The Mode of The Novel Coronavirus 2019 Evolution

Kevin Surya, Jacob Gardner, and Chris Organ

The Coronavirus Disease 2019 (COVID-19), caused by the *Severe acute respiratory syndrome-related coronavirus 2* (SARS-CoV-2)[1], led to a pandemic that originated in Wuhan, China[2,3] in 2019 and spread to all continents except Antarctica by early 2020. The most common clinical symptoms of this disease are fever and cough[4]. Critically ill patients also tend to experience shortness of breath[5]. As of May 2020, SARS-CoV-2 has infected ~5 million people and killed ~300,000 worldwide[6].

SARS-CoV-2 has mutated relatively slowly[7], which is crucial for drug and vaccine development[8]. The effectiveness of COVID-19 drugs and vaccines[9,10] will benefit from minimal changes in the SARS-CoV-2 genome or in individual genes. However, shifts in the mutation rate will dictate how effective these drugs and vaccines are in the future. It is, therefore, essential to understand the mode of evolution[11]—how the mutation rate changes through time and with respect to the splitting (transmission) of viral lineages. Inferring the evolutionary mode among the broader group of SARS-like betacoronaviruses is also necessary for predicting whether COVID-19 drugs and vaccines will be useful for future novel coronavirus outbreaks. To help predict future drug and vaccine effectiveness, we here report on the mode of SARS-CoV-2 and SARS-like betacoronavirus evolution.

There are three possible modes of SARS-CoV-2 evolution. First, SARS-CoV-2 accumulates mutations steadily, consistent with a strict molecular clock[12] (null: gradual evolution). Second, the SARS-CoV-2 mutation rate jumps during transmission events, as the viruses infect new hosts (alternative 1: punctuated evolution[13]). Viral transmissions will only involve a subset of the host's virus population[14]. This scenario is similar to Mayr's founder-effect model of speciation[15], where the small transmitted subpopulation is subject to genetic drift, and therefore, rapid evolution. Third, mutations accumulate faster in SARS-CoV-2 lineages that tend to stay within the same host for a prolonged period than in lineages that frequently diversify (alternative 2: Red Queen-like[16]). Perhaps, the coevolutionary arms race between the viruses and host immune system (biotic interaction) drives evolution more than host-switching (abiotic change).

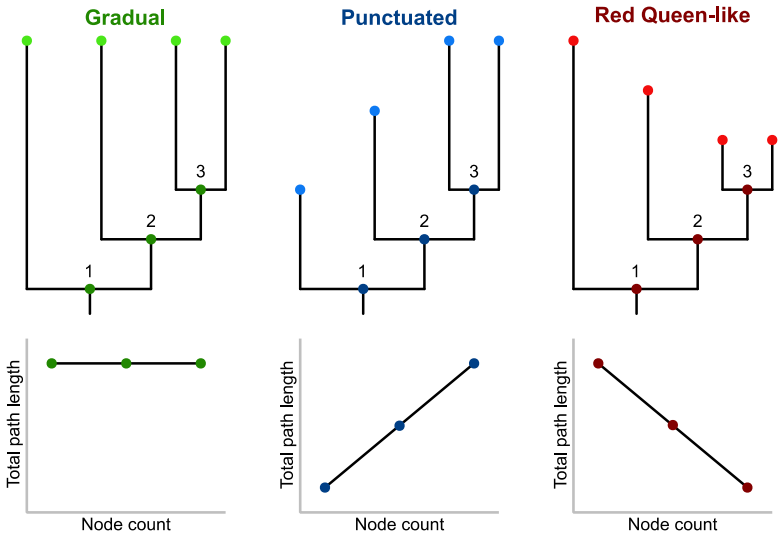

Fig. 1. Expectations regarding the three possible modes of SARS-CoV-2 evolution. The relationship between the total path length (tree root-to-tip distance) and node count (net transmission events) can be represented by a flat, positive, or negative slope. The slopes correspond to the gradual, punctuated, and Red Queen-like scenarios, respectively. Figure modified from Pagel *et al*[18].

To distinguish between the three modes of evolution, we analyzed their signatures in molecular phylogenies (Fig. 1) by regressing the total phylogenetic path length of the SARS-CoV-2 genomes (tree root-to-tip distance) on the net number of transmission events (node count)[17,18]. We acquired a molecular tree of 4,645 SARS-CoV-2 genomes on 6 May 2020, from NEXTSTRAIN[19,20] (Fig. A1), which used sequences kindly shared by labs and researchers at GISAID[21]. For this build, the NEXTSTRAIN team randomly sampled 120 genomes per administrative division (state or county or region or other admin sub-division) per month to achieve an equitable global sequence distribution. We used a restricted maximum likelihood (REML) algorithm under a phylogenetic generalized least squares (PGLS) framework to estimate the likelihood and estimates of the regression model above. The PGLS regression was performed in R[22] with the packages APE[23], PHYTOOLS[24], and NLME[25]. We set Pagel's λ, a measure of phylogenetic signal[26], to 1.

Across the entire tree, we found little evidence for punctuated or Red Queen-like genomic evolution (β = -0.0000000000019 ± 0.000000035, P = 0.99; R^2 = -2.67; Fig. 2). The model with node count is less likely than a mean-only model (ΔBIC = 40.93; Table A1). Two REML runs converged on the same likelihood and estimates. Regression diagnostics do not indicate any severe violations of the normality and equal variance assumptions (Fig. A2). The node-density artifact[17,27], an underestimation of branch lengths in tree regions with fewer taxa, does not seem to be present (δ = -0.046; Fig. A3). This analysis supports a gradual genomic evolution in SARS-CoV-2.

