# Genome-wide data inferring the evolution and population demography of the novel pneumonia coronavirus (SARS-CoV-2)

Bin Fang[1*], Linlin Liu[1*], Xiao Yu[1*], Xiang Li[1*], Guojun Ye[1], Juan Xu[3], Ling Zhang[3] , Faxian Zhan[1], Guiming Liu[4], Tao Pan[2#] , Yilin Shu[2#] , Yongzhong Jiang[1#]

1. Influenza Reference Laboratory , Institute of Health Inspection and Testing , Hubei Provincial Center for Disease Control and Prevention, Wuhan 430079, China
2. Anhui Province Key Laboratory for Conservation and Exploitation of Biological Resource, College of Life Sciences, Anhui Normal University, Wuhu 241000, Anhui, China
3. Hubei Province Key Laboratory of Occupational Hazard Identification and Control, School of Public Health, Wuhan University of Science and Technology, Wuhan, Hubei, 430065, China
4. Beijing Agro-Biotechnology Research Center, Beijing Academy of Agriculture and Forestry Sciences, 100097, Beijing, China

*These authors have equal contribution to this study.

#Correspondence: Tao Pan, pantao20070778@163.com; Yilin Shu, sjshuyilin@163.com; Yongzhong Jiang, hbcdcxd@163.com.

**Other Author's Email:**

Bin Fang: nicolfang@163.com

Guojun Ye: gjy0822@sina.comv

Lin-lin Liu: mice0809@163.com

Xiang Li: li_xiang75@yeah.net

Xiao Yu: fish-yuxiao@hotmail.com

Juan Xu: xujuan9270@163.com

Ling Zhang: zhangling@wust.edu.cn

Faxian Zhan: zhanfx@163.com

Guiming Liu: mingguiliu@aliyun.com

## Abstract

Since December 2019, coronavirus disease 2019 (COVID-19) emerged in Wuhan, Central China and rapidly spread throughout China. Up to March 3, 2020, SARS-CoV-2 has infected more than 89,000 people in China and other 66 countries across six continents. In this study, we used 10 new sequenced genomes of SARS-CoV-2 and combined 136 genomes from GISAID database to investigate the genetic variation and population demography through different analysis approaches (e.g. Network, EBSP, Mismatch, and neutrality tests) in the previous three months. The results showed that eighty haplotypes had 183 substitution sites, including 27 parsimony-informative and 156 singletons. Sliding window analyses of genetic

diversity suggested a certain range in mutations and variation in genomic abundance of SARS-CoV-2, which could be explaining the existing widespread and high adaptation of the deadly virus. Phylogenetic analysis showed that pangolin may not be an intermediate host. The network indicated that, in the original haplotype (H14), one patient sample lived near the Huanan seafood market (approximately 2 km), which indicating high possibility of the patient having a history of unconscious contact with this market. However, based on this clue, we cannot accurately conclude that whether this market was the origin center of SARS-CoV-2. Additionally, 16 genomes, collected from this market, assigned to 10 haplotypes, indicated a circulating infection within the market in a short term, which would have led to the outbreak of SARS-CoV-2 in Wuhan and other areas. The EBSP results showed that the first estimated expansion date began from 7 December 2019, which indicated that the transmission of SARS-CoV-2 could have begun from person to person in mid to late November.

**Key words** SARS-CoV-2; metagenomic next-generation sequencing; virus evolution; population demography; phylogenetic relationship

## Introduction

As the largest non-segmented genomes among all the RNA viruses (about 30 kb in length), Coronaviruses (CoVs) own the plasticity due to the mutation and recombination, which increased the potential risks of spread across species [1, 2]. The

COVID-19 (original named 2019-nCOV) is the seventh member of enveloped RNA coronavirus (subgenus Sarbecovirus, Orthocoronavirinae subfamily) [3]. In early December, 2019, an unexplained pneumonia associated with the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged in Wuhan, China [4, 5]. The emergence and rapid spread of the SARS-CoV-2 between infected and healthy people became so devastating as a large population within Wuhan was getting infected [6, 7]. There's evidence that SARS-CoV-2 may have originated from bats, but there is no clear information about the intermediate host that transferred it to humans [8-10]. On 30 January 2020, the World Health Organization (WHO) declared the outbreak of COVID-19 to be a Public Health Emergency of International Concern. Up to March 3, 2020, SARS-CoV-2 has infected more than 89,000 people in China and other 66 countries across the six continents (source: World Health Organization report).

Despite the worldwide rapid spread, the genomic variation dynamics, evolutionary rate, and virus transmission dynamics of SARS-CoV-2 are not yet well understood. In several recent studies, phylogenetic relationships, variations, evolutionary rates, and propagation dynamics were analyzed using limited genomic data from the SARS-CoV-2 [5, 9-17]. As the epidemic progresses, many research institutes around the world have obtained and submitted SARS-CoV-2 genome sequences, increasing the number of SARS-CoV-2 genome sequences in the Data Centers (i.e, GISAID). Given the extremely rapid spread of SARS-CoV-2, an updated analysis with significantly larger sample sizes by incorporating cases throughout the

world is urgently needed. This would allow for more accurate identification of SARS-CoV-2 variant dynamics, evolutionary rates, virus transmission dynamics, and epidemic history in order to effectively implement public health measures, promote drugs and vaccines development and to efficiently prevent similar epidemics in the future.

Recent research findings have suggested bat and other non-bat intermediate mammals (such as pangolin) as potential intermediate host for the SARS-CoV-2 that have so far been widely transmitted to humans [5, 9, 18]. According to the medical information of the first patients to be infected in Wuhan, 27 out of 41 patients were found to be traders of wild mammals in the Huanan seafood market, which directly suggests that the Huanan seafood market as the possible source of origin of the SARS-CoV-2, and later got transmitted it to other areas by the infected people [4, 19]. However, due to some infected persons having no relation or access to the Huanan food market, some researchers are skeptical of the market as the only actual origin or the source of SARS-CoV-2 transmission to humans [4, 19, 20]. Therefore, up to date, in the absence of key potential intermediary host, the origin and transmission pattern of SARS-CoV-2 still need to be thoroughly probed for a more reliable and accurate information on the origin and transmission mechanisms of the virus.

In this study, 56 full genome sequences of outgroups and 136 genomes of SARS-CoV-2 from GISAID EpiFluTM database (access date 22 February 2020) were collected. Ten new sequenced genomes of SARS-CoV-2 were obtained by macrogenomic sequencing. The time origin, genetic diversity, transmission dynamics

and evolutionary history of SARS-CoV-2 were analyzed based on the genomic data to provide understanding on the origin, transmission pathway, and evolutionary characteristics of SARS-CoV-2 outbreak.

## Materials and methods

### *Patients and samples*

In this study, ten patients with unexplained viral pneumonia from five hospitals in Hubei Province were included. Four samples were collected in December, four in January as well as two samples in February. The December samples were examined by LightCycle 480II fluorescent PCR instrument (Roche, Basel, Switzerland) for investigating 26 respiratory pathogens and detection of the virus using SARS-Cov and MERS-Cov primer probes. All the ten samples were detected by the SARS-CoV-2 primer probes. Two samples were used in the virus isolation. The 10 samples were also subjected to metagenomic next-generation sequencing. Specific sample information is presented in Table 1.

### *Virus isolation*

Virus were isolated by HUH7 cells from one lavage fluid sample and one pharynx swab sample. The cells were monitored daily for cytopathic effects by light microscopy. HUH7 cells were harvested after 6 days of culturing at 37℃. The supernatant was collected to extract the total RNA nucleic acid through EZ1 virus mini kit v2.0 (955134; Qiagen, Heiden, Germany), detection was performed by Light Cycle 480II fluorescent PCR instrument (Roche, Basel, Switzerland) followed by subsequent sequencing.

***Library preparation and sequencing***

Total RNA extracted from 10 samples were subjected to metagenomic next generation sequencing testing by EZ1 virus mini kit v.2.0 (955134; Qiage, Heiden, Germany). TruSeq Stranded Total RNA Library Prep Kit (Illumina, San Diego, CA, USA) was used to remove rRNA, reverse transcription and synthesize the double-stranded DNA. The fragmenting, modifying the ends, connecting the joints and enriching of DNA were done by NexteraXT library prepkit (Illumina, San Diego, CA, USA), then the DNA library was obtained. Iseq$^{TM}$ 100 i1 Cartridge, Miseq v2 reagent kit, and High output Reagent Cartridge (illumina, San Diego, CA, USA) were used for deep sequencing in Iseq100, Miseq, and Miniseq platforms (illumina, San Diego, CA, USA), respectively. About 11.5 GB data were obtained for each sample.

