Analysis of the mutation dynamics of SARS-CoV-2 reveals the spread history and emergence of RBD mutant with lower ACE2 binding affinity

Yong Jia[1,*,†], Gangxu Shen[2,3,*], Yujuan Zhang[1], Keng-Shiang Huang[2,4], Hsing-Ying Ho[5], Wei-Shio Hor[6], Chih-Hui Yang[4], Chengdao Li[1,7,†], Wei-Lung Wang[3,†]

[1]College of Science, Health, Engineering and Education, Murdoch University, Murdoch, WA, 6150, Australia

[2]School of Chinese Medicine for Post-Baccalaureate, I-Shou University, Kaohsiung, Taiwan

[3]National Changhua University of Education, Changhua, Taiwan

[4]College of Medicine, I-Shou University, Kaohsiung, Taiwan

[5]Guo-Yuan Clinic, Taichung, Taiwan

[6]TOPO Pharmaceutical Co., Ltd, Taichung, Taiwan

[7]Department of Primary Industry and Regional Development, Government of Western Australia, South Perth, WA, 6155, Australia

*These authors have contributed equally to the study

[†]Correspondence author:

Dr. Yong Jia: y.jia@murdoch.edu.au

Prof. Chengdao Li: c.li@murdoch.edu.au

Prof. Wei-Lung Wang: wlwang@cc.ncue.edu.tw

Summary

Monitoring the mutation dynamics of SARS-CoV-2 is critical for the development of effective approaches to contain the pathogen. By analyzing 106 SARS-CoV-2 and 39 SARS genome sequences, we provided direct genetic evidence that SARS-CoV-2 has a much lower mutation rate than SARS. Minimum Evolution phylogeny analysis revealed the putative original status of SARS-CoV-2 and the early-stage spread history. The discrepant phylogenies for the spike protein and its receptor binding domain proved a previously reported structural rearrangement prior to the emergence of SARS-CoV-2. Despite that we found the spike glycoprotein of SARS-CoV-2 is particularly more conserved, we identified a mutation that leads to weaker receptor binding capability, which concerns a SARS-CoV-2 sample collected on 27th January 2020 from India. This represents the first report of a significant SARS-CoV-2 mutant, and raises the alarm that the ongoing vaccine development may become futile in future epidemic if more mutations were identified.

Highlights

- Based on the currently available genome sequence data, we proved that SARS-COV-2 genome has a much lower mutation rate and genetic diversity than SARS during the 2002-2003 outbreak.
- The spike (S) protein encoding gene of SARS-COV-2 is found relatively more conserved than other protein-encoding genes, which is a good indication for the ongoing antiviral drug and vaccine development.
- Minimum Evolution phylogeny analysis revealed the putative original status of SARS-CoV-2 and the early-stage spread history.
- We confirmed a previously reported rearrangement in the S protein arrangement of SARS-COV-2, and propose that this rearrangement should have occurred between human SARS-CoV and a bat SARS-CoV, at a time point much earlier before SARS-COV-2 transmission to human.
- We provided first evidence that a mutated SARS-COV-2 with reduced human ACE2 receptor binding affinity have emerged in India based on a sample collected on 27th January 2020.

Introduction

The outbreak of severe acute respiratory syndrome–coronavirus 2 (SARS-CoV-2) has caused an unprecedented pandemic

45 and severe fatality around the world. As of 4th April 2020, the total number of SARS-CoV-2 infection has reached over
46 1.05 million cases globally, claiming 56,985 lives (Coronavirus disease 2019, Situation Report-15, WHO). Most
47 concerning is that this number is predicted to continue rising significantly in the next couple of months. Scientists have
48 been working round the clock to understand how the virus spreads and evolves. There is an imminent challenge to develop
49 effective approaches to contain the rapid spread of this pathogen.
50
51 In addition to the traditional control methods, such as travel ban and house isolation, which have clear negative impact on
52 economy and disrupt normal social activities, the development of antiviral drugs and vaccine should be the ultimate solution
53 to contain the epidemic and reduce the fatality (*1, 2*). Similar to other SARS-like CoVs (*3, 4*), SARS-CoV-2 uses its spike
54 (S) protein to bind and invade human cells (*5, 6*). The S protein and its host receptor are the key targets for drug design and
55 vaccine development (*7, 8*). Recently, several 3D protein structures of the receptor binding domain (RBD) of SARS-CoV-
56 2 spike protein have been determined (*5, 6, 9*). The structural basis of receptor recognition by SARS-CoV-2 has become
57 clear (*6, 9*). This laid the foundation for future vaccine development.
58
59 Vaccine utilizes the human immune system and is specific to the viral-encoded peptides (*10*). One of the major concerns
60 for antiviral vaccine development is the constant emergence of new mutations, which may make vaccine not effective for
61 future epidemic (*7, 10*). A prominent example is that, new Influenza viruses arise every year, requiring new immunization
62 (*11*). SARS-CoV-2 belong to the single-stranded RNA virus, whose genome can readily mutate as virus spreads (*12, 13*).
63 Interestingly, initial assessment of the first 9 SARS-CoV-2 genome sequence revealed a low level of mutation rate (*14*). A
64 recent article by Washington Post reported that the mutation rate in SARS-CoV-2 is relatively low despite its rapid spread,
65 which suggests that only a single vaccine may be required for SARS-CoV-2. However, these results may be based on
66 limited genomic data in the early stage of virus development. It is critical to study and monitor the mutation dynamics of
67 SARS-COV-2.
68
69 Taking advantage of the increasing amount of genomic data collected around the world, we set to explore the current status
70 of SARS-CoV-2 genomic diversity, assess the mutation rate, and potentially identify the emergence of novel mutations that

71 may require close attention. A total of 106 complete or near complete SARS-CoV-2 genome data covering over 12 countries
72 was downloaded from public database. The genetic diversity profile and evolutionary rate for each protein-encoding gene
73 were characterized. Phylogenetic analyses in this study revealed clue to the spread history of SARS-CoV-2 in some
74 countries. Most importantly, we identified a SARS-CoV-2 mutation with likely reduced human angiotensin-converting
75 enzyme 2 (ACE2) binding affinity. We confirmed that SARS-CoV-2 has a relatively low mutation rate but also proved that
76 novel mutation with varied virulence and immune characteristics have already emerged.
77

78 ## Methods

79 **Sequence retrieval**

80 The latest sequence data for SARS-CoV-2 and SARS was retrieved from NCBI public database at
81 https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/. The 5UTR, 3UTR, and CDS sequences of the reference SARS-
82 CoV-2 genome (NC_045512.2) and the human SARS genome (NC_004718.3) were used to blastn against the available
83 genome data. The homology search targets were restricted to the complete or near-complete genomes for further analyses.

