocytopenia (chen et al., 2020). However, we also found a trend of lower diversity
313 of T cell epitopes for common cold coronaviruses, indicating that epitope conservation *per se* is not
314 directly linked to disease severity. Moreover, other SARS-CoV-2 immunogenic proteins such as M and
315 ORF7 did not show differences in T cell epitope conservation, which was instead observed for nsp16
316 and nsp15. These latter are not known to be T cell targets (Grifoni et al., 2020b). Clearly, further
317 analyses will be required to clarify the significance of T cell epitope conservation in SARS-CoV-2. An
318 interesting possibility is that both for SARS-CoV-2 and for common cold coronaviruses, conservation
319 serves to maintain epitopes that elicit tolerizing T cell responses or induce T cells with regulatory
320 activity. Indeed, we considered T cell epitopes as a whole, but differences exist in terms of variability
321 and, most likely, antigenicity. This clearly represents a limitation of this study, but the modest amount
322 of genetic diversity in the SARS-CoV-2 population does not presently allow analysis of single epitope
323 regions. Moreover, more detailed and robust analyses will indubitably require the systematic,
324 experimental definition of T and B cell epitopes in the SARS-CoV-2 proteome.

325

326

327 **Material and Methods**

328 **Epitope Prediction**

329 Epitope prediction was performed using different tools from The Immune Epitope Database (IEDB)
330 (<https://www.iedb.org/>), as previously suggested (Grifoni et al., 2020a). Protein sequences from
331 reference strains of human coronaviruses were used as input for all prediction analyses (SARS-CoV-2,
332 NC_045512; SARS-CoV, NC_004718; Human coronavirus 229E, NC_002645; Human coronavirus
333 NL63, NC_005831; Human coronavirus OC43, NC_006213; Human coronavirus HKU1, NC_006577).
334 In particular, for linear B cell epitope prediction, we used the Bepipred Linear Epitope Prediction 2.0
335 tool (Jespersen et al., 2017) with a cutoff of 0.550 and epitope length > 7. Conformational B epitopes
336 for the S and N proteins of SARS-CoV-2 were calculated using Discotope 2.0 (Kringelum et al., 2012)
337 with a threshold = -2.5 and published 3D protein structures (PDB IDs: 6VSB, spike; 6M3M (N-term)
338 and 7C22 (C-term), nucleocapsid protein).

339 SARS-CoV-2 predicted T cell epitopes were retrieved from a previous work (Grifoni et al., 2020a). For
340 all other coronaviruses, we applied the same methodology used by Grifoni et al. (Grifoni et al., 2020a).
341 CD4⁺ cell epitopes were predicted using TepiTool (Paul et al., 2016) with default parameters. CD8⁺
342 epitopes were predicted by using MHC-I Binding Predictions v2.23 tool (<http://tools.iedb.org/mhci/>).
343 The NetMHCpan EL 4.0 method (Jurtz et al., 2017) was applied and the 12 most frequent HLA class I
344 alleles in human populations (HLA-A01:01, HLA-A02:01, HLA-A03:01, HLA-A11:01, HLA-A23:01,
345 HLA-A24:02, HLA-B07:02, HLA-B08:01, HLA-B35:01, HLA-B40:01, HLA-B44:02, HLA-B44:03)
346 were analyzed with a 8-14 kmer range. Only epitopes with a score rank ≤ 0.1 in one of the 12 HLA
347 classes were selected.

348 **Sequences and alignments**

349

350 SARS-CoV-2 protein sequences were downloaded from the GISAID Initiative (<https://www.gisaid.org>)
351 database (as of June, 5th). All protein sequences were retrieved and several filters were applied. Only
352 complete genomes flagged as “high coverage only” and “human” were selected. Positions
353 recommended to be masked by DeMaio and coworkers ([https://virological.org/t/masking-strategies-for-](https://virological.org/t/masking-strategies-for-sars-cov-2-alignments/480)
354 [sars-cov-2-alignments/480](https://virological.org/t/masking-strategies-for-sars-cov-2-alignments/480), last accessed June, 5th, 2020) were also removed.

355 Finally, for each SARS-CoV-2 protein, we selected only strains that had the same length as the protein
356 in the SARS-CoV-2 reference strain (NC_045512), generating a set of at least 23625 sequences for
357 each ORF. Proteins with less than 60 amino acids were excluded from the analyses.

358 The list of GISAID IDs along with the list of laboratories which generated the data is provided as Table
359 S5.

360 For all the other human coronaviruses, as well as for a set of non-human infecting sarbecoviruses,
361 sequences of either complete genome or single ORFs (i.e. nucleocapsid and spike protein) were
362 retrieved from the National Center for Biotechnology Information database (NCBI,
363 <http://www.ncbi.nlm.nih.gov/>). For all human coronaviruses, the only filter we applied was the host
364 identification as “human”. SARS-CoV strains sampled during the second outbreak were excluded from
365 the analyses. NCBI ID identifiers are listed as Table S2 and Table S4.

366 Alignments were generated using MAFFT (Kato and Standley, 2013).

367

368 **Protein variability and statistical analysis**

369 Variability at each amino acid position was estimated using the Shannon's entropy (H) index using the
370 Shannon Entropy-One tool from the HIV database (<https://www.hiv.lanl.gov/content/index>), with
371 ambiguous character (e.g. gaps) excluded from the analysis. For SARS-CoV-2 strains, H was calculated
372 on alignments of 10000 randomly selected sequences for each protein. For each protein we evaluated
373 the difference D between average H values at epitope and non-epitope positions.

374 Most positions of analyzed viruses are invariable along the alignments, so the distribution of H is zero-
375 inflated. We thus calculated statistical significance by permutations. For each protein, the predicted
376 epitope intervals were collapsed to a single position while non-epitope intervals were left unchanged.
377 After randomly shuffling this collapsed sequence it was expanded back to full length and the difference
378 between shuffled epitope and non-epitope H values was calculated. This procedure was repeated 1000
379 times and the proportion of repetitions showing a difference more extreme than D was reported as p-
380 value. An in house R script was written and is available as supplementary text S1.

