Notes from the Field

Epidemiological Model Suggests D614G Spike Protein Mutation Accelerates Transmission of COVID-19 — Worldwide, 2020

Liang Wang[1]; Yuhai Bi[1,2,#]; George F. Gao[1,2,#]

The coronavirus disease 2019 (COVID-19) pandemic was caused by a novel type of coronavirus also known as SARS-CoV-2, 2019-nCoV, and HCoV-19 (*1*) and quickly spread around the world becoming a major threat to public health (*2–3*). Although COVID-19 virus is characterized as possessing a large genome and having limited genetic diversity (*4*) like other coronaviruses due to the high fidelity of its replication mechanism, many single nucleotide polymorphisms (SNPs) have been detected so far. Among the reported SNPs, the strain with the G614 mutation in the S protein has replaced the strain containing D614 as the world's primary pandemic strain (*5*). However, it was still unclear if COVID-19 virus strains with G614 were more transmissible than those with D614 (*6*).

We reconstructed several transmission chains of COVID-19 during the early epidemic phase for 3 countries (Australia, the UK, and USA) based on genomic data from GISAID (*7*) and Bayesian inference under an epidemiological model for strains with D614 and G614 due to the similar amount of genomes within each country (Figure 1A). Then we inferred the R_0 in those transmission chains to compare the difference of transmissibility among humans between D614 and G614 (see the Supplementary Material for details, available in http://weekly.chinacdc.cn/). The R_0 caused by G614 was significantly higher than that caused by D614 in Australia and USA, as there was no intersection of 95% confidence intervals (CI) of the 2 distributions of R_0 (Figure 1B). For the UK, the mean value of R_0 for G614 was slightly higher than D614. For G614, its mean value of R_0 was outside the 95% CI of the estimated R_0 of D614. The mean value of R_0 of D614 also was outside the 95% CI of the estimated R_0 of G614. In addition, the R_0 for D614 was similar among three countries (from 1.56 for USA to 1.73 for Australia). However, significantly different R_0 was detected among countries (from 1.82 for the UK to 3.87 for USA), indicating that the spatial transmission of G614 had higher variation than D614.

Since the data used in this study were all collected during the early phases of each countries' outbreak and no stringent prevention and control strategies were implemented in these three countries in those phases, the estimation of R_0 would not be affected by non-pharmacological intervention. Thus, the result would better reflect the true nature of transmissibility within humans for strains with different mutations. Our findings demonstrated that the G614 mutation accelerated the transmission of the COVID-19 virus and also had higher spatial transmissibility, indicating that strains with G614, which were the dominant strains around the world, could spread on a larger scale and be more difficult to control. These results also echo those experimental results *in vitro*, in which both clinical samples and pseudoviruses with G614 have higher levels of viral RNA and titers compared to those with D614 (*5*).

As the COVID-19 pandemic spreads and continues, real-time monitoring and evaluation of the impacts of COVID-19 virus strain mutations need to be consistently maintained to provide earlier warnings for the public and provide evidence that supports government-led countermeasures and strategies.

Acknowledgments: We gratefully acknowledge the authors from the originating laboratories and the submitting laboratories where genetic sequence data were generated and shared via GISAID, enabling this research; Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB29010202, XDB29010102 and XDA19090118); National Science and Technology Major Project (Grant No. 2018ZX10733403, and 2018ZX10713001-010); National Natural Science Foundation of China (NSFC) (Grant No. 32041010 and 31900155); NSFC Outstanding Young Scholars (Grant No. 31822055); and Youth Innovation Promotion Association of CAS (Grant No. 2017122).

doi: 10.46234/ccdcw2020.247

[#] Corresponding authors: Yuhai Bi, beeyh@im.ac.cn; George F. Gao, gaof@im.ac.cn.

