Science in One Health 2 (2023) 100041

Contents lists available at ScienceDirect

Science in One Health

journal homepage: www.journals.elsevier.com/science-in-one-health

Full length article

Genome analysis of SARS-CoV-2 haplotypes: separation and parallel evolution of the major haplotypes occurred considerably earlier than their emergence in China

Siqin Guan [a,b,1], Xiaowen Hu [c,1], Guohui Yi [d], Lei Yao [e,**], Jiaming Zhang [a,b,*]

[a] Key Laboratory of Microbiology of Hainan, Institute of Tropical Bioscience and Biotechnology, Chinese Academy of Tropical Agricultural Sciences, Haikou, Hainan 571101, China

[b] College of Animal Sciences, Huazhong Agricultural University, Wuhan, Hubei Province 430070, China

[c] Institute of South Subtropical Crops, Chinese Academy of Tropical Agricultural Sciences, Zhanjiang 524013, China

[d] Public Research Laboratory, Hainan Medical University, Haikou 571199, China

[e] Sichuan Provincial Key Laboratory for Human Disease Gene Study and the Center for Medical Genetics, Department of Laboratory Medicine, Sichuan Academy of Medical Sciences & Sichuan Provincial People's Hospital, University of Electronic Science and Technology, Chengdu 610054, China

ARTICLE INFO

Keywords:
SARS-CoV-2
Haplotype
Genome classification
MRCA
Phylogenetic analysis

ABSTRACT

More than 3 years have passed since the outbreak of COVID-19 and yet, the origin of the causal virus SARS-CoV-2 remains unknown. We examined the evolutionary trajectory of SARS-CoV-2 by analyzing non-redundant genome sets classified based on six closely linked mutations. The results indicated that SARS-CoV-2 emerged in February 2019 or earlier and evolved into three main haplotypes (GL, DS, and DL) before May 2019, which then continued to evolve in parallel. The dominant haplotype GL had spread worldwide in the summer (May to July) of 2019 and then evolved into virulent strains in December 2019 that triggered the global pandemic, whereas haplotypes DL and DS arrived in China in October 2019 and caused the epidemic in China in December 2019. Therefore, haplotype GL neither originated in China nor from the viral strains that caused the epidemic in China. Accordingly, considering data solely from China would be inadequate to reveal the mysterious origin of SARS-CoV-2, emphasizing the necessity of global cooperation.

1. Introduction

Humans have been suffering from coronavirus disease 2019 (COVID-19) for more than three years [1,2]. The rapid spread of this disease caught the entire world off guard. As of July 10, 2023, there have been over 768 million confirmed cases and 6.9 million deaths associated with COVID-19 reported to the World Health Organization (WHO) (https://covid19.who.int/). However, the origin of SARS-CoV-2, which is important to solve to enable effective control of this pandemic, as well as future similar threats, remains a mystery [3–5].

The Huanan Seafood Market in Wuhan, Hubei Province, China, was once suspected to be associated with the origins of COVID-19 [6] and was closed for environmental sanitation and disinfection 3 days after the discovery of the disease, although the causal agent was then unknown. As of January 3, 2020, a total of 44 patients with pneumonia of unknown etiology had been reported to the WHO by the national authorities in China [7]. However, a few cases were not associated with the Huanan Seafood Market, and animal-to-human transmission has never been confirmed [8]. The causal agent of the disease was identified as a novel coronavirus on January 8, 2020 [1], which was later named SARS-CoV-2 by the WHO, and its genome was first submitted to GenBank on January 12, 2020 [2], which enabled quick action in global detection and vaccine development. SARS-CoV-2 was identified to be closely related to a virus in bats, RaTG13 [9], suggesting 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 cross species barriers to infect humans. However, the exact intermediate host of SARS-CoV-2 remains unclear, and the

[*] Corresponding author. Key Laboratory of Microbiology of Hainan, Institute of Tropical Bioscience and Biotechnology, Chinese Academy of Tropical Agricultural Sciences, Haikou, Hainan 57110, China.

[**] Corresponding author.

E-mail addresses: 827049686@qq.com (S. Guan), yuyin110110110@163.com (X. Hu), guohuiyi6@hainmc.edu.cn (G. Yi), yaolei2009@gmail.com (L. Yao), zhangjiaming@itbb.org.cn (J. Zhang).

[1] These authors contributed equally.

https://doi.org/10.1016/j.soh.2023.100041
Received 17 July 2023; Accepted 26 September 2023
Available online 28 September 2023

investigation is ongoing.

Most tracing research to date has focused on the analysis of early genomes [5,14,15], in which the viral genomes collected in China played a major role. Pekar et al. [14] explored the evolutionary dynamics of early SARS-CoV-2 genomes. Specifically, they reconstructed the underlying coalescent processes of the index case for the non-introduced SARS-CoV-2 genomes that were sampled during the first wave of the pandemic (from December 2019 to the end of April 2020) in China, using a Bayesian Skyline phylodynamics approach and applying a strict molecular clock [14]. The mean time to most recent common ancestor (tMRCA) for these early strains was revealed to be December 9, 2019 (95% highest posterior distribution [HPD]: November 17 to December 20, 2019). However, 78.7% of the posterior density postdates the earliest published case on December 1, 2019, and failed to capture the index case in Hubei Province [14]. Therefore, the authors concluded that the earliest diverged SARS-CoV-2 lineages must have been extinct in Hubei, China [14]. However, the genomes they used for these analyses are highly redundant, which is known to influence the accuracy of tMRCA estimation and results in a later tMRCA, as indicated by the separate analysis of high-mutation sequences in a recent study [16].