381

382 **Acknowledgments**

383 We gratefully acknowledge the authors, originating and submitting laboratories of the sequences from
384 GISAID's EpiCoV™ database on which this research is based. This work was supported by the Italian
385 Ministry of Health ("Ricerca Corrente 2019-2020" to MS, "Ricerca Corrente 2018-2020" to DF)

386

387 **Author Contributions**

388 Conceptualization, DF and MS; Formal Analysis, MS, UP, and DF; Investigation, DF, RC, CP, AM,
389 and MS; Visualization, DF, RC; Writing –Original Draft, MS and DF Writing –Review & Editing, MS,
390 MC, RC, UP; Funding Acquisition MS and DF; Supervision, MS and MC.

391

392 **Declaration of Interests**

393 The authors declare no competing interests.

394

395

396 **References**

397

398 Abduriyim, S., Zou, D.H., and Zhao, H. (2019). Origin and evolution of the major histocompatibility
399 complex class I region in eutherian mammals. *Ecol. Evol.* 9, 7861-7874.

400 Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., and Garry, R.F. (2020). The proximal origin
401 of SARS-CoV-2. *Nat. Med.* 26, 450-452.

402 Andersen, K.G., Shapiro, B.J., Matranga, C.B., Sealfon, R., Lin, A.E., Moses, L.M., Folarin, O.A.,
403 Goba, A., Odiya, I., Ehiane, P.E., *et al.* (2015). Clinical sequencing uncovers origins and evolution of
404 lassa virus. *Cell.* 162, 738-750.

405 Boni, M.F., Lemey, P., Jiang, X., Lam, T.T., Perry, B., Castoe, T., Rambaut, A., and Robertson, D.L.
406 (2020). Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19
407 pandemic. *Biorxiv.* , 2020.03.30.015008.

408 Bucknall, R.A., King, L.M., Kapikian, A.Z., and Chanock, R.M. (1972). Studies with human
409 coronaviruses. II. some properties of strains 229E and OC43. *Proc. Soc. Exp. Biol. Med.* 139, 722-727.

410 Burgevin, A., Saveanu, L., Kim, Y., Barilleau, E., Kotturi, M., Sette, A., van Endert, P., and Peters, B.
411 (2008). A detailed analysis of the murine TAP transporter substrate specificity. *PLoS One.* 3, e2402.

412 Cagliani, R., Forni, D., Clerici, M., and Sironi, M. (2020). Computational inference of selection
413 underlying the evolution of the novel coronavirus, SARS-CoV-2. *J. Virol.*

414 Cao, X. (2020). COVID-19: Immunopathology and its implications for therapy. *Nature Reviews*
415 *Immunology.* 20, 269-270.

416 Channappanavar, R., Zhao, J., and Perlman, S. (2014). T cell-mediated immune response to respiratory
417 coronaviruses. *Immunol. Res.* 59, 118-128.

418 chen, y., Feng, Z., Diao, B., Wang, R., Wang, G., Wang, C., Tan, Y., Liu, L., Wang, C., Liu, Y., *et al.*
419 (2020). The novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) directly decimates

420 human spleens and lymph nodes. Medrxiv.

421 Comas, I., Chakravarti, J., Small, P.M., Galagan, J., Niemann, S., Kremer, K., Ernst, J.D., and
422 Gagneux, S. (2010). Human T cell epitopes of mycobacterium tuberculosis are evolutionarily
423 hyperconserved. *Nat. Genet.* *42*, 498-503.

424 Coronaviridae Study Group of the International Committee on Taxonomy,of Viruses. (2020). The
425 species severe acute respiratory syndrome-related coronavirus: Classifying 2019-nCoV and naming it
426 SARS-CoV-2. *Nat Microbiol.* *5*, 536-544.

427 Coscolla, M., Copin, R., Sutherland, J., Gehre, F., de Jong, B., Owolabi, O., Mbayo, G., Giardina, F.,
428 Ernst, J.D., and Gagneux, S. (2015). M. tuberculosis T cell epitope analysis reveals paucity of antigenic
429 variation and identifies rare variable TB antigens. *Cell. Host Microbe.* *18*, 538-548.

430 Cui, J., Li, F., and Shi, Z.L. (2019). Origin and evolution of pathogenic coronaviruses. *Nat. Rev.*
431 *Microbiol.* *17*, 181-192.

432 Denison, M.R., Graham, R.L., Donaldson, E.F., Eckerle, L.D., and Baric, R.S. (2011). Coronaviruses:
433 An RNA proofreading machine regulates replication fidelity and diversity. *RNA Biol.* *8*, 270-279.

434 Elbe, S., and Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative
435 contribution to global health. *Glob. Chall.* *1*, 33-46.

436 Farrera, L., Dagher, J., Barluenga, S., Cohen, P.R., Pagano, S., Yerly, S., Kaiser, L., Vuilleumier, N.,
437 and Winssinger, N. (2020). Identification of immunodominant linear epitopes from SARS-CoV-2
438 patient plasma. Medrxiv.

439 Forni, D., Cagliani, R., Clerici, M., and Sironi, M. (2017). Molecular evolution of human coronavirus
440 genomes. *Trends Microbiol.* *25*, 35-48.

441 Gorse, G.J., Donovan, M.M., and Patel, G.B. (2020). Antibodies to coronaviruses are higher in older
442 compared with younger adults and binding antibodies are more sensitive than neutralizing antibodies in
443 identifying coronavirus-associated illnesses. *J. Med. Virol.* *92*, 512-517.

