Published online 5 November 2018 *Nucleic Acids Research, 2019, Vol. 47, Database issue* ***D23–D28***
doi: 10.1093/nar/gky1069

Database resources of the National Center for Biotechnology Information

Eric W. Sayers*, Richa Agarwala, Evan E. Bolton, J. Rodney Brister, Kathi Canese, Karen Clark, Ryan Connor, Nicolas Fiorini, Kathryn Funk, Timothy Hefferon, J. Bradley Holmes, Sunghwan Kim, Avi Kimchi, Paul A. Kitts, Stacy Lathrop, Zhiyong Lu, Thomas L. Madden, Aron Marchler-Bauer, Lon Phan, Valerie A. Schneider, Conrad L. Schoch, Kim D. Pruitt and James Ostell

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

Received September 19, 2018; Revised October 17, 2018; Editorial Decision October 18, 2018; Accepted October 18, 2018

ABSTRACT

The National Center for Biotechnology Information (NCBI) provides a large suite of online resources for biological information and data, including the GenBank® nucleic acid sequence database and the PubMed database of citations and abstracts published in life science journals. The Entrez system provides search and retrieval operations for most of these data from 38 distinct databases. The E-utilities serve as the programming interface for the Entrez system. Augmenting many of the web applications are custom implementations of the BLAST program optimized to search specialized data sets. New resources released in the past year include PubMed Labs and a new sequence database search. Resources that were updated in the past year include PubMed, PMC, Bookshelf, genome data viewer, Assembly, prokaryotic genomes, Genome, BioProject, dbSNP, dbVar, BLAST databases, igBLAST, iCn3D and PubChem. All of these resources can be accessed through the NCBI home page at www.ncbi.nlm.nih.gov.

INTRODUCTION

NCBI overview

The National Center for Biotechnology Information (NCBI), a center within the National Library of Medicine at the National Institutes of Health, was created in 1988 to develop information systems for molecular biology. Since that time the amount and variety of data that NCBI maintains has expanded enormously and can be generally grouped into six categories: Literature, Health, Genomes, Genes, Proteins and Chemicals (Table 1). NCBI provides facilities for submitting and downloading data, analysis and visualization software, educational events and materials about NCBI products, and software and services to support an expanding developer community. These services, along with all other data resources, are available through the NCBI home page at www.ncbi.nlm.nih.gov/. In most cases, the data underlying these resources and executables for the software described are available for download at ftp.ncbi.nlm.nih.gov.

This article provides a brief overview of the NCBI Entrez system of databases, followed by a summary of resources that were either introduced or significantly updated in the past year. More complete discussions of NCBI resources can be found on the home pages of individual databases, on the NCBI Learn page (www.ncbi.nlm.nih.gov/learn/), or in the NCBI Handbook (www.ncbi.nlm.nih.gov/books/NBK143764/).

The Entrez system

Entrez (1) is an integrated database retrieval system that provides access to a diverse set of 38 databases that together contain 2.5 billion records (Table 1 and Figure 1). Links to the web portal for each of these databases are provided on the Entrez global search page (www.ncbi.nlm.nih.gov/search/). Entrez supports text searching using simple Boolean queries, downloading of data in various formats, and linking records between databases based on asserted relationships. The LinkOut service expands the range of links to include external resources, such as organism-specific genome databases. The records retrieved in Entrez can be displayed in many formats and downloaded singly or in batches. An Application Programming Interface for Entrez functions (the E-utilities) is available, and detailed documentation is provided at eutils.ncbi.nlm.nih.gov.

[*] 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. The Entrez Databases (as of 1 September 2018)

