1 **Full Title:**

2 **Molecular Evolution of SARS-CoV-2 Structural Genes:**

3 **Evidence of Positive Selection in Spike Glycoprotein**

4

5 **Authors:** Xiao-Yong Zhan¹, Ying Zhang¹, Xuefu Zhou¹, Ke Huang¹, Yichao Qian¹, Yang Leng¹,

6 Leping Yan¹, Bihui Huang^{1*}, Yulong He^{1*}

7

8 ¹ The Seventh Affiliated Hospital, Sun Yat-sen University, Shenzhen 517108, China

9

10 **Correspondence authors:**

11 * Bihui Huang

12 ORCID: 0000-0003-3084-7096

13 Address: No.628, Zhenyuan Road, Guangming District, Shenzhen 518107, China

14 Tel: 86-755-81207035, Email: huangbh7@mail.sysu.edu.cn

15 *Yulong He

16 ORCID: 0000-0001-8930-8704

17 Address: No.628, Zhenyuan Road, Guangming District, Shenzhen 518107, China

18 Tel: 86-755-81201030, Email: heyulong@mail.sysu.edu.cn

19 Abstract

20 SARS-CoV-2 caused a global pandemic in early 2020 and has resulted in more than 8,000,000
 21 infections as well as 430,000 deaths in the world so far. Four structural proteins, envelope (E),
 22 membrane (M), nucleocapsid (N) and spike (S) glycoprotein, play a key role in controlling the
 23 entry into human cells and virion assembly of SARS-CoV-2. However, how these genes evolve
 24 during its human to human transmission is largely unknown. In this study, we screened and
 25 analyzed roughly 3090 SARS-CoV-2 isolates from GenBank database. The distribution of the four
 26 gene alleles is determined: 16 for E, 40 for M, 131 for N and 173 for S genes. Phylogenetic
 27 analysis shows that global SARS-CoV-2 isolates can be clustered into three to four major clades
 28 based on the protein sequences of these genes. Intragenic recombination event isn't detected
 29 among different alleles. However, purifying selection has conducted on the evolution of these
 30 genes. By analyzing full genomic sequences of these alleles using codon-substitution models (M8,
 31 M3 and M2a) and likelihood ratio tests (LRTs) of codeML package, it reveals that codon 614 of S
 32 glycoprotein has subjected to strong positive selection pressure and a persistent D614G mutation
 33 is identified. The definitive positive selection of D614G mutation is further confirmed by internal
 34 fixed effects likelihood (IFEL) and Evolutionary Fingerprinting methods implemented in Hyphy
 35 package. In addition, another potential positive selection site at codon 5 in the signal sequence of
 36 the S protein is also identified. The allele containing D614G mutation has undergone significant
 37 expansion during SARS-CoV-2 global pandemic, implying a better adaptability of isolates with
 38 the mutation. However, L5F allele expansion is relatively restricted. The D614G mutation is
 39 located at the subdomain 2 (SD2) of C-terminal portion (CTP) of the S1 subunit. Protein structural
 40 modeling shows that the D614G mutation may cause the disruption of salt bridge among S protein

41 monomers increase their flexibility, and in turn promote receptor binding domain (RBD) opening,
42 virus attachment and entry into host cells. Located at the signal sequence of S protein as it is, L5F
43 mutation may facilitate the protein folding, assembly, and secretion of the virus. This is the first
44 evidence of positive Darwinian selection in the *spike* gene of SARS-CoV-2, which contributes to a
45 better understanding of the adaptive mechanism of this virus and help to provide insights for
46 developing novel therapeutic approaches as well as effective vaccines by targeting on mutation
47 sites.
48

49 Introduction

50 Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of an
 51 emerging coronavirus disease (COVID-19) that has caused more than 430,000 deaths, is still a
 52 serious global pandemic currently. The genome of SARS-CoV-2 is consisting of a single-stranded
 53 and positive-sense RNA of around 30 kb in length with a 5' cap and 3'-polyA tail. It shows that
 54 SARS-CoV-2 genome possesses six major open reading frames (ORFs) that encodes 27 different
 55 proteins, in which four are structural proteins named Envelope (E), Membrane (M), Nucleocapsid
 56 (N) and Spike (S). Many studies have demonstrated important functions of these proteins in virus
 57 entry, transcription and virion particle assembly of SARS-CoV-2. The E protein is a small
 58 envelope protein with 75 amino acids. Given that a close genetic relationship between
 59 SARS-CoV-2 and SARS-CoV, functions of this protein may include virion assembly and
 60 morphogenesis[1]. In addition, induction of apoptosis of host cells might be another crucial
 61 function of SARS-CoV-2 E protein, thus making it a potential determinant of viral pathogenesis
 62 [2]. M protein, consisting of 222 amino acids, is the most abundant component of the viral
 63 envelope and plays a key role in the virion assembly[3]. N protein, composed of 419 amino acids,
 64 may form complexes with genomic RNA, interact with the viral membrane protein, and play a
 65 critical role in enhancing the efficiency of virus transcription and assembly[4]. S protein,
 66 consisting of 1,273 amino acids, is the most important factor that mediates virus entry and a
 67 primary determinant of cell tropism and pathogenesis of SARS-CoV-2[5].
 68 Many studies demonstrated SARS-CoV-2 underwent the evolution and some genetic evolutionary
 69 features have been reported[6]. The whole genomic sequence of SARS-CoV-2 has 79.6% identity
 70 with SARS-CoV and 96% with a bat SARS-related coronavirus (SARSr-CoV), RaTG13.

71 Although no positive time evolution signal was found between SARS-CoV-2 and RaTG13, the
72 SARS-CoV-2 shows a strong positive temporal evolution relationship with bat-SL-CoVZC45,
73 which has a slightly less identical genomic sequence (87.5%) than RaTG13 [7]. Combining the
74 phylogenetic analysis of full-length genomes of coronaviruses, a potential bat origin of
75 SARS-CoV2 is indicated [8]. A recent study reported that *spike* (S) gene (coding gene of S protein)
76 of SARSr-CoVs from their natural reservoir host, the Chinese horseshoe bat (*Rhinolophus sinicus*),
77 has coevolved with *R. sinicus* angiotensin converting enzyme 2 (ACE2) via positive selection[9].
78 A single-stranded positive-sense RNA virus as it is, SARS-CoV-2 causes global pandemic within
79 half a year, suggesting it may evolve rapidly. However, the evolution of SARS-CoV-2 based on
80 structural genes from human to human transmission has not been investigated in detail. The
81 primary purpose of this work is to study the evolutionary pattern of the four structural genes of
82 SARS-CoV-2 derived from a global isolate collection including the E, M, N and S. Various
83 molecular evolution and selection analysis approaches were employed to identify the phylogeny of
84 the four structural proteins and potential selection effects on these genes. Hereby, our study
85 reveals that intragenic recombination does not contribute to the evolution of these genes while
86 purifying selection is the main evolutionary force. Moreover, a D614G mutation in the S protein is
87 operated by strong positive selection and may be responsible for the quick spread of SARS-CoV-2
88 globally. Additionally, another potential L5F mutation may also be operated by positive selection,
89 but with relatively less strong pressure as compared to D614G.

