Cheng & Zhang. *Zool. Res.* 2023, 44(3): 494−504
https://doi.org/10.24272/j.issn.2095-8137.2022.535

ZR | Zoological Research

Article Open Access

Reconstructing early transmission networks of SARS-CoV-2 using a genomic mutation model

Chao-Yuan Cheng[1], Zhi-Bin Zhang[1,2,*]

[1] *State Key Laboratory of Integrated Management of Pest Insects and Rodents*, *Institute of Zoology*, *Chinese Academy of Sciences*, *Beijing* 100101, *China*

[2] *CAS Center for Excellence in Biotic Interactions*, *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*

ABSTRACT

The coronavirus disease 2019 (COVID-19) pandemic has greatly damaged human society, but the origins and early transmission patterns of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pathogen remain unclear. Here, we reconstructed the transmission networks of SARS-CoV-2 during the first three and six months since its first report based on ancestor-offspring relationships using BANAL-52-referenced mutations. We explored the position (i.e., root, middle, or tip) of early detected samples in the evolutionary tree of SARS-CoV-2. In total, 6 799 transmission chains and 1 766 transmission networks were reconstructed, with chain lengths ranging from 1–9 nodes. The root node samples of the 1 766 transmission networks were from 58 countries or regions and showed no common ancestor, indicating the occurrence of many independent or parallel transmissions of SARS-CoV-2 when first detected (i.e., all samples were located at the tip position of the evolutionary tree). No root node sample was found in any sample (n=31, all from the Chinese mainland) collected in the first 15 days from 24 December 2019. Results using six-month data or RaTG13-referenced mutation data were similar. The reconstruction method was verified using a simulation approach. Our results suggest that SARS-CoV-2 may have already been spreading independently worldwide before the outbreak of COVID-19 in Wuhan, China. Thus, a comprehensive global survey of human and animal samples is essential to explore the origins of SARS-CoV-2 and its natural reservoirs and hosts.

Keywords: SARS-CoV-2; Transmission chain; Transmission network; Ancestor-offspring relationship; *De novo* mutation; Back mutation; Secondary mutation

INTRODUCTION

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a novel betacoronavirus, is the pathogenic agent responsible for coronavirus disease 2019 (COVID-19) (Wu et al., 2020), with an estimated 760 million human infections and over 6.8 million deaths as of 12 March 2023 (WHO, 2023). However, the origins of SARS-CoV-2 remain unknown, and it is imperative that we reveal its early transmission patterns to better manage the current pandemic and prevent future ones.

As a novel betacoronavirus, SARS-CoV-2 is distinct from SARS-CoV and Middle East respiratory syndrome coronavirus (MERS-CoV) (Lu et al., 2020; Zhou et al., 2020a, 2020b), sharing only 79% and 50% genome sequence identity with SARS-CoV and MERS-CoV, respectively (Deng & Peng, 2020). Two coronavirus strains isolated in bat populations in Yunnan, China (*Rhinolophus affinis*) and northern Laos (*Rhinolophus malayanus*) share 96.1% (RaTG13) and 96.8% (BANAL-52) genetic similarity to the whole SARS-CoV-2 genome, respectively (Temmam et al., 2022; Zhou et al., 2020b). SARS-CoV-2 relatives are also found in bats from Cambodia and Japan (Mallapaty, 2020). The existence of diverse SARS-CoV-2 relatives suggests that bats could be potential reservoirs of SARS-CoV-2 (Lau et al., 2020; Paraskevis et al., 2020; Wong et al., 2020). Nevertheless, the differences between SARS-CoV-2 and bat coronaviruses are still large. For example, RaTG13 diverged from SARS-CoV-2 approximately 50 years ago (Shan et al., 2021). Thus, they are more likely to be evolutionary precursors than direct progenitors of SARS-CoV-2 (Zhang & Holmes, 2020).

Previous studies have shown that the mutation rate of SARS-CoV-2 is 6×10^{-4} to 1×10^{-3} bp/site/year (Chan et al., 2020; Duchene et al., 2020; Li et al., 2020b, 2020c, 2020d; Lu et al., 2020; van Dorp et al., 2020). As such, SARS-CoV-2 has produced many variants, with some found to exhibit much higher transmission capacities than earlier variants (Abdool Karim & de Oliveira, 2021; Burki, 2021; Lauring & Hodcroft, 2021). Based on molecular clock theory, the time of the most

Received: 17 March 2023; Accepted: 31 March 2023; Online: 31 March 2023

Foundation items: This study was supported by the Ministry of Science and Technology of the People's Republic of China (2021YFC0863400) and Institute of Zoology, Chinese Academy of Sciences (E0517111, E122G611)

*Corresponding author, E-mail: zhangzb@ioz.ac.cn

recent common ancestor (TMRCA) of SARS-CoV-2 is estimated to be November or early December 2019 (Duchene et al., 2020; Giovanetti et al., 2020; Hill & Rambaut, 2020; Lu et al., 2020), or October to early December 2019 (Li et al., 2020b; van Dorp et al., 2020), suggesting SARS-CoV-2 likely originated much earlier than when it was first detected in Wuhan, China. A study suggests that the COVID-19 outbreak in Wuhan could be associated with spillover events from cold-chains (Yu et al., 2022).

Many studies have used phylogenetic trees or haplotype networks to designate lineage or reconstruct the evolutionary patterns of SARS-CoV-2 (Forster et al., 2020; Nie et al., 2020; Song et al., 2020; Tang et al., 2020, 2021; Turakhia et al., 2020; Yu et al., 2020). However, given the molecular clock deviations (Cheng & Zhang, 2023) and high similarity among samples, phylogenetic trees alone cannot reveal the origin of SARS-CoV-2 (Morel et al., 2021; Pipes et al., 2021). Therefore, it is necessary to develop alternative or complementary approaches to reconstruct early transmission patterns of SARS-CoV-2.

In this study, we reconstructed the transmission network of SARS-CoV-2 by identifying ancestor-offspring relationships based on BANAL-52- and RaTG13-referenced mutations (i.e., bases different from BANAL-52 (or RaTG13) at corresponding sites) in samples collected worldwide in the first three and six months from 24 December 2019 to clarify the early transmission patterns of SARS-CoV-2. We tested the following three hypotheses: (1) If a common ancestral sample exists in samples collected during the first three or six months, the reconstructed lineages should be located at the bottom of the evolutionary tree, and the transmission network should consist of a single full tree with one common ancestor or one root node (Original Lineage Hypothesis; Supplementary Figure S1A); (2) If there is no common ancestral sample, but several samples are genetically very close to the ancestral sample, the lineages should be located in the middle of the evolutionary tree, and the transmission network should consist of a few large tree branches with a few root nodes (Intermediate Lineage Hypothesis; Supplementary Figure S1B); and (3) If all samples are genetically distant from the common ancestral sample, the lineages should be located in the tip of the evolutionary tree, and the transmission network should consist of many short and small tree branches or many root nodes (Tip Lineage Hypothesis; Supplementary Figure S1C). The Original Lineage Hypothesis denotes detection of the SARS-CoV-2 ancestor, Intermediate Lineage Hypothesis denotes detection of samples very close to the SARS-CoV-2 ancestor, and Tip Lineage Hypothesis denotes detection of samples very distant from the SARS-CoV-2 ancestor.