Database	Records	Description
Literature		www.ncbi.nlm.nih.gov/home/literature/
PubMed	28 809 515	scientific and medical abstracts/citations
PubMed Central	5 096 212	full-text journal articles
NLM Catalog	1 586 932	index of NLM collections
Books	653 701	books and reports
MeSH	277 030	ontology used for PubMed indexing
Health		www.ncbi.nlm.nih.gov/home/health/
ClinVar	442 601	human variations of clinical significance
dbGaP	344 078	genotype/phenotype interaction studies
MedGen	307 690	medical genetics literature and links
GTR	55 299	genetic testing registry
Genomes		www.ncbi.nlm.nih.gov/home/genomes/
SNP	672 043 185	short genetic variations
Nucleotide	265 485 730	DNA and RNA sequences
GSS	40 713 027	genome survey sequences
Clone	38 325 184	genomic and cDNA clones
Probe	32 407 891	sequence-based probes and primers
BioSample	9 015 281	descriptions of biological source materials
SRA	6 243 265	high-throughput DNA and RNA sequence read archive
dbVar	5 227 838	genome structural variation studies
Taxonomy	1 969 776	taxonomic classification and nomenclature catalog
BioProject	309 309	biological projects providing data to NCBI
Assembly	194 537	genome assembly information
Genome	38 734	genome sequencing projects by organism
BioCollections	7623	museum, herbaria, and other biorepository collections
Genes		www.ncbi.nlm.nih.gov/home/genes/
GEO Profiles	128 414 055	gene expression and molecular abundance profiles
EST	76 990 816	expressed sequence tag sequences
Gene	32 928 347	collected information about gene loci
UniGene	6 473 284	clusters of expressed transcripts
GEO DataSets	2 756 045	functional genomics studies
PopSet	307 577	sequence sets from phylogenetic and population studies
HomoloGene	141 268	homologous gene sets for selected organisms
Proteins		www.ncbi.nlm.nih.gov/home/proteins/
Protein	568 577 026	protein sequences
Identical Protein Groups	182 401 155	protein sequences grouped by identity
Protein Clusters	1 137 329	sequence similarity-based protein clusters
Structure	142 217	experimentally-determined biomolecular structures
Conserved Domains	56 066	conserved protein domains
Chemicals		www.ncbi.nlm.nih.gov/home/chemicals/
PubChem Substance	247 411 095	deposited substance and chemical information
PubChem Compound	96 501 627	chemical information with structures, information and links
PubChem BioAssay	1 252 901	bioactivity screening studies
BioSystems	983 968	molecular pathways with links to genes, proteins and chemicals

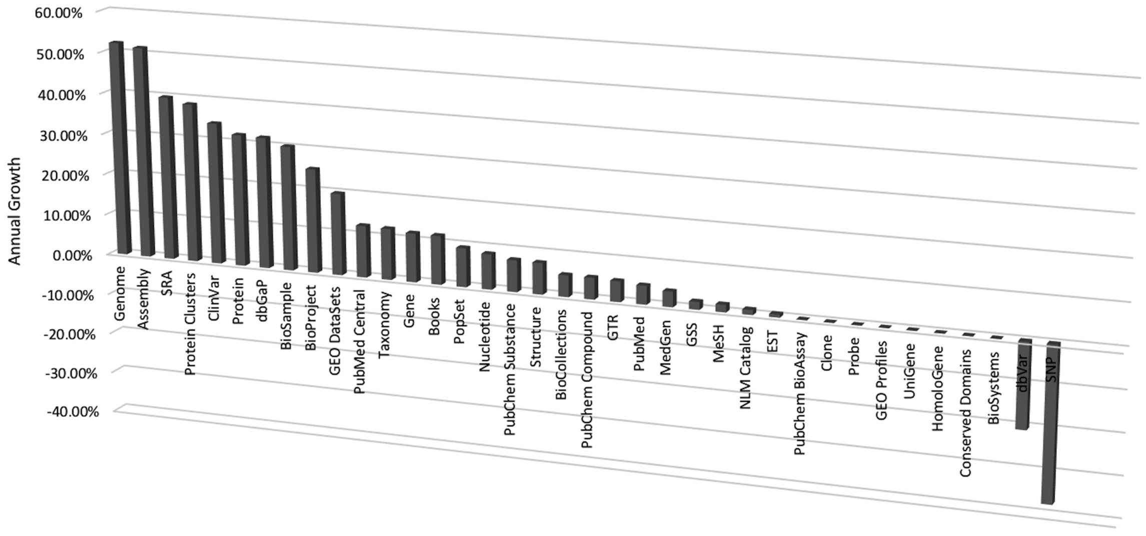

Figure 1. Annual growth rates of the number of records in each Entrez database as of 1 September 2018. Identical Protein Groups is not included since this database was released during the past year. Please see the text for a discussion of a change in scope for dbVar and SNP.

Data sources and collaborations

NCBI receives data from three sources: direct submissions from researchers, national and international collaborations or agreements with data providers and research consortia, and internal curation efforts. For example, NCBI manages the GenBank database (2) and participates with the EMBL-EBI European Nucleotide Archive (ENA) (3) and the DNA Data Bank of Japan (DDBJ) (4) as a partner in the International Nucleotide Sequence Database Collaboration (INSDC) (5). Details about direct submission processes are available from the NCBI Submit page (www.ncbi.nlm.nih.gov/home/submit.shtml) and from the resource home pages (e.g. the GenBank page, www.ncbi.nlm.nih.gov/genbank/). NCBI staff provide identifiers to submitters for their data usually within 2–5 business days, depending on the destination database and the complexity of the submission. More information about the various collaborations, agreements, and curation efforts are also available through the home pages of the individual resources.

