

1 **Computational inference of selection underlying the evolution of the novel coronavirus,**

2 **SARS-CoV-2**

3

4 Rachele Cagliani*^{1#}, Diego Forni*¹, Mario Clerici^{2,3}, Manuela Sironi¹

5

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

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

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

9 * These authors equally contributed to this work. Author order was determined alphabetically.

10

11 # Address for correspondence: Rachele Cagliani (rachele.cagliani@lanostrafamiglia.it)

12

13

14 Running title: Molecular evolution of SARS-CoV-2

15

16

17 Abstract word count: 250

18 Text word count: 3151

19

20 **Abstract**

21

22 The novel coronavirus (SARS-CoV-2) recently emerged in China is thought to have a bat origin, as
23 its closest known relative (BatCoV RaTG13) was described in horseshoe bats. We analyzed the
24 selective events that accompanied the divergence of SARS-CoV-2 from BatCoV RaTG13. To this
25 aim, we applied a population genetics-phylogenetics approach, which leverages within-population
26 variation and divergence from an outgroup. Results indicated that most sites in the viral ORFs
27 evolved under strong to moderate purifying selection. The most constrained sequences
28 corresponded to some non-structural proteins (nsps) and to the M protein. Conversely, nsp1 and
29 accessory ORFs, particularly ORF8, had a non-negligible proportion of codons evolving under very
30 weak purifying selection or close to selective neutrality. Overall, limited evidence of positive
31 selection was detected. The 6 *bona fide* positively selected sites were located in the N protein, in
32 ORF8, and in nsp1. A signal of positive selection was also detected in the receptor-binding motif
33 (RBM) of the spike protein but most likely resulted from a recombination event that involved the
34 BatCoV RaTG13 sequence. In line with previous data, we suggest that the common ancestor of
35 SARS-CoV-2 and BatCoV RaTG13 encoded/encodes an RBM similar to that observed in SARS-
36 CoV-2 itself and in some pangolin viruses. It is presently unknown whether the common ancestor
37 still exists and which animals it infects. Our data however indicate that divergence of SARS-CoV-2
38 from BatCoV RaTG13 was accompanied by limited episodes of positive selection, suggesting that
39 the common ancestor of the two viruses was poised for human infection.

40

41

42

43

44

45

46 **Importance**

47

48 Coronaviruses are dangerous zoonotic pathogens: in the last two decades three coronaviruses
49 have crossed the species barrier and caused human epidemics. One of these is the recently
50 emerged SARS-CoV-2. We investigated how, since its divergence from a closely related bat
51 virus, natural selection shaped the genome of SARS-CoV-2. We found that distinct coding
52 regions in the SARS-CoV-2 genome evolve under different degrees of constraint and are
53 consequently more or less prone to tolerate amino acid substitutions. In practical terms, the
54 level of constraint provides indications about which proteins/protein regions are better suited
55 as possible targets for the development of antivirals or vaccines. We also detected limited
56 signals of positive selection in three viral ORFs. However, we warn that, in the absence of
57 knowledge about the chain of events that determined the human spill-over, these signals should
58 not be necessarily interpreted as evidence of an adaptation to our species.

59

60

61

62 **Introduction**

63

64 In December 2019, a human-infecting coronavirus, now referred to as SARS-CoV-2 (1), emerged in
65 Wuhan, China, causing respiratory disease in a large number of people and being responsible for
66 thousands of deaths (<https://www.who.int/emergencies/diseases/novel-coronavirus-2019>) (2). After
67 SARS-CoV (severe acute respiratory syndrome coronavirus) and MERS-CoV (Middle East
68 respiratory syndrome coronavirus), SARS-CoV-2 is the third coronavirus to cause a human
69 epidemic in the last two decades (3, 4).

70 Coronaviruses (family *Coronaviridae*, order *Nidovirales*) have positive-sense, single stranded RNA
71 genomes, which are unusually long and complex if compared to those of other RNA viruses. Two
72 thirds of the coronavirus genome are occupied by two large overlapping open reading frames
73 (ORF1a and ORF1b), that are translated into the pp1a and pp1ab polyproteins. These are processed
74 to generate 16 non structural proteins (nsp1 to 16) (5). The remaining portion of the genome
75 includes ORFs for the structural proteins: spike (S), envelope (E), membrane (M) and nucleoprotein
76 (N), as well as a variable number of accessory proteins (3-5).

77 Several coronavirus genera and subgenera are recognized (<https://talk.ictvonline.org/ictv-reports/>)
78 (1, 6, 7). Whereas MERS-CoV is a member of the *Merbecovirus* subgenus, phylogenetic analyses
79 indicated that SARS-CoV-2 clusters with SARS-CoV and other bat-derived viruses in the
80 *Sarbecovirus* subgenus (genus *Betacoronavirus*) (1, 8, 9). A recent report by the Coronavirus Study
81 Group of the International Committee on Taxonomy of Viruses (ICTV) indicated that SARS-CoV-2
82 can be assigned to the species *Severe acute respiratory syndrome-related coronavirus* (1).

83 Bats host a large diversity of coronaviruses related to SARS-CoV (5, 10, 11) and, in general, these
84 animals are believed to represent the original reservoir of several human-infecting coronaviruses (3,
85 4). This also seems to be the case for SARS-CoV-2, as analysis of the viral genome indicated that
86 its known closest relative, with an average identity of ~96%, is a virus (BatCoV RaTG13) identified

87 in horseshoe bats (*Rhinolophus affinis*) (8). Two other bat-derived coronaviruses (bat-SL-CoVZC45
88 and bat-SL-CoVZXC21) display high levels of similarity (> 70%) to SARS-CoV-2, with identity
89 varying along the genome (9, 12, 13). However, because both SARS-CoV and MERS-CoV were
90 transmitted to humans via intermediate hosts (3, 4), it remains unclear whether the Wuhan epidemic
91 was initiated by a spill-over from bats or from other animals. Recent data suggested that viruses
92 related to SARS-CoV-2 are found in pangolins (*Manis javanica*), but the role of these animals in
93 fueling the human epidemic remains unclear (14-17).

94 A major determinant of coronavirus host range is represented by the binding affinity between the
95 spike protein and the cognate cellular receptor (18-22). Notably, this was previously shown to be
96 the case for SARS-CoV, which, in analogy to SARS-CoV-2, uses ACE2 (angiotensin-converting
97 enzyme 2) to enter host cells (8, 23). Few amino acid changes in the receptor binding domain
98 (RBD) of SARS-CoV were shown to modulate the binding efficiency to ACE2 from different
99 mammalian species and contributed to the adaptation of the virus to human cells (24-26). However,
100 the SARS-CoV epidemic was characterized by another signature change in the viral genome:
101 relatively early during the human-to-human transmission chain, SARS-CoV strains acquired a 29-
102 nucleotide deletion which split ORF8, encoding an accessory protein, in two functional ORFs (27).
103 Together with the observation that ORF8 is fast evolving in SARS-CoV strains, this finding was
104 taken to imply adaptation to our species (28). The evidence for adaptation was subsequently
105 questioned and recent data indicated that the 29-nucleotide deletion most likely represents a founder
106 effect, which causes fitness loss irrespective of the host species (4, 29). These data underscore the
107 relevance (and possible pitfalls) of evolutionary analyses in the study of viral species emergence
108 and host shifts.

