Contents lists available at ScienceDirect

Biosafety and Health

journal homepage: www.elsevier.com/locate/bsheal

Lack of evolutionary changes identified in SARS-CoV-2 for the re-emerging outbreak of COVID-19 in Beijing, China

Yang Li [a,b,1], Yunjun Zhang [c,1], Mifang Liang [a], Yi Zhang [a,*], Xuejun Ma [a,*], Yong Zhang [a,2,*], Xiaohua Zhou [c,d,e,*]

[a] NHC Key Laboratory of Medical Virology and Viral Diseases, National Institute for Viral Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 102206, China
[b] Chongqing School, University of Chinese Academy of Sciences (UCAS), Chongqing 400020, China
[c] Department of Biostatistics, School of Public Health, Peking University, Beijing 100876, China
[d] Beijing International Center for Mathematical Research, Peking University, Beijing 100876, China
[e] Center for Statistical Science, Peking University, Beijing 100876, China

ARTICLE INFO

Article history:
Received 8 September 2021
Revised 22 December 2021
Accepted 22 December 2021
Available online 25 December 2021

Keywords:
Molecular clock
Frozen virus
Leaf-dating
Bayes factors
SARS-CoV-2

ABSTRACT

Although significant achievements have shown that the coronavirus disease 2019 (COVID-19) resurgence in Beijing, China, was initiated by contaminated frozen products and transported via cold chain transportation, international travelers with asymptomatic symptoms or false-negative nucleic acid may have another possible transmission mode that spread the virus to Beijing. One of the key differences between these two assumptions was whether the virus actively replicated since, so far, no reports showed viruses could stop evolution in alive hosts. We studied severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) sequences in this outbreak by a modified leaf-dating method with the Bayes factor. The numbers of single nucleotide variants (SNVs) found in SARS-CoV-2 sequences were significantly lower than those called from B.1.1 records collected at the matching time worldwide ($P = 0.047$). In addition, results of the leaf-dating method showed ages of viruses sampled from this outbreak were earlier than their recorded dates of collection (Bayes factors > 10), while control sequences (selected randomly with ten replicates) showed no differences in their collection dates (Bayes factors < 10). Our results which indicated that the re-emergence of SARS-CoV-2 in Beijing in June 2020 was caused by a virus that exhibited a lack of evolutionary changes compared to viruses collected at the corresponding time, provided evolutionary evidence to the contaminated imported frozen food should be responsible for the reappearance of COVID-19 cases in Beijing. The method developed here might also be helpful to provide the very first clues for potential sources of COVID-19 cases in the future.

1. Introduction

A super-spreading event of COVID-19 outbreak at Xinfadi (XFD) market in Beijing in June 2020 was supposed to be caused by contaminated imported frozen food. However, this hypothesis resulted in critical issues: 1) Could SARS-CoV-2 be transmitted through environment-to-human transmission? 2) Would infectivity of SARS-CoV-2 be reduced after transportation and storage associated with international cold-food logistics? 3) Could the "frozen" features be observed among those SARS-CoV-2 genomes sequenced from this outbreak?

Significant achievements had been made. Pang and colleagues provided molecular evidence (ancestral sequences were circulating in Europe [B.1.1 lineage]) and epidemiological investigations to conclude that environment-to-human transmission originated from contaminated imported food should be responsible for the COVID-19 resurgence in Beijing [1]. In addition, SARS-CoV-2 was successfully isolated from the imported frozen cod package surface while cytopathic effects (CPE) were observed from Vero-E6 cells inoculated with the isolated virus [2]. Therefore, the first two concerns have been well addressed, while the third remained obscure. Importantly, answers to the third question could provide insights into whether the international travelers with asymptomatic symptoms or false-negative of nucleic acid test spread the virus into Beijing [3].

