

GenBank

**Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell
and David L. Wheeler***

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

Received September 15, 2005; Revised and Accepted October 31, 2005

ABSTRACT

GenBank (R) is a comprehensive database that contains publicly available DNA sequences for more than 205 000 named organisms, obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects. Most submissions are made using the Web-based BankIt or standalone Sequin programs and accession numbers are assigned by GenBank staff upon receipt. Daily data exchange with the EMBL Data Library in Europe and the DNA Data Bank of Japan ensures worldwide coverage. GenBank is accessible through NCBI's retrieval system, Entrez, which integrates data from the major DNA and protein sequence databases along with taxonomy, genome, mapping, protein structure and domain information, and the biomedical journal literature via PubMed. BLAST provides sequence similarity searches of GenBank and other sequence databases. Complete bimonthly releases and daily updates of the GenBank database are available by FTP. To access GenBank and its related retrieval and analysis services, go to the NCBI Homepage at www.ncbi.nlm.nih.gov.

INTRODUCTION

GenBank (1) is a comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotation, 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.

NCBI builds GenBank primarily from the submission of sequence data from authors and from the bulk submission of expressed sequence tag (EST), genome survey sequence

(GSS) and other high-throughput data from sequencing centers. The US Office of Patents and Trademarks also contributes sequences from issued patents. GenBank, the EMBL Data Library (2) in Europe, and the DNA Databank of Japan (DDBJ) (3) comprise the International Nucleotide Sequence Databases and are members of a long-standing collaboration in which data are exchanged daily to ensure a uniform and comprehensive collection of sequence information. NCBI makes the GenBank data available at no cost over the Internet, via FTP and a wide range of Web-based retrieval and analysis services which operate on the GenBank data (4) (info@ncbi.nlm.nih.gov).

ORGANIZATION OF THE DATABASE

From its inception, GenBank has doubled in size about every 18 months. It currently contains over 51 billion nucleotide bases from more than 46 million individual sequences, with 8 million new sequences added in the past year. Contributions from Whole Genome Shotgun (WGS) projects supplement the data in the traditional divisions to bring the total beyond 100 gigabases. Complete genomes (www.ncbi.nlm.nih.gov/Genomes/index.html) represent a growing portion of the database, with over 70 of more than 250 complete microbial genomes in GenBank deposited over the past year. The number of eukaryote genomes for which coverage and assembly are significant continues to increase as well, with over 90 such assemblies now available, including that of the reference human genome.

Sequence-based taxonomy

Database sequences are classified and can be queried using a comprehensive sequence-based taxonomy (www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html) developed by NCBI in collaboration with EMBL and DDBJ and with the valuable assistance of external advisers and curators. Over 205 000 named species are represented in GenBank and new species are being added at the rate of over 3000 per month. About 16% of the sequences in GenBank are of human origin and 11% of

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

all sequences are human ESTs. After *Homo sapiens*, the top species in GenBank in terms of number of bases are *Mus musculus*, *Rattus norvegicus*, *Danio rerio*, *Bos taurus*, *Zea mays*, *Oryza sativa*, *Xenopus tropicalis*, *Canis familiaris* and *Drosophila melanogaster*.

GenBank records and divisions

Each GenBank entry includes a concise description of the sequence, the scientific name and taxonomy of the source organism, bibliographic references and a table of features (www.ncbi.nlm.nih.gov/collab/FT/index.html) listing areas of biological significance, such as coding regions and their protein translations, transcription units, repeat regions and sites of mutations or modifications.

The files in the GenBank distribution have traditionally been partitioned into 'divisions' that roughly correspond to taxonomic groups, such as bacteria (BCT), viruses (VRL), primates (PRI) and rodents (ROD). In recent years, divisions have been added to support specific sequencing strategies. These include divisions for EST, GSS, high throughput genomic (HTG), high throughput cDNA (HTC) and environmental sample (ENV) sequences, making a total of 18 divisions. For convenience in file transfer, the larger divisions, such as the EST and PRI, are partitioned into multiple files for the bimonthly GenBank releases on NCBI's FTP site.

