Biosafety and Health 5 (2023) 204–210

Contents lists available at ScienceDirect

Biosafety and Health

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

Review Article

On the collection and integration of SARS-CoV-2 genome data

Lina Ma [a,b,c,1,*], Wei Zhao [a,b,c,1], Tianhao Huang [a,b,c,1], Enhui Jin [a,b,c,1], Gangao Wu [a,b,c], Wenming Zhao [a,b,c], Yiming Bao [a,b,c,*]

[a] China National Center for Bioinformation, Beijing 100101, China
[b] National Genomics Data Center & CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
[c] University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Article history:
Received 16 April 2023
Revised 3 July 2023
Accepted 3 July 2023
Available online 4 July 2023

Keywords:
SARS-CoV-2 resource
Genome data
Data deposition
Data integration
Data curation

ABSTRACT

Genome data of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is essential for virus diagnosis, vaccine development, and variant surveillance. To archive and integrate worldwide SARS-CoV-2 genome data, a series of resources have been constructed, serving as a fundamental infrastructure for SARS-CoV-2 research, pandemic prevention and control, and coronavirus disease 2019 (COVID-19) therapy. Here we present an overview of extant SARS-CoV-2 resources that are devoted to genome data deposition and integration. We review deposition resources in data accessibility, metadata standardization, data curation and annotation; review integrative resources in data source, de-redundancy processing, data curation and quality assessment, and variant annotation. Moreover, we address issues that impede SARS-CoV-2 genome data integration, including low-complexity, inconsistency and absence of isolate name, sequence inconsistency, asynchronous update of genome data, and mismatched metadata. We finally provide insights into data standardization consensus and data submission guidelines, to promote SARS-CoV-2 genome data sharing and integration.

1. Introduction

The genome data of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), including both genome sequences and the relevant metadata, play essential roles in the design of effective molecular diagnostic tests [1,2], preparation of efficient vaccine [3,4], accurate tracing of the origins and transmission of the outbreak [5–7], and development of efficacious therapeutics [8–10]. Since the outbreak of the coronavirus disease 2019 (COVID-19) in late 2019, genome sequences of SARS-CoV-2 have been generated by a large number of laboratories around the world at an unprecedented rate. As of June 2023, over 16 million SARS-CoV-2 genome sequences from 193 countries/regions have been available to the researchers according to RCoV19 (formerly known as 2019-nCoVR) [11]. Particularly, the United States and United Kingdom have contributed more than 50% of the total sequences.

Given the rapid data generation along with the duration of the pandemic, resources dedicated to SARS-CoV-2 genome data collection and integration continue to increase in importance. These resources could be broadly classified into two categories, viz., those devoted to data deposition, and those that feature genome data integration and value-added curation. Two representative repositories for data deposition are GenBank at National Center for Biotechnology Information (NCBI) [12], part of International Nucleotide Sequence Database Collaboration (INSDC) [13], and EpiCoV™ at Global Initiative on Sharing All Influenza Data (GISAID) [14]. Both serve as indispensable resources for fast deposition and dissemination of SARS-CoV-2 genome data among the scientific community. However, there is no clear evidence that GenBank (unrestricted access) and EpiCoV™ (restricted access) publicly exchange SARS-CoV-2 genome data with each other. Therefore, it is highly desired to integrate a comprehensive list of worldwide SARS-CoV-2 sequences with appropriate metadata, to support pandemic prevention and control. Towards this end, integrative resource such as RCoV19 was launched to integrate genome data, offer cross-references between different data sources, and provide systematic variation annotation and lineage monitoring of SARS-CoV-2 [11]. Also, ViruSurf was built to provide an integrated search of metadata and sequence variants over multiple sources in an effective and scalable way [15]. However, ViruSurf has not been updated since 2022.

Despite the significant efforts made by the existing resources in collecting and integrating SARS-CoV-2 genome data, important issues remain around data standardization and consistency, and connections

* Corresponding authors: National Genomics Data Center & CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China.
E-mail addresses: malina@big.ac.cn (L. Ma), baoym@big.ac.cn (Y. Bao).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.bsheal.2023.07.004
2590-0536/

between different resources are not always seamless. While the pandemic is ongoing with new Omicron subvariants (e.g., BA.4, BA.5, BQ.1.1, and XBB) emerging and becoming globally dominant [16,17], a central collection of data or interoperability among resources is critically important for fostering a better understanding and promoting new discoveries related to SARS-CoV-2. Here, we review the progress made by existing resources for SARS-CoV-2 genome data, discuss their limitations, and point out challenges ahead in terms of data distribution and sharing. We also propose appropriate frameworks for efficient data integration to provide new insights into SARS-CoV-2 data collection and pave the way for studying possible future pandemics.