- 444 Grifoni, A., Sidney, J., Zhang, Y., Scheuermann, R.H., Peters, B., and Sette, A. (2020a). A sequence
445 homology and bioinformatic approach can predict candidate targets for immune responses to SARS-
446 CoV-2. *Cell. Host Microbe*. 27, 671-680.e2.
- 447 Grifoni, A., Weiskopf, D., Ramirez, S.I., Mateus, J., Dan, J.M., Moderbacher, C.R., Rawlings, S.A.,
448 Sutherland, A., Premkumar, L., Jadi, R.S., *et al.* (2020b). Targets of T cell responses to SARS-CoV-2
449 coronavirus in humans with COVID-19 disease and unexposed individuals. *Cell*. 181, 1489.
- 450 Grubaugh, N.D., Petrone, M.E., and Holmes, E.C. (2020). We shouldn't worry when a virus mutates
451 during disease outbreaks. *Nature Microbiology*. 5, 529-530.
- 452 Hammer, G.E., Kanaseki, T., and Shastri, N. (2007). The final touches make perfect the peptide-MHC
453 class I repertoire. *Immunity*. 26, 397-406.
- 454 Hannan, M.A., Rahman, M.A., Rahman, M.S., Sohag, A.A.M., Dash, R., Hossain, K.S., Farjana, M.,
455 and Uddin, M.J. (2020). Intermittent fasting, a possible priming tool for host defense against SARS-
456 CoV-2 infection: Crosstalk among calorie restriction, autophagy and immune response. *Immunol. Lett.*
- 457 Iwasaki, A., and Yang, Y. (2020). The potential danger of suboptimal antibody responses in COVID-19.
458 *Nature Reviews Immunology*. 20, 339-341.
- 459 Jespersen, M.C., Peters, B., Nielsen, M., and Marcatili, P. (2017). BepiPred-2.0: Improving sequence-
460 based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res.* 45, W24-W29.
- 461 Jiang, H., Li, Y., Zhang, H., Wang, W., Men, D., Yang, X., Qi, H., Zhou, J., and Tao, S. (2020). Global
462 profiling of SARS-CoV-2 specific IgG/ IgM responses of convalescents using a proteome microarray.
463 Medrxiv.
- 464 Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017). NetMHCpan-4.0:
465 Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding
466 affinity data. *J. Immunol.* 199, 3360-3368.
- 467 Kang, S., Yang, M., Hong, Z., Zhang, L., Huang, Z., Chen, X., He, S., Zhou, Z., Zhou, Z., Chen, Q., *et*

- 468 *al.* (2020). Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals
469 potential unique drug targeting sites. *Acta Pharm. Sin. B.*
- 470 Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7:
471 Improvements in performance and usability. *Mol. Biol. Evol.* *30*, 772-780.
- 472 Killerby, M.E., Biggs, H.M., Midgley, C.M., Gerber, S.I., and Watson, J.T. (2020). Middle east
473 respiratory syndrome coronavirus transmission. *Emerg. Infect. Dis.* *26*, 191-198.
- 474 Kim, Y., Aigerim, A., Park, U., Kim, Y., Rhee, J., Choi, J., Park, W., Park, S., Kim, Y., Lim, D., *et al.*
475 (2019). Sequential emergence and wide spread of neutralization escape middle east respiratory
476 syndrome coronavirus mutants, south korea, 2015. *Emerging Infectious Diseases.* *25*, 1161-1168.
- 477 Kleine-Weber, H., Elzayat, M.T., Wang, L., Graham, B.S., Müller, M.A., Drosten, C., Pöhlmann, S.,
478 and Hoffmann, M. (2019). Mutations in the spike protein of middle east respiratory syndrome
479 coronavirus transmitted in korea increase resistance to antibody-mediated neutralization. *J. Virol.* *93*,
480 e01381-18. doi: 10.1128/JVI.01381-18. Print 2019 Jan 15.
- 481 Korber, B., Fischer, W., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi,
482 E., Bhattacharya, T., and Foley, B. (2020). Tracking changes in SARS-CoV-2 spike: Evidence that
483 D614G increases infectivity of the COVID-19 virus. *Cell.*
- 484 Kringelum, J.V., Lundegaard, C., Lund, O., and Nielsen, M. (2012). Reliable B cell epitope predictions:
485 Impacts of method development and improved benchmarking. *PLoS Comput. Biol.* *8*, e1002829.
- 486 Kryazhimskiy, S., and Plotkin, J.B. (2008). The population genetics of dN/dS. *PLoS Genet.* *4*,
487 e1000304.
- 488 Lam, T.T., Shum, M.H., Zhu, H., Tong, Y., Ni, X., Liao, Y., Wei, W., Cheung, W.Y., Li, W., Li, L., *et al.*
489 (2020). Identification of 2019-nCoV related coronaviruses in malayan pangolins in southern china.
490 *Biorxiv.* , 2020.02.13.945485.
- 491 Li, X., Wang, W., Zhao, X., Zai, J., Zhao, Q., Li, Y., and Chaillon, A. (2020). Transmission dynamics

492 and evolutionary history of 2019-nCoV. *J. Med. Virol.* *92*, 501-511.

493 Lindestam Arlehamn, C.S., Paul, S., Mele, F., Huang, C., Greenbaum, J.A., Vita, R., Sidney, J., Peters,
494 B., Sallusto, F., and Sette, A. (2015). Immunological consequences of intragenus conservation of
495 mycobacterium tuberculosis T-cell epitopes. *Proceedings of the National Academy of Sciences.* *112*,
496 E147-E155.

497 Liu, P., Jiang, J., Wan, X., Hua, Y., Wang, X., Hou, F., Chen, J., Zou, J., and Chen, J. (2020). Are
498 pangolins the intermediate host of the 2019 novel coronavirus (2019-nCoV) ? *Biorxiv.* ,
499 2020.02.18.954628.

500 Liu, Y., McNevin, J., Zhao, H., Tebit, D.M., Troyer, R.M., McSweyn, M., Ghosh, A.K., Shriner, D.,
501 Arts, E.J., McElrath, M.J., *et al.* (2007). Evolution of human immunodeficiency virus type 1 cytotoxic
502 T-lymphocyte epitopes: Fitness-balanced escape. *J. Virol.* *81*, 12179-12188.

