Landscape of SARS-CoV-2 genomic surveillance, public availability extent of genomic data, and epidemic shaped by variants: a global descriptive study

Zhiyuan Chen[1], Andrew S. Azman[2,3], Xinhua Chen[1], Junyi Zou[1], Yuyang Tian[1], Ruijia Sun[1], Xiangyanyu Xu[1], Yani Wu[1], Wanying Lu[1], Shijia Ge[4], Zeyao Zhao[1], Juan Yang[1], Daniel T. Leung[5,6], Daryl B. Domman[7], and Hongjie Yu[1,4,8]

Affiliations

1. School of Public Health, Fudan University, Key Laboratory of Public Health Safety, Ministry of Education, Shanghai, China
2. Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
3. Institute of Global Health, Faculty of Medicine, University of Geneva, Switzerland
4. Department of Infectious Diseases, Huashan Hospital, Fudan University, Shanghai, China
5. Division of Infectious Diseases, University of Utah School of Medicine, Salt Lake City, UT, USA
6. Division of Microbiology & Immunology, University of Utah School of Medicine, Salt Lake City, UT, USA
7. Center for Global Health, Department of Internal Medicine, University of New Mexico Health Sciences Center, New Mexico, USA
8. Shanghai Institute of Infectious Disease and Biosecurity, Fudan University, Shanghai, China

Corresponding authors: Hongjie Yu, School of Public Health, Fudan University, Shanghai 200032, China; E-mail: yhj@fudan.edu.cn

Word count (abstract): 286

Word count (main text): 3,534

Disclaimer: The views expressed are those of the authors and do not necessarily represent the institutions with which the authors are affiliated.

Abstract

Background

Genomic surveillance has shaped our understanding of SARS-CoV-2 variants, which have proliferated globally in 2021. Characterizing global genomic surveillance, sequencing coverage, the extent of publicly available genomic data coupled with traditional epidemiologic data can provide evidence to inform SARS-CoV-2 surveillance and control strategies.

Methods

We collected country-specific data on SARS-CoV-2 genomic surveillance, sequencing capabilities, public genomic data, and aggregated publicly available variant data. We divided countries into three levels of genomic surveillance and sequencing availability based on predefined criteria. We downloaded the merged and deduplicated SARS-CoV-2 sequences from multiple public repositories, and used different proxies to estimate the sequencing coverage and public availability extent of genomic data, in addition to describing the global dissemination of variants.

Findings

Since the start of 2021, the COVID-19 global epidemic clearly featured increasing circulation of Alpha, which was rapidly replaced by the Delta variant starting around May 2021 and reaching a global prevalence of 96.6% at the end of July 2021. SARS-CoV-2 genomic surveillance and sequencing availability varied markedly across countries, with 63 countries performing routine genomic surveillance and 79 countries with high availability of SARS-CoV-2 sequencing. Less than 3.5% of confirmed SARS-CoV-2 infections were sequenced globally since September 2020, with the lowest sequencing coverage in the WHO regions of Eastern Mediterranean, South East Asia, and Africa. Across different variants, 28-52% of countries with explicit reporting on variants shared less than half of their variant sequences in public repositories. More than 60% of demographic and 95% of clinical data were absent in GISAID metadata accompanying sequences.

Interpretation

Our findings indicated an urgent need to expand sequencing capacity of virus isolates, enhance the sharing of sequences, the standardization of metadata files, and supportive networks for countries with no sequencing capability.

Keywords

Genomic Surveillance, Sequencing, SARS-CoV-2 Variants, Data Sharing

Funding

66 Key Program of the National Natural Science Foundation of China, the US National Institutes of
67 Health.

Research in context

Evidence before this study

On September 3, 2021, we searched PubMed for articles in any language published after January 1, 2020, using the following search terms: ("COVID-19" OR "SARS-CoV-2") AND ("Global" OR "Region") AND ("genomic surveillance" OR "sequencing" OR "spread"). Among 43 papers identified, few papers discussed the global diversity in genomic surveillance, sequencing, public availability of genomic data, as well as the global spread of SARS-CoV-2 variants. A paper from Furuse employed the publicly GISAID data to evaluate the SARS-CoV-2 sequencing effort by country from the perspectives of "fraction", "timeliness", and "openness". Another viewpoint paper by Case Western Reserve University's team discussed the impediments of genomic surveillance in several countries during the COVID-19 pandemic. The paper as reported by Campbell and colleagues used the GISAID data to present the global spread and estimated transmissibility of recently emerged SARS-CoV-2 variants. We also found several studies that reported the country-level genomic surveillance and spread of variants. To our knowledge, no research has quantitatively depicted the global SARS-CoV-2 genomic surveillance, sequencing ability, and public availability extent of genomic data.

Added value of this study

This study collected country-specific data on SARS-CoV-2 genomic surveillance, sequencing capabilities, public genomic data, and aggregated publicly available variant data as of 20 August 2021. We found that genomic surveillance strategies and sequencing availability is globally diverse. Less than 3.5% of confirmed SARS-CoV-2 infections were sequenced globally since September 2020. Our analysis of publicly deposited SARS-CoV-2 sequences and officially reported number of variants implied that the public availability extent of genomic data is low in some countries, and more than 60% of demographic and 95% of clinical data were absent in GISAID metadata accompanying sequences. We also described the pandemic dynamics shaped by VOCs.

Implications of all the available evidence

Our study provides a landscape for global sequencing coverage and public availability extent of sequences, as well as the evidence for rapid spread of SRAS-CoV-2 variants. The pervasive spread of Alpha and Delta variants further highlights the threat of SARS-CoV-2 mutations despite the availability of vaccines in many countries. It raised an urgent need to do more work on defining the ideal sampling schemes for different purposes (e.g., identifying new variants) with an additional call to share these data in public repositories to allow for further rapid scientific discovery.

Introduction

Following the first pandemic wave of coronavirus disease 2019 (COVID-19), the emergence and dissemination of SARS-CoV-2 variants have resulted in new waves of infections across the globe in 2021. Some SARS-CoV-2 variants disappeared immediately, while others characterized by several key mutations adapted well, enabling their rapid spread[1]. WHO has designated four Variants of Concern (VOCs) associated with increased transmissibility and various extents of immune escape[2-4], namely the Alpha, Beta, Gamma, and Delta variants, first detected in the UK, South Africa, Brazil, and India, respectively[5]. Specifically, the Delta variant is highly transmissible, with an estimated transmissibility increase of 40-60% compared with Alpha variant[6-9], while the Beta variant has been shown to have the highest reduction in neutralization activity whether from natural infection or vaccination[10], both reflected in lower vaccine efficacy or effectiveness[11-13].

The identification and classification of SARS-CoV-2 variants mainly relied on partial or whole genome sequencing, although PCR assays have been used to identify specific features relatively unique in specific variants, like spike gene target failure (SGTF)[14]. Since the first SARS-CoV-2 sequence was published in January 2020[15], the unprecedented rate of genome data generation was far greater than any other pathogen[16], with 3.1 million genomes deposited in Global Initiative on Sharing All Influenza Data (GISAID) through August 2021[17]. Genomic data has been vital to the early detection of mutations and monitoring of virus evolution, as well as evaluating the degree of similarities between circulating variants with vaccine strains, especially since the availability of SARS-CoV-2 vaccines[18].

Several studies have employed genomic data to examine the evolution and associated spread of dominant variants in one country or one region, raising a claim of the rapidity of local transmission of SARS-CoV-2 variants and urgency of genomic surveillance[6,19-21]. However, the lack of representation of genomic data from low and middle-income countries (LMICs) was most concerning in these studies [6,19-21]. Indeed, the impact of genome data is dependent on their quality, and the reliability and accuracy of such data may influence the global community's ability to track the spread of variants in a timely manner.

In this study, we aimed to investigate the global diversity of SARS-CoV-2 genomic surveillance and the global sequencing coverage of confirmed cases and public availability extent of genomes. In addition, we sought to map the global identification and spread of SARS-CoV-2 variants. This data can provide evidence to better inform policy on SARS-CoV-2 surveillance.

