GenBank

The **GenBank** sequence database is an open access, annotated collection of all publicly available nucleotide sequences and their protein translations. It is produced and maintained by the National Center for Biotechnology Information (NCBI; a part of the National Institutes of Health in the United States) as part of the International Nucleotide Sequence Database Collaboration (INSDC).

GenBank and its collaborators receive sequences produced in laboratories throughout the world from more than 500,000 formally described species.[2] The database started in 1982 by Walter Goad and Los Alamos National Laboratory. GenBank has become an important database for research in biological fields and has grown in recent years at an exponential rate by doubling roughly every 18 months.[3][4]

Release 250.0, published in June 2022, contained over 17 trillion nucleotide bases in more than 2,45 billion sequences.[5] GenBank is built by direct submissions from individual laboratories, as well as from bulk submissions from large-scale sequencing centers.

Submissions

Only original sequences can be submitted to GenBank. Direct submissions are made to GenBank using BankIt, which is a Web-based form, or the stand-alone submission program, Sequin. Upon receipt of a sequence submission, the GenBank staff examines the originality of the data and assigns an accession number to the sequence and performs quality assurance checks. The submissions are then released to the public database, where the entries are retrievable by Entrez or downloadable by FTP. Bulk submissions of Expressed Sequence Tag (EST), Sequence-tagged site (STS), Genome Survey Sequence (GSS), and High-Throughput Genome Sequence (HTGS) data are most often submitted by large-scale sequencing centers. The GenBank direct submissions group also processes complete microbial genome sequences.[6][7]

History

Walter Goad of the Theoretical Biology and Biophysics Group at Los Alamos National Laboratory (LANL) and others established the Los Alamos Sequence Database in 1979, which culminated in 1982 with the creation of the public GenBank.[8] Funding was provided by the National Institutes of Health, the National Science

GenBank	
Content	
Description	Nucleotide sequences for more than 300,000 organisms with supporting bibliographic and biological annotation.
Data types captured	Nucleotide sequence Protein sequence
Organisms	All
Contact	
Research center	NCBI
Primary citation	PMID 21071399 (https://pubmed.ncbi.nlm.nih.gov/21071399)
Release date	1982
Access	
Data format	XML ASN.1 Genbank format
Website	NCBI (https://www.ncbi.nlm.nih.gov/)
Download URL	ncbi ftp (http://ftp.ncbi.nih.gov/)
Web service URL	eutils (http://eutils.ncbi.nlm.nih.gov/) soap (http://eutils.ncbi.nlm.nih.g

Foundation, the Department of Energy, and the Department of Defense. LANL collaborated on GenBank with the firm Bolt, Beranek, and Newman, and by the end of 1983 more than 2,000 sequences were stored in it.

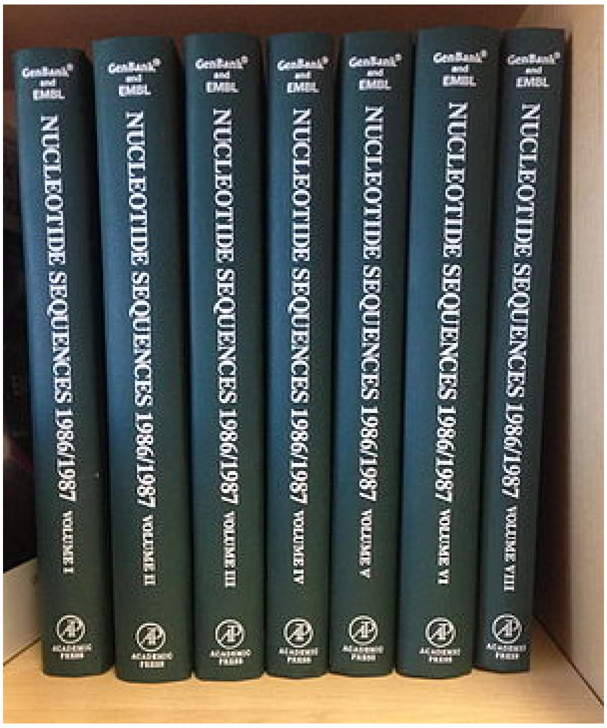
In the mid 1980s, the Intelligenetics bioinformatics company at Stanford University managed the GenBank project in collaboration with LANL.[9] As one of the earliest bioinformatics community projects on the Internet, the GenBank project started BIOSCI/Bionet news groups for promoting open access communications among bioscientists. During 1989 to 1992, the GenBank project transitioned to the newly created National Center for Biotechnology Information (NCBI).[10]

	ov/soap/v2.0/eutils.wsdl)
Tools	
Web	BLAST
Standalone	BLAST
Miscellaneous	
License	Unclear[1]

