Bayesian phylodynamic inferences on the temporal evolution and global transmission of SARS-CoV-2

Jianguo Li (lijg@sxu.edu.cn)
Laboratory of System Biology, Institutes of Biomedical Sciences, Shanxi University
Zhen Li
Laboratory of System Biology, Institutes of Biomedical Sciences, Shanxi University
Xiaogang Cui
Laboratory of System Biology, Institutes of Biomedical Sciences, Shanxi University
Changxin Wu (cxw20@sxu.edu.cn)
Laboratory of System Biology, Institutes of Biomedical Sciences, Shanxi University

Research Article

Keywords: SARS-CoV-2, tMRCA, evolutionary dynamics, Markov Chain Monte Carlo, non-synonymous mutation

DOI: https://doi.org/10.21203/rs.3.rs-23358/v1

Abstract

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has spread widely from China to the world. Although the viral genome has been well characterized, the evolutionary origin and global transmission dynamics of SARS-CoV-2 remain poorly investigated. To address this, we retrieved 313 SARS-CoV-2 genomes from the GISAID database (https://www.gisaid.org), from which 99 genomes generated from original clinical specimens with exact collection dates from 16 countries were selected and enrolled for Bayesian phylodynamic analysis. Here we show that the time to the Most Recent Common Ancestor (tMRCA) of SARS-CoV-2 is Dec 11, 2019 (95%HPD, Nov 21 - Dec 24). Two clades of global circulating strains of SARS-CoV-2 were suggested by Bayesian Maximum Clade Credibility (MCC) tree. The USA circulating strains of SARS-CoV-2 seemed to be from both of the two clades, the UK and Australia circulating strains were from Clade 1, the circulating strains in Singapore, Japan, Germany, France, and Italy were from Clade 2. Although we have not found any obvious bottle-neck-effect from the Bayesian Skyline Plot of the viral population dynamics reconstruction, a sharp reduction of the lower 95% HPD of the relative genetic diversity was observed from Feb 5, 2020, suggesting a possible initiation of a bottle-neck-effect. Thirteen (6 synonymous and 7 non-synonymous) mutations in the viral genome were observed, including two clade-specific mutations (C8782T and T1844C in Clade 1 rather than Clade 2) and eleven sub-clade specific mutations. All of the observed mutations occurred in the USA circulating strains, except one mutation T18488C only occurred in the UK circulating strains. A non-synonymous mutation in the 3’-UTR was also observed, suggesting an altered RNA replication capacity of SARS-CoV-2. We thus came to the conclusion that continuous evolution occurred in almost all regions of the SARS-CoV-2 genome and potentially in a country-specific manner. Further efforts on monitoring the genomic mutations of SARS-CoV-2 from different countries are recommended.

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), first recognized in humans in late 2019 in Wuhan (the capital of middle China’s Hubei province) 1-3, has spread worldwide, developing a global pandemic 4. The initial transmission of

SARS-CoV-2 has been limited in the national wide of China during the first two month 5, while a global spread is establishing with more than 134, 000 laboratory confirmed infections and 5300 deaths from 123 countries by March 14, 2020 (data released by WHO).

SARS-CoV-2 is a betacoronavirus, sharing around 80% whole genome sequence identity with SARS-CoV 6 and about 96% similarity with a bat SARS-related coronavirus (RaTG13, MN996532.1) 7,8. The genome of SARS-CoV-2 exhibited a relative high similarity among the early obtained strains 9,10. However, two key mutations were recently identified, potentially contributing to the sub-lineage classification of SARS-CoV-2 11. Although the genome structure of SARS-CoV-2 has been well documented, the temporal evolution and global transmission of the virus remains poorly investigated.

Herein, we retrieved 313 SARS-CoV-2 genomes from the GISAID database, from which 99 genomes were selected with exact collection dates before Feb 29, 2019. We analyzed the 99 viral genomes by Bayesian phylodynamic approaches to infer the origin time and global transmission of SARS-CoV-2.