84 **Conservation profiling**

85 The assessment of sequence conservation was performed using the PLOTCON tool from the The European Molecular
86 Biology Open Software Suite at https://www.bioinformatics.nl/cgi-bin/emboss/plotcon. The gene model of SARS-CoV-2
87 was generated using the AnnotationSketch (*15*) tool based on the genome annotation data downloaded from NCBI database.

88 **Phylogeny construction**

89 Codon-based sequence alignment was performed for the conserved domain sequences (CDS) using MUSCLE program
90 (limited to 2 iterations for fast alignment of long sequences) (*16*). Alignment of the 5UTR and 3UTR sequences were
91 performed separately. The obtained alignment files were concatenated for final phylogeny construction. The phylogenetic
92 tree was developed in MEGA7.0 (*17*) using the Minimum Evolution method with p-distance substitution model, and the
93 maximum Likelihood method (HKY+G+I, 500 times bootstrap test) for the S protein analyses. Tree annotation was carried
94 out using Figtree software (http://tree.bio.ed.ac.uk/software/figtree/).

95 **Evolutionary rate assessment**

96 The ratio of nonsynonymous mutations (d_N) to synonymous mutations (d_S) was calculated using codeml in the PAML
97 (version 4.7) package (*18*). CDS sequences for each protein encoding gene were filtered to remove redundant identical
98 sequences. Then codon-based CDS sequence alignment was performed using MUSCLE program, and an individual NJ
99 tree was generated using MEGA7.0 (*17*) with p-distance model. The obtained sequence alignment and phylogenetic tree
100 files were used as PAML inputs for d_N and d_S calculations.

101 **Protein structural analyses**

102 3D structure of the SARS-CoV-2 spike glycoprotein in complex with (PDB: 6VW1, 6VW1) has been determined recently
103 (*5, 9*). The structural model for the receptor binding domain (RBD) was extracted from 6VW1 for comparison analysis
104 with human SARS structure (PDB: 2AJF) (*3*), which is in complex with the receptor: human ACE2. Amino acid sequence
105 alignment of the spike glycoprotein was visualized and annotated using ESPript 3.0 tool
106 (http://espript.ibcp.fr/ESPript/ESPript/index.php). Protein hydrophobicity profiles were implemented in PyMOL using
107 the Color_h script (http://www.pymolwiki.org/index.php/Color_h), based on the hydrophobicity scale defined at
108 http://us.expasy.org/tools/pscale/Hphob.Eisenberg.html. All structure visualization was carried out using PyMol (Version
109 1.3r1. Schrodinger, LLC).

110

111 ## Results

112 **Genetic diversity analyses identified a single amino acid mutation in RBD of the spike protein in SARS-CoV-2**

113 As of 24th March 2020, there are a total of 174 nucleotide sequences for SARS-CoV-2 in the NCBI database. By restricting
114 to the complete or near-complete genomes, 106 sequences from 12 countries were obtained and used for further analyses.
115 This encompasses 54 records from USA, 35 from China, and the rest from other countries: Australia (1), Brazil (2), Finland
116 (1), India (2), Italy (1), Nepal (1), Spain (3), South Korea (1), and Sweden (1).

117 Based on the gene model of the reference SARS-CoV-2 genome (GeneBank: NC_045512.2), a total of 12 protein-encoding
118 open reading frames (ORFs), plus 5UTR and 3UTR were annotated (**Figure 1A**). Overall, the gene sequences from
119 different samples are highly homologous, sharing > 99.1% identity, with the exception of 5UTR (96.7%) and 3UTR (98%)

(**Table 1**), which are relatively more divergent. Sequence alignment showed that there is no mutation in ORF6, ORF7a, and ORF7b. The genetic diversity profile across the 106 genomes was displayed in **Figure 1A**. A few nucleotide sites within ORF1a, ORF1b, ORF3a, and ORF8 exhibiting high genetic diversity were identified (**Figure 1A**).

The S protein is critical for virus infection and vaccine development. As shown in **Figure 1B**, 12 single amino acid substitutions in 12 genomes were identified for the spike glycoprotein, only one of which occurs in the receptor binding domain (RBD). Notably, this mutation concerns an accession collected from Kerala State, India on 27th Jan 2020.