Growth

The GenBank release notes for release 250.0 (June 2022) state that "from 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months".[5][11] As of 15 June 2022, GenBank release 250.0 has over 239 million loci, 1,39 trillion nucleotide bases, from 239 million reported sequences.[5]

The GenBank database includes additional data sets that are constructed mechanically from the main sequence data collection, and therefore are excluded from this count.

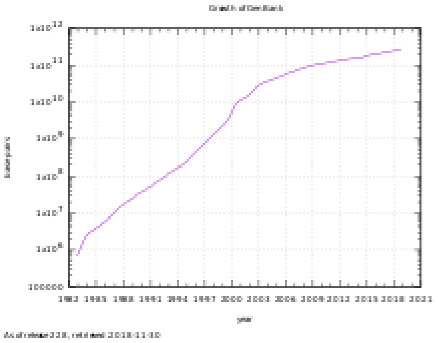


Genbank and EMBL: NucleotideSequences 1986/1987 Volumes I to VII.

CD-ROM of Genbank v100

Top 20 organisms in GenBank (Release 250)[5]

Organism	base pairs
Triticum aestivum	$2.15443744183 \times 10^{11}$
SARS-CoV-2	$1.65771825746 \times 10^{11}$
Hordeum vulgare subsp. *vulgare*	$1.01344340096 \times 10^{11}$
Mus musculus	$3.0614386913 \times 10^{10}$
Homo sapiens	$2.7834633853 \times 10^{10}$
Avena sativa	$2.1127939362 \times 10^{10}$
Escherichia coli	$1.5517830491 \times 10^{10}$
Klebsiella pneumoniae	$1.1144687122 \times 10^{10}$
Danio rerio	$1.0890148966 \times 10^{10}$
Bos taurus	$1.0650671156 \times 10^{10}$
Triticum turgidum subsp. *durum*	$9.981529154 \times 10^{9}$
Zea mays	$7.412263902 \times 10^{9}$
Avena insularis	$6.924307246 \times 10^{9}$
Secale cereale	$6.749247504 \times 10^{9}$
Rattus norvegicus	$6.548854408 \times 10^{9}$
Aegilops longissima	$5.920483689 \times 10^{9}$
Canis lupus familiaris	$5.776499164 \times 10^{9}$
Aegilops sharonensis	$5.272476906 \times 10^{9}$
Sus scrofa	$5.179074907 \times 10^{9}$
Rhinatrema bivittatum	$5.178626132 \times 10^{9}$



Growth in GenBank base pairs, 1982 to 2018, on a semi-log scale

Incomplete identifications

Public databases which may be searched using the National Center for Biotechnology Information Basic Local Alignment Search Tool (NCBI BLAST), lack peer-reviewed sequences of type strains and sequences of non-type strains. On the other hand, while commercial databases potentially contain high-quality filtered sequence data, there are a limited number of reference sequences.

A paper released in the *Journal of Clinical Microbiology*[12] evaluated the 16S rRNA gene sequencing results analyzed with GenBank in conjunction with other freely available, quality-controlled, web-based public databases, such as the EzTaxon-e[13] and the BIBI[14] databases. The results showed that analyses performed using GenBank combined with EzTaxon-e (kappa = 0.79) were more discriminative than using GenBank (kappa = 0.66) or other databases alone.

GenBank, being a public database, may contain sequences wrongly assigned to a particular species, because the initial identification of the organism was wrong. A recent article published in *Genome* showed that 75% of mitochondrial Cytochrome c oxidase subunit I sequences were wrongly assigned to the fish

Nemipterus mesoprion resulting from continued usage of sequences of initially misidentified individuals.[15] The authors provide recommendations how to avoid further distribution of publicly available sequences with incorrect scientific names.