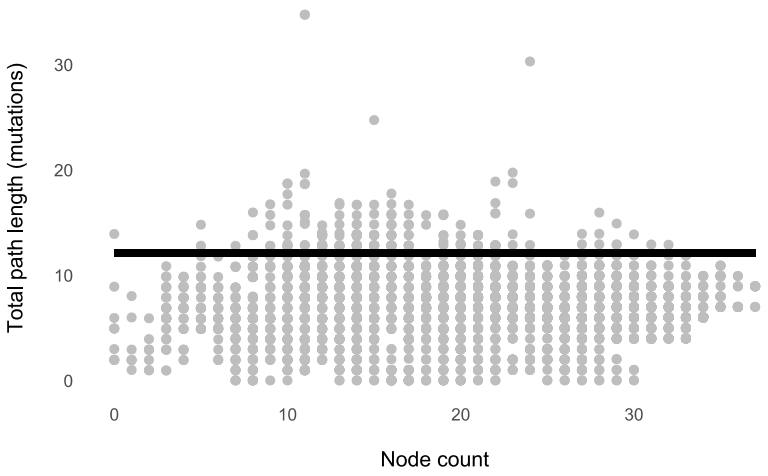

Fig. 2. The node count (net transmission events) along the SARS-CoV-2 lineages does not correlate with (β = 0), nor does it explain (R^2 = -2.67) the variation in total path lengths (accumulated mutations in the genome). PGLS equation: y = 12.05 + 0.00x + ε. The R^2 of a model is negative if the fit is worse than a mean-only model.

Evolution is a heterogeneous process. We therefore, investigated whether different parts of the tree exhibited different modes of evolution (i.e., Simpson's Paradox[28]). Such heterogeneity might be driven by immune selection and spatial dynamics, two variables that shape pathogen phylogenies[29]. Both factors can then be combined into one variable, which is the human population size on the continent scale. Natural selection overpowers the effects of genetic drift in large populations, including selection acting on the immune system[30]. We collected population size estimates from the United Nations World Population Prospects[31]. To test for the heterogeneity in mode, we fitted a regression model with continent-level population size as the second predictor, in addition to node count, plus an interaction term. This interaction allows the slope and the intercept to vary across continents. There is, however, little evidence for an effect of continent-level population size (Fig. A4) on the mode of evolution; the mean-only model still fits best (ΔBIC = 210.71; Table A1).

We then analyzed the mode of evolution of the broader SARS-like betacoronaviruses (SARS-like CoVs). How unexpected is gradual evolution among a larger taxonomic scale? Is gradualism specific to pandemic-related viruses (SARS-CoV-1 and SARS-CoV-2)? We acquired a molecular tree of 52 SARS-like CoV genomes from NEXTSTRAIN (Fig. A5) and fitted a regression model allowing the slope to vary by virus type (SARS-CoV-1, SARS-CoV-2, and SARS-like CoV). We found that the best-fitting model was the one with a single fit line (ΔBIC = 21.13; Table A2), suggesting a similar mode of evolution among the SARS-like CoVs. Moreover, there was little evidence for punctuated or Red Queen-like evolution (β = 0.000022 ± 0.000014, P = 0.115; R^2 = -0.02; Fig. 3). Gradualism is likely ubiquitous among SARS-like CoVs. Diagnostics indicate some violations of linear regression assumptions (Fig. A6). The node-density artifact was present (δ = 9.46), but it did not bias our analysis because we did not detect punctuated evolution (Fig. A7). We further found, using phylogenetic predictions[32], that seven SARS-like CoV genomes (blue data points above the fit line in Figure 3) are outliers (Fig. A8). However, removing them did not change our result as the model with a single fit line was still the best-fitting one (Table A3; Fig. A9). A potentially more serious bias is the undersampling of non-human SARS-like CoVs[33], but researchers are beginning to fill in this gap[34,35].

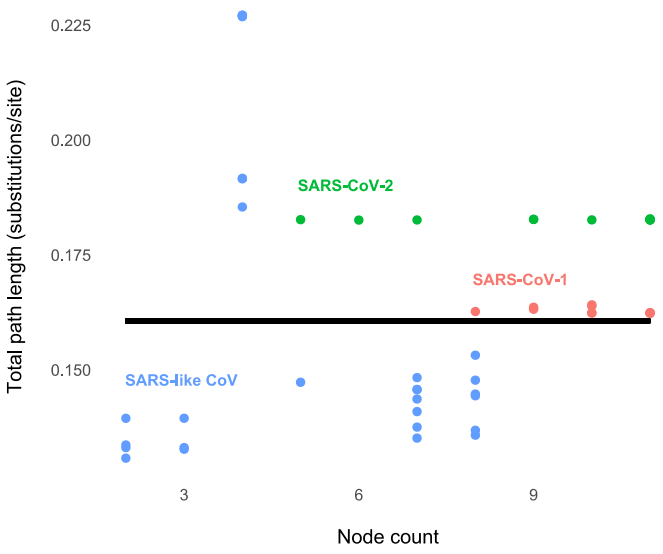

Fig. 3. SARS-like betacoronaviruses likely evolved gradually before and during the two pandemics (β = 0; R^2 = -0.02). Two regression fit lines (SARS-CoVs and SARS-like CoV) and three (SARS-CoV-1, SARS-CoV-2, and SARS-like CoV) did not increase the likelihood enough relative to the model with only one line ($\Delta BIC_{2\text{-line}}$ = 21.13; $\Delta BIC_{3\text{-line}}$ = 24.23). PGLS Equation: y = 0.161 + 0.000022x + ε.

Our findings suggest that SARS-CoV-2 genomes have been mutating gradually, with most mutations occurring in between net transmission events. This mode of evolution was likely the norm for the broader SARS-like betacoronaviruses as well. We, therefore, expect that COVID-19 drugs and vaccines under development to still be effective in the future. Given how the evolutionary epidemiological processes of SARS-CoV-2 happens on an ecological or population genetics scale[29,36], we can observe its broad-scale evolution in real-time and utilize phylogenetic comparative methods to help predict and eradicate COVID-19.

