Molecular Evolution of SARS-CoV-2 Structural Genes: Evidence of Positive Selection in 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 (✉ huangbh7@mail.sysu.edu.cn)

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: July 29th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-42498/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background:

SARS-CoV-2 has caused a global pandemic since early 2020 and is still a serious public health issue world-wide. 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. The evolution of these genes may determine the infectivity of SARS-CoV-2, but is largely unknown.

Results:

We analyzed roughly 3090 SARS-CoV-2 isolates from GenBank database. The distribution of 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 sequence. Although intragenic recombination event isn't detected among different alleles, purifying selection has conducted on the evolution of these genes. By analyzing full genomic sequences of these alleles, it reveals that codon 614 of S glycoprotein has subjected to strong positive selection pressure and a consistent D614G mutation is identified. Additionally, another potential positive selection site at codon 5 in the signal sequence of the S protein is also identified with consistent L5F mutation. The allele containing D614G mutation has undergone significant expansion during SARS-CoV-2 transmission, 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 a salt bridge between S protein monomers and 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.

Conclusions:

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.

Background

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of an emerging coronavirus disease (COVID-19) that has caused more than 430,000 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.

Results

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

The 3090 SARS-CoV-2 isolates harbor only 16 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 comparison of the WH01 isolate with a SARSr-CoV isolate, bat-SL-CoVZC45, shows identity of 100% (75/75) in E, 98.65% (219/222) in M, 94.27% (395/419) in N and 80.06% (1171/1273) in S proteins, respectively. These results imply a close homology between SARS-CoV-2 and bat SARSr-CoV, particularly on E and M proteins. On the other hand, it indicates an extreme conservation of E and M proteins and their functions among coronaviruses[10].

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 component that mediates virus entry by receptor binding and membrane fusion and determines the infectivity of SARS-CoV-2 [11], harbors 155 SNPs on 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).

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

Gene	Sequence, n*	Sequence length	<i>h</i>	π	<i>S</i>	θ	Σ
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

Σ , 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 (Fig. 1a and b). 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 (Fig. 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 harboring different N and S variants (such as those clustered into different clades) may influence their infectivity 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 [12-14], 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 is also assessed through reticulate network tree by phi test in SplitsTree4. Although some internal nodes are noticed in N and S alleles, no clear 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 though a possible genetic interaction of different isolates might have occurred when it became a

global pandemic [15, 16]. 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; Additional file 1: Figure S1). 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; Additional file 1: Figure S1).

Table 2. Summary of neutrality for 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

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 SARS-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). Although a potential positive selection site 208A in the N gene is identified by using M3 model, it is not validated by any other models especially the M8 model, suggesting the evidence for N gene positive selection is limited (Additional file 2: Table S1). For the S gene, we found the average ω is 0.37199 calculated by M0 model of the codeML package, suggesting that purifying selection operates S gene evolution of during SARS-CoV-2 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 ($\omega=18.22175-20.61283$) (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 [17], a clear evidence showing that this site is still experiencing positive selection when the virus is transmitting from human to human. The result is also validated using internal fixed effects likelihood (IFEL) and Evolutionary Fingerprinting methods implemented in HyPhy package (Fig. 3) [18-20]. 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 [21]. 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 (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 $\omega_0=0.26126$, $\omega_1=2.70530$, $\omega_2=20.61283$	M0 vs. M3	0.000000001	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	$\omega_0=0.37199$			Not Allowed
M2a(selection)	-6766.432802	p0=0.81731, p1=0.17872, p2=0.00397 $\omega_0=0.17504$, $\omega_1=1.00000$, $\omega_2=18.76936$	M1a vs. M2a	0.000004385	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 $\omega_0=0.04395$, $\omega_1=1.00000$			Not Allowed
M8(beta& ω)	-6768.829411	p0=0.99578, p=0.40368, q=0.82224 p1= 0.00422, $\omega=18.22175$	M7 vs.M8	0.000030400	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			Not Allowed

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

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. 4a, 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. This result is also supported by the parsimony network of S gene alleles using PopART (<http://popart.otago.ac.nz>) [22]. 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 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 (Additional file 4: Figure S3).

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 (Additional file 5: Figure S4a). Further parsimony network analysis reveals that S alleles with L5F mutation are not closely related, but distribute in both WH01 and GZMU0019 haplotype groups (Additional file 5: Figure S4b). No scattered star structure of these alleles can be formed, indicating L5F mutation might arise from independent origins unlike D614G mutants.

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 is obtained from the haplotype network of S alleles mentioned above (Fig. 4b). 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 [23]. 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 Additional file 6: Table S2 and Additional file 7: Table S3.

