RESEARCH ARTICLE
MICROBIOLOGY

On the origin and continuing evolution of SARS-CoV-2

Xiaolu Tang[1,7], Changcheng Wu[1,7], Xiang Li[2,3,4,7], Yuhe Song[2,5,7], Xinmin Yao[1], Xinkai Wu[1], Yuange Duan[1], Hong Zhang[1], Yirong Wang[1], Zhaohui Qian[6], Jie Cui[2,3,*], and Jian Lu[1,*]

1. State Key Laboratory of Protein and Plant Gene Research, Center for Bioinformatics, School of Life Sciences, Peking University, Beijing, 100871, China
2. CAS Key Laboratory of Molecular Virology & Immunology, Institut Pasteur of Shanghai, Chinese Academy of Sciences, China
3. Center for Biosafety Mega-Science, Chinese Academy of Sciences, China
4. University of Chinese Academy of Sciences, China
5. School of Life Sciences, Shanghai University, China
6. NHC Key Laboratory of Systems Biology of Pathogens, Institute of Pathogen Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing
7. These authors contributed equally to this work.

*Corresponding authors:
Jian Lu, Email: LUJ@pku.edu.cn
Jie Cui, Email: jcui@ips.ac.cn

ABSTRACT

The SARS-CoV-2 epidemic started in late December 2019 in Wuhan, China, and has since impacted a large portion of China and raised major global concern. Herein, we investigated the extent of molecular divergence between SARS-CoV-2 and other related coronaviruses. Although we found only 4% variability in genomic nucleotides between SARS-CoV-2 and a bat SARS-related coronavirus (SARSr-CoV; RaTG13), the difference at neutral sites was 17%, suggesting the divergence between the two viruses is much larger than previously estimated. Our results suggest that the development of new variations in functional sites in the receptor-binding domain (RBD) of the spike seen in SARS-CoV-2 and viruses from pangolin SARSr-CoVs are likely caused by mutations and natural selection besides recombination. Population genetic analyses of 103 SARS-CoV-2 genomes indicated that these viruses evolved into two major types (designated L and S), that are well defined by two different SNPs that show nearly complete linkage across the viral strains sequenced to date. Although the L type (~70%) is more prevalent than the S type (~30%), the S type was found to be the ancestral version. Whereas the L type was more prevalent in the early stages of the outbreak in Wuhan, the frequency of the L type decreased after early January 2020. Human intervention may have placed more severe selective pressure on the L type, which might be more aggressive and spread more quickly. On the other hand, the S type, which is evolutionarily older and less aggressive, might have increased in relative frequency due to relatively weaker selective pressure. These findings strongly support an urgent need for further immediate, comprehensive studies that combine genomic data, epidemiological data, and chart records of the clinical symptoms of patients with coronavirus disease 2019 (COVID-19).

Keywords: SARS-CoV-2, virus, molecular evolution, population genetics

Received: 25-Feb-2020; Revised: 28-Feb-2020; Accepted: 29-Feb-2020.

INTRODUCTION

The coronavirus disease 2019 (COVID-19) epidemic started in late December 2019 in Wuhan, the capital of Central China's Hubei Province. Since then, it has rapidly spread across China and in other countries, raising major global concerns. The etiological agent is a novel coronavirus, SARS-CoV-2, named for the similarity of its symptoms to those induced by the severe acute respiratory syndrome. As of February 28, 2020, 78,959 cases of SARS-CoV-2 infection have been confirmed in China, with 2,791 deaths. Worryingly, there have also been more than 3,664 confirmed cases outside of China in 46 countries and areas (https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/), raising significant doubts about the likelihood of successful containment. Further, the genomic sequences of SARS-CoV-2 viruses isolated from a number of patients share sequence identity higher than 99.9%, suggesting a very recent host shift into humans [1-3].

Coronaviruses are naturally hosted and evolutionarily shaped by bats [4, 5]. Indeed, it has been postulated that most of the coronaviruses in humans are derived from the bat reservoir [6, 7]. Unsurprisingly, several teams have recently confirmed the genetic similarity between SARS-CoV-2 and a bat betacoronavirus of the sub-genus *Sarbecovirus* [8-13]. The whole-genome sequence identity of the novel virus has 96.2% similarity to a bat SARS-related coronavirus (SARSr-CoV; RaTG13) collected in Yunnan province, China [2, 14], but is not very similar to the genomes of SARS-CoV (about 79%) or MERS-CoV (about 50%) [1, 15]. It has also been confirmed that the SARS-CoV-2 uses the same receptor, the angiotensin converting enzyme II (ACE2), as the SARS-CoV [11]. Although the specific route of transmission from natural reservoirs to humans remains unclear [5, 13], several studies have shown that pangolins may have provided a partial *spike* gene to SARS-CoV-2; the critical functional sites in the spike protein of SAR-CoV-2 are nearly identical to one identified in a virus isolated from a pangolin [16-18].

Despite these recent discoveries, several fundamental issues related to the evolutionary patterns and driving forces behind this outbreak of SARS-CoV-2 remain unexplored [19]. Herein, we investigated the extent of molecular divergence between SARS-CoV-2 and other related coronaviruses and carried out population genetic analyses of 103 sequenced genomes of SARS-CoV-2. This work provides new insights into the factors driving the evolution of SARS-CoV-2 and its pattern of spread through the human population.

RESULTS

Molecular phylogeny and divergence between SARS-CoV-2 and related coronaviruses.

For each annotated ORF in the reference genome of SARS-CoV-2 (NC_045512), we extracted the orthologous sequences in human SARS-CoV, four bat

SARS-related coronaviruses (SARSr-CoV: RaTG13, ZXC21, ZC45, and BM48-31), one Pangolin SARSr-CoV from Guangdong (GD) [17], and six Pangolin SARSr-CoV genomes from Guangxi (GX) [18] (Table S1). We aligned the coding sequences (CDSs) based on the protein alignments (see Materials and Methods). Most ORFs annotated from SARS-CoV-2 were found to be conserved in other viruses, except for *ORF8* and *ORF10* (Table 1). The protein sequence of SARS-CoV-2 *ORF8* shared very low similarity with sequences in SARS-CoV and BM48-31, and *ORF10* had a premature stop codon in both SARS-CoV and BM48-31 (Fig. S1). A one-base deletion caused a frame-shift mutation in *ORF10* of ZXC21 (Fig. S1).

To investigate the phylogenetic relationships between these viruses at the genomic scale, we concatenated coding regions (CDSs) of the nine conserved ORFs (*orf1ab, E, M, N, S, ORF3a, ORF6, ORF7a,* and *ORF7b*) and reconstructed the phylogenetic tree using the synonymous sites (Fig. 1A). We also used CODEML in the PAML [20] to infer the ancestral sequence of each node and calculated the dN (nonsynonymous substitutions per nonsynonymous site), dS (synonymous substitutions per synonymous site), and dN/dS (ω) values for each branch (Fig. 1A). In parallel, we also calculated the pairwise dN, dS, and ω values between SARS-CoV-2 and another virus (Table 1).

The genome-wide phylogenetic tree indicated that SARS-CoV-2 was closest to RaTG13, followed by GD Pangolin SARSr-CoV, then by GX Pangolin SARSr-CoVs, then by ZC45 and ZXC21, then by human SARS-CoV, and finally by BM48-31(Fig. 1A). Notably, we found that the nucleotide divergence at synonymous sites between SARS-CoV-2 and other viruses was much higher than previously anticipated. For example, although the overall genomic nucleotides overall differ ~4% between SARS-CoV-2 and RaTG13, the genomic average dS was 0.17, which means the divergence at the neutral sites is 17% between these two viruses (Table 1). This is because the nonsynonymous sites are usually under stronger negative selection than synonymous sites, and calculating sequence differences without separating these two classes of sites may underestimate the extent of molecular divergence by several folds.

Notably, the dS value varied considerably across genes in SARS-CoV-2 and the other viruses analyzed. In particular, the *spike* gene (*S*) consistently exhibited larger dS values than other genes (Table 1). This pattern became clear when we calculated the dS value for each branch in Fig. 1A for the *spike* gene versus the concatenated sequences of the remaining genes (Fig. S2). In each branch, the dS of *spike* was 2.22 ± 1.35 (mean ± SD) times as large as that of the other genes. This extremely elevated dS value of *spike* could be caused either by a high mutation rate or by natural selection that favors synonymous substitutions. Synonymous

substitutions may serve as another layer of genetic regulation, guiding the efficiency of mRNA translation by changing codon usage [21]. If positive selection is the driving force for the higher synonymous substation rate seen in *spike*, we expect the frequency of optimal codons (FOP) of *spike* to be different from that of other genes. However, our codon usage bias analysis (Table S2) suggests the FOP of *spike* was only slightly higher than that of the genomic average (0.717 versus 0.698, see Materials and Methods). Thus, we believe that the elevated synonymous substitution rate measured in *spike* is more likely caused by higher mutational rates; however, the underlying molecular mechanism remains unclear.