2. SARS-CoV-2 genome data deposition and standardization

The repositories for SARS-CoV-2 genome data deposition are mainly designed to accept the original data submitted by users, issue accession numbers for the genome sequences, and standardize the sequences as well as related metadata for further research. A total number of seven repositories have been developed to collect SARS-CoV-2 genome data, including GenBank at NCBI [12], European Nucleotide Archive (ENA) at EMBL's European Bioinformatics Institute (EMBL-EBI) [18], DNA Data Bank of Japan (DDBJ) at National Institute of Genetics (NIG) in Japan [19], EpiCoV™ at GISAID [14], GenBase at National Genomics Data Center (NGDC) [20], China National GeneBank DataBase (CNGBdb) at China National GeneBank (CNGB) [21], and Novel Coronavirus National Science and Technology Resource Service System (hereafter referred to as NCNSTRSS for convenience) at National Microbiology Data Center (NMDC) (Fig. 1). It should be noted that GenBank, ENA, and DDBJ are INSDC member databases, and they employ unique accession prefixes (https://www.ncbi.nlm.nih.gov/genbank/acc_prefix/) to denote their individual data sources. The member databases collect data under INSDC policies, and exchange data daily to keep the identical information and represent the complete INSDC collection [13]. For conveniences, GenBank will be used hereafter to represent all three INSDC databases. GenBase, CNGBdb, and NCNSTRSS are constructed by separate data centers in China for individual demands. However, unlike INSDC members, these databases or data centers do not share data amongst one another.

2.1. Data accessibility

Most of these deposition repositories are public and open to all users (Table 1). Among these resources, GenBank archives SARS-CoV-2 sequences of genomes, other nucleotides, and proteins [22], and NCBI Virus serves as a customized search engine to support retrieval, display and analysis of GenBank SARS-CoV-2 datasets [23]. In line with GenBank, GenBase was released in 2023 to take over Genome Warehouse [24], to accept submissions of SARS-CoV-2 sequences in NGDC. As of June 19, 2023, the total INSDC collection included 7,115,780 SARS-CoV-2 nucleotide sequences (GenBank 3,541,157, ENA 3,557,035, DDBJ 1,479 sequences through direct submission; with the remaining 16,109 sequences derived from genome project data, patents, and other sources); GenBase accommodated 19,795 SARS-CoV-2 genome sequences. In addition, CNGBdb [21] and NCNSTRSS have collected a small number of 87 and 305 SARS-CoV-2 genome sequences respectively. Different from the public open repositories, data are restricted in EpiCoV™, which requires registration for data acquisition and application for data sharing [14]. EpiCoV™ has hosted the largest number of SARS-CoV-2 nucleotide sequences (15,694,915 as of June 19, 2023).

2.2. Metadata standardization

Deposition repositories play a crucial role in standardizing SARS-CoV-2 genome metadata (Table 1). These metadata typically include virus details (virus name, accession ID, passage details/history), sample information (collection date, location, host, outbreak detail, sampling strategy, specimen source, originating lab), patient information (gender, patient age, patient status, last vaccinated, treatment), sequencing and assembly information (sequencing technology, assembly method, coverage), and submission information (submitter, submission date, address, submitting lab). EpiCoV™ has provided the most comprehensive meta information by including all these sections and items, while the remaining resources provide little clinical information. The SARS-CoV-2 isolate name format recommended by International Committee on Taxonomy of Viruses, “SARS-CoV-2/host/location/isolate/date” [25], is adopted by both GenBank and GenBase, while EpiCoV™ uses the format of “hCoV-19/location/isolate/date”. In terms of sequence changes, both INSDC member databases and GenBase use accession.version format to assign identification numbers. If there is any change to the sequence data, the version numbers will be incremented accordingly.

2.3. Data curation and annotation

The deposition repositories provide essential curations and analysis (Table 1). EpiCoV™ evaluates sequence quality based on three aspects: N, Gap, Coverage. GenBank and GenBase assess nucleotide completeness. GenBank and GenBase provide links to raw data and related publication. Furthermore, GenBank and EpiCoV™ assign Pango lineages for genome sequences. Aside from basic curation and annotation, each repository offers unique annotations and tools. GenBase, CNGBdb, and NCNSTRSS provide BLAST tools for sequence retrieval, while GenBank offers gene annotations with its virus genome annotation pipeline, and provides tools for multiple alignment, phylogenetic tree construction and BLAST. Notably, EpiCoV™ offers amino acids variations, as well as global phylogeny, genome diversity, and emerging variants analysis and visualization.