Methods

Data sources and collection

Through extracting country-specific data from multiple publicly available sources, we built three datasets of genomic surveillance, deposited genomic data in publicly repositories, and official aggregated genomic data as of 20 August 2021.

Dataset of genomic surveillance

Each country's genomic surveillance strategy and sequencing availability was gathered from searches of the websites of regional WHO, the country's Ministry of Health, CDC, local academic partners, and official news, supplemented by a literature search (appendix). Data extracted included the overall surveillance strategy, sequencing availability, target population, sampling method, diagnostic criteria, and sequenced volume. Given that the surveillance strategy and density may change with time, we only gathered information on the most recent surveillance strategy (as of 20 August 2021).

Dataset of SARS-CoV-2 sequences in publicly repositories

SARS-CoV-2 sequences along with related metadata file were downloaded from an online coronavirus analysis platform from the National Genomics Data Center (NGDC)[22], where has merged and deduplicated sequences that deposited in GISAID[23], GenBank[24], National Genomics Data Center[25], National Microbiology Data Center[26], and China National GeneBank[27]. Detailed process of integration and deduplication was previously reported[28]. An acknowledgement table for those contributing to this work is available in the Appendix.

Official aggregated dataset

To gain additional insights regarding the public availability (sharing) extent of genomic data of SARS-CoV-2 variants, we extracted country-specific, variant-specific, and time-specific aggregated data on the number of SARS-CoV-2 variant cases from official websites, using the same sources as above, except for the literature source. The search was done by either directly locating to the official website for each country or indirectly searching in search engines (Google, Bing, Baidu) by using the terms "variant" and country name. To supplement the aggregated data of variants that we collected, we also downloaded the aggregated data with a valid denominator (namely, the number of isolates sequenced is reasonable) from the European Surveillance System (TESSy). The aggregated dataset included country name, date of report or collection, new or cumulative numbers of different SARS-CoV-2 variant cases and total sequenced cases, as well as the new number of confirmed cases. COVID-19 epidemic data were derived from WHO[29]. Considering that the diagnostic criteria of SARS-CoV-2 variants varies in different countries, the general principle for collecting aggregated

data was to give priority to the results based on whole and partial genome sequencing instead of those based on a PCR assay.

For countries noted by WHO as having had VOCs identified, but no data in publicly repositories, we collected the information about when VOC/VOCs were first detected from country's Ministry of Health and official media news. The search of media news was also done in search engine queries (Google, Bing, Baidu) using combined terms "first" and "variant" and country name.

All data were entered into a structured database by a trained team (co-authors). All recorded data were cross-checked by coauthors. Details of data sources and data completeness are shown in Appendix (Table S1-S2, S4-S5).

Data analysis

We used the variant naming system proposed by WHO, where four VOCs and five Variants of Interest (VOIs, included Eta, Iota, Kappa, Lambda, and Mu) had been designated as of 2 September 2021[5]. Our study focused on analyses on four VOCs in the 194 Member States of WHO. We did not integrate data from the overseas territories into that country's data. Given most countries weekly released aggregated data of variant, most of our analyses were performed on a weekly basis.

To characterize the global diversity of SARS-CoV-2 genomic surveillance, we classified the surveillance strategy of each country into three categories: 1) routine genomic surveillance; 2) limited routine genomic surveillance; and 3) no routine genomic surveillance. Routine genomic surveillance was defined as conducting nationwide genomic sequencing, coupled with at least 150 specimens per week or 10% of all samples sequenced[30]. The classification of most African countries was from a pre-defined version from African CDC[31], while other countries were subsequently defined according to the definition criteria (Table S3). As we could not identify information on the surveillance strategy for some countries through public sources, we also classified three extra categories according to the public availability/ability of genomic sequencing: 1) high availability; 2) moderate availability; and 3) low availability (Table S3).

The process of data cleaning in publicly repositories and aggregated dataset is shown in the Appendix. For data that was reported in aggregate, when the date of sample collection was not available, we assumed a fixed three-week lag[32] from sample collection to reporting unless other country-specific information was available to inform this extrapolation.

The sequencing coverage was inferred by using the percent of cumulative positives sequenced as a proxy, defined by the ratio of the number of isolates sequenced to the number of confirmed cases in the same unit of time. Given that not all sequences will be uploaded to genomic repositories, we analysed the public availability extent of genomic data by comparing the cumulative number of variants between public repositories and official aggregated datasets. Since the Alpha variant had a characteristic deletion of amino acids 69-70 in the SGTF that can be detected via a widely used PCR assay[33], we performed this analysis across VOCs.

We plotted the earliest time when the first VOC specimen was identified in each country. The earliest identification was defined by the earliest sampling time of sequences deposited in publicly repositories. If a VOC was identified by WHO but no corresponding sequence in publicly repositories for one country, we used the date collected from other sources. The sequences with sampling date earlier than the earliest sample identified in the United Kingdom (for Alpha), South Africa (for Beta), Brazil (for Gamma) and India (for Delta), respectively, were not used in analyses.

We also described the global and regional prevalence trend of variants. The prevalence of variant was defined as the proportion of the variant number to the sequencing number in the same period. When multiple data sources were available for one country, the most abundant dataset was chosen. For example, if there was publicly available genomic data as well as an official aggregated dataset, the priority was given to the one with highest reported sequenced numbers in a specific week.

Statistical significance was tested using the Chi-square test for comparing ratios of two groups, and differences were considered statistically significant at P-value < 0.05. All statistical analyses and visualizations were done using R (version 4.0.2).

Results

We classified genomic surveillance strategies for 105 countries, including 72.3% (34/47) of WHO-defined Africa Region countries, 58.5% (31/53) of European Region countries, 61.9% (13/21) of countries in the Eastern Mediterranean Region, 31.4% (11/35) of countries in Americas Region, 37.0% (10/27) of countries in the Western Pacific Region, and 54.5% (6/11) of countries in the South East Asia Region (Table S4). We downloaded a total of 3.1 million deduplicated SARS-CoV-2 sequence samples from publicly repositories corresponding to samples collected between 1 December 2019 to 20 August 2021 in 163 countries. Additionally, we collected official aggregated data of variants from 55 countries, and extra data for the first identification of COVID-19 variants from 33 countries.

The global diversity of SARS-CoV-2 genomic surveillance strategies and sequencing availability

We observed marked geographical heterogeneity in genomic surveillance of SARS-CoV-2 across countries. Globally, a total of 60.0% (63) countries had performed routine genomic surveillance, 26.7% (28) countries implemented limited routine genomic surveillance, 13.3% (14) countries had no routine genomic surveillance with the remaining countries (89) having no data on genomic surveillance strategy identified (Figure 1A). Surveillance diversity across various countries was also reflected in the context of target populations, sampling method, sequenced proportion, and diagnostic criteria (Table S4). Specifically, 32 countries randomly selected or used all confirmed cases with sufficient quality for sequencing, and 18 countries adopted PCR assay to screen or confirm variants. From the regional perspective, limited or no routine genomic surveillance was common in the Eastern Mediterranean (84.6%, 11/13) and Africa (61.8%, 21/34), followed by the South East Asia (50.0%, 3/6), Americas (36.4%, 4/11), Western Pacific (20.0%, 2/10), and Europe (3.2%, 1/31) (Figure 1A). However, among the 167 Member States with data accessible, the availability of SARS-CoV-2 sequences was high in 79 countries, moderate in 80 countries, and low in 8 countries (Figure 1B).

Sequencing coverage of SARS-CoV-2 confirmed cases

The global volume of genomic data gradually increased over time until April 2021 (Figure 1C). The European (59.4%) and Americas (33.2%) Regions uploaded the most sequences to public repositories, with marked intra-region heterogeneity across countries, with a range from 3 (Saint Kitts and Nevis) to 847,000 (United States) as of 20 August 2021 (Figure 1D).

