D94–D99 *Nucleic Acids Research, 2019, Vol. 47, Database issue* *Published online 26 October 2018*
doi: 10.1093/nar/gky989

GenBank

Eric W. Sayers*, Mark Cavanaugh, Karen Clark, James Ostell, Kim D. Pruitt and Ilene Karsch-Mizrachi

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received September 18, 2018; Revised October 03, 2018; Editorial Decision October 05, 2018; Accepted October 18, 2018

ABSTRACT

GenBank® (www.ncbi.nlm.nih.gov/genbank/) is a comprehensive database that contains publicly available nucleotide sequences for 420 000 formally described species. Most GenBank submissions are made using BankIt, the NCBI Submission Portal, or the tool *tbl2asn*, and are obtained from individual laboratories and batch submissions from large-scale sequencing projects, including whole genome shotgun (WGS) and environmental sampling projects. 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 an expansion of sequence identifier formats to accommodate expected database growth, submission wizards for ribosomal RNA, and the transfer of Expressed Sequence Tag (EST) and Genome Survey Sequence (GSS) data into the Nucleotide database.

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 through the Internet, FTP, and a wide range of Web-based retrieval and analysis services (5).

RECENT DEVELOPMENTS

Database milestones

In the past year GenBank achieved two remarkable milestones. In release 224 (February 2018) GenBank grew to over three terabases in size, including 254 gigabases of traditional records, 2.61 terabases of WGS records, 208 gigabases of transcriptome shotgun assembly (TSA) records, and 4.53 gigabases of targeted locus study (TLS) records. In release 225 (April 2018) GenBank surpassed 1 billion sequence records consisting of 208 million traditional records, 621 million WGS records, 244 million TSA records, and 15 million TLS records. In base pairs, GenBank continues to grow at an annual rate of ∼40%, with the majority of new data in the WGS and TSA divisions (Table 1).

Changes to sequence identifiers

GenBank assigns to each record a unique identifier called an accession number that has a common format used by all three of the collaborating databases (GenBank, DDBJ, ENA). The accession number remains constant over the lifetime of the record. Given the rapid growth of sequence data, the INSDC recently agreed to expand the ranges of accession number formats to accommodate these new data, and plans to implement these changes by the end of 2018. We want to emphasize that none of these new accessions will replace any existing accessions, and all existing sequences will continue to be retrievable using their current acces-

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

Published by Oxford University Press on behalf of Nucleic Acids Research 2018.

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

Division	Description	Release 227 (August 2018)	Annual increase (%)[a]
MAM	Other mammals	6 214 774 850	60.47%
WGS	Whole genome shotgun data	3 204 855 013 281	42.93%
UNA	Unannotated	296 706	42.25%
PLN	Plants	23 027 832 426	37.21%
BCT	Bacteria	53 541 127 504	36.93%
TSA	Transcriptome shotgun data	225 520 004 678	35.01%
PHG	Phages	463 029 085	34.38%
VRL	Viruses	4 073 816 676	16.99%
PAT	Patent sequences	22 019 723 131	14.57%
VRT	Other vertebrates	10 441 689 546	12.90%
ENV	Environmental samples	5 818 999 756	4.09%
HTC	High-throughput cDNA	721 454 983	3.57%
PRI	Primates	8 262 441 252	2.96%
SYN	Synthetic	1 192 279 390	1.62%
GSS	Genome survey sequences	26 339 143 098	1.40%
EST	Expressed sequence tags	42 988 632 150	0.82%
HTG	High-throughput genomic	27 770 730 435	0.45%
ROD	Rodents	4 534 815 151	0.31%
STS	Sequence tagged sites	640 879 986	0.00%
INV	Invertebrates[b]	8 597 126 159	−50.09%
TOTAL	All GenBank sequences	3 677 023 810 243	39.52%

[a] Measured relative to Release 221 (August 2017).
[b] The decrease in INV data resulted from the suppression of 36 nematode-related genomes. See the release notes for Release 227 for more details (ftp.ncbi.nlm.nih.gov/genbank/release.notes/gb227.release.notes).