***Virus genome analysis***

Using CLC Genomics Workbench 12 and Geneious 12.0.1 software (QIAGEN Bioinformatics, Redwood City, CA, USA) and using reference sequence BetaCoV/Wuhan-Hu-2019 (EPI_ISL_402125), the raw sequencing data of illumina were analyzed. The open reading frames of the verified genome sequences were predicted using Geneious and annotated using the Conserved Domain Database [21].

***Phylogenetic reconstructions***

In this study, for confirmation of the evolutionary history of SARS-CoV-2 (Table S1), 136 complete genomes were retrieved from GISAID (Table S1, access on 22 February, 2020), and combined with our sequenced 10 genomes of this virus. In total, 56 outgroups were collected following the previous study [5]. Mafft v.7.450 was used to align the sequence of SARS-CoV-2 with reference sequences [22].

Phylogenetic analyses of the complete genome and major coding regions were done with RAxML v.8.2.9 [23] in 1000 bootstrap replicates, employing the general time reversible nucleotide substitution model. The tree was visualized with FigTree v.1.4.3 [24].

***Population structure***

On this analysis technique, to precisely decode the evolutionary history of SARS-CoV-2, the genome EPI_ISL_402131 (bat-RaTG13-CoV) from GISAID was also included as the outgroup following the procedure outlined in the previous study [25], because it is the closest sister betacoronavirus to SARS-CoV-2.3. The 136 complete genome sequences of SARS-CoV-2 and an outgroup (bat-RaTG13-CoV) were aligned using MAFFT then the alignment was manually checked using Geneious. For probing the haplotypes relationships among localities, the Network v.4.6.1 (Fluxus Technology, Suffolk, UK) was used to construct the minimum-spanning network based on the full median-joining algorithm [26]. During the alignment, 10 genomes containing ambiguous sites or “N” bases or more degenerated bases were excluded in this study (Table S1). In addition, 4 genome sequences (EPI_ISL_409067, EPI_ISL_412981, EPI_ISL_407071 and EPI_ISL_408489) with one degenerated bases were included, which were divided into two sequences without degenerated bases. For ensuring the accuracy of this analysis, the 5’UTR and 3’UTR contain missing and ambiguous sites of both regions were excluded in the following alignment analyses. DnaSP v.5.10 [27] was used to convert the relevant format. Population genetic indices in was estimated in DnaSP, including nucleotide diversity ($\pi$) and haplotype diversity (*Hd*). $F_{ST}$ between haplotypes was calculated in Mega 6.0 based on haplotype frequency differences with 10,000 permutations [28]. Additionally, a sliding window of 500 bp and steps of 50 bp were used to estimate

nucleotide diversity ($\pi$) for the entire alignment this data. Nucleotide diversity for the entire alignment was plotted against midpoint positions of each window. To indicate the relative position of the mutations in the genome, we selected the EPI_ISL_402124 sequence as the reference genome.

***Population demography***

The alignment was then imported into DnaSP for haplotype analyses. Population size changes were estimated based on a constant population size hypothesis using DnaSP, in combination with neutrality tests (Tajima's *D* and *Fu*'s Fs). The mismatch distribution was estimated based on the genomes of SARS-CoV-2 in Arlequin to test the hypothesis of recent population growth. The Harpending's raggedness index (Rag) and the sum of squares deviation (SSD) were used to determine the smoothness of observed mismatch distribution and the degree of fit between observed and simulated data [29]. Because of the sensitivity to demographic changes of neutral tests, Tajima's *D* [30] and *Fu*'s Fs [31] were estimated using 10,000 coalescent simulations to assess the significance in Arlequin. In addition, the Extended bayesian skyline plot (EBSP) analysis was conducted to examine past population dynamics of SARS-CoV-2 based on 146 genomes by BEAST v.1.8 [32]. The divergence times were estimated in BEAST using a Bayesian Markov chain Monte Carlo (MCMC) method with a strict clock. In this study, the substitution rate was set as $0.92\times10^{-3}$ (95% CI, $0.33\times10^{-3}$-$1.46\times10^{-3}$) substitution/site/year based on the most recent estimation for SARS-CoV-2 [25]. The other parameters were set as follows: extended Bayesian skyline process, 10 million MCMC generations, sampling every 1,000th iteration, the initial 25% burn-in. Tracer was used to check the convergence of the MCMC analyses (ESS values >200). Convergence of the two independent MCMC runs was assessed in

Tracer, as was convergence of model parameter values (effective sample size [ESS]) to ensure ESS values >200.

## Results and discussion

### *Genomic variations of SARS-CoV-2*

According to the database, the genome size of SARS-CoV-2 varied from 29,409 bp to 29,911 bp. In the network dataset, the aligned matrix was 29,130 bp in length including 183 variable sites, of which 27 were parsimony-informative and 156 were singletons, which were classified as 80 haplotypes (Table S2). Nucleotide diversity ($\pi$) was $0.16\times10^{-3} \pm 0.02\times10^{-3}$ (Standard Deviation, SD) and haplotype diversity (*Hd*) was 0.954 ± 0.013 (SD) and variance of *Hd* was $0.16\times10^{-3}$ (Table S3). In this study, our new sequenced 10 genomes of SARS-CoV-2 represented 10 haplotypes (H3, H8, H10, H14, H16, H28, H29, H30, H76 and H78; Table S1). Among these 10 genomes, four samples were collected from those patients who had been exposed to the Huanan seafood market in Wuhan. Additionally, the HBCDC-HB-03/2019 belonged to the core H3 haplotype and most of them represented new haplotype. Sliding window analysis of SARS-CoV-2 revealed significant regional variation across the alignment (Fig. 1). The plot readily showed a relative high degree of nucleotide variation amongst the aligned SARS-CoV-2 genomes for any given window of 500 bp and steps of 50 bp, with the $\pi$ ranging from 0.00005 to 0.0014. Additionally, based on the curve, there were several relative high variation area, compared with other fragments (Fig. 1). The $F_{ST}$ analyses among 80 haplotypes ranged from 0.00003 (e.g, H3 and H9; H14 and H15; H3 and H20) to 0.00134 (H19 and H80), which indicated genetic differentiation among these haplotypes (Fig. 2, Table S4). Overall, the H19 showed relatively large genetic distance with other haplotypes, while the H3 showed opposite pattern, and was also confirmed by the network figure (Fig. 2, Table S4). Overall,

SARS-CoV-2 maintain the relatively rich mutations that has occurred across the world, which may be the main reason of the numerous subgenomic RNAs generated during the viral replication [3, 9]. This phenomenon also provides the possibility of widespread and adaptation for this virus. Fortunately, under the strict quarantine policy in China since 23 January 2020, the circulation and spreading of some haplotypes may have relatively reduced in circulation frequency relatively, compared with the early stage of COVID-19.

***Phylogenetic relationship of SARS-CoV-2***

The virus's popularity and its intermediate host has been a topic of great concern. Some researchers have suggested that the SARS-CoV-2 originated from the bats [5,9]. However, some other researchers have proposed that the non-bat intermediate mammals (e.g. pangolins) may be the transmission path of this genus [18]. However, other studies have also failed to confirm the pangolins as the intermediate host to this viral genus [33]. In this study, based on the phylogenetic tree, bat-RaTG13-CoV was the sister group to the SARS-CoV-2 and the two pangolin samples collected from Guangdong which showed a relatively large genetic distance from the SARS-CoV-2 (Fig. 3 and S1). Based on this finding, , the study suggested that pangolin may not be the intermediate host, as had been reiterated by other previous study [33]. As the Huanan seafood market was closed on 1 January 2020, this created many challenges in identifying the first intermediate hosts (whether people or animal) of SARS-CoV-2. Therefore, more sampling and analysis of the sample from this area may suggest clearer results for more concrete conclusions in future studies.

***Evolutionary relationships of SARS-CoV-2 haplotypes***

The evolutionary network of 80 haplotypes of SARS-CoV-2, with bat-RaTG13-CoV as the outgroup, is shown in Figure 4. The network analysis showed typical star

network, with several core haplotype node (H3, H14, H15) and edge haplotype nodes. In the network, fifty-five satellite haplotypes and H14 connected to the H3 haplotype; eighteen satellite haplotypes and H15+H3 connected to the H14 haplotype; and five satellite haplotypes and H14 connected to H15. Most of these haplogroups were separated by one to six mutations, except two haplotypes (H19 and H80, Fig. 4). The evolutionary network showed that bat-RaTG13-CoV to be connected through a hypothesized haplotype (mv6) to the H15 and H31 haplotypes by single mutations (Fig. 4). However, the mutations between mv6 and bat-RaTG13-CoV was more than 800, which indicated that SARS-CoV-2 still has a relative distant kinship with the outgroup (bat-RaTG13-CoV).

