The twin-beginnings of COVID-19 in Asia and Europe – One prevails quickly

Chung-I Wu (ciwu@uchicago.edu)
Sun Yat-Sen University

Research Article

Keywords: COVID-19 evolution

Posted Date: October 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-955853/v1

The twin-beginnings of COVID-19 in Asia and Europe – One prevails quickly

Yongsen Ruan[1], Haijun Wen[1], Mei Hou[1], Ziwen He[1], Xuemei Lu[2], Yongbiao Xue[3], Xionglei He[1], Ya-Ping Zhang[2*], Chung-I Wu[1, 3, 4*]

Affiliations:

[1]State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China.

[2]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Science, Kunming 650223, China.

[3]Beijing Institute of Genomics, Chinese Academy of Sciences, and China National Centre for Bioinformation, Beijing 100101, China.

[4]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA.

*Corresponding author. Email: wzhongyi@mail.sysu.edu.cn, ciwu@uchicago.edu (Chung-I Wu); zhangyp@mail.kiz.ac.cn (Ya-Ping Zhang).

Abstract: In the spread of SARS-CoV-2, there have been multiple waves of replacement between strains, each of which having a distinct set of mutations. The first wave is a group of 4 mutations that includes D614G. This DG (D614G) group, fixed at the start of the pandemic, is the foundation of all subsequent waves of strains. Curiously, the DG group is absent in early Asian samples but present (and likely common) in Europe from the beginning. European data show that the high fitness of DG1111 requires the synergistic effect of all four mutations. However, the European strains would have no time to evolve the 4 DG mutations (0 to 1), had they come directly from the early Asian DG0000 strain. Very likely, the European DG1111 strain had acquired the highly adaptive DG mutations in the pre-pandemic Europe and had been spreading in parallel with the Asian strains. Two recent reports further support this twin-beginning interpretation. There was a period of two-way spread between Asia and Europe but, by May of 2020, the European strains had supplanted the Asian strains globally. This large-scale replacement of one set of mutations for another has since been replayed many times as COVID-19 progresses.

Main Text:

The study of molecular evolution is constrained by the data on extant organisms only. In contrast, the large number of genome sequences of SARS-CoV-2, collected throughout the entire period of the epidemics, has provided an unprecedented opportunity to observe evolution in action [1-6]. In a companion study (Ruan et al. 2021), we leverage this large data set to track the evolution of SARS-CoV-2, focusing on the Delta strain in mid-2021. In this study, we analyze the beginning of the epidemics in early 2020. The methodologies used are developed between the two studies (see the Supplement). Since the results are roughly comparable across regions, we choose the extensive UK data to represent Europe in a comparison between Asia and Europe.

The succession of waves of variants in COVID-19

The evolutionary dynamics of SARS-CoV-2 in UK is given in Fig. 1 where the changes in the variant frequency at each site (e.g., C→T) from March of 2020 to July of 2021 is presented. Briefly, by lining up a large number of viral sequences, we examine each site across sequences using the infinite site model of population genetics [7]. In contrast, virological studies usually examine each sequence across sites [2,4,6,8], akin to using the infinite-allele model of population genetics. Variants are compared with the ancestral state in the outgroup sequences of bats to determine the mutant status. A haplotype is defined as 2 or more variants of the same viral genome. We filtered out sites where the mutant frequency never reached a cutoff value of 0.1, 0.3 or 0.5. Variants that have never been sufficiently common in the population are of lower interest as they have limited impact on the progression of epidemics. Here, we use 0.3 as the cutoff although the conclusion is the same with the other two cutoffs.

In Fig. 1A, one could observe 5 waves of variants (labeled W0 to W4) that rise and fall together in the same wave (data shown in Table S1). Each wave is a composite of multiple overlapping curves with each curve representing a variant at a particular site. The overlap therefore portrays a haplotype that bears multiple variants of the same evolutionary dynamics. Variants of each wave do differ slightly in the low frequency range and, in W2 and W4, the differences can be seen even in higher frequencies. These variants do follow the same trend in the rise and fall. The difference happens when a new variant emerges in the same haplotype, but after others have reached a modest frequency of, say, 10%. The formation of the waves will be discussed in detail below (see also Ruan et al. 2021).

Fig. 1 here

An unexpected and most interesting observation in Fig. 1A is that a new wave usually rises at the expense of the previous wave. This is surprising in the analysis of the site-by-site evolution whereby one might have expected cumulative evolution as shown in Fig. 1B [9]. In this mode of evolution, new mutations are piled on the earlier successful haplotypes, resulting in a series of ascending curves depicted in Fig. 1B. Thus, the rises in Fig. 1A are expected but the falls are not. The overall patterns suggest strong competition and mutual exclusion between different sets of mutations. The competition will later help to explain the evolution of SARS-CoV-2 in the early days.

The only wave that rises and stays at the top is W0. It has four variants (the noncoding site C241T, synonymous site C3037T, two nonsynonymous sites C14408T and A23403G, the latter being D614G site) which will be referred to collectively as the D614G (or DG) group [10-16]. Since the evolution proceeds from CCCA to TTTG, we designate the ancestral haplotype CCCA

as DG0000 and the globally successful haplotype TTTG as DG1111. Depending on the bat species used as the outgroup, the ancestral sequence could be DG0000 or DG0100 (see Supplement). Because the analyses and conclusion are not affected by this synonymous site, we use DG0000 as the ancestor. Usually, all four DG mutations are either all present or absent. This tight association, however, may not be true when the mutations are still in low frequency and partial haplotypes such as DG1001 can occasionally be seen.

The unique strength of the DG group is evident in their becoming fixed very quickly (Fig. 1A). Other subsequent waves, W1 – W3, all went up and down while the latest W4 (the Delta wave) is too recent to judge [17,18].

The anatomy of Wave 0

We shall focus on the first and the only fixed wave, W0, across geographical regions. The numbers of counts of the W0 variants in China and Europe are given in Tables 1 and 2. Fig. 2A tracks the frequency change of the D614G mutation in China, Asia, Europe and North America (data shown in Table S2). In the entire period, the D614G mutation frequencies are higher in Europe than in Asia. After Jan. 31, two forces influence the frequencies of the DG mutations. One is the inflow of the ancestral DG0000 strain from Asia into Europe, thus driving down the frequency of the DG1111 haplotype there. The other one being the fitness advantage that drives up the DG mutation frequency in all regions a few weeks later. Hence, the dynamics in the European and North American samples show a dip due to the import of the Asian strains before gaining in frequency. The Asian trend, in contrast, rose steadily after mid-March with a time-lag behind the European trend. The difference between continents will be informative about the evolution before the epidemics.