90

91 **Materials and Methods**

92 **SARS-CoV-2 isolates**

93 Complete full-length genomic sequences of SARS-CoV-2 were downloaded from 2019 Novel
94 Coronavirus Resource (2019nCoV) in China National Center for Bioinformation. All of which
95 were also uploaded to the NCBI GenBank database. The sequences were manually checked and
96 finally a total of 3090 isolates were selected and verified for the present study. These isolates were
97 collected from December 24, 2019 to April 24, 2020 in the different geographical locations
98 including China, USA, Japan, Pakistan, Australia, Greece, German, Peru, Turkey, Kazakhstan,
99 Iran, Serbia, Thailand, Nederland, Sri Lanka, Czech, Malaysia, India etc. Detailed information of
100 these isolates including the GenBank accession number or biosample number is summarized in **S1**
101 **Table**.

102

103 **Sequence analysis of the four structural genes and proteins**

104 The E, M, N, S gene sequences were extracted from SARS-CoV-2 global isolate collection and
105 aligned by the MEGA X package using Muscle (codons) parameters [10]. Because some regions
106 of genomic sequences of SARS-CoV-2 couldn't be exactly identified, in which nucleic acid bases
107 are shown as degenerate bases (e.g. N, R, Y), we were unable to obtain all of the four structural
108 gene sequences from an isolate sometimes. Allele type and DNA sequence polymorphism analyses
109 were performed using DnaSP 6.12.03[11]. The protein sequences and polymorphism loci of these
110 isolates were also aligned and analyzed with the MEGA X.

111

112 **Molecular evolution analysis**

113 An unrooted phylogenetic tree of the four structural proteins was constructed using the MEGA X
114 package [10], and the evolutionary history was inferred using the Maximum Likelihood method,

115 based on the JTT matrix-based model for E protein sequences, General Reversible Chloroplast +
116 Freq. model for M, JTT matrix-based model for N and Jones et al. w/freq. model for S protein
117 sequences. Model selection was conducted in MEGA X. Bootstrap values were estimated by 1000
118 replications. Initial tree(s) for the heuristic search were obtained automatically by applying
119 Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using each model
120 mentioned above. The tree is drawn to scale, and FigTree V1.4 was utilized to form cladogram
121 branches (<http://tree.bio.ed.ac.uk/software/figtree/>). The aligned DNA sequences were also
122 screened using RDP4 software to detect intragenic recombination among the alleles of each
123 structural gene[12]. Six methods implemented in the RDP4 were utilized. These methods are RDP
124 [12], GENECONV[13], BootScan [14], MaxChi[15], Chimaera [16], and SiScan [17]. Common
125 settings for all methods include considering sequences as linear and setting statistical significance
126 at the $P < 0.05$ with Bonferroni correction for multiple comparisons and requiring phylogenetic
127 evidence and polishing of breakpoints. Potential recombination events (PREs) were considered as
128 those identified by at least two methods. Reticulate network tree of alleles of the four structural
129 genes of SARS-CoV-2 was also generated by Splitstree4 [18]. Phi test implemented in Splitstree4
130 was used to define probable recombination events. Tajima's D, Fu and Li's D* and F* tests were
131 employed to test the mutation neutrality hypothesis of the whole gene as previously described by
132 our research group[19]. These analyses were carried out using DnaSP 6.12.03[11]. A statistical
133 significance level with $P < 0.05$ is acceptable. The false discovery rate and 1000 replications in a
134 coalescent simulation were applied for correcting multiple comparisons. Non-neutrality evolution
135 was considered when identified by at least two out of three tests. Nonsynonymous and

136 synonymous mutations of the alleles of the four structural genes were also calculated using MEGA
137 X package [10].

138

139 **Analysis of positive selection based on codon**

140 The selection pressure operating the four structural genes of SARS-CoV-2 was searched by using
141 the Maximum Likelihood (ML) method. Analyses were performed using a visual tool of codeml
142 program, named EasyCodeML algorithm with site model [20]. Three nested models (M3 vs. M0,
143 M2a vs. M1a, and M8 vs. M7) were compared and likelihood ratio tests (LRTs) were applied to
144 access a better fit of codes. Model fitting was also performed using multiple seed values for dN/dS
145 and assuming the F3x4 model of codon frequencies. Positive selection is inferred when individual
146 site or codon with ratio of nonsynonymous to synonymous mutations (dN/dS ratios) is greater than
147 one ($\omega > 1$). When the LRT is significant ($p < 0.05$), Bayes empirical Bayes (BEB) (M8 model) and
148 Naive Empirical Bayes (NEB) methods (M3 and M2a model) are further employed to identify
149 amino acid residues that likely evolve under positive selection based on a posterior probability
150 threshold of 0.95. Results from M8 model were taken as the standard as Yang *et al.* reported. M3
151 model was used for the frequency distribution of codon class analysis as Yang *et al.*
152 recommended[21]. HyPhy package was used to validate the result obtained by ML method[22].

153

154 **Structural modeling of the protein with positive selection sites**

155 Three-dimensional structures of proteins with positive selection sites were modeled using
156 SWISS-MODEL (<http://swissmodel.expasy.org>) according to the most fitted protein template.
157 Model quality was evaluated by QMEAN while the structure of the model was visualized by using

158 PyMoL [23].

159

160 **Results and Discussion**

161 **Characteristics of SARS-CoV-2 isolates, structural gene and protein** 162 **sequences**

163 The 3090 SARS-CoV-2 isolates harbor only 16 unique alleles of E and 40 alleles of M, but an
164 abundant number of alleles of N and S genes, which contain 131 and 173, respectively. These
165 alleles correspond to 10, 14, 88 and 99 different amino acid sequences of E, M, N, and S proteins,
166 respectively. Protein sequence comparisons of WH01 isolate with SARSr-CoV, bat-SL-CoVZC45
167 isolate show 100% (75/75) identity in E, 98.65% (219/222) identity in M, 94.27% (395/419)
168 identity in N and 80.06% (1171/1273) in S proteins, respectively. These results imply a close
169 kinship between SARS-CoV-2 and bat SARSr-CoV, especially on E and M proteins. On the other
170 hand, it indicates an extreme conservation of E and M proteins and their functions among
171 coronaviruses[24].

172 Further analysis revealed that there are 14 single nucleotide polymorphisms (SNPs) of E gene, but
173 only 5 single amino acid polymorphic (SAP) loci in the E protein. Similar result was observed on
174 M gene and protein, with 37 SNPs and 9 SAPs. In contrast, 126 SNPs and 75 SAPs are detected
175 on N gene and protein, respectively. S protein, the most important factor that mediates virus entry
176 by receptor binding and membrane fusion and determines the infection ability of SARS-CoV-2
177 [25], harbors 155 SNPs on the alleles and 90 SAPs in the protein. Considering the size of
178 nucleotides and amino acid residues, N gene has the maximum sequence variability with 10.02%
179 (126/1257) SNPs and 17.90% (75/419) SAPs, respectively. However, S gene has most pairwise
180 nucleotide differences among the four structural genes, indicating a more genetic diversity of S
181 gene (Table 1). A key player in the virus transcription and assembly as N protein is [26, 27] a high

sequence variability of the N protein may indicate a vast adaption of the virus during host transmission. Previous study shows that high genetic variance has been found among bat SARSr-CoVs, particularly in the S gene[9]. Similar, higher nucleotide diversity (π , a major parameter to define genetic diversity) of S gene is also detected on SARS-CoV-2 isolates, suggesting this may benefit virus survival in the host of human beings.