We previously reported those sequences found in XFD were "older" than Europe's viruses collected at the matching time [4], leading to the "frozen evolution" virus hypothesis. A typical character of frozen viral

[*] Corresponding authors: NHC Key Laboratory of Medical Virology and Viral Diseases, National Institute for Viral Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 102206, China (Y. Zhang, X. Ma, Y. Zhang); Department of Biostatistics, School of Public Health, Peking University, Beijing 100876, China (X. Zhou).

E-mail addresses: zhangyicdc@126.com (Y. Zhang), maxj@ivdc.chinacdc.cn (X. Ma), yongzhang75@sina.com (Y. Zhang), azhou@math.pku.edu.cn (X. Zhou).

[1] These authors contributed to this article equally.

[2] Given his role as Editorial Board Member, Yong Zhang had no involvement in the peer-review of this article and had no access to information regarding its peer-review. Full responsibility for the editorial process for this article was delegated to Editor Di Qu.

https://doi.org/10.1016/j.bsheal.2021.12.001

isolates showed no accumulated mutations while in storage. Given that phylogenetic and evolutionary analyses had been performed to prove "frozen" virus as a potential cause of arbovirus re-emergence in France [5], here we aimed to conduct similar investigations to explore whether XFD SARS-CoV-2 sequences presented "frozen" genomic features.

2. Material and methods

2.1. SARS-CoV-2 sequences

We firstly collected the SARS-CoV-2 sequences from XFD during June 2020 from GISAID (https://www.gisaid.org, 5 records, Accession ID: EPI_ISL_3154875, EPI_ISL_469254, EPI_ISL_469255, EPI_ISL_469256 and EPI_ISL_850948) and Genome Warehouse (https://ngdc.cncb.ac.cn/gwh/, 3 records, Accession ID: GWHANPA 01000001, GWHANPB01000001 and GWHANPC01000001) (Access Date: 21/Oct/2021) (Table S1). Given that the genomes found in this outbreak belonged to the B.1.1 lineage, which was previously circulating outside of China, it should be reasonable to compare those XFD genomes to those records sampled from all over the world at the corresponding time. Therefore, we built a collection including all the B.1.1 sequences regardless of their location, collected during June 2020 (B.1.1 collection) (Table S1). After that, the sequences with < 15 ambiguous N bases in the genome were kept, with 2,355 out of 3,541 remaining (Table S1).

2.2. Single nucleotide variant (SNV) calling

We aligned the XFD SARS-CoV-2 sequences using Burrows-Wheeler Aligner (BWA) [6], regarding the official sequence of SARS-CoV2 (NC_045512.2) as the reference genome. After the alignments, BAM files were sorted then using SAMtools [7].

Command line used in alignment:

```
bwa mem NC_045512.2.fa $seqID.fa | samtools sort -O BAM -o $seqID.sorted.bam
```

Next. the sorted BAM files were analyzed with Bcftools [7] to generate variant call format (VCF) files using the command line:

```
bcftools mpileup -f NC_045512.2.fa $seqID.sorted.bam | bcftools call -c -v -A -o $seqID.bcftool.vcf
```

2.3. Leaf-dating with Bayes factor

The Leaf-dating method was developed to estimate unknown sequences ages [5]. Here, we regarded both XFD and part sequences from B.1.1 collection as the unknown ones to calculate their computational collection time ($Date_E$) through leaf-dating with BEAST v2.6.2 [8]. Briefly, the background sequences with known collection time were randomly selected (seven sequences per week) from lineage B.1.1 collection. The HKY85 nucleotide substitution model with Gamma distributed rate variation applied a strict clock model and exponential population growth. The priors of the sequences to estimate their $Date_E$ were defined with a uniform prior from 1 January 2020 to 30 November 2020. The rest parameters of priors were described in the previous study [9]. The chain length was set to 100 million states with a 10% burn-in. Convergence was evaluated using Tracer v1.7.1 [10]. Ten replicates of Leaf-dating with the Bayes factor were implemented.

