D30–D35 *Nucleic Acids Research, 2015, Vol. 43, Database issue* *Published online 20 November 2014*
doi: 10.1093/nar/gku1216

GenBank

Dennis A. Benson, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell and Eric W. Sayers*

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

Received October 06, 2014; Revised November 05, 2014; Accepted November 06, 2014

ABSTRACT

GenBank® (http://www.ncbi.nlm.nih.gov/genbank/) is a comprehensive database that contains publicly available nucleotide sequences for over 300 000 formally described species. These sequences are obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects, including whole-genome shotgun and environmental sampling projects. Most submissions are made using the web-based BankIt or standalone Sequin programs, and GenBank staff assign accession numbers upon data receipt. Daily data exchange with the European Nucleotide Archive and the DNA Data Bank of Japan ensures worldwide coverage. GenBank is accessible through the NCBI Entrez retrieval system, 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.

INTRODUCTION

GenBank (1) is a comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotation. GenBank is built and distributed by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine, located on the campus of the U.S. National Institutes of Health (NIH) in Bethesda, MD, USA.

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), whole-genome shotgun (WGS) and other high-throughput data from sequencing centers. The U.S. Patent and Trademark Office also contributes sequences from issued patents. GenBank participates with the EMBL European Nucleotide Archive (ENA) (2), and the DNA Data Bank of Japan (DDBJ) (3) as a partner in the International Nucleotide Sequence Database Collaboration (INSDC) (4). The INSDC partners exchange data daily to ensure that a uniform and comprehensive collection of sequence information is available worldwide. NCBI makes GenBank data available at no cost over the Internet, through FTP and a wide range of web-based retrieval and analysis services (5).

RECENT DEVELOPMENTS

Reorganized genome FTP site

In late 2014 NCBI began releasing a major revision to the genome area of the NCBI FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/). This initial release provides a standard set of data files for over 45 000 genomic assemblies, and these are found in three new directories within the genome area: *genbank*, *refseq* and *all*. The *genbank* directory contains data submitted directly to GenBank (or an INSDC database), while the *refseq* directory contains data that are part of the RefSeq project (6). The common data unit within these three directories is a subdirectory corresponding to a record in the Assembly database (5) and given a name consisting of the Assembly accession followed by the name of the assembly. For example, the human GRCh38 release has assembly accession GCF_000001405.26, and so the subdirectory is named GCF_000001405.26_GRCh38. Users are encouraged to search the Assembly database (http://www.ncbi.nlm.nih.gov/assembly/) directly to find these accessions, assembly names and other details about the data sets. The *genbank* and *refseq* directories collect the Assembly subdirectories within broad taxonomic directories (e.g. plant, bacteria, vertebrate_mammalian) and also directories for each species. Each Assembly subdirectory contains a standard set of files including FASTA and GenBank/GenPept data for genome, transcript and protein sequences, along with GFF3 files for annotated genome records. These new directories will co-exist with the older genome FTP data until 1 March 2015, when the older data files will be removed.

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

Identical protein reports

In 2013 NCBI introduced the non-redundant WP protein sequences in response to the anticipated rapid growth in the submission of highly redundant prokaryotic genome sequences from clinical samples (7). Such redundant genomes will result in large numbers of identical protein annotations, and each set of identical proteins will be represented by a single WP sequence. Two consequences of this are (i) that these individual, identical protein annotations will not have separate records or GI numbers at NCBI and (ii) that WP records will typically link to not one but a corresponding set of Nucleotide CDS sequences. To clarify these relationships, the Protein database now provides a new protein record format called an 'Identical Protein Report'. These reports are available from the 'Display Settings' menu and include a table listing all protein accessions identical to the given record along with links to the Nucleotide CDS for each sequence. The report is also available through the E-utility EFetch with *&rettype = ipg* (http://eutils.ncbi.nlm.nih.gov).