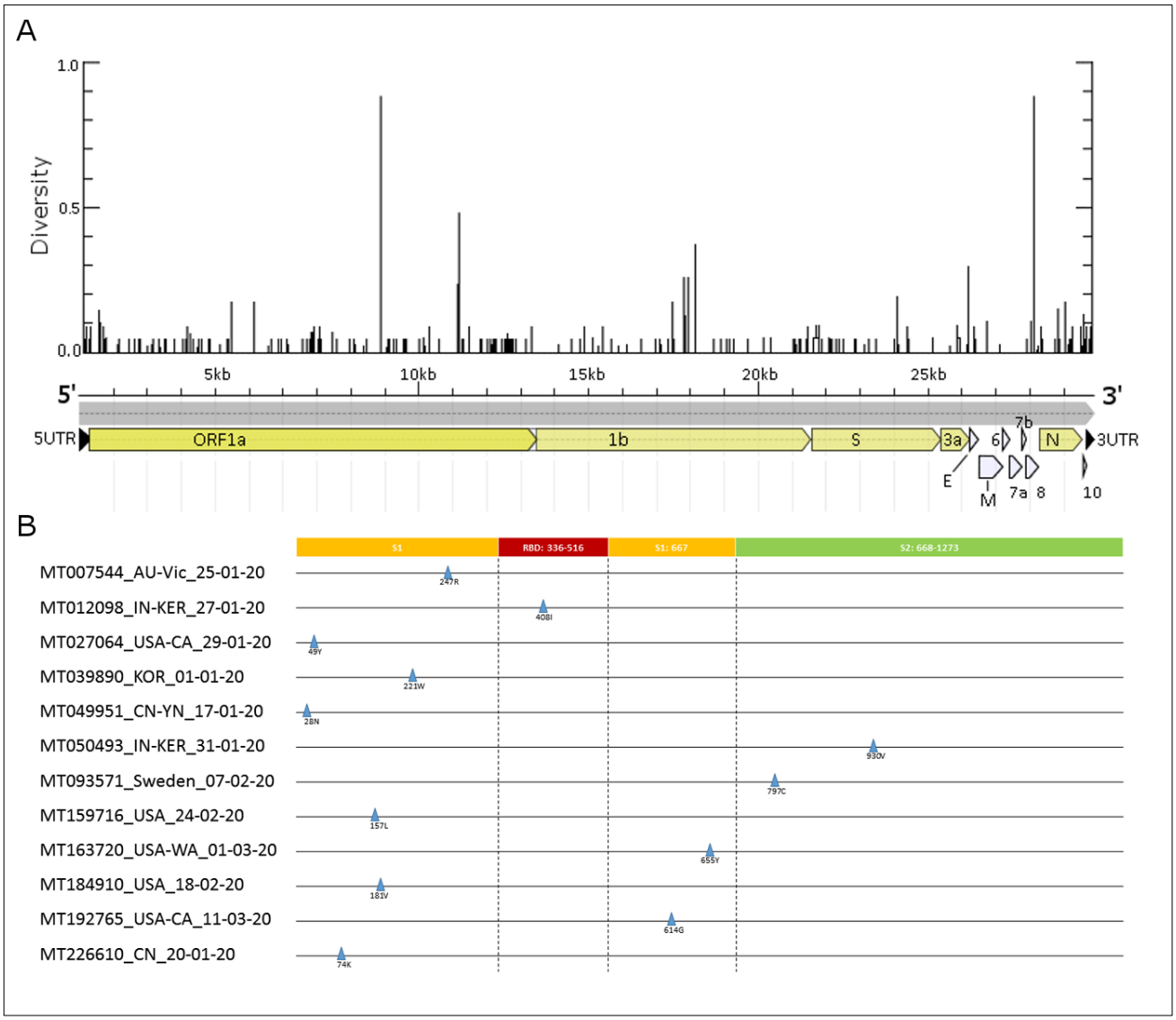


Figure 1. Genetic diversity profile of SARS-CoV-2 genomes and amino acid mutations in the spike glycoprotein. A) Pair-wise genetic distance for each nucleotide site calculated from the 106 SARS-CoV-2 genomes. Gene model is based on the reference genome (GeneBank: NC_045512.2). **B**) Identification of amino acid mutations in the spike glycoprotein. Sequences were named as: Accession name_country_ sample collection time (AU: Australia; IN: India; USA: United States; KOR: South Korea; CN: China; Sweden: Sweden.) Amino acid numbering according to the reference spike protein (Accession ID: YP_009724390.1).

133 **Table 1. Mutation rate analysis on SARS-CoV-2 genes.** Gene model is according to the SARS-CoV-2 reference genome (GeneBank: NC_045512.2).
134 S: spike glycoprotein. "Pair-wise identity" indicate the minimum pair-wise sequence identity among the 106 genomes. d_N : nonsynonymous mutation; d_S :
135 synonymous mutations. "--": not applicable.

Gene name		5UTR	1a	1b	S	ORF3a	ORF4_E	ORF5_M	ORF6	ORF7a	ORF7b	ORF8	ORF9	ORF10	3UTR
Length (bp)		211	13218	8088	3822	828	228	669	186	336	132	366	1260	117	152
Pair-wise identity		96.7%	99.8%	99.9%	99.9%	99.6%	99.1%	99.7%	100%	100%	100%	99.5%	99.7%	99.1%	98%
d_N	SARS-CoV-2	--	0.0081	0.0029	0.0040	0.0074	0.0063	0.0023	0	0	0	0.0111	0.0079	0	--
	SARS	--	0.0119	0.0077	0.0532	0.0331	0.0338	0.023	0.3031	0.0040	0.5339	0.0287	0.0197	0.0135	--
d_S	SARS-CoV-2	--	0.0041	0.0083	0.0055	0	0.0611	0.0046	0	0	0	0	0.0172	0.0326	--
	SARS	--	0.0196	0.0326	0.0442	0.0248	0.0146	0.0928	0.0202	0.0183	0.0005	0.0566	0.9552	0.0341	--

136

137 **SARS-CoV-2 displayed a much lower mutation rate than SARS-CoV, with a highly conserved S gene**

138 To assess how the mutation rate and genetic diversity of SARS-CoV-2, the ratio of nonsynonymous mutations (d_N) and
139 synonymous mutations (d_S), was calculated for each protein-encoding ORF based on the 106 SARS-CoV-2 and 39 SARS
140 genomes. For SARS-CoV-2, the highest d_N was observed for ORF8 (0.0111), followed by ORF1a (0.0081), ORF9 (0.0079),
141 and ORF4 (0.0063) (**Table 1**), indicating these genes may be more likely to accumulate nonsynonymous mutations. In
142 contrast, ORF1b (0.0029), S gene (0.0040) encoding the spike protein, and ORF5 (0.0023) are relatively more conserved
143 in terms of nonsynonymous mutation. Noteworthy, ORF6, ORF7ab and ORF10 are strictly conserved with no
144 nonsynonymous mutation. Compared to SARS-CoV-2, SARS displayed higher mutation rates for all of the ORFs in the
145 virus genome (Table 1), suggesting an overall higher levels of genetic diversity and mutation rate. In particular, the d_N and
146 d_S values for the S gene in SARS-CoV is around 12 and 7 times higher than that for SARS-CoV-2. In contrast, the mutation
147 rate differences for ORF1a and ORF1b between SARS-CoV-2 and SARS are relatively milder, varying from 1.5 times to
148 4.8 times only. In contrast to SARS-CoV-2, which has strictly conserved ORF6, ORF7a, and ORF7b, SARS displayed
149 mutation rates at different levels. Notably, the d_S for ORF10 are comparable between the two genomes at 0.0326 and 0.0341,
150 respectively.

