ZR Zoological Research

Supplementary Materials

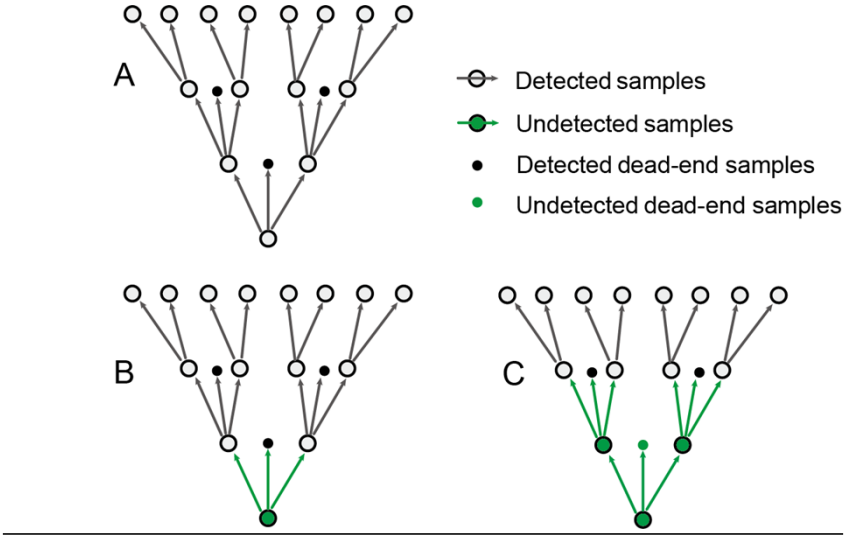
Reconstructing early transmission networks of SARS-CoV-2 using a genomic mutation model

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

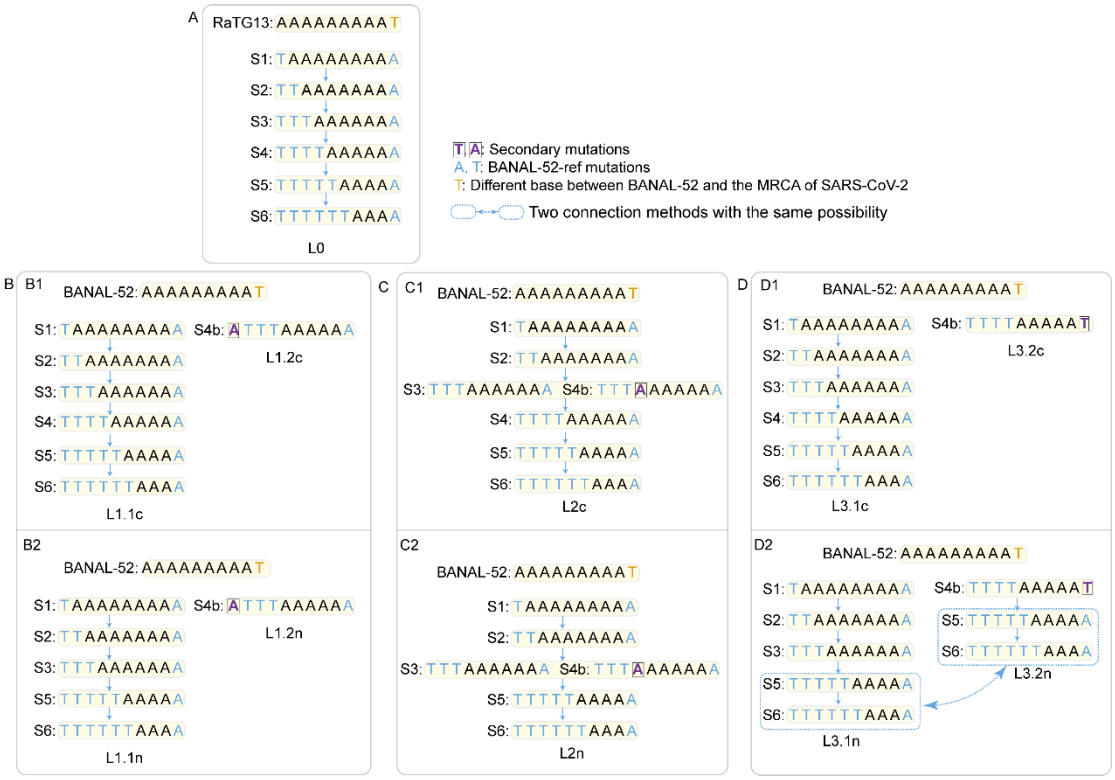
[2]CAS Center for Excellence in Biotic Interactions, University of Chinese Academy of Sciences, Beijing 100049, China

