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

Chaoyuan Cheng
Institute of Zoology, Chinese Academy of Sciences
Zhibin Zhang (zhangzb@ioz.ac.cn)
Institute of Zoology, Chinese Academy of Sciences

Research Article

Keywords: SARS-CoV-2, transmission chain or network, ancestor-offspring relationship, de novo mutation, back mutation, secondary mutation

Posted Date: May 31st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1644027/v1

Abstract

Since its first report in Wuhan city, China in late December of 2019, SARS-CoV-2 has caused huge damage to human society, but its origins and early transmission patterns remain unclear. We reconstructed the transmission network of SARS-CoV-2 within the first three and six months since its first report based on its ancestor-offspring relationship using BANAL-52-referenced mutations. We want to explore the position of the early detected samples in the evolutionary tree of SARS-CoV-2. 19,187 samples (first 3 months) and 84,243 samples (first 6 months) were used for the analysis, respectively. Using data of the first 3 months, BANAL-52-referenced mutations were found in 7062 out of 29,410 nucleotide sites in the SARS-CoV-2 genome. 6,799 transmission chains and 1766 transmission networks were reconstructed with chain lengths ranging from 1–9 nodes. The root node samples of the 1766 transmission networks are from 58 countries or regions, and they have no common ancestor, indicating plenty of independent or parallel transmissions of SARS-CoV-2 have occurred when they were firstly detected. No root node sample was found in the samples (n = 31, all from the mainland of China) that were collected in the first 15 days since December 24, 2019. Results using data of the first 6 months are similar to those of the first 3 months, but more independent transmission chains were revealed in more countries. The reconstruction method was verified by using a simulation approach. Our results suggest that SARS-CoV-2 has long been spread independently in many parts of the world before it was first detected in Wuhan, China. It is essential to have a global survey on human or animal samples to look for the origins of SARS-CoV-2 and its natural hosts.

Introduction

A novel beta-coronavirus, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was detected by late December of 2019 in Wuhan, China, and it was identified to be responsible for the COVID-19 pandemic [1]. Nearly two years after the first reported case, the COVID-19 pandemic has been surging in many parts of the world. By September 20, 2021, it was estimated that the COVID-19 pandemic has caused more than 227 million cases of human infections, and over four million people died (WHO, 2021). However, the origins of SARS-CoV-2 remain unknown. There is an urgent need to reveal the early transmission pattern of SARS-CoV-2 to take more effective measures of controlling this pandemic and preventing the next pandemic.

SARS-CoV-2 is a novel beta-CoV that is distinct from SARS-CoV and MERS-CoV [2–4]. SARS-CoV-2 shares ~ 79% genome sequence identity with SARS-CoV and only ~ 50% with MERS-CoV [5]. Two coronavirus strains isolated from a bat species (*Rhinolophus affinis*) in Yunnan, China and a bat species (*Rhinolophus malayanus*) in the northern part of Laos, has a ~ 96.1% (RaTG13) and 96.8% (BANAL-52) similarity in whole genome with SARS-CoV-2, respectively [4, 6]. The other SARS-CoV-2 relatives are also found in bats from Cambodia and Japan [7]. The discovery of diverse SARS-CoV-2 relatives suggests that bats are potential reservoirs of SARS-CoV-2 [8, 9]. However, the difference between SARS-CoV-2 and these bat coronaviruses is still large. For example, RaTG13 has a divergence time of about 50 years with SARS-

CoV-2 [10]. Thus, they are more likely evolutionary precursors, not as the direct progenitor of SARS-CoV-2 [11].

Previous studies have shown that the mutation rate of SARS-CoV-2 is 6×10^{-4} to 1×10^{-3} bp/site/year [3, 12–17]. The SARS-CoV-2 has produced many variants, and the transmission capacity of some variants is much higher than the original ones[18–20]. Based on molecular clock theory, the time of the most recent common ancestor (TMRCA) of SARS-CoV-2 is estimated to be November or early December 2019 [3, 17, 21–23], or October to early December 2019 [12, 16], suggesting SARS-CoV-2 is likely originated much earlier than the time when it was first detected in Wuhan, China.

There are many studies using phylogenetic trees, or haplotype networks to designate lineage, or to reconstruct the evolutionary patterns of SARS-CoV-2 [24–29]. Because of deviations from a molecular clock and the high similarity between SARS-CoV-2 samples, the phylogenetic tree alone is not able to reveal the origin of SARS-CoV-2 [30, 31]. 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 its ancestor-offspring relationship based on BANAL-52-referenced mutations in samples that were collected from December 24, 2019 to March 22, 2020 around the world, aiming to explore the early transmission patterns of SARS-CoV-2. We want to test the following three hypotheses: (1) If there is a common ancestral sample existed in samples collected during the first three months, the reconstructed lineages should be located in the bottom of the evolution tree, and the transmission network should look like a single full tree with one common ancestor or one root node (Original Lineage Hypothesis, OLH Hypothesis, Fig. 1A), (2) If there is no common ancestral sample, but there are a few samples which are very close to the ancestral sample genetically, the lineages should be located in the middle of the evolution tree, and the transmission network should look like a few large tree-branches with a few root nodes (Intermediate Lineage Hypothesis, ILH Hypothesis, Fig. 1B), (3) If all samples are very far away from the common ancestral sample genetically, the lineages should be located in the tip position of the evolution tree, and the transmission network should have many short and small “tree-branches” or many root nodes (Tip Lineage Hypothesis, TLH Hypothesis, Fig. 1C). OLH Hypothesis indicates that ancestor of SARS-CoV-2 is detected. ILH Hypothesis indicates that SARS-CoV-2 samples very close to ancestor are detected. TLH Hypothesis indicates that SARS-CoV-2 samples very far away from ancestor are detected.

Materials And Methods

Genome sequence processing

The genome sequences of SARS-CoV-2 were downloaded from the Global Initiative of Sharing All Influenza Data (GISAID, https://www.gisaid.org/) on January 25, 2021. The dataset contains 279,411 samples which were sampled from December 24, 2019 to January 21, 2021. Firstly, we used the data collected in the first three months (i.e., December 24, 2019 to March 22, 2020) for reconstructing the early

transmission network of SARS-CoV-2 (including 19,187 samples, covering 95 countries or regions). Secondly, to repeat and validate our results, we used data covering the first 6 months (including 84,243 samples, covering 112 countries or regions). These samples represent the early detected samples in the world covering many countries. Because the results are similar, thus, we mainly reported the results of the first 3 months, but discussed the results of first 6 months. 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 these genome sequences to the reference sequence (Wuhan-Hu-1, GISAID: EPI_ISL_402125) using Muscle-V5. To minimize the potential impacts of sequencing errors, nucleotides at the 5' UTR (sites 1—265) and 3' UTR (sites 29675—29903) were excluded.

Reconstruction of the transmission network

We clustered all samples based on sequence differences. Samples with the same genome sequence were assigned into a node of transmission chains. Each node in the transmission chains represents a unique sequence with distinct mutation sites, which is often composed of multiple samples from the same or different places. We reconstructed the transmission network of SARS-CoV-2 based on differences of BANAL-52-referenced mutations (i.e., BANAL-52-ref mutations) in all sites of each node. The transmission network is composed of transmission chains. The core process of reconstructing each transmission chain is to find the closest ancestor node of each node according to the following mutation model (Fig. 2A).

We defined mutations that occurred after the emergence of the most recent common ancestor (MRCA) of SARS-CoV-2 as the *de novo* mutations. Since it is impossible to directly determine the *de novo* mutations because the common ancestor is unknown, we inferred the mutations by using BANAL-52 as a reference (for details, see *Reference Selection* Section). We assumed that sequence S0 is the MRCA of the other sequences (i.e., S1, S2, S3.1, S3.2, S4.1, S4.2) in Fig. 2. We used the following steps to reconstruct a transmission chain or network of SARS-CoV-2 based on its ancestry relationship (ancestor-offspring relationship) of *de novo* mutations (Fig. 2):

(1) For each node, we identified its ancestor nodes. We performed a pairwise comparison of the sequences. The *de novo* mutations contained in a node X is set $M_X = \{m_1, m_2, ..., m_n\}$, then, for node U and V, 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, the mutations of the offspring node must include all the mutations of its ancestor node. For example, in Fig. 2A, S0, S1, S2, S3.1 are the ancestor nodes of S4.1, and S0, S1, S2, S3.2 are the ancestor nodes of S4.2.

(2) For each node (i.e., a unique sequence), we identified its closest ancestral node from all ancestor nodes. For all ancestor nodes identified in step (1), we selected the node which has the closest similarity of mutations to the focal node as its closest ancestral node. For example, in Fig. 2A, S2 is the closest ancestral node of S3.1 and S3.2; S3.1 is the closest ancestral node of S4.1; 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 (Fig. 2B) or network (Fig. 2C). A transmission network indicates several transmission chains which share a common root node.

By following the above procedure (1), (2), and (3), ancestral nodes of all nodes were determined, and the transmission chains were finally reconstructed (Fig. 2). Because some chains would share a common node (Fig. 2B), thus, we reconstructed the transmission network by merging chains sharing the common node (Fig. 2C).

The sample clustering (i.e., identification of node) and reconstruction process of transmission chains and the networks were implemented by custom scripts in Python-3.7.

Reference selection

To determine the *de novo* mutations of a node, we need to select a reference sequence. Because the most recent common ancestor of SARS-CoV-2 (i.e., MCRA in Fig. 2A) is still unknown, we have to choose a detected sequence that is close to SARS-CoV-2 as the reference sequence.

As shown in Fig. 3, if we use a sequence of SARS-CoV-2 which is an offspring of SARS-CoV-2, the mutations inferred from the reference may be incorrect. For example, according to Fig. 3, using the *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. But if we select an offspring as the reference (e.g., S3), the identification of 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.

