Analysis

Title: Re-insights into origin and adaptation of SARS-CoV-2

Authors: Massimiliano S. Tagliamonte[1,2#], Nabil Abid[3,4#], Giovanni Chillemi[5,6], Marco Salemi[1,2*], Carla Mavian[1,2*]

*= corresponding author, cmavian@ufl.edu

*= corresponding author, salemi@pathology.ufl.edu

#= authors contributed equally to this work

Affiliations: 1. Emerging Pathogen Institute, University of Florida, Gainesville, FL, USA; 2. Department of Pathology, immunology and Laboratory Medicine, University of Florida, Gainesville, FL, USA; 3. Laboratory of Transmissible Diseases and Biological Active Substances LR99ES27, Faculty of Pharmacy, University of Monastir, Rue Ibn Sina, 5000, Monastir, Tunisia 4. High Institute of Biotechnology of Sidi Thabet, Department of Biotechnology, University of Manouba, BP-66, 2020, Ariana-Tunis, Tunisia; 5. Department for Innovation in Biological, Agro-food and Forest systems, DIBAF, University of Tuscia, via S. Camillo de Lellis s.n.c., 01100 Viterbo, Italy; 6. Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, IBIOM, CNR, Via Giovanni Amendola, 122/O, 70126, Bari, Italy

Abstract

Following the recent report of a furin-like cleavage site unique to SARS-CoV-2, we characterized additional amino acid residues, positioned within the crown of the spike (S) glycoprotein, under diversifying selection along the ancestral lineage of currently circulating pandemic strains, and affecting enhanced ability to infect host cells. Molecular dynamic simulations revealed long-range covariant movements of SARS-CoV-2 S glycoprotein monomer in pre-fusion conformation with furin cleavage, and residue T333, under purifying selection, hinge of such movement. Evolutionary analysis revealed that the current human lineage likely emerged after a recombination event between ancestors in bat and pangolin. Recombination impacted residues

located in the binding receptor region of S glycoprotein, identical to pangolin ancestor. Interestingly, several recombination hotspots were detected throughout the genome, with the exception of the S glycoprotein, which exhibits a coldspot in the N terminal region, possibly the result of negative selection against changes that may reduce its optimal fitness.

Main – With a toll of confirmed cases rapidly increasing globally, the COVID-19 pandemic is likely to meet the millions in matter of weeks. One of the features that may explain why this novel betacoronavirus (SARS-CoV-2) has been successfully spreading around the globe, is the newly acquired furin-like cleavage site[1]. Furin is ubiquitously expressed at different levels in all tissues[2,3]. However, kinetics of furin-like enzyme activity are different among species – e. g. human lung cells display quicker kinetic than fruit bat (Pteropus *alecto*) lung cells [4] – and may not fully explain SARS-CoV-2 enhanced infectivity. On the other hand, we detected several unique amino acid residues that distinguish the spike (S) glycoprotein from ancestral bat and pangolin strains, tested for selective pressure acting on these residues, and further performed homology modeling and molecular dynamic simulations to understand the impact of the cleavage site on the S glycoprotein structure, as it might be modulating infectivity and pathogenicity of the virus[7]. Our results may provide important insights on SARS-CoV-2 infectivity and pathogenesis, as well as for the development of targeted therapies and a vaccine.

SARS-CoV-2 spike (S) glycoprotein features unique to COVID-19

We defined all residues in the new SARS-CoV-2 S glycoprotein unique to the strains circulating among humans by comparing SARS-CoV-2 strains obtained from human, bat, and two pangolin lineages (lineage ‘a’ strains isolated in 2017 in China; ‘b’ strains isolated in 2019 in China) (Table S1). Before diving into the new features, we define major domains of SARS-CoV-2 S glycoprotein, based on the available SARS-CoV-1 structure [5]: S_1 subunit, that binds to the receptor on the host cell surface[6], and S_2 subunit which plays a role in membrane fusion [7]. In particular, S_1 is composed of an N-terminal domain (NTD, residues 14-294) and three C-terminal domains (CTD1, CTD2, and CTD3) (Figure 1). The CTD1 (residues 320-516 for SARS-CoV and

59 residue 333-529 for SARS-CoV-2, see Methods) functions as the receptor-binding domain (RBD),
60 which allows SARS-CoV-2 to directly bind to the peptidase domain (PD) of angiotensin-
61 converting enzyme 2 (ACE2) [8] (Figure 1). Within this region, several binding receptor residues
62 and different snapshots of SARS-CoV-2 S glycoprotein have been reported[1,9] that reflect the
63 dynamic. The recognition step is a dynamic process in which the protein surface of both partners
64 adapt one to the other, in line with cryo-EM evidences that shows how the receptor binding can
65 open up CTD1[10]. For simplicity, we show canonical ACE2 binding residues F486, Q493, S494
66 and N501, as reported by Andersen *et al*[1] (Figure 1). The S_2 subunit is located downstream of S_1
67 after the S_1/S_2 cleavage site (residue 667 for SARS-CoV and residue 660 for SARS-CoV-2). The
68 majority of the residues that differentiate human S glycoproteins from bat or pangolin were found
69 within the S_1 subunit (Figure 1a). The S glycoprotein was under strong negative/purifying
70 selection with 775 out of 1273 sites were under negative/purifying selection (Table S3), which
71 suggests the protein has reached a peak on the fitness landscape in terms of structure, stability
72 and folding[11]. We detected six amino acid positions within the S_1 subunit under
73 positive/diversifying selection, likely involved in the adaptation to new host environments (Figure
74 1 a, b): positions 32 (S32F), 50 (L50S), 483 (Q483V), and 519 (N519H). Residues 112 (A112S
75 and 352 (K253D) were also under positive selection, but emerged before the split between
76 SARS-CoV-2 human and bat (Figure 1a). Residue mutation Q483V folds in near proximity to the
77 ACE2 binding domain residues (F496-Q493-S493-N501). It is likely that V483 affects the
78 enhanced ability of currently circulating strains to target human cells (Figure 1a). *In silico*
79 prediction of the three-dimensional structure of SARS-CoV-2 S glycoprotein by homology
80 modeling shows that residue mutation Q483V folds in close proximity with residues in the CTD1
81 domain binding to ACE2 (Figure 1). All other residues are likely to be positioned in the outside of
82 the S_2 subunit crown, and therefore potentially interacting with receptor of the host cell membrane
83 (Figure 1 c, d). Interestingly, canonical ACE2 binding residues, conserved between human and
84 pangolin lineage b strains, do not appear to be under negative selection. These mutations may be
85 involved in host adaption, thus leaving with the question of when and how these residues have
86 been acquired.

