D84–D86 *Nucleic Acids Research, 2020, Vol. 48, Database issue* *Published online 29 October 2019*
doi: 10.1093/nar/gkz956

GenBank

Eric W. Sayers*, Mark Cavanaugh, Karen Clark, James Ostell, Kim D. Pruitt 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 17, 2019; Revised October 08, 2019; Editorial Decision October 09, 2019; Accepted October 09, 2019

ABSTRACT

GenBank® (www.ncbi.nlm.nih.gov/genbank/) is a comprehensive, public database that contains over 6.25 trillion base pairs from over 1.6 billion nucleotide sequences for 450 000 formally described species. Daily data exchange with the European Nucleotide Archive (ENA) and the DNA Data Bank of Japan (DDBJ) ensures worldwide coverage. Recent updates include a new version of Genome Workbench that supports GenBank submissions, new submission wizards for viral genomes, enhancements to BankIt and improved handling of taxonomy for sequences from pathogens.

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 summarizing the growth of GenBank in the past year, this paper will briefly review recent updates and developments.

GROWTH OF THE DATABASE

The size and growth of the various divisions of GenBank are shown in Table 1 and Figure 1. Notable increases in the past year include the submission of 57 synthetic chromosomal constructs in January 2019 to the SYN division (ftp.ncbi.nlm.nih.gov/genbank/release.notes/gb230.release.notes) and the submission of about 60 chromosome-scale eukaryotic sequences to the VRT division as part of Release 231 (ftp.ncbi.nlm.nih.gov/genbank/release.notes/gb231.release.notes). 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 233 there are 2467 files requiring 1057 GB of uncompressed disk storage. In addition, daily GenBank incremental update files containing new and updated records since the most recent release are available in flat file format at ftp.ncbi.nlm.nih.gov/genbank/daily-nc/.

RECENT DEVELOPMENTS

Genome submissions

Advice for submitters. We would like to call attention to two cases where submitters can add additional value to their data. We encourage submitters of genomic sequences, including whole genome shotgun (WGS) sequences, to provide contextual metadata to support further use and analysis of the data. For example, where possible submitters should provide geographical data (e.g. country, latitude and longitude of the sampling location) along with other data such as the isolate name or number plus museum/collection identifiers as applicable. We also urge submitters to use evidence tags to provide information about supporting evidence for annotations of the form '/experimental = *text*' and '/inference = *TYPE*:*text*', where *TYPE* is a standard inference type and *text* consists of structured text, as explained at www.ncbi.nlm.nih.gov/genbank/evidence/. In cases where submitters have used existing public sequencing reads to improve the quality of their assemblies prior to submission, we encourage submitters to cite the accession numbers of these reads within their submission. Regarding prokaryotic genomes, while annotations are not required, we encourage submitters to request that the genome be annotated by the NCBI Prokaryotic Genome Annotation Pipeline (www.ncbi.nlm.nih.gov/genome/annotation_prok/) before being released.

NCBI strongly encourages submitters to register large-scale sequencing projects in the BioProject database (www.ncbi.nlm.nih.gov/bioproject) and to update their BioProject records after relevant publications are available. Doing so provides reliable linkages between sequencing projects and the data they produce, and may also allow links to the BioSample database (2) that provides additional information about the biological materials used in the study.

Taxonomy assignments. For submissions of bacterial genomes, GenBank performs an average nucleotide identity analysis (ANI) (3) to investigate whether the asserted

*To whom correspondence should be addressed. Tel: +1 301 496 2475; Fax: +1 301 480 9241; Email: sayers@ncbi.nlm.nih.gov

