



## OPEN ACCESS

## EDITED BY

Jessica L. Jones,  
United States Food and Drug Administration,  
United States

## REVIEWED BY

Guennadi Kouzaev,  
Norwegian University of Science and  
Technology, Norway  
Ruth Evangeline Timme,  
US Food and Drug Administration,  
United States  
Mohammad Khalid,  
Linköping University Hospital, Sweden  
Otun Saha,  
Washington State University, United States  
Neha Periwal,  
Jamia Hamdard University, India

## \*CORRESPONDENCE

Cao Chen

✉ [chencao@ivdc.chinacdc.cn](mailto:chencao@ivdc.chinacdc.cn)

RECEIVED 05 September 2024

ACCEPTED 07 November 2024

PUBLISHED 20 November 2024

## CITATION

Feng Y, Chen S, Wang A, Zhao Z and  
Chen C (2024) Trends and impacts of  
SARS-CoV-2 genome sharing: a comparative  
analysis of China and the global community,  
2020–2023.

*Front. Public Health* 12:1491623.

doi: 10.3389/fpubh.2024.1491623

## COPYRIGHT

© 2024 Feng, Chen, Wang, Zhao and Chen.  
This is an open-access article distributed  
under the terms of the [Creative Commons  
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,  
distribution or reproduction in other forums is  
permitted, provided the original author(s) and  
the copyright owner(s) are credited and that  
the original publication in this journal is cited,  
in accordance with accepted academic  
practice. No use, distribution or reproduction  
is permitted which does not comply with  
these terms.

# Trends and impacts of SARS-CoV-2 genome sharing: a comparative analysis of China and the global community, 2020–2023

Yenan Feng, Songqi Chen, Anqi Wang, Zhongfu Zhao and  
Cao Chen\*

National Key Laboratory of Intelligent Tracking and Forecasting for Infectious Diseases, 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, China

**Objective:** The global sharing of pathogen genome sequences has been significantly expedited by the COVID-19 pandemic. This study aims to elucidate the global landscape of SARS-CoV-2 genome sharing between 2020 and 2023 with a focus on quantity, timeliness, and quality. Specifically, the characteristics of China are examined.

**Methods:** SARS-CoV-2 genomes along with associated metadata were sourced from GISAID database. The genomes were analyzed to evaluate the quantity, timeliness, and quality across different countries/regions. The metadata characteristics of shared genomes in China in 2023 were examined and compared with the actual demographic data of China in 2023.

**Results:** From 2020 to 2023, European countries consistently maintained high levels of genomic data sharing in terms of quantity, timeliness, and quality. In 2023, China made remarkable improvements in sequence sharing, ranking among the top 3.89% globally for quantity, 22.78% for timeliness, and 17.78% for quality. The genome sharing in China in 2023 covered all provinces with Shanghai Municipality contributing the most genomes. Human samples accounted for 99.73% of the shared genomes and exhibited three distinct peaks in collection dates. Males constituted 52.06%, while females constituted 47.94%. Notably, there was an increase in individuals aged 65 and above within the GISAID database compared to China's overall population in 2023.

**Conclusion:** The global sharing of SARS-CoV-2 genomes in 2020–2023 exhibited disparities in terms of quantity, timeliness, and quality. However, China has made significant advancements since 2023 by achieving comprehensive coverage across provinces, timely dissemination of data, and widespread population monitoring. Strengthening data sharing capabilities in countries like China during the SARS-CoV-2 pandemic will play a crucial role in containing and responding to future pandemics caused by emerging pathogens.

## KEYWORDS

SARS-CoV-2, genomic sequence, data sharing, GISAID, China

## 1 Introduction

The timely sharing of genomic sequences and associated metadata has played a crucial role in promoting global data awareness, enhancing our understanding of pathogenic evolution characteristics, and facilitating the development of detection reagents, vaccines, and drugs (1–4), which was particularly evident during the COVID-19 pandemic (5, 6). Since January 10, 2020, when China released the initial genome sequence of SARS-CoV-2 into the Global Initiative on Sharing All Influenza Data (GISAID) database, more than 16 million SARS-CoV-2 genomes have been stored in GISAID to date (7). In contrast, GenBank, serves as another prominent repository, hosted a collection of over 8 million SARS-CoV-2 genomes (8). It encompasses all other International Nucleotide Sequence Database Collaboration (INSDC) databases (9), including the European Nucleotide Archive (ENA) (10) and the DNA Data Bank of Japan (DDBJ) (11). Other repositories, such as the China National Center for Bioinformation GenBase (12), have stored more than 40,000 SARS-CoV-2 genomes. The China National GeneBank DataBase (CNGDB) (13), along with the Novel Coronavirus National Science and Technology Resource Service System at the National Microbiology Data Center, has, respectively, collected a small number of 87 and 305 SARS-CoV-2 genomes (9). Several analysis platforms such as Outbreak. Info (14), [Cov-Spectrum.org](https://cov-spectrum.org) (15), and [CoVariants.org](https://covariants.org) (16) were opportunistically developed utilizing these publicly available datasets. Additionally, there have been numerous intriguing investigations conducted based on the sharing of SARS-CoV-2 data to further explore the virus's structure, pathogenic mechanisms, mutation biases, and more (17–23). The World Health Organization (WHO) Guiding Principles for Pathogen Genome Data Sharing (24) advocate for the timely and high-quality sharing of genome data; however, there exists significant variation in data sharing levels among different countries/regions (25). The increased sharing of data may further exacerbate these imbalances and discrepancies. Therefore, a comprehensive understanding of global disparities in shared genomes can enhance objectivity when interpreting genomic data-driven analyses.

