Title: Spread dynamics of SARS-CoV-2 epidemic in China: a phylogenetic analysis

Author: Hong GuoHu[1] , Guan Qing[2] , Mao Qing[3]

1 Department of Infectious Disease, Guizhou Provincial People`s Hospital, Guiyang, Guizhou, China

2 Department of Dermatology, The First People`s Hospital of Guiyang, Guiyang, Guizhou, China

3 Department of Infectious Disease, The first hospital affiliated to Army Medical University, Shapingba District, Chongqing, China

Correspondence to: Professor Mao Qing, Department of Infectious Disease, The first hospital affiliated to Army Medical University, 30 Gaotanyan street, Shapingba District, Chongqing 400038,P.R. China. E-mail: qingmao@tmmu.edu.cn

Key word: SARS-CoV-2, molecular epidemiology, bayesian inference, phylogenetic analyses, phylogeographical reconstruction

Abstract

To reveal the detailed spread dynamics of SARS-CoV-2 epidemic in China, a phylogenetic analysis was performed by the Bayesian inference framework tool. 233 strains were retrieved from confirmed cases in China until March 31, 2020. The tMRCA of SARS-CoV-2 strains in China could be traced back to December 9, 2019. According to the effective population size curve reconstructed by Skyline model, this research revealed the influence of travel ban measures on the effective population size in China. Furthermore, we divided the epidemic process of SARS-CoV-2 in China into 4 stages according to the effective population size curve. With the Bayesian stochastic search variable selection method, phylogeographical reconstruction detailedly described the geographic spread behavior of SARS-CoV-2 in each stage and confirmed the importance of travel ban in blocking SARS-CoV-2 cross-regional spread. This article summarizes the influence of prevention and control measures in China, which has a positive impact on the world.

Introduction

Coronavirus (CoV) is a broadly distributed zoonotic virus, some species of them can cause human disease. In a long time, it was generally believed that CoV spread with small-scale and caused mild symptoms in immunocompetent individuals[1]. However, the outbreak of severe acute respiratory syndrome coronavirus (SARS-CoV)[2] and Middle East respiratory syndrome coronavirus (MERS-CoV)[3] in the early 21st century, indicated that coronavirus may cause large-scale outbreaks of serious disease.

Since December 2019, an acute respiratory infectious disease caused by a novel coronavirus[4] spread throughout the world in a short time. WHO named it severe acute respiratory syndrome coronavirus 2(SARS-CoV-2), because of a close genetic relationship with SARS-CoV. Some SARS-CoV-2 infected individuals are suffering from Corona Virus Disease 2019 (COVID-19), which not only cause respiratory system disease but also multiple organs dysfunction[5, 6]. According to the clinical published data, millions of people were infected, and the inflection point of the epidemic has not yet arrived.

As the first country outbreak the SARS-CoV-2 epidemic, China confronted a great challenge in prevention and control. Although there was no experience to refer, a series of strong and effective response measures were implemented in the early period of SARS-CoV-2 epidemic. According to the published data[7] from December 2019 to March 2020, the epidemic was effectively curbed in china, however, the global epidemic was not controlled. Summing up China's experience is beneficial to global prevention and control.

It was generally accepted that the travel ban was an important prevention and control measure taken by China. on January 23, 2020, Hubei Province activated the Level-I alert of public health incidents, and the other Provinces, municipalities, autonomous regions in the China mainland activated the Level - I alert, in the following week. The national response appeared to delay the growth of SARS-CoV-2 infected[8]. Although the clinical published data was detailed and reliable, probably it was not accurately reflected the number of infections and the scale of geographic

spread of SARS-CoV-2, because of the incubation period and lack of nucleic acid detection capacity in the early-period.

To explore more details of SARS-CoV-2 epidemic in China, we used the molecular epidemiological method to investigate the spread dynamics of SARS-CoV-2. Bayesian inference through a Markov chain Monte Carlo (MCMC) framework was employed for estimating the evolutionary characteristic parameters of SARS-CoV-2 and an appropriate non-parametric coalescent model was used to calculate the effective population size (*Ne*), which was an important supplement in estimating the number of infections besides the clinical published data. Based on the calculation results of evolutionary parameters, we reconstructed a phylogeographical distribution among different regions of China, which indicated the dynamic distribution of SARS-CoV-2 in China.