109 Herein, we used available SARS-CoV-2 strains to describe the selective events that accompanied
110 the divergence of this novel human pathogen from its closest known relative (BatCoV RaTG13) (8).
111

112 **Results and Discussion**

113

114 As mentioned above, the closest relative (BatCoV RaTG13) of the novel human-infecting SARS-
115 CoV-2 was identified in bats (8). It is presently unknown whether BatCoV RaTG13 can be
116 transmitted in human populations and if it can infect human cells. Likewise, the reservoir and the
117 animal host that fueled the human transmission of SARS-CoV-2 is presently uncertain. For sure,
118 ample data now indicate that human-to-human transmission has a role in spreading the SARS-CoV-
119 2 epidemic (30-33) and that, in addition to humans, the virus can infect cells from bats, small
120 carnivores, and pigs (8). We thus set out to determine the selective events that accompanied the
121 divergence of the SARS-CoV-2 lineage from BatCoV RaTG13. In doing so, we do not imply that
122 any such event was primarily responsible for human adaptation, as high efficiency of human
123 infection might instead represent an incidental byproduct of adaptation to another host.

124 Based on the alignment of forty-four SARS-CoV-2 genomes and the BatCoV RaTG13 sequence,
125 147 amino acid replacements, unevenly distributed along the genome, were found to separate
126 SARS-CoV-2 from its closest relative. Forty-one amino acid changes are polymorphic in the SARS-
127 CoV-2 population (Fig. 1A).

128 To investigate the selection patterns acting on SARS-CoV-2 genomes, we applied a method that
129 combines analysis of within-population variation (i.e., variation among SARS-CoV-2 strains) and
130 divergence from an outgroup (BatCoV RaTG13). Specifically, nucleotide alignments were analyzed
131 using gammaMap (34), which estimates selection coefficients (γ) along coding regions and allows
132 the detection of fine-scale differences in selective pressures at specific codons. In practical terms, γ
133 values can be considered a measure of the fitness consequences of new nonsynonymous mutations.
134 The method categorizes selection coefficients into 12 predefined classes ranging from -500
135 (inviable) to 100 (strongly beneficial). For gammaMap analysis, we divided the ORF1a and ORF1b
136 alignments into the 16 nsps; because nsp3 is a long, multi-domain protein, it was also split into

137 domains. Likewise, the coronavirus S protein includes two functionally distinct units (S1 and S2),
138 which were separately analyzed. Alignments of more than 80 codons were analyzed with
139 gammaMap (Fig. 1A).

140 As previously shown for several other viruses (35-37), we found that most sites evolved under
141 strong to moderate purifying selection ($\gamma < -5$). However, the strength of purifying selection varied
142 depending on the region. The strongest constraints were observed for nsp6 to 10, for nsp16, and
143 for the M ORF (Fig. 1B). Whereas nsp6 is involved in the formation of the reticulovesicular
144 membrane network where viral RNA replication occurs, nsp7 to nsp10 are small proteins that
145 function as cofactors for viral replicative enzymes, including nsp16, a 2'-O-methyl transferase (38).
146 Conversely, the M ORF encodes a structural protein, which is highly abundant in the in the virion of
147 coronaviruses (39). The M protein interacts with other structural viral proteins and plays an
148 important role in virion morphogenesis (40). Importantly, the M protein is a dominant immunogen
149 for both the humoral and the cellular immune responses (41, 42). These latter features and its high
150 level of constraint suggest that the M protein represents an excellent target for vaccine design.

151 Among the non-accessory ORFs, the lowest levels of constraint were observed for nsp1 and the
152 acidic domain of nsp3 (Fig. 1B and 1C). This is in line with evidences indicating that these regions
153 are fast evolving in coronaviruses at large (see below) (43, 44). Accessory ORFs, and in particular
154 ORF8, had a non-negligible proportion of codons evolving under very weak purifying selection or
155 close to selective neutrality. On one hand, this is in line with the idea that genetic variation in
156 accessory ORFs causes limited fitness consequences, as the above-mentioned case of SARS-CoV
157 ORF8 indicates (4, 29). In fact, gains and losses of accessory proteins have been common during
158 the evolutionary history of coronaviruses and accessory ORFs differ in number and sequence even
159 among coronaviruses belonging to the same genus or subgenus (4). On the other hand, accessory
160 proteins were often shown to contribute to the modulation of immune responses, as well as to
161 virulence (3, 4). It is thus conceivable that their limited constraint maintains variability in

162 coronavirus accessory ORFs, eventually facilitating rapid adaptation when the environment (e.g.,
163 host) changes.

164 We next wished to determine whether positive selection at specific sites also drove the evolution of
165 SARS-CoV-2. We thus estimated codon-wise posterior probabilities for each selection coefficient.
166 Very strong evidence (defined as a posterior probability > 0.80 of $\gamma \geq 1$) of positive selection was
167 detected for seven sites, six in the S1 region of the spike protein and one in N (Fig. 2). When the
168 posterior probability cutoff was lowered to a less stringent value of 0.50, five additional sites in
169 ORF8 (4) and in nsp1 (1) were identified (Fig. 2). It should be noted that this p value cutoff
170 represents a reasonably strong evidence of positive selection. Using these criteria, positively
171 selected sites were estimated to account for the 0.12% of analyzed codons if 0.5 is used as the cutoff
172 (0.07% for a 0.8 cutoff) (34, 45, 46).

173 The S1 region contains the RBD, and crystal structure of the SARS-CoV S protein in complex with
174 human ACE2 showed that, in turn, the RBD is formed by two subdomains, a core structure and the
175 receptor-binding motif (RBM, that directly contacts ACE2) (47, 48). The S2 region includes the
176 fusion machinery (49). We performed homology modeling of the SARS-CoV-2 S protein onto the
177 SARS-CoV structure and we analyzed the distribution of selection coefficients (Fig. 3A). The S2
178 subunit was characterized by stronger constraint than the S1 portion and five out of six putative
179 positively selected sites were found to be located in the RBM, at the binding interface with ACE2
180 (Fig. 3A).

181 When SARS-CoV-2 and BatCoV RaTG13 are compared, the RBM stands out as the single most
182 divergent region (Fig. 1A)(8, 16). Very recent evidence indicated that, although the average genome
183 similarity is lower compared to BatCoV RaTG13, coronaviruses isolated from pangolins have
184 RBMs almost identical to that of SARS-CoV (14-17). This clearly implies that recombination might
185 have inflated the estimation of positive selection in the S1 region. A pangolin virus available in
186 GenBank (isolate MP789) has an RBM with high identity to SARS-CoV-2. Thus, using the genome

187 sequence of isolate MP789, SARS-CoV-2 and BatCoV RaTG13 we searched for recombination
188 events using RDP4 (50). No evidence of recombination was detected, but this finding might be due
189 to the fact that the parental sequence with which BatCoV RaTG13 recombined is presently
190 unsampled. We thus analyzed synonymous substitutions in the RBM alignment for these viruses:
191 we found that 41% (n= 37) of such substitutions are shared between SARS-CoV-2 and isolate
192 MP789, whereas only 27% (n= 10) are shared between SARS-CoV-2 and BatCoV RaTG13.
193 Overall, these findings strongly suggest that recombination rather than positive selection shaped the
194 genetic diversity at the RBM, as previously suggested (16). Recombination is known to affect
195 evolutionary inference (51). In this case, because we used the BatCoV RaTG13 as an outgroup, the
196 spurious signals were generated by considering the selected sites as amino acid replacements that
197 arose and fixed in the SARS-CoV-2 population, whereas they may represent changes that occurred
198 in the outgroup through recombination. We consider that this is not the case for the other signals we
199 detected, as all of them were located in regions of high overall similarity between BatCoV RaTG13
200 and SARS-CoV-2, indicating no evidence of recombination (Fig. 1A).