References

1. Gorbalenya, A. E. *et al.* The species *Severe acute respiratory syndrome-related coronavirus 2*: Classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* **5**, 536–544 (2020).
2. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
3. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
4. Guan, W. *et al.* Clinical characteristics of Coronavirus Disease 2019 in China. *N. Engl. J. Med.* **382**, 1708–1720 (2020).
5. Arentz, M. *et al.* Characteristics and outcomes of 21 critically ill patients with COVID-19 in Washington State. *JAMA* **323**, 1612–1614 (2020).
6. Johns Hopkins University. Coronavirus Research Center. https://coronavirus.jhu.edu/.
7. Korber, B. *et al.* Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv* 2020.04.29.069054 (2020) doi:10.1101/2020.04.29.069054.
8. Lash, N. & Schlossberg, T. The Coronavirus Is Mutating. What Does That Mean for a Vaccine? *The New York Times* (2020).
9. Gao, Q. *et al.* Development of an inactivated vaccine candidate for SARS-CoV-2. *Science* (2020) doi:10.1126/science.abc1932.
10. Sheahan, T. P. *et al.* An orally bioavailable broad-spectrum antiviral inhibits SARS-CoV-2 in human airway epithelial cell cultures and multiple coronaviruses in mice. *Sci. Transl. Med.* **12**, (2020).
11. Simpson, G. G. *Tempo and Mode in Evolution*. (Columbia University Press, 1945).
12. Zuckerkandl, E. & Pauling, L. Evolutionary divergence and convergence in proteins. in *Evolving Genes and Proteins* (eds. Bryson, V. & Vogel, H. J.) 97–166 (Academic Press, 1965). doi:10.1016/B978-1-4832-2734-4.50017-6.
13. Eldredge, N. & Gould, S. J. Punctuated equilibria: An alternative to phyletic gradualism. in *Models in Paleobiology* (ed. Schopf, T. J. M.) 82–115 (Freeman, Cooper, 1972).
14. Bergstrom, C. T., McElhany, P. & Real, L. A. Transmission bottlenecks as determinants of virulence in rapidly evolving pathogens. *Proc. Natl. Acad. Sci.* **96**, 5095–5100 (1999).
15. Mayr, E. Change of genetic environment and evolution. in *Evolution as a Process* (eds. Huxley, J., Hardy, A. C. & Ford, E. B.) 157–180 (Allan & Unwin, 1954).
16. Van Valen, L. A new evolutionary law. *Evol. Theory* **1**, 1–30 (1973).
17. Webster, A. J., Payne, R. J. H. & Pagel, M. Molecular phylogenies link rates of evolution and speciation. *Science* **301**, 478–478 (2003).
18. Pagel, M., Venditti, C. & Meade, A. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. *Science* **314**, 119–121 (2006).
19. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, (2017).
20. Hadfield, J. *et al.* Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
21. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, (2017).
22. R Core Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, 2019).
23. Paradis, E. & Schliep, K. ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
24. Revell, L. J. phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).
25. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team. *nlme: Linear and nonlinear mixed effects models*. (R package, 2019).

26. Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884 (1999).
27. Venditti, C., Meade, A. & Pagel, M. Detecting the node-density artifact in phylogeny reconstruction. *Syst. Biol.* **55**, 637–643 (2006).
28. Blyth, C. R. On Simpson's paradox and the sure-thing principle. *J. Am. Stat. Assoc.* **67**, 364–366 (1972).
29. Grenfell, B. T. *et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332 (2004).
30. Lanfear, R., Kokko, H. & Eyre-Walker, A. Population size and the rate of evolution. *Trends Ecol. Evol.* **29**, 33–41 (2014).
31. United Nations Department of Economic and Social Affairs. World Population Prospects. https://population.un.org/wpp/Download/Standard/Population/ (2020).
32. Organ, C., Nunn, C. L., Machanda, Z. & Wrangham, R. W. Phylogenetic rate shifts in feeding time during the evolution of *Homo*. *Proc. Natl. Acad. Sci.* **108**, 14555–14559 (2011).
33. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nat. Med.* **26**, 450–452 (2020).
34. Zhou, H. *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. *Curr. Biol.* (2020) doi:10.1016/j.cub.2020.05.023.
35. Joffrin, L. *et al.* Bat coronavirus phylogeography in the Western Indian Ocean. *Sci. Rep.* **10**, 6873 (2020).
36. Pybus, O. G. & Rambaut, A. Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* **10**, 540–550 (2009).