[1] CAS Key Laboratory of Pathogenic Microbiology and Immunology,

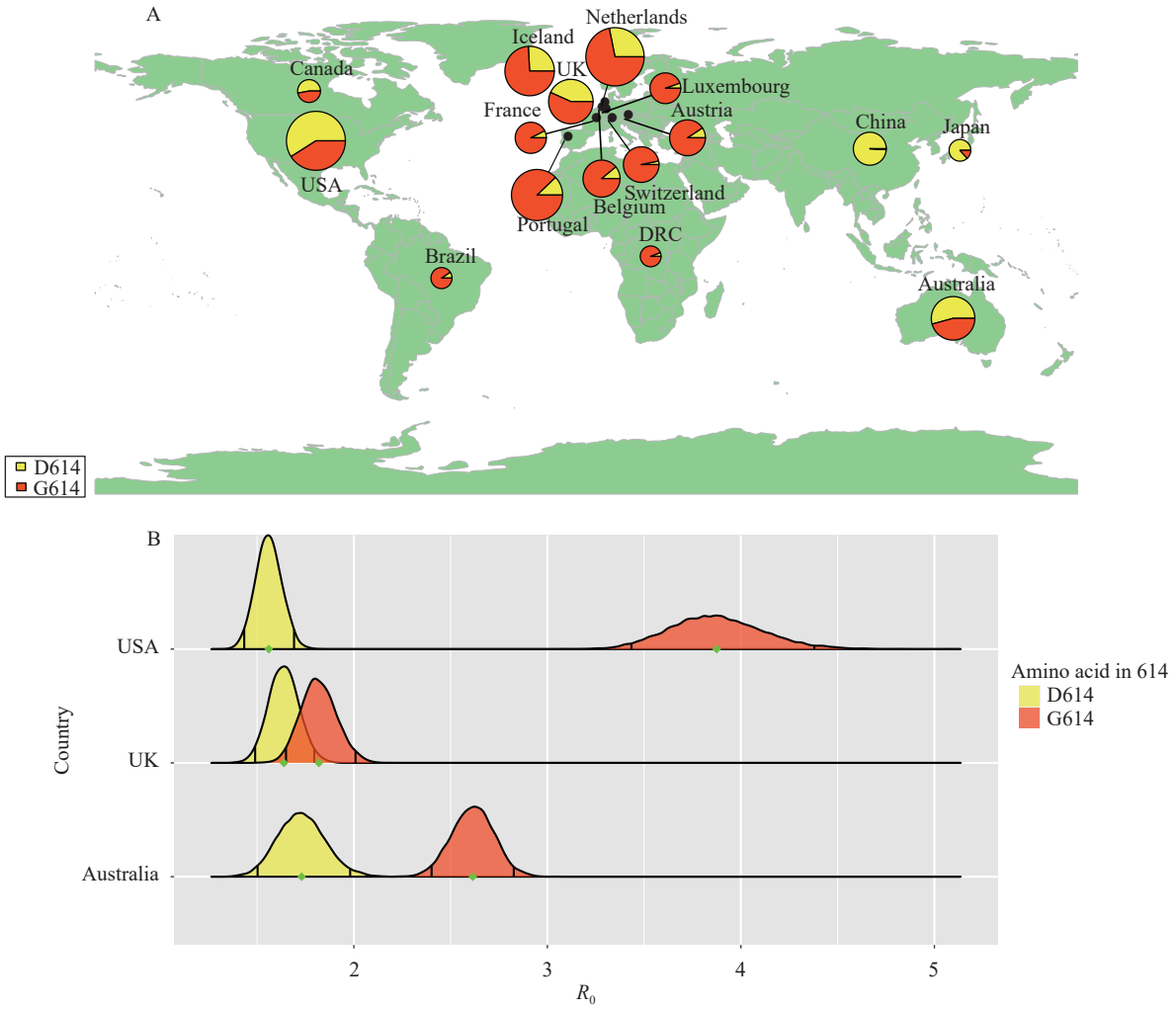


FIGURE 1. The statistics and transmissibility of the COVID-19 virus with D614 and G614 during the early phases of the pandemic within each country. (A) The number of publicly available genomes with D614 and G614 during the first two months after the first confirmed case in each country. The size of the pie chart is associated with the number of genomes. (B) The distribution of R_0 for D614 and G614 during the first two months after the first confirmed case in each country. The mean of R_0 is represented by a green point. The black line within the distribution represents the 95% confidence interval (CI).