Materials and methods

Data sources

Spike (S) gene was defined as the target gene in this research. NCBI and GISAID[9] databases were used to download the data as of March 31, 2020. All SARS-CoV-2 sequences were collected from China in human hosts with the full length of S gene. Those sequences without high coverage sequencing or accurate regional description were eliminated. we also eliminated the duplicate data with the same origin information and nucleic acid in NCBI and GISAID databases. The rest 235 SARS-CoV-2 sequences were performed the multiple alignments by CLUSTAL X program[10] and trimmed into a full-length S gene of 3822 nt. LR757997 and EPI_ISL_417181 were eliminated because of containing over 2% inaccurate nucleic acids in S gene. 233 sequences with 3822nt were selected for this study, including 29 in Hubei, 72 in Shanghai, 44 in Hong Kong, 38 in Guangdong, 21 in Zhejiang, 6 in Taiwan, 6 in Shandong, 5 in Beijing, 3 in Chongqing, 2 in Fujian, 2 in Yunnan, 2 in Jiangxi, 1 in Anhui, 1 in Jiangsu and 1 in Sichuan.

Clinical published data was provided by a open-source R package named nCov2019[7], which described the real-time statistic data with explicit geographic information. New daily confirmed cases data was extracted, and divided into Hubei

Province data, national data, and national data without Hubei Province. Due to the modification of the diagnostic criteria in Wuhan, new daily confirmed cases increased significantly on February 12 and 13, 2020. The data of two days above was eliminated. The trend line was estimated with gamma distribution by R package ggplot2.

Time-resolved Phylogenetic Analyses and *Ne* Estimation

Bayesian inference through a MCMC framework implemented by BEAST v1.10.4 packages[11] was employed to analyze the evolutionary details of S gene, and *Ne* was reconstructed by a non-parametric coalescent model Skyline[12]. General Time Reversible (GTR) was selected as the nucleotide substitution model by calculated AIC score using jModeltest v1.9.1 program[13]. Based on the previous genetic analysis[14-16], an uncorrelated lognormal molecular clock[17] was assumed as the optimization clock mode, given a continuous-time Markov chain reference prior[18]. Collected time of almost sequences was reliable to certainty day except MN908947, which was re-estimated as a prior from December 26 to 31, 2019[19]. Three independent computing processes were performed, and in each process, MCMC chains were set to 200 million and sampling computed every 10000 steps. The output files generated by bayesian computing were discarded the first 10% as burn-in and then combined by LogCombiner tool in BEAST v1.10.4 packages. Tracer software v1.7.1[20] was used to diagnose MCMCs output and estimate *Ne* curve with another 10% burn-in. effective sample sizes (ESS) values greater than 200 for each estimated parameter were accepted.

Phylogeographical Reconstruction and Visualization

To infer the geographical phylogenetic diffusion among the 15 discrete locations in China, we adopt an asymmetric continuous-time Markov chain[21] and inferred spatial-temporal linkage by using Bayesian stochastic search variable selection (BSSVS)[21]. Nucleotide substitution model and tree prior method were kept as GTR and Skyline respectively. The result of time-resolved phylogenetic analyses provided explicit prior information in MN908947 height, uncorrelated lognormal molecular clock setting, and tree root height presupposition. In this computing process we replaced molecular clock rate as a log-normal prior mean and standard deviation value.

Tree root height replaced as an uniform distribution prior with the mean value ± standard error based on the value estimated in time-resolved phylogenetic analyses. There were three independent processes computed as 200 million steps MCMC chains and sampling every 10000 steps. The tree files combined as one file after discarding the first 10%, and which were summarized as maximum clade credibility (MCC) trees using Tree Annotator tool in BEAST v1.10.4 packages. We displayed the discrete geographical phylogenetic MCC trees in FigTree v1.4.3 and generated a visual kml file by spreaD3 v0.9.7[(22)], which was viewed in Google Earth v7.1.8.