Since September 2020, no more than 3.5% of global confirmed SARS-CoV-2 infections were sequenced. In any week, the Africa, South East Asia and Eastern Mediterranean Regions (Figure 1E) sequenced no more than 1.2% of cases. At the end of May 2021(start of Delta widely spreading), Europe had the highest sequenced proportion of 9.3%, followed by Western Pacific (2.1%), Americas (1.4%), Africa (1.0%), South-East Asia (0.1%), and Eastern Mediterranean (0.07%) (Figure 1E). At the country-level, higher rates of sequencing were observed in Iceland, New Zealand, Australia, Denmark, Luxembourg, Finland, United Kingdom, and Norway, all of which had at least 10% reported infections sequenced as of 20 August 2021. In addition, almost all countries in the Africa and Eastern Mediterranean had sequenced less than 2.5% of confirmed infections, except for Gambia (6.6%) and Mauritius (3.5%) (Figure 1F).

Public availability extent of SARS-CoV-2 variant sequences

The public availability extent of SARS-CoV-2 genomic data varied across variants and countries. Overall, among countries with aggregated data on the number of variant infections, less than half of sequences of Alpha, Beta, Gamma, and Delta were publicly available in 45.3% (24/53), 27.9% (12/43), 33.3% (11/33), and 52.3% (23/44) countries, respectively (Figure 2). However, the result for

Alpha might be influenced by SGTF detected via PCR. The public availability extent of Delta variants across countries ranged from 0.0% (Cyprus, Hungary, Iceland, Laos) to 92.9% (Denmark). At the country level, low sharing proportions (less than 50%) across all VOCs was found in several countries, including Austria, Cyprus, Greece, Hungary, Iceland, Philippines, Thailand and Senegal. For example, the sharing proportion of Alpha, Beta, and Delta in Thailand was 5.2%, 13.6%, and 2.0%, respectively, which indicated more than 85.0% genomic data of variants were not uploaded.

Moreover, incomplete metadata attached to GISAID sequences was common globally, with about two thirds of sequences missing demographic information (age and sex), and more than 95% of that missing clinical information (e.g., symptom history, clinical outcome, and vaccination status) (Table S5). We found that high-income regions tend to have lower information completeness ($P < 0.0001$, Chi-square test), especially in the European Region, where less than 25.0% and 3.0% of sequences had demographic and clinical information, respectively. Besides, 92.8% sequences were reported by subnational geographies, with a low proportion in Western Pacific (52.0%) and Eastern Mediterranean (76.9%).

Earliest identification of SARS-CoV-2 variants across regions

Alpha was first identified in the Europe, then in Americas and South-East Asia in September-October 2020, followed by spread to Eastern Mediterranean, Africa, and Western Pacific in November 2020 (Figure 3A). The earliest publicly available sequenced Beta infection was sampled in Africa in May 2020, and subsequently identified in other regions (Figure 3B). Gamma variant detection remained spatially constrained after it was first identified in Brazil (Figure 3C). After the first identification of Delta in October 2020 in South-East Asia, global spread occurred after January 2021 (Figure 3D).

Global and regional spread of SARS-CoV-2 variants

The number of new VOC cases dramatically increased until April 2021, with a peak weekly value of about 100,000 VOC cases sequenced in which most of them were Alpha variants (Figure 5A). Subsequently, another peak of weekly new VOC case occurred in July 2021, but with a large amount of Delta variants. The number of VOC cases may be an underestimate for the most recent weeks due to a collection-to-report time delay. Notably, this increase was also accompanied by the increase in the volume of new sequenced cases and new COVID-19 confirmed cases.

The global prevalence of reference (non-variant) strains fell into a low level of 0.6% in the period of Jul-Aug 2021, compared with 13.4% of that in 2020 (Figure 4). Globally, the COVID-19 pandemic was driven by the circulation of Alpha at the start of 2021, with an average prevalence of 50.6% in the first quarter of 2021. Alpha variants continued to outcompete other strains in the second quarter of

2021, accounting for 60.0% of the contemporary lineages (Figure 4). However, the rapid global rise of the Delta variant began in May 2021, reaching a global prevalence of nearly 96.6% at the end of July 2021 (Figure 5B). In contrast, Beta and Gamma variants remained at low prevalence (Figure 4). Additionally, the shifting of predominant variants from Alpha to Delta first occurred in South-East Asia where the proportion of Delta exceeded 60.0% in April 2021 (Figure 5J).

Discussion

Our study characterized the global diversity of genomic surveillance strategies and sequencing availability, sequenced coverage of SARS-CoV-2 cases, public availability extent of variant sequences, as well as current epidemic situation of SARS-CoV-2 variants. We found that genomic surveillance strategies were globally heterogenous, with limited or no routine surveillance among many countries in the Africa and Eastern Mediterranean Regions. Our analysis of publicly deposited SARS-CoV-2 sequences implied that the sequenced coverage is low in most countries, with a low proportion of VOCs sequences shared to public repositories. The pervasive spread of Alpha and Delta variants further highlights the threat of SARS-CoV-2 mutations despite the availability of vaccines in many countries. Overall, we describe the pandemic dynamics shaped by VOCs and call for more work on defining the ideal sampling schemes for different purposes (e.g., identifying new variants) with an additional call to share these data in public repositories to allow for further rapid scientific discovery.

The diversity of SARS-CoV-2 genomic surveillance between countries is associated with country-specific priorities (e.g., surveillance objectives, targeted monitoring, or event-/risk-based sequencing) and available resources. ECDC recommends population-based and/or targeted sampling strategies (e.g., imported cases, cluster cases, and potential vaccine escapers) for genomic surveillance, which could provide a more representative estimate of the relative prevalence of variants. Notably, several countries, many of which are classified as low- or lower middle-income countries by the World Bank, lack genomic surveillance data, likely due to limitations in infrastructure capacity and resources. However, even some countries classified as high-income, have suffered from a slow and inconsistent process of adopting genomics-based surveillance[34]. Establishment of reference laboratories and networks to provide sequencing services for countries without established sequencing capacity may enable improved detection and tracking of emerging variants worldwide.

The detection of most variants relies on the full-length or partial genomic sequencing, but the sequences only become available for the global community when the laboratories have established sequencing capacity, willing to share, and legally allowed to upload them. The discrepancies in sharing was observed in each region, which confirmed that some countries are sequencing but are not uploading. However, our study observed a sharing extent of exceed 100% exists in some countries, likely due to delays in the official reporting of sequencing results, or the incomplete official reporting system. The timely sharing of those enables to adequately contextualize local data when looking at introductions and examine transmission routes, as well as to look for sites of repeated mutations that can guide laboratory work on characterizing those mutations effects on therapeutics and vaccine

efficacy. The underlying reasons why some countries didn't share might be related with the distrust for publicly repositories in the concept of data security.

We found relatively low completeness of demographic and clinical characteristics in metadata accompanying uploaded sequences. Our analysis revealed that high-income countries frequently did not share demographic information. A possible reason for this is that these regions may having more restrictive data privacy/laws preventing/discouraging the release of this information. Genomic data coupled with those additional data can maximize the utility of genomic sequences in rapid scientific discovery during this pandemic, which are valuable for in-depth epidemiological analyses to characterize risk factors, clinical severity, and other public health risk of variants[35-37]. Therefore, it's vital to optimize the sharing of information in a secure and trusted channel in the context of protect patient anonymity and in accordance of local regulations[38]. In addition, decreasing the lag between sample collection to deposition of these sequences[39], including the timely sharing and standardizing of metadata [18,30,35], may facilitate the design and development of treatment and prevention strategies by policy-makers[40].

An important role of genomic surveillance is to investigate the spread and dynamics of SARS-CoV-2 variants. Amidst the emergence of different variants, the current dominance of the Delta variant suggests that it may possess higher fitness than other variants, which might be associated with a combination of higher virus load, and shorter incubation period and serial intervals[41-43]. The decrease in real-world vaccine effectiveness against Beta variants[11] and increasing breakthrough cases with Delta variants[44] underlines the importance of determining the local or regional patterns of variant spread, including the need to develop new or modified vaccines to achieve adequate protection[45].

