

Re-emergence of COVID-19 in Beijing was triggered by frozen virus: evidence from molecular clock

Yunjun Zhang¹, Yong Zhang², Mifang Liang², Yi Zhang^{2*}, Yang Li^{2*}, Xuejun Ma^{2*}, Xiaohua Zhou^{1,3,4*}

¹Department of Biostatistics, School of Public Health, Peking University, Beijing, China

²NHC Key Laboratory of Medical Virology and Viral Diseases, National Institute for Viral Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China

³Beijing International Center for Mathematical Research, Peking University, Beijing, China;

⁴Center for Statistical Science, Peking University, Beijing, China

*Corresponding author:

Email:

zhangyicdc@126.com(YZ), yeli7068@outlook.com(YL), maxj@ivdc.chinacdc.cn (XJM),
azhou@math.pku.edu.cn (XHZ)

ABSTRACT

The COVID-19 outbreak in Xinfadi (XFD) Wholesale market in Beijing, China in June, 2020 caused 368 reported cases within 39 days. Genetic evidences indicated that imported SARS-CoV-2 (belong to the lineage B1.1.29) initiated this outbreak. However, the transmission route of the virus is still unknown. We obtained from public database three SARS-CoV-2 genomes isolated in XFD (XFD genomes) and adopted the leaf-dating method to calculate their expected collection dates using temporal calibrating information from other 241 genomes collected in mainland of China. All three XFD genomes were calculated to have earlier collection dates than the recorded (Bayes factor >1), and hence exhibited a lack of genetic divergence. We additionally combined the XFD genomes with other 225 genomes subsampled from those of the lineage B1.1.29, among which five sequences were also included for control analysis. Two of three XFD genomes were calculated to have earlier collection dates than the recorded (Bayes factor >1), while no control genomes provided such evidence. According to present understanding of SARS-CoV-2, a lack of genetic divergence is most likely due to being frozen. Considering the fact that the XFD outbreak started from a booth of frozen food, we judged that the XFD outbreak was caused by contaminated frozen food. Our results provided molecular evidence for the source of COVID-19 outbreak in Beijing XFD, which highlights new targets for SARS-CoV-2 surveillance for the public health.

INTRODUCTION

A COVID-19 case was identified in Beijing on 2020-06-11, ending a run of 55 days without local transmission and starting the second outbreak. Since then, the city has responded quickly and fiercely with a series of control measures, such as large scale qRT-PCR screening, locking down affected residential compounds, and closing all schools [1]. With all these efforts, the second outbreak was taken under control within two weeks, and lasted only for 39 days (till 2020-07-20) with 368 qRT-PCR positive cases [2].

Most cases of the outbreak had a history of close contact with Xinfadi (XFD) Wholesale Market. Genetic tracing of SARS-CoV-2 genomes from XFD indicated that the outbreak was triggered by imported virus. Specifically it was a member of the lineage B1.1 (now of the lineage B1.1.29, [3]) predominately circulating in Europe rather than in China [4]. However, how the virus was

transmitted to Beijing has not been clarified. Evidences from epidemiological field investigation showed that a salmon booth (#S14) was the starting point [2], suggesting that the viruses could have been imported with contaminated frozen food. However, some held a belief that the virus was imported by infected human [1].

Molecular clock-based evolutionary analysis of temporally spaced genome data provides a new tool for timing past events and exploring the mechanism and process of evolution [5]. Generally, the molecular clock posits that viral genomes accumulate mutations at roughly a constant rate, and the process could be affected by certain environmental factors [6]. For example, frozen viral isolates do not accumulate genetic mutations while in storage [7]. This environmental factor, frozen in storage, could help test the contaminated seafood hypothesis. We therefore performed evolutionary analysis of the SARS-CoV-2 genome data of the isolates collected in the XFD Wholesale Market (denoted as XFD genomes). We evaluated the amount of genetic changes embedded in these genomes to test whether the contaminated seafood hypothesis worded for the import of virus leading to the second outbreak of COVID-19 in Beijing in 2020.

RESULTS

Three human-source SARS-CoV-2 genomes from the XFD outbreak were obtained from open database of GISAID ([8], accessed on 2021-01-30). Assuming that XFD genomes follow a certain evolutionary rate, the expected collection dates (denoted as $Date_E$) of XFD genomes were calculated with leaf-dating method based on the molecular clock. The virus founded in XFD outbreak belonged to the B1.1.29 lineage which was previously circulating outside of China. The rates were therefore generated with two datasets, one consisted of SARS-CoV-2 genomes sampled from the mainland of China (referred to as domestic dataset) and the other of the B1.1.29 lineage regardless of the region (referred to as lineage-specified dataset). Control analysis was conducted with five randomly selected SARS-CoV-2 genomes (referred to as control genomes) from the lineage of B1.1.29.

