Published online 28 November 2016 *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D37–D42***
doi: 10.1093/nar/gkw1070

GenBank

Dennis A. Benson, Mark Cavanaugh, 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 15, 2016; Revised October 19, 2016; Editorial Decision October 24, 2016; Accepted November 07, 2016

ABSTRACT

GenBank® (www.ncbi.nlm.nih.gov/genbank/) is a comprehensive database that contains publicly available nucleotide sequences for 370 000 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 (WGS) and environmental sampling projects. Most submissions are made using the web-based BankIt or the NCBI Submission Portal. GenBank staff assign accession numbers upon data receipt. 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 NCBI Nucleotide database, which links to related information such as taxonomy, genomes, protein sequences and 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. Recent updates include changes to policies regarding sequence identifiers, an improved 16S submission wizard, targeted loci studies, the ability to submit methylation and BioNano mapping files, and a database of anti-microbial resistance genes.

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 submissions of sequence data from authors and from bulk submissions of 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-EBI 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

Upcoming changes to sequence identifiers

As first described in the release notes for GenBank 199.0 in December 2013, and discussed in more detail previously (1), NCBI is phasing out the practice of assigning GI numbers as sequence identifiers. Since GI numbers were first introduced in GenBank 81.0 (February 1994), GenBank records have had both a GI number and an accession.version identifier. Removing the redundant GI identifiers and retaining the more human-readable accession.version simplifies the process of tracking sequences without any loss of functionality. Therefore, over the coming months we will no longer assign GI numbers to a gradually growing number of new sequences. (Current examples of such sequences are unannotated contigs in WGS and TSA projects.) In addition, GI numbers will no longer appear in the default flat file presentations and FASTA definition lines of sequence data records, whether obtained from the web, API calls, or the NCBI FTP site. Sequence records with existing GI numbers will retain them in XML and Abstract Syntax Notation One (ASN.1) formats, and NCBI services that accept GI numbers as input will continue to be supported. NCBI will continue to add support for accession.version identifiers to all services that currently do not support them. For example, the E-utilities now accept the parameter *idtype*, which when set to 'acc', allows these calls to accept accession.version as input and provide them as output. As this process unfolds, NCBI will provide additional announcements on our social

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

media platforms and news feeds, as well as in GenBank release notes.

16S rRNA submission wizard

The 16S rRNA submission wizard, part of the NCBI submission portal, now offers faster, real-time analysis to assist submitters of 16S rRNA sequences from prokaryotes (submit.ncbi.nlm.nih.gov/genbank/help/). Prokaryotic samples can be from uncultured, environmental sources or pure cultured strains. Moreover, a wizard for submitting rRNA sequences and transcribed spacers from all organisms is being developed. If samples were generated using next-generation technologies, only assembled sequences (two or more reads) will be accepted. Sequences submitted using the wizard will be automatically processed and checked for chimeras, vector contamination, low quality sequence and other problems.

Targeted locus studies (TLS)

GenBank is now accepting sequences from targeted loci studies. These studies often contain large sets of 16S rRNA sequence or ultra-conserved elements (UCEs). TLS sequences are given a 'TLS' keyword and have accessions similar to those of Whole Genome Shotgun (WGS) and Transcriptome Shotgun Assembly (TSA) sequences (see below) where the four-letter accession prefix begins with 'K'. For example, 'KAAA00000000.1' represents the master record for a TLS project that contains contigs with accessions KAAA01000001–KAAA01169849.

Bacterial average nucleotide identity (ANI)

As part of the NCBI bacterial genome submission process, GenBank now performs an average nucleotide identity analysis to investigate whether the asserted organism name may be incorrect. Using the genomes already in GenBank, this ANI analysis can report that a genome submitted as *Escherichia coli* is actually *Salmonella enterica*, for example. However, since the analysis uses the genomes already in GenBank, it cannot necessarily be performed for all new genome submissions.

DNA methylation analysis files

GenBank now accepts analysis files derived from SMRT sequencing provided by Pacific Biosciences (6). These files summarize the observed patterns of methylation and can be included as part of the genome assembly submission or as supplementary file submissions made through the NCBI Submission Portal. The most common file type submitted is the motif_summary.csv file. A complete list of the available base modification files received is provided at ftp.ncbi.nlm.nih.gov/pub/supplementary_data/basemodification.csv. The list provides for each file the Nucleotide or Sequence Read Archive (SRA) accessions for the target sequence, along with BioProject and BioSample accessions, the source organism, and the URI to the data file itself.