Our results should be interpreted in view of several limitations. First, the lack of data from some countries limited our global mapping. The data completeness and quality could be impacted by key steps in the surveillance or reporting, including differences in diagnostic criteria, under-reporting, delayed reporting, and reporting methods. The inconsistent diagnostic criteria of variants might cause sampling bias, especially when adopting PCR assay to detect Alpha variant owing to its non-specificity[46]. We did an extensive search to collect multi-source data and chose the aggregated data with a priority to sequencing results rather than PCR-screening results. Second, the analysis of global and national spread could be biased as data from public repositories or aggregated dataset are not always representative of the variants circulating in the regions, especially for the regions with relatively limited sequencing capacity or with investigating outbreak-based events. Therefore, the global prevalence of variants may be biased due to the uneven sequencing across the regions. Indeed, it's difficult to obtain a truly representative and random sample, and how to understand these biases will become important[36]. Lastly, the detailed demographical, epidemiological and clinical information

about variant cases cannot be accessed, which limited our further epidemiological analysis about variant spread.

In conclusion, our study provides a landscape for genomic surveillance, the global coverage of sequencing and public availability of sequences, as well as the evidence for rapid spread of SARS-CoV-2 variants. Our findings suggest that global SARS-CoV-2 genomic surveillance strategies and capacity are diverse, and may be limited in some regions, especially in the context of the global spread and dominance of variants of concern. The gap still exists in sequencing availability and magnitude, therefore international efforts are needed to address some genomic bottleneck, such as lack of tracking representative samples, lack of sequencing capacity, strict regulations about data sharing, and lack of funding [46,47].

Contributors

H.Y. designed and supervised the study. Z.C., J.Z., Y.T., R.S, X.X., Y.W., W.L., S.G., and Z.Z. collected data. Z.C., J.Z., Y.T., R.S, and X.X. prepared the tables. Z.C. analysed data, prepared the figures, and wrote the first draft of the manuscript. A.S.A., X.C., J.Y., D.T.L., D.B.D, and H.Y. interpreted the results and revised the content critically. All authors contributed to review and revision and approved the final manuscript as submitted and agree to be accountable for all aspects of the work.

Declaration of interests

H.Y. has received research funding from Sanofi Pasteur, GlaxoSmithKline, Yichang HEC Changjiang Pharmaceutical Company, and Shanghai Roche Pharmaceutical Company. None of those research funding is related to COVID-19. All other authors report no competing interests.

Role of the funding source

The funder had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Data sharing

All datasets generated and analysed are available in the Article and Appendix. Any additional information is available from the lead contact upon request.

Acknowledgments

We thank all the authors contributed to generating and sharing sequences to GISAID, GenBank, National Genomics Data Center, National Microbiology Data Center, and China National GeneBank. We thank all the authors contributed to generating and sharing aggregated data to TESSy in ECDC. We thank Dr.Lina Ma, Dr.Yiming Bao, and their team from China National Center for Bioinformation for building 2019nCoVR dataset. We thank Junbo Chen, Xiaowei Deng from Fudan University. This study was funded by Key Program of the National Natural Science Foundation of China (grant no. 82130093 to HJY) and the US National Institutes of Health (R01 AI135115 to DTL and ASA; KL2TR001448 to DD).

Figure legends

Fig 1. Global SARS-CoV-2 genomic surveillance, sequencing availability, and publicly deposited genomic data.

(A) The global distribution of three strategies of SARS-CoV-2 genomic surveillance. (B) The global availability of SARS-CoV-2 sequencing, countries with a high level of availability represent the ability to perform in-country SARS-CoV-2 sequencing alone. (C) The weekly number of publicly deposited SARS-CoV-2 genomic data by region. (D) Cumulative number of publicly deposited SARS-CoV-2 genomic data by countries as of 20 August 2021. (E) The weekly proportion of infections sequenced by region. (F) Cumulative proportion of infections sequenced by countries as of 20 August 2021, defined as the proportion of cumulative isolates sequences to the cumulative confirmed cases. The number of sequences for the most recent weeks might be incomplete due to a time delay between collecting specimens and sequencing submission. Data unavailable, includes those locations that not belonged to 194 Member States or had not applicable data.

Fig 2. The public availability extent of SARS-CoV-2 genomic data to publicly repositories by country and the variants of concern.

The public availability extent was defined as the ratio of the cumulative number of variants in publicly repositories to the official reported number of variants within the same period. The ratio of exceed 100% in some countries might be due to the delay in the official report of the sequencing results or the incomplete official reporting system. In view of the availability of official data, the cumulative number of variants in different countries corresponds to different time periods, with detailed information in Table S6. Sequence without date of specimen collection in publicly repositories is not included in our analysis. The variant data for China only includes cases reported by mainland China. The officially reported number of Alpha variants might contain those confirmed by the PCR screen assay. The values beneath the country names show numbers of cumulative variant as of one specific week: variants in publicly repositories/official reported variants.

Fig 3. The earliest identification of Alpha, Beta, Gamma, and Delta variant in each country.

If information about the earliest sampling date was unavailable but that of the earliest reporting date was available, we extrapolated the sampling date by using a fixed three-week lag from sample collection to reporting. Countries with darker red indicate earlier samples, and with darker blue refers to later samples. The white areas represent countries with variants unreported or data unavailable.

Fig 4. The prevalence and temporal dynamics of reference strains and four SARS-CoV-2 VOCs by country.

The date presented in the top refers to the range of date of specimen collection. The prevalence was defined as the proportion of the strain number (reference strains or variants) to the total number of sequences generated in the same unit of time. Reference strains includes lineage A, A.1, B, and B.1; the sub-lineages of four VOCs are aggregated with the parent lineages. The grey areas represent countries with no COVID-19 epidemic, or no performing sequencing, or no uploading genomic data to publicly repositories.

Fig 5. The number and proportion of SARS-CoV-2 variants by region and time.