MATERIALS AND METHODS

Genome sequence processing

Genome sequences of SARS-CoV-2 were downloaded from the Global Initiative of Sharing All Influenza Data (GISAID, https://www.gisaid.org/) on 25 January 2021. The dataset contained 279 411 samples collected from 24 December 2019 to 21 January 2021. First, we used data collected in the first three months (i.e., 24 December 2019 to 22 March 2020) to reconstruct the early transmission network of SARS-CoV-2 (19 187 samples covering 95 countries or regions) using BANAL-52-referenced mutations. Second, we used data covering the first six months (84 243 samples covering 112 countries or regions) to validate our results. Third, we repeated the analysis based on the first three months of data using RaTG13-referenced mutations to validate the results. The samples represented the earliest detected samples, covering many countries. As the results were similar, we mainly reported findings based on BANAL-52-referenced mutations during the first three months, but also discussed results from the first six months with both BANAL-52- and RaTG13-referenced mutations. The genome sequence of BANAL-52 (GISAID accession number: EPI_ISL_4302644) was downloaded from GISAID and used as a reference for identifying mutations of SARS-CoV-2. We aligned the SARS-CoV-2 genome sequences to the reference sequence using Muscle v5. To minimize the potential impacts of sequencing errors, nucleotides at the 5' untranslated region (UTR) (sites 1–265) and 3' UTR (sites 29 675–29 903) were excluded.

Reconstruction of transmission network

We clustered all samples based on sequence differences. Samples with the same genome sequence were assigned to a transmission chain node. Each node in the transmission chain represented a unique sequence (similar to haplotypes) with distinct mutation sites, which may be composed of multiple samples from the same or different places. We reconstructed the transmission network of SARS-CoV-2 based on differences in BANAL-52-referenced mutations in all sites of each node. The transmission network was composed of many transmission chains, which were reconstructed to identify the closest ancestor node of each node according to the following mutation model (Figure 1A).

We defined mutations that occurred after the emergence of the most recent common ancestor (MRCA) of SARS-CoV-2 as *de novo* mutations. As directly determining *de novo* mutations is not feasible because the common ancestor is unknown, we inferred the mutations using BANAL-52 as an outgroup reference (see *Reference Selection* section). We assumed sequence S0 to be the MRCA of the other sequences (i.e., S1, S2, S3.1, S3.2, S4.1, and S4.2) in Figure 2. The following steps were applied to reconstruct the transmission chains and networks of SARS-CoV-2 based on the ancestor-offspring relationships (with distance of one or a few generations) of nodes (Figure 1A):

(1) We identified the ancestor nodes of each node. If sequence A carries all mutations of sequence B, in addition to its own new mutations, then A is considered the offspring of B. We performed pairwise comparison of sequences. *De novo* mutations contained in node X are set to M_X={m_1, m_2, ..., m_n}, if $M_U \in M_V$, then U is an ancestor node of V. The offspring node should have more mutations than its ancestor node, that is, offspring node mutations must include all ancestor node mutations. For example, as shown in Figure 1A, S0, S1, S2, and S3.1 are ancestor nodes of S4.1, and S0, S1, S2, and S3.2 are ancestor nodes of S4.2.

(2) For each node (i.e., unique sequence), we identified its closest ancestral node from all ancestor nodes. For all ancestor nodes identified in step (1), we selected the node with the closest mutation similarity to the focal node as its closest ancestral node. For example, in Figure 1A, S2 is the closest ancestral node of S3.1 and S3.2, S3.1 is the closest ancestral node of S4.1, and S3.2 is the closest ancestral node of S4.2.

(3) We connected all nodes with their closest ancestral nodes to form a transmission chain (Figure 1A2) or network

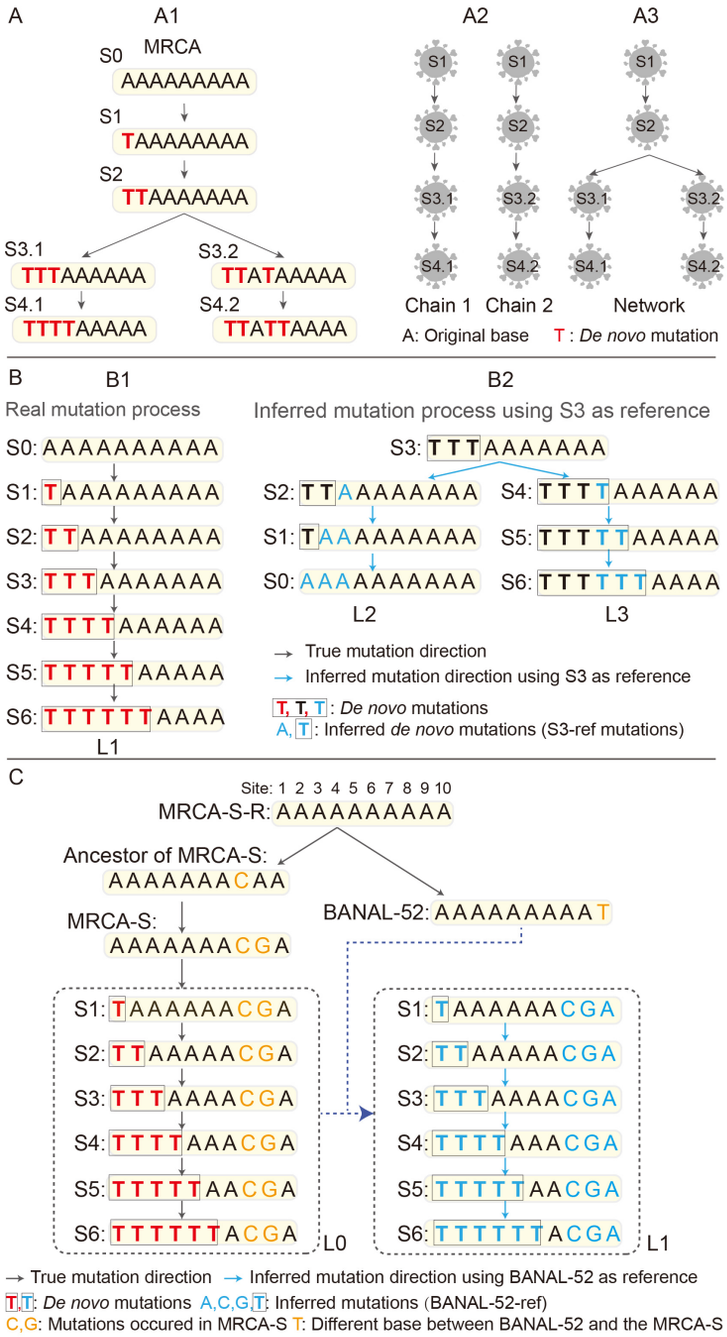

Figure 1 Illustrations of transmission chain and network reconstructions

A: Reconstruction of transmission chains and networks based on mutations using MRCA as a reference. For simplification, original nucleotides are lettered A and *de novo* mutation nucleotides are lettered T. A1: Ancestor-offspring relationship of SARS-CoV-2 virus based on *de novo* mutation sites. MRCA (S0) is the most recent common ancestor of all samples of SARS-CoV-2 (S1, S2, S3.1, S3.2, S4.1, S4.2). A2: Transmission chains (chains 1 and 2) were reconstructed from (A1). A3: Transmission network was reconstructed by merging chains 1 and 2, which share a common node in (A2). B: Definitions of *de novo* mutations using MRCA (S0) as a reference (S0-ref mutations) and inferred mutations using offspring (S3) as a reference (S3-ref mutation), and reconstruction of transmission chains and networks of SARS-CoV-2 based on ancestral (S0) and offspring (S3) sequences. Red and boxed T in L1 are mutated base types based on ancestral sequences (S0). Cyan A and Cyan boxed T in L2 and L3 are inferred mutations based on offspring sequence (S3). L1 is correctly reconstructed using *de novo* mutations. L2 and L3 are incorrectly reconstructed using S3-ref mutations, which appear later than ancestral node. C: Definitions of *de novo* mutations using MRCA-S (MRCA-ref mutations) or inferred mutations using BANAL-52 (BANAL-52-ref mutations) and reconstruction of transmission chains of SARS-CoV-2 using MRCA-S (L0) or an earlier and closer relative (BANAL-52) (L1). Red-boxed T is *de novo* mutation site using MRCA-S. Orange C and G are ancestor mutations of SARS-CoV-2 using MRCA-S-R. Cyan boxed T is *de novo* mutations using BANAL-52 as a reference. Orange T is BANAL-52 mutations using MRCA-S-R as a reference. Cyan C and G are ancestor mutations of SARS-CoV-2 using BANAL-52 as a reference. Cyan A is an incorrect mutation using BANAL-52 as a reference. L0 is the true mutation chain using MRCA-S as a reference. Lineage L1 was correctly reconstructed using a non-ancestor relative (BANAL-52) as a reference, which appeared earlier than the detected samples of SARS-CoV-2. Cyan C, G, and A were not considered in the reconstruction of the transmission chain of L1 if no secondary mutation occurred on these sites.