Genome classification has played a critical role in efforts for tracing the origin of SARS-CoV-2. Ruan et al. [15] discovered a set of mutations driving the waves of strain shifts, demonstrating that four highly linked mutations (C241T, C3037T, C14408T, and A23403G, with the latter causing the amino acid mutation D614G in the spike protein) constituted the first wave of strain shift. These mutations were referred to collectively as the “DG group,” and the haplotype containing all four mutant sites was designated DG1111, whereas the wild type was designated DG0000 [15]. The strains driving the early phase of the COVID-19 pandemic were mainly DG0000 in East Asia and DG1111 in Europe, and the split between the Asian and European lineages occurred before September 2019, suggesting a twin-beginning scenario of the pandemic [15]. In another study, Tang et al. [4] classified the early SARS-CoV-2 genomes into two major lineages (designated L and S) that were well-defined by two genetically linked single-nucleotide polymorphisms (T8782C and C28144T, with the latter causing a substitution from serine to leucine at residue 84 in Orf8). The variant sites of the S lineage (8782T and 28,144C) were identical to the orthologous sites in related animal coronaviruses, representing an ancient haplotype [4]. We previously classified the genomes into 16 haplotypes based on Spike-614 and Orf8-84 sites, in which the GL, DL, and DS haplotypes constituted 99.93% of the total genomes [17]. The main evolutionary trajectory of SARS-CoV-2 was proposed to be DS→DL→GL; however, the most recent haplotype GL had the farthest mean tMRCA of May 1, 2019, whereas the most distant haplotype DS had the closest mean tMRCA of October 17, 2019. This suggests that the GL strain had already spread worldwide before the summer of 2019 and triggered the global pandemic, although it went unnoticed until the virus was declared in China [17].

In the present study, we classified the genomes of SARS-CoV-2 based on six closely linked mutation sites (241, 3,037, 8,782, 14,408, 23,403, and 28,144), and analyzed the geographical distribution of the tMRCAs of each haplotype to reveal the evolutionary trajectory of the main haplotypes.

2. Materials and methods

2.1. Genome sequences and processing

FASTA and meta.tsv files of SARS-CoV-2 genomes were downloaded from the GISAID database [18] on May 1, 2022. The genomes were filtered as previously described [17]. In brief, genomes that were collected from hosts other than humans and/or had a length of less than 29,000 nucleotides and more than 0.05% unknown nucleotides were filtered out. More than eight million genomes were retained for further analysis.

2.2. Classification of genomes

The filtered SARS-CoV-2 genomes were aligned to the reference genome wiv04 (EPI_ISL_402124) using ViralMSA [19]. The nucleotides at positions 241, 3,037, 8,782, 14,408, 23,403, and 28,144 in the alignments were retrieved using a personalized script (fetch_nucleotides_from_alignments.pl, https://github.com/XiaowenH/SARS-CoV-2-Classification). The specific haplotype was sub-pooled using seqkit grep [20].

2.3. Phylogenetic analysis

Genome sequences of the main haplotypes and bat-origin viruses RaTG13 (MN996532.2), BANAL-20-52 (MZ937000.1), SL-CoVZC45 (MG772933), SL-CovZXC21 (MG772934), and PrC31 (EPI_ISL_1098866) were aligned using ViralMSA with wiv04 (EPI_ISL_402124) as a reference [19]. The sites with unknown and/or ambiguous nucleotides were removed using a personalized script (rm_site_with_N_from_alignment.pl, https://github.com/XiaowenH/SARS-CoV-2-Classification). A preliminary phylogenetic tree was generated using MEGA 11 [21]. For simplified presentation, six representative sequences were selected randomly from each haplotype, and the evolutionary history was inferred using the maximum-likelihood method and Tamura–Nei model [22]. The tree was rooted with the bat-origin coronaviruses, and the tree with the highest log-likelihood value is shown.

2.4. Phylodynamic analysis

We applied a Bayesian phylodynamics approach to estimate the tMRCA of the early SARS-CoV-2 genomes (from December 2019 to April 2020) [23]. The genomes that were incomplete (<29,000 nucleotides), had more than 0.05% “N”s, and did not have accurate dates were filtered out. The genomes were then aligned using ViralMSA [19]. Redundant genomes were removed using CD-HIT [24] with a threshold to retain a genome number smaller than 2,000. The outputs were transformed to Nexus format using ALTER [25]. The collection dates of samples were transformed into decimal years using a personalized script (date_convertor.pl, https://github.com/XiaowenH/SARS-CoV-2-Classification) and were combined with the aligned sequences in a BEAST xml file using BEAUti v1.10.4 [23]. Bayesian phylodynamic analysis was performed using Bayesian Evolutionary Analysis Sampling Trees (BEAST) version 1.10 [23]. BEAST is a package for evolutionary inference from molecular sequences based on the Markov chain Monte Carlo (MCMC) inference framework [23]. This method was also used by Pekar et al. [14] to estimate the tMRCA of the early SARS-CoV-2 strains isolated in Hubei Province, China. We applied most of the same settings they used for this analysis; the molecular clock was set to strict or relaxed, and the coalescent model was set to Bayesian Skyline [26]. For each analysis, the MCMC chain was run for 400 to 1200 million generations until the explained sum of squares of all parameters was significant (>200). BEAGLE library version 3 [27] was used for accelerated, parallel-likelihood evaluation. The posterior distribution was summarized using TRACER 1.7 [28].

