Full Title:

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

Authors: Xiao-Yong Zhan[1]**,** Ying Zhang[1], Xuefu Zhou[1]**,** Ke Huang[1], Yichao Qian[1], Yang Leng[1], Leping Yan[1], Bihui Huang[1]*, Yulong He[1*]

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

Correspondence authors:

***** Bihui Huang

ORCID: 0000-0003-3084-7096

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

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

*Yulong He

ORCID: 0000-0001-8930-8704

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

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

Abstract

SARS-CoV-2 caused a global pandemic in early 2020 and has resulted in more than 8,000,000 infections as well as 430,000 deaths in the world so far. Four structural proteins, envelope (E), membrane (M), nucleocapsid (N) and spike (S) glycoprotein, play a key role in controlling the entry into human cells and virion assembly of SARS-CoV-2. However, how these genes evolve during its human to human transmission is largely unknown. In this study, we screened and analyzed roughly 3090 SARS-CoV-2 isolates from GenBank database. The distribution of the four gene alleles is determined:16 for E, 40 for M, 131 for N and 173 for S genes. Phylogenetic analysis shows that global SARS-CoV-2 isolates can be clustered into three to four major clades based on the protein sequences of these genes. Intragenic recombination event isn't detected among different alleles. However, purifying selection has conducted on the evolution of these genes. By analyzing full genomic sequences of these alleles using codon-substitution models (M8, M3 and M2a) and likelihood ratio tests (LRTs) of codeML package, it reveals that codon 614 of S glycoprotein has subjected to strong positive selection pressure and a persistent D614G mutation is identified. The definitive positive selection of D614G mutation is further confirmed by internal fixed effects likelihood (IFEL) and Evolutionary Fingerprinting methods implemented in Hyphy package. In addition, another potential positive selection site at codon 5 in the signal sequence of the S protein is also identified. The allele containing D614G mutation has undergone significant expansion during SARS-CoV-2 global pandemic, implying a better adaptability of isolates with the mutation. However, L5F allele expansion is relatively restricted. The D614G mutation is located at the subdomain 2 (SD2) of C-terminal portion (CTP) of the S1 subunit. Protein structural modeling shows that the D614G mutation may cause the disruption of salt bridge among S protein monomers increase their flexibility, and in turn promote receptor binding domain (RBD) opening, virus attachment and entry into host cells. Located at the signal sequence of S protein as it is, L5F mutation may facilitate the protein folding, assembly, and secretion of the virus. This is the first evidence of positive Darwinian selection in the *spike* gene of SARS-CoV-2, which contributes to a better understanding of the adaptive mechanism of this virus and help to provide insights for developing novel therapeutic approaches as well as effective vaccines by targeting on mutation sites.

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of an emerging coronavirus disease (COVID-19) that has caused more than430,000 deaths, is still a serious global pandemic currently. The genome of SARS-CoV-2 is consisting of a single-stranded and positive-sense RNA of around 30 kb in length with a 5' cap and 3'-polyA tail. It shows that SARS-CoV-2 genome possesses six major open reading frames (ORFs) that encodes 27 different proteins, in which four are structural proteins named Envelope (E), Membrane (M), Nucleocapsid (N) and Spike (S). Many studies have demonstrated important functions of these proteins in virus entry, transcription and virion particle assembly of SARS-CoV-2. The E protein is a small envelope protein with 75 amino acids. Given that a close genetic relationship between SARS-CoV-2 and SARS-CoV, functions of this protein may include virion assembly and morphogenesis[1]. In addition, induction of apoptosis of host cells might be another crucial function of SARS-CoV-2 E protein, thus making it a potential determinant of viral pathogenesis [2]. M protein, consisting of 222 amino acids, is the most abundant component of the viral envelope and plays a key role in the virion assembly[3]. 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[4]. S protein, consisting of 1,273 amino acids, is the most important factor that mediates virus entry and a primary determinant of cell tropism and pathogenesis of SARS-CoV-2[5].

Many studies demonstrated SARS-CoV-2 underwent the evolution and some genetic evolutionary features have been reported[6]. The whole genomic sequence of SARS-CoV-2 has 79.6% identity with SARS-CoV and 96% with a bat SARS-related coronavirus (SARSr-CoV), RaTG13.

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

Materials and Methods

SARS-CoV-2 isolates

Complete full-length genomic sequences of SARS-CoV-2 were downloaded from 2019 Novel Coronavirus Resource (2019nCoVR) in China National Center for Bioinformation. All of which were also uploaded to the NCBI GenBank database. The sequences were manually checked 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 the different geographical locations including China, USA, Japan, Pakistan, Australia, Greece, German, Peru, Turkey, Kazakhstan, Iran, Serbia, Thailand, Nederland, Sri Lanka, Czech, Malaysia, India etc. Detailed information of these isolates including the GenBank accession number or biosample number is summarized in S1 Table.

Sequence analysis of the four structural genes and proteins

The E, M, N, S gene sequences were extracted from SARS-CoV-2 global isolate collection and aligned by the MEGA X package using Muscle (codons) parameters [10]. 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 unable to obtain all of the four structural gene sequences from an isolate sometimes. Allele type and DNA sequence polymorphism analyses were performed using DnaSP 6.12.03[11]. The protein sequences and polymorphism loci of these isolates were also aligned and analyzed with the MEGA X.

Molecular evolution analysis

An unrooted phylogenetic tree of the four structural proteins was constructed using the MEGA X package [10], and the evolutionary history was inferred using the Maximum Likelihood method, based on the JTT matrix-based model for E protein sequences, General Reversible Chloroplast + Freq. model for M, JTT matrix-based model for N and Jones et al. w/freq. model for S protein sequences. Model selection was conducted in MEGA X. Bootstrap values were estimated by 1000 replications. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using each model mentioned above. The tree is drawn to scale, and FigTree V1.4 was utilized to form cladogram branches (http://tree.bio.ed.ac.uk/software/figtree/). The aligned DNA sequences were also screened using RDP4 software to detect intragenic recombination among the alleles of each structural gene[12]. Six methods implemented in the RDP4 were utilized. These methods are RDP [12], GENECONV[13], BootScan [14], MaxChi[15], Chimaera [16], and SiScan [17]. Common settings for all methods include considering sequences as linear and setting statistical significance at the $P < 0.05$ with Bonferroni correction for multiple comparisons and requiring phylogenetic evidence and polishing of breakpoints. Potential recombination events (PREs) were considered as those identified by at least two methods. Reticulate network tree of alleles of the four structural genes of SARS-CoV-2 was also generated by Splitstree4 [18]. Phi test implemented in Splitstree4 was 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 the whole gene as previously described by our research group[19]. These analyses were carried out using DnaSP 6.12.03[11]. A statistical significance level with $P < 0.05$ is acceptable. The false discovery rate and 1000 replications in a coalescent simulation were applied for correcting multiple comparisons. Non-neutrality evolution was considered when identified by at least two out of three tests. Nonsynonymous and synonymous mutations of the alleles of the four structural genes were also calculated using MEGA X package [10].

Analysis of positive selection based on codon

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

Structural modeling of the protein with positive selection sites

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

PyMoL [23].