BioNano genome map files

GenBank now accepts whole genome maps produced by BioNano mapping technology (7,8). These files can be used in a variety of genomic analyses, including *de novo* assembly, structural variant detection, and assembly curation, and they can be submitted as supplementary file submissions through the NCBI Submission Portal. A complete list of the available BioNano map files is provided at ftp.ncbi.nlm.nih.gov/pub/supplementary_data/bionanomaps.csv. The list provides for each file the supplementary files accession number (e.g. SUPPF_0000000066), the genome accession number (if there is a public genome assembly in GenBank), the BioProject and BioSample accessions, the source organism, and the URI to the data file itself.

Anti-microbial resistance data

As part of the NCBI Pathogen Detection project, NCBI is now accepting submissions of beta-lactamase sequences as supplementary data for either WGS genome submissions or submissions of novel beta-lactamase sequences (www.ncbi.nlm.nih.gov/pathogens/submit_beta_lactamase/). Beta-lactamase antibiograms should also be submitted, and these will be linked to the BioSample record associated with the submission (www.ncbi.nlm.nih.gov/biosample/docs/beta-lactamase/).

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 five high-throughput divisions (EST, GSS, HTC, HTG, STS). In addition, the PAT division contains records supplied by patent offices, the TSA division contains sequences from transcriptome shotgun assembly projects, 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 (9,10). About 370 000 formally described species are represented in GenBank, and the top species (not including those in the WGS and TSA 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

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

Division	Description	Release 215 (August 2016)	Annual Increase (%)*
TSA	Transcriptome shotgun data	103 399 724 586	49.1%
WGS	Whole genome shotgun data	1 637 224 970 324	40.7%
BCT	Bacteria	26 474 028 571	36.9%
PHG	Phages	270 541 687	28.7%
PLN	Plants	14 705 679 094	22.9%
VRL	Viruses	2 973 938 989	19.2%
PRI	Primates	7 802 428 126	14.6%
PAT	Patent sequences	17 128 458 325	10.2%
UNA	Unannotated	204 984	9.3%
ENV	Environmental samples	5 218 628 157	7.7%
INV	Invertebrates	16 241 123 317	5.4%
SYN	Synthetic	1 045 567 653	4.4%
VRT	Other vertebrates	6 917 600 814	4.1%
HTG	High-throughput genomic	27 630 729 177	2.1%
MAM	Other mammals	3 647 546 848	1.5%
HTC	High-throughput cDNA	682 400 482	1.3%
ROD	Rodents	4 502 193 236	0.4%
EST	Expressed sequence tags	42 516 725 239	0.4%
GSS	Genome survey sequences	25 696 517 526	0.3%
STS	Sequence tagged sites	640 833 351	0.0%
TOTAL	All GenBank sequences	1 944 719 840 486	36.8%

*Measured relative to Release 209 (8/2015).

Table 2. Top organisms in GenBank (release 215)

Organism	Base pairs*
Homo sapiens	18 313 373 647
Mus musculus	10 031 175 251
Rattus norvegicus	6 528 259 145
Bos taurus	5 414 550 206
Zea mays	5 207 478 336
Sus scrofa	4 896 632 524
Hordeum vulgare	3 235 262 275
Danio rerio	3 183 146 925
Ovis canadensis	2 590 574 434
Triticum aestivum	1 941 609 064
Cyprinus carpio	1 836 265 727
Solanum lycopersicum	1 745 709 667
Oryza sativa Japonica Group	1 641 275 254
Apteryx australis	1 595 400 865
Strongylocentrotus purpuratus	1 436 120 930
Macaca mulatta	1 335 500 855
Spirometra erinaceieuropaei	1 264 190 782
Xenopus tropicalis	1 250 099 171
Arabidopsis thaliana	1 204 613 994
Nicotiana tabacum	1 202 887 576

*Excludes sequences from chloroplasts, mitochondria, metagenomes, uncultured organisms, WGS and TSA.

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 suffix of the accession number, and this *Accession.version* identifier appears on the VERSION line of the GenBank flat file. Beginning with an initial version of '.1', each change to the sequence data causes the version suffix to increment. 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 default record view in the Nucleotide database (www.ncbi.nlm.nih.gov/nuccore/), summarizes the various updates for a given record, including non-sequence changes. A similar system tracks changes in the corresponding protein translations in the Protein database (www.ncbi.nlm.nih.gov/protein/). 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 (11). The individual, identical protein annotations represented by a WP sequence do not

have separate records at NCBI, and so many WP records link to a set of Nucleotide CDS sequences. To clarify these relationships, the Protein database provides a record format called an 'Identical Protein Report' linked from the top of a protein record page. These reports list all protein annotations identical to the given record along with the Nucleotide CDS for each sequence. The report is also available through the E-utility EFetch with *&rettype = ipg* (eutils.ncbi.nlm.nih.gov).