201 The positively selected site (A267) in the nucleocapsid protein is located in the C-terminal domain.
202 Homology modeling using the SARS-CoV N protein as a template indicated that A267 is located on
203 an exposed loop on the protein surface (Fig. 3B)(52). The N protein is the most abundant protein in
204 coronavirus-infected cells (53, 54). Its primary function is to package the viral genome into a
205 ribonucleoprotein complex. In addition, the N protein performs non-structural functions, as it
206 regulates the host cell cycle and the stress response, it acts as a molecular chaperone, and it
207 interferes with the host immune response (53, 54). Because these activities are mediated by
208 interaction with different cellular proteins, the positively selected site might be evolving to
209 establish, maintain, or avoid the binding of different host molecules.

210 Another positively selected site was detected in the nsp1 region, which also displayed relatively
211 weak selective constraint. In SARS-CoV and other betacoronaviruses, nsp1 is a virulence factor and

212 is essential for viral replication at least in the presence of an intact host interferon (IFN) response
213 (55-57). Despite their relevant role for viral fitness *in vivo*, nsp1 proteins tend to be variable in
214 sequence both within and among coronavirus genera. Detailed analysis of SARS-CoV nsp1
215 indicated that the protein plays multiple roles during viral infection, including inhibition of host
216 protein synthesis, antagonism of IFN responses, modulation of the calcineurin/NFAT pathway, and
217 induction of chemokine secretion (43). Homology modeling using the SARS-CoV nsp1 structure
218 indicated that the positively selected site (E93) is exposed on the protein surface (Fig. 3C).
219 Extensive mutagenesis of SARS-CoV nsp1 showed that exposed charged residues, including the
220 positively selected site, mediate inhibition of gene expression and antiviral signaling (58).
221 Moreover, the N-terminal half of SARS-CoV nsp1 interacts with immunophilins and calcipressins
222 to modulate the calcineurin/NFAT pathway (59). Overall, these observations suggest that the
223 diversity of coronavirus nsp1 proteins is driven by the need to establish interactions with multiple
224 cellular partners and to evade immune surveillance. This is also likely to explain the positive
225 selection signal we detected. In general, a better understanding of the evolutionary constraints and
226 forces acting on coronavirus nsp1 proteins may be extremely relevant, as the generation of viruses
227 carrying nsp1 mutations was regarded as a potential strategy to generate attenuated vaccine strains
228 (57, 60), and inhibitors of cyclophilins were considered as potential antivirals for coronavirus
229 treatment (59).
230 Finally, the selected sites we identified in ORF8 (F3, I10, A14, T26) are all located in the N-
231 terminal portion of the protein (Fig. 2). The SARS-CoV-2 ORF8 protein displays 30% identity to
232 the intact ORF8 from the SARS-CoV GZ02 strain. It is presently unsure whether the SARS-CoV
233 ORF8 N-terminus is cleaved as a signal peptide or inserted in the endoplasmic reticulum membrane
234 (61, 62). Using computational methods to predict signal peptides and transmembrane helices we
235 found evidence for both in the case of the N-terminus of SARS-CoV-2 ORF8 (not shown). Clearly,

236 experimental analyses will be required to determine the function of the N-terminal region of ORF8,
237 and, more generally the relevance of the selected sites on virus fitness or pathogenicity.

238 Overall, our analyses indicate that distinct coding regions in the SARS-CoV-2 genome evolve under
239 different degrees of constraint and are consequently more or less prone to tolerate amino acid
240 substitutions. In practical terms, the level of constraint can provide indications concerning which
241 specific proteins or protein regions are better suited as possible targets for the development of
242 antivirals or vaccines. Conversely, the current available knowledge and the analyses reported here
243 allow no inference on the selective events (or lack thereof) that turned SARS-CoV-2 into a human
244 pathogen. Recent analyses paid much attention to changes in the RBM. This is indeed expected to
245 represent a major determinant of host range and its sequence is highly variable among SARS-CoV-
246 related viruses (as also evident in Fig. 2). Albeit preliminary and necessarily limited to currently
247 sampled genomes, our analyses suggest that recombination had a role in shaping the diversity of the
248 RBMs in these viruses. Our data also indicate that divergence of SARS-CoV-2 from BatCoV
249 RaTG13 was accompanied by limited episodes of positive selection, suggesting that the common
250 ancestor of the two viruses was poised for human infection. We also emphasize that lack of
251 knowledge about the reservoir host and the chain of events that determined the human spill-
252 over prevent us from drawing any conclusion on the selective pressure underlying the limited
253 positive selection events we detected. These will need to be interpreted in the future, by
254 incorporating epidemiological, biochemical, and additional genetic data.

255 Clearly, a caveat of our analyses lies in the quality and paucity of SARS-CoV-2 genomes, as well as
256 in the limited availability of genomes of other coronaviruses closely related to SARS-CoV-2.

257 Available sequences were obtained using different methods and most likely contain errors. This is
258 unlikely to strongly affect inference of positive selection, as the frequency of all selected sites is
259 high in the SARS-CoV-2 population. Also, the SARS-CoV-2 sequences we analyzed display limited
260 diversity (with only 41 nonsynonymous polymorphisms, most of them present in one or a few

261 sequences). Thus, although the availability of additional genomes may increase the power to detect
262 selective events and the confidence with which evolutionary patterns are inferred, simply increasing
263 the number of genomes is unlikely to change the bulk of our results. However, sustained viral
264 spread in the human population will necessarily introduce new mutations in the viral population.
265 Thus, data reported herein can only depict the situation of the early phases of the human epidemic.
266 Follow-up analyses of the SARS-CoV-2 population will be required to determine the evolutionary
267 trajectories of new mutations and to assess whether and how they affect viral fitness in the human
268 hots.

269

270 **Materials and Methods**

271

272 **Sequences and alignments**

273 Genome sequences were retrieved from the National Center for Biotechnology Information
274 database (NCBI, <http://www.ncbi.nlm.nih.gov/>). Only complete or almost complete genome
275 sequences were included in the analysis (Table 1).

276 Alignments were generated using MAFFT (63), setting sequence type as codons.

277

278 **Population genetics-phylogenetic analysis**

279 Analyses were performed with gammaMap, that uses intra-species variation and inter-species
280 diversity to estimate, along coding regions, the distribution of selection coefficients (γ). In this
281 framework, γ is defined as $2PN_e s$, where P is the ploidy, N_e is effective population size, and s is the
282 fitness advantage of any amino acid-replacing derived allele (34).

283 For the eight longest ORFs in the SARS-CoV-2 genome, the corresponding coding sequence of

284 BatCoV RaTG13 was used as the outgroup.

285 We assumed θ (neutral mutation rate per site), k (transitions/transversions ratio), and T (branch
286 length) to vary within genes following log-normal distributions, whereas p (probability of adjacent
287 codons to share the same selection coefficient) following a log-uniform distribution. For each ORF
288 we set the neutral frequencies of non-STOP codons (1/61). For selection coefficients, we considered
289 a uniform Dirichlet distribution with the same prior weight for each selection class. For each ORF
290 we performed 2 runs with 100,000 iterations each and with a thinning interval of 10 iterations. Runs
291 were merged after checking for convergence.

292 The similarity plot was computed using a Kimura (two-parameter) distance model with SimPlot
293 version 3.5.1 (64). The strip gap option was set at the 50% default value. Similarity scores were
294 calculated in sliding windows of 250 bp moving with a step of 50 bp.

295

296 **Protein 3D structures and homology modeling**

297

298 The structures of SARS-CoV N (PDB ID:2CJR) (65) and S (PDB ID: 6ACG)(48) proteins were
299 obtained from the Protein Data Bank (PDB).