Institute of Microbiology, Center for Influenza Research and Early-warning (CASCIRE), CAS-TWAS Center of Excellence for Emerging Infectious Diseases (CEEID), Chinese Academy of Sciences, Beijing, China; [2] University of Chinese Academy of Sciences, Beijing, China.

Submitted: November 21, 2020; Accepted: November 25, 2020

REFERENCES

1. Jiang SB, Shi ZL, Shu YL, Song JD, Gao GF, Tan WJ, et al. A distinct name is needed for the new coronavirus. Lancet 2020;395(10228):949. http://dx.doi.org/10.1016/S0140-6736(20)30419-0.
2. Tan WJ, Zhao X, Ma XJ, Wang WL, Niu PH, Xu WB, et al. A novel coronavirus genome identified in a cluster of pneumonia cases — Wuhan, China 2019–2020. China CDC Weekly 2020;2(4):61 – 2. http://dx.doi.org/10.46234/ccdcw2020.017.
3. Zhu N, Zhang DY, Wang WL, Li XW, Yang B, Song JD, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 2020;382(8):727 – 33. http://dx.doi.org/10.1056/NEJMoa2001017.
4. Fauver JR, Petrone ME, Hodcroft EB, Shioda K, Ehrlich HY, Watts AG, et al. Coast-to-coast spread of SARS-CoV-2 during the early epidemic in the United States. Cell 2020;181(5):990 – 6.e5. http://dx.doi.org/10.1016/j.cell.2020.04.021.
5. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell 2020;182(4):812 – 27.e19. http://dx.doi.org/10.1016/j.cell.2020.06.043.
6. Grubaugh ND, Hanage WP, Rasmussen AL. Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear. Cell 2020;182(4):794 – 5. http://dx.doi.org/10.1016/j.cell.2020.06.040.
7. Shu YL, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. EuroSurveillance 2017;22(13):pii=30494. http://dx.doi.org/10.2807/1560-7917.ES.2017.22.13.30494..