3. SARS-CoV-2 genome data integration and curation

To provide a comprehensive collection of worldwide SARS-CoV-2 genomes and perform value-added curations and analysis, several databases have been built, including CoV-Seq, RCoV19, VirusDIP, and ViruSurf (Fig. 1). Databases that rely solely on one primary data source have not been included in this review. These integration resources are different in data sources, curation strategies, and variation analysis (Table 1). VirusDIP is a one-stop service platform for virus data archive, integration, and analysis [26]. CoV-Seq is an integrated web server for SARS-CoV-2 genome data visualization and analysis [27]. ViruSurf is an integrated database to enable effective search over large and curated sequence data from heterogeneous sources [15]. RCoV19 is a comprehensive database for global landscape of SARS-CoV-2 genomes, mutations, and haplotypes [28]. It is noted that RCoV19 and VirusDIP provide daily updates. As of June 19, 2023, RCoV19 and VirusDIP accommodated genome data for 16,182,187 and 15,709,705 SARS-CoV-2 sequences, respectively. However, CoV-Seq and ViruSurf have not been updated since 2020–09-28 and 2022–01-13, respectively.

3.1. Data source

All of these integration resources integrate genome data from both GenBank and EpiCoV™ (Table 1). In addition, VirusDIP incorporates

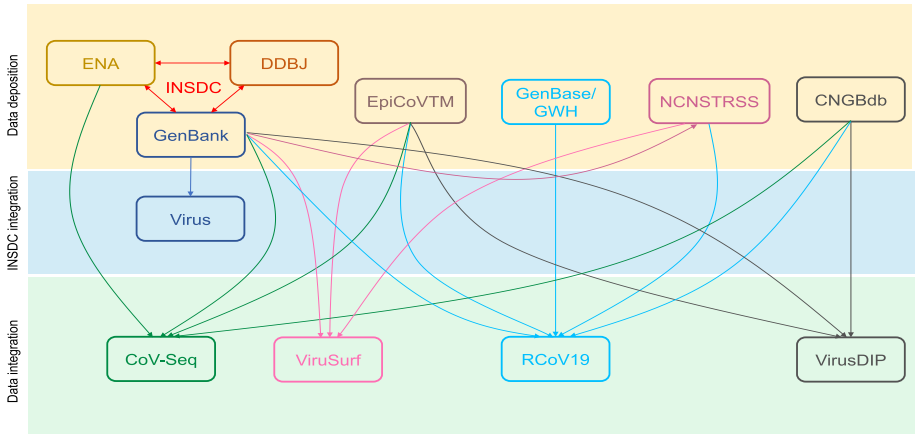

Fig. 1. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genome data resources and their relationships. SARS-CoV-2 genome data resources fall into two main categories: those focused on data deposition and those dedicated to data integration. Among the deposition repositories, GenBank (NCBI), ENA (EMBL-EBI), and DDBJ (NIG) are International Nucleotide Sequence Database Collaboration (INSDC) member databases. They three exchange data on a daily basis to ensure identical information and represent the complete INSDC collection. In contrast, GenBase (NGDC), CNGBdb (CNGB), and NCNSTRSS (which stands for "Novel Coronavirus National Science and Technology Resource Service System") (NMDC) are from separate data centers in China and do not share data with each other like the INSDC member databases do. Additionally, NCBI Virus functions as a customized search engine for INSDC SARS-CoV-2 datasets. Databases stemming from the same data center are highlighted with matching fonts and borders. Double arrows illustrate data exchange, while single arrows denote data collection.

data from CNGBdb, ViruSurf includes data from COG-UK (https://www.cogconsortium.uk/) [29] and NCNSTRSS, CoV-Seq integrates data from CNGBdb and ENA, RCoV19 collects data from CNGBdb, NCNSTRSS and GenBase. As a result, RCoV19 has collected the most comprehensive list of SARS-CoV-2 genomes worldwide, and databases such as Coronavirus GenBrowser obtains high-quality SARS-CoV-2 variants from RCoV19 to carry out global/local transmission and evolution analysis [28]. It is noted that VirusDIP provides data mirroring of EpiCoV™, while sequences from EpiCoV™ couldn't be downloaded in VirusDIP as well as other related resources due to usage rights.

3.2. De-redundancy processing

Most integration resources perform de-redundancy processing among different data sources, to provide a non-redundant list of worldwide SARS-CoV-2 genomes (Table 1). Specifically, ViruSurf identifies redundancies by comparing strain/isolate names and sequence lengths [15]; CoV-Seq identifies identical sequences between EpiCoV™ and INSDC member databases (it is unclear how) [27]; VirusDIP identifies redundancies that have identical sequences (it is unclear how); RCoV19 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. Once redundancies are identified, RCoV19 selects the first publicly released sequence to represent this genome, cross-referencing other IDs in the "Related ID" field. However, the remaining resources lack cross-referencing for redundancies.