During the COVID-19 pandemic from 2020 to 2023, China endeavored to disseminate newly identified SARS-CoV-2 genomes through public databases, encompassing those obtained from the initial COVID-19 patient and the first SARS-CoV-2 genome isolated from the external packaging of cold-chain products (26). However, there is a limited scope for comprehensive assessment and comparison China with other global regions during this timeframe. Therefore, this study aims to comprehensively analyze the global landscape of SARS-CoV-2 genome sharing between 2020 and 2023, focusing on quantity, timeliness, and quality of shared genomes. Additionally, it will specifically examine China's characteristics in terms of sharing SARS-CoV-2 genomes. Considering the challenges associated with integrating SARS-CoV-2 genome data from diverse repositories, including sequence discrepancies and inconsistent metadata, we opted to utilize the GISAID database as our source for this study due to its extensive collection of SARS-CoV-2 sequences and comprehensive meta-information that surpasses other available resources (9).

## 2 Materials and methods

### 2.1 Data source

The genome and metadata of SARS-CoV-2 were obtained from the GISAID database on October 7, 2024. The total population data of China in 2023 was extracted from National Bureau of Statistics of China.<sup>1</sup> The standard map [No. GS (2023) 2767] was downloaded without modification from the standard map service website of the National Administration of Surveying, Mapping and Geographic Information. The global and Chinese count of reported cases was sourced from John Hopkins University and the WHO<sup>2</sup> via the Global Epidemic Analysis and Risk Assessment Platform of China CDC.

### 2.2 Inclusion criteria and data management

The period for genome submission ranged from January 1, 2020 to December 31, 2023. Genomes meeting the criteria of providing complete country of origin information and sampling date. The sampling date no later than the submit date were included in the analysis. The genomes were classified according to the continent and country/region, based on the information provided in the "Location" field of the metadata associated with each genome, indicating the geographical locations where samples were collected. The high-quality whole genome sequences were filtered with a length above 29,000 nt and Ns  $\leq 5\%$  in the entire genome. Genomes from China does not include the Hong Kong Special Administrative Region (SAR), Macau SAR, and Taiwan, China. To investigate the characteristics of shared SARS-CoV-2 genomes in China during 2023, we extracted genomes from GISAID submissions originating from China, covering the period from January 1 to December 31, 2023. The analysis included only individuals classified as "male" or "female" for gender, and age was limited to numeric values ranging from 0 to 200, excluding any symbols other than the decimal point. Genomes meeting both the sex and age criteria were selected for inclusion in the analysis. The prevalence of variants in each year were analyzed based on the information provided in the "Variant" field of the metadata associated with each genome. The proportion of each variant to the total number of shared genomes in each year was calculated.

### 2.3 Statistical analysis

Descriptive analysis was conducted to present the general characteristics of the genomes sharing. Continuous variables were reported using the median and interquartile range (IQR), while categorical variables were presented as counts and proportions. Structured Query Language and Python were used for data cleaning, processing, and generating descriptive statistics, while both Python and GraphPad Prism 9 (GraphPad Software, Inc., LaJolla, CA, United States) were employed for data visualization. Detailed information and codes can be found.<sup>3</sup>

