Science Bulletin 66 (2021) 2297–2311

Contents lists available at ScienceDirect

Science Bulletin

journal homepage: www.elsevier.com/locate/scib

ELSEVIER

Science Bulletin
www.scibull.com

Article

Evolutionary analysis and lineage designation of SARS-CoV-2 genomes

Xiaolu Tang [a,1], Ruochen Ying [a,1], Xinmin Yao [a,1], Guanghao Li [b,c], Changcheng Wu [a], Yiyuli Tang [b], Zhida Li [d], Bishan Kuang [d], Feng Wu [d], Changsheng Chi [d], Xiaoman Du [d], Yi Qin [d], Shenghan Gao [e], Songnian Hu [e], Juncai Ma [f], Tiangang Liu [g], Xinghuo Pang [h], Jianwei Wang [i,j], Guoping Zhao [k], Wenjie Tan [l,*], Yaping Zhang [b,*], Xuemei Lu [b,*], Jian Lu [a,*]

[a] *State Key Laboratory of Protein and Plant Gene Research, Center for Bioinformatics, School of Life Sciences, Peking University, Beijing 100871, China*
[b] *State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China*
[c] *Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*
[d] *Yuxi Rongjian Information Technology Co., Ltd., Yuxi 653100, China*
[e] *State Key Laboratory of Microbial Resources (SKLMR), The Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China*
[f] *The Microresource and Big Data Center, The Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China*
[g] *Key Laboratory of Combinatorial Biosynthesis and Drug Discovery, Ministry of Education and Wuhan University School of Pharmaceutical Sciences, Wuhan 430071, China*
[h] *Beijing Center for Disease Prevention and Control (CDC) & Research Center for Preventive Medicine of Beijing, Beijing 100013, China*
[i] *NHC Key Laboratory of Systems Biology of Pathogens and Christophe Mérieux Laboratory, Institute of Pathogen Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China*
[j] *Key Laboratory of Respiratory Disease Pathogenomics, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China*
[k] *Key Laboratory of Synthetic Biology, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China*
[l] *NHC Key Laboratory of Biosafety, National Institute for Viral Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 102206, China*

ARTICLE INFO

Article history:
Received 3 November 2020
Received in revised form 3 January 2021
Accepted 1 February 2021
Available online 6 February 2021

Keywords:
COVID-19
Evolutionary analysis
Compensatory advantageous mutation
Adaptive evolution
Lineage designation

ABSTRACT

The pandemic due to the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the etiological agent of coronavirus disease 2019 (COVID-19), has caused immense global disruption. With the rapid accumulation of SARS-CoV-2 genome sequences, however, thousands of genomic variants of SARS-CoV-2 are now publicly available. To improve the tracing of the viral genomes' evolution during the development of the pandemic, we analyzed single nucleotide variants (SNVs) in 121,618 high-quality SARS-CoV-2 genomes. We divided these viral genomes into two major lineages (L and S) based on variants at sites 8782 and 28144, and further divided the L lineage into two major sublineages (L1 and L2) using SNVs at sites 3037, 14408, and 23403. Subsequently, we categorized them into 130 sublineages (37 in S, 35 in L1, and 58 in L2) based on marker SNVs at 201 additional genomic sites. This lineage/sublineage designation system has a hierarchical structure and reflects the relatedness among the subclades of the major lineages. We also provide a companion website (www.covid19evolution.net) that allows users to visualize sublineage information and upload their own SARS-CoV-2 genomes for sublineage classification. Finally, we discussed the possible roles of compensatory mutations and natural selection during SARS-CoV-2's evolution. These efforts will improve our understanding of the temporal and spatial dynamics of SARS-CoV-2's genome evolution.

1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the etiological agent of Coronavirus Disease 2019 (COVID-19), has rapidly developed into a global pandemic. SARS-CoV-2 is a positive-sense, single-stranded RNA virus [1–4] whose genome has a similarity of ~96% to the bat coronavirus RaTG13 [4], and of ~85%–90% to coronaviruses isolated from Malayan pangolins [5–8]. With the rapid accumulation of SARS-CoV-2 genomes deposited in the Global Initiative on Sharing All Influenza Data (GISAID, https://www.epicov.org) [9,10], thousands of single nucleotide variants (SNVs) have been identified across different isolated SARS-CoV-2 strains [11–20].

Previously, Tang et al. [12] analyzed 103 SARS-CoV-2 genomes and demonstrated that SARS-CoV-2 could be divided into two major lineages (denoted as Tang's "L" and "S" hereafter) based on two SNVs at sites 8782 and 28144 that had nearly complete

[*] Corresponding authors.
E-mail addresses: tanwj@ivdc.chinacdc.cn (W. Tan), zhangyp@mail.kiz.ac.cn (Y. Zhang), xuemeilu@mail.kiz.ac.cn (X. Lu), luj@pku.edu.cn (J. Lu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.scib.2021.02.012
2095-9273/

linkage (reference genome: NC_045512 [1]). Tang's L strains exhibited a "CU" haplotype (defined as "L" lineage because U28144 is in the codon of leucine), and S strains exhibited a "UC" haplotype (defined as "S" lineage because C28144 is in the codon of serine) at these two sites. Tang et al. [12] introduced the principle of outgroup rooting using coronaviruses isolated from bats and pangolins, which distinguished ancestral S from derived L viral genomes. Forster et al. [16] published another analysis almost contemporaneously that confirmed the bat outgroup rooting and introduced a nomenclature based on the ancestral type "A" and the derived types "B" and "C". These two sites have also been used to delineate SARS-CoV-2 lineages in other studies [13,15,17,21,22]. For instance, the designation of A and B lineages by Rambaut et al. [21] was based on these SNVs at sites 8782 and 28144 (Rambaut's A corresponded to Tang's S and Forster's A, and Rambaut's B corresponded to Tang's L and Forster's B and C). Based on SNVs at these two sites and other sites, GISAID (http://gisaid.org) divided SARS-CoV-2 genomes into four major groups (S, L, V, and G), and Nextstrain (https://nextstrain.org) [22] categorized them into five major clades (19A, 19B, 20A, 20B, and 20C). It is worth noting that the outgroup rooting was not considered in the lineage/clade designation by Rambaut et al. [21], GISAID, or Nextstrain [22]. The detailed relationships among these nomenclature systems were presented in Table S1 (online).

During the development of the COVID-19 pandemic, numerous subtypes have arisen which may deserve further labels. Although the phylogenetic method is powerful for revealing the genetic relatedness of SARS-CoV-2 strains, phylogeny alone is not sufficient for tracing viral genealogies when both the ancestral and descendent samples are analyzed [23,24]. Moreover, the similarities within a large number of viral sequences have posed a significant challenge to inferring a reliable phylogeny of SARS-CoV-2 genomes [25]. To better trace the evolution of the viral genomes and facilitate comparison of patient samples taken at different stages of the pandemic, we analyzed SNVs in 121,618 high-quality SARS-CoV-2 genomes and expanded the previous L and S lineage designation in this study. Inspired by the nomenclature systems of the human Y chromosomes [26,27] and mitochondrial genomes [28], we refined our lineage designation system into 130 sublineages based on marker SNVs at 206 genomic sites, most of which exhibited strong linkage. Our hierarchical lineage/sublineage classification system, which is rooted with the outgroup and coupled with the haplotype network analysis, allows us to trace the circulation of SARS-CoV-2. A user-friendly website (www.covid19evolution.net) that enables easy, detailed visualization of the global distribution of SARS-CoV-2 lineages and sublineages was constructed. Finally, we discussed the possible roles of compensatory mutations and natural selection during SARS-CoV-2 evolution.

2. Materials and methods

2.1. Genome sequence processing

We downloaded 217,305 SARS-CoV-2 genomes from the GISAID database (http://gisaid.org, as of December 2, 2020; the detailed information for the authors and originating and submitting laboratories of the sequences were acknowledged in Table S2 online). We preserved the genome that was longer than 29,000 nucleotides (after removing Ns) and aligned these 202,679 genome sequences to the reference sequence (Wuhan-Hu-1, GenBank: NC_045512, GISAID: EPI_ISL_402125) using MAFFT v7.453 (--auto) [29]. We used snp-sites (-v) [30] to identify SNVs and BCFtools v1.8 (merge --force-samples -O v) [31] to merge the vcf files. To minimize the impact of sequencing errors, nucleotides at the 5′ end (sites 1–220) and 3′ end (sites 29675–29903) relative to the reference genome were masked from further analysis. To control the potential influence of sampling bias, we used USEARCH (v11, 64-bit) [32] to group the genome sequences and obtained 158,151 unique clusters under the threshold of 99.99% sequence identity. We selected one representative genome with the least ambiguous nucleotides (gaps and degenerate nucleotides) in each cluster. To further reduce the possible impact of sequencing errors, we identified the SNV sites that had MAF $\geq$ 0.1% in the genomic regions spanning coding regions (CDSs, 265−29674) and required each genome to have $\leq$5 ambiguous nucleotides and $\leq$50 SNVs in these regions.

2.2. Constructing the phylogenetic tree

MAFFT was used to align the 121,618 genomes after trimming part of the 5′ end (sites 1–220) and the entire 3′ end (sites 29675–29903; relative to the reference genome). To reduce computational time, we further grouped the 121,618 high-quality genomes with a sequence identity cutoff of 99.9% using USEARCH, and obtained 10,061 non-redundant genomes for tree reconstruction. The genome sequence of bat coronavirus RaTG13 (GenBank accession number: MN996532), and GD Pangolin-CoV (the SARS-CoV-2-related viruses in Malayan pangolin samples obtained by anti-smuggling operations by the Guangdong (GD) customs [7]; merged from GISAID: EPI_ISL_410544 and Genome Warehouse: GWHABKW00000000 as previously described [12]) was sequentially added as outgroups by MAFFT (--auto --add). IQ-TREE v 2.1.2 (-m GTR + G-B 1000) [33] was used to construct the maximum likelihood phylogenetic tree. Interactive Tree Of Life (iTOL) v 5 (https://itol.embl.de) was used to visualize the tree. The L and S lineages were defined based on the SNVs at sites 8782 and 28144, as previously described [12]. The L lineage was further divided into L1 and L2 major sublineages by three tightly linked genomic variants (C3037U, C14408U, and A23403G) considering the topology of the phylogenetic tree.

2.3. Linkage disequilibrium (LD) analysis

Haploview [34] was employed to analyze and visualize the LD patterns. Occasionally, an LD pair in a certain sublineage was not detected in the global population analysis because the variants had very low frequencies and were neglected by Haploview, or the LD detection was interfered by recurrent mutations. Therefore, to recover the LD pairs significantly linked in a sublineage but failed to be detected in the global analysis, besides the global viral population, we also analyzed the LD patterns between SNVs in the S, L1, and L2 clades, respectively. In case one site has multiple variants, only the reference allele and the most abundant alternative allele were considered in the LD analysis. The vcftools v 0.1.15 (--plink) [35] and plink v 1.90b3.46 (--recode HV --snps-only just-acgt) [36] were used to transform vcf file to linkage format in the LD analysis.