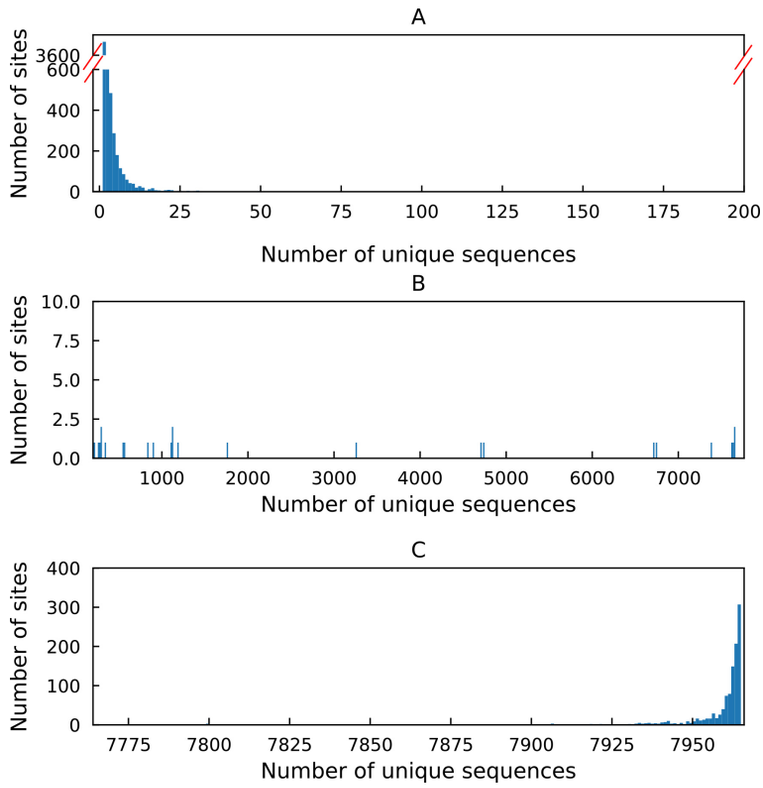

Figure 2 Frequency of number of BANAL-52-ref mutation sites in SARS-CoV-2 genome against number of unique sequences (nodes) that contain these sites

A: 1 to 200 unique sequences. B: 201 to 7 718 unique sequences. C: 7 719 to 7 918 unique sequences.

(Figure 1A3). A transmission network contains several transmission chains that share a common root node.

Following procedures (1), (2), and (3), we determined the ancestral nodes of all nodes, and finally reconstructed the transmission chains (Figure 1A). Because some chains shared a common node (Figure 1A2), we reconstructed the transmission network by merging chains sharing common nodes (Figure 1A3). Of note, the ancestor-offspring relationship may have distance of one (i.e., parent-offspring relationship) or a few generations due to missing samples.

Sample clustering (i.e., identification of nodes) and reconstruction of transmission chains and networks were implemented using custom scripts in Python v3.7.

Reference selection

A reference sequence is required to determine the *de novo* mutations of a node. As the MRCA of SARS-CoV-2 (Figure 1A) is unknown, we chose a detected sequence close to SARS-CoV-2 as the reference sequence.

If we use a reference that is an offspring of SARS-CoV-2 (i.e., in-group sequence), the mutations inferred from the reference may be incorrect. As shown in Figure 1B, using *de novo* mutations, the correct transmission chain based on the ancestor-offspring relationship defined above should be L1: S0→S1→S2→S3→S4→S5→S6. However, if we select an offspring as the reference (e.g., S3), identification of the S3 inferred mutation (S3-ref mutation) sites may be incorrect for some sequences (i.e., cyan A in L2), resulting in two short chains (L2, L3). The transmission direction or ancestor-offspring relationship after S3 is still correct (i.e., L3), but incorrect for sequences before S3 (i.e., L2). Some S3-ref mutations (cyan T) in L3 are correct *de novo* mutations, while other S3-ref mutations (cyan A) are not. Thus, using an in-group sequence (as commonly used) will cause errors in the identification of *de novo* mutations, and thus ancestor-offspring relationships.

Alternatively, we selected the closest non-SARS-CoV-2 relative sequence (i.e., BANAL-52) as the out-group reference to infer mutations in SARS-CoV-2 (i.e., BANAL-52-ref mutations), and assessed its potential in transmission chain reconstruction. BANAL-52 was sampled in 2020 and is the closest known relative to SARS-CoV-2, with a 3.2% difference in the whole genome (Temmam et al., 2022). All or most SARS-CoV-2 sequences should carry BANAL-52-referenced mutations from the 3.2% differential sites in the two genomes, but only a limited number of sequences will carry *de nova* mutations at these sites. As shown in Figure 1C, we assumed that, compared to the MRCA between SARS-CoV-2 and BANAL-52 (i.e., MRCA-S-R), BANAL-52 has a mutation (orange T) at site 10 (from left to right), ancestor of MRCA of SARS-CoV-2 has a mutation at site 8 (orange C), and MRCA of SARS-CoV-2 (i.e., MRCA-S) has two mutations (orange C, G) at sites 8 and 9; compared to MRCA-S, SARS-CoV-2 has 1–6 *de novo* mutations (boxed red T) at sites 1–6, and two inherited ancestor mutations (orange C, G) from MRCA-S and ancestor of MRCA-S at sites 8 and 9. Thus, if we use MRCA-S as the reference, the transmission chain of SARS-CoV-2 should be reconstructed as L0: S1→S2→S3→S4→S5→S6. If we use BANAL-52 as the reference, the *de novo* mutations would be correctly identified in L1 (i.e., boxed cyan T), and are the same as the red and boxed T in L0. The ancestor mutations of SARS-CoV-2 in L1 (cyan C, G) would be correctly identified. However, the mutation of BANAL-52 (orange T) would be identified as an incorrect or background mutation in L1 (cyan A) because mutation of the BANAL-52-reference results in a base difference at site 10 between

BANAL-52 and SARS-CoV-2. Because both inherited ancestor mutations of SARS-CoV-2 (sites 8 and 9) and incorrectly inferred mutations using BANAL-52 (site 10, named as incorrect or background mutation) would be carried by all SARS-CoV-2 samples, they are not considered in transmission chain reconstruction based our method if no secondary mutation occurs on these sites (see below). Using the BANAL-52-ref mutations, the transmission chain can be reconstructed as L1: S1→S2→S3→S4→S5→S6, which is the same as S0 (Figure 1C). This lays a solid foundation for reconstructing the transmission chains and networks using BANAL-52 as the reference. As shown in Figure 1C, we introduced mutations of C and G to illustrate the ancestor mutation of SARS-CoV-2.

Model errors

If using BANAL-52-referenced mutations to reconstruct the transmission chains and networks, secondary mutations at the BANAL-52-referenced mutation sites in L1 (Figure 1C) would cause errors in the reconstruction, e.g., a secondary mutation can change the base back into its original (i.e., back mutation). Our model contains three kinds of BANAL-52-referenced mutations in L1: i.e., *de novo* (cyan T), ancestor (cyan C, G), and incorrect mutations (cyan A), used for reconstructing transmission chains and networks (Figure 1C). If no secondary mutation occurs in these SARS-CoV-2 mutation sites during the study period, the reconstructed transmission chain (L1) is the same to the true one (L0) (Figure 1C); otherwise, it will cause reconstruction errors (see Supplementary Figure S2).