Numerous published manuscripts have identified erroneous sequences on GenBank.[16][17][18] These are not only incorrect species assignments (which can have different causes) but also include chimeras and accession records with sequencing errors. A recent manuscript on the quality of all Cytochrome b records of birds further showed that 45% of the identified erroneous records lack a voucher specimen that prevents a reassessment of the species identification.[19]

See also

- Ensembl
- Human Protein Reference Database (HPRD)
- Sequence analysis
- UniProt
- List of sequenced eukaryotic genomes
- List of sequenced archaeal genomes
- RefSeq — the Reference Sequence Database
- Geneious — includes a GenBank Submission Tool
- Open science data
- Open Standard

References

1. The download page (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/) at UCSC says "NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank."
2. Eric W Sayers; Mark Cavanaugh; Karen Clark; Kim D Pruitt; Conrad L Schoch; Stephen T Sherry; Ilene Karsch-Mizrachi (7 January 2022). "GenBank" (https://doi.org/10.1093%2Fnar%2Fgkab1135). *Nucleic Acids Archive*. **50** (D1): D161–D164. doi:10.1093/nar/gkab1135 (https://doi.org/10.1093%2Fnar%2Fgkab1135).
3. Benson D; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Wheeler, D. L.; et al. (2008). "GenBank" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2238942). *Nucleic Acids Research*. **36** (Database): D25–D30. doi:10.1093/nar/gkm929 (https://doi.org/10.1093%2Fnar%2Fgkm929). PMC 2238942 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2238942). PMID 18073190 (https://pubmed.ncbi.nlm.nih.gov/18073190).
4. Benson D; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Sayers, E. W.; et al. (2009). "GenBank" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2686462). *Nucleic Acids Research*. **37** (Database): D26–D31. doi:10.1093/nar/gkn723 (https://doi.org/10.1093%2Fnar%2Fgkn723). PMC 2686462 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2686462). PMID 18940867 (https://pubmed.ncbi.nlm.nih.gov/18940867).
5. "GenBank release notes (Release 250)" (http://ftp.ncbi.nih.gov/genbank/gbrel.txt). NCBI. 15 June 2022. Retrieved 20 July 2022.
6. "How to submit data to GenBank" (https://www.ncbi.nlm.nih.gov/genbank/submit/). *NCBI*. Retrieved 20 July 2022.