151

152 **Phylogeny analysis revealed the original status of SARS-CoV-2 and its spread history**

153 To trace the potential spread history of SARS-CoV-2 across the world, an unrooted Minimum Evolution (ME) tree of the
154 106 genomes was developed based on whole-genome sequence alignment. The clustering pattern of the ME phylogeny

155 (**Figure 2**) shed light on how the virus may have spread at the early stage. At the center of the ME tree, a number of virus
156 accessions collected from China (including the reference genome NC_045512.2) and USA have the shortest branch
157 (marked by red and black dots), thus may indicate the original status of SARS-CoV-2. The radial pattern, instead of
158 clustering together, of these accessions and other accessions derived from the tree center (highlighted in yellow color) with
159 longer branches, implies the independent mutations occurring during the virus spread (**Figure 2**). However, the longer
160 branch may not be always associated with a longer evolution time, as some accessions collected in December 2019 have
161 equal or even longer branch that those collected in January and February 2020.
162 Due to the data availability, virus accessions collected from China and USA are dominant in the ME tree and constantly
163 group with accessions from other countries. Overall, the target SARS-CoV-2 genomes tend to separate into two major
164 clusters (highlighted in yellow dots, **Figure 2**), suggesting these SARS-CoV-2 may have originated from two major spread
165 sources. Of particular interest is the observation of several phylogenetic clades encompassing samples collected from more
166 than one countries, which may provide clue to track the spread history of SARS-CoV-2 in these countries. For example, a
167 notable clade (clade a) containing accessions collected from USA, Brazil, Italy, Australia, Sweden and South Korea was
168 identified. The only Brail accession (MT126808.1) in this study is found to be clustered with one accession from USA
169 (MT163716.1) with strong support. Whilst the virus accessions from China are prevalent in the ME tree, it is intriguing
170 that no correlated accession from China is found in this clade. An additional clade include accessions collected from China,
171 USA and Finland were found together (clade b). In another notable clade (clade c), 2 of the 3 accessions (LC528232.1 and
172 LC528233.1) collected from the cruise ship in Japan were grouped with several accessions from USA. Two accessions
173 (MT198651.1 and MT198652.1) collected in March 2020 from Spain were grouped (clade f) with one accession collected
174 in January 2020 from China. The additional Spain accession (MT198653.1) was clustered with one from USA
175 (MT192765.1). One India accession (MT012098.1) was found together (clade g) with an accession from Wuhan, China,
176 collected in December 2019. Interestingly, the single Nepal accession (MT072688.1) seems to be closely related (clade d)
177 to several accessions from USA.