Unverified sequences

As reported previously (12), 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.

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. Therefore, sequence data retrieved today by an accession may 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 BankIt or the NCBI Submission Portal (submit.ncbi.nlm.nih.gov). 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 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* 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 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 tbl2asn and the submission portal. 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 record suitable for submission to GenBank. These files for WGS genome and TSA submissions are then transmitted to GenBank through the Submission Portal. Alternatively, the Submission Portal provides an interface that accepts WGS and TSA data in FASTA format using a set of online forms. Now *tbl2asn* also accepts data in the GFF3 format.

Notes on particular divisions

Environmental sample sequences (ENV). The ENV division of GenBank accommodates sequences obtained us-

ing environmental sampling methods in which the sequence is derived directly from the isolate. Records in the ENV division contain 'ENV' keywords 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 a Nucleotide accession number 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 one million or more contigs. Master records contain no sequence data; rather, they include links to displays of the 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'. The complete list of WGS projects is available at www.ncbi.nlm.nih.gov/Traces/wgs/.

Many WGS project sequences do not contain annotation, and those that do may not have these annotations tracked from one assembly version to the next, and so 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 raw sequence reads deposited in the Sequence Read Archive (SRA). While SRA is not part of GenBank, it is part of the INSDC and provides access to the data underlying these assemblies (13). TSA records have 'TSA' as their keyword and can be retrieved with the query 'tsa[properties]'.

RETRIEVING GENBANK DATA

The Entrez system

The sequence records in GenBank are accessible through the NCBI Entrez retrieval system (14). 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 BioProject database (www.ncbi.nlm.nih.gov/bioproject) allows submitters to register large-scale sequencing projects under a unique project identifier, enabling reliable linkage between sequencing projects and the data they produce. A 'DBLINK' line appearing in GenBank flat files identifies the sequencing projects associated with a GenBank sequence record. In addition, sequence records may have a link to the BioSample database (15) that provides additional information about the biological materials used in the study. 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 to the corresponding record in the Assembly database (16). 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 (17,18). BLAST searches may be performed on the NCBI Web site (19) 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 215 there are 2695 files requiring 790 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. Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2016) GenBank. *Nucleic Acids Res.*, **44**, D67–D72.
2. Gibson,R., Alako,B., Amid,C., Cerdeno-Tarraga,A., Cleland,I., Goodgame,N., Ten Hoopen,P., Jayathilaka,S., Kay,S., Leinonen,R. *et al.* (2016) Biocuration of functional annotation at the European nucleotide archive. *Nucleic Acids Res.*, **44**, D58–D66.
3. Mashima,J., Kodama,Y., Kosuge,T., Fujisawa,T., Katayama,T., Nagasaki,H., Okuda,Y., Kaminuma,E., Ogasawara,O., Okubo,K. *et al.* (2016) DNA data bank of Japan (DDBJ) progress report. *Nucleic Acids Res.*, **44**, D51–D57.
4. Cochrane,G., Karsch-Mizrachi,I., Takagi,T. and International Nucleotide Sequence Database, C. (2016) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.*, **44**, D48–D50.
5. NCBI Resource Coordinators. (2016) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **44**, D7–D19.
6. Flusberg,B.A., Webster,D.R., Lee,J.H., Travers,K.J., Olivares,E.C., Clark,T.A., Korlach,J. and Turner,S.W. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods*, **7**, 461–465.
7. Das,S.K., Austin,M.D., Akana,M.C., Deshpande,P., Cao,H. and Xiao,M. (2010) Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucleic Acids Res.*, **38**, e177.
8. Stankova,H., Hastie,A.R., Chan,S., Vrana,J., Tulpova,Z., Kubalakova,M., Visendi,P., Hayashi,S., Luo,M., Batley,J. *et al.* (2016) BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. *Plant Biotechnol. J.*, **14**, 1523–1531.
9. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
10. Federhen,S. (2015) Type material in the NCBI Taxonomy Database. *Nucleic Acids Res.*, **43**, D1086–D1098.
11. NCBI Resource Coordinators. (2014) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **42**, D7–D17.
12. 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.
13. Kodama,Y., Shumway,M. and Leinonen,R. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
14. Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.
15. 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.
16. Kitts,P.A., Church,D.M., Thibaud-Nissen,F., Choi,J., Hem,V., Sapojnikov,V., Smith,R.G., Tatusova,T., Xiang,C., Zherikov,A. *et al.* (2016) Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.*, **44**, D73–D80.
17. 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.
18. 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.
19. 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.