Alternatively, we selected an earlier but the closest relative sequence (i.e., BANAL-52) to SARS-CoV-2 as the reference to infer the mutations in SARS-CoV-2 (i.e., BANAL-52-ref mutations), and assessed its potential in reconstruction of transmission chains. BANAL-52 was sampled in 2020, it is a closest relative to SARS-CoV-2 with an about 3.2% difference with SARS-CoV-2 in the whole genome [6]. As shown in Fig. 4, we assume that, as 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), MRCA of SARS-CoV-2 (i.e. MRCA-S) has two mutations (orange C,G) at site 8 & 9; as compared to MRCA-S, SARS-CoV-2 has 1–6 *de novo* mutations (boxed red T) at site 1–6, and two inherited ancestor mutations (orange C,G) from MRCA-S and ancestor of MRCA-S at site 8 & 9. If we use the 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 identified correctly in L1 (i.e., boxed cyan T at site from 1–6) which are same to the red and boxed T in L0. The ancestor mutations of SARS-CoV-2 in L1 (cyan C, G) would be correctly identified. But mutation of BANAL-52 (orange T) would be identified as an incorrect mutation in L1 (cyan A) because the mutation

of the BANAL-52 reference results in difference of base at site 10 between BANAL-52 and SARS-CoV-2. Because both inherited ancestor mutations of SARS-CoV-2 (site 8,9) and incorrectly inferred mutation using BANAL-52 (site 10, named as incorrect mutation) would be carried by all samples of SARS-CoV-2, they are not considered in the reconstruction of the transmission chain 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 same to S0 (Fig. 4). This lays the solid foundation of reconstructing the transmission chains or networks using BANAL-52 as the reference. In Fig. 4, we introduced mutations of C, G to illustrate ancestor mutation of SARS-CoV-2.

Model errors

Because BANAL-52-ref mutations are used to reconstruct the transmission chains or networks, secondary mutation at the BANAL-52-ref mutation sites in L1 (Fig. 4) would cause errors in reconstruction of transmission chains and networks. If the secondary mutation changes the base back into its original base, it is called a back mutation. In our model, there are three kinds of BANAL-52-ref mutations in L1: *de novo* mutations (cyan T), ancestor mutations (cyan C, G) and incorrect mutations (cyan A) which were used for reconstructing the transmission chains or networks (Fig. 4). If there is no secondary mutation occurring on these mutation sites of SARS-CoV-2 during the study period, the reconstructed transmission chain (L1) is same to the true one (L0) (Fig. 4); otherwise, it will cause errors in reconstructing the transmission chains or networks (see Fig. 5).

As shown in Fig. 5, a secondary mutation on the BANAL-52-ref mutation sites during the study period of three months would cause biases in the reconstruction of transmission chains. If no secondary mutation occurs, the original transmission chain should be: L0 = S1→S2→S3→S4→S5→S6. If the secondary mutation (T→A, here, it is 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 Fig. 5B2) will become an isolated chain (L1.2n = S4b, Fig. 5B2), while the relationship of other samples on the chain remains unchanged except for the absence of S4, i.e., L1.1n = S1→S2→S3→S5→S6 (Fig. 5B2). However, if S4 has extra copies, it will not affect the original chain: L1.1c = S1→S2→S3→S4→S5→S6, while the secondary mutation would produce an isolated chain: L1.2c = S4b (Fig. 5B1). If the 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), this will cause no change of original chain if S4 has extra copies (i.e., L2c = S1→S2→S3→S4→S5→S6) or it would produce a shorter chain with the absence of S4 (i.e., L2n = S1→S2→S3→S5→S6) (Fig. 5C). If the secondary mutation occurs in the site corresponding to the BANAL-52 mutation site (cyan A →purple T, here, it is not a back mutation because 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 relationship of the other samples on the chain remains unchanged: L3.1c = S1→S2→S3→S4→S5→S6 (Fig. 5D1). However, if the sequence (S4) has no extra copies, it would break the original chain into two chains with the same probability: L3.1n = S1→S2→S3→S5→S6, L3.2n = S4b→S5→S6 (Fig. 5D2). Secondary mutation in the ancestor mutation sites (i.e., C, G in Fig. 4) would have similar results to cyan A (for simplification, we did not show these

sites in Fig. 5). Because the proportion of sequence with two or more secondary mutations was extremely low in this study, thus, we only presented illustrations of one secondary mutation in Fig. 5.

Considering the model error which could alter the ancestral-offspring of about 20% sequences of the first 3 months (higher for sequences of first 6 months) (For details, see below), we emphasized on analysis and discussion of general transmission patterns (such as number of chains, networks ect.) of SARS-CoV-2, not on the specific sequences.

Error probability caused by secondary mutations

If there is no secondary mutation during the study period of three months, the transmission chain (L1) is same to the true one (L0) (Fig. 4). However, if secondary mutation occurs on these sites of BANAL-52-ref mutations, error of reconstructed transmission chain would arise. The BANAL-52-ref mutations are composed of *de novo* mutations (cyan T), ancestor mutations (cyan C, G) and incorrect mutations (cyan A) in L1 (Fig. 4).

In theory, the probability of secondary mutation on *de novo* mutations (p_d) is determined by the mutation probability of the *de novo* mutations of SARS-CoV-2 within three months. The mutation rate was assumed as 6×10^{-4} substitutions per site annually[12], and the mutation probability of SARS-CoV-2 within 3 months was calculated as: $P_T = 6 \times 10^{-4} / 4 = 1.5 \times 10^{-4}$. The proportion of *de novo* mutation within three months can be assumed to be 1.5×10^{-4}. Therefore, the error probability of secondary mutation on *de novo* sites was calculated as: $p_d = (P_T)^2 = (1.5 \times 10^{-4})^2 = 2.25 \times 10^{-8}$. The probability of samples with such secondary mutations (i.e., number of secondary mutation ≥ 1) was estimated as $1-(1-p)^n$ (p is the secondary mutation rate, n is the number of bases of SARS-CoV-2). Let $p_d = 2.25 \times 10^{-8}$, n = 29,410, then, the probability of samples with such secondary mutations was estimated to be 6.6×10^{-4}. Thus, we predict about 0.066% of samples with such secondary mutations on *de novo* mutation sites would be reconstructed into short chains. Because each node sequence has 2.4 copies (= 19187/7918) in our study, the probability of breaking the original chains was calculated as: $(6.6 \times 10^{-4})^{2.4} = 2.3 \times 10^{-8}$, which is very small.

BANAL-52 has about 3.2% difference with MRCA of SARS-CoV-2 in the whole genome (P_R), which is likely composed of ancestor mutation sites (cyan C, G) and incorrect mutation site (cyan A). Secondary mutation on these sites would cause model errors in reconstructing the transmission chains or networks of SARS-CoV-2. The mutation rate during three months of study period is assumed as: $P_T = 1.5 \times 10^{-4}$ substitutions per site within three months. Thus, the error probability caused by secondary mutations on ancestor and incorrect mutation sites was calculated as: $p_a = P_R \times P_T = 3.2 \times 10^{-2} \times 1.5 \times 10^{-4} = 4.8 \times 10^{-6}$. The probability of a sample that has secondary mutations on ancestor or BANAL-52 mutation sites was estimated as $1-(1-p)^n$ (p is the total error probability caused by secondary mutations on mutation sites of ancestor of SARS-CoV-2 and BANAL-52, n is the number of bases of SARS-CoV-2 genome). Let $p = 4.8 \times 10^{-6}$, n = 29,410, then, the probability of a sample that has secondary mutations on ancestor and

incorrect mutation sites (cyan C, G, A in Fig. 4) was estimated to be 13.2%. Because each node sequence has 2.4 copies in our study, the probability of breaking the original chains was calculated as: $(13.2\%)^{2.4}$ = 0.78%. Thus, using BANAL-52 as reference would have little influence on the original transmission chains or network, although it may produce 13.2% short transmission chains.

Error probability caused by sequence gap or uncertainty

Similar to secondary mutations, degenerate or gaps (missing bases in the genome sequence) due to sequencing error or uncertainty may cause biased estimation of reconstruction of the evolutionary tree. These uncertain bases are often treated as none mutations in evolutionary studies (same in our study). In our study, there were 1.9 degenerate (or missing) bases per sequence on average. Thus, the proportion of degenerate or missing bases was calculated as: $P_D = 2.9/29410 = 9.8 \times 10^{-5}$. Therefore, the model error probability caused by degenerate or missing bases on *de novo* mutations was calculated as: $p_s = P_D \times P_T = 9.8 \times 10^{-5} \times 1.5 \times 10^{-4} = 1.47 \times 10^{-8}$. Similarly, the model error probability on ancestor and incorrect mutations was calculated as: $p_s = P_D \times P_R = 9.8 \times 10^{-5} \times 3.2 \times 10^{-2} = 3.1 \times 10^{-6}$. The probability of samples with degenerate or missing bases (i.e. number of degenerate or missing bases ≥ 1) was estimated as $1-(1-p)^n$ (p is the degenerate mutation rate, n is the number of bases of SARS-CoV-2). Let $p = 1.47 \times 10^{-8}$, or, 3.1×10^{-6}, n = 29,410, then $p = 4.3 \times 10^{-4}$, or, 8.7%. Thus, we predict a total of 8.7% samples would be reconstructed into short transmission chains caused by degenerate or missing bases. Because each node sequence has 2.4 copies in our study, the probability of breaking the original chains by degenerate base was calculated as: $(8.7\%)^{2.4}$ = 0.29%. Therefore, degenerate mutations or missing bases would have little influence on the reconstruction of original transmission chains or network, but they would produce 7.13% short transmission chains.