Results

The mean value of S gene evolutionary rate was inferred to 4.0520×10^{-3} substitutions per site per year, and standard deviation value was 1.9239×10^{-3}, with 95% confidence interval (CI) 1.3268×10^{-3} to 7.7691×10^{-3}. According to the prior model, the most recent common ancestor (tMRCA) of SARS-CoV-2 in China was probably dated back to December 9,2019, with 95% CI November 11 to December 22, 2019, and the standard error was 6 days.

The trend line of clinical published data illustrated that the growth period of new daily confirmed began on January 19, 2020. The data of Hubei Province dominated China's and gradually increased until Feb 6, 2020. In other regions of China expect Hubei, the declined inflection point appeared on January 1, 2020. The *Ne* curve ascended rapidly from December 25, 2019 to the inflection point on January 25, 2020, and then declined until February 6, 2020. Compared with clinical published data, the *Ne* curve estimated by the Skyline model probably provided more information. We found that the *Ne* of viral strain in China had expanded before Jan 2020. The inflection point of *Ne* curve emerged on the third day after travel ban policy implemented (25 Jan 2020), which indicated that China's control measures had curbed the growth trend expeditiously and effectively. Besides, the rapid decline of viral *Ne* between Jan 25 to February 6, 2020 indicated that the epidemic in China was controlled in a short time, however, the trend was not judged by clinical published data until the mid of Feb 2020.

The epidemic process in China was divided into 4 stages according to the *Ne* curve. Stage □ was defined as the period before December 25, 2019, considering as the stage before SARS-CoV-2 spread. Stage □, from December 25, 2019 to January 25, 2020, was regarded as the key stage of SARS-CoV-2 expansion in China. In approaching Spring Festival, billions of travel movements provided the opportunity to viral strains increase, furthermore, due to the lack of effective control measures in early-stage □, the viral *Ne* increased rapidly. In stage □, form January 25 to February 6, 2020, with the implementation of Level - I alert and travel ban in almost regions of China, the viral *Ne* decreased to a stable situation in about 10 days. In stage □, the viral *Ne* kept at a low-level state after February 6, 2020, meanwhile, the decline of new daily confirmed number in China was also observed.

The phylogeographic reconstruction MCC tree was displayed by FigTree package (Fig. 2), after an initial spread in Hubei Province, 2 lineages emerged on December 23, 2019 simultaneously, and the ancestors of those 2 lineages were traced in Hubei strains. Lineage A covered 160 strains and distributed in 12 regions of China. Lineage B contained 51 strains distributed in 4 regions of China. We identified 4 phylogenetic clusters as posterior probabilities over 0.85 in those two lineages, one of them was identified in lineage A distributed in Shanghai, and the others originated from lineage B. We noticed that cluster d contained 5 strains in Hong Kong, Beijing and Sichuan were traced in stage □. The common ancestor time of other 3 phylogenetic clusters (a, b, c) distributed in Shanghai, Hong Kong and Beijing independently were traced in stage □. Those results suggested that the cross-regional cluster in China related to population migration, and be blocked by travel ban.

To display the geographic spread of SARS-CoV-2 strains in China more clearly, a visualized dynamic graph was generated by spreaD3 and viewed by Google Earth. We displayed the inter-regional spread situation in some key time points of the 4 epidemic stages we divided (Fig. 3). In the stage □, although the increased trend of strain was not detected, the inter-regional spread to Guangdong existed (Fig. 3A), which was probably considered as the first region influenced by SARS-CoV-2 outside Hubei. The analysis of viral *Ne* had been confirmed the importance of stage □, and the

result of geographical reconstruction revealed the complication of SARS-CoV-2 strains spread in China (Fig. 3B), compared with stage □ (Fig. 3C), the inter-regional propagation path had not been changed. Even the observation time was extended to the stage □, the inter-regional cross mode remained stable, although two inter-regional spread path was observed, considered with the population migration had been almost paused, we believed that those two suspicious spreads probably be related to sampling error.