187

188 **Table 1. Summary of genetic diversity of the 4 structural genes of the SARS-CoV-2 isolates**

Gene	Sequence, n*	Sequence length	h	π	S	θ	$\square$
E	2928	228	16	0.00012	14	0.00475	15
M	2891	669	40	0.00018	37	0.00665	40
N	2253	1260	131	0.00056	126	0.01081	130
S	2339	3825	173	0.00075	155	0.00753	169

189

190 h , Haplotypes,

191 π , Nucleotide diversity

192 S , Polymorphic sites

193 θ , Theta (per site) from S , population mutation ration

194 $\square$, Total number of mutations

195 * Some bases of SARS-CoV-2 genomic sequences are not exactly identified; thus, the number of
196 gene sequences were less than 3090.

197

198 **Distinct phylogenetic patterns of the four structural genes**

199 The phylogenetic analysis revealed that all SARS-CoV-2 E proteins form three clusters. Similar to
200 E protein, phylogenetic tree of SARS-CoV-2 M proteins is formed by three clusters with few
201 branches (Figs 1A and 1B). The results suggest both E and M genes may display a relatively high
202 conservation during coronavirus evolution. In contrast, SARS-CoV-2 N and S proteins show
203 distinct phylogenetic pattern as compared with that of E and M. Four and three main phylogenetic
204 clusters with various branches are identified in the N and S proteins, respectively (Figs 1C and D).

205 Given the crucial roles of N and S proteins in virus transcription, assembly, and entry to host cells,
206 whether SARS-CoV-2 isolates harbor different N and S variants (such as those clustered into
207 different clades) may influence their infection efficiency remains unknown, and requires further
208 study.

209 **Purifying selection drives the evolution at whole structural gene** 210 **levels of SARS-CoV-2 during its human to human transmission**

211 Although many studies demonstrated that recombination plays an important role on the emergence
212 of SARS-CoV-2 and its contribution to admit SARS-CoV-2 as a human infectious pathogen
213 [28-30], how this virus evolves during its global transmission has not been profiled yet. Therefore,
214 we first analyzed intragenic recombination events of each structural gene using RDP4. The results
215 indicate there were no recombination events occurred among the alleles of each gene (data not
216 shown). Recombination event were also assessed through reticulate network tree by phi test in
217 SplitsTree4. Although some internal nodes are noticed in N and S alleles, no significant evidence
218 for recombination is validated of each gene by Phi test ($p > 0.05$) (Fig 2). It indicates a relative
219 stable state of SARS-CoV-2 during its transmission although a possible genetic interaction of
220 different isolates might have occurred when it became a global pandemic [31, 32]. In addition,
221 Tajima's D, Fu and Li's D* and F* statistics were calculated to examine the mutation neutrality
222 hypothesis of the four structural genes of SARS-CoV-2. The results reveal that the evolution of all
223 four genes does not match the neutral hypothesis, but favor purifying selection (Table 2 and Fig 3).
224 The average of all pairwise dN/dS ratios (ω) among the alleles of each structural gene of
225 SARS-CoV-2 is 0.5443 in E, 0.1562 in M, 0.07978 in N, and 0.4980 in S gene, respectively. All
226 together, these results suggest that at the whole gene level, inconsistent purifying selection is the

227 main evolution force (Table 2). Li et al studied the origin of SARS-CoV-2 and showed evidence
228 of strong purifying selection in the S and other genes among bat, pangolin and human
229 coronaviruses, indicating similar strong evolutionary constraints in different host species [33].
230 Similarly, our results suggest purifying selection drives the evolution at the whole structural gene
231 level of SARS-CoV-2 during its transmission from human to human. This result also implies that
232 in general, the genetic variation on these structural genes will not confer a significant disadvantage
233 on the virus survival, and ratios reflect general variability of these genes and proteins. Considering
234 that no recombination happened, nonsynonymous mutations would be removed at a great rate
235 during the virus transmission [34].

236

237 **Table 2. Summary of neutrality for the four structural genes in SARS-CoV-2 isolates**

Gene	Tajima'sD	Fu and Li's D* test t	Fu and Li's F * test	dN	dS	dN/dS (ω)	Selection
E	-2.29974, P<0.01	-3.18477, P<0.02	-3.38505, P<0.02	0.006836	0.1256	0.5443	Purifying selection
M	-2.74611, P<0.001	-5.64276, P<0.02	-5.50855, P<0.02	0.001294	0.008296	0.1562	Purifying selection
N	-2.87598, P<0.001	-9.67153, P<0.02	-7.95879, P<0.02	0.000251	0.003146	0.07978	Purifying selection
S	-2.87646, P<0.001	-11.01171, P<0.02	-8.59037, P<0.02	0.000609	0.001223	0.4980	Purifying selection

238

239 **SARS-CoV2 S gene is operated by positive selection at a definitive**
240 **codon located at the C-terminal portion of S1 subunit and a potential**
241 **codon located at the signal sequence**

242 Guo et al. reported that the S gene of SARS-CoV populations in their natural host, Chinese
243 horseshoe bat (*Rhinolophus sinicus*), has evolved through positive selection at some codons[9]. As
244 mentioned above, at the whole gene level, purifying selection is the main force driving the

245 evolution of studied genes. Whether positive selection pressure accelerates the diversification of
246 the structural genes of SARS-CoV-2 remains unclear. Therefore, we used codon-substitution
247 models to estimate the ratio of nonsynonymous over synonymous substitutions (dN/dS), also
248 known as ω . The role of recombination in the polymorphism of four genes is excluded because no
249 intragenic recombination was detected (Fig 2). By using ML model, we don't find any codon of E
250 and M gene subjecting to positive selection obviously (data not shown). However, a potential
251 positive selection site 208A in N gene is identified by using M3 model, but not by any other
252 models especially the M8 model, suggesting a limited amount of evidence of positive selection in
253 N gene (S1 Table). For the S gene, we found the average ω is 0.37199 calculated by M0 model of
254 the codeML package, suggesting that purifying selection was a major force operating the
255 evolution of the S gene during its transmission among human beings. In three LRTs, all alternative
256 models (M3, M2a, M8) are significantly better fit ($P < 10^{-4}$) than relevant null models (M0, M1a,
257 M7), indicating that some sites of S were subjected to strong positive selection
258 ($\omega = 18.22175$ - 20.61283) (Table 3). A single positive selection site (614D) is identified in the S
259 gene with posterior probability of 1.000 in all the three models [21], a clear evidence showing that
260 this site is still experiencing positive selection when the virus transmitted from human to human.
261 The result is also validated using internal fixed effects likelihood (IFEL) and Evolutionary
262 Fingerprinting methods implemented in HyPhy package (Fig 4) [35-37]. To our surprise, the
263 positive selection site is not located at the receptor binding domain (RBD) or receptor binding
264 motif (RBM) as we anticipated, which play the most important role in virus-receptor interaction
265 and virus entry into host cells [38]. This result suggests that a relatively genetic stability of this
266 motif would benefit the virus survival. Intriguingly, the site under positive selection pressure