Materials And Methods

Viral genome sequence acquisition and alignment

We retrieved 313 complete SARS-CoV-2 genomes (> 29000 bp) uploaded to the GISAID database (www.gisaid.org) with exact collection dates before Feb 29, 2019. After removal of the viral genome with multiple stretches of NNNs (> 0.1% of overall sequence), or with multiple gaps aligned against the reference genome (GenBank Acc. Num. MN908947.3), a total of 99 SARS-CoV-2 complete genomes from original clinical specimens with exact collection date were selected and enrolled for further analysis (Table S1). The enrolled SARS-CoV-2 genomes were aligned by ClustalW implemented in MEGA 6.0 12.

Molecular clock and evolutionary dynamics analysis

We used BEAST v1.10.4 13 to perform Markov Chain Monte Carlo (MCMC) analysis to infer the molecular clock and evolutionary dynamics of SARS-CoV-2. The program BEAUti (Bayesian Evolutionary analysis Utility, included in BEAST software package) was applied for setting the evolutionary model and parameters for the MCMC analysis. An exponential growth coalescent tree prior was selected as the demographic model to estimate the epidemic growth rate, whist a Bayesian skyline coalescent tree prior was selected to infer the viral population dynamics. The tip date was set as the actual sampling time of the corresponding clinical specimen. The best-fit nucleotide substitution model was selected by jModelTest 14. A relaxed (uncorrelated lognormal) clock was chosen as the clock model. Priors for model parameters and statistics were set by default. A classic operator (specifying how the parameters change as the MCMC runs) mix was selected and auto-optimized by BEAUti. A length of 500, 000,000 MCMC chain was set with a sampling frequency of 2000 to ensure stationarity and convergence.

The created BEAST XML file was performed using BEAST. The program Tracer v1.7.1 15 was used to analyze the BEAST output. The parameter effective sample size (ESS) was applied to evaluate if the calculation intensity was enough (ESS > 200). The demographic analysis module implemented in Tracer was applied for reconstruction of viral population dynamics based on the sampled trees from the Bayesian skyline coalescent prior. TreeAnnotator included in BEAST software package was used to summarize and annotate the information contained within the sampled trees. FigTree v1.4.4 13 was applied to view the annotated tree.

Results

Temporal evolution of SARS-CoV-2

To gain insight into the temporal evolutionary dynamics of SARS-CoV-2, we performed Markov Chain Monte Carlo (MCMC) algorithms implemented in BEAST 1.10.4 package with the 99 enrolled SARS-CoV-2 genomes. Generalized Time Reversible (GTR) with invariant sites as site heterogeneity model (GTR+I) was selected as the best-fit nucleotide substitution model by the Akaike Information Criterion (AIC) implemented in jModelTest. The estimated mean evolutionary rate of SARS-CoV-2 was estimated to be 6.14 × 10-6 subs/site/day (95% HPD: 3.61 × 10^{-6} – 8.68 × 10^{-6} subs/site/day), corresponding to 2.24 × 10^{-3} subs/site/year (95% HPD: 1.32 × 10^{-3} –3.17 × 10^{-3} subs/site/year).

The information of MCMC reconstruction was summarized and recorded into a Maximum Clade Credibility (MCC) tree by the program TreeAnnotator. From the MCC tree (Figure 1), the tMRCA of the SARS-CoV-2 was dated back to Dec 11, 2019 (95%HPD, Nov 21, 2019 – Dec 24, 2019). Two major clades were also observed from the MCC tree, with a divergence time at Dec 23, 2019 (95%HPD, Dec 18, 2019– Dec 29, 2019), both of which consist strains of SARS-CoV-2 from Wuhan and other regions of China.