In 173 S gene alleles, 85 carry D614G mutation, accounting for 49.13% of all. Similarly, 47 out of 99 S proteins carry D614G mutation, accounting for 47.47% of all. The first two isolates, GWHABKF000000001 and WH01 in our data collect (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 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 (Fig 5a; Additional file 6: Table S2). Further analysis reveals that the frequency of D614G mutation in the S gene was steadily increasing when combining data from week 6 to 17 (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 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 (Fig. 5c and d; Additional file 7: Table S3). 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

We found that the D614G mutation is located at the subdomain 2 (SD2) that at the C-terminus of RBD and close to the two potential cleavage sites between S1 and S2 [24] (Fig. 6a). 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, 25], and in turn improving the infectivity. 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. 6b) [24]. The distance between the closest atoms of the two residues is 2.6 Å, which is an optimal distance to form salt bridge (Fig. 6c). From the modelled structure with 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 S protein, where RBD move away from the core structure and expose 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 [26, 27]. 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 [24, 28]. 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.

Discussion

Many studies demonstrated the continuing evolution of SARS-CoV-2 [29–31]. Four structural genes of SARS-CoV-2, E, M, N and S, may determine the infectivity or pathogenesis of this persistent transmission virus, but the molecular evolution patterns of these structural genes remain largely unknown.

Among four structural genes, E and M show the highest homology and few polymorphisms of both SNP and SAP are found, indicating the importance of the conservation of these two genes for the virus survival. A key factor for virus transcription and assembly as N protein is [32, 33], high sequence variability of the N protein is found in this study, suggesting a vast adaption of the virus during host transmission. Previous study shows that high genetic variation has been found among bat SARS-CoVs, particularly in the S gene [9]. Similarly, high nucleotide diversity (π , a major parameter to define genetic diversity, Table 1) of the S gene is also detected on SARS-CoV-2 isolates, suggesting this may benefit virus survival in the host of human beings. Recombination is an important evolutionary event during the emergence of SARS-CoV-2 [34, 35]. It is reported that SARS-CoV-2 is a recombinant virus of bat and pangolin CoVs, suggesting the most critical role of recombination [34]. However, 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), indicating the evolution of these genes is not predominately 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 similar strong evolutionary constraints in different host species [35]. Similarly, our results also show purifying selection drives the evolution at the whole structural gene level of SARS-CoV-2 during its transmission from human to human (Table 2; Additional file 1: Figure S1). This result also implies that in general, the genetic variation on these structural genes will not confer a significant disadvantage on the virus survival. Because no recombination happened during SARS-CoV-2 evolution, nonsynonymous mutations would be removed at a great rate during the virus transmission [36], while positive selection site with mutation will be fixed. The frequency of S alleles with D614G mutation is increasing during SARS-CoV-2 spread proves the case (Fig. 3; Fig. 5; Additional file 6: Table S2; Additional file 7: Table S3).

We also identified another potential positive selection site at codon 5 of S genes with 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 protein into the endoplasmic reticulum (ER) for proper folding and assembly [37], 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. In addition, our results show that majority of S

alleles with D614G mutation are clustered in one clade and make a distinct star scattering network group (Fig. 4), suggesting a potential common ancestor for these mutants. However, sporadic alleles with L5F mutation identified so far indicates that L5F mutation might subject to relatively less strong pressure and is still at early stage of positive selection (Additional file 5: Figure S4).

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[38]. 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[39]. 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 [25, 40, 41]. 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[24, 42]. From the S protein structural modeling (Fig. 6), we hypothesize that D614G mutation driven by the positive selection through molecular evolution promotes accessibility of RBD by losing a critical salt bridge between the S protein monomers (Fig. 6c and d), which subsequently triggers membrane fusion upon ACE2 binding. However, the exact influence and detailed mechanism of D614G mutation on SARS-CoV-2 infectivity and expansion need further investigation and empirical evidence is required. It is to be noted that a consistent L5F mutation is always found in a potential positive selection site of S gene (codon 5). The low frequency (3.82%, 5/131) of S alleles with L5F mutation does not show clear increasing pattern yet, possibly due to relatively less strong positive selection pressure as compared to codon 614. Because the terminal point of the studied isolates is late April, 2020, the persistent monitoring of frequency of L5F in S alleles is required to determine whether it experiences expansion and potential effect of L5F mutation to SARS-CoV-2 need to be documented.

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 SARS-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 definitely 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 and L5F mutations of S should be paid more attention. Owing 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.

Methods

SARS-CoV-2 isolates

Complete full-length genomic sequences of SARS-CoV-2 were downloaded from 2019 Novel Coronavirus Resource (2019nCoV) in China National Center for Bioinformatics. 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, Germany, Peru, Turkey, Kazakhstan, Iran, Serbia, Thailand, Netherlands, Sri Lanka, Czech, Malaysia, India etc. Detailed information of these isolates including the GenBank accession number or biosample number is summarized in Additional file 8: Table S4.

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 [43]. 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[44]. 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 [43], 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[45]. Six methods implemented in the RDP4 were utilized. These methods are RDP [45], GENECONV[46], BootScan [47], MaxChi[48], Chimaera [49], and SiScan [50]. 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 [51]. 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[52]. These analyses were carried out using DnaSP 6.12.03[44]. 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 [43].

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 [53]. 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 ($\omega > 1$). 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[17]. HyPhy package was used to validate the result obtained by ML method[54].