267 always has a D614G (for the S gene is 1841A>G) mutation, implying such mutation may enhance
 268 virus adaptability in human hosts. Another potential positive selection site at codon 5 is also
 269 identified, and a L5F mutation (for the S gene is 13C>T) is always found, with posterior
 270 probabilities greater than 0.95, 0.93 and 0.92 (critical values) calculated by M3, M2a and M8
 271 models (Table 3), respectively. Similar result was also confirmed by Evolutionary Fingerprinting
 272 method (S1 Fig). Considering signal sequence (SS) is a short hydrophobic peptide that plays an
 273 important role in guiding viral protein into the endoplasmic reticulum (ER) for proper folding and
 274 assembly [39], we postulate that L5F mutation may increase hydrophobicity of the SS, thus
 275 facilitating the entry of S protein into ER for folding and assembly, and in turn secretion of the
 276 virus.

277

Table 3. Log-likelihood values and parameter estimates for the SARS-CoV-2 S gene sequences

Model	Ln L	Estimates of parameters	Model compared	LRT P-value	Positive sites
		p0=0.96797, p1=0.02883, p2=0.00320			5 L 0.958* ,28 Y 0.850,221 S 0.901, 614 D
M3 (discrete)	-6766.339162	ω 0=0.26126, ω 1= 2.70530, ω2=20.61283			1.000*** ,677 Q 0.891
M0 (one ratio)	-6790.072925	ω 0=0.37199	M0 vs. M3	0.000000001	Not Allowed
		p0=0.81731, p1=0.17872, p2=0.00397			5 L 0.9258,28 Y 0.812,221 S 0.832, 614 D
M2a(selection)	-6766.432802	ω 0=0.17504, ω 1=1.00000, ω2=18.76936			1.000*** ,677 Q 0.828
		p0=0.70461, p1=0.29539			
M1a (neutral)	-6778.770190	ω 0=0.04395, ω 1=1.00000	M1a vs. M2a	0.000004385	Not Allowed
		p0=0.99578, p=0.40368, q=0.82224			5 L 0.931,28 Y 0.817,221 S 0.831, 614 D
M8(beta& ω)	-6768.829411	p1= 0.00422, ω= 18.22175			1.000*** ,677 Q 0.828
M7(beta)	-6779.230494	p=0.00857, q=0.02623	M7 vs.M8	0.000030400	Not Allowed

LnL is the log likelihood; ω is ratio of dN/dS , LRT P-value indicates the value of chi-square test; Parameters indicating positive selection are presented in bold; Positive selection sites were identified by the Bayes empirical Bayes (BEB) methods under M8 model. The posterior probabilities (p) $\geq$ 0.80 are shown, (p) $\geq$ 0.95 (p) $\geq$ 0.99, and (p)=1.000 are indicated by *, ** and ***, respectively. Yang *et al.* recommended that results from M8 model were preferred to find sites under positive selection pressure.

284 **Evolutionary relationship of S gene alleles with or without D614G** 285 **and L5F mutation**

286 Phylogenetic tree of S gene alleles was derived to test the evolutionary relationship among the
287 alleles with or without D614G mutation. As shown in Fig 5A, the 173 alleles of the S gene could
288 be clustered into four clades. Alleles with D614G mutation could be found in all 4 clades, among
289 which a dominant one contains 79 out of 85 alleles with such mutation. The remaining 6 mutated
290 S alleles are distributed in other 3 clades. The result suggests a potential common ancestor for the
291 majority of S alleles with D614G mutation, while some other maybe derived from alternative
292 ancestors. This result is also supported by the parsimony network of S gene alleles using PopART
293 (<http://popart.otago.ac.nz>) [40]. Two central alleles (representative virus isolates are WH01 and
294 GZMU0019) and associated alleles around them form a star scattering network, suggesting that
295 the S gene may have two potential origins (Fig 5B). All S alleles with D614G mutation are closely
296 related (with a few point mutations), and comprise a scattered star structure, suggesting the
297 expansion of SARS-CoV-2 population with D614G mutation on S gene. In contrast, alleles of the
298 N gene show a single ancestor analyzed by parsimony network though 3 phylogenetic clades are
299 identified (S2 Fig).

300 A total of 5 alleles with L5F mutation are found and all of them are in one clade, accounting for
301 83.33% of all alleles in the clade (S3A Fig). Further parsimony network analysis reveals that S
302 alleles with L5F mutation are not closely related, but distribute in both WH01 and GZMU0019
303 haplotype groups (S3B Fig). No scattered star structure of these alleles can be formed, indicating
304 L5F mutation might arise from independent origins other than that of D614G mutants. Limited

number of alleles with L5F mutation identified so far also suggests that L5F might subject to relatively less strength of the pressure and is still at early stage of positive selection.

Frequency of S allele with D614G mutation increased in SARS-CoV-2 isolates during human to human transmission

Considering that mutation of a positive selection site should be beneficial to the survival of the individuals carrying the mutation, we postulate that the D614G (1841A>G) mutation may help the spread of SARS-CoV-2. Some evidence has been obtained from the haplotype network of S alleles mentioned above (Fig 5B). S gene haplotypes (alleles) with D614G mutation (representative isolate GZMU0019) have evolved many subtypes and comprise a star structure with GZMU0019 in the center. This starburst pattern with one haplotype in the center and many other haplotypes surrounding the central haplotype suggests a signature of rapid population expansion [41]. To further study whether SARS-CoV-2 isolates with D614G mutation have advantage in survival during its transmission among human beings, we calculated the frequencies of S alleles carrying D614G mutation in each week from the collected SARS-CoV-2 isolates from December 24, 2019 to April 20, 2020 (17 weeks). Detailed information of these isolates including collection date, collection region and accession or biosample numbers is summarized on S3 and S4 Tables.

In 173 S gene alleles, 85 carry D614G mutation, accounting for 49.13% of all. Similarly, 47 out 99 S proteins carry D614G mutation, accounting for 47.47% of all. The first two isolates, GWHABKF000000001 and WH01 (isolated in December 24, 2019 and December 26, 2019, respectively), carry 614D in the S protein, while the first SARS-CoV-2 isolate with a D614G mutation is GZMU0019 in our collected dataset, isolated from a patient with COVID-19 on

February 5, 2020 (week 7 in our dataset). After that, except for week 9 and week 10 (possibly due to the small number of samples and sampling deviation), a spread trend that more and more proportion of isolates carry the D614G mutation in the S protein stands out. In the week 17, the last week of our dataset, 91.11% of SARS-CoV-2 isolates carry this mutation (S3 Table, Fig 6A). Further analysis reveals that the frequency of D614G mutation in the S gene was steadily increasing when combining data from week 6 to 17 (S3 Table, Fig 6B). To exclude the influence of sample size on the result (in some weeks, only 4-6 isolates were collected in the dataset), we reorganized the dataset by taking both the sample size and sampling time into account. Various panels of 200-300 isolates were studied and similar results were observed (S4 Table, Figs 6C and D). Taken together, these results suggest that SARS-CoV-2 isolates with D614G mutation may increase their ability to transmit, and contribute to the rapid spread of this virus to the world.