Results and Discussion

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

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

Further analysis revealed that there are 14 single nucleotide polymorphisms (SNPs) of E gene, but only 5 single amino acid polymorphic (SAP) loci in the E protein. Similar result was observed on M gene and protein, with 37 SNPs and 9 SAPs. In contrast, 126 SNPs and 75 SAPs are detected on N gene and protein, respectively. S protein, the most important factor that mediates virus entry by receptor binding and membrane fusion and determines the infection ability of SARS-CoV-2 [25], harbors 155 SNPs on the alleles and 90 SAPs in the protein. 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 most pairwise nucleotide differences among the four structural genes, indicating a more genetic diversity of S gene (Table 1). A key player in the virus transcription and assembly as N protein is [26, 27] , 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.

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*	□
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

h, Haplotypes,

π, Nucleotide diversity

S, Polymorphic sites

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

□, Total number of mutations

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

Distinct phylogenetic patterns of the four structural genes

The phylogenetic analysis revealed that all SARS-CoV-2 E proteins form three clusters. Similar to E protein, phylogenetic tree of SARS-CoV-2 M proteins is formed by three clusters with few branches (Figs 1A and 1B). The results suggest 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 pattern as compared with that of E and M. Four and three main phylogenetic clusters with various branches are identified in the N and S proteins, respectively (Figs 1C and D).

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

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

Although many studies demonstrated that recombination plays an important role on the emergence of SARS-CoV-2 and its contribution to admit SARS-CoV-2 as a human infectious pathogen [28-30], how this virus evolves during its global transmission has not been profiled yet. Therefore, we first analyzed intragenic recombination events of each structural gene using RDP4. The results indicate there were no recombination events occurred among the alleles of each gene (data not shown). Recombination event were also assessed through reticulate network tree by phi test in SplitsTree4. Although some internal nodes are noticed in N and S alleles, no significant evidence for recombination is validated of each gene by Phi test ($p>0.05$) (Fig 2). It indicates a relative stable state of SARS-CoV-2 during its transmission although a possible genetic interaction of different isolates might have occurred when it became a global pandemic [31, 32]. In addition, Tajima's D, 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. The results reveal that the evolution of all four genes does not match the neutral hypothesis, but favor purifying selection (Table 2 and Fig 3). The average of all pairwise *dN/dS* ratios (ω) among the 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 gene, respectively. All together, these results suggest that at the whole gene level, inconsistent purifying selection is the main evolution force (Table 2). 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 similar strong evolutionary constraints in different host species [33]. Similarly, our results suggest purifying selection drives the evolution at the whole structural gene level of SARS-CoV-2 during its transmission from human to human. This result also implies that in general, the genetic variation on these structural genes will not confer a significant disadvantage on the virus survival, and ratios reflect general variability of these genes and proteins. Considering that no recombination happened, nonsynonymous mutations would be removed at a great rate during the virus transmission [34].

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

Gene	Tajima'sD	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

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

Guo et al. reported that the S gene of SARSr-CoV populations in their natural host, Chinese horseshoe bat (*Rhinolophus sinicus*), has evolved through positive selection at some codons[9]. 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. Therefore, we used codon-substitution models to estimate the ratio of nonsynonymous over synonymous substitutions (*dN/dS*), also known as ω. The role of recombination in the polymorphism of four genes is excluded because no intragenic recombination was detected (Fig 2). By using ML model, we don't find any codon of E and M gene subjecting to positive selection obviously (data not shown). However, a potential positive selection site 208A in N gene is identified by using M3 model, but not by any other models especially the M8 model, suggesting a limited amount of evidence of positive selection in N gene (S1 Table). For the S gene, we found the average ω is 0.37199 calculated by M0 model of the codeML package, suggesting that purifying selection was a major force operating the evolution of the S gene during its transmission among human beings. In three LRTs, all alternative models (M3, M2a, M8) are significantly better fit ($P<10^{-4}$) than relevant null models (M0, M1a, M7), indicating that some sites of S 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 posterior probability of 1.000 in all the three models [21], a clear evidence showing that this site is still experiencing positive selection when the virus transmitted from human to human. The result is also validated using internal fixed effects likelihood (IFEL) and Evolutionary Fingerprinting methods implemented in HyPhy package (Fig 4) [35-37]. 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 play the most important role in virus-receptor interaction and virus entry into host cells [38]. This result suggests that a relatively genetic stability of this motif would benefit the virus survival. Intriguingly, the site under positive selection pressure always has a D614G (for the S gene is 1841A>G) mutation, implying such mutation may enhance virus adaptability in human hosts. Another potential positive selection site at codon 5 is also identified, and a L5F mutation (for the S gene is 13C>T) is 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. Similar result was also confirmed by Evolutionary Fingerprinting method (S1 Fig). Considering signal sequence (SS) is a short hydrophobic peptide that plays an important role in guiding viral protein into the endoplasmic reticulum (ER) for proper folding and assembly [39], we postulate that L5F mutation may increase hydrophobicity of the SS, thus facilitating the entry of S protein into ER for folding and assembly, and in turn secretion of the virus.

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(beta&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(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)≥0.80 are shown, (p)≥0.95 (p)≥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.

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

Phylogenetic tree of S gene alleles was derived to test the evolutionary relationship among the alleles with or without D614G mutation. As shown in Fig 5A, the 173 alleles of the S gene could be clustered into four clades. Alleles with D614G mutation could be found in all 4 clades, among which a dominant one contains 79 out of 85 alleles with such mutation. The remaining 6 mutated S alleles are distributed in other 3 clades. The result suggests a potential common ancestor for the majority of S alleles with D614G mutation, while some other maybe derived from alternative ancestors. This result is also supported by the parsimony network of S gene alleles using PopART (http://popart.otago.ac.nz) [40]. 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 5B). All S alleles with D614G mutation are closely related (with a few point mutations), and comprise a scattered star structure, suggesting the expansion of SARS-CoV-2 population with D614G mutation on S gene. In contrast, alleles of the N gene show a single ancestor analyzed by parsimony network though 3 phylogenetic clades are identified (S2 Fig).

A total of 5 alleles with L5F mutation are found and all of them are in one clade, accounting for 83.33% of all alleles in the clade (S3A Fig). Further parsimony network analysis reveals that S alleles with L5F mutation are not closely related, but distribute in both WH01 and GZMU0019 haplotype groups (S3B Fig). No scattered star structure of these alleles can be formed, indicating 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, GWHABKF00000001 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.