Appendix

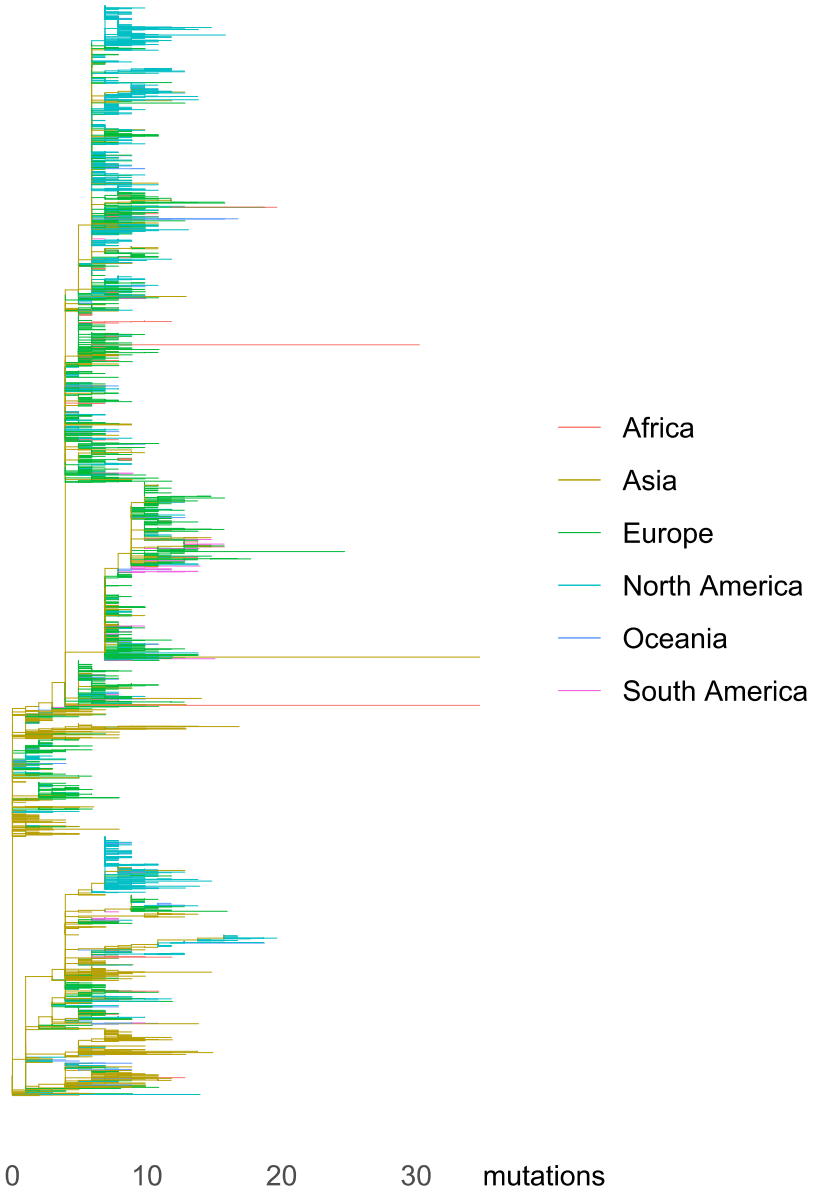

Fig. A1. The molecular tree of 4,645 SARS-CoV-2 genomes, downloaded from NEXTSTRAIN on 6 May 2020, shows that every continent listed above had multiple introductions of the virus. While this build has an equitable global sequence distribution, it also undersamples regions that are doing a lot of sequencing. The tree was rooted relative to early samples from Wuhan, China. We did not infer the continent assignment of the internal branches using a formal ancestral state reconstruction method. Geographic distributions of the genomes: Africa (*n* = 121; 2.60%), Asia (*n* = 913; 19.66%), Europe (*n* = 1,992; 42.89%), North America (*n* = 1,311; 28.22%), Oceania (*n* = 190; 4.09%), and South America (*n* = 118; 2.54%). We plotted the tree using the R packages GGTREE[37,38] and GGIMAGE[39].

Model	Pagel's λ	BIC	ΔBIC
$y = \beta_0 + \varepsilon$	0	23,481.92	58,553.21
$y = \beta_0 + \varepsilon$	1	-35,071.29	0.00
$y = \beta_0 + \beta_1 x_1 + \varepsilon$	0	23,486.07	58,557.36
$y = \beta_0 + \beta_1 x_1 + \varepsilon$	1	-35,030.36	40.93
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$	0	23,142.71	58,214.00
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$	1	-34,947.61	123.68
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$	0	23,156.59	58,227.88
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$	1	-34,860.58	210.71

Table A1. The most likely regression model, accounting for the number of free parameters (i.e., model complexity), is the mean-only model with Pagel's $\lambda = 1$ (the model with the lowest Bayesian Information Criterion [BIC][40] score). y: total path length. x_1: node count. x_2: continent-level human population size.

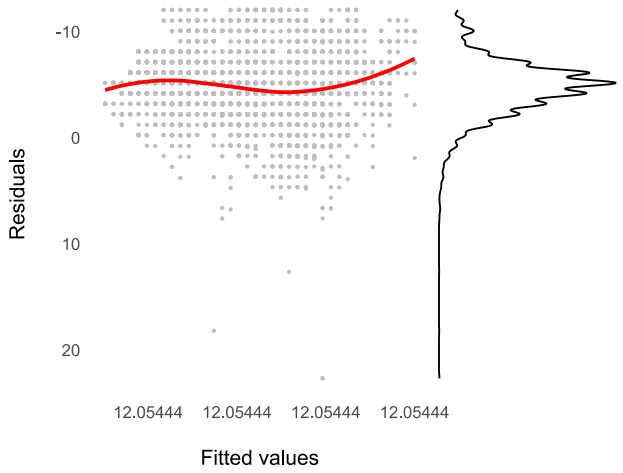

Fig. A2. Diagnostics indicate violations of the normality assumption, but not of the equal variance assumption. The distribution of the residuals is right-skewed (Shapiro-Wilk $W = 0.98$, $P < 0.0001$). This non-normality, however, is caused by three data points near the bottom of the plot, which are unlikely to be influential with respect to the slope. There is no evidence of unequal variance as the variability of the residuals is roughly constant across the fitted values. We created the plot above using the R packages CAIRO[41], GGEXTRA[42], GGPLOT2[43], GGTHEMES[44], and SVGLITE[45].

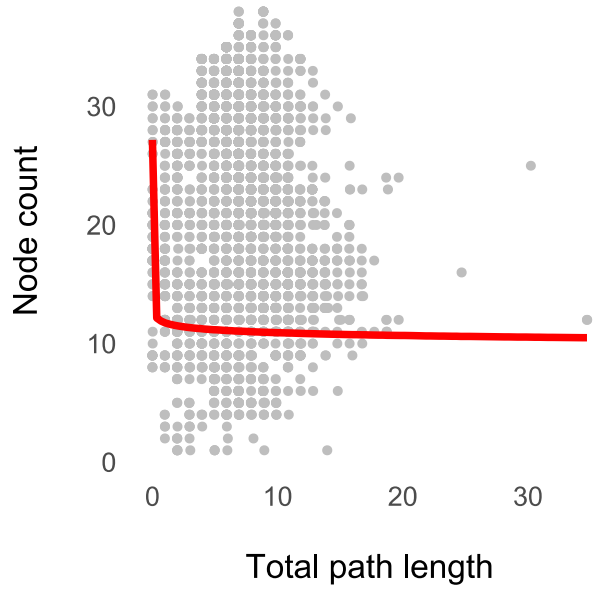

Fig. A3. The node-density artifact, which can bias our analysis, does not seem to be present ($\delta = -0.032$). A symptom of this artifact, in this case, is a positive curvilinear relationship ($\delta > 1$). PGLS equation: $y = 11.75x^{-0.032} + \varepsilon$.