338

D614G mutation of S gene may destabilize S protein trimer and promote receptor binding and membrane fusion

The positive selected D614G mutation might play an important role for the adaptability of SARS-CoV-2 in both the host and the virus population[42]. Another explanation is that the mutation is driven by specific interaction between high level of virus sequence divergence and polymorphic host receptors or interacting proteins[43]. S protein is the key determinant for the tissue tropism and host range and specificity of coronavirus such as SARS-CoV-2. The virus infects host cells through the interaction between the S protein and its cellular receptor, named ACE2 [8]. In this process, virus entry requires the precursor S protein cleaved by cellular proteases including trypsin, furin, transmembrane serine protease 2 (TMPRSS2), or endosomal

349 cathepsin L, which generate the receptor binding subunit S1 and the membrane fusion S2 [44-46].

350 From structural studies in both SARS-CoV and SARS-CoV-2, receptor binding domain (RBD)

351 located at the C-terminal of S1 and the adjacent N-terminal domain (NTD) are relatively flexible,

352 which is the feature required for receptor recognition and subsequent membrane fusion[47, 48].

353 We found that the D614G mutation is located at the subdomain 2 (SD2) that at the C-terminal of

354 RBD and close to the two potential cleavage sites between S1 and S2 [48] (Fig 7A). Considering

355 that positive selection is usually beneficial to the survival of the individual carrying the mutation,

356 we speculate that the D614G mutation may facilitate structural conformation change to promote

357 receptor binding or membrane fusion[5, 44], and in turn improving the infection efficiency. From

358 the latest cryo-electron microscopy (cryo-EM) structure of SARS-CoV-2 S protein, the negatively

359 charged sidechain of D614 points towards the positively charged sidechain of K854 from the

360 neighboring monomer (Fig 7B) [48] . The distance between the closest atoms of the two residues

361 is 2.6 Å, which is an optimal distance to form salt bridge (Fig 7C). From the modelled structure

362 with D614G mutation, the distance is increased to 5.2 Å (Fig 7D), which would potentially abolish

363 the salt bridge and destabilize the integrity of the S trimer in wild type. It has been reported that

364 human receptor ACE2 binds to an “open” conformation of S protein, where RBD move away from

365 the core structure and expose its receptor binding surface. The entire S trimer then undergoes a

366 series of dramatic conformation changes, including cleavages between S1 and S2, disassociation

367 of S1 and post-fusion transformation of S2 [49, 50]. Changes including mutations at cleavage sites

368 and adding internal crosslinks in S trimer would keep the protein in a stable and “closed”

369 conformation where the receptor binding surface of RBD is inaccessible [48, 51]. Therefore, we

370 hypothesize that the highly transmissible D614G mutation driven by the positive selection through

371 evolution promotes accessibility of RBD by losing a critical salt bridge between the S protein
372 monomers, which subsequently triggers membrane fusion upon ACE2 binding.

373

374 **Conclusions**

375 We present modern molecular evolution analyses on a large and comparative set of SARS-CoV-2
376 structural gene sequences, derived from an international collection of SARS-CoV-2 isolates.
377 Distinct phylogenetic patterns of four structural proteins of SARS-CoV-2 are depicted. Protein
378 sequence comparisons show E and M genes exhibit a relatively close relationship to bat
379 SARSr-CoV, suggesting the evolution conservation of these two genes. In contrast, relatively high
380 genetic variation is observed in N and S proteins among SARS-CoV-2 isolates, implying extensive
381 adaptability of N and S genes. No clear intragenic recombination is detected of these four genes,
382 suggesting that it is not the major force to drive the evolution of the four genes. However, our
383 analyses show purifying selection pressure may be the main force operating the evolution at whole
384 gene levels of SARS-CoV-2 during its human to human transmission. We also identify a codon in
385 S gene definitively experiencing positive selection pressure, and always leads to the D614G
386 mutation in S proteins. S alleles with D614G mutation have expanded rapidly among
387 SARS-CoV-2 isolates. D614G mutation significantly extends the distance between monomers in
388 the S protein trimer, which may disrupt the salt bridge formed by D614 and K854 between
389 monomers, promote RBD opening, and facilitate the entry of the virus into host cells, thus
390 contributing to the diffusion of this mutated alleles. Codon 5 of S gene is another potential positive
391 selection site. Although a limited number of alleles with L5F mutation is identified, it may
392 potentially affect the assembly and secretion of SARS-CoV-2. A close eye on L5F mutation may

393 be required in case another expansion occurs. As S protein is a key target for SARS-CoV-2
 394 vaccines, therapeutic antibodies, and diagnostics, the D614G mutation of S should be paid more
 395 attention. Owing that the exact mechanism remains unclear, further study should focus on the
 396 exact function of these mutation sites and how they affect the expansion of these mutated alleles
 397 on SARS-CoV-2.

398

399 **Acknowledgements**

400 This research was supported by National Natural Science Foundation of China (grant number
 401 31870001) to X.Y.Z.

402 **Conflict of Interest** The authors have declared no conflict of interests.

403 **References**

- 404 1. Liu DX, Yuan Q, Liao Y. Coronavirus envelope protein: a small membrane protein with
405 multiple functions. Cellular and molecular life sciences : CMLS. 2007;64(16):2043-8. Epub
406 2007/05/29. doi: 10.1007/s00018-007-7103-1. PubMed PMID: 17530462; PubMed Central
407 PMCID: PMCPMC7079843.
- 408 2. Jimenez-Guardeno JM, Nieto-Torres JL, DeDiego ML, Regla-Nava JA,
409 Fernandez-Delgado R, Castano-Rodriguez C, et al. The PDZ-binding motif of severe acute
410 respiratory syndrome coronavirus envelope protein is a determinant of viral pathogenesis.
411 PLoS pathogens. 2014;10(8):e1004320. Epub 2014/08/15. doi: 10.1371/journal.ppat.1004320.
412 PubMed PMID: 25122212; PubMed Central PMCID: PMCPMC4133396.
- 413 3. Arndt AL, Larson BJ, Hogue BG. A conserved domain in the coronavirus membrane
414 protein tail is important for virus assembly. Journal of virology. 2010;84(21):11418-28. Epub
415 2010/08/20. doi: 10.1128/JVI.01131-10. PubMed PMID: 20719948; PubMed Central PMCID:
416 PMCPMC2953170.
- 417 4. McBride R, van Zyl M, Fielding BC. The coronavirus nucleocapsid is a multifunctional
418 protein. Viruses. 2014;6(8):2991-3018. Epub 2014/08/12. doi: 10.3390/v6082991. PubMed
419 PMID: 25105276; PubMed Central PMCID: PMCPMC4147684.
- 420 5. Belouzard S, Millet JK, Licitra BN, Whittaker GR. Mechanisms of coronavirus cell entry
421 mediated by the viral spike protein. Viruses. 2012;4(6):1011-33. Epub 2012/07/21. doi:
422 10.3390/v4061011. PubMed PMID: 22816037; PubMed Central PMCID: PMCPMC3397359.
- 423 6. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated
424 with human respiratory disease in China. Nature. 2020;579(7798):265-9. Epub 2020/02/06.

doi: 10.1038/s41586-020-2008-3. PubMed PMID: 32015508; PubMed Central PMCID: PMCPMC7094943.