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 cathepsin L, which generate the receptor binding subunit S1 and the membrane fusion S2 [44-46]. From structural studies in both SARS-CoV and SARS-CoV-2, receptor binding domain (RBD) located at the C-terminal of S1 and the adjacent N-terminal domain (NTD) are relatively flexible, which is the feature required for receptor recognition and subsequent membrane fusion[47, 48]. We found that the D614G mutation is located at the subdomain 2 (SD2) that at the C-terminal of RBD and close to the two potential cleavage sites between S1 and S2 [48] (Fig 7A). Considering that 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[5, 44], and in turn improving the infection efficiency. From the latest cryo-electron microscopy (cryo-EM) structure of SARS-CoV-2 S protein, the negatively charged sidechain of D614 points towards the positively charged sidechain of K854 from the neighboring monomer (Fig 7B) [48] . The distance between the closest atoms of the two residues is 2.6 Å, which is an optimal distance to form salt bridge (Fig 7C). From the modelled structure with D614G mutation, the distance is increased to 5.2 Å (Fig 7D), 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 S protein, where RBD move away from the core structure and expose its receptor binding surface. The entire S trimer then undergoes a serial of dramatic conformation changes, including cleavages between S1 and S2, disassociation of S1 and post-fusion transformation of S2 [49, 50]. Changes including mutations at cleavage sites and adding internal crosslinks in S trimer would keep the protein in a stable and “closed” conformation where the receptor binding surface of RBD is inaccessible [48, 51]. Therefore, we hypothesize that the highly transmissible D614G mutation driven by the positive selection through evolution promotes accessibility of RBD by losing a critical salt bridge between the S protein monomers, which subsequently triggers membrane fusion upon ACE2 binding.

Conclusions

We present modern molecular evolution 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 depicted. Protein sequence comparisons show E and M genes exhibit a relatively close relationship to bat SARSr-CoV, suggesting the evolution 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 of N and S genes. No clear intragenic recombination is detected of these four genes, suggesting that it is not the major force to drive the evolution of the four genes. However, our analyses show purifying selection pressure may be the main force operating the evolution at whole gene levels of SARS-CoV-2 during its human to human transmission. We also identify a codon in S gene definitively experiencing positive selection pressure, and always leads to the D614G mutation in S proteins. S alleles with D614G mutation have expanded rapidly among SARS-CoV-2 isolates. 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, promote RBD opening, and facilitate the entry of the virus into host cells, thus contributing to the diffusion of this mutated alleles. Codon 5 of S gene is another potential positive selection site. Although a limited number of alleles with L5F mutation is identified, it may potentially affect the assembly and secretion of SARS-CoV-2. A close eye on L5F mutation may be required in case another expansion occurs. As S protein is a key target for SARS-CoV-2 vaccines, therapeutic antibodies, and diagnostics, the D614G mutation of S should be paid more attention. Owning that the exact mechanism remains unclear, further study should focus on the exact function of these mutation sites and how they affect the expansion of these mutated alleles on SARS-CoV-2.

Acknowledgements

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

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

References

1. 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):2043-8. Epub 2007/05/29. doi: 10.1007/s00018-007-7103-1. PubMed PMID: 17530462; PubMed Central PMCID: PMCPMC7079843.

2. Jimenez-Guardeno JM, Nieto-Torres JL, DeDiego ML, Regla-Nava JA, Fernandez-Delgado R, Castano-Rodriguez C, et al. The PDZ-binding motif of severe acute respiratory syndrome coronavirus envelope protein is a determinant of viral pathogenesis. PLoS pathogens. 2014;10(8):e1004320. Epub 2014/08/15. doi: 10.1371/journal.ppat.1004320. PubMed PMID: 25122212; PubMed Central PMCID: PMCPMC4133396.

3. 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):11418-28. Epub 2010/08/20. doi: 10.1128/JVI.01131-10. PubMed PMID: 20719948; PubMed Central PMCID: PMCPMC2953170.

4. McBride R, van Zyl M, Fielding BC. The coronavirus nucleocapsid is a multifunctional protein. Viruses. 2014;6(8):2991-3018. Epub 2014/08/12. doi: 10.3390/v6082991. PubMed PMID: 25105276; PubMed Central PMCID: PMCPMC4147684.

5. Belouzard S, Millet JK, Licitra BN, Whittaker GR. Mechanisms of coronavirus cell entry mediated by the viral spike protein. Viruses. 2012;4(6):1011-33. Epub 2012/07/21. doi: 10.3390/v4061011. PubMed PMID: 22816037; PubMed Central PMCID: PMCPMC3397359.

6. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated 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, SÃ nchez-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.

13. Padidam M, Sawyer S, Fauquet CM. Possible emergence of new geminiviruses by frequent recombination. Virology. 1999;265(2):218-25. Epub 1999/12/22. doi: 10.1006/viro.1999.0056. PubMed PMID: 10600594.

14. 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. doi: 10.1089/aid.2005.21.98. PubMed PMID: 15665649.

15. Smith JM. Analyzing the mosaic structure of genes. Journal of molecular evolution. 1992;34(2):126-9. PubMed PMID: 1556748.

16. Posada D. Evaluation of methods for detecting recombination from DNA sequences: empirical data. Molecular biology and evolution. 2002;19(5):708-17. PubMed PMID: 11961104.

17. Gibbs MJ, Armstrong JS, Gibbs AJ. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics. 2000;16(7):573-82. PubMed PMID: 11038328.

18. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Molecular biology and evolution. 2006;23(2):254-67. Epub 2005/10/14. doi: 10.1093/molbev/msj030. PubMed PMID: 16221896.

19. 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. Epub 2017/12/12. doi: 10.7717/peerj.4114. PubMed PMID: 29226035; PubMed Central PMCID: PMCPMC5719964.

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

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

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

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

24. 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):9059-67. Epub 2001/09/05. doi: 10.1128/JVI.75.19.9059-9067.2001. PubMed PMID: 11533169; PubMed Central PMCID: PMCPMC114474.

25. 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):562-9. Epub 2020/02/26. doi: 10.1038/s41564-020-0688-y. PubMed PMID: 32094589; PubMed Central PMCID: PMCPMC7095430.

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

27. 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):12862-72. Epub 2010/02/16. doi: 10.1074/jbc.M109.030270. PubMed PMID: 20154085; PubMed Central PMCID: PMCPMC2857088.

28. Wong MC, Javornik Cregeen SJ, Ajami NJ, Petrosino JF. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. 2020:2020.02.07.939207. doi: 10.1101/2020.02.07.939207 bioRxiv.

29. Wu Y. Strong evolutionary convergence of receptor-binding protein spike between COVID-19 and SARS-related coronaviruses. 2020:2020.03.04.975995. doi: 10.1101/2020.03.04.975995. bioRxiv.

30. Wu A, Niu P, Wang L, Zhou H, Zhao X, Wang W, et al. Mutations, Recombination and Insertion in the Evolution of 2019-nCoV. 2020:2020.02.29.971101. doi: 10.1101/2020.02.29.971101 bioRxiv.

31. 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.

32. Mallapaty S. How sewage could reveal true scale of coronavirus outbreak. Nature. 2020;580(7802):176-7. Epub 2020/04/05. doi: 10.1038/d41586-020-00973-x. PubMed PMID: 32246117.

33. Li X, Giorgi EE, Marichann MH, Foley B, Xiao C, Kong X-P, et al. Emergence of SARS-CoV-2 through Recombination and Strong Purifying Selection. 2020:2020.03.20.000885. doi: 10.1101/2020.03.20.000885. bioRxiv.

34. Hughes AL, Hughes MA. More effective purifying selection on RNA viruses than in DNA viruses. Gene. 2007;404(1-2):117-25. Epub 2007/10/12. doi: 10.1016/j.gene.2007.09.013. PubMed PMID: 17928171; PubMed Central PMCID: PMCPMC2756238.