3. Results and discussion

3.1. Influence of genome redundancy on tMRCA estimation

To estimate the timing of the SARS-CoV-2 index case in China, Pekar et al. [14] performed Bayesian phylodynamics analysis using the early genomes (from December 2019 to the end of April 2020) obtained in China. The mean tMRCA was revealed to be December 9, 2019 (95% HPD: November 17 to December 20); however, 78.7% of the posterior density postdates the earliest published case on December 1, and they failed to capture the index case in Hubei Province. Therefore, the authors concluded that the earliest diverged SARS-CoV-2 lineages must have been extinct in Wuhan, China [14].

We repeated their analysis using the same dataset (582 genomes; one genome in their dataset was not found in the GISAID database) and method and obtained a similar result (Fig. 1A). However, the genomes that they used are highly redundant; in particular, 38 genomes were identical (100% identities) in cluster 1 and 66 genomes were identical in cluster 30 (Supplementary Data 1). Based on our experience, redundancy typically influences the accuracy of tMRCA estimation. That is, an older strain that has become less abundant than a newer strain would have a more recent tMRCA. Accordingly, if the redundant genomes are not removed, the estimated tMRCA may only reflect the latest outbreak by covering up the rare ancient sequences. Removal of redundant sequences will reduce the weight of a major outbreak and highlight the weight of rare sequences to obtain an estimation closer to reality, as indicated by the separate analysis of high-mutation sequences demonstrated in a recent study [16].

Therefore, we removed redundant genomes from the early genome set obtained in China using CD-HIT [24] with a threshold of 99.99% and obtained a non-redundant dataset of 212 genomes. As expected, the inferred tMRCA was much earlier than the estimation obtained by Pekar et al. (Fig. 1A, Supplementary Table S1), which was November 26, 2019 (95% HPD: November 4 to December 16), a few days before the earliest reported infection (December 1, 2019). This result suggests that Chinese doctors likely encountered the disease very shortly after its initial occurrence.

One of the major flaws in previous efforts to trace the origin of SARS-CoV-2 is that only the viral strains from China were used for analysis [14, 29,30], omitting the massive global data. We estimated the tMRCA of the viral strains collected worldwide from December 2019 through to the end of April 2020 and compared the tMRCA estimates of redundant and non-redundant genome sets. A total of 80,317 complete genomes were used as the redundant genome set, and 1,338 genomes were retained in the non-redundant genome set according to the threshold of 99.97%. The non-redundant genome set had a mean tMRCA of February 19, 2019 (95% HPD interval: November 5, 2018, to June 4, 2019), whereas the redundant genome set had a mean tMRCA of June 25, 2019 (95% HPD: April 7 to September 6, 2019), representing an approximate 4-month difference between the two estimates (Fig. 1B).

The estimation of a mean tMRCA of February 19, 2019 for the non-redundant genome set is similar to that estimated in two recent studies: one study employed haplotype reconstruction [17], while the other used a separate analysis of high-mutation sequences [16]. Therefore, the proper removal of redundant sequences is necessary to obtain more accurate results. Moreover, the much earlier tMRCA of the global genomes than that based only on genomes isolated from China suggests that China was not the actual place of origin, but rather a secondary center for viral spread. This assumption conforms to the genetic drift theory, according to which the highest genetic diversity is found at the site of origin; when a portion of strains is introduced to a new territory, the diversity in the new territory depends on the abundance of the introduced sample, which is usually lower than that of the original population (Fig. 1C). For example, if the number of introduced genotypes is one, the initial diversity in the new territory will be zero. If the evolutionary rates in the original and secondary sites are similar, the diversity difference will remain at the time of detection, which provides the genetic basis for tMRCA estimation.

Based on the estimated tMRCA of the non-redundant global genomes, SARS-CoV-2 appears to have already been circulating globally for at least 10 months before its first discovery in Wuhan. Therefore, focusing only on the genomes in China would make it impossible to reveal the actual origin of the pandemic.

3.2. Genome classification

In our previous research, the early genomes of SARS-CoV-2 were classified into 16 haplotypes based on two amino acid residues, Spike_614 (S_614) and Orf8_84, wherein GL (S_614G-Orf8_84L), DL (S_614D-Orf8_84L), and DS (S_614D-Orf8_84S) were the three major haplotypes constituting 99.5% of the total genomes [17]. To eliminate the contamination of the haplotypes by recombination and reverse mutation, in the present study, we further characterized the haplotypes according to six highly linked mutations (C241T, C3037T, T8782C, C14408T, A23403G, and C28144T; Fig. 2A, Supplementary Table S2), wherein A23403G corresponds to the amino acid mutation S_D614G and was linked to C241T, C3037T, and C14408T with a linkage rate of 99.6% (Supplementary Table S3), whereas C28144T corresponds to the amino acid mutation Orf8_S84L and was linked to T8782C with a linkage rate of 99.8% (Supplementary Table S4). This suggests a two-step evolutionary process in the evolutionary history, either from DS to GL or from GL to DS.