300 Homology modeling analysis was performed through the SWISS-MODEL server (66). The
301 accuracy of the models was examined through the GMQE (Global Model Quality Estimation) and
302 QMEAN (Qualitative Model Energy ANalysis) scores (67).

303 3D structures were rendered using PyMOL (The PyMOL Molecular Graphics System, Version
304 1.8.4.0 Schrödinger, LLC).

305

306

307 **Acknowledgments**

308

309 This work was supported by the Italian Ministry of Health (“Ricerca Corrente 2019-2020” to MS,
310 “Ricerca Corrente 2018-2020” to DF)

311

312

313 **References**

314

- 315 1. **Coronaviridae Study Group of the International Committee on Taxonomy,of Viruses.** 2020.
316 The species *Severe acute respiratory syndrome-related coronavirus*: classifying 2019-nCoV and
317 naming it SARS-CoV-2. *Nature Microbiology.* **5**:536-544. doi: 10.1038/s41564-020-0695-z.
- 318 2. **Zhu, N., D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P.**
319 **Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G. F. Gao, W. Tan, and China Novel**
320 **Coronavirus Investigating and Research Team.** 2020. A Novel Coronavirus from Patients with
321 Pneumonia in China, 2019. *N. Engl. J. Med.* **382**:727-733. doi: 10.1056/NEJMoa2001017.
- 322 3. **Cui, J., F. Li, and Z. L. Shi.** 2019. Origin and evolution of pathogenic coronaviruses. *Nat. Rev.*
323 *Microbiol.* **17**:181-192. doi: 10.1038/s41579-018-0118-9.
- 324 4. **Forni, D., R. Cagliani, M. Clerici, and M. Sironi.** 2017. Molecular Evolution of Human
325 Coronavirus Genomes. *Trends Microbiol.* **25**:35-48. doi: S0966-842X(16)30133-0.
- 326 5. **Luk, H. K. H., X. Li, J. Fung, S. K. P. Lau, and P. C. Y. Woo.** 2019. Molecular epidemiology,
327 evolution and phylogeny of SARS coronavirus. *Infect. Genet. Evol.* **71**:21-30. doi: S1567-
328 1348(19)30031-0.
- 329 6. **de Groot, R. J., S. C. Baker, R. S. Baric, C. S. Brown, C. Drosten, L. Enjuanes, R. A.**
330 **Fouchier, M. Galiano, A. E. Gorbalenya, Z. A. Memish, S. Perlman, L. L. Poon, E. J. Snijder,**
331 **G. M. Stephens, P. C. Woo, A. M. Zaki, M. Zambon, and J. Ziebuhr.** 2013. Middle East
332 respiratory syndrome coronavirus (MERS-CoV): announcement of the Coronavirus Study Group. *J.*
333 *Virol.* **87**:7790-7792. doi: 10.1128/JVI.01244-13.

- 334 7. **Gorbalenya, A. E., E. J. Snijder, and W. J. Spaan.** 2004. Severe acute respiratory syndrome
335 coronavirus phylogeny: toward consensus. *J. Virol.* **78**:7863-7866. doi: 10.1128/JVI.78.15.7863-
336 7866.2004.
- 337 8. **Zhou, P., X. L. Yang, X. G. Wang, B. Hu, L. Zhang, W. Zhang, H. R. Si, Y. Zhu, B. Li, C. L.**
338 **Huang, H. D. Chen, J. Chen, Y. Luo, H. Guo, R. D. Jiang, M. Q. Liu, Y. Chen, X. R. Shen, X.**
339 **Wang, X. S. Zheng, K. Zhao, Q. J. Chen, F. Deng, L. L. Liu, B. Yan, F. X. Zhan, Y. Y. Wang, G.**
340 **F. Xiao, and Z. L. Shi.** 2020. A pneumonia outbreak associated with a new coronavirus of probable
341 bat origin. *Nature.* **579**:270-273. doi: 10.1038/s41586-020-2012-7.
- 342 9. **Wu, F., S. Zhao, B. Yu, Y. M. Chen, W. Wang, Z. G. Song, Y. Hu, Z. W. Tao, J. H. Tian, Y. Y.**
343 **Pei, M. L. Yuan, Y. L. Zhang, F. H. Dai, Y. Liu, Q. M. Wang, J. J. Zheng, L. Xu, E. C. Holmes,**
344 **and Y. Z. Zhang.** 2020. A new coronavirus associated with human respiratory disease in China.
345 *Nature.* **579**:265-269. doi: 10.1038/s41586-020-2008-3.
- 346 10. **Hu, B., L. P. Zeng, X. L. Yang, X. Y. Ge, W. Zhang, B. Li, J. Z. Xie, X. R. Shen, Y. Z.**
347 **Zhang, N. Wang, D. S. Luo, X. S. Zheng, M. N. Wang, P. Daszak, L. F. Wang, J. Cui, and Z. L.**
348 **Shi.** 2017. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights
349 into the origin of SARS coronavirus. *PLoS Pathog.* **13**:e1006698. doi:
350 10.1371/journal.ppat.1006698.
- 351 11. **Wang, L., S. Fu, Y. Cao, H. Zhang, Y. Feng, W. Yang, K. Nie, X. Ma, and G. Liang.** 2017.
352 Discovery and genetic analysis of novel coronaviruses in least horseshoe bats in southwestern
353 China. *Emerg. Microbes Infect.* **6**:e14. doi: 10.1038/emi.2016.140.
- 354 12. **Lu, R., X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, Y. Bi,**
355 **X. Ma, F. Zhan, L. Wang, T. Hu, H. Zhou, Z. Hu, W. Zhou, L. Zhao, J. Chen, Y. Meng, J.**
356 **Wang, Y. Lin, J. Yuan, Z. Xie, J. Ma, W. J. Liu, D. Wang, W. Xu, E. C. Holmes, G. F. Gao, G.**
357 **Wu, W. Chen, W. Shi, and W. Tan.** 2020. Genomic characterization and epidemiology of 2019

- 358 novel coronavirus: implications for virus origins and receptor binding. *Lancet*. **395**:565-574. doi:
359 S0140-6736(20)30251-8.
- 360 13. **Paraskevis, D., E. G. Kostaki, G. Magiorkinis, G. Panayiotakopoulos, G. Sourvinos, and S.**
361 **Tsiodras**. 2020. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects
362 the hypothesis of emergence as a result of a recent recombination event. *Infect. Genet. Evol.*
363 **79**:104212. doi: S1567-1348(20)30044-7.
- 364 14. **Lam, T. T., M. H. Shum, H. Zhu, Y. Tong, X. Ni, Y. Liao, W. Wei, W. Y. Cheung, W. Li, L.**
365 **Li, G. M. Leung, E. C. Holmes, Y. Hu, and Y. Guan**. 2020. Identification of 2019-nCoV related
366 coronaviruses in Malayan pangolins in southern China. *Biorxiv*. 2020.02.13.945485. doi:
367 10.1101/2020.02.13.945485.
- 368 15. **Xiao, K., J. Zhai, Y. Feng, N. Zhou, X. Zhang, J. Zou, N. Li, Y. Guo, X. Li, X. Shen, Z.**
369 **Zhang, F. Shu, W. Huang, Y. Li, Z. Zhang, R. Chen, Y. Wu, S. Peng, M. Huang, W. Xie, Q.**
370 **Cai, F. Hou, Y. Liu, W. Chen, L. Xiao, and Y. Shen**. 2020. Isolation and Characterization of 2019-
371 nCoV-like Coronavirus from Malayan Pangolins. *Biorxiv*. 2020.02.17.951335. doi:
372 10.1101/2020.02.17.951335.
- 373 16. **Wong, M. C., S. J. Javornik Cregeen, N. J. Ajami, and J. F. Petrosino**. 2020. Evidence of
374 recombination in coronaviruses implicating pangolin origins of nCoV-2019. *Biorxiv*.
375 2020.02.07.939207. doi: 10.1101/2020.02.07.939207.
- 376 17. **Liu, P., J. Jiang, X. Wan, Y. Hua, X. Wang, F. Hou, J. Chen, J. Zou, and J. Chen**. 2020. Are
377 pangolins the intermediate host of the 2019 novel coronavirus (2019-nCoV) ? *Biorxiv*.
378 2020.02.18.954628. doi: 10.1101/2020.02.18.954628.
- 379 18. **Haijema, B. J., H. Volders, and P. J. Rottier**. 2003. Switching species tropism: an effective
380 way to manipulate the feline coronavirus genome. *J. Virol.* **77**:4528-4538. doi:
381 10.1128/jvi.77.8.4528-4538.2003.