As the most abundant haplotype, H3 included 28 samples, while 55 satellite haplotypes are directly derived from the H3 haplotype (Fig. 4 and Table S1). Moreover, many haplotypes from other countries should also be derived from the H3 haplotype. The current samplings showed that the H3 haplotype has been found to be in 28 samples, but 12 of 28 samples were collected from Wuhan in Hubei Province. Fifteen of the 55 satellite haplotypes were also collected from Hubei Province (Fig. 4). One possible explanation was that a common haplotype from the Huanan seafood market (Fig. 4) was rapidly circulated at an early stage of human-to-human transmissions. It is worth noting that in the H14, there are two samples (EPI_ISL_406801 and EPI_ISL_ 412979) from Wuhan. Although these two hosts didn’t directly link with the Huanan market, one of the host (EPI_ISL_ 412979) having lived in a residential area about 2 kilometers from the Huanan seafood market. This also provides the possibility of them having indirect possible contact with the Huanan seafood market.

Additionally, the networks of 24$^{th}$ December, 2019 - 30$^{th}$ December, 2019 and

24th December, 2019 - 6th January, 2020 also suggested that the most frequent haplotype (H3) occurred. All the haplotypes were collected from Wuhan, which indicated that Wuhan may have acted as the important origin center (Fig. 4). On the other hand, in this study, 16 sequences of Huanan seafood market (4 new sequenced and 12 GISAID data) were collected. Fifteen of out of the sixteen samples were collected before 1st Januray, 2020 and one samples was collected on 8th Januray, 2020. These samples were represented ten haplotypes (H2, H3, H4, H5, H6, H7, H8, H10, H11, H16), which indicated the high proportion of haplotype diversity in Huanan seafood market. Among those haplotypes, the H3 haplotype, the most abundant, was present in 6 samples from Huanan seafood market, while the other haplotypes were directly derived from the H3 haplotype. All the samples from the Huanan seafood market had the H3 haplotype and other 9 derived haplotypes (Fig. 4), indicating that there were circulated infections within the market in a short term. Noteworthy, in the network, a total of 65 virus samples from 15 other countries were assigned to 43 haplotypes. Among them, 27 haplotypes were satellite haplotypes of H3 haplotype, 11 haplotypes were satellite haplotypes of H14 haplotype, and 3 haplotypes were satellite haplotype of H15 haplotype. Indeed, most of haplotypes originated from the H3, which may hint the phenomenon was related to the input of virus carriers from Wuhan.

Based on the outgroup (bat-RaTG13-CoV) having a possible direct connection with outgroup H15 and H31, it indicates that both the H15 and H31 were the suggested ancestral haplotypes (Fig. 4), as had also been suggested by other previous study [25]. The H15 was only recovered from five Shenzhen (Guangdong) samples, while H31 included 3 United States samples and 1 Fujian sample from China (Fig. 4). Based on the epidemiological statistics, all the patients from the H15 were ever

traveled to Wuhan [16]. As to the H31, three genomes from the same patient in Unite State [34] and 1 genome from Fujian of China (Table S1) were included. For the patient in Unite State, He may have infected during the period of visiting his family in China [25]. The travel history of Fujian sample was unclear, but we cannot deny that it has a history of unconscious Wuhan contacts. Based on the theory in the previous study [25], this current study also supported two main evolutionary paths, that the available haplotypes could be from H15 through H14 to H3, or from H31 through H14 to H3 (Fig. 4). Both these two path demonstrate that H14 was the key connection from an ancestral haplotype to H3. However, unlike the previous research, in H14, two individuals (EPI_ISL_406801 and EPI_ISL_ 412979) were collected from Wuhan. Additionally, there was another new haplotype (H28, EPI_ISL_412980), which originated from the H14, and was also collected from Wuhan. Although the three hosts of these two haplotypes didn't directly link with the Huanan seafood market, one of the hosts (EPI_ISL_ 412979) lived in a residential area about 2 kilometers from the Huanan seafood market. This also provides the possibility for that host to have unconsciously indirect contact with the Huanan seafood market. Overall, it cannot further be proved the SARS-CoV-2 in the Huanan seafood market had been transmitted from other places or that Huanan seafood market did not host the original source of SARS-CoV-2. Due to the closing up of the Huanan seafood market on 1$^{st}$ of January, 2020, it made it difficult to determine whether there were intermediate hosts in the market. To accurately determine whether this market was the central origin, further collection of early SARS-CoV-2 samples within/around the Huanan seafood market is needed to provide genomic information and epidemiological survey data needed in integration analysis. However, based on the facts provided in the current study undertaken in the early days of the outbreak of the pneumonia infection in

Wuhan, the Huanan seafood market promoted the spread of SARS-CoV-2, and since then, infected travelers have spread to throughout China and other countries.

***Population size expansion of SARS-CoV-2***

As to population demography of SARS-CoV-2, the EBSP results revealed that it experienced an expansion of effective population size for the last 66 days prior to 11$^{th}$ of February, 2020 (Fig. 5). With the latest one being sampled on 11$^{th}$ February 2020, the first expansion date is estimated to had begun from 7$^{th}$ December 2019. The mismatch in the distribution of the total populations was also clearly shown through unimodal (Fig. 6) and the neutral tests (Tajima's D and Fu's Fs) revealed statistically significant population expansion of SARS-CoV-2 (Table S3), which also supported the EBSP results. In addition, with the change of time, multiple evidence (i.e. haplotype number, Tajima's D, Fu's FS) indicated a relatively highest population growth at around 21$^{st}$ - 27$^{th}$ of January (Table S3). The demographic expansion of SARS-CoV-2 during this period was consistent with the patients first diagnosed on 8$^{th}$ December, 2019 [4,19] and the previously suggested most recent common ancestor (TMRCA) dates for SARS-CoV-2 at 6$^{th}$ December, 2019) [17]. Additionally, considering that the virus has longest reported incubation period of up to 24 days [6], the virus may had first infected humans in mid to late November, which is basically consistent with the results of previous studies [25]. Therefore, in this study, EBSP revealed that SARS-CoV-2 experienced an effective population size expansion since 7$^{th}$ December 2019, which was also supported by a star-like network, EBSP, the neutral tests (Fu's and Tajima's D test) and mismatch analysis.

**Contributors**

Yilin Shu, Tao Pan, and Yongzhong Jiang conceived the research, analyzed the data, interpreted the results, and wrote the draft manuscript; Bin Fang, Linlin Liu, Xiao Yu

and Xiang Li performed macrogenome sequencing experiments. Xiao Yu and Xiang Li performed virus isolation experiments. Guojun Ye, Bin Fang, Juan Xu, Ling, Zhang, Yilin Shu and Faxian Zhan collected data. All authors reviewed and approved the final version of the manuscript.

**Declaration of interests**

We declare no competing interests.

**Acknowledgements**

We thank scientists and researchers for depositing the whole genomic sequences of Novel Pneumonia Coronavirus (SARS-CoV-2) in the Global Initiative on Sharing All Influenza Data (GISAID) EpiFlu™; and the GISAID database for allowing us access to sequence for non-commercial scientific research. We also are grateful to Dr. Oscar Omondi Donde (Egerton University, Kenya) for polishing the manuscript language. Thanks to Yifa Zhu, Yangyang Tao and Xierong Li (Tianmen Center for Disease Control and Prevention) for collecting samples and thanks to Maoyi Chen, Jie Hu, Chunlin Mao (Jingzhou Center for Disease Control and Prevention) for sampling.

## Reference


[1]. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol. 2019;17(3):181-192.

[2]. Gorbalenya AE. Severe acute respiratory syndrome-related coronavirus–The species and its viruses, a statement of the Coronavirus Study Group. BioRxiv. 2020;DOI: https://doi.org/10.1101/2020.02.07.937862.

[3]. Zhu N, Zhang DY, Wang WL, et al. A novel coronavirus from patients with pneumonia in China, 2019. New Engl J Med. 2020;382(8):727-733.

[4]. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet. 2020;395(10223):497-506

[5]. Lu RJ, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 2020; DOI: 10.1016/S0140-6736(20)30251-8.

[6]. Guan WJ, Ni ZY, Hu Y, et al. Clinical characteristics of 2019 novel coronavirus infection in China. N Engl J Med. 2020; DOI: 10.1056/NEJMoa2002032.

[7]. Yang Y, Lu QB, Liu MJ, et al. Epidemiological and clinical features of the 2019 novel coronavirus outbreak in China. MedRxiv. 2020;DOI: https://doi.org/10.1101/2020.02.10.20021675.