Expressed sequence tags

ESTs continue to be the major source of new sequence records and gene sequences, comprising over 15 billion nucleotide bases in GenBank release 149. Over the past year, the number of ESTs has increased by over 21% to a total of 28.4 million sequences representing more than 740 different organisms. The top five organisms represented in the EST division are *H.sapiens* (6.1 million records), *M.musculus* (4.3 million records), *X.tropicalis* (963 000 records), *Rattus* sp. (701 000 records), *Ciona intestinalis* (684 000 records) and *D.rerio* (635 000 records). As part of its daily processing of GenBank EST data, NCBI identifies through BLAST searches all homologies for new EST sequences and incorporates that information into the companion database, dbEST (www.ncbi.nlm.nih.gov/dbEST/index.html) (5). The data in dbEST are processed further to produce the UniGene database (www.ncbi.nlm.nih.gov/UniGene/) of more than 806 000 gene-oriented sequence clusters representing over 50 organisms, described more fully in (4).

Sequence-tagged sites (STSs), GSSs and ENV sequences

The STS division of GenBank (www.ncbi.nlm.nih.gov/dbSTS/index.html) contains over 875 000 sequences, double last year's count, including anonymous STSs based on genomic sequence as well as gene-based STSs derived from the 3' ends of genes and ESTs. These STS records usually include primer sequences, annotations and PCR reaction conditions.

The GSS division of GenBank (www.ncbi.nlm.nih.gov/dbGSS/index.html) has grown over the past year by 27% to a total of 12.2 million records for over 500 organisms and comprises over 7.6 billion nucleotide bases. GSS records are predominantly single reads from Bacterial Artificial

Chromosomes ('BAC-ends') used in a variety of genome sequencing projects. The most highly represented species in the GSS division are *Zea mays* (1.9 million records), *M.musculus* (1.5 million records), *H.sapiens* (906 000 records) and *C.familiaris* (854 000 records). The human data have been used (www.ncbi.nlm.nih.gov/genome/clone) along with the STS records in tiling the BACs for the Human Genome Project (6).

The ENV division of GenBank, for non-WGS sequences obtained via environmental sampling methods in which the source organism is unknown, debuted with release 147 in April 2005. Records in the ENV division contain 'ENV' in the keyword field and use an 'environmental_sample' qualifier in the source feature. As of GenBank release 149, the ENV division of GenBank contained over 175 000 sequences, comprising 136 million base pairs, representing more than 3500 studies.

HTG and HTC sequences

The HTG division of GenBank (www.ncbi.nlm.nih.gov/HTGS/) contains unfinished large-scale genomic records that are in transition to a finished state (7). These records are designated as Phase 0-3 depending on the quality of the data. Upon reaching Phase 3, the finished state, HTG records are moved into the appropriate organism division of GenBank. As of release 149 of GenBank, the HTG division comprised almost 13 billion base pairs of sequence.

The HTC division of GenBank accommodates high-throughput cDNA sequences. HTCs are of draft quality but may contain 5' untranslated regions (5' UTRs) and 3' UTRs, partial coding regions and introns. HTC sequences which are finished and of high quality are moved to the appropriate organism GenBank division. GenBank release 149 contained more than 380 000 HTC sequences totaling over 422 million bases. One project generating HTC data is described in (8).

Whole genome shotgun sequence

Over 50 million bases of WGS sequence appears in GenBank as sets of WGS contigs, many of them bearing annotations, originating from a single sequencing project. These sequences are issued accession numbers consisting of a 4-letter project ID, followed by a two-digit version number and a 6-digit contig ID. Hence, the WGS accession number 'AAAA 01072744' is assigned to contig number '072744' of the first version of project 'AAAA'. WGS sequencing projects have contributed over 11 million contigs to GenBank, a 3-fold increase over past year's total. These primary sequences have been used to construct some 332 000 large-scale assemblies of scaffolds and chromosomes. WGS project contigs for *H.sapiens*, *C.familiaris*, *Pan troglodytes*, *Drosophila*, *Saccharomyces*, and more than 200 other organisms and environmental samples are available. For a complete list of WGS projects with links to the data, see www.ncbi.nlm.nih.gov/projects/WGS/WGSprojectlist.cgi.

Submitters of WGS sequences, and genomic sequences in general, are urged to use a new set of evidence tags, described below, in their annotations. In the case of WGS records, these annotations are not tracked from one assembly version to the next and should be considered preliminary.