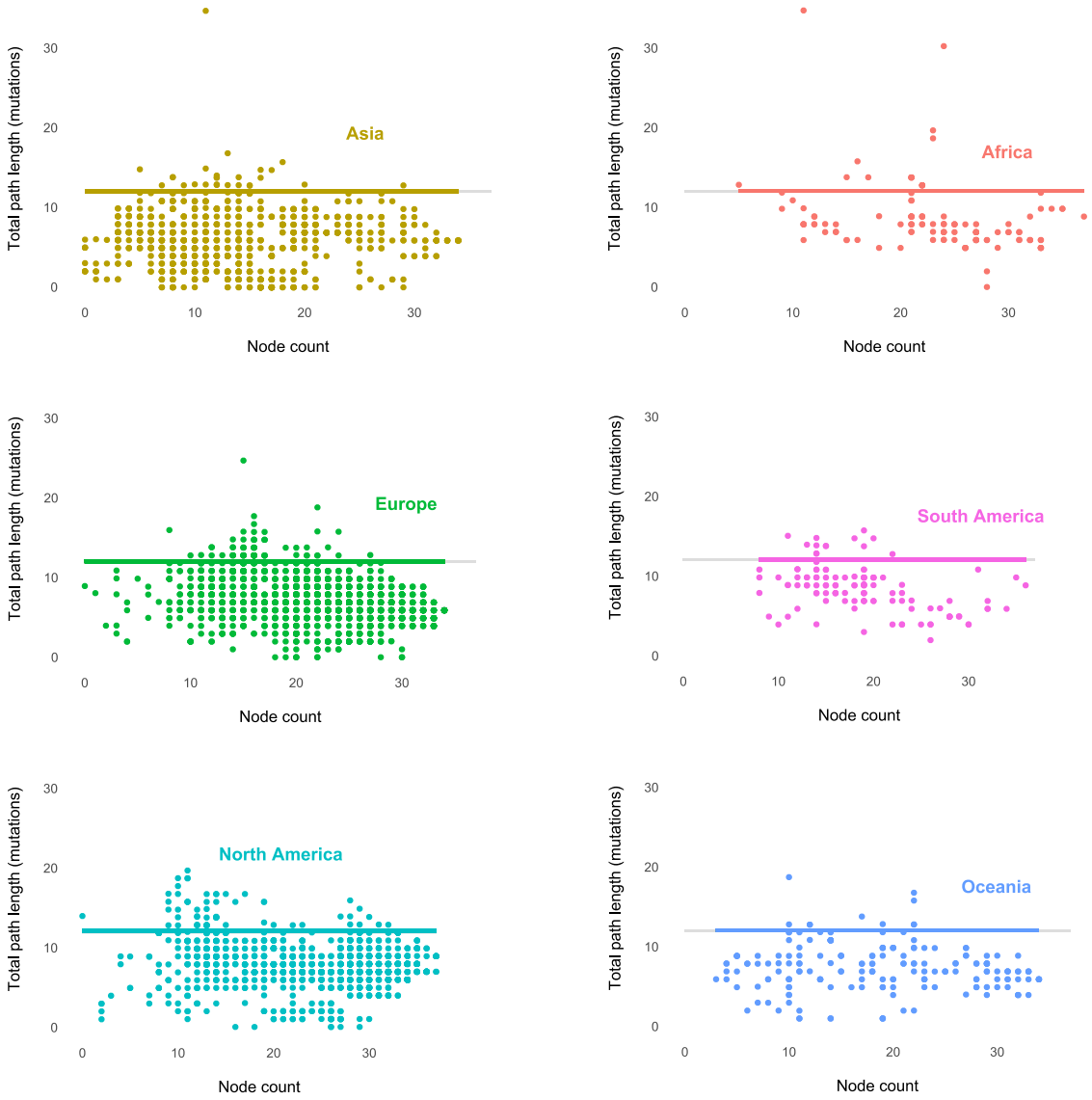

Fig. A4. The mode of SARS-CoV-2 evolution did not vary by continents. The node count (net transmission events) does not correlate with the total path length (accumulated mutations in the genome). The gray fit lines are from the single-line regression model. We order the continents according to their population sizes (descending). We created the scatter plot using CAIRO, GGPLOT2, GGTHEMES, and SVGLITE.

