Published online 20 November 2015 *Nucleic Acids Research, 2016, Vol. 44, Database issue* ***D67–D72***
doi: 10.1093/nar/gkv1276

GenBank

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 September 18, 2015; Revised October 29, 2015; Accepted November 03, 2015

ABSTRACT

GenBank® (www.ncbi.nlm.nih.gov/genbank/) is a comprehensive database that contains publicly available nucleotide sequences for over 340 000 formally described species. Recent developments include a new starting page for submitters, a shift toward using accession.version identifiers rather than GI numbers, a wizard for submitting 16S rRNA sequences, and an Identical Protein Report to address growing issues of data redundancy. GenBank organizes the sequence data received from individual laboratories and large-scale sequencing projects into 18 divisions, and GenBank staff assign unique accession.version identifiers upon data receipt. Most submitters use the web-based BankIt or standalone Sequin programs. Daily data exchange with the European Nucleotide Archive (ENA) and the DNA Data Bank of Japan (DDBJ) ensures worldwide coverage. GenBank is accessible through the nuccore, nucest, and nucgss databases of the Entrez retrieval system, which integrates these records with a variety of other data including taxonomy nodes, genomes, protein structures, and biomedical journal literature in 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 (NLM), located on the campus of the US 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 through a wide range of Web-based retrieval and analysis services (5).

RECENT DEVELOPMENTS

New submit page

In 2014, NCBI released a revised version of the main NCBI home page that includes six new buttons, one of which is labeled 'Submit'. This button leads to a new Submit page that serves as a unified starting point for any type of data submission to NCBI. If a user is logged in to NCBI, the main banner of the Submit page links to that user's submissions, making it easy to track ongoing submissions or begin new ones. The page provides a QuickStart menu that navigates users to the various submission sites, along with a 'wizard' that allows users to browse all submission resources. Future versions of this wizard will include a questionnaire to guide users to the most appropriate submission site.

Upcoming changes to sequence identifiers

As first described in the release notes for GenBank 199.0 in December 2013, and discussed in more detail in the release notes for GenBank 209.0, NCBI is in the process of phasing out the use of GI numbers as sequence identifiers. GI numbers were first introduced in GenBank 81.0 (February 1994) as an additional identifier to the accession number that would stably refer to a particular version of a sequence record. In 1997, such version tracking was added to accession numbers in the form of an integer suffix that would increment with each update to the sequence data within a record. For example, AC020606.7 refers to a record that has had its sequence data updated six times. In this way both the GI and the accession.version identifier uniquely refer to

*To whom correspondence should be addressed. Tel: +1 301 496 2475; Fax: +1 301 480 9241; Email: sayers@ncbi.nlm.nih.gov a given version of a sequence record, and both identifiers have been included in GenBank records for years to support both approaches. Given the rapidly increasing number of data submissions, it has become clear that it is now time for us to take the next step and remove the older, redundant GI identifiers and retain a single identifier for sequences, the more human-readable accession.version. This change will simplify the process of tracking sequences without any loss of functionality. Therefore, over the coming months we will no longer assign GI's to a gradually growing number of new sequences. (Current examples of such sequences are unannotated contigs in WGS and TSA projects.) Sequence records with existing GI's will retain them, and NCBI services that accept GI's as input will continue to be supported. NCBI will be adding support for accession.version identifiers to all services that currently do not support them. As NCBI makes this transition, we encourage any users who have workflows that depend on GI's to begin planning to use accession.version identifiers instead. As this process unfolds, NCBI will provide additional announcements on our social media platforms and news feeds, as well as in GenBank release notes.

16S rRNA submission wizard

The NCBI submission portal now offers a new wizard to assist submitters of 16S rRNA sequences from microbes (submit.ncbi.nlm.nih.gov/genbank/help/). This wizard is intended for bacterial or archaeal samples that are either from uncultured, environmental sources, or from pure cultured strains. Samples should be only 16S rRNA sequences and should not be raw reads from next-generation technologies. Sequences submitted using the wizard will be automatically processed and checked for chimeras, vector contamination, low quality sequence, and other problems.