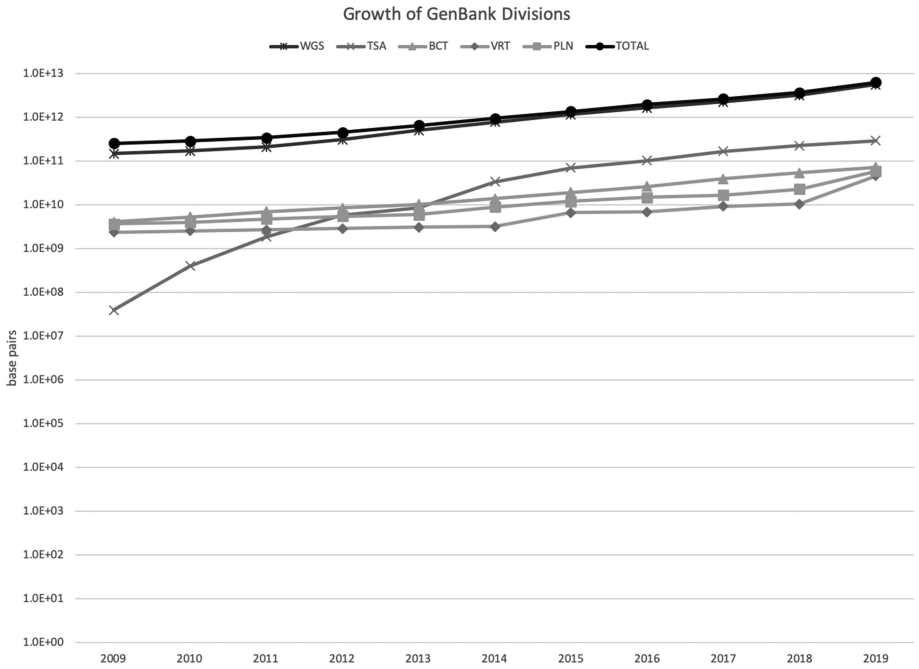

Figure 1. Growth of the five GenBank divisions that received the most sequence data in 2019, along with the growth of GenBank as a whole.

Table 1. Growth of GenBank Divisions (nucleotide base-pairs)

Division	Description	Release 233 (8/2019)	Annual Increase (%)[a]
SYN	Synthetic	7 701 613 755[b]	545.96%
VRT	Other vertebrates	46 205 911 214[b]	342.51%
PLN	Plants	59 248 524 178	157.29%
UNA	Unannotated	548 041	84.71%
WGS	Whole genome shotgun data	5 585 922 333 160	74.30%
TLS	Targeted locus studies	10 531 800 829	73.28%
INV	Invertebrates	12 578 394 104	46.31%
PHG	Phages	637 015 044	37.58%
BCT	Bacteria	72 495 994 966	35.40%
TSA	Transcriptome shotgun data	294 727 165 179	30.69%
VRL	Viruses	4 782 719 535	17.40%
PAT	Patent sequences	24 715 727 030	12.24%
ENV	Environmental samples	6 139 560 312	5.51%
PRI	Primates	8 491 950 612	2.78%
HTC	High-throughput cDNA	728 868 423	1.03%
MAM	Other mammals	6 258 926 080	0.71%
EST	Expressed sequence tags	43 280 039 563	0.68%
ROD	Rodents	4 554 525 905	0.43%
HTG	High-throughput genomic	27 774 725 922	0.01%
STS	Sequence tagged sites	640 918 572	0.01%
GSS	Genome survey sequences	26 339 260 641	0.00%
TOTAL	All GenBank sequences	6 233 224 722 236	69.52%

[a]Measured relative to Release 227 (8/2018).
[b]See the text for descriptions of these large increases.

organism name may be incorrect. GenBank also regularly scans existing data using this ANI analysis to identify other possibly erroneous taxonomic assignments. In cases where the taxonomic assignment is a generic species (e.g. *Genus sp.*), GenBank staff will update the organism name with the calculated binomial name. If the record was submitted with an incorrect binomial name (e.g. *Genus species*), GenBank staff will consult with the original submitters before updating the name.