Next, the Bayes factor, which was to test the discrimination between the $Date_E$ and recorded collection date ($Date_R$) of each sequence, was calculated with the Savage-Dickey ratio [11] based on the prior and posterior distribution of sequences ages generated from the Leaf-dating method. The interpretation of the Bayes factor was guided as follows, where a Bayes factor of at least 10 indicated "strong" support for $Date_E \neq Date_R$, a value of 3.2 showed "positive" support for $Date_E \neq Date_R$, a value of 1 indicated "not worth" support for $Date_E \neq Date_R$ [12].

2.4. Statistic analysis

Mann-Whitney *U* test was applied on non-normally distributed variables, with a two-tailed $P < 0.05$ defined as statistically significant.

3. Results

To determine whether the SARS-CoV-2 genomes sequenced from XFD in Beijing were intrinsically differed from the B.1.1 collection, XFD sequences and B.1.1 records were analyzed for single nucleotide variants (SNVs). Firstly, it should be noted that all the XFD genomes shared the mutations 28881G > A, 28882G > A, 28883G > C those had been considered as typical molecular features for B.1.1 European lineage (https://cov-lineages.org/) (Fig. 1). The differences in evolutionary patterns in SARS-CoV-2 with the different hosts could also be observed. In addition to common mutations, viruses from the environment had mutations of 11910A > G, 29868G > A, 29874A > G, and viruses from humans tolerated mutations of 2560A > C, 12085C > T, 23282G > T, and 24621C > T. It has been reported that SNVs (C > T) were the leading group of changes and could derive from APOBEC-mediated C-to-U deamination in human [13]. In addition, those two sequences (Beijing_BJ0617-01 and Beijing/IVDC-02–06) were identical. Next, we observed the median (interquartile range [IQR]) of mutations found in XFD sequences were 10 [8.5–10], significantly lower than that called from B.1.1 records worldwide collected during 10 June 2020 and 18 June 2020 (10 [8.5–10] vs. 11 [9–13], $P = 0.047$), especially in Asia group (Fig. 2, Figure S1).

We performed the leaf-dating method over all the XFD records and B.1.1 collection. As some of the sequences in the B.1.1 collection showed an uncommon number of SNVs (≤7 or ≥ 17), the B.1.1 collection was divided into three groups: low mutations (≤7 SNVs), high mutations (≥17 SNVs), and regular controls. Ten replicates of Bayesian phylodynamic inferences of XFD sequences, low mutations, high mutations, and regular controls with the fixed sampling date of 11 June 2020 (randomly selecting 3, 3, and 5 sequences, respectively) were performed to estimate the ages of sequences by BEAST 2.6.2. The outputs were further subject to python to calculate the Bayes factor of each record. The estimated dates ($Date_E$) did not contain the recorded dates ($Date_R$) in XFD genomes and vice versa in both low mutations and normal controls (Fig. 3), which might indicate a problem with the true age [5]. We next calculated the Bayes factors over these outputs to test this hypothesis. The results showed the XFD virus causing the outbreak in Beijing could be earlier than its $Date_R$ (Bayes factor > 10). In contrast, groups of low mutation and regular control showed no significant difference to their $Date_R$ (Bayes factor < 10). Our results were robust to different sampling datasets (Fig. 4).

In addition, the ages of SARS-CoV-2 with high mutations (≥17 SNVs) might be later than its $Date_R$ (Figure S2). It was as expected that sequences with high mutations showed delays in $Date_R$. The mechanism behind such high mutations accumulated in such a short time needed to be studied in the future.