[8]. Wassenaar TM, Zou Y. 2019_nCoV: Rapid classification of betacoronaviruses and identification of traditional Chinese medicine as potential origin of zoonotic coronaviruses. Lett Appl Microbiol. 2020; DOI: 10.1111/lam.13285.

[9]. Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020; DOI: 10.1038/s41586-020-2012-7.

[10]. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020; DOI: 10.1038/s41586-020-2008-3.

[11]. Benvenuto D, Giovannetti M, Ciccozzi A, et al. The 2019-new coronavirus epidemic: evidence for virus evolution. J Med Virol. 2020; DOI: 10.1101/2020.01.24.915157

[12]. Ceraolo C, Giorgi FM. Genomic variance of the 2019-nCoV coronavirus. J Med Virol. 2020; DOI:10.1002/jmv.25700.

[13]. Chen LJ, Liu WY, Zhang Q, et al. RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. Emerg Microbes Infec. 2020; 9(1):313-319.

[14]. Paraskevis D, Kostaki EG, Magiorkinis G, 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. 2020; 79:104212.

[15]. Chan JF, Kok KH, Zhu Z, et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg Microbes Infec. 2020;9(1):221-236.

[16]. Chan JFW, Yuan SF, Kok KH, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a

study of a family cluster. Lancet. 2020; 395(10223):514-523.

[17]. Li XW, Wang W, Zhao XF, et al. Transmission dynamics and evolutionary history of 2019-nCoV. J Med Virol. 2020; DOI: 10.1002/jmv.25701.

[18]. Lam TTY, Shum MHH, Zhu HC, et al. Identification of 2019-nCoV related coronaviruses in Malayan pangolins in southern China. bioRxiv. 2020; DOI: https://doi.org/10.1101/2020.02.13.945485.

[19]. Li Q, Guan XH, Wu P, et al. Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. New Engl J Med. 2020; DOI: 10.1056/NEJMoa2001316.

[20]. Cohen J. Wuhan seafood market may not be source of novel virus spreading globally. Science. 2020;10.1126/science.abb0611.

[21]. Marchler-Bauer A, Bo Y, Han LY, et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 2017;45(D1):D200-D203.

[22]. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772-780.

[23]. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–1313.

[24]. Rambaut A. FigTree v1.4.3; 2016. Available at: http://tree.bio.ed.ac.uk/software/figtree/ 2016.

[25]. Yu WB, Tang GD, Zhang L, et al. Decoding the evolution and transmissions of

the novel pneumonia coronavirus (SARS-CoV-2) using whole genomic data. ChinaXiv. 2020;202002.00033v2.

[26]. Bandelt H, Forster P, Röhl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol. 1999;16(1):37-48.

[27]. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009;25(11):1451–1452.

[28]. Tamura K, Stecher G, Peterson D, et al. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013;30(12):2725–2729.

[29]. Harpending HC. Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. Hum Biol. 1994 ;66(4):591-600.

[30]. Tajima F. The effect of change in population size on DNA polymorphism. Genetics. 1989;123(3):597–601.

[31]. Fu YX. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics. 1997;147(2):915–925.

[32]. Drummond AJ, Suchard MA, Xie D, et al. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29(8):1969–1973.

[33]. Liu P, Jiang JZ, Wan XF, et al. Are pangolins the intermediate host of the 2019 novel coronavirus (2019-nCoV) ? bioRxiv. 2020;DOI: https://doi.org/10.1101/2020.02.18.954628.

[34]. Holshue ML, DeBolt C, Lindquist S, et al. First Case of 2019 Novel Coronavirus in the United States. N Engl J Med 2020; 10.1056/NEJMoa2001191.

**Table 1. Sample information for 10 SARS-Cov-2 infected patients.**

| Sample name | Sample type | Collection date | Collection site | Exposure to huanan seafood marke |
|---|---|---|---|---|
| BetaCoV/Wuhan/HBCDC-HB-02/2019 | Bronchoalveolar lavage fluid | 30th Dec,2019 | Wuhan, Hubei Province | Yes |
| BetaCoV/Wuhan/HBCDC-HB-01/2019 | Bronchoalveolar lavage fluid | 30th Dec,2019 | Wuhan, Hubei Province | Yes |
| BetaCoV/Wuhan/HBCDC-HB-03/2019 | Bronchial scraping | 30th Dec,2019 | Wuhan, Hubei Province | Yes |
| BetaCoV/Wuhan/HBCDC-HB-04/2019 | Sputum | 30th Dec,2019 | Wuhan, Hubei Province | No |
| BetaCoV/Wuhan/HBCDC-HB-04/2020 | Throat swab | 18th Jan,2020 | Wuhan, Hubei Province | No |
| BetaCoV/Wuhan/HBCDC-HB-03/2020 | Throat swab | 18th Jan,2020 | Wuhan, Hubei Province | No |
| BetaCoV/Wuhan/HBCDC-HB-02/2020 | Throat swab | 17th Jan,2020 | Wuhan, Hubei Province | No |
| BetaCoV/Jingzhou/HBCDC-HB-01/2020 | Throat swab and cultured virus | 8th Jan,2020 | Jingzhou,Hubei Province | Yes |
| BetaCoV/Wuhan/HBCDC-HB-06/2020 | Faeces | 7th Feb,2020 | Wuhan, Hubei Province | No |
| BetaCoV/Tianmen/HBCDC-HB-07/2020 | Throat swab | 8th Feb,2020 | Tianmen, Hubei Province | No |

## Figure Legend

**Fig. 1. Sliding window analyses showing the nucleotide diversity based on alignment of genomes of SARS-CoV-2.** The red line shows the value of nucleotide diversity ($\pi$) in a sliding window analysis of window size 500 bp with step size 50, the value is inserted at its mid-point. EPI_ISL_402124 sequence was selected as the reference genome to indicate the mutation position.

**Fig. 2. Pairwise $F_{ST}$ among 80 haplotypes o**f SARS-CoV-2. Consistent with the Table S4.

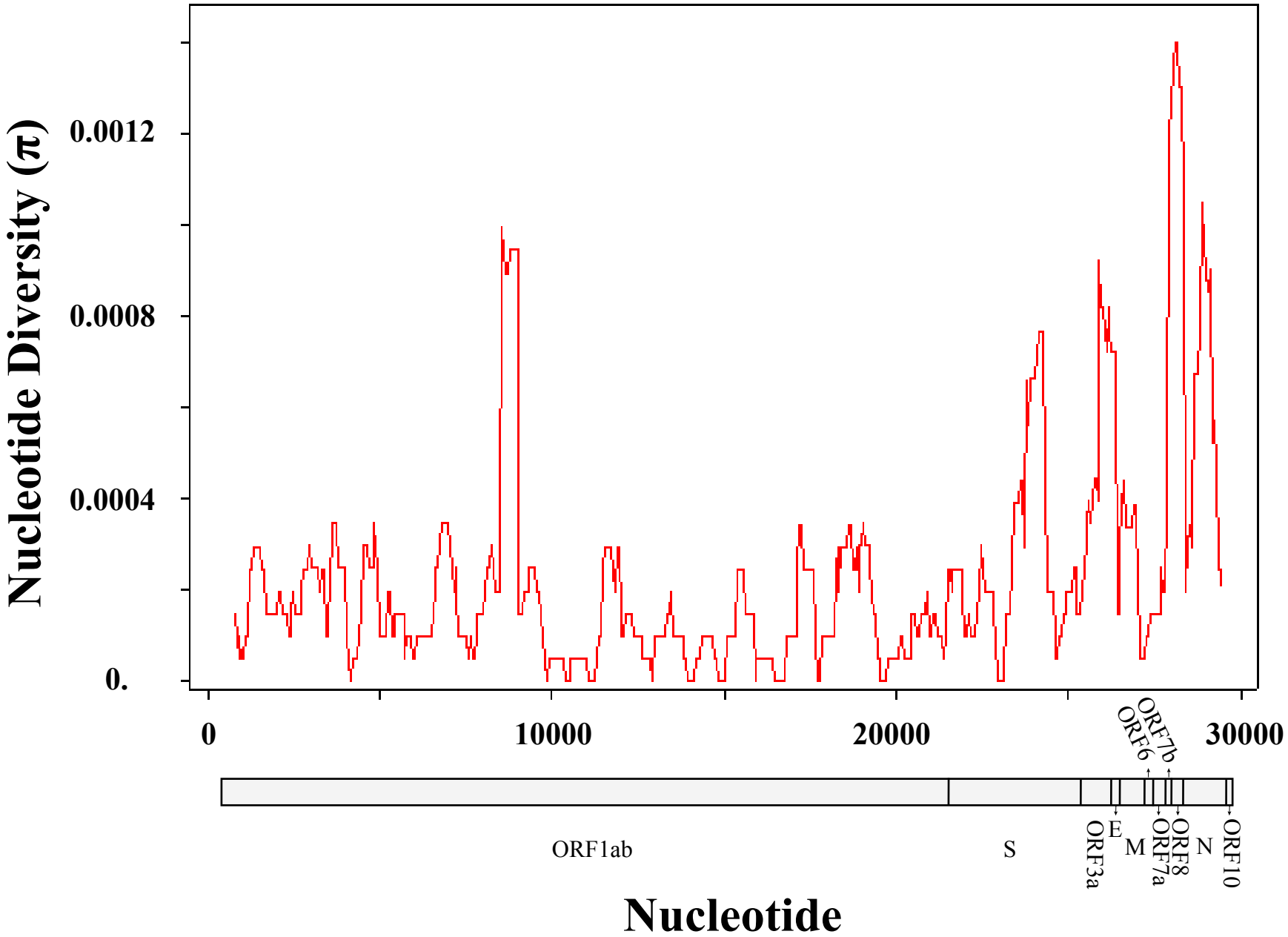
**Fig. 3. Simplify phylogenetic tree of SARS-CoV-2 form Fig. S1.** The nodal numbers are ML bootstrap values.