2.4. Polarizing mutations in SARS-CoV-2

We first inferred the ancestral states of the 206 marker SNV sites in SARS-CoV-2 using the maximum parsimony method based on the multiple sequence alignment results of SARS-CoV-2 and coronaviruses in bats and pangolins, which was recently evaluated with molecular evolution simulations [37]. In addition, the 44-way whole-genome sequence alignments of SARS-CoV-2 and bat coronaviruses in UCSC Genome Browser (http://genome.ucsc.edu/covid19.html) were also considered in the ancestral inference. Occasionally, the ancestral state of an SNV site could not be unambiguously inferred based on the nucleotides in the outgroups. In such cases, ancestral inference in other sites that exhibited strong LD with the site of interest was considered.

Overall, among the 206 marker SNV sites, except for sites 8782 and 28144 (the “UC” haplotype was ancestral as for sites 8782 and 28144), the nucleotides of the reference genome (NC_045512) were inferred to be ancestral at the other 204 sites (see Fig. S1 online for details). Of note, there were two variants on site 29095 (reference: C, alternative: U). Although the U29095 variant was observed in the orthologous sites of many SARS-CoV-2 related coronaviruses, its frequency was very low in both S (1.3%) and L (0.7%) lineages. Hence, we inferred the reference allele C29095 to be the ancestral one, and recurrent mutations (C→U) occurred in the S and L lineage independently.

2.5. Haplotype network analysis

For each of the 130 sublineages, the major haplotype sequence was inferred for the 206 marker SNV sites. The nucleotides in the 206 orthologous sites of RaTG13 were used to root the haplotype network. DnaSP v 6.12.03 [38] was used to generate the haplotype data format, and PopART v 1.7 [39] was used to draw haplotype networks. The haplotype network was inferred with the TCS Networks [40] and Median Joining Network [41] methods. Note that an edge linking RaTG13 and the S7 node (distinct from S2 by the U29095 variant) was manually removed in the haplotype network because it was likely caused by a recurrent mutation on site 29095 in S7, which resembled the same state as RaTG13 on the orthologous site.

2.6. Temporal and spatial distributions of SARS-CoV-2 lineages

We extracted the detailed information of the high-quality SARS-CoV-2 genomes (the dates and locations the viruses were isolated) from GISAID for the temporal and spatial distribution analysis. Samples without detailed date information were not considered. We summarized the numbers and proportions of genomes at a two-week interval. This analysis was carried out for the worldwide samples and samples for each individual continent. We present further detailed information on our website (www.covid19evolution.net).

3. Results

3.1. The spectrum of SARS-CoV-2 variants

We downloaded 217,305 SARS-CoV-2 genomes from the GISAID database (http://gisaid.org, as of December 2, 2020; the detailed information for the authors and originating and submitting laboratories of the sequences were acknowledged in Table S2 online). An intense sampling of viruses in a specific location or during a short period of time would cause an excess of highly similar viral genomes in the GISAID database, and potentially leads to biased estimations of the global frequencies of variants. Thus, we grouped the viruses with at least 99.99% sequence identity to reduce the potential influence of sampling bias. In total, we obtained 121,618 high-quality SARS-CoV-2 genomes after redundancy filtering and quality control.

Among the 121,618 genomes, we identified 29,091 SNVs at 20,487 genomic sites after trimming the 5′ (1–220) and 3′ (29675–29903) ends. The number of SNVs identified in the untrimmed region of a single genome (relative to the reference genome NC_045512) was 12 ± 5 (mean ± standard deviation; ranging from 0 to 198). Of these identified SNV sites, 13,001 (63.5%) were bi-allelic and 7486 (36.5%) were multi-allelic. The majority of these SNVs had very low minor allele frequencies (MAF), including 9839 singletons (33.8%). Only 28 variants had a MAF of >5%, including 14 nonsynonymous (nonsyn) sites (1059, 1163, 11083, 14408, 21614, 22227, 22992, 23403, 25563, 28854, 28881, 28883, 28932, and 29645), 13 synonymous (syn) sites (445, 3037, 6286, 7540, 16647, 18555, 18877, 20268, 21255, 23401, 26801, 27944, and 28882), and one noncoding site (241). Fig. S2 (online) presented the frequency spectra of the minor alleles in the CDSs across these genomes.

3.2. The phylogeny of SARS-CoV-2 strains

Previous studies used bat coronavirus RaTG13 or pangolin coronavirus to root SARS-CoV-2 evolution [12–16,19]. At that time, the accuracy of such ancestral inferences remained uncertain [42–45]. Recently, molecular evolution simulations have demonstrated that using these animal coronaviruses as outgroups can yield an accuracy of >95.98% for inferring the ancestral state for a variant of SARS-CoV-2 [37]. Here, we first reconstructed the phylogenetic tree of the SARS-CoV-2 genomes using the maximum-likelihood method, with RaTG13 and a pangolin coronavirus as outgroups. Since phylogeny reconstruction with all the genomes is computationally challenging, we clustered the 121,618 high-quality genomes with a sequence identity cutoff of 99.9%. This yielded 10,061 non-redundant genomes that were used for tree reconstruction. Out of the 10,061 genomes, 9698 (96.4%) belonged to the L lineage (C8782 and U28144), 339 (3.4%) belonged to the S lineage (U8782 and C28144), while 24 (0.2%) could not be categorized as either L or S lineage. These results demonstrated the robustness of delineating the L and S lineages despite the extensive accumulations of the sequenced viral genomes during the development of the COVID-19 pandemic.

As expected, the phylogenetic tree showed a clear delineation between the L and S lineages (Fig. 1). Consistent with previous observations [12–16,19], S was more closely related to RaTG13 and GD Pangolin-CoV than L. The L lineage could be further divided into two major sublineages (L1 and L2) using three tightly-linked SNVs (C3037U, C14408U, and A23403G). Specifically, out of the 9698 L-lineage genomes in Fig. 1, 524 (5.4%) belonged to the L1 sublineage (C3037, C14408, and A23403), and 9127 (94.1%) belonged to the L2 sublineage (U3037, U14408, and G23403). The remaining 47 (0.5%) could be assigned to neither the L1 nor L2 sublineage.

3.3. Extensive linkage of genetic variants among SARS-CoV-2 genomes

For a pair of bi-allelic SNVs, there are four possible haplotypes, namely AB, Ab, aB, and ab, where A and B are the ancestral alleles, and a and b are the derived alleles at the two sites, respectively. There are several possible evolutionary paths that lead to the four observed haplotypes, such as recombination following mutations (Fig. S3a online), multiple independent mutations (Fig. S3b online), stepwise mutations followed by reverse mutations (Fig. S3c, d online), or reverse mutations following simultaneous mutations (Fig. S3e online). In principle, recombination is invoked to explain the four haplotypes only when the recombination rate is substantially higher than the mutation rate, which might be violated for the SARS-CoV-2 viruses. The extent of non-random association of two variants in a given population can be measured with the r^2 metric, which is routinely used in population genetics. Briefly, let p denote the frequency of an allele, then the LD coefficient $D = p_{AB}p_{ab} - p_{Ab}p_{aB}$ and $r^2 = D^2/(p_Ap_ap_Bp_b)$. The log of the likelihood odds ratio (LOD) value can be used to measure the confidence in the non-random association of alleles.

Consistent with the clear delineation between S and L in the phylogenetic tree (Fig. 1), results of the LD analysis revealed nearly

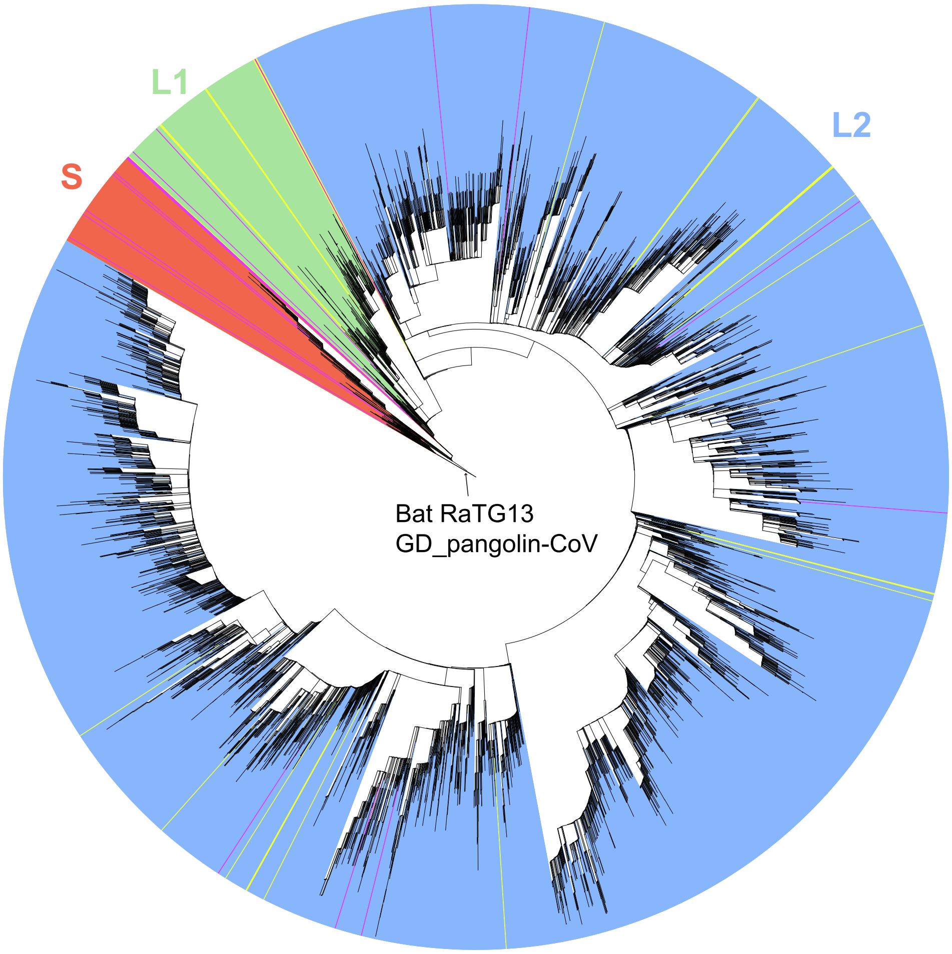


Fig. 1. The phylogenetic tree of 10,061 SARS-CoV-2 genomes. The phylogenetic tree was rooted with the bat coronavirus RaTG13 and GD Pangolin-CoV (the SARS-CoV-2-related viruses in Malayan pangolin samples obtained by anti-smuggling operations by the Guangdong (GD) customs). Note that S was clearly delineated from L, and L further separated into L1 and L2 major sublineages. Genomes from each lineage are colored (S: red; L1: green; L2: blue). The genomes that could not be assigned to S or L are in purple, and the L-lineage genomes that could not be assigned to L1 or L2 are in yellow. Long branches are pruned for better visualization.

complete linkage between the SNVs at sites 8782 and 28144 (Table 1). Likewise, the distinction of L1 and L2 in the phylogenetic tree (Fig. 1**)** is also congruent with the strong LD among sites 3037, 14408, and 23403 (Table 1). These observations inspired us to identify the SNV pairs that were in significant LD systematically. Specifically, we required a significant LD pair to meet three criteria: 1) $r^2 \geq 0.9$, 2) LOD $\geq$ 150 (equivalent to $P \leq 10^{-150}$), and 3) the minor allele frequencies of both sites were no less than 0.5% in at least one major clade (S, L1, or L2). Besides 8782/28144 and the 3037/14408/23403 linkage group (Table 1), we identified another 128 SNVs in CDSs that formed 198 significant LD pairs (53 pairs in S, 20 pairs in L1, and 125 pairs in L2; Fig. S4 online). Thus, in total, we obtained 202 significant LD pairs (133 sites), including 63 nonsyn/nonsyn, 106 syn/nonsyn, and 33 syn/syn pairs (see Tables S3 and S4 online for details).