Updates to the submission portal and related tools

NCBI continues to provide additional submission tools as part of a unified submission portal that will ultimately provide a single access point for GenBank submitters (http://submit.ncbi.nlm.nih.gov). For example, a new wizard assists submission of 16S rRNA sequences from uncultured bacteria. For WGS submissions, a new wizard now accepts FASTA data as input. Additional wizards and related tools are being planned, and submitters should continue to monitor the Submission Portal and GenBank release notes for updates.

Changes to indexing of microbial strains

Last year we indicated that GenBank would no longer assign taxonomy IDs at the strain level for microbes, except for informal strain-specific names for genomes from specimens that have not been identified to the species level, e.g. '*Rhizobium* sp. CCGE 510'. This change went into effect in February 2014 (8). Bacterial and other microbial strains that already have a unique taxonomy ID will retain them, but now any submitted genome for a new strain will not be given a unique taxonomy ID; rather, GenBank will assign the taxonomy ID of the species to these data. Because these new strains will also be entered in the BioSample database as unique records, BioSample is now the source of unique NCBI identifiers for individual bacteria strains and isolates.

ORGANIZATION OF THE DATABASE

GenBank divisions

GenBank assigns sequence records to various divisions based either on the source taxonomy or the sequencing strategy used to obtain the data. There are 12 taxonomic divisions (BCT, ENV, INV, MAM, PHG, PLN, PRI, ROD, SYN, UNA, VRL, VRT) and 5 high-throughput divisions (EST, GSS, high-throughput cDNA (HTC), high-throughput genomic (HTG), STS). Finally, the PAT division contains records supplied by patent offices, the transcriptome shotgun assembly (TSA) division contains sequences from TSA projects and the WGS division contains sequences from WGS projects. The size and growth of these divisions, and of GenBank as a whole, are shown in Table 1.

Sequence-based taxonomy

Database sequences are classified and can be queried using a comprehensive sequence-based taxonomy (http://www.ncbi.nlm.nih.gov/taxonomy/) developed by NCBI in collaboration with ENA and DDBJ and with the valuable assistance of external advisers and curators (9). Over 300 000 formally described species are represented in GenBank, and the top species in the non-WGS GenBank divisions are listed in Table 2.

Sequence identifiers and accession numbers

Each GenBank record, consisting of both a sequence and its annotations, is assigned a unique identifier called an accession number that is shared across the three collaborating databases (GenBank, DDBJ, ENA). The accession number appears on the ACCESSION line of a GenBank record and remains constant over the lifetime of the record, even when there is a change to the sequence or annotation. Changes to the sequence data itself are tracked by an integer extension of the accession number, and this *Accession.version* identifier appears on the VERSION line of the GenBank flat file. The initial version of a sequence has the extension '.1'. In addition, each version of the DNA sequence is also assigned a unique NCBI identifier called a *GI* number that also appears on the VERSION line following the *Accession.version*:

```
ACCESSION AF000001
VERSION AF000001.5 GI: 7274584
```

When a change is made to a sequence in a GenBank record, a new GI number is issued to the updated sequence and the version extension of the *Accession.version* identifier is incremented. The accession number for the record as a whole remains unchanged, and will always retrieve the most recent version of the record; the older versions remain available under the old *Accession.version* identifiers and their original GI numbers. The Revision History report, available from the 'Display Settings' menu on the sequence record view, summarizes the various updates for that GenBank record.