4. Discussion and conclusion

One of the significant conclusions about the potential source of the XFD outbreak was that the outbreak originated from a seafood booth contaminated by SARS-CoV-2 based on the epidemiological data [1]. The related facts led to the hypothesis that the virus triggered the XFD outbreak: 1) was imported by the infected patient and somehow contaminated the frozen food; 2) was imported by the frozen food and spread in Beijing [3]. The hypothesis stated above could be simplified to whether the frozen food was contaminated before or after its appearance in China. This problem was equivalent to whether the

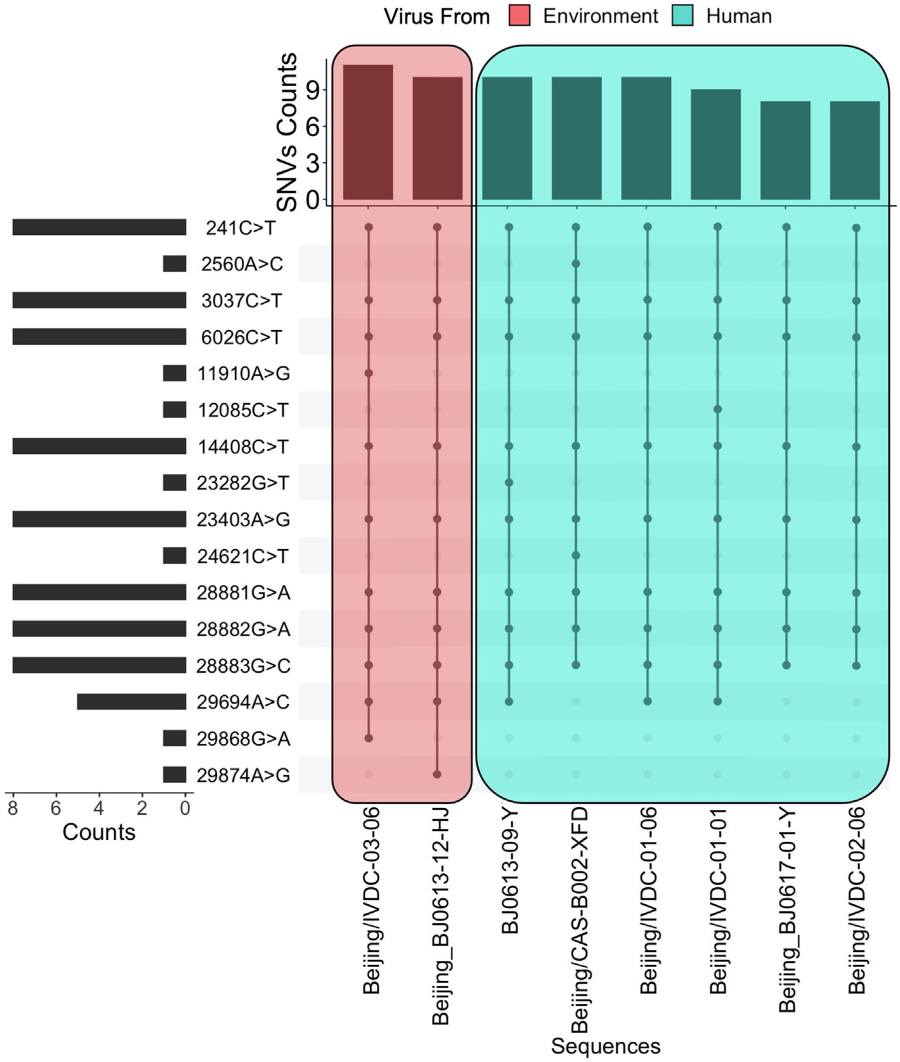

Fig. 1. Single nucleotide variants (SNVs) called from SARS-CoV-2 sequences sampled from COVID-19 outbreak in Beijing in June 2020. The reference was the official sequence of SARS-CoV2 (NC_045512.2).

virus was frozen or not. Based on the modified leaf-dating method developed here, we showed the re-emergence of SARS-CoV-2 in Beijing in June 2020 was caused by a virus that exhibits a lack of evolutionary changes compared to viruses collected at the corresponding time (Figs. 2–4). In other words, we did reveal the “frozen” genomic features in SARS-CoV-2 sequences found in the COVID-19 outbreak at Xinfadi market in Beijing in June 2020.