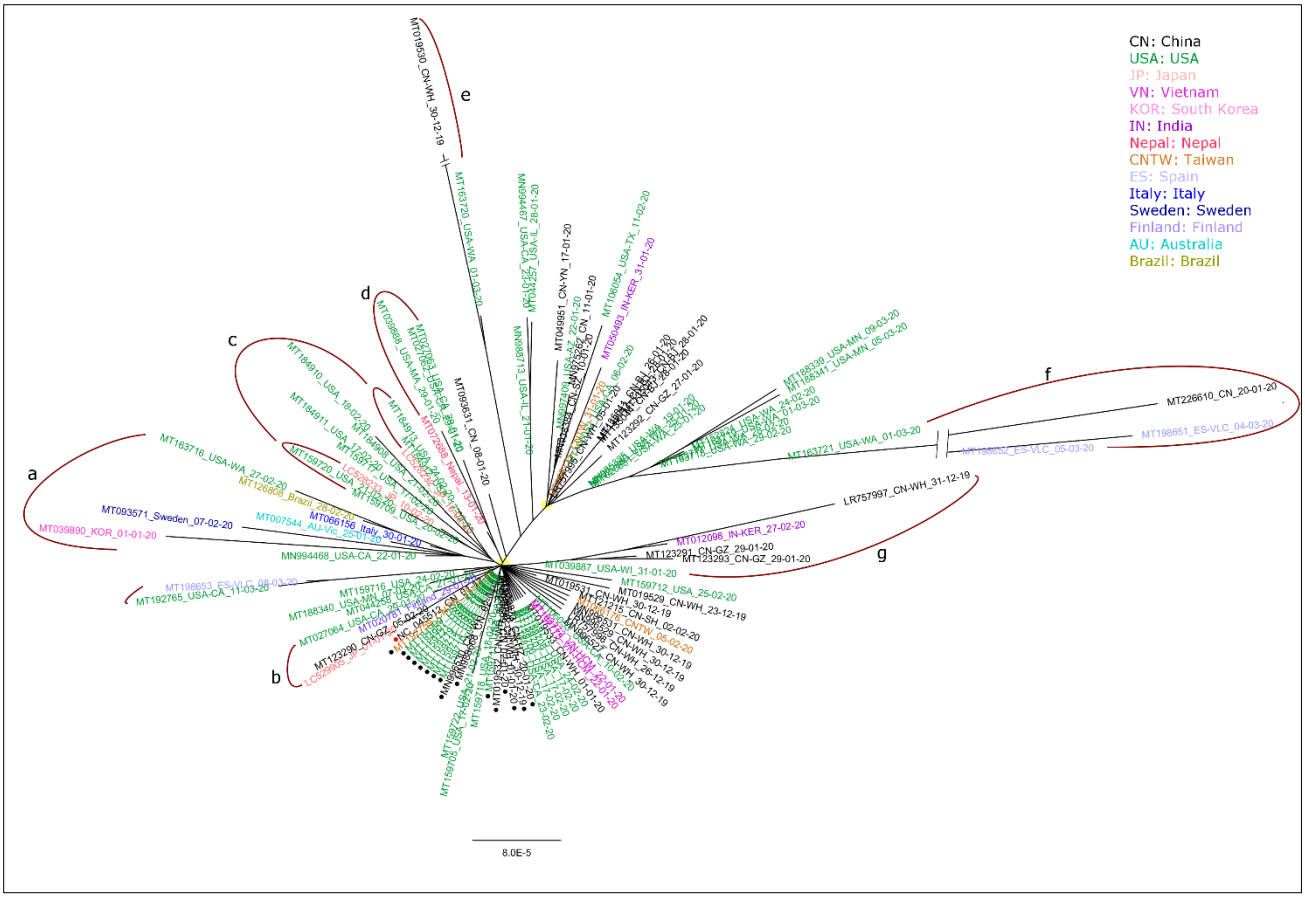
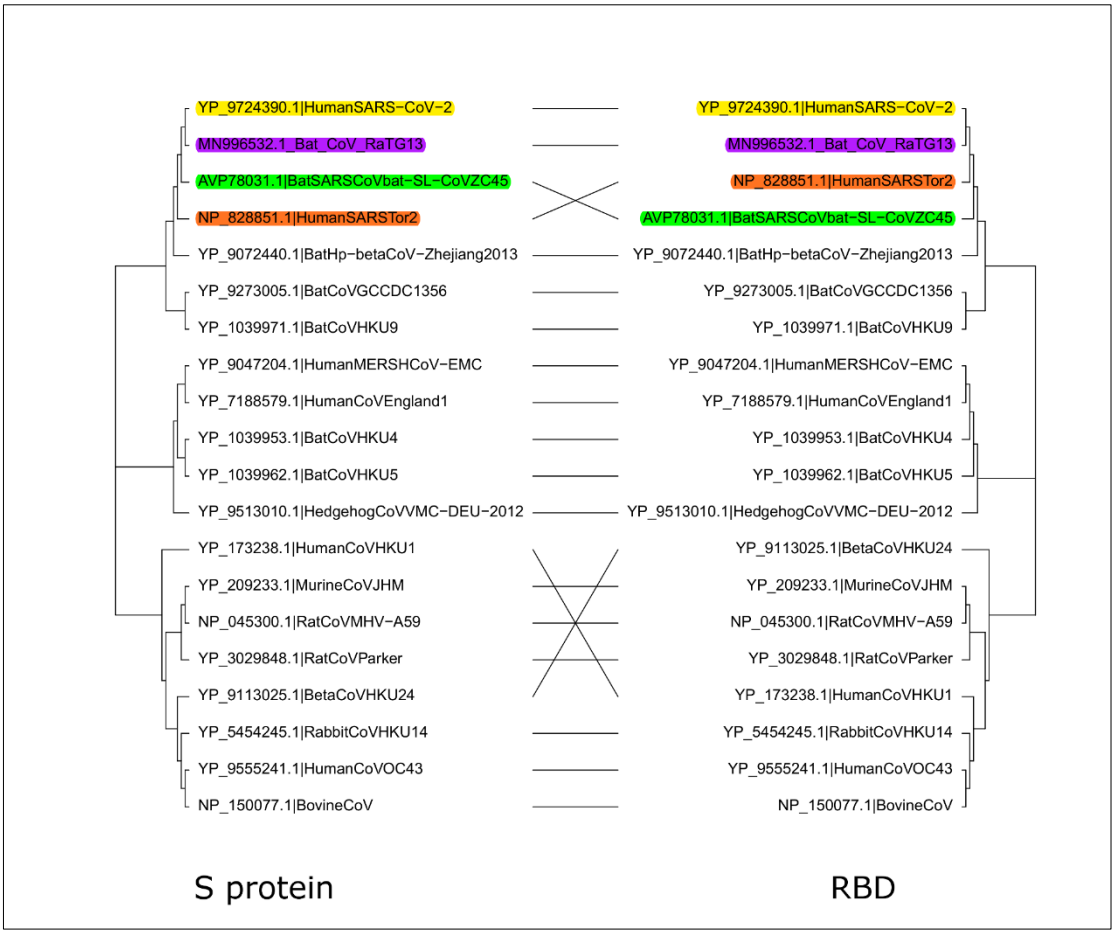


Figure 2. Phylogeny clustering analyses of the 106 SARS-CoV-2 genomes. Results were based on whole genome sequence alignment using Minimum Evolution method. Each accession was named in the "accession ID, country, sample collection time" format. Samples collected from different countries were highlighted in different colors. Red dot indicated the reference SARS-CoV-2 genome (GeneBank: NC_045512.2), which together with black dots indicated the original status of SARS-CoV-2 (branch length = 0). The putative two types of SARS-CoV-2 were highlighted in yellow shades Notable clade containing sequences from more than one countries were highlighted in curved line (magenta).

Spike protein of SARS-CoV-2 has underwent a structural rearrangement

The spike glycoprotein is critical for the virus infection. Recent study suggested that the S protein in SARS-CoV-2 may has underwent a structural rearrangement(*13*). To investigate this hypothesis, two separate phylogenies were developed based on the full-S and RBD sequences, respectively. Overall, the two phylogenies displayed similar clustering patterns, separating into three major clades (**Figure 3)**. SARS-CoV-2 was identified in the same major clade, and was clustered most closely with two bat SARS CoVs (highlighted in purple and green colors, **Figure 3**) and the human SARS-CoV (orange color, **Figure 3**). In both phylogenies, SARS-CoV-2 is most closely related to bat_CoV_RaTG13, suggesting SARS-CoV-2 may have originated from bat. However, the evolutionary positions of human SARS-CoV and bat-SL-CoVZ45 were swapped between the full-S and RBD-only phylogenies. In the full-S phylogeny, bat-SL-CoVZ45 is relatively more similar to human SARS-CoV-2, whilst human SARS-CoV is closer to SARS-CoV-2 than bat-SL-CoVZ45. Taken together, these

194 results suggested that the RBD of SARS-CoV-2 is more likely originated from human SARS-CoV, whilst the rest part of
195 the S protein in SARS-CoV-2 may have originated from bat-SL-CoVZ45, supporting the potential structural rearrangement
196 of S protein in SARS-CoV-2. bat_CoV_RaTG13 is similar to SARS-CoV-2, indicating the proposed structural
197 rearrangement may have occurred in bat first before its transmission to human.