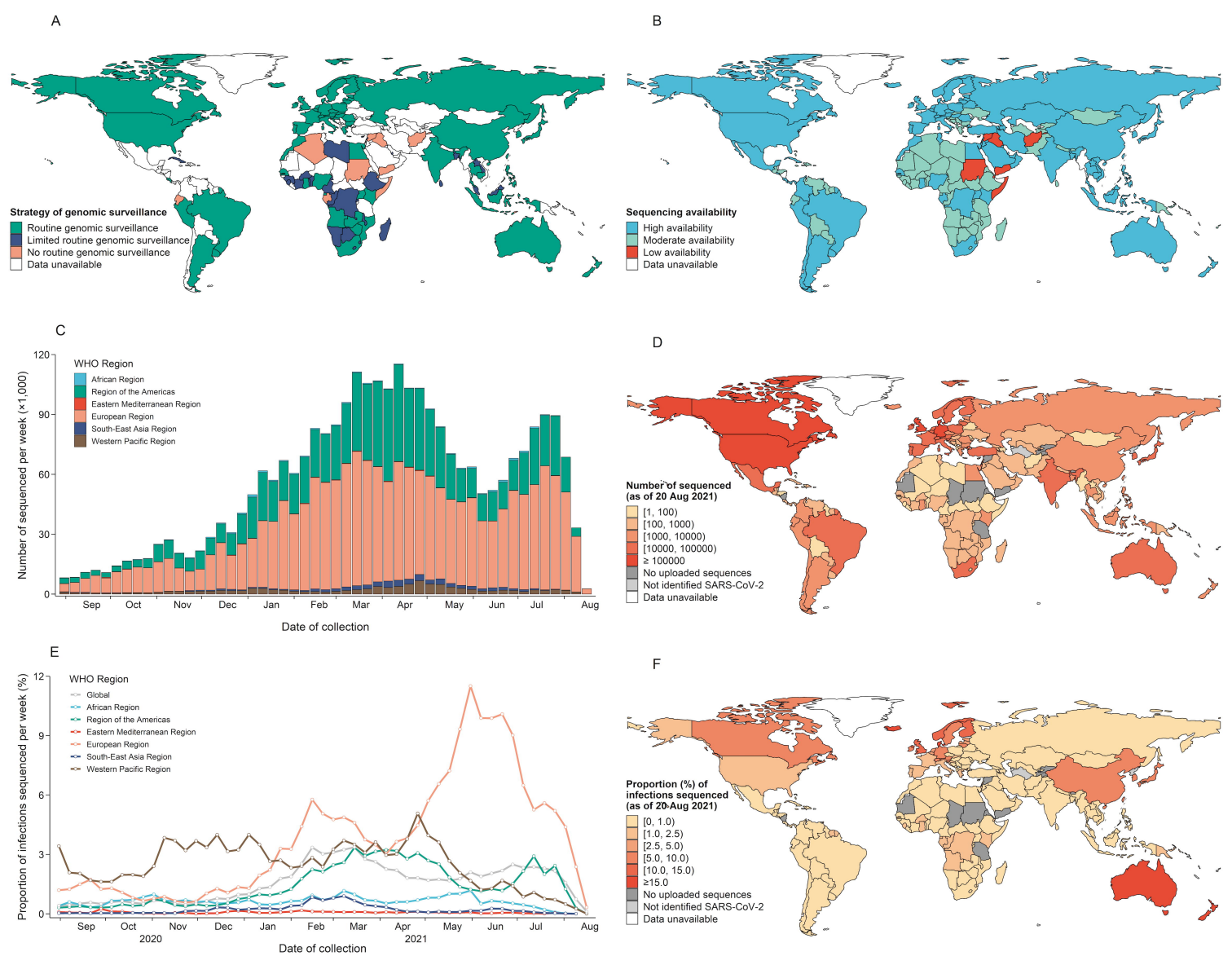
The line and point in the left figure correspond to the y-axis on the right. The sub-lineages of four VOCs are aggregated with the parent lineages; designated Variants of Interest (VOIs) included lineage B.1.525, B.1.526, B.1.617.1, C.37, B.1.621 and their sub-lineages; other lineages included reference strains and other variants. Data used here are derived from the publicly repositories and aggregated dataset, with priority given to the one with highest sequenced numbers in a specific week.