Using Genome Workbench for preparing submissions

Genome Workbench 3.0 offers a new interface for creating a genome submission for GenBank. This new interface supports both prokaryotic and eukaryotic genomes and allows submitters to enter information directly into dialog boxes and then generate a finished submission file. Using this tool, submitters can also edit sequence features and validate the data against GenBank submission standards. More infor-

mation is available at www.ncbi.nlm.nih.gov/tools/gbench/releasenotes/.

Submission enhancements

Influenza and Norovirus submissions. The Submission Portal for GenBank (submit.ncbi.nlm.nih.gov/subs/genbank/) provides two new wizards to streamline submissions of *Influenza* and *Norovirus* genomes. These wizards accelerate the submission process and provide automatic feature annotation and validation functions. The wizards accept FASTA formatted sequences and require the following source information: isolate, serotype/genotype, collection date, host, and country of collection. In addition, all of the sequences in a submission must be derived from one virus subtype. Looking forward, we plan to continue releasing similar tools for additional marker genes. More information is available at submit.ncbi.nlm.nih.gov/about/genbank/.

Feature propagation. Submitters often need to deposit a large set of related sequences that typically share a common set of feature annotations. In such cases it is convenient to provide annotations for one sequence in the set and then automatically propagate these annotations onto the remaining members of the set. BankIt now supports this feature propagation function, greatly easing the handling of large sequence sets.

Taxonomy handling in BLAST

Recent enhancements to the BLAST+ command-line programs leverage the new version 5 BLAST databases (ftp.ncbi.nlm.nih.gov/blast/db/v5/blastdbv5.pdf) to provide important new functions. In particular, the version 5 databases (BLASTDBv5) index proteins by their accession.version identifiers and also include taxonomic information. This allows users to restrict protein BLAST searches by taxonomy and also to retrieve sequences from these databases using taxonomic limits. Users can also efficiently limit searches by lists of accession.version identifiers.

Pathogen sequences

Genome sequences from the NCBI Pathogen Detection Project (4,5) are now being deposited in GenBank. Given the large amount of data being submitted by these surveillance efforts, the Assembly resource (www.ncbi.nlm.nih.gov/assembly/) now provides an easy way to exclude such sequences from searches of the database. The new filter, available on the left side bar, is 'exclude derived from surveillance project' and is checked by default for all Assembly searches.

Expanded sequence identifier formats

As announced in the notes for GenBank release 226, in 2018 the INSDC expanded the ranges of accession number formats to accommodate the rapid growth of sequence databases. We want to emphasize that none of these new accessions will replace any existing accession, and all existing sequences will continue to be retrievable using their current accession.version identifiers. In some cases existing (and exhausted) prefixes have been reactivated with extended numerical suffixes (e.g. JG00000001–JG99999999 versus the exhausted JG000001–JG999999). To be clear, JG000001 (existing accession) and JG00000001 (new accession) will refer to two distinct sequences.

MAILING ADDRESS

GenBank, National Center for Biotechnology Information, Building 45, Room 6AN12D-37, 45 Center Drive, Bethesda, MD 20892, USA.

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.

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., Ostell,J., Pruitt,K.D. and Karsch-Mizrachi,I. (2019) GenBank. *Nucleic Acids Res.*, **47**, D94–D99.
2. 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.
3. Ciufo,S., Kannan,S., Sharma,S., Badretdin,A., Clark,K., Turner,S., Brover,S., Schoch,C.L., Kimchi,A. and DiCuccio,M. (2018) Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int. J. Syst. Evol. Microbiol.*, **68**, 2386–2392.
4. Timme,R.E., Rand,H., Shumway,M., Trees,E.K., Simmons,M., Agarwala,R., Davis,S., Tillman,G.E., Defibaugh-Chavez,S., Carleton,H.A. *et al.* (2017) Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance. *PeerJ*, **5**, e3893.
5. NCBI Resource Coordinators. (2017) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **45**, D12–D17.