Supplementary Material

Data collection, classification, and preliminary process

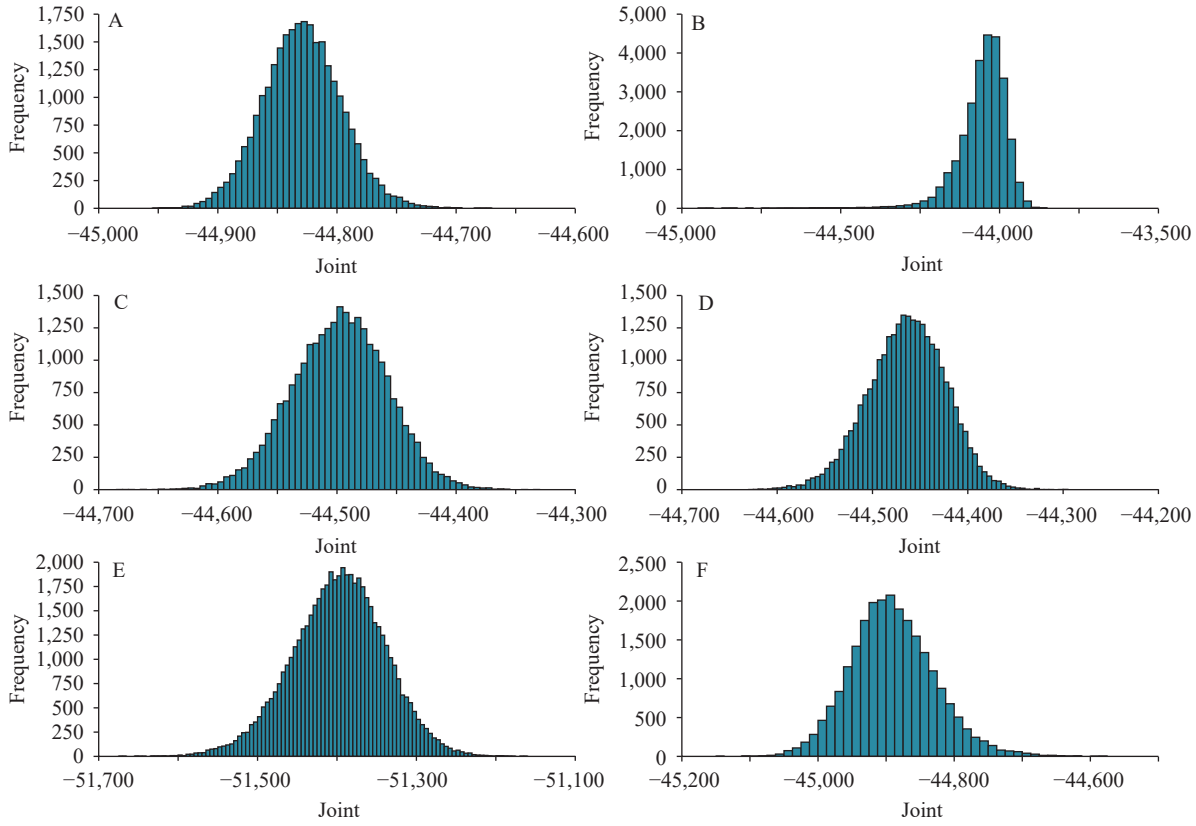
In order to minimize the impact of stringent prevention and control measures on the transmission of the COVID-19 virus and strike a balance between the small amount of variation among viral genomes during the early stages of the outbreak and sufficient variation to support this study, we defined the scope of the study to focus on the first 2 months after the start of COVID-19 outbreak within each country. All genomic sequences, their spike protein sequences, and collection dates of the COVID-19 virus were retrieved from GISAID on April 9, 2020. Since there were some countries [e.g. Austria, Brazil, the Democratic Republic of the Congo (DRC), Iceland, Luxembourg, Portugal, and Switzerland] in which the outbreak had not existed for two months before April 9, 2020, we collected additional data for these countries on June 1, 2020 (Table S1). Only complete genomic sequences with high coverage and exact collection dates (accurate to days) were used in this study.Only countries with more than 80 COVID-19 virus genomes were shown in Figure 1A. Genomic and corresponding spike protein sequences from each country were aligned using Mafft v7.310 (*1*). The genomic sequences were split into two datasets (D and G) based on the amino acid at 614 of the spike protein sequence (based on Wuhan reference sequence) for each country. To eliminate the potential impact on results due to different regions, we only compared the 2 datasets (G614 and D614) within each country. In addition, only countries where the difference in the number of genomes between the 2 datasets was less than 50% and both datasets having contained more than 100 genomes were used in the subsequent analysis to minimize the potential impact of the difference in the amount of data on results. In this case, datasets from Australia, the UK, and USA were used in the final analysis. We trimmed uncertain regions in 3' and 5' terminals and also masked 30 sites (Table S2) that are highly homoplastic and have no phylogenetic signal as previous noted (https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473). Finally, we obtained 197, 166, 164, 215, 393, and 273 genomic sequences, with aligned genomic length of 29,381, 29,381, 29,582, 29,582, 29,498, and 29,498 from Australia (D614), Australia (G614), the UK (D614), the UK (G614), USA (D614), and USA (G614), respectively. We used jModelTest v2.1.6 (*2*) to find the best substitution model for each dataset according to the Bayesian Information Criterion (Table S3). The list of genomic sequences used in this study were openly shared via the GISAID initiative (*3*) (see the Acknowledgement Table for details).

SUPPLEMENTARY TABLE S1. The time point for each country used in this study.