XFD genomes

Based on the domestic dataset, all three XFD genomes were expected to have earlier collection dates than their recorded collection (Fig 1). We calculated that the XFD genome IVDC-02-06

was sampled on 2020-03-11 (95%HPD: 2020-01-21, 2020-05-08), being 92 days earlier than its recorded collection on 2020-06-11. The consequent Bayes factor was 11.27, suggesting strong evidence in favour of the hypothesis (H_2 : $\text{Date}_E \neq \text{Date}_r$, defined in the Fig 1) that the expected collection date of the genome is unequal to the recorded date.

The other two genomes, IVDC-01-01 and IVDC-01-06 were calculated to be 68 and 49 days earlier than their recorded collections, respectively, providing moderate and weak level of bayes factor in favour of the hypothesis H_2 . Meanwhile, the evolutionary rate of SARS-CoV-2 genome was estimated to be 5.94×10^{-4} substitutions per site per year (95% HPD: 3.94×10^{-4} , 7.85×10^{-4}), which was consistent with the previous study [9].

Figure 1 Calculated and recorded collection dates of XFD genomes based on the domestic dataset. Bayes factor is defined as the marginal likelihood of H_2 : $\text{Date}_E \neq \text{Date}_r$ over that of H_1 : $\text{Date}_E = \text{Date}_r$, where Date_E is the expected collection date, and Date_r the recorded collection date of a particular genome. The bayes factor exceeds 1, 3, or 10, suggesting weak, moderate, and strong evidence in favor of H_2 , respectively.

Based on the lineage-specified dataset, two XFD genomes (IVDC-01-01 and IVDC-02-06) provided weak evidence in favor of hypothesis H_2 with the bayes factor 2.98 and 1.07, respectively (Fig 2). Compared with the results obtained from the domestic dataset, the calculated collection dates of the two XFD genomes were 11~17 days later (but still earlier than the recorded collection dates) and with broader 95%HPD.

Meanwhile, the evolutionary rate of SARS-CoV-2 genome was estimated to be 3.17×10^{-4} substitutions per site per year (95% HPD: 2.38×10^{-4} , 4.02×10^{-4}), which was overlapped with the estimation from domestic dataset.

Control genomes

In addition, the expected collection dates of the five control genomes were simultaneously calculated on the basis of lineage-specific dataset (Fig 2). Compared with those of the XFD genomes, the calculated collection dates of control genomes were consistently later and closer to

their recorded collection dates, leading to consistently smaller bayes factors and weak evidence in favor of the hypothesis ($H_1: \text{Date}_E = \text{Date}_r$) that the expected collection date of the genome is equal to its recorded date.

Figure 2 Calculated and recorded collection dates of XFD genomes and control genomes based on the lineage-specified dataset. The bayes factor suggests weak, moderate, and strong evidence in favor of H_1 if it is smaller than 1, 1/3, or 1/10, respectively.

DISCUSSIONS

The evolutionary analysis of SARS-CoV-2 genomes sampled in XFD revealed that the two XFD genomes (IVDC-02-06 and IVDC-01-1) lacked genetic divergence. If they followed the previous evolutionary rate, the calculated collection date of these two XFD genomes were earlier than the record date.

This lack of genetic divergence was attributed to frozen storage for the following reasons. First, SARS-CoV-2 biology suggests it be impossible for some viruses staying in a host or reservoir for several months with little or no replication before reactivation and a second outbreak [6]. Second, it has been reported that SARS-CoV-2 can survive the time and temperatures associated with transportation and storage conditions associated with international food trade [10]. Besides, epidemiological evidence revealed that the outbreak started from a booth of frozen food in the market [2]. Therefore, we hypothesized that the contaminated imported food carried the live virus from the supply chain, survived in the frozen storage, and transferred the virus to the XFD market, thus leading the outbreak.

In addition, when analyzing the control genomes from lineage B1.1.29 with the same collection date of 2020-06-11 as that of the two XFD genomes, their calculated collection dates skewed to the right (Fig 2). Four of five genomes had the peak (of the posterior distribution of the calculated collection date) later than the actual collection date. This bias arises from the used leaf-dating method which exhibited moderate bias when the true age was small [7]. But when this method was applied to the XFD genomes, they showed peaks clearly ahead of their recorded

collection dates. This further supported the XFD genomes being earlier position in the evolutionary process of SARS-CoV-2.