Simulation analysis

To validate our method of reconstructing the transmission network of SARS-CoV-2 based on the paternity relationship as described above, we simulated the occurrence of mutations based on the way that the virus sequence replicates in nature. We chose a sequence with a length of 1000 bp as the starting sequence to simulate the proliferation process of this sequence. Each sequence produced three progeny sequences in one replication cycle (one generation). We set the mortality rate of sequences as 20%, and the mutation rate as 0.1% for each base within a replication cycle. We chose a shorter sequence and a higher mutation rate for saving computation time.

We simulated the evolution process of a sequence for ten 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 of 2019) for subsequent analysis (Fig. S2). We chose the last sequence in the 9th generation as the reference sequence (similar to BANAL-52) (Fig. S2). We selected three-group data from the detected samples: (1) the detected samples from 0th to 6th generation (the 0th generation represents the starting sequence of the simulated data) which includes the common ancestor, (2) the detected samples from the 2nd to 6th generation with missing data (similar to the undetected

data) from 0th to 2nd generation, and (3) the detected samples from the 4th to 6th generation with missing data from 0th to 3rd generations. By using these three-group data, we reconstructed the transmission chains and networks based on the paternity relationship, tested the three hypotheses (Fig. 1), and validated the method we used in this study (Table 2).

The simulation process was implemented by custom scripts in Python-3.7.

Results

Frequency of BANAL-52-ref mutations

Among 29,410 nucleotide sites in the SARS-CoV-2 genome we used, 7062 sites (Fig. 6) have different bases from BANAL-52. These different bases are defined as BANAL-52-ref mutations and were used to reconstruct the transmission chains and networks. We identified 7918 unique sequences (equivalent to haplotypes) from 19,187 samples according to the composition of the BANAL-52-ref *de novo* mutations and each unique sequence represents a node in the transmission chain or network. The frequency of BANAL-52-ref mutations shows a U-shaped relationship with the number of unique sequences. A majority of mutations have less than 20 unique sequences (Fig. 6A), indicating these mutations are more likely *de novo* mutations of SARS-CoV-2. Some mutations occur in the unique sequences with their sample size larger than 7910 (Fig. 6C), indicating these mutations are likely ancestor mutations and incorrect mutations inferred using BANAL-52 which are carried by nearly all samples. Mutation sites in the middle are more likely early *de novo* mutations or secondary mutations on ancestor or incorrect mutations of SARS-CoV-2 (Fig. 6B).

Frequency of transmission chains with different length

The frequency of transmission chains with different lengths (i.e., number of nodes) is shown in Fig. 7 and Table 1. Among 7918 nodes of the transmission networks, there are 1766 root nodes (having 2847 samples) from 58 countries and these root nodes have no common ancestor of SARS-CoV-2 (Table 1). We reconstructed 6799 transmission chains with the number of nodes ranging from 1 to 9, and these chains form 1766 networks.

The average length of transmission chains was estimated to be 3.7 ± 2.1 (Fig. 7A). If only the longest chain with the common root sample in a transmission network is counted, the number of transmission chains decreases rapidly with the increase of the chain length (Fig. 7B).

Table 1
Statistics of nodes of the reconstructed transmission chains with different lengths within the first three months.

Chain length	No. chains	No. root nodes	No. root samples	No. root country/regions	First sampling time	Last sampling time
1	1515	1515	2207	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	1002	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	6799	1766	2847	58	16	90

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

There are several large networks in terms of number of nodes for both using data of the first 3 months and the first 6 months. For networks reconstructed using data of the first 3 months, there are 5 large networks with length ≥ 7, with nodes and samples accounting for 55% and 66% of the total sample size, respectively.

Frequency of sampling time of root-node samples

The sampling time of root-node samples ranges from day 16 to 90 (Table 1, Fig. 8). None of the samples collected from the first 15 days (i.e., from 2019/12/24 to 2020/1/8) are root-node samples of the transmission chains (Table 1, Fig. 8). The samples of root nodes with chain length ranging from 1–4 were mainly collected after and before day 60, while chains with length ≥ 5 collected during Day 30 to 50, respectively (Fig. 8A). In Fig. 8A, some chains may share the same root node, which could result in duplicate counting of the root nodes. If each root node and the longest chain are counted in the transmission network (corresponding to Fig. 7B), 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 Day 20 to 70 (Fig. 8B). There are on transmission networks with the longest chain length of 8.

Illustration of transmission networks

For simplification, we only presented the reconstructed transmission networks with chain length of four nodes (Fig. S4A), five nodes (Fig. S4B), six nodes (Fig. S4C), seven nodes (Fig. S4D), eight nodes (Fig. S4E), and nine nodes (Fig. S4F). These figures show that the detected transmission network of SARS-CoV-2 is composed of many short transmission chains, like the short fragmented “tree branches”. Many spiking nodes were detected which obviously caused super transmissions.

Simulation results

Simulation results indicate that the reconstructed transmission network using the detected data covering 0th to 6th generation have two independent networks with an average length of 4.8; the transmission network covering the 2nd to 6th generation have 13 independent networks with an average length of 3.8; the network covering the 4th to 6th generation have 52 independent networks with an average length of 2.6 (Table 2, fig. S3). These results clearly demonstrate that shorter transmission chains and more root nodes (number of root nodes is equivalent to number of transmission networks) would be reconstructed using samples collected in later stage of the evolutionary tree.

Table 2
Parameters of transmission chains or networks reconstructed using simulated data.

Data range	Number of root nodes	Number of chains	Average chain length
0–6	2	267	4.8
2–6	13	264	3.8
4–6	52	251	2.6

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

Discussion

Currently, the origins and early transmission of SARS-CoV-2 remain unclear. By using 19,187 samples of SARS-CoV-2 collected during the first three months since December 24, 2019, we reconstructed the transmission chains and networks based on ancestor-offspring relationship of all samples by using BANAL-52-ref mutations. We found there are over thousands of independent transmission chains of SARS-CoV-2 without common ancestor sample. No root-node sample of the first 15 days was found from the mainland of China. Our results indicate that all detected samples during the study period of three months are not common ancestors or close to the common ancestor of SARS-CoV-2, supporting the Tip

Lineage Hypothesis (TLH). Simulation analysis indicates that our reconstruction method is robust, and the results are consistent with predictions of our hypotheses. Our study suggests that SARS-CoV-2 would have long been spread in many parts of the world before its first report in Wuhan, China. It is necessary to have a global survey for looking for the origins and natural reservoir of SARS-CoV-2 in the world.

Revealing early transmission patterns of SARS-CoV-2 is important for preventing future spillover of the virus, but its origins and natural hosts are still unknown [32, 33]. Because bats are natural reservoirs of many kinds of coronavirus, and the coronavirus closest to the SARS-CoV-2 genome was found to come from bats (i.e., BANAL-52), it is suggested that bat may be the natural host of the SARS-CoV-2 [8, 11, 34, 35]. Recent studies found several bat species from East Asia and Southeast Asia carry SARS-CoV-2-like viruses [36–39]. Some SARS-CoV-2-like viruses collected from Laos had a high similarity to SARS-CoV-2, and the RBDs of these viruses bind to human ACE2 protein as efficiently as the SARS-CoV-2 [39]. 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, which suggest that pangolin may be an intermediate host of SARS-CoV-2 [40, 41]. A joint-report by WHO and China research teams concluded that SARS-CoV-2 very likely originated from nature, extremely unlikely from the laboratory [42, 43]. But some people claimed the SARS-CoV-2 may come from a laboratory in Wuhan, China due to an accident leak [44]. Ruan et al explored the SARS-CoV-2 evolution in the first wave of the pandemic (early 2020) by using SARS-CoV-2 genome data, and they suggested Europe strains may spread in parallel with Asian strains; the Europe strains had supplanted the Asian strains globally by May of 2020 [45]. Our results revealed 1766 independent transmission networks which are widely distributed in 58 countries of the world, and they do not share a common ancestor, indicating independent parallel spread of SARS-CoV-2 might have occurred in many parts of world before it was detected in Wuhan, China. Notably, there is no root node from samples of the first 15 days (n = 31, all from mainland of China, for details, see below), indicating the early detected samples in mainland of China are not the ancestors of SARS-CoV-2 samples of the world collected during the first 3 months. These results suggest that SARS-CoV-2 would have long been circulating in many places of the world before it was detected in Wuhan, China.

There is a large proportion of short transmission chains, probably because their root-node samples were collected in a later stage (Table 1, Fig. 8). Indeed, samples of root nodes with chain length ranging from 1 to 4 were mainly collected after day 60 (Fig. 8A). Because SARS-CoV-2 samples of most countries or regions outside the mainland of China were mainly reported in the third month, the missing data before the third month could attribute to the large number of short chains. As shown in Methods, secondary mutation and sequencing gaps or uncertainty would cause 13.2% and 8.7% short transmission chains, thus, the observed short chains should be incorrect transmission chains. We should be cautious in explaining the biological meaning of the short transmission chains. Long transmission chains were likely caused by the early appearance of their root samples which survived longer (Fig. 8). Samples of root nodes with chain length larger than 5 were mainly collected before day 60 (Fig. 8A). Our approach only causes a small influence on reconstruction of the original transmission chains considering the extra copies of SARS-CoV-2 samples. The error probability caused by both secondary mutations and sequencing error or uncertainty was estimated to be 0.78% and 0.29%, respectively.

Secondary mutation means that the bases at the BANAL-52-ref sites have been mutated twice or more during the study period. It can be called the back mutation if the base is mutated back to its original base [46]. Back mutations may obscure the estimation of the true evolutionary distances between the sequences when building phylogenetic trees, which is known as homoplasy [47]. Similar, the other kinds of secondary mutations would cause similar errors. According to our estimation, within three months, secondary mutations on sites of d*e novo* mutations and ancestor or incorrect mutations would produce 13.2% and 8.7% short transmission chains, which need to pay high attention when building evolutionary tree of SARS-CoV-2. Fortunately, extra copies of SARS-CoV-2 samples would significantly reduce the model error in revealing the original transmission chains or networks in our study. There are only about 0.78% and 0.29% original chains which would be broken into short chains.