Both SARS-CoV and SARS-CoV-2 bind to ACE2 through the RBD of spike protein in order to initiate membrane fusion and enter human cells [1, 2, 22-26]. Five out of the six critical amino acid (AA) residues in RBD were different between SARS-CoV-2 and SARS-CoV (Fig. 1B), and a 3D structural analysis indicated that the spike of SARS-CoV-2 has a higher binding affinity to ACE2 than SARS-CoV [23]. Intriguingly, these same six critical AAs are identical between GD Pangolin-CoV and SARS-CoV-2 [16]. In contrast, although the genomes of SARS-CoV-2 and RaTG13 are more similar overall, only one out of the six functional sites are identical between the two viruses (Fig. 1B). It has been proposed that the SARS-CoV-2 RBD region of the spike protein might have resulted from recent recombination events in pangolins [16-18]. Although several ancient recombination events have been described in *spike* [27, 28], it also seems likely that the identical functional sites in SARS-CoV-2 and GD Pangolin-CoV may actually the result of coincidental convergent evolution [18].

If the functional AA residues in the SARS-CoV-2 RBD region were acquired from GD Pangolin-CoV in a very recent recombination event, we would expect the nucleotide sequences of this region to be nearly identical between the two viruses. However, for the CDS sequences that span five critical AA sites in the SARS-CoV-2 spike (ranging from codon 484 to 507, covering five adjacent functional sites: F486, Q493, S494, N501, and Y505; Fig. S3), we estimated dS = 0.411, dN = 0.019, and ω= 0.046 between SARS-CoV-2 and GD Pangolin-CoV. By assuming the synonymous substitution rate (u) of 1.67-4.67 x 10^{-3}/site/year, as estimated in SARS-CoV [29], the recombination/introgression, if it occurred at all, would be estimated to happen approximately 19.8-55.4 years ago. Here, the formula $t = dS/(u \times 2 \times 2.22)$ was used to calculate divergence time; note that the increased mutational rate of *spike* was considered for this calculation. Thus, it seems very unlikely that SARS-CoV-2 originated from the GD Pangolin-CoV due to a very recent recombination event. Alternatively, it seems more likely that a high mutation rate in *spike*, coupled with strong natural selection, has shaped the identical functional AA residues between these two viruses, as proposed previously [18]. Although these sites are maintained in SARS-CoV-2 and GD

Pangolin-CoV, mutations may have changed the residues in the RaTG13 lineage after it diverged from SARS-CoV-2 (the blue arrow in Fig. 1A). In summary, it seems that the shared identity of critical AA sites between SARS-CoV-2 and GD Pangolin-CoV might be due to random mutations coupled with natural selection, and not necessarily recombination.

Selective constraints and positive selection during the evolution of SARS-CoV-2 and related coronaviruses

The genome-wide ω value between SARS-CoV-2 and other viruses ranged from 0.044 to 0.124 (Table 1), indicative of strong negative selection on the nonsynonymous sites. In other words, 87.6% to 95.6% of the nonsynonymous mutations were removed by negative selection during viral evolution. To determine the extent of positive selection, we concatenated the CDS sequences of 9 conserved ORFs in all the viruses in Fig. 1A and fitted the M7 (beta: neutral and negative selection) and M8 (beta + ω>1:neutral, negative selection, and positive selection) model using CODEML (Materials and Methods). The M8 model (*lnL* = -104,813.732, *np* =18) was a significantly better fit than the M7 (*lnL* = -105,063.284, *np* = 16) model ($P < 10^{-10}$), suggesting that some AA substitutions were favored by positive Darwinian selection (but not necessarily in the SARS-CoV-2 lineage).Under the M8 model, 98.48% (p_0) of the nonsynonymous substitutions were estimated under neutral evolution or purifying selection ($0 \leqslant \omega \leqslant 1$), and 1.52% ($p_1$) of the nonsynonymous substitutions were under positive selection ($\omega = 1.50$). A Bayes Empirical Bayes (BEB) analysis suggested that 10 AA sites showed strong signals of positive selection, and, interestingly, three of those were located in the RBD of spike, including at one critical site (Fig. 1C and Fig. S4). Thus, although these coronaviruses were generally under very strong negative selection, positive selection was also responsible for the evolution of protein sequences. The putatively positively-selected sites might serve as candidates for further functional studies.

Mutations in 103 SARS-CoV-2 genomes

We downloaded 103 publicly available SARS-CoV-2 genomes, aligned the sequences, and identified the genetic variants. For ease of visualization, we marked each virus strain based on the location and date the virus was isolated with the format of "Location_Date" throughout this study (see Table S1 for details; Each ID did not contain information of the patient's race or ethnicity). Although SARS-CoV-2 is an RNA virus, for simplicity, we presented our results based on DNA sequencing results throughout this study (*i.e.*, the nucleotide T (thymine) means U (uracil) in SARS-CoV-2). For each variant, the ancestral state was inferred based on the genome and CDS alignments of SARS-CoV-2 (NC_045512), RaTG13, and GD Pangolin-CoV (Materials and Methods). In total, we identified mutations in 149 sites across the 103 sequenced strains. Ancestral states for 43 synonymous, 83 non-synonymous, and two stop-gain mutations were unambiguously inferred. The frequency spectra of synonymous and nonsynonymous mutations are shown in Fig. 2.

Most derived mutations were singletons (67.4% (29/43) of synonymous mutations and 84.3% (70/83) of nonsynonymous mutations), indicating either a recent origin [30] or population growth [31]. In general, the derived alleles of synonymous mutations were significantly skewed towards higher frequencies than those of nonsynonymous ones ($P < 0.01$, Wilcoxon rank-sum test; Fig. 2), suggesting the nonsynonymous mutations tended to be selected against. However, 16.3% (7 out of 43) synonymous mutations, and one nonsynonymous (ORF8 (L84S, 28,144)) mutation had a derived frequency of ≥ 70% across the SARS-CoV2 strains. The nonsynonymous mutations that had derived alleles in at least two SARS-CoV-2 strains affected six proteins: orf1ab (A117T, I1607V, L3606F, I6075T), S (H49Y, V367F), ORF3a (G251V), ORF7a (P34S), ORF8 (V62L, S84L), and N (S194L, S202N, P344S).

Two major types of SARS-CoV-2 are defined by two SNPs that show complete linkage

To detect the possible recombination among SARS-CoV2 viruses, we used Haploview [32] to analyze and visualize the patterns of linkage disequilibrium (LD) between variants with minor alleles in at least two SARS-CoV-2 strains (Fig. 3A). Since most mutations were at very low frequencies, it is not surprising that many pairs had a very low r^2 or LOD value (Fig. 3B-C). Consistent with another recent report [31], we did not find evidence of recombination between the SARS-CoV2 strains.

However, we found that SNPs at location 8,782 (*orf1ab*: T8517C, synonymous) and 28,144 (*ORF8*: C251T, S84L) showed significant linkage, with an r^2 value of 0.954 (Fig. 3B, red) and a LOD value of 50.13 (Fig. 3C, red). Among the 103 SARS-CoV-2 virus strains, 101 of them exhibited complete linkage between the two SNPs: 72 strains exhibited a "CT" haplotype (defined as "L" type because T28,144 is in the codon of Leucine) and 29 strains exhibited a "TC" haplotype (defined as "S" type because C28,144 is in the codon of Serine) at these two sites. Thus, we categorized the SARS-CoV-2 viruses into two major types, with L being the major type (~70%) and S being the minor type (~30%).

The evolutionary history of L and S types of SARS-CoV-2

Although we defined the L and S types based on two tightly linked SNPs, strikingly, the separation between the L (blue) and S (red) types was maintained when we reconstructed the haplotype networks using all the SNPs in the SARS-CoV-2 genomes (Fig. 4A; the number of mutations between two neighboring haplotypes was inferred parsimoniously). This analysis further supports the idea that the two linked SNPs at sites 8,782 and 28,144 adequately define the L and S types of SARS-CoV-2.

To determine whether L or S type is ancestral, we examined the genomic alignments of SARS-CoV-2 and other highly related viruses. Strikingly, nucleotides of the S type at sites 8,782 and 28,144 were identical to the orthologous sites in the most closely related viruses (Fig. 4B). Remarkably, both sites were highly conserved in other viruses as well. Hence, although the L type (~70%) was more prevalent than the S type (~30%) in the SARS-CoV-2 viruses we examined, the S type is actually the ancestral version of SARS-CoV-2.

To further examine the relationship among the strains in the L and S types, we reconstructed a phylogenetic tree of all the 103 SARS-CoV-2 viruses based on their whole-genome sequences. Our phylogenetic tree also clearly shows the separation of the two types (Fig. 5). Viruses of the L type (blue) first clustered together, and likewise, viruses of the S type (red) were also more closely related to each other. Therefore, our whole-genome comparisons further confirm the separation of the L and S types.