Four hypotheses for this two-step evolution model are shown in Fig. 2B. Hypotheses 2 and 4 assume the GS haplotype as an intermediate haplotype, which are not supported, as GS strains are not found in the genomes (Supplementary Table S2). The GS strains identified in our previous study were derived from recombination between GL and DS haplotypes [17]. Hypothesis 1 assumes DS to be the ancient haplotype, whereas hypothesis 3 assumes GL to be the ancient haplotype. As DS is the closest to the bat viruses by six nucleotides (Fig. 2A) and DS is located

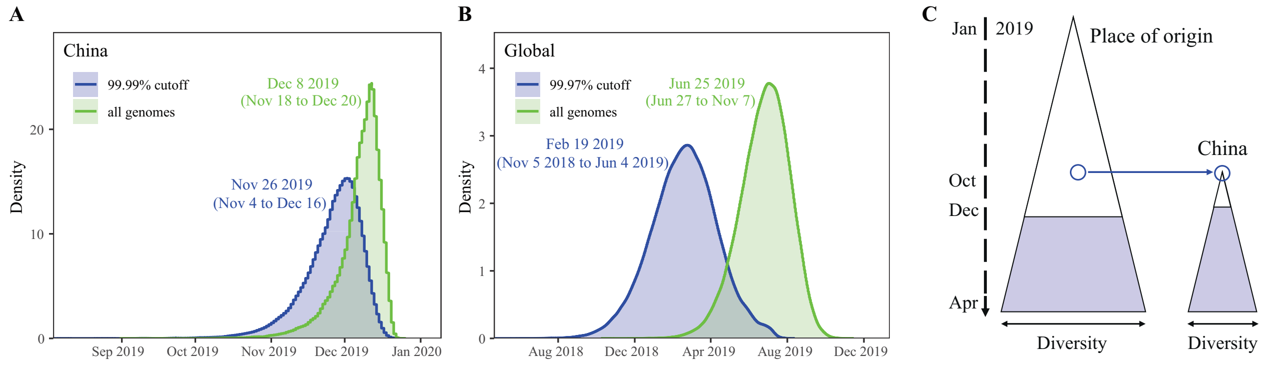

Fig. 1. Influence of genome redundancy on tMRCA estimation. A, Posterior distribution of the tMRCA of the early genomes (all genomes) in the study by Pekar et al. [14] and non-redundant genomes in China, with a redundancy cut-off value of 99.99%. B, tMRCA inferred using global genomes with (all genomes) and without redundant genomes (99.97% cut-off). C, Proposed evolutionary model of SARS-CoV-2 from the place of origin to China; the tips of the triangles represent the tMRCA, the widths from the tip to the bottom indicate expansion of genetic diversity, the white parts of the triangle indicate undetected periods, the shaded areas indicate detected periods, and the circles indicate a portion of strains introduced from the origin to China. Genomes collected in the early phase of the pandemic (from the beginning to the end of April 2019) were used in this analysis.

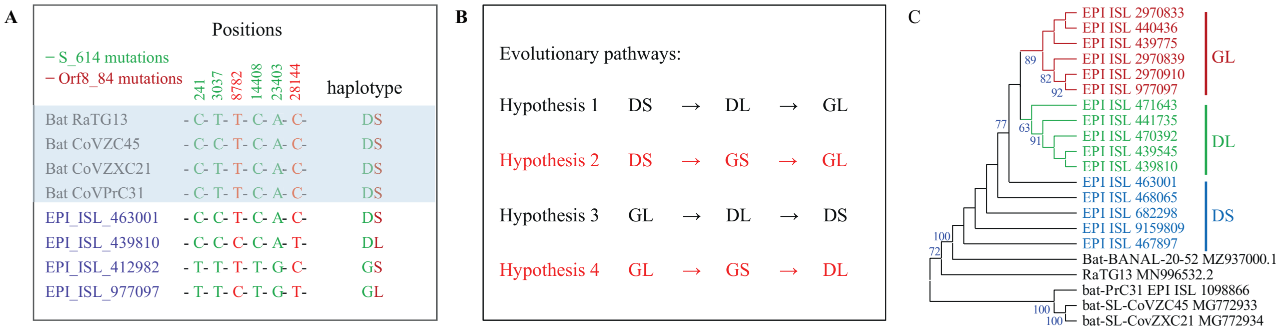
A
Positions
S_614 mutations
Orf8_84 mutations
241 3037 8782 14408 23403 28144
haplotype
Bat RaTG13 - C- T- T- C- A- C- DS
Bat CoVZC45 - C- C- T- C- A- C- DS
Bat CoVZXC21 - C- T- T- C- A- C- DS
Bat CoVPrC31 - C- T- T- C- A- C- DS
EPI_ISL_463001 - C- C- T- C- A- C- DS
EPI_ISL_439810 - C- C- C- C- A- T- DL
EPI_ISL_412982 - T- T- T- T- G- C- GS
EPI_ISL_977097 - T- T- C- T- G- T- GL
B
Evolutionary pathways:
Hypothesis 1 DS → DL → GL
Hypothesis 2 DS → GS → GL
Hypothesis 3 GL → DL → DS
Hypothesis 4 GL → GS → DL
C
EPI ISL 2970833
EPI ISL 440436
EPI ISL 439775
EPI ISL 2970839
EPI ISL 2970910
EPI ISL 977097
GL
EPI ISL 471643
EPI ISL 441735
EPI ISL 470392
EPI ISL 439545
EPI ISL 439810
DL
EPI ISL 463001
EPI ISL 468065
EPI ISL 682298
EPI ISL 9159809
EPI ISL 467897
DS
Bat-BANAL-20-52 MZ937000.1
RaTG13 MN996532.2
bat-PrC31 EPI ISL 1098866
bat-SL-CoVZC45 MG772933
bat-SL-CovZXC21 MG772934
89
82
92
77
63
91
100
72
100
100