A similar system tracks changes in the corresponding protein translations. These identifiers appear as qualifiers for CDS features in the FEATURES portion of a GenBank entry, e.g. /protein_id = 'AAF14809.1'. Protein sequence translations also receive their own unique GI number, which appears as a second qualifier on the CDS feature:

```
/db_xref = 'GI : 6513858'
```

Citing GenBank records

Besides being the primary identifier of a GenBank sequence record, GenBank accessions are also the most efficient and reliable way to cite a sequence record in publications. We

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

Division	Description	Release 203 (8/2014)	Annual increase (%)[a]
WGS	Whole genome shotgun data	774 052 098 731	54.7%
PLN	Plants	9 012 205 825	51.1%
UNA	Unannotated	187 345	43.5%
BCT	Bacteria	13 722 041 634	33.5%
PHG	Phages	146 804 958	22.5%
VRL	Viruses	2 125 907 663	21.0%
ENV	Environmental samples	4 297 282 924	14.8%
INV	Invertebrates	3 085 847 038	12.7%
PAT	Patented sequences	14 647 872 659	10.2%
TSA	Transcriptome shotgun data	9 323 352 861	8.0%
MAM	Other mammals	951 689 720	4.4%
PRI	Primates	6 697 769 597	4.2%
VRT	Other vertebrates	3 188 004 508	3.9%
SYN	Synthetic	976 696 717	3.8%
HTC	High-throughput cDNA	671 972 485	2.4%
GSS	Genome survey sequences	24 293 870 378	2.4%
EST	Expressed sequence tags	42 086 482 490	1.0%
HTG	High-throughput genomic	25 386 830 568	0.8%
STS	Sequence tagged sites	640 701 468	0.7%
ROD	Rodents	4 467 459 537	0.4%
TOTAL	All GenBank sequences	939 775 079 106	43.6%

[a]Measured relative to Release 197 (8/2013).

Table 2. Top organisms in GenBank (Release 203)

Organism	Non-WGS base pairs
Homo sapiens	17 575 474 103
Mus musculus	9 993 232 725
Rattus norvegicus	6 525 559 108
Bos taurus	5 391 699 711
Zea mays	5 079 812 801
Sus scrofa	4 894 315 374
Danio rerio	3 128 000 237
Triticum aestivum	1 925 428 081
Solanum lycopersicum	1 764 995 265
Hordeum vulgare	1 617 554 059
Strongylocentrotus purpuratus	1 435 261 003
Macaca mulatta	1 297 237 624
Oryza sativa Japonica Group	1 265 215 013
Xenopus tropicalis	1 249 788 384
Nicotiana tabacum	1 200 025 462
Arabidopsis thaliana	1 165 816 533
Drosophila melanogaster	1 155 228 906
Vitis vinifera	1 071 458 039
Glycine max	1 020 646 789
Pan troglodytes	1 010 316 029

certainly encourage submitters and other authors to cite GenBank data using these accessions. However, as discussed above, since searching with a GenBank accession number will retrieve the most recent version of a record, the data returned from such searches will change over time if the record is updated. It is quite possible, therefore, for the sequence data retrieved today by an accession to be different from that discussed or analyzed in a paper published several years ago. We therefore recommend that authors include the version suffix when citing a GenBank accession (e.g. AF000001.5), so that future readers can easily retrieve the data in question.

BUILDING THE DATABASE

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

Direct electronic submission

Virtually all records enter GenBank as direct electronic submissions (http://www.ncbi.nlm.nih.gov/genbank/), 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 sequence 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 ∼3500 per day. The accession number serves as confirmation that the sequence has been submitted and provides a means for 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 the 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 submitter is permitted to modify sequence data or annotations, all users are encouraged to report lags in releas-

ing data or possible errors or omissions to GenBank at 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 http://www.ncbi.nlm.nih.gov/genbank/tbl2asn2.html. Submitters can keep abreast of updates to *tbl2asn* and Sequin by subscribing to the NCBI submissions RSS feed (http://www.ncbi.nlm.nih.gov/feed/rss.cgi?ChanKey=genbanksubmissiontoo).

Submission using BankIt. About a third of author submissions are received through an NCBI web-based data submission tool named BankIt. Using BankIt, authors enter sequence information and biological annotations, such as coding regions or mRNA features, directly into a series of tabbed forms that allow the submitter to describe the sequence further without having to learn formatting rules or controlled vocabularies. Additionally, BankIt allows submitters to upload source and annotation data using tab-delimited tables. Before creating a draft record in the GenBank flat file format for the submitter to review, BankIt validates the submissions by flagging many common errors and checking for vector contamination using a variant of Basic Local Alignment Search Tool (BLAST) called Vecscreen.

