Genome charaterization based on the Spike-614 and NS8-84 loci of SARS-CoV-2 reveals two major onsets of the COVID-19 pandemic

Xiaowen Hu[1,2#], Yaojia Mu[1#], Ruru Deng[1#], Guohui Yi[3], Lei Yao[*4], Jiaming Zhang[1*]

[1]Key Laboratory of Microbiology of Hainan Province, Institute of Tropical Bioscience and Biotechnology, Chinese Academy of Tropical Agricultural Sciences, Haikou, Hainan, 57110, China

[2]Zhanjiang Experimental Station, Chinese Academy of Tropical Agricultural Sciences, Zhanjiang, 524013, China

[3]Public Research Laboratory, Hainan Medical University, Haikou571199, China

[4]Experiment Medicine Center, The Affiliated Hospital of Southwest Medical University, 646000, Luzhou, Sichuan, China

[#]These authors contributed equally

* Correspondence should be addressed to J.Z. (zhangjiaming@itbb.org.cn) and L.Y. (yaolei2009@gmail.com)

Xiaowen Hu: yuyin110110110@163.com

Yaojia Mu: muyaojia@foxmail.com

Ruru Deng: 601859770@qq.com

Guohui Yi: guohuiyi6@hainmc.edu.cn

Lei Yao: yaolei2009@gmail.com

Jiaming Zhang: zhangjiaming@itbb.org.cn

Abstract

The global COVID-19 pandemic has lasted for three years since its outbreak, however its origin is still unknown. Here, we analyzed the genotypes of 3.14 million SARS-CoV-2 genomes based on the amino acid 614 of the Spke (S) and the amino acid 48 of NS8 (nonstructural protein 8), and identified 16 linkage haplotypes. The GL haplotype (S_614G and NS8_48L) was the major haplotype driving the global pandemic and accounted for 99.2% of the sequenced genomes, while the DL haplotype (S_614D and NS8_48L) caused the pandemic in China in the spring of 2020 and accounted for approximately 60% of the genomes in China and 0.45% of the global genomes. The GS (S_614G and NS8_48S), DS (S_614D and NS8_48S) and NS (S_614N and NS8_48S) haplotypes accounted for 0.26%, 0.06%, and 0.0067% of the genomes, respectively. The main evolutionary trajectory of SARS-CoV-2 is DS→DL→GL, whereas the other haplotypes are minor byproducts in the evolution. Surprisingly, the newest haplotype GL had the oldest time of most recent common ancestor (tMRCA), which was May 1 2019 by mean, while the oldest haplotype had the newest tMRCA with a mean of October 17, indicating that the ancestral strains that gave birth to GL had been extinct and replaced by the more adapted newcomer at the place of its origin, just like the sequential rise and fall of the delta and omicron variants. However, they arrived and evolved into toxic strains and ignited a pandemic in China where the GL strains did not exist at the end of 2019. The GL strains had spread all over the world before they were discovered, and ignited the global pandemic, which had not been noticed until the pandemic was declared in China. However, the GL haplotype had little influence in China during the early phase of the pandemic due to its late arrival as well as the strict transmission controls in China. Therefore, we propose two major onsets of the COVID-19 pandemic, one was mainly driven by the haplotype DL in China, the other was driven by the haplotype GL globally.

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

Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the causal agent of COVID-19, a disease first reported in Wuhan, China [1-3]. This virus differs from all known coronaviruses, including the six coronavirus species that are known to cause human disease [4], and has been classified as the seventh coronavirus that can infect humans by the International Committee on Taxonomy of Viruses [5]. SARS-CoV-2 quickly spread within China. A range of public health interventions was used to control the epidemic, including isolation of suspected and confirmed cases, transport prohibition in and out of epidemic centers, suspension of public transport, closing of schools and entertainment venues, and public gathering bans and health checks [6, 7]. These measures eventually brought about an end to sustained local transmission across China in April 2019. However, COVID-19 expands its territories, and became a global pandemic. As of 8 September 2021, there have been over 266 million confirmed cases and 5.2 million deaths of COVID-19, reported to WHO (https://covid19.who.int/).

To our knowledge, the origins of SARS-CoV-2 remain elusive. Understanding how, when, and where the virus was transmitted from its natural reservoir to humans is crucial for preventing future coronavirus outbreaks [8]. Progresses have been made in this aspect. A bat-origin virus RaTG13 that has 96% sequence identities with SARS-CoV-2 has been published [9], which suggests bats as a likely natural reservoir of this virus. Pangolin [10], snakes [11], turtles [12], and/or Bovidae and Cricetidae [13] have been suggested to act as potential intermediate hosts that helped the virus to cross the species barriers to infect humans. According to the clinical summary of the earliest cases of COVID-19 (also known as 2019-nCoV), the majority of cases were exposed to the Huanan Seafood Market [14], which also had wild animals. Therefore, the market was considered to be an obvious candidate location of the initial zoonotic transmission event of COVID-19 [15], and both epidemiological and phylogenetic approaches suggested late 2019 the occurring time of the pandemic [14, 16, 17]. However, none of the wild animals from the Huanan Seafood Market were tested positive for SARS-CoV-2 [18]. Some environmental samples were positive, but their viral genomes were not located at the basal position of the phylogeny [19]. Moreover, some of the early cases were not epidemiologically linked to the Huanan Seafood Market [16], but linked to other markets [20], and animal-to-human transmission in the market has never been confirmed and should not be overemphasized [21].

The origin and evolution of SARS-CoV-2 has been well reviewed [22, 23]. Most tracing research focused on analysis of the early genomes [23-25]. For example, Tang and colleagues classified the early SARS-CoV-2 genomes into two major lineages (L and S lineages), and the two lineages were well defined by just two SNPs that show complete linkage across SARS-CoV-2 strains[26]. The L lineage constituted 70% of the sequenced genome, while the S lineage constituted approximately 30% of the sequenced genome in the 103 genomes in the early phase of epidemic [26]. However, higher percentage is not sufficient to substantiate a more aggressive type [27], until the emergence of the delta variant [28]. Evolutionary analyses suggested the S lineage appeared to be more related to coronaviruses in animals [26]. Li and colleagues outlined the early viral spread in Wuhan and its transmission patterns in China and across the rest of the world [23]. Pekar and colleagues analyzed the early viral genomes (before the end of April 2020) in China, and defined a period between mid-October and mid-November 2019 as the plausible interval when the first case of SARS-CoV-2 emerged in Hubei province, China [24]. Ruan and colleagues discovered distinct set of mutations driving the waves of replacements of strains, and the split between the Asian and European lines occurred before September of 2019, suggesting twin-beginning scenario of the pandemic [25]. Due to the presumed undetected transmission and insufficient genome sequences in the early phase of the pandemic, the origin of the virus remains uncovered. It is true that coronaviruses evolve rapidly [29], and the mutation rate of SARS-CoV-2 was predicted to be approximately 8.0×10^{-4} subs/site/year [30]. However, the inheritance rate of each nucleotide should be higher than 99.9992%, taken the negative selection pressure into account [26]. Upon the increased coverage of the sequenced genomes in the late phase of the pandemic, we believe that the secret of the viral origin may be buried in the genomes, not necessarily in the early genomes.

In this study, we analyzed 3.14 million genomes obtained from GISAID database, and used two gene loci to anchor the haplotypes, and identified the haplotype for the most recent common ancestor (MRCA) and the proximal geographical origin of SARS-CoV-2. The evolutionary trajectory of the SARS-CoV-2 haplotypes is proposed.

Materials and methods

SARS-CoV-2 genome sequences were obtained from the GISAID database (https://www.gisaid.org) on 18 October 2021. The genomes that were incomplete <29,000 nt, had low coverage with >0.5% unique amino acid mutations and more than 1% 'N's were filtered. A total of 3,140,626 genome sequences of SARS-CoV-2 were remained for further analysis. All the genomes were isolated from human patients. To analyze the temporal-dependent evolutionary pattern of SARS-CoV-2, the genome sequences were divided into groups by months according to their sample collection date, and were aligned using ViralMSA [31] and Clustal-Omega [32]. The unknown and/or low quality nucleotides were removed from the alignments using personalized scripts, which is available (https://github.com/XiaowenH/SeqSC.git). Nucleotide diversity (Pi) and population mutation rate (Theta-W) of each subgroup was calculated using Pairwise Deletion Model with DnaSP 6 [33]. The Pearson correlation coefficient [34] and significant analysis was calculated by cor.test in the R package (version 4.1.1) developed by the R Development Core Team (https://www.r-project.org).

To genotype the genomes based on the spike gene, the spike coding sequences of the 3.14 million genomes were aligned to the reference genome Wuhan-Hu-1 (MN908947.3), and the relevant CDS was extracted and translated into peptides by using SeqKit package [35]. The peptides were aligned with the reference using Clustal-Omega [32], and the site aligned with the site 614 of the reference was identified using personalized script (https://github.com/XiaowenH/SeqSC.git). The NS8 gene was genotyped using the same method, except that NS8 gene was used as the reference.

To analyze the mutation accumulation pattern of each haplotype, the genome sequences were divided by month, and were aligned using ViralMSA[31] and/or Clustal-Omega [32]. Phylogenetic analysis was performed using Mega 11 [36]. The evolutionary network analysis was performed by using the Median-joining Networks in Potpart 1.7 [37].

To explore the evolutionary dynamics, genomes of each haplotype were aligned with ViralMSA[31] after removal of redundants by using CD-HIT [38] with a threshhold of 99.98%. The outputs were transformed to Nexus format by using ALTER [39]. Bayesian phylodynamics analysis was performed by using BEAST 1.10 [40] with molecular clock set to strict and coalescent prior set to Bayesian Skyline [24]. The posterior distribution was summarized by using TRACER 1.7 [41].

Results

Genome and mutation accumulation patterns of SARS-CoV-2

A total of 3,140,626 SARS-CoV-2 genomes designated as complete and high coverage in the GISAID database (www.gisaid.org) were obtained on 2021-10-18. The average length of the genomes is 29,781.5 bases with the minimum and maximum lengths of 29,000 and 31,579 bases, respectively. Both the sequenced genomes and confirmed cases increased exponentially during 2020, and remained at high level during 2021 (Fig. 1A). Regression analysis shows that the sequenced genomes and confirmed cases are positively correlated with a linear coefficient r=0.75 (p-value = 5.5e-05, Fig. 1B), indicating that the sequenced genomes sufficiently represented the dynamic population.

The nucleotide diversity (Pi) was accumulated in a temporal-dependent pattern (Fig.1C, supplementary Table S1). Both Pi and Theta-W (population mutation rate) increased linearly in 2020 with correlation coefficients r=0.99 (p=5.349e-10, Fig. 1C) and r=0.975 (p=7.331e-08, Fig.1D) for Pi and Theta-W, respectively. They then varied up and down in 2021, possibly due to the shift of dominant viral strains, such as the delta and the omicron variants (submitted to PlosOne).

Mutations of the amino acid 84 of NS8 in the 3.14 million genomes

The mutation at site 84 of the nonstructural protein ORF8 (NS8) was one of the earliest sites to be used to classified the early SARS-CoV-2 strains, and they were classified into two major lineages (designated L and S) that were well defined by two different SNPs (T8782C and C28144T, the latter being Orf8-S84L) [26]. The NS8 contains 121 amino acids, and promotes the expression of the ER unfolded protein response factor ATF6 in human cells [42]. We extracted the NS8 coding DNA sequence from all the 3.14 million SARS-CoV-2 genomes, and classified the genomes according to the identities at the amino acid 84. Besides the two known alleles, four additional mutations were identified, including NS8_84V, NS8_84I, NS8_84F, NS8_84C (Table 1). NS8_84L was the first to have been collected. Its first collection date was 2019-12-24 (EPI_ISL_402123), and a total of 21 NS8_84L samples were collected in Wuhan, China in December 2019, and constituted 95.5% of the early samples (Supplementary Table S3). This genotype constituted 99.66% of the 3.13 million genomes that were successfully classified (Table 1). NS8_84S was first collected on 2019-12-30 and constituted approximately 0.34% of the total genomes. The numbers of the newly identified genomes were very few (Table 1).

Approximately 0.4% genomes were unidentified due to low quality in the region (Table 1).

Table 1. Alleles of the amino acid 84 in the nonstructural protein 8 (NS8).

Allele	Sequence (75-93)	Genomes	Percentage	First collect	Collection place
NS8_TG13	DIGNYTVSC**S**PFTINCQEP	RaTG13 ref.		2013-07-24	Yunan,China
NS8_wiv04	DIGNYTVSC**L**PFTINCQEP	GISAID ref.		2019-12-30	Wuhan,China
NS8_84L	DIGNYTVSC**L**PFTINCQEP	3,115,900	99.6564*	2019-12-24	Wuhan,China
NS8_84S	DIGNYTVSC**S**PFTINCQEP	10,732	0.3432	2019-12-30	Wuhan,China
NS8_84V	DIGNYTVSC**V**PFTINCQEP	3	0.0001	2020-09-15	South Korea
NS8_84I	DIGNYTVSC**I**PFTINCQEP	4	0.0001	2021-05-05	USA
NS8_84F	DIGNYTVSC**F**PFTINCQEP	2	0.00006	2020-02-02	Sichuan,China
NS8_84C	DIGNYTVSC**C**PFTSNCQEP	1	0.00003	2020-03-12	USA
unknown		13,985			
Total		3,140,626			

*Note: Percentage in the genomes that have been successfully classified.

Genome and mutation accumulation and phylogenetic analysis of NS8 gene

NS8_84L was dominant over NS8_84S all the time (Supplementary Table S3). The genome number of NS8_84L increased exponentially from December 2019 to March 2021, and kept at high numbers (Fig. 2A), while NS8_84S had a peak in March 2020, and remained at low numbers the rest of time (Fig. 2B). Even in its peak month, its genome number was only 12% of NS8_84L (supplementary Table S3). The genome number of NS8_84L is positively correlated with the number of the confirmed cases with a coefficient r=0.88 (p=8.914e-05, Fig. 2C), while the genome number of NS8_84S is not (r=-0.32, p=0.29, Fig.2D). Therefore, NS8_84L was the main genotype driving the growth of confirmed cases worldwide.