sion.version identifiers. As announced in the notes for GenBank release 226, these changes expand the alphabetic prefix and/or the numeric suffix of accession numbers. For traditional records (those not in the WGS, TSA or TLS divisions), the numeric suffix will grow from six to eight digits, allowing previously exhausted prefixes to be reactivated (e.g. JG00000001-JG99999999). To be clear, JG000001 (existing accession) and JG00000001 (new accession) will refer to two distinct sequences. Protein annotations, which have had accessions with three-letter prefixes and five-digit numeric suffixes, will now have seven-digit suffixes (e.g. EZZ0000001-EZZ9999999). Finally, accessions for WGS, TSA and TLS projects will have expanded prefixes and suffixes. Prefixes will expand to six letters, and suffixes to seven, eight or nine digits, depending on the size of the project (e.g. AAAAAA020000001). The version number between the prefix and suffix will continue to have two digits.

Ribosomal RNA submissions

The rRNA submission wizard, part of the NCBI submission portal, should now be used for all rRNA sequences from both prokaryotes and eukaryotes (submit.ncbi.nlm.nih.gov/subs/genbank/). Prokaryotic samples can be from uncultured, environmental sources, or pure cultured strains, and can include 16S rRNA, 23S rRNA or 16S-23S rRNA intergenic spacers. Eukaryotic samples can include both large and small subunit rRNA, nuclear rRNA-ITS regions or internal transcribed spacers, and mitochondrial or chloroplast rRNA. 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.

Ribosomal RNA TLS submissions

The NCBI submission portal now supports submissions of large scale ribosomal RNA TLS projects (www.ncbi.nlm.nih.gov/genbank/tlsguide/). This process includes registration of both BioSample and BioProject IDs, and can also accept existing BioSample/BioProject IDs. TLS projects must contain at least 2 500 sequences, and each sequence must have a unique ID within the submitted file. We encourage submitters to include contextual metadata such as the isolation source or host, the collection date, and collection site (country and latitude/longitude). These submissions will be processed by the same rRNA wizard described above, and will be subject to the same quality checks.

Changes to the EST and GSS databases

To better align our services to the needs of the bioinformatics community, we will be moving all records in the Expressed Sequence Tag (EST) and Genome Survey Sequence (GSS) databases to the Nucleotide database as of 1 December 2018. These EST and GSS records will retain their accession.version identifiers and by default will be displayed in the standard GenBank flat file format on the web. The specialized (and current default) EST and GSS record formats will no longer be available. We will add new filters to the Nucleotide database to allow users easily to exclude or specifically include EST and GSS records from search results. These changes will affect the E-utilities API in that *nucest* and *nucgss* values of the *db* parameter will be replaced with *nuccore* after 1 December. EST and GSS data will continue to be part of GenBank and will continue to be available for download at ftp.ncbi.nlm.nih.gov/genbank. In addition, BLAST will continue to support the current *est* and *gss* databases.

Table 2. Top organisms in GenBank (Release 227)

Organism	Base pairs[a]
Homo sapiens	19 752 523 722
Mus musculus	10 246 475 076
Rattus norvegicus	6 530 046 440
Bos taurus	5 431 692 037
Zea mays	5 245 788 885
Sus scrofa	5 075 446 882
Hordeum vulgare	3 237 283 130
Escherichia coli	3 220 757 391
Danio rerio	3 191 415 637
Oryzias latipes	2 836 938 628
Arachis hypogaea	2 682 391 941
Triticum aestivum	2 636 490 116
Ovis canadensis	2 590 574 434
Solanum lycopersicum	2 572 291 998
Bos mutus	2 290 216 303
Cyprinus carpio	1 836 731 087
Oryza sativa	1 727 115 789
Apteryx australis	1 595 510 956
Bordetella pertussis	1 456 386 736
Strongylocentrotus purpuratus	1 436 247 256

[a] Excludes sequences from chloroplasts, mitochondria, metagenomes, uncultured organisms, WGS, TSA, and the CON division.

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 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 (6,7). About 420 000 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.