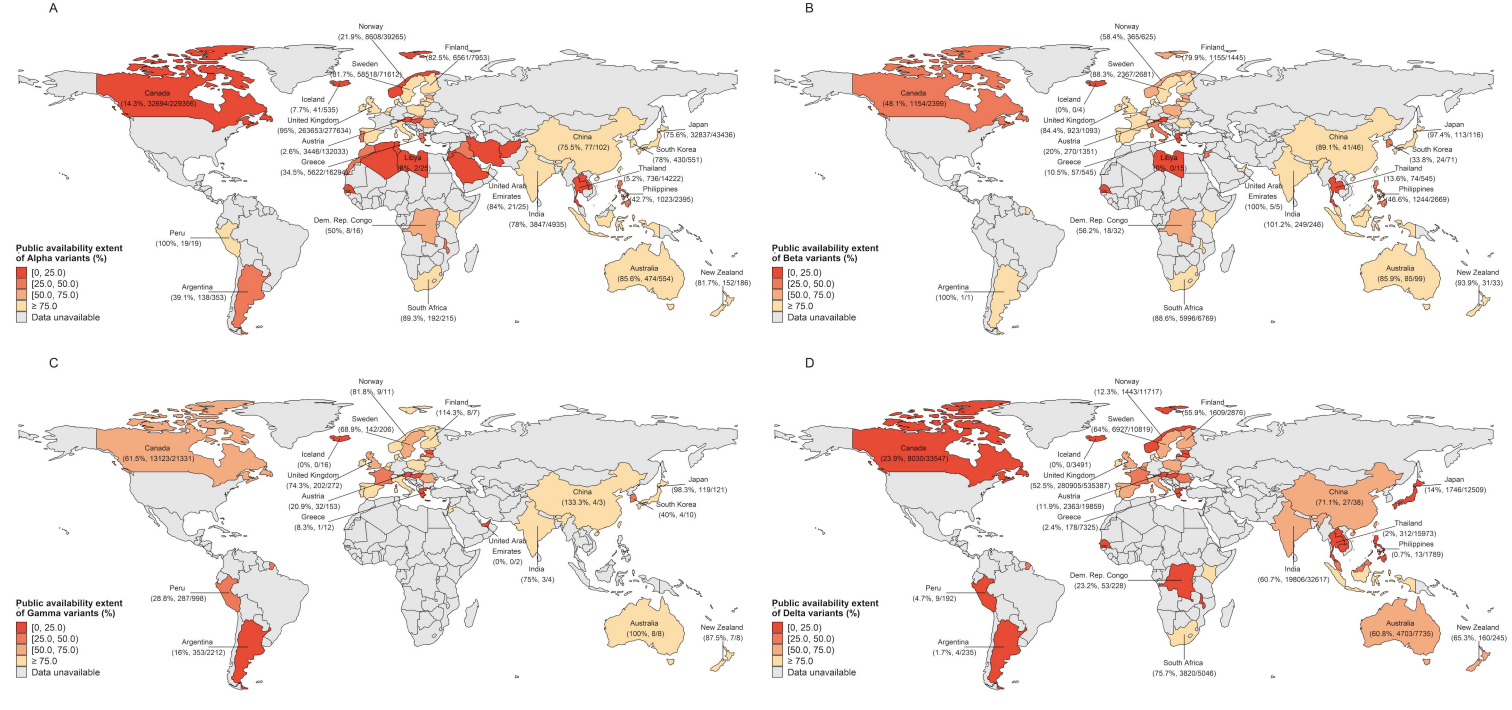
Reference

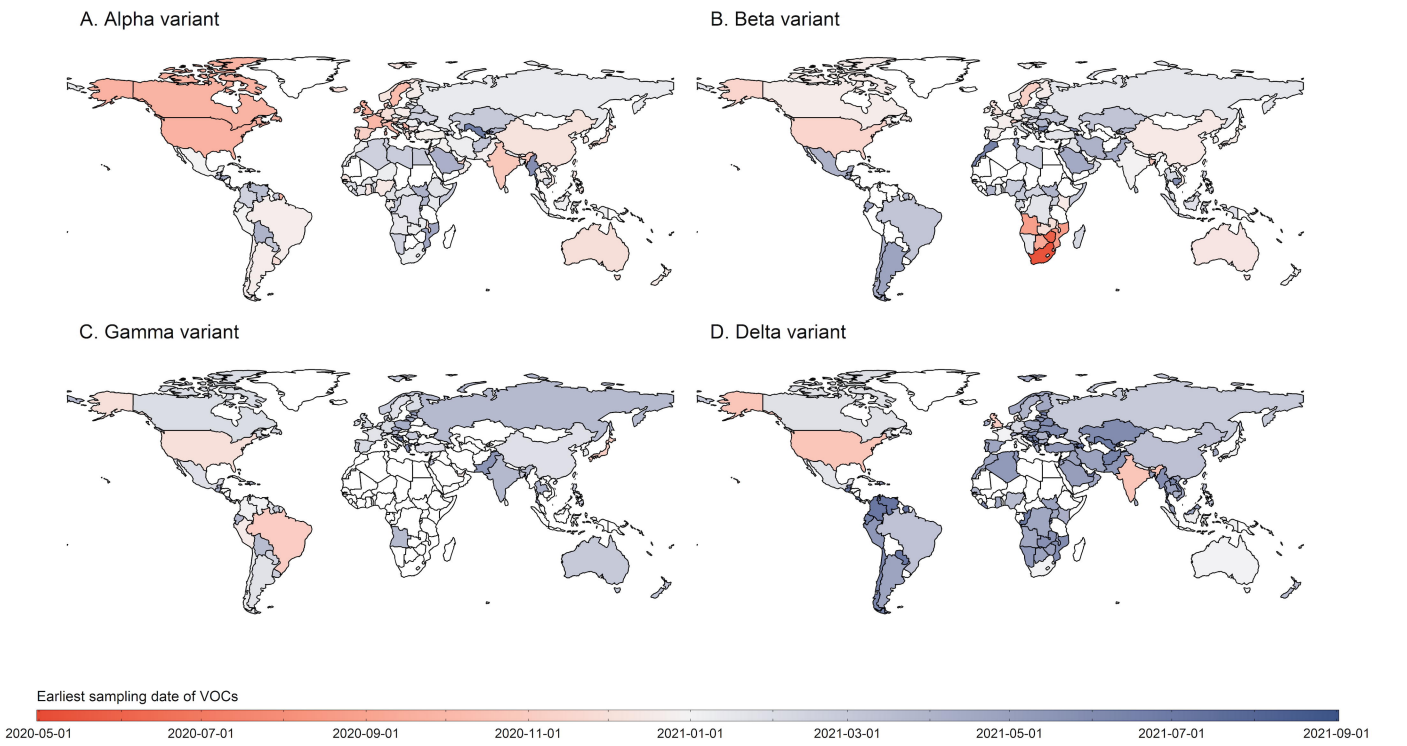
1. Galloway SE, Paul P, MacCannell DR, et al. Emergence of SARS-CoV-2 B.1.1.7 Lineage - United States, December 29, 2020-January 12, 2021. *MMWR Morbidity and mortality weekly report* 2021; **70**(3): 95-9.
2. Davies NG, Abbott S, Barnard RC, et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science (New York, NY)* 2021; **372**(6538).
3. Volz E, Mishra S, Chand M, et al. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* 2021; **593**(7858): 266-9.
4. Wall EC, Wu M, Harvey R, et al. AZD1222-induced neutralising antibody activity against SARS-CoV-2 Delta VOC. *Lancet (London, England)* 2021; **398**(10296): 207-9.
5. World Health Organization. Tracking SARS-CoV-2 variants. 2021. https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/ (accessed 18 Jul 2021).
6. Campbell F, Archer B, Laurenson-Schafer H, et al. Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2021; **26**(24).
7. Ito K, Piantham C, Nishiura H. Predicted dominance of variant Delta of SARS-CoV-2 before Tokyo Olympic Games, Japan, July 2021. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2021; **26**(27).
8. Scientific Advisory Group for Emergencies (SAGE). SPI-M-O: Consensus Statement on COVID-19. 2021. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/986709/S1237_SPI-M-O_Consensus_Statement.pdf.
9. Dhar MS, Marwal R, Radhakrishnan VS, et al. Genomic characterization and Epidemiology of an emerging SARS-CoV-2 variant in Delhi, India. *medRxiv* 2021: 2021.06.02.21258076.
10. Chen X, Chen Z, Azman AS, et al. Neutralizing antibodies against SARS-CoV-2 variants induced by natural infection or vaccination: a systematic review and pooled meta-analysis. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2021.
11. Abu-Raddad LJ, Chemaitelly H, Butt AA. Effectiveness of the BNT162b2 Covid-19 Vaccine against the B.1.1.7 and B.1.351 Variants. *The New England journal of medicine* 2021; **385**(2): 187-9.
12. Sheikh A, McMenamin J, Taylor B, Robertson C. SARS-CoV-2 Delta VOC in Scotland: demographics, risk of hospital admission, and vaccine effectiveness. *The Lancet* 2021; **397**(10293): 2461-2.
13. Li XN, Huang Y, Wang W, et al. Efficacy of inactivated SARS-CoV-2 vaccines against the Delta variant infection in Guangzhou: A test-negative case-control real-world study. *Emerging microbes & infections* 2021: 1-32.
14. Vogels CBF, Breban MI, Ott IM, et al. Multiplex qPCR discriminates variants of concern to enhance global surveillance of SARS-CoV-2. *PLoS biology* 2021; **19**(5): e3001236.
15. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. *Nature* 2020; **579**(7798): 265-9.
16. Hodcroft EB, Zuber M, Nadeau S, et al. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* 2021.
17. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global challenges (Hoboken, NJ)* 2017; **1**(1): 33-46.
18. European Centre for Disease Prevention and Control. Sequencing of SARS-CoV-2: first update, 2021.
19. Bugembe DL, Phan MVT, Ssewanyana I, et al. Emergence and spread of a SARS-CoV-2 lineage A variant (A.23.1) with altered spike protein in Uganda. *Nature microbiology* 2021.
20. Hodcroft EB, Zuber M, Nadeau S, et al. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* 2021.
21. Wilkinson E, Giovanetti M, Tegally H, et al. A year of genomic surveillance reveals how the SARS-CoV-2 pandemic unfolded in Africa. *medRxiv* 2021: 2021.05.12.21257080.
22. Gong Z, Zhu JW, Li CP, et al. An online coronavirus analysis platform from the National Genomics Data Center. *Zoological research* 2020; **41**(6): 705-8.

