SCIENCE CHINA
Life Sciences

•LETTER TO THE EDITOR•

October 2020 Vol.63 No.10: 1608–1611
https://doi.org/10.1007/s11427-020-1764-2

The use of SARS-CoV-2-related coronaviruses from bats and pangolins to polarize mutations in SARS-Cov-2

Tao Li[1], Xiaolu Tang[2], Changcheng Wu[2], Xinmin Yao[2], Yirong Wang[2], Xuemei Lu[1*] & Jian Lu[2*]

[1]*State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology; Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China;*
[2]*State Key Laboratory of Protein and Plant Gene Research, Center for Bioinformatics, School of Life Sciences, Peking University, Beijing 100871, China*

Received May 20, 2020; accepted June 22, 2020; published online July 1, 2020

Citation: Li, T., Tang, X., Wu, C., Yao, X., Wang, Y., Lu, X., and Lu, J. (2020). The use of SARS-CoV-2-related coronaviruses from bats and pangolins to polarize mutations in SARS-Cov-2. Sci China Life Sci 63, 1608–1611. https://doi.org/10.1007/s11427-020-1764-2

Dear Editor,

The coronavirus disease 2019 (COVID-19) caused by the SARS-CoV-2 coronavirus has become a global pandemic. The SARS-CoV-2 genome has a similarity of 96.2% to that of RaTG13, a bat SARS-CoV-2-related coronavirus detected in *Rhinolophus affinis* (Paraskevis et al., 2020; Zhou et al., 2020). The SARS-CoV-2 genome also has 85.5%−92.4% sequence similarity to SARS-CoV-2-related coronaviruses from Malayan pangolins that have been seized in anti-smuggling operations in southern China (Guangdong-Pangolin (GD-Pangolin-CoV) and Guangxi-Pangolin (GX-Pangolin-CoV) genomes) (Liu et al., 2019; Lam et al., 2020). Although the genomic sequences of SARS-CoV-2 viruses share a similarity of greater than 99.9% (Lu et al., 2020; Ren et al., 2020; Zhou et al., 2020), hundreds of genetic variants have been identified across different SARS-CoV-2 strains (Forster et al., 2020; Tang et al., 2020; Yu et al., 2020). Several groups have used SARS-CoV-2-related coronavirus from bats and pangolins as outgroups to polarize the ancestral and derived mutations across SARS-CoV-2 strains (Forster et al., 2020; Tang et al., 2020; Yu et al., 2020); however, the accuracy of such ancestral inferences remains unclear. To address this issue, we conducted forward simulations of the molecular evolution of viral genomes by incorporating mutations and natural selection.

We modeled viral evolution as a stochastic Markov chain process. We assumed that: (i) an ancestral virus (N0) split into two lineages, one leading to N1 (resembling the outgroup) and the other leading to N2, resembling the most recent common ancestor of the viral strains of interest; (ii) N3 and N4 are two randomly chosen strains descending from N2; and (iii) the viruses evolved in a stochastic process, and both mutation and selection occurred during each time unit (Figure 1A). The nucleotides of N1, N3, and N4 can be determined by genome sequencing. Although the nucleotides of both N0 and N2 are unknown, we can infer the N2 nucleotide states by comparing N1, N3, and N4 using the maximum parsimony (MP) method (N2′) and subsequently compare the inferred N2′ to the actual N2 nucleotides recorded in the simulations to assess the accuracy (Figure 1B). To evaluate the effect of sequence similarity between the outgroup and ingroup lineages on the accuracy of ancestral allele inference, we modeled a series of divergence periods between N1 and N3 (or N4) and estimated the accuracy rates.

Since ~98% of the SARS-CoV-2 genome encodes proteins (coding regions; CDS), we considered only the CDS se-

*Corresponding authors (Jian Lu, email: LUJ@pku.edu.cn; Xuemei Lu, email: xuemeilu@mail.kiz.cas.cn)