35. Pond SLK, Muse SV. HyPhy: Hypothesis Testing Using Phylogenies: Springer New York; 2005. 676-9 p.

36. Pond SL, Scheffler K, Gravenor MB, Poon AF, Frost SD. Evolutionary fingerprinting of genes. Molecular biology and evolution. 2010;27(3):520-36. Epub 2009/10/30. doi: 10.1093/molbev/msp260. PubMed PMID: 19864470; PubMed Central PMCID: PMCPMC2877558.

37. 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):1208-22. Epub 2005/02/11. doi: 10.1093/molbev/msi105. PubMed PMID: 15703242.

38. 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). Epub 2020/01/31. doi: 10.1128/JVI.00127-20. PubMed PMID: 31996437; PubMed Central PMCID: PMCPMC7081895.

39. 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):281-92 e6. Epub 2020/03/11. doi: 10.1016/j.cell.2020.02.058. PubMed PMID: 32155444; PubMed Central PMCID: PMCPMC7102599.

40. Clement M, Snell Q, Walker P, Posada D, Crandall KJP, Distributed Processing

Symposium IP. TCS: Estimating gene genealogies. 2002;2:184.

41. Bubac CM, Spellman GMJTAOA. How connectivity shapes genetic structure during range expansion: Insights from the Virginia's Warbler. 2016;(2):2.

42. Duxbury EM, Day JP, Maria Vespasiani D, Thuringer Y, Tolosana I, Smith SC, et al. Host-pathogen coevolution increases genetic variation in susceptibility to infection. eLife. 2019;8. Epub 2019/05/01. doi: 10.7554/eLife.46440. PubMed PMID: 31038124; PubMed Central PMCID: PMCPMC6491035.

43. Meyerson NR, Sawyer SL. Two-stepping through time: mammals and viruses. Trends in microbiology. 2011;19(6):286-94. Epub 2011/05/03. doi: 10.1016/j.tim.2011.03.006. PubMed PMID: 21531564; PubMed Central PMCID: PMCPMC3567447.

44. 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):468-78. Epub 2015/07/25. doi: 10.1016/j.tim.2015.06.003. PubMed PMID: 26206723; PubMed Central PMCID: PMCPMC7125587.

45. Bestle D, Heindl MR, Limburg H, van TVL, Pilgram O, Moulton H, 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. 2020:2020.04.15.042085. doi: 10.1101/2020.04.15.042085. bioRxiv.

46. Ou X, Liu Y, Lei X, Li P, Mi D, Ren L, 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. Epub 2020/03/30. doi: 10.1038/s41467-020-15562-9. PubMed PMID: 32221306; PubMed Central PMCID: PMCPMC7100515.

47. Gui M, Song W, Zhou H, Xu J, Chen S, Xiang Y, et al. Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. Cell research. 2017;27(1):119-29. Epub 2016/12/23. doi: 10.1038/cr.2016.152. PubMed PMID: 28008928; PubMed Central PMCID: PMCPMC5223232.

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

49. Walls AC, Xiong X, Park YJ, Tortorici MA, Snijder J, Quispe J, et al. Unexpected Receptor Functional Mimicry Elucidates Activation of Coronavirus Fusion. Cell. 2019;176(5):1026-39 e15. Epub 2019/02/05. doi: 10.1016/j.cell.2018.12.028. PubMed PMID: 30712865; PubMed Central PMCID: PMCPMC6751136.

50. Walls AC, Tortorici MA, Snijder J, Xiong X, Bosch BJ, Rey FA, et al. 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):11157-62. Epub 2017/10/27. doi: 10.1073/pnas.1708727114. PubMed PMID: 29073020; PubMed Central PMCID: PMCPMC5651768.

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

Supporting information

S1 Table. SARS-CoV-2 isolates information.

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

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

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

S1 Fig. The evolutionary relationship of N alleles. A. Phylogenetic tree of N gene based on nucleotide sequences of 131 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. Bootstrap values more than 0.5 are shown. **B.** Parsimony network of SARS-CoV-2 N 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. The ancestral haplotype, or root of the network, is labeled with a square, and represent haplotype name is marked red.

S2 Fig. Evolutionary relationship of S alleles with or without L5F mutation. A. Phylogenetic tree of S gene based on nucleotide sequences of 173 alleles. Each clade is highlighted with different color. Alleles are shown with their representative isolate names, and alleles with L5F mutation are highlighted in blue. 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 blue nodes indicate haplotypes with L5F mutation. Dotted boxes indicate major haplotype groups. Haplotypes include in red dotted boxes are with D614G mutation while those included in black dotted boxes are without D614G mutation.

S3 Fig. Positive selection analysis of S gene codon 5 by Evolutionary Fingerprinting method. Log (Bayes Factor) for positive selection at codon 5 of S gene and its frequencies. The cut-off value for the Bayes factor (BF) in the Evolutionary Fingerprinting method was set at 25 to reflect a positive selection at a given site (Posterior probability>0.95). Pr {BF>25} indicates posterior probability of Bayes Factor >25.

Figure legends

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 E (A), M (B), N (C) and S (D) alleles of SARS-CoV-2 analyzed by the neighbor-net algorithm of SplitsTree4. 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 are too few informative characters to use the Phi test for E and M genes, *p-values* of Phi test of N and S genes are shown.

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

Figure 4. Positive selection analysis of S gene codons by IFEL and Evolutionary Fingerprinting methods. A. Diagram of selection analysis result of S codons by IFEL method. Asterisk indicates the positive selection site with statistical significance (*p<0.01*). **B.** Log (Bayes Factor) for positive selection at codon 614 of S gene and its frequencies. The cut-off value for the Bayes factor (BF) in the Evolutionary Fingerprinting method was set at 25 to reflect a positive