As shown in Supplementary Figure S2, a secondary mutation in the BANAL-52-referenced mutation sites during the three-month study period would cause biases in transmission chain reconstruction. If no secondary mutation occurs, the original transmission chain should be: L0=S1→S2→S3→S4→S5→S6. If a secondary mutation (T→A, back mutation) occurs in the *de novo* mutation sites of a sequence (S4), and the sequence has no extra copies, the mutated sequence (S4b in Supplementary Figure S2B2) would become an isolated chain (L1.2n=S4b, Supplementary Figure S2B2), with the relationships of other samples on the chain remaining unchanged, except for the absence of S4, i.e., L1.1n=S1→S2→S3→S5→S6 (Supplementary Figure S2B2). If S4 has extra copies, however, the original chain would not be affected, i.e., L1.1c=S1→S2→S3→S4→S5→S6, but the secondary mutation would still produce an isolated chain, L1.2c=S4b (Supplementary Figure S2B1). If a secondary mutation occurs in the *de novo* mutation sites of a sequence (S4), and the mutated sequence (S4b) is the same as its ancestor sequence (S3), the original chain would remain unchanged if S4 has extra copies (i.e., L2c=S1→S2→S3→S4→S5→S6) or produce a shorter chain in the absence of S4 (i.e., L2n=S1→S2→S3→S5→S6) (Supplementary Figure S2C). If a secondary mutation occurs in the site corresponding to the BANAL-52 mutation site (cyan A→purple T, cyan A is an incorrect mutation) and the sequence (S4) has extra copies, the mutated sequence (S4b) would become an independent chain: L3.2c=S4b, while the relationships of other samples on the chain would remain unchanged: L3.1c=S1→S2→S3→S4→S5→S6 (Supplementary Figure S2D1). However, if the sequence (S4) has no extra copies, the original chain would break into two chains with the same probability, i.e., L3.1n=S1→S2→S3→S5→S6 and L3.2n=S4b→S5→S6 (Supplementary Figure S2D2). A secondary mutation in the ancestor mutation sites (i.e., C, G in Figure 1C) would produce similar results to cyan A (for simplification, these sites are not shown in Supplementary Figure S2). Because the proportion of the sequence with two or more secondary mutations was extremely low in this study, we only presented illustrations of one secondary mutation in Supplementary Figure S2.

Considering the potential influence of model error on ancestor-offspring relationships in about 20% of all sequences during the first three months (higher for sequences over the first six months) (see below for details), we focused on the analysis and discussion of general transmission patterns of SARS-CoV-2, such as number of chains and networks, rather than specific sequences.

Error probability caused by secondary mutations

If no secondary mutations occur during the first three months, the transmission chain (L1) is the same as the true one (L0) (Figure 1C). However, if a secondary mutation occurs in the BANAL-52-ref mutation sites, transmission chain reconstruction errors will arise. BANAL-52-ref mutations can include *de novo* (cyan T), ancestor (cyan C, G), and incorrect mutations (cyan A) in L1 (Figure 1C).

In theory, the probability of secondary mutations in *de novo* mutations (p_d) can be determined by the mutation probability of *de novo* mutations of SARS-CoV-2 within three months. Here, the mutation rate was assumed to be 6×10^{-4} substitutions per site annually (van Dorp et al., 2020), and the mutation probability of SARS-CoV-2 within three months was calculated as: $P_T=6\times10^{-4}/4=1.5\times10^{-4}$. The proportion of *de novo* mutation within three months was assumed to be 1.5×10^{-4}. Therefore, the error probability of secondary mutation in *de novo* sites was calculated as: $p_d=(P_T)^2=(1.5\times10^{-4})^2=2.25\times10^{-8}$. The probability of samples with such secondary mutations (i.e., number of secondary mutations ≥1) was estimated as $1-(1-p)^n$ (where p is the secondary mutation rate and n is the number of bases of SARS-CoV-2). For $p_d=2.25\times10^{-8}$ and n=29 410, the probability of samples with secondary mutations was estimated to be 6.6×10^{-4}. Thus, we predicted that 0.066% of samples containing secondary mutations in *de novo* mutation sites would be reconstructed into short chains. In our study, because each node sequence contained 2.4 copies (=19 187/7 918), the probability of breaking the original chains was calculated as: $(6.6\times10^{-4})^{2.4}=2.3\times10^{-8}$, which is very small.

Based on the whole genome (P_R), BANAL-52 exhibits a 3.2% dissimilarity to the MRCA of SARS-CoV-2, which can be attributed to ancestor (cyan C and G) and incorrect mutation sites (cyan A). Secondary mutation in these sites would cause model errors during reconstruction of the SARS-CoV-2 transmission chains and networks. The mutation rate during the three-months study period was assumed as: $P_T=1.5\times10^{-4}$ substitutions per site within three months. Thus, the error probability caused by secondary mutations in ancestor and incorrect mutation sites was calculated as: $p_a=P_R\times P_T=3.2\times10^{-2}\times1.5\times10^{-4}=4.8\times10^{-6}$. The probability of a sample having a secondary mutation at the ancestor or BANAL-52 mutation sites was estimated as $1-(1-p)^n$ (where p is the total error probability caused by secondary mutations in mutation sites of the ancestor of SARS-CoV-2 and BANAL-52 and n is the number of SARS-CoV-2 genome bases). For $p=4.8\times10^{-6}$ and n=29 410, the probability of a sample having secondary

mutations in ancestor and incorrect mutation sites (cyan C, G, A in Figure 1C) was estimated to be 13.2%. In our study, because each node sequence contained 2.4 copies, the probability of breaking the original chains was calculated as: $(13.2\%)^{2.4}=0.78\%$. Thus, using BANAL-52 as the reference would have minimal influence on the original transmission chains and networks but may produce 13.2% short transmission chains.

Error probability caused by sequence gap or uncertainty

Similar to secondary mutations, degenerate bases or gaps in the genome sequence due to sequencing error or uncertainty may cause biased estimation of evolutionary tree reconstruction. Base uncertainty is often treated as no mutation in evolutionary studies (as in our study). Here, there were 2.9 degenerate (or missing) bases per sequence, on average, and the proportion of degenerate or missing bases was calculated as: $P_D=2.9/29410=9.8\times10^{-5}$. Therefore, the model error probability caused by degenerate or missing bases in *de novo* mutations was calculated as: $p_s=P_D\times P_T=9.8\times10^{-5}\times1.5\times10^{-4}=1.47\times10^{-8}$. Similarly, model error probability of ancestor and incorrect mutations was calculated as: $p_s=P_D\times P_R=9.8\times10^{-5}\times3.2\times10^{-2}=3.1\times10^{-6}$. The probability of samples with degenerate or missing bases (i.e., number of degenerate or missing bases ≥1) was estimated as $1-(1-p)^n$ (where p is the degenerate mutation rate and n is the number of bases of SARS-CoV-2). Letting $p=1.47\times10^{-8}$ or 3.1×10^{-6} and $n=29\,410$, then $p=4.3\times10^{-4}$ or 8.7%. Thus, we predicted a total of 8.7% samples would be reconstructed into short transmission chains caused by degenerate or missing bases. Based on each node sequence containing 2.4 copies, the probability of breaking the original chain due to a degenerate base was calculated as: $(8.7\%)^{2.4}=0.29\%$. Therefore, degenerate mutations or missing bases would have minimal influence on the reconstruction of original transmission chains or networks, but may produce 7.13% short transmission chains.

Simulation analysis

To validate our SARS-CoV-2 transmission network reconstruction method based on the paternity relationship described above, we simulated the occurrence of mutations based on how virus sequences replicate in nature. We chose a sequence with a length of 1 000 bp as the starting sequence to simulate the proliferation process of the sequence. Each sequence produced three progeny sequences in one replication cycle (one generation). We set the mortality rate of sequences to 20% and the mutation rate to 0.1% for each base within a replication cycle. We chose a shorter sequence and a higher mutation rate to save computation time.