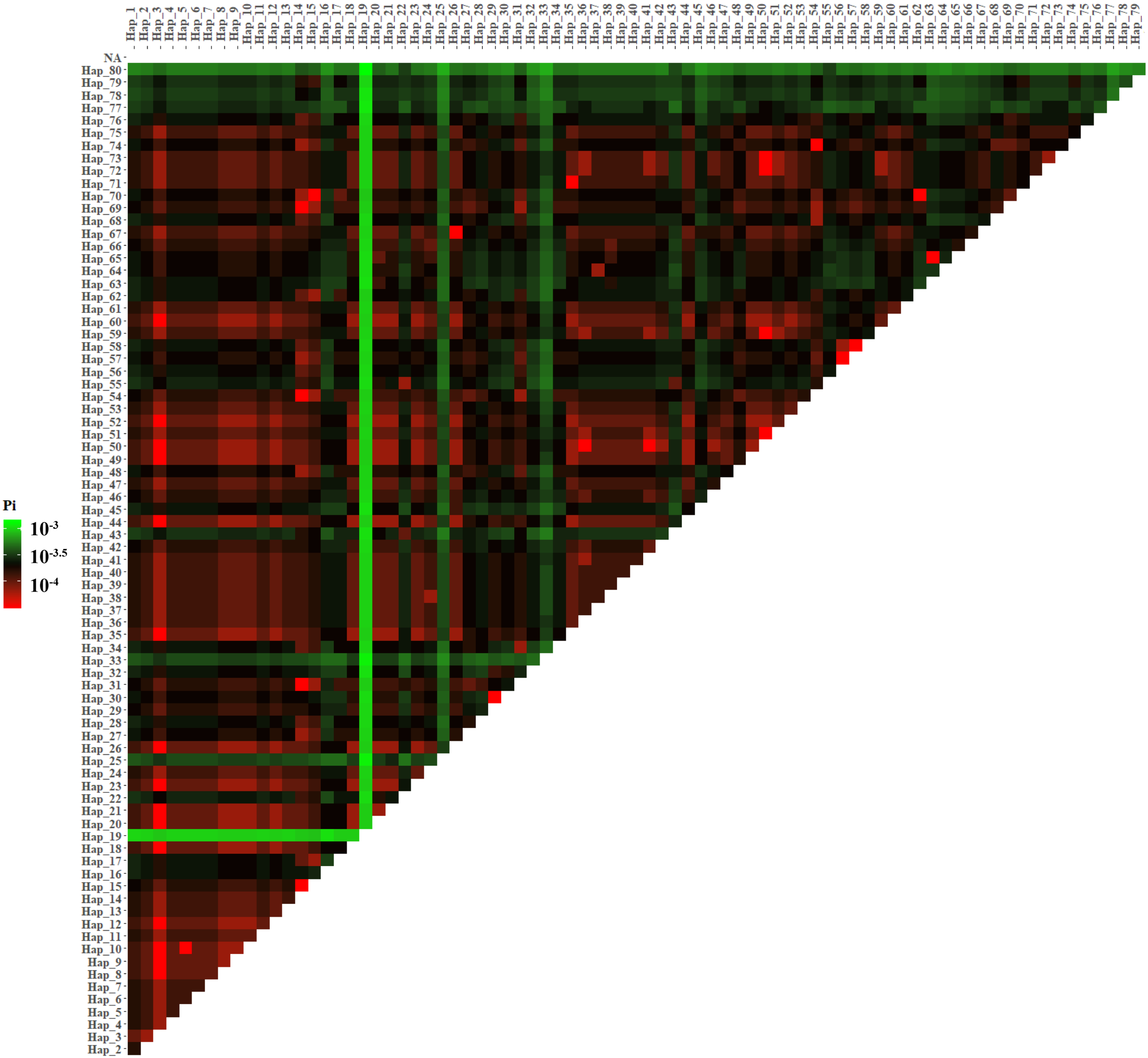
**Fig. 4. Median-joining network with node sizes proportional to the frequencies of haplotypes in SARS-CoV-2 (A, B and C).** (A)All the individuals collected during 24$^{th}$ December to 30$^{th}$ December in 2019; (B) All the individuals collected during 24$^{th}$ December in 2019 to 6$^{th}$ January in 2020; (C) All the individuals collected before 11$^{th}$ February in 2019. The numbers of mutations separating the haplotypes are shown on the branches, except for the lower than seven-step mutations. The little red diamond nodes indicate undetected haplotypes. The sampling areas are indicated by different colors .

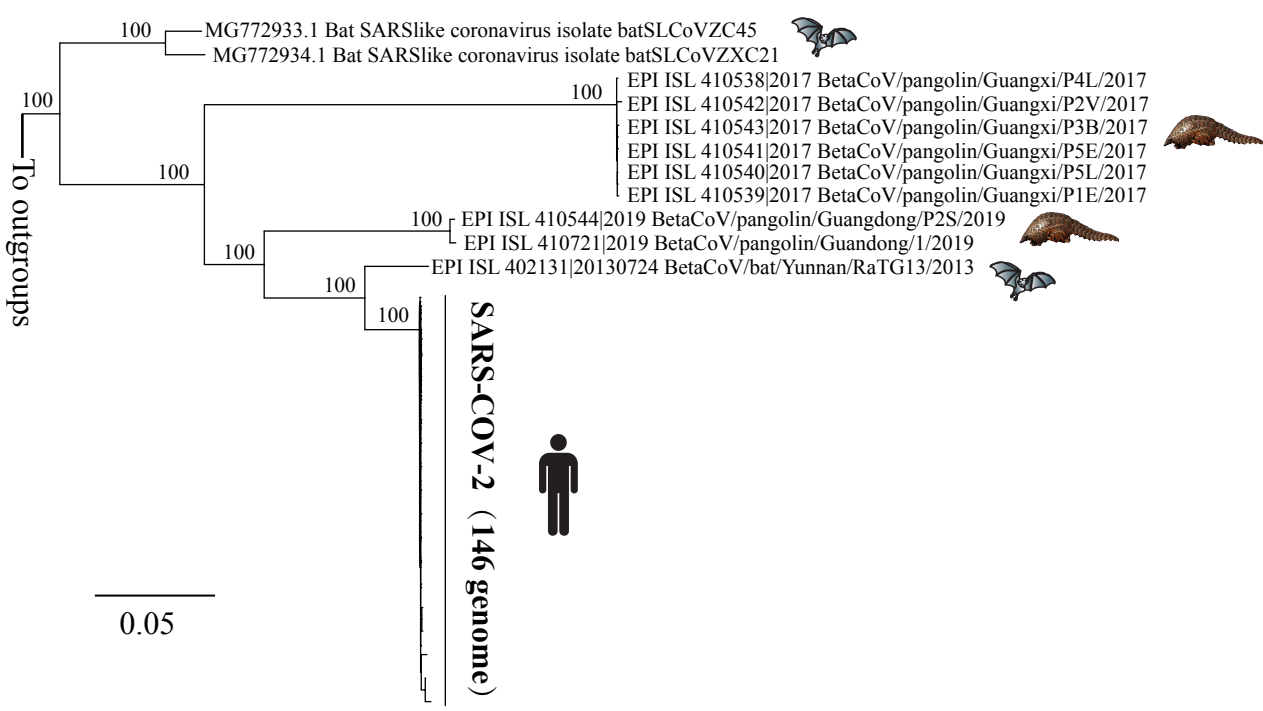
**Fig. 5. Population fluctuation infered by Extended bayesian skyline plot (EBSP) of SARS-CoV-2.** The x-axis indicates time in days BP, and the y-axis indicates the effective population size divided by generation time in units of $N_{e\tau}$ (the product of effective population size and generation time in days). The blue areas represent 95 % highest posterior density. Time is expressed in days.