Country	Continent	Date of first confirmed case	The first two months
Australia	Oceania	2020-01-25	2020-03-25
Austria	Europe	2020-02-25	2020-04-25
Belgium	Europe	2020-02-04	2020-04-04
Brazil	South America	2020-02-25	2020-04-25
Canada	North America	2020-01-27	2020-03-27
China	Asia	2019-12-24	2020-02-24
DRC	Africa	2020-03-10	2020-05-10
France	Europe	2020-01-24	2020-03-28
Iceland	Europe	2020-02-28	2020-04-28
Japan	Asia	2020-01-14	2020-03-14
Luxembourg	Europe	2020-02-29	2020-04-29
Portugal	Europe	2020-03-02	2020-05-02
Switzerland	Europe	2020-02-25	2020-04-25
UK	Europe	2020-01-31	2020-03-31
USA	North America	2020-01-20	2020-03-20

Reconstruction of dated phylogenies

Since recombination could impact the evolutionary signal, we first detected the recombination events in these COVID-19 virus genomes by RDP4 (*4*). No evidence of recombination was found in any dataset. We then used the



SUPPLEMENTARY FIGURE S1. The posterior distributions of phylogenies in the posterior tree space for each dataset. (A) D614 dataset in Australia; (B) G614 dataset in Australia; (C) D614 dataset in UK; (D) G614 dataset in UK; (E) D614 dataset in USA; (F) G614 dataset in USA

Bayesian Markov Chain Monte Carlo (MCMC) approach implemented in Bayesian Evolutionary Analysis Sampling Trees (BEAST) v1.10.4 (*5*) to derive an accurate, dated phylogeny for COVID-19 under the best substitution model for each dataset. The result of model comparison was listed in Table S4. Analyses were performed with at least 3 independent replicates of 100 million MCMC steps each and sampling parameters and trees every 10,000 steps. The estimation of the most appropriate combination of molecular clock and coalescent models for Bayesian phylogenetic analysis was determined using both path-sampling (PS) and stepping-stone (SS) models (*6*). Tracer 1.7.1 (*7*) was then used to check the convergence of MCMC chains (effective sample size >200) and to compute marginal posterior distributions of parameters after discarding 10% of the MCMC chain as burn-in (Figure S1). TreeAnnotator was used to summarize a maximum clade credibility (MCC) tree from the posterior distribution of trees after discarding 10% of the MCMC chain as burn-in (Figure S2). We determined whether there was a sufficient temporal signal in each dataset as it was the prerequisite for getting a reliable inference when performing phylodynamic analysis. Bayesian evaluation of temporal signal (BETS) (*8*) was used to evaluate the temporal signal in each dataset. BETS relied on the comparison of marginal likelihoods of two models: the heterochronous (with tip date) and isochronous (without tip date) models. Analyses were performed with at least 3 independent replicates of 100 million MCMC steps each and sampling parameters and trees every 10,000 steps with the best substitution model and most appropriate combination of molecular clock and coalescent models determined above for each dataset. The marginal likelihoods were estimated by PS. The Bayes factor (BF) was then calculated based on the likelihoods of two models (heterochronous and isochronous). If the log BF>5 (heterochronous model against isochronous model), it indicated there were sufficient temporal signals in this dataset. The results of BETS for each country are listed in Table S5. All datasets had log BF>5, suggesting that the temporal signal was sufficiently strong.

Inferring the transmission chain and its parameters

Because viral genomes were incompletely sampled and the epidemic is still ongoing, TransPhylo v1.3.19 (*9*) was

SUPPLEMENTARY TABLE S2. List of 30 masked sites in the COVID-19 virus genome.

ID	Site
1	187
2	1,059
3	2,094
4	3,037
5	3,130
6	4,050
7	6,990
8	8,022
9	10,323
10	10,741
11	11,074
12	11,083
13	13,402
14	13,408
15	14,786
16	15,324
17	19,684
18	20,148
19	21,137
20	21,575
21	24,034
22	24,378
23	25,563
24	26,144
25	26,461
26	26,681
27	28,077
28	28,826
29	28,854
30	29,700

SUPPLEMENTARY TABLE S3. The best substitution model from each dataset.