1 <https://data.stats.gov.cn/english/easyquery.htm?cn=C01>

2 <https://covid19.who.int/data>

3 <https://github.com/SongqiChen/covid19-genome-sharing-analysis>

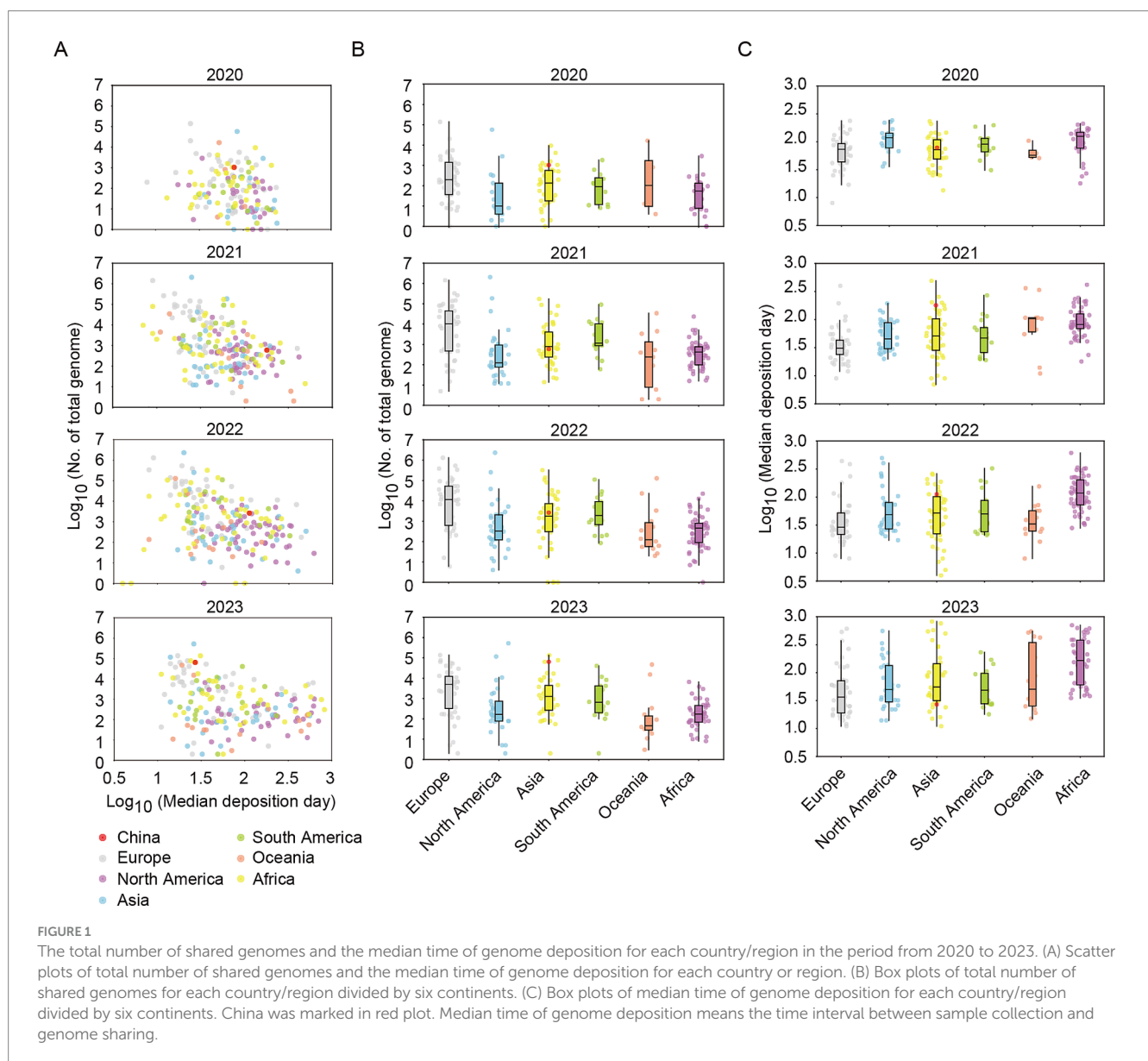
### 3 Results

#### 3.1 The sharing of SARS-CoV-2 genomes exhibited worldwide variation in quantity and timeliness across continents

From 2020 to 2023, by searching in the GISAID database, a total of 222 countries/regions actively contributed 16,001,611 SARS-CoV-2 genomes. The annual counts were as follows: 142, 205, 210, and 180 countries/regions with genome contributions amounting to 307,565; 6,205,472; 7,623,101; and 1,865,473, respectively. Compared to the number of reported cases each year, we observed a similar trend between the number of cases and the sharing of genomes. Since the emergence of variants of concern (VOCs), variants of interest (VOIs) and variants under monitoring (VUMs) of the WHO, there has been a global increase in both reported cases and shared genomes. The peak in both

case numbers and genome sharing occurred with the Omicron variant in 2022 (Supplementary Figure S1A).

Next, the numbers and median deposition days of genomes were analyzed across six continents. Certain European, North American, and Asian countries/regions, such as United Kingdom, United States of America, and Singapore et al., exhibited both substantial numbers of shared genomes and short median deposition days (Figure 1A). Overall, the European region consistently contributed a high median number of shared genomes and short median deposition days throughout the years from 2020 to 2023, indicating the continuity and timeliness of genome sharing in Europe (Figures 1B,C). Compared to the prior to 2023, the number and timeliness of shared genomes in China in 2023 (total number: 64302; median deposition days: 27, IQR: 16–52) were far higher than those of countries/regions in Asian (median total numbers: 1258.5, IQR: 265.3–4391.5; median deposition days: 55.3, IQR: 31.5–145.3) and were among the forefront of the world (Figures 1A–C).



### 3.2 The levels in SARS-CoV-2 genome sharing from China remarkable strengthened in 2023

To further investigate the features of Chinese SARS-CoV-2 genome sharing, we analyzed the SARS-CoV-2 genomes shared by China on GISAID database. The results showed that the number of shared sequences increased significantly in 2023 compared to the period of 2020–2022 (Figures 2A,B). The highest number of shared sequences occurred in January 2023 (total numbers: 9862), with two small peaks observed in April (total numbers: 7963) and June (total numbers: 8136), followed by a smaller peak in November (total numbers: 3198). Although the number of shared genomes in China from 2020 to 2022 was very similar, the relative ranking of China's shared genomes in the global countries/regions was lower than the median level in 2021 and 2022, indicating a relatively lower level of genome sharing compared to other regions worldwide. The number of genomes shared by China in 2023 significantly exceeded the median number of shared genomes of countries/regions globally (median total numbers: 379.5, IQR: 94.8–4049.3), and the relative ranking of China had risen to the top 3.89% globally (Figure 2C). Similarly to the global, there were much closer trend between the number of reported cases and the sharing of genomes in China (Supplementary Figure S1B).

The timeliness trend of genome sharing also demonstrates similar patterns, with a more consistent and shorter time interval in 2023 compared to the period of 2020–2022 in China (Figure 2D). In 2023, the median deposition period for genome sharing in China was notably shorter compared to the global median (median deposition days: 56.3, IQR: 28.8–173.0; Figure 2E), ranking among the top 22.78% worldwide (Figure 2F). Therefore, in contrast to the increased global median deposition days for genomes in 2023, China has achieved advancements in both quantity and timeliness.

### 3.3 Noticeable disparity existed in the quality of globally shared genomes

Subsequently, we conducted a comprehensive analysis on the quality of the SARS-CoV-2 genomes shared via GISAID. Genomes with a length above 29,000 nt were selected, while low coverage sequences ( $N_s > 5\%$ ) were excluded to obtain high-quality whole genome sequences. The ratio of these high-quality sequences to the total number of sequences was then calculated. The findings indicated that there was high median ratio and minimal degree of dispersion in genome quality among the shared genomes from European countries/regions from 2020 to 2023, demonstrating consistently high quality in European countries/regions (Figure 3A). Conversely, African exhibited a lower median ratio and a wider dispersion, suggesting an overall lower quality of genomic sequences with significant disparities among countries/regions. Notably, the quality of genomes shared from China in 2020 falls below the median level for Asia and globally. However, it demonstrated steady improvement over subsequent years, surpassing the median levels of both Asia and global by 2022 (Figures 3A,B). By 2023, the relative ranking of quality for shared genomes by China has risen to the top 17.78% worldwide (Figure 3C).