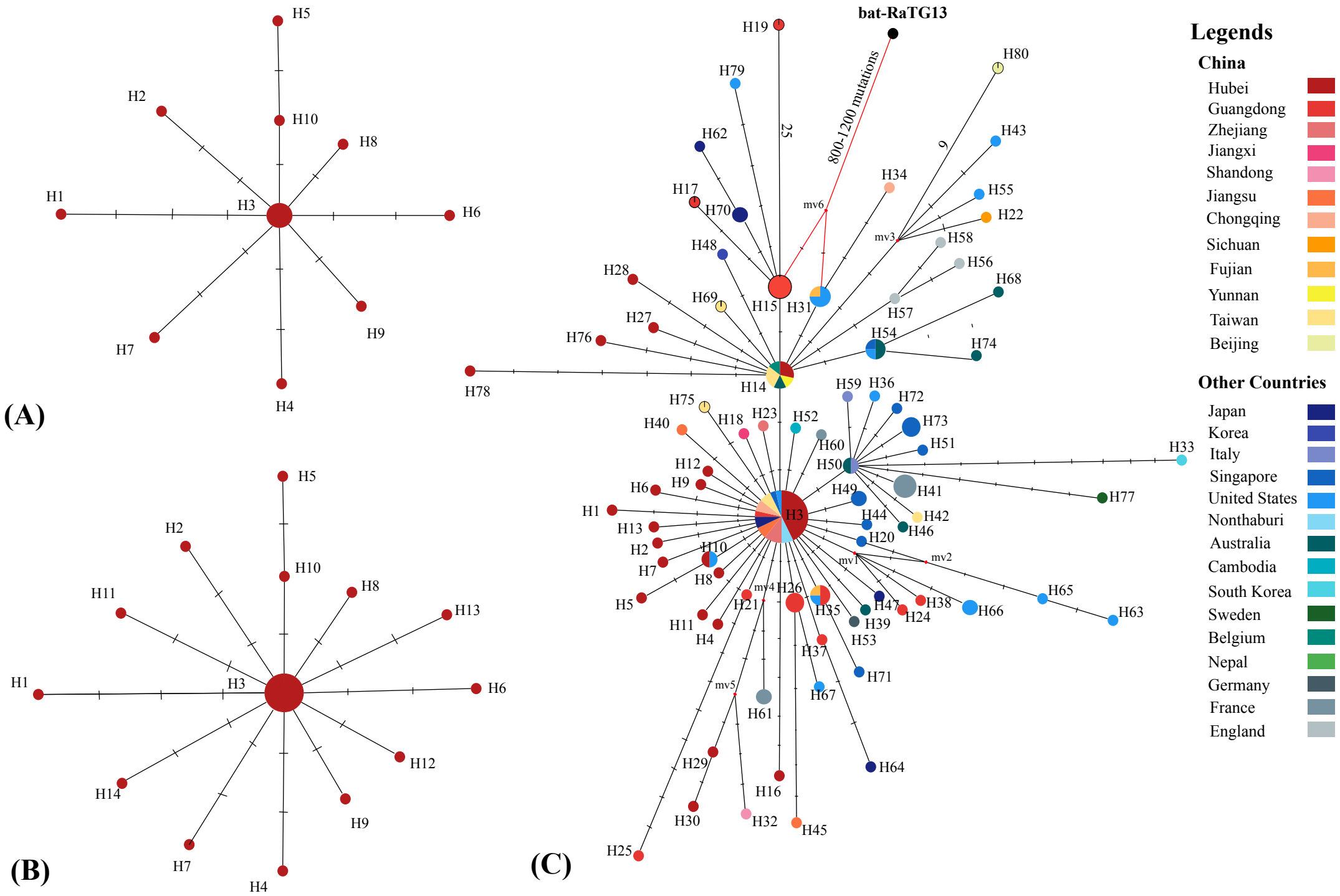
**Fig. 6: Mismatch distributions analyses for SARS-CoV-2 in different period (A, B, C, D, E, F, G).** (A) All the individuals collected during 24$^{th}$ December to 30$^{th}$ December, 2019; (B) 24$^{th}$ December in 2019 to 6$^{th}$ January in 2020; (C) 24$^{th}$ December in 2019 to 13$^{th}$ January in 2020; (D) 24$^{th}$ December in 2019 to 20$^{th}$ January in 2020; (E) 24$^{th}$ December in 2019 to 27$^{th}$ January in 2020; (F) 24$^{th}$ December in 2019 to 3$^{th}$ February in 2020; (G) 24$^{th}$ December in 2019 to 11$^{th}$ February in 2020. The x coordinate represents the number of differences in each pair of sequence comparisons; the y coordinate represents the frequencies of pairwise differences. The blue histogram are the observed frequencies of pairwise divergences among sequences and the red line refers to the expectation under the model of population expansion.

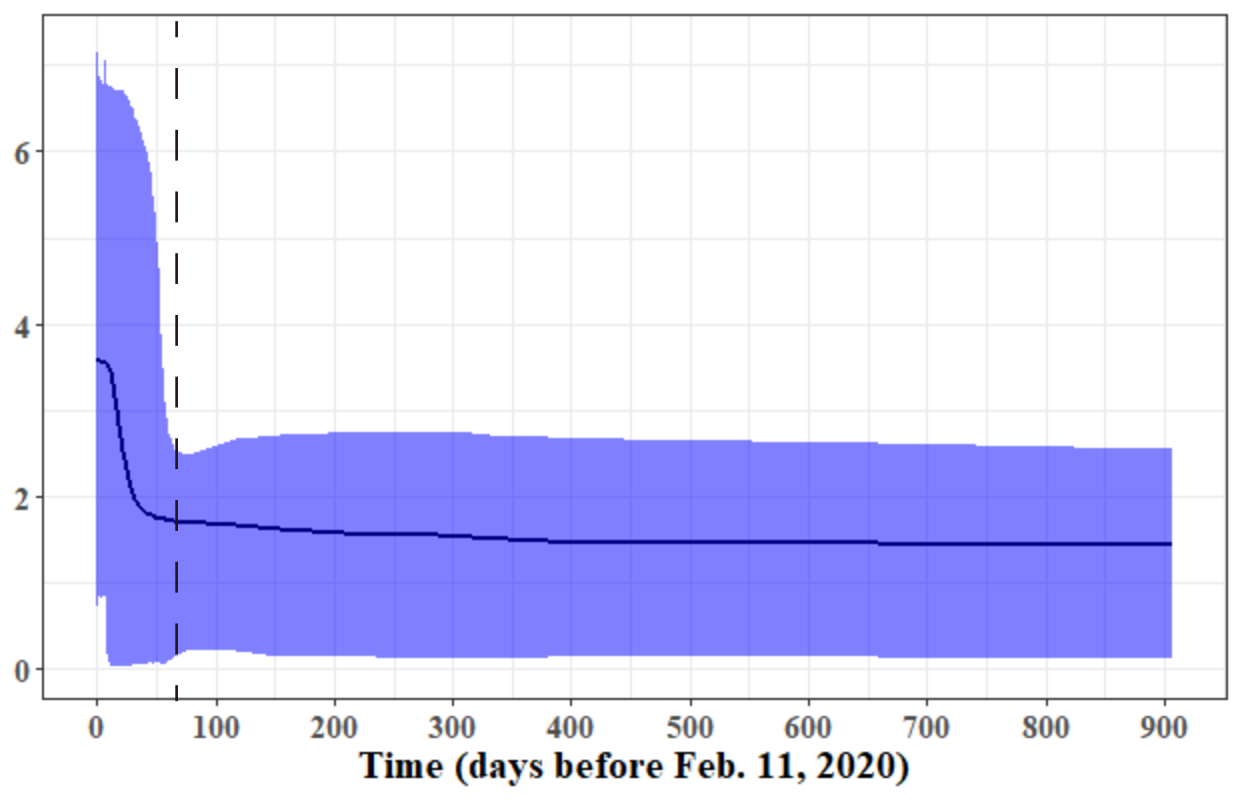
**Fig. S1. Phylogenetic tree of** SARS-CoV-2**.** The nodal numbers are ML bootstrap values.

