Predicted Impact of the Viral Mutational Landscape on the Cytotoxic Response against SARS-CoV-2

Anna Foix^a, Daniel López^b, Michael J. McConnell^c and Antonio J. Martín-Galiano^{c*}

4

^a European Bioinformatic Institute, European Molecular Biology Laboratory. Hinxton. United Kingdom. UK.

^b Presentation and Immune Regulation Unit, Centro Nacional de Microbiología, Instituto de Salud Carlos III. Majadahonda 28220. Spain.

^c Laboratory of Intrahospital Infections, Centro Nacional de Microbiología, Instituto de Salud Carlos III. Majadahonda 28220. Spain.

11

*Corresponding author: mgaliano@isci.es

13

ABSTRACT

The massive assessment of immune evasion due to viral mutations that potentially increase COVID-19 susceptibility can be computationally facilitated. The adaptive cytotoxic T response is critical during primary infection and the generation of long-term protection. Potential epitopes in the SARS-CoV-2 proteome were predicted for 2,915 human alleles of 71 HLA class I families. Allele families showed extreme differences in number of recognized epitopes, underscoring genetic variability of protective capacity between humans. Up to 1,222 epitopes were associated with any of the twelve supertypes, that is, allele clusters covering 90% population. Among them, the B27 supertype showed the lowest number of epitopes. Epitope escape mutations identified in ~118,000 NCBI isolates mainly involved non-conservative substitutions at the second and C-terminal position of the ligand core, or total ligand removal by large recurrent deletions. Escape mutations affected 47% of supertype epitopes, which in 21% of cases concerned isolates from two or more sub-continental areas. Some of these changes were coupled, but never surpassed 15% evaded epitopes for the same supertype in the same isolate, except for B27, which reached up to 33%. In contrast to most supertypes, eight particular allele families mostly contained alleles with few SARS-CoV-2 ligands. Isolates harboring cytotoxic escape mutations for these families co-existed geographically within sub-Saharan and Asian populations enriched in these alleles. Collectively, these data indicate that independent escape mutation events have already occurred for half of HLA class I supertype epitopes. However, it is presently unlikely that, overall, it poses a threat to the global population. In contrast, single and double mutations for susceptible alleles may be associated with viral selective pressure and alarming local outbreaks. This study highlights the automated integration of genomic, geographical and immunoinformatic information for surveillance of SARS-CoV-2 variants potentially affecting the population as a whole, as well as minority subpopulations.

40

AUTHOR SUMMARY

The cytotoxic T response, a type of immune response dependent upon an individual's genetics that does not require antibodies, is critical for neutralizing SARS-CoV-2 infection. The potential bypass of the cytotoxic T response by mutations acquired by the virus after one year of the pandemic is therefore of maximal concern. We have approached the complexity of human variability and more than 100.000 viral genomes in this respect using a computational strategy. We have detected numerous mutations in these genomes that mask some viral regions involved in the cytotoxic response. However, the accumulation of these changes in independent isolates is still too low to threaten the global human population. In contrast, our protocol has identified mutations that may be relevant for specific populations and minorities with cytotoxic genetic backgrounds susceptible to SARS-CoV-2 infection. Some viral variants co-existed in the same country with these human communities which warrants deeper surveillance in these cases to prevent local outbreaks. Our study support the integration of massive data of different natures in the surveillance of viral pandemics.

55

56 **Introduction**

57 Mutations in the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)
58 leading to increased susceptibility are of extreme concern. Given the slow pace of vaccination in
59 some geographic regions, enhanced primary infection by strains that evade immune detection
60 might worsen the significant health and socioeconomic burden caused by the COVID-19
61 pandemic.

62 Long term protection from viral infection relies on a competent adaptive response.
63 Adaptive protection includes the coordinated activation and memory of three adaptive response
64 compartments. These branches consist of the humoral response, driven by antibodies
65 synthesized by B cells, and the two types of cellular response, driven by CD8⁺ and CD4⁺
66 lymphocytes that recognize viral peptides bound to human leukocyte antigen (HLA) class I and
67 II molecules, respectively (1). Antibodies neutralize the virus, for example, by specific binding
68 to the spike protein and inhibit binding to the ACE2 receptor expressed in the lung (2). CD8⁺, or
69 cytotoxic, T lymphocytes directly kill SARS-CoV-2 infected cells through the secretion of pore-
70 forming proteases and the induction of programmed cell death (3). Finally, CD4⁺, T helper
71 lymphocytes, play a pivotal stimulatory role for both antibody-led and cytotoxic activities. An
72 effective cellular response is associated with prompt and efficient protection during primary and
73 successive SARS infections (4–6). Moreover, the cellular and humoral responses are long-
74 lasting (7) and elicit immunoprotective memory (8,9).

75 Naïve cytotoxic lymphocytes are stimulated through the presentation of specific
76 proteolyzed fragments of antigens, or epitopes, bound to HLA class I molecules in the
77 membrane of infected cells. Some peptides of approximately nine residues are generated by
78 cleavage of intracellular pathogen proteins and bound in the endoplasmic reticulum to the apical
79 antigenic groove of the monomeric chain of HLA class I. Once in the membrane, these ligands
80 can be recognized by the T-cell receptor of T CD8⁺ lymphocytes that start the maturation
81 process. After activation and division of a sufficient subsets of mature CD8⁺ T cells, the subject
82 may be protected against the particular virus as SARS-CoV-2 at a cellular level (3).

83 Severe COVID-19 outcomes have been associated with aging and co-morbidities such
84 as hypertension and diabetes mellitus (10). However, how host genetic factors influence the
85 disease is still largely unknown. In this respect, the adaptive cellular response is strongly
86 influenced by host genetics. Notably, HLA class I genes are among the most multiple and
87 variable genes in humans. The HLA class I system consists of three loci for which over 17,000
88 alleles have been reported (11). These alleles are further grouped into phylogenetic families and,
89 some of them, into supertypes that shared comparable ligands (12). Overall, this huge allelic
90 diversity provides the human species with an enormous capacity to detect different antigens
91 from virtually any pathogen.

92 HLA class I epitope pool screenings with SARS-CoV-2 sequences have been carried
93 out for specific countries (13–15). This kind of experimental information is stored in
94 repositories like the Immune Epitope Database and Analysis Resource (IEDB) (16), which
95 allows for the global analysis of potential mutation-evasion events. Certain HLA alleles have
96 been associated with permissiveness to SARS-CoV-2 infection, such as B*44 and C*01 families
97 in Italy (17), and B*15:27 and C*07:29 in China (18). However, these data still do not evenly
98 represent genetic differences in susceptibility at the global population level. Since analyzing
99 thousands of HLA alleles is experimentally unrealistic (19), the confines of the human SARS-
100 CoV-2 cytotoxic ligandome can be explored by *bona fide* computation approaches in a neutral
101 manner. For instance, Nguyen *et al.* identified HLA-B*46:01 as the less efficient allele for
102 presenting SARS-CoV-2 epitopes among 145 alleles by using an immunoinformatic approach
103 (20).

104 Widespread infection with SARS-CoV-2 at the global level provides the virus with
105 great opportunity to explore the mutational space. Some of these changes may be selected based
106 on immunological evasive advantages. In this respect, how the genetic variability of the virus
107 can affect individuals carrying different HLA class I alleles is currently unknown. Viral
108 mutations that dramatically decrease binding affinity of epitopes to HLA class I molecules can
109 act as escape mutants and alter the cellular immunity, with important implications for clinical
110 evolution of the infection (21). How many and which mutations an isolate must acquire in order

111 to evade the adaptive cytotoxic responses of the general population remains an open question.
112 Such emerging capacity to bypass the cytotoxic response would presumably not follow a
113 categorical binary pattern but a gradient dependant on underlying individual genetics.

114 The goal of this study is two-fold. First, we have interrogated how SARS-CoV-2
115 mutations can affect predicted HLA class I binding at the global population level, and second,
116 how existing mutations influence the response of specific sub-populations that harbor alleles
117 with few SARS-CoV-2 epitopes. For that, we have taken advantage of the strength of
118 computational methods to design a protocol that generates and operates on formatted data. This
119 allowed us to conduct an exhaustive analysis that involved over 2,900 human HLA class I
120 alleles and ~118,000 viral genomes. The knowledge acquired here may help to understand the
121 current status of the human cytotoxic defense in the context of the pandemic and to promptly
122 identify emerging strains that require close monitoring.