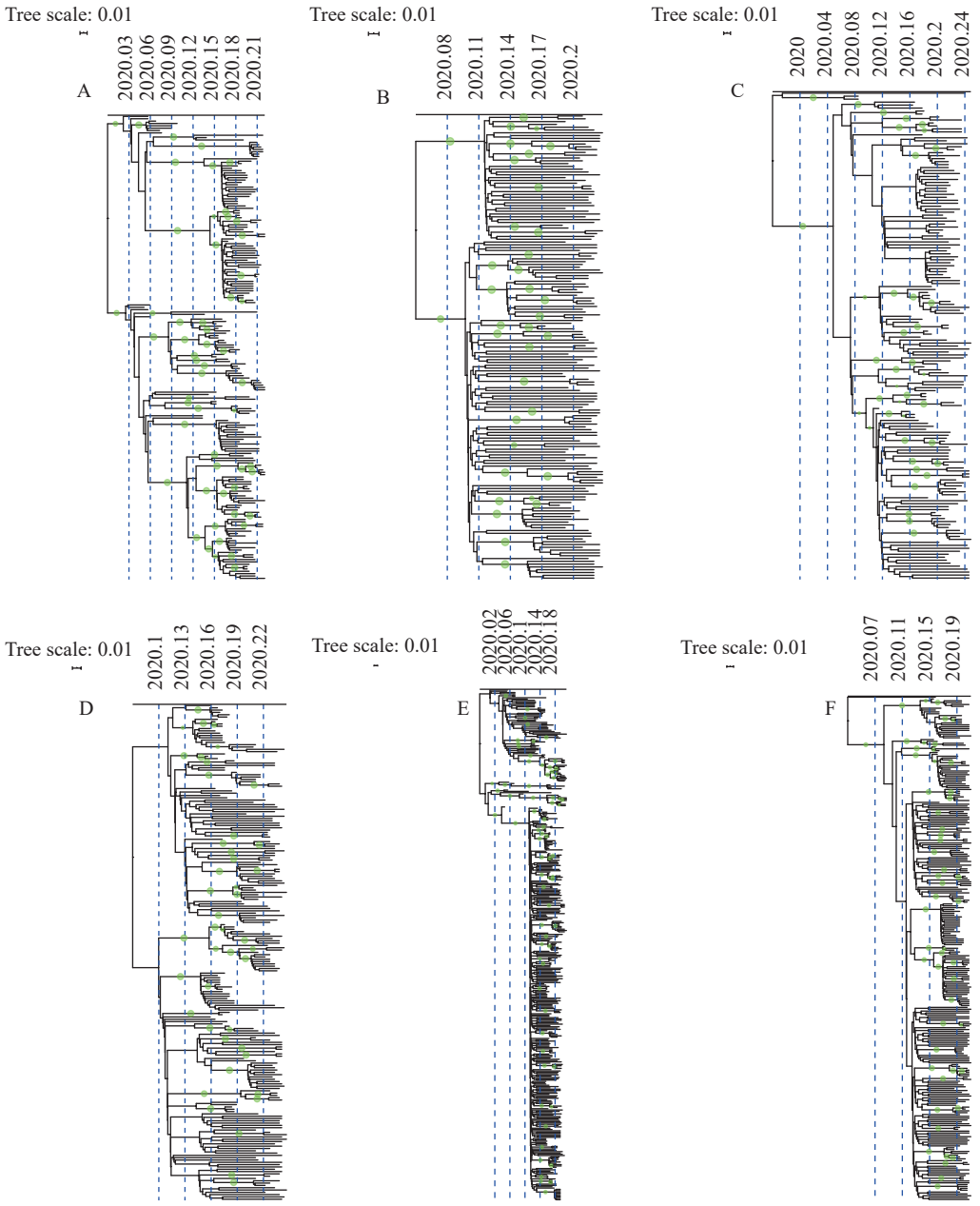
Country	Category	Best substitution model
Australia	D614	GTR+I
	G614	GTR
UK	D614	GTR
	G614	GTR
USA	D614	GTR
	G614	GTR

used to infer the transmission tree using the MCC dated phylogeny generated above as input, as a previous study demonstrated that using an MCC tree will greatly reduce the impact of phylogenetic uncertainty on the results (*10*). The generation time (i.e. the time interval from infection onward to transmission, denoted G) of COVID-19 was previously estimated as 7.5±3.4 days (*11*), we used these values to compute the shape and scale parameter of a gamma distribution of G with the R package epitrix (*12*). The distribution of sampling time (i.e. the time gap from

SUPPLEMENTARY TABLE S4. Log-marginal likelihood estimates from model selection by using the path-sampling (PS) and stepping-stone (SS) approaches.