New evidence qualifiers

The International Nucleotide Sequence Databases have adopted two new qualifiers to describe the evidence for features annotated in sequence records. The new qualifiers have the form `/experimental=text` and `/inference=TYPE:text`, where *TYPE* is one of a number of standard inference types and *text* is made up of structured text. These new qualifiers replace 'evidence=experimental' and 'evidence=non-experimental', respectively, which are no longer supported. New versions of the 'tbl2asn' and 'Sequin' sequence submission programs, described below, support the new qualifiers. For details about the new qualifiers and examples of their use, see <http://www.ncbi.nlm.nih.gov/Genbank/evidence.html>.

BUILDING THE DATABASE

The data in GenBank, and the collaborating databases EMBL and DDBJ, are submitted primarily by individual authors to one of the three databases, or by sequencing centers as batches of EST, STS, GSS, HTC, WGS or HTG sequences. Data are exchanged daily with DDBJ and EMBL so that the daily updates from NCBI servers incorporate the most recently available sequence data from all sources.

Direct submission

Virtually all records enter GenBank as direct electronic submissions (www.ncbi.nlm.nih.gov/Genbank/index.html), with the majority of authors using the BankIt or Sequin programs. Many journals require authors with sequence data to submit the data to a public database as a condition of publication.

GenBank staff can usually assign an accession number to a sequence submission within two working days of receipt, and do so at a rate of almost 1600 per day. The accession number serves as confirmation that the sequence has been submitted and allows readers of articles in which the sequence is cited to retrieve the data. Direct submissions receive a quality assurance review that includes checks for vector contamination, proper translation of coding regions, correct taxonomy and correct bibliographic citations. A draft of the GenBank record is passed back to the author for review before it enters the database. Authors may ask that their sequences be kept confidential until the time of publication. Since GenBank policy requires that deposited sequence data be made public when the sequence or accession number is published, authors are instructed to inform GenBank staff of the publication date of the article in which the sequence is cited in order to ensure a timely release of the data. Although only the submitting scientist is permitted to modify sequence data or annotations, all users are encouraged to report lags in releasing data or possible errors or omissions to GenBank.update@ncbi.nlm.nih.gov.

NCBI works closely with sequencing centers to ensure timely incorporation of bulk data into GenBank for public release. GenBank offers special batch procedures for large-scale sequencing groups to facilitate data submission, including the program 'tbl2asn', described at www.ncbi.nlm.nih.gov/Sequin/table.html.

Sequence identifiers and accession numbers

Each GenBank record, consisting of both a sequence and its annotations, is assigned a stable and unique identifier, the accession number, which is shared across the three collaborating databases (GenBank, DDBJ and EMBL) and remains constant over the lifetime of the record even when there is a change to the sequence or annotation. The DNA sequence within a GenBank record is also assigned a unique NCBI identifier, called a 'gi', that appears on the VERSION line of GenBank flatfile records following the accession number. A third identifier of the form 'Accession.version', also displayed on the VERSION line of flatfile records, consolidates the information present in both the gi and accession numbers. An entry appearing in the database for the first time has an 'Accession.version' identifier equivalent to the ACCESSION number of the GenBank record followed by '.1' to indicate the first version of the sequence for the record, e.g. Accession no. AF000001; Version AF000001.1 GI: 987654321.

When a change is made to a sequence given in a GenBank record, a new gi number is issued to the sequence and the version extension of the 'Accession.version' identifier is incremented. The accession number for the record as a whole remains unchanged and the older sequence remains available under the old 'Accession.version' identifier and gi.

A similar system tracks changes in the corresponding protein translations using 'Accession.version' identifiers comprising a protein accession number, e.g. AAA00001, followed by a version number. These identifiers appear as qualifiers for CDS features in the FEATURES portion of a GenBank entry, e.g. `/protein_id='AAA00001.1'` Protein sequence translations also receive their own unique gi number, which appears as a second qualifier on the CDS feature, e.g. `/db_xref='GI:1233445'`.