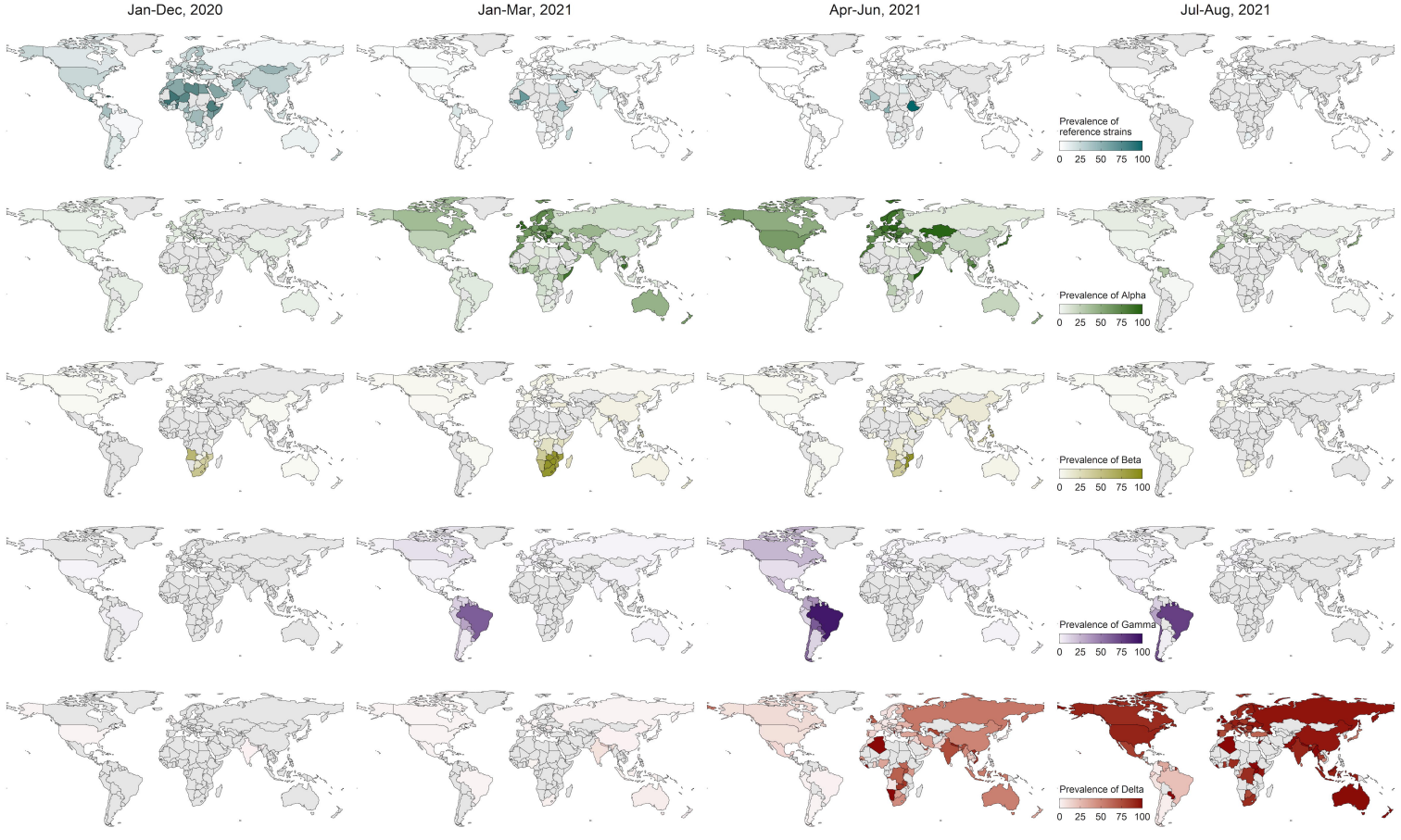
23. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 2017; **22**(13).
24. Hatcher EL, Zhdanov SA, Bao Y, et al. Virus Variation Resource - improved response to emergent viral outbreaks. *Nucleic acids research* 2017; **45**(D1): D482-d90.
25. National Genomics Data Center Members and Partners. Database Resources of the National Genomics Data Center in 2020. *Nucleic acids research* 2020; **48**(D1): D24-d33.
26. Shi W, Qi H, Sun Q, et al. gcMeta: a Global Catalogue of Metagenomics platform to support the archiving, standardization and analysis of microbiome data. *Nucleic acids research* 2019; **47**(D1): D637-d48.
27. Xiao SZ, Armit C, Edmunds S, et al. Increased interactivity and improvements to the GigaScience database, GigaDB. *Database : the journal of biological databases and curation* 2019; **2019**.
28. Song S, Ma L, Zou D, et al. The Global Landscape of SARS-CoV-2 Genomes, Variants, and Haplotypes in 2019nCoVR. *Genomics, proteomics & bioinformatics* 2020; **18**(6): 749-59.
29. World Health Organization. WHO Coronavirus (COVID-19) Dashboard. https://covid19.who.int/info/.
30. World Health Organization. Operational considerations to expedite genomic sequencing component of GISRS surveillance of SARS-CoV-2, 2021.
31. Africa CDC. Africa Pathogen Genomics Initiative (PGI). 2021. https://africacdc.org/institutes/africa-pathogen-genomics-initiative/ (accessed Jul 18 2021).
32. Paul P, France AM, Aoki Y, et al. Genomic Surveillance for SARS-CoV-2 Variants Circulating in the United States, December 2020-May 2021. *MMWR Morbidity and mortality weekly report* 2021; **70**(23): 846-50.
33. Brown KA, Gubbay J, Hopkins J, et al. S-Gene Target Failure as a Marker of Variant B.1.1.7 Among SARS-CoV-2 Isolates in the Greater Toronto Area, December 2020 to March 2021. *Jama* 2021; **325**(20): 2115-6.
34. Crawford DC, Williams SM. Global variation in sequencing impedes SARS-CoV-2 surveillance. *PLoS genetics* 2021; **17**(7): e1009620.
35. World Health Organization. Genomic sequencing of SARS-CoV-2: A guide to implementation for maximum impact on public health, 2021.
36. World Health Organization. Guidance for surveillance of SARS-CoV-2 variants: Interim guidance. 2021. https://www.who.int/publications/i/item/WHO_2019-nCoV_surveillance_variants (accessed August 26 2021).
37. Page AJ, Mather AE, Le-Viet T, et al. Large-scale sequencing of SARS-CoV-2 genomes from one region allows detailed epidemiology and enables local outbreak management. *Microbial genomics* 2021; **7**(6).
38. Black A, MacCannell DR, Sibley TR, Bedford T. Ten recommendations for supporting open pathogen genomic analysis in public health. *Nature medicine* 2020; **26**(6): 832-41.
39. Capoferri AA, Shao W, Spindler J, Coffin JM, Rausch JW, Kearney MF. 2020 SARS-CoV-2 diversification in the United States: Establishing a pre-vaccination baseline. *medRxiv* 2021: 2021.06.01.21258185.
40. Robishaw JD, Alter SM, Solano JJ, et al. Genomic surveillance to combat COVID-19: challenges and opportunities. *The Lancet Microbe* 2021.
41. Li B, Deng A, Li K, et al. Viral infection and transmission in a large, well-traced outbreak caused by the SARS-CoV-2 Delta variant. *medRxiv* 2021: 2021.07.07.21260122.
42. Meng Z, Jianpeng X, Aiping D, et al. Transmission Dynamics of an Outbreak of the COVID-19 Delta Variant B.1.617.2 — Guangdong Province, China, May–June 2021. *China CDC Weekly* 2021; **3**(27): 584-6.
43. Pung R, Mak TM, Kucharski AJ, Lee VJ. Serial intervals in SARS-CoV-2 B.1.617.2 variant cases. *The Lancet*.
44. Brown CM, Vostok J, Johnson H, et al. Outbreak of SARS-CoV-2 Infections, Including COVID-19 Vaccine Breakthrough Infections, Associated with Large Public Gatherings - Barnstable County, Massachusetts, July 2021. *MMWR Morbidity and mortality weekly report* 2021; **70**(31): 1059-62.

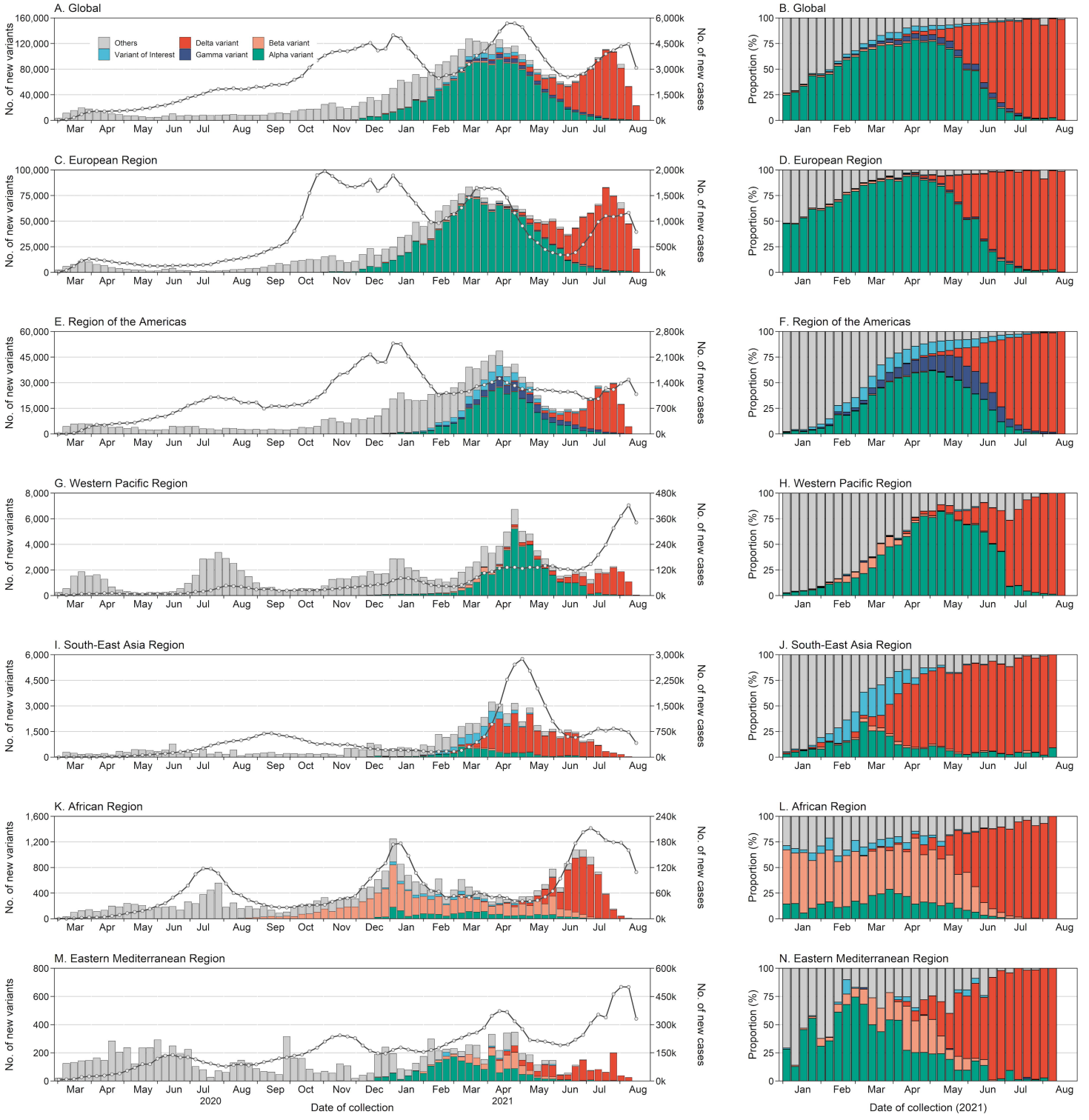
588 45. Krause PR, Fleming TR, Longini IM, et al. SARS-CoV-2 Variants and Vaccines. *New*
589 *England Journal of Medicine* 2021; **385**(2): 179-86.
590 46. Hodcroft EB, De Maio N, Lanfear R, et al. Want to track pandemic variants faster? Fix the
591 bioinformatics bottleneck. *Nature* 2021; **591**(7848): 30-3.
592 47. Wadman M. United States rushes to fill void in viral sequencing. 2021; **371**(6530): 657-8.
593