We simulated the evolutionary process of a sequence for 10 generations and chose the first sequence in the 10th generation as a template to produce six generations as the detected samples (similar to the detected SARS-CoV-2 samples after late December 2019) for subsequent analysis (Supplementary Figure S3). We chose the last sequence in the 9th generation as the reference sequence (similar to BANAL-52) (Supplementary Figure S3). We selected three-group data from the detected samples: (1) detected samples from the 0th to 6th generation (0th generation represents starting sequence of the simulated data), which includes the common ancestor; (2) detected samples from the 2nd to 6th generation, with missing data (similar to the undetected data) from the 0th to 1st generation; and (3) detected samples from the 4th to 6th generation, with missing data from the 0th to 3rd generations. Using these three-group data, we reconstructed the transmission chains and networks based on paternity relationships, tested the three hypotheses (Supplementary Figure S1), and validated our method. The simulation process was implemented using custom scripts in Python v3.7.

Validation using RaTG13-referenced mutations

To validate our findings, we repeated the same analyses on data from the first three months using RaTG13-referenced mutations. The bat coronavirus RaTG13 genome sequence (GenBank accession number: MN996532) was downloaded from GenBank (https://www.ncbi.nlm.nih.gov/genbank/).

RESULTS

Frequency of BANAL-52-referenced mutations

Among the 29 410 nucleotide sites in the SARS-CoV-2 genomes, 7 062 (Figure 2) exhibited different bases from BANAL-52 (from 19 187 sequences). These different bases were defined as BANAL-52-referenced mutations and were used to reconstruct the transmission chains and networks. We identified 7 918 unique sequences (equivalent to haplotypes) from the 19 187 samples according to the composition of the BANAL-52-referenced *de novo* mutations, with each unique sequence representing a node in the transmission chain or network. The frequency of BANAL-52-referenced mutations showed a U-shaped relationship with the number of unique sequences. Most mutations contained less than 20 unique sequences (Figure 2A), indicating these mutations were more likely *de novo* mutations of SARS-CoV-2. Some mutations occurred in the unique sequences with sample sizes larger than 7 910 (Figure 2C), indicating these mutations were likely ancestor and incorrect mutations inferred using BANAL-52 and were carried by nearly all samples. Mutation sites in the middle were more likely early *de novo* mutations or secondary mutations in ancestor or incorrect mutations of SARS-CoV-2 (Figure 2B).

Frequency of transmission chains with different length

The frequency of transmission chains with different lengths (i.e., number of nodes) is shown in Figure 3 and Table 1. Among the 7 918 nodes of the transmission networks, there are 1 766 root nodes (2 847 samples) from 58 countries/regions, which had no common ancestor of SARS-CoV-2 (Table 1). We reconstructed 6 799 transmission chains with number of nodes ranging from 1 to 9, forming 1 766 networks.

The average transmission chain length was estimated to be 3.7±2.1 (Figure 3A). If only the longest chain with a common root sample in a transmission network was counted, the number of transmission chains decreased rapidly with the increase in chain length (Figure 3B).

There were several large networks in terms of number of nodes when using data of the first three months and the first six months. Samples of SARS-CoV-2 were not evenly distributed in the different transmission networks. For networks reconstructed using data from the first three months, there were five large networks (length ≥7), with nodes and samples accounting for 55% and 66% of all study samples, respectively.

Frequency of sampling time of root-node samples

The sampling time of root-node samples ranged from days 16

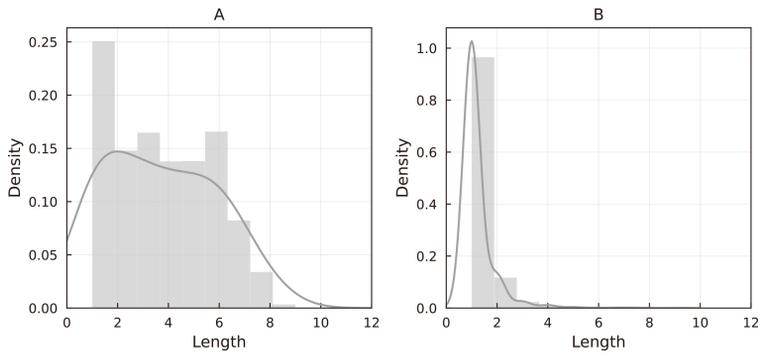

Figure 3 Kernel density of transmission chains with different chain lengths (i.e., number of nodes) in transmission chains and networks of SARS-CoV-2

A: Transmission chains. B: Transmission network counting only the longest transmission chain.

Table 1 Statistics of reconstructed transmission chains and networks with different lengths within first three months

Chain length	No. chains	No. root nodes (networks)	No. root samples	No. root countries/regions	First sampling time	Last sampling time
1	1 515	1 515	2 207	55	16	90
2	894	219	585	31	18	90
3	996	64	191	22	18	90
4	833	30	116	17	18	89
5	835	13	70	13	18	89
6	1 002	7	37	9	18	75
7	498	5	28	8	30	75
8	205	2	14	4	30	63
9	21	2	14	4	30	63
Total	6 799	1 766	2 847	58	16	90

No. chains represents number of chains of corresponding length. No. root nodes represents number of root nodes of corresponding chains (one root node corresponds to one transmission network). No. root samples represents number of samples of root nodes. No. root country/regions represents number of sampling countries and regions of root nodes. First sampling time represents first sampling time of the root nodes (i.e., number of days since 24 December 2019 (set as day 1)). Last sampling time represents final sampling time of samples in root nodes.

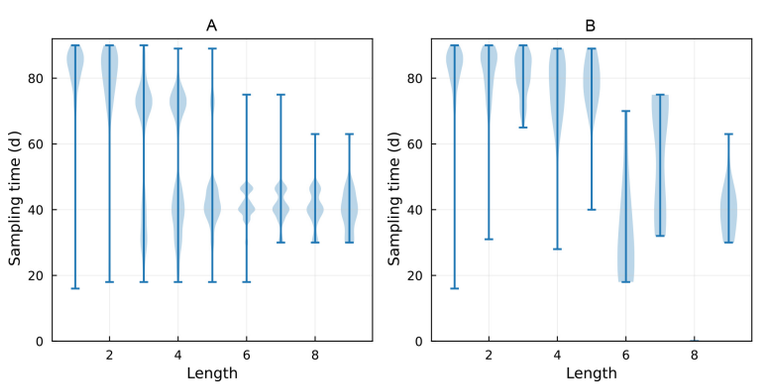

Figure 4 Frequency of sampling time (day) of root node samples of different chain lengths

A: Transmission chains. B: Transmission networks (length is length of longest chain it contains). Lower end of the line represents the earliest sampling time of root nodes, upper end represents the latest sampling time of root nodes, and shade width indicates the density of samples corresponding to the sampling time.

to 90 (Table 1; Figure 4). None of the samples collected from the first 15 days (i.e., from 2019/12/24 to 2020/1/8) were root-node samples of the transmission chains (Table 1; Figure 4). Samples of root nodes with chain lengths of 1–4 were primarily collected before day 50 and after day 65, while chains with lengths ≥5 were collected from days 30 to 50 (Figure 4A). As shown in Figure 4A, some chains shared the same root node, which could result in duplicate counting of root nodes. If each root node and the longest chain were counted in the transmission network (corresponding to Figure 3B), samples of root nodes in the transmission network with the longest length ranging from 1 to 5 were mainly collected after day 60, while those with the longest length of ≥ 6 were collected between days 20 to 70 (Figure 4B).