3.3. Data curation and quality assessment

These integration resources standardize metadata terms, unify data formats, and assess sequence quality, to facilitate data screening before downstream analysis (Table 1). RCoV19 and ViruSurf use virus names and accession IDs of data sources, standardize host terms, and unify the formats of collection date and location. In addition, ViruSurf provides sequencing and assembly information (coverage of the raw data, sequencing technology, and assembly method), calculates GC% and N %, and determines sequence completeness. RCoV19 assesses sequence quality in five aspects (Ns, degenerate bases, total gaps, mutation number, and mutation density), defines high-quality sequences by considering Ns and degenerate bases, and determines completeness of the coding region.

3.4. Variation annotation

These integration resources perform variation calling, visualization, and monitoring based on data integration and curation. CoV-Seq and RCoV19 utilize in-house tools to filter out low-quality sequences and call variants. ViruSurf provides association analysis between variants and epitopes. RCoV19 offers spatiotemporal dynamics visualization, high-risk variants monitoring, and online analysis with multiple tools [11,30].

4. Issues on SARS-CoV-2 genome data integration

As noted, most of the databases identify the same sequences submitted to different sources and perform de-redundancy processing, based on the comparison of meta information and sequences. However, data integration is significantly challenged by several factors such as low-complexity, inconsistency and absence of isolate name, sequence inconsistency, asynchronous update of genome data, and mismatched metadata (Table 2, Supplementary Data 1).

4.1. Low-complexity isolate name

A significant proportion of isolate names exhibit low character diversity or are short in length, which poses challenges in ensuring their uniqueness. According to our statistics, at least 390,101 genome sequences use exclusively numerical characters as isolate names, with at least 82,176 isolate names being fewer than five characters in length. For instance, 7 genomes (from Iceland, Finland, Croatia, Norway, and Singapore) use "3000" as their isolate names. Furthermore, at least 38,005 genomes use exclusively alphabetical characters as isolate names, such as the isolate name "JFEG" in virus name "SARS-CoV-2/human/USA/JFEG/2020".

Table 1
Resources for SARS-CoV-2 genome data deposition and integration.

	URL	Data type	Data source	Virus detail	Sample	Patient	Sequencing and assembly	Submission	Sequence quality	Completeness	Related links	De-redundancy	Lineage	Variant annotation	Update	Update frequency	Data accessibility
GenBank	https://www.ncbi.nlm.nih.gov/genbank/	Genome, other nucleotide, protein	Direct submission, INSDC datasets	Name, Accession ID	Collection date, Location, Host, Isolation source	–	Sequencing technology, Assembly method	Submitter, Date, Lab	–	Complete / partial	Raw data, Publications	–	Pango lineages	–	Version	Real time	Unrestricted
ENA	https://www.ebi.ac.uk/ena/	Genome, other nucleotide	Direct submission, INSDC datasets	Name, Accession ID	Collection date, Location, Host, Isolation source	–	–	Submitter, Date, Lab	–	–	Raw data, Publications	–	–	–	Version	Real time	Unrestricted
DDBJ	https://www.ddbj.nig.ac.jp	Genome, other nucleotide	Direct submission, INSDC datasets	Name, Accession ID	Collection date, Location, Host, Isolation source	–	–	Submitter, Date, Lab	–	–	Raw data, Publications	–	–	–	Version	Real time	Unrestricted
GenBase	https://ngdc.cncb.ac.cn/genbase/	Genome, other nucleotide, protein	Direct submission	Name, Accession ID, Passage details/ history	Collection date, Location, Host, Sampling strategy, Isolation source	Sex, Age, Status	Sequencing technology, Assembly method	Submitter, Date, Lab	-	Complete / partial	Raw data, Publications	–	–	–	Version	Real time	Unrestricted
EpiCoV™	https://www.epicov.org/epi3/frontend#3543d8	Genome	Direct submission	Name, Accession ID, Passage details/ history	Collection date, Location, Host, Outbreak detail, Sampling strategy, Specimen source, Lab	Gender, Age, Status, Specimen source, Last vaccinated, Treatment	Sequencing technology, Assembly method, Coverage	Submitter, Date, Lab	N, Gap, Coverage	–	–	–	Pango lineages, WHO label	Yes	–	Real time	Restricted
CNGBdb	https://db.cngb.org/virusdip/ncov	Genome, other nucleotide, protein	Direct submission	Name, Accession ID	Collection date, Location, Host, Lab	–	Sequencing technology/platform, Assembly method	Submitter, Date, Lab	–	–	–	–	–	–	–	Real time	Unrestricted
NCNSTRSS	https://nmdc.cn/nCoV	Genome, other nucleotide, protein	Direct submission, GenBank	Name, Accession ID	Collection date, Location, Host, Isolation source	–	Sequencing technology	Submitter, Lab	–	–	–	–	–	–	–	Real time	Unrestricted
CoV-Seq	https://covseq.baidu.com/	Genomes	GenBank, EpiCoV™, ENA, CNGBdb	Name, Accession ID	Collection date, Location	–	–	–	–	–	–	Identify redundancies	–	Yes	–	Last update: 2020–09-28	Unrestricted
VirusDIP	https://db.cngb.org/virusdip/	Genomes, proteins	GenBank, EpiCoV™, CNGBdb	Name, Accession ID	Collection date, Location, Host, Lab	–	Sequencing technology,	Submitter, Date, Lab	–	–	Publication	Identify redundancies	–	Yes	–	Daily	Unrestricted
ViruSurf	https://gmql.eu/virusurf/, https://cerilab.deib.polimi.it/virusurf_gisaid/	Genomes	GenBank, EpiCoV™, COG-UK, NCNSTRSS	Name, Accession ID	Collection date, Location, Host, Specimen source	Gender	Sequencing technology, Assembly method, Coverage	Date	GC%, N%	Complete / partial	–	Identify redundancies	Pango lineage	Yes	–	Last update: 2022–01-13	Unrestricted
RCoV19	https://ngdc.cncb.ac.cn/ncov/	Genomes	GenBank, EpiCoV™, GenBase, CNGBdb, NCNSTRSS	Name, Accession ID	Collection date, Location, Host, Lab	Gender, Age	–	Date, Lab	Sequence quality control and quality Assessment	Complete / partial	Raw data	Identify redundancies, and provide cross-references	Pango lineage, WHO label	Yes	–	Daily	Unrestricted