Country	Category	Clock	Coalescent	Log marginal likelihood	
				Path-sampling (PS)	Stepping-stone (SS)
Australia	D614	Strict	Constant	−43,161.09724	−43,161.78956
		Strict	Exponential	−43,142.87942	−43,137.39070
		Strict	Skyline	−43,129.81686	−43,131.14703
		UCLN	Constant	−43,161.86138	−43,161.88169
		UCLN	Exponential	−43,159.59768	−43,160.25060
		UCLN	Skyline	−43,137.20113	−43,137.40271
	G614	Strict	Constant	−42,114.88661	−42,115.21617
		Strict	Exponential	−42,083.25716	−42,082.87311
		Strict	Skyline	−42,077.37285	−42,077.67556
		UCLN	Constant	−42,117.49317	−42,117.88757
		UCLN	Exponential	−42,091.95378	−42,092.27226
		UCLN	Skyline	−42,083.47337	−42,083.78117
UK	D614	Strict	Constant	−42,608.42147	−42,608.81099
		Strict	Exponential	−42,597.33016	−42,597.92089
		Strict	Skyline	−42,590.39452	−42,590.98556
		UCLN	Constant	−42,611.70853	−42,611.82690
		UCLN	Exponential	−42,602.15788	−42,602.19886
		UCLN	Skyline	−42,593.34587	−42,593.38936
	G614	Strict	Constant	−42,723.66227	−42,723.33748
		Strict	Exponential	−42,713.28573	−42,713.22968
		Strict	Skyline	−42,689.38860	−42,689.92583
		UCLN	Constant	−42,728.86399	−42,730.02142
		UCLN	Exponential	−42,717.35538	−42,717.62522
		UCLN	Skyline	−42,697.11254	−42,698.05113
USA	D614	Strict	Constant	−44,757.71082	−44,761.16177
		Strict	Exponential	−44,742.61060	−44,746.51211
		Strict	Skyline	−44,716.67090	−44,719.31089
		UCLN	Constant	−44,747.41638	−44,748.70688
		UCLN	Exponential	−44,733.82629	−44,736.89652
		UCLN	Skyline	−44,708.56055	−44,709.34196
	G614	Strict	Constant	−43,222.54858	−43,222.84280
		Strict	Exponential	−43,196.32220	−43,196.83572
		Strict	Skyline	−43,168.56326	−43,170.73422
		UCLN	Constant	−43,216.36134	−43,217.73716
		UCLN	Exponential	−43,198.82606	−43,199.01257
		UCLN	Skyline	−43,180.20141	−43,178.26833

infection to detection and sampling) was set to equal the distribution of generation time. We performed TransPhylo with at least 500,000 iterations (and sampling parameters every at least 50 steps) by simultaneously estimating the transmission tree, the proportion of sampling, the within-host coalescent time Neg, and the two parameters of the negative binomial of offspring distribution (which represented the number of secondary cases caused by each infection). All results were generated by discarding the first part of the MCMC chains as burn-in. The MCMC mixing and convergence was assessed based on the effective sample size of each parameter (>200). The estimated parameters for each country and each time point were listed in Table S6.



SUPPLEMENTARY FIGURE S2. The maximum clade credibility (MCC) tree of COVID-19 virus during the first two months of the COVID-19 outbreak in each country. Posterior probabilities greater than 0.6 are shown with a green circle. The size of the circle is proportional to the posterior probability. (A) MCC tree for D614 dataset in Australia; (B) MCC tree for G614 dataset in Australia; (C) MCC tree for D614 dataset in UK; (D) MCC tree for G614 dataset in UK; (E) MCC tree for D614 dataset in USA; (F) MCC tree for G614 dataset in USA

SUPPLEMENTARY TABLE S5. Bayesian evaluation for the temporal signal of each dataset.