Third Party Annotation (TPA)

TPA records currently support the reporting of published, experimentally confirmed sequence annotation by a scientist other than the original submitter of the primary sequence record in DDBJ/EMBL/GenBank. The scope of the annotations permitted will be expanded in the future to include those derived by inference. TPA sequences may be created by assembling a number of primary sequences. The format of a TPA record (e.g. BK000016) is similar to that of a conventional GenBank record but includes the label 'TPA:' at the beginning of each Definition Line and the keywords 'Third Party Annotation; TPA' in the Keywords field. The Comment field of TPA records lists the primary sequences used to assemble the TPA sequence; the Primary field provides the base ranges of the primary sequences that contribute to the TPA sequence.

Over 4500 TPA records are contained in GenBank release 149, including over 2000 for *D.melanogaster*, 900 for *H.sapiens*, 600 for *O.sativa* and 200 for *M.musculus*. TPA submissions to GenBank may be made using either BankIt, or Sequin but TPA sequences are not released to the public until their accession numbers or sequence data and annotation appear in a peer-reviewed biological journal. For more information on TPA, see www.ncbi.nlm.nih.gov/Genbank/TPA.html.

GenBank CON records for assemblies of smaller records

In 1995, the DDBJ/EMBL/GenBank International Nucleotide Sequence Collaboration Databases agreed to a 350 kb limit on the size of most database sequence records in order to conform to the limitations on sequence length of existing molecular biology software. Exceptions were made in the cases of HTG sequence, assemblies of WGS project data and for large eukaryotic genes. The large records that were broken into multiple 350 kb segments to conform to the standard were represented in the GenBank 'CON' division as sets of assembly instructions to allow the transparent display and download of the full record using tools such as NCBI's Entrez. Because of the greater ability of current software to efficiently handle long sequences the 350 kb limit was removed by the Database Collaborators as of June 2004. Although the removal of the limit has immediately allowed many genomes, such as bacterial genomes, to be represented in GenBank as single sequences, it will still be desirable from the standpoints of data transfer and analysis to break some very long sequences, such as portions of eukaryotic genomes, into smaller segments. In these cases, CON division records for the entire sequence will continue to contain assembly instructions to allow the seamless display and download of the sequence.

BankIt

About one-third of author submissions are received through NCBI's Web-based data submission tool, BankIt (www.ncbi.nlm.nih.gov/BankIt). Using BankIt, authors enter sequence information directly into a form and add biological annotation, such as coding regions or mRNA features. Free-form text boxes list boxes, and pull-down menus allow the submitter to further describe the sequence without having to learn formatting rules or restricted vocabularies. BankIt validates submissions, flagging many common errors, and checks for vector contamination using a variant of BLAST called Vecscreen, before creating a draft record in GenBank flat file format for the submitter to review. BankIt is the tool of choice for simple submissions, especially when only one or a small number of records is to be submitted (7). BankIt can also be used by submitters to update their existing GenBank records.

Sequin and tbl2asn

NCBI also offers a standalone multi-platform submission program called Sequin (www.ncbi.nlm.nih.gov/Sequin/index.html) that can be used interactively with other NCBI sequence retrieval and analysis tools. Sequin handles simple sequences such as a cDNA, as well as segmented entries, phylogenetic studies, population studies, mutation studies, environmental samples, and alignments for which BankIt and other Web-based submission tools are not well-suited. Sequin has convenient editing and complex annotation capabilities and contains a number of built-in validation functions for quality assurance. In addition, Sequin is able to accommodate large sequences, such as that of the 5.6 Mb *Escherichia coli* genome, and read in a full complement of annotations via simple tables. Versions for Macintosh, PC and Unix computers are available via anonymous FTP at 'ftp.ncbi.nih.gov' in the 'sequin' directory. Once a submission is completed, submitters can e-mail the Sequin file to the address gb-sub@ncbi.nlm.nih.gov.

Submitters of large, heavily annotated genomes may find it convenient to use 'tbl2asn', referenced above under 'Direct submission', to convert a table of annotations generated via an annotation pipeline, into an ASN.1 record suitable for submission to GenBank.

RETRIEVING GENBANK DATA

The ENTREZ system

The sequence records in GenBank are accessible via Entrez (www.ncbi.nlm.nih.gov/Entrez/), a robust and flexible database retrieval system that covers over 30 biological databases containing DNA and protein sequence data, genome mapping data, population sets, phylogenetic sets, environmental sample sets, gene expression data, the NCBI taxonomy, protein domain information, protein structures from the Molecular Modeling Database, MMDB (9), and MEDLINE references via PubMed. The Entrez sequence databases are taken from a variety of sources and therefore include more sequence data than is available within GenBank alone.