The nucleotide diversity of both genotypes was accumulated linearly in 2020. Regression analysis revealed first occurrence dates of approximately 14 October 2019 (95% confidence interval: 18 September to 9 November) for the genotype NS8_84L (Fig. 2E), and 28 September 2019 (95%

confidence interval: 6 August to 21 November) for genotype and NS8_84S (Fig.2F), respectively. Therefore, NS8_84S occurred earlier than NS8_84L. Phylogenetic analysis using both peptide and RNA sequences of the NS8 showed that NS8_84S was closest to the root (Fig. 2G, H), and the phylogenies by using different methods all suggested an earlier occurrence of NS8_84S than NS8_84L consistent (Supplementary Figure S2 and S3), which was supported by that the bat SARS-CoV related viruses all have NS8_84S [26].

Taken together, NS8_84S is an earlier genotype than NS8_84L, and the latter may be originated from mutation of NS8_84S

Mutations of the amino acid 614 of the spike protein in the 3.14 million genomes

Mutation of D614G in Spike protein was the major driving force in the COVID-19 pandemic, the virulent strains alpha, delta, and omicron all carried this mutation. The Spike (S) is a surface projection glycoprotein. Mutations in this protein have previously been associated with altered pathogenesis and virulence in other coronaviruses [43]. We extracted the S gene sequences from the 3.14 genomes, and classified the genomes based on mutations at the amino acid 614. Seven alleles were identified, including S_614D, S_614G, S_614N, S_614V, S_614A, S_614V, S_614C (Table 2).

The S_614D was the earliest to have been collected, and was first collected in Wuhan, China on 2019-12-24 (EPI_ISL_402123, Table 2). In outside China, it was first collected in Thailand on 2020-01-08, Nepal on 2020-01-13, USA on 2020-01-19, and France on 2020-01-23. This genotype constituted the second largest number of the genomes with a percentage of 0.7%.

The S_614G accounted for 99.69% of the genomes (Table 2), and was first collected on 2020-01-01 in Argentina (EPI_ISL_4405694), followed by nine samples collected on 2020-01-03 in USA (EPI_ISL_3537067, EPI_ISL_3537066, EPI_ISL_3537065, EPI_ISL_3537064, EPI_ISL_3537063, EPI_ISL_3537062, EPI_ISL_3537061, EPI_ISL_3537060, EPI_ISL_3537059). It was first collected in Australia (EPI_ISL_3568416) on 2020-01-08, in Japan (EPI_ISL_2671842) on 2020-01-09, in India on 2020-01-12, and in Africa (EPI_ISL_2716636) on 2020-01-14. This variant was collected in Zhejiang (EPI_ISL_422425) and Sichuan (EPI_ISL_451345) provinces of China on 2020-01-24.

The S_614N had the third largest number with 200 genomes. It was first collected in England on 2020-03-24. The S_614S was characterized by 61 genomes, while the other genotypes had very few numbers (Table 2).

Table 2. Alleles of the amino acid 614 in the spike protein (spike_614) of SARS-CoV-2.

Allele	Sequence (605-623)	Genomes	Percentage	First collect	Collection place
S_TG13	SNQVAVLYQ**D**VNCTEVPVA	reference		2013-07-24	Yunnan, China
S_wiv04	SNQVAVLYQ**D**VNCTEVPVA	reference		2019-12-30	Wuhan, China
S_614G	SNQVAVLYQ**G**VNCTEVPVA	2,846,092	99.293	2020-01-01	France, Argentina
S_614D	SNQVAVLYQ**D**VNCTEVPVA	19,992	0.697	2019-12-24	Wuhan, China
S_614N	SNQVAVLYQ**N**VNCTEVPVA	200	0.00698	2020-03-24	England
S_614S	SNQVAVLYQ**S**VNCTEVPVA	61	0.00213	2020-05-21	USA
S_614A	SNQVAVLYQ**A**VNCTEVPVA	11	0.00038	2020-07-13	South Korea
S_614V	SNQVAVLYQ**V**VNCTEVPVA	2	0.00007	2020-07-10	USA
S_614C	SNQVAVLYQ**C**VNCTEVPVA	1	0.00003	2021-07-30	Cambodia
unknown		274,267			
Total		3,140,626			

Mutation accumulation pattern and phylogenetic analysis of Spike variants

Similar to NS8 gene, S_614G and S_614D constituted 99.99% of the sequenced genomes (Supplementary Table S4). At the same time, S_614G was dominant over S_614D all the time. The genome number of S_614G increased exponentially from December 2019 to March 2021, and kept at high numbers (Fig.3A), while S_614D reached a peak in March 2020, and remained at low numbers for the rest of time (Fig. 3B， supplementary Table S4). Regression analysis shows that the genome number of S_614G is positively correlated with the number of confirmed cases globally with a coefficient r= 0.69 (p= 0.0031, Fig.3C), while the genome number of S_614D had no correlation (r= -0.39, p=0.14, Fig.3D). Therefore, S_614G was the main haplotype driving the growth of the confirmed cases worldwide.

The nucleotide diversity of S_614G genomes was accumulated linearly in 2020 with

R^2=0.9733 (Fig. 3E), while the nucleotide diversity of S_614D increased linearly from December 2019 to August 2020, but decreased sharply in October 2020 (Fig.3F), which was probably resulted by the strict lockdown in China, since this variant mainly present in China (see below).

Phylogenetic analysis using genome sequences of S_614 alleles indicated that S_614N and S_614S were the closest relatives to the root in the ML tree (Fig. 3G), however, consistent results were not obtained by using the NJ and ME methods (supplementary Figure S4), and/or by the phylogenenies of peptide sequences (Fig. 3H, supplementary Figure S5). Since S_614S is used in the bat spikes, S_614S should be the ancestral allele, and S_614N belongs to an allele that evolved early in evolutionary history.

Genome classification based on the two loci of NS8 and S

The six NS8_84 alleles and the seven spike_614 alleles should theoretically form 42 linkage types. However, only 16 haplotypes were identified in the 3.14 million genomes (Table 3). These haplotypes are termed by two capital letters, the first letter represents the amino acid 614 of spike, while the second letter represents the amino acid 84 of NS8. For example, S_614G+NS8_84L is termed GL, S_614G+NS8_84S is termed GS. Four main haplotypes GL, GS, DL, and DS accounted for 99.99% of the total classified genomes. The other haplotypes included GV, GI, GF, DF, DC, NL, NS, SL, AL, AS, VL, CL (Table 3).

GL haplotype was dominant over the other main haplotypes with 2.8 million sequenced genomes (99.24% of all genomes). It was dominant all the time since March 2020 with hundreds of thousands of genomes (Figure 4A). GL was first collected in Argentina (EPI_ISL_4405694) on 2020-01-01 (submission dates are 2021-09-22), and was collected in USA (EPI_ISL_3537059, EPI_ISL_3537060, EPI_ISL_3537061, EPI_ISL_3537062) on 2020-01-03. This haplotype was first collected in Sichuan (EPI_ISL_451345) and Zhejiang (EPI_ISL_422425) China on 2020-01-24.

GS haplotype was first collected in Wuhan, China (EPI_ISL_412982) on 2020-02-07, which was the only sample collected globally in February 2020. It did not show up in China again, due to its rareness. Genome analysis showed that this genotype came from a recombination between GL and DS haplotype, since the S variant came from a variation of C28144T, which is closely linked with a variation of C8782T [26], however, the variation C28144T in the GS haplotype is not linked with C8782T. The GS haplotype showed up again in Spain on 2020-03-03, and a total of 19 GS genomes were identified in 2020-03 globally, including 10 in Spain, seven in USA, one in Scotland, and one in Belgium. Its number increased to a few hundreds each month in 2021 (Fig. 4B). All these strains did not have the linked variation C8782T, therefore, they may have all come from recombination between GL and DS haplotypes, or reverse mutations.

The DS and DL genomes reached more than a thousand only in March and April 2020 (Fig. 4C), and kept at very low numbers after May 2020, possibly due to the strict lockdown policy in mainland China, since DL and DS contributed 92.6% of the genomes in mainland China before June 2020, but contributed very few proportions outside China (Fig. 4D).

The genome numbers and the first occurrence of other genotypes are provided in Table 3.

Regression analysis showed that GL was the major haplotype driving the increase of confirmed cases worldwide with a coefficient r=0.77, p= 2.571e-05 (Fig. 4E), while GS was weak (r=0.53, p= 0.01152, Fig. 4F). The genome numbers of DL and DS were negatively correlated with the confirmed cases, however, not significant at 5% significance level (r=-0.4, p>0.05, Fig. 4G, H).

Table 3. Genome numbers of Spike_614 and NS8_84 mutation linkage types and their first collection date.

	NS8_84L	NS8_84S	NS8_84V	NS8_84I	NS8_84F	NS8_84C	total
S_614G	2,831,567 2020-01-01 Argentina	1,622 2020-02-07 Wuhan,China	3 2020-09-15 South Korea	4 2021-05-05 USA	1 2021-05-06 Sweden	0	2,833,197
S_614D	12,865 2019-12-24 Wuhan,China	7,016 2019-12-30 Wuhan,China	0	0	1 2020-02-02 Sichuan,China	1 2020-03-12 USA	19,883
S_614N	9 2020-03-24 England	191 2020-06-10 Mali	0	0	0	0	200
S_614S	61 2020-05-21 USA	0	0	0	0	0	61
S_614A	4 2020-12-12 USA	7 2020-07-13 South Korea	0	0	0	0	11
S_614V	2 2020-07-10 USA	0	0	0	0	0	2
S_614C	1 2021-07-30 Cambodia	0	0	0	0	0	1
total	2,844,509	8,836	3	4	2	1	2,853,355

Diversity dynamics of the main haplotypes

Mutation accumulation curve showed that the nucleotide diversity of haplotypes GL, DL, and DS all had a period of linear increase. The difference was that GL had a high initial Pi in January 2020 followed by a sharp decrease in February 2020 before the linear increase (Figure 4I). The nonsynonymous and synonymous mutation ratios (Ka/Ks) are between 0.65 to 0.29 (Supplementary Fig.S6), indicating the presence of a strong purify selection, and many strains may have failed in further transmission till an adaptive strain emerged, which was also pointed out by Pekar and colleagues [24].

In contrast to GL haplotype, DL and DS haplotypes went straight to the linear period without initial high diversity, with correlation coefficients 0.995 (p= 3.723e-06) and 0.961 (p= 1.461e-5) for DL (Fig. 4J) and DS (Fig. 4K), respectively. These results suggested that DL and DS may have gone through the selection period when SARS-CoV-2 was discovered. However, the linear accumulation periods for DL and DS haplotypes were relatively shorter compared to that of GL, possibly due to the strong transmission control measures carried out in China, since these two haplotypes were the majority haplotypes in China (Fig. 4D).

The mutation accumulation curve of GS did not show any linear period, but its diversity was mostly high (Fig. 4L), which was even the highest among the main haplotypes most of the months in 2020 (Supplementary Figure S7). These results suggested that GS may have evolved many times by recombination or reverse mutation, and had never obtained strong transmission ability in the human population for some reason, which was also supported by the lowest number of the sequenced genomes in the four main haplotypes (Fig. 4B, supplementary Table S5).

Phylogenetic and phylodynamics analysis of the haplotypes

Among the three main haplotypes (GL, DL, and DS), DS is the closest haplotype to bat genomes, since most bat genomes have Spike-614D and NS8-84S [25]. The evolutionary trajectory of SARS-CoV-2 should be DS → DL → GL. The rest haplotypes are minor haplotypes resulting from recombination, reverse mutation, and/or mutation in the adjacent sites. Phylogenetic analysis revealed that the NS haplotype together with AS and DS are the closest to the root by using all the ML, MP, NJ, and ME methods (Fig. 5, supplementary Figure S8), followed DL and GL. Therefore, the NS, AS, and DC may have arrived from mutations of DS. The GS haplotype are distributed widely in the DS, DL, and GL haplotypes, indicating their multiple origins (recombination between GL and DS, or mutation from either DS or GL), which is in agreement with previous observation of its always high Pi (Fig. 4L). NL and DF haplotypes are located in the clade of DL, indicating its origin from DL, whereas GF, GV, SL, VL, and AL were derived from GL.

Bayesian phylodynamics analysis [40] was performed to explored the evolutionary dynamics of the major haplotypes by using a Bayesian Skyline approach. Surprisingly, the newest haplotype GL in the main haplotypes was inferred to have the oldest tMRCA with a mean of May 1 (95% HPD interval: Feb 8 to Aug 4 2019), whereas the oldest haplotype DS had the newest tMRCA with a mean of October 17 2019 (95% HPD: September 11 to November 23 2019), while the tMRCA of DL was inferred to locate in September with a mean of September 20 (95% HPD: August 11 to October 24 2019). The tMRCA of the three haplotypes had a mean of February 18 (95% HPD: Nov 5 2018 to Jun 4 2019). These results indicate that the common ancestor of the three haplotypes may have evolved before February 2019, and the three haplotypes co-existed before April 2019. The ancestor strains of DS and DL that gave birth to GL have been extinct upon the outbreak of COVID-19, therefore, much recent tMRCAs were estimated by using the extant genome sequences.

Proposed evolutionary trajectory of SARS-CoV-2

Based on the results of mutation accumulation curves, phylogenetic and phylodynamics analysis, and combined with codon usage of the mutants by single nucleotide mutation principle, we propose a possible evolutionary trajectory of SARS-CoV-2 haplotypes as shown in Figure 7.

The DS haplotypes was the ancestral haplotype of SARS-CoV-2, and may have occurred in January 2019, and evolved to DL in February 2019, and further evolved to GL in April 2019. The ancestral strains that gave birth to GL went distinct and replaced by the more adapted newcomer at the place of its origin. However, they arrived and evolved into toxic strains in China where the GL strains did not exist, and ignited the COVID-19 pandemic in China at the end of 2019. The GL strains had spread all over the world before the pandemic began in China, and ignited the global pandemic, which had not been noticed until it was reported in China.

