Molecular Evolution of SARS-CoV-2 Structural Genes: Evidence of Positive Selection in the Spike Glycoprotein

Xiao-Yong Zhan
The Seventh Affiliated Hospital, Sun Yat-sen University
Ying Zhang
The Seventh Affiliated Hospital, Sun Yat-sen University
Xuefu Zhou
Sun Yat-sen University Cancer Center
Ke Huang
Sun Yat-sen University Cancer Center
Yichao Qian
The Seventh Affiliated Hospital, Sun Yat-sen University
Yang Leng
The Seventh Affiliated Hospital, Sun Yat-sen University
Leping Yan
Sun Yat-sen University Cancer Center
Bihui Huang (bihui_huang@126.com)
The Seventh Affiliated Hospital, Sun Yat-sen University
Yulong He
The Seventh Affiliated Hospital, Sun Yat-sen University

Research article

Keywords: SARS-CoV-2, Structural genes, Molecular evolution, Positive selection, Spike glycoprotein

Posted Date: August 31st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-42498/v2

Abstract

Background: SARS-CoV-2 has caused a global pandemic since early 2020 and remains a serious public health issue worldwide. Four structural genes, envelope (E), membrane (M), nucleocapsid (N) and spike (S), play a key role in controlling entry into human cells and virion assembly of SARS-CoV-2. The evolution of these genes may determine infectivity of SARS-CoV-2, but thus far, little is known about them.

Methods: We analyzed 3090 SARS-CoV-2 isolates from the GenBank database to determine the evolutionary patterns of the four structural genes by employing various molecular evolution algorithms.

Results: Phylogenetic analyses showed that global SARS-CoV-2 isolates can be clustered into three to four major clades based upon protein sequence. Although intragenic recombination was not detected among different alleles, purifying selection has affected the evolution of these genes. By analyzing full genomic sequences of these alleles, our result revealed that codon 614 of the S glycoprotein has been subjected to a strong positive selection pressure, and a consistent D614G mutation was identified. Additionally, another potentially positive selection site at codon 5 in the signal sequence of the S protein was also identified with a consistent L5F mutation. The allele containing the D614G mutation has undergone significant expansion during SARS-CoV-2 transmission, implying a better adaptability of isolates with the mutation. Nevertheless, L5F allele expansion was found to be relatively restricted. The D614G mutation is located at subdomain 2 (SD2) of the C-terminal portion (CTP) of the S1 subunit. Protein structural modeling showed that the D614G mutation may cause the disruption of a salt bridge between S protein monomers and increase their flexibility, consequently promoting receptor binding domain (RBD) opening, virus attachment, and ultimately entry into host cells. Located at the signal sequence of S protein, the L5F mutation may facilitate protein folding, assembly, and secretion of the virus.

Conclusions: This is the first reported evidence of positive Darwinian selection in the spike gene of SARS-CoV-2. This finding contributes to a broader understanding of the adaptive mechanisms of this virus, and provide insight for the development of novel therapeutic approaches, as well as the creation of effective vaccines, through targeting mutation sites.

Background

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of an emerging coronavirus disease (COVID-19) that has thus far caused more than 430,000 deaths, is still a serious global threat. The genome of SARS-CoV-2 consists of a single-stranded, positive-sense RNA strand, approximately 30 kb in length, with a 5’ cap and 3’-polyA tail [1]. The SARS-CoV-2 genome possesses six major open reading frames (ORFs) that encode 27 different proteins, in which four are structural proteins: Envelope (E), membrane (M), nucleocapsid (N) and spike (S) [1, 2]. Previous research has demonstrated the important functions of these proteins in virus entry, transcription, and virion particle assembly of

SARS-CoV-2[1, 3-6]. The E protein is a small envelope protein of 75 amino acids. Given the close genetic relationship between SARS-CoV-2 and SARS-CoV, functions of this protein may include virion assembly and morphogenesis[3]. In addition, induction of apoptosis of host cells may be another crucial function of the SARS-CoV-2 E protein, thus making it a potential determinant of viral pathogenesis [7]. The M protein, consisting of 222 amino acids, is the most abundant component of the viral envelope and plays a key role in virion assembly [4]. The N protein, composed of 419 amino acids, may form complexes with genomic RNA, interact with the viral membrane protein, and play a critical role in enhancing the efficiency of virus transcription and assembly[5]. The S protein, consisting of 1,273 amino acids, is the most important factor for mediating virus entry, and is a primary determinant of cell tropism and pathogenesis of SARS-CoV-2 [6, 8, 9]. Prior studies have demonstrated that SARS-CoV-2 has evolved and some genetic evolutionary features have been reported[10-13]. The complete genomic sequence of SARS-CoV-2 is 79.6% identical to SARS-CoV, and 96.0% identical to a SARS-related bat (Rhinolophus affinis) coronavirus (SARSr-CoV), RaTG13 [11]. Although no positive temporal evolution signal was found between SARS-CoV-2 and RaTG13, SARS-CoV-2 exhibits a strong positive temporal evolutionary relationship with bat-SL-CoVZC45, which has a slightly less similar genomic sequence (87.5% identity) than RaTG13 [12]. By combining the phylogenetic analyses of full-length coronavirus genomes, a potential bat (Order: Chiroptera) origin of SARS-CoV-2 is indicated [11]. A recent study reported that the spike (S) gene (the coding gene of the S protein) of SARSr-CoV's from their natural reservoir host, the Chinese horseshoe bat (Rhinolophus sinicus), has coevolved with R. sinicus angiotensin converting enzyme 2 (ACE2) via positive selection [13]. As a single-stranded positive-sense RNA virus, SARS-CoV-2 has caused a pandemic within half of a year, suggesting it may evolve rapidly. The evolution of SARS-CoV-2 based on structural genes from human to human transmission has, however, not been investigated thoroughly. The primary purpose of this study was to investigate the evolutionary pattern of the four structural genes of SARS-CoV-2, derived from a global isolate collection including the E, M, N and S genes. Various molecular evolution and selection analysis approaches were employed to identify the phylogeny of the four structural proteins and potential selection effects on genes. Our study reveals that intragenic recombination does not contribute to the evolution of these genes, but rather purifying selection is the main evolutionary force at work. Moreover, a D614G mutation in the S protein is affected by strong positive selection and may be partially responsible for the rapid spread of SARS-CoV-2. Lastly, another potential L5F mutation may be being operated on by positive selection, but with relatively weaker pressure compared to D614G.

Results

Characteristics of SARS-CoV-2 isolates, structural genes, and protein sequences

The 3090 SARS-CoV-2 isolates harbor 16 E alleles, 40 M alleles, 131 N alleles, and 173 S alleles. These alleles correspond to 10, 14, 88 and 99 different amino acid sequences of E, M, N, and S proteins, respectively. Protein sequence comparison of the WH01 isolate with a SARSr-CoV isolate, bat-SL-CoVZC45, shows a similarity of 100% (75/75) in E protein, 98.65% (219/222) in M protein, 94.27% (395/419) in N protein, and 80.06% (1171/1273) in S protein. These results imply a close homology

between SARS-CoV-2 and bat SARSr-CoV, particularly with E and M proteins. Alternatively, it indicates an extreme conservation of E and M proteins and their functions among coronaviruses[14].

Further analysis revealed that there are 14 single nucleotide polymorphisms (SNPs) of E gene, but only 5 single amino acid polymorphic (SAP) loci on the E protein. A similar result was observed for the M gene and protein, with 37 SNPs and 9 SAPs. In contrast, 126 SNPs and 75 SAPs were detected on N gene and protein, respectively. S protein, the most important component that mediates virus entry by receptor binding and membrane fusion and determines the infectivity of SARS-CoV-2 [15], harbors 155 SNPs and 90 SAPs. Considering the size of nucleotides and amino acid residues, N gene has the maximum sequence variability with 10.02% (126/1257) SNPs and 17.90% (75/419) SAPs, respectively. However, S gene has the most pairwise nucleotide differences among the four structural genes, indicating a relatively greater genetic diversity (Table 1).

Table 1. Genetic diversity indices of the 4 structural genes of SARS-CoV-2 isolates

ene	Sequence, n*	Sequence length	*h*	*π*	*S*	*θ*	SD of *θ*	*k*
E	2928	228	16	0.00012	14	0.00475	0.00232	15
M	2891	669	40	0.00018	37	0.00665	0.00145	40
N	2253	1260	131	0.00056	126	0.01081	0.00432	130
S	2339	3825	173	0.00075	155	0.00753	0.00093	169

h, Haplotypes,

π, Nucleotide diversity

S, Polymorphic sites

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

k, Total number of mutations

SD, Standard deviation

* Some bases of SARS-CoV-2 genomic sequences are not exactly identified; thus, the number of gene sequences (n) is less than 3090.