87

88 **Polybasic furin cleavage and prefusion conformation**

89 While the S glycoproteins of MERS-CoV, HcoV-OC43, and HKU1 display the canonical (R/K)-
90 (2X)n-(R/K)* (or RXXR*SA) motif for the S1/S2 cleavage[12,13], the S glycoprotein of the recently
91 emerged SARS-CoV-2 presents a solvent exposed furin-like cleavage sequence (PRRARS*V) in
92 positions 681-684[1]. We carried out a 90 ns-long simulation to assess the stability of the trimeric
93 structure of SARS-CoV-2 and whether the new gained favorable furin-like cleavage site alters the
94 overall structure of the S glycoprotein.

95 As recently reported [14], the CTD1 domain of SARS-CoV-2 S glycoprotein for the pre-fusion
96 conformation rotates outward from center of the triangular head to an “up” position, while this
97 conformation for SARS-CoV is “down” [5] (Figure 1f). One monomer of the SARS-CoV-2 S
98 glycoprotein, in fact, acquired the ability of being already in a pre-fusion conformation pointing the
99 ACE2 binding site in a different angle than its predecessor, thus giving SARS-CoV-2
100 advantageous ability to bind and enter into the cell. To test whether specific evolutionary
101 pressures allowed SARS-CoV-2 to display enhanced binding to the human binding receptor, we
102 performed molecular dynamic (MD) simulations. Projection of the MD trajectory along the first
103 eigenvector, carried out by Essential Dynamics analysis [15], shows a clear up/down
104 (opening/closing) mechanism of the SARS-CoV-2 monomer by means of the rotation of CTD1
105 (see Figure 1e and Suppl. Movies 1-2). MD simulations highlighted T333 (under
106 negative/purifying selection), its first residue, as the hinge of the movement. It is worth noting that
107 the region 325-344 (mostly under negative/purifying selection) is completely conserved between
108 SARS-CoV-2 and closely related bat and pangolin strains (Figure 1a), in line with a key role of
109 this region in guiding the opening/closing of CTD1. The same analysis shows a correlated
110 movement between CTD1 (residues between red lines in Figure 1f) and the region 828-852
111 (highlighted by cyan lines in Figure 1 f), encompassing a portion of the fusion peptide (Suppl.
112 Movie 3, in which residues 828-852 are highlighted in cyan). This co-variant movement between
113 CTD1 and the fusion peptide region is indicative of long range coordinated movements, functional
114 to the protein rearrangement during the fusion step (Figure 1f, Suppl. Movie 3). A correlated

movement (peak of fluctuation) is also observed in the residues 699-718 (orange lines in Figure 1f) that comprehend furin-like cleavage site typical of SARS-CoV-2. This region shows a covariance in its movement with residues 460-496 in CTD1 and 250-263 in NTD (Suppl. movie 4, in which residues 828-852 are highlighted in orange color).

On the origin of SARS-CoV-2 lineage: recombination

While the origin of this virus strain remains unknown, numerous scenarios have been proposed[1,16,17]. Recombination is an important factor driving viral diversity [18,19], and has been reported as a possible mechanism favoring cross-species transmission of CoV and increasing its adaptability to a new host [20,21]. Recombination was a critical factor in the emergence of Middle East respiratory syndrome coronavirus (MERS-CoV) in 2012 from bat reservoirs [22,23]. There is currently conflicting evidence on whether or not the new coronavirus originated from recombination events [24-26]. Since the human lineages is 96% identical to bat-SARS-CoV-2 [1], but ACE2 binding residues F486, Q493, S494 and N501 (Figure 1a) are conserved between pangolin-SARS-CoV-2 lineage b strains, we investigated the evolutionary relationships among available full genome sequences (bat and pangolin-related SARS-CoV-2 strains, bat-SL-CoVs, SARS-CoV-1 and SARS-like strains) using a neighbor network (NNet) based algorithm [27]. NNets have the advantage over phylogeny inference algorithms of allowing internal nodes to be cyclically connected (nets) to represent the presence of conflicting phylogenetic signal in the data (i.e. data support for different tree topologies), as in fact we found in the analysis of the coronavirus genomes (Fig. 2a). Of course, conflicting phylogenetic signal in the data is not sufficient to infer recombination [28]. To test whether the NNet topology is due to recombination we used the pairwise homoplasy index (PHI) test[29], which detected highly significant signal for recombination ($p < 10^{-99}$) among all included strains (Figure 1a). SARS-CoV-1 and human SARS-CoV-2 monophyletic clades do not display net-like topology, and the PHI test on a sub-dataset including only these lineages was not significant ($p = 0.08$). SARS-like and bat-SL-CoVs, on the other hand, show a high degree of conflicting phylogenetic signals reflected by intricate networks (Fig. 1a). While SARS-like strains are ancestors of SARS-CoV-1 strains, bat-SL-CoVs stand as

143 intermediate link between SARS-CoV-1 and SARS-CoV-2 strains. Bat-SARS-CoV-2
144 (EPI_ISL_402131, isolated from bat in 2013 in China) and pangolin-SARS-CoV-2 strains
145 (EPI_ISL_410538-43 strains isolated in 2017 in China, and EPI_ISL_412860 and
146 EPI_ISL_410721 strains isolated in 2019 in China, indicated with “a” and “b”, respectively, in
147 Figure 2a) are ancestors of SARS-CoV-2 circulating in humans. When testing sub-sets of the
148 alignment including only SARS-CoV-2 and its ancestors, the PHI test still indicated strong
149 evidence of recombination ($p = 1.9x10^{-8}$), although a weaker signal for recombination was
150 detected for SARS-CoV-2 with pangolin-SARS-CoV-2 ancestral strains ($p = 0.013$). Significant
151 recombination signal was also detected ($p = 6.6x10^{-4}$) testing SARS-CoV-2 bat
152 (EPI_ISL_402131) and pangolin strains. As it is not possible to conclude whether pangolin, bat or
153 human SARS-CoV-2 strains have been recombining based on PHI test, we scanned whole
154 genomes for putative recombination breakpoints using six algorithms[30] that allow the identification
155 of specific sets of recombinant as well as parental strains (when present in the sampled
156 sequences) . A total of 68 potential recombination events distributed along the genome were
157 found to affect bat-SL-CoVs, SARS-CoVs and SARS-like CoVs strains (Table S4). Human SARS-
158 CoV-2 strains were identified as potential parental sequences recombining with pangolin-SARS-
159 CoV-2 isolated in 2019 and bat-SARS-CoV-2 strains; with a recombination event involving the
160 region of the S glycoprotein that binds to the cell receptor (Table 1). This finding matches the
161 previous observation that ACE2 binding residues are shared between pangolin and human, and
162 the recent analysis by Li *et al* [26]. While our analysis could not confidently discern the recombinant
163 from the parent sequences, signature pattern analysis (Figure 1a) and a Maximum likelihood tree
164 calculated using this recombinant region only (data not shown), clearly showed pangolin and
165 COV-2 clustering together. Our overall findings point to the conclusion that a recombination event
166 between ancestors of human SARS-CoV-2, or a closely related to bat-SARS-CoV-2, and
167 ancestors of pangolin-SARS-CoV-2 lineage b strains, is the likely origin of the currently circulating
168 pandemic strains.
169