Fig. 2 and Tables 1-2 here

While the trend in Fig. 2A tracks the last one of the four sites (i.e., DG***0 vs DG***1 where * indicates any nucleotide), the evolution of DG1111 involves all four mutations. The fitness of DG1111 has been suggested to be a function of the entire DG group mutations [19]. In this section, we examine the fitness evolution from DG0000 to DG1111. Table 1 shows the haplotype distribution in China which is uncomplicated. First, the full DG1111 has not evolved in China and its presence from March on is, on the record, due to the inflow from abroad. Second, in the absence of DG1111, one may look for the partial haplotypes between DG0000 and DG1111. Among the 3-mutant configurations, 6 DG1101, one DG-111 and one DG1110 are found among 783 sequences before the presence of DG1111. Third, there are a few 1- or 2-mutant haplotypes scattered in the background (Table S3).

In contrast, the haplotypes in Europe show a broader distribution. Table 2 shows the number of each haplotype at each given time. We now focus on DG0000, DG1111 and the 3-mutant haplotypes DG 0111, 1011, 1101 and 1110. The occurrences of DG1111 as well as each partial haplotype is compared with that of DG0000. The changes in the relative abundance manifest the fitness differences among haplotypes. Most noteworthy is the abundance of DG1111 vs. DG0000 that rises from a ratio of 0.21 on Feb. 10, 2020, to 1.0, 1.47. 2.2., 2.36 and then arrives at 3.12 on March 31. We assume that each strain grows in number by $N_t = N_0 e^{R \times t}$ where N_t is the number at time t. The fitness advantage, D_w, of strain X over strain Y is then

represented by $R_X - R_Y$. Here, strain *Y* is always DG0000 and strain *X* is DG1111 or a partial haplotype. The calculation of D_w is based the sample of each time point against the latest Mar-31 sample in Table 2 and is measured over a 10-day interval.

It is most interesting that the only haplotype with a higher fitness than DG0000 is DG1111 which shows $D_w \sim 0.30$ for the average. For all 3-mutant configurations, D_w is negative meaning lower proliferation than the wildtype DG0000. In particular, the most common partial haplotype, DG1101, has a $D_w \sim -0.82$ on average. Given the estimated D_w, the frequency of DG1111 relative to that of DG0000 over a 3-month time would be $e^{0.30\times 9} = 14.9$ whereas DG1101 would be reduced to 0.0006 of the DG0000. The higher fitness of DG1111 over all partial haplotypes suggests some fitness effect for all 4 mutations. The first of the four mutations is the noncoding site C241T which is surprisingly important as D_w would change from 0.30 to -0.42 between DG1111 and DG01111. The second mutation is a synonymous site C3037T, which should be the least important one and indeed DG1011 appears to be most fit among the partial haplotypes.

The formation of a multi-variant haplotype, DG1111

The analysis above provides us the means to roughly estimate the time it may take for DG0000 to evolve to DG1111 by the routes of Fig. 3A, each route passing through 3 interior nodes. The rate of evolution between two nodes of Fig. 3A, say D0000 (*A*) and D1000 (*B*), is given by

$$R = N_A\, u\, f_B \qquad \text{(Eq. 1)}$$

where N_A is the number of individuals in node *A*, *u* is the mutation rate and $f_B = f(N_A, s) = \frac{1-e^{-2s}}{1-e^{-N_A s}}$ is the fixation probability of haplotype *B* which has a selective advantage of *s* over haplotype *A*.

If the variant is neutral, $f_B = 1/N_A$ and $R = u$. Hence, the "waiting time" between two successive mutations (e.g., from 1000 to 1100) would be $1/u$. Given the per base mutation rate of $< 10^{-5}$/bp/day [20,21], each step would take on average $> 10^5$ days. (Of course, if the mutation can fall on any site of the ~30 kb genome, the waiting time would only be a few days.) With the four sites, the evolution from DG0000 to DG1111 would take much longer than one year.

Fig. 3 here

With the slow rate of neutral evolution, speedier evolution at specific sites must be driven by natural selection. If both *N* and *s* are large, *R* can even be higher than 10^{-2} and it may take only 2-3 months to evolve from DG0000 to DG1111. However, the measurements of *s* (expressed as D_w in Table 2) for the four partial haplotypes (DG0111, 1011, 1101 and 1110) are nowhere near the level that can reduce the evolutionary time to within a year. In fact, since these partial haplotypes all appear even less fit than the wildtype DG0000 (Table 2), the evolution of DG1111 from DG0000 would be blocked. This may imply that the evolution of DG0000 to DG1111 most likely happened in a genetic background different from that of the current DG0000; in other words, the evolution may have happened in an unknown haplotype

background of a more distant past than shown in Fig. 2. In this background, at least one of the routes in Fig. 3A is not blocked.

Testing the conventional one-beginning view - All strains spread initially from one place

The main message of Tables 1 and 2 is about the emergence of DG1111 in Europe as well as the non-emergence in China. The conventional view is that a (pre-DG1111) strain spread from China to Europe where it evolved into DG1111 (Fig. 2B). This view raises two related questions: 1) Which one is this pre-DG1111? 2) How much time is available for this pre-DG1111 strain to evolve into DG1111 in Europe? (Here, we may also ask why DG1111 did not emerge in China and the simplest answer is the stochastic nature of evolution – not everything that could happen had happened.)

In answering the questions, we divide the evolution of the DG haplotypes in China into 3 stages in Fig. 2B (see also Table 1). In the Early stage before Jan. 21, 2020, the haplotypes are entirely DG0000 (91/91). In the Middle stage between 1/21/20 and 3/1/20, there is no DG1111 haplotype yet but partial haplotypes (such as DG0001, DG1001 and DG1101) began to emerge albeit collectively accounting for only 12 out of 692 samples (1.7%). In the Late stage after 3/1/20, DG1111 appeared and increased to 50% by the end of March. The increase happened with the inflow from abroad. In contrast, the DG1111 haplotype is present in *all* European and North American samples including the earliest collection dated Jan. 21 of 2020 (or even Jan. 1 in an unusual Canadian sample, see Table S4 and Table S5).