Submission using Sequin and tbl2asn. NCBI also offers a standalone multiplatform submission program called Sequin (http://www.ncbi.nlm.nih.gov/projects/Sequin/) that can be used interactively with other NCBI sequence retrieval and analysis tools. Sequin handles simple sequences (such as a single cDNA), phylogenetic studies, population studies, mutation studies, environmental samples with or without alignments and sequences with complex annotation. Sequin is available for Macintosh, PC and Unix computers by anonymous FTP from ftp://ftp.ncbi.nlm.nih.gov/sequin. Once a submission is completed, submitters can e-mail the Sequin file to gb-sub@ncbi.nlm.nih.gov or upload the Sequin file to http://www.ncbi.nlm.nih.gov/LargeDirSubs/dir_submit.cgi. Submitters of large, heavily annotated genomes may find it convenient to use the command line tool *tbl2asn* to convert a table of annotations generated from an annotation pipeline into an ASN.1 (Abstract Syntax Notation One) record suitable for submission to GenBank. These files for WGS genome and TSA submissions are then transmitted to GenBank through the Submission Portal.

Notes on particular divisions

Environmental sample sequences (ENV). The ENV division of GenBank accommodates sequences obtained via environmental sampling methods in which the source organism is unknown. Many ENV sequences arise from metagenome samples derived from microbiota in various animal tissues, such as within the gut or skin, or from particular environments, such as freshwater sediment, hot springs or areas of mine drainage. Records in the ENV division contain 'ENV' in the keyword field and use an '/environmental_sample' qualifier in the source feature. Environmental sample sequences are generally submitted for whole metagenomic shotgun sequencing experiments or surveys of sequences from targeted genes, like 16S rRNA. NCBI continues to support BLAST searches (see below) of metagenomic ENV sequences, but sequences within WGS projects are now part of the WGS BLAST database.

WGS sequences. WGS sequences appear in GenBank as groups of sequence-overlap contigs collected under a master WGS record. Each master record represents a WGS project and has an accession number in the Nucleotide database consisting of a four-letter prefix followed by eight zeroes and a version suffix as found in standard GenBank records. The number of zeroes increases to nine for WGS projects with 1 million or more contigs. Master records contain no sequence data; rather, links appear at the bottom of these records that provide displays of individual contigs in the WGS browser. Contig records have accessions consisting of the same four-letter prefix as their master accession, followed by a two-digit version number and a six-digit contig ID. For example, the WGS accession number 'AAAA02002744' is assigned to contig number '002744' of the second version of project 'AAAA', whose accession number is 'AAAA00000000.2'. Currently, there are 27 000 WGS sequencing projects, many of whose data have been used to build over 22 million scaffolds and chromosomes for genome assemblies. For a complete list of WGS projects with links to the data, see http://www.ncbi.nlm.nih.gov/Traces/wgs/.

Although WGS project sequences may be annotated, many low-coverage genome projects do not contain annotation. Because these sequence projects are ongoing and incomplete, these annotations may not be tracked from one assembly version to the next and should be considered preliminary. Submitters of genomic sequences, including WGS sequences, are urged to use evidence tags of the form '/experimental = *text*' and '/inference = *TYPE*:*text*', where *TYPE* is one of a number of standard inference types and *text* consists of structured text. Annotation is no longer required for complete genomes, although submitters can request that the genome be annotated by NCBI's Prokaryotic Genome Annotation Pipeline (http://www.ncbi.nlm.nih.gov/genome/annotation_prok/) before being released.