Phylogenetic tree methods are often used to analyze the evolutionary relationship between viruses [48–50]. Some studies used unrooted trees to explore the phylogeny of SARS-CoV-2 [51, 52]. Due to lack of root information, these studies usually refer to the sampling time of the sample 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, it is not adequate for tracing the origin of SARS-CoV-2. Most phylogenetic trees are established based on the genetic similarity between sequences and rely on the assumption of a constant mutation rate. If the mutation rate has a very large variation, genetic distance does not necessarily mean the divergent time or ancestral-offspring relationship, which would impact the results of the phylogenetic tree. Thus, the phylogenetic tree alone is not able to reveal the origins of SARS-CoV-2 [30]. Different from phylogenetic tree method, our approach neither relies on the assumptions of strict molecular clock theory, nor on the sampling time. Primitive strains could be detected in later stage if they have a low mutation rate.

The haplotype network method has been widely used to study the relationship between different clades or lineages of the SRAS-CoV-2 [24, 53, 54]. The nodes in our transmission network represent unique sequences, which are the same as the haplotypes. But the connecting principle of our nodes and the meaning of the transmission chains are completely different from the haplotype network method. There are three major differences between our transmission network and the haplotype network methods. First, our transmission network is constructed based on the ancestor-offspring relationship between haplotypes in the transmission chain. While, the commonly used haplotype network method connects haplotypes by using distance matrix-based algorithms such as median-joining networks [55], which mainly rely on the overall similarity between haplotypes [56]. Second, the haplotype network method does not give the direction of transmission, and some studies may designate the node of the earliest sample as the origin node. Our transmission network method gives the direction of transmission from an ancestor to an offspring sequence and can be used to trace the source and transmission of the viruses. Third, all haplotypes are connected in a single network in the haplotype network method, even if these haplotypes do not have an ancestor-offspring relationship. In our transmission network method, only sequences with the ancestor-offspring relationships were connected. If no common ancestor was detected, there would be several or many independent transmission chains. As shown in Fig. S1, the similarity between sequence S0 and S1 is same to that between sequence S0 and S2. Our method would

link S0 and S1 with transmission direction from S0 to S1 and would not link S0 and S2 based on the ancestor-offspring principle. However, the other methods including phylogenetic tree or haplotype network methods would link both S0 and S1, and S0 and S2 together based on the similarity between every two sequences.

Simulation results have demonstrated that our method can accurately reconstruct the evolutionary tree based on the ancestor-offspring relationship between samples. By using incomplete data of sequences and a reference not being the direct ancestor, our method could reveal the early transmission patterns of SARS-CoV-2. This method would have broad applications in studying the origins and transmission patterns of various viruses.

We validated our model results using data of the first 6 months. A total of 33,818 chains with 8426 root nodes or networks were reconstructed (Table S1). These root nodes contain 13,069 samples from 88 countries. There are 5 large networks with length ≥ 10, and their nodes and samples accounting for 48% and 55% of the total sample size, respectively. Similar to results using data of the first 3 months, there are plenty of independent transmission networks, much higher than those of the first 3 months, indicating many transmission chains or networks were not detected in the first 3 months. There is also no root sample of the first 15 days. The length of transmission chain of the first 6 months ranged from 1 to 13, longer than that of the first 3 months. The time of reconstructing the network using data of the first 6 months is much longer than using data of the first 3 months.

In summary, our results suggest that long time before the first COVID-19 case detected in Wuhan, China, SARS-CoV-2 would have been widely spread globally and independently. Global cooperation is essential in searching for the origins and its natural hosts of SARS-CoV-2 in the world, and in preventing the occurrence of next pandemic.

Declarations

Competing interests: The authors declare that they have no competing interests.

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Funding: This study is supported by the grant support from the Ministry of Science and Technology of the People's Republic of China (2021YFC0863400), the Institute of Zoology, Chinese Academy of Sciences (E0517111; E122G611).

Acknowledgment: We are grateful to Prof. Jian Lu and Dr. Xiaolu Tang from Beijing University; Prof. Hua Chen from The Beijing Genome Institute, Chinese Academy of Sciences for their kind help in data analysis and editing, and valuable comments to this manuscript.

Authors' contributions: Conceptualization: Z.Z., C.C.; Investigation: Z.Z, C.C.; Data curation: C.C.; Formal analysis: C.C., Z.Z.; Methodology: Z.Z, C.C.; Software: C.C.; Visualization: C.C, Z.Z.; Writing, original draft: Z.Z., C.C.; Writing, review & editing: Z.Z., C.C;

Availability of data and materials: The SARS-CoV-2 sequences used in this study were all downloaded from the GISAID (https://www.gisaid.org/) database, and the accession number of each genomic sequence of this study is available in the ScienceDB (https://www.scidb.cn/en) repository, at https://dx.doi.org/10.57760/sciencedb.01771.

References

1. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579:265–9.
2. Zhou H, Chen X, Hu T, Li J, Song H, Liu Y, et al. 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. 2020;30:2196–2203.e3.
3. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet. 2020;395:565–74.
4. Zhou P, Yang X Lou, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579:270–3.
5. Deng S-Q, Peng H-J. Characteristics of and Public Health Responses to the Coronavirus Disease 2019 Outbreak in China. Journal of Clinical Medicine. 2020;9:575.
6. Temmam S, Vongphayloth K, Baquero E, Munier S, Bonomi M, Regnault B, et al. Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. Nature. 2022;604:330–6.
7. Mallapaty S. Coronaviruses closely related to the pandemic virus discovered in Japan and Cambodia. Nature. 2020;588:15–6.
8. Lau SKP, Luk HKH, Wong ACP, Li KSM, Zhu L, He Z, et al. Possible Bat Origin of Severe Acute Respiratory Syndrome Coronavirus 2. Emerging Infectious Diseases. 2020;26:1542–7.
9. Paraskevis D, Kostaki EG, Magiorkinis G, Panayiotakopoulos G, Sourvinos G, Tsiodras S. 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. 2020;79 January:104212.
10. Shan K, Wei C, Wang Y, Huan Q, Qian W. Host-specific asymmetric accumulation of mutation types reveals that the origin of SARS-CoV-2 is consistent with a natural process. The Innovation. 2021;:100159.
11. Zhang YZ, Holmes EC. A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. Cell. 2020;181:223–7.
12. Dorp L van, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infection, Genetics and Evolution. 2020;83:104351.

13. Li X, Zai J, Zhao Q, Nie Q, Li Y, Foley BT, et al. Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. Journal of Medical Virology. 2020;92:602–11.
14. Chan JFW, Kok KH, Zhu Z, Chu H, To KKW, Yuan S, et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerging Microbes and Infections. 2020;9:221–36.
15. Li X, Zai J, Wang X, Li Y. Potential of large "first generation" human-to-human transmission of 2019-nCoV. Journal of Medical Virology. 2020;92:448–54.
16. Li X, Wang W, Zhao X, Zai J, Zhao Q, Li Y, et al. Transmission dynamics and evolutionary history of 2019-nCoV. Journal of Medical Virology. 2020;92:501–11.
17. Duchene S, Lemey P, Stadler T, Ho SYW, Duchene DA, Dhanasekaran V, et al. Bayesian Evaluation of Temporal Signal in Measurably Evolving Populations. Molecular Biology and Evolution. 2020;37:3363–79.
18. Lauring AS, Hodcroft EB. Genetic Variants of SARS-CoV-2—What Do They Mean? JAMA. 2021;325:529.
19. Burki T. Understanding variants of SARS-CoV-2. The Lancet. 2021;397:462.
20. Abdool Karim SS, de Oliveira T. New SARS-CoV-2 Variants — Clinical, Public Health, and Vaccine Implications. New England Journal of Medicine. 2021;384:1866–8.
21. Rambaut A. Phylodynamic Analysis | 176 genomes | 6 Mar 2020 - Novel 2019 coronavirus / nCoV-2019 Genomic Epidemiology - Virological. 2020. https://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356. Accessed 15 Feb 2021.
22. Hill V, Rambaut A. Phylodynamic analysis of SARS-CoV-2 | Update 2020-03-06 - SARS-CoV-2 coronavirus. 2020. https://virological.org/t/phylodynamic-analysis-of-sars-cov-2-update-2020-03-06/420. Accessed 17 Feb 2021.
23. Giovanetti M, Benvenuto D, Angeletti S, Ciccozzi M. The first two cases of 2019-nCoV in Italy: Where they come from? Journal of Medical Virology. 2020;92:518–21.
24. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. National Science Review. 2020;7:1012–23.
25. Turakhia Y, De Maio N, Thornlow B, Gozashti L, Lanfear R, Walker CR, et al. Stability of SARS-CoV-2 phylogenies. PLOS Genetics. 2020;16:e1009175.
26. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proceedings of the National Academy of Sciences. 2020;117:9241–3.
27. Nie Q, Li X, Chen W, Liu D, Chen Y, Li H, et al. Phylogenetic and phylodynamic analyses of SARS-CoV-2. Virus Research. 2020;287 July:198098.
28. Tang X, Ying R, Yao X, Li G, Wu C, Tang Y, et al. Evolutionary analysis and lineage designation of SARS-CoV-2 genomes. Science Bulletin. 2021. https://doi.org/10.1016/j.scib.2021.02.012.
29. Song S, Ma L, Zou D, Tian D, Li C, Zhu J, et al. The Global Landscape of SARS-CoV-2 Genomes, Variants, and Haplotypes in 2019nCoVR. Genomics, Proteomics & Bioinformatics. 2020;18:749–59.