Potentially due to the strong interventions of Wuhan city lockdown by the Chinese government, the Wuhan circulating viral strains were well controlled, while the viral strains outside Wuhan spread to worldwide. To emphasize the viral global transmission from outside Wuhan of China, we excluded the Wuhan circulating strains and observed that the circulating viral strains outside Wuhan could be separated into four sub-clades (Figure 1a). The two sub-clades from Clade 1 was diverged at Jan 1, 2020 (95%HPD, Dec 27, 2019 – Jan 5, 2020), while the two sub-clades from Clade 2 was diverged at Jan 8, 2020 (95%HPD, Jan 3 – Jan 13). With respect to the country-specific strains of SARS-CoV-2, we observed that the circulating strains in USA were from both of the two clades, the UK and Australia circulating strains were from Clade 1, the circulating strains in Singapore, Japan, Germany, France and Italy seemed to be from Clade 2 (Figure 1a, Table S1).

Population dynamics of SARS-CoV-2

To infer the population growth dynamics of SARS-CoV-2, the viral relative genetic diversity was reconstructed by Bayesian Skyline Plot (BSP) analysis 16. BSP analysis suggested that SARS-CoV-2 possessed a relative stable effective population size (*Ne*) during the first month (Dec 23, 2019 to Jan 22, 2020) of the virus outbreak (Figure 1b). A slow but accelerating reduction in the *Ne* was observed from Jan 22, 2020, with a sharp reduction of the lower 95% HPD of the *Ne* from Feb 5, 2020. A sharp reduction in the *Ne* suggests the initiation of a bottle-neck-effect in the virus population size. The bottle-neck-effect indicates that the current circulating virus strain was trapped, and more mutations in the virus genome will occur to help the virus escape, resulting in a leap in the virus population. Despite the BSP was generated from a limited sample size, the results suggested a possible initiation of a bottle-neck-effect in the population size of SARS-CoV-2, indicating more infected cases will occur in the near future due to the increased mutations in the virus genome.

Clade-/Sub-clade specific genomic mutations of SARS-CoV-2

Despite the SARS-CoV-2 remains relative stable, thirteen clade/sub-clade-specific mutations were observed in the present study (Figure 1a). The mutations at nt 8782 and nt 28144 were clade specific, i.e., C8782T and T28144C were only occurred in Clade 1, rather than in Clade 2. Only a viral strain (EPI_ISL_406592 from Guangdong, China) in Clade 1 did not possess C8782T, while all strains in Clade 1 possess T28144C. Eleven out of the thirteen sub-clade specific mutations were also observed (Figure 1a). Seven mutations were located in Clade 1, among which C29095G and C24034T/T26729C were observed in a sub-clade consisting of viral strains from China (outside Wuhan) and USA, respectively. G28878Aand G29742A were observed in a subclade of viral strains from Australia and USA. Four mutations were located in Clade 2, among which C21707T and C28854T were observed in a sub-clade consisting of viral strains from China (outside Wuhan) and USA. C17373T was observed in a sub-clade of viral strains from China (outside Wuhan), USA and Singapore. G26144T was observed in a sub-clade of viral strains from USA, Taiwan, Australia, Sweden, Italy, and Singapore.

Seven of the observed mutations resulted in non-synonymous mutations in the translated viral protein, including two mutations in nucleocapsid phosphoprotein (C28854T: Ser-Phe; G28878A: Ser-Asn), one mutation in ORF1ab polyprotein (T18488C: Ile-Thr), Surface glycoprotein (C21707T: His-Tyr), ORF3a protein (G26144T: Gly-Val), ORF8 protein (T28144C: Leu-Ser), and ORF10 protein (G29742A: Arg-His). Notably, most of the sub-clades possessed one non-synonymous mutation, while one sub-clade consisting viral strains from Australia and USA possessed two non-synonymous mutations (G28878A and G29742A). One non-synonymous mutation (G29742A) occurred in 3’- untranslated region (3’-UTR) of an Australia and USA circulating strains (Figure 1a, Table 1).