503 Long, Q.X., Tang, X.J., Shi, Q.L., Li, Q., Deng, H.J., Yuan, J., Hu, J.L., Xu, W., Zhang, Y., Lv, F.J., *et*
504 *al.* (2020). Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections. *Nat.*
505 *Med.*

506 Lu, Dan AND Liu, Kefang AND Zhang, Di AND Yue, Can AND Lu, Qiong AND Cheng, Hao AND
507 Wang, Liang AND Chai, Yan AND Qi, Jianxun AND Wang, Lin-Fa AND Gao, George F. AND
508 Liu, William J. (2019). Peptide presentation by bat MHC class I provides new insight into the antiviral
509 immunity of bats. *PLOS Biology.* *17*, 1-24.

510 Martinez-Picado, J., Prado, J.G., Fry, E.E., Pfafferott, K., Leslie, A., Chetty, S., Thobakgale, C.,
511 Honeyborne, I., Crawford, H., Matthews, P., *et al.* (2006). Fitness cost of escape mutations in p24 gag
512 in association with control of human immunodeficiency virus type 1. *J. Virol.* *80*, 3617-3623.

513 McElduff, F., Cortina-Borja, M., Chan, S.K., and Wade, A. (2010). When t-tests or wilcoxon-mann-
514 whitney tests won't do. *Adv. Physiol. Educ.* *34*, 128-133.

515 Ng, J.H., Tachedjian, M., Deakin, J., Wynne, J.W., Cui, J., Haring, V., Broz, I., Chen, H., Belov, K.,

- 516 Wang, L.F., *et al.* (2016). Evolution and comparative analysis of the bat MHC-I region. *Sci. Rep.* *6*,
517 21256.
- 518 Ni, L., Ye, F., Cheng, M.L., Feng, Y., Deng, Y.Q., Zhao, H., Wei, P., Ge, J., Gou, M., Li, X., *et al.*
519 (2020). Detection of SARS-CoV-2-specific humoral and cellular immunity in COVID-19 convalescent
520 individuals. *Immunity.* *52*, 971-977.e3.
- 521 OKBA, N.M.A., Muller, M.A., Li, W., Wang, C., GeurtsvanKessel, C.H., Corman, V.M., Lamers,
522 M.M., Sikkema, R.S., de Bruin, E., Chandler, F.D., *et al.* (2020). SARS-CoV-2 specific antibody
523 responses in COVID-19 patients. *Medrxiv.* , 2020.03.18.20038059.
- 524 Papenfuss, A.T., Baker, M.L., Feng, Z.P., Tachedjian, M., Cramer, G., Cowled, C., Ng, J., Janardhana,
525 V., Field, H.E., and Wang, L.F. (2012). The immune gene repertoire of an important viral reservoir, the
526 Australian black flying fox. *BMC Genomics.* *13*, 261-2164-13-261.
- 527 Paul, S., Sidney, J., Sette, A., and Peters, B. (2016). TepiTool: A pipeline for computational prediction
528 of T cell epitope candidates. *Curr. Protoc. Immunol.* *114*, 18.19.1-18.19.24.
- 529 Peng, Y., Mentzer, A.J., Liu, G., Yao, X., Yin, Z., Dong, D., Dejnirattisai, W., Rostron, T., Supasa, P.,
530 Liu, C., *et al.* (2020a). Broad and strong memory CD4 (+) and CD8 (+) T cells induced by SARS-CoV-
531 2 in UK convalescent COVID-19 patients. *Biorxiv.*
- 532 Peng, Y., Mentzer, A.J., Liu, G., Yao, X., Yin, Z., Dong, D., Dejnirattisai, W., Rostron, T., Supasa, P.,
533 Liu, C., *et al.* (2020b). Broad and strong memory CD4+ and CD8+ T cells induced by SARS-CoV-2 in
534 UK convalescent COVID-19 patients. *Biorxiv.*
- 535 Petrosillo, N., Viceconte, G., Ergonul, O., Ippolito, G., and Petersen, E. (2020). COVID-19, SARS and
536 MERS: Are they closely related? *Clin. Microbiol. Infect.* *26*, 729-734.
- 537 Poh, C.M., Carissimo, G., Wang, B., Amrun, S.N., Lee, C.Y., Chee, R.S., Fong, S.W., Yeo, N.K., Lee,
538 W.H., Torres-Ruesta, A., *et al.* (2020). Two linear epitopes on the SARS-CoV-2 spike protein that elicit
539 neutralising antibodies in COVID-19 patients. *Nat. Commun.* *11*, 2806-020-16638-2.

540 Rockx, B., Donaldson, E., Frieman, M., Sheahan, T., Corti, D., Lanzavecchia, A., and Baric, R.S.
541 (2010). Escape from human monoclonal antibody neutralization affects in vitro and in vivo fitness of
542 severe acute respiratory syndrome coronavirus. *J. Infect. Dis.* *201*, 946-955.

543 Sanjuán, Rafael AND Nebot, Miguel R. AND Peris, Joan B. AND Alcamí, José. (2013). Immune
544 activation promotes evolutionary conservation of T-cell epitopes in HIV-1. *PLOS Biology.* *11*, 1-10.

545 Schneidewind, A., Brockman, M.A., Sidney, J., Wang, Y.E., Chen, H., Suscovich, T.J., Li, B., Adam,
546 R.I., Allgaier, R.L., Mothé, B.R., *et al.* (2008). Structural and functional constraints limit options for
547 cytotoxic T-lymphocyte escape in the immunodominant HLA-B27-restricted epitope in human
548 immunodeficiency virus type 1 capsid. *J. Virol.* *82*, 5594-5605.

549 Schneidewind, A., Brockman, M.A., Yang, R., Adam, R.I., Li, B., Le Gall, S., Rinaldo, C.R., Craggs,
550 S.L., Allgaier, R.L., Power, K.A., *et al.* (2007). Escape from the dominant HLA-B27-restricted
551 cytotoxic T-lymphocyte response in gag is associated with a dramatic reduction in human
552 immunodeficiency virus type 1 replication. *J. Virol.* *81*, 12382-12393.

553 Shi, Z., and Wang, L.F. (2017). Evolution of SARS coronavirus and the relevance of modern molecular
554 epidemiology. *Genetics and Evolution of Infectious Diseases.* , 601-619.