Thus far, we found that, although the L type is derived from the S type, L (~70%) is more prevalent than S (~30%) among the sequenced SARS-CoV-2 genomes we examined. This pattern suggests that L has a higher transmission rate than the S type. Furthermore, our mutational load analysis indicated that the L type had accumulated a significantly higher number of derived mutations than S type ($P < 0.0001$, Wilcoxon rank-sum test; Fig. S5). We propose that, although the L type newly evolved from the ancient S type, it transmits faster or replicates faster in human populations, causing it to accumulate more mutations than the S type. Thus, our results suggest the L might be more aggressive than the S type due to the potentially higher transmission and/or replication rates.

To test whether the two types of SARS-CoV-2 had differences in temporal and spatial distributions, we stratified the viruses based on the locations and dates they were isolated (Table S1). Among the 27 viruses isolated from Wuhan, 26 (96.3%) were L type, and only 1 (3.7%) was S type. However, among the other 73 viruses isolated outside Wuhan, 45 (61.6%) were L type, and 28 (38.4%) were S type. This comparison suggests that the L type is significantly more prevalent in Wuhan than in other places ($P = 0.0004$, Fisher's exact test, Fig. 6 and Table S3). All of the 26 samples isolated before January 7, 2020, were from Wuhan, and among the 74 samples collected from January 7, 2020, only one was from Wuhan, 33 were from other places in China, and 40 were from patients outside China. Thus, it is not surprising that the L type was significantly more prevalent before January 7, 2020 (96.2%, 25 L and 1 S) than after January 7, 2020 (62.2%, 46 L and 28 S) ($P = 0.0008$, Fisher's exact test, Fig. 6 and Table S3).

If the L type is more aggressive than the S type, why did the relative frequency of the L type decrease compared to the S type in other places after the initial breakout in Wuhan? One possible explanation is that, since January 2020, the Chinese central and local governments have taken rapid and comprehensive prevention and control measures. These human intervention efforts might have caused severe selective pressure against the L type, which might be more aggressive and spread more quickly. The S type, on the other hand, might have experienced weaker selective pressure by human intervention, leading to an increase in its relative abundance among the SARS-CoV-2 viruses. Thus, we hypothesized that the two types of SARS-CoV-2 viruses might have experienced different selective pressures due to different epidemiological features. Of note, the above analyses were based on very patchy SARS-CoV-2 genomes that were collected from different locations and time points. More comprehensive genomic data is required for further testing of our hypothesis.

Heteroplasmy of SARS-CoV-2 viruses in patients

It is currently unclear how the L type specifically evolved from the S type during the development of SARS-CoV-2. However, we found that the sequence of viruses isolated from one patient that lived in the United States on January 21 (USA_2020/01/21.a, GISAID ID: EPI_ISL_404253) had the genotype Y (C or T) at both positions 8,782 and 28,144, differing from the general trend of having either C or T. Although novel mutations could lead to this result, the most parsimonious explanation is that this patient may have been infected by both the L and S types (Fig. 7A). The sample of USA_2020/01/21.a was collected from a 63-year-old female patient living in Chicago (from GISAID). Based on the report from the United States Centers for Disease Control and Prevention (https://www.cdc.gov/media/releases/2020/p0124-second-travel-coronavirus.html), we inferred this patient returned to the United States from Wuhan on January 13, 2020. However, whether the co-existence of L and S types in this patient was due to multiple-time infections during her visit to Wuhan is currently unclear. Notably, the viruses identified from a patient in Australia on January 28, 2020 (Australia_2020/01/28.a, GISAID ID: EPI_ISL_407894) had multiple degenerate nucleotides. This sample was collected from a 44-year-old male patient in Gold Cost, Australia (from GISAID). Based on the report from the Courier Mail (January 30, 2020), we inferred this patient had the history of traveling from Wuhan to the Gold Coast before the diagnosis of infection. As shown in Fig. 7B, we inferred this patient might have been infected by at least two different strains of SARS-CoV-2 (Fig. 7B).

To further investigate the heteroplasmy of SARS-CoV-2 viruses in patients, we searched 12 deep-sequencing libraries of SARS-CoV-2 genomes that were deposited in the Sequence Read Archive (SRA) (Table S4, Materials and Methods). We found 17 genomic sites that showed evidence of heteroplasmy of SARS-CoV-2 virus in five patients, but we did not find

any other instances of the co-existence of L and S types in any patient (Table 2). These findings evince the developing complexity of the evolution of SARS-CoV-2 infections. Further studies investigating how the different alleles of SARS-CoV-2 viruses compete with each other will be of significant value.

DISCUSSION

In this study, we investigated the patterns of molecular divergence between SARS-CoV-2 and other related coronaviruses. Although the genomic analyses suggested that SARS-CoV-2 was closest to RaTG13, their difference at neutral sites was much higher than previously realized. Our results provide novel insights into tracing the intermediate natural host of SARS-CoV-2. With population genetic analyses of 103 genomes of SARS-CoV-2, we found that SARS-CoV-2 viruses evolved into two major types (L and S types), and the two types were well defined by just two SNPs that show nearly complete linkage across SARS-CoV-2 strains. Although the L type (~70%) was more prevalent than the S type (~30%) in the SARS-CoV-2 viruses we examined, our evolutionary analyses suggested the S type was most likely the more ancient version of SARS-CoV-2. Our results also support the idea that the L type is more aggressive than the S type.

Since nonsynonymous sites are usually under stronger negative selection than synonymous sites, calculating sequence differences without separating these two classes of sites could lead to a potentially significant underestimate of the degree of molecular divergence. For example, although the overall nucleotides only differed by ~4% between SARS-CoV-2 and RaTG13, the genomic average dS value, which is usually a neutral proxy, was 0.17 between these two viruses (Table 1). Of note, the genome-wide dS value is 0.012 between humans and chimpanzees [33], and 0.08 between humans and rhesus macaques [34]. Thus, the neutral molecular divergence between SARS-CoV-2 and RaTG13 is 14 times larger than that between humans and chimpanzees, and twice as large as that between humans and macaques. The genomic average dS value between SARS-CoV-2 and GD Pangolin-CoV is 0.475, which is comparable to that between humans and mice (0.5) [35], and the dS value between SARS-CoV-2 and GX Pangolin-Cov is even larger (0.722). The scale of these measures suggests that we should perhaps consider the difference in the neutral evolving site rather than the difference in all nucleotide sequences when tracing the origin and natural intermediate host of SARS-CoV-2.

Our analyses of molecular evolution and population genetics suggested that some amino acid changes might be favored by natural selection during the evolution of SARS-CoV-2 and other related viruses. However, negative selection appears to be the predominant force acting on these viruses. Interestingly, the virus isolated from one patient in Shenzhen on January 13,

2020 (SZ_2020/01/13.a, GISAID ID: EPI_ISL_406592) had C at both positions 8,782 and 28,144 in the genome, belonging to neither L nor S type (Fig. 4A and 5). Notably, this strain had one stop-gain mutation in *orf1ab* and had accumulated 20 silent and 5 nonsynonymous mutations after diverging from the ancestor haplotype (Fig. 4A). Thus, it is possible that functional constraints on the genomic sequence were weakened after the disruption of *orf1ab* in this strain. Notably, on viruses isolated from a patient living in South Korean (Skorea_2020/01.a, GISAID: EPI_ISL_411929), acquired six nonsynonymous mutations that were different from the most recent common ancestor of SARS-CoV-2: orf1ab (M902I and T6891M), S (S221W), ORF3a (W128L and G251V), and E (L37H). If these changes are not due to sequencing errors, it would be interesting to test whether and how these mutations affect the transmission and pathogenesis of SARS-CoV-2.

In this work, we propose that SARS-CoV-2 can be divided into two major types (L and S types): the S type is ancestral, and the L type evolved from S type. Intriguingly, the S and L types can be clearly defined by just two tightly linked SNPs at positions 8,782 (*orf1ab*: T8517C, synonymous) and 28,144 (*ORF8*: C251T, S84L). However, it is currently unclear whether L type evolved from the S type in humans or in the intermediate hosts. It is also unclear whether the L type is more virulent than the S type. *orf1ab*, which encodes replicase/transcriptase, is required for viral genome replication and might also be important for viral pathogenesis [36]. Although the T8517C mutation in *orf1ab* does not change the protein sequence (it changes the codon AGT (Ser) to AGC (Ser)), we hypothesized this mutation might affect *orf1ab* translation since AGT is preferred while AGC is unpreferred (Table S2). ORF8 promotes the expression of ATF6, the ER unfolded protein response factor, in human cells [37]. Thus, it will be interesting to investigate the function of the S84L AA change in ORF8, as well as the combinatory effect of these two mutations in SARS-CoV-2 pathogenesis.

In summary, our analyses of 103 sequenced SARS-CoV-2 genomes suggest that the L type is more aggressive than the S type and that human interference may have shifted the relative abundance of L and S type soon after the SARS-CoV-2 outbreak. As previously noted [19], the data examined in this study are still very limited, and follow-up analyses of a larger set of data are needed to have a better understanding of the evolution and epidemiology of SARS-CoV-2. There is a strong need for further immediate, comprehensive studies that combine genomic data, epidemiological data, and chart records of the clinical symptoms of patients with SARS-CoV-2.