Discussion

Tracing the origin of SARS-CoV-2 is critical for preventing future spillover of the virus [44]. This is a routine process in infectious disease prevention and control. Despite of tremendous efforts, the origin of SARS-CoV-2 remains unclear. Bats have been recognized as the natural reservoirs of a large variety of viruses [45]. SARS-CoV-1 that caused a pandemic in China in 2003 and Middle

East Respiratory Syndrome Coronavirus (MERS-CoV) are both suggested to be originated from bats [46-48]. A bat-origin coronavirus RaTG13 has the most similar genome compared to SARS-CoV-2 with 96.2% identities on whole genome level [49]. Several other bat-origin coronavirus with highly similar genomic sequences compared to SARS-CoV-2 have also been found in different countries [50-52]. Although the receptor binding domains (RBD) of the spike proteins of RaTG13 was only 89.2% compared to SARS-CoV-2, it could bind to human ACE2 (hACE2), and RaTG13 pseudovirus could transduce cells expressing hACE2 with low efficiency [53]. Conversely, the SARS-CoV-2 spike protein RBD could bind to bACE2 from *Rhinolophus macrotis* with substantially lower affinity compared with that to hACE2, and its infectivity to host cells expressing bACE2 was confirmed with pseudotyped SARS-CoV-2 virus and SARS-CoV-2 wild virus [54]. Based on these facts, SARS-CoV-2 could have more likely originated from bats. However, intermediate hosts are needed for bat-origin coronaviruses to acquire sufficient mutations so as to infect humans [55]. Two SARS-CoV-2-related coronaviruses with RBD highly similar to that of SARS-CoV-2 were found in Malayan pangolins (*Manis javanica*)[10, 56]. However, their overall genomic similarity compared to SARS-CoV-2 was both low (<93%), which suggested that pangolins were unlikely to be the intermediate host for SARS-CoV-2. So far, the direct evolutionary progenitor of SARS-CoV-2 remains unclear, whether bats are the original reservoir hosts, how, when and where SARS-CoV-2 was transmitted from animals to humans remain mysteries.

In this study, we used two gene loci to classify the haplotypes of 3.14 million SARS-CoV-2 genomes, and identified seven S_614 alleles and six NS8_48 alleles, and 16 linkage types. Phylogenetic and phylodynamics, mutation accumulation curve and codon usage analysis proposed the evolutionary trajectory of the 16 haplotypes (Fig. 7). The DS haplotype was proposed to be the oldest haplotype and probably represents the haplotype of the most recent common ancestor (MRCA) of all SARS-CoV-2 strains. However, the ancestral strain of the DS haplotype was not found due to its low adaptability compared to GL haplotype. However, its possible occurring time was estimated by using the currently available genomes of the three main haplotypes, and the mean tMRCA was estimated to be February 18 2019. Although the estimated tMRCAs of SARS-CoV-2 found in humans means the time to the most recent common ancestor of the viral variants, and should not be interpreted as the timing of the viral jump from animal hosts to humans, the tMRCA can be regarded as a most recent time, and the actual occurring time may be much earlier. Imagine that an ancestral SARS-CoV-2 invaded human populations, e.g. 20 years ago. Since then, the viral population may have undergone a series of strain replacements due to genetic drift and selective sweep, just like the sequential rise and fall of the delta and omicron variants. The continual losses and gains of genetic diversity may result in the re-building of the extant diversity, which may result in a very recent tMRCA. Anyway, the actual time of emergence should be much earlier than the estimated tMRCA based on the extant genome diversity.

The evolutionary trajectory we proposed suggested two major onsets of the COVID-19 pandemic (Fig. 7), one was in China mainly driven by the haplotype DL, the other was driven by the haplotype GL outside China. Since GL had already spread all over the world before the pandemic was delcared, its place of origin is not known, however, this haplotype was suggested to have emerged in Europe in a recent report by Wu and colleagues [25]. The GL haplotype had little influence in China during the early phase of the pandemic due to its late arrival and the strict transmission controls in China.

Acknowledgements

We are grateful to the referees for their comments to improve the manuscript. We are also grateful to those who have contributed genome sequences to the GISAID database. This research was supported by grants from National Key R&D Program of China and the Central Public-interest Scientific Institution Basal Research Fund to J.Z. (1630052020022), and the Project of Science and Technology Department of Sichuan Provincial of China to L.Y. (2019JDJQ0035).

Author contributions: Jiaming Zhang and Lei Yao conceived and designed the experiments. Yaojia Mu, Xiaowen Hu, Ruru Deng, Xuepiao Sun, Guohui Yi, and Lei Yao performed the analysis. Xiaowen Hu wrote the personalized scripts. Jiaming Zhang wrote the draft. All authors revised and approved the manuscript.

Conflicts of interest statement: The authors declare no conflict of interest.

Supplementary data

Figure S1. Temporal curve of Theta-W (population mutation rate) of the SARS-CoV-2 globally.

Figure S2. Molecular phylogenetic trees inferred by using peptide sequences of NS8; A, NJ tree, B, ME tree, C, MP tree, D, ML tree. All ambiguous positions were removed for each sequence pair in NJ and ME tree, or positions with less than 95% site coverage were eliminated in MP and ML tree. The percentage in 1000 replicates of trees in which the associated taxa clustered together is shown next to the branches. There were a total of 121 positions (NJ and ME) or 118 positions (MP and ML) in the final dataset. Evolutionary analyses were conducted in MEGA11[36].

Figure S3. Molecular phylogenetic trees inferred by using RNA sequences of NS8; A, NJ tree, B, ME tree, C, MP tree, D, ML tree. Codon positions included were 1st+2nd+3rd. All ambiguous positions were removed for each sequence pair in NJ and ME tree, or positions with less than 95% site coverage were eliminated in MP and ML tree. The percentage in 1000 replicates of trees in which the associated taxa clustered together is shown next to the branches. There were a total of 366 positions (NJ and ME) or 354 positions (MP and ML) in the final dataset.

Figure S4. Figure S4. Molecular phylogenetic trees inferred by using RNA sequences of Spike; A, NJ tree, B, ME tree, C, MP tree, D, ML tree. Codon positions included were 1st+2nd+3rd. All ambiguous positions were removed for each sequence pair in NJ and ME tree, or positions with less than 95% site coverage were eliminated in MP and ML tree. The percentage in 1000 replicates of trees in which the associated taxa clustered together is shown next to the branches.

Figure S5. Molecular phylogenetic trees inferred by using peptide sequences of Spike; A, NJ tree, B, ME tree, C, MP tree, D, ML tree. All ambiguous positions were removed for each sequence pair in NJ and ME tree, or positions with less than 95% site coverage were eliminated in MP and ML tree. The percentage in 1000 replicates of trees in which the associated taxa clustered together is shown next to the branches. There were a total of 121 positions (NJ and ME) or 118 positions (MP and ML) in the final dataset.

Figure S6. Accumulation curves of nonsynonymous (Ka) and synonymous (Ks) mutations. A, Ka and Ks mutation rates; B, Ka/Ks ratios.

Figure S7. Temporal-dependent diversity (Pi) of four main genotypes.

Figure S8. Molecular phylogenetic trees inferred by using the genome sequence; A, ME tree, B, MP tree. All ambiguous positions were removed for each sequence pair in ME tree, or positions

with less than 95% site coverage were eliminated in MP tree. The percentage in 1000 replicates of trees in which the associated taxa clustered together is shown next to the branches.

Table S1. Temporal dependent nucleotide diversity (Pi) of genomes by collection date.

Table S2. Uneven distribution of diversity in the SARS-CoV-2 genome.

Table S3. Genome mutation accumulation of main NS8_84 genotypes.

Table S4. Genome and mutation accumulation of main spike genotypes to the GISAID database.

Table S5. Genome and mutation accumulation of main spike genotypes.

Table S6. Distribution of NS genotypes worldwide

References

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med. 2020;382(8):727-33. doi: 10.1056/NEJMoa2001017. PubMed PMID: 31978945.

2. Working Group of Novel Coronavirus PUMCH. Diagnosis and clinical management of 2019 novel coronavirus infection: an operational recommendation of Peking Union Medical College Hospital (V2.0). Zhonghua Nei Ke Za Zhi. 2020;59(3):186-8. doi: 10.3760/cma.j.issn.0578-1426.2020.03.003. PubMed PMID: 32023681.

3. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265-9. doi: 10.1038/s41586-020-2008-3. PubMed PMID: 32015508.

4. Su S, Wong G, Shi W, Liu J, Lai ACK, Zhou J, et al. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol. 2016;24(6):490-502.

5. Coronaviridae Study Group of the International Committee on Taxonomy of V. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020;5(4):536-44. doi: 10.1038/s41564-020-0695-z. PubMed PMID: 32123347.

6. Tian H, Liu Y, Li Y, Wu C-H, Chen B, Kraemer MUG, et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. Science. 2020;368(6491):638-42. doi: DOI: 10.1126/science.abb6105.

7. Chen S, Yang J, Yang W, Wang C, Barnighausen T. COVID-19 control in China during mass population movements at New Year. Lancet. 2020;395(10226):764-6. Epub 2020/02/28. doi: 10.1016/S0140-6736(20)30421-9. PubMed PMID: 32105609; PubMed Central PMCID: PMCPMC7159085.

8. Wang Q, Chen H, Shi Y, Hughes AC, Liu WJ, Jiang J, et al. Tracing the origins of SARS-CoV-2: lessons learned from the past. Cell Res. 2021;31(11):1139-41. Epub 2021/10/01. doi: 10.1038/s41422-021-

00575-w. PubMed PMID: 34588626; PubMed Central PMCID: PMCPMC8480455.

9. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579:270-3.

10. Zhang T, Wu Q, Zhang Z. Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. Curr Biol. 2020. doi: 10.1016/j.cub.2020.03.022. PubMed PMID: 32197085.

11. Ji W, Wang W, Zhao X, Zai J, Li X. Cross-species transmission of the newly identified coronavirus 2019-nCoV. J Med Virol. 2020;92(4):433-40. doi: 10.1002/jmv.25682. PubMed PMID: 31967321; PubMed Central PMCID: PMCPMC7138088.

12. Liu Z, Xiao X, Wei X, Li J, Yang J, Tan H, et al. Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2. J Med Virol. 2020; 92(6):595-601. doi: 10.1002/jmv.25726. PubMed PMID: 32100877.

13. Luan J, Jin X, Lu Y, Zhang L. SARS-CoV-2 spike protein favors ACE2 from Bovidae and Cricetidae. J Med Virol. 2020;92(9):1649-56. doi: 10.1002/jmv.25817. PubMed PMID: 32239522; PubMed Central PMCID: PMCPMC7228376.

14. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020;395(10223):497-506. doi: 10.1016/S0140-6736(20)30183-5. PubMed PMID: 31986264.

15. 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;395(10224):565-74. doi: 10.1016/S0140-6736(20)30251-8. PubMed PMID: 32007145.

16. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med. 2020;382(13):1199-207. Epub 2020/01/30. doi: 10.1056/NEJMoa2001316. PubMed PMID: 31995857; PubMed Central PMCID: PMCPMC7121484.

17. Zhang X, Tan Y, Ling Y, Lu G, Liu F, Yi Z, et al. Viral and host factors related to the clinical outcome of COVID-19. Nature. 2020;583(7816):437-40. Epub 2020/05/21. doi: 10.1038/s41586-020-2355-0. PubMed PMID: 32434211.

18. Wang H, Zhao W. WHO-Convened Global Study of Origins of SARS-CoV-2: China Part (Text Extract). Infectious Diseases & Immunity. 2021;1(3):125-32. doi: doi: 10.1097/ID9.0000000000000017.

19. Hill V, Rambaut A. Phylodynamic analysis of SARS-CoV-2 | Update 2020-03-06. 2020.

20. Holmes EC, Goldstein SA, Rasmussen AL, Robertson DL, Crits-Christoph A, Wertheim JO, et al. The origins of SARS-CoV-2: A critical review. Cell. 2021;184(19):4848-56. Epub 2021/09/05. doi: 10.1016/j.cell.2021.08.017. PubMed PMID: 34480864; PubMed Central PMCID: PMCPMC8373617.

21. Nishiura H, Linton NM, Akhmetzhanov AR. Initial cluster of novel coronavirus (2019-nCoV) infections in Wuhan, China is consistent with substantial human-to-human transmission. J Clin Med. 2020;9(2):488;doi:10.3390/jcm9020488. doi: 10.3390/jcm9020488. PubMed PMID: 32054045.

22. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol.

2019;17(3):181-92. Epub 2018/12/12. doi: 10.1038/s41579-018-0118-9. PubMed PMID: 30531947; PubMed Central PMCID: PMCPMC7097006.

23. Li J, Lai S, Gao GF, Shi W. The emergence, genomic diversity and global spread of SARS-CoV-2. Nature. 2021;600(7889):408-18. Epub 2021/12/10. doi: 10.1038/s41586-021-04188-6. PubMed PMID: 34880490.

24. Pekar J, Worobey M, Moshiri N, Scheffler K, Wertheim JO. Timing the SARS-CoV-2 index case in Hubei province. Science. 2021;372(6540):412-7. Epub 2021/03/20. doi: 10.1126/science.abf8003. PubMed PMID: 33737402; PubMed Central PMCID: PMCPMC8139421.

25. Ruan Y, Wen H, Hou M, He Z, Lu X, Xue Y, et al. The twin-beginnings of COVID-19 in Asia and Europe – One prevails quickly. National Science Review. 2021. doi: 10.1093/nsr/nwab223.

26. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev. 2020;7(6):1012-23. Epub 2020/06/01. doi: 10.1093/nsr/nwaa036. PubMed PMID: 34676127; PubMed Central PMCID: PMCPMC7107875.