555 Sironi, M., Hasnain, S.E., Rosenthal, B., Phan, T., Luciani, F., Shaw, M.A., Sallum, M.A., Mirhashemi,
556 M.E., Morand, S., González-Candelas, F., *et al.* (2020). SARS-CoV-2 and COVID-19: A genetic,
557 epidemiological, and evolutionary perspective. *Infect. Genet. Evol.* *84*, 104384.

558 St John, A.L., and Rathore, A.P.S. (2020). Early insights into immune responses during COVID-19. *J.*
559 *Immunol.*

560 Su, Y.C.F., Bahl, J., Joseph, U., Butt, K.M., Peck, H.A., Koay, E.S.C., Oon, L.L.E., Barr, I.G.,
561 Vijaykrishna, D., and Smith, G.J.D. (2015). Phylodynamics of H1N1/2009 influenza reveals the
562 transition from host adaptation to immune-driven selection. *Nat. Commun.* *6*, 7952.

563 Tse, L.V., Klinc, K.A., Madigan, V.J., Castellanos Rivera, R.M., Wells, L.F., Havlik, L.P., Smith, J.K.,

- 564 Agbandje-McKenna, M., and Asokan, A. (2017). Structure-guided evolution of antigenically distinct
565 adeno-associated virus variants for immune evasion. *Proceedings of the National Academy of Sciences*.
566 *114*, E4812-E4821.
- 567 Vabret, N., Britton, G.J., Gruber, C., Hegde, S., Kim, J., Kuksin, M., Levantovsky, R., Malle, L.,
568 Moreira, A., Park, M.D., *et al.* (2020). Immunology of COVID-19: Current state of the science.
569 *Immunity*. *52*, 910-941.
- 570 van Dorp, L., Acman, M., Richard, D., Shaw, L.P., Ford, C.E., Ormond, L., Owen, C.J., Pang, J., Tan,
571 C.C.S., Boshier, F.A.T., *et al.* (2020). Emergence of genomic diversity and recurrent mutations in
572 SARS-CoV-2. *Infect. Genet. Evol.* *83*, 104351.
- 573 Vita, R., Mahajan, S., Overton, J.A., Dhanda, S.K., Martini, S., Cantrell, J.R., Wheeler, D.K., Sette, A.,
574 and Peters, B. (2019). The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res.* *47*,
575 D339-D343.
- 576 Wang, M., Yan, M., Xu, H., Liang, W., Kan, B., Zheng, B., Chen, H., Zheng, H., Xu, Y., Zhang, E., *et*
577 *al.* (2005). SARS-CoV infection in a restaurant from palm civet. *Emerg. Infect. Dis.* *11*, 1860-1865.
- 578 Wiehe, K., Easterhoff, D., Luo, K., Nicely, N.I., Bradley, T., Jaeger, F.H., Dennison, S.M., Zhang, R.,
579 Lloyd, K.E., Stolarchuk, C., *et al.* (2014). Antibody light-chain-restricted recognition of the site of
580 immune pressure in the RV144 HIV-1 vaccine trial is phylogenetically conserved. *Immunity*. *41*, 909-
581 918.
- 582 Wong, M.C., Javornik Cregeen, S.J., Ajami, N.J., and Petrosino, J.F. (2020). Evidence of recombination
583 in coronaviruses implicating pangolin origins of nCoV-2019. *Biorxiv.* , 2020.02.07.939207.
- 584 Woo, P.C., Lau, S.K., Tsoi, H.W., Huang, Y., Poon, R.W., Chu, C.M., Lee, R.A., Luk, W.K., Wong,
585 G.K., Wong, B.H., *et al.* (2005). Clinical and molecular epidemiological features of coronavirus HKU1-
586 associated community-acquired pneumonia. *J. Infect. Dis.* *192*, 1898-1907.
- 587 Wrapp, D., Wang, N., Corbett, K.S., Goldsmith, J.A., Hsieh, C.L., Abiona, O., Graham, B.S., and

- 588 McLellan, J.S. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation.
589 *Science*. 367, 1260-1263.
- 590 Wu, F., Wang, A., Liu, M., Wang, Q., Chen, J., Xia, S., Ling, Y., Zhang, Y., Xun, J., Lu, L., *et al.*
591 (2020). Neutralizing antibody responses to SARS-CoV-2 in a COVID-19 recovered patient cohort and
592 their implications. *Medrxiv*.
- 593 Wu, Z., and McGoogan, J.M. (2020). Characteristics of and important lessons from the coronavirus
594 disease 2019 (COVID-19) outbreak in china: Summary of a report of 72□314 cases from the chinese
595 center for disease control and prevention. *Jama*.
- 596 Wynne, J.W., Woon, A.P., Dudek, N.L., Croft, N.P., Ng, J.H., Baker, M.L., Wang, L.F., and Purcell,
597 A.W. (2016). Characterization of the antigen processing machinery and endogenous peptide
598 presentation of a bat MHC class I molecule. *J. Immunol.* 196, 4468-4476.
- 599 Xiao, K., Zhai, J., Feng, Y., Zhou, N., Zhang, X., Zou, J., Li, N., Guo, Y., Li, X., Shen, X., *et al.* (2020).
600 Isolation and characterization of 2019-nCoV-like coronavirus from malayan pangolins. *Biorxiv.* ,
601 2020.02.17.951335.
- 602 Ye, Z.W., Yuan, S., Yuen, K.S., Fung, S.Y., Chan, C.P., and Jin, D.Y. (2020). Zoonotic origins of human
603 coronaviruses. *Int. J. Biol. Sci.* 16, 1686-1697.
- 604 Yurkovetskiy, L., Pascal, K.E., Tompkins-Tinch, C., Nyalile, T., Wang, Y., Baum, A., Diehl, W.E.,
605 Dauphin, A., Carbone, C., Veinotte, K., *et al.* (2020). SARS-CoV-2 spike protein variant D614G
606 increases infectivity and retains sensitivity to antibodies that target the receptor binding domain.
607 *Biorxiv*.
- 608 Zhang, B., Zhou, X., Zhu, C., Feng, F., Qiu, Y., Feng, J., Jia, Q., Song, Q., Zhu, B., and Wang, J.
609 (2020a). Immune phenotyping based on neutrophil-to-lymphocyte ratio and IgG predicts disease
610 severity and outcome for patients with COVID-19. *Medrxiv*.
- 611 Zhang, L., Jackson, C.B., Mou, H., Ojha, A., Rangarajan, E.S., Izard, T., Farzan, M., and Choe, H.