Distinct phylogenetic patterns of the four structural genes

Phylogenetic analysis revealed that all SARS-CoV-2 E proteins form three clusters. Similar to E protein, a phylogenetic tree of SARS-CoV-2 M proteins is formed from three clusters with few branches (Fig. 1a and b). Results suggest that both E and M genes may display a relatively high conservation during coronavirus evolution. In contrast, SARS-CoV-2 N and S proteins show distinct phylogenetic patterns

compared with E and M proteins. Four and three main phylogenetic clusters with various branches are identified in the N and S proteins, respectively (Fig. 1c and d). Given the crucial roles of N and S proteins in virus transcription, assembly, and entry to host cells, the influence on infectivity based on whether SARS-CoV-2 isolates harbor different N and S variants (such as those clustered into different clades) remains unknown.

Purifying selection drives evolution at a whole structural gene level of SARS-CoV-2 during human to human transmission

Although numerous studies have demonstrated that recombination plays an important role on the emergence of SARS-CoV-2 [16-18], how this virus evolves during its global transmission has not yet been profiled. Therefore, we first analyzed intragenic recombination events of each structural gene using RDP4. Results indicate that no recombination events occurred among the alleles of each gene (data not shown). Potential recombination events were also assessed by reticulate network tree—phi test, using SplitsTree4 software. Although some internal nodes are present in N and S alleles, no clear evidence for recombination can be validated for each gene via phi test ($P > 0.05$) (Fig. 2). This result indicates a relatively stable state of SARS-CoV-2 during transmission, though a possible genetic interaction of different isolates may have occurred when it became a global pandemic [19, 20]. Tajima's D, and Fu and Li's D* and F* statistics were calculated to examine the mutation neutrality hypothesis of the four structural genes of SARS-CoV-2. Results reveal that the evolution of all four genes does not support population neutrality, but rather favors purifying selection (Table 2; Additional file 1: Figure S1). The average of all pairwise *dN/dS* ratios (ω) among alleles of each structural gene of SARS-CoV-2 is 0.5443 in E, 0.1562 in M, 0.07978 in N, and 0.4980 in S, respectively. In aggregate, these results suggest that at the whole gene level, inconsistent purifying selection is the main evolutionary force at work given the experimental conditions (Table 2; Additional file 1: Figure S1).

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

Gene	Tajima's D	Fu and Li's D* test	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

For Tajima's D, Fu and Li's D* and F* tests, the cutoff was set at 0.05.

The SARS-CoV2 S gene is operated on by positive selection at a definitive codon located at the C-terminal portion of the S1 subunit, and another potential codon located at the signal sequence

Guo et al. (2020) reported that the S gene of SARSr-CoV populations in their natural host, Chinese horseshoe bats, has evolved through positive selection at some codons[13]. As mentioned above, at the whole gene level, purifying selection is the main force driving the evolution of studied genes. Whether positive selection pressure accelerates the diversification of the structural genes of SARS-CoV-2 remains unclear. We therefore used codon-substitution models to estimate the ratio of nonsynonymous over synonymous substitutions (*dN/dS*), also known as "ω". These codon substitution models include M0 (one-ratio, one ω for all sites), M1a (nearly neutral, two classes of sites, ω0 < 1 and ω1 = 1), M2a (positive selection, allows three site classes including ω0 < 1, ω1 =1, and ω2 > 1), M3 (discrete, allows unconstrained discrete distribution of ω among sites), M7 (β, fit to a β distribution for ω among sites) and M8 (β and ω > 1, fit to a beta distribution with an extra rate that allows ω > 1) and can be typed into a null model group (M0, M1a, M7), and a positive selection model group (M3, M2a, and M8) [21]. The role of recombination in the polymorphisms of the four analyzed genes is excluded because no intragenic recombination was detected (Fig. 2). By using a Maximum Likelihood (ML) method, we did not find any codon of either the E or M gene subject to positive selection (data not shown). Although a potential positive selection site in the N gene, 208A, is identified using an M3 model, but the LRT *P*-value is more than 0.05 (M0 Vs. M3) and the result is not validated by other models including the M8 model, suggesting that evidence for N gene positive selection is limited (Additional file 2: Table S1). For the S gene, we found the average ω to be 0.37199, calculated using the one-ratio (M0) model of the codeML software package, suggesting that purifying selection operates S gene evolution during SARS-CoV-2 transmission among humans. In three LRTs, all alternative models (M3, M2a, and M8) are significantly better fits ($P < 10^{-4}$) than relevant null models (M0, M1a, and M7), indicating that some S sites were subjected to strong positive selection (ω = 18.22175–20.61283) (Table 3). A single positive selection site (614D) is identified in the S gene with a posterior probability of 1.000 in all three models [22], clear evidence for this site experiencing positive selection while the virus transmits between human hosts. This result is furthermore validated using internal fixed effects likelihood (IFEL) and evolutionary fingerprinting methodology, implemented using the HyPhy software package (Fig. 3) [23-25]. To our surprise, the positive selection site is not located at the receptor binding domain (RBD) or receptor binding motif (RBM), as we anticipated, which generally play the most integral role in virus-receptor interaction and virus entry into host cells [26]. This result suggests that relative genetic stability of this motif benefits virus survival. Interestingly, the site under positive selection pressure always has a D614G (for the S gene is 1841A>G) mutation, implying that such a mutation may enhance virus adaptability in human hosts. Another potential positive selection site at codon 5 was also identified, and a L5F mutation (for the S gene is 13C>T) was always found, with posterior probabilities greater than 0.95, 0.93 and 0.92 (critical values) calculated by M3, M2a and M8 models (Table 3), respectively. A similar result was also confirmed using an evolutionary fingerprinting method (Additional file 3: Figure S2).

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
M3 (discrete)	-6766.339162	p0=0.96797, p1=0.02883, p2=0.00320 ω0=00.26126, ω1= 2.70530, **ω2=20.61283**			**5 L 0.958***, 28 Y 0.850, 221 S 0.901, **614 D 1.000*****, 677 Q 0.891
M0 (one ratio)	-6790.072925	ω0=0.37199	M0 vs. M3	0.000000001	Not Allowed
M2a(selection)	-6766.432802	p0=0.81731, p1=0.17872, p2=0.00397 ω0=0.17504, ω1=1.00000, **ω2=18.76936**			5 L 0.9258, 28 Y 0.812, 221 S 0.832, **614 D 1.000*****, 677 Q 0.828
M1a (neutral)	-6778.770190	p0=0.70461, p1=0.29539 ω0=0.04395, ω1=1.00000	M1a vs. M2a	0.000004385	Not Allowed
M8(β & ω)	-6768.829411	p0=0.99578, p=0.40368, q=0.82224 p1= 0.00422, **ω= 18.22175**			5 L 0.931, 28 Y 0.817 ,221 S 0.831, **614 D 1.000*****, 677 Q 0.828
M7(β)	-6779.230494	p=0.00857, q=0.02623	M7 vs.M8	0.000030400	Not Allowed

LnL is log likelihood; ω is ratio of *dN*/*dS*, LRT *P*-value indicates the value of the 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) ≥ 0.80 are shown, (p) ≥ 0.95 (p) ≥ 0.99, and (p) = 1.000 are indicated by *, ** and ***, respectively.

The evolutionary relationship of S gene alleles with and without D614G and L5F mutations

Phylogenetic trees of S gene alleles were derived to test the evolutionary relationship among alleles with and without the D614G mutation. As shown in Fig. 4a, the 173 alleles of the S gene clustered into four clades. Alleles with a D614G mutation could be found in all the 4 clades. A dominant clade contains 79 out of total 85 alleles with this mutation. The remaining 6 mutated S alleles were distributed among the other 3 clades. This result is supported by a parsimony network of S gene alleles created using PopART software (http://popart.otago.ac.nz) [27]. Two central alleles (representative virus isolates are WH01 and GZMU0019) and associated alleles around them form a star-scattering network, suggesting that the S gene may have two potential origins (Fig. 4b). All S alleles with a D614G mutation are closely related with a few point mutations, and comprise a scattered star structure, suggesting the expansion of the SARS-CoV-2 population with a D614G mutation on the S gene. In contrast, alleles of the N gene suggest a single ancestor, as analyzed by parsimony networking, though 3 phylogenetic clades were identified (Additional file 4: Figure S3).

A total of 5 alleles with L5F mutations were discovered, and all of them were located within a single clade, accounting for 83.33% of all alleles in the clade (Additional file 5: Figure S4a). Further parsimony network analyses revealed that S alleles with a L5F mutation are not closely related, but rather are

distributed in both WH01 and GZMU0019 haplotype groups (Additional file 5: Figure S4b). No scattered star structure could be formed, indicating that the L5F mutation might arise from independent origins, unlike the D614G mutation.

Frequency of S alleles with D614G mutations increases in SARS-CoV-2 isolates during human to human virus transmission

Considering that mutations in a positive selection site should be beneficial to the survival of the individuals carrying the mutation, we postulate that the D614G (1841 A>G) mutation may increase the spread of SARS-CoV-2. Evidence is obtained from the haplotype network analysis aforementioned of S alleles (Fig. 4b). S gene haplotypes (alleles) with a 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 [28]. To further study whether SARS-CoV-2 isolates with a D614G mutation have a selective advantage in survival during transmission among humans we calculated the frequency of S alleles carrying a D614G mutation in each week from collected SARS-CoV-2 isolates from December 24, 2019 through April 20, 2020 (17 epidemiological weeks). Detailed information on these isolates including collection date, collection region, and accession or biosample number is available in Additional file 6: Table S2, and Additional file 7: Table S3.