Country	Category	log likelihood with sampling time (PS)	log likelihood without sampling time (PS)	logBF
Australia	D614	−43,129.81686	−43,225.16828	95.35142
	G614	−42,077.37285	−42,085.17860	7.80576
UK	D614	−42,590.39452	−42,615.28006	24.88554
	G614	−42,689.38860	−42,731.79239	42.40379
USA	D614	−44,708.56055	−44,864.82497	156.26440
	G614	−43,168.56326	−43,196.49259	27.92932

SUPPLEMENTARY TABLE S6. The parameters of offspring distribution estimated for different countries and at different times.

Country	Category	off.r				off.p			pi			R_0	
		Burnin %	ESS	Mean	95% CI	ESS	Mean	95% CI	ESS	Mean	95% CI	Mean	95% CI
Australia	D614	20	413.8905	0.354063	0.1982752−0.5891458	391.9605	0.831348	0.7441796−0.8986794	363.7455	0.125216	0.08353388−0.17563576	1.73018	1.502513−1.980841
	G614	20	322.2057	1.461322	0.7348804−2.7670567	320.9891	0.651306	0.4905768−0.7752587	1,077.3300	0.011942	0.01006814−0.01636301	2.614771	2.401631−2.826343
UK	D614	20	380.9515	0.66069	0.3205959−1.2570012	353.6261	0.719412	0.5635651−0.8372063	220.9361	0.04233	0.02596552−0.06322041	1.638472	1.489879−1.795124
	G614	20	222.4615	0.470768	0.2569896−0.7838550	212.8004	0.796703	0.6957713−0.8770333	219.8597	0.054564	0.03487453−0.07792511	1.818953	1.649302−2.008619
USA	D614	20	522.5207	0.407595	0.2852916−0.5612393	525.152	0.79371	0.7348665−0.8461921	341.9888	0.098685	0.07512361−0.12470692	1.560207	1.433389−1.690855
	G614	20	587.0347	0.983375	0.5873045−1.5882248	572.285	0.798982	0.7012245−0.8728161	342.5892	0.053212	0.03578729−0.07459420	3.874898	3.434469−4.379678

REFERENCES

1. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002;30(14):3059 – 66. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4594756/.
2. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 2012;9(8):772.
3. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges, 1:33-46. https://doi.org/10.1002/gch2.1018.
4. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. RDP4: detection and analysis of recombination patterns in virus genomes. Virus Evol 2015;1(1):vev003. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5014473/.
5. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol 2018;4(1):vey016. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6007674/.
6. Baele G, Li WLS, Drummond AJ, Suchard MA, Lemey P. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. Mol Biol Evol 2013;30(2):239 – 43. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3548314/.
7. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst Biol 2018;67(5):901 – 4. http://dx.doi.org/10.1093/sysbio/syy032.
8. Duchene S, Lemey P, Stadler T, et al. Bayesian evaluation of temporal signal in measurably evolving populations. Mol Biol Evol 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7454806/.
9. Didelot X, Fraser C, Gardy J, Colijn C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. Mol Biol Evol 2017;34(4):997 – 1007. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5850352/.
10. Wang L, Didelot X, Yang J, Wong G, Shi Y, Liu WJ, et al. Inference of person-to-person transmission of COVID-19 reveals hidden super-spreading events during the early outbreak phase. Nat Commun 2020;11(1):5006. http://dx.doi.org/10.1038/s41467-020-18836-4.
11. Li Q, Guan X, Wu P, Wang XY, Zhou L, Tong YQ, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med 2020;382(13):1199 – 207. http://dx.doi.org/10.1056/NEJMoa2001316.
12. Thibaut J, Anne C. Epitrix: small helpers and tricks for epidemics analysis. 2019. https://CRAN.R-project.org/package=epitrix.