MATERIALS AND METHODS

Molecular evolution of SARS-CoV-2 and other related viruses

The set of 103 complete genome sequences were downloaded from GISAID (Global Initiative on Sharing All Influenza Data; https://www.gisaid.org/) with acknowledgment, GenBank (https://www.ncbi.nlm.nih.gov/genbank), and NMDC (http://nmdc.cn/#/nCoV). Sequences and annotations of the reference genome of SARS-CoV-2 (NC_045512) and other related viruses were downloaded from GenBank or GISAID (Table S1). The genomic sequences of SARS-CoV-2 were aligned using MUSCLE v3.8.31 [38].

The annotated CDSs of other viruses were downloaded from GenBank. To avoid missing annotations in other viruses, we also annotated the ORFs using CDSs annotated in SARS-CoV-2 using Exonerate (--model protein2genome:bestfit --score 5 -g y) [39]. The protein sequences of SARS-CoV-2 and other related viruses were aligned with MUSCLE v3.8.31 [38], and the codon alignments were made based on the protein alignment with RevTrans [40]. The codon alignments of the conserved ORFs were further concatenated for down-stream evolutionary analysis. The phylogenetic tree was constructed by the neighbor-joining method in MEGA-X [41] using the parameters of Kimura 2-parameter model, and only the third positions of codons were considered. YN00 from PAML v4.9a [20] was used to calculate the pairwise divergence between SARS-CoV-2 and other viruses for each individual gene or for the concatenated sequences. The free-ratio model in CODEML in the PAML [20] package was used to calculate the dN, dS, and ω values for each branch.

Positively selected amino acids

Positive selection was detected using EasyCodeML [42], a recently published wrapper of CODEML [20]. The M7 and M8 models were compared. In the M7 model, ω follows a beta distribution such that $0 \leqslant \omega \leqslant 1$, and in the M8 model, a proportion p_0 of sites have ω drawn from the beta distribution, and the remaining sites with proportion p_1 are positively selected and have $\omega_1 > 1$. The LRTs between M7 and M8 models were conducted by comparing twice the difference in log-likelihood values ($2 \ln \Delta l$) against a χ^2-distribution (df=2). The positively selected sites were identified with the Bayes Empirical Bayes (BEB) score larger than 0.95.

Haplotype network

DnaSP v6.12.03 [43] was used to generate multi-sequence aligned haplotype data, and PopART v1.7 [44] was used to draw haplotype networks based on the haplotypes generated by DnaSP. RAxML v8.2.12 [45] was used to build the maximum likelihood phylogenetic tree of 103 aligned SARS-CoV-2 genomes with theparameters “-p 1234 -m GTRCAT”.

SNP calling process

We downloaded 12 SARS-CoV-2 metagenomic sequencing libraries (Table S2), and mapped the NGS reads to the reference genome of SARS-CoV-2 (NC_045512) using BWA (0.7.17-r1188) [46] with the default parameters. SNP calling was done using bcftools mpileup (bcftools 1.9) [47].

Codon usage bias analysis

We calculated the RSCU (Relative Synonymous Codon Usage) value of each codon in the SARS-CoV-2 reference genome (NC_045512). The RSCU value for each codon was the observed frequency of this codon divided by its expected frequency under equal usage among the amino acid [48]. The codons with RSCU > 1 were defined as preferred codons, and those with RSCU < 1 were defined as unpreferred codons. The FOP (frequency of optimal codons) value of each gene was calculated as the number of preferred codons divided by the total number of preferred and unpreferred codons.

Conflict of interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors thank the researchers who generated and shared the sequencing data from GISAID (https://www.gisaid.org/) on which this research is based. We thank Dr. Chung-I Wu, Hong Wu, Hongya Gu, Liping Wei, Xuemei Lu, Weiwei Zhai, Guodong Wang, Xiaodong Su, Keping Hu, and Leiliang Zhang for suggestive comments to this study. This work was supported by grants from the National Natural Science Foundation of China (No. 91731301) to J.L. JC is supported by CAS Pioneer Hundred Talents Program.

References

1. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 2020. Epub 2020/02/03. doi: 10.1016/S0140-6736(20)30251-8. PubMed PMID: 32007145.
2. 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. doi: 10.1038/s41586-020-2012-7. PubMed PMID: 32015507.
3. Ren L-L, Wang Y-M, Wu Z-Q, Xiang Z-C, Guo L, Xu T, et al. Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. Chinese Medical Journal. 2020.
4. Cui J, Li F, Shi Z-L. Origin and evolution of pathogenic coronaviruses. Nature Reviews Microbiology. 2019;17(3):181-92. doi: 10.1038/s41579-018-0118-9.
5. Li X, Song Y, Wong G, Cui J. Bat origin of a new human coronavirus: there and back again. Science China Life Sciences. 2020. doi: 10.1007/s11427-020-1645-7.
6. Li W, Shi Z, Yu M, Ren W, Smith C, Epstein JH, et al. Bats are natural reservoirs of SARS-like coronaviruses. Science. 2005;310(5748):676-9. Epub 2005/10/01. doi: 10.1126/science.1118391. PubMed PMID: 16195424.

7. Dominguez SR, O'Shea TJ, Oko LM, Holmes KV. Detection of group 1 coronaviruses in bats in North America. Emerg Infect Dis. 2007;13(9):1295-300. Epub 2008/02/07. doi: 10.3201/eid1309.070491. PubMed PMID: 18252098; PubMed Central PMCID: PMCPMC2857301.
8. Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P, et al. Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. Cell Host Microbe. 2020. Epub 2020/02/09. doi: 10.1016/j.chom.2020.02.001. PubMed PMID: 32035028.
9. Xu X, Chen P, Wang J, Feng J, Zhou H, Li X, et al. Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. Sci China Life Sci. 2020. Epub 2020/02/06. doi: 10.1007/s11427-020-1637-5. PubMed PMID: 32009228.
10. Benvenuto D, Giovanetti M, Ciccozzi A, Spoto S, Angeletti S, Ciccozzi M. The 2019-new coronavirus epidemic: Evidence for virus evolution. J Med Virol. 2020. Epub 2020/01/30. doi: 10.1002/jmv.25688. PubMed PMID: 31994738.
11. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin. bioRxiv. 2020.
12. Chan JF, Kok KH, Zhu Z, Chu H, To KK, Yuan S, et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg Microbes Infect. 2020;9(1):221-36. Epub 2020/01/29. doi: 10.1080/22221751.2020.1719902. PubMed PMID: 31987001.
13. Wei X, Li X, Cui J. Evolutionary Perspectives on Novel Coronaviruses Identified in Pneumonia Cases in China. National Science Review. 2020.
14. Paraskevis D, Kostaki EG, Magiorkinis G, Panayiotakopoulos G, Sourvinos G, Tsiodras S. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. Infect Genet Evol. 2020;79:104212. Epub 2020/02/01. doi: 10.1016/j.meegid.2020.104212. PubMed PMID: 32004758.
15. Gralinski LE, Menachery VD. Return of the Coronavirus: 2019-nCoV. Viruses. 2020;12(2). Epub 2020/01/30. doi: 10.3390/v12020135. PubMed PMID: 31991541.
16. Wong MC, Cregeen SJJ, Ajami NJ, Petrosino JF. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. bioRxiv. 2020.
17. Xiao K, Zhai J, Feng Y, Zhou N, Zhang X, Zou J-J, et al. Isolation and Characterization of 2019-nCoV-like Coronavirus from Malayan Pangolins. bioRxiv. 2020:2020.02.17.951335. doi: 10.1101/2020.02.17.951335.
18. Lam TT-Y, Shum MH-H, Zhu H-C, Tong Y-G, Ni X-B, Liao Y-S, et al. Identification of 2019-nCoV related coronaviruses in Malayan pangolins in southern China. bioRxiv. 2020:2020.02.13.945485. doi: 10.1101/2020.02.13.945485.
19. Wu C-I, Poo M-m. Moral imperative for the immediate release of 2019-nCoV sequence data. National Science Review. 2020. doi: 10.1093/nsr/nwaa030.
20. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24(8):1586-91. Epub 2007/05/08. doi: 10.1093/molbev/msm088. PubMed PMID: 17483113.
21. Hanson G, Coller J. Codon optimality, bias and usage in translation and mRNA decay. Nature reviews Molecular cell biology. 2018;19(1):20-30. Epub 2017/10/11. doi: 10.1038/nrm.2017.91. PubMed PMID: 29018283.
22. Wan Y, Shang J, Graham R, Baric RS, Li F. Receptor recognition by novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS. J Virol. 2020. Epub 2020/01/31. doi: 10.1128/JVI.00127-20. PubMed PMID: 31996437.
23. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O, et al. Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation. bioRxiv. 2020:2020.02.11.944462. doi: 10.1101/2020.02.11.944462.
24. Ou X, Liu Y, Lei X, Li P, Mi D, Ren L, et al. Characterization of spike glycoprotein of 2019-nCoV on virus entry and its immune cross-reactivity with spike glycoprotein of SARS-CoV. 2020:10.21203/rs.2.4016/v1. doi: 10.21203/rs.2.24016/v1.
25. Qu X-X, Hao P, Song X-J, Jiang S-M, Liu Y-X, Wang P-G, et al. Identification of Two Critical Amino Acid Residues of the Severe Acute Respiratory Syndrome Coronavirus Spike Protein for Its Variation in Zoonotic Tropism Transition via a Double Substitution Strategy. Journal of Biological Chemistry. 2005;280(33):29588-95.
26. Ren W, Qu X, Li W, Han Z, Yu M, Zhou P, et al. Difference in Receptor Usage between Severe Acute Respiratory Syndrome (SARS) Coronavirus and SARS-Like Coronavirus of Bat Origin. Journal of Virology. 2008;82(4):1899. doi: 10.1128/JVI.01085-07.