170 Using coordinates of identified recombination breakpoints, we plotted their distribution and
171 analyzed it through permutations to predict hot- and coldspots of recombination across the
172 coronavirus genomes (Figure 2b). A first iteration, performed with a sliding window of size 100,
173 identified nine potential recombination hotspots. Seven of them were found in the polyprotein
174 segment (Figure 2b), which spans three quarters of the viral RNA; the remaining two involved
175 protein E, *orf8*, and protein N. It has been reported that the *orf8* gene plays a role in regulating
176 the viral replication in the host cell[31], and was probably acquired by the whole lineage of SARS-
177 CoV-1 from the SARS-related CoV from Greater Horseshoe bats (SARSr-Rf-BatCoVs) following
178 a recombination event with a Chinese horseshoe bat coronavirus (SARSr-Rs-BatCoV)[32]. A
179 second analysis, with a sliding window of size 200, detected two recombination coldspots on the
180 N terminal region of the glycoprotein (between residues S31 to V47, and Y145 to G184, $p \leq 0.01$)
181 (Figure 2b). Under more relax assumptions ($0.01 \leq p \leq 0.05$) coldspot region spans from Q14 to
182 S71, and from V127 to T240 (Figure 2b). Residues S32F, A112S, N519H (Figure 1) are outside
183 the coldspot region, however L50S is inside the first coldspot region of SARS-CoV-2 S
184 glycoprotein. Our finding seems to be in conflict with high recombination rates being found in non-
185 human coronaviruses[33], yet, it is not surprising that part of a gene playing a fundamental role in
186 cell entry and fusion would sit into a recombination coldspot. While the majority of the CoV
187 genome seems prone to recombination, the S glycoprotein is the only one that harbored
188 recombination coldspots. This might depend on functional costraints on specific domains of the
189 protein [34], or other factors impacting the probability of recombination events, such as RNA
190 secondary structure [35,36]. Another possibility is that the N terminal, under negative selection,
191 reached a peak on the adaptive landscape and new mutations/rearrangements are unlikely to be
192 fixed in the next generations due to negative selection. These findings correlate with mobility of
193 MD simulations, where flexible stretches are less conserved, while hinge regions remained
194 unchanged during evolution of the lineage.
195
196 **4 – Conclusions**

197 We provided molecular insights on SARS-CoV-2 infectivity and pathogenesis and identified
198 potential targets for therapies, in the context of coronaviruses evolutionary trajectory that has led
199 to the current pandemic.
200 SARS-CoV-2 S glycoprotein presents two notable changes: a favorable furin-like cleavage site [1],
201 and CTD1 up in pre-fusion mode [14]. On the contrary S glycoprotein of previously circulating
202 SARS-CoV strains do not have a favorable furin-like cleavage site [1] and remain un-cleaved after
203 biosynthesis[37]. Furin has the potential to cleave specifically viral envelope glycoproteins, to
204 enhance viral fusion with host cell membranes[38], thereby any variation/mutational changes in
205 amino acid composition at cleavage sites may impact tissue and cell tropism, host range, and
206 pathogenesis[39], as reported for other respiratory viruses[40-42]. Moreover, our MD simulations
207 correlated furin-like cleavage site correlated with pre-fusion conformation. Simulations indicated
208 that the motion of CTD1 is quite visible and is the most relevant motion in the protein: a
209 conserved region that acts as a hinge in the opening/closing mechanism. We described residue
210 T333, under purifying selection, as the hinge of the rotation movement that gives further
211 advantage to bind cell receptors facilitating fusion and entry. Further analysis will be done to
212 analyze the conformation change of the S glycoprotein and the residues interface in nearly real
213 conditions. Development of inhibitors of such motion may reduce substantially pre-fusion
214 conformation and therefore the infectivity of SARS-CoV-2.
215 SARS-CoV-2 S glycoprotein gained specific features that have facilitated the successful spread
216 of the virus as compared to its predecessors. The assumption that the insertion of a furin
217 cleavage site allowed a bat CoV to gain the ability to infect humans facilitating the host specie
218 jump once the conditions were favorable cannot be excluded. By structure similarity with S
219 glycoprotein of mouse hepatitis virus, it is possible that SARS-CoV-2 also uses distal loops of the
220 N-terminal domain to interact with cell receptor [43]. Although more evidence is needed, it is
221 conceivable that residues in this region that are under positive selection may have been selected
222 to immune-modulate the host. Several recombination events involving ancestral pangolin- and
223 bat- SARS-CoV-2 lineages. According to our results, one of these events led to the current
224 configuration of SARS-CoV-2 ACE2 binding residues, including recently described residues

225 forming hydrogen bonds with the host cell receptor [9]. In spite of the frequency of recombination
226 events involving the S glycoprotein, the 5' end of the gene might sit into a recombination coldspot.
227 While mutation/recombination events in the glycoprotein RBD might expand the virus tropism,
228 recombination coldspots and negative/diversifying selective pressure limit its variability to
229 maintain correct structure, and therefore function. RDB is definitively a good target for therapeutic
230 development, while regions in coldspots or undergoing negative/diversifying selective pressure
231 may be helpful for successful vaccine design. Our candidate *in silico* recombination hotspots and
232 coldspots findings need to be further investigated with *in vitro* experiments.
233 We conclude that SARS-CoV-2 likely emerged from recombination events that involves ancestral
234 strains circulating among bat and pangolin. However, we could not completely distinguish the
235 recombinant sequence from the parental ones in other cases. Recombination signal and selective
236 pressures can interfere with each other [44], with positive selected mutations near potential
237 recombination breakpoints [45]. More extensive sampling of wildlife CoVs would help in clarifying
238 the events that led to the jump to humans. Future emergence of novel recombinant CoVs, and
239 other viruses as well, should be a constant concern; strengthening surveillance of viral
240 populations and improving our understanding of their dynamics will help the prevention of new
241 zoonotic outbreaks and spread of emerging pathogens.
242
243

Methods

Genetic data

Publicly available genomic CoVs sequences were obtained from GenBank and GISAID (Table S1 and S5). To gather closely related CoVs BLASTn was performed using the ViPR database and keeping the first 100 hits. Genome sequences were aligned using MAFFT and refined manually[49]. Alignments for the glycoprotein were extracted from the genome alignment.