RECENT DEVELOPMENTS

Literature updates

PubMed and PubMed Labs. As the biomedical literature keeps growing at an exponential rate in PubMed (over 28 million articles in August 2018), we have continuously experimented and investigated ways to improve the overall search quality and user experience for the biomedical literature. As part of our efforts to move towards PubMed 2.0 (6), PubMed now offers Best Match: a new relevance sort option—as an alternative to the default date sort—that uses a state-of-the-art machine learning algorithm trained on aggregated past user searches. The new Best Match algorithm ranks search results according to a number of relevance signals, including an article's popularity, its publication date and type, and its query-document relevance score (7).

In late 2017, we also launched PubMed Labs (www.pubmed.gov/labs), an experimental system that exposes new literature search features and tools such as informative article snippets in the search results and a convenient toggle switch between two sort orders: Best Match and Most Recent. PubMed Labs (8) also features a clean and mobile-friendly design tailored specifically towards the small screen devices increasingly popular with our users in recent years. PubMed Labs is also a platform for users to provide feedback, which allows us to make more informed decisions about potential changes to improve the search quality and overall usability of PubMed, ultimately providing a better user experience.

PubMed Central (PMC). The PubMed Central archive grew to 5 million articles in July 2018. This growth is supported by the ongoing Biomedical Journal Digitization project (www.ncbi.nlm.nih.gov/pmc/about/scanning/), the continued expansion of funder policies requiring public access to research, and the steady increase in journals that use PMC as a digital archive. PMC continues to receive >500 applications each year from journals interested in joining the archive. To address questions about the NLM's standards for including a journal in PMC as well as ongoing publisher and journal expectations, PMC released a full statement on journal selection, including scientific and editorial quality considerations (www.ncbi.nlm.nih.gov/pmc/pub/journalselect/) in addition to an overview of its reevaluation process (www.ncbi.nlm.nih.gov/pmc/about/guidelines/#standards). This new guidance aims to increase the transparency of NLM selection processes and to ensure the continued public trust in NLM resources.

Although many of the 5 million articles in PMC are subject to traditional copyright restrictions and are not available for bulk downloading, there are several collections within PMC where bulk retrieval is permitted for text mining and other purposes (www.ncbi.nlm.nih.gov/pmc/tools/textmining/). The largest of these collections is the PMC Open Access Subset, which surpassed 2 million articles in May 2018. The Biomedical Journal Digitization project has continued to add content from historically significant biomedical journals, some dating back to the 18th century, to these text mining collections, which now span >200 years of scientific research. The files for these collections can be obtained in XML or text formats, and together represent some of the largest and most diverse biomedical text corpuses available in machine-readable format. License terms for re-use may vary by collection or even within a collection.

Bookshelf. The NCBI Bookshelf now provides free online access to over 6000 books and documents in life science healthcare from over 150 content providers. In the past year, Bookshelf has better defined its scope and process for selecting titles for the resource, including scientific and technical quality criteria (www.ncbi.nlm.nih.gov/books/about/publishers/). In addition, Bookshelf added support for searching by MeSH (Medical Subject Headings) fields, including MeSH Major Topics [MAJR], MeSH Subheadings [SH], and MeSH Terms [MH]. Bookshelf populates these MeSH fields in its index from the MeSH assignments in the NLM Catalog (www.ncbi.nlm.nih.gov/nlmcatalog). Bookshelf also supports searching by author-supplied keywords. These keywords are indexed along with autogenerated concept phrases in the Bookshelf Concept Phrases and Keywords field [KYWD]. For more information about using these and other Bookshelf search fields, see the Search Field Descriptions and Tags section of Bookshelf Help (www.ncbi.nlm.nih.gov/books/NBK45615/). Finally, Bookshelf has added an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) service that provides access to metadata of all items in the Bookshelf archive, as well as to the full text of a subset of these items (www.ncbi.nlm.nih.gov/books/about/oai/).