Unverified sequences

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 (8). 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. 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 sequence version suffix when citing a GenBank record (e.g. AF000001.5), since this ensures that the citation refers to a specific version, in time, of the sequence data for that accession.

BUILDING THE DATABASE

The data in GenBank and the collaborating databases, ENA and DDBJ, are submitted by investigators to one of the three databases. Data are exchanged daily between GenBank, DDBJ and ENA so that 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 (www.ncbi.nlm.nih.gov/WebSub/?tool=genbank) or the NCBI Submission Portal (submit.ncbi.nlm.nih.gov). BankIt allows authors to enter and/or upload sequence information and biological annotations directly into a series of tabbed forms without the need to learn formatting rules or controlled vocabularies. The Submission Portal is a centralized system that supports submissions of prokaryotic and eukaryotic genomes and a variety of specialized sequence types, such as ribosomal RNA, TSA and SRA. It can also manage BioSample and BioProject submissions. GenBank also offers the command line program *tbl2asn* (www.ncbi.nlm.nih.gov/genbank/tbl2asn2.html) to facilitate submissions from large-scale sequencing groups. This program converts a table of annotations generated from an annotation pipeline into an ASN.1 record suitable for submission to GenBank. A version of *tbl2asn* called *table2asn_GFF* also accepts data in the GFF3 format (ftp.ncbi.nih.gov/toolbox/ncbi_tools/converters/by_program/table2asn_GFF/).

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.

Notes on particular sequence types

Third Party Annotation (TPA). TPA records are either derived or assembled from primary records in the INSDC databases that were sequenced by other investigators or sequencing centers (www.ncbi.nlm.nih.gov/genbank/tpa/). These primary records may be genomic, mRNA, WGS or SRA sequences. Reasonable candidates for TPA submissions include assemblies constructed from overlapping next-generation reads, gene annotations on previously unannotated genomic sequences, and mRNAs derived from unannotated genomic sequences by comparisons with known mRNAs from other organisms. TPA submissions must be accompanied by a peer-reviewed publication that describes the assembly or annotation.

Genomes. 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 a standard inference type and *text* consists of structured text. Annotations are not required for prokaryotic 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. Users should be aware that annotations on WGS project sequences may not be tracked from one assembly version to the next, and so should be considered preliminary. As part of the bacterial genome submission process, GenBank performs an average nucleotide identity analysis (9) to investigate whether the asserted organism name may be incorrect. For example, this ANI analysis can report that a genome submitted as *Escherichia coli* is actually *Salmonella enterica*. Since the analysis uses the genomes already in GenBank, it cannot necessarily be performed for all new genome submissions where there are no close taxonomic neighbors.

Targeted locus studies (TLS). Targeted locus studies often contain large sets of 16S rRNA sequence or ultra-conserved elements (UCEs). Similar to TSA records, TLS sequences are given a 'TLS' keyword and can be retrieved with the query 'tls[properties]' in the Nucleotide database.

Anti-microbial resistance data. As part of the NCBI Pathogen Detection project, NCBI accepts submissions of beta-lactamase sequences as supplementary data for either 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/).

RETRIEVING GENBANK DATA

The Entrez system

The sequence records in GenBank are accessible from the Nucleotide database in the NCBI Entrez retrieval system (10,11). GenBank sequences that are part of population or phylogenetic studies are also collected together in the PopSet database, and protein translations for coding region (CDS) features 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).

As discussed previously (1), given that a growing number of GenBank records (including those in many WGS, TSA and TLS projects) do not have GI numbers, NCBI has been transitioning to using accession.version identifiers as the primary identifiers for sequence records. As part of this effort, the Nucleotide database now supports retrieval of records without GI numbers by using their accession.version. For example, the query 'ULVQ01000101' in Nucleotide will retrieve that record, which can then be viewed and/or downloaded in the various formats supported by the Nucleotide database.

Sequence set browser

While sequences without GI numbers can now be retrieved in Nucleotide using their accession.version, these sequences cannot be retrieved by any other text queries. Thus, NCBI provides the Sequence Set Browser to allow enhanced access to such sequences (www.ncbi.nlm.nih.gov/Traces/wgs/). This interface serves both as a browser that can restrict a list of projects by facets such as taxonomy, source and BioProject ID, and also as a downloading tool that can provide either metadata tables or actual sequence data from selected projects.