Unverified sequences

As reported previously (6), as part of the standard review process for new submissions, GenBank staff may label sequences as unverified if the accuracy of the submitted sequence data or annotations cannot be confirmed. Until the submitter is able to resolve these problems, the definition line of the sequence will begin with 'UNVERIFIED': and the sequence will not be included in BLAST databases. This treatment is being extended to genomic submissions where the source organism is uncertain, there is evidence of contamination, or there are other problems with the data. In addition to the UNVERIFIED label in the definition line, a short description of the problems will be entered in the COMMENT field of the record.

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 twelve taxonomic divisions (BCT, ENV, INV, MAM, PHG, PLN, PRI, ROD, SYN, UNA, VRL, VRT) and six high-throughput divisions (EST, GSS, HTC, HTG, STS, TSA). Finally, the PAT division contains records supplied by patent offices and the WGS division contains sequences from whole genome shotgun 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 (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 (7). Over 340 000 formally described species are represented in GenBank, and the top species in the non-WGS GenBank divisions are listed in Table 2.

Sequence identifiers

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'. When a change is made to a sequence in a GenBank record, the version extension of the *Accession.version* identifier is incremented. The accession portion of the identifier remains unchanged and will always retrieve the most recent version of the record; the older versions remain available under the old *Accession.version* identifiers. The Revision History report, available from the 'Display Settings' menu on the sequence record view, summarizes the various updates for that GenBank record, including non-sequence changes that do not result in the version suffix being incremented. 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'.

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 (8). 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. These individual, identical protein annotations will not have separate records at NCBI, and so a WP record may link to not one but a corresponding set of Nucleotide CDS sequences. To clarify these relationships, the Protein database provides a record format called an 'Identical Protein Report.' These reports are available from the top of a protein record page 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

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

Division	Description	Release 209 (8/2015)	Annual Increase (%)[a]
INV	Invertebrates	15 413 731 414	399.5%
MAM	Other mammals	3 592 838 191	277.5%
VRT	Other vertebrates	6 643 601 831	108.4%
WGS	Whole genome shotgun data	1 163 275 601 001	50.3%
PHG	Phages	210 143 517	43.1%
BCT	Bacteria	19 331 233 520	40.9%
PLN	Plants	11 966 142 676	32.8%
TSA	Transcriptome shotgun data	11 171 215 516	19.8%
VRL	Viruses	2 493 936 092	17.3%
ENV	Environmental samples	4 845 868 034	12.8%
HTG	High-throughput genomic	27 057 268 218	6.6%
PAT	Patented sequences	15 549 880 984	6.2%
GSS	Genome survey sequences	25 607 093 540	5.4%
SYN	Synthetic	1 001 954 270	2.6%
PRI	Primates	6 808 335 498	1.7%
EST	Expressed sequence tags	42 333 093 845	0.6%
ROD	Rodents	4 482 375 973	0.3%
HTC	High-throughput cDNA	673 910 306	0.3%
UNA	Unannotated	187 511	0.1%
STS	Sequence tagged sites	640 833 351	0.0%
TOTAL	All GenBank sequences	1 363 099 245 288	45.0%

[a] Measured relative to Release 203 (8/2014).

Table 2. Top organisms in GenBank (release 203)

Organism	Non-WGS base pairs
Homo sapiens	17 791 718 636
Mus musculus	10 004 995 614
Rattus norvegicus	6 526 314 722
Bos taurus	5 412 338 175
Zea mays	5 203 408 728
Sus scrofa	4 895 555 549
Hordeum vulgare	3 229 866 896
Danio rerio	3 151 064 646
Ovis canadensis	2 590 569 059
Triticum aestivum	1 937 727 565
Cyprinus carpio	1 835 902 375
Solanum lycopersicum	1 744 606 771
Apteryx australis	1 595 383 668
Strongylocentrotus purpuratus	1 435 471 103
Macaca mulatta	1 297 900 273
Oryza sativa Japonica Group	1 267 676 263
Spirometra erinaceieuropaei	1 264 189 828
Xenopus tropicalis	1 249 270 365
Arabidopsis thaliana	1 202 220 229
Nicotiana tabacum	1 200 826 354

is also available through the E-utility EFetch with *&rettype = ipg* (eutils.ncbi.nlm.nih.gov).