Fig. 2. Classification of the early SARS-CoV-2 genomes. A, Main haplotypes classified based on six sites; the positions of the six sites in the reference genome (wiv04, EPI_ISL_402124) are shown in green for the S_614 mutation groups and in red for the Orf8_84 mutation groups. B, Evolutionary hypotheses of the three main haplotypes. C, Maximum-likelihood tree of representative genomes in the three main haplotypes using bat-origin coronaviruses as outgroups.

close to the root of the phylogenetic tree based on bat-origin coronaviruses (Fig. 2C), hypothesis 1 is supported, and the evolutionary trajectory should be DS → DL → GL. This hypothesis is also supported by the fact that the most adapted haplotype that caused the global pandemic is GL, with 77.8% of the genomes sequenced in the early phase of the pandemic belonging to GL (Fig. 3A), and the percentage increased to 96.2% by July 31, 2021 (Supplementary Table S2). However, the haplotypes that had the greatest contribution to the early epidemic in China are DL (58.3%) and DS (28.4%, Fig. 3A). Overall, haplotypes GL, DL, and DS constituted 97.2% of the early genomes globally (Supplementary Table S2).

3.3. Distinct chronological order between tMRCA and evolutionary trajectory

Bayesian phylodynamics analysis of the main haplotypes in the early phase of the pandemic revealed a chronological order that is completely different to the otherwise accepted evolutionary trajectory. The newest haplotype GL had the farthest tMRCA among the three main haplotypes (Fig. 3B), with a mean of May 1, 2019 (95% HPD: February 8 to August 4, 2019). The tMRCAs of DL and DS were in September and October, respectively. Considering the ancient status of DS and DL compared with the status of GL, these results indicate that the three haplotypes were already separated before April 2019 and have evolved in parallel since then. In other words, GL that caused the global pandemic did not originate from the viral strains that caused the epidemic in China. The ancestor strains (DS_0 and DL_0) of DS and DL that gave birth to GL were already extinct before the outbreak of COVID-19; therefore, much more recent tMRCAs have been estimated by using only the current genome sequences.

3.4. Geographical distribution of tMRCAs of the main haplotypes

To explore the geographical distribution of the tMRCA of SARS-CoV-2, Bayesian phylodynamics analysis was performed using the early genomes collected in different regions. Surprisingly, the order of the tMRCAs of the three main haplotypes in China was distinct from the global time order (Fig. 4A) as well as most regional orders (Fig. 4B–I). Within China, the tMRCA of DL came first with a mean of October 25, 2019 (for 95% HPD values for all of these estimates, please refer to Supplementary Table S5); followed by DS, with a mean of October 20, 2019; and GL, with a mean of December 13, 2019. These results suggest that DS and DL may have arrived in China in October 2019. GL did not originate in China; instead, it arrived in China over 6 months after its initial occurrence.

The time order of tMRCAs of the main haplotypes in China was supported by the actual emergence of infections by haplotypes. The first DL (EPI_ISL_402123) and DS (EPI_ISL_529213) samples were collected on December 24, 2019, and December 30, 2019, respectively, in Wuhan, China, while the first GL sample of China (EPI_ISL_6951092) was collected on February 10, 2020, in Nanning, Guangxi Province. Among the patients infected with the GL haplotype, 42% had well-documented records of either traveling overseas or having close contacts with individuals who had traveled internationally, as indicated in the GISAID database (Table 1; for details, please refer to Supplementary Table S6). By contrast, outside of China, none of the 62,453 cases infected by GL had records of travelling in China or other associations with China, whereas 0.11% and 0.16% of cases infected by DS and DL, respectively, were associated with China (Table 1; for details, please refer to Supplementary Table S7).

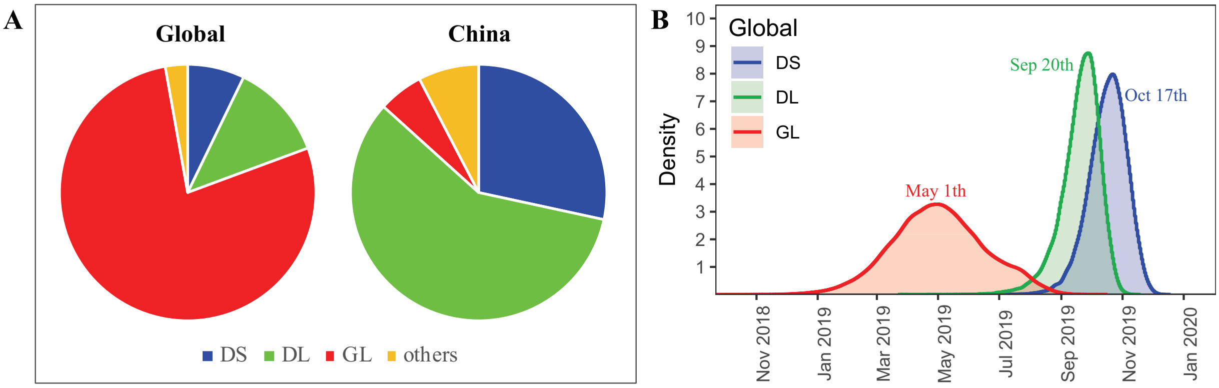

Fig. 3. Composition (A) and posterior distribution (B) of tMRCA of the three main haplotypes in the early phase of the pandemic. Genomes were downloaded from the GISAID database on May 1, 2022, and genomes collected in the early phase of the pandemic (from the beginning to the end of April 2020) were used in this study. For phylodynamic analysis, redundancy was removed using a cut-off of 99.97%.