Discussion

SARS-CoV-2 has caused a serious public health event in the world since December 2019. The epidemic has a huge negative impact on the global economy and human health[23, 24]. Chloroquine[25] and Remdesivir[26] are considered as potential specific drugs, but the efficacy needs to be confirmed by more clinical trials[27]. Safety, effectiveness and virus variation are still bottlenecks to the vaccine developed[28]. As there is currently neither a vaccine nor a specific drug, formulating effective virus spread blocking measures probably is an achievable option.

Benefited from a series of prevention and control measures in early period, the epidemic in China was curbed effectively in 2 or 3 months, and the infection rate was lower than other comparable countries. As the first country confronted the epidemic, the experiences of China should be appreciated and summarized. The clinical published data in China has basically summarized from the statistical data of the National Centers for Disease Control (CDC) epidemic reporting system. In the early stage of epidemic, lack of nucleic acid detection, the actual number of infections is hard to be evaluated by the clinical published data. To infer the spread dynamics of SARS-CoV-2 phylogenetic Bayesian method is an important supplement of epidemiological investigation.

S gene codes the SARS-CoV-2 spike glycoprotien and plays an important role in human susceptibility[29], antibody marker[30], antiviral target[31] and vaccine design[32]. So We choose the S gene as the target in this research. The molecular clock rate estimating provide a vital practical basis for the further study of the S gene evolutionary characteristics. The most recent common ancestor of China strains was

traced to 9 Dec 2019, which was very similar to Li`s research[33]. Considered with the incubation period[34], we infer that the first patient hospitalized on December 12, 2019[19]. It probably is close to the China strains ancestor.

Dividing the spread process into 4 stages can reveals the changes of viral *Ne* and geographic spread behavior in each stage distinctly.Combined with the preventive and control measures implemented in China, the importance of travel ban in blocking the spread of SARS-CoV-2 is revealed. We can not accurately assess the influence of SARS-CoV-2 on China without travel ban. However, the research demonstrates that the SARS-CoV-2 has spread to almost all regions in this research in the early period.

After travel ban implemented on January 23,2020, in 3 days, the viral *Ne* declined rapidly, and inter-regional spread almost was blocked. But base on the clinical published data, this change detail was not observed, and the inflection point was observed until mid-February 2020. This research provides a good example that the viral epidemiological survey based on phylogenetic analysis probably provides more information. During the latest two months, there is not any large-scale spread of SARS-CoV-2 in China, which is closely related to the continued travel ban. Considered with the number of tourists during the Spring Festival in other years, the travel ban in 2020 markedly reduced the number of domestic spread and potential exportations, which is not only important to China but also the whole world.

Acknowledgements:

Thanks to the researchers who share the genome data of SARS-CoV-2 to GISAID and NCBI.

Availability of data and materials:

All data generated or analyzed during this study are included in this published article.

Author's contributions:

HGH was the major contributor in designing the research, performing phylogenetic analyses and writing the manuscript. GQ collated and analyzed the sequence information and clinical published data. MQ performed the phylogeographical reconstruction and supervised the study. All authors have read and approved the final manuscript.

Ethics approval and consent to participate:

Not applicable.

Consent for publication:

Not applicable.

Competing interests:

The authors declare that they have no competing interests.

Reference

1. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol. 2019 Mar;17(3):181-92.

2. Peiris JS, Lai ST, Poon LL, Guan Y, Yam LY, Lim W, et al. Coronavirus as a possible cause of severe acute respiratory syndrome. Lancet. 2003 Apr 19;361(9366):1319-25.

3. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med. 2012 Nov 8;367(19):1814-20.

4. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. 2020 Feb 20;382(8):727-33.

5. Zhang C, Shi L, Wang FS. Liver injury in COVID-19: management and challenges. The lancet Gastroenterology & hepatology. 2020 May;5(5):428-30.

6. Zheng YY, Ma YT, Zhang JY, Xie X. COVID-19 and the cardiovascular system. Nature reviews Cardiology. 2020 May;17(5):259-60.