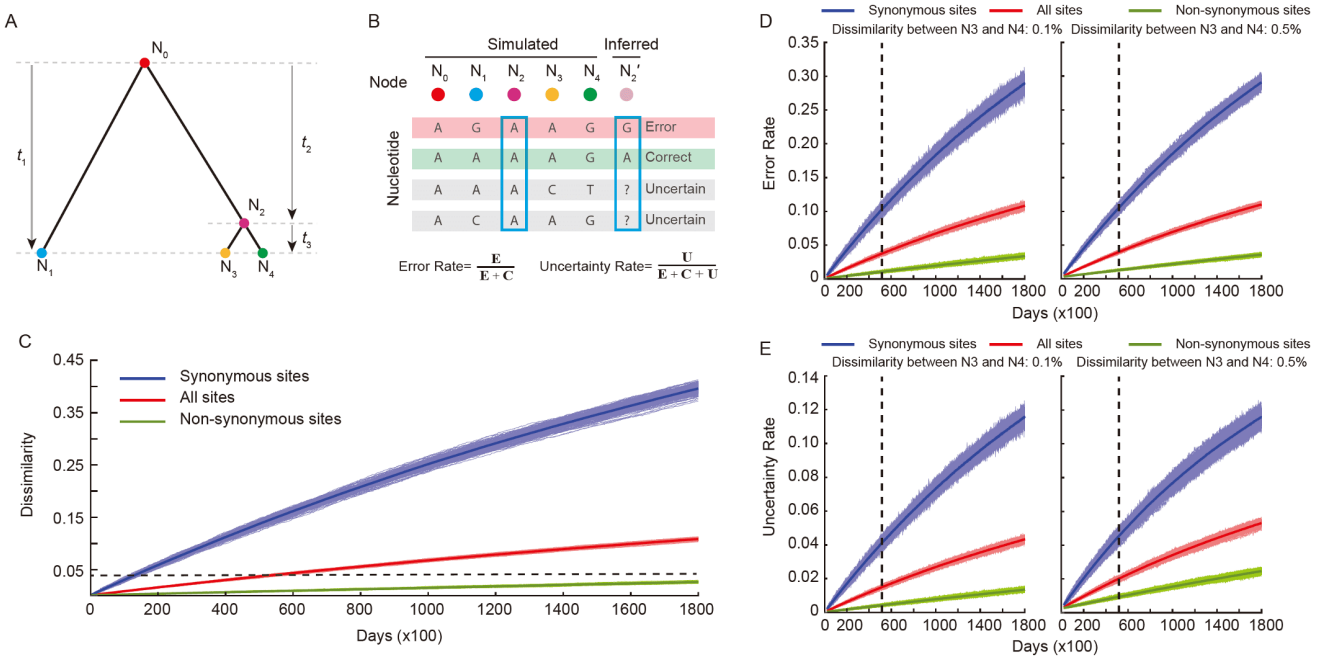

Figure 1 The molecular evolution simulation and ancestral nucleotide inference. A, In the simulation, an ancestral virus N0 split into two lineages, one leading to N1 (resembling the outgroup) and the other leading to N2 (resembling the most recent common ancestor of the viral strains of interest). After t_2 days, N2 split into two lineages, N3 and N4. B, Examples showing whether the inferred N2 nucleotide state (N2′) based on the comparison of the nucleotides of N1, N3, and N4 matched the N2 nucleotide in the simulation. Correct (C), N2=N2′; Error (E), N2≠N2′; Uncertain (U), N2 cannot be inferred using the MP method. C, Sequence divergence between N1 and N3/N4 (*y* axis) increases as the evolutionary period (days) between N1 and N3/N4 (*x* axis) increases. D, The error rate for inferring the most recent common ancestor of N3 and N4 (*y* axis) increases as the divergence period (days) between N1 and N3/N4 increases (*x* axis). E, The uncertainty rate for inferring the most recent common ancestor of N3 and N4 (*y* axis) increases as the divergence period (days) between N1 and N3/N4 increases (*x* axis). The blue lines represent the synonymous sites, the red lines represent the overall sites, and the green lines represent the nonsynonymous sites. Lines in a darker color in (C–E) represent the mean value for 200 replications of the simulations. The left and right panels of (D) and (E) represent the results when the difference between N3 and N4 (θ) was 0.1% and 0.5%, respectively. The dashed lines represent the overall genomic similarity equivalent to that between RaTG13 and SARS-CoV-2.

quences in our simulations. Our simulations also considered the nucleotide mutational bias, which was inferred by comparing the synonymous sites of extant SARS-CoV-2, RaTG13, GD-Pangolin-CoV, and GX-Pangolin-CoV CDSs (Supplementary Methods and Figure S1 in Supporting Information). We assumed both synonymous and nonsynonymous sites had the mutation rate (u) of 1.04×10^{-3} mutation/site/year, as previously described (Wang et al., 2020). We assumed synonymous mutations were neutral, and only 5% of the nonsynonymous mutations had a chance of being preserved in each time unit, as shown previously (Tang et al., 2020). Very similar results in ancestral inferences were obtained when purifying selection was not considered between SARS-CoV-2 strains (i.e., in the branches leading from N2 to N3 or N4, see below) or was considered (see Supporting Information for details).