- 612 (2020b). The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases
613 infectivity. Biorxiv.
- 614 Zhao, J., Yuan, Q., Wang, H., Liu, W., Liao, X., Su, Y., Wang, X., Yuan, J., Li, T., Li, J., *et al.* (2020).
615 Antibody responses to SARS-CoV-2 in patients of novel coronavirus disease 2019. *Clinical Infectious*
616 *Diseases*.
- 617 Zhou, P., Yang, X.L., Wang, X.G, Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L.,
618 *et al.* (2020a). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*.
619 *579*, 270-273.
- 620 Zhou, P., Yang, X., Wang, X., Hu, B., Zhang, L., Zhang, W., Si, H., Zhu, Y., Li, B., Huang, C., *et al.*
621 (2020b). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. *579*,
622 270-273.

623

624 **Figure Legends**

625 **Figure 1. Amino acid variability of the SARS-CoV-2 spike protein.** Shannon's entropy (H) values
626 for each amino acid position calculated using 10000 SARS-CoV-2 spike proteins are shown. B cell
627 predicted epitopes and T cell predicted epitopes are also reported in blue and green, respectively. B cell
628 epitopes identified in the sera of COVID-19 patients (Farrera et al., 2020; Poh et al., 2020) are also
629 reported in red.

630

631 **Figure 2. Variability of epitope and non-epitope positions among SARS-CoV-2 proteins.**

632 Shannon's entropy (H) mean values along with standard errors are shown for all SARS-CoV-2 proteins
633 longer than 60 residues. Epitope positions are shown in dark gray, non-epitopes in light gray.
634 Significant comparisons, calculated by a permutation approach, are indicated with asterisks (*, $P <$
635 0.05; **, $P <$ 0.01; ***, $P <$ 0.001). Immunogenic proteins are shown in blue and the length of each
636 protein is reported in the bottom panel.

637

638 **Figure 3. Variability of epitope and non-epitope positions among sarbecoviruses.** Shannon's
639 entropy (H) mean values along with standard errors are shown for a set of sarbecovirus ORFs. SARS-
640 CoV-2 epitope positions are shown in dark gray, non-epitopes in light gray. Significant comparisons,
641 calculated by a permutation approach, are indicated with asterisks (*, $P <$ 0.05; **, $P <$ 0.01; ***, $P <$
642 0.001).

643

644

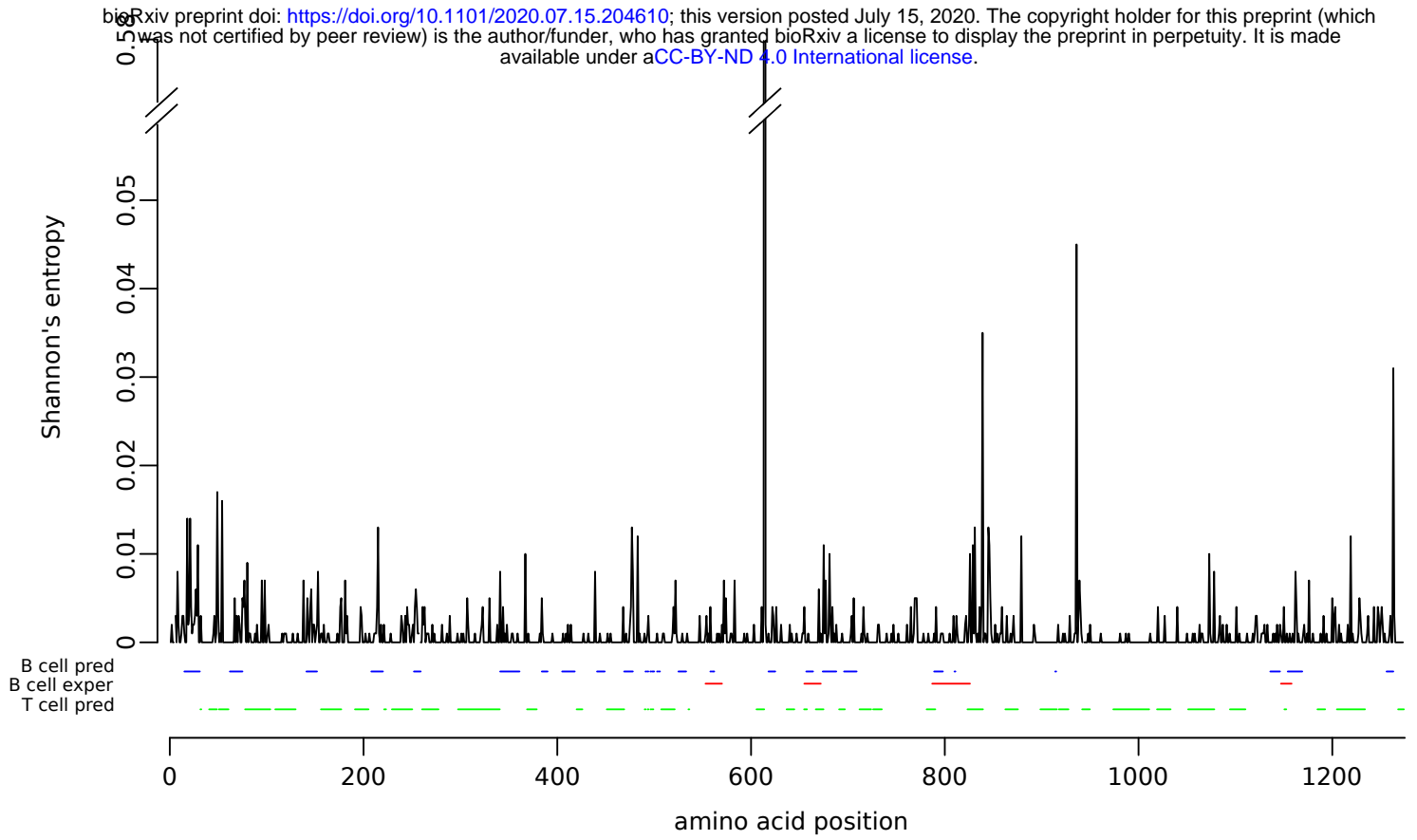
645 **Figure 4. Variability of epitope and non-epitope positions among human coronaviruses.** Shannon's
646 entropy (H) mean values along with standard errors are shown for human coronavirus spike and