Although frozen viral isolate would not accumulate the mutations while in storage, it should be noted that the viral strain with low mutations could be necessary but not sufficient condition to “frozen virus.” Our results from the modified Leaf-dating method demonstrated that Date_E of the sequences with low mutations (≤ 7 SNVs) showed no differences to Date_R (Bayes factors < 10) (Fig. 2, 3). In other words, “frozen virus” showed fewer SNVs and more complex evolutionary features (e.g., mutation position, substitution pattern, etc.), which required further studies. Thus, the SARS-CoV-2 in the XFD outbreak in Beijing showed a lack of evolutionary changes.

The modified leaf-dating method proposed in this study could provide a quantitative way through the Bayes factor to show the discrimination between the computational age (estimated from leaf-dating) and collection date with two steps. In the first step, the ages of target sequences (e.g., XFD sequences and control sequences in this study) were assumed to be unknown and estimated through the original leaf-dating method [5]. The second step was to test the gap between the computational age (Date_E) and recorded collection date (Date_R) of target sequences. Finally, the Bayes factor was applied and calculated through the Savage-Dickey ratio. The codes were available at https://github.com/yunPKU/BayesFactorCalculation.

We were noted that this method carried out in this study could provide alternative insights to identify the possible source of re-emergency of COVID-19 cases. Although this study focused on the last question of the potential source of the XFD outbreak, this method could deliver the very first clues in cases of re-emergency of COVID-19 in the future. For example, a resurgence of COVID-19 cases in Anhui Province in East China and Northeast China’s Liaoning Province raised public alarms in May 2021. Furthermore, after obtaining the complete genomes of SARS-CoV-2, we found out those virus genomes belonged to lineage B.1.1.317 exhibited a lack of evolutionary changes (Manuscript under preparation). In the meantime, we were also introducing this method to disease prevention and control centers at all levels in China.

There were some limitations in this study. First, given that the power of molecular clock analysis could be reduced when the study period was limited from years/ decades to only months (in this study), the specific how earlier the XFD sequences than expected might be biased and misleading when the Date_E could be large of uncertainty.