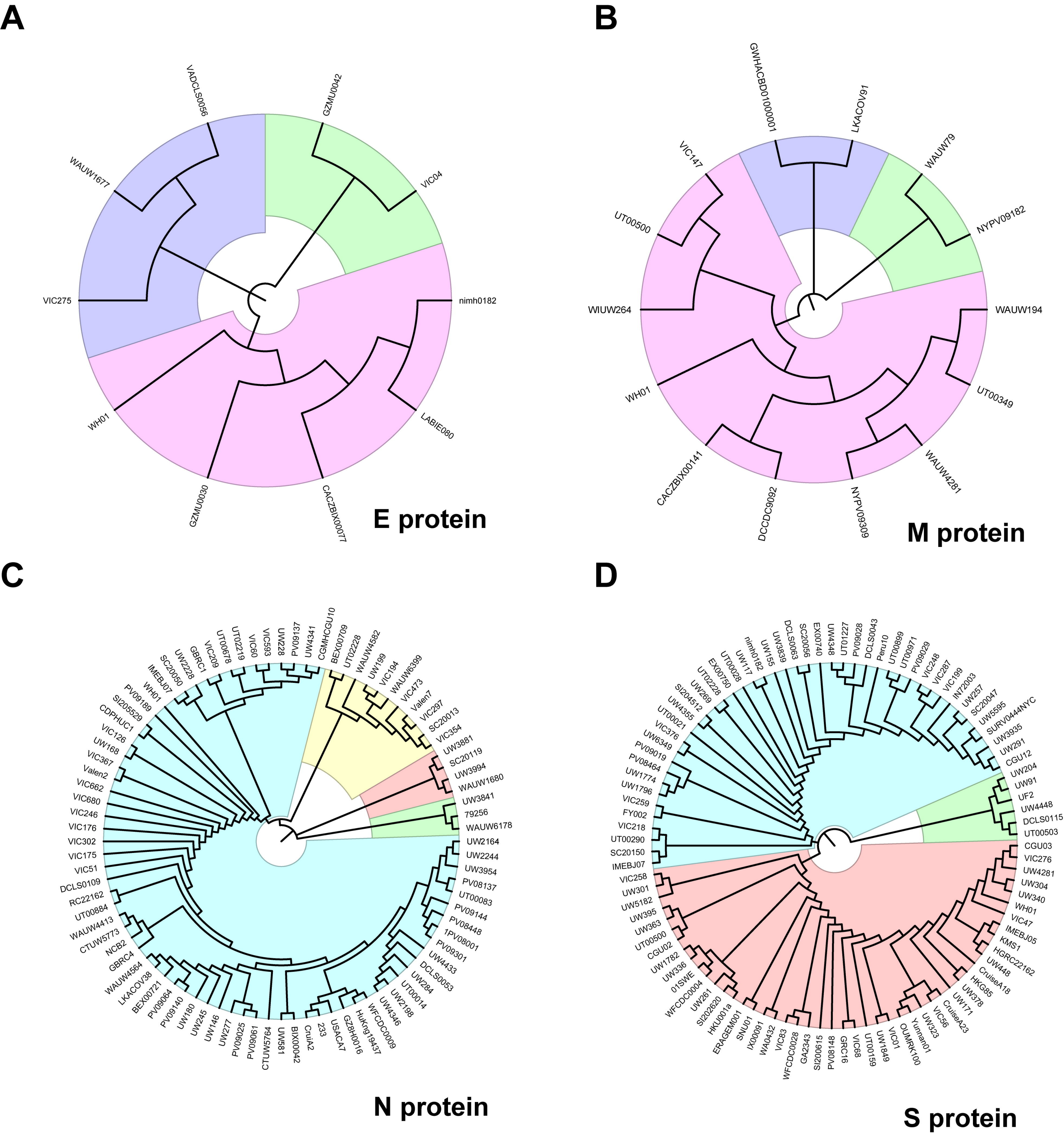
selection at a given site (Posterior probability>0.95). Pr {BF>25} indicates posterior probability of Bayes Factor >25.

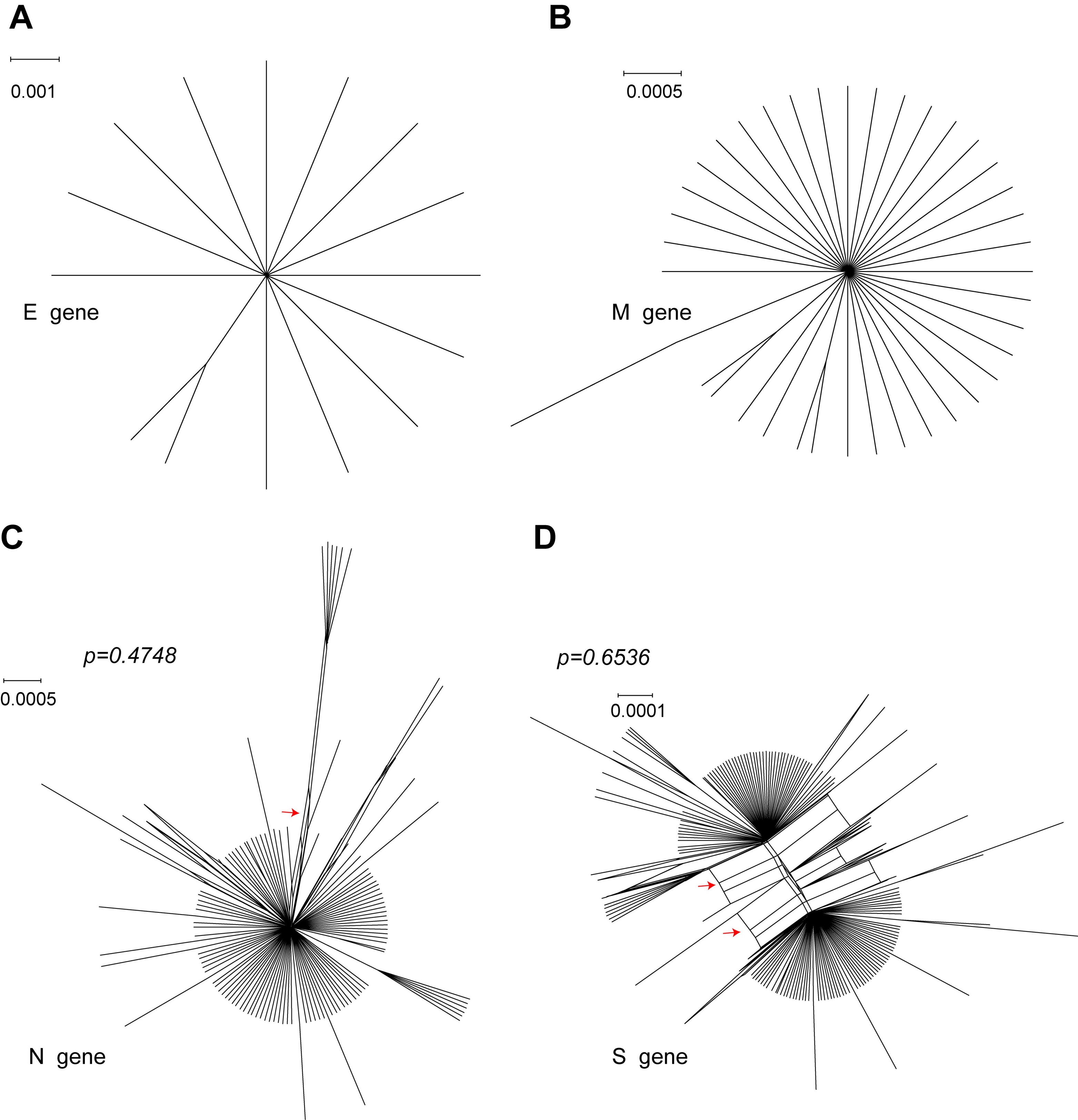
Figure 5. Evolutionary relationship of S alleles with or without D614G mutation. A. Phylogenetic tree of S gene based on nucleotide sequences of 173 alleles. The evolutionary history is inferred using the Maximum Likelihood method and Tamura-Nei model. The tree is drawn to scale, with branch lengths 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.