7. Wu T, Ge X, Yu G, Hu E. Open-source analytics tools for studying the COVID-19 coronavirus outbreak. medRxiv. 2020;doi.org/10.1101/2020.02.25.20027433.

8. Tian H, Liu Y, Li Y, Wu CH, Chen B, Kraemer MUG, et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. Science. 2020 May 8;368(6491):638-42.

9. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global challenges. 2017 Jan;1(1):33-46.

10. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic acids research. 1997 Dec 15;25(24):4876-82.

11. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus evolution. 2018 Jan;4(1):vey016.

12. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol. 2005 May;22(5):1185-92.

13. Posada D. jModelTest: phylogenetic model averaging. Mol Biol Evol. 2008 Jul;25(7):1253-6.

14. Vijgen L, Keyaerts E, Lemey P, Maes P, Van Reeth K, Nauwynck H, et al. Evolutionary history of the closely related group 2 coronaviruses: porcine hemagglutinating encephalomyelitis virus, bovine coronavirus, and human coronavirus OC43. J Virol. 2006 Jul;80(14):7270-4.

15. Cotten M, Watson SJ, Zumla AI, Makhdoom HQ, Palser AL, Ong SH, et al. Spread, circulation, and evolution of the Middle East respiratory syndrome coronavirus. mBio. 2014 Feb 18;5(1).

16. Kim JI, Kim YJ, Lemey P, Lee I, Park S, Bae JY, et al. The recent ancestry of Middle East respiratory syndrome coronavirus in Korea has been shaped by recombination. Scientific reports. 2016 Jan 6;6:18825.

17. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS biology. 2006 May;4(5):e88.

18. Ferreira M, Suchard M. Bayesian analysis of elapsed times in continuous - time markov chains. Canadian Journal of Statistics. 2008;36:355-68.

19. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020 Mar;579(7798):265-9.

20. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Systematic biology. 2018 Sep 1;67(5):901-4.

21. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS computational biology. 2009 Sep;5(9):e1000520.

22. Bielejec F, Baele G, Vrancken B, Suchard MA, Rambaut A, Lemey P. SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes. Mol Biol Evol. 2016 Aug;33(8):2167-9.

23. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet. 2020 Mar 28;395(10229):1054-62.

24. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. N Engl J Med. 2020 Apr 30;382(18):1708-20.

25. Yao X, Ye F, Zhang M, Cui C, Huang B, Niu P, et al. In Vitro Antiviral Activity and Projection of Optimized Dosing Design of Hydroxychloroquine for the Treatment of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2020 Mar 9.

26. Ledford H. Hopes rise for coronavirus drug remdesivir. Nature. 2020 Apr 29.

27. Li G, De Clercq E. Therapeutic options for the 2019 novel coronavirus (2019-nCoV). Nature reviews Drug discovery. 2020 Mar;19(3):149-50.

28. Chen WH, Strych U, Hotez PJ, Bottazzi ME. The SARS-CoV-2 Vaccine Pipeline: an Overview. Current tropical medicine reports. 2020 Mar 3:1-4.

29. Cao Y, Li L, Feng Z, Wan S, Huang P, Sun X, et al. Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations. Cell discovery. 2020;6:11.

30. Qu J, Wu C, Li X, Zhang G, Jiang Z, Li X, et al. Profile of IgG and IgM antibodies against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2020 Apr 27.

31. Tang T, Bidon M, Jaimes JA, Whittaker GR, Daniel S. Coronavirus membrane fusion mechanism offers a potential target for antiviral development. Antiviral research. 2020 Apr 6;178:104792.

32. Watanabe Y, Allen JD, Wrapp D, McLellan JS, Crispin M. Site-specific glycan analysis of the SARS-CoV-2 spike. Science. 2020 May 4.

33. Li J, Li Z, Cui X, Wu C. Bayesian phylodynamic inference on the temporal evolution and global transmission of SARS-CoV-2. The Journal of infection. 2020 Apr 20.

34. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. N Engl J Med. 2020 Mar 26;382(13):1199-207.