Table 2
Issues on SARS-CoV-2 genome data integration.

Type	Description	Count	Statistical time
Low-complexity isolate name	Low character diversity	428,106	2022–11-13
Inconsistent isolate name	Isolate name inconsistency across different sources	68,802	End of 2022
Isolate name absence	Absent of isolate name	3,176,537	2023–04-10
Sequence inconsistency	Varying lengths due to different processing	2,756,473	2023–04-07
Asynchronous update	Asynchronous update across data sources or between data sources and integration resources	640,092	End of 2022
Mismatched metadata	Mismatched names	404	2023–01-11

Note: These data were collected at different points in time and are evidently underestimated. Details for each type could be viewed in Supplementary Data 1.

4.2. Inconsistent isolate name

Inconsistency of isolate naming across different sources is a prevalent issue. We discovered at least 68,802 pairs of genomes show different isolate names between GenBank and EpiCoV™, despite being identical genomes sharing the same meta information such as collection dates, locations, and hosts. Examples of such inconsistencies include MZ472651.1 (AZ-1773) and EPI_ISL_2716369 (AZ-ASPHL-1773), OK549954.1 (CA-SEARCH-104691) and EPI_ISL_3373690 (CA-ALSR-104691), and MZ637560.1 (NY-PRL-2021_0205_08D12) and EPI_ISL_1041222 (NY-PRL-2021_02_05_08D12), representing the major categories of “keyword missing”, “keyword variation”, and “format variation”.

4.3. Isolate name absence

At least 3,176,537 genomes in GenBank lack isolate names, primarily exchanged from ENA and originated from the UK. We sourced their sample names from BioSample and cross-referenced them with EpiCoV™ based on isolate name, collection date, country, and sequence, ultimately discovering at least 1,882,388 pairs of identical genomes between GenBank and EpiCoV™. For instance, GenBank sequence OX243287.1, named “COG-UK/QEUH-3ED15C6” in BioSample, corresponds to EpiCoV™ sequence EPI_ISL_13875089, with a virus name of “hCoV-19/England/QEUH-3ED15C6/2022”.

Unfortunately, we were unable to find matching EpiCoV™ sequences for 450,491 GenBank sequences from Germany by comparing BioSample sample name with EpiCoV™ virus name. This is because the sample name listed in BioSample don’t correspond with the virus name present in EpiCoV™, but instead reflect the Sample ID allocated by the originating laboratory. As an example, GenBank sequence OU366026.1, with a sample name of “IMS-10209-CVDP-2FB19292-0D4A-486F-AAA4-2393066873F0” in BioSample, corresponds to EpiCoV™ sequence EPI_ISL_1153242. However, the virus name is actually “hCoV-19/Germany/BW-RKI-I-023076/2021”, and the Sample ID is the same with the sample name.

4.4. Sequence inconsistency

The same genome sequences submitted to GenBank and EpiCoV™ sometimes display differing lengths, primarily due to inconsistent processing of Ns. For example, the GenBank sequence OW028255.1 (29885 nt) and EpiCoV™ sequence EPI_ISL_10485469 (28771 nt) are from the same sample and are identical after all Ns are removed. We have identified at least 2,756,473 pairs of sequences exhibiting varying lengths.

4.5. Asynchronous update

Asynchronous update of sequences and metadata across different deposition repositories, or between such deposition repositories and integration resources, pose a significant hindrance to data integration. To illustrate this, the collection date for EPI_ISL_13344383 was changed from 2022 to 06-06 to 2022–06-03. However, as of June 19, 2023, the corresponding GenBank sequence ON797504.1 still had a collection date of 2022-06-06.