TSA sequences. The TSA division contains TSA sequences that are assembled from sequences deposited in the NCBI Trace Archive, the Sequence Read Archive (SRA) and the EST division of GenBank. While neither the Trace Archive nor SRA is a part of GenBank, they are part of the INSDC and provide access to the data underlying these assemblies (5,10). TSA records have 'TSA' as their keyword and can be retrieved with the query 'tsa[properties]'.

HTG and HTC sequences. The HTG division of GenBank (http://www.ncbi.nlm.nih.gov/genbank/htgs/) contains unfinished large-scale genomic records, which are in transition to a finished state (11). These records are designated as belonging to Phases 0–3 depending on the quality of the data, with Phase 3 being the finished state. Upon reaching Phase 3, HTG records are moved into the appropriate organism division of GenBank.

The HTC division of GenBank contains HTC sequences that are of draft quality but may contain 5′ UTRs, 3′ UTRs, partial coding regions and introns. HTC sequences which are finished and of high quality are moved to the appropriate organism division of GenBank. A project generating HTC data is described in (12).

Special record types

Third-party annotation (TPA). TPA records are sequence annotations published by someone other than the original submitter of the primary sequence record in DDBJ/ENA/GenBank (http://www.ncbi.nlm.nih.gov/genbank/TPA). Each of the 245 000 TPA records falls into one of three categories: *experimental*, in which case there is direct experimental evidence for the existence of the annotated molecule; *inferential*, in which case the experimental evidence is indirect; and *assembly*, where the focus is on providing a better assembly of the raw reads. 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_exp:', 'TPA_inf:' or 'TPA_asm:' at the beginning of each Definition Line as well as corresponding keywords. TPA experimental and inferential records also contain a Primary block that provides the base ranges and identifier for the sequences used to build the TPA. TPA sequences are not released to the public until their accession numbers or sequence data and annotation appear in a peer-reviewed biological journal. TPA submissions to GenBank may be made using either BankIt or Sequin.

Contig (CON) records for assemblies of smaller records. Within GenBank, CON records are used to represent very long sequences, such as a eukaryotic chromosome, where the sequence is not complete but consists of several contig records with uncharacterized gaps between them. Rather than listing the sequence itself, CON records contain assembly instructions involving the several component sequences. An example of such a CON record is CM000663 for human chromosome 1.

RETRIEVING GENBANK DATA

The Entrez system

The sequence records in GenBank are accessible through the NCBI Entrez retrieval system (5). Records from the EST and GSS divisions of GenBank are stored in the EST and GSS databases, while all other GenBank records are stored in the Nucleotide database. GenBank sequences that are part of population or phylogenetic studies are also collected together in the PopSet database, and conceptual translations of CDS sequences annotated on GenBank records are available in the Protein database. Each of these databases is linked to the scientific literature in PubMed and PubMed Central. Additional information about conducting Entrez searches is found in the NCBI Help Manual (http://www.ncbi.nlm.nih.gov/books/NBK3831/) and links to related tutorials are provided on the NCBI Education page (http://www.ncbi.nlm.nih.gov/education/).

Associating sequence records with sequencing projects

The ability to identify all GenBank records submitted by a specific group or those with a particular focus, such as metagenomic surveys, is essential for the analysis of large volumes of sequence data. The use of organism or submitter names as a means to define such a set of sequences is unreliable. The BioProject database (http://www.ncbi.nlm.nih.gov/bioproject), developed at NCBI and subsequently adopted across the INSDC, allows submitters to register large-scale sequencing projects under a unique project identifier, enabling reliable linkage between sequencing projects and the data they produce. BioProject includes pointers to data from a wide variety of projects deposited in any NCBI primary data archive. Sequencing projects focus on genomes, metagenomes, transcriptomes, comparative genomics as well as on particular loci, such as 16S ribosomal RNA. A 'DBLINK' line appearing in GenBank flat files identifies the sequencing projects with which a GenBank sequence record is associated. In addition, sequence records may have a link to the BioSample database (13) that provides additional information about the biological materials used in the study that produced the sequence data. Such studies include genome-wide association studies, high-throughput sequencing, microarrays and epigenomic analyses. As an example, the TSA project GAAA contains DBLINK lines that associate the GenBank sequence record with BioProject record PRJNA77699 and BioSample record SRS283232, as well as the SRA record containing the raw data, SRR401852:

```
BioProject: PRJNA77699
BioSample: SRS283232
Sequence Read Archive: SRR401852
```