We polarized the SARS-CoV-2 mutations and summarized the frequencies of the four possible haplotypes for each LD pair. Intriguingly, for 179 out of the 202 significant LD pairs (101 out of 105 pairs if only sites whose ancestral states could be inferred with high confidence were considered), the haplotypes that had both ancestral alleles (AB) or both derived alleles (ab) had higher frequencies than the other two haplotypes across the 121,618 genomes (i.e., p_{AB} and p_{ab} were greater than p_{Ab} and p_{aB}, see Fig. 2 for all 179 pairs and Fig. S5 online for 101 pairs in which the ancestral states at both sites could be confidently inferred; see Table 2 for some examples; see Tables S3 and S4 online for details). Overall, these results suggest the existence of a potential functional association between the derived variants in an LD pair.

Table 1
Pairwise LD analysis for the marker SNVs at sites 8782/28144 (S/L delineation) and sites 3037/14408/23403 (L1/L2 delineation).

Pair of sites	r^2 (LOD)		
	n = 10,061	n = 121,618	n = 202,679
8782, 28144	0.939 (1207)	0.953 (14,874)	0.952 (24,003)
3037, 14408	0.965 (2492)	0.957 (32,084)	0.953 (51,414)
3037, 23403	0.966 (2464)	0.963 (32,168)	0.958 (51,531)
14408, 23403	0.941 (2396)	0.947 (31,574)	0.948 (50,951)

LD was analyzed for the four pairs of sites in three datasets: (1) 10,061 genomes used for the construction of phylogenetic tree; (2) 121,618 genomes obtained after redundancy filtering and quality control; (3) 202,679 genomes obtained after initial quality control. The LOD value is presented in parentheses.

Strikingly, 88.4% (178/202) of the significant LD pairs were further grouped into haplotypes consisting of multiple SNVs. For example, among the 53 LD pairs that had derived haplotypes within S, 48 pairs formed six linkage groups ($\geq$3 SNVs). Among the linkage groups that had derived alleles in the L lineage, the SNVs at sites 3037, 14408, and 23403 had an average pairwise r^2 value of 0.956 ± 0.007 and a LOD value of 31,942 ± 263. These three nearly completely linked SNVs were used to classify the L lineage into L1 (7720 genomes) and L2 (108,833 genomes) lineages, with

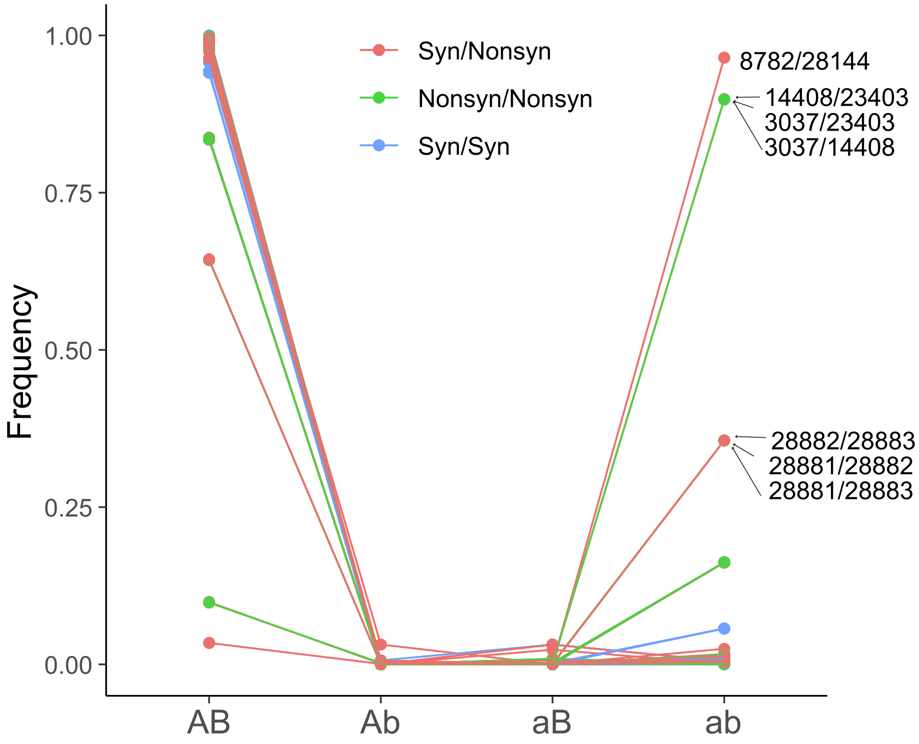


Fig. 2. The frequencies of haplotypes for LD pairs. The normalized frequencies of the four haplotypes (namely AB, Ab, aB, and ab; A and B are the ancestral alleles, and a and b are the derived alleles) for the 202 significant LD pairs. Each dot means the frequency of a certain haplotype for a pair, and the four haplotypes for an LD pair are connected with lines.

Table 2
The observed numbers of haplotypes for seven pairs of sites.

Pair of sites	AB	Ab	aB	ab	n_{AB}	n_{Ab}	n_{aB}	n_{ab}	Sum	
8782/28144	UC	UU	CC	CU	4138	84	80	117,236	121,538	Global
3037/23403	CA	CG	UA	UG	11,947	210	155	109,224	121,536	Global
					(7750)	(159)	(142)	(109,115)	(117,166)	(L)
14408/23403	CA	CG	UA	UG	11,947	260	161	109,051	121,419	Global
					(7750)	(201)	147	(108,949)	(117,047)	(L)
3037/14408	CC	CU	UC	UU	12,027	136	180	109,081	121,424	Global
					(7784)	(126)	(163)	(108,975)	(117,048)	(L)
4402/5062	UG	UU	CG	CU	121,461	0	3	50	121,514	Global
					(4089)	(0)	(0)	(49)	(4138)	(S)
1440/2891	GG	GA	AG	AA	120,930	11	9	606	121,556	Global
					(7069)	(1)	(1)	(603)	(7674)	(L1)
1513/22377	CC	CU	UC	UU	120,069	3	6	746	120,824	Global
					(107,384)	(3)	(6)	(746)	(108,139)	(L2)

Global: all the 121,618 genomes were considered. For the pairs other than 8782/28144, the sizes of the haplotypes (the numbers of genomes) in a major clade are also given in parentheses. The inferred ancestral nucleotides are in black, and the derived variants are in red.

L1 carrying the A23403 (*S*: A1841, D614) and L2 carrying the G23403 (*S*: G1841, G614) allele. Of the 20 LD pairs with derived haplotypes primarily within L1, 11 pairs formed three linkage groups. Similarly, of the 125 LD pairs that had derived haplotypes primarily within L2, 116 pairs formed nine linkage groups (Table S5 and Fig. S6 online).

Altogether, a salient feature of the SARS-CoV-2 viral population is the existence of strong haplotype blocks that are characterized by tightly or even completely linked SNVs, which facilitates lineage designation. We then utilized SNVs at these 133 sites in strong LD, together with SNVs at 73 other sites that reached a MAF $\geq$ 1% in at least one clade (S, L1, or L2) as markers for lineage and sublineage designation. Besides sites 8782/28144 (used for L/S delineation) and sites 3037/14408/23403 (used for L1/L2 delineation), SNVs at 195 sites were used as markers only within one of the three major clades (S, L1, or L2), and SNVs at six sites (11230, 14805, 15324, 15406, 28854, and 28311) were used within two major clades (see Table S5 online for details).

3.4. Defining lineages of SARS-CoV-2 genomes

To better trace the genealogies of the SARS-CoV-2 genomes, we designated the sublineages within the S, L1, or L2 clade based on marker SNVs at 201 genomic sites as described in the previous section. In total, we designated 130 sublineages, including 35 sublineages in L1 (based on 44 sites), 58 sublineages in L2 (based on 105 sites), and 37 sublineages in S (based on 58 sites) (Fig. S6 online). The nomenclature of a sublineage was in the format of L*x* or S*x*, where *x* was an integer starting from 1. A sublineage could be divided into 2nd-tier subclades, each of which ended with a lower-case letter (e.g., L2b), which could be further divided into 3rd-tier haplotypes that ended with an integer (e.g., L2b5). Occasionally, a 3rd-tier sublineage was divided into 4th-tier subclades (ending with a lower-case letter, e.g., L2b5d), 5th-tier haplotypes (ending with an integer, e.g., L2b5d2), and even 6th-tier sublineages (ending with a lower-case letter, e.g., L2b5d2b). Thus, our nomenclature system, which was based on nested or high-frequency marker SNVs, was hierarchical.

The S lineage was divided into ten sublineages that were termed S1–S10. Further classification of L1 yielded L1a–L1j sublineages, whiles that of L2 resulted in L2a–L2g sublineages. To obtain a fine-scale sublineage designation based on these selected marker SNVs, we defined a derived haplotype within a certain clade as a lower-tier subclade if that haplotype carried a marker

SNV in at least ten viral genomes within S or L1, or in at least 100 genomes within L2. In Fig. 3, we presented the characteristic variants for a lineage/sublineage in a hierarchical order. Briefly, for a specific clade (either a lineage or sublineage) to be further divided into subclades, we looked out for specific characteristic variants that uniquely defined each subclade. Except for sites 8782 and 28144, the nucleotides in the reference genome at all the other 204 marker SNV sites were inferred to be the ancestral states. Therefore, for any given clade, the subclade that carried the ancestral alleles was inferred to be the ancestral form to all the other subclades within that clade. For instance, when the sublineage L2c was further categorized into smaller sublineages based on