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 [55].

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. Liu DX, Yuan Q, Liao Y. Coronavirus envelope protein: a small membrane protein with multiple functions. *Cell Mol Life Sci.* 2007;64(16):2043–8.
2. 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.
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.
4. McBride R, van Zyl M, Fielding BC. The coronavirus nucleocapsid is a multifunctional protein. *Viruses.* 2014;6(8):2991–3018.
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.
6. 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):265–9.
7. Z Y, Z S, J. C CW, Z W. B. Z: Analysis of variation and evolution of SARS-CoV-2 genome. *Journal of Southern Medical University.* 2020;02:152–8.
8. 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):270–3.
9. 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.
10. 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.
11. 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.
12. 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.
13. Wu Y. **Strong evolutionary convergence of receptor-binding protein spike between COVID-19 and SARS-related coronaviruses.** 2020:2020.2003.2004.975995.
14. 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.

15. **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>.
16. Mallapaty S. How sewage could reveal true scale of coronavirus outbreak. *Nature*. 2020;580(7802):176–7.
17. Yang Z, Wong WS, Nielsen R. Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular biology evolution*. 2005;22(4):1107–18.
18. Pond SLK, Muse SV: **HyPhy: Hypothesis Testing Using Phylogenies**: Springer New York; 2005.
19. Pond SL, Scheffler K, Gravenor MB, Poon AF, Frost SD. Evolutionary fingerprinting of genes. *Molecular biology evolution*. 2010;27(3):520–36.
20. Kosakovsky Pond SL, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular biology evolution*. 2005;22(5):1208–22.
21. 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).
22. Clement M, Snell Q, Walker P, Posada D, Crandall KJP, Distributed Processing Symposium IP: **TCS: Estimating gene genealogies.** 2002, 2:184.
23. Bubac CM, Spellman GMJTAOA. **How connectivity shapes genetic structure during range expansion: Insights from the Virginia's Warbler.** 2016(2):2.
24. 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):1260–3.
25. 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.
26. 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):1026–39 e1015.
27. Walls AC, Tortorici MA, Snijder J, Xiong X, Bosch BJ, Rey FA, Veasler D. Tectonic conformational changes of a coronavirus spike glycoprotein promote membrane fusion. *Proc Natl Acad Sci USA*. 2017;114(42):11157–62.
28. 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.
29. 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.
30. Phan T. Genetic diversity and evolution of SARS-CoV-2. *Infection genetics evolution: journal of molecular epidemiology evolutionary genetics in infectious diseases*. 2020;81:104260.
31. 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.
32. Voss D, Kern A, Traggiai E, Eickmann M, Stadler K, Lanzavecchia A, Becker S. Characterization of severe acute respiratory syndrome coronavirus membrane protein. *FEBS Lett*. 2006;580(3):968–73.
33. Tseng YT, Wang SM, Huang KJ, Lee AI, Chiang CC, Wang CT. Self-assembly of severe acute respiratory syndrome coronavirus membrane protein. *J Biol Chem*. 2010;285(17):12862–72.
34. 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.

35. 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.
36. Hughes AL, Hughes MA. More effective purifying selection on RNA viruses than in DNA viruses. *Gene*. 2007;404(1–2):117–25.
37. 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 e286.
38. 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**.
39. Meyerson NR, Sawyer SL. Two-stepping through time: mammals and viruses. *Trends in microbiology*. 2011;19(6):286–94.
40. Bestle D, Heindl MR, Limburg H, van TVL, Pilgram, Moulton O, Stein H, Harges DA, Eickmann K, Dolnik M. 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.
41. 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.
42. 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):119–29.
43. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular biology evolution*. 2018;35(6):1547–9.
44. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE. Sánchez-GAJMB, Evolution: **DnaSP 6: DNA Sequence Polymorphism Analysis of Large Datasets.** 2017, 34(12).
45. Martin DP, Murrell B, Khoosal A, Muhire B. Detecting and Analyzing Genetic Recombination Using RDP4. *Methods in molecular biology*. 2017;1525:433–60.
46. Padidam M, Sawyer S, Fauquet CM. Possible emergence of new geminiviruses by frequent recombination. *Virology*. 1999;265(2):218–25.
47. 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.
48. Smith JM. Analyzing the mosaic structure of genes. *Journal of molecular evolution*. 1992;34(2):126–9.
49. Posada D. Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Molecular biology evolution*. 2002;19(5):708–17.
50. Gibbs MJ, Armstrong JS, Gibbs AJ. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics*. 2000;16(7):573–82.
51. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. *Molecular biology evolution*. 2006;23(2):254–67.
52. 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.
53. Gao F, Chen C, Arab DA, Du Z, He Y, Ho SYWJE. Evolution: **EasyCodeML: A visual tool for analysis of selection using CodeML.** 2019.

Figure 2

Reticulate network trees of E (a), M (b), N (c) and S (d) 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 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

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 .

a **b** **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 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 6

Structure of the S protein 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. 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 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.

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](#)