Protein selection and signature pattern

Because of the fast rate at which new genomes have been deposited in GISAID, this analysis was updated to 728 genomes (Table S5). While in the recombination analysis the number of genomes is not relevant, as they are very similar to each other, for selection analysis and identification of new features, a higher number of strains permits to explore a larger evolutionary landscape. HyPhy [46] algorithm mixed effects model of evolution (MEME) [47] and Fast Unconstrained Bayesian AppRoximation (FUBAR)[48] were used to estimate the ratio of non-synonymous (*dN*) to synonymous (*dS*) substitution rates for each codon. MEME was employed to identify individual sites subject to episodic diversifying selection, and FUBAR for inferring site-specific pervasive selection. Sites were considered to have experienced statistically significant positive or negative selection based on the following cut-offs: LRT ≤ 0.05 or posterior probability (PP) > 0.90, for MEME and FUBAR respectively. VESPA algorithm[49] was used to detect the residue signature pattern distinguishing SARS-CoV-2 strains from the closely related bat (EPI_ISL_402131, 2013, Accession MN996532.1) and pangolin (EPI_ISL_410538, 2017; EPI_ISL_410539, 2017; EPI_ISL_410540, 2017; EPI_ISL_410541, 2017; EPI_ISL_410542, 2017; EPI_ISL_410543, 2017; EPI_ISL_412860, 2019; EPI_ISL_410721, 2019) strains. Positions are numbered with SARS-CoV-2 isolate Wuhan-Hu-1 (Accession MN908947.3) as reference.

Structure homology modeling

As a first step, the amino acid sequences of SARS-CoV-2 Wuhan-Hu-1 (Accession QHD43416) and EPI_ISL_402131 (bat, 2013, Accession MN996532.1) were submitted to Swiss-model [50]. The

272 best models showing high values of Global Model Quality Estimation (GMQE) were selected and
273 used as templates. Amino acid alignments of targets and templates were performed using
274 EMBOSS Needle [51]. The 3D homology homo-trimer models were created with the Modeller v9.23
275 [52,53] *automodel* class using the alignment as a guide. Then Discreet Optimised Protein Energy
276 Score (DOPE) based model selection [54] and refinements were conducted using Modeller v9.23
277 scripts. The final structure model validation was conducted with ProSA [55,56] and QMEAN servers
278 [57,58]. Visualization of the atomic models, including figures and movies, is made with Chimera
279 v1.12[59].
280
281 **Molecular Dynamics simulation**
282 The structure of SARS-CoV-2, obtained for homology modeling, was used as starting structure
283 for Molecular Dynamics (MD) simulations. In detail, topology file was generated with the pdb2gmx
284 Gromacs tool[60], using the amber99sb forcefield [61].The protein was embedded in a triclinic box,
285 extending up to 15 Å from the solute, and immersed in TIP3P water molecules [62]. Counter ions
286 were added to neutralize the overall charge with the genion gromacs tool. After energy
287 minimizations, the systems were slowly relaxed for 5 ns by applying positional restraints of 1000
288 kJ mol^{-1} nm^{-2} to the protein atoms. Then unrestrained MD simulations were carried out for a
289 length of 90 ns with a time step of 2 fs with GROMACS 2018.3 simulation package (Pall et al.,
290 2015) (supercomputer Galileo, CINECA, Bologna, Italy). V-rescale temperature coupling was
291 employed to keep the temperature constant at 300 K [63]. The Particle-Mesh Ewald method was
292 used for the treatment of the long-range electrostatic interactions [64]. The first 5 ns portion of the
293 trajectory was excluded from the analysis.
294 We carried out the Essential Dynamics analysis to identify the main 3N directions along which the
295 majority of the protein motion is defined [15]. This analysis is based on the diagonalization of the
296 covariance matrix built from the atomic fluctuations after the removal of the translational and
297 rotational movement and we applied it on the 1120 c-alpha atoms of each monomer (residue 27-
298 1146 in our model), after concatenating their trajectories (255 ns total sampling).
299

300 **Recombination breakpoint and hotspot analyses**

301 The presence of conflicting phylogenetic signals (i.e. distinct tree topologies compatible with the
302 same set of aligned sequences) in the full genome data set was investigated, first, by inferring
303 networks using neighbor-net [27] implemented in SplitsTree4 [28]. Significant presence of
304 recombination signal was then assessed with the pairwise homoplasy index (Phi) test [29].
305 Identification of putative recombinant events and associated breakpoints was performed using the
306 RDP, GENECONV, MaxChi, CHIMAERA, SIScan, and 3Seq algorithms implemented in the
307 RDP4 [30] software (http://web.cbio.uct.ac.za/~darren/rdp.html). Statistical evidence of
308 recombination was indicated by *p*-values < 0.05, and a putative recombination events were
309 considered as such if supported by at least five of the six algorithms used. Full table with
310 recombination results id given in Table S4.
311 Recombination breakpoints hotspot analysis was also performed using RDP4 through a
312 permutation test, which takes into account number of variable positions between the three
313 sequences involved in the event when performing simulations. Two iteration were performed, with
314 sliding windows of size 100 and 200, with different sensitivity to hotspots and coldspots.
315 Segments that had more observed breakpoints than the local maximum number of breakpoints of
316 99% of 1,000 permutaions were considered hotspots (p=0.01); in the same fashion, segments
317 that had less breakpoints than the local minimum of 99% of permutations were considered
318 coldspots [65].
319
320