In 173 S gene alleles, 85 carried a D614G mutation, accounting for 49.13% of the total. Similarly, 47 of 99 S proteins were found to carry a D614G mutation, accounting for 47.47% of the total. The first two isolates of our collected data, GWHABKF00000001 and WH01 (isolated in December 24, 2019 and December 26, 2019, respectively), carried 614D mutations in the S protein. The first SARS-CoV-2 isolate with a D614G mutation was isolated from a patient with COVID-19 on February 5, 2020 (week 7 in our dataset). After that, except for weeks 9 and 10 (possibly due to the small number of samples and sampling deviation), an increasing trend in the proportion of isolates carrying the D614G mutation in the S protein was noteworthy. In week 17, the last week of our dataset, 91.11% of SARS-CoV-2 isolates carried this mutation (Fig 5a; Additional file 6: Table S2). Further analyses revealed that the frequency of the D614G mutation in the S gene steadily increased when data from weeks 6 through 17 were combined (Fig 5b; Additional file 6: Table S2). To exclude the influence of sample size on the result (in some weeks, only 4–6 isolates were collected) we reorganized the dataset by taking both sample size and sampling time into account. Various panels of 200–300 isolates were studied, with similar results observed (Fig. 5c and d; Additional file 7: Table S3). Taken together, these results suggested that SARS-CoV-2 isolates with a D614G mutation might increase the virus's ability to transmit and hence rapidly spread around the world.

D614G mutations in the S gene may destabilize the S protein trimer and promote receptor binding and membrane fusion

We found that the D614G mutation is located at the subdomain 2 (SD2) at the C-terminus of RBD and close to the two potential cleavage sites between S1 and S2 [29] (Fig. 6a). Considering positive selection

is usually beneficial to the survival of the individual carrying the mutation, we speculate that the D614G mutation may facilitate structural conformation change to promote receptor binding or membrane fusion[8, 30], and in turn, improving infectivity. From the latest cryo-electron microscopy-determined (cryo-EM) structure of the SARS-CoV-2 S protein, the negatively charged sidechain of D614 points toward the positively charged sidechain of K854 from the neighboring monomer (Fig. 6b) [29] . The distance between the closest atoms of the two residues is 2.6 Å, which is an optimal distance to form a salt bridge (Fig. 6c). From the modelled structure with a D614G mutation, the distance is increased to 5.2 Å (Fig. 6d), which would potentially abolish the salt bridge and destabilize the integrity of the S trimer in wild type. It has been reported that human receptor ACE2 binds to an “open” conformation of the S protein, where RBD moves away from the core structure, and exposes its receptor binding surface. The entire S trimer then undergoes a series of dramatic conformation changes, including cleavages between S1 and S2, disassociation of S1, and post-fusion transformation of S2 [31, 32]. Mutations at cleavage sites and adding internal crosslinks to the S trimer would keep the protein in a stable and “closed” conformation, where the receptor binding surface of RBD is inaccessible [29, 33]. Therefore, we hypothesize that the highly transmissible D614G mutation, driven by positive evolutionary selection, promotes the accessibility of RBD by eradicating a critical salt bridge between S protein monomers, which subsequently triggers membrane fusion upon ACE2 binding.

Discussion

Many recent studies have demonstrated the continual evolution of SARS-CoV-2 [34-36]. Four structural genes of SARS-CoV-2, E, M, N and S, may determine the infectivity or pathogenesis of this persistent virus, but the molecular evolution patterns of these structural genes remain largely unknown. Among the four aforementioned structural genes, E and M show the highest homology, with relatively few SNPs and SAPs, indicating the importance of the conservation of these two genes for virus survival. A key factor for viral transcription and assembly is the N protein [37, 38]. High sequence variability was found in the N protein, suggesting a vast adaption of the virus during host transmission. Previous studies shows that elevated genetic variation has been found among bat SARSr-CoVs, particularly in the S gene[13]. High nucleotide diversity (π; Table 1) of the S gene was detected in SARS-CoV-2 isolates, suggesting that it may benefit virus survival in human hosts. Recombination has played an important evolutionary role during the emergence of SARS-CoV-2[39, 40]. It has been reported that SARS-CoV-2 is a recombinant virus of bat and pangolin (*Manis javanica*) CoVs, suggesting a critical role of recombination [39]. When this zoonotic virus transfers from animal to human and leads to continuous human to human transmission, no clear evidence of recombination is found among the alleles from the four structural genes in our study (Fig. 2), suggesting that the evolution of these genes is not driven by recombination. Li et al. studied the origin of SARS-CoV-2 and showed evidence of strong purifying selection in the S and other genes among bat, pangolin and human coronaviruses, indicating similarly strong evolutionary constraints in different host species[40]. Likewise, our results show that purifying selection drives evolution at the whole structural gene level of SARS-CoV-2 during its transmission between human hosts (Table 2; Additional file 1: Figure S1). This result implies that in general, genetic variation of these structural genes will not confer

a significant disadvantage to virus survival. Because no recombination events occurred during SARS-CoV-2 evolution, nonsynonymous mutations would be removed at a high rate during virus transmission [41], and positive selection sites with mutations will be fixed. The frequency of S alleles with a D614G mutation is increasing during the SARS-CoV-2 spread (Fig. 3; Fig. 5; Additional file 6: Table S2; Additional file 7: Table S3).

We identified another potentially positive selection site at codon 5 of the S gene, with a consistent L5F mutation (Table 3; Additional file 3: Figure S2). Considering that signal sequence (SS) is a short hydrophobic peptide that plays an important role in guiding viral proteins into endoplasmic reticulum (ER) for proper folding and assembly [9], we postulate that L5F mutations may increase hydrophobicity of the SS, thus facilitating the entry of the S protein into ER for folding and assembly, and in turn, secretion of the virus. Our results moreover show that majority of S alleles with a D614G mutation cluster in one clade, and make a distinct star-scattering network (Fig. 4), suggesting a potentially recent common ancestor of these mutants. Nevertheless, sporadic alleles with an L5F mutation identified thus far indicate that the L5F mutation might be subject to relatively weaker pressure, still at an early stage of positive selection (Additional file 5: Figure S4).

The positive selected D614G mutation may play an important role in the adaptability of SARS-CoV-2 in both the host and virus populations [42]. Another explanation is that the mutation is driven by specific interactions between high levels of viral sequence divergence and polymorphic host receptors or interacting proteins [43]. The S protein is the key determinant for tissue tropism, and host range and specificity of coronaviruses such as SARS-CoV-2. The virus infects host cells through the interaction between the S protein and its cellular receptor, ACE2 [11]. In this process, viral entry requires the precursor S protein cleaved by cellular proteases including trypsin, furin, transmembrane serine protease 2 (TMPRSS2), or endosomal cathepsin L, which generate receptor binding subunit S1 and membrane fusion S2 [30, 44, 45]. From structural studies of both SARS-CoV and SARS-CoV-2, the receptor binding domain (RBD) located at the C-terminal of S1 and the adjacent N-terminal domain (NTD) are relatively flexible, the feature required for receptor recognition and subsequent membrane fusion [29, 46]. From S protein structural modeling (Fig. 6), we hypothesize that the D614G mutation, driven by positive selection through molecular evolution, promotes accessibility of RBD through the loss of a critical salt bridge between S protein monomers (Fig. 6c and d), which subsequently triggers membrane fusion upon ACE2 binding. The exact influence and detailed mechanism of the D614G mutation on SARS-CoV-2 infectivity and expansion require further investigation. It should be noted that a L5F mutation is always found in a potentially positive selection site of the S gene (codon 5). The low frequency (3.82%, 5/131) of S alleles with an L5F mutation does not show a clearly increasing pattern, possibly due to relatively weaker positive selection pressure compared to codon 614. Because the end date of the studied isolates is late April, 2020, the persistent frequency monitoring of L5F in S alleles is required to determine whether it experiences expansion. Potential effects of the L5F mutation to SARS-CoV-2 need to be documented.

Conclusions

We present modern evolutionary virology analyses on a large and comparative set of SARS-CoV-2 structural gene sequences derived from an international collection of SARS-CoV-2 isolates. Distinct phylogenetic patterns of four structural proteins of SARS-CoV-2 are illustrated. Protein sequence comparisons show that E and M genes exhibit a relatively close genetic affinity to bat SARSr-CoV, suggesting the evolutionary conservation of these two genes. In contrast, relatively high genetic variation is observed in N and S proteins among SARS-CoV-2 isolates, implying extensive adaptability. No clear intragenic recombination is detected among the four genes, suggesting that recombination is not the primary evolutionary force driving the evolution of these genes. Our analysis, however, shows that purifying selection pressure may be the main force operating on a whole gene level of SARS-CoV-2 during interhuman transmission.