A real-world support was from a genome collected in Singapore (Singapore/487/2020 (EPI_ISL_476821)), which was identical to the XFD genome IVDC-02-06. The Singapore genome was collected from an international traveler on 2020-03-26, which was very close to our calculation (2020-03-22) of the XFD IVDC-02-06 with leaf-dating method. This lends proof to the validity of our estimation of collection date and provides further evidence that the genome should have appeared much earlier than the recorded 2020-06-11.

Previous studies have attributed the recurrence of infectious diseases to frozen viruses via phylogenetic and evolutionary analysis. Previous researches were based on either the phylogenetic clustering of viral genomes [11] or on the discontinuity of the evolutionary rates between the first and second outbreaks [6]. In this study, the leaf-dating method was used to estimate the collection time, and the Bayes factor was used to explore the gap between the actual collection time and the calculated collection time. A quantitative measure of the evidence for whether the virus being frozen and, if so, a composite estimation of the time length of being frozen were provided by our method. In addition, our method required less genome data from the second outbreak and could be performed independently for each genome collected in the second outbreak.

The conclusion of our study was made based on a relative homogeneous model setting: i.e., assuming the evolutionary rate be constant (strict molecular clock) and using the coalescent-based tree prior regardless of population structure. In recent years, it is more appealing to incorporate heterogeneity both in the molecular clock, such as relaxed molecular clock, and in the tree prior model, such as structured coalescent model. However, a previous study found that very similar estimation of collection date would be obtained under the strict or relaxed molecular model [7]. Hence our evidence for XFD genomes may not have been affected if adopting the relaxed molecular clock model. In addition, incorporating population structure into the tree prior model was demonstrated to considerably decrease the estimation of evolutionary rate [12]. This

effect of a slower rate, had we adopted population structure, would have strengthened our evidence for earlier expected collection date.

In conclusion, we conclude that the outbreak in XFD market in Beijing was caused by contaminated frozen food (or its packaging). Our findings, coupled with the reports from China of SARS-CoV-2 being detected on imported frozen food, should alert food safety authorities and the food industry where this virus could pose a non-traditional food safety risk.

MATERIALS AND METHODS

XFD genomes

The XFD genomes included three human-sourced SARS-CoV-2 genomes, namely, Beijing/IVDC-01-06/2020 (EPI_ISL_469254), Beijing/IVDC-02-06/2020 (EPI_ISL_469255), and Beijing/IVDC-01-01/2020 (EPI_ISL_850948) (or briefly as IVDC-01-06, IVDC-02-06, and IVDC-01-01). The genomes were sampled at the very beginning of the XFD outbreak, one (IVDC-01-01) sampled on 2020-06-10, and the other two (IVDC-01-06 and IVDC-02-06) on 2020-06-11.

Datasets

All the SARS-CoV-2 genomes in the background datasets were downloaded from the GISAID. We only analyzed human-sourced SARS-CoV-2 genomes with at least 29,000 bases and with complete date information.

The first dataset included the three XFD genomes and 241 genomes (referred to as domestic dataset) that were subsampled --22 genomes per division (the median genome size of divisions)-- from 893 genomes in mainland of China between 2019-12-24 (the time of the first reported genome) and 2020-07-15. The second dataset (referred to as lineage-specified dataset) included the three XFD genomes, the five control genomes (mentioned later), and 215 genomes were subsampled (10 genomes per week as in [4]) from 14,946 genomes in the lineage B1.1.29 till 2020-07-15. The second dataset also included the earliest two genomes in Wuhan China (Wuhan/WH01/2019(EPI_ISL_406798) and Wuhan/Hu-1/2019 (EPI_ISL_402125)) as the root

reference. The ending date (i.e., 2020-07-15) was chosen to keep the XFD genomes away from the boundary of the study period and to make accurate estimates of their collection dates [7].

All the genomes were processed using the Nextstrain [13] bioinformatics pipeline Augur to align genomes via MAFFT v7.4 [14] and masked spurious SNPs. The masked genomes from the Nextstrain bioinformatics pipeline (i.e., the ‘masked.fasta’) worked as input for the following leaf-dating analysis.