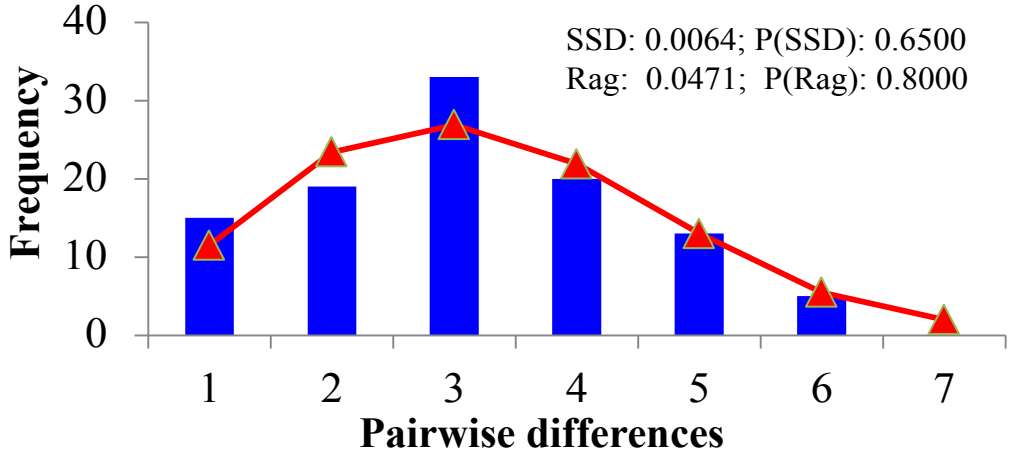

Nucleotide Diversity (π)
0.0012
0.0008
0.0004
0.
0
10000
20000
30000
ORF1ab
S
ORF3a
E
M
ORF6
ORF7a
ORF7b
ORF8
N
ORF10
Nucleotide


Pi
$10^{-3}$
$10^{-3.5}$
$10^{-4}$
NA
Hap_1 Hap_2 Hap_3 Hap_4 Hap_5 Hap_6 Hap_7 Hap_8 Hap_9 Hap_10 Hap_11 Hap_12 Hap_13 Hap_14 Hap_15 Hap_16 Hap_17 Hap_18 Hap_19 Hap_20 Hap_21 Hap_22 Hap_23 Hap_24 Hap_25 Hap_26 Hap_27 Hap_28 Hap_29 Hap_30 Hap_31 Hap_32 Hap_33 Hap_34 Hap_35 Hap_36 Hap_37 Hap_38 Hap_39 Hap_40 Hap_41 Hap_42 Hap_43 Hap_44 Hap_45 Hap_46 Hap_47 Hap_48 Hap_49 Hap_50 Hap_51 Hap_52 Hap_53 Hap_54 Hap_55 Hap_56 Hap_57 Hap_58 Hap_59 Hap_60 Hap_61 Hap_62 Hap_63 Hap_64 Hap_65 Hap_66 Hap_67 Hap_68 Hap_69 Hap_70 Hap_71 Hap_72 Hap_73 Hap_74 Hap_75 Hap_76 Hap_77 Hap_78 Hap_79
Hap_80 Hap_79 Hap_78 Hap_77 Hap_76 Hap_75 Hap_74 Hap_73 Hap_72 Hap_71 Hap_70 Hap_69 Hap_68 Hap_67 Hap_66 Hap_65 Hap_64 Hap_63 Hap_62 Hap_61 Hap_60 Hap_59 Hap_58 Hap_57 Hap_56 Hap_55 Hap_54 Hap_53 Hap_52 Hap_51 Hap_50 Hap_49 Hap_48 Hap_47 Hap_46 Hap_45 Hap_44 Hap_43 Hap_42 Hap_41 Hap_40 Hap_39 Hap_38 Hap_37 Hap_36 Hap_35 Hap_34 Hap_33 Hap_32 Hap_31 Hap_30 Hap_29 Hap_28 Hap_27 Hap_26 Hap_25 Hap_24 Hap_23 Hap_22 Hap_21 Hap_20 Hap_19 Hap_18 Hap_17 Hap_16 Hap_15 Hap_14 Hap_13 Hap_12 Hap_11 Hap_10 Hap_9 Hap_8 Hap_7 Hap_6 Hap_5 Hap_4 Hap_3 Hap_2

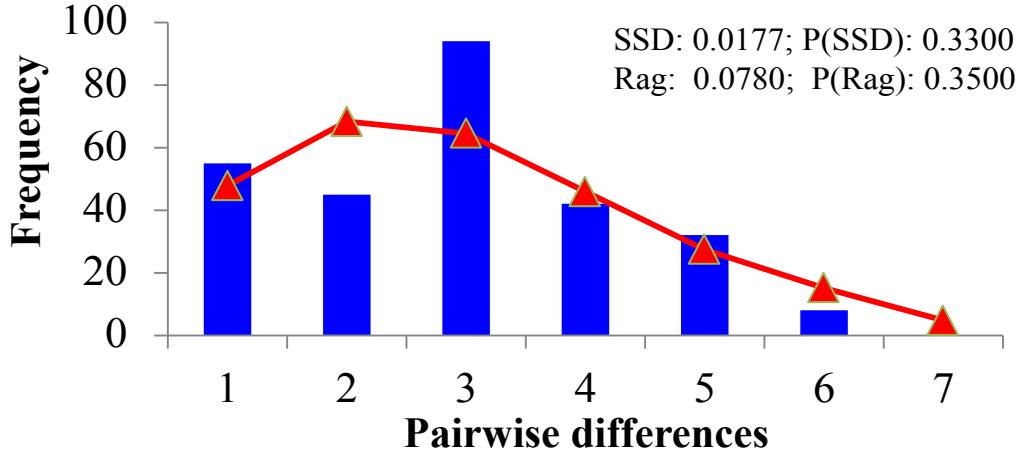

100
MG772933.1 Bat SARSlike coronavirus isolate batSLCoVZC45
MG772934.1 Bat SARSlike coronavirus isolate batSLCoVZXC21
100
100
100
EPI ISL 410538|2017 BetaCoV/pangolin/Guangxi/P4L/2017
EPI ISL 410542|2017 BetaCoV/pangolin/Guangxi/P2V/2017
EPI ISL 410543|2017 BetaCoV/pangolin/Guangxi/P3B/2017
EPI ISL 410541|2017 BetaCoV/pangolin/Guangxi/P5E/2017
EPI ISL 410540|2017 BetaCoV/pangolin/Guangxi/P5L/2017
EPI ISL 410539|2017 BetaCoV/pangolin/Guangxi/P1E/2017
100
EPI ISL 410544|2019 BetaCoV/pangolin/Guangdong/P2S/2019
EPI ISL 410721|2019 BetaCoV/pangolin/Guandong/1/2019
100
EPI ISL 402131|20130724 BetaCoV/bat/Yunnan/RaTG13/2013
100
100
To outgroups
SARS-COV-2（146 genome）
0.05


(A)
H1
H2
H3
H4
H5
H6
H7
H8
H9
H10
(B)
H1
H2
H3
H4
H5
H6
H7
H8
H9
H10
H11
H12
H13
H14
(C)
bat-RaTG13
800-1200 mutations
25
6
mv1
mv2
mv3
mv4
mv5
mv6
H1
H2
H3
H4
H5
H6
H7
H8
H9
H10
H11
H12
H13
H14
H15
H16
H17
H18
H19
H20
H21
H23
H24
H25
H26
H27
H28
H29
H30
H31
H32
H33
H34
H35
H36
H37
H38
H39
H40
H41
H42
H43
H44
H45
H46
H47
H48
H49
H50
H51
H52
H53
H54
H55
H56
H57
H58
H59
H60
H61
H62
H63
H64
H65
H66
H67
H68
H69
H70
H71
H72
H73
H74
H75
H76
H77
H78
H79
H80
Legends
China
Hubei
Guangdong
Zhejiang
Jiangxi
Shandong
Jiangsu
Chongqing
Sichuan
Fujian
Yunnan
Taiwan
Beijing
Other Countries
Japan
Korea
Italy
Singapore
United States
Nonthaburi
Australia
Cambodia
South Korea
Sweden
Belgium
Nepal
Germany
France
England