123

124 **Results**

125 **Predicted vs validated SARS-CoV-2 HLA class I epitopes**

126 To obtain a more complete insight of the cellular cytotoxic response to SARS-CoV-2, HLA
127 class I epitopes from the SARS-CoV-2 reference proteome were predicted by the universal
128 netMHCpan 4.1 EL algorithm (22). These included all medium-strong peptide binders for 2,915
129 human alleles grouped into 71 families of the HLA-A (21 families, 886 alleles), HLA-B (36
130 families, 1412 alleles), HLA-C (14 families, 617 alleles) loci available in this software. The
131 predicted full SARS-CoV-2 ligandome for HLA class I reached 5,224 independent epitopes.

132 Data complexity reduction by clustering alleles into families can cause some
133 information loss. However, the degree of intra-family coherence, that is, the percentage of
134 matching epitopes between two alleles of the same family with respect to the total predicted by
135 both alleles, reached $61.3 \pm 19.2\%$ (mean $\pm$ SD). In contrast, the inter-family coherence, that is,
136 the average matching epitopes after all-against-all family comparison), was only $3.0 \pm 1.6\%$.

137 This supports that, despite the existence of intra-family differences, the allele family cluster
138 stratum is acceptable for a global view of HLA epitopes.

139 Families showed a drastic difference in the number of predicted epitopes (Fig. 1A).
140 Globally, families of A and C loci showed higher values than B loci families. In particular,
141 A*01, A*23, A*24, C*12 and C*14 families surpassed 300 epitopes on average, whereas B*46,
142 B*82 and B*83 were below five. Twenty-six alleles, several from the B*46 family, were not
143 associated with any predicted epitope. These computational predictions are in line with
144 antecedent observations concerning great differences between HLA class I alleles in the
145 response to the SARS-CoV-2 reference strain (18,20). Some families were linked to exclusive
146 epitope pools but others shared overlapping SARS-CoV-2 ligandomes (Fig. 1B).

147 All viral proteins theoretically generated HLA class I epitopes. On average, 1.19
148 epitopes per allele family (those identified for $\geq 50\%$ alleles in the family) and 100 residues
149 were identified in the SARS-CoV-2 proteome. Among polypeptides with ≥ 75 residues, the M
150 protein carried higher (1.41 epitopes per family and 100 residues) and N lower (0.71) epitope
151 densities, respectively.

152 Epitope predictions were compared to 760 experimentally validated 8-12mer epitopes
153 for HLA class I included in the IEDB dataset. Up to 90% of validated epitopes perfectly
154 matched predicted epitopes, for at least one allele, at the stringent thresholds applied. There was
155 high correlation ($r^2 = 0.87$, polynomial fit) between the number of predicted and validated
156 epitopes for the allele (Fig. 2A). However, several alleles from the A*02 family were
157 comparatively over-represented while B*27 and B*39 alleles were under-represented in the
158 validated dataset. Differences for sequence and number of validated ligand datasets were
159 evident between whole families (Fig. 2B). Globally, the ratio between validated and predicted
160 epitopes was significantly higher for families of the A locus (0.35 ± 0.12) with respect to those
161 of the B (0.31 ± 0.13 , $P < 0.001$ Student's t-test) and the C (0.29 ± 0.10 , $P < 0.001$) loci. Forty-
162 four alleles from 13 families did not show any validated experimental epitope, a higher number
163 than allele and families without predicted epitopes. Despite invaluable studies that contributed
164 data with relatively large and distributed datasets (23), experimental screenings may be slightly

165 biased by the low frequencies of some alleles in the cohorts analyzed. Overall, computational
166 and experimental approaches may be complementary and beneficial for the global
167 characterization of the SARS-CoV-2 cytotoxic response.

168

169 **Supertypes show very different number of SARS-CoV-2 supermotifs**

170 Alleles, from the same or different families, that bind similar epitopes are functionally grouped
171 into twelve, so-called, supertypes (12). Supertypes cover >90% of the world population
172 regardless of ethnicity. In our dataset, 1,222 (23.4%) of all non-redundant epitopes were able to
173 cover $\geq 50\%$ of the alleles associated with at least one supertype, i.e. are supermotifs (24) (Fig.
174 3A, Supplementary Table S2). Moreover, twenty supermotifs covered three or more supertypes
175 (Table 1). On average, 11.1 supermotifs were identified per 100 residues of the viral proteome.
176 Among ORFs with ≥ 75 residues, the M (13.5 supermotifs per 100 residues) and the N (5.0
177 supermotifs per 100 residues) proteins showed the highest and the lowest concentration of
178 supermotifs, respectively. The number of supertypes was unevenly represented since, for
179 instance, the "A01 A24" and "A24" supertypes were associated with >250 supermotifs, while
180 others as "A01" or "B07" with around 50 supermotifs, and the "B27" supertype showed only 12
181 (Fig. 3B).

182

183 **Table 1. SARS-CoV-2 HLA class I supermotifs involving three or more supertypes.**

Protein	Supermotif sequence	% allele supertype coverage												Number of covered supertypes
		A01	A01 A03	A01 A24	A02	A03	A24	B07	B08	B27	B44	B58	B62	
ORF1ab	5533-VVYRGTTTY-5541	76		100		53						72	62	5
ORF1ab	1582-QVVDMSMTY-1590	78		100		52							62	4
ORF1ab	2273-STNVTIATY-2281	80		100								67	58	4
ORF1ab	4072-VVIPDYNTY-4080	82		100								67	62	4
ORF1ab	4673-KLFDYFKY-4681	50		100		67							54	4
ORF1ab	6154-HSIGFDYVY-6162	74		100								61	50	4
ORF1ab	77-RTAPGHVM-85		60									89	60	3
ORF1ab	110-HVGEIPVAY-118	72		100									60	3
ORF1ab	568-TILDGISQY-576	74		100									60	3
ORF1ab	906-YLFDESCEF-914			78	70								68	3
ORF1ab	1768-VMYMGTLTY-1776	52		100									62	3

ORFlab	1806-MMSAPPAQY-1814	62	100			62	3
ORFlab	2876-TTNGDFLHF-2884	58	78			78	3
ORFlab	2960-SIIQFPNTY-2968	74	100			60	3
ORFlab	3103-GVYSVIYLY-3111	70	100	61			3
ORFlab	4533-TLKEILVTY-4541	58	100			60	3
ORFlab	5267-QEYADVFLHY-5276		100	74		57	3
ORFlab	5981-SMMGFKMNY-5989		100	64		60	3
S	192-FVFKNIDGY-200	72	100			52	3
S	269-YLQPRTFLL-277			88	74	87	3

184

185 Recurrent mutations can affect HLA class I epitopes

186 Calculations performed up to this point have included only the original Wu-han-1 reference
187 strain. However, viruses are continuously evolving entities where HLA class I ligand
188 recognition can be dynamically subjected to extensive mutation-selection processes. Key
189 mutations could produce cytotoxic escape variants by reducing affinity or even deleting HLA
190 class I ligands and, then, influencing the ability of CD8+ lymphocytes to clear the infection
191 (25). To assess this possibility, mutations were identified in 117,811 SARS-CoV-2 isolates from
192 87 countries covering 21 out of the 22 sub-continental areas of the United Nations M49
193 geoscheme (<https://unstats.un.org/unsd/methodology/m49/>). A total of 1,128,631 genetic
194 alterations with respect to the Wu-han-1 reference strain were identified. These involved 28,512
195 unique residue substitutions in 9,723 positions. A total of 78% of unique substitutions were non-
196 conservative, those concerning distinct amino acid classes (Fig. 4A, left). Up to 26,231
197 deletions of 1 to 193 residues and 127 insertions, ranging between 1 and 8 residues, were also
198 detected (Fig. 4B). Substitutions were much more prevalent than deletions (Fig. 4C). In
199 contrast, deletions affected a higher number of epitopes (Fig. 4C). Nevertheless, the
200 degeneration of ligand binding is expected to counteract many substitutions and point deletions
201 affecting epitopes, leading to a negligible effect on the cytotoxic response.