We identified a codon in the S gene that is definitively experiencing positive selection pressure, and always leads to a D614G mutation in S proteins. S alleles with a D614G mutation have expanded rapidly among SARS-CoV-2 isolates. The D614G mutation significantly extends the distance between monomers in the S protein trimer, which may disrupt the salt bridge formed by D614 and K854 between monomers. Such disruption may promote RBD opening and facilitate the entry of the virus into host cells, in turn contribute to the diffusion of these mutated alleles. Codon 5 of the S gene is another potential positive selection site. Although a limited number of alleles with an L5F mutation were identified, this mutation may potentially affect the assembly and secretion of SARS-CoV-2. A close examination of the L5F mutation may be required in case another expansion occurs. Because the S protein is a key target for SARS-CoV-2 vaccines, therapeutic antibodies, and diagnostics, the D614G and L5F mutations should be paid close attention to. Owning to the fact that the exact mechanism remains unclear, further research should focus on the function of these mutation sites, particularly how they affect the expansion of these mutated alleles in SARS-CoV-2.

Methods

SARS-CoV-2 isolates

Complete full-length genomic sequences of SARS-CoV-2 were downloaded from the 2019 Novel Coronavirus Resource (2019nCoVR), China National Center for Bioinformation (all of which were also uploaded to the NCBI GenBank database). Sequences were manually checked for the structural gene integrity and sequencing accuracy and finally a total of 3090 isolates were selected and verified for the present study. These isolates were collected from December 24, 2019 to April 24, 2020 in China, USA, Japan, Pakistan, Australia, Greece, Germany, Peru, Turkey, Kazakhstan, Iran, Serbia, Thailand, Netherlands, Sri Lanka, Czech Republic, Malaysia, India, et al. Detailed information on these isolates, including GenBank accession number or biosample number is summarized in Additional file 8: Table S4.

Sequence analysis of four structural genes and proteins

The E, M, N, and S gene sequences were extracted from the SARS-CoV-2 global isolate collection and aligned using the MEGA X software package using Muscle (codons) parameters [47]. Because some

regions of genomic sequences of SARS-CoV-2 couldn't be exactly identified (in which nucleic acid bases are shown as degenerate bases [e.g. N, R, Y]), we were sometimes unable to obtain all of the four structural gene sequences from an isolate. Allele type and DNA sequence polymorphism analyses were performed using software DnaSP 6.12.03[48]. Protein sequences and polymorphism loci of these isolates were also aligned and analyzed using the MEGA X software package [47].

Molecular evolution analysis

An unrooted phylogenetic tree of the four structural proteins was constructed using the MEGA X software package [47], with evolutionary history inferred using the Maximum Likelihood method, based on the JTT matrix-based model for E protein sequences, General Reversible Chloroplast + Freq. model for M protein sequences, JTT matrix-based model for N protein sequences, and the Jones et al. w/freq. model for S protein sequences. Model selection was conducted in MEGA X [43]. Bootstrap values were estimated using 1000 replications. Initial trees for the heuristic search were obtained automatically by applying Neighbor-Joining and BioNJ algorithms to a matrix of pairwise distances estimated using each model mentioned above. Trees were drawn to scale, and FigTree V1.4.4 was utilized to form cladogram branches [49]. The aligned DNA sequences were also screened using RDP4 software to detect intragenic recombination among the alleles of each structural gene [50]. Six methods implemented in RDP4 were utilized. These methods include RDP [50], GENECONV[51], BootScan [52], MaxChi [53], Chimaera [54], and SiScan [55]. Common settings for all methods include considering sequences as linear and setting the statistical significance cutoff was at 0.05, with Bonferroni correction for multiple comparisons. Phylogenetic evidence and polishing of breakpoints were also required. Potential recombination events (PREs) were classified as those identified by at least two methods. A reticulate network tree of the alleles of the four structural genes of SARS-CoV-2 was also generated using Splitstree4 [56]. Phi tests implemented in Splitstree4 were used to define probable recombination events. Tajima's D, Fu and Li's D* and F* tests were employed to test the mutation neutrality hypothesis of a whole gene as previously described by our research group[57]. These analyses were carried out using DnaSP 6.12.03[48]. A statistical significance cutoff was set at 0.05 for Tajima's D test, Fu and Li's D* and F* tests. The false discovery rate method and 1000 replications in a coalescent simulation were applied for correcting multiple comparisons. Non-neutrality evolution was determined when identified by at least two of three tests. Nonsynonymous and synonymous mutations of the alleles of the four structural genes were calculated using the MEGA X software package [47].

Analysis of positive selection at codon level

The selection pressure operating on the four structural genes of SARS-CoV-2 was investigated using the Maximum Likelihood (ML) method. Analyses were performed using a visual tool from the codeml software program, named EasyCodeML [58]. Three nested models (M3 vs. M0, M2a vs. M1a, and M8 vs. M7) were compared and likelihood ratio tests (LRTs) were applied to assess a best fit of codes. Model fitting was performed using multiple seed values for *dN/dS*, assuming the F3x4 model of codon frequencies. Positive selection was inferred when the individual site or codon had a ratio of

nonsynonymous to synonymous mutations (*dN/dS* ratios) of greater than one (ω > 1). When the LRT was significant (P < *0.05*), Bayes empirical Bayes (BEB) (M8 model) and Naive Empirical Bayes (NEB) methods (M3 and M2a models) were further employed to identify amino acid residues that likely evolved under positive selection based on a posterior probability threshold of 0.95. Results from the M8 model were taken as the standard, in accordance with Yang et al. [17]. An M3 model was used for the frequency distribution of codon class analysis also as recommended by Yang et al. [22]. The HyPhy software package was used to validate the results obtained using the ML method [59].

Structural modeling of proteins with positive selection sites

Three-dimensional structures of proteins with positive selection sites were modeled using SWISS-MODEL [60]. according to the best-fit protein template. Model quality was evaluated by QMEAN, while the structure of the model was visualized using the PyMoL software package [61].

Abbreviations

SARS-CoV-2: Severe acute respiratory syndrome coronavirus 2; SARSr-CoV: SARS-related coronavirus; ORFs: open reading frames; E: Envelope; M: Membrane; N: Nucleocapsid; S: Spike; SNP: single nucleotide polymorphisms; SAP: amino acid polymorphic; ACE2: angiotensin converting enzyme 2; TMPRSS2: transmembrane serine protease 2; RBD: receptor binding domain; NTD: N-terminal domain; RBM: receptor binding motif; SD1: subdomain 1; SD2: subdomain 2; CTP :C-terminal portion; cryo-EM: cryo-electron microscopy; ER: endoplasmic reticulum; PREs: Potential recombination events.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable

Availability of data and materials All data generated or analyzed during this study are included in this published article and its additional files.

Competing interests The authors have declared no conflict of interests.

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

Availability of data and materials All data generated during this study are included in this published article and its Additional files 1 and 2.

Authors' contributions XYZ designed, carried out, and analyzed the data and wrote the manuscript. YZ designed, carried out, and analyzed the structure of SARS-CoV-2 S protein. KH, XZ, YQ, YL and LeY collect

the genomic data of the isolates. YH and BH supervised and assisted in research planning and also supervised the manuscript. All authors read and approved final manuscript.

Acknowledgements Not applicable

References

1. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N *et al*: **Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding**. *Lancet* 2020, **395**(10224):565-574.
2. Khailany RA, Safdar M, Ozaslan M: **Genomic characterization of a novel SARS-CoV-2**. *Gene Rep* 2020:100682.
3. Liu DX, Yuan Q, Liao Y: **Coronavirus envelope protein: a small membrane protein with multiple functions**. *Cellular and molecular life sciences : CMLS* 2007, **64**(16):2043-2048.
4. Arndt AL, Larson BJ, Hogue BG: **A conserved domain in the coronavirus membrane protein tail is important for virus assembly**. *Journal of virology* 2010, **84**(21):11418-11428.
5. McBride R, van Zyl M, Fielding BC: **The coronavirus nucleocapsid is a multifunctional protein**. *Viruses* 2014, **6**(8):2991-3018.
6. Shang J, Wan Y, Luo C, Ye G, Geng Q, Auerbach A, Li F: **Cell entry mechanisms of SARS-CoV-2**. *Proceedings of the National Academy of Sciences of the United States of America* 2020, **117**(21):11727-11734.
7. Jimenez-Guardeno JM, Nieto-Torres JL, DeDiego ML, Regla-Nava JA, Fernandez-Delgado R, Castano-Rodriguez C, Enjuanes L: **The PDZ-binding motif of severe acute respiratory syndrome coronavirus envelope protein is a determinant of viral pathogenesis**. *PLoS pathogens* 2014, **10**(8):e1004320.
8. Belouzard S, Millet JK, Licitra BN, Whittaker GR: **Mechanisms of coronavirus cell entry mediated by the viral spike protein**. *Viruses* 2012, **4**(6):1011-1033.
9. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D: **Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein**. *Cell* 2020, **181**(2):281-292 e286.
10. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY *et al*: **A new coronavirus associated with human respiratory disease in China**. *Nature* 2020, **579**(7798):265-269.
11. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL *et al*: **A pneumonia outbreak associated with a new coronavirus of probable bat origin**. *Nature* 2020, **579**(7798):270-273.
12. 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**:152-158.
13. Guo H, Hu B-J, Yang X-L, Zeng L-P, Li B, Ouyang S-Y, Shi Z-L: **Evolutionary arms race between virus and host drives genetic diversity in bat SARS related coronavirus spike genes**. 2020:2020.2005.2013.093658.