27. MacLean OA, Orton RJ, Singer JB, Robertson DL. No evidence for distinct types in the evolution of SARS-CoV-2. Virus Evolution. 2020;6(1). doi: 10.1093/ve/veaa034.

28. Davis C, Logan N, Tyson G, Orton R, Harvey WT, Perkins JS, et al. Reduced neutralisation of the Delta (B.1.617.2) SARS-CoV-2 variant of concern following vaccination. PLoS Pathog. 2021;17(12):e1010022. Epub 2021/12/03. doi: 10.1371/journal.ppat.1010022. PubMed PMID: 34855916; PubMed Central PMCID: PMCPMC8639073.

29. Hanada K, Suzuki Y, Gojobori T. A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. Mol Biol Evol. 2004;21(6):1074-80. doi: 10.1093/molbev/msh109. PubMed PMID: 15014142; PubMed Central PMCID: PMCPMC7107514.

30. Lai A, Bergna A, Acciarri C, Galli M, Zehender G. Early phylogenetic estimate of the effective reproduction number of SARS-CoV-2. J Med Virol. 2020:DOI:10.1002/jmv.25723. doi: 10.1002/jmv.25723. PubMed PMID: 32096566.

31. Moshiri N. ViralMSA: massively scalable reference-guided multiple sequence alignment of viral genomes. Bioinformatics (Oxford, England). 2021;37(5):714-6. Epub 2020/08/21. doi: 10.1093/bioinformatics/btaa743. PubMed PMID: 32814953.

32. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci. 2018;27(1):135-45. doi: 10.1002/pro.3290. PubMed PMID: 28884485; PubMed Central PMCID: PMCPMC5734385.

33. 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.

34. Pearson WH. Estimation of a correlation coefficient from an uncertainty measure. Psychometrika. 1966;31(3):421-33. Epub 1966/09/01. doi: 10.1007/BF02289473. PubMed PMID: 5221136.

35. Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File

Manipulation. PLoS One. 2016;11(10):e0163962. Epub 2016/10/06. doi: 10.1371/journal.pone.0163962. PubMed PMID: 27706213; PubMed Central PMCID: PMCPMC5051824.

36. Tamura K, Stecher G, Kumar S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol Biol Evol. 2021;38(7):3022-7. Epub 2021/04/24. doi: 10.1093/molbev/msab120. PubMed PMID: 33892491; PubMed Central PMCID: PMCPMC8233496.

37. Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol. 1999;16(1):37-48. Epub 1999/05/20. doi: 10.1093/oxfordjournals.molbev.a026036. PubMed PMID: 10331250.

38. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics (Oxford, England). 2012;28(23):3150-2. Epub 2012/10/13. doi: 10.1093/bioinformatics/bts565. PubMed PMID: 23060610; PubMed Central PMCID: PMCPMC3516142.

39. Glez-Pena D, Gomez-Blanco D, Reboiro-Jato M, Fdez-Riverola F, Posada D. ALTER: program-oriented conversion of DNA and protein alignments. Nucleic Acids Res. 2010;38(Web Server issue):W14-8. Epub 2010/05/05. doi: 10.1093/nar/gkq321. PubMed PMID: 20439312; PubMed Central PMCID: PMCPMC2896128.

40. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 2018;4(1):vey016. Epub 2018/06/27. doi: 10.1093/ve/vey016. PubMed PMID: 29942656; PubMed Central PMCID: PMCPMC6007674.

41. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst Biol. 2018;67(5):901-4. Epub 2018/05/03. doi: 10.1093/sysbio/syy032. PubMed PMID: 29718447; PubMed Central PMCID: PMCPMC6101584.

42. Hu B, Zeng LP, Yang XL, Ge XY, 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 Pathog. 2017;13(11):e1006698. Epub 2017/12/01. doi: 10.1371/journal.ppat.1006698. PubMed PMID: 29190287; PubMed Central PMCID: PMCPMC5708621.

43. Weiss SR, Leibowitz JL. Chapter 4 - Coronavirus Pathogenesis. In: Maramorosch K, Shatkin AJ, Murphy FA, editors. Advances in Virus Research. 81: Academic Press; 2011. p. 85-164.

44. Tong Y, Liu W, Liu P, Liu WJ, Wang Q, Gao GF. The origins of viruses: discovery takes time, international resources, and cooperation. Lancet. 2021;398(10309):1401-2. Epub 2021/10/04. doi: 10.1016/S0140-6736(21)02180-2. PubMed PMID: 34600605; PubMed Central PMCID: PMCPMC8483647 Correspondence.

45. Hu B, Ge X, Wang LF, Shi Z. Bat origin of human coronaviruses. Virol J. 2015;12:221. Epub 2015/12/23. doi: 10.1186/s12985-015-0422-1. PubMed PMID: 26689940; PubMed Central PMCID: PMCPMC4687304.

46. 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.

47. Lau SK, Woo PC, Li KS, Huang Y, Tsoi HW, Wong BH, et al. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. Proc Natl Acad Sci U S A. 2005;102(39):14040-5. Epub 2005/09/20. doi: 10.1073/pnas.0506735102. PubMed PMID: 16169905; PubMed Central PMCID: PMCPMC1236580.

48. Corman VM, Ithete NL, Richards LR, Schoeman MC, Preiser W, Drosten C, et al. Rooting the phylogenetic tree of middle East respiratory syndrome coronavirus by characterization of a conspecific virus from an African bat. J Virol. 2014;88(19):11297-303. Epub 2014/07/18. doi: 10.1128/JVI.01498-14. PubMed PMID: 25031349; PubMed Central PMCID: PMCPMC4178802.

49. Wrobel AG, Benton DJ, Xu P, Roustan C, Martin SR, Rosenthal PB, et al. SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. Nat Struct Mol Biol. 2020;27(8):763-7. Epub 20200709. doi: 10.1038/s41594-020-0468-7. PubMed PMID: 32647346; PubMed Central PMCID: PMCPMC7610980.

50. Murakami S, Kitamura T, Suzuki J, Sato R, Aoi T, Fujii M, et al. Detection and Characterization of Bat Sarbecovirus Phylogenetically Related to SARS-CoV-2, Japan. Emerg Infect Dis. 2020;26(12):3025-9. Epub 2020/11/22. doi: 10.3201/eid2612.203386. PubMed PMID: 33219796; PubMed Central PMCID: PMCPMC7706965.

51. Delaune D, Hul V, Karlsson EA, Hassanin A, Ou TP, Baidaliuk A, et al. A novel SARS-CoV-2 related coronavirus in bats from Cambodia. Nat Commun. 2021;12(1):6563. Epub 2021/11/11. doi: 10.1038/s41467-021-26809-4. PubMed PMID: 34753934; PubMed Central PMCID: PMCPMC8578604.

52. Zhou H, Ji J, Chen X, Bi Y, Li J, Wang Q, et al. Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. Cell. 2021;184(17):4380-91 e14. Epub 2021/06/21. doi: 10.1016/j.cell.2021.06.008. PubMed PMID: 34147139; PubMed Central PMCID: PMCPMC8188299.

53. Liu K, Pan X, Li L, Yu F, Zheng A, Du P, et al. Binding and molecular basis of the bat coronavirus RaTG13 virus to ACE2 in humans and other species. Cell. 2021;184(13):3438-51 e10. Epub 20210524. doi: 10.1016/j.cell.2021.05.031. PubMed PMID: 34139177; PubMed Central PMCID: PMCPMC8142884.

54. Liu K, Tan S, Niu S, Wang J, Wu L, Sun H, et al. Cross-species recognition of SARS-CoV-2 to bat ACE2. Proc Natl Acad Sci U S A. 2021;118(1):DOI: 10.1073/pnas.2020216118. Epub 2020/12/19. doi: 10.1073/pnas.2020216118. PubMed PMID: 33335073; PubMed Central PMCID: PMCPMC7817217.

55. Gao GF, Wang L. COVID-19 Expands Its Territories from Humans to Animals. China CDC Wkly. 2021;3(41):855-8. Epub 2021/10/28. doi: 10.46234/ccdcw2021.210. PubMed PMID: 34703641; PubMed Central PMCID: PMCPMC8521158.

56. Liu P, Jiang JZ, Wan XF, Hua Y, Li L, Zhou J, et al. Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? PLoS Pathog. 2020;16(5):e1008421. Epub 20200514. doi: 10.1371/journal.ppat.1008421. PubMed PMID: 32407364; PubMed Central PMCID:

PMCPMC7224457.

57. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 1992;8(3):275-82. Epub 1992/06/01. doi: 10.1093/bioinformatics/8.3.275. PubMed PMID: 1633570.

58. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol. 1993;10(3):512-26. Epub 1993/05/01. doi: 10.1093/oxfordjournals.molbev.a040023. PubMed PMID: 8336541.

59. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4):406-25. Epub 1987/07/01. doi: 10.1093/oxfordjournals.molbev.a040454. PubMed PMID: 3447015.

60. Tamura K, Nei M, Kumar S. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proceedings of the National Academy of Sciences (USA). 2004;101:11030-5.

Figure legend

Figure 1. Accumulation of sequenced genomes, confirmed cases and mutations of SARS-CoV-2. A, Accumulation of sequenced genomes, confirmed cases. B, Correlation of sequenced genomes and confirmed cases. C, Accumulation curve of nucleotide diversity (Pi) in genomes as calculated using pairwise deletion (PD) model; D, Accumulation curve of theta(W).

Figure 2. Mutation accumulation and phylogenetic analysis of NS8 variants in SARS-CoV-2. A and B, Genome accumulation curve of NS8_84L and NS8_84S as counted by the sample collection date; C and D, Correlation analysis between genomes of genotypes and confirmed cases; E and F, Mutation (Pi) accumulation curves of genotype NS8_84L (E) and NS8_84S (F); G, Maximum Likelihood tree of peptide sequences of NS8 as inferred by using the JTT matrix-based model [57]. The bootstrap supports by 1000 replicates are shown next to the branches. H, Maximum Likelihood tree of gene sequences of NS8 as inferred by using the Tamura-Nei model [58]. The bootstrap supports are shown above the branches. All positions with less than 95% site coverage were eliminated. Evolutionary analyses were conducted in MEGA11 [36].

Figure 3. Mutation accumulation and phylogenetic analysis of Spike variants in SARS-CoV-2. A and B, Genome accumulation curve of S_614G and S_614D as counted by the sample collection date; C and D, Correlation analysis between genomes of S_614G and S_614D and

confirmed cases; E and F, Mutation (Pi) accumulation curves of genotype S_614G (E) and S_614D (F); G, Maximum Likelihood tree of gene sequences of Spike as inferred by using the Tamura-Nei model [58]; H, Evolutionary tree of peptide sequences of Spike as inferred by using the Neighbor-Joining method [59]. The bootstrap supports by 1000 replicates are shown next to the branches. All positions with less than 95% site coverage were eliminated. Evolutionary analyses were conducted in MEGA11 [36].

Figure 4. Analysis of main haplotypes of SARS-CoV-2. A-C, Genome accumulation curves of haplotype GL (A), GS (B), DL (C), and DS (C); D, Haplotype profile in China, Europe, USA, and other parts of the world; E-H, Correlation analysis between genomes of main haplotypes and confirmed cases; I-L, Mutation (Pi) accumulation curves of main genotypes.

Figure 5. Phylogenetic analysis of main haplotypes of SARS-CoV-2. The trees were inferred by using Maximum Likelihood (A) and Neighbor-Joining (B) methods and rooted with the bat SARS-related viral genome RaTG13. For ML tree, Tamura-Nei model [58] was used, all positions with less than 95% site coverage were eliminated, and the tree with the highest log likelihood (-48997.72) is shown. For NJ tree, the evolutionary distances were computed using the Maximum Composite Likelihood method [60], and all ambiguous positions were removed for each sequence pair (pairwise deletion option). The bootstrap supports by 1000 replicates are shown next to the branches. All positions with less than 95% site coverage were eliminated. The percentages of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. Evolutionary analyses were conducted in MEGA11 [36].

Figure 6. Posterior distribution for the tMRCA of haplotypes DL, DS, GL, and their combination (DS+DL+GL). Redundant genomes were removed by using CD-HIT {Fu, 2012 #1002} with a threshhold of 99.97% identities. Inference was performed by using a strict molecular clock and a Bayesian Skyline coalescent prior in BEAST 1.10 [40], and summarized with Tracer 1.7 [41]. The mean tMRCA is shown above or beside the curves.

Figure 7. Evolutionary trajectory of SARS-CoV-2 haplotypes. The codons of S_614 and NS8_84 are provided below the haplotypes. The single nucleotides mutated from previous haplotypes were high-lighted in green. The proposed time axis is provided below.