- 382 19. **Kuo, L., G. J. Godeke, M. J. Raamsman, P. S. Masters, and P. J. Rottier.** 2000. Retargeting
383 of coronavirus by substitution of the spike glycoprotein ectodomain: crossing the host cell species
384 barrier. *J. Virol.* **74**:1393-1406. doi: 10.1128/jvi.74.3.1393-1406.2000.
- 385 20. **McCray, P. B., Jr, L. Pewe, C. Wohlford-Lenane, M. Hickey, L. Manzel, L. Shi, J. Netland,**
386 **H. P. Jia, C. Halabi, C. D. Sigmund, D. K. Meyerholz, P. Kirby, D. C. Look, and S. Perlman.**
387 2007. Lethal infection of K18-hACE2 mice infected with severe acute respiratory syndrome
388 coronavirus. *J. Virol.* **81**:813-821. doi: JVI.02012-06.
- 389 21. **Moore, M. J., T. Dorfman, W. Li, S. K. Wong, Y. Li, J. H. Kuhn, J. Coderre, N. Vasilieva,**
390 **Z. Han, T. C. Greenough, M. Farzan, and H. Choe.** 2004. Retroviruses pseudotyped with the
391 severe acute respiratory syndrome coronavirus spike protein efficiently infect cells expressing
392 angiotensin-converting enzyme 2. *J. Virol.* **78**:10628-10635. doi: 10.1128/JVI.78.19.10628-
393 10635.2004.
- 394 22. **Schickli, J. H., L. B. Thackray, S. G. Sawicki, and K. V. Holmes.** 2004. The N-terminal
395 region of the murine coronavirus spike glycoprotein is associated with the extended host range of
396 viruses from persistently infected murine cells. *J. Virol.* **78**:9073-9083. doi:
397 10.1128/JVI.78.17.9073-9083.2004.
- 398 23. **Li, W., M. J. Moore, N. Vasilieva, J. Sui, S. K. Wong, M. A. Berne, M. Somasundaran, J. L.**
399 **Sullivan, K. Luzuriaga, T. C. Greenough, H. Choe, and M. Farzan.** 2003. Angiotensin-
400 converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature.* **426**:450-454. doi:
401 10.1038/nature02145.
- 402 24. **Li, W., C. Zhang, J. Sui, J. H. Kuhn, M. J. Moore, S. Luo, S. K. Wong, I. C. Huang, K. Xu,**
403 **N. Vasilieva, A. Murakami, Y. He, W. A. Marasco, Y. Guan, H. Choe, and M. Farzan.** 2005.
404 Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. *Embo J.*
405 **24**:1634-1643. doi: 7600640.

- 406 25. **Wu, K., G. Peng, M. Wilken, R. J. Geraghty, and F. Li.** 2012. Mechanisms of host receptor
407 adaptation by severe acute respiratory syndrome coronavirus. *J. Biol. Chem.* **287**:8904-8911. doi:
408 10.1074/jbc.M111.325803.
- 409 26. **Qu, X. X., P. Hao, X. J. Song, S. M. Jiang, Y. X. Liu, P. G. Wang, X. Rao, H. D. Song, S. Y.**
410 **Wang, Y. Zuo, A. H. Zheng, M. Luo, H. L. Wang, F. Deng, H. Z. Wang, Z. H. Hu, M. X. Ding,**
411 **G. P. Zhao, and H. K. Deng.** 2005. Identification of two critical amino acid residues of the severe
412 acute respiratory syndrome coronavirus spike protein for its variation in zoonotic tropism transition
413 via a double substitution strategy. *J. Biol. Chem.* **280**:29588-29595. doi: M500662200.
- 414 27. **Chinese SARS Molecular Epidemiology Consortium.** 2004. Molecular evolution of the
415 SARS coronavirus during the course of the SARS epidemic in China. *Science.* **303**:1666-1669. doi:
416 10.1126/science.1092002.
- 417 28. **Lau, S. K., Y. Feng, H. Chen, H. K. Luk, W. H. Yang, K. S. Li, Y. Z. Zhang, Y. Huang, Z. Z.**
418 **Song, W. N. Chow, R. Y. Fan, S. S. Ahmed, H. C. Yeung, C. S. Lam, J. P. Cai, S. S. Wong, J. F.**
419 **Chan, K. Y. Yuen, H. L. Zhang, and P. C. Woo.** 2015. Severe Acute Respiratory Syndrome
420 (SARS) Coronavirus ORF8 Protein Is Acquired from SARS-Related Coronavirus from Greater
421 Horseshoe Bats through Recombination. *J. Virol.* **89**:10532-10547. doi: 10.1128/JVI.01048-15.
- 422 29. **Muth, D., V. M. Corman, H. Roth, T. Binger, R. Dijkman, L. T. Gottula, F. Gloza-Rausch,**
423 **A. Balboni, M. Battilani, D. Rihtaric, I. Toplak, R. S. Ameneiros, A. Pfeifer, V. Thiel, J. F.**
424 **Drexler, M. A. Muller, and C. Drosten.** 2018. Attenuation of replication by a 29 nucleotide
425 deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission.
426 *Sci. Rep.* **8**:15177. doi: 10.1038/s41598-018-33487-8.
- 427 30. **Chan, J. F., S. Yuan, K. H. Kok, K. K. To, H. Chu, J. Yang, F. Xing, J. Liu, C. C. Yip, R. W.**
428 **Poon, H. W. Tsoi, S. K. Lo, K. H. Chan, V. K. Poon, W. M. Chan, J. D. Ip, J. P. Cai, V. C.**
429 **Cheng, H. Chen, C. K. Hui, and K. Y. Yuen.** 2020. A familial cluster of pneumonia associated