202

203 **There are mutations for most supermotifs but only a fraction causes epitope escape**
204 **and are geographically distributed**

205 All 1,224 supermotifs carried some type of mutation in least one isolate. Nevertheless, a central
206 question is to what extend these changes have high impact in the context of the cytotoxic
207 response of the worldwide population.

208 Mutations were scrutinized using three criteria: recurrence, binding affinity reduction
209 and geographical dissemination of isolates carrying them. For this, a series of incremental
210 selective criteria on all the genetic changes observed was applied: (Filter 1) presence of the
211 mutation in ≥ 2 isolates if the mutation was a substitution, or ≥ 5 isolates if the mutation was an
212 insertion or a deletion, since these are more be resultant of sequencing errors; (Filter 2) drastic
213 reduction of supermotif binding. Changes in the second (P2) and C-terminal (P9 in core
214 nonamers) positions preferentially perturb binding affinity, in particular when these changes
215 involve amino acids of different physicochemical classes (non-conservative) (Fig 5A). This can
216 be explained by residues in these positions intimately interact with the selective B and F pockets
217 in the groove of the HLA molecule. However, the epitope disturbing capacity of mutations in
218 these positions is not an exact rule (Fig. 5A). Thus, the actual impact on binding was explicitly
219 recalculated in the mutated sequence and quantified using recommended thresholds (see
220 Materials and Methods); and (Filter 3) detection of the mutation in isolates from different M49
221 world regions.

222 Expectedly, the fraction of escape supermotifs substantially decreased as more selective
223 criteria were applied (Fig. 5B). Only 22.1% of supermotifs contained mutations that satisfied all
224 the stringent criteria, that is, show recurrent mutations that cancel the HLA class I binding and
225 are found in isolates over several sub-continental zones. Such high-impact changes affected
226 differently to various supertypes. The A03 supertype was more affected with 36.7% while only
227 3.3% of B08 supertype supermotifs showed escape and disseminated mutations (Fig. 5C, Table
228 S2). The "Australia and New Zealand" and, in particular, "Northern America" M49 zones
229 presented isolates with mutations in a substantial part of supermotifs (Fig. 5D).

230 Recurrent mutations may have great relevance if they affect more than one supertype.
231 Thirty substitutions that disabled binding affinity of universal supermotifs appeared in 70 or
232 more isolates (Table 2, Supplementary Table S3). Among them, was the W152C change in the

spike protein in 3,455 USA isolates, which removed four supermotifs of two supertypes. The W6974V (ORF1ab) substitution found in 211 isolates of four M49 areas, destroyed three supermotifs of two supertypes. Notably, the P5828 and K6980 (ORF1ab) positions showed two different recurrent escape mutations each.

Comparatively, insertions played an almost global negligible effect. Only 6965::SKF and 6981::KEGQ (ORF1ab) decreased the HLA binding, affected one supermotif and supertype each, and were in more than 10 isolates (Table 3).

Deletions showed a binary pattern (Table 4 and Supplementary Table S4). On the one hand, short deletions (≤ 3 residues) affected many proteome zones and mostly concerned a single supermotif. For instance, the predominant $\Delta 145$ (S protein), $\Delta 363$ (N protein) and $\Delta 141$ -143 (ORF1ab) deletions were in this group. On the other hand, long recurrent deletions (>80 residues) removed up to twelve supermotifs from seven supertypes and tended to occur in discrete proteome hotspots, namely the 2791-2883 (28-120 of nsp4) and the 6338-6436 (413-511 of nsp14A2_ExonN) regions of ORF1ab.

Table 2. Top-30 recurrent supermotif escape substitutions.

Protein	Mutation	Number of isolates (number of countries)	M49 areas*	Escape supermotif(s)**	Affected supertypes
S	W152C	3455 (1)	Northern America	4	A01 A24, A24
ORF1ab	L3606F	2986 (49)	18	3605-FLYENAF-3612	A02
ORF1ab	P5828L	2147 (6)	6	5827-NPAWRKAVF-5835	B07
ORF1ab	P4619L	1023 (5)	5	4618-TPGSGVPV-4626	B07
ORF1ab	L6102F	797 (7)	7	6100-KNLSDRVVFV-6109	A02
ORF1ab	V3595G	304 (1)	Northern America	3587-ILTSLLVLV-3595	A02
ORF1ab	L6981Q	258 (4)	4	6973-SWNADLYKL-6981	A24
ORF1ab	K2511N	251 (7)	6	2511-KTYERHSL-2519	A01 A03
ORF1ab	K6980G	223 (4)	4	6972-HSWNADLYK-6980	A03
ORF1ab	P5828S	214 (5)	4	5827-NPAWRKAVF-5835	B07
ORF1ab	W6974V	211 (4)	4	3	A24, B58
ORF1ab	S2625F	181 (2)	2	2624-VSLDNVLSTF-2633	B58
ORF3a	K16N	148 (4)	4	2	A03
ORF1ab	S2535L	147 (5)	5	2534-GSLPINVIVF-2543	B58
ORF1ab	K2200N	145 (4)	4	2191-KASMPTTIK-2200	A03
ORF1ab	L642F	124 (3)	3	641-FLRDGWEIV-649	A02
ORF1ab	L2122F	108 (1)	Northern America	2121-TLATHGLAAV-2130	A02

ORF1ab	V627F	104 (5)	3	619-TVYEKLKPV-627	A02
S	K1073N	100 (3)	3	2	A01 A03, A03
ORF1ab	K6464N	94 (6)	6	6456-NVAFNVVNK-6464	A03
S	W152L	91 (6)	5	151-SWMESEFRV-159	A24
ORF1ab	K6980S	83 (2)	2	6972-HSWNADLYK-6980	A03
S	W152R	83 (4)	4	143-VYYHKNNKSW-152	A24
ORF3a	K67N	79 (4)	4	59-ASKIITLKK-67	A03
ORF1ab	K2497N	75 (5)	4	2489-YIVDSVTVK-2497	A03
ORF3a	Y107H	75 (1)	Northern America	2	A01 A24
ORF1ab	P1659S	73 (2)	2	1658-YPQVNGTSL-1667	B07
ORF3a	L52F	73 (2)	2	51-ALLAVFQSA-59	A02
ORF1ab	L446F	71 (3)	3	445-GLNDNLLEIL-454	A02
ORF1ab	L681F	70 (3)	3	680-KLVNKFLAL-688	A02

249

250 * If the number of M49 areas is higher than one, the number is given instead.

251 ** If the number of supermotifs with escape substitutions is higher than one, the number is
252 given instead.

253

254 **Table 3. Recurrent supermotif escape insertions.**

Protein	Insertion	Number of isolates (number of countries)	M49 areas	Escape supermotif(s)	Affected supertypes
ORF1ab	6965::SKF	31 (2)	Northern America, Western Africa	6958-KLALGGSVAI-6967	A02
ORF1ab	6981::KEGQ	24 (1)	Northern America	6973-SWNADLYKL-6981	A24
ORF1ab	6980::EG	9 (1)	Northern America	6972-HSWNADLYK-6980	A03
				6973-SWNADLYKL-6981	A24
ORF1ab	6980::G	9 (1)	Northern America	6972-HSWNADLYK-6980	A03
				6973-SWNADLYKL-6981	A24

255

256 **Table 4. Top-30 recurrent supermotif escape deletions.**

Protein	Deletion location (length)	Number of isolates (number of countries)	M49 areas*	Escape supermotifs**	Affected supertypes***
S	Δ145 (1)	2766 (24)	16	142-GVYYHKNNK-150	A03
ORF1ab	Δ6343-6429 (87)	1695 (2)	2	9	5
N	Δ363 (1)	833 (1)	Northern America	355-KHIDAYKTF-363	A24
ORF1ab	Δ6338-6436 (99)	774 (8)	7	2	A02
ORF1ab	Δ6342-6432 (91)	705 (3)	3	7	3
ORF1ab	Δ6343-6432 (90)	493 (2)	2	6	A02, A24
ORF1ab	Δ6343-6431 (89)	492 (2)	2	6	A02, A24