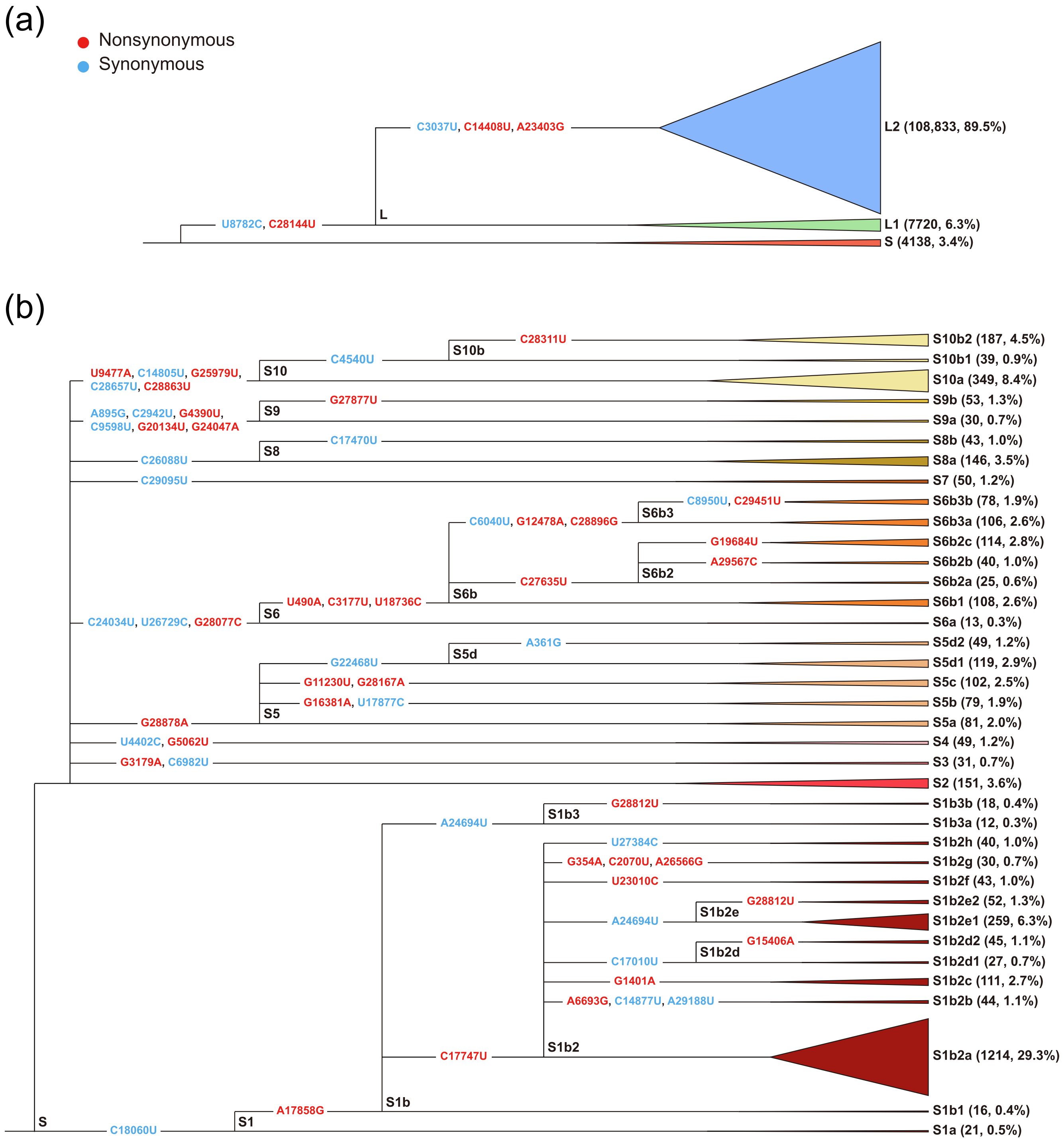


Fig. 3. The hierarchical structure of the sublineage designation system based on marker SNVs. (a) The hierarchical structure of S, L1, and L2 lineages/sublineages. (b–d) Hierarchical structures within S, L1, and L2, respectively. In (a–d), a colored triangle represented a subclade lineage, and the width of the triangle was in scale to the number of the genomes in a clade. For a sublineage, the number of genomes, as well as its percentage in the major clades ((a) for all the genomes; (b–d) for S, L1, and L2, respectively), were given in parentheses. All the SNVs were in coding regions, and the derived alleles (nonsyn, red; syn, blue) labeled in each branch were shared by all the descendant subclades. Except sites 8782 and 28144, the nucleotides in the reference genome at all the other 204 marker SNV sites were inferred to be the ancestral states. All the variants were given in the ancestral/position/derived format. The detailed information for the SNVs that specifically define each lineage or sublineage is given in Table S6 (online). Note that these schemes illustrate how the lineages and sublineages are defined based on the derived variants in a hierarchical manner, and they are not presented in the strict formats of phylogenetic trees.

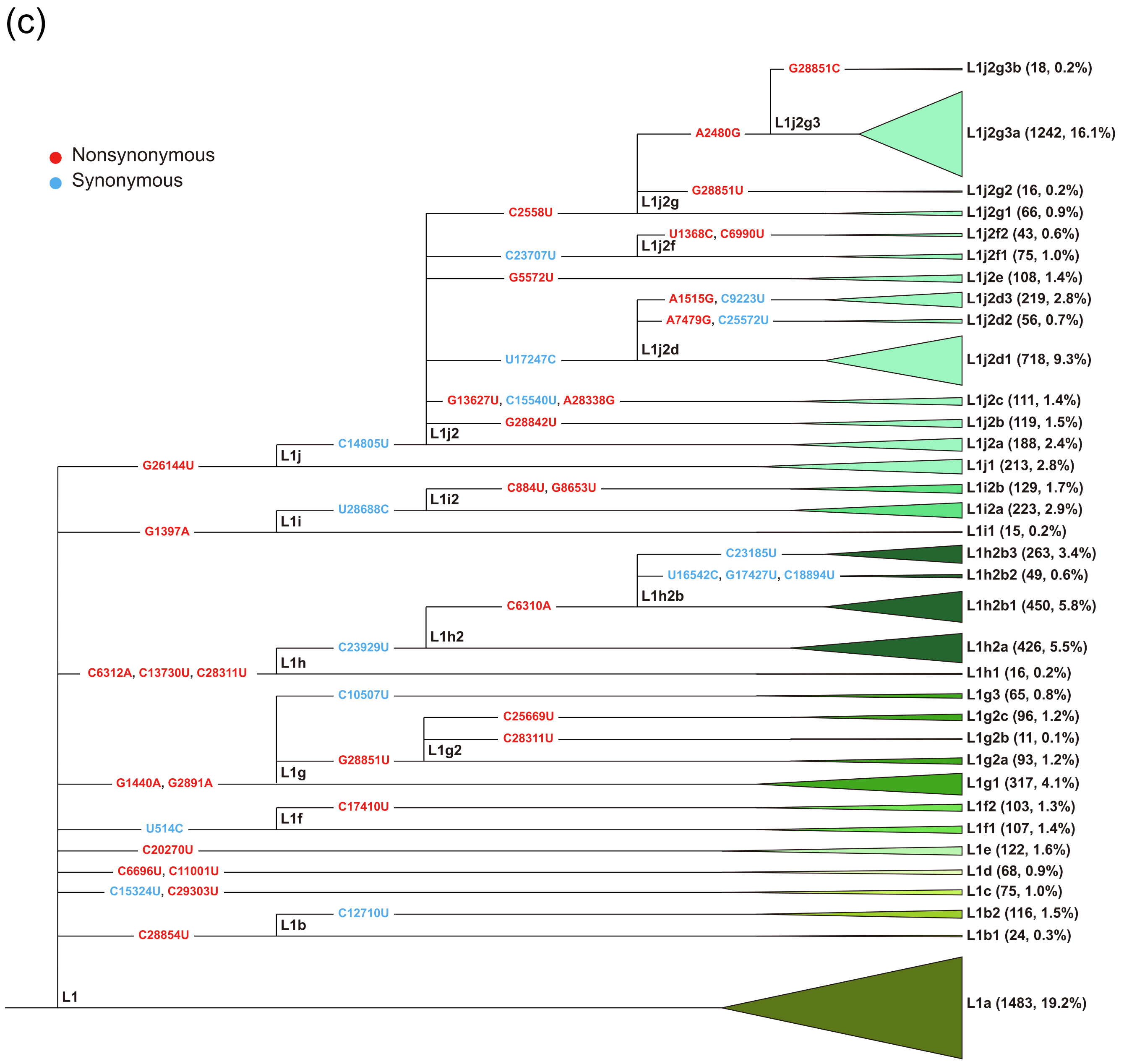


Fig. 3 (*continued*)

SNVs at seven sites (10870, 25505, 25906, 25996, 28651, 28869, and 28975), the smaller sublineage L2c1 had the same nucleotides as the reference genome at all seven sites (G10870, A25505, G25906, G25996, C28651, C28869, and G28975). Whiles L2c2 was characterized with two derived variants (U10870 and U28975), and the sublineage L2c3 carried five derived variants (G25505, C25906, U25996, U28651, and U28869). As shown in Fig. 3, it can be inferred that within L2c, L2c1 was ancestral, whiles L2c2 and L2c3 independently evolved from it. We fully described all the sublineage designation in Supplementary Information (online). The detailed information for the characteristic SNVs used to designate a lineage or sublineage is given in Table S6 (online), and the consensus sequence of each sublineage across the marker SNV sites is presented in Fig. S6 (online). The correspondence of our nomenclature system to other studies is shown in Table S1 (online).

Our analysis revealed that, out of the 121,618 genomes used for lineage and sublineage designation, 4138 (3.4%) belonged to S, 117,236 (96.4%) belonged to L, and 244 (0.2%) belonged to Other (O). Within L and S lineages, 5360 genomes (157 in S, and 5203 in L; 4.4% out of the total analyzed genomes) were labeled with * (asterisk) symbols to represent uncertain belongingness of a subclade in a given clade. For instance, S* belonged to the S lineage but did not fall into S1–S10; S1b* belonged to S1b but did not fall into any subclades in S1b (i.e., S1b1, S1b2, or S1b3). Generally, the * strains were few in number in the viral population, presumably due to mutations at the marker SNV sites, or due to sequencing errors. It is also plausible that some * strains may represent the transitional stages between two sublineages during viral evolution, but they were under-represented in the GISAID dataset either due to sampling bias or reduced fitness of the SARS-CoV-2 (see Supplementary materials online for a detailed description). A small frac-