To answer the first question, we can conclude that the pre-DG1111 haplotype from Asia is not DG1111 itself, which was absent in China. The most common partial haplotype in China and Europe in the Middle stage is DG1101. However, as DG1101 appears at the same time in both continents and is far more common in Europe (~ 20%) than in China (~ 1%), DG1101 is unlikely to be the putative pre-DG1111 from China either. The other two partial haplotypes, DG-111 and DG1110, are both < 0.2% in China in early February and have not been seen later (Table 1). They are even less likely to be the direct ancestor of DG1111 in Europe. (DG-111, in this interpretation, is most likely DG0111, as DG1111 would not have died out almost immediately.) In short, the import from China should be of only one possible haplotype – that of DG0000.

For a visual impression, genomic sequences of the 3 major haplotypes collected before 1/30/2020 are presented in Fig. 3B. The common DG0000, DG1111 and the partial haplotype DG1101 are shown. While the age should be in the order of DG0000, DG1101 and DG1111, the within-haplotype diversity appears the highest in the presumably youngest DG1111. The comparison implies that DG1111, when first detected, is already diverse, hinting a period of mutation accumulation prior to the onset of the epidemics. This issue is pursued below.

The second question is about how much time is available for the evolution of DG1111 in Europe. The time span should be between the presumed arrival of DG0000 from Asia and the first appearance of DG1111 in Europe. We shall allow the earliest possible time for the arrival from Asia which is Dec. 15, 2019. The earliest possible time of the appearance of DG1111 is uncertain, but definitely no later than Feb. 10, 2020. By that date, 7 of the 32 European samples are DG1111 and, 10 days later (2/20/2020), it is 18 of the 37 samples (Table 2). Nevertheless, the most interesting samples are those collected in January 2020, marked by the blue circled

number in Fig. 2B (also see Table S6). As annotated in the legends, the sequence deposition in this period is unusually prone to revisions and retractions. Still, unless all these reports are false, the full DG1111 haplotype has been formed, likely in Europe, by the beginning of January 2020. Given the presence of DG1111 in small samples, often one out of one, the frequency may be quite high. Two recent reports on even earlier appearances of SARS-CoV-2 in Italy will be discussed in the next section.

The distribution of DG1111 across continents in the beginning of the epidemics rejects the single-beginning scenario depicted in Fig. 2B whereby DG1111 has to evolve from DG0000 to DG1111 in an impossibly small timespan. Furthermore, the earliest appearance of DG1111 is associated with very high nucleotide diversity as shown in Fig. 3B, indicating the real age of DG1111 being much older than its first detection in humans. To conclude, if the spread from Asia to Europe has to be months before there was any sign of an impending epidemic, we would effectively be considering multiple beginnings of the epidemics.

The multiple-beginning scenario – Evidence from Europe

In discussing the beginning of an epidemic, we should note the distinction between the beginning and the origin. The place where SARS-CoV-2 originated is referred to PL0. It has been suggested that PL0 is not the same as PL1, the first place that reports the impending epidemic [22,23]. To allow the multi-step adaptive shift from animal to human hosts, PL0 needs to possess several characteristics that are distinct from those conducive for the first epidemic [22]. The *beginning* of the epidemics is hence at PL1, which receives the infections directly from PL0 and spread the virus to other places, collectively referred to as PL2's. It is often assumed that there is only one PL1 and it is in China, even though the theory supports multiple PL1's. This assumption will be tested here.

A parsimonious explanation for the emergence of DG1111 may be the twin-beginning scenario shown in Fig. 4A. In this scenario, the virus spread to both continents quite early, presumably from the yet unidentified PL0. With the independent beginning in Europe, the virus would have sufficient time to evolve the 4 DG mutations.

There have been many reports that SARS-CoV-2 was circulating in Europe (and perhaps in the US) in late 2019 [24-31]. Among the most convincing studies is that of Amendola et al. [27]. In this report, a 4-year-old child has been shown, by PCR and DNA sequencing, to be infected with SARS-CoV-2 in November of 2019 in northern Italy [27]. More recently, Amendola et al. (2021b) further show that SARS-CoV-2 was indeed circulating in the Lombardy region in September to December 2019 with 11 of the 44 suspected patients yielding SARS-CoV-2 sequences [31]. They further demonstrate that 6 out of the 11 patients have sequence reads covering at least one DG site. All 6 patients thus appear to have the DG group mutations when data are available. By early December of 2019, the DG1111 haplotype had already been assembled in Europe but remained absent in Asia. If we consider the strong association among the DG group mutations, the first appearance of DG1111 in Europe could be as early as September of 2019, a timeline implied in Amendola et al. (2021b) and adopted in Fig. 4B.

Fig. 4 Here

Fig. 4B summarizes the evolution of the Asian and European lines of SARS-CoV-2 proposed in Fig. 2B. The two lineages compete in the spread much like the competition among sets of mutations depicted in Fig. 1. Although the Asian lineage has begun the spread slightly earlier, the main trunk of SARS-CoV-2 evolution is dominated by the European lineage (see the W0 wave of Fig. 1). Since April 2020, the Asian lineage has nearly disappeared from the epidemics. The period between late January and April of 2020 is when the competition between the two lineages could be observed as shown by the pie charts of Fig. 4B. These pie charts illustrate the haplotype compositions in, as well as the viral exchanges between, Asia and Europe.

Discussion

As emphasized above, the *beginning* of epidemics is at PL1. In theory, once the virus evolved to its epidemic form in PL0, it can spread simultaneously to multiple PL1's [22]. This study shows that Asia and Europe could both have PL1s' within its boundary (designated with the red and blue colors in Fig. 4). However, only the PL1 that first reports the impending epidemic is identified as such and all other PL1's would be recognized as PL2. This is the caveat against the single-beginning view as the first one is perceived as the only one.

For COVID-19, the single-beginning view has been widely accepted but never been tested. In this study, we provide the evidence from sequence evolution to reject this view. In brief, the full DG1111 haplotype emerged in Europe at about the same time as (or even earlier than) the arrival of Asian strains in Europe. Crucially, the Asian strains arriving in Europe must be of the DG0000 type because even partial haplotypes (such as DG1011) were not seen in China in the early samples.

The necessary condition for the W0 wave to rise and usher in the global pandemic is the assembly of the full DG1111 haplotype. Table 2 shows that DG1111 has a much higher fitness than the wildtype DG0000 whereas all other partial haplotypes are no better, and often worse, than DG0000. Our analysis shows that the Chinese haplotypes are far from the successful assembly of DG1111. Since the strains in China fail to evolve to DG1111 in situ, it is unlikely that a small cohort would succeed in this evolution immediately upon the arrival in Europe.