- 430 with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family
431 cluster. *Lancet*. **395**:514-523. doi: S0140-6736(20)30154-9.
- 432 31. **Li, Q., X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. M. Leung, E. H. Y. Lau,**
433 **J. Y. Wong, X. Xing, N. Xiang, Y. Wu, C. Li, Q. Chen, D. Li, T. Liu, J. Zhao, M. Li, W. Tu, C.**
434 **Chen, L. Jin, R. Yang, Q. Wang, S. Zhou, R. Wang, H. Liu, Y. Luo, Y. Liu, G. Shao, H. Li, Z.**
435 **Tao, Y. Yang, Z. Deng, B. Liu, Z. Ma, Y. Zhang, G. Shi, T. T. Y. Lam, J. T. K. Wu, G. F. Gao, B.**
436 **J. Cowling, B. Yang, G. M. Leung, and Z. Feng.** 2020. Early Transmission Dynamics in Wuhan,
437 China, of Novel Coronavirus-Infected Pneumonia. In press. *N. Engl. J. Med.* . doi:
438 10.1056/NEJMoa2001316.
- 439 32. **Phan, L. T., T. V. Nguyen, Q. C. Luong, T. V. Nguyen, H. T. Nguyen, H. Q. Le, T. T.**
440 **Nguyen, T. M. Cao, and Q. D. Pham.** 2020. Importation and Human-to-Human Transmission of a
441 Novel Coronavirus in Vietnam. *N. Engl. J. Med.* **382**:872-874. doi: 10.1056/NEJMc2001272.
- 442 33. **Chinazzi, M., J. T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, A. Pastore Y**
443 **Piontti, K. Mu, L. Rossi, K. Sun, C. Viboud, X. Xiong, H. Yu, M. E. Halloran, I. M. Longini Jr,**
444 **and A. Vespignani.** 2020. The effect of travel restrictions on the spread of the 2019 novel
445 coronavirus (COVID-19) outbreak. *Science*. In press. doi: eaba9757.
- 446 34. **Wilson, D. J., R. D. Hernandez, P. Andolfatto, and M. Przeworski.** 2011. A population
447 genetics-phylogenetics approach to inferring natural selection in coding sequences. *PLoS Genet.*
448 **7**:e1002395. doi: 10.1371/journal.pgen.1002395.
- 449 35. **Ho, S. Y., R. Lanfear, L. Bromham, M. J. Phillips, J. Soubrier, A. G. Rodrigo, and A.**
450 **Cooper.** 2011. Time-dependent rates of molecular evolution. *Mol. Ecol.* **20**:3087-3101. doi:
451 10.1111/j.1365-294X.2011.05178.x.
- 452 36. **Wertheim, J. O., and S. L. Kosakovsky Pond.** 2011. Purifying selection can obscure the
453 ancient age of viral lineages. *Mol. Biol. Evol.* **28**:3355-3365. doi: 10.1093/molbev/msr170.

- 454 37. Wertheim, J. O., D. K. Chu, J. S. Peiris, S. L. Kosakovsky Pond, and L. L. Poon. 2013. A
455 case for the ancient origin of coronaviruses. *J. Virol.* **87**:7039-7045. doi: 10.1128/JVI.03273-12.
- 456 38. Snijder, E. J., E. Decroly, and J. Ziebuhr. 2016. The Nonstructural Proteins Directing
457 Coronavirus RNA Synthesis and Processing. *Adv. Virus Res.* **96**:59-126. doi: S0065-
458 3527(16)30047-1.
- 459 39. Armstrong, J., H. Niemann, S. Smeekens, P. Rottier, and G. Warren. 1984. Sequence and
460 topology of a model intracellular membrane protein, E1 glycoprotein, from a coronavirus. *Nature.*
461 **308**:751-752. doi: 10.1038/308751a0.
- 462 40. Siu, Y. L., K. T. Teoh, J. Lo, C. M. Chan, F. Kien, N. Escriou, S. W. Tsao, J. M. Nicholls, R.
463 Altmeyer, J. S. M. Peiris, R. Bruzzone, and B. Nal. 2008. The M, E, and N structural proteins of
464 the severe acute respiratory syndrome coronavirus are required for efficient assembly, trafficking,
465 and release of virus-like particles. *J. Virol.* **82**:11318-11330. doi: 10.1128/JVI.01052-08.
- 466 41. Liu, J., Y. Sun, J. Qi, F. Chu, H. Wu, F. Gao, T. Li, J. Yan, and G. F. Gao. 2010. The
467 membrane protein of severe acute respiratory syndrome coronavirus acts as a dominant immunogen
468 revealed by a clustering region of novel functionally and structurally defined cytotoxic T-
469 lymphocyte epitopes. *J. Infect. Dis.* **202**:1171-1180. doi: 10.1086/656315.
- 470 42. Pang, H., Y. Liu, X. Han, Y. Xu, F. Jiang, D. Wu, X. Kong, M. Bartlam, and Z. Rao. 2004.
471 Protective humoral responses to severe acute respiratory syndrome-associated coronavirus:
472 implications for the design of an effective protein-based vaccine. *J. Gen. Virol.* **85**:3109-3113. doi:
473 10.1099/vir.0.80111-0.
- 474 43. Narayanan, K., S. I. Ramirez, K. G. Lokugamage, and S. Makino. 2015. Coronavirus
475 nonstructural protein 1: Common and distinct functions in the regulation of host and viral gene
476 expression. *Virus Res.* **202**:89-100. doi: 10.1016/j.virusres.2014.11.019.

- 477 44. **Neuman, B. W.** 2016. Bioinformatics and functional analyses of coronavirus nonstructural
478 proteins involved in the formation of replicative organelles. *Antiviral Res.* **135**:97-107. doi:
479 10.1016/j.antiviral.2016.10.005.
- 480 45. **Brand, C. L., M. V. Cattani, S. B. Kingan, E. L. Landeen, and D. C. Presgraves.** 2018.
481 Molecular Evolution at a Meiosis Gene Mediates Species Differences in the Rate and Patterning of
482 Recombination. *Curr. Biol.* **28**:1289-1295.e4. doi: S0960-9822(18)30241-0.
- 483 46. **Hemmer, L. W., and J. P. Blumenstiel.** 2016. Holding it together: rapid evolution and positive
484 selection in the synaptonemal complex of *Drosophila*. *BMC Evol. Biol.* **16**:91. doi:
485 10.1186/s12862-016-0670-8.
- 486 47. **Li, F., W. Li, M. Farzan, and S. C. Harrison.** 2005. Structure of SARS coronavirus spike
487 receptor-binding domain complexed with receptor. *Science.* **309**:1864-1868. doi: 309/5742/1864.
- 488 48. **Song, W., M. Gui, X. Wang, and Y. Xiang.** 2018. Cryo-EM structure of the SARS coronavirus
489 spike glycoprotein in complex with its host cell receptor ACE2. *PLoS Pathog.* **14**:e1007236. doi:
490 10.1371/journal.ppat.1007236.
- 491 49. **Graham, R. L., and R. S. Baric.** 2010. Recombination, reservoirs, and the modular spike:
492 mechanisms of coronavirus cross-species transmission. *J. Virol.* **84**:3134-3146. doi:
493 10.1128/JVI.01394-09.
- 494 50. **Martin, D. P., B. Murrell, A. Khoosal, and B. Muhire.** 2017. Detecting and Analyzing
495 Genetic Recombination Using RDP4. *Methods Mol. Biol.* **1525**:433-460. doi: 10.1007/978-1-4939-
496 6622-6_17.
- 497 51. **Martin, D. P., P. Lemey, and D. Posada.** 2011. Analysing recombination in nucleotide
498 sequences. *Mol. Ecol. Resour.* **11**:943-955. doi: 10.1111/j.1755-0998.2011.03026.x.
- 499 52. **Takeda, M., C. K. Chang, T. Ikeya, P. Guntert, Y. H. Chang, Y. L. Hsu, T. H. Huang, and**
500 **M. Kainosho.** 2008. Solution structure of the c-terminal dimerization domain of SARS coronavirus