ORF1ab	Δ141-143 (3)	404 (9)	9	135-SYGADLKSF-143	A24
ORF1ab	Δ6342-6429 (88)	303 (1)	Northern America	9	6
ORF1ab	Δ7014-7096 (83)	298 (4)	4	11	7
ORF1ab	Δ4714 (1)	261 (9)	8	4710-STVFPPTS-4718	A01
ORF1ab	Δ6341-6432 (92)	228 (2)	2	6	2
ORF1ab	Δ6345-6429 (85)	200 (1)	Northern America	8	5
ORF1ab	Δ2797-2877 (81)	173 (2)	2	2	A03, B44
ORF1ab	Δ6345-6428 (84)	164 (2)	2	6	3
ORF1ab	Δ2276-2356 (81)	150 (2)	2	4	4
ORF1ab	Δ6656-6686 (31)	150 (1)	Northern America	6669-AMDEFIER-6677	A01
ORF1ab	Δ3705-3705 (1)	149 (2)	2	3699-TVYDDGARR-3707	A03
ORF1ab	Δ2796-2877 (82)	148 (2)	2	3	A03, B44
ORF1ab	Δ84-85 (2)	129 (8)	7	2	3
ORF1ab	Δ6969-7036 (68)	126 (4)	4	8	4
ORF1ab	Δ2791-2883 (93)	108 (5)	5	4	3
ORF1ab	Δ85 (1)	107 (3)	3	77-RTAPGHVM-85	B62
S	Δ143-144 (2)	98 (3)	3	142-GVYYHKNNK-150	A03
ORF1ab	Δ768-862 (95)	92 (4)	4	12	6
ORF1ab	Δ2797-2876 (80)	92 (2)	2	3	2
ORF1ab	Δ6341-6436 (96)	88 (1)	Northern America	5	4
ORF1ab	Δ6158 (1)	84 (2)	2	6154-HSIGFDYVY-6162	4
ORF1ab	Δ6345-6430 (86)	83 (1)	Northern America	7	4
ORF1ab	Δ6956 (1)	81 (6)	5	6954-FIQQLAL-6961	B08

257

258 * If the number of M49 areas is higher than one, the number is given instead.

259 ** If the number of supermotifs with escape deletions is higher than one, the number is given

260 instead.

261 *** If the number of affected supertypes is higher than two, the number is given instead.

262

263 Only a few supermotif escape mutations coexist in the same isolate

264 Beyond prevalence, these mutations may show distinct combinatorial preferences for
265 simultaneously co-occurring in the same isolates. This information was utilized to detect nine
266 independent mutation networks of 2-44 mutations. Several supermotif mutations were linked
267 through a few spread mutations acting as hubs: W152C (S protein), L3606F (ORF1ab, L37 in
268 nsp6_TM) and four long recurrent deletions in the 6342-6432 range (ORF1ab, nsp14A2_ExonN
269 protein)(Fig. 6).

270 Mutated lineages were also analyzed at the isolate level. Ultimately, isolates enriched in
271 supermotif escape mutations may evade immune system response and disseminate quickly.
272 There was a direct relationship between the number of substitutions in an isolate and the number
273 of supertype alleles altered, which may be mostly attributed to neutral RNA replication errors.
274 Importantly, 7347 isolates conveyed mutations in ≥ 5 supermotifs (Table 5 and Supplementary
275 Table S5) and in 1027 cases affected ≥ 5 superotypes.

276 The origin of most supertype-mutated isolates were USA and Australia (Fig. 7A).
277 Among emergent isolates, a strongly mutated isolate (Assembly database entry: "MT577016",
278 297 mutations) from India stood out with 18 escape supermotifs corresponding to 7 superotypes.
279 Notably, 16.4% of isolates, mostly showing only moderate mutational profiles (≥ 5
280 substitutions), presented ≥ 0.5 negated supermotifs per mutation. This feature suggests potential
281 pressure for cytotoxic evasion by precise supermotif mutation in a subpopulation of isolates in
282 the later cases. For instance, the MW586153 isolate collected in USA:SC showed 16
283 substitutions where twelve of them removed supermotifs from four superotypes. Other
284 remarkable cases were three isolates (MW702787, MW702788 and MW702806) from the same
285 county in USA:CA that showed eleven escape supermotifs from seven superotypes with only
286 fourteen substitutions, suggesting an incipient evasive lineage.

287 However, no isolates carried escape mutations for $>15\%$ supermotifs of specific
288 superotypes (Fig. 7B). The only exceptions were three USA isolates with <25 mutations
289 (including deletions) that invalidated four out of the twelve B*27 supermotifs. Seventeen USA
290 isolates with ≤ 20 substitutions and no indels invalidated up to three B*27 supermotifs.

291

292 **Table 5. Top isolates showing supermotif escape mutations.**

Accession	Total number of mutations	Number of escape supermotifs	Number of affected superotypes	Ratio escape supermotifs per mutation	Collection date	Country
MW673525	40	19	5	0.475	08/02/2021	USA
MT577016	297	18	7	0.061	2020	India
MW694016	31	15	9	0.484	11/02/2021	USA
MT451283	276	14	8	0.051	24/03/2020	Australia

MW156473	51	14	7	0.275	28/07/2020	Australia
MT451279	142	13	8	0.092	24/03/2020	Australia
MT451436	134	13	8	0.097	26/03/2020	Australia
MW689154	62	13	7	0.210	13/02/2021	USA
MW653643	21	13	6	0.619	07/12/2020	USA
MW525102	27	13	4	0.481	10/01/2021	USA
MW406716	58	12	9	0.207	24/06/2020	USA
MW190139	49	12	8	0.245	16/07/2020	USA
MW228176	68	12	8	0.176	16/06/2020	USA
MW406699	62	12	8	0.194	24/06/2020	USA
MW542158	75	12	8	0.160	13/01/2021	USA
MW725850	27	12	8	0.444	24/02/2021	USA
MW449384	63	12	7	0.190	30/11/2020	USA
MW474268	54	12	7	0.222	12/11/2020	USA
MW617514	62	12	7	0.194	02/02/2021	USA
MW704295	50	12	7	0.240	19/01/2021	Bahrain
MW715548	71	12	7	0.169	19/02/2021	USA
MW673420	38	12	6	0.316	09/02/2021	USA
MW741583	26	12	6	0.462	23/02/2021	USA
MW751588	30	12	6	0.400	04/03/2021	USA
MW783199	68	12	6	0.176	02/03/2021	USA
MW518131	34	12	5	0.353	03/01/2021	USA
MW693032	18	12	5	0.667	05/11/2020	USA
MW586153	16	12	4	0.750	28/01/2021	USA
MW596067	32	12	4	0.375	29/01/2021	USA
MW673042	33	12	3	0.364	07/02/2021	USA

293

294 **Epitope escape mutations in families with scarce SARS-CoV-2 ligandomes**

295 In contrast to most supertypes, some alleles did not shown affinity to any SARS-CoV-2 peptide
296 or showed scarce SARS-CoV-2 peptide repertoires. In this light, 246 alleles (8.4%) of the three
297 loci (HLA-A: 39 alleles; HLA-B: 143 alleles; HLA-C: 64 alleles) were predicted to bind with
298 high affinity to twenty or less epitopes. These alleles belonged to 48 families which showed
299 three possible patterns depending on their alleles with few SARS-CoV-2 epitopes was either the
300 norm or the exception (Fig. 8A). Firstly, eight families contained $\geq 81\%$ of alleles with few
301 predicted epitopes and ≤ 22 epitopes per allele on average, and were deemed poor SARS-CoV-
302 2-repertoire families. These families were A*74, B*46, B*52, B*73, B*82, B*83, C*01 and
303 C*18. Remarkably, the combination of alleles of inefficient families for the three loci, the
304 A*74:02-B*46:01-C*01:02 haplotype, has been detected with a 0.02% frequency in a Hong

Kong sample. Secondly, and in contrast, most families analyzed contained only <17% alleles linked to few epitopes and ≥ 34 epitopes per allele on average. However, three of these families (B*08, B*15 and C*07) were large families that included ≥ 10 alleles with limited SARS-CoV-2 epitope sets. Finally, just two families behaved in a hybrid manner: B*14 (18% alleles with few epitopes; 23.7 epitopes/allele) and B*78 (43% alleles with few epitopes, 31.1 epitopes/allele).