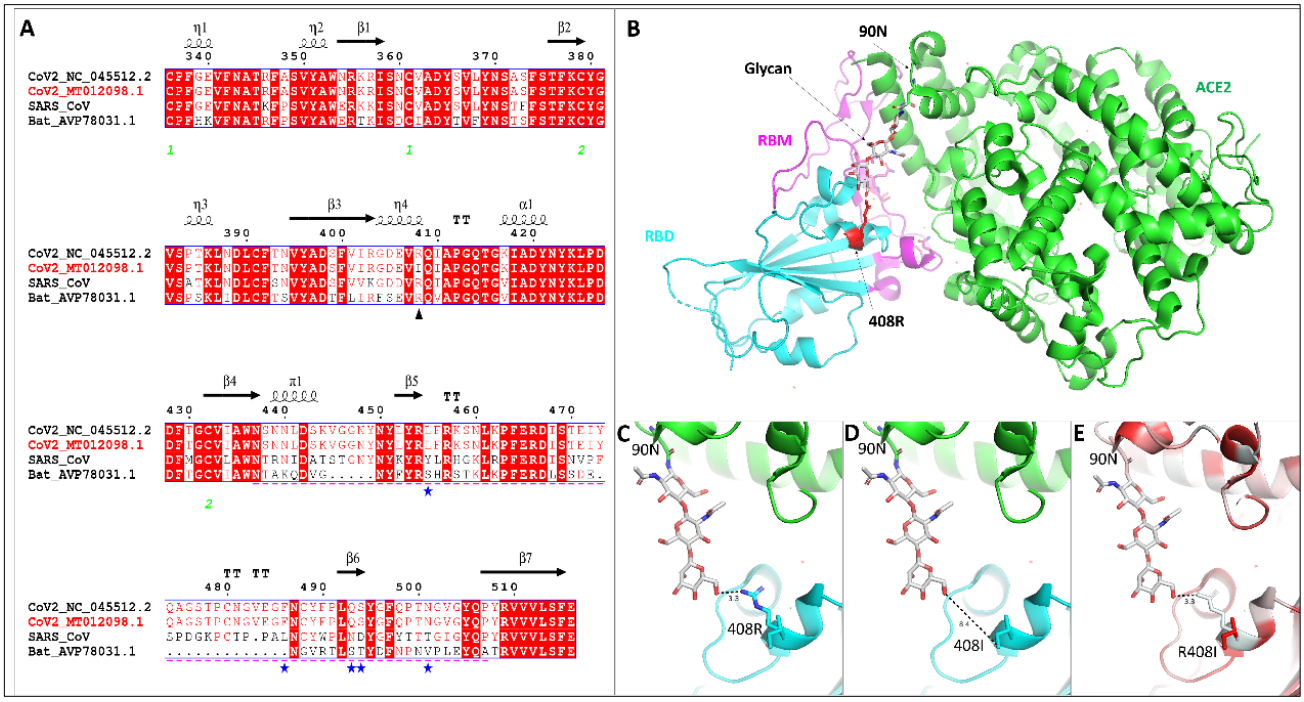
198

199 **Figure 3. Displays the phylogeny discrepancy of the full-S and RBD sequences.** Maximum Likelihood phylogenies based on the full-S protein (left)
200 and RBD (right) sequences of SARS-like CoVs. Taxa names were in the "Accession Id, host organism, sample name" format. Human SARS-CoV-2
201 and its close relatives were highlighted in different colors.
202

203 **A single amino acid mutation in RBD results in reduced receptor binding affinity on human ACE2**

204 The RBD of virus S protein binds to a receptor in host cells, and is responsible for the first step of CoV infection (*3*). Thus,
205 amino acid mutation to RBD may have significant impact on receptor binding and vaccine development. The 3D structure
206 of the spike protein RBD of SARS-CoV-2 (PDB: 6VW1) has recently been determined in complex with human ACE2
207 receptor (*6*). One of the12 amino acid mutations in the RBD of S protein (R408I) was identified among the 106 SARS-
208 CoV-2 genomes. Sequence alignment showed that 408R is strictly conserved in SARS-CoV-2, SARS-CoV and bat SARS-

209 like CoV (**Figure 4A**). Based on the determined CoV2_RBD-ACE2 complex structure, 408R is located at the interface
210 between RBD and ACE2, but is positioned relatively far away from ACE2, thus does not have direct interaction with ACE2
211 (**Figure 4B**). However, the determined RBD0-ACE2 structure showed that 408R forms a hydrogen bond (3.3 Å in length)
212 with the glycan attached to 90N from ACE2 (**Figure 4C**) (*6*). The hydrogen bond may have contributed to the exceptionally
213 higher ACE2 binding affinity. In contrast, despite this arginine residue is also conserved in human SARS-CoV
214 (corresponding to 395R in PDB: 2AJF), it is positioned relatively distant (6.1 Å) from the glycan bound to 90N from ACE2
215 (**Figure S1**). Interestingly, the 408R-glycan hydrogen bond seem to be disrupted by the R408I mutation in one SARS-CoV-
216 2 accession (GeneBank ID: MT012098.1) (**Figure 4D**), which was collected from India on 27th Jan 2020. Furthermore, in
217 contrast to the arginine residue, which is electrically charged and highly hydrophilic, the mutated isoleucine residue has a
218 highly hydrophobic side chain with no hydrogen-bond potential (**Figure 4E**). To sum up, the R408I mutation identified
219 from the SARS-CoV-2 strain in India represents a SARS-CoV-2 mutant with potentially reduced ACE2 binding affinity.
220



221

222 **Figure 4. Sequence alignment and protein structural analyses of the mutation in RBD of SARS-CoV-2. A**) Sequence alignment of RBDs. ▲:
223 R408I mutation; --- : receptor binding motif (RBM). ⋆ : RBD-interacting sites. **B**) Overall position of the identified mutation relative to: RBD (cyan),
224 ACE2 (green) with RBM (pink) and Glycan (grey). **C,D**) Display the disrupted hydrogen bond by the R408I mutation. "---": distance in Å. **E**)
225 Hydrophobic profile changes due to R408I mutation, with with red and white colours representing the highest hydrophobicity and the lowest
226 hydrophobicity respectively. All amino acid number according to the S protein of SARS-CoV-2 (NC_045512.2) and human ACE2, respectively.
227

228 **Discussions**

229 Based on the currently available genome sequence data, our results showed that the mutation rate of SARS-CoV-2 is much
230 lower than that for SARS, which caused the 2002-2003 outbreak. Our study is the first to provide a direct quantitative
231 comparison between SARS-COV-2 and SARS. A relatively stable genome of SARS-CoV-2 is a good indication for the
232 epidemic control, as less mutation raises the hope of the rapid development of validate vaccine and antiviral drugs. Our
233 results are consistent with several recent genetic variance analyses on SARS-CoV-2 (*19, 20*), which suggested the SARS-
234 CoV-2 genomes are highly homogeneous. Molecular geneticists closely monitoring the virus development also suggested
235 that the mutation rate of SARS-CoV-2 maintains at a low level. Whilst it is generally safe to say that SARS-CoV-2 tends
236 to mutate at a low rate, all current analyses are merely based on data collected at the early stage of this pandemic. As the
237 virus continues to spread rapidly around the world, and more genomic data is accumulated, the evolution and mutation
238 dynamics of SARS-CoV-2 still need to be monitored closely.