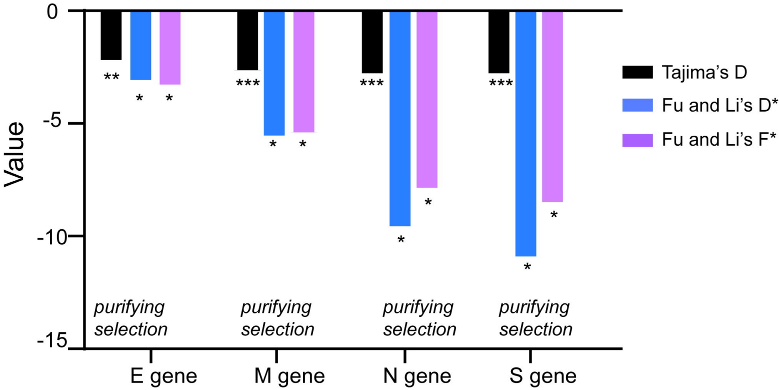
Figure 6. Expansion of S alleles with D614G mutation during SARS-CoV-2 human to human transmission. **A**. Percentage of SARS-CoV-2 isolates carrying the alleles of D614G mutation in each week collected. **B.** Frequencies of D614G mutation in the S gene in each period of time (Four to five weeks' data are combined). **C.** Percentage of SARS-CoV-2 isolates carrying the alleles with

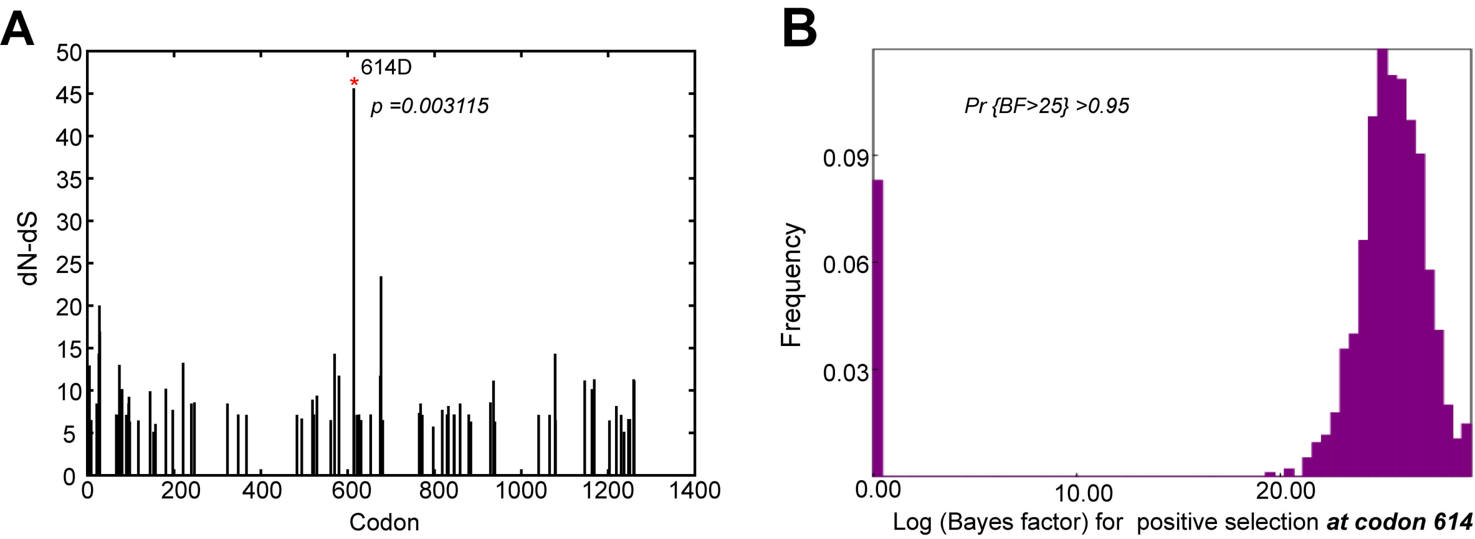
D614G mutation in each period of time. **D.** Frequencies of D614G mutation in the S gene in each period of time. *p < 0.05; **p < 0.01.

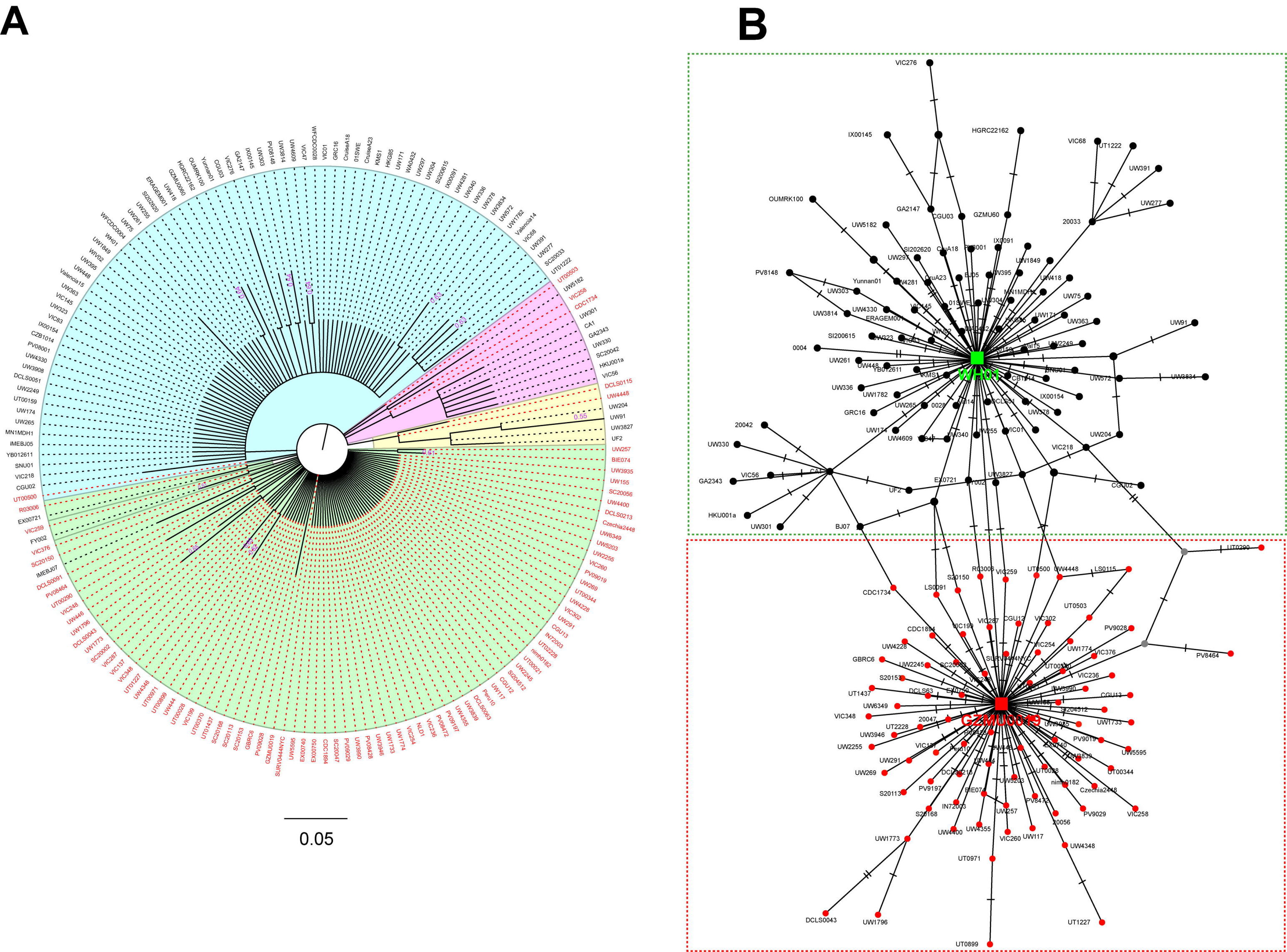
Figure 7. The structure of the S protein of SARS-CoV-2 and potential influence of D614G mutation on its structural change. **A.** Schematic of the primary structure of SARS-CoV-2 S protein colored by domains. Some boundary-residues are listed. The S1/S2 cleavage sites are indicated by arrows. RBD: receptor 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; CTD: C-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 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 mutation as shown.