30. Pipes L, Wang H, Huelsenbeck JP, Nielsen R. Assessing Uncertainty in the Rooting of the SARS-CoV-2 Phylogeny. Molecular Biology and Evolution. 2021;38:1537–43.
31. Morel B, Barbera P, Czech L, Bettisworth B, Hübner L, Lutteropp S, et al. Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult. Molecular Biology and Evolution. 2021;38:1777–91.
32. Tong Y, Liu W, Liu P, Liu WJ, Wang Q, Gao GF. The origins of viruses: discovery takes time, international resources, and cooperation. The Lancet. 2021;6736:2–3.
33. Lundstrom K, Seyran M, Pizzol D, Adadi P, Mohamed Abd El-Aziz T, Hassan SkS, et al. The Importance of Research on the Origin of SARS-CoV-2. Viruses. 2020;12:1203.
34. Fan Y, Zhao K, Shi Z-L, Zhou P. Bat Coronaviruses in China. Viruses. 2019;11:210.
35. Li W, Shi Z, Yu M, Ren W, Smith C, Epstein JH, et al. Bats are natural reservoirs of SARS-like coronaviruses. Science (1979). 2005;310:676–9.
36. Zhou H, Ji J, Chen X, Bi Y, Li J, Wang Q, et al. Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. Cell. 2021;184:4380–4391.e14.
37. Hul V, Delaune D, Karlsson EA, Hassanin A, Tey PO, Baidaliuk A, et al. A novel SARS-CoV-2 related coronavirus in bats from Cambodia. bioRxiv. 2021. https://doi.org/10.1101/2021.01.26.428212.
38. Wacharapluesadee S, Tan CW, Maneeorn P, Duengkae P, Zhu F, Joyjinda Y, et al. Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia. Nature Communications. 2021;12:972.
39. Temmam S, Vongphayloth K, Salazar EB, Munier S, Bonomi M, Régnault B, et al. Coronaviruses with a SARS-CoV-2-like receptor- binding domain allowing ACE2-mediated entry into human cells isolated from bats of Indochinese peninsula. Research Square. 2021; September 17th. https://doi.org/10.21203/rs.3.rs-871965/v1.
40. Lam TT-Y, Jia N, Zhang Y-W, Shum MH-H, Jiang J-F, Zhu H-C, et al. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature. 2020;583:282–5.
41. Xiao K, Zhai J, Feng Y, Zhou N, Zhang X, Zou J-J, et al. Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. Nature. 2020;583:286–9.
42. Wu C-I, Wen H, Lu J, Su X, Hughes AC, Zhai W, et al. On the origin of SARS-CoV-2—The blind watchmaker argument. Science China Life Sciences. 2021;64:1560–3.
43. Joint WHO-China Study Team. WHO-Convened Global Study of Origins of SARS-CoV-2: China Part (Text Extract). Infectious Diseases & Immunity. 2021;Publish Ah.
44. Segreto R, Deigin Y. The genetic structure of SARS-CoV-2 does not rule out a laboratory origin. BioEssays. 2021;43:2000240.
45. Ruan Y, Wen H, Hou M, He Z, Lu X, Xue Y, et al. The twin-beginnings of COVID-19 in Asia and Europe – One prevails quickly. National Science Review. 2021;186:227–36.
46. Ellis N, Ciocci S, German J. Back mutation can produce phenotype reversion in Bloom syndrome somatic cells. Human Genetics. 2001;108:167–73.

47. Amit Roy SR. Molecular Markers in Phylogenetic Studies-A Review. Journal of Phylogenetics & Evolutionary Biology. 2014;02.
48. Lanciotti RS, Ebel GD, Deubel V, Kerst AJ, Murri S, Meyer R, et al. Complete Genome Sequences and Phylogenetic Analysis of West Nile Virus Strains Isolated from the United States, Europe, and the Middle East. Virology. 2002;298:96–105.
49. Poon AFY, Walker LW, Murray H, McCloskey RM, Harrigan PR, Liang RH. Mapping the Shapes of Phylogenetic Trees from Human and Zoonotic RNA Viruses. PLoS ONE. 2013;8:e78122.
50. Holmes EC, Nee S, Rambaut A, Garnett GP, Harvey PH. Revealing the history of infectious disease epidemics through phylogenetic trees. Philos Trans R Soc Lond B Biol Sci. 1995;349:33–40.
51. Li J, Li Z, Cui X, Wu C. Bayesian phylodynamic inference on the temporal evolution and global transmission of SARS-CoV-2. Journal of Infection. 2020;81:318–56.
52. Nabil B, Sabrina B, Abdelhakim B. Transmission route and introduction of pandemic SARS-CoV-2 between China, Italy, and Spain. Journal of Medical Virology. 2021;93:564–8.
53. Sekizuka T, Itokawa K, Kageyama T, Saito S, Takayama I, Asanuma H, et al. Haplotype networks of SARS-CoV-2 infections in the Diamond Princess cruise ship outbreak. Proceedings of the National Academy of Sciences. 2020;117:20198–201.
54. Liu Q, Zhao S, Shi C-M, Song S, Zhu S, Su Y, et al. Population Genetics of SARS-CoV-2: Disentangling Effects of Sampling Bias and Infection Clusters. Genomics, Proteomics & Bioinformatics. 2020; xxxx:4–11.
55. Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. Molecular Biology and Evolution. 1999;16:37–48.
56. Kong S, Sánchez-Pacheco SJ, Murphy RW. On the use of median-joining networks in evolutionary biology. Cladistics. 2016;32:691–9.

Figures

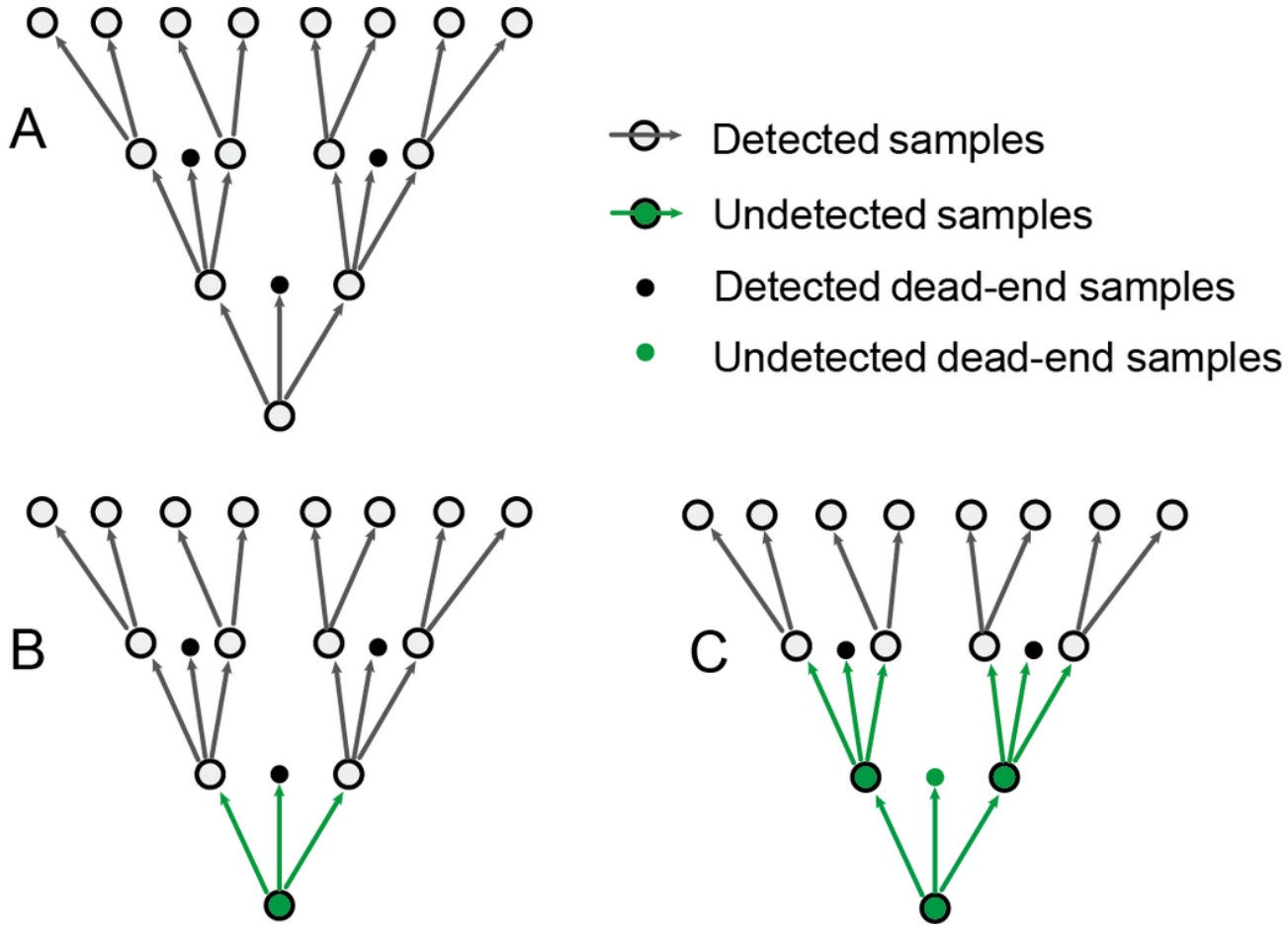

Figure 1

Three hypotheses on the transmission network of SARS-CoV-2 which is reconstructed by using detected samples. (A) Originating Lineage Hypothesis (OLH). (B) Intermediate Lineage Hypothesis (ILH). (C) Tip Lineage Hypothesis (TLH).