Illustration of transmission networks

For simplification, we only presented reconstructed transmission networks with chain lengths consisting of four (Supplementary Figure S4A), five (Supplementary Figure S4B), six (Supplementary Figure S4C), seven (Supplementary Figure S4D), eight (Supplementary Figure S4E), and nine nodes (Supplementary Figure S4F). Based on these figures, the detected transmission network of SARS-CoV-2 was

composed of many short transmission chains, like short fragmented "tree branches". Many spiking nodes were also detected, which caused super transmissions.

Simulation results

The simulation results indicated that the reconstructed transmission network using the detected data covering the 0th to 6th generations consisted of two independent networks, with an average length of 4.8. The transmission network covering the 2nd to 6th generations had 13 independent networks, with an average length of 3.8. The network covering the 4th to 6th generations had 52 independent networks, with an average length of 2.6 (Table 2; Supplementary Figure S5). These results indicate that collecting samples in the later stages of the evolutionary tree enables the reconstruction of transmission networks with shorter chains and more root nodes (number of root nodes is equivalent to number of transmission networks).

DISCUSSION

Currently, the origins and early transmission of SARS-CoV-2 remain unclear. Here, using 19 187 samples of SARS-CoV-2 collected over the first three months from 24 December 2019, we reconstructed the transmission chains and networks based on ancestor-offspring relationship of all samples using BANAL-52-referenced mutations. We identified 1 766 independent transmission networks of SARS-CoV-2 without a common ancestor sample, with the five largest networks (length ≥7) accounting for 66% of all samples. No root-node sample from the first 15 days was found from the Chinese mainland. Our results indicate that all detected samples during the three-month study period were not common ancestors or close to the common ancestor of SARS-CoV-2, supporting the Tip Lineage Hypothesis. We performed the same analysis using RaTG-13-referenced mutations with the three-month data and BANAL-52-referenced mutations with the six-month data and obtained similar results to those obtained using BANAL-52-referenced mutations with three-month data (see Supplementary Tables S1, S2). Simulation analysis indicated that our reconstruction method was robust, and the results were consistent with our hypothesis predictions. Our findings suggest that SARS-CoV-2 may have been spreading globally long before its first report in Wuhan, China. As such, a comprehensive worldwide survey is required to further explore the origins and natural reservoirs of SARS-CoV-2.

Revealing the early transmission patterns of SARS-CoV-2 is important for preventing future viral spillovers. However, the origins and natural hosts of SARS-CoV-2 are still unknown (Lundstrom et al., 2020; Peng et al., 2021; Tong et al., 2021). As bats are natural reservoirs of many coronaviruses, and the closest known coronavirus to the SARS-CoV-2 genome is from bats (i.e., BANAL-52), bats are suggested as the potential natural host of SARS-CoV-2 (Fan et al., 2019; Lau et al., 2020; Li et al., 2005; Zhang & Holmes, 2020). Recent studies found several bat species from East Asia and Southeast Asia that carry SARS-CoV-2-like viruses (Delaune et al., 2021; Temmam et al., 2022; Wacharapluesadee et al., 2021; Zhou et al., 2021). Notably, some SARS-CoV-2-like viruses collected from Laos exhibit high similarity to SARS-CoV-2, with their receptor-binding domains binding to the human ACE2 protein as efficiently as SARS-CoV-2 (Temmam et al., 2022). Although the pangolin coronavirus is not as similar to SARS-CoV-2 as BANAL-52 at the whole-genome level, its receptor-binding domain (e.g., GD410721) is closer to SARS-CoV-2, suggesting that pangolins may be an intermediate host (Lam et al., 2020; Xiao et al., 2020). A joint-report by the World Health Organization (WHO) and Chinese research teams concluded that SARS-CoV-2 very likely originated from nature, extremely unlikely from a laboratory (Wang & Zhao, 2021; Wu et al., 2021). Ruan et al. (2022) explored the evolution of SARS-CoV-2 during the first wave of the pandemic (early 2020) based on genomic data and suggested that European strains may have spread in parallel with Asian strains, with the European strains globally supplanting the Asian strains by May of 2020. Based on the number of cumulative mutations, Cheng & Zhang (2023) reported that SAR-CoV-2 showed much earlier global divergence than in the Chinese mainland. Here, our results revealed 1 766 independent transmission networks widely distributed in 58 countries, which did not share a common ancestor, indicating possible independent parallel spread of SARS-CoV-2 in many parts of the world before detection in Wuhan, China. Notably, no root node from samples collected in the first 15 days (n=31, all from the Chinese mainland, see below) was detected, indicating that the early detected samples from the Chinese mainland were not the ancestors of the global SARS-CoV-2 samples collected during the first three months. Using mutation network analysis, these results suggest that SARS-CoV-2 may have been circulating worldwide long before it was detected in Wuhan, China, consistent with our previous study using the cumulative mutation method (Cheng & Zhang, 2023).

The results of our analysis identified a considerable number of short transmission chains, which may be attributed to the root-node samples being collected at a later stage (Table 1; Figure 4). Indeed, samples of root nodes with chain lengths ranging from 1 to 4 were mainly collected after day 60 (Figure 4A). Given that samples of SARS-CoV-2 from most countries or regions outside the Chinese mainland were primarily reported in the third month, the absence of data before this period may account for the observation of many short chains. As described in the Methods section, a proportion of short transmission chains can be attributed to secondary mutations and sequencing gaps or uncertainty (13.2% and 8.7%, respectively); thus, some of the observed short chains will be incorrect transmission chains. Conversely, long transmission chains were likely due to the early appearance of root samples that survived for a longer period (Figure 4). Specifically,

Table 2 Parameters of transmission chains and networks reconstructed using simulated data

Data range	No. root nodes	No. chains	Average chain length
0–6	2	267	4.8
2–6	13	264	3.8
4–6	52	251	2.6

Data range indicates generations used to reconstruct transmission chains. Number of root nodes indicates number of different sequences at root node, equivalent to the number of independent networks.

samples of root nodes with chain lengths >5 were mainly collected before day 60 (Figure 4A). Despite the potential for error caused by secondary mutations and sequencing uncertainty, our approach only has a small influence on the reconstruction of the original transmission chains when considering the extra copies of SARS-CoV-2 samples. Notably, the error probabilities caused by secondary mutations and sequencing error or uncertainty were estimated to be 0.78% and 0.29%, respectively, suggesting that the general pattern of the transmission networks revealed by our approach is highly reliable.

A secondary mutation was deemed to have occurred when the bases at the BANAL-52-referenced sites underwent two or more mutations during the study period, referred to as a back mutation if the base mutates back to its original base (Ellis et al., 2001). Back mutations may obscure true evolutionary distance estimations between sequences when building a phylogenetic tree, known as homoplasy (Patwardhan et al., 2014). Other types of secondary mutations can cause similar errors. According to our estimation, within three months, secondary mutations on de *novo* mutation sites and ancestor or incorrect mutation sites could account for 13.2% and 8.7% of short transmission chains, highlighting the need for caution when constructing evolutionary trees of SARS-CoV-2. However, the presence of extra copies of SARS-CoV-2 samples significantly reduced model error in revealing the original transmission chains and networks in our study, with only 0.78% and 0.29% of original chains broken into short chains.

Sampling bias is an important problem encountered in data analysis, including of SARS-CoV-2 samples (Liu et al., 2020). Indeed, some lineages may not have been detected in the early stage of the pandemic because early SARS-CoV-2 infection may have been diagnosed as influenza. As shown in Supplementary Table S1, more independent transmission networks were detected when using samples from the first six months than when using samples from the first three months.

Phylogenetic trees are often applied to analyze the evolutionary relationships among viruses (Holmes et al., 1995; Lanciotti et al., 2002; Poon et al., 2013). Several studies have explored the phylogeny of SARS-CoV-2 using unrooted trees (Li et al., 2020a; Nabil et al., 2021), in which sampling time is used to specify a hypothetical root, resulting in the first detected sample being at the root. Although these unrooted trees may be suitable for clade/lineage classification, they are not adequate for tracing the origin of SARS-CoV-2. Phylogenetic trees typically rely on the assumption of a constant mutation rate and genetic similarity between sequences to establish ancestor-offspring relationships, which may not hold if the mutation rate varies significantly. Therefore, phylogenetic trees alone cannot reveal the origins of SARS-CoV-2 (Pipes et al., 2021). Different from phylogenetic tree approaches, our method does not rely on the assumptions of a strict molecular clock nor on sampling time.