A small number of key viral changes may be sufficient to completely negate the contribution of these families, nearly devoid of SARS-CoV-2 ligands, to the cytotoxic protection. Substitutions and deletions (but no insertions) negated the binding of 42% and 37% epitopes (averaged by family), respectively, for weak alleles in the eight families with fewest alleles (Fig. 8B).

A pending issue is whether these SARS-CoV-2 isolates with changes that remove the HLA binding were collected from geographical zones with populations expressing these alleles. According to the Allele Frequency Database, the A*74, B*82 and C*18 families were prevalent in Africa whereas B*46 is common in Eastern Asia. These allele families were also common in minorities within these origins in other countries such as USA. By comparison, the B*52, B*73 and C*01 families were globally disseminated whereas B*83 was extremely rare. When isolates carrying mutations were geographically mapped using sample metadata and superimposed onto allele distribution, co-localization was observed in several cases (Fig. 8C). For instance, nine isolates from Ghana and USA showed the K369D (N protein) change, which canceled the binding of the 361-KTFPPTEPK-369 of eight A*74 alleles. Five isolates from Ghana and Kenya conveyed the ORF1ab deletion $\Delta 6656-6744$ (corresponding to $\Delta 204-292$ of nsp15_A1), which erased the 6669-AMDEFIERY-6677 epitope of the A*74:10 allele. Another example is constituted by nine isolates from India carrying the Q575R mutation in ORF1ab (Q395R in nsp2). This change invalidated binding to eight alleles of B*52 family, being India one of the countries with population samples enriched in this family. Likewise, the $\Delta 6342-6432$ deletion in ORF1ab ($\Delta 417-507$ in nsp14A2) was found in 17 isolates collected in Ghana and negated the

332 6353-TPAFDKSAF-6361 epitope of the B*82:03 allele. The deletion Δ872-966 (equivalent to
333 Δ54-148 of nsp3) of ORF1ab underwent by two Hong-Kong isolates erased the 906-
334 YLFDESGEF-914 epitope associated to three B*46 alleles. Finally, the M85Q (ORF1ab)
335 substitution overrode five B*46 alleles and was found in Bahrain and USA isolates.

336 Another intriguing question is whether independent changes destroying two epitopes
337 bound to alleles of any of these family tend to accumulate in the same isolates. Although it was
338 a rare event, isolates with mutations negating two epitopes were identified in five out of the
339 eight poor SARS-CoV-2-ligandome families. These isolates reached 2.59% of the total carrying
340 at least one mutation negating B*52 epitopes. A prominent example is embodied by twenty-six
341 isolates that combined alterations of ORF1ab Δ5828 (P504 in nsp13_ZBD protein) and large
342 deletions in the ORF1ab 6341-6436 range (416-511 in nsp14A2_ExonN protein). These
343 changes inactivated the 5827-NPAWRKAVF-5835 and 6353-TPAFDKSAF-6361 epitopes,
344 respectively, of the B*82:03 allele. These isolates were collected from 22/03/20 to 09/02/21, in
345 six USA states with different percentages of Afro-American population, suggesting some
346 maintained dissemination degree and potential convergent selective pressure. The fact that these
347 changes were also detected in isolation in several samples from the same country (22 and 62
348 isolates, respectively), indicates double mutants may have arisen by recombination.

349

350 Discussion

351 This study aims to determine to what extent the mutations observed in large SARS-CoV-2
352 genome datasets can perturb the human cytotoxic response against this virus. This impact was
353 studied in HLA class I molecules that practically cover the human population as a whole and,
354 with special attention, to subsets with reduced SARS-CoV-2-ligand repertoires. In general,
355 human and pathogen variability can greatly influence the CD8⁺ response, which may affect the
356 outcome of infection. Some combinations of HLA class I haplotypes and viral genomes appear
357 to further offset the balance towards an insufficient cytotoxic response and, thus, a probable bad
358 prognosis. The surveillance of escape viral variants carried out in this study might therefore help

359 to ameliorate enhanced susceptibility to COVID-19 in sub-populations by designing appropriate
360 countermeasures.

361 The experimental evaluation of the immune response of every human allele associated
362 to each viral variant is not feasible. Computational methods can facilitate this task and generate
363 new, otherwise overlooked, hypotheses. Pioneering bioinformatic studies focused on predicting
364 cytotoxic epitopes of a limited subset of common HLA alleles against the reference viral strain
365 (13,20,23,26). However, SARS-CoV-2 has substantially evolved after more than a year of
366 pandemic, resulting in a human-viral combination landscape of immense scale only
367 approachable using automated techniques.

368 Bioinformatic approaches suffer from intrinsic limitations. These include the possible
369 application of biologically inappropriate thresholds and potentially low predictive performance.
370 Furthermore, alleles considered in algorithms as much as the priceless genome sampling by the
371 worldwide sequencing effort still represent an underestimation of biological variability. Such
372 obstacles were addressed in this study by: (i) utilizing an state-of-the-art algorithm that permits
373 nearly universal fine-grained predictions (~3000 alleles); (ii) the application of stringent cutoffs
374 that reflect the natural strictness of the ligand-HLA binding; (iii) the re-calculation of peptide
375 binding affinity for each mutation; and (iv) the utilization of a large dataset of ~118,000 viral
376 genomes and their corresponding metadata. Mutations were stratified by occurrence, reduction
377 of HLA-binding affinity and geographical dissemination. Thus, the integration of omic data and
378 immunoinformatics in this study very likely capture, despite drawbacks, the principal trends that
379 respond to the posed questions.

380 Large epitope numbers were computationally predicted to be presented by most
381 supertypes. Although all these supermotifs appeared mutated in at least one isolate, most of
382 these mutations did not overcome the supermotif degeneracy. In most cases, the HLA binding
383 affinity was reasonably maintained except from (i) residue substitutions in the second and C-
384 terminal positions of the ligand core, amino acids that usually are anchor motifs; and (2) large
385 deletions that fully removed the epitope. For instance, the Spike-W152C mutation and deletions
386 in the 6342-6432 range in ORF1ab removed several epitopes at the level of supermotifs, and

387 were coupled to several other changes. Respect to the persistence of these escape mutations,
388 point substitutions are likely less prone to impose a dramatic fitness although some extensive
389 deletions have been also been shown to be compatible with infection and transmission (27,28).
390 Large deletions have been related to progressive adaptation to host and reduced virulence
391 (29,30), but their middle-term stability should be analyzed case-to-case.

392 A central question is whether escape mutations have longitudinally accumulated in
393 genomes of individual isolates. If so, such emerging strains would have acquired, or be in the
394 process of acquiring, enhanced capacity to infect individuals previously able to mount an
395 effective cytotoxic response. However, the emergence of this challenging phenotype would not
396 be expected after the examination of the genomic space of the virus carried out in this study.
397 Even the forward line of mutated variants in this respect only combined low numbers (<15%) of
398 escape supermotifs of a given supertype. The remaining intact supermotifs, other HLA class I
399 loci and heterozygosity should compensate escape mutations, provided that the pool of naïve
400 lymphocyte is high enough and the innate-to-adaptive response priming correctly coordinated.
401 Notably, the humoral and CD4⁺ responses would likely remain active and be sufficient in many
402 cases. Therefore, we conclude that the systemic nature of the immune response translates into
403 most healthy subjects remaining competent to respond against variants. The only exception that
404 moderately threatens supertype redundancy was the B27 supertype with isolates that convey
405 evading mutations for up to 33% of these supermotifs. This supertype is common in many
406 populations such, in particular, in Eskimo (31), which may be exposed to "Northern America"
407 isolates with disabled B27 supermotifs.

408 The emergence of isolates that undergo the step-wise accumulation of genetic markers
409 to achieve extended cytotoxic resistance should not be ruled out. This may be favored by
410 considering the explosive expansion of the virus worldwide. However, the mutational space
411 would be reduced in practice due to potential antagonism between cytotoxic evasion pressure
412 and structural-functional restrictions of proteins. However, a sizable fraction of the human
413 population has been infected with the virus, which represents innumerable replication cycles
414 and infection attempts. Some variants have been linked by other scientific groups to different

clinical phenotypes such as increased mortality (32) and antibody escape (33). Likewise, progressive mutation and recombination in SARS-CoV-2 may conceivably achieve a critical number of supermotif escape mutations that collectively constitute a selective advantage. Some identified isolates appeared to have experienced a higher-than-expected number of these changes over the genetic noise, and may have initiated the evasion-driven process.