Figures and Tables

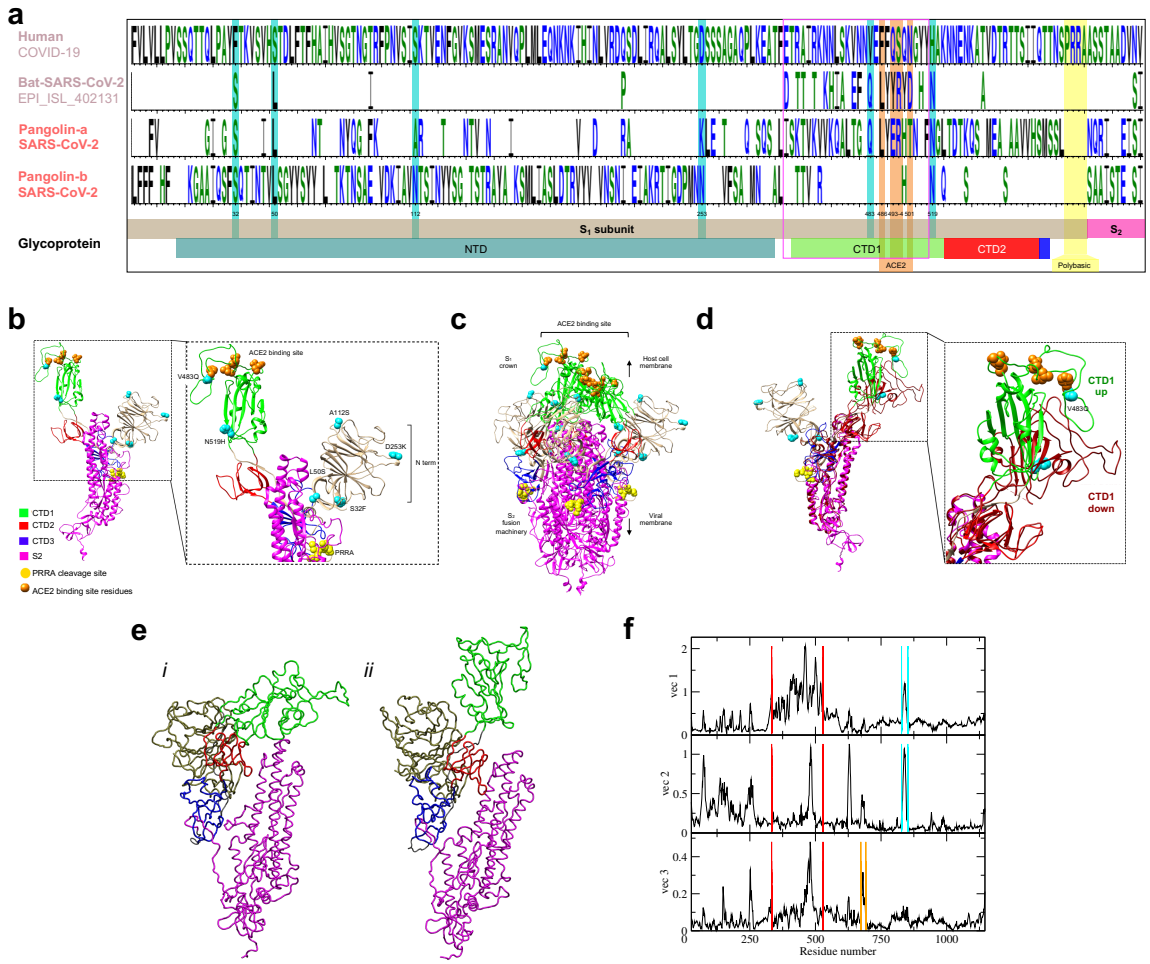

Figure 1. Signature residues in the S glycoprotein of SARS-CoV-2 under positive selection, ribbon representations of the S protomer and MD simulations. (a) Signature residues that distinguish human SARS-COV-2 from closely related bat and pangolin (lineage 'a' strains isolated in 2017 in China; 'b' strains isolated in 2019 in China) strains; residues are colored according their charge: hydrophilic (RKDENQ) residues in blue, neutral (SGHTAP) in green, and hydrophobic (YVMCLFIW) in black. Numbers correspond to the residue position in Wuhan-Hu-1 MN908947.3 strain. In the box below we show a schematic representation of the spike glycoprotein is given: S_1 and the S_2 subunits of the S glycoprotein are indicated in grey and fuchsia, while CTD1 is shown green, CTD2 in red, CTD3 in blue. Fuchsia box around the residues between 333-516 are indicating the residues involved into rotation of CTD1. Residues

333 under positive selection are shaded in cyan, residues that form the ACE receptor binding are
334 shaded in orange, and the acquired polybasic residues (PRRA) in the S_1/S_2 cleavage site in
335 yellow. (b) The monomer of SARS-CoV-2 S glycoprotein and its three S1 domains (CTD1, CTD2,
336 and CTD3) and S2 subunit (pink) are shown in different colors: CTD1 in green, CTD2 in red,
337 CTD3 in blue. The N-terminal domain was shown is colored tan. The residues under selection,
338 ACE2 biding site residues, and S1/S2 cleavage sites were shown with cyan, orange and yellow
339 spheres, respectively. (c) Homo-trimer structure of the modeled SARS-CoV-2. Residues, S_1
340 (NTD, CTD1, CTD2, and CTD3) and the S_2 subunits are colored corresponding to panel b. (d)
341 The superposition of the modeled S protomer of SARS-CoV-2 with the S protomer of SARS-CoV
342 (PDB 5WRG). The missing NTD domain for SARS-CoV was not shown whereas the NTD domain
343 of SARS-CoV-2 is shown by tan color. Residues and domains are colored corresponding to panel
344 b. (e) Projection of the SARS-CoV-2 S glycoprotein MD trajectory along Essential Dynamics
345 eigenvector 1 (see also Supplementary movies 1-2). The two extreme conformations are shown
346 in i and ii. Domain colors as in panel b. The opening/closing mechanism of the SARS-CoV-2 S
347 glycoprotein monomer by means of the rotation of CTD1 is evident. (f) Per-residue Root Mean
348 Square Fluctuations in nm of the filtered S glycoprotein trajectory along eigenvectors 1-3. The
349 region with higher fluctuations along eigenvector 1 is the CTD1, highlighted by red lines, since its
350 rotation in the opening/closing mechanism is the most correlated motion in the protein. The peak
351 in fluctuations observed in eigenvectors 1 and 2, involving residues 828-852, is highlighted by
352 cyan lines and is indicative of a correlated motion of this region with the opening/closing motion of
353 CTD1. Note that this region includes a portion of the fusion peptide (Suppl. Movie 3). The peak in
354 fluctuations observed in eigenvector 3 involving residues 699-718 is highlighted by orange lines.
355 Note that this region includes all the furin-like cleavage site typical of SARS-CoV-2. Fluctuations
356 along eigenvector 3 shows a correlated motion between the furin-like cleavage site, residues 460-
357 496 in CTD1, and 250-263 in NTD (Suppl. Movie 4).
358
359

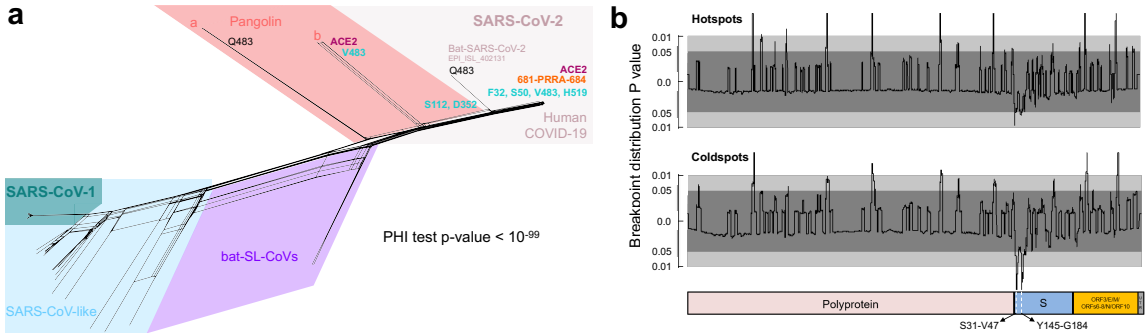

Figure 2. Recombination hot- and cold- spots prediction across CoV genomes. (a) NNet graphs obtained using split-decomposition and based on the full genome sequence to detect recombination. P-value of PHI test of recombination is reported. Pangolin taxa: (a) EPI_ISL_410538-43 strains isolated in 2017 in China; (b) EPI_ISL_412860 and EPI_ISL_410721 strains isolated in 2019 in China. (b) Hotspot (upper graph) or coldspot (lower graph) presence across the genome. P-values ≤ 0.01 indicate presence of either hotspots (upper graph) or coldspots (lower graph).