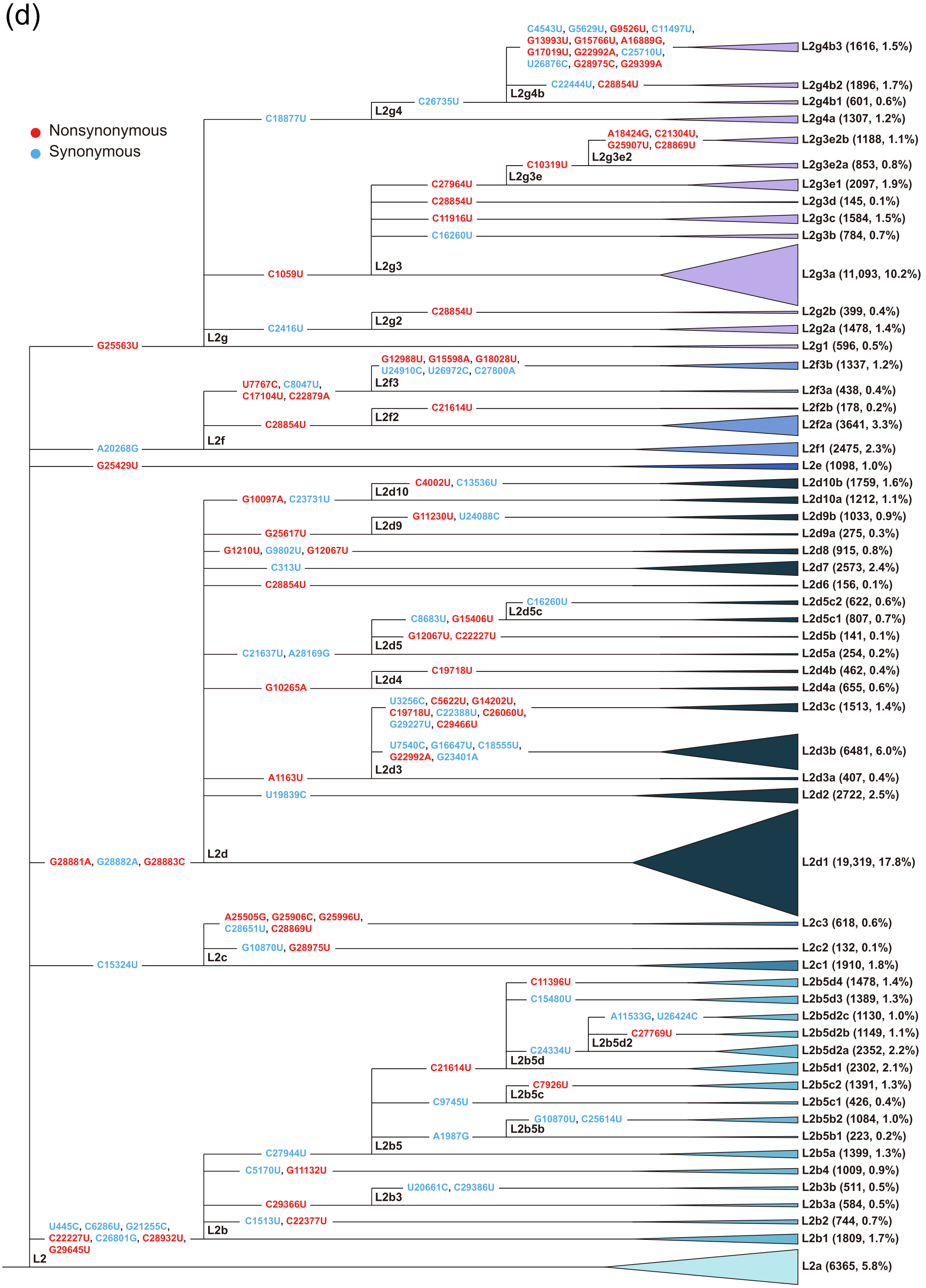


Fig. 3 (*continued*)

tion of the analyzed genomes (7 in S and 475 in L lineage) had ambiguity in sublineage assignment due to the presence of marker SNVs for more than one sublineage. The presence of multiple sublineage SNVs in the genome of a single strain could be due to recombination, recurrent mutations, or sequencing errors.

3.5. Haplotype network analysis of SARS-CoV-2 genomes

The sublineages varied considerably in size within the L or S lineage. This observed variation in the size of sublineages could be attributed to at least three factors: 1) whether the sublineage has an early or late origin, 2) whether the sublineage is under natural selection, and 3) sampling bias. Within L, the largest sublineage was L2d1 (19,319 genomes, 15.9% of all genomes), followed by L2g3a (11,093, 9.1%), and the smallest sublineage was L1g2b (11, < 0.1%). The S lineage was dominated by the S1b2a sublineage (1214 genomes, ~1.0% of all genomes). Each of the other S sublineages accounted for < 0.3% of all SARS-CoV-2 genomes we analyzed.

The haplotype network analysis is powerful for tracing viral genealogies when both the ancestral and descendent samples are analyzed [23,24]. To trace the evolutionary trends of the SARS-CoV-2 genomes, we reconstructed the haplotype networks of the sublineages using all 206 marker SNV sites (Fig. 4). We used the nucleotides in the 206 orthologous sites of RaTG13 in the haplotype network analysis. The same haplotype network topology was obtained when the TCS Networks [40] and Median Joining Network [41] methods were used. As expected, the network analysis showed a distinct separation between the L and S lineages, as well as the delineation between L1 and L2 sublineages. Within S, the sublineages S3–S10 were likely derived from S2 (Fig. 4). Within L, the separation between L1 and L2 lineages was also clearly shown in the network analysis results, with L1 and L2 designated as the ancestral and derived forms, respectively. L1a and L2a were inferred to be the ancestral forms in L1 and L2, respectively. The haplotype network analysis, therefore, provided important insights into the genealogies of the SARS-CoV-2 genomes.

3.6. The continuing evolution of SARS-CoV-2 genomes

Our preliminary analysis showed that 119,168 (98.0%) out of the 121,618 non-redundant genomes had detailed date information for virus isolation. When the lineages were traced over time, we found that the L lineage kept increasing as the pandemic progressed (Fig. 5a). We observed a substantial increase in the fraction of the L2 genomes, all of which carried the U3037, U14408, and G23403 (*S*: G614) variants at the global level. Interestingly, although the frequency of L2d (characterized with A28881, A28882, and C28883) kept increasing until the end of July 2020, the L2b (characterized with C445, U6286, C21255, U22227, G26801, U28932, and U29645) genomes became dominant from the beginning of August 2020. The frequency of L2c (characterized with U15324) became higher from November 2020, although the overall frequency of L2c was still relatively low (Fig. 5a, see Fig. S7c online for other marker SNVs for these lineages).

The lineages were strongly biased in spatial distributions due to high rates of strain isolation and sequencing in some locations as compared to others. In Europe (n = 71,120), the majority (70,434, 99.0%) of the genomes belonged to the L lineage, with L2d (26,206, 36.8%) and L2b (19,416, 27.3%) being the two largest sublineages (Fig. 5b). Since the majority (59.7%, 71,120/119,168) of the viral genomes in GISAID were sequenced in Europe, the frequencies of viral sublineages in Europe were very similar to those at the global level (Fig. 5b). However, the spatial distributions of lineages/sublineages exhibited dramatically different patterns in other continents. For instance, in Asia (n = 8066), 91.8% (7404) of the viral genomes belonged to the L, and 7.1% (576) belong to the S lineage; the L lineage was initially dominated by L2d (2381, 29.5%), but the frequency of L2g substantially increased recently (Fig. 5c). In North America (n = 27,603), the majority of the genomes belonged to the L lineage (25,279, 91.6%), of which L2g (15,149, 54.9%) was the dominant sublineage (Fig. 5d). In Oceania (n = 8973), L2d and L2g were the two dominant sublineages after June 2020, and their frequencies alternated (Fig. 5e). In South America and Africa, where the numbers of genomes in GISAID were relatively small (n = 1307 and 2099, respectively), L2d seemed to be the dominant sublineage (Fig. 5f, g). Similar patterns were observed when we considered all the genomes deposited in GISIAD (198,752 of the 202,679 genomes had detailed dates of virus isolation, as of December 2, 2020; see Fig. S8 online for details). In Figs. S9 and S10 (online), we presented the detailed numbers of genomes in each lineage/sublineage at both global and at continental levels. A more detailed distribution of the viral lineages and sublineages is shown on our user-friendly website (www.covid19evolution.net).

The frequency of a viral variant can fluctuate temporally or spatially due to sampling bias (i.e., the founder effect) [46,47]. Nevertheless, the frequencies of some viral variants may have changed due to the transmission or pathogenicity of the virus. For example, the A82V amino acid change in the glycoprotein of the Ebola virus spread rapidly during the 2013–2016 Ebola outbreak, and the frequency of this variant eventually reached over 90% among all sequenced Ebola genomes [48]. Consequently, studies showed that the A82V change increased viral infectivity in human and primate cells [49,50]. Several recent studies have shown that some variants of SARS-CoV-2 might be associated with viral transmissibility [51] or pathogenicity [52]. All the sublineages in L2 carried the G23403 (*S*: G614) variant, which is known to be associated with increased infectivity of SARS-CoV-2 [20,51,53–56] and mortality of COVID-19 [57–59]. Remarkably, as revealed by our analysis, the L2 sublineage had exhibited a notable increase in frequency over time (Figs. 5 and S11a online; see Figs. S8 and S11b online for all the sequences). This is consistent with a recent study that reported that the G614 variant was driven by adaptive evolution [60]. The pattern has been observed in multiple regions (Figs. 5 and S11a online), indicating that adaptive evolution might be an important force driving the prevalence of the L2 sublineage. These results suggested the SARS-CoV-2 genomes have been continuingly evolving during the COVID-19 pandemic. One caveat is that the frequency of a sublineage might be caused by sampling bias or founder effects. How to separate the effect of natural selection from these confounding factors requires further studies.

3.7. Possible epistatic effects between tightly linked variants

A salient observation in this study is that dozens of SNVs exhibit nearly complete linkage among the examined SARS-CoV-2 genomes. This is surprising since there is very limited evidence yet demonstrating that SARS-CoV-2 has undergone recombination events. In principle, for a pair of bi-allelic SNVs that has four possible haplotypes (AB, Ab, aB, and ab; A and B are ancestral, and a and b are derived at the two sites, as described above), one would not expect to observe p_{ab} to be greater than both p_{Ab} and p_{aB} under neutral evolution. However, among the 202 significant LD pairs, 179 of them exhibited a pattern opposite to neutral expectation at the global scale (Fig. 2a). One hypothesis to account for such observed patterns is that there is extensive epistasis between the SNVs in such LD pairs. Specifically, under the multiple independent mutations model (Fig. S2b online), both A→a and B→b mutations are deleterious, and hence both Ab and aB have reduced fitness than AB. Nevertheless, the subsequent compensatory mutation (e.g., A→a in the Ab molecule, or B→b in the aB molecule) produces