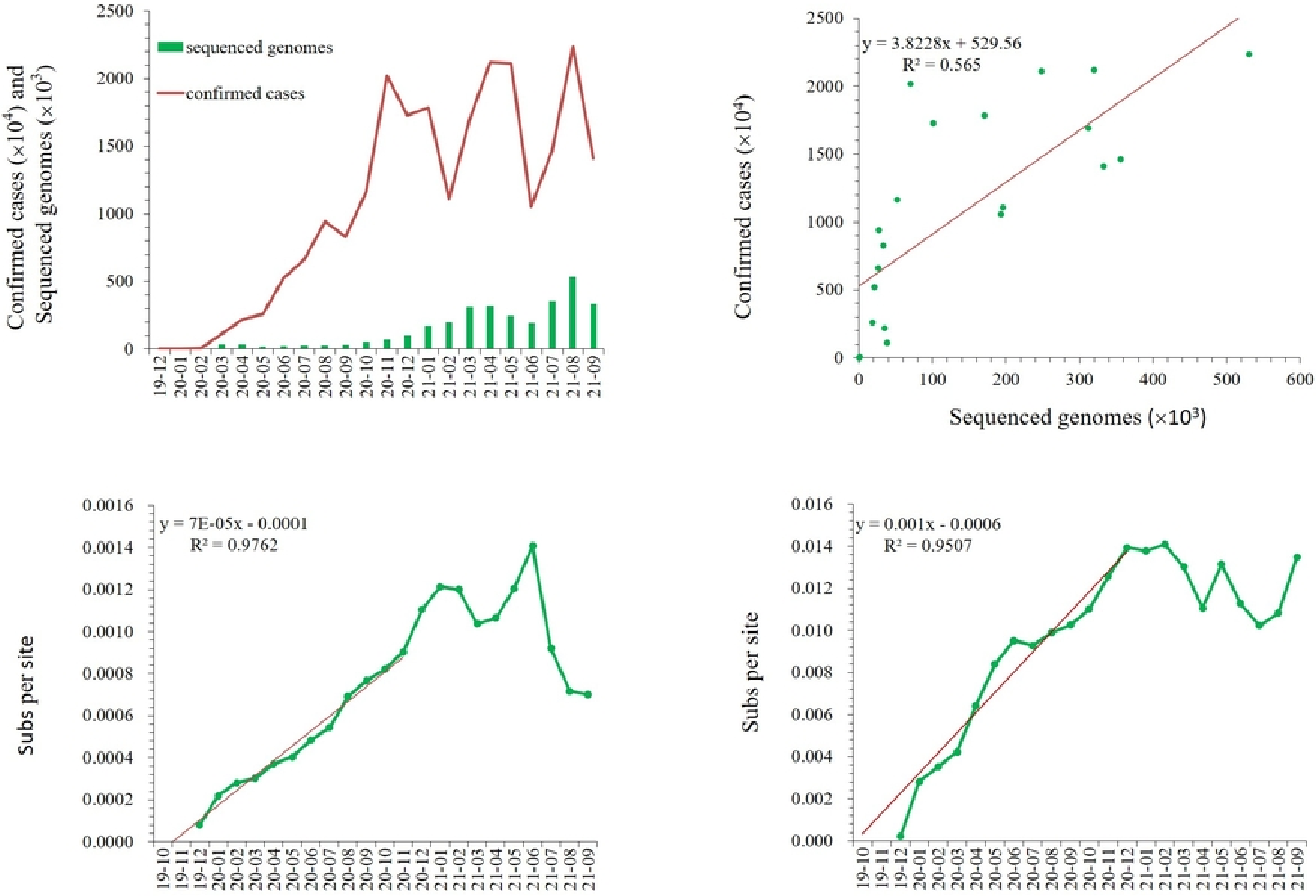
sequenced genomes
confirmed cases
Confirmed cases (×10⁴) and Sequenced genomes (×10³)
2500
2000
1500
1000
500
0
19-12
20-01
20-02
20-03
20-04
20-05
20-06
20-07
20-08
20-09
20-10
20-11
20-12
21-01
21-02
21-03
21-04
21-05
21-06
21-07
21-08
21-09
y = 3.8228x + 529.56
R² = 0.565
Confirmed cases (×10⁴)
Sequenced genomes (×10³)
0
100
200
300
400
500
600
y = 7E-05x - 0.0001
R² = 0.9762
Subs per site
0.0016
0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
19-10
19-11
y = 0.001x - 0.0006
R² = 0.9507
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000

Figure 1

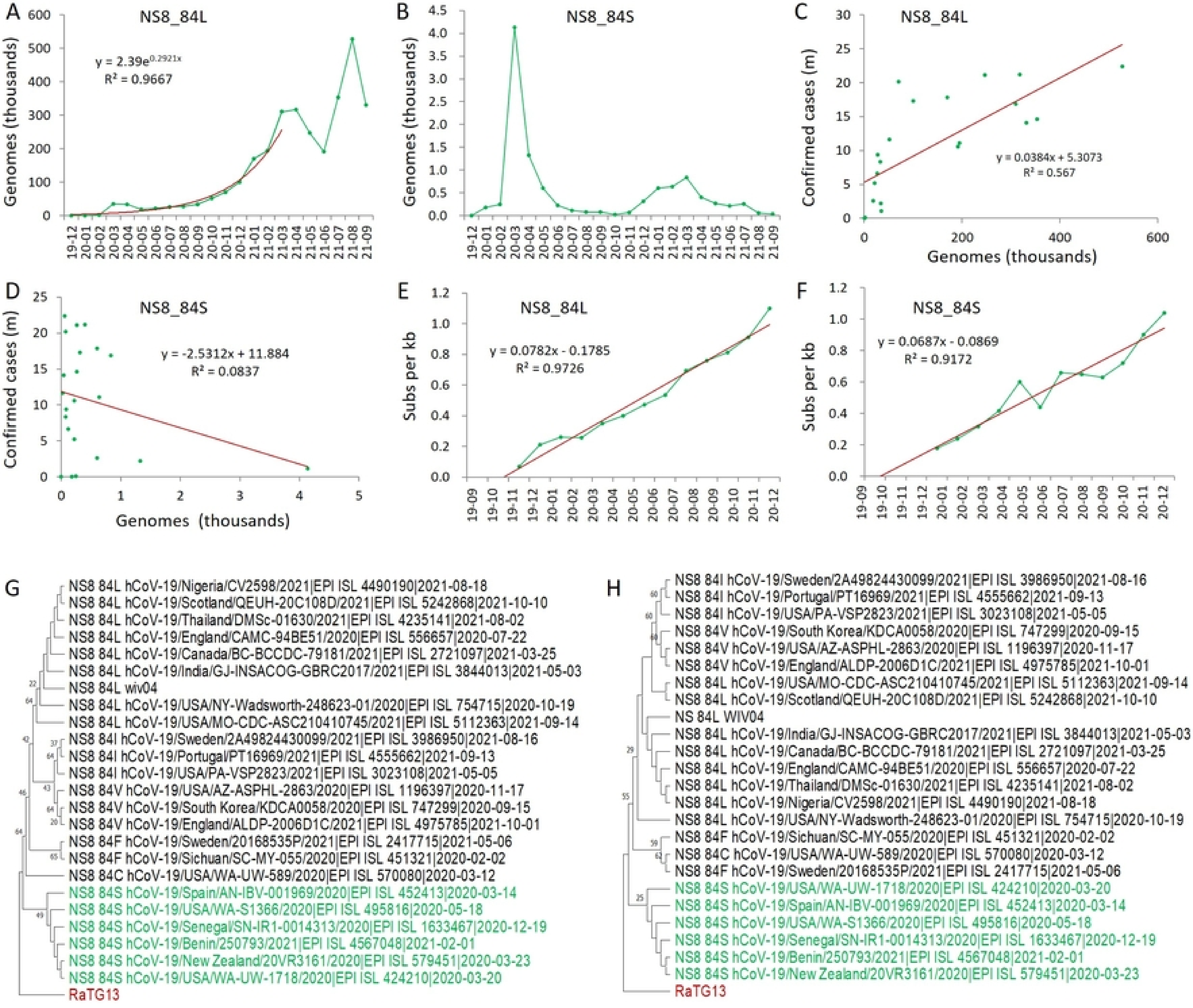
A
NS8_84L
Genomes (thousands)
0
100
200
300
400
500
600
y = 2.39e^0.2921x
R² = 0.9667
19-12
20-01
20-02
20-03
20-04
20-05
20-06
20-07
20-08
20-09
20-10
20-11
20-12
21-01
21-02
21-03
21-04
21-05
21-06
21-07
21-08
21-09
B
NS8_84S
Genomes (thousands)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
19-12
20-01
20-02
20-03
20-04
20-05
20-06
20-07
20-08
20-09
20-10
20-11
20-12
21-01
21-02
21-03
21-04
21-05
21-06
21-07
21-08
21-09
C
NS8_84L
Confirmed cases (m)
0
5
10
15
20
25
30
y = 0.0384x + 5.3073
R² = 0.567
0
200
400
600
Genomes (thousands)
D
NS8_84S
Confirmed cases (m)
0
5
10
15
20
25
y = -2.5312x + 11.884
R² = 0.0837
0
1
2
3
4
5
Genomes (thousands)
E
NS8_84L
Subs per kb
0.0
0.2
0.4
0.6
0.8
1.0
1.2
y = 0.0782x - 0.1785
R² = 0.9726
19-09
19-10
19-11
19-12
20-01
20-02
20-03
20-04
20-05
20-06
20-07
20-08
20-09
20-10
20-11
20-12
F
NS8_84S
Subs per kb
0.0
0.2
0.4
0.6
0.8
1.0
1.2
y = 0.0687x - 0.0869
R² = 0.9172
19-09
19-10
19-11
19-12
20-01
20-02
20-03
20-04
20-05
20-06
20-07
20-08
20-09
20-10
20-11
20-12
G
NS8 84L hCoV-19/Nigeria/CV2598/2021|EPI ISL 4490190|2021-08-18
NS8 84L hCoV-19/Scotland/QEUH-20C108D/2021|EPI ISL 5242868|2021-10-10
NS8 84L hCoV-19/Thailand/DMSc-01630/2021|EPI ISL 4235141|2021-08-02
NS8 84L hCoV-19/England/CAMC-94BE51/2020|EPI ISL 556657|2020-07-22
NS8 84L hCoV-19/Canada/BC-BCCDC-79181/2021|EPI ISL 2721097|2021-03-25
NS8 84L hCoV-19/India/GJ-INSACOG-GBRC2017/2021|EPI ISL 3844013|2021-05-03
NS8 84L wiv04
NS8 84L hCoV-19/USA/NY-Wadsworth-248623-01/2020|EPI ISL 754715|2020-10-19
NS8 84L hCoV-19/USA/MO-CDC-ASC210410745/2021|EPI ISL 5112363|2021-09-14
NS8 84I hCoV-19/Sweden/2A49824430099/2021|EPI ISL 3986950|2021-08-16
NS8 84I hCoV-19/Portugal/PT16969/2021|EPI ISL 4555662|2021-09-13
NS8 84I hCoV-19/USA/PA-VSP2823/2021|EPI ISL 3023108|2021-05-05
NS8 84V hCoV-19/USA/AZ-ASPHL-2863/2020|EPI ISL 1196397|2020-11-17
NS8 84V hCoV-19/South Korea/KDCA0058/2020|EPI ISL 747299|2020-09-15
NS8 84V hCoV-19/England/ALDP-2006D1C/2021|EPI ISL 4975785|2021-10-01
NS8 84F hCoV-19/Sweden/20168535P/2021|EPI ISL 2417715|2021-05-06
NS8 84F hCoV-19/Sichuan/SC-MY-055/2020|EPI ISL 451321|2020-02-02
NS8 84C hCoV-19/USA/WA-UW-589/2020|EPI ISL 570080|2020-03-12
NS8 84S hCoV-19/Spain/AN-IBV-001969/2020|EPI ISL 452413|2020-03-14
NS8 84S hCoV-19/USA/WA-S1366/2020|EPI ISL 495816|2020-05-18
NS8 84S hCoV-19/Senegal/SN-IR1-0014313/2020|EPI ISL 1633467|2020-12-19
NS8 84S hCoV-19/Benin/250793/2021|EPI ISL 4567048|2021-02-01
NS8 84S hCoV-19/New Zealand/20VR3161/2020|EPI ISL 579451|2020-03-23
NS8 84S hCoV-19/USA/WA-UW-1718/2020|EPI ISL 424210|2020-03-20
RaTG13
22
64
42
37
64
43
46
64
20
64
65
49
H
NS8 84I hCoV-19/Sweden/2A49824430099/2021|EPI ISL 3986950|2021-08-16
NS8 84I hCoV-19/Portugal/PT16969/2021|EPI ISL 4555662|2021-09-13
NS8 84I hCoV-19/USA/PA-VSP2823/2021|EPI ISL 3023108|2021-05-05
NS8 84V hCoV-19/South Korea/KDCA0058/2020|EPI ISL 747299|2020-09-15
NS8 84V hCoV-19/USA/AZ-ASPHL-2863/2020|EPI ISL 1196397|2020-11-17
NS8 84V hCoV-19/England/ALDP-2006D1C/2021|EPI ISL 4975785|2021-10-01
NS8 84L hCoV-19/USA/MO-CDC-ASC210410745/2021|EPI ISL 5112363|2021-09-14
NS8 84L hCoV-19/Scotland/QEUH-20C108D/2021|EPI ISL 5242868|2021-10-10
NS 84L WIV04
NS8 84L hCoV-19/India/GJ-INSACOG-GBRC2017/2021|EPI ISL 3844013|2021-05-03
NS8 84L hCoV-19/Canada/BC-BCCDC-79181/2021|EPI ISL 2721097|2021-03-25
NS8 84L hCoV-19/England/CAMC-94BE51/2020|EPI ISL 556657|2020-07-22
NS8 84L hCoV-19/Thailand/DMSc-01630/2021|EPI ISL 4235141|2021-08-02
NS8 84L hCoV-19/Nigeria/CV2598/2021|EPI ISL 4490190|2021-08-18
NS8 84L hCoV-19/USA/NY-Wadsworth-248623-01/2020|EPI ISL 754715|2020-10-19
NS8 84F hCoV-19/Sichuan/SC-MY-055/2020|EPI ISL 451321|2020-02-02
NS8 84C hCoV-19/USA/WA-UW-589/2020|EPI ISL 570080|2020-03-12
NS8 84F hCoV-19/Sweden/20168535P/2021|EPI ISL 2417715|2021-05-06
NS8 84S hCoV-19/USA/WA-UW-1718/2020|EPI ISL 424210|2020-03-20
NS8 84S hCoV-19/Spain/AN-IBV-001969/2020|EPI ISL 452413|2020-03-14
NS8 84S hCoV-19/USA/WA-S1366/2020|EPI ISL 495816|2020-05-18
NS8 84S hCoV-19/Senegal/SN-IR1-0014313/2020|EPI ISL 1633467|2020-12-19
NS8 84S hCoV-19/Benin/250793/2021|EPI ISL 4567048|2021-02-01
NS8 84S hCoV-19/New Zealand/20VR3161/2020|EPI ISL 579451|2020-03-23
RaTG13
60
60
60
29
55
59
62
25