Citing GenBank records

Besides being the primary identifier of a GenBank sequence record, GenBank accession.version identifiers are also the most efficient and reliable way to cite a sequence record in publications. Because searching with a GenBank accession number (without the version suffix) 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 encourage submitters and other authors to include the version suffix when citing a GenBank accession (e.g. AF000001.5), since this ensures that the citation refers to a specific version in time.

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 WGS, TSA, HTG, EST, or GSS 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 (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 approximately 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 releasing 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 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 (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. Using BankIt, submitters can submit sets of sequences as well as single sequences. 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 BLAST called Vecscreen.

Submission using Sequin, tbl2asn, and the submission portal. NCBI also offers a standalone multi-platform submission program called Sequin (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 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 (submit.ncbi.nlm.nih.gov). Alternatively, the Submission Portal provides an interface that accepts WGS data in FASTA format using a set of online forms.

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.

Whole genome shotgun sequences. Whole Genome Shotgun (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 4-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 one 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 4-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'. For a complete list of the more than 30 000 WGS projects, along with links to the data, see 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, but we encourage submitters to request that the genome be annotated by NCBI's Prokaryotic Genome Annotation Pipeline (www.ncbi.nlm.nih.gov/genome/annotation_prok/) before being released.

Transcriptome shotgun assembly (TSA) sequences. The TSA division contains transcriptome shotgun assembly 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 (5,9). TSA records have 'TSA' as their keyword and can be retrieved with the query 'tsa[properties]'.

Special record types

Third party annotation. Third Party Annotation (TPA) records are sequence annotations published by someone other than the original submitter of the primary sequence record in DDBJ/ENA/GenBank (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 (www.ncbi.nlm.nih.gov/books/NBK3831/) and links to related tutorials are provided on the NCBI Learn page (www.ncbi.nlm.nih.gov/home/learn.shtml).

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 (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, and 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 (10) 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 GBJS contains DBLINK lines that associate the GenBank sequence record with BioProject record PRJNA255770 and BioSample record SAMN02928618 as well as the two SRA records containing the raw data, SRR1522120 and SRR1522122:

BioProject: PRJNA255770
BioSample: SAMN02928618
Sequence Read Archive: SRR1522120, SRR1522122

In addition to the DBLINK lines for BioProject and BioSample, GenBank records that represent genome assemblies will also have a link on the right side of the page to a corresponding record in the Assembly database (11). Assembly records not only collect metadata and statistics for these genome assemblies, but also provide a stable accession for the assembly along with a link to the FTP directory containing the sequence data for the assembly in GenBank, FASTA, and GFF3 formats.

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 (blast.ncbi.nlm.nih.gov) to detect similarities between a query sequence and database sequences (12,13). BLAST searches may be performed on the NCBI Web site (14) 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 209 there are 2451 files requiring 735 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

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 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. (2015) GenBank. *Nucleic Acids Res.*, **43**, D30–D35.
2. Silvester,N., Alako,B., Amid,C., Cerdeno-Tarraga,A., Cleland,I., Gibson,R., Goodgame,N., Ten Hoopen,P., Kay,S., Leinonen,R. *et al.* (2015) Content discovery and retrieval services at the European Nucleotide Archive. *Nucleic Acids Res.*, **43**, D23–D29.
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. (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, doi:10.1093/nar/gkv1290.
6. Benson,D.A., Karsch-Mizrachi,I., Clark,K., Lipman,D.J., Ostell,J. and Sayers,E.W. (2012) GenBank. *Nucleic Acids Res.*, **40**, D48–D53.
7. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
8. Tatusova,T., Ciufo,S., Fedorov,B., O’Neill,K. and Tolstoy,I. (2014) RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.*, **42**, D553–D559.
9. Kodama,Y., Shumway,M. and Leinonen,R. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
10. 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.
11. Kitts,P.A., Church,D.M., Choi,J., Hem,V., Smith,R., Tatusova,T., Thibaud-Nissen,F., DiCuccio,M., Murphy,T.D., Pruitt,K.D. *et al.* (2016) Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.*, doi:10.1093/nar/gkv1226.
12. 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.
13. 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.
14. 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.