On the other hand, according to our computational study, a worrisome scenario has already occurred for around ~10% of alleles able to bind a reduced number of ligands from the SARS-CoV-2 proteome. Among them, the A*74, B*82 and C*18 allele families, with sub-Saharan African origin, and the C*46 family, with Far East origin, excelled. Lost or debilitation of the cytotoxic response would make these individuals too dependent upon the humoral response, which can be inefficient during primary infection in some cases (1). This may be very relevant when these alleles are combined into the same haplotype, in particular when in homozygosis.

Underprotection may become exacerbated if these individuals become infected with these escape variants. Given their low epitope redundancy, a very few number of viral mutations, such as those identified in this study, may suffice to circumvent both the cytotoxic primary and memory T responses. The geographical co-existence of viral variants that experience epitope switch respect to some HLA class I molecules and individuals expressing these alleles may exert immediate selective pressure. This may cause rampant dissemination of emergent strains in these niches with local clinical consequences. Most isolates at great risk of achieving critical mutations to impair the CD8⁺ ligand repertoires in these families were found in “Northern America” where some African Americans and Asian subpopulations carried these alleles. Whether these immunotypes with further diminished SARS-CoV-2 ligandomes have undergone positive selection warrants massive local HLA genotyping and viral sequencing. Some of these alleles may be ancestrally specialized in single pathogens, but unable to be effective against international viral infections as reported for Dengue (34), HIV (35) and influenza (36).

In conclusion, here we provide a complete repository of the predicted escape mutations in a recent NCBI genome sampling of SARS-CoV-2. Fortunately, accumulation of these mutations in single isolates does not appear close enough yet to be alarming at the global population level. However, isolates carrying mutations able to override limited CD8⁺ response in some alleles and haplotypes are already co-circulating with individuals carrying these HLA class I molecules. Emerging SARS-CoV-2 variants may further increment the susceptibility of highly vulnerable communities and should be actively surveyed to coordinate appropriate countermeasures. In this respect, bioinformatic pipelines operating on a timely basis may play an irreplaceable role in the protection against this and other pandemic threats.

451

452 **Materials and Methods**

453 **Data acquisition**

SARS-CoV-2 coding sequences and isolate metadata were downloaded from the NCBI repository (Last accession: 19/03/2021) (37). Protein sequences of clinical isolates showing length differences >3% respect to the reference variant were considered anomalous and rejected. The country of origin of isolates were assigned to sub-continental regions following the M49 United Nation geoscheme. Experimental epitopes were downloaded from the IEDB (Last accession: 19/03/2021)(16) using the following search terms: Epitopes: "Any epitopes"; Assay: "T Cell", "MHC Ligand" and outcome: "Positive"; MHC Restriction: "MHC Class I"; Host: "Human"; Disease "COVID-19 (ID: DOID: 0080600)".

Alleles for the twelve HLA class I supertypes were acquired from the original publication (12).

Geographical localization of populations with allele families with few epitopes was carried out using the Allele Frequency Net Database (38). Only samples with at least 50 individuals and $\geq 1\%$ frequency for the given allele family were considered.

467

468 **HLA class I epitope prediction and analysis**

HLA class I epitopes between 8-12 residues in 11 viral proteins, and the ORF1ab polyprotein, of the SARS-CoV-2 reference proteome (Wuhan-1; RefSeq: NC_045512.2) were predicted for 2,915 alleles using NetMHCipan EL 4.1 (22). Binding epitopes were considered those that satisfy the rank ≤ 0.5 (EL rank) and score (EL score) ≥ 0.5 estimations provided by this neural network method. The predictive performance of this algorithm was superior when trained with mass spectrometry elution (EL) data than when trained with binding affinity (BA) data and therefore the former is recommended by the developers for general applications. However, the "EL score" quantifies biologically meaningless abstract units whereas the score of the BA version "BA score" reflects the IC50 in nM. Thus, to take advantage of the strengths of both strategies, the approximate equivalences between EL and BA scores were assessed by exponential regression ($r^2=0.69$) (Supplementary Fig. S1). This comparison resulted in a value for "EL score" ≥ 0.5 was roughly equivalent to an IC50 ≤ 500 nM. This affinity threshold is satisfied by most medium to high-affinity real ligands (39). Redundant epitopes with distinct lengths and lower "EL scores" but sharing the same peptide core and allele were ignored.

Intra-family coherence was calculated by comparing the non-redundant epitope pools between all alleles of the same family and calculate the average Jaccard coefficient (intersection divided by the union) of all families. For that, the intersection was constituted by the number of epitopes that showed a perfect coordinate match for two alleles among all epitopes identified by each allele. Inter-family epitope correlation was calculated by comparing family epitope pools, i.e. those shared by at least half of the alleles of the alleles in the family, like explained above between families. A matrix with all inter-family Jaccard coefficients was used for agglomerative hierarchical clustering by *clustermap* function of *seaborn* data visualization Python library with default options.

Supermotifs, or supertype-associated epitopes, were those predicted as non-redundant epitopes showing perfect coordinates for $\geq 50\%$ alleles in the supertype (12). Only alleles which motifs were experimentally established or shared exact match(es) to second and C-terminal peptide positions, i.e. B and F pockets of the HLA class I groove, in the original reference were considered.

497

498 **Mutation analyses**

499 Mutations respect to the proteins of the reference Wuhan-1 strain (RefSeq: NC_045512.2) were
500 identified by aligning with Clustal Omega 1.2.1 (40) with an in-house perl script. All adjacent
501 insertion or deletion runs were collapsed into single events. Non-conservative mutations were
502 deemed those involving distinct physicochemical classes: acidic (D, E), amide (N, Q), basic (H,
503 K, R), cysteine (C), glycine (G), hydroxyl (S, T), hydrophobic aliphatic (A, I, L, M, V),
504 hydrophobic aromatic (F, W, Y) and proline (P) residues.

505 The impact of point substitutions on epitope binding was assessed by recalculating the
506 "EL score" and "EL rank" of the mutated peptide. For insertions, flanking regions to the insertion
507 limits were taken to complete 22mer sections and binding also recalculated. Likewise, for
508 deletions, the resulting 11mers flanking the deletion limits were merged into 22mer sections.
509 Based on the "BA score" and "EL score" correspondences, "EL scores" of < 0.1 roughly
510 corresponded to BA scores of >5000nM (Supplementary Fig. S1), associated to non-binders
511 according to the IEDB curators. Thus, mutations causing medium-strong peptide binders
512 decreases to EL scores < 0.1 besides EL rank ≥ 1 were deemed epitope escape mutations. For
513 supermotif escape mutations, the allele of the supermotif showing the highest "EL score" for the
514 wild-type epitope was tested. For the eight families with fewest epitopes, escape mutations were
515 calculated for each allele.

516 Network graphs of coupled mutations were carried out using the NetworkX (41) Python
517 library.

518

519 **Funding**

520 This research was supported by Acción Estratégica en Salud from the ISCIII
521 (<https://www.isciii.es>), grants MPY 380/18 (to MJM), 388/18 (to DL) and 509/19 (to AJM-G).
522 AJM-G is the recipient of a Miguel Servet contract by the ISCIII. The funders had no role in
523 study design, data collection and analysis, decision to publish, or preparation of the manuscript.

524

525 **Author Contributions**

526 **Conceptualization:** Daniel López, Michael J. McConnell, Antonio J. Martín-Galiano.

527 **Data curation:** Anna Foix, Antonio J. Martín-Galiano.

528 **Formal analysis:** Anna Foix, Antonio J. Martín-Galiano.

529 **Funding acquisition:** Daniel López, Michael J. McConnell, Antonio J. Martín-Galiano.

530 **Investigation:** Anna Foix, Daniel López, Michael J. McConnell, Antonio J. Martín-Galiano.