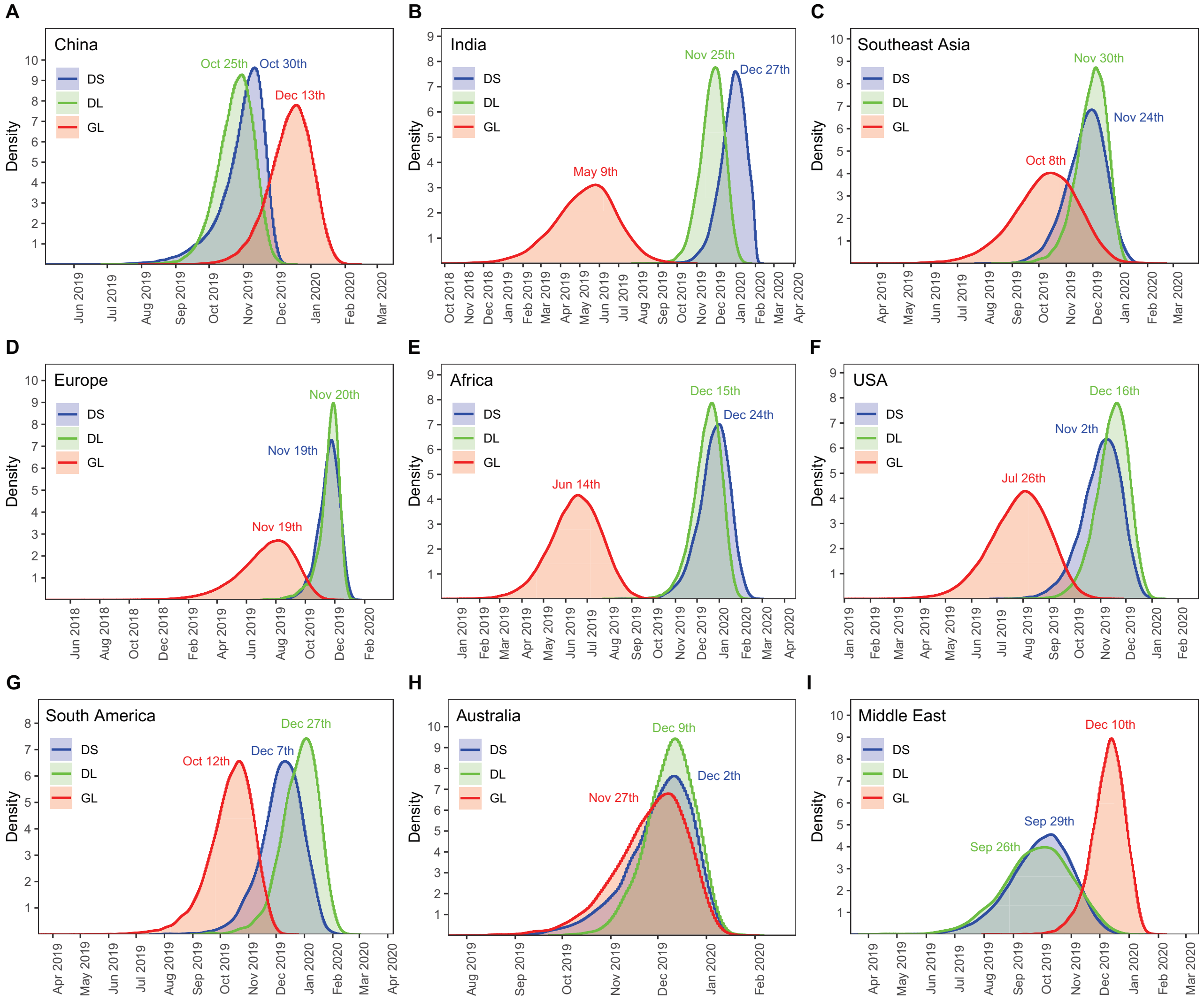

Fig. 4. Posterior distribution of the tMRCA of the main haplotypes in representative regions of the world in the early phase of the pandemic (from the beginning to the end of April 2019). The means of the tMRCA are shown with the same color as each haplotype; for corresponding 95% HPDs, please refer to Supplementary Table S5.

In contrast to the time order in China, the tMRCA of GL was considerably earlier than that of the two older haplotypes in most regions of the world, including India, Southeast Asia, Europe, Africa, the USA, and South America (Fig. 4). The earliest tMRCA of GL was identified in India, which had a mean of May 9, 2019, followed by Africa, Europe, and the USA with tMRCAs in May to July, 2019, suggesting that GL had spread globally at least 6 months before its discovery. The only region that showed some similarity with China was the Middle East, where the tMRCA of GL was the most recent (December 10, 2019; Fig. 4) among the three haplotypes. The Middle East is also the only identified region where DL and DS showed an earlier tMRCA than that in China, which were September 26 and September 29, 2019, respectively, approximately 1 month earlier than the estimated tMRCA in China. The time order in Australia was somewhat unique, showing similar tMRCA estimates for all three haplotypes (Fig. 4); however, the epidemiologic significance of this pattern is unknown.