A
Strategy of genomic surveillance
Routine genomic surveillance
Limited routine genomic surveillance
No routine genomic surveillance
Data unavailable
B
Sequencing availability
High availability
Moderate availability
Low availability
Data unavailable
C
WHO Region
African Region
Region of the Americas
Eastern Mediterranean Region
European Region
South-East Asia Region
Western Pacific Region
Number of sequenced per week (×1,000)
120
90
60
30
0
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Date of collection
D
Number of sequenced (as of 20 Aug 2021)
[1, 100)
[100, 1000)
[1000, 10000)
[10000, 100000)
≥ 100000
No uploaded sequences
Not identified SARS-CoV-2
Data unavailable
E
WHO Region
Global
African Region
Region of the Americas
Eastern Mediterranean Region
European Region
South-East Asia Region
Western Pacific Region
Proportion of infections sequenced per week (%)
12
9
6
3
0
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
2020
2021
Date of collection
F
Proportion (%) of infections sequenced (as of 20 Aug 2021)
[0, 1.0)
[1.0, 2.5)
[2.5, 5.0)
[5.0, 10.0)
[10.0, 15.0)
≥15.0
No uploaded sequences
Not identified SARS-CoV-2
Data unavailable


A
Norway
(21.9%, 8608/39265)
Finland
(82.5%, 6561/7953)
Sweden
(61.7%, 58518/71612)
Canada
(14.3%, 32694/229356)
Iceland
(7.7%, 41/535)
United Kingdom
(95%, 263653/277634)
Austria
(2.6%, 3446/132033)
Greece
(34.5%, 5622/16294)
Libya
(0%, 0/20)
China
(75.5%, 77/102)
Japan
(75.6%, 32837/43436)
South Korea
(78%, 430/551)
Thailand
(5.2%, 736/14222)
Philippines
(42.7%, 1023/2395)
United Arab
Emirates
(84%, 21/25)
India
(78%, 3847/4935)
Peru
(100%, 19/19)
Dem. Rep. Congo
(50%, 8/16)
Australia
(85.6%, 474/554)
New Zealand
(81.7%, 152/186)
Argentina
(39.1%, 138/353)
South Africa
(89.3%, 192/215)
Public availability extent
of Alpha variants (%)
[0, 25.0)
[25.0, 50.0)
[50.0, 75.0)
≥ 75.0
Data unavailable
B
Norway
(58.4%, 365/625)
Finland
(79.9%, 1155/1445)
Sweden
(88.3%, 2367/2681)
Canada
(48.1%, 1154/2399)
Iceland
(0%, 0/4)
United Kingdom
(84.4%, 923/1093)
Austria
(20%, 270/1351)
Greece
(10.5%, 57/545)
Libya
(0%, 0/15)
China
(89.1%, 41/46)
Japan
(97.4%, 113/116)
South Korea
(33.8%, 24/71)
Thailand
(13.6%, 74/545)
Philippines
(46.6%, 1244/2669)
United Arab
Emirates
(100%, 5/5)
India
(101.2%, 249/246)
Dem. Rep. Congo
(56.2%, 18/32)
Australia
(85.9%, 85/99)
New Zealand
(93.9%, 31/33)
Argentina
(100%, 1/1)
South Africa
(88.6%, 5996/6769)
Public availability extent
of Beta variants (%)
[0, 25.0)
[25.0, 50.0)
[50.0, 75.0)
≥ 75.0
Data unavailable
C
Norway
(81.8%, 9/11)
Finland
(114.3%, 8/7)
Sweden
(68.9%, 142/206)
Canada
(61.5%, 13123/21331)
Iceland
(0%, 0/16)
United Kingdom
(74.3%, 202/272)
Austria
(20.9%, 32/153)
Greece
(8.3%, 1/12)
China
(133.3%, 4/3)
Japan
(98.3%, 119/121)
South Korea
(40%, 4/10)
United Arab
Emirates
(0%, 0/2)
India
(75%, 3/4)
Peru
(28.8%, 287/998)
Australia
(100%, 8/8)
New Zealand
(87.5%, 7/8)
Argentina
(16%, 353/2212)
Public availability extent
of Gamma variants (%)
[0, 25.0)
[25.0, 50.0)
[50.0, 75.0)
≥ 75.0
Data unavailable
D
Norway
(12.3%, 1443/11717)
Finland
(55.9%, 1609/2876)
Sweden
(64%, 6927/10819)
Canada
(23.9%, 8030/33547)
Iceland
(0%, 0/3491)
United Kingdom
(52.5%, 28909/55087)
Austria
(11.9%, 2363/19859)
Greece
(2.4%, 178/7325)
China
(71.1%, 27/38)
Japan
(14%, 1746/12509)
Thailand
(12%, 312/15973)
Philippines
(0.7%, 13/1789)
India
(60.7%, 19806/32617)
Peru
(4.7%, 9/192)
Dem. Rep. Congo
(23.2%, 53/228)
Australia
(60.8%, 4703/7735)
New Zealand
(65.3%, 160/245)
Argentina
(1.7%, 4/235)
South Africa
(75.7%, 3820/5046)
Public availability extent
of Delta variants (%)
[0, 25.0)
[25.0, 50.0)
[50.0, 75.0)
≥ 75.0
Data unavailable


A. Alpha variant
B. Beta variant
C. Gamma variant
D. Delta variant
Earliest sampling date of VOCs
2020-05-01
2020-07-01
2020-09-01
2020-11-01
2021-01-01
2021-03-01
2021-05-01
2021-07-01
2021-09-01


Jan-Dec, 2020
Jan-Mar, 2021
Apr-Jun, 2021
Jul-Aug, 2021
Prevalence of reference strains
0 25 50 75 100
Prevalence of Alpha
0 25 50 75 100
Prevalence of Beta
0 25 50 75 100
Prevalence of Gamma
0 25 50 75 100
Prevalence of Delta
0 25 50 75 100


A. Global
B. Global
C. European Region
D. European Region
E. Region of the Americas
F. Region of the Americas
G. Western Pacific Region
H. Western Pacific Region
I. South-East Asia Region
J. South-East Asia Region
K. African Region
L. African Region
M. Eastern Mediterranean Region
N. Eastern Mediterranean Region
Others
Variant of Interest
Delta variant
Gamma variant
Beta variant
Alpha variant
No. of new variants
No. of new cases
Proportion (%)
Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug
Jan Feb Mar Apr May Jun Jul Aug
2020
2021
Date of collection
Date of collection (2021)