27. 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. Epub 2020/02/06. doi: 10.1038/s41586-020-2008-3. PubMed PMID: 32015508.
28. Ji W, Wang W, Zhao X, Zai J, Li X. Homologous recombination within the spike glycoprotein of the newly identified coronavirus may boost cross‐species transmission from snake to human. Journal of medical virology. 2020.
29. Zhao Z, Li H, Wu X, Zhong Y, Zhang K, Zhang Y-P, et al. Moderate mutation rate in the SARS coronavirus genome and its implications. BMC Evolutionary Biology. 2004;4(1):21. doi: 10.1186/1471-2148-4-21.
30. Zhang C, Wang M. Origin time and epidemic dynamics of the 2019 novel coronavirus. bioRxiv. 2020.
31. Yu W-B, Tang G-D, Zhang L, Corlett RT. Decoding evolution and transmissions of novel pneumonia coronavirus using the whole genomic data. ChinaXiv. 2020:202002.00033. doi: 10.12074/202002.00033.
32. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21(2):263-5. Epub 2004/08/07. doi: 10.1093/bioinformatics/bth457. PubMed PMID: 15297300.
33. Waterson RH, Lander ES, Wilson RK, The Chimpanzee S, Analysis C. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature. 2005;437(7055):69-87. doi: 10.1038/nature04072.
34. Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, et al. Evolutionary and Biomedical Insights from the Rhesus Macaque Genome. Science. 2007;316(5822):222. doi: 10.1126/science.1139247.
35. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002;420(6915):520-62. Epub 2002/12/06. doi: 10.1038/nature01262. PubMed PMID: 12466850.
36. Graham RL, Sparks JS, Eckerle LD, Sims AC, Denison MR. SARS coronavirus replicase proteins in pathogenesis. Virus Res. 2008;133(1):88-100. Epub 2007/04/03. doi: 10.1016/j.virusres.2007.02.017. PubMed PMID: 17397959; PubMed Central PMCID: PMCPMC2637536.
37. Hu B, Zeng L-P, Yang X-L, Ge X-Y, Zhang W, Li B, et al. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. PLOS Pathogens. 2017;13(11):e1006698. doi: 10.1371/journal.ppat.1006698.
38. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792-7. Epub 2004/03/23. doi: 10.1093/nar/gkh340. PubMed PMID: 15034147; PubMed Central PMCID: PMCPMC390337.
39. Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6:31. doi: 10.1186/1471-2105-6-31. PubMed PMID: 15713233; PubMed Central PMCID: PMCPMC553969.
40. Wernersson R, Pedersen AG. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res. 2003;31(13):3537-9. Epub 2003/06/26. PubMed PMID: 12824361; PubMed Central PMCID: PMCPMC169015.
41. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 2018;35(6):1547-9. Epub 2018/05/04. doi: 10.1093/molbev/msy096. PubMed PMID: 29722887; PubMed Central PMCID: PMCPMC5967553.
42. Gao F, Chen C, Arab DA, Du Z, He Y, Ho SYW. EasyCodeML: A visual tool for analysis of selection using CodeML. Ecol Evol. 2019;9(7):3891-8. Epub 2019/04/25. doi: 10.1002/ece3.5015. PubMed PMID: 31015974; PubMed Central PMCID: PMCPMC6467853.
43. Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, et al. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. Mol Biol Evol. 2017;34(12):3299-302. doi: 10.1093/molbev/msx248. PubMed PMID: 29029172.
44. Leigh JW, Bryant D. popart: full-feature software for haplotype network construction. Methods in Ecology and Evolution. 2015;6(9):1110-6. doi: 10.1111/2041-210x.12410.
45. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312-3. Epub 2014/01/24. doi: 10.1093/bioinformatics/btu033. PubMed PMID: 24451623; PubMed Central PMCID: PMCPMC3998144.

46. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60. Epub 2009/05/20. doi: 10.1093/bioinformatics/btp324. PubMed PMID: 19451168; PubMed Central PMCID: PMCPMC2705234.
47. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. Epub 2009/06/10. doi: 10.1093/bioinformatics/btp352. PubMed PMID: 19505943; PubMed Central PMCID: PMCPMC2723002.
48. Sharp PM, Li WH. Codon usage in regulatory genes in Escherichia coli does not reflect selection for 'rare' codons. Nucleic Acids Res. 1986;14(19):7737-49. Epub 1986/10/10. doi: 10.1093/nar/14.19.7737. PubMed PMID: 3534792; PubMed Central PMCID: PMCPMC311793.

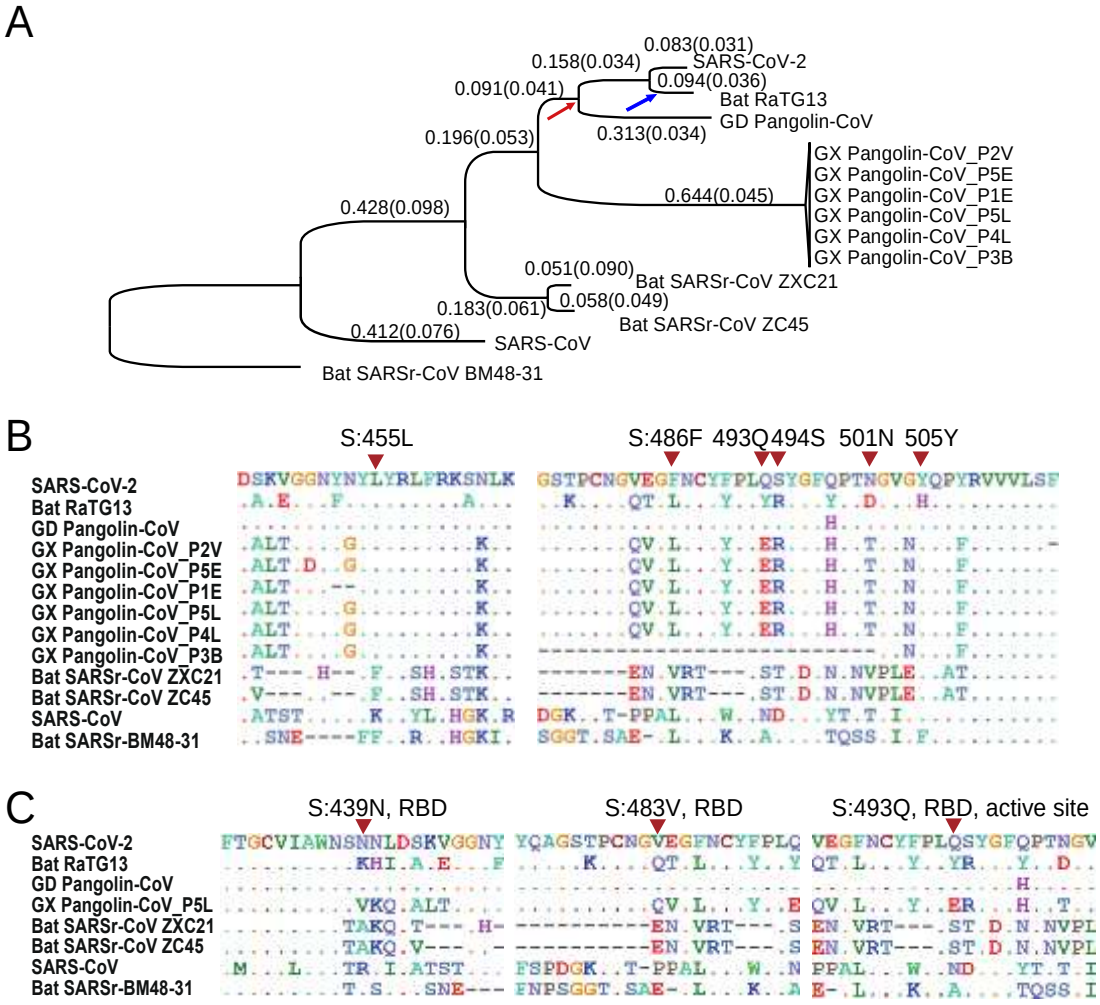

Figure 1. Molecular divergence and selective pressures during the evolution of SARS-CoV-2 and related viruses.