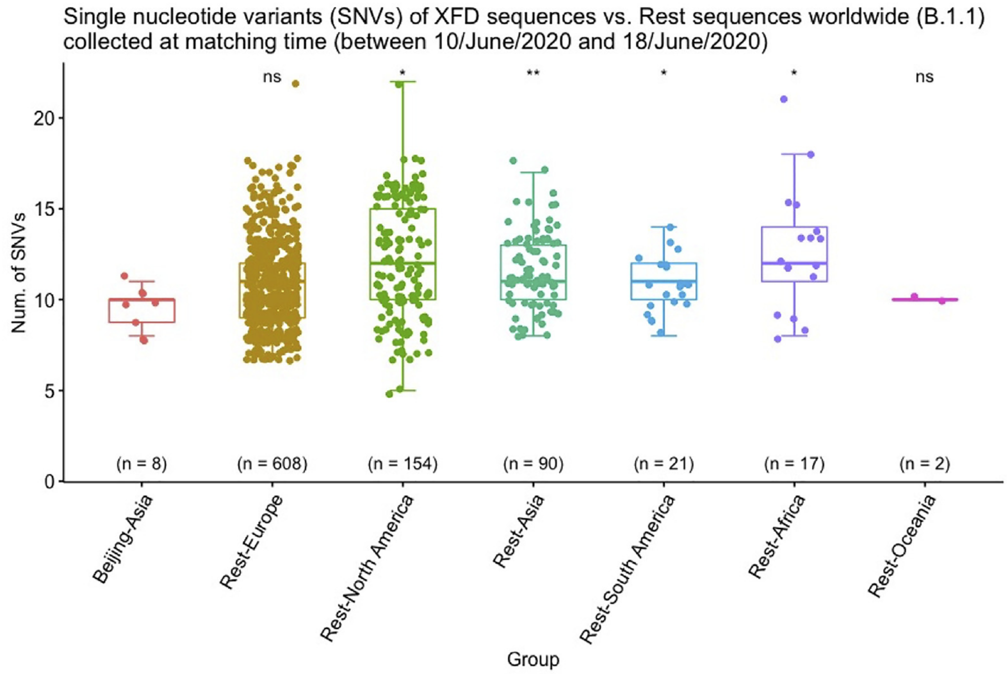

Fig. 2. Median of SNVs in SARS-CoV-2 genomes. Each dot represented a sequence. The reference group in the comparison was Beijing-Asia. *P*-value was determined by the Mann-Whitney *U* test for two-group comparisons with median reported, *: *P*-value < 0.05; **: *P*-value < 0.01; ***: *P*-value < 0.001; ns: *P*-value > 0.05.

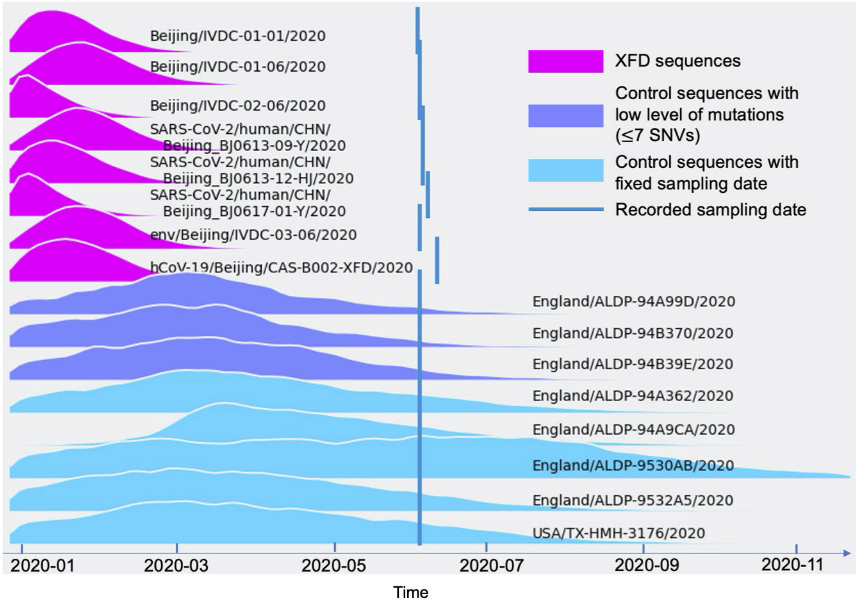

Fig. 3. An example of sampling datasets showed the estimated ages ($Date_E$) of SARS-CoV-2 sequences from the leaf-dating method were compared to recorded dates ($Date_R$).

Therefore, we applied the test ($Date_E \neq Date_R$) with the Bayes factor instead of a specific $Date_{E.}$ Secondly, the method could not be used for the virus that underwent recombination.

Border control and quarantine have effectively prevented the spread of SARS-CoV-2 by infected travelers in China. However, strict strategies of the monitor for imported goods, especially those cold-chained products, need to be developed accordingly, to prevent the potential secondary extensive outbreak in this country, while emerging variants of SARS-CoV-2, such as Delta variant, were still spreading worldwide.

Acknowledgements

We appreciate the helpful discussion with Dr. Tao Zhao from Nvidia and Prof. Li Li from Southwest University. This work was supported by the National Natural Science Foundation of China (Grant number: 82041023), the Bill & Melinda Gates Foundation (Grant number: INV-016826), and China Mega-Projects for Infectious Disease (2018ZX10711001, 2017ZX10104001). However, the funders had no role in study design, data collection, analysis, decision to publish, or manuscript preparation.