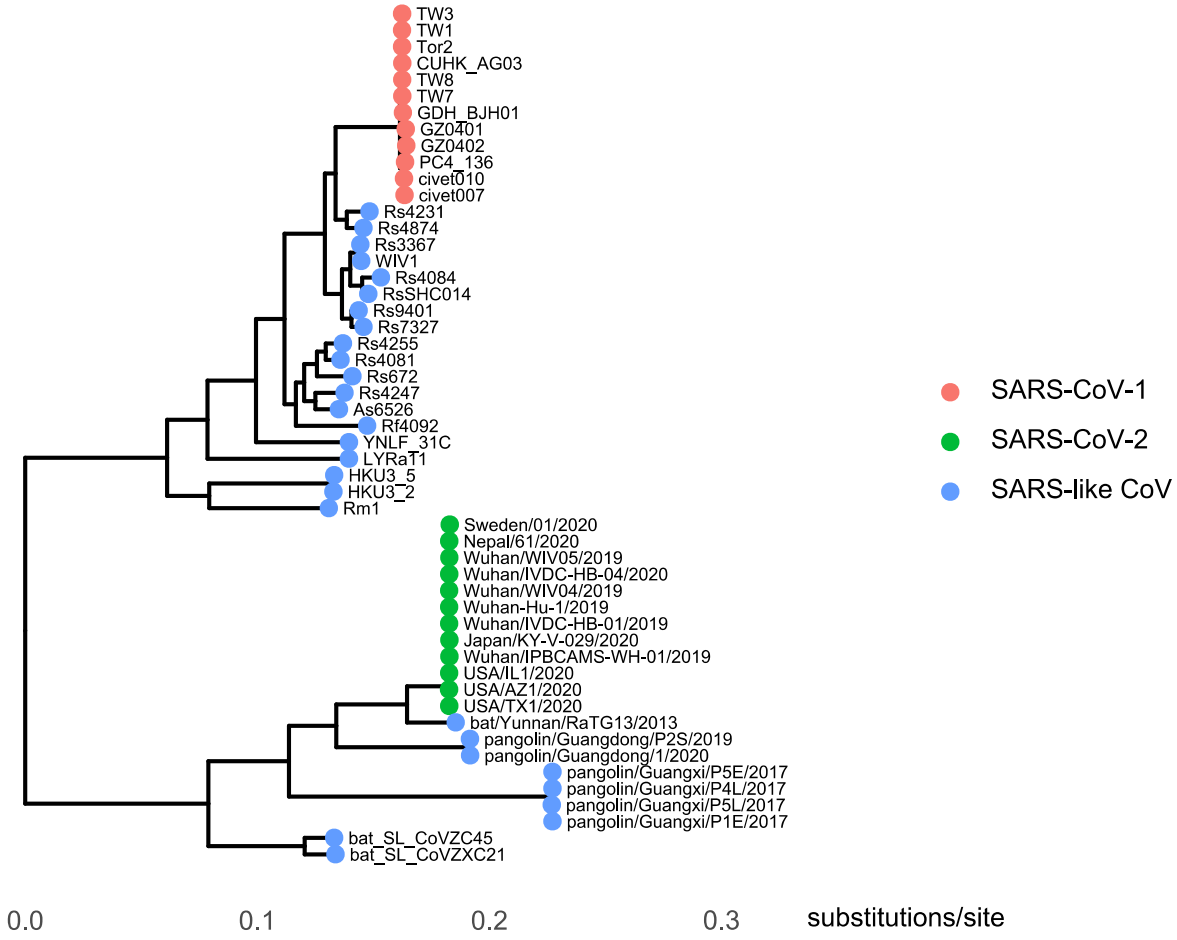

Fig. A5. Molecular phylogeny of SARS-like betacoronaviruses downloaded from NEXTSTRAIN. This tree contains 52 genomes (12 SARS-CoV-1, 12 SARS-CoV-2, and 28 SARS-like CoV).

Model	BIC	ΔBIC
$y = \beta_0 + \beta_1 x + \varepsilon$	-504.47	0.00
$y = \beta_0 + \beta_1 x + \beta_2 I_{SARS\text{-}like\ CoV} + \beta_3 x . I_{SARS\text{-}like\ CoV} + \varepsilon$	-483.34	21.13
$y = \beta_0 + \beta_1 x + \beta_2 I_{SARS\text{-}CoV\text{-}2} + \beta_3 I_{SARS\text{-}like\ CoV} + \beta_4 x . I_{SARS\text{-}like\ CoV} + \varepsilon$	-480.24	24.23

Table A2. The most likely regression model, according to the BIC model selection, is one where the mode of evolution did not vary between SARS-CoV-1, SARS-CoV-2, and SARS-like CoV genomes. For the most complex model (third), we did not allow the slope to vary between SARS-CoV-1 and SARS-CoV-2. For all models, Pagel's $\lambda = 1$. *y:* total path length. *x*: node count. $I_{SARS\text{-}like\ CoV}$*:* Indicator variable for SARS-like CoV (0 = no; 1 = yes). $I_{SARS\text{-}CoV\text{-}2}$*:* Indicator variable for SARS-CoV-2.

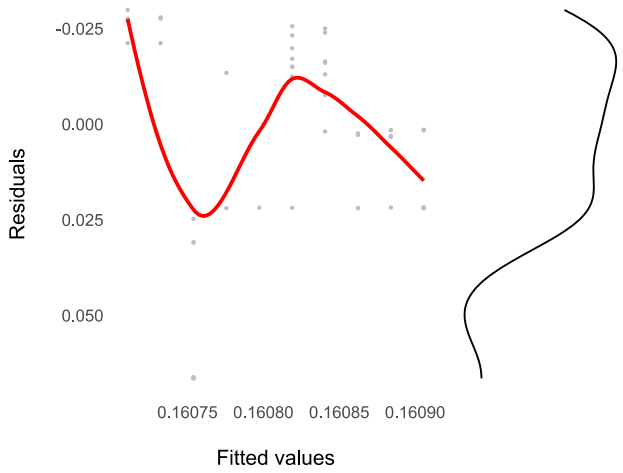

Fig. A6. Regression diagnostics indicate assumption violations. The residuals vs. fitted values plot shows that residuals are not constant across the fitted values (i.e., unequal variance). Also, the distribution of the residuals is right-skewed. Log-transformation did not significantly change the distributions of the response (total path length) and predictor (node count).

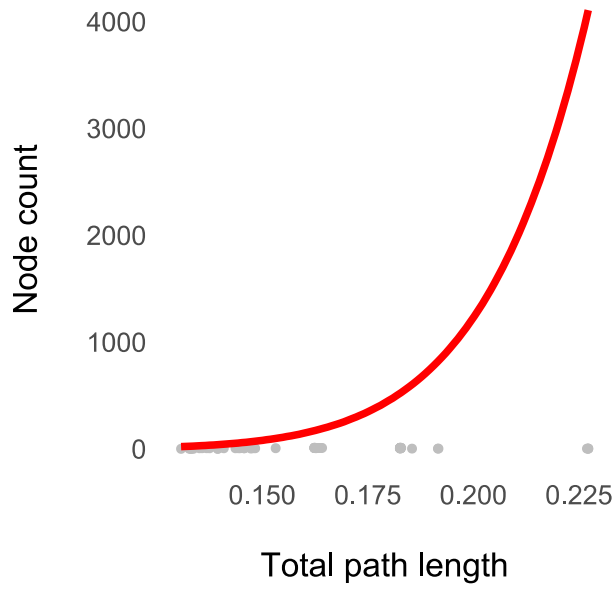

Fig. A7. The node-density artifact is obvious (δ = 9.46). But, it did not bias our analysis because we did not find evidence for punctuation. PGLS equation: $y = 5{,}039{,}309{,}975x^{9.46}$.