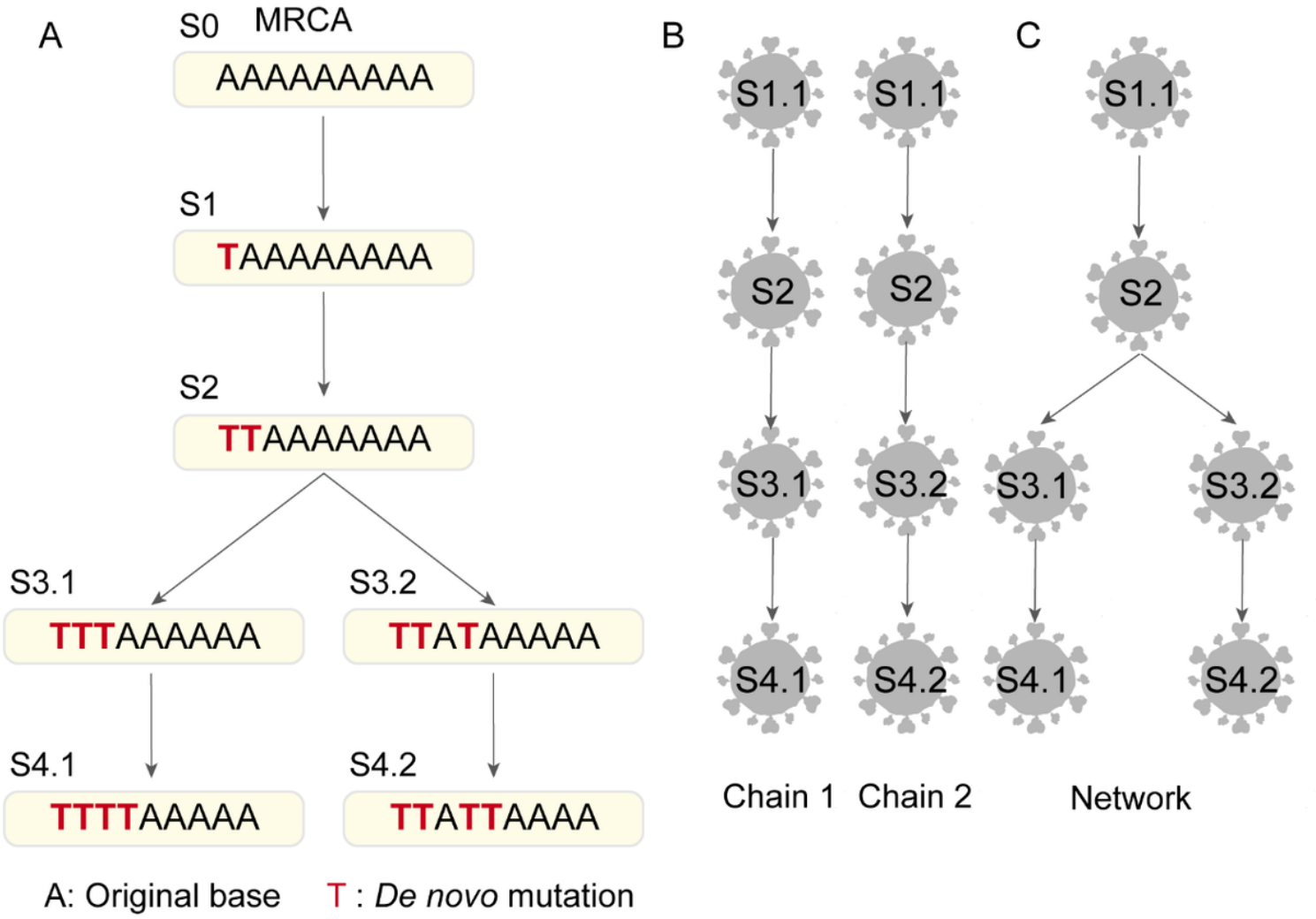

Figure 2

Illustrations of reconstructing the transmission chains and networks based on mutation using MRCA as the reference. For simplification, all nucleotide sequences are assumed to have a chain length of 10 (the original nucleotide is black letter A, and the *de novo* mutation nucleotides is red letter T). (A) The 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). (B) The transmission chains were reconstructed from Panel A. (C) The transmission network was reconstructed based on panel B by merging transmission chains sharing a common node.

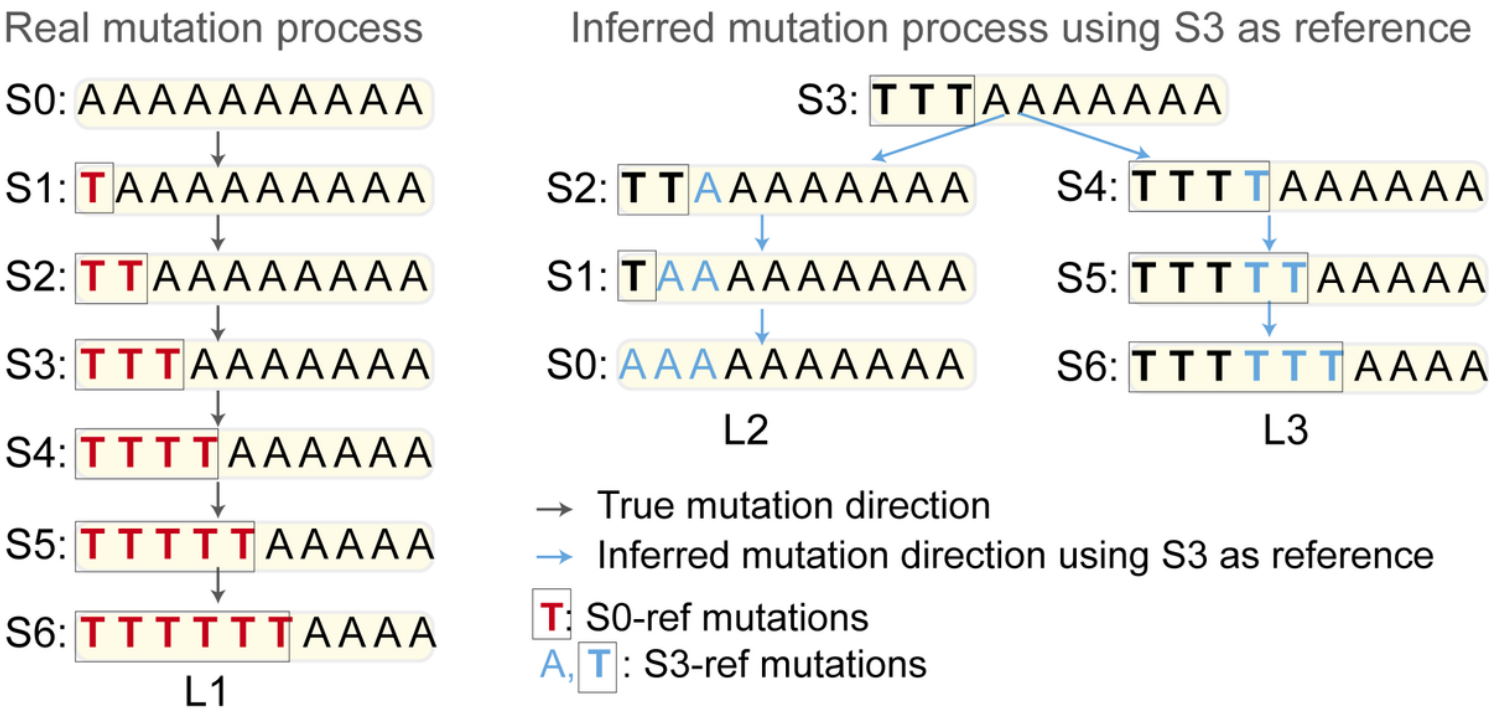

Figure 3

Illustrations of defining the *de novo* mutations using the most recent common ancestor (S0) as the reference (S0-ref mutations) and inferred mutations using an offspring (S3) as the reference (S3-ref mutation), and reconstructing transmission chains or networks of SARS-CoV-2 based on ancestral (S0) and offspring (S3) sequences. For simplification, all nucleotide sequences are assumed to have a chain length of 10 (the original nucleotide is A, and the mutation nucleotides is T). Letter A is the original nucleotide base type, red and boxed letter T in L1 is the mutated base type by referring to the ancestral sequence (S0). Cyan letter A and Cyan boxed letter T in L2 and L3 are the inferred mutations based on the offspring sequence (S3). L1 is correctly reconstructed by using *de novo* mutations. L2 and L3 are incorrectly reconstructed by using S3-ref mutations which appears later than the ancestral node.