369

Breakpoint in codon coordinates in aligned SARS-COV-2		Breakpoints codon coordinates in recombinant sequence				
Start	End	Start	End	Recombinant Sequence	Minor Parental Sequence	Major Parental Sequence
in poly-protein	A626	A6892 (poly-protein)	A601	^MG772934_Bat_coV_SARS-like_bat-SL-CoVZXC21_Host:Ba	KY417146_Bat_coV_SARS-like_Rs4231_Host:Bat	Unknown (KP886809_Bat_coV_SARS-likeYNLF_34C_Host:Bat)
D111	K304	D109	K298	EPI_ISL_410721_BetaCoV_pangolin_Guandong_1_2019	Unknown (EPI_ISL_410543_BetaCoV_pangolin_Guangxi_P3B_2017)	EPI_ISL_402131_BetaCoV_bat_Yunnan_RaTG13_2013
G339	P579	H335	P556	MG772933_Bat_coV_SARS-like_bat-SL-CoVZC45_Host:Bat	GQ153542_Bat_coV_SARS_HKU3-7_Host:Unknown	EPI_ISL_410721_BetaCoV_pangolin_Guandong_1_2019
T430	V511	T429	V510	^EPI_ISL_402131_BetaCoV_bat_Yunnan_RaTG13_2013	Unknown (EPI_ISL_412860_BetaCoV_pangolin_China_MP789_2019)	EPI_ISL_410535_BetaCoV_Singapore_4_2020
T768	I993	S741	I966	^MG772933_Bat_coV_SARS-like_bat-SL-CoVZC45_Host:Bat	Unknown (EPI_ISL_410538_BetaCoV_pangolin_Guangxi_P4L_2017)	EPI_ISL_412860_BetaCoV_pangolin_China_MP789_2019
L849	Q992	L818	Q961	^KF294457_SARS-relatedHost:Bat	MG772934_Bat_coV_SARS-like_bat-SL-CoVZXC21_Host:Ba	KJ473814_BtRs-BetaCoV_HuB2013Host:Bat

370

371

372 **Table 1.** Potential recombination breakpoints of CoVs detected using the RDP, GENECONV, BootScan, MaxChi, CHIMAERA, SIScan, and 3Seq

373 algorithms affecting ancestors of SARS-CoV-2 S protein. Significant p-values are indicated (alpha = 0.05). ~ = It is possible that this apparent

recombination signal could have been caused by an evolutionary process other than recombination. * = The actual breakpoint position is undetermined (it was most likely overprinted by a subsequent recombination event). ^ = The recombinant sequence may have been misidentified (one of the identified parents might be the recombinant) Minor Parent = Parent contributing the smaller fraction of sequence. Major Parent = Parent contributing the larger fraction of sequence. Unknown = Only one parent and a recombinant need be in the alignment for a recombination event to be detectable. The sequence listed as unknown was used to infer the existence of a missing parental sequence. NS = No significant P-value was recorded for this recombination event using this method.

380 **Supplementary information**

381

382 **Table S1.** List of strains sequences from GISAID and GenBank and used for recombination
383 analysis. See Excel file

384

385 **Table S2**. Signature pattern analyses for Figure 1 panel a. See Excel file

386

387 **Table S3**. List of protein S glycoprotein codons under negative selection. See Excel file

388

389 **Table S4**. Full list of potential recombination breakpoints of CoVs detected by the RDP,
390 GENECONV, BootScan, MaxChi, CHIMAERA, SIScan, and 3Seq algorithms. See Excel file

391

392 **Table S5.** List of strains sequences from GISAID and GenBank and used for selection and
393 signature pattern analyses. See Excel file

394

395 **Supplementary Movie S1**. Projection of the S glycoprotein monomer trajectory along
396 eigenvector 1. Tube visualization of c-alpha atoms. Domain colors as in Figure 1. The
397 opening/closing mechanism of the SARS-CoV-2 monomer by means of the rotation of CTD1 is
398 evident.

399

400 **Supplementary Movie S2**. Projection of the S glycoprotein monomer trajectory along
401 eigenvector 1 in a different orientation.

402

403 **Supplementary Movie S3**. Projection of the S glycoprotein monomer trajectory along
404 eigenvector 1. Residues 828-852 (cyan color), including a portion of the fusion peptide, show
405 correlated motion with the opening/closing rotation of CTD1.

406

407 **Supplementary Movie S4**. Projection of the S glycoprotein monomer trajectory along
408 eigenvector 3. Residues 699-718 (orange color), comprehending all the furin-like cleavage site
409 typical of SARS-CoV-2, show correlated motion with residues 460-496 in CTD1, and 250-263 in
410 NTD.
411
412

Acknowledgements

MS is supported in part by the Stephany W. Holloway University Chair in AIDS Research. This work was supported by the Tunisian Ministry of Higher Education and Scientific Research, the 'Departments of Excellence-2018' Program (Dipartimenti di Eccellenza) of the Italian Ministry of Education, University and Research, DIBAF-Department of University of Tuscia, Project 'Landscape 4.0–food, wellbeing and environment'. We thank all those who have contributed SARS-CoV-2 genome sequences to the GISAID database (https://www.gisaid.org/). We acknowledge Cineca and ELIXIR-IIB for computing resources, and David Ostrov for his helpful suggestions and comments.

References

1 Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nature Medicine*, doi:10.1038/s41591-020-0820-9 (2020).

2 Schalken, J. A. *et al.* fur gene expression as a discriminating marker for small cell and nonsmall cell lung carcinomas. *J Clin Invest* **80**, 1545-1549, doi:10.1172/JCI113240 (1987).

3 Hatsuzawa, K. *et al.* Structure and expression of mouse furin, a yeast Kex2-related protease. Lack of processing of coexpressed prorenin in GH4C1 cells. *J Biol Chem* **265**, 22075-22078 (1990).

4 El Najjar, F., Lampe, L., Baker, M. L., Wang, L. F. & Dutch, R. E. Analysis of cathepsin and furin proteolytic enzymes involved in viral fusion protein activation in cells of the bat reservoir host. *PLoS One* **10**, e0115736, doi:10.1371/journal.pone.0115736 (2015).