### 3.4 The genomic metadata shared in China in 2023 exhibited distinct characteristics

In 2023, China made great efforts in genome sharing. Although all provinces in China shared genomic sequences, there were regional differences. Shanghai Municipality, Guangdong Province, and Beijing Municipality shared the most sequences, with 9,970, 6,310, and 4,693, respectively, (Figure 4A). From the species composition of the shared sequence, the vast majority were human samples (total number:

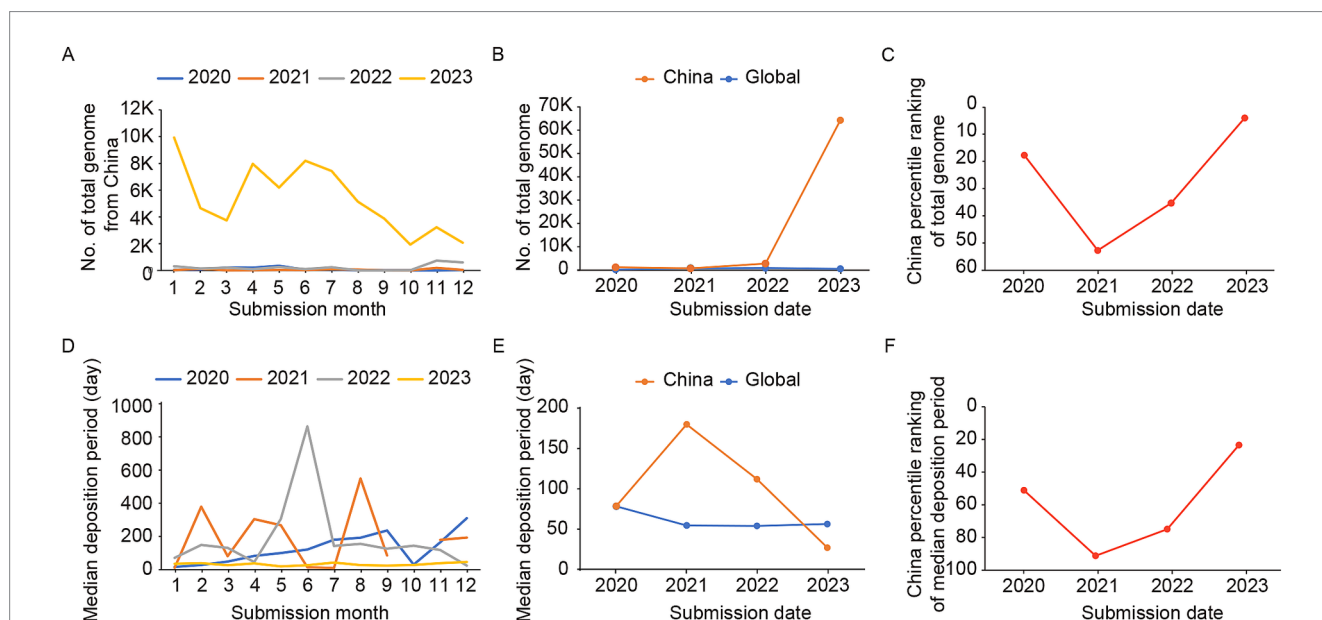
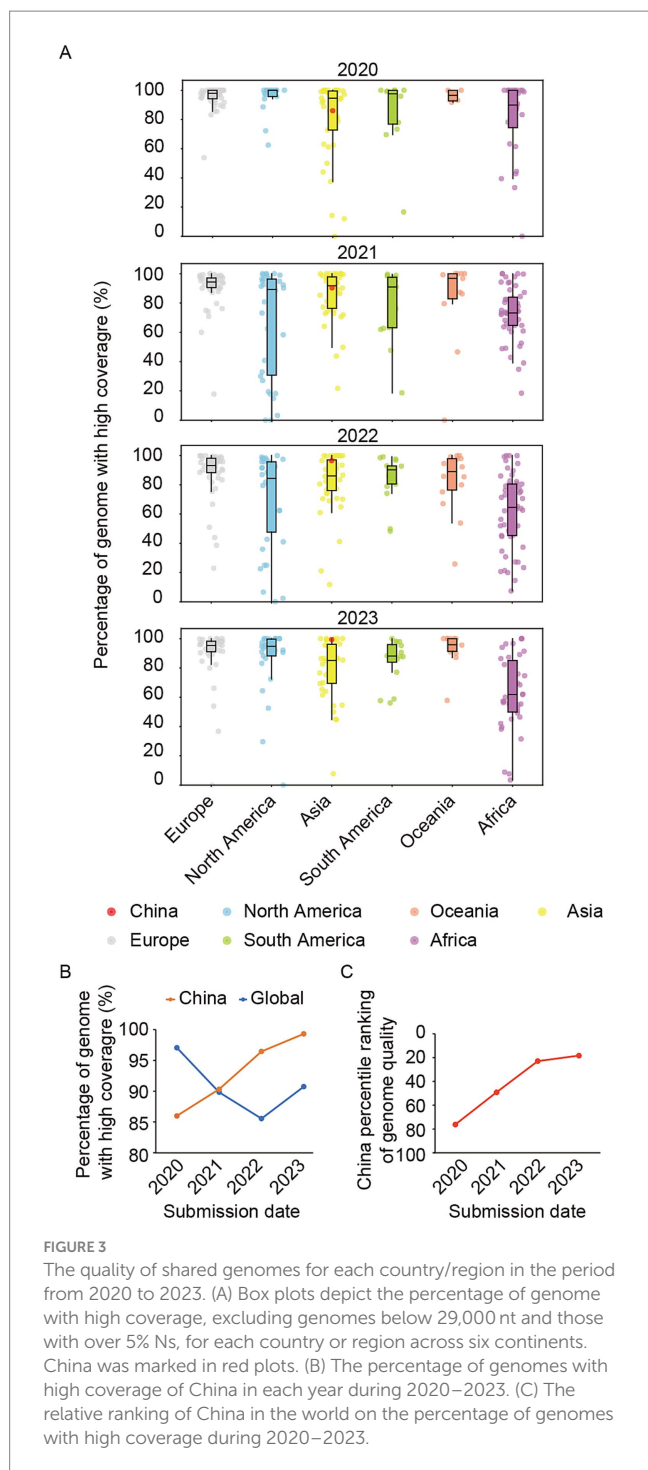


FIGURE 2

The number of shared genomes and the median time of genome deposition in China from 2020 to 2023. (A) The number of shared genomes in China each month during 2020–2023. (B) The median number of shared genomes in China and among countries/regions worldwide from 2020 to 2023. (C) The relative ranking of China in the world on the number of shared genomes during 2020–2023. (D) The median time of genome deposition in China during 2020–2023. (E) The median time of genome deposition in China and among countries/regions worldwide from 2020 to 2023. (F) The relative ranking of China in the world on the median time of genome deposition during 2020–2023.