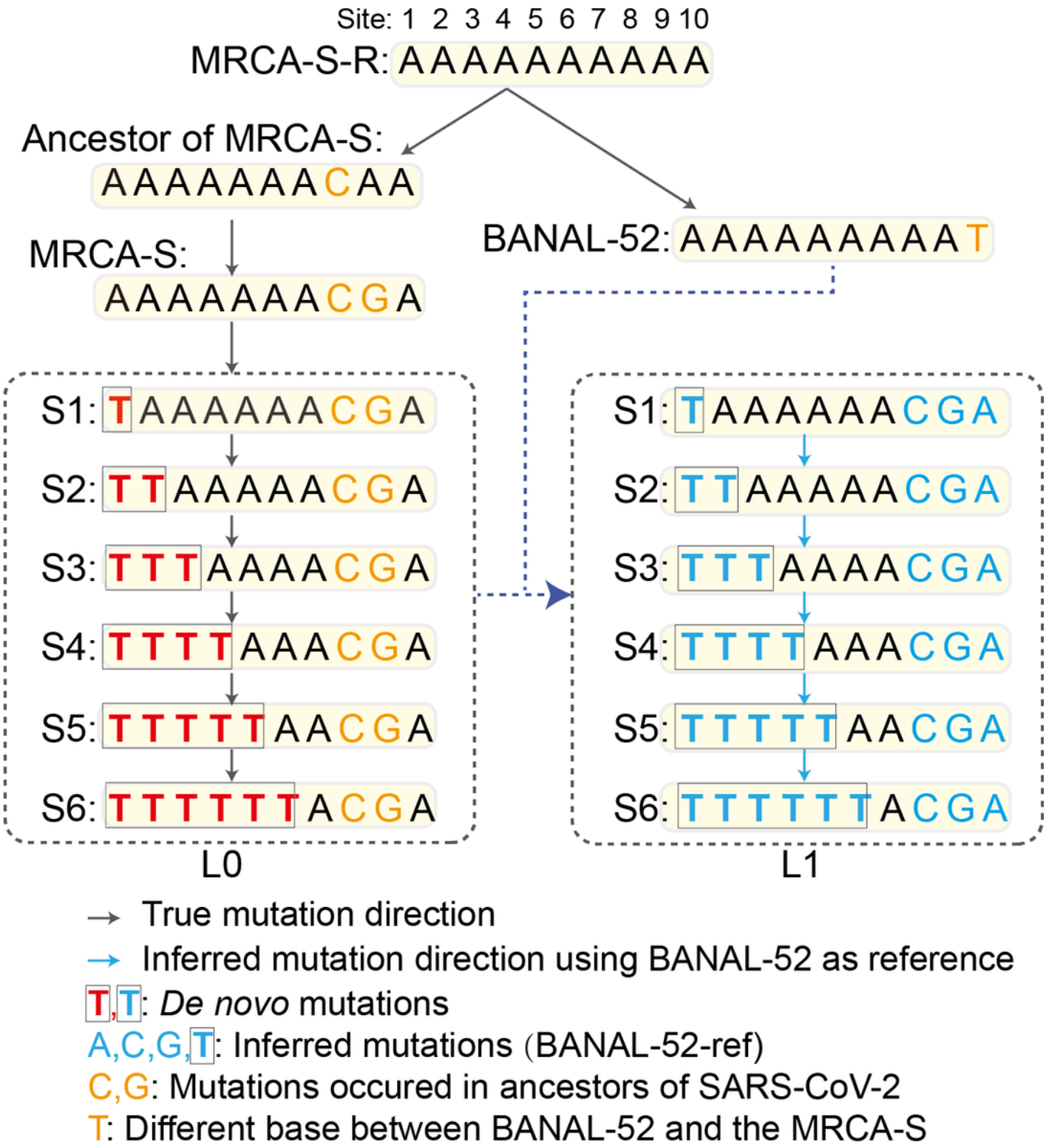

Figure 4

Illustration of defining the *de novo* mutations using MRCA-S (MRCA-ref mutations) or inferred mutations using BANAL-52 (BANAL-52-ref mutations) and reconstructing transmission chains of SARS-CoV-2 by using MRCA-S (L0) or an earlier and closest relative (BANAL-52) (L1). Black A is the original nucleotide base. Red boxed T is the *de novo* mutation site using MRCA-S, orange C, G are ancestor mutations of SARS-CoV-2 using MRCA-S-R. Cyan boxed T is *de novo* mutations using BANAL-52 as the reference.

Orange T is the mutation of BANAL-52 using MRCA-S-R as the reference. Cyan C, G are ancestor mutations of SARS-CoV-2 using BANAL-52 as the reference. Cyan A is the incorrect mutation using BANAL-52 as the reference. L0 is the true mutation chain using MCRA-S as the reference. Lineage L1 was correctly reconstructed by using a non-ancestor relative (BANAL-52) as the reference which appeared earlier than the detected samples of the SARS-CoV-2. Cyan C, G and A were not considered in reconstruction of transmission chain of L1 if no secondary mutation occurs on these sites (see below).

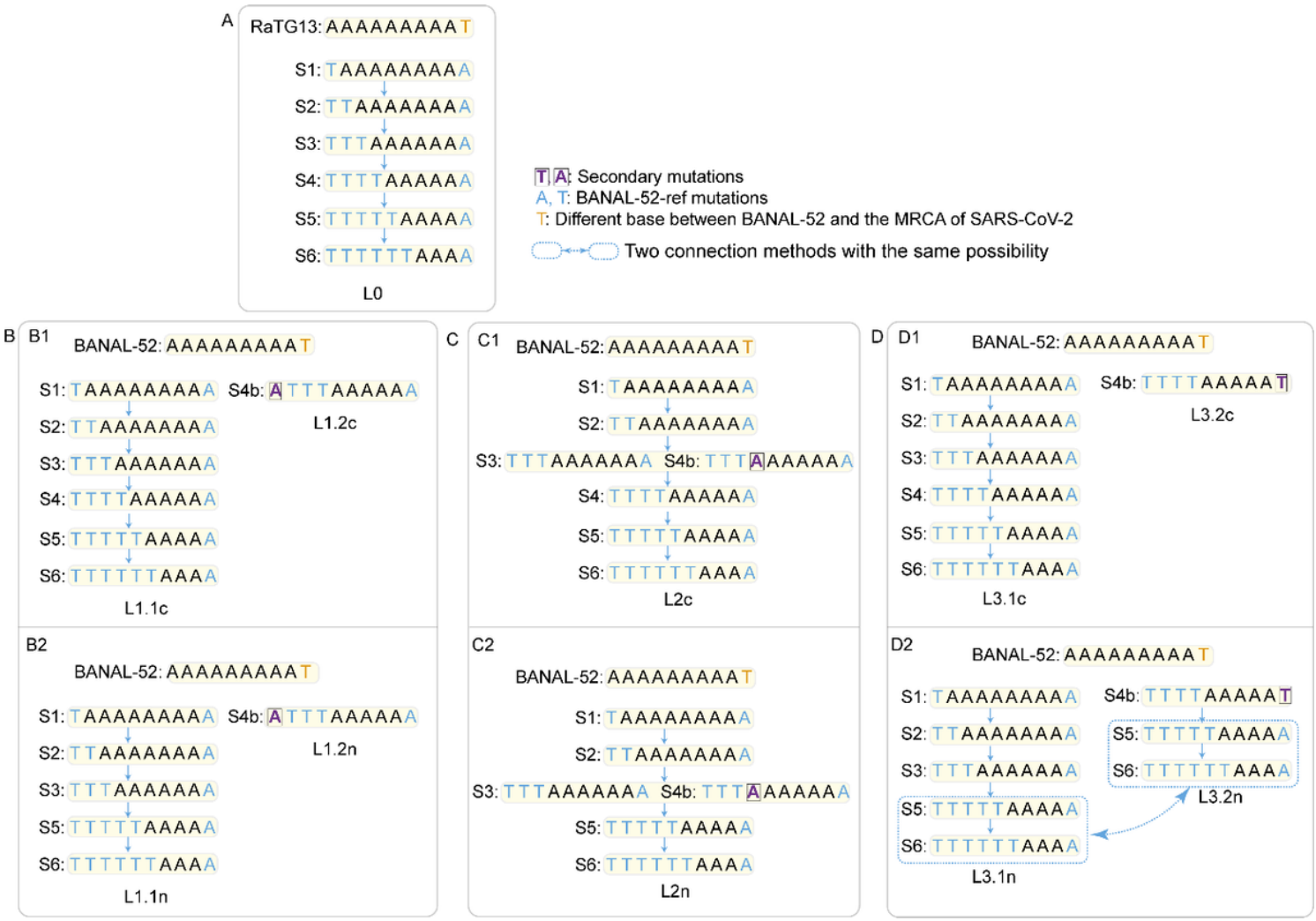

Figure 5

Illustration of model errors caused by the secondary mutations during the study period of three months (i.e., bold purple boxed T or A in sequence S4) in the reconstruction of transmission chains and networks by using BANAL-52 as the reference. S4 is assumed to be mutated into S4b. (A) Transmission chains reconstructed by using samples with no secondary mutations: L0 = S1→S2→S3→S4→S5→S6. (B) Secondary mutations occur in the *de novo* mutation sites of sequence S4 with or without extra copies, resulting in L1.1c = S1→S2→S3→S4→S5→S6 and L1.2c = S4b (B1, with extra copies), or L1.1n = S1→S2→S3→S5→S6 and L1.2n = S4b (B2, without extra copies). (C) Secondary mutations occur in the *de novo* mutation site of S4, resulting in a sequence same to S3. If S4 has extra copies, L2c = S1→S2→S3→S4→S5→S6 (C1), otherwise, L2n = S1→S2→S3→S5→S6 (C2). (D) Secondary mutations

occur in the site that BANAL-52 is different from that of the MRCA of SARS-CoV-2 (equivalent to the 3.2% difference region in genome between SARS-CoV-2 and BANAL-52). L3.1c = L1.1n = S1→S2→S3→S4→S5→S6, and L3.2c = S4b (D1, with extra copies), or L3.1n = S1→S2→S3→S5→S6, L3.2n = S4b→S5→S6 (D2, without extra copies).