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



Supplementary Figure S1 Three hypotheses of SARS-CoV-2 transmission network, reconstructed using detected samples

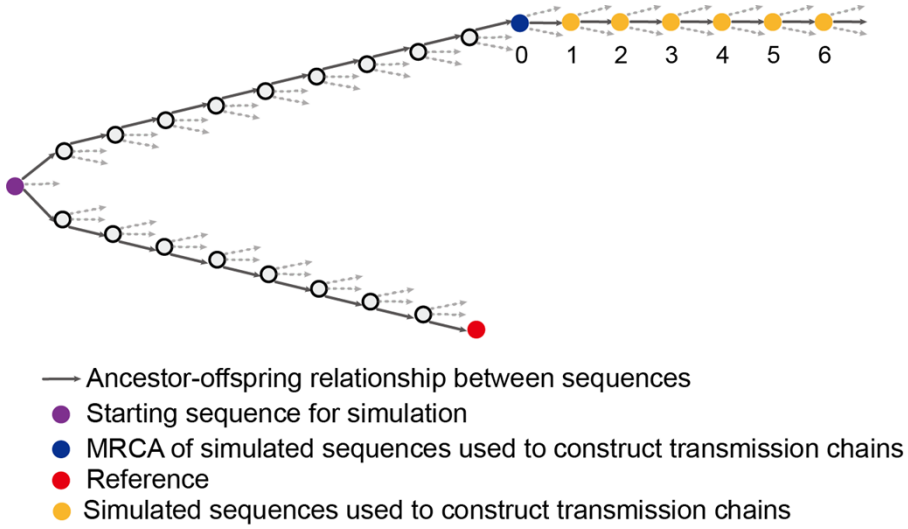
A: Originating Lineage Hypothesis. B: Intermediate Lineage Hypothesis. C: Tip Lineage Hypothesis.



Supplementary Figure S2 Illustration of model errors caused by secondary mutations during three-month study period (i.e., bold purple boxed T or A in sequence S4) in reconstruction of transmission chains and networks using BANAL-52 as a reference