14. Narayanan K, Makino S: **Cooperation of an RNA packaging signal and a viral envelope protein in coronavirus RNA packaging**. *Journal of virology* 2001, **75**(19):9059-9067.
15. Letko M, Marzi A, Munster V: **Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses**. *Nat Microbiol* 2020, **5**(4):562-569.
16. Wong MC, Javornik Cregeen SJ, Ajami NJ, Petrosino JF: **Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019**. 2020:2020.2002.2007.939207.
17. Wu Y: **Strong evolutionary convergence of receptor-binding protein spike between COVID-19 and SARS-related coronaviruses**. 2020:2020.2003.2004.975995.
18. Wu A, Niu P, Wang L, Zhou H, Zhao X, Wang W, Wang J, Ji C, Ding X, Wang X *et al*: **Mutations, Recombination and Insertion in the Evolution of 2019-nCoV**. 2020:2020.2002.2029.971101.
19. **Iceland patient infected by two strains**. *The Standard* 2020, https://www.thestandard.com.hk/section-news/section/11/217711/Iceland-patient–infected-by–two-strains.
20. Mallapaty S: **How sewage could reveal true scale of coronavirus outbreak**. *Nature* 2020, **580**(7802):176-177.
21. Yang Z, Nielsen R, Goldman N, Pedersen AM: **Codon-substitution models for heterogeneous selection pressure at amino acid sites**. *Genetics* 2000, **155**(1):431-449.
22. Yang Z, Wong WS, Nielsen R: **Bayes empirical bayes inference of amino acid sites under positive selection**. *Molecular biology and evolution* 2005, **22**(4):1107-1118.
23. Pond SLK, Muse SV: **HyPhy: Hypothesis Testing Using Phylogenies**: Springer New York; 2005.
24. Pond SL, Scheffler K, Gravenor MB, Poon AF, Frost SD: **Evolutionary fingerprinting of genes**. *Molecular biology and evolution* 2010, **27**(3):520-536.
25. Kosakovsky Pond SL, Frost SD: **Not so different after all: a comparison of methods for detecting amino acid sites under selection**. *Molecular biology and evolution* 2005, **22**(5):1208-1222.
26. Wan Y, Shang J, Graham R, Baric RS, Li F: **Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus**. *Journal of virology* 2020, **94**(7).
27. Clement M, Snell Q, Walker P, Posada D, Crandall KJP, Distributed Processing Symposium IP: **TCS: Estimating gene genealogies**. 2002, **2**:184.
28. Bubac CM, Spellman GMJTAOA: **How connectivity shapes genetic structure during range expansion: Insights from the Virginia's Warbler**. 2016(2):2.
29. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, Graham BS, McLellan JS: **Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation**. *Science* 2020, **367**(6483):1260-1263.
30. Lu G, Wang Q, Gao GF: **Bat-to-human: spike features determining 'host jump' of coronaviruses SARS-CoV, MERS-CoV, and beyond**. *Trends in microbiology* 2015, **23**(8):468-478.

31. Walls AC, Xiong X, Park YJ, Tortorici MA, Snijder J, Quispe J, Cameroni E, Gopal R, Dai M, Lanzavecchia A *et al*: **Unexpected Receptor Functional Mimicry Elucidates Activation of Coronavirus Fusion**. *Cell* 2019, **176**(5):1026-1039 e1015.
32. Walls AC, Tortorici MA, Snijder J, Xiong X, Bosch BJ, Rey FA, Veesler D: **Tectonic conformational changes of a coronavirus spike glycoprotein promote membrane fusion**. *Proceedings of the National Academy of Sciences of the United States of America* 2017, **114**(42):11157-11162.
33. Xiong X, Qu K, Ciazynska KA, Hosmillo M, Carter AP, Ebrahimi S, Ke Z, Scheres SHW, Bergamaschi L, Grice GL *et al*: **A thermostable, closed, SARS-CoV-2 spike protein trimer**. 2020:2020.2006.2015.152835.
34. Xiaolu T, Changcheng W, Xiang L, Yuhe S, Xinmin Y, Xinkai W, Yuange D, Hong Z, Yirong W, Review QZJNS: **On the origin and continuing evolution of SARS-CoV-2**. 2020.
35. Phan T: **Genetic diversity and evolution of SARS-CoV-2**. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* 2020, **81**:104260.
36. Kasibhatla SM, Kinikar M, Limaye S, Kale MM, Kulkarni-Kale U: **Understanding evolution of SARS-CoV-2: A perspective from analysis of genetic diversity of RdRp gene**. *Journal of medical virology* 2020.
37. Voss D, Kern A, Traggiai E, Eickmann M, Stadler K, Lanzavecchia A, Becker S: **Characterization of severe acute respiratory syndrome coronavirus membrane protein**. *FEBS letters* 2006, **580**(3):968-973.
38. Tseng YT, Wang SM, Huang KJ, Lee AI, Chiang CC, Wang CT: **Self-assembly of severe acute respiratory syndrome coronavirus membrane protein**. *The Journal of biological chemistry* 2010, **285**(17):12862-12872.
39. Huang J-M, Jan SS, Wei X, Wan Y, Ouyang S: **Evidence of the Recombinant Origin and Ongoing Mutations in Severe Acute Respiratory Syndrome 2 (SARS-COV-2)**. 2020:2020.2003.2016.993816.
40. Li X, Giorgi EE, Marichannegowda MH, Foley B, Xiao C, Kong X-P, Chen Y, Gnanakaran S, Korber B, Gao F: **Emergence of SARS-CoV-2 through recombination and strong purifying selection**. 2020:eabb9153.
41. Hughes AL, Hughes MA: **More effective purifying selection on RNA viruses than in DNA viruses**. *Gene* 2007, **404**(1-2):117-125.
42. Duxbury EM, Day JP, Maria Vespasiani D, Thuringer Y, Tolosana I, Smith SC, Tagliaferri L, Kamacioglu A, Lindsley I, Love L *et al*: **Host-pathogen coevolution increases genetic variation in susceptibility to infection**. *eLife* 2019, **8**.
43. Meyerson NR, Sawyer SL: **Two-stepping through time: mammals and viruses**. *Trends in microbiology* 2011, **19**(6):286-294.
44. Bestle D, Heindl MR, Limburg H, van TVL, Pilgram O, Moulton H, Stein DA, Hardes K, Eickmann M, Dolnik O *et al*: **TMPRSS2 and furin are both essential for proteolytic activation and spread of SARS-CoV-2 in human airway epithelial cells and provide promising drug targets**. 2020:2020.2004.2015.042085.

45. Ou X, Liu Y, Lei X, Li P, Mi D, Ren L, Guo L, Guo R, Chen T, Hu J *et al*: **Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV**. *Nature communications* 2020, **11**(1):1620.
46. Gui M, Song W, Zhou H, Xu J, Chen S, Xiang Y, Wang X: **Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding**. *Cell research* 2017, **27**(1):119-129.
47. 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):1547-1549.
48. Rozas J, Ferrer-Mata A, SÃ nchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sã n-GAJMB, Evolution: **DnaSP 6: DNA Sequence Polymorphism Analysis of Large Datasets**. 2017, **34**(12).
49. Rambaut A: **FigTree v1.4.4. Institute of Evolutionary Biology, University of Edinburgh, Edinburgh. http://tree.bio.ed.ac.uk/software/figtree/**. 2018.
50. Martin DP, Murrell B, Khoosal A, Muhire B: **Detecting and Analyzing Genetic Recombination Using RDP4**. *Methods in molecular biology* 2017, **1525**:433-460.
51. Padidam M, Sawyer S, Fauquet CM: **Possible emergence of new geminiviruses by frequent recombination**. *Virology* 1999, **265**(2):218-225.
52. Martin DP, Posada D, Crandall KA, Williamson C: **A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints**. *AIDS Res Hum Retroviruses* 2005, **21**(1):98-102.
53. Smith JM: **Analyzing the mosaic structure of genes**. *Journal of molecular evolution* 1992, **34**(2):126-129.
54. Posada D: **Evaluation of methods for detecting recombination from DNA sequences: empirical data**. *Molecular biology and evolution* 2002, **19**(5):708-717.
55. Gibbs MJ, Armstrong JS, Gibbs AJ: **Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences**. *Bioinformatics* 2000, **16**(7):573-582.
56. Huson DH, Bryant D: **Application of phylogenetic networks in evolutionary studies**. *Molecular biology and evolution* 2006, **23**(2):254-267.
57. Zhan XY, Zhu QY: **Molecular evolution of virulence genes and non-virulence genes in clinical, natural and artificial environmental Legionella pneumophila isolates**. *PeerJ* 2017, **5**:e4114.
58. Gao F, Chen C, Arab DA, Du Z, He Y, Ho SYWJE, Evolution: **EasyCodeML: A visual tool for analysis of selection using CodeML**. 2019.
59. Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, Shank SD, Magalis BR, Bouvier D, Nekrutenko A *et al*: **HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies**. *Molecular biology and evolution* 2019, **37**(1):295-299.
60. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L *et al*: **SWISS-MODEL: homology modelling of protein structures and complexes**. *Nucleic*