Figure 2

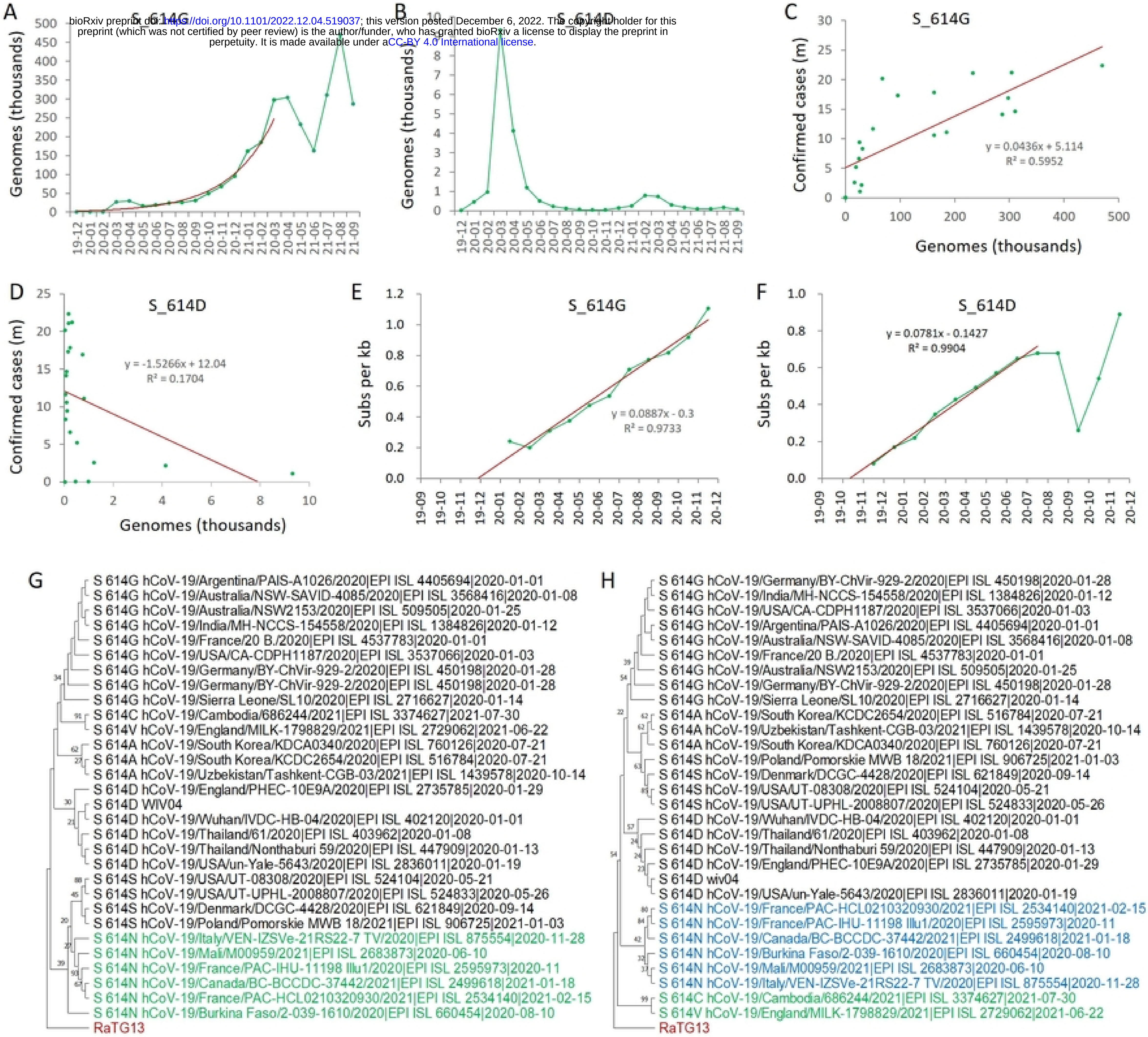
A
S_614G
Genomes (thousands)
500
450
400
350
300
250
200
150
100
50
0
19-12 20-01 20-02 20-03 20-04 20-05 20-06 20-07 20-08 20-09 20-10 20-11 20-12 21-01 21-02 20-03 20-04 21-05 21-06 21-07 21-08 21-09
B
S_614D
Genomes (thousands)
10 9 8 7 6 5 4 3 2 1 0
19-12 20-01 20-02 20-03 20-04 20-05 20-06 20-07 20-08 20-09 20-10 20-11 20-12 21-01 21-02 20-03 20-04 21-05 21-06 21-07 21-08 21-09
C
S_614G
Confirmed cases (m)
30 25 20 15 10 5 0
y = 0.0436x + 5.114
R² = 0.5952
0 100 200 300 400 500
Genomes (thousands)
D
S_614D
Confirmed cases (m)
25 20 15 10 5 0
y = -1.5266x + 12.04
R² = 0.1704
0 2 4 6 8 10
Genomes (thousands)
E
S_614G
Subs per kb
1.2 1.0 0.8 0.6 0.4 0.2 0.0
y = 0.0887x - 0.3
R² = 0.9733
19-09 19-10 19-11 19-12 20-01 20-02 20-03 20-04 20-05 20-06 20-07 20-08 20-09 20-10 20-11 20-12
F
S_614D
Subs per kb
1.0 0.8 0.6 0.4 0.2 0.0
y = 0.0781x - 0.1427
R² = 0.9904
19-09 19-10 19-11 19-12 20-01 20-02 20-03 20-04 20-05 20-06 20-07 20-08 20-09 20-10 20-11 20-12
G
S 614G hCoV-19/Argentina/PAIS-A1026/2020|EPI ISL 4405694|2020-01-01
S 614G hCoV-19/Australia/NSW-SAVID-4085/2020|EPI ISL 3568416|2020-01-08
S 614G hCoV-19/Australia/NSW2153/2020|EPI ISL 509505|2020-01-25
S 614G hCoV-19/India/MH-NCCS-154558/2020|EPI ISL 1384826|2020-01-12
S 614G hCoV-19/France/20 B./2020|EPI ISL 4537783|2020-01-01
S 614G hCoV-19/USA/CA-CDPH1187/2020|EPI ISL 3537066|2020-01-03
S 614G hCoV-19/Germany/BY-ChVir-929-2/2020|EPI ISL 450198|2020-01-28
S 614G hCoV-19/Germany/BY-ChVir-929-2/2020|EPI ISL 450198|2020-01-28
S 614G hCoV-19/Sierra Leone/SL10/2020|EPI ISL 2716627|2020-01-14
S 614C hCoV-19/Cambodia/686244/2021|EPI ISL 3374627|2021-07-30
S 614V hCoV-19/England/MILK-1798829/2021|EPI ISL 2729062|2021-06-22
S 614A hCoV-19/South Korea/KDCA0340/2020|EPI ISL 760126|2020-07-21
S 614A hCoV-19/South Korea/KCDC2654/2020|EPI ISL 516784|2020-07-21
S 614A hCoV-19/Uzbekistan/Tashkent-CGB-03/2021|EPI ISL 1439578|2020-10-14
S 614D hCoV-19/England/PHEC-10E9A/2020|EPI ISL 2735785|2020-01-29
S 614D WIV04
S 614D hCoV-19/Wuhan/IVDC-HB-04/2020|EPI ISL 402120|2020-01-01
S 614D hCoV-19/Thailand/61/2020|EPI ISL 403962|2020-01-08
S 614D hCoV-19/Thailand/Nonthaburi 59/2020|EPI ISL 447909|2020-01-13
S 614D hCoV-19/USA/un-Yale-5643/2020|EPI ISL 2836011|2020-01-19
S 614S hCoV-19/USA/UT-08308/2020|EPI ISL 524104|2020-05-21
S 614S hCoV-19/USA/UT-UPHL-2008807/2020|EPI ISL 524833|2020-05-26
S 614S hCoV-19/Denmark/DCGC-4428/2020|EPI ISL 621849|2020-09-14
S 614S hCoV-19/Poland/Pomorskie MWB 18/2021|EPI ISL 906725|2021-01-03
S 614N hCoV-19/Italy/VEN-IZSVe-21RS22-7 TV/2020|EPI ISL 875554|2020-11-28
S 614N hCoV-19/Mali/M00959/2021|EPI ISL 2683873|2020-06-10
S 614N hCoV-19/France/PAC-IHU-11198 Illu1/2020|EPI ISL 2595973|2020-11
S 614N hCoV-19/Canada/BC-BCCDC-37442/2021|EPI ISL 2499618|2021-01-18
S 614N hCoV-19/France/PAC-HCL0210320930/2021|EPI ISL 2534140|2021-02-15
S 614N hCoV-19/Burkina Faso/2-039-1610/2020|EPI ISL 660454|2020-08-10
RaTG13
34 91 62 27 30 21 88 45 20 27 39 93 62
H
S 614G hCoV-19/Germany/BY-ChVir-929-2/2020|EPI ISL 450198|2020-01-28
S 614G hCoV-19/India/MH-NCCS-154558/2020|EPI ISL 1384826|2020-01-12
S 614G hCoV-19/USA/CA-CDPH1187/2020|EPI ISL 3537066|2020-01-03
S 614G hCoV-19/Argentina/PAIS-A1026/2020|EPI ISL 4405694|2020-01-01
S 614G hCoV-19/Australia/NSW-SAVID-4085/2020|EPI ISL 3568416|2020-01-08
S 614G hCoV-19/France/20 B./2020|EPI ISL 4537783|2020-01-01
S 614G hCoV-19/Australia/NSW2153/2020|EPI ISL 509505|2020-01-25
S 614G hCoV-19/Germany/BY-ChVir-929-2/2020|EPI ISL 450198|2020-01-28
S 614G hCoV-19/Sierra Leone/SL10/2020|EPI ISL 2716627|2020-01-14
S 614A hCoV-19/South Korea/KCDC2654/2020|EPI ISL 516784|2020-07-21
S 614A hCoV-19/Uzbekistan/Tashkent-CGB-03/2021|EPI ISL 1439578|2020-10-14
S 614A hCoV-19/South Korea/KDCA0340/2020|EPI ISL 760126|2020-07-21
S 614S hCoV-19/Poland/Pomorskie MWB 18/2021|EPI ISL 906725|2021-01-03
S 614S hCoV-19/Denmark/DCGC-4428/2020|EPI ISL 621849|2020-09-14
S 614S hCoV-19/USA/UT-08308/2020|EPI ISL 524104|2020-05-21
S 614S hCoV-19/USA/UT-UPHL-2008807/2020|EPI ISL 524833|2020-05-26
S 614D hCoV-19/Wuhan/IVDC-HB-04/2020|EPI ISL 402120|2020-01-01
S 614D hCoV-19/Thailand/61/2020|EPI ISL 403962|2020-01-08
S 614D hCoV-19/Thailand/Nonthaburi 59/2020|EPI ISL 447909|2020-01-13
S 614D hCoV-19/England/PHEC-10E9A/2020|EPI ISL 2735785|2020-01-29
S 614D wiv04
S 614D hCoV-19/USA/un-Yale-5643/2020|EPI ISL 2836011|2020-01-19
S 614N hCoV-19/France/PAC-HCL0210320930/2021|EPI ISL 2534140|2021-02-15
S 614N hCoV-19/France/PAC-IHU-11198 Illu1/2020|EPI ISL 2595973|2020-11
S 614N hCoV-19/Canada/BC-BCCDC-37442/2021|EPI ISL 2499618|2021-01-18
S 614N hCoV-19/Burkina Faso/2-039-1610/2020|EPI ISL 660454|2020-08-10
S 614N hCoV-19/Mali/M00959/2021|EPI ISL 2683873|2020-06-10
S 614N hCoV-19/Italy/VEN-IZSVe-21RS22-7 TV/2020|EPI ISL 875554|2020-11-28
S 614C hCoV-19/Cambodia/686244/2021|EPI ISL 3374627|2021-07-30
S 614V hCoV-19/England/MILK-1798829/2021|EPI ISL 2729062|2021-06-22
RaTG13
39 54 22 62 62 63 85 57 24 24 23 54 80 84 42 32 33 99