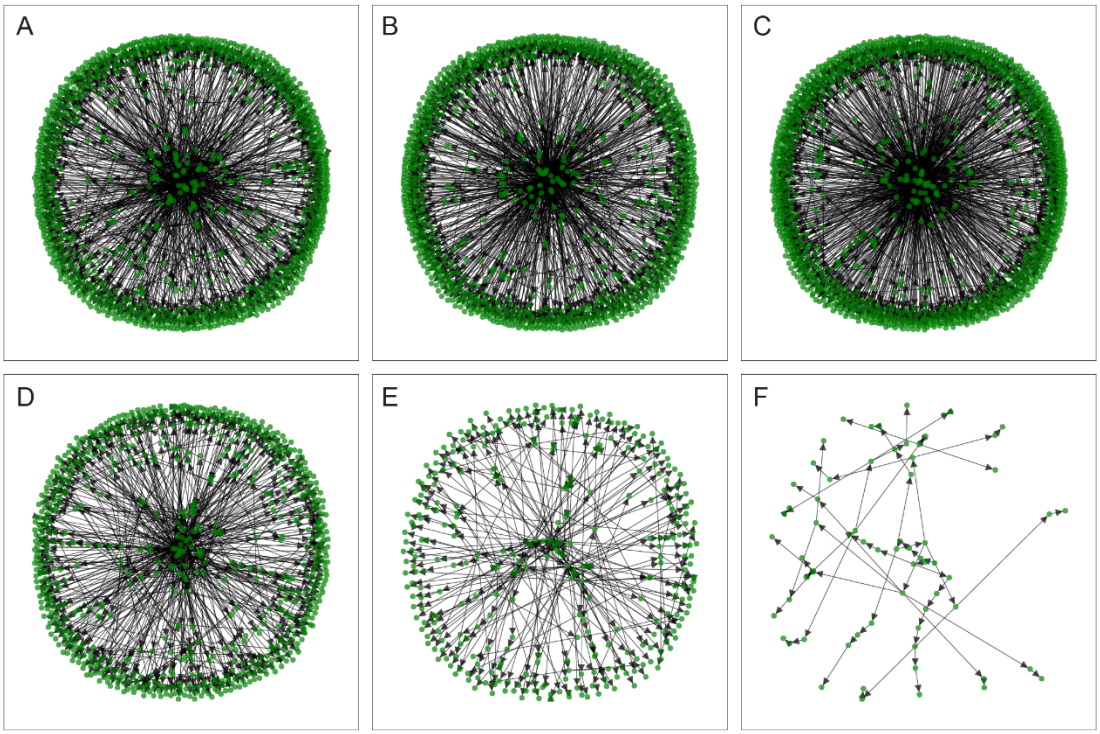
S4 is assumed to be mutated into S4b and S5. A: Transmission chains reconstructed using samples with no secondary mutations: L0=S1→S2→S3→S4→S5→S6. B: Secondary mutations occur in *de novo* mutation sites of sequence S4 with or without extra copies, resulting in L1.1c=S1→S2→S3→S4→S5→S6 and L1.2c=S4b (B1, with extra copies), or L1.1n=S1→S2→S3→S5→S6 and L1.2n=S4b (B2, without extra copies). C: Secondary mutations occur in *de novo* mutation site of S4, resulting in the same sequence as 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

where BANAL-52 differs from the MRCA of SARS-CoV-2 (equivalent to 3.2% differentiated region between SARS-CoV-2 and BANAL-52 genomes). 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).



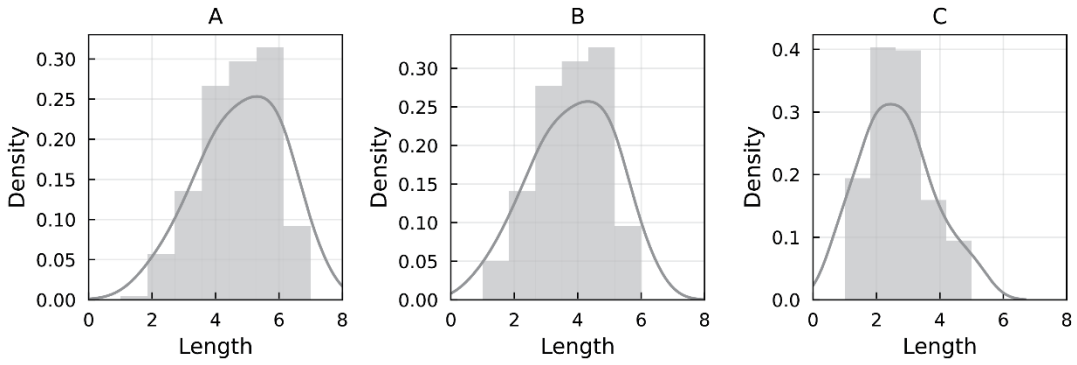
Supplementary Figure S3 Simulated evolutionary process and reconstruction of transmission network of viruses based on ancestor-offspring relationship between sequences and outgroup-ref mutations

Circles represent sequences, with only one sequence shown in each generation due to space limitations. Dashed arrows represent clades (and offspring sequences) not shown due to space limitations. Positions of reference and detected samples were determined by roughly referring to the evolutionary pattern of SARS-CoV-2 from bat coronavirus.