acids research 2018, **46**(W1):W296-W303.
/ol>

Figures

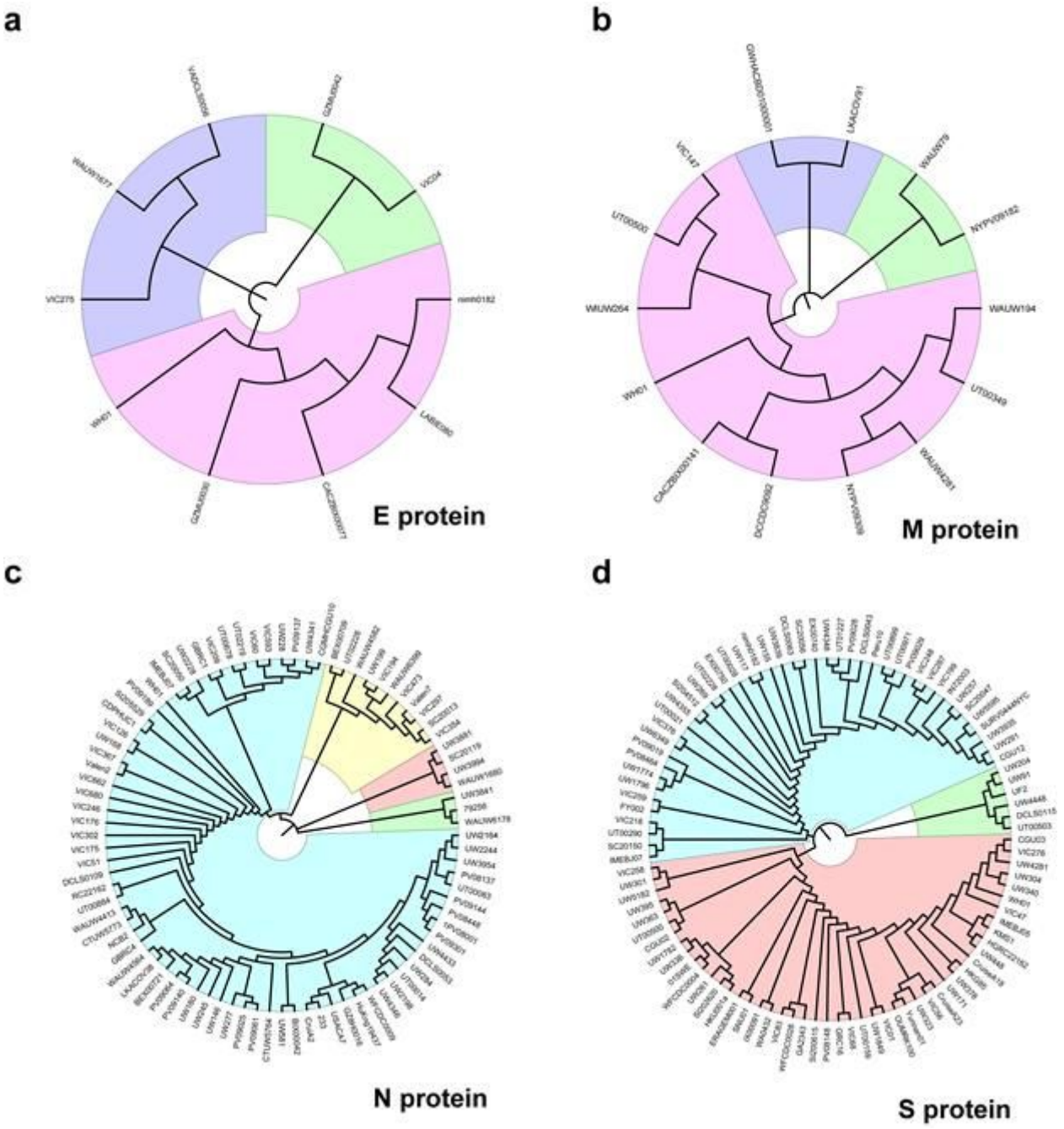

Figure 1

Phylogenetic tree of E (a), M (b), N (c), and S (d) proteins of SARS-CoV-2. Major clades are highlighted with different color. The tree shows topology of the protein of each allele, named by their representative isolates.

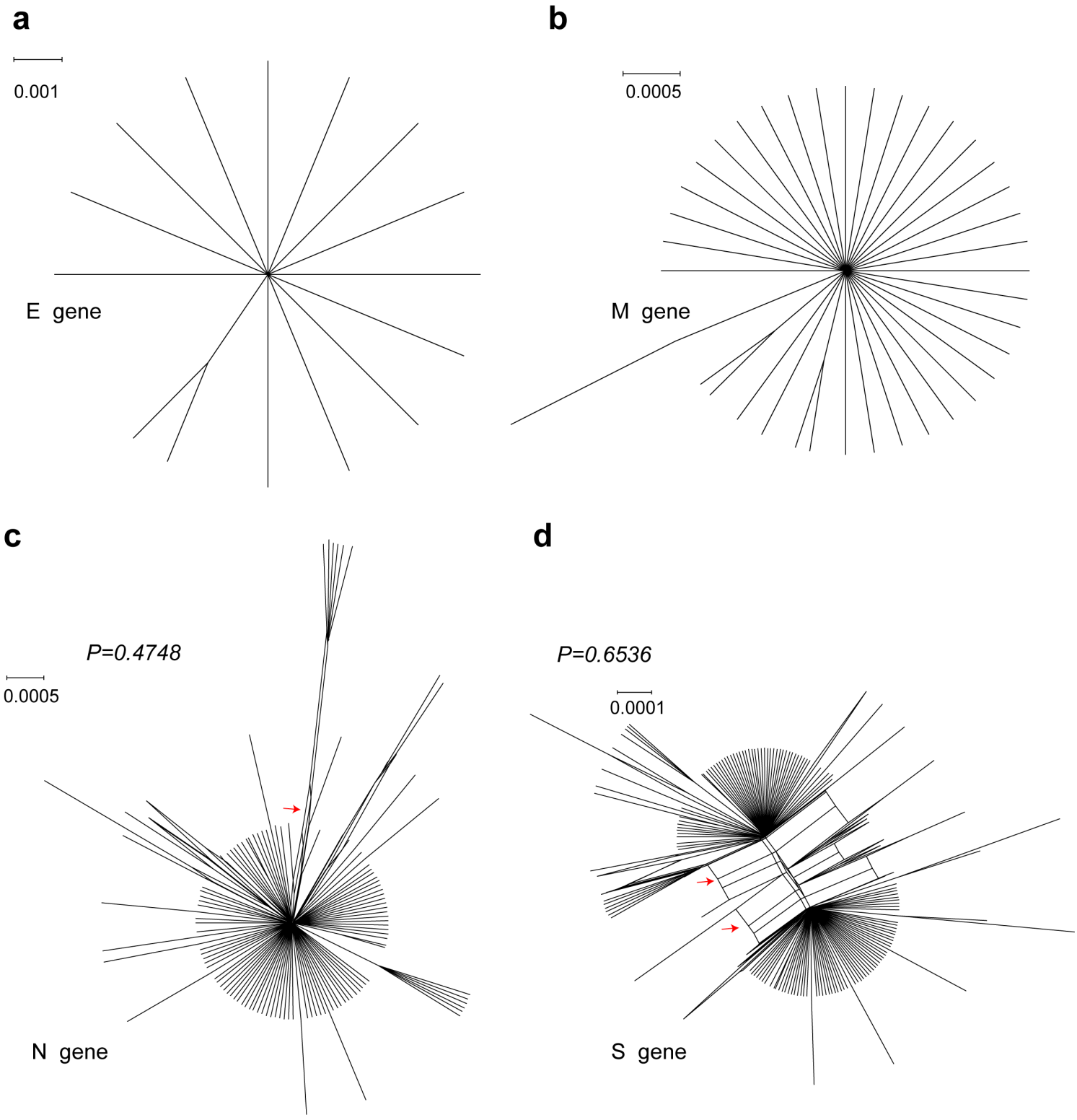

Figure 2

Reticulate network trees of (a) E, (b) M, (c) N, and (d) S alleles of SARS-CoV-2. Scale bars indicate number of substitutions per site. All internal nodes represent hypothetical ancestral alleles and edges that correspond to reticulate events, such as recombination. Red arrows indicate edges. Because there were too few informative characters to use the phi test for E and M genes, the P-values of the phi tests of N and S genes are shown.