BLAST sequence-similarity searching

Sequence-similarity searches are the most frequent and basic type of analysis performed on the GenBank data. NCBI offers the BLAST (www.ncbi.nlm.nih.gov/BLAST/) family of programs to locate regions of similarity between a query sequence and database sequences (10,11). BLAST searches may be performed on NCBI's Web site, or using a set of stand-alone programs distributed by FTP. BLAST is discussed in a separate article in this issue (4).

Obtaining GenBank by FTP

NCBI distributes the GenBank releases in the traditional flat-file format as well as in the Abstract Syntax Notation (ASN.1) format used for internal maintenance. The full bimonthly GenBank release and the daily updates, which also incorporate sequence data from EMBL and DDBJ, are available by anonymous FTP from NCBI at (<ftp://ftp.ncbi.nih.gov>) as well as from a mirror site at the University of Indiana (<ftp://bio-mirror.net/biomirror/genbank/>). The full release in flat-file format is available as compressed files in the directory, 'genbank' with a non-cumulative set of updates contained in 'daily-nc'. A script is provided in the 'tools' directory of the GenBank FTP site to convert a set of daily updates into a cumulative update.

Plans for the future

NCBI has been working with the Consortium for the Barcode of Life (CBOL) (http://barcoding.si.edu/index_detail.htm) to create a new tool for the bulk submission of sequences to GenBank. CBOL is an international initiative devoted to developing DNA barcoding as a tool for characterizing species of organisms using a short DNA sequence derived from a portion of the cytochrome oxidase subunit I gene. Barcode submissions to GenBank include the cytochrome oxidase subunit I sequence combined with a standard set of elements to describe the organism. NCBI offers a new submission tool for Barcode sequences (<http://www.ncbi.nlm.nih.gov/BankIt/barcode/>). Unlike Bankit, which is form based, this web-based submission tool allows users to upload files containing

a batch of sequences with associated source information. It is anticipated that this tool will be used for other types of bulk submissions in the near future.

CITING GENBANK

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

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by the National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2005) GenBank: Update. *Nucleic Acids Res.*, **33**, D34–D38.
2. Kanz,C., Aldebert,P., Althorpe,N., Baker,W., Baldwin,A., Bates,K., Browne,P., van den Broek,A., Castro,M., Cochrane,G. *et al.* (2005) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **33**, D29–D33.
3. Tateno,Y., Saitou,N., Okubo,K., Sugawara,H. and Gojobori,T. (2005) DDBJ in collaboration with mass-sequencing teams on annotation. *Nucleic Acids Res.*, **33**, D25–D28.
4. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **34**, D173–D180.
5. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—database for “expressed sequence tags”. *Nature Genet.*, **4**, 332–333.
6. Smith,M.W., Holmsen,A.L., Wei,Y.H., Peterson,M. and Evans,G.A. (1994) Genomic sequence sampling: a strategy for high resolution sequence-based physical mapping of complex genomes. *Nature Genet.*, **7**, 40–47.
7. Kans,J.A. and Ouellette,B.F.F. (2001) In Baxevanis,A. and Ouellette,B.F.F. (eds), *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. John Wiley and Sons, Inc., New York, NY, pp. 65–81.
8. Hayashizaki,Y., Kawai,J., Shinagawa,A., Shibata,K., Yoshino,M., Itoh,M., Ishii,Y., Arakawa,T., Hara,A., Fukunishi,Y. *et al.* (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**, 685–690.
9. Marchler-Bauer,A., Anderson,J., Fedorova,N., DeWeese-Scott,C., Geer,L.Y., Hurwitz,D., Jackson,J.J., Jacobs,A., Lanczycki,C., Liebert,C. *et al.* (2005) MMDB: Entrez’s 3D-structure database. *Nucleic Acids Res.*, **33**, D192–D196.
10. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
11. Zhang,Z., Schaffer,A.A., Miller,W., Madden,T.L., Lipman,D.J., Koonin,E.V. and Altschul,S.F. (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.*, **26**, 3986–3991.