531 **Methodology:** Anna Foix, Antonio J. Martín-Galiano.

532 **Project administration:** Antonio J. Martín-Galiano.

533 **Resources:** Antonio J. Martín-Galiano.

534 **Software:** Anna Foix, Antonio J. Martín-Galiano.

535 **Supervision:** Daniel López, Michael J. McConnell, Antonio J. Martín-Galiano.

536 **Validation:** Antonio J. Martín-Galiano.

537 **Visualization:** Anna Foix, Antonio J. Martín-Galiano.

538 **Writing—original draft:** Antonio J. Martín-Galiano.

539 **Writing—review & editing:** Daniel López, Michael J. McConnell, Antonio J. Martín-Galiano.

540

541 **References**

- 542 1. Rydzynski Moderbacher C, Ramirez SI, Dan JM, Grifoni A, Hastie KM, Weiskopf D, et
543 al. Antigen-Specific Adaptive Immunity to SARS-CoV-2 in Acute COVID-19 and
544 Associations with Age and Disease Severity. *Cell*. 2020 Nov 12;183(4):996-1012.e19.
- 545 2. Liu L, Wang P, Nair MS, Yu J, Rapp M, Wang Q, et al. Potent neutralizing antibodies
546 against multiple epitopes on SARS-CoV-2 spike. *Nature*. 2020 Aug;584(7821):450–6.
- 547 3. Sette A, Crotty S. Adaptive immunity to SARS-CoV-2 and COVID-19. *Cell*. 2021 Jan 12;
- 548 4. Hellerstein M. What are the roles of antibodies versus a durable, high quality T-cell
549 response in protective immunity against SARS-CoV-2? *Vaccine X*. 2020 Dec
550 11;6:100076.
- 551 5. Liu L, Wei Q, Lin Q, Fang J, Wang H, Kwok H, et al. Anti-spike IgG causes severe acute
552 lung injury by skewing macrophage responses during acute SARS-CoV infection. *JCI*
553 *Insight*. 2019 Feb 21;4(4).

- 554 6. Tang F, Quan Y, Xin Z-T, Wrammert J, Ma M-J, Lv H, et al. Lack of peripheral memory
555 B cell responses in recovered patients with severe acute respiratory syndrome: a six-year
556 follow-up study. *J Immunol Baltim Md* 1950. 2011 Jun 15;186(12):7264–8.
- 557 7. Le Bert N, Tan AT, Kunasegaran K, Tham CYL, Hafezi M, Chia A, et al. SARS-CoV-2-
558 specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls.
559 *Nature*. 2020 Aug;584(7821):457–62.
- 560 8. Dan JM, Mateus J, Kato Y, Hastie KM, Yu ED, Faliti CE, et al. Immunological memory
561 to SARS-CoV-2 assessed for up to 8 months after infection. *Science*. 2021 Feb
562 5;371(6529).
- 563 9. Schulien I, Kemming J, Oberhardt V, Wild K, Seidel LM, Killmer S, et al.
564 Characterization of pre-existing and induced SARS-CoV-2-specific CD8(+) T cells. *Nat*
565 *Med*. 2021 Jan;27(1):78–85.
- 566 10. Mudatsir M, Fajar JK, Wulandari L, Soegiarto G, Ilmawan M, Purnamasari Y, et al.
567 Predictors of COVID-19 severity: a systematic review and meta-analysis. *F1000Research*.
568 2020;9:1107.
- 569 11. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. IPD-IMGT/HLA
570 Database. *Nucleic Acids Res*. 2020 Jan 8;48(D1):D948–55.
- 571 12. Sidney J, Peters B, Frahm N, Brander C, Sette A. HLA class I supertypes: a revised and
572 updated classification. *BMC Immunol*. 2008 Jan 22;9:1.
- 573 13. Nelde A, Bilich T, Heitmann JS, Maringer Y, Salih HR, Roerden M, et al. SARS-CoV-2-
574 derived peptides define heterologous and COVID-19-induced T cell recognition. *Nat*
575 *Immunol*. 2021 Jan;22(1):74–85.
- 576 14. Kiyotani K, Toyoshima Y, Nemoto K, Nakamura Y. Bioinformatic prediction of potential
577 T cell epitopes for SARS-Cov-2. *J Hum Genet*. 2020 Jul;65(7):569–75.
- 578 15. Peng Y, Mentzer AJ, Liu G, Yao X, Yin Z, Dong D, et al. Broad and strong memory
579 CD4(+) and CD8(+) T cells induced by SARS-CoV-2 in UK convalescent individuals
580 following COVID-19. *Nat Immunol*. 2020 Nov;21(11):1336–45.
- 581 16. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The Immune
582 Epitope Database (IEDB): 2018 update. *Nucleic Acids Res*. 2019 Jan 8;47(D1):D339–43.
- 583 17. Correale P, Mutti L, Pentimalli F, Baglio G, Saladino RE, Sileri P, et al. HLA-B*44 and
584 C*01 Prevalence Correlates with Covid19 Spreading across Italy. *Int J Mol Sci*. 2020 Jul
585 23;21(15).
- 586 18. Wang W, Zhang W, Zhang J, He J, Zhu F. Distribution of HLA allele frequencies in 82
587 Chinese individuals with coronavirus disease-2019 (COVID-19). *HLA*. 2020
588 Aug;96(2):194–6.
- 589 19. Harjanto S, Ng LFP, Tong JC. Clustering HLA class I superfamilies using structural
590 interaction patterns. *PloS One*. 2014;9(1):e86655.
- 591 20. Nguyen A, David JK, Maden SK, Wood MA, Weeder BR, Nellore A, et al. Human
592 Leukocyte Antigen Susceptibility Map for Severe Acute Respiratory Syndrome
593 Coronavirus 2. *J Virol*. 2020 Jun 16;94(13).

- 594 21. Du VY, Bansal A, Carlson J, Salazar-Gonzalez JF, Salazar MG, Ladell K, et al. HIV-1-
595 Specific CD8 T Cells Exhibit Limited Cross-Reactivity during Acute Infection. *J Immunol*
596 *Baltim Md* 1950. 2016 Apr 15;196(8):3276–86.
- 597 22. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and
598 NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent
599 motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.*
600 2020 Jul 2;48(W1):W449–54.
- 601 23. Tarke A, Sidney J, Kidd CK, Dan JM, Ramirez SI, Yu ED, et al. Comprehensive analysis
602 of T cell immunodominance and immunoprevalence of SARS-CoV-2 epitopes in COVID-
603 19 cases. *Cell Rep Med.* 2021 Feb 16;2(2):100204.
- 604 24. Sette A, Sidney J. HLA supertypes and supermotifs: a functional perspective on HLA
605 polymorphism. *Curr Opin Immunol.* 1998 Aug;10(4):478–82.
- 606 25. Rousseau CM, Daniels MG, Carlson JM, Kadie C, Crawford H, Prendergast A, et al. HLA
607 class I-driven evolution of human immunodeficiency virus type 1 subtype c proteome:
608 immune escape and viral load. *J Virol.* 2008 Jul;82(13):6434–46.
- 609 26. Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. A Sequence
610 Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune
611 Responses to SARS-CoV-2. *Cell Host Microbe.* 2020 Apr 8;27(4):671-680.e2.
- 612 27. Lam J-Y, Yuen C-K, Ip JD, Wong W-M, To KK-W, Yuen K-Y, et al. Loss of orf3b in the
613 circulating SARS-CoV-2 strains. *Emerg Microbes Infect.* 2020 Dec;9(1):2685–96.
- 614 28. Su YCF, Anderson DE, Young BE, Linster M, Zhu F, Jayakumar J, et al. Discovery and
615 Genomic Characterization of a 382-Nucleotide Deletion in ORF7b and ORF8 during the
616 Early Evolution of SARS-CoV-2. *mBio.* 2020 Jul 21;11(4).
- 617 29. Benedetti F, Pachetti M, Marini B, Ippodrino R, Ciccozzi M, Zella D. SARS-CoV-2:
618 March toward adaptation. *J Med Virol.* 2020 Nov;92(11):2274–6.
- 619 30. Peacock TP, Penrice-Randal R, Hiscox JA, Barclay WS. SARS-CoV-2 one year on:
620 evidence for ongoing viral adaptation. *J Gen Virol.* 2021 Apr;102(4).
- 621 31. Peschken CA, Esdaile JM. Rheumatic diseases in North America's indigenous peoples.
622 *Semin Arthritis Rheum.* 1999 Jun;28(6):368–91.
- 623 32. Toyoshima Y, Nemoto K, Matsumoto S, Nakamura Y, Kiyotani K. SARS-CoV-2
624 genomic variations associated with mortality rate of COVID-19. *J Hum Genet.* 2020
625 Dec;65(12):1075–82.
- 626 33. McCarthy KR, Rennick LJ, Nambulli S, Robinson-McCarthy LR, Bain WG, Haidar G, et
627 al. Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape.
628 *Science.* 2021 Mar 12;371(6534):1139–42.
- 629 34. Vejbaesya S, Thongpradit R, Kalayanaroj S, Luangtrakool K, Luangtrakool P, Gibbons
630 RV, et al. HLA Class I Supertype Associations With Clinical Outcome of Secondary
631 Dengue Virus Infections in Ethnic Thais. *J Infect Dis.* 2015 Sep 15;212(6):939–47.
- 632 35. Li S, Jiao H, Yu X, Strong AJ, Shao Y, Sun Y, et al. Human leukocyte antigen class I and
633 class II allele frequencies and HIV-1 infection associations in a Chinese cohort. *J Acquir*
634 *Immune Defic Syndr* 1999. 2007 Feb 1;44(2):121–31.