64,302), followed by environmental samples (total number: 130), and 43 samples of unidentified species (Figure 4B), suggesting that China's monitoring strategy in 2023 focused primarily on population surveillance with secondary emphasis on environmental monitoring. The sampling collection dates of population data showed a concentration after December 2022 with three prominent peaks: December 2022 to January 2023, May to June 2023, and August to September 2023 (Figure 4C), indicating that concentrated data sharing during these periods may be associated with clustered outbreaks. Gender distribution was evenly balanced, with 52.06% male and

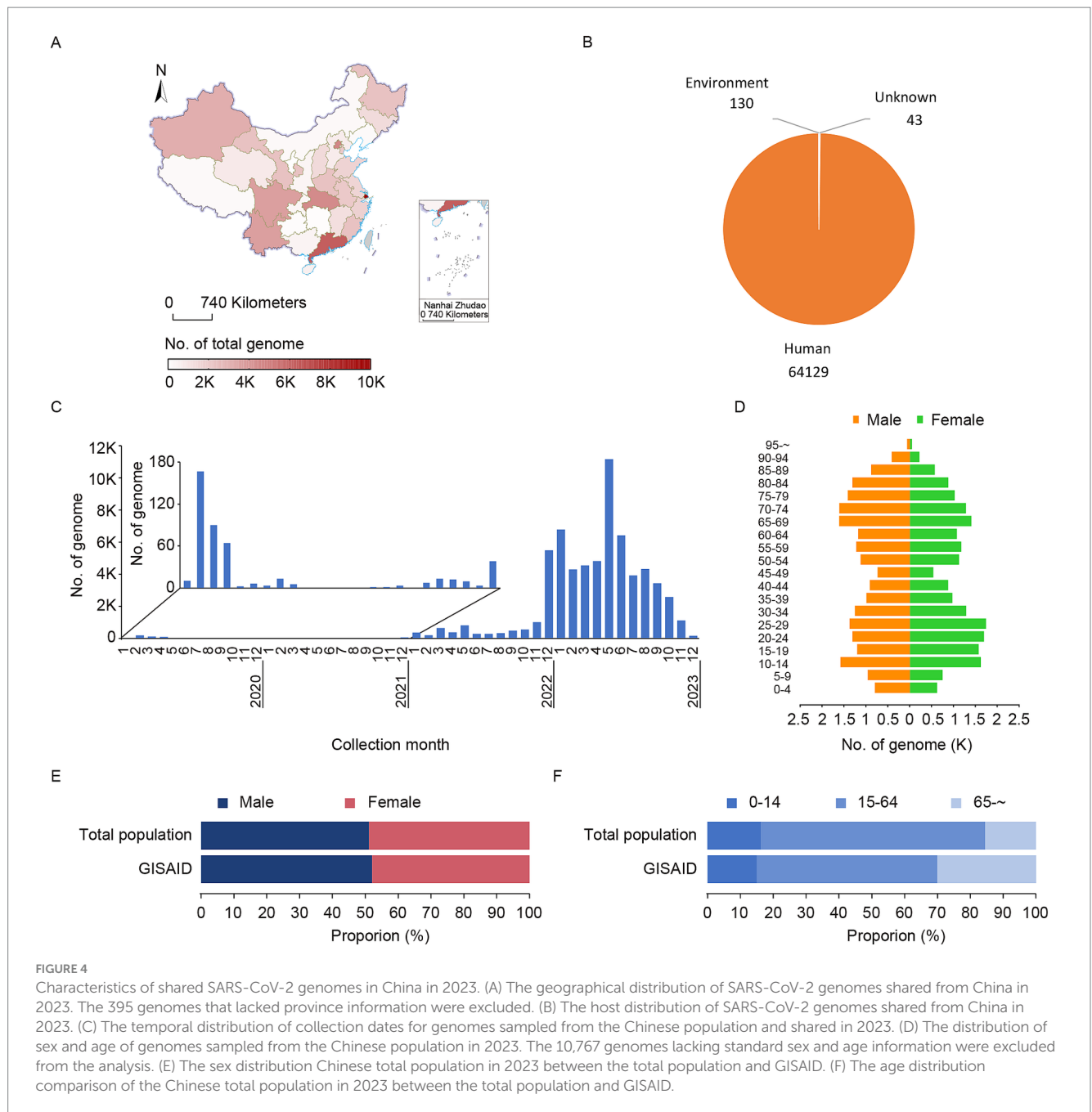
47.94% female. The gender composition spanned all age groups but was mainly concentrated in the age ranges of 10–29 and 65–74 (Figure 4D). A comparison between GISAID's sampled population distribution and China's total population revealed similar gender ratios but marked differences in age structure (Figure 4E, male: total population vs. GISAID: 51.10% vs. 52.06%). Notably, a significantly higher proportion of individuals over 65 years old among GISAID's shared genomic data (Figure 4F, total population vs. GISAID: 15.40% vs. 30.01%), potentially attributed to specific surveillance targeting this older adult population.

## 4 Discussion

This study systematically analyzed and compared the differences in SARS-CoV-2 genome sharing among countries/regions across continents during the pandemic from 2020 to 2023. The study findings demonstrated that the European countries/regions had superior performance in terms of the quantity, timeliness, and quality of shared genomes compared to other countries/regions. Starting from 2023, China has made significant improvements in all these aspects, with full coverage of provinces, timely sharing, and widespread monitoring of the population. The findings of our research suggested an expansion of data sharing capacity during the SARS-CoV-2 pandemic. It will be critical in containing and responding to future pandemics caused by novel pathogens.