- 501 nucleocapsid protein solved by the SAIL-NMR method. *J. Mol. Biol.* **380**:608-622. doi:
502 10.1016/j.jmb.2007.11.093.
- 503 53. **Chang, C. K., M. H. Hou, C. F. Chang, C. D. Hsiao, and T. H. Huang.** 2014. The SARS
504 coronavirus nucleocapsid protein--forms and functions. *Antiviral Res.* **103**:39-50. doi:
505 10.1016/j.antiviral.2013.12.009.
- 506 54. **Surjit, M., and S. K. Lal.** 2008. The SARS-CoV nucleocapsid protein: a protein with
507 multifarious activities. *Infect. Genet. Evol.* **8**:397-405. doi: S1567-1348(07)00102-5.
- 508 55. **Wathelet, M. G., M. Orr, M. B. Frieman, and R. S. Baric.** 2007. Severe acute respiratory
509 syndrome coronavirus evades antiviral signaling: role of nsp1 and rational design of an attenuated
510 strain. *J. Virol.* **81**:11620-11633. doi: JVI.00702-07.
- 511 56. **Brockway, S. M., and M. R. Denison.** 2005. Mutagenesis of the murine hepatitis virus nsp1-
512 coding region identifies residues important for protein processing, viral RNA synthesis, and viral
513 replication. *Virology.* **340**:209-223. doi: S0042-6822(05)00377-6.
- 514 57. **Zust, R., L. Cervantes-Barragan, T. Kuri, G. Blakqori, F. Weber, B. Ludewig, and V. Thiel.**
515 2007. Coronavirus non-structural protein 1 is a major pathogenicity factor: implications for the
516 rational design of coronavirus vaccines. *PLoS Pathog.* **3**:e109. doi: 07-PLPA-RA-0063.
- 517 58. **Jauregui, A. R., D. Savalia, V. K. Lowry, C. M. Farrell, and M. G. Wathelet.** 2013.
518 Identification of residues of SARS-CoV nsp1 that differentially affect inhibition of gene expression
519 and antiviral signaling. *PLoS One.* **8**:e62416. doi: 10.1371/journal.pone.0062416.
- 520 59. **Pfefferle, S., J. Schopf, M. Kogl, C. C. Friedel, M. A. Muller, J. Carbajo-Lozoya, T.**
521 **Stellberger, E. von Dall'Armi, P. Herzog, S. Kallies, D. Niemeyer, V. Ditt, T. Kuri, R. Zust, K.**
522 **Pumpor, R. Hilgenfeld, F. Schwarz, R. Zimmer, I. Steffen, F. Weber, V. Thiel, G. Herrler, H. J.**
523 **Thiel, C. Schwegmann-Wessels, S. Pohlmann, J. Haas, C. Drosten, and A. von Brunn.** 2011.
524 The SARS-coronavirus-host interactome: identification of cyclophilins as target for pan-
525 coronavirus inhibitors. *PLoS Pathog.* **7**:e1002331. doi: 10.1371/journal.ppat.1002331.

- 526 60. **Jimenez-Guardeno, J. M., J. A. Regla-Nava, J. L. Nieto-Torres, M. L. DeDiego, C.**
527 **Castano-Rodriguez, R. Fernandez-Delgado, S. Perlman, and L. Enjuanes.** 2015. Identification
528 of the Mechanisms Causing Reversion to Virulence in an Attenuated SARS-CoV for the Design of a
529 Genetically Stable Vaccine. *PLoS Pathog.* **11**:e1005215. doi: 10.1371/journal.ppat.1005215.
- 530 61. **Oostra, M., C. A. de Haan, and P. J. Rottier.** 2007. The 29-nucleotide deletion present in
531 human but not in animal severe acute respiratory syndrome coronaviruses disrupts the functional
532 expression of open reading frame 8. *J. Virol.* **81**:13876-13888. doi: JVI.01631-07.
- 533 62. **Sung, S. C., C. Y. Chao, K. S. Jeng, J. Y. Yang, and M. M. Lai.** 2009. The 8ab protein of
534 SARS-CoV is a luminal ER membrane-associated protein and induces the activation of ATF6.
535 *Virology.* **387**:402-413. doi: 10.1016/j.virol.2009.02.021.
- 536 63. **Katoh, K., and D. M. Standley.** 2013. MAFFT multiple sequence alignment software version
537 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**:772-780. doi:
538 10.1093/molbev/mst010; 10.1093/molbev/mst010.
- 539 64. **Lole, K. S., R. C. Bollinger, R. S. Paranjape, D. Gadkari, S. S. Kulkarni, N. G. Novak, R.**
540 **Ingersoll, H. W. Sheppard, and S. C. Ray.** 1999. Full-length human immunodeficiency virus type
541 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype
542 recombination. *J. Virol.* **73**:152-160.
- 543 65. **Chen, C., C. Chang, Y. Chang, S. Sue, H. Bai, L. Riang, C. Hsiao, and T. Huang.** 2007.
544 Structure of the SARS coronavirus nucleocapsid protein RNA-binding dimerization domain
545 suggests a mechanism for helical packaging of viral RNA. *J. Mol. Biol.* **368**:1075-1086. doi:
546 10.1016/j.jmb.2007.02.069.
- 547 66. **Biasini, M., S. Bienert, A. Waterhouse, K. Arnold, G. Studer, T. Schmidt, F. Kiefer, T. Gallo**
548 **Cassarino, M. Bertoni, L. Bordoli, and T. Schwede.** 2014. SWISS-MODEL: modelling protein
549 tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* **42**:W252-8.
550 doi: 10.1093/nar/gku340.

- 551 67. **Benkert, P., M. Biasini, and T. Schwede.** 2011. Toward the estimation of the absolute quality
552 of individual protein structure models. *Bioinformatics.* **27**:343-350. doi:
553 10.1093/bioinformatics/btq662.
- 554 68. **Lei, J., Y. Kusov, and R. Hilgenfeld.** 2018. Nsp3 of coronaviruses: Structures and functions of
555 a large multi-domain protein. *Antiviral Res.* **149**:58-74. doi: 10.1016/j.antiviral.2017.11.001.
- 556 69. **Chan, J. F., K. H. Kok, Z. Zhu, H. Chu, K. K. To, S. Yuan, and K. Y. Yuen.** 2020. Genomic
557 characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with
558 atypical pneumonia after visiting Wuhan. *Emerg. Microbes Infect.* **9**:221-236. doi:
559 10.1080/22221751.2020.1719902.
- 560 70. **Coutard, B., C. Valle, X. de Lamballerie, B. Canard, N. G. Seidah, and E. Decroly.** 2020.
561 The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent
562 in CoV of the same clade. *Antiviral Res.* **176**:104742. doi: S0166-3542(20)30052-8.

563
564

565 **Figure legends**

566

567 **Figure 1. Selective patterns of SARS-CoV-2.** (A) Similarity plot (generated with SimPlot) of
568 BatCoV RaTG13 relative to SARS-CoV-2 (Wuhan-Hu-1 reference strain, NC_045512.2).
569 Similarity (Kimura distance) was calculated within sliding windows of 250 bp moving with a step
570 of 50 bp. A schematic representation of the SARS-CoV-2 genome is also shown. ORF and nsp (non-
571 structural protein) names, lengths, and relative positions are in accordance with the annotation for
572 the reference Wuhan-Hu-1 sequence. Box colors indicate the level of amino acid identity between
573 the SARS-CoV-2 and BatCoV RaTG13 sequences. Black triangles indicate amino acid changes that
574 are polymorphic in the analyzed SARS-CoV-2 genomes. Asterisks denote positively selected sites
575 and their size is proportional to the number of selected sites/region. Short ORFs with names in red

576 were not analyzed with gammaMap. Violin plots (median, white dot; interquartile range, black bar)
577 of selection coefficients (γ) for the longest (more than 80 codons) ORFs (B) and nsp3 sub-domains
578 (C) are shown. Nsp3 domains were retrieved from the SARS-CoV annotation (68).