647 nucleocapsid proteins. Epitope positions are shown in dark gray, non-epitopes in light gray. Significant
648 comparisons, calculated by a permutation approach, are indicated with asterisks (*, $P < 0.05$; **, $P <$
649 0.01 ; ***, $P < 0.001$).

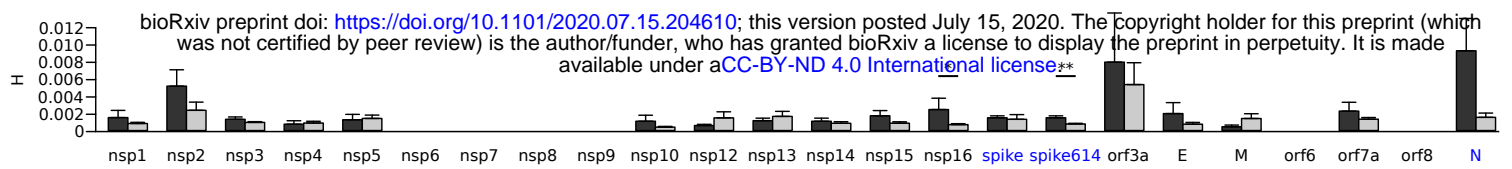
650

651

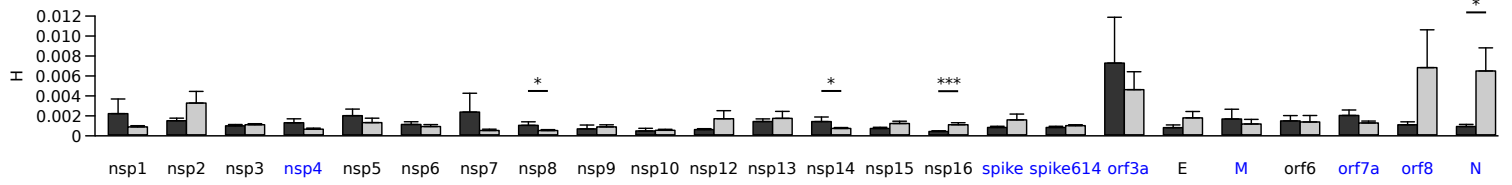
652



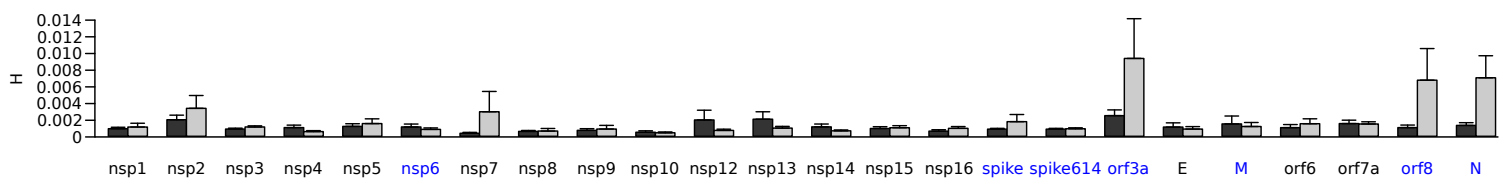