Our analysis reveals substantial disparities in global genome sharing among different countries/regions. In view of the extensive incorporation of genomic data within GISAID, conducting comprehensive genome quality control measures, such as identifying frameshifts, posed a challenge. In our study, a threshold exceeding 29,000 nt coverage along with less than 5% Ns were utilized to define high quality whole genome sequencing. Notably, we found that Europe demonstrated a higher level of genome sharing compared to other continents, while Africa exhibits relatively lower performance. These variations may stem from diverse factors encompassing discrepancies in sequencing capacities as well as policy and financial support across nations. Consistent with our findings, most East African Community nations encountered challenges including insufficient local NGS equipment, limited bioinformatics expertise, inadequate computational resources, and ineffective data-sharing mechanisms (27). However, the Public Health England has been recognized as an early leader at a national level for employing high-throughput sequencing for pathogen surveillance (28, 29). The quality of related metadata, in addition to sequence quality, was also deemed important. A previous study revealed a prevalent occurrence of incomplete metadata worldwide for GISAID sequences. Specifically, approximately 63% of the sequences lacked demographic information, 84% were devoid of sampling strategy details, and patient-level clinical information was missing in over 95% of the cases (30). One limitation of our study is that we used the sample collection location as the country for analyzing genome sharing levels. However, there may be potential bias in assessing actual sharing performance among countries/regions due to inter-regional scientific projects leading to differences between the submitting country and sample collection country/region. Regardless, the COVID-19 pandemic undeniably propels pathogen whole-genome sequencing endeavors and facilitates data sharing.

The growing prevalence of shared genomes presents several challenges, such as the management of extensive public databases, and



the issue of duplicate data uploading. The RCoV19 database, for instance, possesses the capability to integrate and eliminate redundant genomes as well as annotate database sources (31, 32). RCoV19 offers a comprehensive integration of data and identifies the same genome sequences submitted to different sources by comparing key meta information (virus name, collection date, and location) as well as sequences after removing Ns and unifying the letter case (9). Besides RCoV19, the VirusDIP (33), ViruSurf (34), and CoV-Seq (35) databases also perform data integration and de-redundancy processing. However, it is worth noting that ViruSurf and CoV-Seq have not been updated since January 2022 and September 2020, respectively. On the other hand, VirusDIP integrates data from GISAID, GenBank, and CNGBdb but does not include information from GenBase and NCNSTRSS. These databases's efforts have greatly

improved the accessibility of comprehensive datasets for users. Even then, data incompleteness is an unavoidable limitation for integration, potentially resulting in information loss due to format discrepancies across different databases. Moreover, this limitation also hampers genomic surveillance as the representation of virus distribution may be skewed due to information incompleteness on local or travel-related cases in the majority of genomes.

Our findings demonstrate the substantial progress made by China in sharing SARS-CoV-2 genomic data. As demonstrated in a previous study, achieving a sequencing turnaround time of less than 21 days could serve as a benchmark for effective SARS-CoV-2 genomic surveillance (36). Here, the median turnaround time for China in 2023 was 27 days, which closely approached the aforementioned threshold, indicating a significant improvement

and underscoring the imperative for sustained efforts. Before 2023, China's robust prevention and control measures, coupled with successful vaccination campaigns, led to a minimal incidence of cases, with the majority of domestic outbreaks attributed to imported infections from overseas. This correlation was consistent with the limited number of shared genomes by China during the period spanning 2020–2022. Due to the adjustment of COVID-19 prevention and control policy at the end of 2022 (37), coupled with the continuous evolution of Omicron variant, there was an increase in reported cases can be observed in China. Simultaneously, there has been a corresponding rise in shared genomes showing consistency. Upon analyzing the temporal distribution of shared genomes of China sampled in 2023, we observed three distinct peaks in the epidemic: December 2022 to January 2023, May 2023 to June 2023, and August 2023 to September 2023. Remarkably, these peaks closely align with the positive rate of COVID-19 among influenza-like cases reported by the China CDC (38). However, there were variations in peak intensity, particularly observed during the May to June 2023 peak. Therefore, genomic data sharing can only serve as a reference for rough estimating the actual epidemic. Given that China made adjustments to its epidemic prevention and control policy at the end of 2022, timely sharing of genomic data will facilitate comprehensive and expeditious analysis of circulating variants within China by both domestic and international researchers.

All in all, the analysis of SARS-CoV-2 genomic data sharing during 2020–2023 reveals significant advancements, particularly in countries like China. The efforts made by China and the global community in sequencing and sharing genome sequences during the COVID-19 pandemic undeniably contribute to advancing the One Health objective's requirements of ensuring discoverable, accessible, interoperable, and reusable data (39). These endeavors also facilitated further researches and the evidence-based policies in response to the spread of VOCs and VOIs (40–44). However, it is crucial to acknowledge the variations in the extent of global genomic data sharing across different regions worldwide. The timely sharing of data is crucial for effectively addressing the current COVID-19 situation, as well as enhancing our preparedness for future outbreaks of emerging pathogens. Further efforts are warranted to address the disparity in global genomic data sharing and establish a universally standardized platform for data utilization, in order to promote scientific collaboration and advance research progress.

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: <https://github.com/SongqiChen/covid19-genome-sharing-analysis>.

