Published online 1 December 2021 *Nucleic Acids Research, 2022, Vol. 50, Database issue* ***D161–D164***
https://doi.org/10.1093/nar/gkab1135

GenBank

Eric W. Sayers*, Mark Cavanaugh, Karen Clark, Kim D. Pruitt, Conrad L. Schoch, Stephen T. Sherry and Ilene Karsch-Mizrachi

National Center for Biotechnology Information, National Library of Medicine , National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received September 18, 2021; Revised October 20, 2021; Editorial Decision October 21, 2021; Accepted November 08, 2021

ABSTRACT

GenBank® (https://www.ncbi.nlm.nih.gov/genbank/) is a comprehensive, public database that contains 15.3 trillion base pairs from over 2.5 billion nucleotide sequences for 504 000 formally described species. Recent updates include resources for data from the SARS-CoV-2 virus, including a SARS-CoV-2 landing page, NCBI Datasets, NCBI Virus and the Submission Portal. We also discuss upcoming changes to GI identifiers, a new data management interface for BioProject, and advice for providing contextual metadata in submissions.

INTRODUCTION

GenBank (1) is a comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotations built and distributed by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM), located on the campus of the US National Institutes of Health (NIH) in Bethesda, MD, USA. After discussing updates to SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) resources, this paper summarizes the growth of GenBank in the past year and briefly reviews recent updates and developments.

SARS-COV-2 RESOURCES

As part of our ongoing response to the COVID-19 pandemic that emerged in early 2020, NCBI continues to update several tools and interfaces to support both submitters and consumers of sequence data for SARS-CoV-2. These include the SARS-CoV-2 landing page, NCBI Datasets, NCBI Virus, and the Submission Portal.

SARS-CoV-2 landing page

The SARS-CoV-2 landing page (https://www.ncbi.nlm.nih.gov/sars-cov-2/) collects a wide variety of data and resources related to SARS-CoV-2, including all relevant data in GenBank. Of particular interest to users seeking GenBank data are links to NCBI Datasets and NCBI Virus (see below) along with a link to download the full list of nucleotide accessions for SARS-CoV-2.

NCBI Datasets

NCBI Datasets is an experimental product that allows users to download complex genomic datasets easily using either a web interface, an API or a UNIX/LINUX command-line tool (https://www.ncbi.nlm.nih.gov/datasets/). The specialized coronavirus page released last year now provides genome downloads for almost 430 000 complete SARS-CoV-2 genomes, an annual increase of 29-fold (https://www.ncbi.nlm.nih.gov/datasets/coronavirus/genomes). This page provides downloads of metadata tables for SARS-CoV-2 genomes as well as complete genomic datasets. Users interested in SARS-CoV-2 proteins can access these data on a separate specialized page (https://www.ncbi.nlm.nih.gov/datasets/coronavirus/proteins/). Finally, NCBI Datasets also includes a new genome interface that supports taxonomic searches and selection based on the taxonomic tree. This page may be of interest to users seeking data for other coronaviruses not included on the specialized SARS-CoV-2 pages.

NCBI Virus

The NCBI Virus resource contains an SARS-CoV-2 Hub (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/sars-cov-2) that organizes an extensive set of data and visualizations about SARS-CoV-2 data (Figure 1), including data from the Sequence Read Archive (SRA). The visualizations on the default "Dashboard" view include a world map showing the geographical distribution of SARS-CoV-2 collection locations. Two interactive filters allow users to

*To whom correspondence should be addressed. Tel: +1 301 496 2475; Fax: +1 301 480 9241; Email: sayers@ncbi.nlm.nih.gov
Electronic addresses:
www.ncbi.nlm.nih.gov - NCBI Home Page. gb-sub@ncbi.nlm.nih.gov - Submission of sequence data to GenBank.
update@ncbi.nlm.nih.gov - Revisions to, or notification of release of, 'confidential' GenBank entries.
info@ncbi.nlm.nih.gov - General information about NCBI resources.