Figure 3

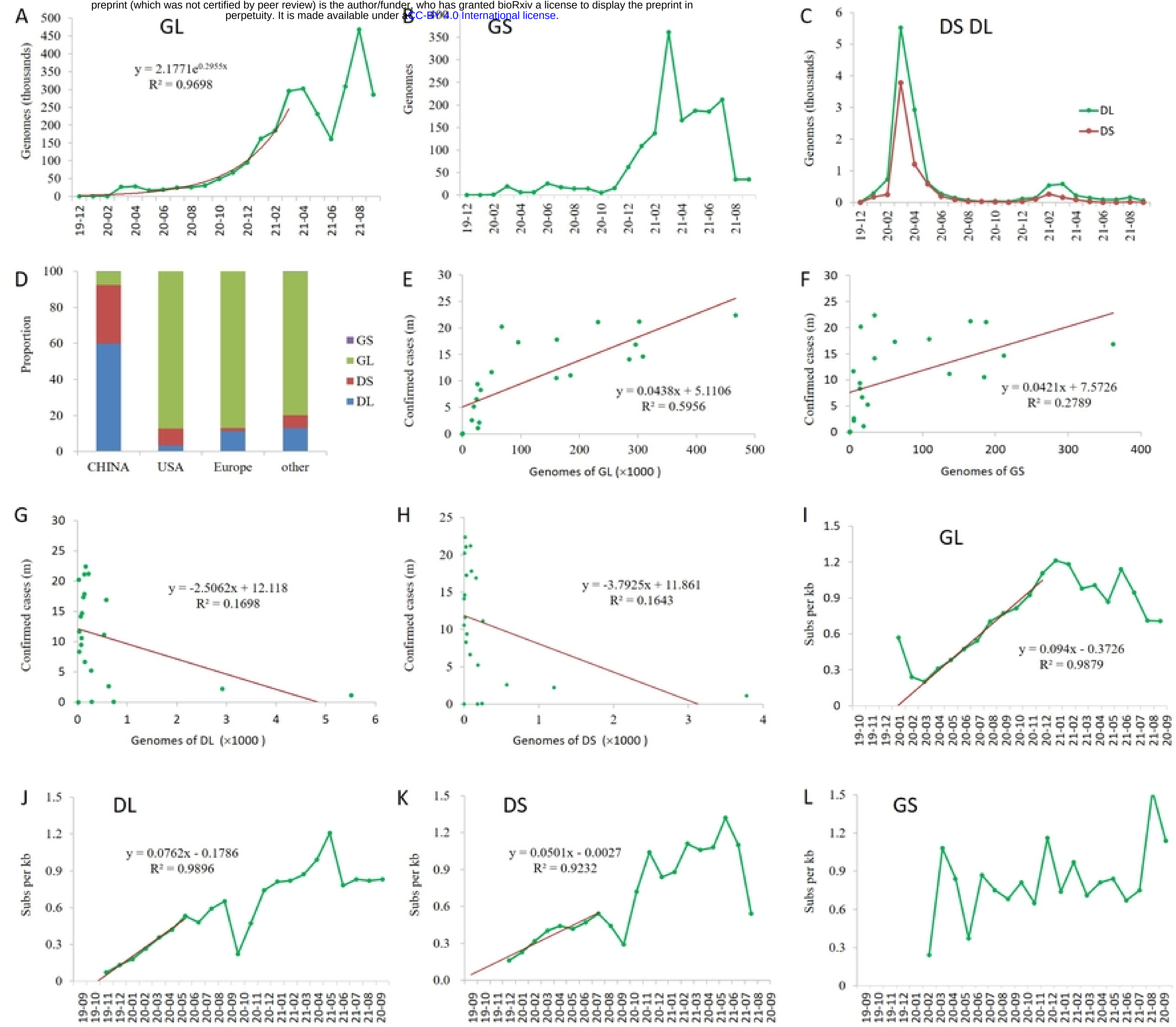
A
GL
Genomes (thousands)
y = 2.1771e$^{0.2955x}$
R² = 0.9698
19-12 20-02 20-04 20-06 20-08 20-10 20-12 21-02 21-04 21-06 21-08
B
GS
Genomes
C
DS DL
Genomes (thousands)
DL
DS
D
Proportion
GS
GL
DS
DL
CHINA
USA
Europe
other
E
Confirmed cases (m)
y = 0.0438x + 5.1106
R² = 0.5956
Genomes of GL (×1000)
F
Confirmed cases (m)
y = 0.0421x + 7.5726
R² = 0.2789
Genomes of GS
G
Confirmed cases (m)
y = -2.5062x + 12.118
R² = 0.1698
Genomes of DL (×1000)
H
Confirmed cases (m)
y = -3.7925x + 11.861
R² = 0.1643
Genomes of DS (×1000)
I
GL
Subs per kb
y = 0.094x - 0.3726
R² = 0.9879
J
DL
Subs per kb
y = 0.0762x - 0.1786
R² = 0.9896
K
DS
Subs per kb
y = 0.0501x - 0.0027
R² = 0.9232
L
GS
Subs per kb
19-09 19-10 19-11 19-12 20-01 20-02 20-03 20-04 20-05 20-06 20-07 20-08 20-09 20-10 20-11 20-12 21-01 21-02 21-03 21-04 21-05 21-06 21-07 21-08 21-09

Figure 4

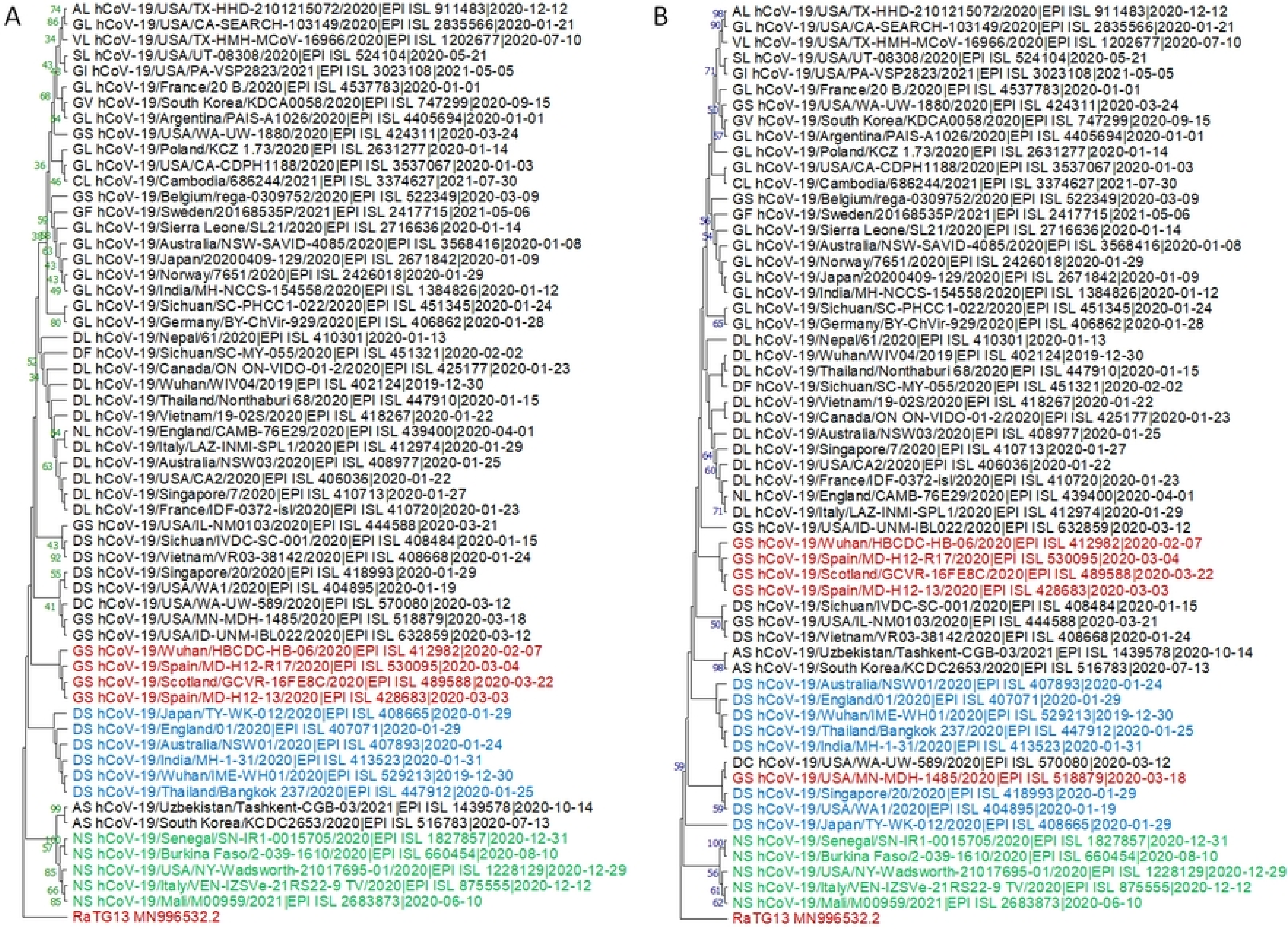
A
AL hCoV-19/USA/TX-HHD-2101215072/2020|EPI ISL 911483|2020-12-12
GL hCoV-19/USA/CA-SEARCH-103149/2020|EPI ISL 2835566|2020-01-21
VL hCoV-19/USA/TX-HMH-MCoV-16966/2020|EPI ISL 1202677|2020-07-10
SL hCoV-19/USA/UT-08308/2020|EPI ISL 524104|2020-05-21
GI hCoV-19/USA/PA-VSP2823/2021|EPI ISL 3023108|2021-05-05
GL hCoV-19/France/20 B./2020|EPI ISL 4537783|2020-01-01
GV hCoV-19/South Korea/KDCA0058/2020|EPI ISL 747299|2020-09-15
GL hCoV-19/Argentina/PAIS-A1026/2020|EPI ISL 4405694|2020-01-01
GS hCoV-19/USA/WA-UW-1880/2020|EPI ISL 424311|2020-03-24
GL hCoV-19/Poland/KCZ 1.73/2020|EPI ISL 2631277|2020-01-14
GL hCoV-19/USA/CA-CDPH1188/2020|EPI ISL 3537067|2020-01-03
CL hCoV-19/Cambodia/686244/2021|EPI ISL 3374627|2021-07-30
GS hCoV-19/Belgium/rega-0309752/2020|EPI ISL 522349|2020-03-09
GF hCoV-19/Sweden/20168535P/2021|EPI ISL 2417715|2021-05-06
GL hCoV-19/Sierra Leone/SL21/2020|EPI ISL 2716636|2020-01-14
GL hCoV-19/Australia/NSW-SAVID-4085/2020|EPI ISL 3568416|2020-01-08
GL hCoV-19/Japan/20200409-129/2020|EPI ISL 2671842|2020-01-09
GL hCoV-19/Norway/7651/2020|EPI ISL 2426018|2020-01-29
GL hCoV-19/India/MH-NCCS-154558/2020|EPI ISL 1384826|2020-01-12
GL hCoV-19/Sichuan/SC-PHCC1-022/2020|EPI ISL 451345|2020-01-24
GL hCoV-19/Germany/BY-ChVir-929/2020|EPI ISL 406862|2020-01-28
DL hCoV-19/Nepal/61/2020|EPI ISL 410301|2020-01-13
DF hCoV-19/Sichuan/SC-MY-055/2020|EPI ISL 451321|2020-02-02
DL hCoV-19/Canada/ON ON-VIDO-01-2/2020|EPI ISL 425177|2020-01-23
DL hCoV-19/Wuhan/WIV04/2019|EPI ISL 402124|2019-12-30
DL hCoV-19/Thailand/Nonthaburi 68/2020|EPI ISL 447910|2020-01-15
DL hCoV-19/Vietnam/19-02S/2020|EPI ISL 418267|2020-01-22
NL hCoV-19/England/CAMB-76E29/2020|EPI ISL 439400|2020-04-01
DL hCoV-19/Italy/LAZ-INMI-SPL1/2020|EPI ISL 412974|2020-01-29
DL hCoV-19/Australia/NSW03/2020|EPI ISL 408977|2020-01-25
DL hCoV-19/USA/CA2/2020|EPI ISL 406036|2020-01-22
DL hCoV-19/Singapore/7/2020|EPI ISL 410713|2020-01-27
DL hCoV-19/France/IDF-0372-isl/2020|EPI ISL 410720|2020-01-23
GS hCoV-19/USA/IL-NM0103/2020|EPI ISL 444588|2020-03-21
DS hCoV-19/Sichuan/IVDC-SC-001/2020|EPI ISL 408484|2020-01-15
DS hCoV-19/Vietnam/VR03-38142/2020|EPI ISL 408668|2020-01-24
DS hCoV-19/Singapore/20/2020|EPI ISL 418993|2020-01-29
DS hCoV-19/USA/WA1/2020|EPI ISL 404895|2020-01-19
DC hCoV-19/USA/WA-UW-589/2020|EPI ISL 570080|2020-03-12
GS hCoV-19/USA/MN-MDH-1485/2020|EPI ISL 518879|2020-03-18
GS hCoV-19/USA/ID-UNM-IBL022/2020|EPI ISL 632859|2020-03-12
GS hCoV-19/Wuhan/HBCDC-HB-06/2020|EPI ISL 412982|2020-02-07
GS hCoV-19/Spain/MD-H12-R17/2020|EPI ISL 530095|2020-03-04
GS hCoV-19/Scotland/GCVR-16FE8C/2020|EPI ISL 489588|2020-03-22
GS hCoV-19/Spain/MD-H12-13/2020|EPI ISL 428683|2020-03-03
DS hCoV-19/Japan/TY-WK-012/2020|EPI ISL 408665|2020-01-29
DS hCoV-19/England/01/2020|EPI ISL 407071|2020-01-29
DS hCoV-19/Australia/NSW01/2020|EPI ISL 407893|2020-01-24
DS hCoV-19/India/MH-1-31/2020|EPI ISL 413523|2020-01-31
DS hCoV-19/Wuhan/IME-WH01/2020|EPI ISL 529213|2019-12-30
DS hCoV-19/Thailand/Bangkok 237/2020|EPI ISL 447912|2020-01-25
AS hCoV-19/Uzbekistan/Tashkent-CGB-03/2021|EPI ISL 1439578|2020-10-14
AS hCoV-19/South Korea/KCDC2653/2020|EPI ISL 516783|2020-07-13
NS hCoV-19/Senegal/SN-IR1-0015705/2020|EPI ISL 1827857|2020-12-31
NS hCoV-19/Burkina Faso/2-039-1610/2020|EPI ISL 660454|2020-08-10
NS hCoV-19/USA/NY-Wadsworth-21017695-01/2020|EPI ISL 1228129|2020-12-29
NS hCoV-19/Italy/VEN-IZSVe-21RS22-9 TV/2020|EPI ISL 875555|2020-12-12
NS hCoV-19/Mali/M00959/2021|EPI ISL 2683873|2020-06-10
RaTG13 MN996532.2
B
AL hCoV-19/USA/TX-HHD-2101215072/2020|EPI ISL 911483|2020-12-12
GL hCoV-19/USA/CA-SEARCH-103149/2020|EPI ISL 2835566|2020-01-21
VL hCoV-19/USA/TX-HMH-MCoV-16966/2020|EPI ISL 1202677|2020-07-10
SL hCoV-19/USA/UT-08308/2020|EPI ISL 524104|2020-05-21
GI hCoV-19/USA/PA-VSP2823/2021|EPI ISL 3023108|2021-05-05
GL hCoV-19/France/20 B./2020|EPI ISL 4537783|2020-01-01
GS hCoV-19/USA/WA-UW-1880/2020|EPI ISL 424311|2020-03-24
GV hCoV-19/South Korea/KDCA0058/2020|EPI ISL 747299|2020-09-15
GL hCoV-19/Argentina/PAIS-A1026/2020|EPI ISL 4405694|2020-01-01
GL hCoV-19/Poland/KCZ 1.73/2020|EPI ISL 2631277|2020-01-14
GL hCoV-19/USA/CA-CDPH1188/2020|EPI ISL 3537067|2020-01-03
CL hCoV-19/Cambodia/686244/2021|EPI ISL 3374627|2021-07-30
GS hCoV-19/Belgium/rega-0309752/2020|EPI ISL 522349|2020-03-09
GF hCoV-19/Sweden/20168535P/2021|EPI ISL 2417715|2021-05-06
GL hCoV-19/Sierra Leone/SL21/2020|EPI ISL 2716636|2020-01-14
GL hCoV-19/Australia/NSW-SAVID-4085/2020|EPI ISL 3568416|2020-01-08
GL hCoV-19/Norway/7651/2020|EPI ISL 2426018|2020-01-29
GL hCoV-19/Japan/20200409-129/2020|EPI ISL 2671842|2020-01-09
GL hCoV-19/India/MH-NCCS-154558/2020|EPI ISL 1384826|2020-01-12
GL hCoV-19/Sichuan/SC-PHCC1-022/2020|EPI ISL 451345|2020-01-24
GL hCoV-19/Germany/BY-ChVir-929/2020|EPI ISL 406862|2020-01-28
DL hCoV-19/Nepal/61/2020|EPI ISL 410301|2020-01-13
DL hCoV-19/Wuhan/WIV04/2019|EPI ISL 402124|2019-12-30
DL hCoV-19/Thailand/Nonthaburi 68/2020|EPI ISL 447910|2020-01-15
DF hCoV-19/Sichuan/SC-MY-055/2020|EPI ISL 451321|2020-02-02
DL hCoV-19/Vietnam/19-02S/2020|EPI ISL 418267|2020-01-22
DL hCoV-19/Canada/ON ON-VIDO-01-2/2020|EPI ISL 425177|2020-01-23
DL hCoV-19/Australia/NSW03/2020|EPI ISL 408977|2020-01-25
DL hCoV-19/Singapore/7/2020|EPI ISL 410713|2020-01-27
DL hCoV-19/USA/CA2/2020|EPI ISL 406036|2020-01-22
DL hCoV-19/France/IDF-0372-isl/2020|EPI ISL 410720|2020-01-23
NL hCoV-19/England/CAMB-76E29/2020|EPI ISL 439400|2020-04-01
DL hCoV-19/Italy/LAZ-INMI-SPL1/2020|EPI ISL 412974|2020-01-29
GS hCoV-19/USA/ID-UNM-IBL022/2020|EPI ISL 632859|2020-03-12
GS hCoV-19/Wuhan/HBCDC-HB-06/2020|EPI ISL 412982|2020-02-07
GS hCoV-19/Spain/MD-H12-R17/2020|EPI ISL 530095|2020-03-04
GS hCoV-19/Scotland/GCVR-16FE8C/2020|EPI ISL 489588|2020-03-22
GS hCoV-19/Spain/MD-H12-13/2020|EPI ISL 428683|2020-03-03
DS hCoV-19/Sichuan/IVDC-SC-001/2020|EPI ISL 408484|2020-01-15
GS hCoV-19/USA/IL-NM0103/2020|EPI ISL 444588|2020-03-21
DS hCoV-19/Vietnam/VR03-38142/2020|EPI ISL 408668|2020-01-24
AS hCoV-19/Uzbekistan/Tashkent-CGB-03/2021|EPI ISL 1439578|2020-10-14
AS hCoV-19/South Korea/KCDC2653/2020|EPI ISL 516783|2020-07-13
DS hCoV-19/Australia/NSW01/2020|EPI ISL 407893|2020-01-24
DS hCoV-19/England/01/2020|EPI ISL 407071|2020-01-29
DS hCoV-19/Wuhan/IME-WH01/2020|EPI ISL 529213|2019-12-30
DS hCoV-19/Thailand/Bangkok 237/2020|EPI ISL 447912|2020-01-25
DS hCoV-19/India/MH-1-31/2020|EPI ISL 413523|2020-01-31
DC hCoV-19/USA/WA-UW-589/2020|EPI ISL 570080|2020-03-12
GS hCoV-19/USA/MN-MDH-1485/2020|EPI ISL 518879|2020-03-18
DS hCoV-19/Singapore/20/2020|EPI ISL 418993|2020-01-29
DS hCoV-19/USA/WA1/2020|EPI ISL 404895|2020-01-19
DS hCoV-19/Japan/TY-WK-012/2020|EPI ISL 408665|2020-01-29
NS hCoV-19/Senegal/SN-IR1-0015705/2020|EPI ISL 1827857|2020-12-31
NS hCoV-19/Burkina Faso/2-039-1610/2020|EPI ISL 660454|2020-08-10
NS hCoV-19/USA/NY-Wadsworth-21017695-01/2020|EPI ISL 1228129|2020-12-29
NS hCoV-19/Italy/VEN-IZSVe-21RS22-9 TV/2020|EPI ISL 875555|2020-12-12
NS hCoV-19/Mali/M00959/2021|EPI ISL 2683873|2020-06-10
RaTG13 MN996532.2