Discussion

SARS-CoV-2 has spread from China to more than 123 countries 17. Inference of the temporal evolution dynamics and global transmission route of SARS-CoV-2 is essential in understanding the genetic diversification and in monitoring the key mutations of the virus 18. In the present study, we applied Bayesian MCMC analysis to reveal the origin time and temporal evolution dynamics of SARS-CoV-2. We observed that the current global circulating viral strains could be separated into two clades and four sub-clades from the Maximum Clade Credibility tree. Several clade- or sub-clade specific mutations were also observed.

The origin time of a novel virus is always a primary interest, which could be of help in answering when the virus was transmitted from animal to human. Bayesian estimation of the time to the most common recent ancestor is a widely accepted approach in inferring the origin time of viruses. Lai et al 19 dated the origin time of SARS-CoV-2 to Nov 18, 2019 using 52 viral genomes. Li et al 7 dated the origin time of SARS-CoV-2 to Nov 22-24, 2019, using 70 viral genomes. In the present study, we used 99 viral genomes to date the origin time of SARS-CoV-2 to Dec 11, 2019, with a relative narrower 95% HPD range (33 days VS 63 days 7 or 110 days 19) than the above two previous reports.

Despite only 4% genomic nucleotide variability occurred between SARS-CoV-2 and a bat SARS-related coronavirus (RaTG13) 7, the circulating strains of SARS-CoV-2 was believed to evolve into two major types (L and S type) supported by a population genetics approach 11. The two types of SARS-CoV-2 were well defined by two mutations at nt 8782 and nt 28144. In the present study, we also observed two major clades (Figure 1) of the circulating strains of SARS-CoV-2 by a time-scaled Maximum Clade Credibility (MCC) tree. The two clades could be well defined by mutations C8782T and T28144C. Meanwhile, the viral strains in Clade 1 and Clade 2 were almost equivalent to the viral strains in S type and L type, respectively. Because the MCC tree was based on a Bayesian phylogenetic inference, differing from the haplotype-based population genetics approach, we thus concluded that the circulating SARS-CoV-2 has evolved into two groups, that were well defined by two mutations C8782T and T28144C.

Besides the two clade-specific mutations, we also observed eleven sub-clade specific mutations. All of the sub-clade specific mutations could be found in the USA circulating SARS-CoV-2 strains (Table 1). While one non-synonymous mutation T18488C was only observed in the UK circulating strains. Two non-synonymous mutations G28878A and G29742A were only observed in the circulating strains in Australia and USA. These results suggested that differential evolution trends of SARS-CoV-2 may be occurred in different countries. Notably, one non-synonymous mutation G29742A occurred in 3'UTR of the viral genome, suggesting a possible altered RNA replication capacity in the Australia and USA circulating strains of SARS-CoV-2 20,21.

In conclusion, we performed Bayesian MCMC analysis on 99 SARS-CoV-2 genomes from original clinical specimen with exact collection dates to infer the temporal evolutionary dynamics and global transmission of the virus. Our results suggested that the tMRCA of the virus is Dec 11, 2019. Two major clades of the circulating strains were observed in a maximum clade credibility tree, which were well defined by two mutations G28878A and G29742A. Eleven sub-clade specific mutations were also observed, all of which could be found in the USA circulating strains. One mutation (G29742A) occurred in the 3'-UTR of SARS-CoV-2. Our results suggested that the circulating SARS-CoV-2 is under continuing evolution. Further more attention should be paid on the genomic mutations of SARS-CoV-2.

Declarations

Data availability

The SARS-CoV-2 genomes enrolled in the present study could be retrieved from the GISAID database (https://www.gisaid.org) with the corresponding information included in Table S1.

Declaration of Competing Interest

The authors declare that they have no conflicts of interest.

Acknowledgements

The authors thank the researchers who share the genome data of SARS-CoV-2 to GISAID (http://www.gisaid.org/). This work was supported by a grant from Key R & D plan of Shanxi Province (No. 202003D31003/GZ to J. L.) and a grant from Major National Science and Technology Projects of the Ministry of Science and Technology of China (No. 2018ZX10305409-001-007 to J. L.).