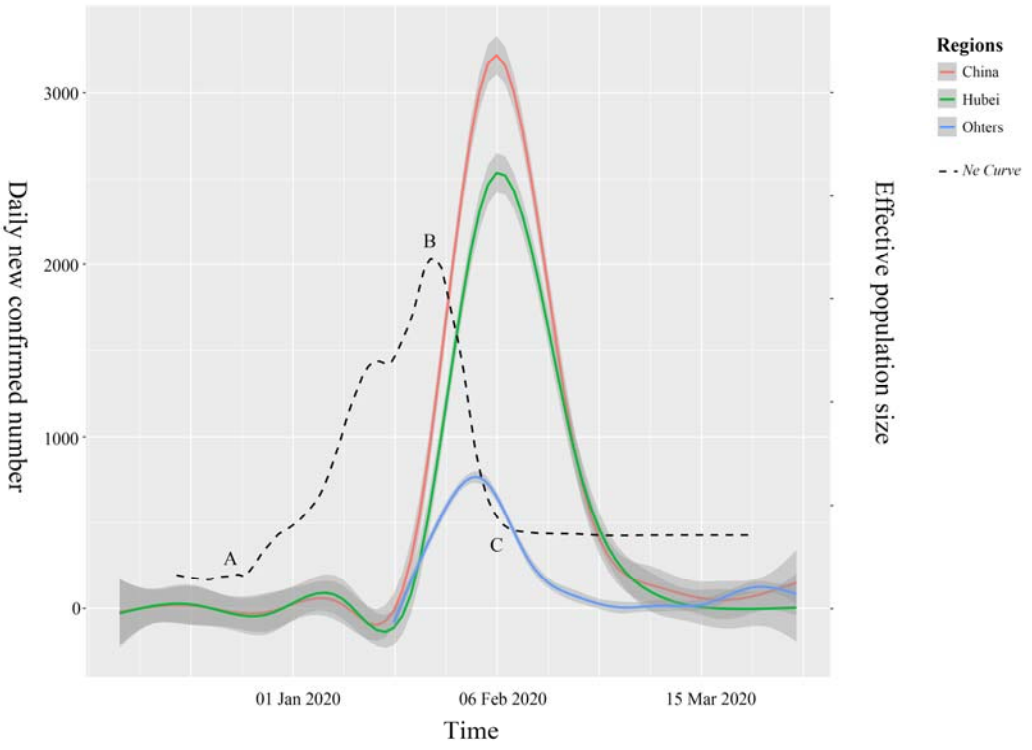

Fig 1: *The trend lines of new daily confirmed number in* Hubei, China, and China without Hubei *are displayed by the red, green and blue line respectively, and the shaded area represents the 95% confidence interval. The mean value of effective population size (Ne) is displayed by the black dotted line. A indicates the beginning of the Ne curve increased is on December 25, 2019. B indicates the Ne curve decreased inflection point is on January 25, 2020. C indicates the time of the Ne curve decreased to a stable situation is on February 6, 2020.*