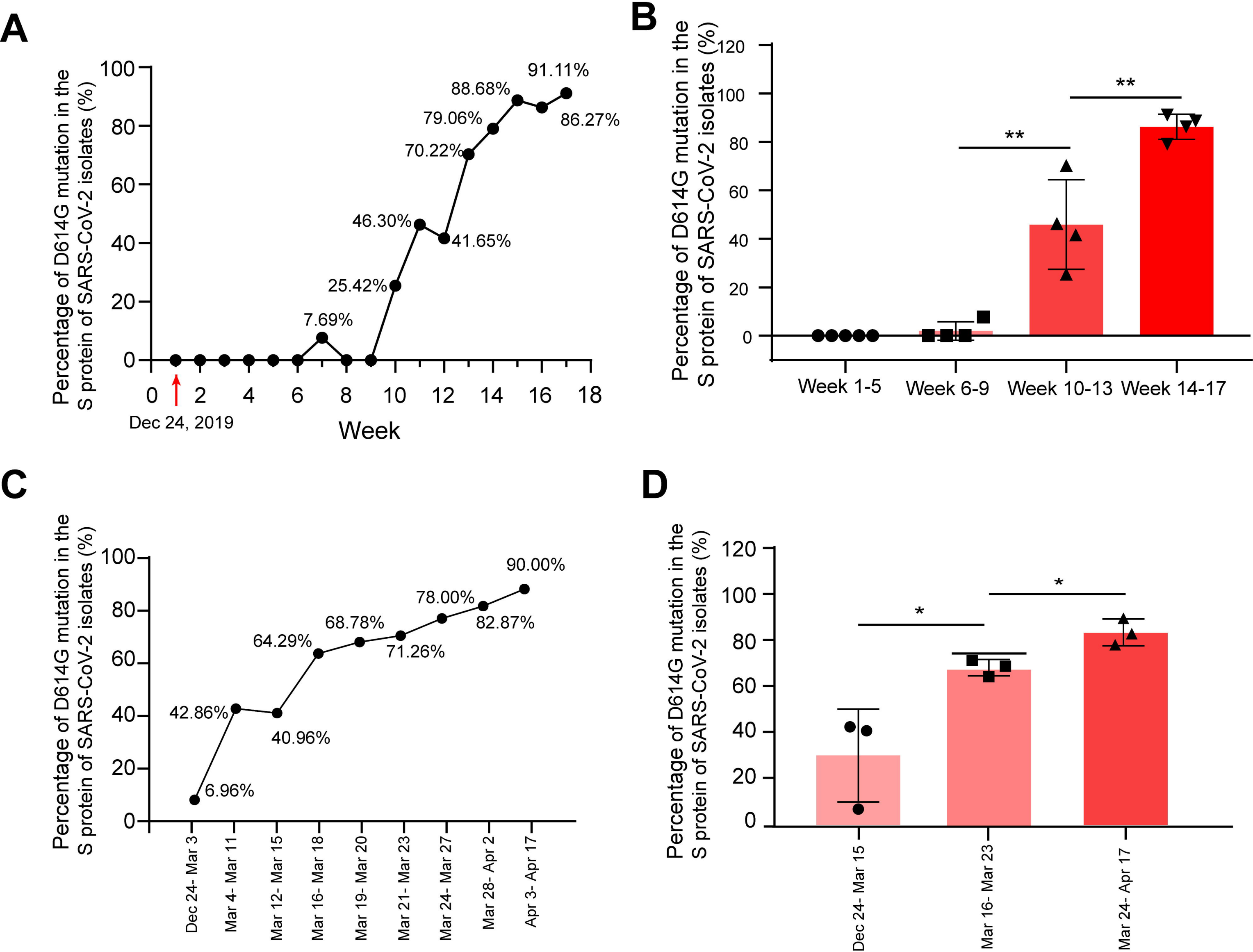

A
VADCLS0056
GZMU0042
VIC04
nimh0182
LABIE080
CACZBIX00077
GZMU0030
WH01
VIC275
WAUW1677
E protein
B
GWHACBDD01000001
LKACOV91
WAUW79
NYPV09182
WAUW194
UT00349
WAUW4281
NYPV09309
DCCDC9092
CACZBIX00141
WH01
WIUW264
UT00500
VIC147
M protein
C
CGMHCGU10
BEX00709
UT02228
WAUW4582
UW199
VIC194
WAUW6399
VIC473
Valen7
VIC297
SC20013
VIC354
UW3881
SC20119
UW3994
WAUW1680
UW3841
79256
WAUW6178
UW2164
UW2244
UW3954
PV08137
UT00083
PV09144
PV08448
1PV08001
PV09301
UW4433
DCLS0053
UW284
UT00014
UW2198
UW4346
WFCDC0009
HuKng19437
GZ8H0016
USACA7
233
CruiA2
BIX00042
UW581
CTUW5764
PV09061
PV09025
UW277
UW146
UW245
UW180
PV09140
PV09064
BEX00721
LKACOV38
WAUW4564
GBRC4
NCB2
CTUW5773
WAUW4413
UT00884
RC22162
DCLS0109
VIC51
VIC175
VIC302
VIC176
VIC246
VIC680
VIC662
Valen2
VIC367
UW168
VIC126
CDPHUC1
SI205529
PV09189
WH01
IMEBJ07
SC20050
UW2228
GBRC1
VIC209
UT00678
UT02219
VIC60
VIC593
UW228
PV09137
UW4341
N protein
D
UW4348
UT01227
PV09028
DCLS0043
Peru10
UT00899
UT00971
PV09029
VIC248
VIC287
VIC199
IN72003
UW257
SC20047
UW5595
SURV0444NYC
UW3935
UW291
CGU12
UW204
UW91
UF2
UW4448
DCLS0115
UT00503
CGU03
VIC276
UW4281
UW304
UW340
WH01
VIC47
IMEBJ05
KMS1
HGRC22162
UW448
CruiseA18
HKG85
UW378
UW171
CruiseA23
VIC56
UW323
Yunnan01
OUMRK100
VIC01
UW1849
UT00159
VIC68
GRC16
PV08148
SI200615
GA2343
WFCDC0028
VIC83
WA0432
IX00091
SNU01
ERAGEM001
HKU001a
SI202620
UW261
WFCDC0004
01SWE
UW336
UW1782
CGU02
UT00500
UW363
UW395
UW5182
UW301
VIC258
IMEBJ07
SC20150
UT00290
VIC218
FY002
VIC259
UW1796
UW1774
PV08464
PV09019
UW6349
VIC376
UT00021
UW4355
SI204512
UW269
UT02228
EX00750
UT00028
UW117
nimh0182
UW155
UW3839
DCLS0063
SC20056
EX00740
S protein