It is worth discussing why only the tMRCA of the Middle East appears to be similar with that of China, given that India is geographically closer to China. However, this is a difficult question to answer owing to the limited evidence available. India is physically isolated from China by the Himalaya Mountains, and its economic relationship with China is comparatively less extensive than that between the Middle East and China. For example, in 2022, the total trade between China and the

Table 1
Travel records of infected cases by haplotypes in the early phase of the pandemic[a].

Haplotypes	China			Global (outside China)		
	Total	Imported	Proportion (%)	Total	Imported	Proportion (%)
DS	241	0	0	5,300	6	0.11
DL	495	8	1.62	8,785	14	0.16
GL	49	21	42.86	62,453	0	0

[a] Data were retrieved from the metadata in the GISAID database. Only genomes with complete sequences that can be classified correctly were included in this analysis.

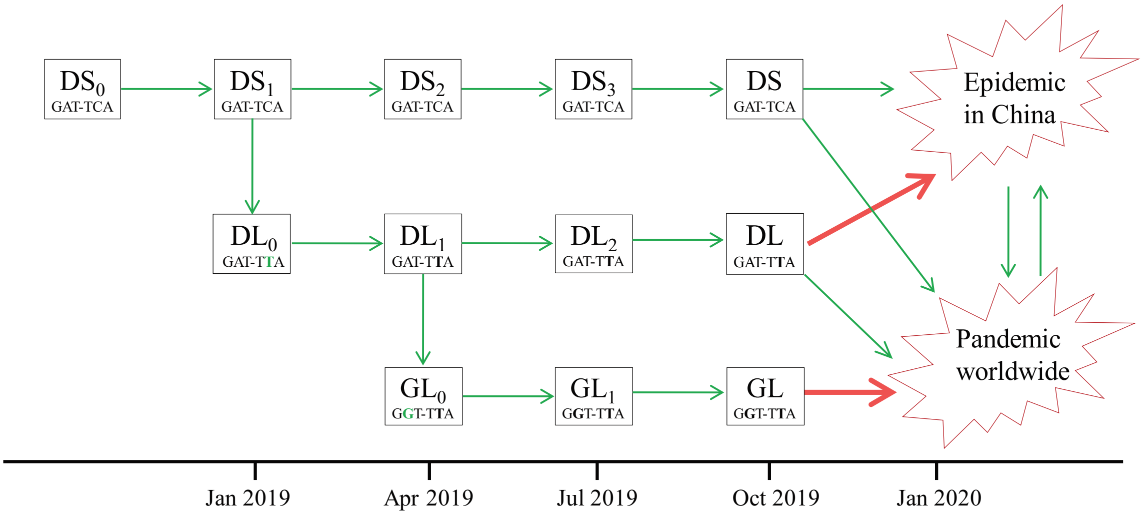

Fig. 5. Evolutionary trajectories of SARS-CoV-2 haplotypes. The codons of S_614 and Orf8_84 are provided below the haplotypes; the nucleotides mutated from previous strains are highlighted in green, the proposed time axis is provided below, and the thick red arrows indicate major events. This diagram is modified from our previous report [17].

Middle East was US $507 billion, while the total trade between China and India was only US $87 billion. However, the distance and relationship between countries may not be a decisive explanation for this pattern. Another possible explanation is that the Middle East experienced the MERS-CoV outbreak in 2012, and China faced the SARS-CoV-1 outbreak in 2003, which may have resulted in certain immunological similarities between these two regions.

3.5. *Proposed evolutionary trajectory*

Taking all of the evidence together, SARS-CoV-2 emerged in February 2019 or earlier and evolved into three main haplotypes (GL, DS, and DL) before May 2019. The main haplotypes continued to evolve in parallel from then on. The dominant haplotype GL had spread worldwide in May to July, 2019, and then evolved into virulent strains in December 2019 that triggered the global pandemic, whereas DL and DS arrived in China in October 2019 and caused the epidemic in December 2019. The proposed evolutionary trajectory is shown in Fig. 5. The GL haplotype that caused the global pandemic neither originated in China nor originated from the viral strains that caused the epidemic in China.

4. Conclusions

SARS-CoV-2 emerged much earlier than reported and had already spread worldwide before its discovery in China. The early genomes of SARS-CoV-2 can be classified into three main haplotypes: GL, DS, and DL. The three haplotypes were separated before May 2019 and then continued to evolve in parallel. The dominant haplotype GL had spread worldwide in May to July 2019, and then evolved into virulent strains in December 2019 to ignite the global pandemic. However, the haplotypes that caused the epidemic in China were DL and DS, which have a much more recent tMRCA than GL, suggesting that GL, which caused the global pandemic, did not originate from the viral strains that caused the epidemic in China, but rather from the ancestral strain DL_0, the origin of which remains unknown. Therefore, restricting the investigation solely to data within China is insufficient to uncover the mysterious origin of the pandemic. Global cooperation is imperative to solve this mystery.

Author contributions

Jiaming Zhang and Lei Yao conceived and designed the experiments. Siqin Guan, Xiaowen Hu, Guohui Yi, and Lei Yao performed the analyses. Xiaowen Hu wrote the custom scripts. Jiaming Zhang performed the formal analysis and wrote the manuscript draft. All authors revised and approved the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We gratefully acknowledge all data contributors, including the authors and their originating laboratories responsible for obtaining the specimens and their submitting laboratories for generating the genetic sequence and metadata and for sharing these data via the GISAID Initiative on which this research is based. We are also grateful to the referees for their comments to improve the manuscript. This research was supported by grants from the National Key R&D Program of China and the Central Public-interest Scientific Institution Basal Research Fund to J.Z. (1630052020022) and by the Project of Science and Technology Department of Sichuan Provincial of China to L.Y. (2019JDJQ0035).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.soh.2023.100041.