239

240 One critical aim of our study is to identify the original status of SARS-CoV-2 before its wide transmission across different
241 countries. Due to the short time space of sample collection and a relatively low mutation rate for SARS-CoV-2, we believe
242 that a Minimum Evolution phylogeny may outperform other phylogenetic methods to achieve this aim. As expected, the
243 earliest few reported SARS-CoV-2 accessions collected from Wuhan China were identified at the center of the phylogenetic
244 tree with the shortest branch. Interestingly, a number of virus genomes from USA were found almost identical to these
245 putative original versions of virus from Wuhan. However, according to public media, the outbreak of SARS-CoV-2 in USA
246 occurred relatively later than other countries. One possible explanation for this observation is that, the spread of SARS-
247 CoV-2 in USA might start much earlier than previously thought or reported. Due to a dominant proportion of the samples
248 in this study were collected from China and USA, we observed a significantly higher level of genetic diversity from these
249 two countries. Most SARS-CoV-2 accessions from the other countries can find their closely related sisters from either
250 China or USA. This data bias, on the other hand, may give us an advantage to trace the spread history of SARS-CoV-2 in
251 different countries. This suggestion is reliable because all of the samples studies in this study were collected at the early
252 stage of the pandemic, which may avoid the potential data noise caused by recent published genomes of complex spread
253 background. One notable finding in our phylogenetic tree is that, the singleton SARS-CoV-2 accessions collected from

254 Australia, Brazil, South Korea, Italy and Sweden were clustered together with two USA samples but without a Chinese
255 version, suggesting that these infection cases may be somehow related. In addition, one of the three samples collected from
256 the cruise ship stranded in Japan was found closely related to a sample collected from Guangzhou, China, whilst the other
257 two were grouped with several cases from USA. Noteworthy, out phylogeny seems to support the presence of two major
258 types of SARS-CoV-2 in the target samples, suggesting the potential existence of two spread sources. Interestingly, this
259 speculation is corroborated by an independent clustering analyses using different phylogeny method (*20*).
260
261 Until now, the origin of SARS-CoV-2, and how it has been transmitted to human remains largely a mystery. Early genomic
262 data proved that human SARS-CoV-2 is an enveloped, positive-sense, and single-stranded RNA virus in the subgenus
263 *Sarbecovirus* of the genus *Betacoronavirus* (*13, 14*). Evolutionarily, SARS-CoV-2 is most closely related to bat SARS-like
264 CoV (88% genome sequence identity) and human SARS CoV (79%), the latter of which has caused world pandemic in
265 2003 (*13*). Based on the strong genome sequence identity between SARS-CoV-2 and bat SARS-like COVs, it was initially
266 speculated that SARS-CoV-2 may have originated from bat (*14, 21*). However, a more recent study proposed that pangolin
267 may be the most likely reservoir hosts due to the identification of closely related SARS-COVs from this species as well
268 (*22*). Both of these two animals can harbor coronaviruses related to SARS-CoV-2. However, direct evidence of the
269 transmission of SARS-CoV-2 from either bat or pangolin to human is still missing.
270
271 Prior to this study, several publications have suggested that SARS-CoV-2 may have originated from the genome
272 recombination of SARS-like CoVs from different animal hosts, as evidenced by the discrepant clustering patterns for the
273 phylogenies using different genetic regions. Lu (*13*) first observed that the RBD of S protein in SARS-CoV-2 is more
274 closely related to human SARS-CoV, whilst the other part of its genome is more similar to bat SARS-CoV. Later Peng (*23*)
275 identified a bat CoV_RaTG13 and several pangolin SARS-CoVs that are consistently closer to SARS-CoV-2 than human
276 SARS-CoV in either full-S protein or RBD. By combining the data from these two studies, our study confirmed the
277 observations reported in both studies, and further determined that the S protein recombination actually happened between
278 human SARS-CoV and a bat SARS-CoV, much earlier before its transmission to human, with the newly identified bat
279 SARS-CoV-RaTG13 as an intermediate.

Another notable finding in this study corresponds to the identification of an amino acid mutation in the RBD of S protein in SARS-CoV-2. Mostly importantly, we showed that this amino acid mutation is very likely to cause a reduced binding affinity to human ACE2 receptor. The RBD of S protein binds to a receptor in host cells, and is responsible for the first step of CoV infection. The receptor binding affinity of RBD directly affects virus transmission rate. Thus, it has been the major target for antiviral vaccine and therapeutic development such as SARS (*8*). Despite the S protein gene seems to be more conserved than the other protein-encoding genes in the SARS-CoV-2 genome, our study provide direct evidences that a mutated version of SARS-CoV-2 S protein with varied transmission rate may have already emerged. Based on the close relationship of SARS-CoV-2 to SARS, current vaccine and drug development for SARS-CoV-2 has also focused on the S protein and its human binding receptor ACE2 (*7, 24*). Thus, the observation in this study raised the alarm that SARS-CoV-2 mutation with varied epitope profile could arise at any time, which means current vaccine development against SARS-CoV-2 is at great risk of becoming futile. Because the receptor recognition mechanism seems to be highly conserved between SARS-CoV-2 and SARS-CoV, which have been proved to share the common human cell receptor ACE2. One suggestion for the next step of therapeutic development is probably to focus on the identification of potential human ACE2 receptor blocker, as suggested in a recent commentary (*7*). This approach will avoid the above-mentioned challenge faced by vaccine development.