Importance of associating sequence records with sequencing projects

NCBI strongly encourages submitters to register large-scale sequencing projects in the BioProject database (www.ncbi.nlm.nih.gov/bioproject). We also encourage submitters to update their BioProject records after relevant publications are available. Doing so provides reliable linkages between sequencing projects and the data they produce. Another benefit is that submitters can include a relevant grant in their BioProject that can then appear in their My Bibliography. In addition, sequence records may have a link to the

Table 3. Selected BLAST nucleotide databases[a]

Database	Contents
nr/nt	Taxonomic GenBank divisions
env_nt	ENV division
tsa_nt	TSA division
wgs	WGS sequences
16SMicrobial	Bacterial and archaeal 16S rRNA

[a] For more databases, see ftp.ncbi.nlm.nih.gov/blast/documents/blastdb.html.

BioSample database (12) that provides additional information about the biological materials used in the study.

GenBank records that contribute to genome assemblies will also have a link to the corresponding record in the Assembly database (13). 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 (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). Users should be aware that, because of the enormous diversity of available nucleotide sequence, it is not possible to search all NCBI sequence data at once. Rather, there are several BLAST databases, each suited to a particular type of sequence (Table 3).

Obtaining GenBank by FTP

NCBI provides sequence records in both the traditional GenBank flat file format and in a structured ASN.1 format. The ASN.1 format is used for the actual storage and maintenance of records at NCBI, while the flat file format is a rendering of the underlying ASN.1 representation. Full bimonthly GenBank releases in the flat file format, which includes sequence data from ENA and DDBJ, are available as a set of compressed files by anonymous FTP at ftp.ncbi.nlm.nih.gov/genbank. For convenience in file transfer, the data are partitioned into multiple files; for release 227 there are 3168 files requiring 893 GB of uncompressed disk storage. In addition, daily GenBank incremental update files containing new and updated records since the most recent release are available in flat file format at ftp.ncbi.nlm.nih.gov/genbank/daily-nc/.

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

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., Cavanaugh,M., Clark,K., Karsch-Mizrachi,I., Ostell,J., Pruitt,K.D. and Sayers,E.W. (2018) GenBank. *Nucleic Acids Res.*, **46**, D41–D47.
2. Silvester,N., Alako,B., Amid,C., Cerdeno-Tarraga,A., Clarke,L., Cleland,I., Harrison,P.W., Jayathilaka,S., Kay,S., Keane,T. *et al.* (2018) The European nucleotide archive in 2017. *Nucleic Acids Res.*, **46**, D36–D40.
3. Kodama,Y., Mashima,J., Kosuge,T., Kaminuma,E., Ogasawara,O., Okubo,K., Nakamura,Y. and Takagi,T. (2018) DNA data bank of Japan: 30th anniversary. *Nucleic Acids Res.*, **46**, D30–D35.
4. Karsch-Mizrachi,I., Takagi,T., Cochrane,G. and International Nucleotide Sequence Database, C. (2018) The international nucleotide sequence database collaboration. *Nucleic Acids Res.*, **46**, D48–D51.
5. NCBI Resource Coordinators. (2018) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **46**, D8–D13.
6. Federhen,S. (2012) The NCBI taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
7. Federhen,S. (2015) Type material in the NCBI taxonomy database. *Nucleic Acids Res.*, **43**, D1086–D1098.
8. 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.
9. Ciufo,S., Kannan,S., Sharma,S., Badretdin,A., Clark,K., Turner,S., Brover,S., Schoch,C.L., Kimchi,A. and DiCuccio,M. (2018) Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int. J. Syst. Evol. Microbiol.*, **68**, 2386–2392.
10. 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.
11. Gibney,G. and Baxevanis,A.D. (2011) Searching NCBI databases using entrez. *Curr. Protoc. Hum. Genet.*, doi:10.1002/0471142905.hg0610s71.
12. 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.
13. 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.
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.