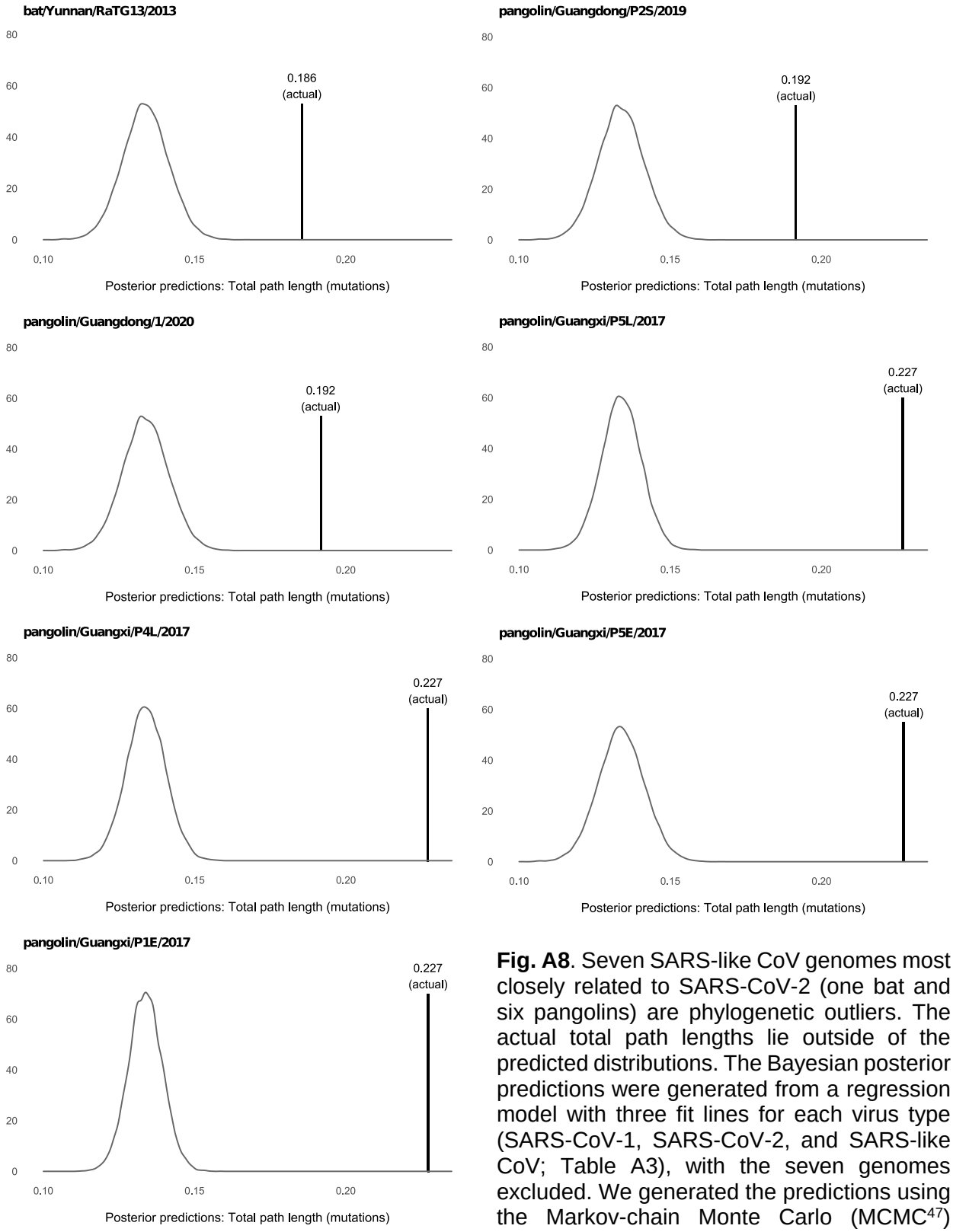

Fig. A8. Seven SARS-like CoV genomes most closely related to SARS-CoV-2 (one bat and six pangolins) are phylogenetic outliers. The actual total path lengths lie outside of the predicted distributions. The Bayesian posterior predictions were generated from a regression model with three fit lines for each virus type (SARS-CoV-1, SARS-CoV-2, and SARS-like CoV; Table A3), with the seven genomes excluded. We generated the predictions using the Markov-chain Monte Carlo (MCMC[47]) algorithm in BAYESTRAITS[26].

Model	BIC	ΔBIC
$y = \beta_0 + \beta_1 x + \varepsilon$	-435.22	0.00
$y = \beta_0 + \beta_1 x + \beta_2 I_{SARS\text{-}like\ CoV} + \beta_3 x . I_{SARS\text{-}like\ CoV} + \varepsilon$	-425.59	9.64
$y = \beta_0 + \beta_1 x + \beta_2 I_{SARS\text{-}CoV\text{-}2} + \beta_3 I_{SARS\text{-}like\ CoV} + \beta_4 x . I_{SARS\text{-}like\ CoV} + \varepsilon$	-417.46	17.76

Table A3. Without the SARS-like CoV phylogenetic outliers (Fig. A8), the most likely regression model is still the one where the mode of evolution did not vary between SARS-CoV-1, SARS-CoV-2, and SARS-like CoV genomes. For the most complex model (third), we did not allow the slope to vary between SARS-CoV-1 and SARS-CoV-2. For all models, Pagel's $\lambda = 1$. *y:* total path length. *x*: node count. $I_{SARS\text{-}like\ CoV}$*:* Indicator variable for SARS-like CoV. $I_{SARS\text{-}CoV\text{-}2}$*:* Indicator variable for SARS-CoV-2. We used the Bayesian PGLS in BAYESTRAITS for model-fitting. The MCMC simulation ran for 11,000,000 iterations with 1,000,000 burn-ins, and a sampling period of 1,000. Marginal likelihoods were sampled using 100 stepping stones[48], one every 10,000 iterations. BIC = -2 × log marginal likelihood. All MCMC runs converged, checked using TRACER[49].