## References

1. Armstrong GL, MacCannell DR, Taylor J, Carleton HA, Neuhaus EB, Bradbury RS, et al. Pathogen genomics in public health. *N Engl J Med.* (2019) 381:2569–80. doi: 10.1056/NEJMs1813907
2. Leguia M, Vila-Sanjurjo A, Chain PSG, Berry IM, Jarman RG, Pollett S. Precision medicine and precision public health in the era of pathogen next-generation sequencing. *J Infect Dis.* (2020) 221:S289–91. doi: 10.1093/infdis/jiz424
3. Bonam SR, Hu H. Next-generation vaccines against COVID-19 variants: beyond the spike protein. *Zoonoses (Burlingt).* (2023) 3. doi: 10.15212/ZOONOSES-2023-0003
4. Lin L, Pei Y, Li Z, Luo D. Progress and challenges of mRNA vaccines. *Interdis Med.* (2023) 1:e20220008. doi: 10.1002/INMD.20220008
5. Oude Munnink BB, Nieuwenhuijse DF, Stein M, O'Toole Á, Haverkate M, Mollers M, et al. Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. *Nat Med.* (2020) 26:1405–10. doi: 10.1038/s41591-020-0997-y
6. Cuypers L, Dellicour S, Hong SL, Potter BI, Verhasselt B, Vereecke N, et al. Two years of genomic surveillance in Belgium during the SARS-CoV-2 pandemic to attain

## Author contributions

YF: Conceptualization, Funding acquisition, Writing – original draft. SC: Data curation, Methodology, Software, Writing – review & editing. AW: Investigation, Writing – original draft. ZZ: Software, Visualization, Writing – review & editing. CC: Conceptualization, Funding acquisition, Writing – review & editing.

## Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Natural Science Foundation of China (82341035, 82341034) and National key research and development program of China (2023YFC2306000).

## Acknowledgments

We gratefully acknowledge all data contributors, i.e., the authors and their originating laboratories responsible for obtaining the specimens, and their submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpubh.2024.1491623/full#supplementary-material>