It would thus appear that SARS-CoV-2 had been in Europe long enough to evolve DG1111 in separation from the viral evolution in Asia. Indeed, Amendola et al. (2021a, b) provide the data supporting this conjecture. With the twin-beginnings, the place in China that takes the credit of being PL1 (or even PL0) is not even the one that dominates in the subsequent global spread; it is merely the first one to be out of the gate. The crucial timeframe to see the competition is the few months before and after the onset of the epidemics. In this period, two concurrent lineages compete and spread in waves. The earlier wave is weaker, yielding to the complete replacement of the Asian variants by the European ones. Epidemics in waves are not uncommon. The 1918 flu is one earlier example [32,33] and COVID-19 has revealed this pattern with clarity.

References

1 Martin, M. A., VanInsberghe, D. & Koelle, K. Insights from SARS-CoV-2 sequences. *Science* **371**, 466-467, doi:10.1126/science.abf3995 (2021).

2 Rochman, N. D. *et al.* Ongoing global and regional adaptive evolution of SARS-CoV-2. *Proc Natl Acad Sci U S A* **118**, e2104241118, doi:10.1073/pnas.2104241118 (2021).

3 Ruan, Y. *et al.* On the founder effect in COVID-19 outbreaks: how many infected travelers may have started them all? *Natl. Sci. Rev.* **8**, nwaa246, doi:10.1093/nsr/nwaa246 (2021).

4 Forster, P., Forster, L., Renfrew, C. & Forster, M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci U S A* **117**, 9241-9243, doi:10.1073/pnas.2004999117 (2020).

5 Kumar, S. *et al.* An Evolutionary Portrait of the Progenitor SARS-CoV-2 and Its Dominant Offshoots in COVID-19 Pandemic. *Mol. Biol. Evol.* **38**, 3046-3059, doi:10.1093/molbev/msab118 (2021).

6 Tang, X. *et al.* On the origin and continuing evolution of SARS-CoV-2. *Natl. Sci. Rev.* **7**, 1012-1023, doi:10.1093/nsr/nwaa036 (2020).

7 Hartl, D. L. & Clark, A. G. *Principles of Population Genetics*. (Sinauer Associates, 1997).

8 Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* **5**, 1403-1407, doi:10.1038/s41564-020-0770-5 (2020).

9 Kimura, M. *The neutral theory of molecular evolution*. (Cambridge University Press, 1983).

10 Volz, E. *et al.* Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell* **184**, 64-75 e11, doi:10.1016/j.cell.2020.11.020 (2021).

11 Plante, J. A. *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* **592**, 116-121 (2021).

12 Zhou, B. *et al.* SARS-CoV-2 spike D614G change enhances replication and transmission. *Nature* **592**, 122-127, doi:10.1038/s41586-021-03361-1 (2021).

13 Zhang, J. *et al.* Structural impact on SARS-CoV-2 spike protein by D614G substitution. *Science* **372**, 525-530, doi:10.1126/science.abf2303 (2021).

14 Yurkovetskiy, L. *et al.* Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. *Cell* **183**, 739-751. e738 (2020).

15 Korber, B. *et al.* Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182**, 812-827. e819 (2020).

16 Hou, Y. J. *et al.* SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* **370**, 1464-1468, doi:10.1126/science.abe8499 (2020).

17 Cherian, S. *et al.* Convergent evolution of SARS-CoV-2 spike mutations, L452R, E484Q and P681R, in the second wave of COVID-19 in Maharashtra, India. *bioRxiv*, 2021.2004.2022.440932, doi:10.1101/2021.04.22.440932 (2021).

18 Planas, D. *et al.* Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature* **596**, 276-280, doi:10.1038/s41586-021-03777-9 (2021).

19 Yang, H. C. *et al.* Analysis of genomic distributions of SARS-CoV-2 reveals a dominant strain type with strong allelic associations. *Proc Natl Acad Sci U S A* **117**, 30679-30686, doi:10.1073/pnas.2007840117 (2020).

20 Bar-On, Y. M., Flamholz, A., Phillips, R. & Milo, R. SARS-CoV-2 (COVID-19) by the numbers. *Elife* **9**, doi:10.7554/eLife.57309 (2020).
21 van Dorp, L. *et al.* No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nat. Commun.* **11**, 5986, doi:10.1038/s41467-020-19818-2 (2020).
22 Ruan, Y., Wen, H., He, X. & Wu, C. I. A theoretical exploration of the origin and early evolution of a pandemic. *Sci Bull (Beijing)* **66**, 1022-1029, doi:10.1016/j.scib.2020.12.020 (2021).
23 Wu, C. I. *et al.* On the origin of SARS-CoV-2-The blind watchmaker argument. *Sci China Life Sci* **64**, 1560-1563, doi:10.1007/s11427-021-1972-1 (2021).
24 La Rosa, G. *et al.* SARS-CoV-2 has been circulating in northern Italy since December 2019: Evidence from environmental monitoring. *Sci. Total Environ.* **750**, 141711 (2021).
25 Randazzo, W. *et al.* SARS-CoV-2 RNA in wastewater anticipated COVID-19 occurrence in a low prevalence area. *Water Res.* **181**, 115942, doi:10.1016/j.watres.2020.115942 (2020).
26 Althoff, K. N. *et al.* Antibodies to SARS-CoV-2 in All of Us Research Program Participants, January 2-March 18, 2020. *Clin. Infect. Dis.* (2021).
27 Amendola, A. *et al.* Evidence of SARS-CoV-2 RNA in an Oropharyngeal Swab Specimen, Milan, Italy, Early December 2019. *Emerg Infect Dis* **27**, 648-650, doi:10.3201/eid2702.204632 (2021).
28 Basavaraju, S. V. *et al.* Serologic Testing of US Blood Donations to Identify Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2)-Reactive Antibodies: December 2019-January 2020. *Clin. Infect. Dis.* **72**, e1004-e1009, doi:10.1093/cid/ciaa1785 (2021).
29 Deslandes, A. *et al.* SARS-CoV-2 was already spreading in France in late December 2019. *Int. J. Antimicrob. Agents* **55**, 106006, doi:10.1016/j.ijantimicag.2020.106006 (2020).
30 Apolone, G. *et al.* Unexpected detection of SARS-CoV-2 antibodies in the prepandemic period in Italy. *Tumori* **0**, 300891620974755, doi:10.1177/0300891620974755 (2020).
31 Amendola, A. *et al.* Molecular Evidence for SARS-CoV-2 in Samples Collected From Patients With Morbilliform Eruptions Since Late Summer 2019 in Lombardy, Northern Italy. *Preprint* (2021).
32 Humphreys, M. The influenza of 1918: Evolutionary perspectives in a historical context. *Evol Med Public Health* **2018**, 219-229, doi:10.1093/emph/eoy024 (2018).
33 Crosby, A. W. *America's forgotten pandemic: the influenza of 1918*. (Cambridge University Press, 2003).
34 Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* **1**, 33-46, doi:10.1002/gch2.1018 (2017).