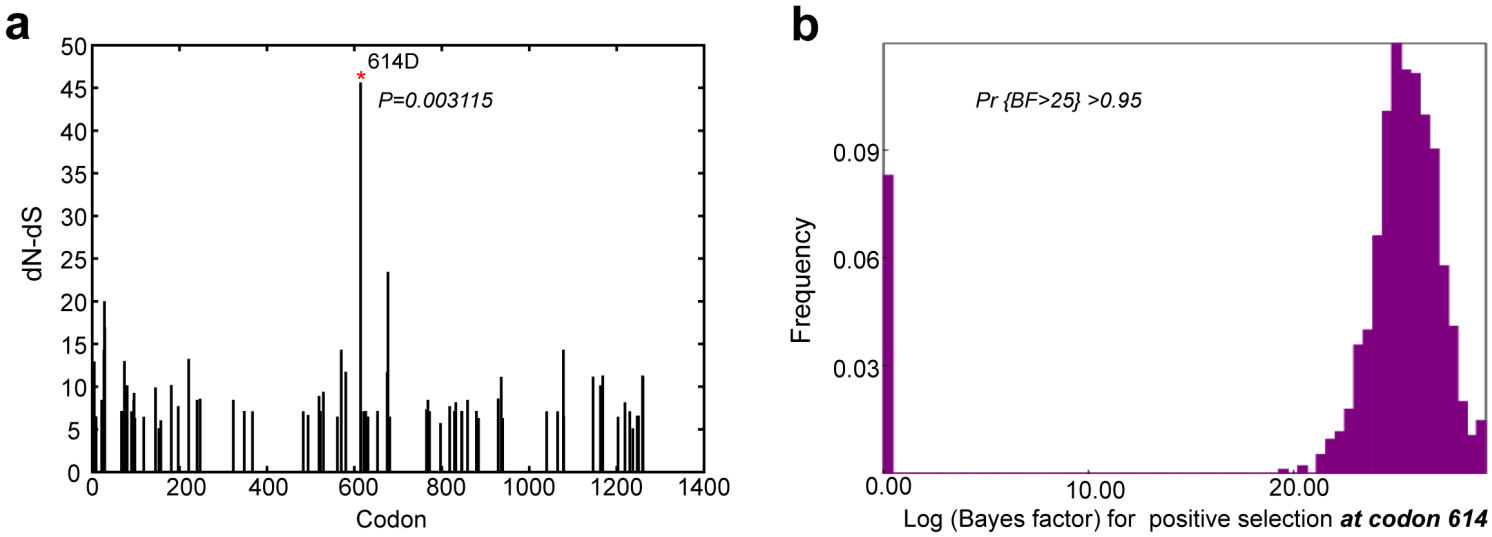

Figure 3

Positive selection analysis of S gene codons by IFEL and evolutionary fingerprinting methods. (a) Diagram of selection analysis results of S codons using the IFEL method. Asterisk (*) indicates a positive selection site with statistical significance (P < 0.01). (b) Log (Bayes factor [BF]) for positive selection at codon 614 of the S gene, and its frequencies. The cutoff value for the Bayes factor in the evolutionary fingerprinting method was set to 25 to reflect a positive selection at a given site (Posterior probability > 0.95). Pr {BF > 25} indicates a posterior probability of Bayes factor > 25.

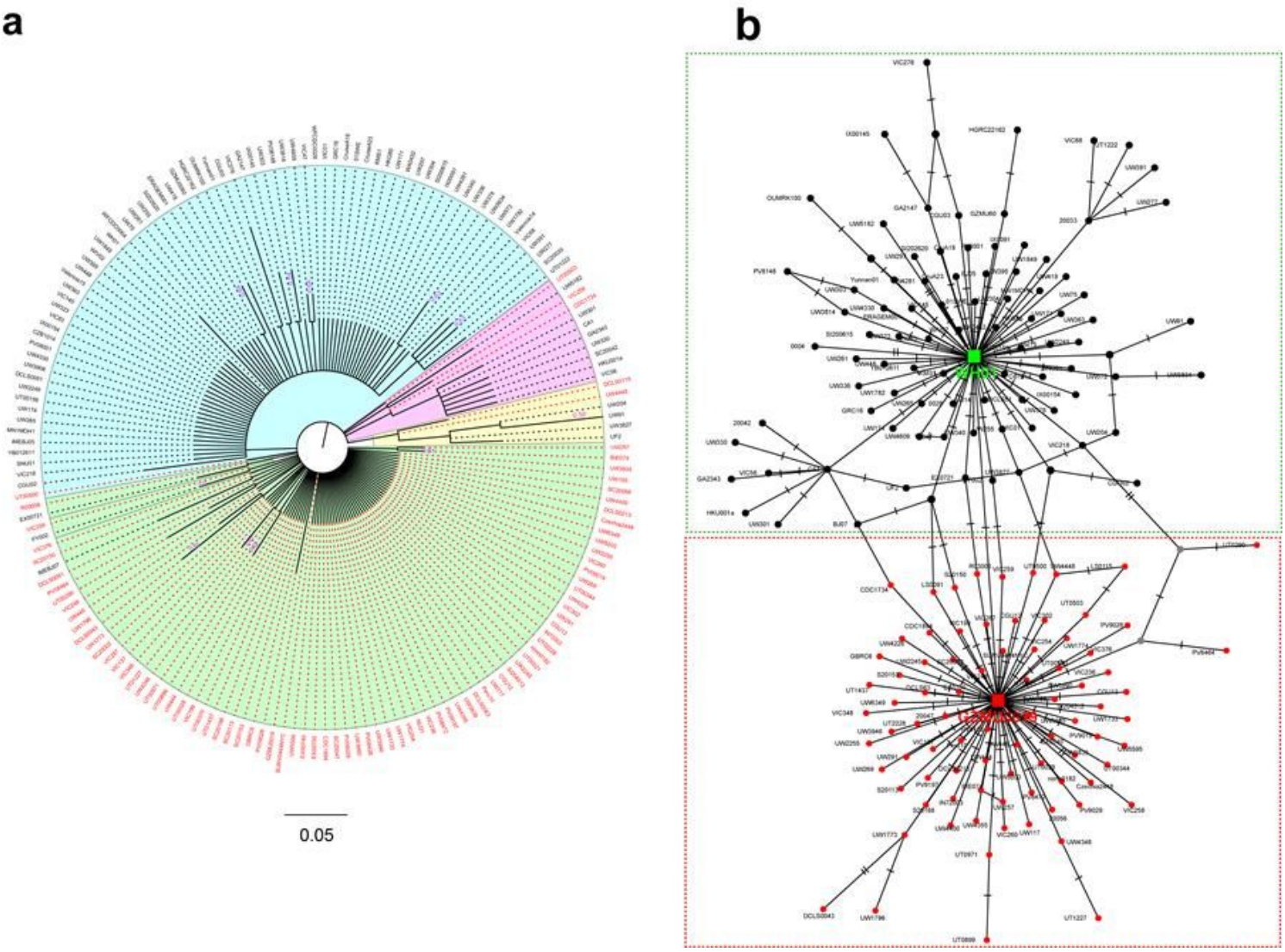

Figure 4

Evolutionary relationship of S alleles with or without D614G mutation. a Phylogenetic tree of S gene based on nucleotide sequences of 173 alleles. The evolutionary history is inferred using the Maximum Likelihood method and Tamura-Nei model. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Each clade is highlighted with different color. Alleles are shown with their representative isolate names, and alleles with D614G mutation are highlighted in red. Bootstrap values more than 0.5 are shown. b Parsimony network of SARS-CoV-2 S gene haplotype (allele) diversity obtained from 3090 isolates worldwide. Each oblique line linking between haplotypes (haplotype name is shown as its representative isolate name) represents one mutational difference. Unlabeled nodes (Gray circle) indicate inferred steps have not found in the sampled populations yet. The ancestral haplotype, or root of the network, is labeled with a square, and represent haplotype name is marked green or red. The red nodes indicate haplotypes with D614G mutation, while green or black nodes indicate haplotypes without D614G mutation. Dotted boxes indicate major haplotype groups.