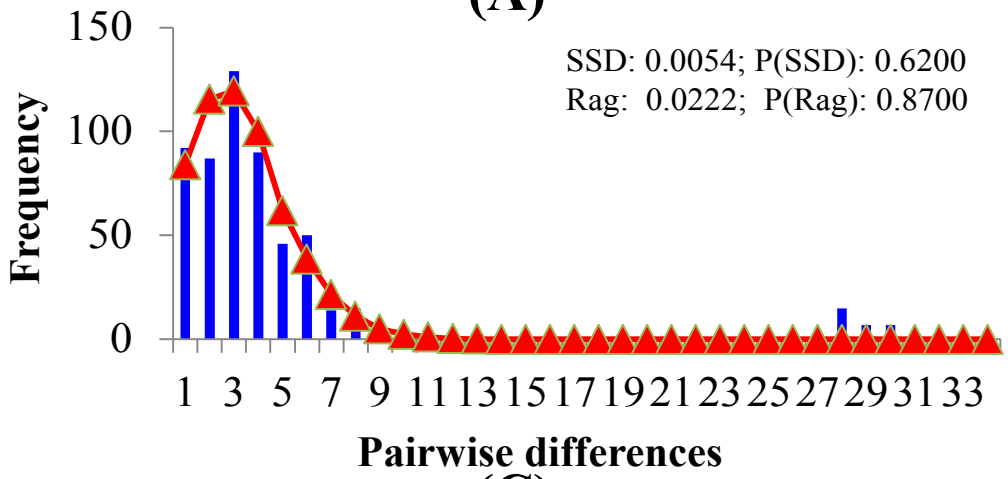

6
4
2
0
0
100
200
300
400
500
600
700
800
900
Time (days before Feb. 11, 2020)


SSD: 0.0064; P(SSD): 0.6500
Rag: 0.0471; P(Rag): 0.8000
Frequency
Pairwise differences

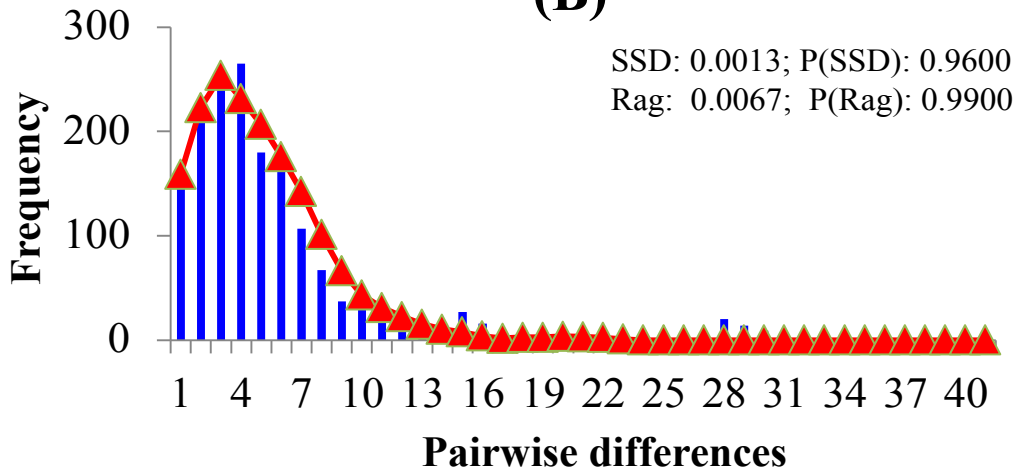
(A)


SSD: 0.0177; P(SSD): 0.3300
Rag: 0.0780; P(Rag): 0.3500
Frequency
Pairwise differences

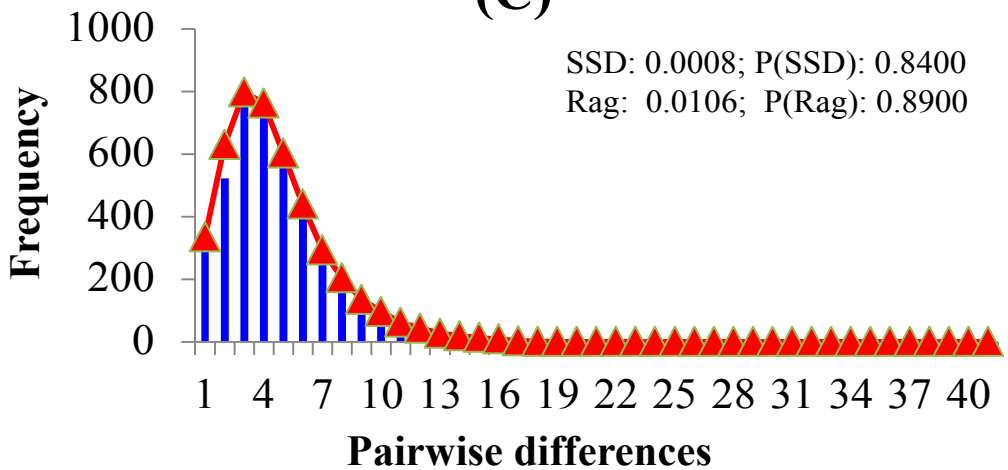
(B)


SSD: 0.0054; P(SSD): 0.6200
Rag: 0.0222; P(Rag): 0.8700
Frequency
Pairwise differences

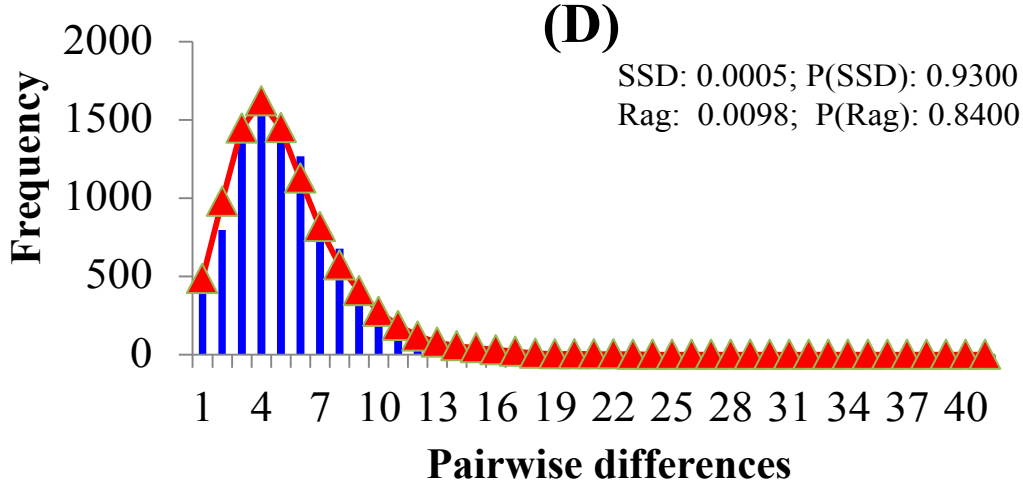
(C)


SSD: 0.0013; P(SSD): 0.9600
Rag: 0.0067; P(Rag): 0.9900
Frequency
Pairwise differences

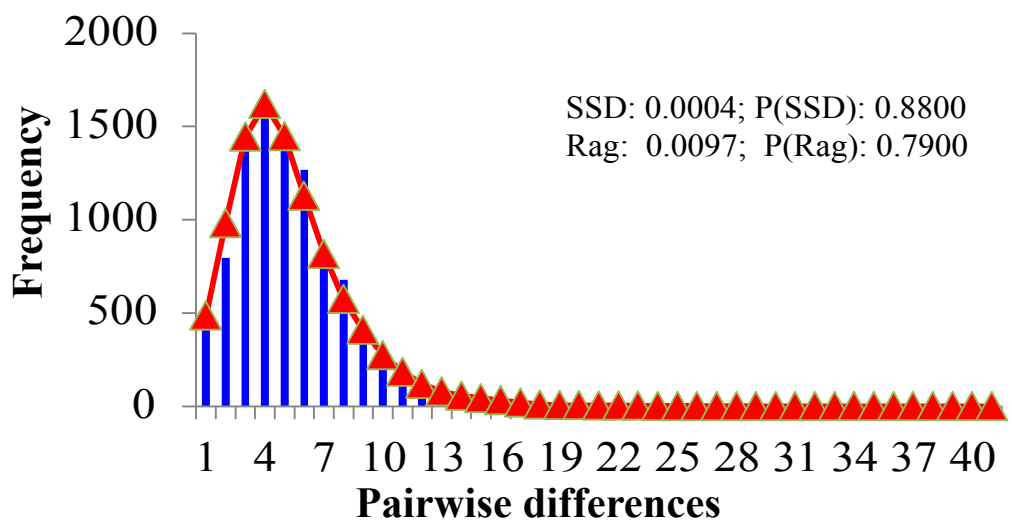
(D)


SSD: 0.0008; P(SSD): 0.8400
Rag: 0.0106; P(Rag): 0.8900
Frequency
Pairwise differences

(E)


SSD: 0.0005; P(SSD): 0.9300
Rag: 0.0098; P(Rag): 0.8400
Frequency
Pairwise differences

(F)


SSD: 0.0004; P(SSD): 0.8800
Rag: 0.0097; P(Rag): 0.7900
Frequency
Pairwise differences

(G)