A. The phylogenetic tree of SARS-CoV-2 and the related Coronaviruses. The branch length (dS) is presented, and the dN/dS (ω) value is given in the parenthesis. The phylogenetic tree was reconstructed with the synonymous sites in the concatenated CDSs of nine conserved ORFs (*orf1ab*, *E*, *M*, *N*, *S*, *ORF3a*, *ORF6*, *ORF7a* and *ORF7b*).

B. Conservation of 6 critical amino acid residues in the spike (S) protein. The critical active sites are Y442, L472, N479, D480, T487, and Y491 in SARS-CoV, and they correspond to L455, F486, Q493, S494, N501, and Y505 in SARS-CoV-2 (marked with inverted triangles), respectively.

C. Three candidate positively selected sites (marked with inverted triangles) in the receptor-binding domain (RBD) of spike protein (S:439N, S:483V and S:493Q) and the surrounding 10 amino acids.

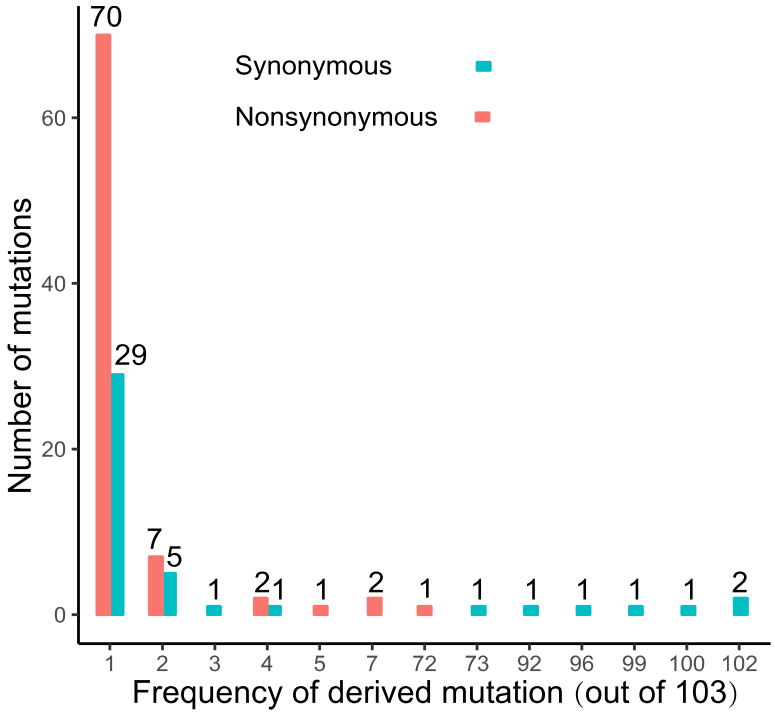

Figure 2. The frequency spectra of derived mutations in 103 SARS-CoV-2 viruses. Note the derived alleles of synonymous mutations are skewed towards higher frequencies than those of nonsynonymous mutations.

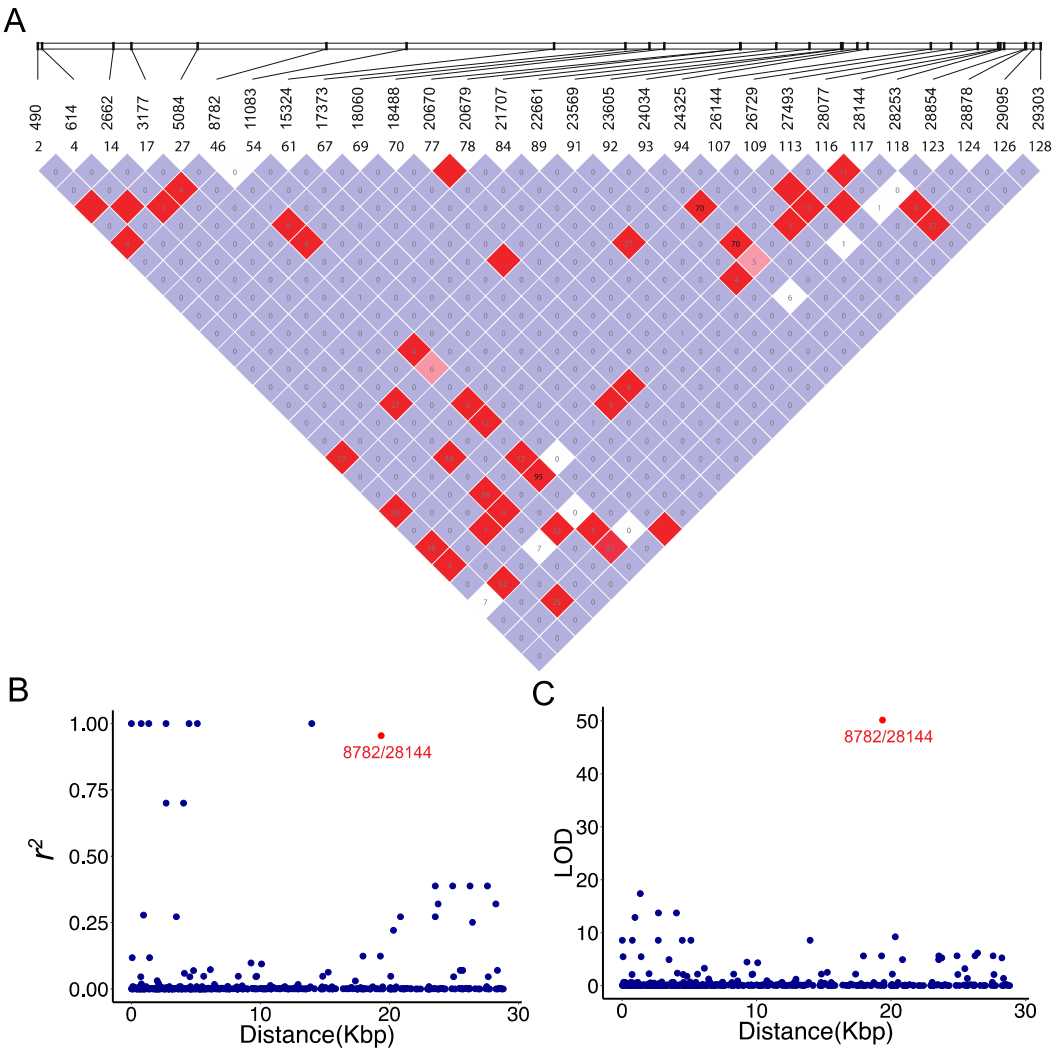

Figure 3. Linkage disequilibrium between SNPs in the SARS-CoV-2 viruses.
A. LD plot of any two SNP pairs among the 29 sites that have minor alleles in at least two strains. The number near slashes at the top of the image shows the coordinate of sites in the genome. Color in the square is given by standard (D'/LOD), and the number in square is r^2 value.
B. The r^2 of each pair of SNPs (*y*-axis) against the genomic distance between that pair (*x*-axis).
C. The LOD of each pair of SNPs (*y*-axis) against the genomic distance between that pair (*x*-axis).
Note that in both **B** and **C**, the red point represents the LD between SNPs at 8,782 and 28,144.

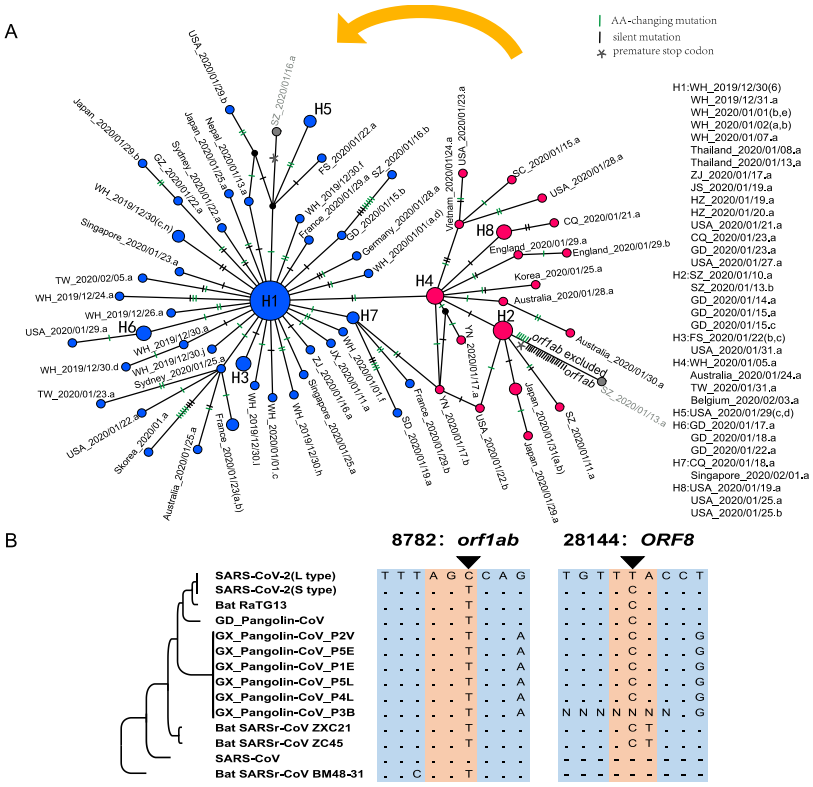

Figure 4. Haplotype analysis of SARS-CoV-2 viruses.