Acknowledgement

The authors would like to thanks the relevant research community for making the genomic data available to the public. We thanks Dr Yinchuan Zhang from Maternal and Child Health Hospital, Dingxi, Gansu, China, and Mrs Hong Ma from Centre of Disease Control (CDC), Dingxi, Gansu, China for providing critical comments on the manuscript.

Author Contribution

WLW, CL and YJ conceived the study. YJ, GS, YZ, KSH, HYH, WSH, CHY performed data analyses. YJ&GS wrote the

manuscript. All authors have read the manuscript.

Conflict of interest

The authors declare no conflict of interest.

Supplementary figure

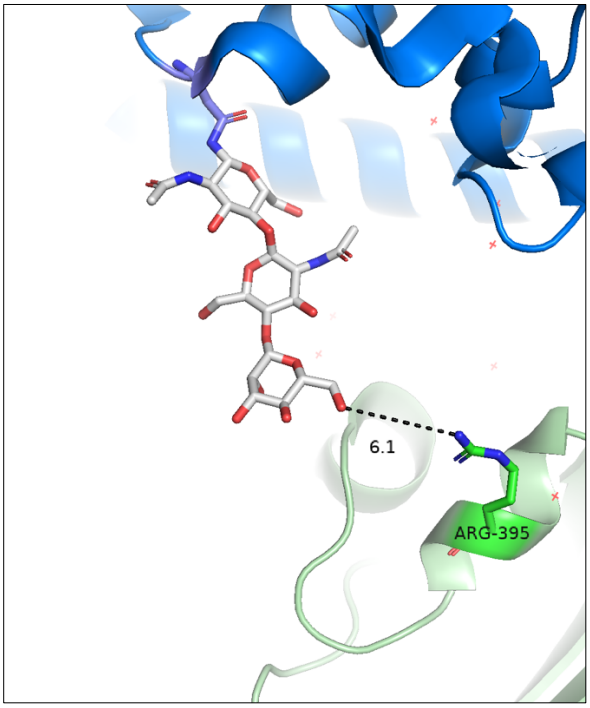


Figure S1. Displays the position of 395R in human SARS-CoV (PDB: 2AJF). Dash line indicates the measured distance in Å.

References:

1. L. Zhang, Y. H. Liu, Potential interventions for novel coronavirus in China: A systematic review. *J Med Virol* **92**, 479-490 (2020).
2. S. Lu, Timely development of vaccines against SARS-CoV-2. *Emerg Microbes Infec* **9**, 542-544 (2020).
3. F. Li, W. H. Li, M. Farzan, S. C. Harrison, Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. *Science* **309**, 1864-1868 (2005).
4. F. Li, Evidence for a Common Evolutionary Origin of Coronavirus Spike Protein Receptor-Binding Subunits. *J Virol* **86**, 2856-2858 (2012).
5. D. Wrapp *et al.*, Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260-+ (2020).
6. J. Shang *et al.*, Structural basis of receptor recognition by SARS-CoV-2. *Nature*, (2020).
7. D. Gurwitz, Angiotensin receptor blockers as tentative SARS-CoV-2 therapeutics. *Drug*

327 *development research*, (2020).
328 8. L. Y. Du *et al.*, The spike protein of SARS-CoV - a target for vaccine and therapeutic
329 development. *Nat Rev Microbiol* **7**, 226-236 (2009).
330 9. R. Yan *et al.*, Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2.
331 *Science* **367**, 1444-1448 (2020).
332 10. B. Correia *et al.*, Proof of principle for epitope-focused vaccine design. *Protein Sci* **24**, 181-184
333 (2015).
334 11. A. Huckriede, L. Bungener, T. Daemen, J. Wilschut, Influenza Virosomes in Vaccine
335 Development. *Liposomes, Pt C* **373**, 74-91 (2003).
336 12. M. R. Denison, R. L. Graham, E. F. Donaldson, L. D. Eckerle, R. S. Baric, Coronaviruses An RNA
337 proofreading machine regulates replication fidelity and diversity. *Rna Biol* **8**, 270-279 (2011).
338 13. R. J. Lu *et al.*, Genomic characterisation and epidemiology of 2019 novel coronavirus:
339 implications for virus origins and receptor binding. *Lancet* **395**, 565-574 (2020).
340 14. X. T. Xu *et al.*, Evolution of the novel coronavirus from the ongoing Wuhan outbreak and
341 modeling of its spike protein for risk of human transmission. *Sci China Life Sci* **63**, 457-460
342 (2020).
343 15. S. Steinbiss, G. Gremme, C. Schrfer, M. Mader, S. Kurtz, AnnotationSketch: a genome
344 annotation drawing library. *Bioinformatics* **25**, 533-534 (2009).
345 16. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput.
346 *Nucleic Acids Res* **32**, 1792-1797 (2004).
347 17. S. Kumar, G. Stecher, K. Tamura, MEGA7: molecular evolutionary genetics analysis version 7.0
348 for bigger datasets. *Mol Biol Evol* **33**, 1870-1874 (2016).
349 18. Z. H. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591
350 (2007).
351 19. C. Ceraolo, F. M. Giorgi, Genomic variance of the 2019-nCoV coronavirus. *J Med Virol*, (2020).
352 20. X. Tang *et al.*, On the origin and continuing evolution of SARS-CoV-2. *National Science Review*,
353 (2020).
354 21. P. Zhou *et al.*, A pneumonia outbreak associated with a new coronavirus of probable bat origin.
355 *Nature* **579**, 270-+ (2020).
356 22. M. H.-H. S. Tommy Tsan-Yuk Lam, Hua-Chen Zhu, Yi-Gang Tong, Xue-Bing Ni,, W. W. Yun-Shi
357 Liao, William Yiu-Man Cheung, Wen-Juan Li, Lian-Feng Li, Gabriel M. Leung,, Y.-L. H. Y. G.
358 Edward C. Holmes, Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. *Nature*,
359 (2020).
360 23. T. T.-Y. Lam *et al.*, Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. *Nature*,
361 1-6 (2020).
362 24. S. F. Ahmed, A. A. Quadeer, M. R. McKay, Preliminary identification of potential vaccine targets
363 for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies.

364 *Viruses* **12**, 254 (2020).
365