Leaf-dating method

The leaf-dating method was originally developed to estimate the collection date of a particular genome with unknown collection date [7]. In this study, the expected collection dates (referred to as calculated collection dates) of the XFD genomes were estimated with leaf-dating method and were further compared with the real collection dates (or recorded collection dates) so as to provide information for inference on possible import route. This analysis was performed with the BEAST2 software [15] and used a HKY85 nucleotide substitution model with Gamma distributed rate variation and used an exponential growth coalescent model for the tree prior and assumed the strict molecular clock. The collection dates of the XFD genomes were estimated in the MCMC with a uniform prior over the period from 2020-01-01 to 2020-07-15. All the other priors used in this analysis were the same as previous studies [16]. The corresponding xml files used in the analysis were provided as supplementary files.

Bayes factor based comparison between the expected collection date and the recorded

We introduced the bayes factor to determine if the expected collection date (denoted as $Date_E$) differed from the recorded collection date (denoted as $Date_r$) of a particular genome.

Specifically, we constructed two hypotheses as $H_1: Date_E = Date_r$ and $H_2: Date_E \neq Date_r$. The bayes factor of the hypothesis H_2 over the hypothesis H_1 (i.e., BF_{21}) was calculated with the method of Savage-Dickey ratio [17] on the basis of the used prior distribution of the collection date as well as its posterior distribution generated from the leaf dating method.

Control analysis

To further verify the validity of the analysis, we randomly selected five SARS-CoV-2 genomes from the lineage of B.1.1.29 for control analysis. They were CostaRica/INC-0055/2020 (EPI_ISL_512661), England/ALDP-952FFA/2020 (EPI_ISL_553677), England/MILK-56B89D/2020 (EPI_ISL_558380), England/MILK-957274/2020 (EPI_ISL_554214), and UnitedArabEmirates/1062/2020 (EPI_ISL_698137). These control genomes had the same recorded collection date of 2020-06-11, equal to that of two XFD genomes (IVDC-01-06 and IVDC-02-06). We incorporated the control genomes into the lineage-specified dataset, calculated their expected collection dates simultaneously with those of the XFD genomes, and made comparison with their recorded collection dates based on the bayes factor.

Supporting information

S1 Data. BEAST 2.6 XMLs and the related metadata for the two Bayesian analyses in a zipped folder.

FUNDING

This work was supported from 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). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

The authors have declared that no competing interests exist.

Data Availability

All relevant data are within the paper.

ACKNOWLEDGEMENTS

We appreciate helpful discussion with Dr. Tao ZHAO from Nvidia.

REFERENCES

- [1] D. Normile, "Source of Beijing's big new COVID-19 outbreak is still a mystery," *Science*, 2020.
- [2] X. Pang, L. Ren, S. Wu, W. Ma, J. Yang, L. Di, J. Li, Y. Xiao, L. Kang, S. Du, J. Du, J. Wang, G. Li, S. Zhai, L. Chen, W. Zhou, S. Lai, L. Gao, Y. Pan, Q. Wang, M. Li, J. Wang, Y. Huang and J. Wang, "Cold-chain food contamination as the possible origin of COVID-19 resurgence in Beijing," *National Science Review*, vol. 7, no. 12, pp. 1861-1864, 2020.
- [3] A. Rambaut, E. C. Holmes, Á. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. d. Plessis and O. G. Pybus, "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology," *Nature microbiology*, vol. 5, no. 11, pp. 1403-1407, 2020.
- [4] J. Yang, P. Niu, L. Chen, L. Wang, L. Zhao, B. Huang, J. Ma, S. Hu, L. Wu, G. Wu, C. Huang, Y. Bi and W. Tan, "Genetic tracing of HCoV-19 for the re-emerging outbreak of COVID-19 in Beijing, China," *Protein & Cell*, vol. 12, no. 1, pp. 4-6, 2021.
- [5] L. Bromham and D. Penny, "The modern molecular clock.," *Nature Reviews Genetics*, vol. 4, no. 3, pp. 216-224, 2003.
- [6] D. J. Pascall, K. Nomikou, E. Bréard, S. Zientara, A. d. S. Filipe, B. Hoffmann, M. Jacquot, J. B. Singer, K. D. Clercq, A. Bøtner, C. Sailleau, C. Viarouge, C. Batten, G. Puggioni, C. Ligios, G. Savini, P. A. v. Rijn, P. P. C. Mertens, R. Biek and M. Palmarini, "'Frozen evolution" of an RNA virus suggests accidental release as a potential cause of arbovirus re-emergence," *PLOS Biology*, vol. 18, no. 4, 2020.
- [7] B. Shapiro, S. Y. W. Ho, A. J. Drummond, M. A. Suchard, O. G. Pybus and A. Rambaut, "A Bayesian Phylogenetic Method to Estimate Unknown Sequence Ages," *Molecular Biology and Evolution*, vol. 28, no. 2, pp. 879-887, 2011.
- [8] Y. Shu and J. McCauley, "GISAI: Global initiative on sharing all influenza data - from vision to reality.," *Eurosurveillance*, vol. 22, no. 13, p. 30494, 2017.
- [9] S. Duchene, L. Featherstone, M. Haritopoulou-Sinanidou, A. Rambaut, P. Lemey and G. Baele, "Temporal signal and the phylodynamic threshold of SARS-CoV-2," *Virus Evolution*, vol. 6, no. 2, 2020.
- [10] D. Fisher, A. Reilly, A. K. E. Zheng, A. R. Cook and D. E. Anderson, "Seeding of outbreaks of COVID-19 by contaminated fresh and frozen food," *bioRxiv*, 2020.
- [11] M. Worobey, "Phylogenetic evidence against evolutionary stasis and natural abiotic reservoirs of influenza A virus," *Journal of Virology*, vol. 82, no. 7, pp. 3769-3774, 2008.
- [12] S. Möller, L. d. Plessis and T. Stadler, "Impact of the tree prior on estimating clock rates during epidemic outbreaks.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 16, pp. 4200-4205, 2018.
- [13] J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford and R. A. Neher, "Nextstrain: real-time tracking of pathogen evolution," *Bioinformatics*, vol. 34, no. 23, pp. 4121-4123, 2018.