Acknowledgments: We thank all those who have contributed sequences to the GISAID (Global Initiative on Sharing All Influenza Data) database (https://www.gisaid.org/). We thank Drs. Xuhua Xia and Jian Lu for comments on and suggestions for this manuscript.

Funding:

National Natural Science Foundation of China

Author contributions:

Conceptualization: C.I.W, Y.P.Z, Y.R.

Methodology: C.I.W, Y.R.

Investigation: C.I.W, Y.P.Z, Y.R.

Visualization: C.I.W, Y.R.

Funding acquisition: C.I.W.

Project administration: Y.R, H.W, M.H., Z.H., X.L., Y.X., X.H., Y.P.Z., C.I.W.

Supervision: C.I.W, Y.P.Z.

Writing – original draft: C.I.W, Y.R.

Writing – review & editing: C.I.W, Y.R, Y.P.Z, H.W, X.H., X.L.

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: All data were downloaded from the GISAID (Global Initiative on Sharing All Influenza Data) database (https://www.gisaid.org/). And other data are available in the main text or the supplementary materials.

Supplementary Materials

Materials and Methods

Figs. S1 to S2

Tables S1 to S7

Table 1. Haplotype frequencies of the D614G group mutations in early samples from China.

Haplotype	Jan-01 2020 (27)	Jan-11 2020 (11)	Jan-21 2020 (53)	Jan-31 2020 (312)	Feb-10 2020 (235)	Feb-20 2020 (80)	Mar-01 2020 (65)	Mar-11 2020 (62)	Mar-21 2020 (83)	Mar-31 2020 (23)
CCCA (0000)	27	11	53	306	228	80	65	53	53	8
TTTG (1111)	0	0	0	0	0	0	0	7	23	11
TTCG (1101)	0	0	0	3	3	0	0	0	4	3
-TTG (-111)	0	0	0	1	0	0	0	0	0	0
TTTA (1110)	0	0	0	0	1	0	0	0	0	0
Others	0	0	0	2	3	0	0	2	3	1

The haplotype frequencies are showed by a 10-day interval. The first period includes all sequences up to Jan-1,2020. The total number of sequences are show in parenthesis in the first row.

Table 2. Haplotype frequencies of the D614G group mutations in early samples from Europe

Haplotype	Jan-01 2020 (0)	Jan-11 2020 (0)	Jan-21 2020 (1)	Jan-31 2020 (31)	Feb-10 2020 (32)	Feb-20 2020 (37)	Mar-01 2020 (592)	Mar-11 2020 (4049)	Mar-21 2020 (8323)	Mar-31 2020 (13677)
CCCA (0000)	0	0	0	23	19	18	234	1236	2432	3259
TTTG (1111)	0	0	1	1	7 (9)	18	344	2723	5741	10184
vs. CCCA					0.21	1.00	1.47	2.20	2.36	3.12
D_w					0.54	0.28	0.25	0.17	0.28	AVG ~ +0.30
TTCG (1101)	0	0	0	7	6	1	5 (6)	7	4	7 (11)
vs. CCCA				0.30	0.32		0.024	0.0057		0.0019
D_w				-0.84	-1.03		-0.85	-0.55		AVG ~ -0.82
TCTG (1011)	0	0	0	0	0	0	2	4 (6)	9	11
vs. CCCA								0.0041	0.0037	0.0034
D_w								-0.094	-0.085	AVG ~ -0.090
CTTG (0111)	0	0	0	0	0	0	0	5	11	8
vs. CCCA								0.0040	0.0045	0.0025
D_w								-0.24	-0.59	AVG ~ -0.42

Table 2 is the counterpart of Table 1 in Europe with the additional information on the relative fitness of different haplotypes. D_w is the fitness advantage of the focus haplotype over DG0000 with the fitness defined by R in $N_t = N_0 e^{R \times t}$. $D_w = R_X - R_Y$ where X is for any haplotype X and R_Y is for DG0000. AVG is D_w averaged across time periods.