It is worth noting that genome metadata and sequences in both GenBank and EpiCoV™ are subject to frequent updates, making it difficult for integration resources to achieve synchronicity, except in the case of mirror database. We have detected at least 10,728 sequences that were removed from the data source, while updates have been made to 88,345 sequences in the original databases with respect to virus name, 405,261 sequences to sequence length, 45,479 sequences to collection date, 1,380 sequences to host, and 88,899 sequences to collection location.

4.6. Mismatched metadata

Last but not least, some sequences are associated with mismatched virus names and sample names. For instance, the isolate name of OU058581.1 is “hCoV-19/Switzerland/ZH-UZH-IMV-3ba472d2/2021”, whereas the corresponding BioSample sample name is “hCoV-19/Switzerland/ZH-UZH-IMV-3ba47365/2021”, which correspond to two different EpiCoV™ sequences: EPI_ISL_2102278 and EPI_ISL_2102222.

5. Perspectives

Current resources and initiatives have been instrumental in promoting and facilitating the sharing of SARS-CoV-2 genome data on a global scale. Nonetheless, the vast amount of SARS-CoV-2 genome data are scattered across different repositories without centralized management, hindering comprehensive analysis that could lead to a better understanding of virus transmission and the development of clinical and epidemiological mitigation strategies. However, the deposition repositories have different standardization and curation criteria, and data integration is significantly impeded by low-complexity, inconsistency, and absence of isolate name, sequence inconsistency, asynchronous update, and metadata mismatch.

To increase the precision and potential benefits of genome data in the public health response, deposition repositories ought to collect as much comprehensive metadata as possible and establish a universe consensus regarding metadata formats (including isolate name, accession.version format ID, sample information, patient information, sequencing and assembly information, and submission information). As an example, the Public Health Alliance for Genomic Epidemiology (PH4GE, https://pha4ge.org/) has developed a contextual data speci-

fication for SARS-CoV-2 based on INSDC pathogen package to enhance consistency and reusability of collected data [31]. Additionally, data submitters must ensure consistency in sequences and metadata submitted to distinct resources and make sure that the data are updated synchronously across resources. It would be preferable if the submitters could specify the corresponding accession IDs when resubmitting the data to different resources. Meanwhile, deposition repositories ought to designate a specific field for archiving the "Related ID" so that cross-referencing can be established for easy tracking. Furthermore, deposition repositories should introduce suitable frameworks for indicating metadata updates, which effectively communicate any significant changes to users (such as collection date). For example, GenBank offers "Revision History" that facilitates tracking metadata updates and enables comparisons between different revisions. With these endeavors and enhancements, compiling the scattered data into larger data sets would be considerably simpler and thus facilitates a range of value-added curations and analysis.

Acknowledgements

This work was supported by Strategic Priority Research Program of the Chinese Academy of Sciences [XDB38030201, XDB38030400, XDB38050300]; Youth Innovation Promotion Association of Chinese Academy of Sciences [2019104]. Genome sequence data were obtained from CNCB, CNGBdb, GenBank, GISAID, and NMDC resources. We acknowledge all sample providers and data submitters.

Conflict of interest statement

The authors declare that there are no conflicts of interest.

Author contributions

Lina Ma: Supervision, Writing – original draft, Writing – review & editing, Data curation. **Wei Zhao:** Writing – original draft, Data curation. **Tianhao Huang:** Writing – original draft, Data analysis. **Enhui Jin:** Writing – original draft, Data curation. **Gangao Wu:** Writing – original draft. **Wenming Zhao:** Writing – review & editing. **Yiming Bao:** Supervision, Writing – review & editing.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bsheal.2023.07.004.

References

[1] L.J. Carter, L.V. Garner, J.W. Smoot, Y. Li, Q. Zhou, C.J. Saveson, J.M. Sasso, A.C. Gregg, D.J. Soares, T.R. Beskid, S.R. Jervey, C. Liu, Assay techniques and test development for COVID-19 diagnosis, ACS Cent. Sci. 6 (5) (2020) 591–605, https://doi.org/10.1021/acscentsci.0c00501.

[2] V.M. Corman, O. Landt, M. Kaiser, R. Molenkamp, A. Meijer, D.K. Chu, T. Bleicker, S. Brunink, J. Schneider, M.L. Schmidt, et al., Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR, Euro Surveill. 25 (3) (2020) 2000045, https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045.

[3] T. Thanh Le, Z. Andreadakis, A. Kumar, R. Gomez Roman, S. Tollefsen, M. Saville, S. Mayhew, The COVID-19 vaccine development landscape, Nat. Rev. Drug Discov. 19 (5) (2020) 305–306, https://doi.org/10.1038/d41573-020-00073-5.