- 635 36. Falfán-Valencia R, Narayanankutty A, Reséndiz-Hernández JM, Pérez-Rubio G, Ramírez-
636 Venegas A, Nava-Quiroz KJ, et al. An Increased Frequency in HLA Class I Alleles and
637 Haplotypes Suggests Genetic Susceptibility to Influenza A (H1N1) 2009 Pandemic: A
638 Case-Control Study. *J Immunol Res*. 2018;2018:3174868.
- 639 37. Database resources of the National Center for Biotechnology Information. *Nucleic Acids*
640 *Res*. 2018 Jan 4;46(D1):D8–13.
- 641 38. Gonzalez-Galarza FF, McCabe A, Santos EJMD, Jones J, Takeshita L, Ortega-Rivera ND,
642 et al. Allele frequency net database (AFND) 2020 update: gold-standard data
643 classification, open access genotype data and new query tools. *Nucleic Acids Res*. 2020
644 Jan 8;48(D1):D783–8.
- 645 39. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. NetMHC-3.0:
646 accurate web accessible predictions of human, mouse and monkey MHC class I affinities
647 for peptides of length 8-11. *Nucleic Acids Res*. 2008 Jul 1;36(Web Server issue):W509-
648 512.
- 649 40. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein
650 sequences. *Protein Sci Publ Protein Soc*. 2018 Jan;27(1):135–45.
- 651 41. Aric A, Schult DA, Swart PJ. “Exploring network structure, dynamics, and function using
652 NetworkX” in *Proceedings of the 7th Python in Science Conference (SciPy2008)*.
653 Pasadena, CA USA: Gäel Varoquaux, Travis Vaught, and Jarrod Millman; 2008. 11–15 p.

654

655 SUPPORTING INFORMATION CAPTIONS

656 Figure S1. Correspondence between netMHCpan 4.1 EL and BA scores.

657 Table S1. Intra-family conserved epitopes.

658 Table S2. Supertype-associated epitopes.

659 Table S3. Supermotif escape substitutions.

660 Table S4. Supermotif escape deletions.

661 Table S5. Isolates carrying escape mutations for five or more supermotifs.

662 Table S6. Escape substitutions for allele families with few epitopes.

663 Table S7. Escape deletions for allele families with few epitopes.

664

665

LEGENDS TO FIGURES

666 **Fig 1. Number and degree of overlap between SARS-CoV-2 epitopes for different HLA-**

667 **class I allelic families. (A)** Average number of predicted HLA class I epitopes by allele family

668 and protein. The standard deviation resulting from all proteins is indicated as a single error bar.

669 **(B)** Hierarchical clustering and associated heatmap indicating the degree of inter-family epitope

670 correlation. Color intensity expresses the Jaccard index for the epitope intersection between all

671 family pairs. Perfect location match between epitopes calculated by netMHCipan 4.1 EL with

672 score ≥ 0.5 and rank ≤ 0.5 were utilized to calculate intersection and union. Intra-family

673 conserved epitopes ($\geq 50\%$ alleles in the family by exact match) are in Supplementary Table S1.

674

675 **Fig 2. Comparison between predicted and validated epitopes. (A)** Number of predicted

676 epitopes (score ≥ 0.5 and rank ≤ 0.5) versus validated epitopes per allele. **(B)** Heatmap showing

677 the family average score (any score, rank ≤ 2) for validated HLA class I epitopes. Predicted

678 epitopes with perfect matches with validated epitopes stored in the IEDB are indicated in

679 Supplementary Table S1.

680

681 **Fig 3. SARS-CoV-2 supermotifs. (A)** Distribution of supermotifs according to the number of

682 supertypes covered. **(B)** Number of supermotifs per supertype detailed by protein antigen.

683

684 **Fig 4. Global mutation analysis in NCBI SARS-CoV-2 genomes. (A)** Proportion of

685 cumulative and unique residue mutation events in SARS-CoV-2. **(B)** Length distribution of

686 insertions (left) and deletions (right). **(C)** Number of isolates and number of epitopes which

687 location overlap to substitutions (left), insertions (center) and deletions (right).

688

689 **Fig 5. Supermotif escape mutations. (A)** Influence of supermotif core position and residue

690 conservation in the epitope escape capacity of substitutions. **(B)** Average percentage of escape

691 supermotifs by any mutation type after incremental filter application. **(C)** Absolute number of

692 mutated supermotifs for each supertype after incremental filter application. **(D)** Nightingale rose

charts indicating the percentage of escape supermotifs in prevalent M49 zones. Only mutations involving ≥ 2 isolates in the M49 were considered. Only M49 zones with $\geq 5\%$ escape supermotifs for at least one supertype are shown.

696

Fig 6. Networks of coupled supermotif escape mutations. Undirected unweighted graphs showing coupled supermotif escape mutations. Sub-networks are named with roman numbers. Nodes correspond to mutations that were substitutions (position and residue change) or deletions (residue range). No coupled insertions were detected. The node color indicates the antigen protein. The sphere diameter reflects the amount of isolates harboring the mutation. Nodes represent mutations carried by ≥ 25 isolates. Edges represent co-existence of a mutation pair in $\geq 20\%$ isolates of all those carrying at least one of the mutations.

704

Fig 7. Isolates carrying different combinations of escape mutations. (A) Each point represents an isolate plotted according to the total number of mutations, the number of escape supermotifs and number of affected superotypes. Only isolates harboring three or more escape supermotifs are represented. (B) Chart panel indicating mutated isolates according to the number of escape supermotifs for each supertype. Isolates are colored by M49 zone of collection.

711

Fig 8. Escape mutations in allele families with fewest epitopes. (A) Number of alleles with ≤ 20 epitopes versus the total number of alleles for HLA families of the three loci. Families without any allele with ≤ 20 epitopes are not represented. (B) Average number of escape epitopes, either by substitutions or deletions, respect to the average total number of epitopes for the eight allele families with the fewest epitopes. (C) World map panel indicating the presence of population samples carrying alleles of the eight families with fewest epitopes and isolates with escape mutations for these families. Family allele frequencies are color ranked for both the majority population (red scale) and sub-population (blue scale) samples. Only the highest frequency sample per country was considered. B*83 data is not shown due to its extremely low

721 prevalence. Spheres in green indicate the presence of isolates with escape mutations for the
 722 allele family collected in that country. The sphere diameter is proportional to the total number of
 723 these isolates. Epitope escape substitutions and deletions for the eight allele family with fewest
 724 epitopes are listed on Supplementary Tables S6 and S7, respectively.

725

A

B

A **B**

A **B**

A

Cumulative:
(all events)

Unique:

B

Insertions:

Deletions:

C

A **B** **C** **D**

Australia and New Zealand

Northern America

Southern Asia

A

B

A **B** **C**