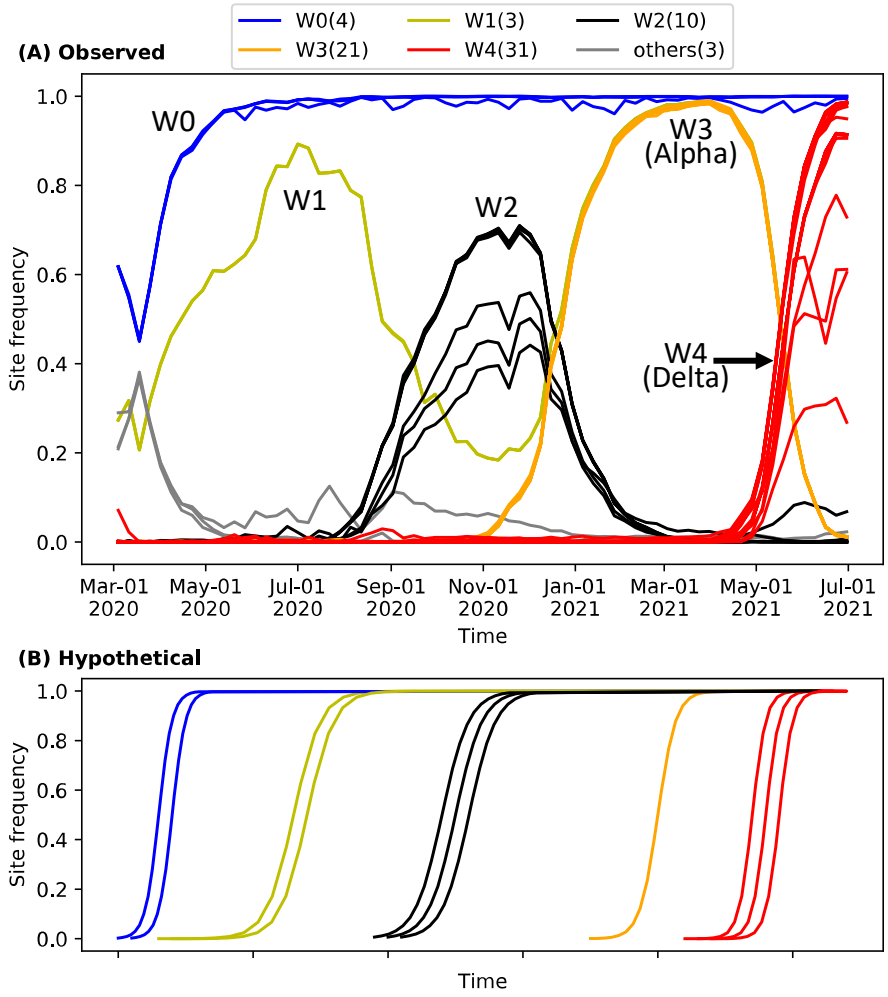
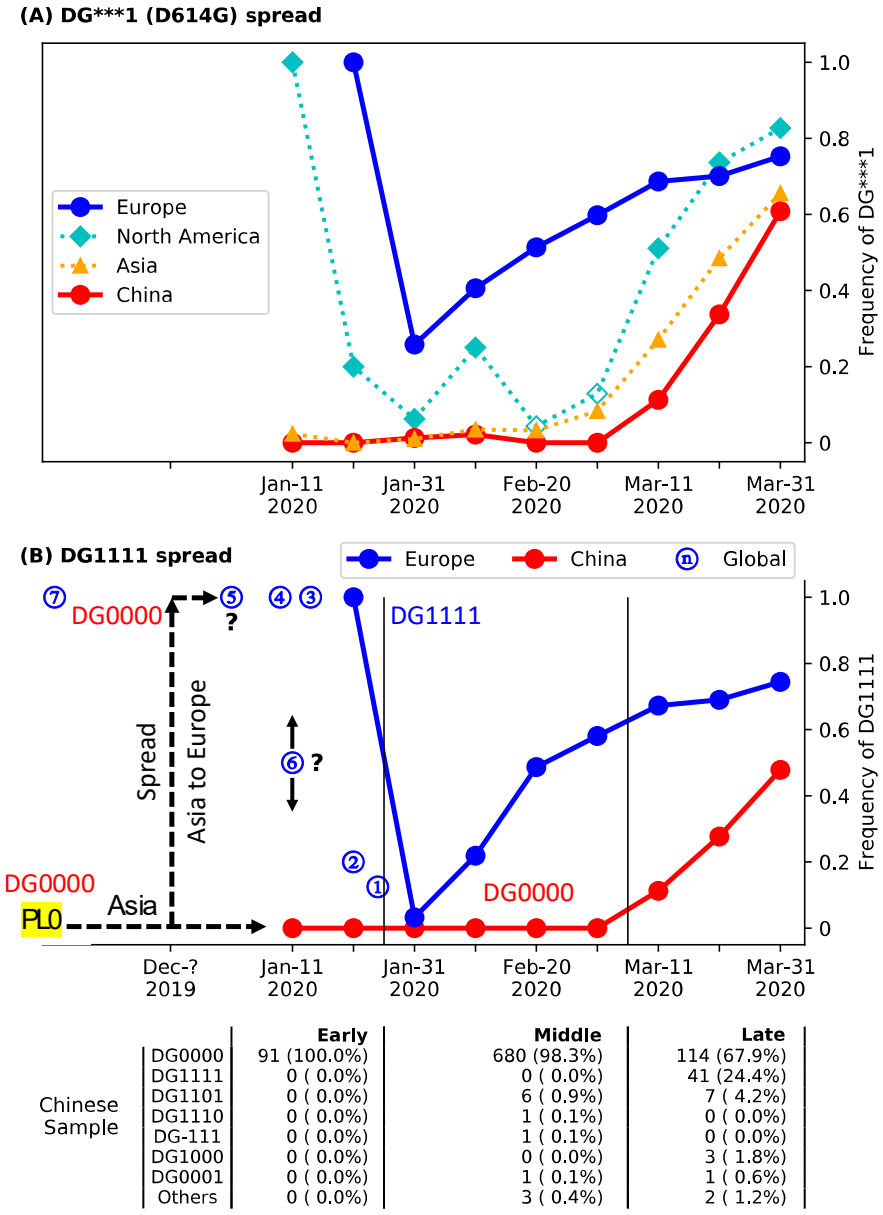


Fig. 1. The observed vs. hypothetical evolution of site frequency in SARS-CoV-2. (A) Evolution of SARS-CoV-2 between Jan. 1, 2020 and July 1, 2021 depicted by waves (i.e., successions of "mutation groups") in UK. Sequencing data were obtained from the GISAID database [34]. The frequency of the mutant at each variable site (e.g., C→ T) is tracked but only variants that reach the frequency cutoff of 0.3 at their peaks are presented. While a curve represents the rise and fall of a variant, each observed curve usually represents multiple curves that overlap completely. In COVID-19, there are 5 waves (W0 to W4). In W2, the curves do not overlap completely, thus revealing multiple variants of similar dynamics. The numbers of variants [non-synonymous: synonymous: non-coding] for the 5 waves of W0 to W4 are: 2:1:1, 2:1:0, 4:5:1, 15:5:1 and 26:3:2 for W0 to W4, respectively. The total is given in the parentheses next to the wave label above the figure. W3 and W4 correspond to the Alpha and Delta strain while the focus of this study is the W0 wave. W0 has 4 mutations, collectively referred to as the DG group. **(B)** The hypothetical site-frequency evolution depicted in the conventional model of molecular evolution whereby mutations are sequentially fixed. Note the contrast between the two panels with only the W0 wave behaving as conventionally portrayed.