[4] H. Wang, Y. Zhang, B. Huang, W. Deng, Y. Quan, W. Wang, W. Xu, Y. Zhao, N. Li, J. Zhang, H. Liang, L. Bao, et al., Development of an inactivated vaccine candidate, BBIBP-CorV, with potent protection against SARS-CoV-2, Cell 182 (3) (2020) 713–721.e9, https://doi.org/10.1016/j.cell.2020.06.008.

[5] K.G. Andersen, A. Rambaut, W.I. Lipkin, E.C. Holmes, R.F. Garry, The proximal origin of SARS-CoV-2, Nat. Med. 26 (4) (2020) 450–452, https://doi.org/10.1038/s41591-020-0820-9.

[6] X. Pang, L. Ren, S. Wu, W. Ma, J. Yang, L. Di, J. Li, Y. Xiao, L. Kang, S. Du, et al., Cold-chain food contamination as the possible origin of COVID-19 resurgence in Beijing, Natl. Sci. Rev. 7 (12) (2020) 1861–1864, https://doi.org/10.1093/nsr/nwaa264.

[7] S. Song, C. Li, L. Kang, D. Tian, N. Badar, W. Ma, S. Zhao, X. Jiang, C. Wang, Y. Sun, et al., Genomic epidemiology of SARS-CoV-2 in Pakistan, Genom. Proteom. Bioinform. 19 (5) (2021) 727–740, https://doi.org/10.1016/j.gpb.2021.08.007.

[8] A. Baum, C.A. Kyratsous, SARS-CoV-2 spike therapeutic antibodies in the age of variants, J. Exp. Med. 218 (5) (2021) e20210198, https://doi.org/10.1084/jem.20210198.

[9] P. Qu, J.P. Evans, J.N. Faraone, Y.M. Zheng, C. Carlin, M. Anghelina, P. Stevens, S. Fernandez, D. Jones, G. Lozanski, et al., Enhanced neutralization resistance of SARS-CoV-2 Omicron subvariants BQ.1, BQ.1.1, BA.4.6, BF.7, and BA.2.75.2, Cell Host Microbe 31 (1) (2023) 9–17, https://doi.org/10.1016/j.chom.2022.11.012.e3.

[10] A. Singh, V. Gupta, SARS-CoV-2 therapeutics: how far do we stand from a remedy?, Pharmacol Rep. 73 (3) (2021) 750–768, https://doi.org/10.1007/s43440-020-00204-0.

[11] W.M. Zhao, S.H. Song, M.L. Chen, D. Zou, L.N. Ma, Y.K. Ma, R.J. Li, L.L. Hao, C.P. Li, D.M. Tian, B.X. Tang, et al., The 2019 novel coronavirus resource, Yi Chuan 42 (2) (2020) 212–221, https://doi.org/10.16288/j.yczz.20-030.

[12] B.A. Underwood, L. Yankie, E.P. Nawrocki, V. Palanigobu, S. Gotvyanskyy, V.C. Calhoun, M. Kornbluh, T.G. Smith, L. Fleischmann, D. Sinyakov, et al., Rapid automated validation, annotation and publication of SARS-CoV-2 sequences to GenBank, Database (Oxford) (2022) baac006, https://doi.org/10.1093/database/baac006.

[13] M. Arita, I. Karsch-Mizrachi, G. Cochrane, The international nucleotide sequence database collaboration, Nucleic Acids Res. 49 (D1) (2021) D121–D124, https://doi.org/10.1093/nar/gkaa967.

[14] S. Khare, C. Gurry, L. Freitas, M.B. Schultz, G. Bach, A. Diallo, N. Akite, J. Ho, R.T. Lee, W. Yeo, G.C. Curation Team, S. Maurer-Stroh, GISAID's Role in Pandemic Response, China CDC Wkly 3 (49) (2021) 1049–1051, https://doi.org/10.46234/ccdcw2021.255.

[15] A. Canakoglu, P. Pinoli, A. Bernasconi, T. Alfonsi, D.P. Melidis, S. Ceri, ViruSurf: an integrated database to investigate viral sequences, Nucleic Acids Res. 49 (D1) (2021) D817–D824, https://doi.org/10.1093/nar/gkaa846.

[16] P.A. Desingu, K. Nagarajan, The emergence of omicron lineages BA.4 and BA.5, and the global spreading trend, J. Med. Virol. 94 (11) (2022) 5077–5079, https://doi.org/10.1002/jmv.27967.

[17] R. Uraki, M. Ito, Y. Furusawa, S. Yamayoshi, K. Iwatsuki-Horimoto, E. Adachi, M. Saito, M. Koga, T. Tsutsumi, S. Yamamoto, et al., Humoral immune evasion of the omicron subvariants BQ.1.1 and XBB, Lancet Infect. Dis. 23 (1) (2023) 30–32, https://doi.org/10.1016/S1473-3099(22)00816-7.