A. The haplotype networks of SARS-CoV-2 viruses. Blue represents the L type, and red is the S type. The orange arrow indicates that the L type evolved from the S type. Note that in this study, we marked each sample with a unique ID that starting with the geological location, followed by the date the virus was isolated (see Table S1 for details). Each ID did not contain information of the patient's race or ethnicity. ZJ, Zhejiang; YN, Yunnan; WH, Wuhan; USA, United States of America; TW, Taiwan; SZ, Shenzhen; SD, Shandong; SC, Sichuan; JX, Jiangxi; JS, Jiangsu; HZ, Hangzhou; GZ, Guangzhou; GD, Guangdong; FS, Foshan; CQ, Chongqing.

B. Evolution of the L and S types of SARS-CoV-2 viruses. Genome sequence alignments with the seven most closely related viruses indicated that the S type was most likely the ancient version of SARS-CoV-2. ".", The nucleotide sequence is identical; "-", gap.

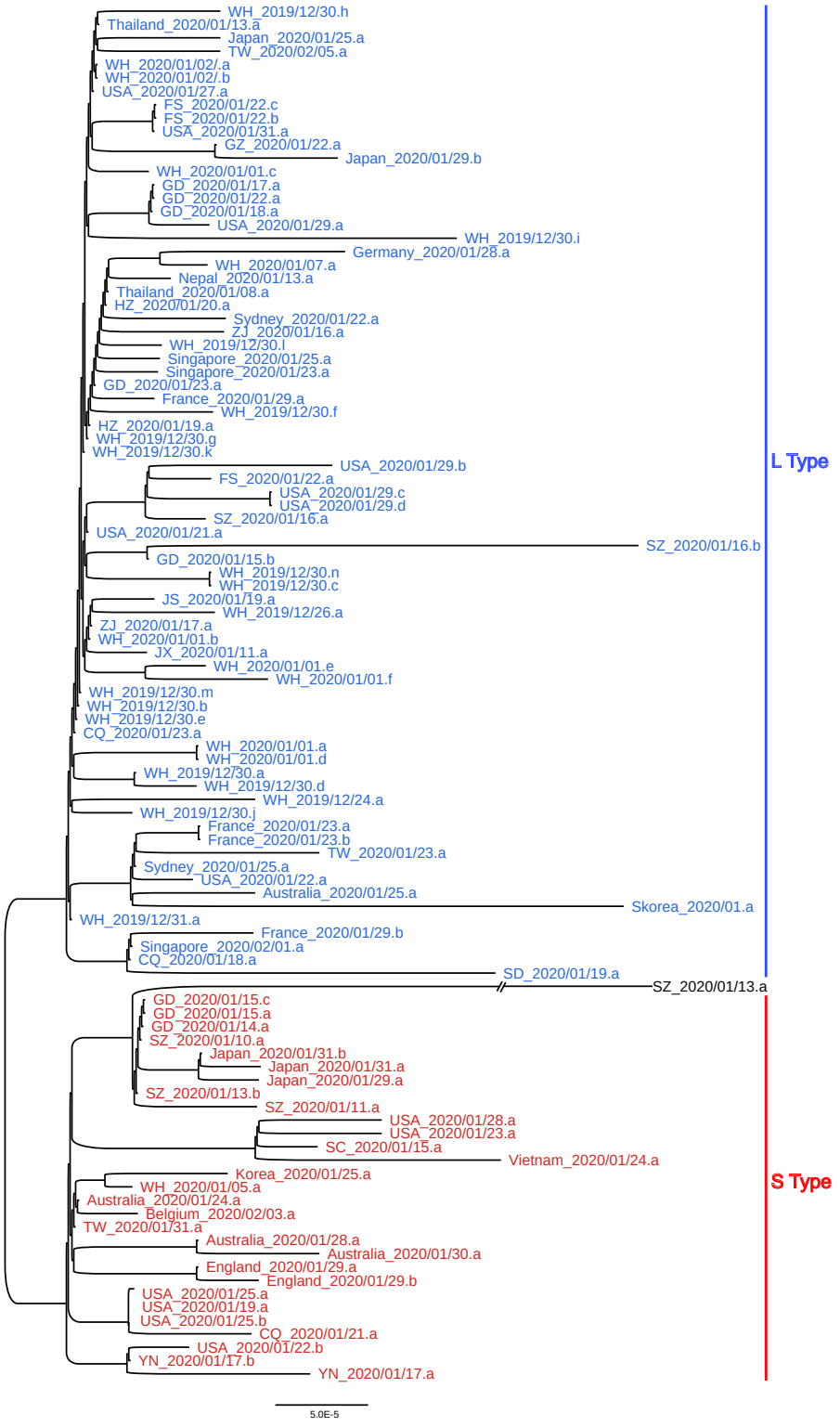

Figure 5. The unrooted phylogenetic tree of the 103 SARS-CoV-2 genomes. The ID of each sample is the same as in Fig. 4A. Note WH_2019/12/31.a represents the reference genome (NC_045512). Note SZ_2020/01/13.a had C at both positions 8,782 and 28,144 in the genome, belonging to neither L nor S type.

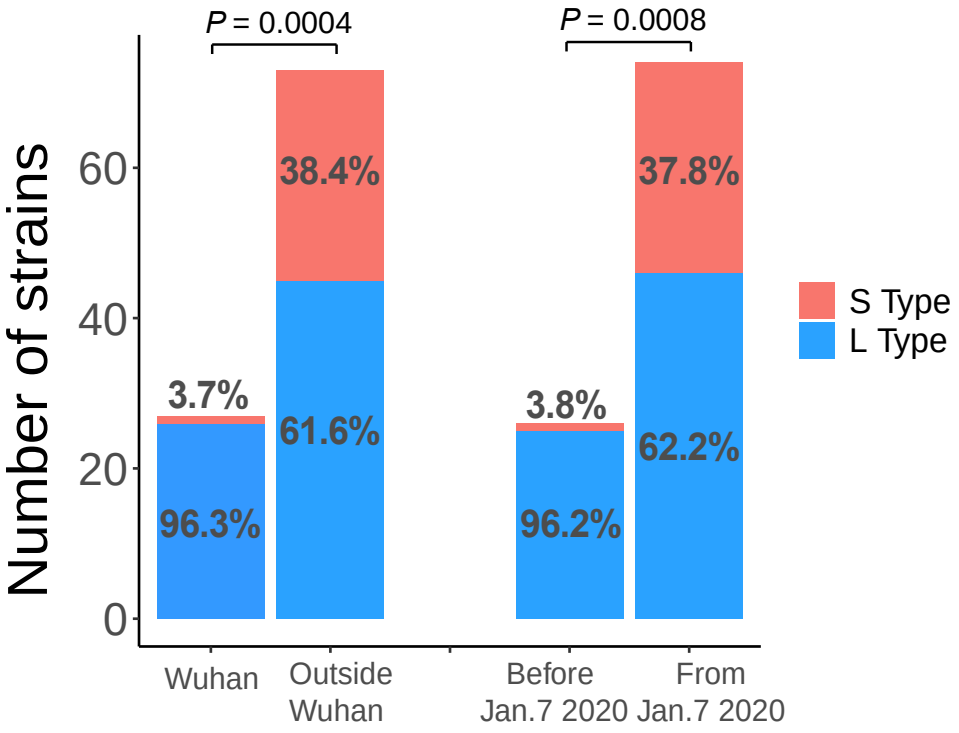

Figure 6. The two types of SARS-CoV-2 showed differences in temporal and spatial distributions.