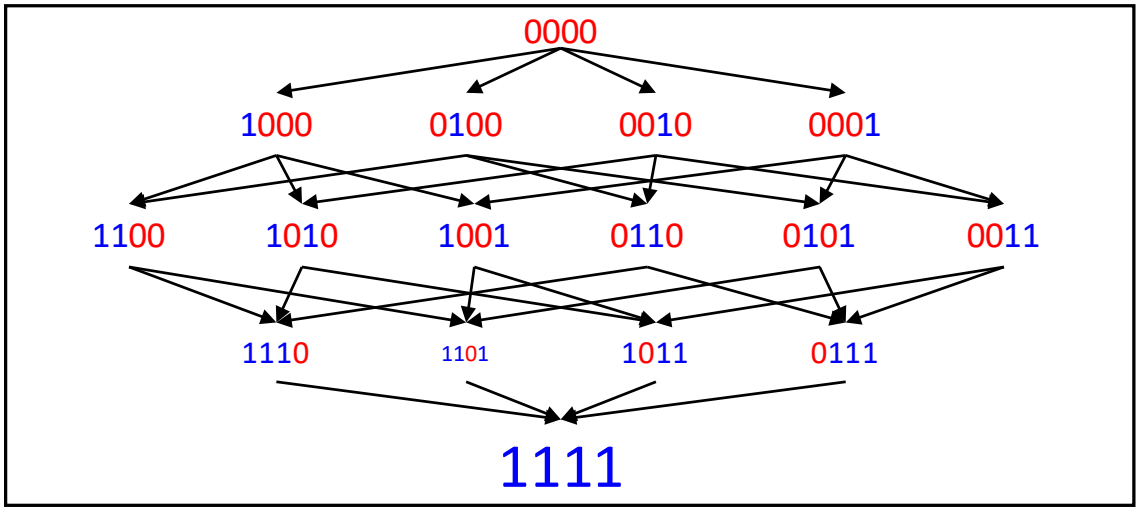


Chinese Sample	Early	Middle	Late
DG0000	91 (100.0%)	680 (98.3%)	114 (67.9%)
DG1111	0 (0.0%)	0 (0.0%)	41 (24.4%)
DG1101	0 (0.0%)	6 (0.9%)	7 (4.2%)
DG1110	0 (0.0%)	1 (0.1%)	0 (0.0%)
DG-111	0 (0.0%)	1 (0.1%)	0 (0.0%)
DG1000	0 (0.0%)	0 (0.0%)	3 (1.8%)
DG0001	0 (0.0%)	1 (0.1%)	1 (0.6%)
Others	0 (0.0%)	3 (0.4%)	2 (1.2%)

Fig. 2. Detailed analyses of the W0 wave before March 2020 under the one-beginning scenario. (A) The frequency of the D614G mutation is tracked among the samples from China, Asia, Europe, and North America. The European samples have higher frequencies of the D614G mutation than the Asian samples at every time point with a drop in February due to the influx of the Asian strains. **(B)** Frequency change of the DG1111 haplotype (with all 4 mutations) in China and Europe. The evolution of the DG mutations is divided into three stages: Early (up to 1/21/2020)– all samples in China being DG0000; Middle (1/21 – 3/1) – a mixture but without DG1111; Late (after 3/1) – DG1111 present in China. While DG1111 is absent in China in the Early stage, it is present, and likely common, in Europe North America, and Africa as shown by circled numbers. If the one-beginning scenario is correct, DG1111 must have descended from the imported DG0000 of Asia (the black dotted arrow) and would have less than a month to make the transition. Since the transition should have taken > 1 year at the minimum, the one-beginning

scenario in untenable. Complete information on partial haplotypes (between DG0000 and DG1111) are given in Table 2. Circled numbers (see Table S6) – 1. Australia (1 DG1111 out of 8); 2. Canada (1/5); 3. Sierra Leone (2/2); 4. Japan (Norway type, 1/1); 5. Canada (31/31); 6. Utah, USA (X/14); 7. Italy (Amendola et al. 2021). Samples of 5 and 6 show irregularity in data deposition with either sequences or dates removed without explanation.

(A) Evolutionary routes to DG1111



(B) Haplotypes of samples of Jan. 2020

	Reference Position (28 sites)
	1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
	1 3 6 8 4 4 7 8 9 9 9 2 3 3 3 5 5 6 7 8 8 8 8 8 8 9 9
	2 0 0 4 7 3 4 5 8 1 2 8 0 0 4 8 5 7 6 2 1 8 8 8 8 8 0 5
	4 5 3 4 8 4 0 5 7 6 7 3 1 1 0 6 6 1 0 1 4 3 6 8 8 8 9 2
	1 9 7 6 2 0 8 1 7 7 4 9 7 2 3 8 3 3 4 3 4 3 9 1 2 3 5 7
EPI_ISL_402125|**Wuhan-Hu-1**	C C C G C C C G C A C T G G A G G T T C T C C G G G C G
DG0000 samples	
EPI_ISL_529213|China|2019-12-30	. . . . T . . . . C . . . .
EPI_ISL_406801|China|2020-01-05	. . . . T . . . . C . . . .
EPI_ISL_406030|China|2020-01-10	. . . . T . . . . C . . . . . T .
EPI_ISL_405839|China|2020-01-11	. . . . T . . . . C . . . . . T .
EPI_ISL_406593|China|2020-01-13	. . . . T . . . . C . . . . . T .
EPI_ISL_403932|China|2020-01-14	. . . . T . . . . C . . . . . T .
EPI_ISL_403935|China|2020-01-15	. . . . T . . . . C . . . . . T .
EPI_ISL_403933|China|2020-01-15	. . . . T . . . . C . . . . . T .
DG1111 samples	
EPI_ISL_2671842|Japan|2020-01-09	T . T . . . T . . . . G . . . . A A C . .
EPI_ISL_2716636|Sierra Leone|2020-01-14	T . T . . T T A T G A C T A G T . C C T . T . A A C . T
EPI_ISL_2716627|Sierra Leone|2020-01-14	T . T . . T T A T G A C T A G T . C C T . T . A A C . T
EPI_ISL_2631277|Poland|2020-01-14	T . T . . . T . . . . G . . . . T . . . . .
EPI_ISL_2835566|USA|2020-01-21	T T T . . . T . . . . G . T . . . . . T . . . . .
EPI_ISL_509505|Australia|2020-01-25	T . T . . . T . . . . G . . . .
EPI_ISL_3364539|USA|2020-01-28	T T T . . . T . . . . G . T . . . .
EPI_ISL_2426018|Norway|2020-01-29	T . T . . . T . . . . G . . . . A A C . .
DG1101 samples	
EPI_ISL_451345|China|2020-01-24	T . T . . . . G . . . .
EPI_ISL_416327|China|2020-01-28	T . T . . . . G . . . .
EPI_ISL_450198|Germany|2020-01-28	T . T . . . . G . . . .
EPI_ISL_450200|Germany|2020-01-28	T . T . . . . G . . . .
EPI_ISL_406862|Germany|2020-01-28	T . T . . . . G . . . .
EPI_ISL_450199|Germany|2020-01-29	T . T A . . . . G . . . .
EPI_ISL_1143993|Germany|2020-01-29	T . T A . . . . G . . . .
EPI_ISL_1143994|Germany|2020-01-30	T . T A . . . . G . . . .