[18] J. Burgin, A. Ahamed, C. Cummins, R. Devraj, K. Gueye, D. Gupta, V. Gupta, M. Haseeb, M. Ihsan, E. Ivanov, et al., The European Nucleotide Archive in 2022, Nucleic Acids Res. 51 (D1) (2023) D121–D125, https://doi.org/10.1093/nar/gkac1051.

[19] T. Okido, Y. Kodama, J. Mashima, T. Kosuge, T. Fujisawa, O. Ogasawara, DNA Data Bank of Japan (DDBJ) update report 2021, Nucleic Acids Res. 50 (D1) (2022) D102–D105, https://doi.org/10.1093/nar/gkab995.

[20] CNCB-NGDC Members and Partners, Database resources of the national genomics data center, china national center for bioinformation in 2023, Nucleic Acids Res. 51 (D1) (2023) D18–D28, https://doi.org/10.1093/nar/gkac1073.

[21] F.Z. Chen, L.J. You, F. Yang, L.N. Wang, X.Q. Guo, F. Gao, C. Hua, C. Tan, L. Fang, R.Q. Shan, W.J. Zeng, et al., CNGBdb: China National GeneBank DataBase, Yi Chuan 42 (8) (2020) 799–809, https://doi.org/10.16288/j.yczz.20-080.

[22] E.W. Sayers, M. Cavanaugh, K. Clark, J. Ostell, K.D. Pruitt, I. Karsch-Mizrachi, GenBank, Nucleic Acids Res. 47 (D1) (2019) D94–D99, https://doi.org/10.1093/nar/gky989.

[23] E.L. Hatcher, S.A. Zhdanov, Y. Bao, O. Blinkova, E.P. Nawrocki, Y. Ostapchuck, A. A. Schaffer, J.R. Brister, Virus Variation Resource - improved response to emergent viral outbreaks, Nucleic Acids Res. 45 (D1) (2017) D482–D490, https://doi.org/10.1093/nar/gkw1065.

[24] M. Chen, Y. Ma, S. Wu, X. Zheng, H. Kang, J. Sang, X. Xu, L. Hao, Z. Li, Z. Gong, et al., Genome Warehouse: a public repository housing genome-scale data, Genom. Proteom. Bioinform. 19 (4) (2021) 584–589, https://doi.org/10.1016/j.gpb.2021.04.001.

[25] Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2, Nat. Microbiol 5 (4) (2020) 536–544, https://doi.org/10.1038/s41564-020-0695-z.

[26] L. Wang, F. Chen, X. Guo, L. You, X. Yang, F. Yang, T. Yang, F. Gao, C. Hua, Y. Ding, J. Cai, L. Yang, W. Huang, et al., VirusDIP: virus data integration platform [Preprint], bioRxiv (2020), https://doi.org/10.1101/2020.06.08.139451.

[27] B. Liu, K. Liu, H. Zhang, L. Zhang, Y. Bian, L. Huang, CoV-Seq, a new tool for SARS-CoV-2 genome analysis and visualization: development and usability study, J. Med. Internet Res. 22 (10) (2020) e22299, https://doi.org/10.2196/22299.

[28] D. Yu, X. Yang, B. Tang, Y.H. Pan, J. Yang, G. Duan, J. Zhu, Z.Q. Hao, H. Mu, L. Dai, et al., Coronavirus GenBrowser for monitoring the transmission and evolution of SARS-CoV-2, Brief. Bioinform. 23 (2) (2022) bbab583, https://doi.org/10.1093/bib/bbab583.

[29] S. Marjanovic, R.J. Romanelli, G.C. Ali, B. Leach, M. Bonsu, D. Rodriguez-Rincon, T. Ling, COVID-19 Genomics UK (COG-UK) consortium: final report, Rand Health Q 9 (4) (2022) 24.

[30] Z. Gong, J.W. Zhu, C.P. Li, S. Jiang, L.N. Ma, B.X. Tang, D. Zou, M.L. Chen, Y.B. Sun, S.H. Song, Z. Zhang, et al., An online coronavirus analysis platform from the National Genomics Data Center, Zool. Res. 41 (6) (2020) 705–708, https://doi.org/10.24272/j.issn.2095-8137.2020.065.

[31] E.J. Griffiths, R.E. Timme, C.I. Mendes, A.J. Page, N.F. Alikhan, D. Fornika, F. Maguire, J. Campos, D. Park, I.B. Olawoye, et al., Future-proofing and maximizing the utility of metadata: The PHA4GE SARS-CoV-2 contextual data specification package, GigaScience 11 (2022) giac003, https://doi.org/10.1093/gigascience/giac003.