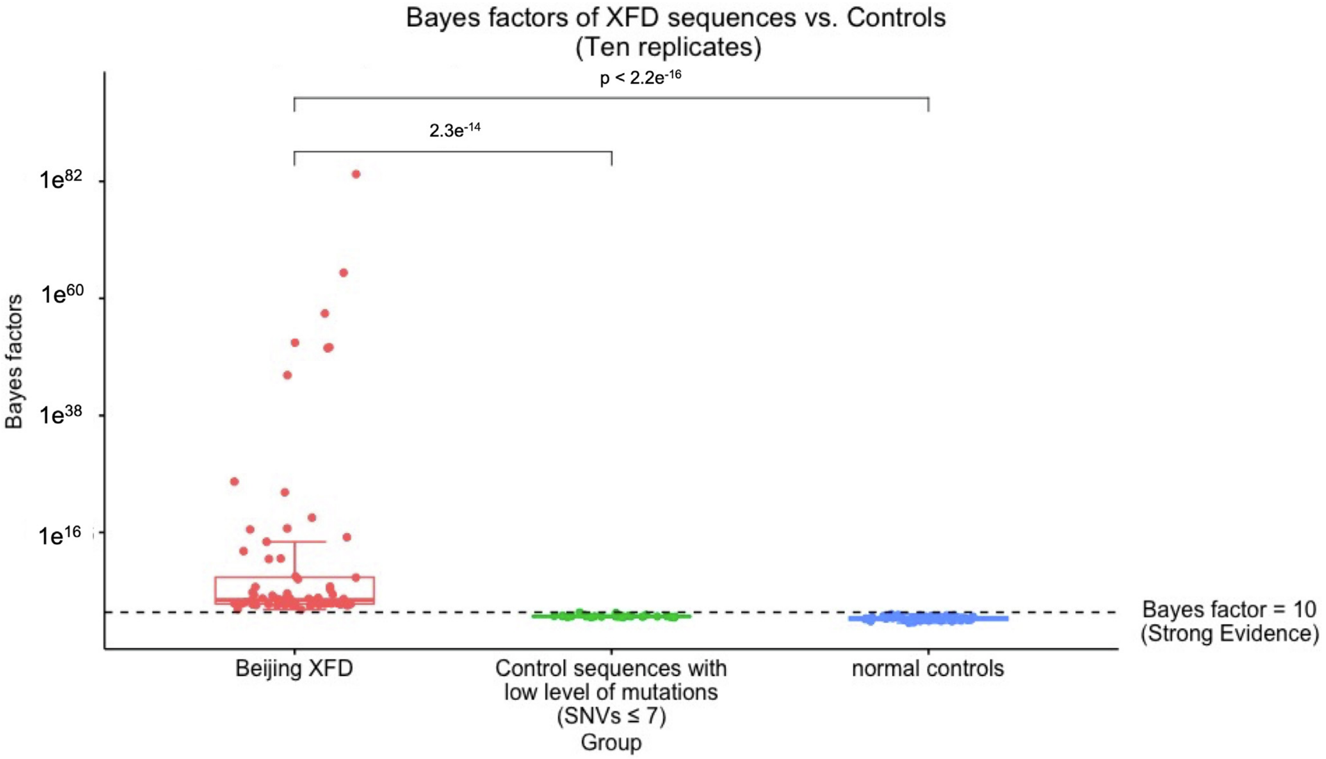

Fig. 4. Comparison of Bayes factors between sequences from XFD and B.1.1 collection based on ten replicates of the leaf-dating method. Each dot represented a sequence. Bayes factor is defined as the marginal likelihood of Hypothesis$_2$ (H_2): $Date_E \neq Date_r$ over that of H_1: $Date_E = Date_r$. The Bayes factor suggested strong evidence in favor of H_2 if being >10. *P*-value was determined by the Mann-Whitney *U* test for two-group comparisons with median reported.

Conflict of interest statement

The authors declare that there are no conflicts of interest.