5 Gui, M. *et al.* Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. *Cell Res* **27**, 119-129, doi:10.1038/cr.2016.152 (2017).

6 Li, F. Structure, Function, and Evolution of Coronavirus Spike Proteins. *Annu Rev Virol* **3**, 237-261, doi:10.1146/annurev-virology-110615-042301 (2016).

7 Bosch, B. J., van der Zee, R., de Haan, C. A. & Rottier, P. J. The coronavirus spike protein is a class I virus fusion protein: structural and functional characterization of the fusion core complex. *J Virol* **77**, 8801-8811, doi:10.1128/jvi.77.16.8801-8811.2003 (2003).

8 Li, F., Li, W., Farzan, M. & Harrison, S. C. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science* **309**, 1864-1868, doi:10.1126/science.1116480 (2005).

9 Yan, R. *et al.* Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* **367**, 1444-1448, doi:10.1126/science.abb2762 (2020).

10 Song, W., Gui, M., Wang, X. & Xiang, Y. Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. *PLoS Pathog* **14**, e1007236, doi:10.1371/journal.ppat.1007236 (2018).

11 Hoffmann, J., Wrabl, J. O. & Hilser, V. J. The role of negative selection in protein evolution revealed through the energetics of the native state ensemble. *Proteins* **84**, 435-447, doi:10.1002/prot.24989 (2016).

12 Chan, C. M. *et al.* Spike protein, S, of human coronavirus HKU1: role in viral life cycle and application in antibody detection. *Exp Biol Med (Maywood)* **233**, 1527-1536, doi:10.3181/0806-RM-197 (2008).

13 Millet, J. K. & Whittaker, G. R. Host cell entry of Middle East respiratory syndrome coronavirus after two-step, furin-mediated activation of the spike protein. *Proc Natl Acad Sci U S A* **111**, 15214-15219, doi:10.1073/pnas.1407087111 (2014).

14 Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260-1263, doi:10.1126/science.abb2507 (2020).

15 Amadei, A., Linssen,A.B. and Berendsen,H.J. . Essential dynamics of proteins. . *Proteins,* **17,** , 412–425. (1993).

16 MacLean, O. A., Orton, R., Singer, J. B. & Robertson, D. L. *Response to "On the origin and continuing evolution of SARS-CoV-2"*, <http://virological.org/t/response-to-on-the-origin-and-continuing-evolution-of-sars-cov-2/418> (2020).

17 Tang, X. *et al.* On the origin and continuing evolution of SARS-CoV-2. *National Science Review*, doi:10.1093/nsr/nwaa036 (2020).

18 Simon-Loriere, E. & Holmes, E. C. Why do RNA viruses recombine? *Nature Reviews Microbiology* **9**, 617-626, doi:10.1038/nrmicro2614 (2011).

19 Mavian, C. *et al.* Emergence of recombinant Mayaro virus strains from the Amazon basin. *Sci Rep* **7**, 8718, doi:10.1038/s41598-017-07152-5 (2017).

20 Graham, R. L. & Baric, R. S. Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. *J Virol* **84**, 3134-3146, doi:10.1128/JVI.01394-09 (2010).

21 Su, S. *et al.* Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. *Trends Microbiol* **24**, 490-502, doi:10.1016/j.tim.2016.03.003 (2016).

22 Dudas, G. & Rambaut, A. MERS-CoV recombination: implications about the reservoir and potential for adaptation. *Virus Evol* **2**, vev023, doi:10.1093/ve/vev023 (2016).

23 Anthony, S. J. *et al.* Further Evidence for Bats as the Evolutionary Source of Middle East Respiratory Syndrome Coronavirus. *mBio* **8**, e00373-00317, doi:10.1128/mBio.00373-17 (2017).

24 Paraskevis, D. *et al.* Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. *Infect Genet Evol* **79**, 104212, doi:10.1016/j.meegid.2020.104212 (2020).

25 Ji, W., Wang, W., Zhao, X., Zai, J. & Li, X. Homologous recombination within the spike glycoprotein of the newly identified coronavirus may boost cross-species transmission from snake to human. *J Med Virol*, doi:10.1002/jmv.25682 (2020).

26 Li, X. *et al.* Emergence of SARS-CoV-2 through Recombination and Strong Purifying Selection. *bioRxiv*, 2020.2003.2020.000885, doi:10.1101/2020.03.20.000885 (2020).

503 27 Bryant, D. & Moulton, V. Neighbor-net: an agglomerative method for the
504 construction of phylogenetic networks. *Mol Biol Evol* **21**, 255-265,
505 doi:10.1093/molbev/msh018 (2004).
506 28 Huson, D. H. & Bryant, D. Application of phylogenetic networks in
507 evolutionary studies. *Mol Biol Evol* **23**, 254-267,
508 doi:10.1093/molbev/msj030 (2006).
509 29 Bruen, T. C., Philippe, H. & Bryant, D. A simple and robust statistical test for
510 detecting the presence of recombination. *Genetics* **172**, 2665-2681,
511 doi:10.1534/genetics.105.048975 (2006).
512 30 Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: Detection
513 and analysis of recombination patterns in virus genomes. *Virus Evol* **1**,
514 vev003, doi:10.1093/ve/vev003 (2015).
515 31 Muth, D. *et al.* Attenuation of replication by a 29 nucleotide deletion in SARS-
516 coronavirus acquired during the early stages of human-to-human
517 transmission. *Sci Rep* **8**, 15177, doi:10.1038/s41598-018-33487-8 (2018).
518 32 Lau, S. K. *et al.* Severe Acute Respiratory Syndrome (SARS) Coronavirus ORF8
519 Protein Is Acquired from SARS-Related Coronavirus from Greater Horseshoe
520 Bats through Recombination. *J Virol* **89**, 10532-10547,
521 doi:10.1128/JVI.01048-15 (2015).
522 33 Rowe, C. L. *et al.* Generation of coronavirus spike deletion variants by high-
523 frequency recombination at regions of predicted RNA secondary structure. *J*
524 *Virol* **71**, 6183-6190 (1997).
525 34 Simon-Loriere, E. *et al.* Molecular mechanisms of recombination restriction in
526 the envelope gene of the human immunodeficiency virus. *PLoS Pathog* **5**,
527 e1000418, doi:10.1371/journal.ppat.1000418 (2009).
528 35 Svarovskaia, E. S., Delviks, K. A., Hwang, C. K. & Pathak, V. K. Structural
529 determinants of murine leukemia virus reverse transcriptase that affect the
530 frequency of template switching. *J Virol* **74**, 7171-7178,
531 doi:10.1128/jvi.74.15.7171-7178.2000 (2000).
532 36 Gao, L., Balakrishnan, M., Roques, B. P. & Bambara, R. A. Insights into the
533 multiple roles of pausing in HIV-1 reverse transcriptase-promoted strand
534 transfers. *J Biol Chem* **282**, 6222-6231, doi:10.1074/jbc.M610056200 (2007).
535 37 Belouzard, S., Chu, V. C. & Whittaker, G. R. Activation of the SARS coronavirus
536 spike protein via sequential proteolytic cleavage at two distinct sites. *Proc*
537 *Natl Acad Sci U S A* **106**, 5871-5876, doi:10.1073/pnas.0809524106 (2009).
538 38 Izaguirre, G. The Proteolytic Regulation of Virus Cell Entry by Furin and
539 Other Proprotein Convertases. *Viruses* **11**, doi:10.3390/v11090837 (2019).
540 39 Klenk, H. D. & Garten, W. Host cell proteases controlling virus pathogenicity.
541 *Trends Microbiol* **2**, 39-43, doi:10.1016/0966-842x(94)90123-6 (1994).
542 40 Cheng, J. *et al.* The S2 Subunit of QX-type Infectious Bronchitis Coronavirus
543 Spike Protein Is an Essential Determinant of Neurotropism. *Viruses* **11**,
544 doi:10.3390/v11100972 (2019).
545 41 Claas, E. C. *et al.* Human influenza A H5N1 virus related to a highly pathogenic
546 avian influenza virus. *Lancet* **351**, 472-477, doi:10.1016/S0140-
547 6736(97)11212-0 (1998).