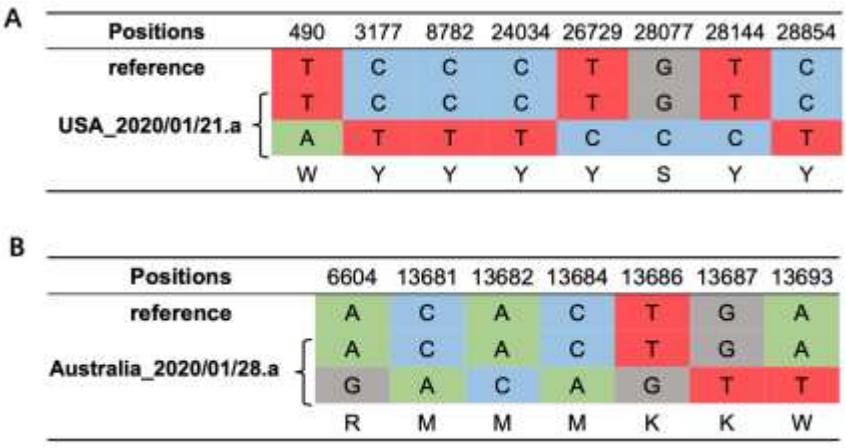

Figure 7. The heteroplasmy of SARS-CoV-2 viruses in human patients.
A. The viruses isolated from a patient that lived in the United States (USA_2020/01/21.a, GISAID ID: EPI_ISL_404253) had the genotype Y (C or T) at both 8,782 and 28,144. The most likely explanation is that this patient was infected by both the L and S types. Note the reference is L type.
B. The viruses Australia_2020/01/28.a (GISAID ID:EPI_ISL_407894) identified from a patient in Australia had multiple degenerated nucleotides, and the best explanation is that this patient was infected by at least two different strains of SARS-CoV-2 viruses.

Table 1 The molecular divergence between SARS-CoV-2 and related viruses

Gene	Aligned Length (nt)	RaTG13	GD Pangolin-CoV	GX Pangolin-CoV	SARSr-CoV ZC45	SARS-CoV	SARSr-CoV BM48-31
Genomic Average	28734	0.008/0.17 (0.044)	0.026/0.475 (0.054)	0.055/0.722 (0.076)	0.044/0.549 (0.081)	0.113/0.926 (0.122)	0.143/1.15 (0.124)
ORF10	114	0.011/0 (NA)	0.011/0 (NA)	0.072/0.044 (1.637)	0.011/0 (NA)	-	-
ORF3a	825	0.009/0.157 (0.06)	0.019/0.291 (0.065)	0.066/0.518 (0.128)	0.052/0.508 (0.102)	0.188/0.918 (0.205)	0.271/0.923 (0.294)
ORF6	183	0/0.098 (0)	0.014/0.217 (0.062)	0.038/0.491 (0.077)	0.027/0.173 (0.158)	0.191/0.913 (0.209)	0.393/1.512 (0.26)
ORF7a	363	0.011/0.177 (0.061)	0.018/0.275 (0.066)	0.073/0.477 (0.153)	0.066/0.351 (0.188)	0.088/0.697 (0.126)	0.337/1.14 (0.296)
ORF7b	129	0.01/0 (NA)	0.02/0.455 (0.043)	0.17/0.436 (0.39)	0.029/0.181 (0.162)	0.155/0.401 (0.387)	0.264/NA (NA)
ORF8	363	0.021/0.07 (0.303)	0.032/0.303 (0.105)	0.099/1.015 (0.098)	0.03/0.603 (0.05)	-	-
E	225	0/0.018 (0)	0/0.037 (0)	0.006/0.096 (0.063)	0/0.056 (0)	0.027/0.166 (0.164)	0.043/0.352 (0.121)
M	666	0.004/0.186 (0.021)	0.014/0.298 (0.046)	0.025/0.372 (0.067)	0.016/0.283 (0.055)	0.07/0.576 (0.121)	0.109/1.292 (0.085)
N	1257	0.005/0.131 (0.039)	0.011/0.144 (0.076)	0.04/0.304 (0.132)	0.036/0.333 (0.108)	0.059/0.381 (0.155)	0.102/1.197 (0.085)
orf1a	13215	0.009/0.167 (0.054)	0.026/0.488 (0.053)	0.073/0.811 (0.09)	0.026/0.405 (0.063)	0.148/1.141 (0.129)	0.174/1.199 (0.145)
orf1ab	21288	0.007/0.152 (0.044)	0.019/0.495 (0.039)	0.055/0.776 (0.071)	0.031/0.527 (0.058)	0.105/0.962 (0.109)	0.125/1.108 (0.113)
S (*spike*)	3819	0.014/0.321 (0.043)	0.075/0.69 (0.108)	0.06/0.86 (0.07)	0.138/1.063 (0.13)	0.172/1.265 (0.136)	0.217/1.518 (0.143)

For each gene, the dN and dS values between SARS-CoV-2 and another virus are given, and the dN/dS (ω) ratio is given in the parenthesis.

Table 2. The heteroplasmy of SARS-CoV-2 viruses in human patients

Accession number	Genomic position	Ref allele	Alt allele	Ref reads	Alt reads	Location_date	GISAID ID
SRR10903401	1821	G	A	52	5	WH_2020/01/02.a	EPI_ISL_406716
SRR10903401	19164	C	T	40	12	WH_2020/01/02.a	EPI_ISL_406716
SRR10903401	24323	A	C	102	67	WH_2020/01/02.a	EPI_ISL_406716
SRR10903401	26314	G	A	15	2	WH_2020/01/02.a	EPI_ISL_406716
SRR10903401	26590	T	C	10	2	WH_2020/01/02.a	EPI_ISL_406716
SRR10903402	11563	C	T	164	26	WH_2020/01/02.b	EPI_ISL_406717
SRR11092057	9064	TTAT	TT	13	2	WH_2019/12/30.e	EPI_ISL_402124
SRR11092057	17825	C	T	19	5	WH_2019/12/30.e	EPI_ISL_402124
SRR11092059	4795	C	T	10	4	WH_2019/12/30.h	EPI_ISL_402130
SRR11092059	6360	A	G	39	5	WH_2019/12/30.h	EPI_ISL_402130
SRR11092059	7042	G	A	5	3	WH_2019/12/30.h	EPI_ISL_402130
SRR11092059	12153	C	T	15	13	WH_2019/12/30.h	EPI_ISL_402130
SRR11092059	15921	G	T	19	2	WH_2019/12/30.h	EPI_ISL_402130
SRR11092059	16474	A	G	11	2	WH_2019/12/30.h	EPI_ISL_402130
SRR11092059	20344	C	T	19	2	WH_2019/12/30.h	EPI_ISL_402130
SRR11092062	565	T	C	64	23	WH_2019/12/30.e	EPI_ISL_402124
SRR11092062	17825	C	T	141	34	WH_2019/12/30.e	EPI_ISL_402124
SRR11092063	29441	C	A	6	2	WH_2019/12/30.d	EPI_ISL_402127

Addendum to

On the origin and continuing evolution of SARS-CoV-2 (DOI: 10.1093/nsr/nwaa036)

Xiaolu Tang[1], Changcheng Wu[1], Xiang Li[2,3,4], Yuhe Song[2,5], Xinmin Yao[1], Xinkai Wu[1], Yuange Duan[1], Hong Zhang[1], Yirong Wang[1], Zhaohui Qian[6], Jie Cui[2,3,*], and Jian Lu[1,*]
1. State Key Laboratory of Protein and Plant Gene Research, Center for Bioinformatics, School of Life Sciences, Peking University, Beijing, 100871, China
2. CAS Key Laboratory of Molecular Virology & Immunology, Institut Pasteur of Shanghai, Chinese Academy of Sciences, China
3. Center for Biosafety Mega-Science, Chinese Academy of Sciences, China
4. University of Chinese Academy of Sciences, China
5. School of Life Sciences, Shanghai University, China
6. NHC Key Laboratory of Systems Biology of Pathogens, Institute of Pathogen Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing

*Corresponding authors: Jian Lu, Email: LUJ@pku.edu.cn; Jie Cui, Email: jcui@ips.ac.cn

In our recent publication (https://doi.org/10.1093/nsr/nwaa036), we showed that among circulating SARS-CoV-2 (with 103 genomes analyzed) two different viral genomes co-exist. We identified them as lineages L and S. The concerned amino acid we used to define the L and S lineages is located in ORF8 (open reading frame 8), which plays a yet undefined role in the viral life cycle. Based on the finding that "L" lineage has a higher frequency than lineage S, we described the L lineage as aggressive. We now recognize that within the context of our study the term "aggressive" is misleading and should be replaced by a more precise term "a higher frequency". In short, while we have shown that the two lineages naturally co-exist, we provided no evidence supporting any epidemiological conclusion regarding the virulence or pathogenicity of SARS-CoV-2. By saying so, corrections will be made in the print version of this paper to avoid being misleading.

中文

在我们最近发表的文章（https://doi.org/10.1093/nsr/nwaa036）中，分析结果显示，103个 SARS-CoV-2 病毒基因组存在两种不同的谱系；分别称之为 "L"和"S"谱系。我们用来定义 L 和 S 谱系的氨基酸位点位于 ORF8(开放阅读框 8)基因，这个基因还没有发现具有任何已知的重要功能。基于" L"谱系的频率高于谱系 S 的发现，我们将 L 谱系描述为"aggressive"（具有侵略性）。我们现在认识到，在本研究阐述的内容中，"侵略性"一词

会具有误导性，应该用更精确的术语“更高的频率”代替。 简而言之，尽管我们已经发现这两个谱系自然并存，但我们没有提供任何证据支持关于 SARS-CoV-2 毒力或致病性的任何流行病学结论。 因此，我们将在本文的印刷版本中进行更正，以避免产生误导。