Author contributions

Yang Li: Data Curation, Formal Analysis, Visualization, Writing – Original Draft. **Yunjun Zhang:** Investigation, Formal Analysis, Visualization, Writing – Original Draft. **Mifang Liang:** Investigation. **Yi Zhang:** Investigation, Writing – Review & Editing. **Xuejun Ma:** Investigation, Writing – Review & Editing. **Yong Zhang:** Conceptualization, Supervision, Methodology, Writing – Review & Editing. **Xiaohua Zhou:** Conceptualization, Supervision, Methodology, Writing – Review & Editing.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bsheal.2021.12.001.

References

[1] X. Pang, L. Ren, S. Wu, W. Ma, G. Li, M. Li, J. Wang, Y. Huang, J. Wang, Cold-chain food contamination as the possible origin of COVID-19 resurgence in Beijing, Natl. Sci. Rev. 7 (12) (2020) 1861–1864, https://doi.org/10.1093/nsr/nwaa264.

[2] P. Liu, M. Yang, X. Zhao, Y. Guo, L. Wang, J. Zhang, W. Lei, W. Han, F. Jiang, W.J. Liu, G.F. Gao, G. Wu, Cold-chain transportation in the frozen food industry may have caused a recurrence of COVID-19 cases in destination: Successful isolation of SARS-CoV-2 virus from the imported frozen cod package surface, Biosaf. Health 2 (4) (2020) 199–201, https://doi.org/10.1016/j.bsheal.2020.11.003.

[3] D. Normile, Source of Beijing's big new COVID-19 outbreak is still a mystery, Science (2020), https://doi.org/10.1126/science.abd3890.

[4] Y. Zhao, China's CDC experts investigate Xinfadi market three times, announce groundbreaking virus tracing discovery. https://www.globaltimes.cn/content/1192146.shtml, 2020 (accessed 14 August 2021).

[5] B. Shapiro, S.Y. Ho, A.J. Drummond, M.A. Suchard, O.G. Pybus, A. Rambaut, A Bayesian phylogenetic method to estimate unknown sequence ages, Mol. Biol. Evol. 28 (2) (2011) 879–887, https://doi.org/10.1093/molbev/msq262.

[6] H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Preprint], arXiv (2013) https://doi.org/arXiv:1303.3997.

[7] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and SAMtools, Bioinformatics 25 (16) (2009) 2078–2079, https://doi.org/10.1093/bioinformatics/btp352.

[8] R. Bouckaert, T.G. Vaughan, J. Barido-Sottani, S. Duchêne, M. Fourment, N.F. Müller, H.A. Ogilvie, C. Zhang, T. Stadler, A.J. Drummond, M. Pertea, et al, An advanced software platform for Bayesian evolutionary analysis, PLoS Comput. Biol. 15 (4) (2019), e1006650. https://doi.org/10.1371/journal.pcbi.1006650.

[9] J. Scire, T.G. Vaughan, T. Stadler, Evolutionary & epidemiological analysis of 93 genomes. https://virological.org/t/evolutionary-epidemiological-analysis-of-93-genomes/405, 2020 (accessed 15 August 2021).

[10] A. Rambaut, A.J. Drummond, D. Xie, G. Baele, M.A. Suchard, Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7, Syst. Biol. 67 (5) (2018) 901–904, https://doi.org/10.1093/sysbio/syy032.

[11] E.J. Wagenmakers, T. Lodewyckx, H. Kuriyal, R. Grasman, Bayesian hypothesis testing for psychologists: a tutorial on the Savage-Dickey method, Cogn. Psychol. 60 (3) (2010) 158–189, https://doi.org/10.1016/j.cogpsych.2009.12.001.

[12] R.E. Kass, A.E. Raftery, Bayes factors, J. Am. Stat. Assoc. 90 (430) (1995) 773–795, https://doi.org/10.2307/2291091.

[13] S. Di Giorgio, F. Martignano, M.G. Torcia, G. Mattiuz, S.G. Conticello, Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2, Sci. Adv. 6 (25) (2020), eabb5813. https://doi.org/10.1126/sciadv.abb5813.