Fig. 3. (A) The routes of evolution from DG0000 to DG1111. The route map shows the complexity of evolving a full haplotype. Note the low fitness of the 4 haplotypes that are one mutation short of DG1111. The fitness is reflected in the font size with DG1111 being the largest and DG1101 being the smallest (see Table 2). Hence, the DG0000 to DG1111 evolution might not have taken place in the current genetic background, thus hinting earlier evolution of DG1111

in an older background than portrayed in Fig. 2. **(B)** The genomic sequences of the 3 major haplotypes collected before 1/30/2020. The yellow highlight paints the 4 DG group mutations. Note that the youngest haplotype DG1111 is the most diverse. The diversity suggests that DG1111 may have been relatively common in Europe no later than late 2019.

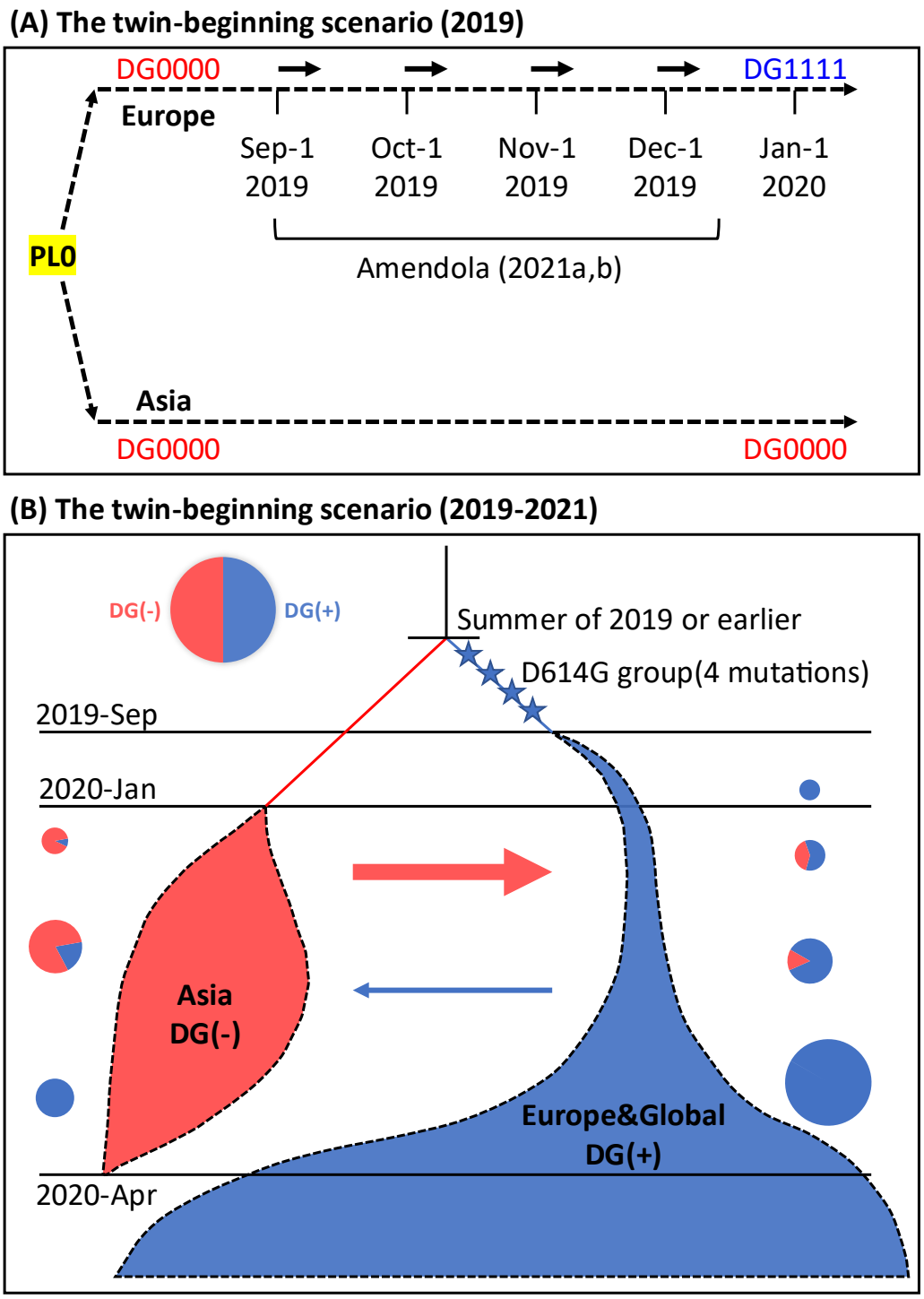


Fig. 4. The twin-beginning scenario. (A) In this scenario, the split between the Asian and European lines occurred before September of 2019. This scenario would allow more time for the evolution from DG0000 to DG1111 in Europe. The time period when SARS-CoV-2 was found in Italy is marked. It suggests that the evolution from DG0000 to DG1111 may have happened even earlier than indicated (see Amendola et al. 2021b). **(B)** The evolution of SARS-CoV-2 from the beginning. All time points are the sampling dates reported. The Asian and European

lineages may have coexisted but unnoticed before September of 2019. By Oct. 1, 2019, the DG group of mutations had already been assembled in Europe. In the subsequent period, from Oct. 1, 2019 to April 1, 2020, the two lineages interact via gene flow (shown by the arrows) and competition. The frequency of the DG(+) vs. that of DG(-) is shown in the pie charts for both populations. DG(+) or DG(-) means “predominantly DG1111 or DG0000”, respectively. The size of the pie chart reflects the number of infections at that time. From April 1, 2020 on, the global sweep of DG1111 mutations is nearly complete.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplement.pdf
- TableS1S7.xls