Author contribution JL and CW conceived and designed the research. JL, ZL and XC processed and analyzed the data. JL and CW wrote the manuscript. All authors read and approved the manuscript.

References

1. Wu, *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269, doi:10.1038/s41586-020-2008-3 (2020).
2. Zhou, *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273, doi:10.1038/s41586-020-2012-7 (2020).
3. Ren, L. L. *et al.* Identification of a novel coronavirus causing severe pneumonia in human: a descriptive *Chin Med J (Engl)*, doi:10.1097/CM9.0000000000000722 (2020).
4. Biondi-Zoccai, G. *et al.* SARS-CoV-2 and COVID-19: facing the pandemic together as citizens and cardiovascular *Minerva Cardioangiol*, doi:10.23736/S0026-4725.20.05250-0 (2020).
5. Zhang, S., Diao, Y., Duan, L., Lin, Z. & Chen, D. The novel coronavirus (SARS-CoV-2) infections in China: prevention, control and challenges. *Intensive Care Med*, doi:10.1007/s00134-020-05977-9 (2020).
6. Phan, Genetic diversity and evolution of SARS-CoV-2. *Infect Genet Evol* **81**, 104260, doi:10.1016/j.meegid.2020.104260 (2020).
7. Li, X. *et al.* Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. *J Med Virol*, doi:10.1002/jmv.25731 (2020).
8. Paraskevis, *et al.* 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, doi:10.1016/j.meegid.2020.104212 (2020).
9. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor *Lancet* **395**, 565-574, doi:10.1016/S0140-6736(20)30251-8 (2020).
10. Wu, *et al.* Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host Microbe* **27**, 325-328, doi:10.1016/j.chom.2020.02.001 (2020).
11. Wilen, C. B. *et al.* Tropism for tuft cells determines immune promotion of norovirus pathogenesis. *Science* **360**, 204-208, doi:10.1126/science.aar3799 (2018).
12. Tamura, , Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* **30**, 2725-2729, doi:10.1093/molbev/mst197 (2013).
13. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 10. *Virus Evol* **4**, vey016, doi:10.1093/ve/vey016 (2018).

14. Darriba, , Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* **9**, 772, doi:10.1038/nmeth.2109 (2012).

15. Rambaut, A., Drummond, A. J., Xie, , Baele, G. & Suchard, M. A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol* **67**, 901-904, doi:10.1093/sysbio/syy032 (2018).

16. Heled, & Drummond, A. J. Bayesian inference of population size history from multiple loci. *BMC Evol Biol* **8**, 289, doi:10.1186/1471-2148-8-289 (2008).

17. Benvenuto, *et al.* The global spread of 2019-nCoV: a molecular evolutionary analysis. *Pathog Glob Health*, 1-4, doi:10.1080/20477724.2020.1725339 (2020).

18. Volz, M., Koelle, K. & Bedford, T. Viral phylodynamics. *PLoS Comput Biol* **9**, e1002947, doi:10.1371/journal.pcbi.1002947 (2013).

19. Lai, A., Bergna, , Acciarri, C., Galli, M. & Zehender, G. Early phylogenetic estimate of the effective reproduction number of SARS-CoV-2. *J Med Virol*, doi:10.1002/jmv.25723 (2020).

20. Sola, I., Mateos-Gomez, A., Almazan, F., Zuniga, S. & Enjuanes, L. RNA-RNA and RNA-protein interactions in coronavirus replication and transcription. *RNA Biol* **8**, 237-248, doi:10.4161/rna.8.2.14991 (2011).

21. Lin, J., Zhang, X. M., Wu, R. C. & Lai, M. M. C. The 3' untranslated region of coronavirus RNA is required for subgenomic mRNA transcription from a defective interfering RNA. *Journal of Virology* **70**, 7236-7240, doi:Doi 10.1128/Jvi.70.10.7236-7240.1996 (1996).