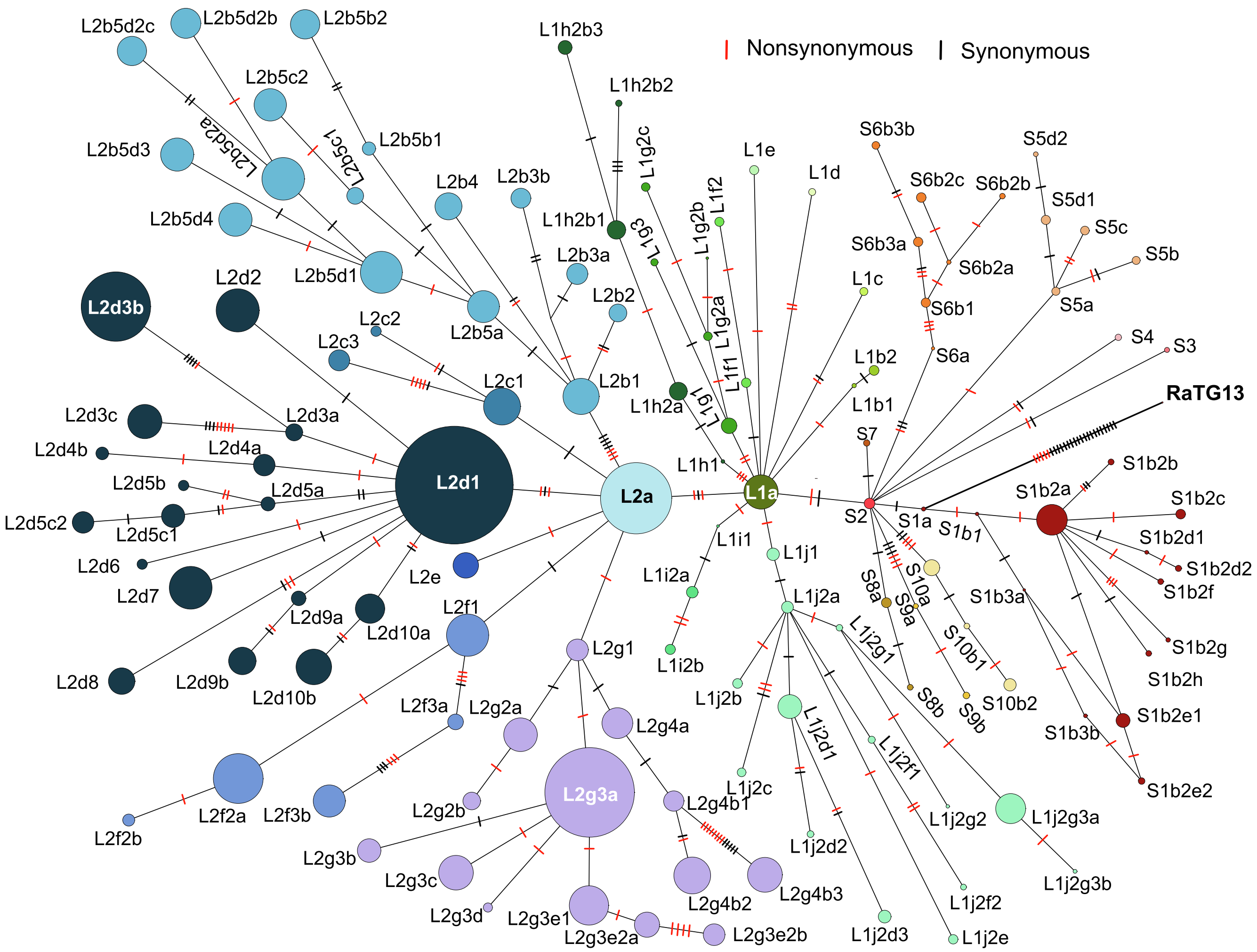


Fig. 4. The haplotype network of the 130 sublineages. The 206 marker SNVs were considered in the haplotype network analysis, and the major haplotype of each sublineage was used as the representative of that sublineage. The size of each sublineage was scaled to the number of genomes in that sublineage. The number of variants (out of 206 sites) between two neighboring sublineages is labeled (red, nonsyn; black, syn). Note that, (1) although SARS-CoV-2 and RaTG13 differed by at least 1000 nucleotides at the genome level, only 206 orthologous sites were considered in the haplotype network analysis; (2) the haplotype network reflected the relative relatedness between the haplogroups but did not necessarily mean one haplotype directly evolved from the neighboring ancestral haplotype because some of the intermediate genomes might be missing in the genomes so far sequenced; and (3) an edge linking RaTG13 and the S7 node (distinct from S2 by the U29095 variant) was manually removed in the haplotype network because it was likely caused by a recurrent mutation on site 29095 in S7, which resembled the same state as RaTG13 on the orthologous site.

the ab haplotype with normal or even higher fitness. Similarly, epistasis can also cause ab to have a similar or even higher fitness than both Ab and aB in other scenarios such as recombination (Fig. S2a online) or recurrent mutations (Fig. S2c, d online).

The possible epistasis between tightly linked variants was more pronounced when we grouped the linked pairs into haplotypes that carried multiple linked variants. For instance, for the 3037/14408/23403 linkage group which defined L1 and L2, 7720 genomes carried the ancestral allele (C3037, C14408, and A23403), 108,833 genomes carried the triple-mutant allele (U3037, U14408, and G23403), while only 22~141 genomes carried the possible transitional haplotypes (Fig. 6a). In other words, the A23403G mutation, which gives rise to the D614G variant in the S protein that has been intensively studied [51,53–55,57–59], was also tightly linked with the variants at sites 3037 (*orf1ab*: C2772U, syn) and 14,408 (*orf1ab*: C14144U, P4715L). Similarly, for the 28881/28882/28883 linkage group within the L2 lineage, 65,529 genomes carried the ancestral allele (G28881, G28882, and G28883), 42,985 genomes carried the triple-mutant allele (A28881, A28882, and C28883), while only 1–107 genomes carried the possible transitional haplotypes (Fig. 6b). Similar patterns were observed for other linkage groups in L1 (Fig. 6c–e) or the S lineage (Fig. S11 online).

More than three decades ago, Motoo Kimura proposed that compensatory neutral mutants (i.e., two mutations that are deleterious individually but jointly restore normal fitness) may be an important driving force of molecular evolution [61]. Here, our observations suggested there might be extensive epistasis and compensatory advantageous mutations between the tightly linked variants. However, at this moment, we cannot rule out the possibilities that other factors (such as sampling bias and founder effects) shaped the observed patterns. Moreover, simultaneous mutations followed by reverse mutations (Fig. S2e online) might explain the non-random associations between the variants. For instance, the A28881/A28882/C28883 variants likely resulted from one replication event and were then maintained by natural selection during evolution. Deciphering the effects of individual and combinatorial variants will be of great value for a deeper understanding of the genome evolution of the SARS-CoV-2 virus.

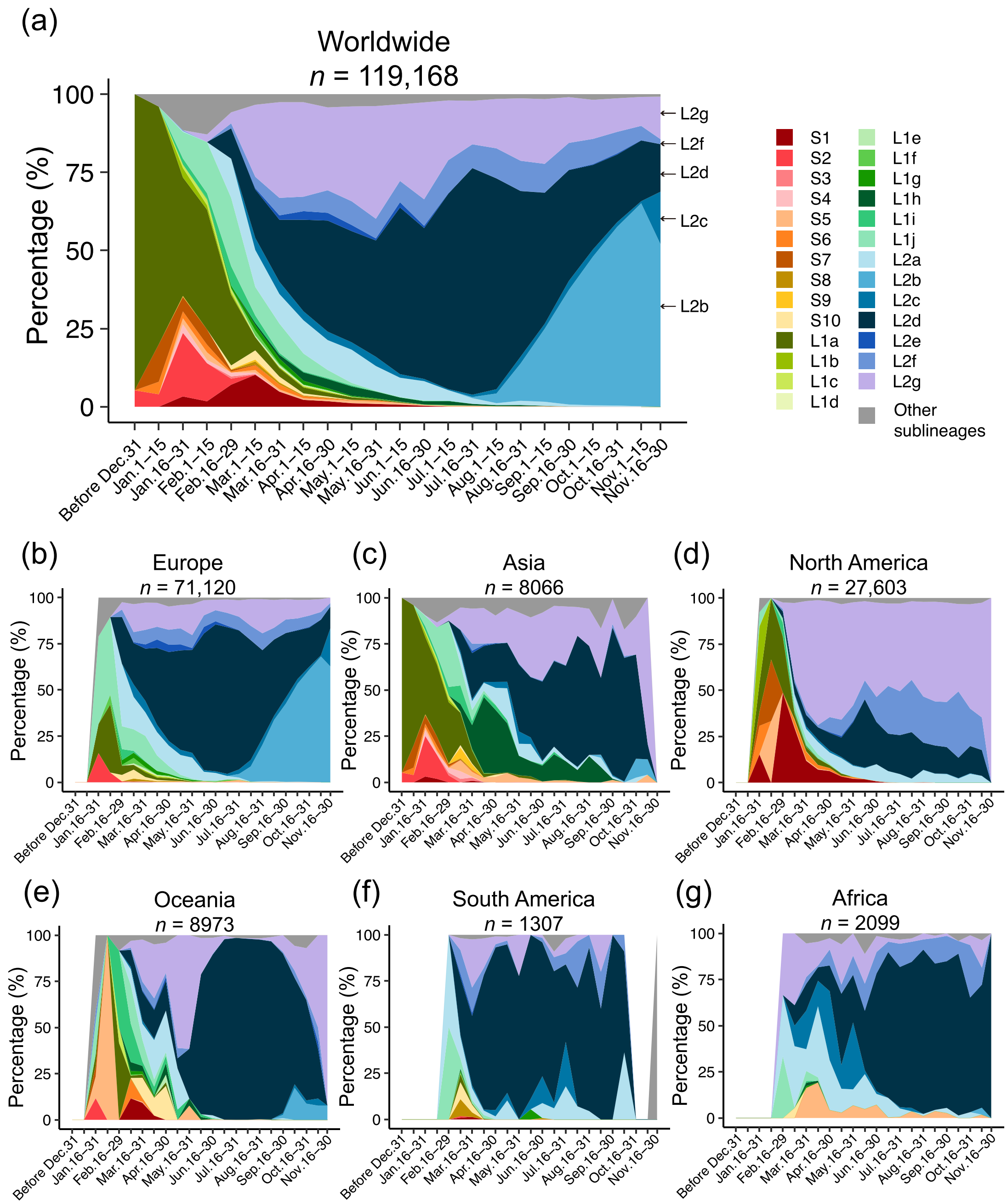
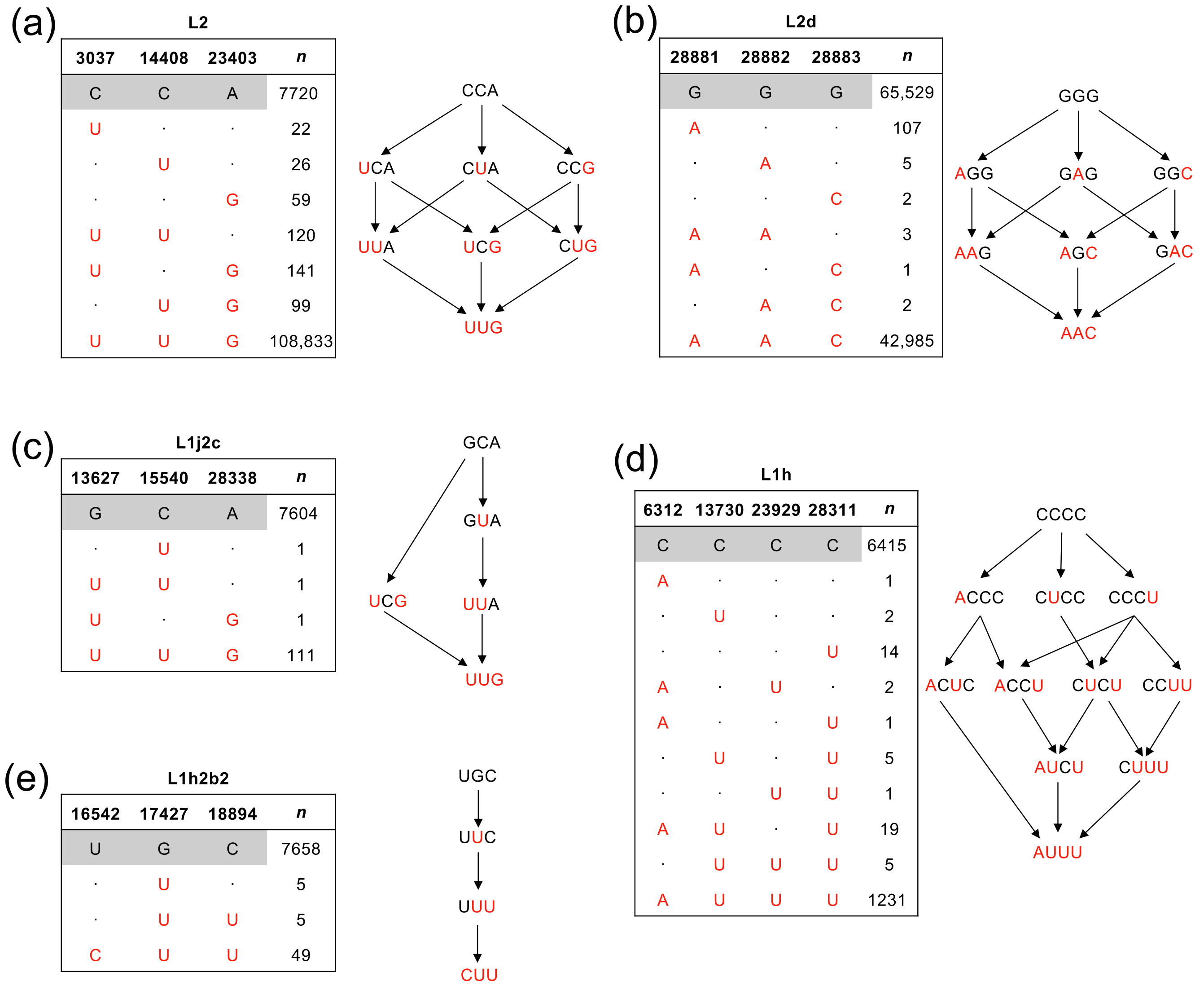