The haplotype network method has been widely used to study the relationships between different clades and lineages of SARS-CoV-2 (Liu et al., 2020; Sekizuka et al., 2020; Tang et al., 2020). The nodes in our transmission network represent unique sequences, similar to haplotypes. However, our transmission network differs from haplotype networks in the connecting principle of nodes and the meaning of transmission chains. There are three key distinctions between our approach and haplotype networks. First, haplotype networks generally connect haplotypes using distance matrix-based algorithms, e.g., median-joining networks (Bandelt et al., 1999), which rely on overall similarity between haplotypes (Kong et al., 2016), whereas our transmission network is constructed based on ancestor-offspring relationships between haplotypes in the transmission chain. Second, haplotype networks do not provide information on transmission direction and may designate the node of the earliest sample as the origin node. In contrast, our transmission network provides transmission direction from ancestor to offspring sequences and can be used to trace the source and transmission of viruses. Third, all haplotypes in a haplotype network are connected to a single network, even if the haplotypes do not exhibit an ancestor-offspring relationship. In contrast, in our transmission network, only sequences with an ancestor-offspring relationship are connected, with several or many independent transmission chains constructed if no common ancestor is detected. As shown in Supplementary Figure S6, the similarity between sequences S0 and S1 is the same as that between sequences S0 and S2. Based on our approach, S0 and S1 are linked with the transmission direction from S0 to S1 but S0 and S2 are not linked based on the ancestor-offspring principle. However, phylogenetic trees and haplotype networks would link both S0 and S1 and S0 and S2 based on the similarity between each pair of sequences. Simulation results demonstrated that our method can accurately reconstruct the evolutionary tree based on the ancestor-offspring relationships between samples. Using incomplete sequencing data and an out-group reference, our approach can provide novel insight into the early transmission patterns of SARS-CoV-2, with broad application potential for studying the origins and transmission patterns of various viruses.

We further compared our results with the phylogenetic tree approach (Supplementary Figure S7). For a clear vision, we only used node samples (haplotypes) from the two longest networks (chain length=9) (Supplementary Figure S4F) and reconstructed a phylogenetic tree using maximum-likelihood method. Results indicated that the positions of the two networks corresponded well with the clades defined in GISAID (Supplementary Figure S7). Samples in the first network overlapped well with clades G and GH (GH accounted for 92%), and most samples in the second network overlapped well with clades V, O, and L (V accounted for 77%), except for some samples that appeared in clades G and GH (Supplementary Figure S7). All samples were forcibly connected using the phylogenetic tree approach (similar to the haplotype network method). According to our model, however, there was no common ancestor in the samples of the two networks.

We validated our model results using BANAL-52-referenced mutation data from the first six months. A total of 33 818 chains with 8 426 root nodes or networks were reconstructed (Supplementary Table S1). These root nodes contained 13 069 samples from 88 countries/regions. There were five large networks (length ≥10) and their nodes and samples accounted for 48% and 55% of the total sample size, respectively. More independent transmission networks were detected in the first six months compared to the first three months, indicating that many transmission chains or networks were not detected in the first three months. Additionally, there was no root sample in the first 15 days. The lengths of the transmission chains during the first six months ranged from 1

to 13, longer than those obtained during the first three months. Furthermore, the process of network reconstruction was considerably more time-consuming for the six-month dataset compared to the three-month dataset.

We further validated our model using RaTG-13-referenced mutation data from the first three months. In total, 6 737 transmission chains ranging from 1–9 nodes were reconstructed, consisting of 2 041 root nodes from 69 countries/regions. These root nodes and transmission networks showed no common ancestor. No sample in the first 15 days (n=31, all from the Chinese mainland) was a root-node sample (Supplementary Table S2). These results are similar to those obtained using the BANAL-52-referenced mutations, indicating our model remains robust with a different out-group reference.

In summary, our results suggest that SARS-CoV-2 may have already been spreading globally and independently long before the first COVID-19 case was detected in Wuhan, China. Thus, global cooperation is essential in searching for the origins and natural hosts of SARS-CoV-2, and in preventing the occurrence of the next pandemic.

DATA AVAILABILITY

All SARS-CoV-2 sequences used in this study were downloaded from the GISAID database (https://www.gisaid.org/). The accession numbers of each genomic sequence of this study are available in the ScienceDB (https://www.scidb.cn/en) repository at https://dx.doi.org/10.57760/sciencedb.01771.

SUPPLEMENTARY DATA

Supplementary data to this article can be found online.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

Z.B.Z.: Conceptualization, investigation, formal analysis, methodology, visualization, writing-original draft, writing-review & editing. C.Y.C.: Investigation, data curation, formal analysis, methodology, software, visualization, writing-original draft, writing-review & editing. All authors read and approved the final version of the manuscript.

ACKNOWLEDGMENTS

We are grateful to Prof. Jian Lu and Dr. Xiao-Lu Tang from Beijing University and Prof. Hua Chen from the Beijing Genome Institute, Chinese Academy of Sciences, for their kind help in data analysis and editing and for valuable comments on this manuscript.

REFERENCES

Abdool Karim SS, de Oliveira T. 2021. New SARS-CoV-2 variants — clinical, public health, and vaccine implications. *New England Journal of Medicine*, **384**(19): 1866−1868.

Bandelt HJ, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, **16**(1): 37−48.

Burki T. 2021. Understanding variants of SARS-CoV-2. *The Lancet*, **397**(10273): 462.

Chan JFW, Kok KH, Zhu Z, et al. 2020. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerging Microbes & Infections*, **9**(1): 221−236.

Cheng CY, Zhang ZB. 2023. SARS-CoV-2 shows a much earlier divergence in the world than in the Chinese mainland. *Science China Life Sciences*,doi: 10.1007/s11427-023-2294-5.

Delaune D, Hul V, Karlsson EA, et al. 2021. A novel SARS-CoV-2 related coronavirus in bats from Cambodia. *Nature Communications*, **12**(1): 6563.

Deng SQ, Peng HJ. 2020. Characteristics of and public health responses to the coronavirus disease 2019 outbreak in China. *Journal of Clinical Medicine*, **9**(2): 575.

Duchene S, Lemey P, Stadler T, et al. 2020. Bayesian evaluation of temporal signal in measurably evolving populations. *Molecular Biology and Evolution*, **37**(11): 3363−3379.

Ellis NA, Ciocci S, German J. 2001. Back mutation can produce phenotype reversion in Bloom syndrome somatic cells. *Human Genetics*, **108**(2): 167−173.

Fan Y, Zhao K, Shi ZL, et al. 2019. Bat coronaviruses in China. *Viruses*, **11**(3): 210.

Forster P, Forster L, Renfrew C, et al. 2020. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **117**(17): 9241−9243.

Giovanetti M, Benvenuto D, Angeletti S, et al. 2020. The first two cases of 2019-nCoV in Italy: where they come from?. *Journal of Medical Virology*, **92**(5): 518−521.

Hill V, Rambaut A. 2020[2021-02-17]. Phylodynamic analysis of SARS-CoV-2 | Update 2020-03-06.https://virological.org/t/phylodynamic-analysis-of-sars-cov-2-update-2020-03-06/420.

Holmes EC, Nee S, Rambaut A, et al. 1995. Revealing the history of infectious disease epidemics through phylogenetic trees. *Philosophical Transactions of the Royal Society B:Biological Sciences*, **349**(1327): 33−40.

Kong S, Sánchez-Pacheco SJ, Murphy RW. 2016. On the use of median-joining networks in evolutionary biology. *Cladistics*, **32**(6): 691−699.