7. Y. Z, S. Z, J. C, C. W, W. Z, B. Z. Analysis of variation and evolution of SARS-CoV-2 genome. Journal of Southern Medical University. 2020;02:152-8. doi: 10.12122/j.issn.1673-4254.2020.02.23.

8. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579(7798):270-3. Epub 2020/02/06. doi: 10.1038/s41586-020-2012-7. PubMed PMID: 32015507; PubMed Central PMCID: PMCPMC7095418.

9. Guo H, Hu B-J, Yang X-L, Zeng L-P, Li B, Ouyang S-Y, et al. Evolutionary arms race between virus and host drives genetic diversity in bat SARS related coronavirus spike genes. 2020;2020.05.13.093658. doi: 10.1101/2020.05.13.093658. bioRxiv.

10. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular biology and evolution. 2018;35(6):1547-9. Epub 2018/05/04. doi: 10.1093/molbev/msy096. PubMed PMID: 29722887; PubMed Central PMCID: PMCPMC5967553.

11. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, et al. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Datasets. 2017;34(12).

12. Martin DP, Murrell B, Khoosal A, Muhire B. Detecting and Analyzing Genetic Recombination Using RDP4. Methods in molecular biology. 2017;1525:433-60. doi: 10.1007/978-1-4939-6622-6_17. PubMed PMID: 27896731.

- 447 13. Padidam M, Sawyer S, Fauquet CM. Possible emergence of new geminiviruses by
448 frequent recombination. *Virology*. 1999;265(2):218-25. Epub 1999/12/22. doi:
449 10.1006/viro.1999.0056. PubMed PMID: 10600594.
- 450 14. Martin DP, Posada D, Crandall KA, Williamson C. A modified bootscan algorithm for
451 automated identification of recombinant sequences and recombination breakpoints. *AIDS Res*
452 *Hum Retroviruses*. 2005;21(1):98-102. doi: 10.1089/aid.2005.21.98. PubMed PMID:
453 15665649.
- 454 15. Smith JM. Analyzing the mosaic structure of genes. *Journal of molecular evolution*.
455 1992;34(2):126-9. PubMed PMID: 1556748.
- 456 16. Posada D. Evaluation of methods for detecting recombination from DNA sequences:
457 empirical data. *Molecular biology and evolution*. 2002;19(5):708-17. PubMed PMID:
458 11961104.
- 459 17. Gibbs MJ, Armstrong JS, Gibbs AJ. Sister-scanning: a Monte Carlo procedure for
460 assessing signals in recombinant sequences. *Bioinformatics*. 2000;16(7):573-82. PubMed
461 PMID: 11038328.
- 462 18. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies.
463 *Molecular biology and evolution*. 2006;23(2):254-67. Epub 2005/10/14. doi:
464 10.1093/molbev/msj030. PubMed PMID: 16221896.
- 465 19. Zhan XY, Zhu QY. Molecular evolution of virulence genes and non-virulence genes in
466 clinical, natural and artificial environmental *Legionella pneumophila* isolates. *PeerJ*.
467 2017;5:e4114. Epub 2017/12/12. doi: 10.7717/peerj.4114. PubMed PMID: 29226035; PubMed
468 Central PMCID: PMC5719964.

469 20. Gao F, Chen C, Arab DA, Du Z, He Y, Ho SYWJE, et al. EasyCodeML: A visual tool for
470 analysis of selection using CodeML. 2019.

471 21. Yang Z, Wong WS, Nielsen R. Bayes empirical bayes inference of amino acid sites under
472 positive selection. *Molecular biology and evolution*. 2005;22(4):1107-18. doi:
473 10.1093/molbev/msi097. PubMed PMID: 15689528.

474 22. Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, et al.
475 HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies.
476 *Molecular biology and evolution*. 2019;37(1):295-9. doi: 10.1093/molbev/msz197 *Molecular*
477 *Biology and Evolution*.

478 23. The PyMOL Molecular Graphics System, Version 1.5.X Schrödinger, LLC.

479 24. Narayanan K, Makino S. Cooperation of an RNA packaging signal and a viral envelope
480 protein in coronavirus RNA packaging. *Journal of virology*. 2001;75(19):9059-67. Epub
481 2001/09/05. doi: 10.1128/JVI.75.19.9059-9067.2001. PubMed PMID: 11533169; PubMed
482 Central PMCID: PMCPMC114474.

483 25. Letko M, Marzi A, Munster V. Functional assessment of cell entry and receptor usage for
484 SARS-CoV-2 and other lineage B betacoronaviruses. *Nat Microbiol*. 2020;5(4):562-9. Epub
485 2020/02/26. doi: 10.1038/s41564-020-0688-y. PubMed PMID: 32094589; PubMed Central
486 PMCID: PMCPMC7095430.

487 26. Voss D, Kern A, Traggiai E, Eickmann M, Stadler K, Lanzavecchia A, et al.
488 Characterization of severe acute respiratory syndrome coronavirus membrane protein. *FEBS*
489 *letters*. 2006;580(3):968-73. Epub 2006/01/31. doi: 10.1016/j.febslet.2006.01.026. PubMed
490 PMID: 16442106; PubMed Central PMCID: PMCPMC7094741.

- 491 27. Tseng YT, Wang SM, Huang KJ, Lee AI, Chiang CC, Wang CT. Self-assembly of severe
492 acute respiratory syndrome coronavirus membrane protein. The Journal of biological
493 chemistry. 2010;285(17):12862-72. Epub 2010/02/16. doi: 10.1074/jbc.M109.030270.
494 PubMed PMID: 20154085; PubMed Central PMCID: PMC2857088.
- 495 28. Wong MC, Javornik Cregeen SJ, Ajami NJ, Petrosino JF. Evidence of recombination in
496 coronaviruses implicating pangolin origins of nCoV-2019. 2020:2020.02.07.939207. doi:
497 10.1101/2020.02.07.939207 bioRxiv.
- 498 29. Wu Y. Strong evolutionary convergence of receptor-binding protein spike between
499 COVID-19 and SARS-related coronaviruses. 2020:2020.03.04.975995. doi:
500 10.1101/2020.03.04.975995. bioRxiv.
- 501 30. Wu A, Niu P, Wang L, Zhou H, Zhao X, Wang W, et al. Mutations, Recombination and
502 Insertion in the Evolution of 2019-nCoV. 2020:2020.02.29.971101. doi:
503 10.1101/2020.02.29.971101 bioRxiv.
- 504 31. Iceland patient infected by two strains. The Standard.
505 2020;<https://www.thestandard.com.hk/section-news/section/11/217711/Iceland-patient--infected-by--two-strains>
506 [ed-by--two-strains](https://www.thestandard.com.hk/section-news/section/11/217711/Iceland-patient--infected-by--two-strains).
- 507 32. Mallapaty S. How sewage could reveal true scale of coronavirus outbreak. Nature.
508 2020;580(7802):176-7. Epub 2020/04/05. doi: 10.1038/d41586-020-00973-x. PubMed PMID:
509 32246117.
- 510 33. Li X, Giorgi EE, Marichann MH, Foley B, Xiao C, Kong X-P, et al. Emergence of
511 SARS-CoV-2 through Recombination and Strong Purifying Selection.
512 2020:2020.03.20.000885. doi: 10.1101/2020.03.20.000885. bioRxiv.