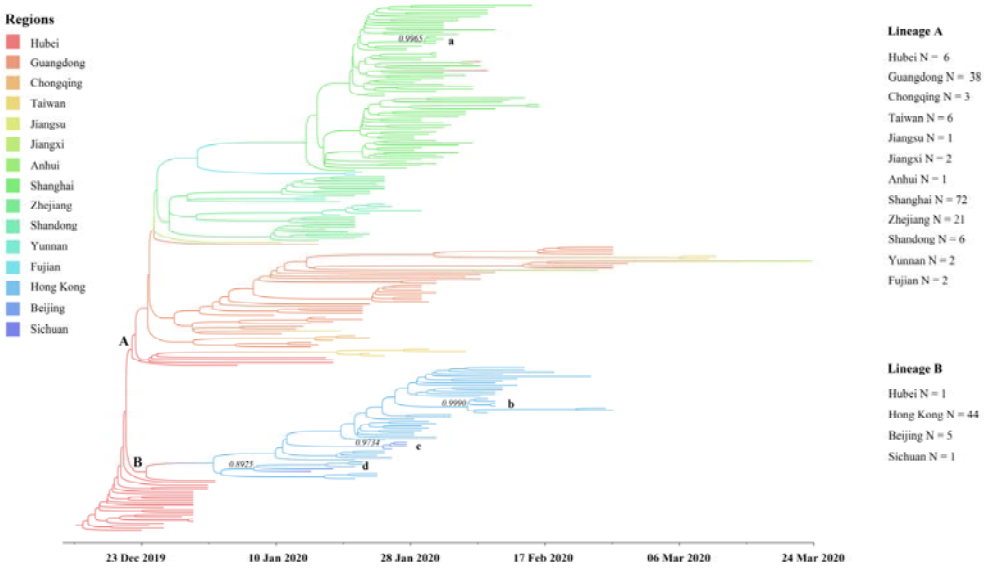

Fig 2: *Maximum clade credibility (MCC) phylogenetic tree of the 233 SARS-CoV-2 S gene sequences were isolated from China reconstructed by geographical phylogenetic. Colors indicate different sampling regions. Uppercase A and B indicate the root of 2 lineages, and the inferred geographical probabilities are equal to 1. The geographical information of strains is listed beside. The 4 clusters we identified are marked by lowercase a, b, c, d and the posterior probabilities are listed above the relevant node.*

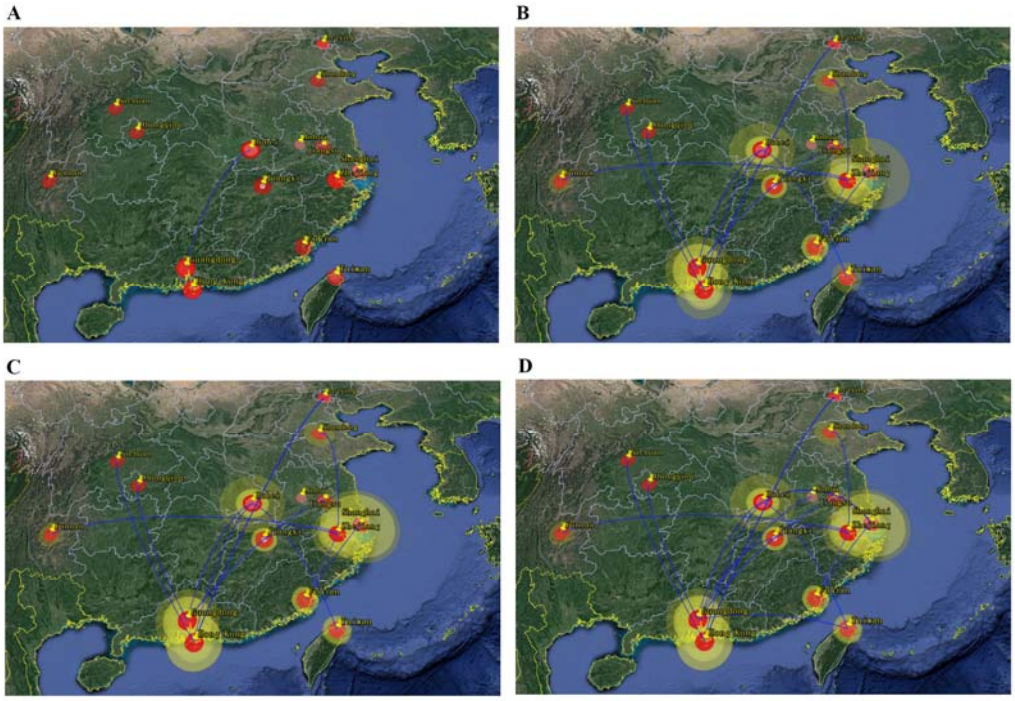

Fig 3: *Visual phylogeographical reconstruction results are displayed by Google Earth. The red dots represent the height of location strains roots, yellow areas represent strain counts in each region, and blue lines indicate the cross-regional pathway. The situations are shown according to different time points as A (December 25, 2019), B (January 25, 2020), C (February 6, 2020), D (March 23, 2020).*