Lam TTY, Jia N, Zhang YW, et al. 2020. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature*, **583**(7815): 282−285.

Lanciotti RS, Ebel GD, Deubel V, et al. 2002. Complete genome sequences and phylogenetic analysis of west Nile virus strains isolated from the United States, Europe, and the Middle East. *Virology*, **298**(1): 96−105.

Lau SKP, Luk HKH, Wong ACP, et al. 2020. Possible bat origin of severe acute respiratory syndrome coronavirus 2. *Emerging Infectious Diseases*, **26**(7): 1542−1547.

Lauring AS, Hodcroft EB. 2021. Genetic variants of SARS-CoV-2—What do they mean?. *JAMA*, **325**(6): 529−531.

Li JG, Li Z, Cui XG, et al. 2020a. Bayesian phylodynamic inference on the temporal evolution and global transmission of SARS-CoV-2. *Journal of Infection*, **81**(2): 318−356.

Li WD, Shi ZL, Yu M, et al. 2005. Bats are natural reservoirs of SARS-like coronaviruses. *Science*, **310**(5748): 676−679.

Li XG, Wang W, Zhao XF, et al. 2020b. Transmission dynamics and evolutionary history of 2019-nCoV. *Journal of Medical Virology*, **92**(5): 501−511.

Li XG, Zai JJ, Wang XM, et al. 2020c. Potential of large "first generation" human-to-human transmission of 2019-nCoV. *Journal of Medical Virology*, **92**(4): 448−454.

Li XG, Zai JJ, Zhao Q, et al. 2020d. Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. *Journal of Medical Virology*, **92**(6): 602−611.

Liu Q, Zhao SL, Shi CM, et al. 2020. Population genetics of SARS-CoV-2: disentangling effects of sampling bias and infection clusters. *Genomics, Proteomics & Bioinformatics*, **18**(6): 640−647.

Lu RJ, Zhao X, Li J, et al. 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet*, **395**(10224): 565−574.

Lundstrom K, Seyran M, Pizzol D, et al. 2020. The importance of research on the origin of SARS-CoV-2. *Viruses*, **12**(11): 1203.

Mallapaty S. 2020. Coronaviruses closely related to the pandemic virus discovered in Japan and Cambodia. *Nature*, **588**(7836): 15−16.

Morel B, Barbera P, Czech L, et al. 2021. Phylogenetic analysis of SARS-

CoV-2 data is difficult. *Molecular Biology and Evolution*, **38**(5): 1777−1791.
Nabil B, Sabrina B, Abdelhakim B. 2021. Transmission route and introduction of pandemic SARS-CoV-2 between China, Italy, and Spain. *Journal of Medical Virology*, **93**(1): 564−568.
Nie Q, Li XG, Chen W, et al. 2020. Phylogenetic and phylodynamic analyses of SARS-CoV-2. *Virus Research*, **287**: 198098.
Paraskevis D, Kostaki EG, Magiorkinis G, et al. 2020. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. *Infection, Genetics and Evolution*, **79**: 104212.
Patwardhan A, Ray S, Roy A. 2014. Molecular markers in phylogenetic studies-a review. *Journal of Phylogenetics & Evolutionary Biology*, **2**(2): 1000131.
Peng MS, Li JB, Cai ZF, et al. 2021. The high diversity of SARS-CoV-2-related coronaviruses in pangolins alerts potential ecological risks. *Zoological Research*, **42**(6): 834−844.
Pipes L, Wang HR, Huelsenbeck JP, et al. 2021. Assessing uncertainty in the rooting of the SARS-CoV-2 phylogeny. *Molecular Biology and Evolution*, **38**(4): 1537−1543.
Poon AFY, Walker LW, Murray H, et al. 2013. Mapping the shapes of phylogenetic trees from human and zoonotic RNA viruses. *PLoS One*, **8**(11): e78122.
Ruan YS, Wen HJ, Hou M, et al. 2022. The twin-beginnings of COVID-19 in Asia and Europe - one prevails quickly. *National Science Review*, **9**(4): nwab223.
Sekizuka T, Itokawa K, Kageyama T, et al. 2020. Haplotype networks of SARS-CoV-2 infections in the *Diamond Princess* cruise ship outbreak. *Proceedings of the National Academy of Sciences of the United States of America*, **117**(33): 20198−20201.
Shan KJ, Wei CS, Wang Y, et al. 2021. Host-specific asymmetric accumulation of mutation types reveals that the origin of SARS-CoV-2 is consistent with a natural process. *The Innovation*, **2**(4): 100159.
Song SH, Ma LN, Zou D, et al. 2020. The global landscape of SARS-CoV-2 genomes, variants, and haplotypes in 2019nCoVR. *Genomics, Proteomics & Bioinformatics*, **18**(6): 749−759.
Tang XL, Wu CC, Li X, et al. 2020. On the origin and continuing evolution of SARS-CoV-2. *National Science Review*, **7**(6): 1012−1023.
Tang XL, Ying RC, Yao XM, et al. 2021. Evolutionary analysis and lineage designation of SARS-CoV-2 genomes. *Science Bulletin*, **66**(22): 2297−2311.
Temmam S, Vongphayloth K, Baquero E, et al. 2022. Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. *Nature*, **604**(7905): 330−336.
Tong YG, Liu WL, Liu PP, et al. 2021. The origins of viruses: discovery takes time, international resources, and cooperation. *The Lancet*, **398**(10309): 1401−1402.
Turakhia Y, de Maio N, Thornlow B, et al. 2020. Stability of SARS-CoV-2 phylogenies. *PLoS Genetics*, **16**(11): e1009175.
van Dorp L, Acman M, Richard D, et al. 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infection, Genetics and Evolution*, **83**: 104351.
Wacharapluesadee S, Tan CW, Maneeorn P, et al. 2021. Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in southeast Asia. *Nature Communications*, **12**(1): 972.
Wang HJ, Zhao W. 2021. WHO-convened global study of origins of SARS-CoV-2: China part (text extract). *Infectious Diseases & Immunity*, **1**(3): 125−132.
WHO Team. 2023. Weekly epidemiological update on COVID-19 - 16 March 2023. https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---16-march-2023.
Wong G, Bi YH, Wang QH, et al. 2020. Zoonotic origins of human coronavirus 2019 (HCoV-19 / SARS-CoV-2): why is this work important?. *Zoological Research*, **41**(3): 213−219.
Wu CI, Wen HJ, Lu J, et al. 2021. On the origin of SARS-CoV-2—the blind watchmaker argument. *Science China Life Sciences*, **64**(9): 1560−1563.
Wu F, Zhao S, Yu B, et al. 2020. A new coronavirus associated with human respiratory disease in China. *Nature*, **579**(7798): 265−269.
Xiao KP, Zhai JQ, Feng YY, et al. 2020. Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature*, **583**(7815): 286−289.
Yu D, Zhu J, Yang J, et al. 2022. Global cold-chain related SARS-CoV-2 transmission identified by pandemic-scale phylogenomics. *Zoological Research*, **43**(5): 871−874.
Yu WB, Tang GD, Zhang L, et al. 2020. Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2 / HCoV-19) using whole genomic data. *Zoological Research*, **41**(3): 247−257.
Zhang YZ, Holmes EC. 2020. A genomic perspective on the origin and emergence of SARS-CoV-2. *Cell*, **181**(2): 223−227.
Zhou H, Chen X, Hu T, et al. 2020a. A novel bat coronavirus closely related to SARS-CoV-2 contains natural insertions at the S1/S2 cleavage site of the spike protein. *Current Biology*, **30**(11): 2196−2203.e3.
Zhou H, Ji JK, Chen X, et al. 2021. Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. *Cell*, **184**(17): 4380−4391.e14.
Zhou P, Yang XL, Wang XG, et al. 2020b. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, **579**(7798): 270−273.