As shown in Figure 1C, the sequence difference between N1 and N3 (or N4) increased linearly as the divergence time t (t_1+t_2+t_3 days) increased for both the synonymous and nonsynonymous sites. Moreover, the nonsynonymous substitution rate was much lower than the synonymous substitution rate due to strong purifying selection. Of note, the accuracy of inferring the ancestral vs. derived state for a variant between N3 and N4 also decreased almost linearly as the overall sequence divergence between N1 and N3 (or N4) increased (Figure 1D). Since the divergence between N1 and N3 (or N4) was considerably higher in synonymous than nonsynonymous sites, accordingly, the accuracy of ancestral inference was much higher in nonsynonymous than synonymous sites for each divergence period (Figure 1D). Meanwhile, the proportion of sites with the uncertainty of ancestral inference increased as the sequence similarity between N1 and N3 (or N4) decreased, and the occurrence of uncertainty was slightly higher in the synonymous than nonsynonymous sites (Figure 1E). Furthermore, the level of genetic difference between N3 and N4 (θ) had a negligible effect on the accuracy of the inference, since we obtained similar results when θ was set at 0.1% or 0.5% (Figure 1D).

When the divergence period t (t=t_1+t_2+t_3, Figure 1A) reached 55,600 days (~152.33 years), we obtained a sequence similarity of 95.95% between N1 and N3 (or N4), which resembled the genome divergence between RaTG13 and SARS-CoV-2. For the sites at which ancestral states could be unambiguously inferred using the MP method, the accuracy of polarizing the ancestral vs. derived states between N3 and N4 for a site was roughly 95.98% (95% CI,

95.96%–96.00%). Specifically, the accuracy rate for ancestral inference at a synonymous site was 89.01% (95% CI, 88.93%–89.07%) and that at a nonsynonymous site was 98.85% (95% CI, 98.84%–98.86%) (Figure 1D). Of note, the accuracy rates were similar when θ was set at 0.1% or 0.5%. Moreover, 1.94% of the variant sites (95% CI, 1.93%–1.96%) could not be unambiguously inferred using the MP method at such a divergence level (Figure 1E). Specifically, the uncertainty rate for ancestral inference at the synonymous sites was 4.63% (95% CI, 4.59%–4.67%) and that at the nonsynonymous sites was 0.8% (95% CI, 0.79%–0.81%).

One caveat in the above analysis is that mutations specifically occurring in the lineage leading from N0 to N1 (i.e., the RaTG13 lineage) will either cause errors or uncertainty in inferring ancestral vs. derived mutations between N3 and N4 (i.e., between two SARS-CoV-2 strains; Figure 1). One solution to correct such errors is to use multiple outgroups in the analysis (Barriel and Tassy, 1998). Hence, in our simulations, we further incorporated another virus that resembles GD-Pangolin-CoV (Supplementary Methods and Figure S2 in Supporting Information). Using the two outgroups, our simulations revealed the accuracy rate for ancestral inference was 97.42% (95% CI, 97.40%–97.44%) for a variant site, specifically, 92.72% (95% CI, 92.67%–92.78%) at a synonymous site and 99.35% (95% CI, 99.34%–99.36%) at a nonsynonymous site (Figure S3A in Supporting Information). Moreover, the use of two outgroups in the ancestral inference had an uncertainty rate of 0.78% (95% CI, 0.77%–0.79%) (Figure S3B in Supporting Information), which was much lower than that obtained using one outgroup (1.94%).