Fig. 5. Temporal and spatial distributions of the sublineages in the whole world (a) and individual continents (b–g) based on 119,168 genomes that had detailed date information. The number of genomes was summarized at a two-week interval, and the frequency of each sublineage (S1–S10, L1a–L1j, L2a–L2g, and other sublineages) in each interval was calculated.

4. Discussion and conclusion

With the rapid increase in publicly available SARS-CoV-2 genome sequences, there are thousands of genetic variants of SARS-CoV-2 available for analysis. In this study, we designated SARS-CoV-2 lineages with 206 marker SNV sites, the majority of which were in strong LD. Our nomenclature system of lineage and sublineage designation has a hierarchical structure and is reflective of the relative relatedness among the subclades of the major clades. The accompanying website that we produced allows users to visualize detailed lineage information and categorize analyses based on SARS-CoV-2 genomes of interest.

Phylogenetic inferences are usually made under the assumption of hierarchical bifurcating trees (i.e., one lineage splits into two descendant lineages). However, the evolution of viruses often violates the bifurcating assumption and evolves in the form of multifurcation, especially in the existence of the super-spreaders. In addition, the phylogenetic analysis can be complicated when both the ancestral and descendent samples are analyzed [23,24]. Thus, phylogeny alone might not be appropriate for tracing viral genealogies. For instance, although the phylogenetic analysis revealed the clear delineation between L and S lineages and the distinction between L1 and L2 clades (Fig. 1), we obtained very complicated results when we analyzed the phylogenetic relation-



(a) L2

3037	14408	23403	*n*
C	C	A	7720
U	·	·	22
·	U	·	26
·	·	G	59
U	U	·	120
U	·	G	141
·	U	G	99
U	U	G	108,833

(b) L2d

28881	28882	28883	*n*
G	G	G	65,529
A	·	·	107
·	A	·	5
·	·	C	2
A	A	·	3
A	·	C	1
·	A	C	2
A	A	C	42,985

(c) L1j2c

13627	15540	28338	*n*
G	C	A	7604
·	U	·	1
U	U	·	1
U	·	G	1
U	U	G	111

(d) L1h

6312	13730	23929	28311	*n*
C	C	C	C	6415
A	·	·	·	1
·	U	·	·	2
·	·	·	U	14
A	·	U	·	2
A	·	·	U	1
·	U	·	U	5
·	·	U	U	1
A	U	·	U	19
·	U	U	U	5
A	U	U	U	1231

(e) L1h2b2

16542	17427	18894	*n*
U	G	C	7658
·	U	·	5
·	U	U	5
C	U	U	49

Fig. 6. The possible evolutionary paths for the haplotypes in five representative linkage groups (a–e). In each linkage group, the inferred ancestral nucleotides are in black, and the derived variants are in red. The haplotypes that carry the reference alleles are presented in a grey background. The numbers of the haplotypes and the possible evolutionary paths from the ancestral to derived via the transitional haplotypes in a clade are given.

ships of the viruses within each of the three major clades (S, L1, or L2). As shown in Fig. S13a (online), on the phylogenetic tree of the S genomes, S1 was overall delineated from the S2–S10 genomes. However, within S1, the S1b2a sublineages were scattered on multiple branches. Similarly, the S2 genomes were scattered on the phylogenetic tree as well. On the other hand, the network analysis revealed that within S1, other subclades radiated from S1b2a; and that S3–S10 sublineages radiated from S2 (Fig. S13b online). Similarly, the L1a genomes, which were inferred to be the ancestral form within the L1 clade, were scattered on the phylogenetic tree of the L1 genomes (Fig. S13c online; see Fig. S13d online for the haplotype network of L1 clade); and the L2a genomes, which were inferred to be the ancestral form within the L2 clade, were also scattered on the phylogenetic tree of the L2 genomes (Fig. S12e online; see Fig. S12f online for the haplotype network of L2 clade). A possible explanation to reconcile these discrepant results is that during the continuing evolution of the viral genomes, the SARS-CoV-2 viruses experienced multifurcating forms of evolution in each major clade. This inference was well congruent with the previously reported super-spreader effect of this virus [62]. Since most of the marker SNVs used to designate the sublineages were in strong LD, we hypothesize that extensive compensatory nucleotide changes occurred during the continuing evolution processes. Altogether, our results supported the notion that combining the phylogenetic and haplotype network analyses better traces the genealogies of the SARS-CoV-2 genomes during evolution.

Recurrent variants (homoplasies) are common in SARS-CoV-2 strains, although most of such variants tend to have very low frequencies (usually < 1%) in SARS-CoV-2 populations [11,63]. In this study, several marker SNVs (at sites 11230, 14805, 15324, 15406, 28311, and 28854), which might be due to recurrent variants, were used to designate sublineages. One might question whether homoplasies may complicate our lineage designation. As shown in Fig. 3 and Fig. S6 (online), the nomenclature system in this study is hierarchical, and the majority of lineages were defined based on nested SNVs that usually exhibit strong (or complete) linkage. For example, recurrent variants are very common at site 28854 [11]. Within L2, the subclade L2d6 was characterized by the C28854U variant.

However, the pre-requisite was that only strains that carry two groups of linked variants (U3037, U14408, and G23403 that defined L2, and A28881, A28882, and C28883 that further defined L2d) simultaneously would be further examined for whether they carry the C28854U variant for L2d6 designation. Although recurrent variants at site 28854 might be common in lineages other than L2d6, they would have a very limited effect on the designation of L2d6. Therefore, our lineage nomenclature system is hierarchical and robust to individual recurrent variants.

The sublineages exhibited substantial differences spatially and temporally. Our analysis showed that adaptive evolution is likely to drive certain sublineages, such as L2, to increase the frequency in multiple areas over the development of the pandemic. Besides, our LD analysis also suggests the existence of possible compensatory substitutions between tightly linked variants during SARS-CoV-2 evolution. The molecular mechanisms underlying these epistatic interactions in viral transmission or pathogenicity are largely unknown. The impact of individual variants and the combined effects for the tightly linked variants on the transmission and pathogenicity of SARS-CoV-2 need further studies.

Our lineage nomenclature system covered most of the major variants in the SARS-CoV-2 genomes currently identified. Nevertheless, given the worryingly increasing number of COVID-19 patients across the globe, it will not be surprising that novel variants appear and get more prevalent in the SARS-CoV-2 populations. Thus, we will readily incorporate such variants and modify the nomenclature of certain sublineages if necessary. For instance, while this work was under review, VUI-202012/01 (also known as VOC-202012/01 or lineage B.1.1.7) [64–66], a new strain of SARS-CoV-2 that might have higher transmissibility [67,68], rapidly increased its frequency, especially in the United Kingdom (Fig. S14 online). The VUI-202012/01 variant, which first appeared on September 20, 2020, had a frequency of ~0.2% (410/202,679) as of December 2, 2020, and its frequency increased to 4.84% (17,043/351,918) as of January 14, 2021 (based on GISAID's data). The VUI-202012/01 variant carried all the eight tightly linked variants used to define the L2d sublineage in our system (Table S7 online). It carried 22 other variants/deletions, and none of them overlapped with the marker SNVs we used for lineage/sublineage designation (Table S7 online). We labeled this variant as L2d11 in our system. The analysis will be regularly updated based on the new sequences released in GISAID and other relevant databases. The progress can be followed by searching on the website (www.covid19evolution.net). Taken together, we believe this study will improve our understanding of the evolutionary dynamics of SARS-CoV-2 genomes at different temporal and spatial scales.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

We thank the researchers who generated and shared the sequencing data from GISAID (https://www.gisaid.org) on which this research is based. We thank Drs. Chung-I Wu, Jindong Zhao, Peter Forster, Ting Wang, Xionglei He, and Weiwei Zhai for suggestive comments to this study. This work was supported by the National Natural Science Foundation of China (91731301 and U1902201), the Ministry of Science and Technology of the People's Republic of China (2020YFC0847000), and the Light of West China Program of the Chinese Academy of Sciences.

Author contributions

Jian Lu, Xuemei Lu, Yaping Zhang, and Wenjie Tan conceived the presented idea and supervised the project with input from Xinghuo Pang, Jianwei Wang, and Guoping Zhao. Xiaolu Tang, Ruochen Ying, Xinmin Yao, Guanghao Li, Changcheng Wu, Xuemei Lu, and Jian Lu analyzed the data and interpreted the results. Shenghan Gao, Songnian Hu, Juncai Ma, and Tiangang Liu contributed to data interpretation. Changcheng Wu, Yiyuli Tang, Zhida Li, Bishan Kuang, Feng Wu, Changsheng Chi, Xiaoman Du, and Yi Qin developed the accompanying website with the suprevison of Jian Lu and Xuemei Lu. Jian Lu, Xuemei Lu, Yaping Zhang, and Wenjie Tan wrote the paper.

Appendix A. Supplementary materials

Supplementary materials to this article can be found online at https://doi.org/10.1016/j.scib.2021.02.012.