Genome updates

Sequence database search. NCBI is updating the sequence search experience, offering support for more natural language queries for gene, transcript, protein, and assembly data and returning results highlighting high value content. A new search service, running in parallel to NCBI's Entrez search, recognizes queries that frequently fail to return results in Entrez or that only return results when performed in specific Entrez databases. These include several query types: 'organism-gene' (e.g. human BRCA1), 'organism-transcript' (e.g. mouse p53 transcripts), and 'organism-assembly' (e.g. dog reference genome). In addition, NCBI now offers an improved ability to find curated gene sets that are part of the RefSeq Targeted Loci Project, which includes ribosomal RNA genes in bacteria, archaea, and fungi, as well as internal transcribed spacer regions in fungi and oomycetes. Featured results from these searches are presented at the top of results pages for NCBI's global search, as well as for the Gene, Nucleotide, Protein, Assembly and Genome database pages. Work is ongoing to continue improving the search experience, expanding both query recognition and featured content.

Assembly. To better support viewing and downloading of viral and viroid genomes, NCBI has made several classes of viral and viroid genome data available through the NCBI Assembly Resource (www.ncbi.nlm.nih.gov/assembly/). This resource supports retrieval of all nucleotide records that together comprise a single genome. This feature is particularly useful for segmented viruses because individual segment sequences are aggregated into a single genome constellation that is represented by a single accession.

The Assembly resource (9) supports searches based on organism and taxonomy names, and viral assemblies can be found at www.ncbi.nlm.nih.gov/assembly/?term=viruses%5Borgn%5D. These include viral RefSeq sequences (10) and GenBank sequences designated as viral species exemplars by the International Committee for Taxonomy (11). A 'reference' subset of these RefSeq assemblies includes experimentally supported and manually curated annotations and is intended to provide high quality reference templates for viral annotation. GenBank assemblies include species exemplars selected by the International Committee on Taxonomy of Viruses (ICTV), assemblies (in GenBank) from which RefSeq assemblies were built, and a set of complete viral genomes that have been validated by NCBI processes. The initial scope of these NCBI-validated GenBank genomes is limited to a few viral taxa but is being expanded to ultimately include all viruses.

In addition to including more viral genomes, NCBI has made several improvements to the Assembly resource that facilitate finding and downloading genome data sets of interest. Annotation status filters have been added that enable users to select genome assemblies that have annotations. Filters are now exposed that make it easy to limit search results to assemblies derived from type strains or ICTV species exemplars. UCSC assembly names have been added as searchable synonyms for most of the recent assemblies in the UCSC Genome Browser. In addition, new file types have been added to the 'Download Assemblies' menu, including a 'Feature count' file with counts of gene, RNA, and CDS features of specific types and a 'Translated CDS' file with conceptual translations of each CDS feature on the genome.

Prokaryotic genomes. NCBI now uses an Average Nucleotide Identity method (12) with optimum threshold ranges for prokaryotic taxa to review all prokaryotic genome assemblies in GenBank and to adjust incorrectly assigned names when compared to genomes from type strains. This is the result of a project initiated after a 2015 NCBI workshop involving several parties in the bacteriology community (13) and was recently described in more detail (12).

SKESA de-novo assemblies in SRA. SKESA (14) is a DeBruijn graph-based *de-novo* assembler developed at NCBI for assembling Illumina reads of microbial genomes. NCBI is using SKESA to support SRA as well as the Pathogen Detection project (www.ncbi.nlm.nih.gov/pathogens/). Over 270 000 read sets within SRA now contain SKESA assemblies, and these assemblies are available for download. SRA runs that have SKESA assemblies will have a *run.realign* file listed on the *Download* tab on the SRA Run Browser. For example, run SRR498276 has a *run.realign* file named *SRR498276.realign* listed on this page: trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR498276. The source code for SKESA is freely available at github.com/ncbi/SKESA/releases.

Genome data viewer. New functionalities in NCBI's genome browser, Genome Data Viewer (GDV) (www.ncbi.nlm.nih.gov/genome/gdv), offer additional means to analyze genomic data. GDV has added more options to support the analysis of user-supplied data. In addition to uploading files, users now have the option to connect to files hosted on remote servers or that are part of track data hubs (15). Once connected, these externally provided data appear as tracks alongside NCBI's own track offerings and can be included in publication-quality PDF downloads from the browser. The BLAST widget integrates genomic BLAST into GDV's graphical display, allowing users to view their existing results as browser tracks or to run new queries directly from the browser. The associated BLAST Alignment Inspector provides a graphical view that highlights the relationship of alignment results to NCBI RefSeq annotations. In addition, BLAST results pages now provide links to GDV views of sequences aligned to either a genome assembly or to RefSeq annotations on an assembly.