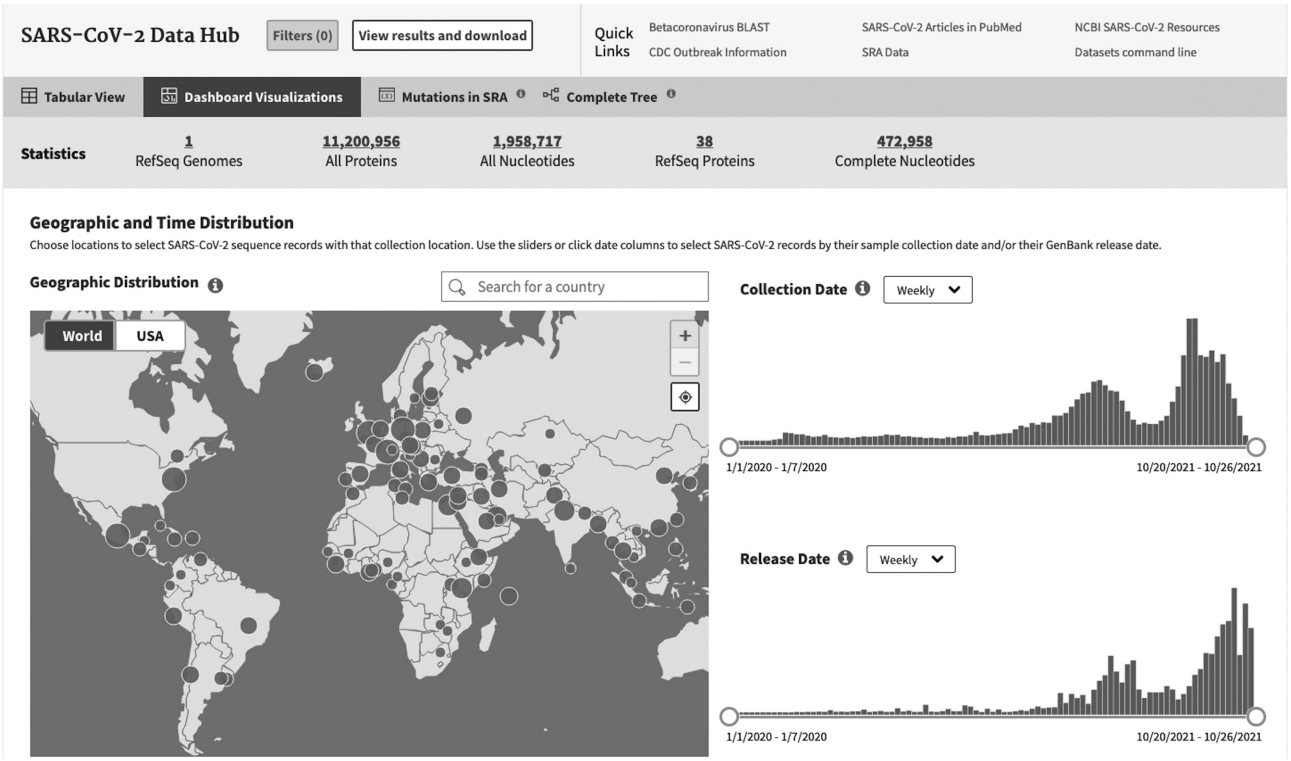

Figure 1. SARS-CoV-2 Data Hub in the NCBI Virus resource.

subset these data by collection date and release date, and these filters update the map display. A 'Tabular View' option loads an interactive table listing all SARS-CoV-2 sequences with 21 filters, including sequence length, collection date, and geographic region. Conveniently, any filters set on the Dashboard transfer to the table, allowing easy exploration of the data. Moreover, users can spawn alignments from this table, and can also build phylogenetic trees.

Submission Portal

NCBI continues to update a customized submission portal for both assembled and unassembled SARS-CoV-2 sequences (https://submit.ncbi.nlm.nih.gov/sarscov2/). On average this portal provides accessions back to submitters in 1–2 h, and assembled sequences will be annotated with VADR (2). We encourage submitters to use these portals, as this ensures not only that sequence data are made available through the INSDC databases, but also through the NCBI Virus resource (3), RefSeq (4), and BLAST (5). We also encourage submitters to submit both reads and traditional GenBank sequences, and to submit data to BioProject and BioSample. We are actively updating all of these resources to support novel variants and to update the content of the SARS-CoV-2 pages discussed above.

GROWTH OF THE DATABASE

Divisions with notable increases

GenBank sequences are organized into 21 divisions, each of which is represented by a three-letter abbreviation (Table 1).

Table 1. Growth of GenBank Divisions

Division	Description	Base pairs[a]	Annual increase[b]
VRL	Viruses	39 351 597 469	575.68%
UNA	Unannotated	4 421 782	550.93%
INV	Invertebrates	108 680 334 593	450.00%
ROD	Rodents	23 336 550 435	93.02%
PRI	Primates	15 165 437 356	72.97%
WGS	Whole genome shotgun data	13 888 187 863 722	57.08%
TLS	Targeted Loci Studies	39 930 167 315	43.50%
MAM	Other mammals	28 568 850 588	37.06%
VRT	Other vertebrates	85 320 979 451	34.22%
BCT	Bacteria	130 518 385 589	32.07%
PLN	Plants	350 590 744 188	30.12%
TSA	Transcriptome shotgun data	454 757 992 932	19.31%
PHG	Phages	935 884 237	19.59%
PAT	Patent sequences	29 588 418 021	11.85%
ENV	Environmental samples	7 394 414 660	9.46%
SYN	Synthetic	7 994 601 379	0.78%
HTC	High-throughput cDNA	737 423 641	0.57%
HTG	High-throughput genomic	27 800 219 072	0.07%
EST	Expressed sequence tags	43 324 455 796	0.05%
GSS	Genome survey sequences	26 380 049 011	0.01%
STS	Sequence tagged sites	640 923 137	0.00%
TOTAL	All GenBank sequences	15 309 209 714 374	54.79%

[a]Release 245 (8/2021).
[b]Relative to release 239 (8/2020).