- [14] K. Katoh and D. M. Standley, "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability," *Molecular Biology and Evolution*, vol. 30, no. 4, pp. 772-780, 2013.
- [15] R. Bouckaert, A. Drummond, A. Rambaut, M. Suchard, T. Vaughan and Y. Heled, "BEAST2: Bayesian evolutionary analysis sampling trees," 2014.
- [16] J. Scire, T. G. Vaughan and T. Stadler, "Evolutionary & epidemiological analysis of 93 genomes," Feb 2020. [Online]. Available: <https://virological.org/t/evolutionary-epidemiological-analysis-of-93-genomes/405>. [Accessed 2020].
- [17] E.-J. Wagenmakers, T. Lodewyckx, H. Kuriyal and R. Grasman, "Bayesian Hypothesis Testing for Psychologists: A Tutorial on the Savage-Dickey Method.," *Cognitive Psychology*, vol. 60, no. 3, pp. 158-189, 2010.

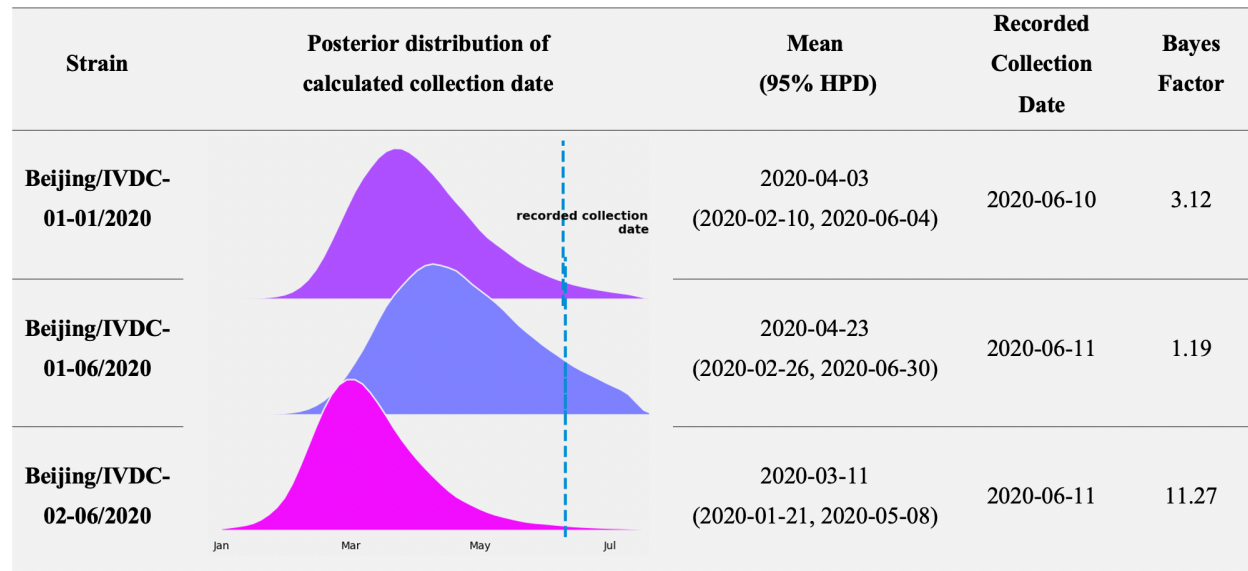