BLAST sequence-similarity searching

Sequence-similarity searches are the most fundamental and frequent type of analysis performed on GenBank data. NCBI offers the BLAST family of programs (http://blast.ncbi.nlm.nih.gov) to detect similarities between a query sequence and database sequences (14,15). BLAST searches may be performed on the NCBI web site (16) or by using a set of standalone programs distributed by FTP (5).

Obtaining GenBank by FTP

NCBI distributes GenBank releases in the traditional flat file format as well as in the ASN.1 format used for internal maintenance. The full bimonthly GenBank release along with the daily updates, which incorporate sequence data from ENA and DDBJ, is available by anonymous FTP from NCBI at ftp://ftp.ncbi.nlm.nih.gov/genbank. The full release in flat file format is available as a set of compressed files with a non-cumulative set of updates at ftp://ftp.ncbi.nlm.nih.gov/genbank/daily-nc/. For convenience in file transfer, the data are partitioned into multiple files; for release 203 there are 2093 files requiring 653 GB of uncompressed disk storage. A script is provided in ftp://ftp.ncbi.nlm.nih.gov/genbank/tools/ to convert a set of daily updates into a cumulative update.

MAILING ADDRESS

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

ELECTRONIC ADDRESSES

http://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

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

REFERENCES

1. Benson,D.A., Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2014) GenBank. *Nucleic Acids Res.*, **42**, D32–D37.
2. Cochrane,G., Alako,B., Amid,C., Bower,L., Cerdeno-Tarraga,A., Cleland,I., Gibson,R., Goodgame,N., Jang,M., Kay,S. *et al.* (2013) Facing growth in the European Nucleotide Archive. *Nucleic Acids Res.*, **41**, D30–D35.
3. Kosuge,T., Mashima,J., Kodama,Y., Fujisawa,T., Kaminuma,E., Ogasawara,O., Okubo,K., Takagi,T. and Nakamura,Y. (2014) DDBJ progress report: a new submission system for leading to a correct annotation. *Nucleic Acids Res.*, **42**, D44–D49.
4. Nakamura,Y., Cochrane,G., Karsch-Mizrachi,I. and International Nucleotide Sequence Database, C. (2013) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.*, **41**, D21–D24.
5. NCBI Resource Coordinators. (2015) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res., doi:10.1093/nar/gku1130.
6. Pruitt,K.D., Tatusova,T., Brown,G.R. and Maglott,D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
7. NCBI Resource Coordinators. (2014) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **42**, D7–D17.
8. Federhen,S., Clark,K., Barrett,T., Parkinson,H., Ostell,J., Kodama,Y., Mashima,J., Nakamura,Y., Cochrane,G. and Karsch-Mizrachi,I. (2014) Toward richer metadata for microbial sequences: replacing strain-level NCBI taxonomy taxids with BioProject, BioSample and Assembly records. *Stand. Genom. Sci.*, **9**, 1275–1277.
9. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
10. Kodama,Y., Shumway,M. and Leinonen,R. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
11. Kans,J.A. and Ouellette,B.F.F. (2001) Submitting DNA sequences to the databases. In: Baxevanis,AD and Ouellette,BFF (eds). *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. John Wiley and Sons, Inc., NY, pp. 65–81.
12. Kawai,J., Shinagawa,A., Shibata,K., Yoshino,M., Itoh,M., Ishii,Y., Arakawa,T., Hara,A., Fukunishi,Y., Konno,H. *et al.* (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**, 685–690.
13. 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.
14. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., 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.
15. 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**, 3986–3990.
16. 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.