Figure 5

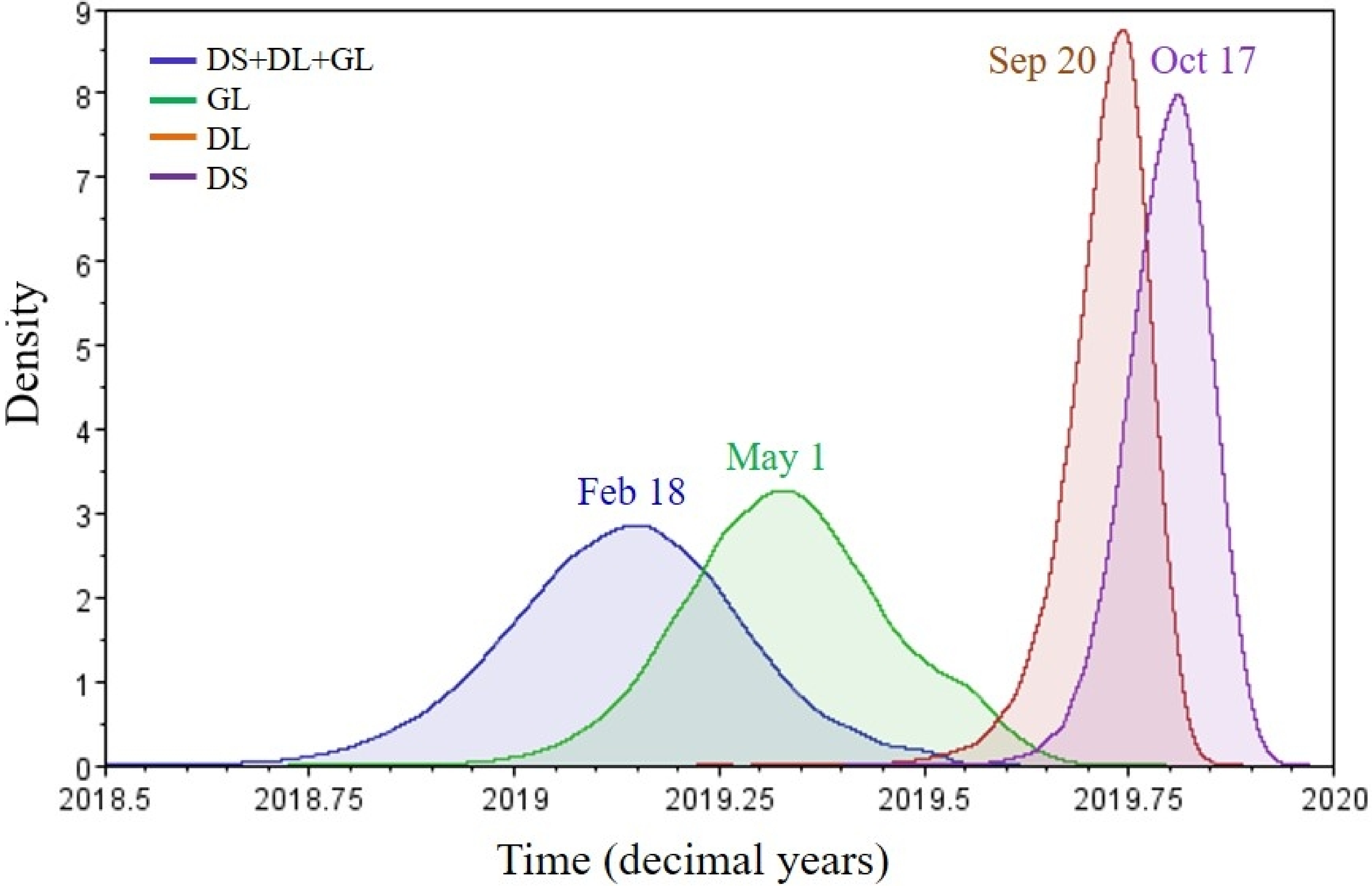
DS+DL+GL
GL
DL
DS
Sep 20
Oct 17
May 1
Feb 18
Density
0
1
2
3
4
5
6
7
8
9
2018.5
2018.75
2019
2019.25
2019.5
2019.75
2020
Time (decimal years)

Figure 6

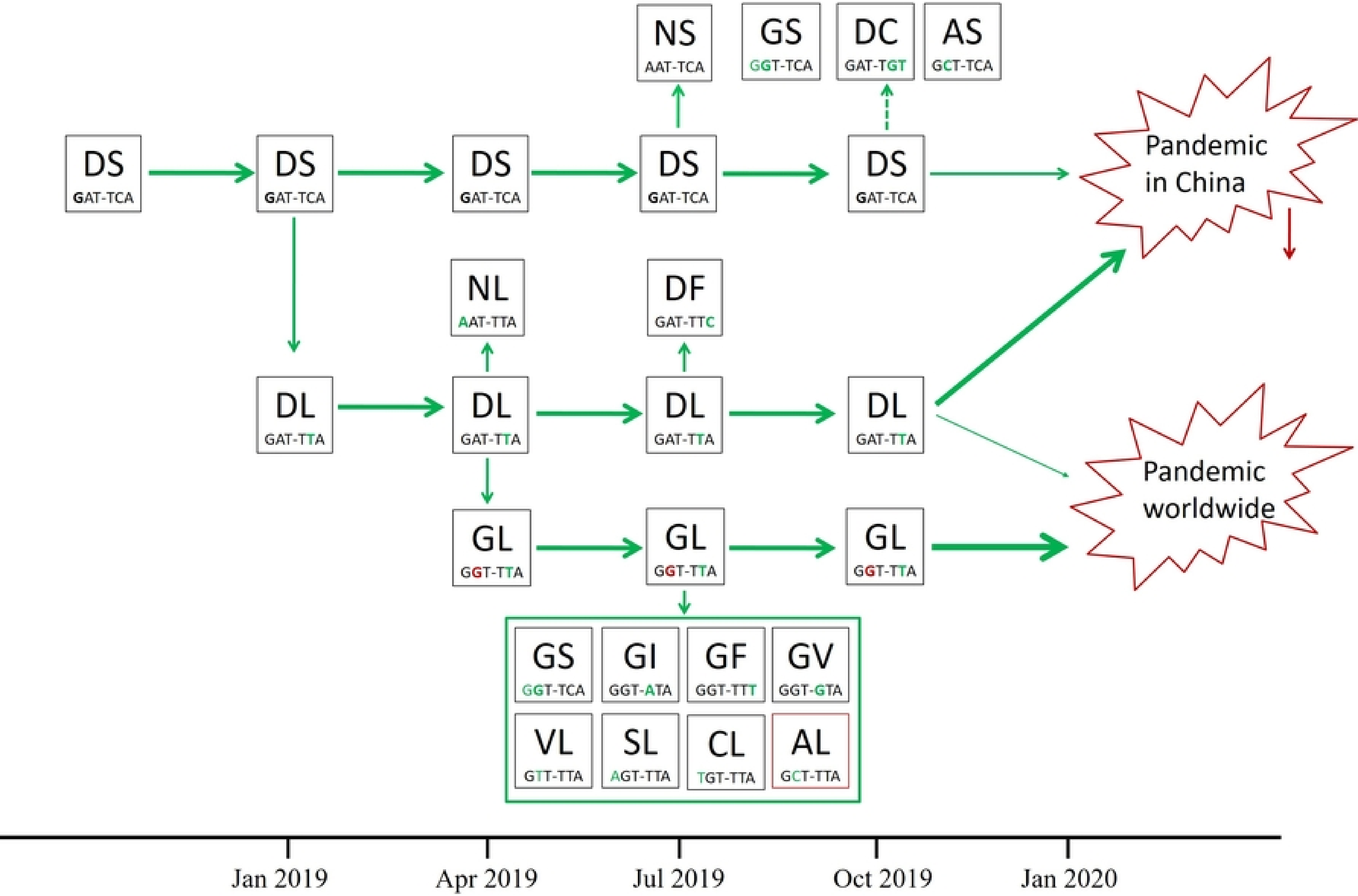
NS
AAT-TCA
GS
GGT-TCA
DC
GAT-TGT
AS
GCT-TCA
DS
GAT-TCA
DS
GAT-TCA
DS
GAT-TCA
DS
GAT-TCA
DS
GAT-TCA
Pandemic
in China
NL
AAT-TTA
DF
GAT-TTC
DL
GAT-TTA
DL
GAT-TTA
DL
GAT-TTA
DL
GAT-TTA
Pandemic
worldwide
GL
GGT-TTA
GL
GGT-TTA
GL
GGT-TTA
GS
GGT-TCA
GI
GGT-ATA
GF
GGT-TTT
GV
GGT-GTA
VL
GTT-TTA
SL
AGT-TTA
CL
TGT-TTA
AL
GCT-TTA
Jan 2019
Apr 2019
Jul 2019
Oct 2019
Jan 2020

Figure 7