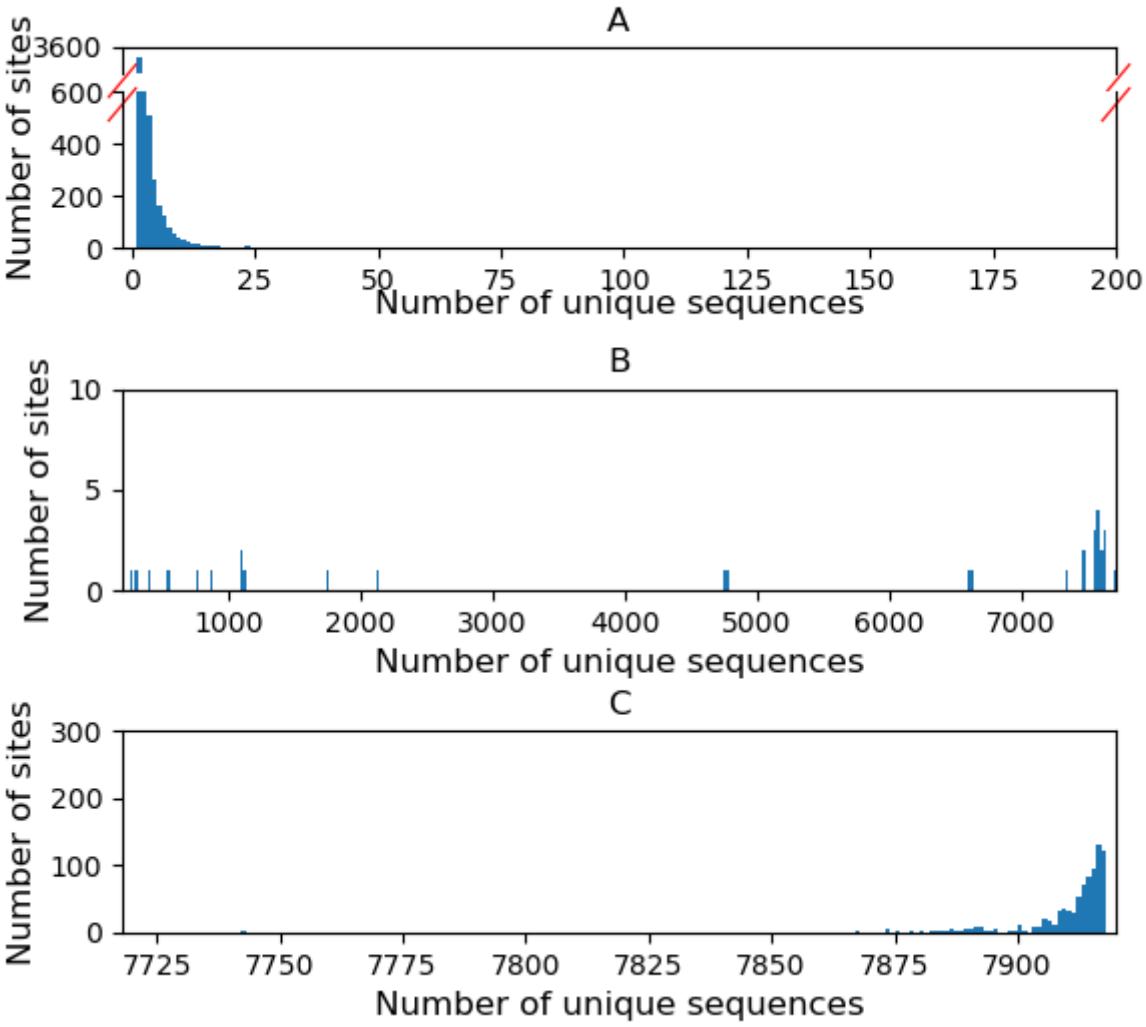

Figure 6

Frequency in the number of BANAL-52-ref mutation sites in SARS-CoV-2 genome against the number of unique sequences (nodes) that contain these sites. A. 1 to 200 unique sequences. B. 201 to 7718 unique sequences. C. 7719 to 7918 unique sequences.

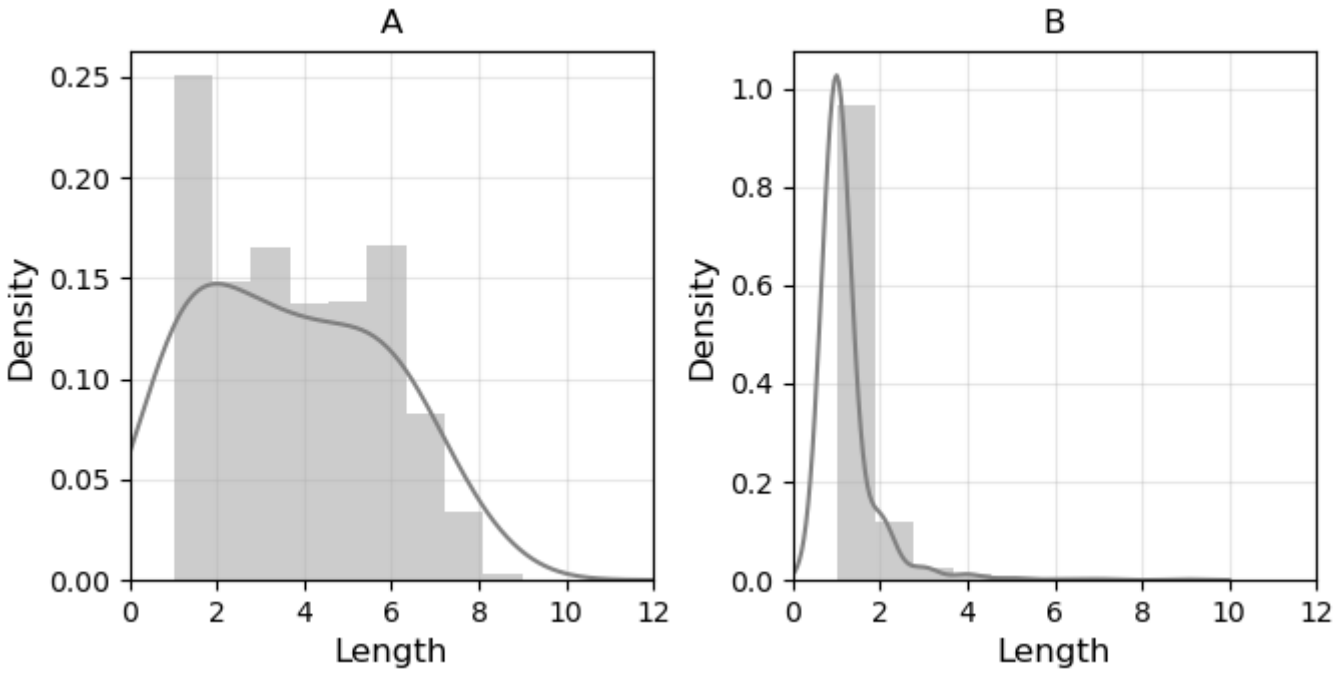

Figure 7

Kernel density of transmission chains with different chain lengths (i.e., number of nodes) in the transmission chains or networks of SARS-CoV-2. (A) Transmission chains. (B) Transmission network by only counting the longest transmission chain.

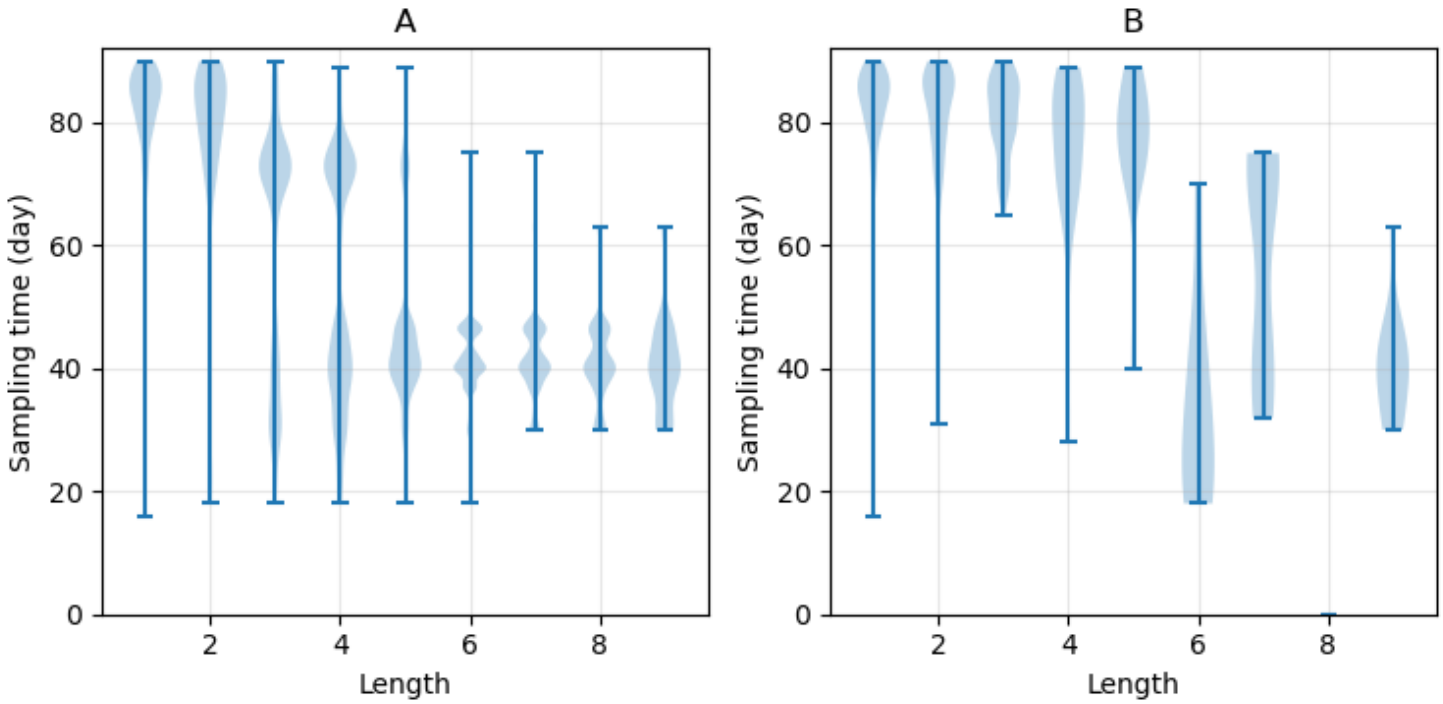

Figure 8

Frequency of sampling time (day) of root node samples with different chain lengths. (A) Transmission chains. (B) Transmission network by keeping chains with the longest chain length. The lower end of the line represents the earliest sample sampling time of root nodes, the upper end represents the latest sample sampling time of root nodes, and the width of the shade indicates the number of samples corresponding to the sampling time.

Supplementary Files

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

- supportinginformation.docx