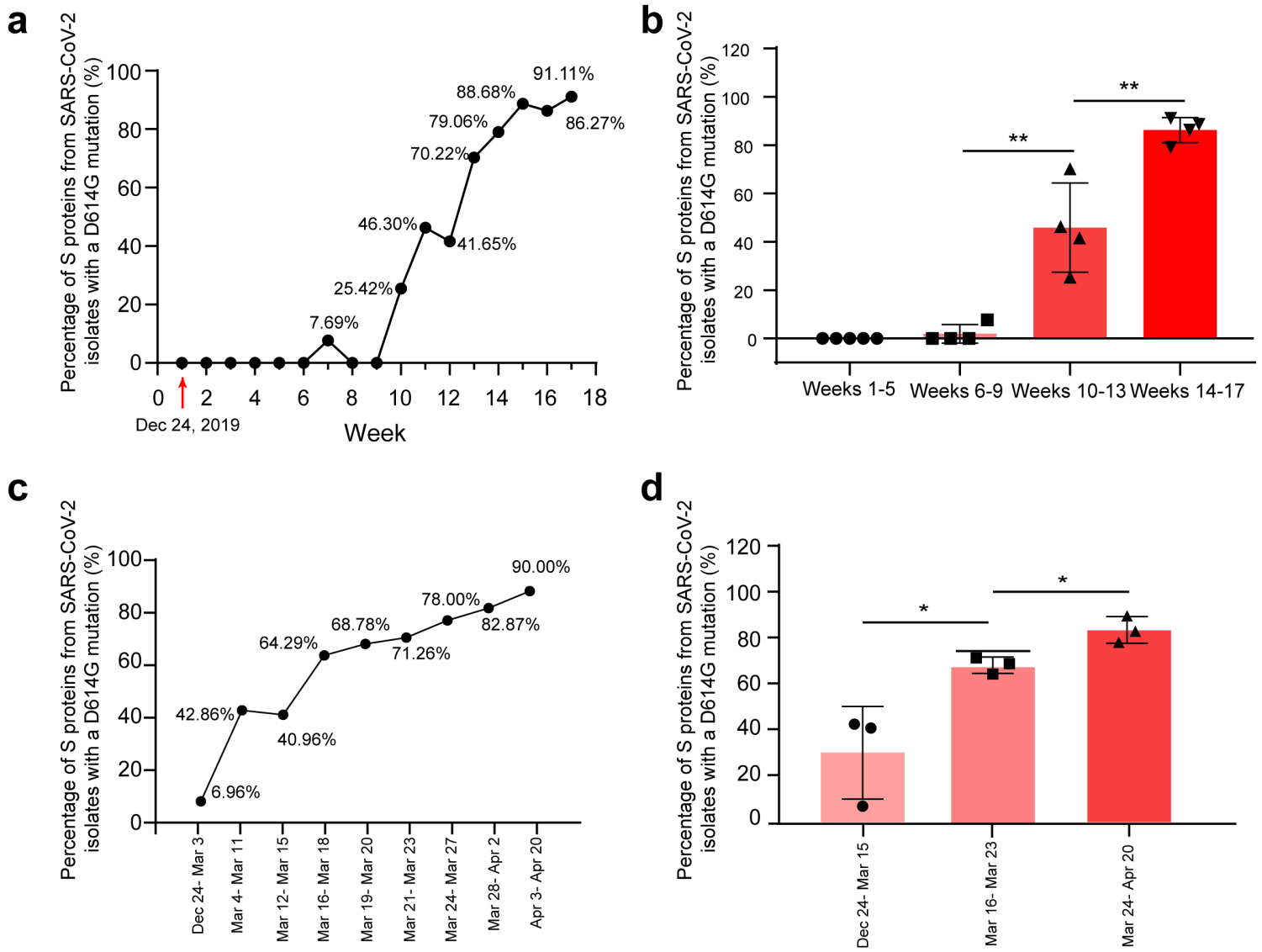

Figure 5

Expansion of S alleles with D614G mutations during SARS-CoV-2 human to human transmission. (a) Percentage of SARS-CoV-2 isolates carrying the alleles of a D614G mutation in each week collected. (b) Frequencies of D614G mutations in the S gene in each period of time (four to five weeks' data are combined). (c) Percentage of SARS-CoV-2 isolates carrying the alleles of a D614G mutation in each period of time. (d) Frequencies of D614G mutations in the S gene in each period of time. *P < 0.05; **P < 0.01.

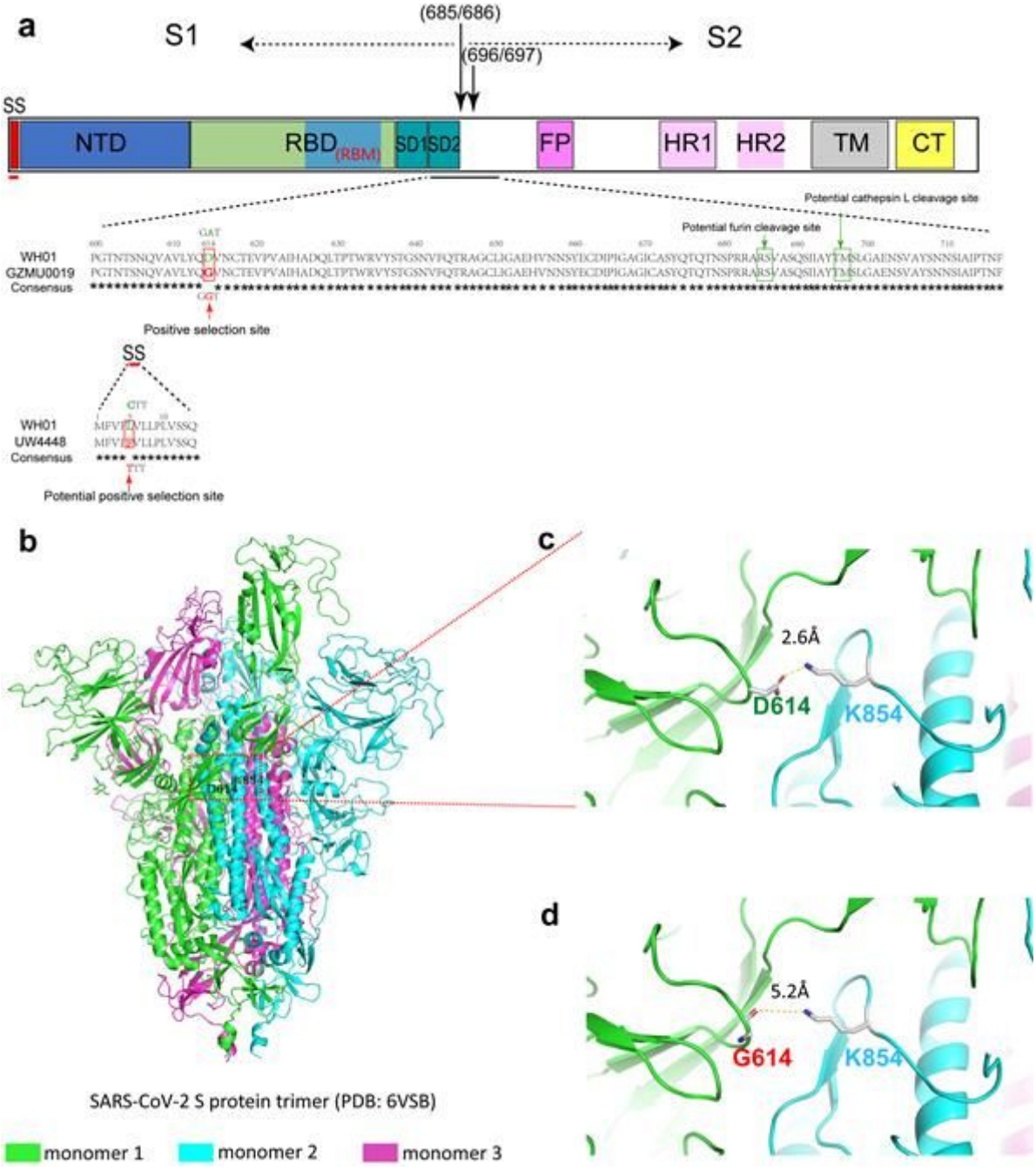

Figure 6

Structure of the S protein and the potential influence of D614G mutations on its structural change. (a) Schematic of the primary structure of the SARS-CoV-2 S protein colored by domains. Some boundary-residues are listed. The S1/S2 cleavage sites are indicated by arrows, SS: signal sequence; RBD: receptor binding domain; RBM: receptor of binding motif; FP: fusion peptide, HR1/2: heptad repeat 1/2; TM: transmembrane domain; CT: cytoplasmic tail; NTD: N-terminal domain; SD1: subdomain 1; SD2: subdomain 2. The structure of the S protein trimer of SARS-CoV-2 and potential influence of D614G mutation on its structural change. (b) Experimentally determined structure of SARS-CoV-2 S protein trimer (PDB ID is 6VSB and the amino acid sequences is the same as the WH01 isolate). (c) D614-K854 inter-

monomer salt bridge. (d) G614-K854 inter-monomer salt bridge. The distance of the salt bridge is increased from 2.6 to 5.2 Å in D614G mutations, as shown.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile8TableS4.xlsx
- Additionalfile7TableS3.xlsx
- Additionalfile6TableS2.xlsx
- Additionalfile5FigureS4.tif
- Additionalfile4FigureS3.tif
- Additionalfile3FigureS2.tif
- Additionalfile2TableS1.docx
- Additionalfile1FigureS1.tif