Tables

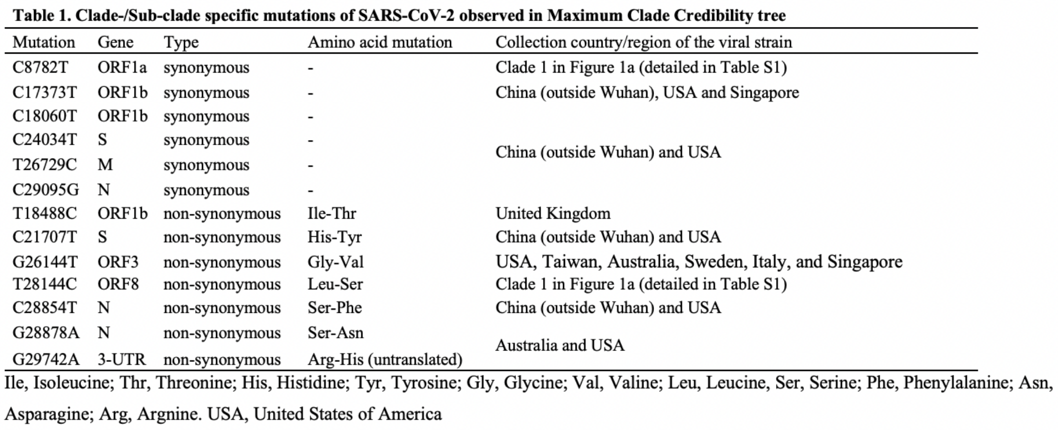

Table 1. Clade-/Sub-clade specific mutations of SARS-CoV-2 observed in Maximum Clade Credibility tree

Mutation	Gene	Type	Amino acid mutation	Collection country/region of the viral strain
C8782T	ORF1a	synonymous	-	Clade 1 in Figure 1a (detailed in Table S1)
C17373T	ORF1b	synonymous	-	China (outside Wuhan), USA and Singapore
C18060T	ORF1b	synonymous	-	China (outside Wuhan) and USA
C24034T	S	synonymous	-	
T26729C	M	synonymous	-	
C29095G	N	synonymous	-	
T18488C	ORF1b	non-synonymous	Ile-Thr	United Kingdom
C21707T	S	non-synonymous	His-Tyr	China (outside Wuhan) and USA
G26144T	ORF3	non-synonymous	Gly-Val	USA, Taiwan, Australia, Sweden, Italy, and Singapore
T28144C	ORF8	non-synonymous	Leu-Ser	Clade 1 in Figure 1a (detailed in Table S1)
C28854T	N	non-synonymous	Ser-Phe	China (outside Wuhan) and USA
G28878A	N	non-synonymous	Ser-Asn	Australia and USA
G29742A	3-UTR	non-synonymous	Arg-His (untranslated)	

Ile, Isoleucine; Thr, Threonine; His, Histidine; Tyr, Tyrosine; Gly, Glycine; Val, Valine; Leu, Leucine, Ser, Serine; Phe, Phenylalanine; Asn, Asparagine; Arg, Argnine. USA, United States of America