548 42 Kido, H. *et al.* Role of host cellular proteases in the pathogenesis of influenza
549 and influenza-induced multiple organ failure. *Biochim Biophys Acta* **1824**,
550 186-194, doi:10.1016/j.bbapap.2011.07.001 (2012).
551 43 Peng, G. *et al.* Structural and Molecular Evidence Suggesting Coronavirus-
552 driven Evolution of Mouse Receptor. *J Biol Chem* **292**, 2174-2181,
553 doi:10.1074/jbc.M116.764266 (2017).
554 44 Simon-Loriere, E., Martin, D. P., Weeks, K. M. & Negroni, M. RNA structures
555 facilitate recombination-mediated gene swapping in HIV-1. *J Virol* **84**, 12675-
556 12682, doi:10.1128/JVI.01302-10 (2010).
557 45 Dalmon, A. *et al.* Evidence for positive selection and recombination hotspots
558 in Deformed wing virus (DWV). *Scientific Reports* **7**, 41045,
559 doi:10.1038/srep41045 (2017).
560 46 Pond, S. L., Frost, S. D. & Muse, S. V. HyPhy: hypothesis testing using
561 phylogenies. *Bioinformatics* **21**, 676-679, doi:10.1093/bioinformatics/bti079
562 (2005).
563 47 Murrell, B. *et al.* Detecting individual sites subject to episodic diversifying
564 selection. *PLoS Genet* **8**, e1002764, doi:10.1371/journal.pgen.1002764
565 (2012).
566 48 Murrell, B. *et al.* FUBAR: a fast, unconstrained bayesian approximation for
567 inferring selection. *Mol Biol Evol* **30**, 1196-1205,
568 doi:10.1093/molbev/mst030 (2013).
569 49 Korber, B. & Myers, G. Signature pattern analysis: a method for assessing
570 viral sequence relatedness. *AIDS Res Hum Retroviruses* **8**, 1549-1560,
571 doi:10.1089/aid.1992.8.1549 (1992).
572 50 Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein
573 structures and complexes. *Nucleic Acids Res* **46**, W296-W303,
574 doi:10.1093/nar/gky427 (2018).
575 51 Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in
576 2019. *Nucleic Acids Res* **47**, W636-W641, doi:10.1093/nar/gkz268 (2019).
577 52 Webb, B. & Sali, A. Protein structure modeling with MODELLER. *Methods Mol*
578 *Biol* **1137**, 1-15, doi:10.1007/978-1-4939-0366-5_1 (2014).
579 53 Eswar, N. *et al.* Comparative protein structure modeling using Modeller. *Curr*
580 *Protoc Bioinformatics* **Chapter 5**, Unit-5 6,
581 doi:10.1002/0471250953.bi0506s15 (2006).
582 54 Shen, M. Y. & Sali, A. Statistical potential for assessment and prediction of
583 protein structures. *Protein Sci* **15**, 2507-2524, doi:10.1110/ps.062416606
584 (2006).
585 55 Wiederstein, M. & Sippl, M. J. ProSA-web: interactive web service for the
586 recognition of errors in three-dimensional structures of proteins. *Nucleic*
587 *Acids Res* **35**, W407-410, doi:10.1093/nar/gkm290 (2007).
588 56 Sippl, M. J. Recognition of errors in three-dimensional structures of proteins.
589 *Proteins* **17**, 355-362, doi:10.1002/prot.340170404 (1993).
590 57 Benkert, P., Biasini, M. & Schwede, T. Toward the estimation of the absolute
591 quality of individual protein structure models. *Bioinformatics* **27**, 343-350,
592 doi:10.1093/bioinformatics/btq662 (2011).

58 Benkert, P., Kunzli, M. & Schwede, T. QMEAN server for protein model quality estimation. *Nucleic Acids Res* **37**, W510-514, doi:10.1093/nar/gkp322 (2009).

59 Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612, doi:10.1002/jcc.20084 (2004).

60 Páll, S. A., M.J.; Kutzner, C.; Hess, B.; Lindahl, E. . Tackling exascale software challenges in molecular dynamics simulati ons with GROMACS. In Solving Software Challenges for Exascale (eds. Markadis, S. & Laure, E.). *Springer International Publishing: Switzerland, London* **Volume 8759,** , pp. 3–27. (2015).

61 Lindahl, V., Villa, A. & Hess, B. Sequence dependency of canonical base pair opening in the DNA double helix. *PLoS Comput Biol* **13**, e1005463, doi:10.1371/journal.pcbi.1005463 (2017).

62 Jorgensen, W. L. C., J.; Madura,J.D.; Impey,R.W.; Klein,M.L. . Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 79, 926–935. (1983).

63 Bussi, G. D., D.; Parrinello, M. Canonical sampling through velocity rescaling. . *J Chem Phys* **126, 014101, doi:10.1063/1.2408420.** (2007).

64 Darden, T. Y., D.; Pedersen, L. . Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *J Chem Phys* **98, 10089, doi: 10.1063/1.464397** (1993).

65 Heath, L., van der Walt, E., Varsani, A. & Martin, D. P. Recombination patterns in aphthoviruses mirror those found in other picornaviruses. *J Virol* **80**, 11827-11832, doi:10.1128/JVI.01100-06 (2006).