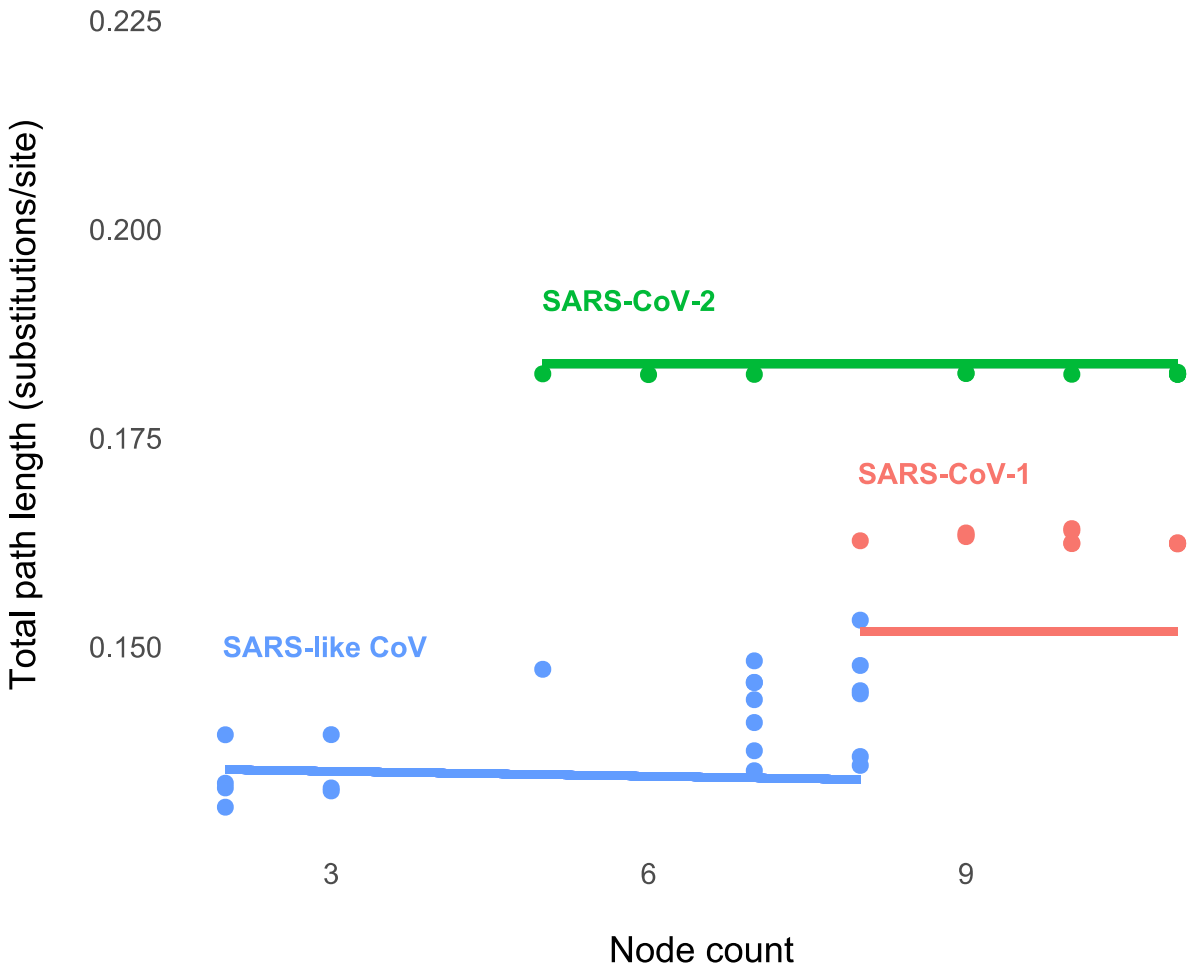

Fig. A9. Without the SARS-like CoV phylogenetic outliers (Fig. A8), the mode of evolution for SARS-like CoV genomes (blue) was most likely gradual ($\beta_{avg.}$ = -0.00020, pMCMC = 0.39). Note that this three-line regression model is less likely than the single-line one (Table A3).

Appendix References

37. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
38. Yu, G., Lam, T. T.-Y., Zhu, H. & Guan, Y. Two methods for mapping and visualizing associated data on phylogeny using ggtree. *Mol. Biol. Evol.* **35**, 3041–3043 (2018).
39. Yu, G. *ggimage: Use Image in 'ggplot2'*. (R package, 2020).
40. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
41. Urbanek, S. & Horner, J. *Cairo: R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output*. (R package, 2019).
42. Attali, D. & Baker, C. *ggExtra: Add marginal histograms to 'ggplot2', and more 'ggplot2' enhancements*. (R package, 2019).
44. Wickham, H. *ggplot2: Elegant graphics for data analysis*. (Springer-Verlag, 2009).
45. Arnold, J. B. *ggthemes: Extra themes, scales and geoms for 'ggplot2'*. (R package, 2019).
46. Wickham, H., Henry, L., Luciani, T. J., Decorde, M. & Lise, V. *svglite: An 'SVG' graphics device*. (R package, 2019).
47. Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. Introducing Markov chain Monte Carlo. in *Markov chain Monte Carlo in Practice* vol. 1 19 (Chapman and Hall, 1996).
48. Xie, W., Lewis, P. O., Fan, Y., Kuo, L. & Chen, M.-H. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* **60**, 150–160 (2011).
49. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).