- 513 34. Hughes AL, Hughes MA. More effective purifying selection on RNA viruses than in DNA
514 viruses. *Gene*. 2007;404(1-2):117-25. Epub 2007/10/12. doi: 10.1016/j.gene.2007.09.013.
515 PubMed PMID: 17928171; PubMed Central PMCID: PMCPMC2756238.
- 516 35. Pond SLK, Muse SV. HyPhy: Hypothesis Testing Using Phylogenies: Springer New York;
517 2005. 676-9 p.
- 518 36. Pond SL, Scheffler K, Gravenor MB, Poon AF, Frost SD. Evolutionary fingerprinting of
519 genes. *Molecular biology and evolution*. 2010;27(3):520-36. Epub 2009/10/30. doi:
520 10.1093/molbev/msp260. PubMed PMID: 19864470; PubMed Central PMCID:
521 PMCPMC2877558.
- 522 37. Kosakovsky Pond SL, Frost SD. Not so different after all: a comparison of methods for
523 detecting amino acid sites under selection. *Molecular biology and evolution*.
524 2005;22(5):1208-22. Epub 2005/02/11. doi: 10.1093/molbev/msi105. PubMed PMID:
525 15703242.
- 526 38. Wan Y, Shang J, Graham R, Baric RS, Li F. Receptor Recognition by the Novel
527 Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS
528 Coronavirus. *Journal of virology*. 2020;94(7). Epub 2020/01/31. doi: 10.1128/JVI.00127-20.
529 PubMed PMID: 31996437; PubMed Central PMCID: PMCPMC7081895.
- 530 39. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veasler D. Structure, Function, and
531 Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell*. 2020;181(2):281-92 e6. Epub
532 2020/03/11. doi: 10.1016/j.cell.2020.02.058. PubMed PMID: 32155444; PubMed Central
533 PMCID: PMCPMC7102599.
- 534 40. Clement M, Snell Q, Walker P, Posada D, Crandall KJP, Distributed Processing

- 535 Symposium IP. TCS: Estimating gene genealogies. 2002;2:184.
- 536 41. Bubac CM, Spellman GMJTAOA. How connectivity shapes genetic structure during range
537 expansion: Insights from the Virginia's Warbler. 2016;(2):2.
- 538 42. Duxbury EM, Day JP, Maria Vespasiani D, Thuringer Y, Tolosana I, Smith SC, et al.
539 Host-pathogen coevolution increases genetic variation in susceptibility to infection. eLife.
540 2019;8. Epub 2019/05/01. doi: 10.7554/eLife.46440. PubMed PMID: 31038124; PubMed
541 Central PMCID: PMCPMC6491035.
- 542 43. Meyerson NR, Sawyer SL. Two-stepping through time: mammals and viruses. Trends in
543 microbiology. 2011;19(6):286-94. Epub 2011/05/03. doi: 10.1016/j.tim.2011.03.006. PubMed
544 PMID: 21531564; PubMed Central PMCID: PMCPMC3567447.
- 545 44. Lu G, Wang Q, Gao GF. Bat-to-human: spike features determining 'host jump' of
546 coronaviruses SARS-CoV, MERS-CoV, and beyond. Trends in microbiology.
547 2015;23(8):468-78. Epub 2015/07/25. doi: 10.1016/j.tim.2015.06.003. PubMed PMID:
548 26206723; PubMed Central PMCID: PMCPMC7125587.
- 549 45. Bestle D, Heindl MR, Limburg H, van TVL, Pilgram O, Moulton H, et al. TMPRSS2 and
550 furin are both essential for proteolytic activation and spread of SARS-CoV-2 in human airway
551 epithelial cells and provide promising drug targets. 2020:2020.04.15.042085. doi:
552 10.1101/2020.04.15.042085. bioRxiv.
- 553 46. Ou X, Liu Y, Lei X, Li P, Mi D, Ren L, et al. Characterization of spike glycoprotein of
554 SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. Nature
555 communications. 2020;11(1):1620. Epub 2020/03/30. doi: 10.1038/s41467-020-15562-9.
556 PubMed PMID: 32221306; PubMed Central PMCID: PMCPMC7100515.

557 47. Gui M, Song W, Zhou H, Xu J, Chen S, Xiang Y, et al. Cryo-electron microscopy
558 structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for
559 receptor binding. Cell research. 2017;27(1):119-29. Epub 2016/12/23. doi:
560 10.1038/cr.2016.152. PubMed PMID: 28008928; PubMed Central PMCID: PMC5223232.

561 48. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, et al. Cryo-EM
562 structure of the 2019-nCoV spike in the prefusion conformation. Science.
563 2020;367(6483):1260-3. Epub 2020/02/23. doi: 10.1126/science.abb2507. PubMed PMID:
564 32075877.

565 49. Walls AC, Xiong X, Park YJ, Tortorici MA, Snijder J, Quispe J, et al. Unexpected Receptor
566 Functional Mimicry Elucidates Activation of Coronavirus Fusion. Cell. 2019;176(5):1026-39
567 e15. Epub 2019/02/05. doi: 10.1016/j.cell.2018.12.028. PubMed PMID: 30712865; PubMed
568 Central PMCID: PMC6751136.

569 50. Walls AC, Tortorici MA, Snijder J, Xiong X, Bosch BJ, Rey FA, et al. Tectonic
570 conformational changes of a coronavirus spike glycoprotein promote membrane fusion.
571 Proceedings of the National Academy of Sciences of the United States of America.
572 2017;114(42):11157-62. Epub 2017/10/27. doi: 10.1073/pnas.1708727114. PubMed PMID:
573 29073020; PubMed Central PMCID: PMC5651768.

574 51. Xiong X, Qu K, Ciazynska KA, Hosmillo M, Carter AP, Ebrahimi S, et al. A thermostable,
575 closed, SARS-CoV-2 spike protein trimer. 2020:2020.06.15.152835. doi:
576 10.1101/2020.06.15.152835. bioRxiv.

577
578

579 **Supporting information**

580 **S1 Table.** SARS-CoV-2 isolates information.

581 **S2 Table.** Log-likelihood values and parameter estimates for the SARS-CoV-2 N gene sequences.