As shown in Table 1, especially large increases occurred in the VRL, UNA and INV divisions. Not surprisingly, the large increase in the VRL division resulted from the many submissions of SARS-CoV-2 sequences (Figure 2).

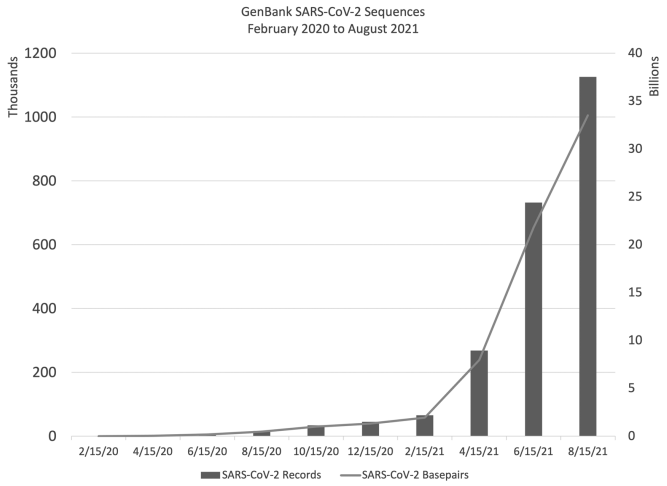

Figure 2. Growth of SARS-CoV-2 sequence data in GenBank. Each data point represents the cumulative number of records (left axis) or base pairs (right axis) at each date.

Handling long sequence records

As previously discussed (1), improving sequencing technologies are now capable of producing very long sequences, some of which are longer than what signed 32-bit integers can represent (about 2.1 Gbp). In such cases, submitters must split such records in order to submit them to GenBank. A recent example of such a case is chromosome 1 from the West African lungfish, *Protopterus annectens*. The total length of this chromosome is 5.26 Gbp, and so in GenBank it is represented by three records: CM033073 (2.00 Gbp), CM033074 (2.00 Gbp) and CM033075 (1.26 Gbp). We encourage GenBank users and developers of products that rely on GenBank data to be aware of the implications involved in representing very long sequences and to consider preparing their own tools for sequence lengths and feature locations that will require 64-bit integers.

RECENT DEVELOPMENTS

Updates to integer sequence identifiers

In addition to the above issue of handling very long individual sequences, GenBank is approaching a point where the number of sequences will exhaust the space of GI identifiers provided by 32-bit integers. To mitigate this, we are taking multiple approaches. First, we continue to recommend that users shift to using accession.version identifiers to refer to all GenBank data (6). Most external NCBI interfaces, including the Entrez web interface and the E-utilities API, now accept and return accession.version identifiers for all sequences. Second, we are transitioning our internal software to use 64-bit integers for GI identifiers. Once the transition occurs, GenBank users will encounter these identifiers in the XML and ASN.1 presentations of GenBank data provided through the Entrez web interface and in GenBank FTP products (https://ncbiinsights.ncbi.nlm.nih.gov/2021/09/02/64-bit-gis/). We encourage developers who rely on GenBank data to ensure that their software is capable of handling these 64-bit identifiers. Such identifiers are easy to recognize, as they are any integer greater than 2 147 483 647.

BioProject data management

When submitters register sequencing projects in the BioProject database (https://www.ncbi.nlm.nih.gov/bioproject), we can create reliable linkages between such sequencing projects and the data they produce, and in many cases to the BioSample database (7) that provides additional information about the biological materials used in the study. In many cases, submitters create BioProject records before they have collected all relevant data and published the results of the study. We have now made it easier for submitters to update their BioProject records with such information by offering a 'Manage Data' interface in the Submission portal (https://dataview.ncbi.nlm.nih.gov/?archive=bioproject). Using this interface, submitters can add publications and grants or edit text metadata such as the BioProject title and description. We hope this will allow BioProject to reflect better the current state of these projects and provide a better service to the community.