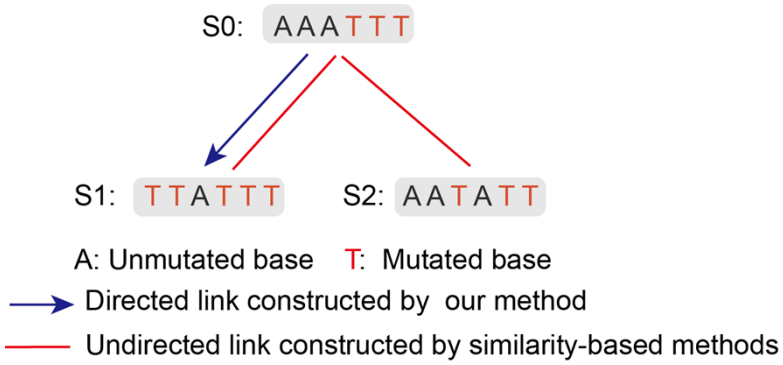
Supplementary Figure S4 Reconstructed transmission networks of SARS-CoV-2 with four nodes (A), five nodes (B), six nodes (C), seven nodes (D), eight nodes (E), and nine nodes (F) using BANAL-52-referenced mutations

Figure was plotted using *NetworkX* package in Python v3.7.



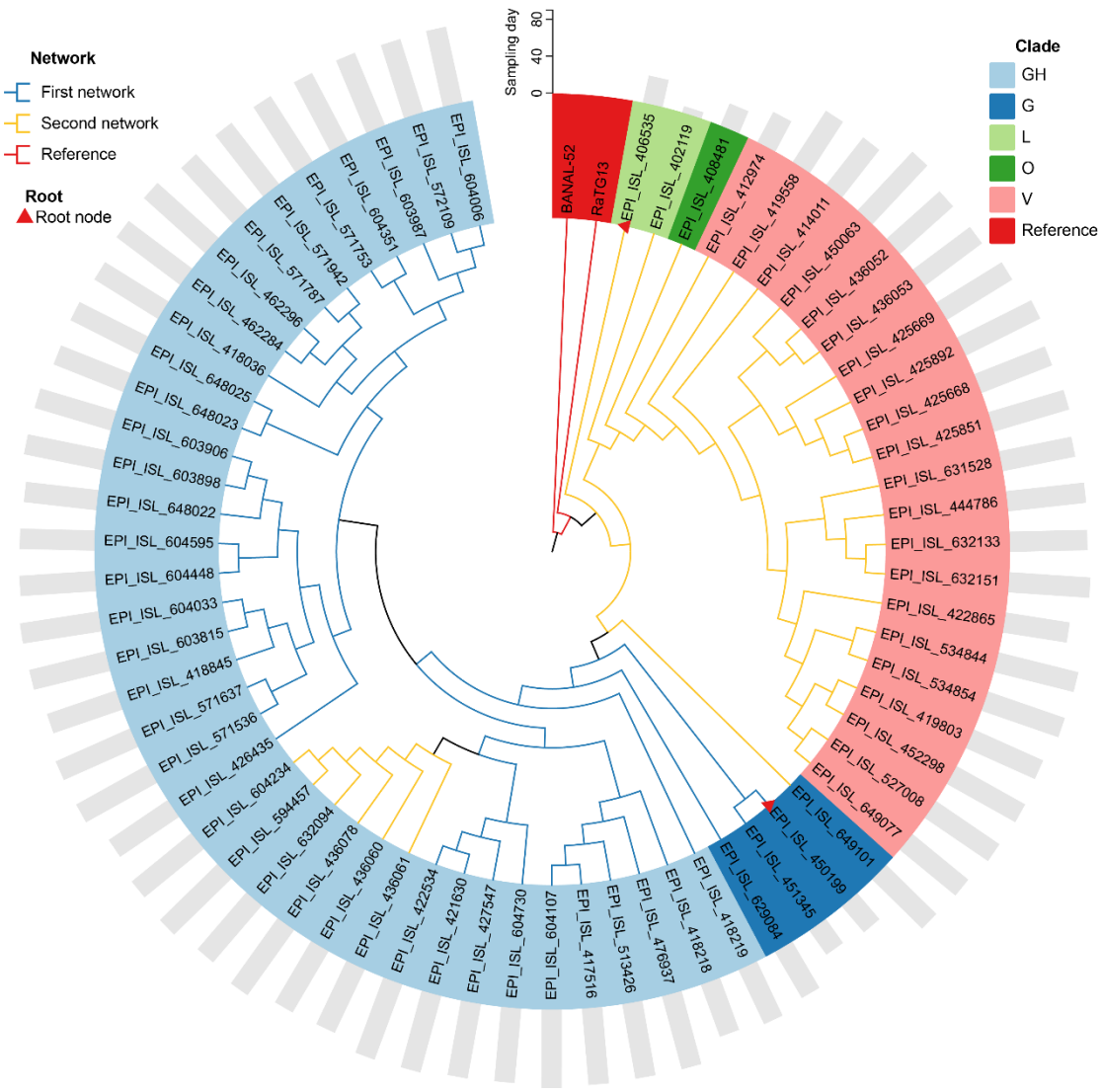
Supplementary Figure S5 Kernel density of transmission chains with different chain lengths using simulation data