7. "GenBank Submission Types" (https://www.ncbi.nlm.nih.gov/genbank/submit_types/). *NCBI*. Retrieved 20 July 2022.
8. Hanson, Todd (2000-11-21). "Walter Goad, GenBank founder, dies" (http://www.lanl.gov/orgs/pa/News/112100.html). *Newsbulletin: obituary*. Los Alamos National Laboratory.
9. LANL GenBank History (http://www.bio.net/bionet/mm/bionews/1994-January/000877.html)
10. Benton D (1990). "Recent changes in the GenBank On-line Service" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC330520). *Nucleic Acids Research*. **18** (6): 1517–1520. doi:10.1093/nar/18.6.1517 (https://doi.org/10.1093%2Fnar%2F18.6.1517). PMC 330520 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC330520). PMID 2326192 (https://pubmed.ncbi.nlm.nih.gov/2326192).
11. Benson, D. A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Sayers, E. W. (2012). "GenBank" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531190). *Nucleic Acids Research*. **41** (Database issue): D36–D42. doi:10.1093/nar/gks1195 (https://doi.org/10.1093%2Fnar%2Fgks1195). PMC 3531190 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531190). PMID 23193287 (https://pubmed.ncbi.nlm.nih.gov/23193287).
12. Kyung Sun Park; Chang-Seok Ki; Cheol-In Kang; Yae-Jean Kim; Doo Ryeon Chung; Kyong Ran Peck; Jae-Hoon Song; Nam Yong Lee (May 2012). "Evaluation of the GenBank, EzTaxon, and BIBI Services for Molecular Identification of Clinical Blood Culture Isolates That Were Unidentifiable or Misidentified by Conventional Methods" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3347139). *J. Clin. Microbiol*. **50** (5): 1792–1795. doi:10.1128/JCM.00081-12 (https://doi.org/10.1128%2FJCM.00081-12). PMC 3347139 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3347139). PMID 22403421 (https://pubmed.ncbi.nlm.nih.gov/22403421).
13. EzTaxon-e Database (https://web.archive.org/web/20130928154318/http://eztaxon-e.ezbiocloud.net/) *eztaxon-e.ezbiocloud.net* (archive accessed 25 March 2021)
14. leBIBI V5 (https://web.archive.org/web/20151001000357/http://pbil.univ-lyon1.fr/bibi/) *pbil.univ-lyon1.fr* (archive accessed 25 March 2021)
15. Ogwang, Joel; Bariche, Michel; Bos, Arthur R. (2021). "Genetic diversity and phylogenetic relationships of threadfin breams (*Nemipterus* spp.) from the Red Sea and eastern Mediterranean Sea" (https://cdnsciencepub.com/doi/full/10.1139/gen-2019-0163). *Genome*. **64** (3): 207–216. doi:10.1139/gen-2019-0163 (https://doi.org/10.1139%2Fgen-2019-0163).
16. van den Burg, Matthijs P.; Herrando-Pérez, Salvador; Vieites, David R. (13 August 2020). "ACDC, a global database of amphibian cytochrome-b sequences using reproducible curation for GenBank records" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7426930). *Scientific Data*. **7** (1). doi:10.1038/s41597-020-00598-9 (https://doi.org/10.1038%2Fs41597-020-00598-9). eISSN 2052-4463 (https://www.worldcat.org/issn/2052-4463). PMC 7426930 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7426930). PMID 32792559 (https://pubmed.ncbi.nlm.nih.gov/32792559).
17. Li, Xiaobing; Shen, Xuejuan; Chen, Xiao; Xiang, Dan; Murphy, Robert W.; Shen, Yongyi (6 February 2018). "Detection of Potential Problematic Cytb Gene Sequences of Fishes in GenBank" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5808227). *Frontiers in Genetics*. **9**. doi:10.3389/fgene.2018.00030 (https://doi.org/10.3389%2Ffgene.2018.00030). eISSN 1664-8021 (https://www.worldcat.org/issn/1664-8021). PMC 5808227 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5808227). PMID 29467794 (https://pubmed.ncbi.nlm.nih.gov/29467794).

18. Heller, Philip; Casaletto, James; Ruiz, Gregory; Geller, Jonathan (7 August 2018). "A database of metazoan cytochrome c oxidase subunit I gene sequences derived from GenBank with CO-ARBitrator" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6080493). *Scientific Data*. **5** (1). doi:10.1038/sdata.2018.156 (https://doi.org/10.1038%2Fsdata.2018.156). eISSN 2052-4463 (https://www.worldcat.org/issn/2052-4463). PMC 6080493 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6080493). PMID 30084847 (https://pubmed.ncbi.nlm.nih.gov/30084847).
19. Van Den Burg, Matthijs P.; Vieites, David R. (22 September 2022). "Bird genetic databases need improved curation and error reporting to <scp>NCBI</scp>" (https://doi.org/10.1111%2Fibi.13143). *Ibis*. doi:10.1111/ibi.13143 (https://doi.org/10.1111%2Fibi.13143). eISSN 1474-919X (https://www.worldcat.org/issn/1474-919X). ISSN 0019-1019 (https://www.worldcat.org/issn/0019-1019).

- This article incorporates public domain material from *NCBI Handbook* (https://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowTOC&rid=handbook.TOC&depth=2). National Center for Biotechnology Information.

External links

- GenBank (https://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide)
- Example sequence record, for hemoglobin beta (https://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=28302128)
- BankIt (https://www.ncbi.nlm.nih.gov/BankIt/)
- Sequin (https://www.ncbi.nlm.nih.gov/Sequin/index.html) — a stand-alone software tool developed by the NCBI for submitting and updating entries to the GenBank sequence database.
- EMBOSS (https://emboss.sourceforge.net) — free, open source software for molecular biology
- GenBank, RefSeq, TPA and UniProt: What's in a Name? (https://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.section.GenBank_ASM)