582 **S3 Table.** Detailed information of SARS-CoV-2 isolates with full length sequence of S gene. The
583 data are organized by weekly.

584 **S4 Table.** Detailed information of SARS-CoV-2 isolates with full length sequence of S gene. The
585 data are organized by panels. Each panel contains 200-300 isolates by combining isolates from
586 several days.

587 **S1 Fig. The evolutionary relationship of N alleles. A.** Phylogenetic tree of N gene based on
588 nucleotide sequences of 131 alleles. The evolutionary history is inferred using the Maximum
589 Likelihood method and Tamura-Nei model. The tree is drawn to scale, with branch lengths
590 measured in the number of substitutions per site. Bootstrap values more than 0.5 are shown. **B.**
591 Parsimony network of SARS-CoV-2 N gene haplotype (allele) diversity obtained from 3090
592 isolates worldwide. Each oblique line linking between haplotypes (haplotype name is shown as its
593 representative isolate name) represents one mutational difference. The ancestral haplotype, or root
594 of the network, is labeled with a square, and represent haplotype name is marked red.

595

596 **S2 Fig. Evolutionary relationship of S alleles with or without L5F mutation. A.** Phylogenetic
597 tree of S gene based on nucleotide sequences of 173 alleles. Each clade is highlighted with
598 different color. Alleles are shown with their representative isolate names, and alleles with L5F
599 mutation are highlighted in blue. Bootstrap values more than 0.5 are shown. **B.** Parsimony
600 network of SARS-CoV-2 S gene_haplotype (allele) diversity obtained from 3090 isolates

601 worldwide. Each oblique line linking between haplotypes (haplotype name is shown as its
602 representative isolate name) represents one mutational difference. Unlabeled nodes (Gray circle)
603 indicate inferred steps have not found in the sampled populations yet. The ancestral haplotype, or
604 root of the network, is labeled with a square, and represent haplotype name is marked green or red.
605 The blue nodes indicate haplotypes with L5F mutation. Dotted boxes indicate major haplotype
606 groups. Haplotypes include in red dotted boxes are with D614G mutation while those included in
607 black dotted boxes are without D614G mutation.

608

609 **S3 Fig. Positive selection analysis of S gene codon 5 by Evolutionary Fingerprinting method.**

610 Log (Bayes Factor) for positive selection at codon 5 of S gene and its frequencies. The cut-off
611 value for the Bayes factor (BF) in the Evolutionary Fingerprinting method was set at 25 to reflect
612 a positive selection at a given site (Posterior probability>0.95). Pr {BF>25} indicates posterior
613 probability of Bayes Factor >25.

614

615

616 **Figure legends**

617

618 **Figure 1.** Phylogenetic tree of E (A), M (B), N (C), and S(D) proteins of SARS-CoV-2. Major
619 clades are highlighted with different color. The tree shows topology of the protein of each allele,
620 named by their representative isolates.

621

622 **Figure 2.** Reticulate network trees of E (A), M (B), N (C) and S (D) alleles of SARS-CoV-2
623 analyzed by the neighbor-net algorithm of SplitsTree4. Scale bars indicate number of substitutions
624 per site. All internal nodes represent hypothetical ancestral alleles and edges that correspond to
625 reticulate events such as recombination. Red arrows indicate edges. Because there are too few
626 informative characters to use the Phi test for E and M genes, *p-values* of Phi test of N and S genes
627 are shown.

628

629 **Figure 3.** Tajima's D, Fu and Li's D* and F* test for the four structural gene alleles of
630 SARS-CoV-2. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

631

632

633 **Figure 4. Positive selection analysis of S gene codons by IFEL and Evolutionary**

634 **Fingerprinting methods.** **A.** Diagram of selection analysis result of S codons by IFEL method.

635 Asterisk indicates the positive selection site with statistical significance ($p < 0.01$). **B.** Log (Bayes

636 Factor) for positive selection at codon 614 of S gene and its frequencies. The cut-off value for the

637 Bayes factor (BF) in the Evolutionary Fingerprinting method was set at 25 to reflect a positive

638 selection at a given site (Posterior probability>0.95). Pr {BF>25} indicates posterior probability of
639 Bayes Factor >25.

640

641 **Figure 5. Evolutionary relationship of S alleles with or without D614G mutation. A.**

642 Phylogenetic tree of S gene based on nucleotide sequences of 173 alleles. The evolutionary history
643 is inferred using the Maximum Likelihood method and Tamura-Nei model. The tree is drawn to
644 scale, with branch lengths measured in the number of substitutions per site. Each clade is
645 highlighted with different color. Alleles are shown with their representative isolate names, and
646 alleles with D614G mutation are highlighted in red. Bootstrap values more than 0.5 are shown. **B.**

647 Parsimony network of SARS-CoV-2 S gene haplotype (allele) diversity obtained from 3090
648 isolates worldwide. Each oblique line linking between haplotypes (haplotype name is shown as its
649 representative isolate name) represents one mutational difference. Unlabeled nodes (Gray circle)
650 indicate inferred steps have not found in the sampled populations yet. The ancestral haplotype, or
651 root of the network, is labeled with a square, and represent haplotype name is marked green or red.
652 The red nodes indicate haplotypes with D614G mutation, while green or black nodes indicate
653 haplotypes without D614G mutation. Dotted boxes indicate major haplotype groups.

654

655 **Figure 6. Expansion of S alleles with D614G mutation during SARS-CoV-2 human to human**

656 transmission. **A.** Percentage of SARS-CoV-2 isolates carrying the alleles of D614G mutation in
657 each week collected. **B.** Frequencies of D614G mutation in the S gene in each period of time (Four
658 to five weeks' data are combined). **C.** Percentage of SARS-CoV-2 isolates carrying the alleles with

659 D614G mutation in each period of time. **D.** Frequencies of D614G mutation in the S gene in each
660 period of time. * $p < 0.05$; ** $p < 0.01$.

661

662

663 **Figure 7.** The structure of the S protein of SARS-CoV-2 and potential influence of D614G
664 mutation on its structural change. **A.** Schematic of the primary structure of SARS-CoV-2 S protein
665 colored by domains. Some boundary-residues are listed. The S1/S2 cleavage sites are indicated by
666 arrows. RBD: receptor binding domain; RBM: receptor of binding motif; FP: fusion peptide,
667 HR1/2: heptad repeat 1/2; TM: transmembrane domain; CT: cytoplasmic tail; NTD: N-terminal
668 domain; CTD: C-terminal domain; SD1: subdomain 1; SD2: subdomain 2. The structure of the S
669 protein trimer of SARS-CoV-2 and potential influence of D614G mutation on its structural change.
670 **B.** Experimentally determined structure of SARS-CoV-2 S protein trimer (PDB ID is 6VSB and
671 the amino acid sequences is the same as WH01 isolate). **C.** D614-K854 inter-monomer salt bridge.
672 **D.** G614-K854 inter-monomer salt bridge. The distance of the salt bridge is increased from 2.6 to
673 5.2 Å in D614G mutation as shown.

674

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

A **B**

A

B

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