- country-wide coverage and monitor the introduction and spread of emerging variants. *Viruses*. (2022) 14:2301. doi: 10.3390/v14102301
7. Khare S, Gurry C, Freitas L, Schultz MB, Bach G, Diallo A, et al. GISAIID's role in pandemic response. *China CDC Wkly*. (2021) 3:1049–51. doi: 10.46234/ccdcw2021.255
  8. Brister JR, Ako-Adjei D, Bao Y, Blinkova O. NCBI viral genomes resource. *Nucleic Acids Res*. (2015) 43:D571–7. doi: 10.1093/nar/gku1207
  9. Ma L, Zhao W, Huang T, Jin E, Wu G, Zhao W, et al. On the collection and integration of SARS-CoV-2 genome data. *Biosafety and Health*. (2023) 5:204–10. doi: 10.1016/j.bsheal.2023.07.004
  10. Tanizawa Y, Fujisawa T, Kodama Y, Kosuge T, Mashima J, Tanjo T, et al. DNA data Bank of Japan (DDBJ) update report 2022. *Nucleic Acids Res*. (2023) 51:D101–5. doi: 10.1093/nar/gkac1083
  11. Okido T, Kodama Y, Mashima J, Kosuge T, Fujisawa T, Ogasawara O. DNA data Bank of Japan (DDBJ) update report 2021. *Nucleic Acids Res*. (2022) 50:D102–5. doi: 10.1093/nar/gkab995
  12. Bu C, Zheng X, Zhao X, Xu T, Bai X, Jia Y, et al. Gen Base: A Nucleotide Sequence Database. *Genomics Proteomics Bioinformatics*. (2024) 22. doi: 10.1093/gpbjnl/qzae047
  13. Chen FZ, You LJ, Yang F, Wang LN, Guo XQ, Gao F, et al. CNGBdb: China national GeneBank DataBase. *Yi Chuan*. (2020) 42:799–809. doi: 10.16288/j.ycz.20-080
  14. Gangavarapu K, Latif AA, Mullen JL, Alkuzweny M, Hufbauer E, Tsueng G, et al. Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. *Nat Methods*. (2023) 20:512–22. doi: 10.1038/s41592-023-01769-3
  15. Chen C, Nadeau S, Yared M, Voinov P, Xie N, Roemer C, et al. CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. *Bioinformatics*. (2022) 38:1735–7. doi: 10.1093/bioinformatics/btab856
  16. Hodcroft EB. CoVariants: SARS-CoV-2 mutations and variants of interest. (2021). Available from: <https://covariants.org/> [Accessed October 7, 2024].
  17. Ansari S, Gupta N, Verma R, Singh ON, Gupta J, Kumar A, et al. Antiviral activity of the human endogenous retrovirus-R envelope protein against SARS-CoV-2. *EMBO Rep*. (2023) 24:e55900. doi: 10.15252/embr.202255900
  18. Periwal N, Rathod SB, Pal R, Sharma P, Nebhani L, Barnwal RP, et al. In silico characterization of mutations circulating in SARS-CoV-2 structural proteins. *J Biomol Struct Dyn*. (2022) 40:8216–31. doi: 10.1080/07391102.2021.1908170
  19. Periwal N, Rathod SB, Sharma S, Johar GS, Jain A, Barnwal RP, et al. Time series analysis of SARS-CoV-2 genomes and correlations among highly prevalent mutations. *Microbiol Spectr*. (2022) 10:e0121922. doi: 10.1128/spectrum.01219-22
  20. Rice AM, Castillo Morales A, Ho AT, Mordstein C, Mühlhausen S, Watson S, et al. Evidence for strong mutation Bias toward, and selection against, U content in SARS-CoV-2: implications for vaccine design. *Mol Biol Evol*. (2021) 38:67–83. doi: 10.1093/molbev/msaa188
  21. Periwal N, Bhardwaj U, Sharma S, Arora P, Sood V. In silico analysis of SARS-CoV-2 genomes: insights from SARS encoded non-coding RNAs. *Front Cell Infect Microbiol*. (2022) 12:966870. doi: 10.3389/fcimb.2022.966870
  22. Wang R, Chen J, Gao K, Hozumi Y, Yin C, Wei GW. Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. *Commun Biol*. (2021) 4:228. doi: 10.1038/s42003-021-01754-6
  23. Chen Y, Li S, Wu W, Geng S, Mao M. Distinct mutations and lineages of SARS-CoV-2 virus in the early phase of COVID-19 pandemic and subsequent 1-year global expansion. *J Med Virol*. (2022) 94:2035–49. doi: 10.1002/jmv.27580
  24. WHO. WHO guiding principles for pathogen genome data sharing. (2022). Available from: <https://iris.who.int/bitstream/handle/10665/364222/9789240061743-eng.pdf?sequence=1&isAllowed=y> [Accessed October 7, 2024].
  25. Khan W, Kabir F, Kanwar S, Aziz F, Muneer S, Kalam A, et al. Building up a genomic surveillance platform for SARS-CoV-2 in the middle of a pandemic: a true north-south collaboration. *BMJ Glob Health*. (2023) 8:e012589. doi: 10.1136/bmjgh-2023-012589
  26. Ma H, Wang Z, Zhao X, Han J, Zhang Y, Wang H, et al. Long distance transmission of SARS-CoV-2 from contaminated cold Chain products to humans-Qingdao City, Shandong Province, China, September 2020. *China CDC Wkly*. (2021) 3:637–44. doi: 10.46234/ccdcw2021.164
  27. Nguinkal JA, Zoclanclounon YAB, Molina A, Roba A, Nyakio NM, Lokamar PN, et al. Assessment of the pathogen genomics landscape highlights disparities and challenges for effective AMR surveillance and outbreak response in the east African community. *BMC Public Health*. (2024) 24:1500. doi: 10.1186/s12889-024-18990-0
  28. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect Dis*. (2013) 13:137–46. doi: 10.1016/S1473-3099(12)70277-3
  29. Jenkins C, Dallman TJ, Grant KA. Impact of whole genome sequencing on the investigation of food-borne outbreaks of Shiga toxin-producing *Escherichia coli* serogroup O157: H7, England, 2013 to 2017. *Euro Surveill*. (2019) 24:1800346. doi: 10.2807/1560-7917.ES.2019.24.4.1800346
  30. Chen Z, Azman AS, Chen X, Zou J, Tian Y, Sun R, et al. Global landscape of SARS-CoV-2 genomic surveillance and data sharing. *Nat Genet*. (2022) 54:499–507. doi: 10.1038/s41588-022-01033-y
  31. Cheng Y, Ji C, Zhou HY, Zheng H, Wu A. Web resources for SARS-CoV-2 genomic database, annotation. *Analysis and Variant Tracking Viruses*. (2023) 15:1158. doi: 10.3390/v15051158
  32. Gong Z, Zhu JW, Li CP, Jiang S, Ma LN, Tang BX, et al. An online coronavirus analysis platform from the National Genomics Data Center. *Zool Res*. (2020) 41:705–8. doi: 10.24272/zj.issn.2095-8137.2020.065
  33. Yu D, Yang X, Tang B, Pan YH, Yang J, Duan G, et al. Coronavirus gen browser for monitoring the transmission and evolution of SARS-CoV-2. *Brief Bioinform*. (2022) 23. doi: 10.1093/bib/bbab583
  34. Canakoglu A, Pinoli P, Bernasconi A, Alfonsi T, Melidis DP, Ceri S. ViruSurf: an integrated database to investigate viral sequences. *Nucleic Acids Res*. (2021) 49:D817–24. doi: 10.1093/nar/gkaa846
  35. Liu B, Liu K, Zhang H, Zhang L, Bian Y, Huang L. CoV-Seq, a new tool for SARS-CoV-2 genome analysis and visualization: development and usability study. *J Med Internet Res*. (2020) 22:e22299. doi: 10.2196/22299
  36. Brito AF, Semenova E, Dudas G, Hassler GW, Kalinich CC, Kraemer MUG, et al. Global disparities in SARS-CoV-2 genomic surveillance. *Nat Commun*. (2022) 13:7003. doi: 10.1038/s41467-022-33713-y
  37. China enters new phase of COVID response [press release]. (2023). [http://english.www.gov.cn/statecouncil/ministries/202302/24/content\\_WS63f7f52cc6d0a757729e726e.html](http://english.www.gov.cn/statecouncil/ministries/202302/24/content_WS63f7f52cc6d0a757729e726e.html) [Accessed October 7, 2024]
  38. CDC C. National situation of COVID-19 infection. (2024). Available from: <https://www.chinaccdc.cn/jksj/xgbdyq/> [Accessed October 7, 2024].
  39. Timme RE, Wolfgang WJ, Balkey M, Venkata SLG, Randolph R, Allard M, et al. Optimizing open data to support one health: best practices to ensure interoperability of genomic data from bacterial pathogens. *One Health Outlook*. (2020) 2:20. doi: 10.1186/s42522-020-00026-3
  40. ECDC. Risk related to spread of new SARS-CoV-2 variants of concern in the EU/EEA. (2020). Available from: <https://www.ecdc.europa.eu/sites/default/files/documents/COVID-19-risk-related-to-spread-of-new-SARS-CoV-2-variants-EU-EEA.pdf> [Accessed October 7, 2024].
  41. Ye S, Zhang Y, Zhao X, Yu Z, Song Y, Tan Z, et al. Emerging variants of B.1.617 lineage identified among returning Chinese employees working in India-Chongqing municipality, China, April 2021. *China CDC Wkly*. (2021) 3:409–10. doi: 10.46234/ccdcw2021.109
  42. National Center for Immunization and Respiratory Diseases (NCIRD), Division of Viral Diseases. National Center for Immunization and Respiratory Diseases (NCIRD), Division of Viral Diseases. CDC COVID-19 Science Briefs [Internet]. Atlanta (GA): Centers for Disease Control and Prevention (US); 2020–. Science Brief: Omicron (B.1.1.529) Variant. (2021).
  43. Grubaugh ND, Hodcroft EB, Fauver JR, Phelan AL, Cevik M. Public health actions to control new SARS-CoV-2 variants. *Cell*. (2021) 184:1127–32. doi: 10.1016/j.cell.2021.01.044
  44. Liang J, Zeng Z, Li Q, Guan W, Yang Z, Hon C. Challenge on prediction of influenza virus and SARS-CoV-2 virus co-circulation. *Interdis Med*. (2023) 1:e20220006. doi: 10.1002/INMD.20220006