Advice for submitters

Contextual metadata. As discussed previously (1), we continue to encourage submitters to provide contextual metadata, particularly data that specifies the sampling location (e.g. country, latitude, and longitude). The importance of such basic geographic information, such as country codes displayed on public sequence records (https://insdc.org/country), has only grown with the urgency to verify and track distribution of biodiversity in the current era. Including other data such as the isolate name or number and applicable museum/collection identifiers is also helpful. Where possible, adding links to permanent samples or vouchers at biorepositories provides access to sources with important, richly populated information. This facilitates replication and validation, while also allowing for analyses across scientific disciplines (8). GenBank has long followed the standards of structuring vouchers using Darwin Core formats (9) that allows us to link to specimen pages at external biorepositories using URLs curated in the NCBI BioCollections database (10). Recently, BioCollections included a new category, 'digital repository', that will include online data aggregators of collection data that do not include physical specimens. Additionally, to prepare the way for a more comprehensive treatment of these data elements and to make their presence required as part of the submission process, GenBank and the INSDC developed a set of standardized terms to clearly indicate when submitters cannot provide voucher information, for example in cases where the data were not collected or cannot be reported because of privacy concerns (https://www.insdc.org/missing-value-reporting).

In addition to the above, there are other ways in which submitters can enhance their data. Submitters can use evidence tags to provide information about supporting evidence for annotations (https://www.ncbi.nlm.nih.gov/genbank/evidence/). They can cite within their submission the accession numbers of any publicly available sequencing reads they used to improve the quality of their

assemblies. When submitting prokaryotic genomes, they can create annotated genomes with NCBI's Prokaryotic Genome Annotation Pipeline (PGAP; https://www.ncbi.nlm.nih.gov/genome/annotation_prok/) either by submitting FASTA files and requesting PGAP during submission of the genomes to GenBank or by running the public version of PGAP themselves and then submitting the GenBank-ready ASN.1 output file.

Acquiring the database

NCBI provides GenBank sequence records in both the traditional flat file format and in a structured ASN.1 format by anonymous FTP at ftp.ncbi.nlm.nih.gov/genbank. For release 245 (15 August 2021) there are 4032 files requiring 1888 GB of uncompressed disk storage. In addition, daily GenBank incremental update files containing new records and those updated since the most recent release are available in flat file format at ftp.ncbi.nlm.nih.gov/genbank/daily-nc/.

CITING GENBANK

If you use the GenBank database in your published research, we ask that this article be cited.

FUNDING

Funding for open access charge: Intramural Research Program of the National Library of Medicine, National Institutes of Health.
Conflict of interest statement. None declared.

REFERENCES

1. Sayers,E.W., Cavanaugh,M., Clark,K., Pruitt,K.D., Schoch,C.L., Sherry,S.T. and Karsch-Mizrachi,I. (2021) GenBank. *Nucleic Acids Res.*, **49**, D92–D96.
2. Schaffer,A.A., Hatcher,E.L., Yankie,L., Shonkwiler,L., Brister,J.R., Karsch-Mizrachi,I. and Nawrocki,E.P. (2020) VADR: validation and annotation of virus sequence submissions to GenBank. *BMC Bioinformatics*, **21**, 211.
3. Brister,J.R., Ako-Adjei,D., Bao,Y. and Blinkova,O. (2015) NCBI viral genomes resource. *Nucleic Acids Res.*, **43**, D571–D577.
4. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
5. Boratyn,G.M., Camacho,C., Cooper,P.S., Coulouris,G., Fong,A., Ma,N., Madden,T.L., Matten,W.T., McGinnis,S.D., Merezhuk,Y. *et al.* (2013) BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.*, **41**, W29–W33.
6. Sayers,E.W., Cavanaugh,M., Clark,K., Ostell,J., Pruitt,K.D. and Karsch-Mizrachi,I. (2020) GenBank. *Nucleic Acids Res.*, **48**, D84–D86.
7. Barrett,T., Clark,K., Gevorgyan,R., Gorelenkov,V., Gribov,E., Karsch-Mizrachi,I., Kimelman,M., Pruitt,K.D., Resenchuk,S., Tatusova,T. *et al.* (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.*, **40**, D57–D63.
8. Cook,J.A., Arai,S., Armien,B., Bates,J., Bonilla,C.A.C., Cortez,M.B.S., Dunnum,J.L., Ferguson,A.W., Johnson,K.M., Khan,F.A.A. *et al.* (2020) Integrating biodiversity infrastructure into pathogen discovery and mitigation of emerging infectious diseases. *Bioscience*, **70**, 531–534.
9. Wieczorek,J., Bloom,D., Guralnick,R., Blum,S., Doring,M., Giovanni,R., Robertson,T. and Vieglais,D. (2012) Darwin Core: an evolving community-developed biodiversity data standard. *PLoS One*, **7**, e29715.
10. Sharma,S., Ciufo,S., Starchenko,E., Darji,D., Chlumsky,L., Karsch-Mizrachi,I. and Schoch,C.L. (2018) The NCBI BioCollections Database. *Database (Oxford)*, **2018**, bay006.