Figure 1 Calculated and recorded collection dates of XFD genomes based on the domestic dataset. Bayes factor is defined as the marginal likelihood of $H_2: \widehat{Date} \neq Date_r$ over that of $H_1: \widehat{Date} = Date_r$, where $\widehat{Date}$ is the calculated collection date, and $Date_r$ is the recorded collection date of a particular genome. The bayes factor exceeds 1, 3, or 10, suggesting weak, moderate, and strong evidence in favor of H_2 , respectively.

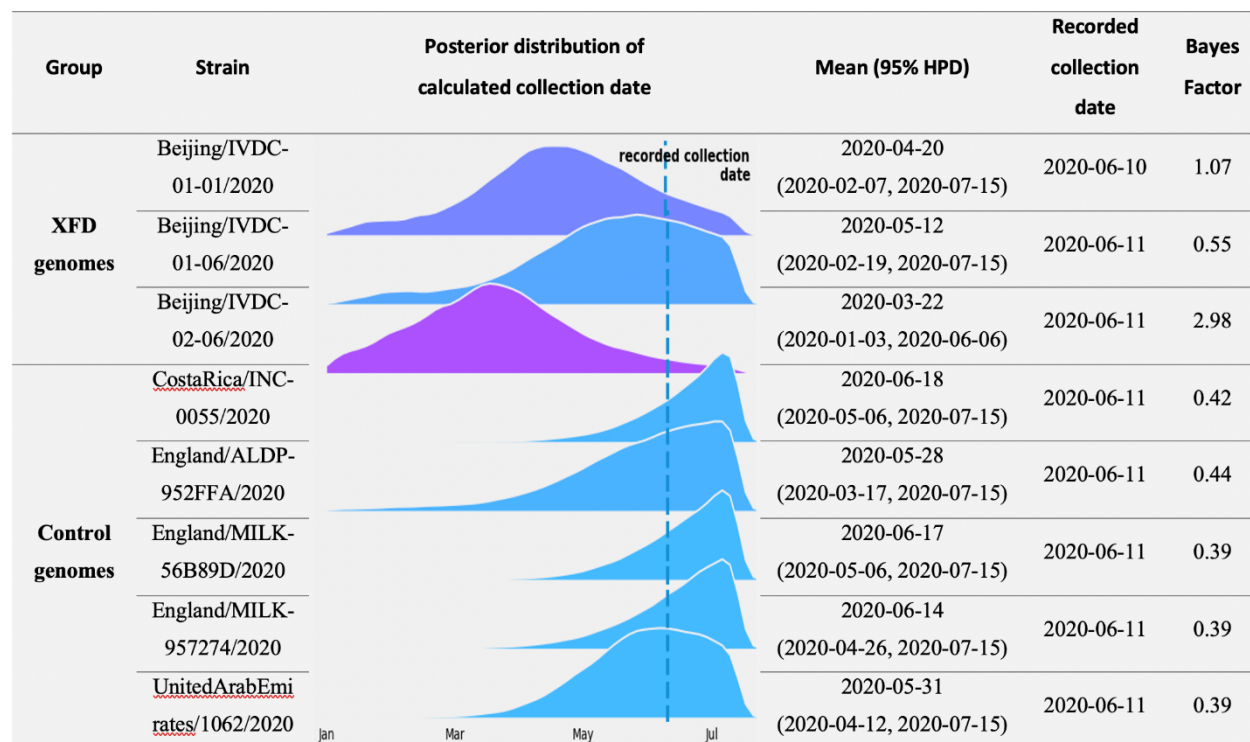


Figure 2 Calculated and recorded collection dates of XFD genomes and control genomes based on the lineage-specified dataset. The bayes factor suggests weak, moderate, and strong evidence in favor of H_1 if being smaller than 1, 1/3, or 1/10, respectively.