Figures

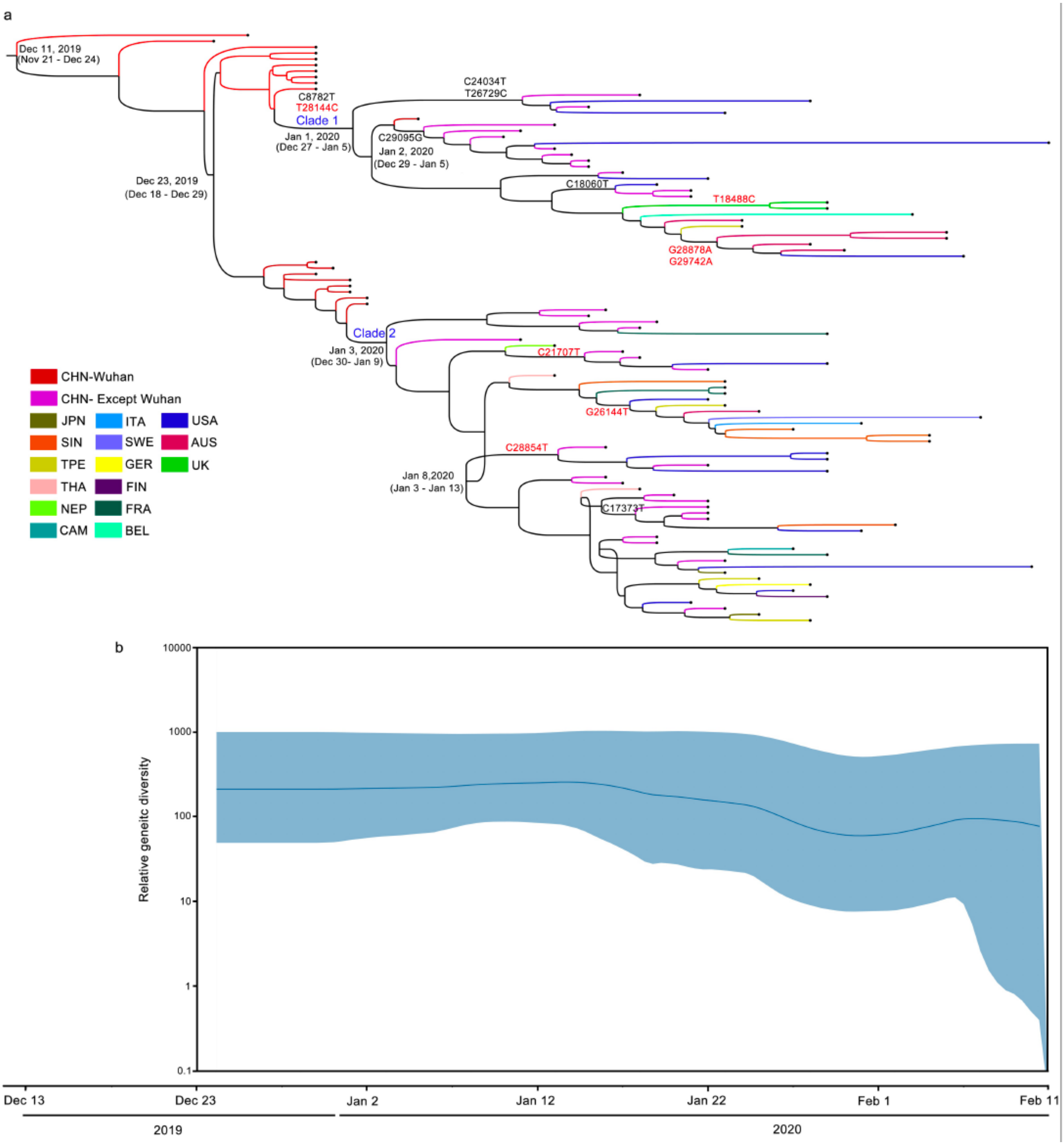

Figure 1

A. The information of Markov Chain Monte Carlo (MCMC) reconstruction was summarized and recorded into a Maximum Clade Credibility (MCC) tree by the program TreeAnnotator. Two major clades were observed from the MCC tree, with a divergence time at Dec 23, 2019 (95%HPD, Dec 18, 2019– Dec 29, 2019), both of which consist strains of SARS-CoV-2 from Wuhan and other regions of China. B. To infer the population growth dynamics of SARS-CoV-2, the viral relative genetic diversity was reconstructed by Bayesian Skyline Plot (BSP) analysis.

Supplementary Files

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

- TableS1.pdf