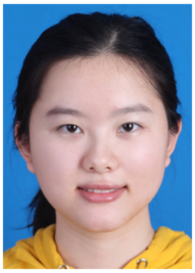
References

[1] Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. Nature 2020;579:265–9.
[2] Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020;382:727–33.
[3] Ren LL, Wang YM, Wu ZQ, et al. Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. Chin Med J (Engl) 2020;133:1015–24.
[4] Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020;579:270–3.
[5] Lam TT, Jia N, Zhang YM, et al. Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. Nature 2020;583:282–5.
[6] Liu P, Chen W, Chen JP. Viral metagenomics revealed Sendai virus and coronavirus infection of Malayan pangolins (*Manis javanica*). Viruses 2019;11:979.
[7] Liu P, Jiang JZ, Wan XF, et al. Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? PLoS Pathog 2020;16:e1008421.
[8] Xiao K, Zhai J, Feng Y, et al. Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. Nature 2020;583:286–9.
[9] Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall 2017;1:33–46.
[10] Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data - from vision to reality. Euro Surveill 2017;22:30494.
[11] van Dorp L, Acman M, Richard D, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol 2020;83:104351.
[12] Tang X, Wu C, Li X, et al. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev 2020;7:1012–23.
[13] Zhang L, Yang JR, Zhang Z, et al. Genomic variations of SARS-CoV-2 suggest multiple outbreak sources of transmission. medRxiv 2020. https://doi.org/10.1101/2020.02.25.20027953.
[14] Yu WB, Tang GD, Zhang L, et al. Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2/HCoV-19) using whole genomic data. Zool Res 2020;41:247–57.
[15] Matsuda T, Suzuki H, Ogata N. Phylogenetic analyses of the severe acute respiratory syndrome coronavirus 2 reflected the several routes of introduction to Taiwan, the United States, and Japan. arXiv: 2020. arXiv: 2002. 08802 [q-bio.GN].
[16] Forster P, Forster L, Renfrew C, et al. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci USA 2020;117:9241–3.
[17] Wu A, Niu P, Wang L, et al. Mutations, recombination and insertion in the evolution of 2019-nCoV. bioRxiv 2020. https://doi.org/10.1101/2020.02.29.971101.
[18] Flynn JA, Purushotham D, Choudhary MNK, et al. Exploring the coronavirus pandemic with the WashU Virus Genome Browser. Nat Genet 2020;52:986–91.
[19] Zhang X, Tan Y, Ling Y, et al. Viral and host factors related to the clinical outcome of COVID-19. Nature 2020;583:437–40.
[20] Li Q, Wu J, Nie J, et al. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. Cell 2020;182:1284–94.
[21] Rambaut A, Holmes EC, O'Toole A, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol 2020;5:1403–7.
[22] Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 2018;34:4121–3.
[23] Jombart T, Eggo RM, Dodd PJ, et al. Reconstructing disease outbreaks from genetic data: a graph approach. Heredity 2011;106:383–90.
[24] Paradis E. Analysis of haplotype networks: the randomized minimum spanning tree method. Methods Ecol Evol 2018;9:1308–17.

[25] Morel B, Barbera P, Czech L, et al. Phylogenetic analysis of SARS-CoV-2 data is difficult. Mol Biol Evol 2020;38:1777–91.
[26] Y Chromosome Consortium. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res 2002;12:339–48.
[27] Poznik GD, Xue Y, Mendez FL, et al. Punctuated bursts in human male demography inferred from 1244 worldwide Y-chromosome sequences. Nat Genet 2016;48:593–9.
[28] van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 2009;30:E386–94.
[29] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013;30:772–80.
[30] Page AJ, Taylor B, Delaney AJ, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom 2016;2:e000056.
[31] Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078–9.
[32] Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010;26:2460–1.
[33] Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 2020;37:1530–4.
[34] Barrett JC, Fry B, Maller J, et al. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21:263–5.
[35] Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. Bioinformatics 2011;27:2156–8.
[36] Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75.
[37] Li T, Tang X, Wu C, et al. The use of SARS-CoV-2-related coronaviruses from bats and pangolins to polarize mutations in SARS-CoV-2. Sci China Life Sci 2020;63:1608–11.
[38] Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, et al. DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol Biol Evol 2017;34:3299–302.
[39] Leigh JW, Bryant D. POPART: full-feature software for haplotype network construction. Methods Ecol Evol 2015;6:1110–6.
[40] Clement M, Posada D, Crandall KA. TCS: a computer program to estimate gene genealogies. Mol Ecol 2000;9:1657–9.
[41] Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 1999;16:37–48.
[42] Mavian C, Pond SK, Marini S, et al. Sampling bias and incorrect rooting make phylogenetic network tracing of SARS-CoV-2 infections unreliable. Proc Natl Acad Sci USA 2020;117:12522–3.
[43] Sánchez-Pacheco SJ, Kong S, Pulido-Santacruz P, et al. Median-joining network analysis of SARS-CoV-2 genomes is neither phylogenetic nor evolutionary. Proc Natl Acad Sci USA 2020;117:12518–9.
[44] MacLean OA, Orton RJ, Singer JB, et al. No evidence for distinct types in the evolution of SARS-CoV-2. Virus Evol 2020;6:veaa034.
[45] Forster P, Forster L, Renfrew C, et al. Reply to Sánchez-Pacheco et al., Chookajorn, and Mavian et al.: explaining phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci USA 2020;117:12524–5.
[46] Rambaut A, Posada D, Crandall KA, et al. The causes and consequences of HIV evolution. Nat Rev Genet 2004;5:52–61.
[47] Ruan Y, Luo Z, Tang X, et al. On the founder effect in COVID-19 outbreaks: how many infected travelers may have started them all? Natl Sci Rev 2021;8:nwaa246.
[48] Bedford T, Malik HS. Did a single amino acid change make Ebola virus more virulent? Cell 2016;167:892–4.
[49] Diehl WE, Lin AE, Grubaugh ND, et al. Ebola virus glycoprotein with increased infectivity dominated the 2013–2016 epidemic. Cell 2016;167:1088–98.
[50] Urbanowicz RA, McClure CP, Sakuntabhai A, et al. Human adaptation of Ebola virus during the West African outbreak. Cell 2016;167:1079–87.
[51] Korber B, Fischer WM, Gnanakaran S, et al. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 2020;182:812–27.
[52] Yao H, Lu X, Chen Q, et al. Patient-derived SARS-CoV-2 mutations impact viral replication dynamics and infectivity *in vitro* and with clinical implications *in vivo*. Cell Discov 2020;6:76.
[53] Daniloski Z, Jordan TX, Ilmain JK, et al. The Spike D614G mutation increases SARS-CoV-2 infection of multiple human cell types. bioRxiv 2020. https://doi.org/10.1101/2020.06.14.151357.
[54] Zhang L, Jackson CB, Mou H, et al. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. Nat Commun 2020;11:6013.
[55] Ozono S, Zhang Y, Ode H, et al. Naturally mutated spike proteins of SARS-CoV-2 variants show differential levels of cell entry. bioRxiv 2020. https://doi.org/10.1101/2020.06.15.151779.
[56] Yurkovetskiy L, Wang X, Pascal KE, et al. Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. Cell 2020;183:739–51.
[57] Becerra-Flores M, Cardozo T. SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. Int J Clin Pract 2020;74:e13525.
[58] Eaaswarkhanth M, Al Madhoun A, Al-Mulla F. Could the D614G substitution in the SARS-CoV-2 spike (S) protein be associated with higher COVID-19 mortality? Int J Infect Dis 2020;96:459–60.
[59] Toyoshima Y, Nemoto K, Matsumoto S, et al. SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. J Hum Genet 2020;65:1075–82.
[60] Garvin MR, Prates ET, Pavicic M, et al. Potentially adaptive SARS-CoV-2 mutations discovered with novel spatiotemporal and explainable AI models. Genome Biol 2020;21:304.
[61] Kimura M. The role of compensatory neutral mutations in molecular evolution. J Genet 1985;64:7.
[62] Gómez-Carballa A, Bello X, Pardo-Seco J, et al. Mapping genome variation of SARS-CoV-2 worldwide highlights the impact of COVID-19 super-spreaders. Genome Res 2020;30:1434–48.
[63] Rice AM, Castillo-Morales A, Ho AT, et al. Evidence for strong mutation bias towards, and selection against, U content in SARS-CoV-2: implications for vaccine design. Mol Biol Evol 2020;38:67–83.
[64] https://www.who.int/csr/don/21-december-2020-sars-cov2-variant-united-kingdom/en/; 2020. [Accessed 15 January 2021].
[65] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/947048/Technical_Briefing_VOC_SH_NJL2_SH2.pdf; 2020. [Accessed 15 January 2021].
[66] https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563; 2020. [Accessed 15 January 2021].
[67] Grabowski F, Preibisch G, Kochańczyk M, et al. SARS-CoV-2 Variant Under Investigation 202012/01 has more than twofold replicative advantage. medRxiv 2021. https://doi.org/10.1101/2020.12.28.20248906.
[68] Davies NG, Barnard RC, Jarvis CI, et al. Estimated transmissibility and severity of novel SARS-CoV-2 Variant of Concern 202012/01 in England. medRxiv 2020. https://doi.org/10.1101/2021.02.01.21250959.

Xiaolu Tang is a Ph.D. candidate in Bioinformatics at the School of Life Sciences, Peking University, China. She received a bachelor's degree from Northwest A&F University in 2017. Her research interest includes the evolution of viral genomes and the translational regulation of eukaryotes.

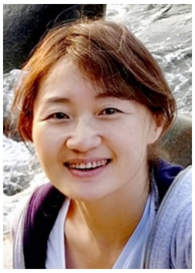
Ruochen Ying is a Ph.D. candidate in Bioinformatics at the School of Life Sciences, Peking University, China. She received a bachelor's degree from Zhejiang University in 2019. Her research interest includes small RNA-mediated gene regulation and the evolution of viral genomes.



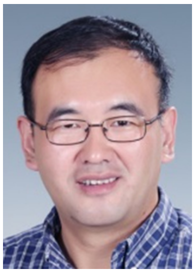
Xinmin Yao is a Ph.D. candidate in Bioinformatics at the School of Life Sciences, Peking University, China. She received a bachelor's degree from Peking University in 2018. Her research interest includes intragenomic conflicts of selfish elements and the evolution of SARS-CoV-2.

Wenjie Tan is a professor at National Institute for Viral Disease Control and Prevention, Chinese Center for Disease Control and Prevention (China CDC). He received his Ph.D. degree from the Chinese Academy of Preventive Medicine in July 1998. In 2008, he became Chief of the Biotech Center for Viral Disease Emergency in China CDC. His current research interest focuses on pathogen biology and immunology of human coronaviruses (including SARS-CoV, SARS-CoV-2, and MERS-CoV) and other emerging viral diseases.

Yaping Zhang is a professor and principal investigator of Molecular Evolution and Genome Diversity, Kunming Institute of Zoology, Chinese Academy of Sciences. His research interest includes molecular phylogenetics, biodiversity, origin of domestic animals and artificial selection, and genome diversity and evolution.



Xuemei Lu is a professor at the State Key Laboratory of Genetic Resource and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences. She obtained her Ph.D. degree from Kunming Institute of Zoology, and conducted postdoctoral training at the Department of Ecology and Evolution, The University of Chicago. She joined the School of Life Science, Sun Yet-sen University as an associate professor in 2005, and moved to Beijing Institute of Genomics, Chinese Academy of Sciences as a professor in 2009. Her research interest focuses on the ecology and evolution of somatic cells and natural populations.



Jian Lu is a principal investigator at the School of Life Sciences & State Key Laboratory of Protein and Plant Gene Research, Peking University. He received his Ph.D. degree in Ecology and Evolution from the University of Chicago in 2008. His research interest includes the mechanisms and evolutionary principles of post-transcriptional gene expression regulation, the genetic basis of adaption, and the evolution of viruses.