A: Using 0^{th} to 6^{th} generation data (including ancestor). B Using 2^{nd} to 6^{th} generation data (not including ancestor). C: Using 4^{th} to 6^{th} generation data (not including ancestor).



Supplementary Figure S6 Differences between our ancestor-offspring network method and similarity-based methods

S0, S1, and S2 are sequences (six bases long). Black letter A is original base and red letter T in each sequence indicates mutations. According to our method, there is a directional link between sequences S0 and S1 (i.e., S0→S1), but no link relationship between S0 and S2 (i.e., S0 and S2 are not ancestors of each other). Based on haplotype network methods, S0 is linked to both S1 and S2, respectively, given the same distance between them (i.e., genetic distances between S0 and S1 and between S0 and S2 are the same), and there is no direction of transmission.



Supplementary Figure S7 Comparison between our approach and maximum-likelihood phylogenetic tree using node samples from two transmission networks with longest chain lengths

Colors of tree branch indicate samples of our transmission networks and colors of tree leaf name indicate clades in GISAID (https://gisaid.org/). Red triangle indicates root node sample of networks; blue lines represent samples of first transmission network; yellow lines represent samples of second transmission network; red lines represent outgroup sequences; and gray bars represent sampling day (24 December 2019 set as day 1). Tree was constructed using IQ-TREE v1.6.12 (http://www.iqtree.org/) with default settings and visualized in ChiPlot (https://www.chiplot.online/).

Supplementary Table S1. Statistics of nodes of the reconstructed transmission chains and networks with different lengths using data of the first six months of BANAL-52-referenced mutations.

Chain length	No. chains	No. root nodes (networks)	No. root samples	No. root country/ regions	First sampling time	Last sampling time
1	7087	7087	9697	86	16	180
2	5548	1151	3107	55	18	180
3	3590	478	1285	44	18	180
4	2557	191	582	36	18	180
5	3491	77	189	27	18	180
6	4782	38	112	20	18	158
7	3977	20	58	12	18	158
8	1803	10	26	9	18	113
9	680	6	18	7	18	113
10	227	5	17	6	18	113
11	64	3	8	3	37	113
12	10	2	7	2	37	113
13	2	1	6	1	37	47
Total	33818	8426	13069	88	16	180

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.

Supplementary Table S2. Statistics of nodes of the reconstructed transmission chains with different lengths within the first three months using RaTG13-referenced mutations as reference.

Chain length	No. chains	No. root nodes (networks)	No. root samples	No. root country/ regions	First sampling time	Last sampling time
1	1742	1742	2379	62	16	90
2	836	242	728	42	18	90
3	1228	96	255	33	18	90
4	959	36	90	22	18	90
5	1056	19	54	14	18	88
6	676	9	39	12	18	88
7	209	6	20	8	18	87
8	30	3	8	3	31	75
9	1	1	1	1	31	31
Total	6737	2041	3193	69	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. *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., the 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.