linear B cell epitope



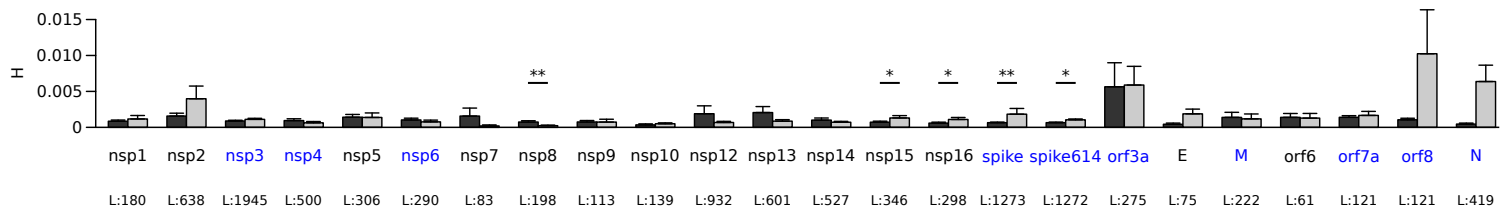
CD4⁺ T cell epitopes

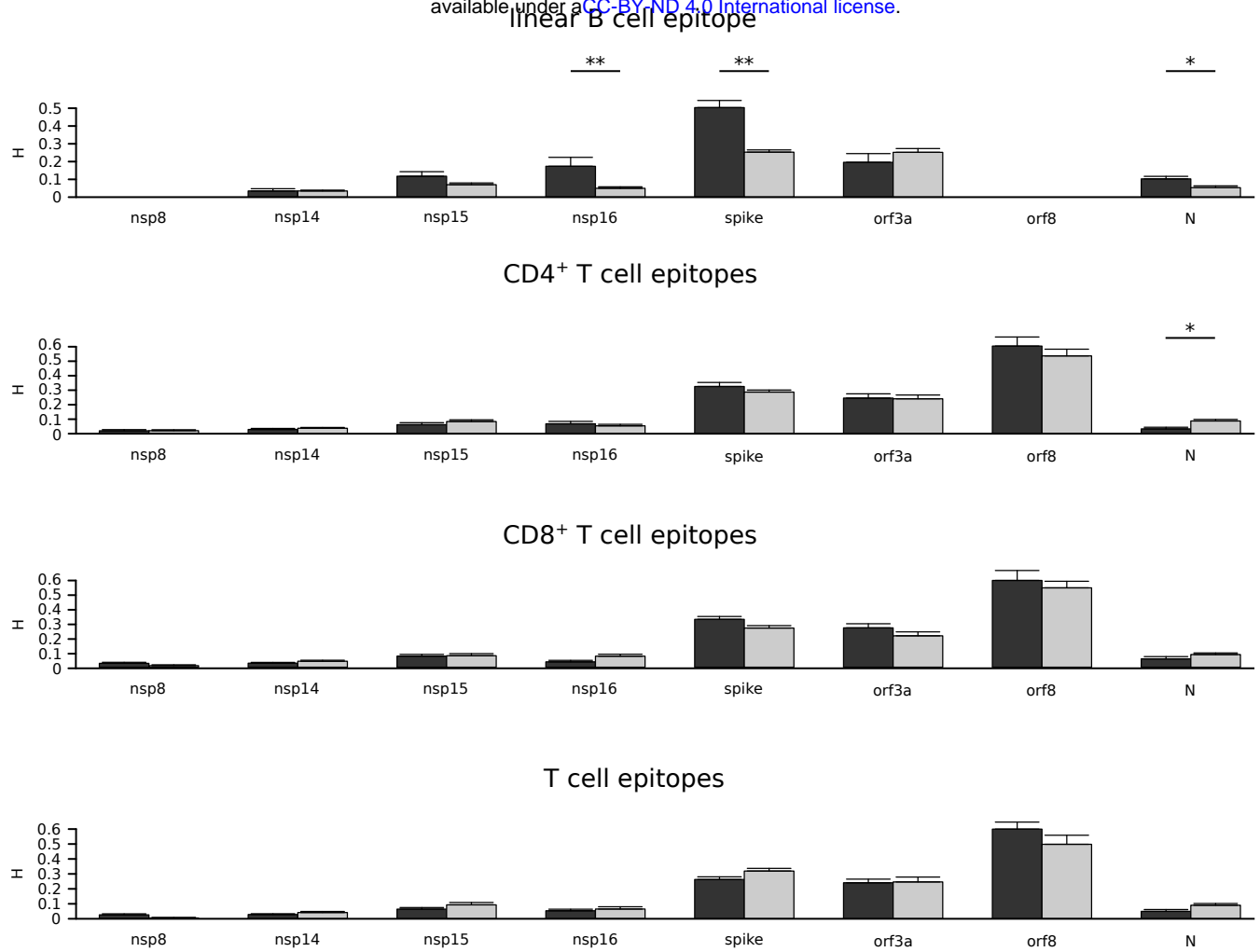


CD8⁺ T cell epitopes

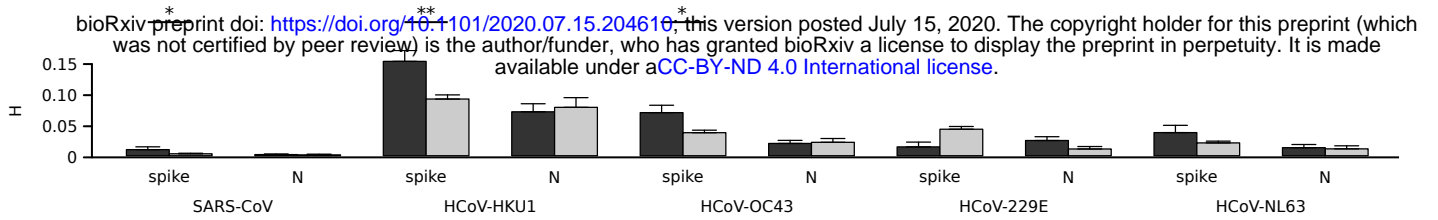


T cell epitopes

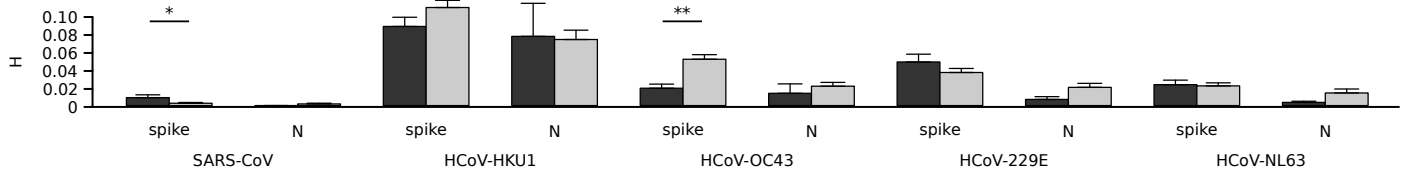




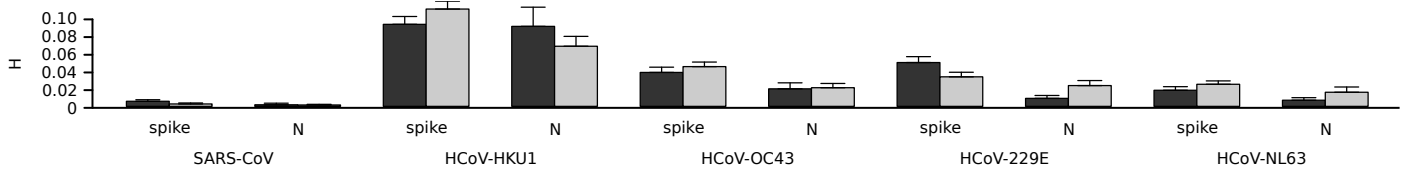
linear B cell epitope



CD4⁺ T cell epitopes



CD8⁺ T cell epitopes



T cell epitopes