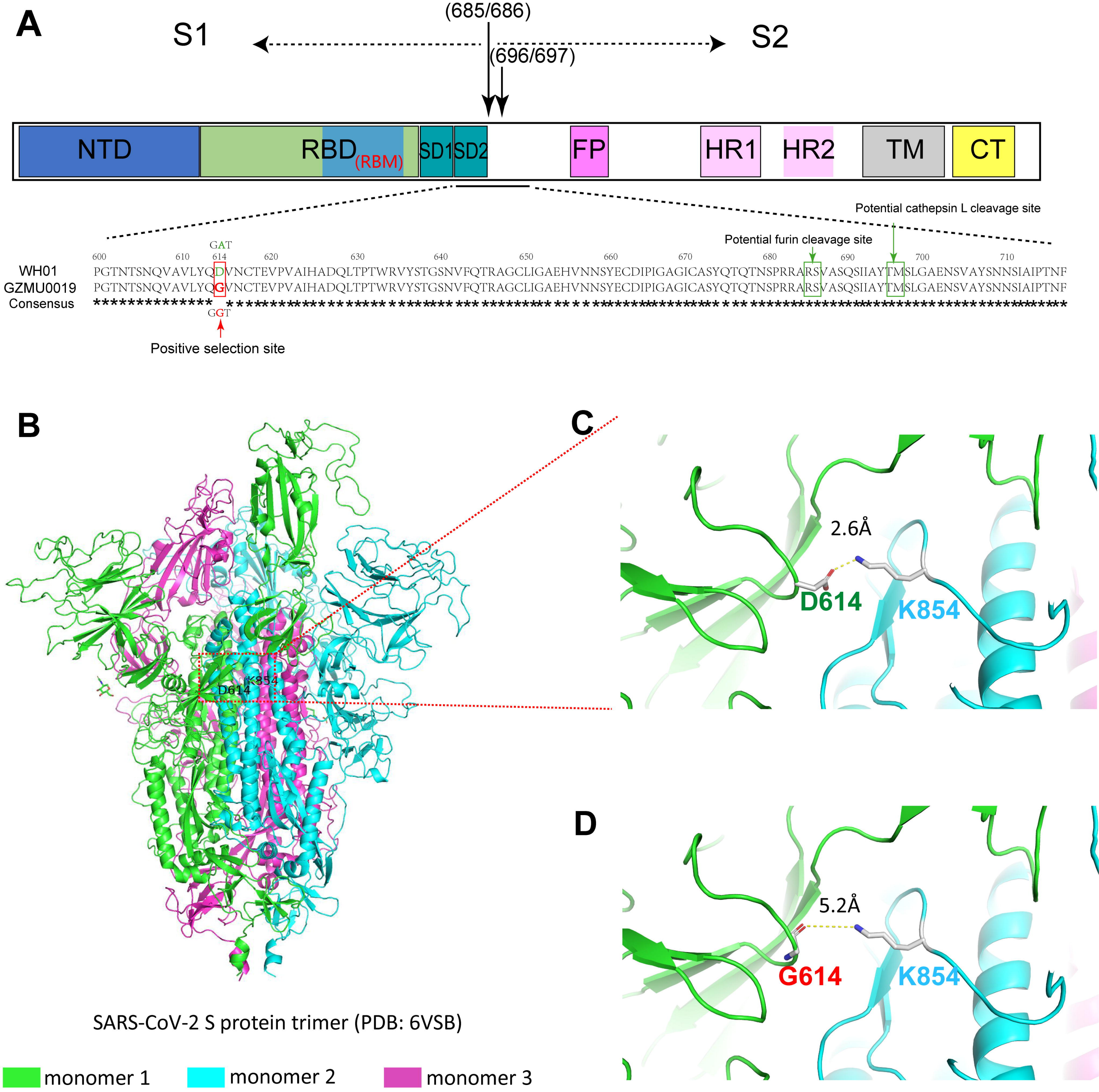

A
0.001
E gene
B
0.0005
M gene
C
p=0.4748
0.0005
N gene
D
p=0.6536
0.0001
S gene


0
-5
-10
-15
Value
**
*
*
***
*
*
***
*
*
***
*
*
purifying selection
purifying selection
purifying selection
purifying selection
E gene
M gene
N gene
S gene
Tajima's D
Fu and Li's D*
Fu and Li's F*


A
B
50
45
40
35
30
25
20
15
10
5
0
dN-dS
0
200
400
600
800
1000
1200
1400
Codon
614D
*
p =0.003115
Pr {BF>25} >0.95
0.09
0.06
0.03
Frequency
0.00
10.00
20.00
Log (Bayes factor) for positive selection at codon 614


A
B
0.05
WH01
GZMU0019


A
Percentage of D614G mutation in the S protein of SARS-CoV-2 isolates (%)
100
80
60
40
20
0
0
2
4
6
8
10
12
14
16
18
Week
Dec 24, 2019
7.69%
25.42%
46.30%
41.65%
70.22%
79.06%
88.68%
91.11%
86.27%
B
Percentage of D614G mutation in the S protein of SARS-CoV-2 isolates (%)
120
100
80
60
40
20
0
Week 1-5
Week 6-9
Week 10-13
Week 14-17
**
**
C
Percentage of D614G mutation in the S protein of SARS-CoV-2 isolates (%)
100
80
60
40
20
0
Dec 24- Mar 3
Mar 4- Mar 11
Mar 12- Mar 15
Mar 16- Mar 18
Mar 19- Mar 20
Mar 21- Mar 23
Mar 24- Mar 27
Mar 28- Apr 2
Apr 3- Apr 17
6.96%
42.86%
40.96%
64.29%
68.78%
71.26%
78.00%
82.87%
90.00%
D
Percentage of D614G mutation in the S protein of SARS-CoV-2 isolates (%)
120
100
80
60
40
20
0
Dec 24- Mar 15
Mar 16- Mar 23
Mar 24- Apr 17
*
*


A
(685/686)
S1
S2
(696/697)
NTD
RBD
(RBM)
SD1
SD2
FP
HR1
HR2
TM
CT
Potential cathepsin L cleavage site
Potential furin cleavage site
GAT
600
610
614
620
630
640
650
660
670
680
690
700
710
WH01 PGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNF
GZMU0019 PGTNTSNQVAVLYQGVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNF
Consensus
GGT
Positive selection site
B
K854
D614
SARS-CoV-2 S protein trimer (PDB: 6VSB)
monomer 1
monomer 2
monomer 3
C
2.6Å
D614
K854
D
5.2Å
G614
K854