In summary, our simulations suggest that, using RaTG13 alone as the outgroup, the accuracy of inferring ancestral vs. derived mutations for a variant site in SARS-CoV-2 was 95.98%. Further, the use of both RaTG13 and GD-Pangolin-CoV as outgroups further increased the accuracy and sensitivity of ancestral inference. Animal coronaviruses might have undergone frequent recombination during evolution (Zhang and Holmes, 2020). Moreover, different regions of the viral genomes might differ in mutational rates. Both factors potentially cause heterogeneity in sequence divergence between SARS-CoV-2 and the outgroups. Since the accuracy of ancestral inference is mainly affected by the divergence between the ingroup and outgroup (Figure 1D), we reason that the ancestral inference will be more accurate in regions that have higher sequence similarities. In our simulations, we only focused on CDS, which accounts for ~98% of the SARS-CoV-2 genome. We expect the accuracy of ancestral inference in the non-coding regions to be similar to that in the synonymous sites with the assumption that both categories of sites are evolving neutrally or under a similar level of selective constraints. Of note, despite its wide usage, the MP method is prone to error in ancestral inference, especially when the divergence of the outgroup to the ingroup is high (Hernandez et al., 2007; Keightley and Jackson, 2018). The maximum likelihood (ML) method potentially improves the performance of ancestral inference, but it requires two or more outgroups in the analysis (Keightley and Jackson, 2018). Thus, it would be interesting to investigate how well the ML method can improve the accuracy of the ancestral inference when more coronaviruses that are suitable to be used as outgroups of SARS-CoV-2 are identified in the future.

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *We would like to thank Drs. Chung-I Wu, Yaping Zhang, and Jindong Zhao for suggestive comments regarding this study. This work was supported by grants from the National Natural Science Foundation of China (U1902201) and the CAS Light of West China Program to X.L, and from the National Natural Science Foundation of China (91731301) to J.L.*

References

Barriel, V., and Tassy, P. (1998). Rooting with multiple outgroups: consensus versus parsimony. Cladistics 14, 193–200.

Forster, P., Forster, L., Renfrew, C., and Forster, M. (2020). Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci USA 117, 9241–9243.

Hernandez, R.D., Williamson, S.H., and Bustamante, C.D. (2007). Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol Biol Evol 24, 1792–1800.

Keightley, P.D., and Jackson, B.C. (2018). Inferring the probability of the derived versus the ancestral allelic state at a polymorphic site. Genetics 209, 897–906.

Lam, T.T.Y., Jia, N., Zhang, Y.W., Shum, M.H.H., Jiang, J.F., Zhu, H.C., Tong, Y.G., Shi, Y.X., Ni, X.B., Liao, Y.S., et al. (2020). Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature https://doi.org/10.1038/s41586-020-2169-0.

Liu, P., Chen, W., and Chen, J.P. (2019). Viral metagenomics revealed Sendai virus and coronavirus infection of Malayan pangolins (*Manis javanica*). Viruses 11, 979.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., et al. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 395, 565–574.

Paraskevis, D., Kostaki, E.G., Magiorkinis, G., Panayiotakopoulos, G., Sourvinos, G., and Tsiodras, S. (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. Infect Genet Evol 79, 104212.

Ren, L.L., Wang, Y.M., Wu, Z.Q., Xiang, Z.C., Guo, L., Xu, T., Jiang, Y.Z., Xiong, Y., Li, Y.J., Li, X.W., et al. (2020). Identification of a novel coronavirus causing severe pneumonia in human. Chin Med J 133, 1015–1024.

Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., Duan, Y., Zhang, H., Wang, Y., Qian, Z., et al. (2020). On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev 7, 1012–1023.

Wang, H., Pipes, L., and Nielsen, R. (2020). Synonymous mutations and the molecular evolution of SARS-Cov-2 origins. bioRxiv, https://doi.org/10.1101/2020.04.20.052019.

Yu, W.B., Tang, G.D., Zhang, L., and T. Corlett, R. (2020). Decoding the evolution and transmissions of the novel pneumonia coronavirus

(SARS-CoV-2 / HCoV-19) using whole genomic data. Zool Res 41, 247–257.
Zhang, Y.Z., and Holmes, E.C. (2020). A genomic perspective on the origin and emergence of SARS-CoV-2. Cell 181, 223–227.
Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270–273.

SUPPORTING INFORMATION

The supporting information is available online at https://doi.org/10.1007/s11427-020-1764-2. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.