579

580

581 **Figure 2. SARS-CoV-2 positively selected sites.** Schematic representation of the nsp1, ORF8,
582 Spike (S), and nucleocapsid (N) proteins. Positively selected sites (magenta), amino acid
583 substitutions between SARS-CoV-2 and BatCoV RaTG13 (red), and between SARS-CoV-2 and
584 pangolin-CoV MP789 (blue) are reported in the alignments.

585 The location of an insertion (insPRRA) in the spike glycoprotein is also shown. This insertion is
586 predicted to occur in the S1/S2 furin-like cleavage site (69, 70).

587

588 **Figure 3. Homology modeling of positively selected SARS-CoV-2 proteins.** Selected sites are
589 mapped onto the 3D structure of models obtained using SARS-CoV proteins as a templates (PDB
590 ID: 6ACG for panel A, 2CJR for panel B, 2HSX for panel C). Coronavirus proteins are colored in
591 hues of blue based on the most likely selection coefficient. Positively selected sites are marked in
592 red. (A) Ribbon representation of the spike glycoprotein model (one monomer is shown) in
593 complex with human ACE2 (green) (48). The binding interface is shown in the enlargement. (B)
594 Ribbon representation of the C-terminal domain of the nucleocapsid protein. (C) Ribbon
595 representation of the N-terminal portion of nsp1. Note that although some sites had the highest
596 posterior probability for $\gamma = 1$ (yellow), they were not called as positively selected because the 0.5
597 threshold was not reached.

598

599

600

601 **Table 1. List of analyzed strains.**

Strain Name	GenBank ID
Wuhan-Hu-1	NC_045512.2
2019-nCoV WHU01	MN988668.1
2019-nCoV WHU02	MN988669.1
2019-nCoV_HKU-SZ-005b_2020	MN975262.1
2019-nCoV_HKU-SZ-002a_2020	MN938384.1
SARS-CoV-2/WH-09/human/2020/CHN	MT093631.1
SARS-CoV-2/IQTC01/human/2020/CHN	MT123290.1
HZ-1	MT039873.1
BetaCoV/Wuhan/IPBCAMS-WH-01/2019	MT019529.1
BetaCoV/Wuhan/IPBCAMS-WH-03/2019	MT019531.1
BetaCoV/Wuhan/IPBCAMS-WH-02/2019	MT019530.1
BetaCoV/Wuhan/IPBCAMS-WH-04/2019	MT019532.1
BetaCoV/Wuhan/IPBCAMS-WH-05/2020	MT019533.1
WIV02	MN996527.1
WIV04	MN996528.1
WIV05	MN996529.1
WIV06	MN996530.1
WIV07	MN996531.1
SARS-CoV-2/Yunnan-01/human/2020/CHN	MT049951.1
nCoV-FIN-29-Jan-2020	MT020781.1
SARS0CoV-2/61-TW/human/2020/ NPL	MT072688.1
SNU01	MT039890.1
SARS-CoV-2/01/human/2020/SWE	MT093571.1
SARS-CoV-2/NTU01/2020/TWN	MT066175.1
SARS-CoV-2/NTU02/2020/TWN	MT066176.1
2019-nCoV/USA-WA1/2020	MN985325.1
2019-nCoV/USA-AZ1/2020	MN997409.1
2019-nCoV/USA-CA1/2020	MN994467.1
2019-nCoV/USA-CA2/2020	MN994468.1
2019-nCoV/USA-CA3/2020	MT027062.1
2019-nCoV/USA-CA4/2020	MT027063.1
2019-nCoV/USA-CA5/2020	MT027064.1
2019-nCoV/USA-CA6/2020	MT044258.1
2019-nCoV/USA-CA7/2020	MT106052.1

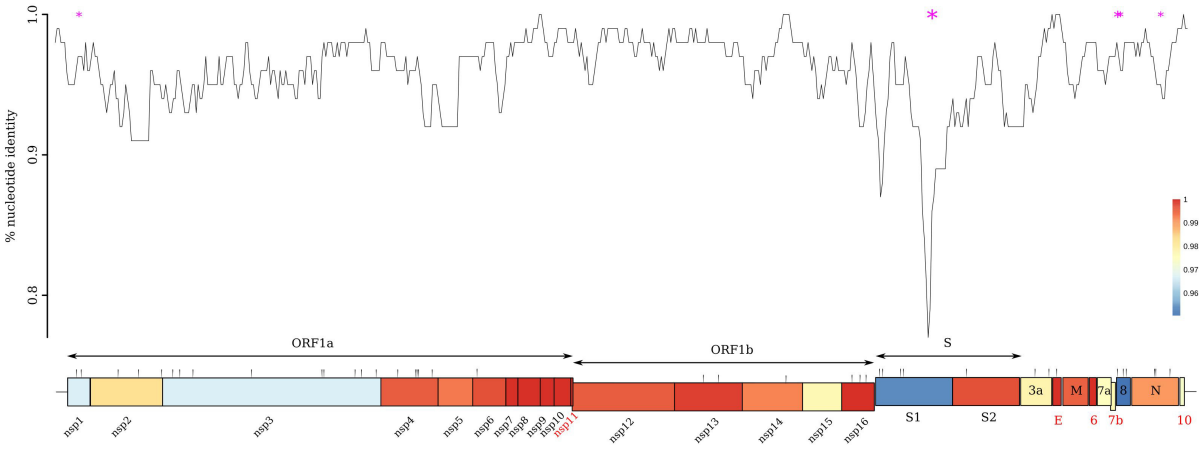
2019-nCoV/USA-CA8/2020	MT106053.1
2019-nCoV/USA-CA9/2020	MT118835.1
2019-nCoV/USA-IL2/2020	MT044257.1
2019-nCoV/USA-IL1/2020	MN988713.1
2019-nCoV/USA-MA1/2020	MT039888.1
2019-nCoV/USA-TX1/2020	MT106054.1
2019-nCoV/USA-WA1-A12/2020	MT020880.1
2019-nCoV/USA-WA1-F6/2020	MT020881.1
2019-nCoV/USA-WI1/2020	MT039887.1
Australia/VIC01/2020	MT007544.1
Bat coronavirus RaTG13	MN996532.1
Pangolin coronavirus isolate MP789	MT084071.1
Bat SARS-like coronavirus isolate bat-SL-CoVZC45	MG772933.1
Bat SARS-like coronavirus isolate bat-SL-CoVZXC21	MG772934.1
SARS-CoV tor2	NC_004718.3
SARS-CoV GZ02	AY390556.1
Bat SARS coronavirus HKU3-1	DQ022305.2
Rhinolophus affinis coronavirus isolate LYRa11	KF569996.1

602

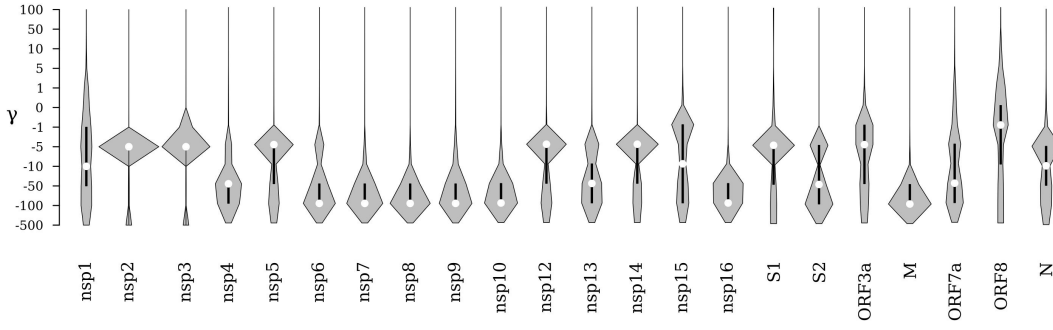
603

604

A



B



C