References

[1] N. Zhu, D. Zhang, W. Wang, et al., A novel coronavirus from patients with pneumonia in China, 2019, N. Engl. J. Med. 382 (2020) 727–733.

[2] F. Wu, S. Zhao, B. Yu, et al., A new coronavirus associated with human respiratory disease in China, Nature 579 (2020) 265–269.

[3] Y. Tong, W. Liu, P. Liu, et al., The origins of viruses: discovery takes time, international resources, and cooperation, Lancet 398 (2021) 1401–1402.

[4] X. Tang, C. Wu, X. Li, et al., On the origin and continuing evolution of SARS-CoV-2, Natl. Sci. Rev. 7 (2020) 1012–1023.

[5] J. Li, S. Lai, G.F. Gao, et al., The emergence, genomic diversity and global spread of SARS-CoV-2, Nature 600 (2021) 408–418.

[6] Q. Li, X. Guan, P. Wu, et al., Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia, N. Engl. J. Med. 382 (2020) 1199–1207.

[7] WHO, Pneumonia of Unknown Cause in China, 2020. https://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/.

[8] H. Nishiura, N.M. Linton, A.R. Akhmetzhanov, Initial cluster of novel coronavirus (2019-nCoV) infections in Wuhan, China is consistent with substantial human-to-

human transmission, J. Clin. Med. 9 (2020) 488, https://doi.org/10.3390/jcm9020488.

[9] P. Zhou, X.-L. Yang, X.-G. Wang, et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature 579 (2020) 270–273.

[10] T. Zhang, Q. Wu, Z. Zhang, Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak, Curr. Biol. 30 (2020) 1346–1351.

[11] W. Ji, W. Wang, X. Zhao, et al., Cross-species transmission of the newly identified coronavirus 2019-nCoV, J. Med. Virol. 92 (2020) 433–440.

[12] Z. Liu, X. Xiao, X. Wei, et al., Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2, J. Med. Virol. 92 (2020) 595–601.

[13] J. Luan, X. Jin, Y. Lu, et al., SARS-CoV-2 spike protein favors ACE2 from Bovidae and Cricetidae, J. Med. Virol. 92 (2020) 1649–1656.

[14] J. Pekar, M. Worobey, N. Moshiri, et al., Timing the SARS-CoV-2 index case in Hubei province, Science 372 (2021) 412–417.

[15] Y. Ruan, H. Wen, M. Hou, et al., The twin-beginnings of COVID-19 in Asia and Europe – one prevails quickly, Natl. Sci. Rev. 9 (nwab223) (2021) 1–9.

[16] C. Cheng, Z. Zhang, SARS-CoV-2 shows a much earlier divergence in the world than in the Chinese mainland, Sci. China Life Sci. 66 (2023) 1440–1443, https://doi.org/10.1007/s11427-023-2294-5.

[17] X. Hu, Y. Mu, R. Deng, et al., Genome characterization based on the Spike-614 and NS8-84 loci of SARS-CoV-2 reveals two major possible onsets of the COVID-19 pandemic, PLoS One 18 (2023) e0279221.

[18] S. Khare, C. Gurry, L. Freitas, et al., GISAID's role in pandemic response, China CDC Wkly 3 (2021) 1049–1051.

[19] N. Moshiri, ViralMSA: massively scalable reference-guided multiple sequence alignment of viral genomes, Bioinformatics 37 (2021) 714–716.

[20] W. Shen, S. Le, Y. Li, et al., SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation, PLoS One 11 (2016) e0163962.

[21] K. Tamura, G. Stecher, S. Kumar, MEGA11: molecular evolutionary genetics analysis version 11, Mol. Biol. Evol. 38 (2021) 3022–3027.

[22] K. Tamura, M. Nei, Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees, Mol. Biol. Evol. 10 (1993) 512–526.

[23] M.A. Suchard, P. Lemey, G. Baele, et al., Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10, Virus Evol 4 (2018) vey016.

[24] L. Fu, B. Niu, Z. Zhu, et al., CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (2012) 3150–3152.

[25] D. Glez-Peña, D. Gómez-Blanco, M. Reboiro-Jato, et al., ALTER: program-oriented conversion of DNA and protein alignments, Nucleic Acids Res. 38 (2010) W14–W18.

[26] A.J. Drummond, A. Rambaut, B. Shapiro, et al., Bayesian coalescent inference of past population dynamics from molecular sequences, Mol. Biol. Evol. 22 (2005) 1185–1192.

[27] D.L. Ayres, M.P. Cummings, G. Baele, et al., BEAGLE 3: improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics, Syst. Biol. 68 (2019) 1052–1061.

[28] A. Rambaut, A.J. Drummond, D. Xie, et al., Posterior summarization in Bayesian phylogenetics using Tracer 1.7, Syst. Biol. 67 (2018) 901–904.

[29] J.E. Pekar, A. Magee, E. Parker, et al., The molecular epidemiology of multiple zoonotic origins of SARS-CoV-2, Science 377 (2022) 960–966.

[30] M. Worobey, J.I. Levy, L. Malpica Serrano, et al., The huanan Seafood wholesale Market in wuhan was the early epicenter of the COVID-19 pandemic, Science 377 (2022) 951–959.