Genome and BioProject data browsers. The BioProject and Genome databases have new interfaces and backends that provide improved browsing of their respective datasets. BioProject, which organizes metadata associated with research projects, and Genome, which aggregates genome-associated data, both serve as entry points for exploration of many other NCBI resources. The new targeted search interfaces for these two resources (www.ncbi.nlm.nih.gov/bioproject/browse and www.ncbi.nlm.nih.gov/genome/browse#!/overview/) share a similar look and feel, and allow users to begin exploring either by text searching or by filtering the data according to relevant categories, such as various taxonomic restrictions. The tabular display of the results is highly customizable, with sortable columns and a variety of fields available for display. Results can be downloaded to tab-delimited files, and each retrieved record is linked to other NCBI data, such as specific records in BioProject, Genome, Taxonomy, and PubMed. More details for the BioProject browser are available in the documentation (www.ncbi.nlm.nih.gov/bioproject/docs/faq/#questions-about-the-browse-page).

dbSNP. The Database of Single Nucleotide Polymorphisms (dbSNP) is a repository of short genetic variations less than 50 base pairs in length. The database has been growing rapidly: the human data alone quadrupled in size in less than a year, growing from 150 million Reference SNPs (RS) in Build 149 to more than 650 million RS records in Build 151. In addition, more than 580 million of these RS records have frequency data in Build 151. To address the challenges of processing, annotating, and exchanging the growing volume of data, we made two major changes to dbSNP in the past year. First, dbSNP entered a new agreement with EMBL-EBI to share responsibility for managing data from genetic variation experiments worldwide. The outcome of this agreement is that dbSNP now only manages human data, while all non-human organisms have been moved to the EMBL-EBI European Variation Archive (EVA) (ncbiinsights.ncbi.nlm.nih.gov/2017/05/09/phasing-out-support-for-non-human-genome-organism-data-in-dbsnp-and-dbvar/). Second, dbSNP represents variants using the new SPDI data model (www.ncbi.nlm.nih.gov/variation/notation/), and exposes a new API based on this data model. In addition, dbSNP released a new RefSNP page for displaying variants in web browsers (ncbiinsights.ncbi.nlm.nih.gov/2017/07/07/dbsnp-redesign-supports-future-data-expansion/).

dbVar. The NCBI dbVar Structural Variant database houses human genomic structural variants (SV) greater than 50 base pairs in length. From the dbVar homepage (www.ncbi.nlm.nih.gov/dbvar) users can search, view, and download variant data from over 150 studies, including 1000 Genomes Phase 3 (estd219), Simons Genome Diversity Project (nstd128), ClinGen (nstd45), ExAC (nstd151) and many more. Users can access the variants using the Study Browser or the graphical Genome Browser. Individual study and variant pages include links to raw data as well as to related information at other NCBI and external resources. Bulk data downloads are available via FTP (ftp.ncbi.nlm.nih.gov/pub/dbVar/data).

In 2018, dbVar introduced a new comprehensive set of non-redundant structural variants (NR SV) consisting of unique insertions, duplications, and deletions. These compact files are suitable for use as references in the analysis of human SV such as filtering and annotating other SV datasets, SV discovery, and identifying rare and/or clinical SV. The dbVar NR SV currently includes >2.2 million deletions, 1.1 million insertions, and 300 thousand duplications, and will be updated regularly as new variants are added to dbVar. Users can find more information about NR SV, including brief tutorials and access to the NR SV FTP files at github.com/ncbi/dbvar/tree/master/Structural_Variant_Sets.

BLAST updates

BLAST databases. NCBI has released a new version of the BLAST databases (version 5) with several enhancements. First, the stand-alone BLAST+ executables (16) can now limit a search by taxonomy without downloading extra files. Subject sequences can be both included or excluded from the search based on taxonomy. Second, the new database version makes uses of LMDB (Lightning Memory-Mapped Database) to perform faster sequence lookups by accession. The version 5 databases can only be used with BLAST+ 2.8.0 or newer.

IgBLAST. IgBLAST (17), the NCBI tool to analyze immunoglobulin and T cell receptors, has received important updates in the last year. First, IgBLAST can now more efficiently process a large batch of queries using a multi-threading approach. Second, IgBLAST can fetch reads from the SRA database by specifying an SRA accession on the command-line, so that the sequences do not need to be downloaded by the user. Finally, IgBLAST now supports the AIRR (Adaptive Immune Receptor Repertoire) rearrangement format. This format is a standard supported by the adaptive immune receptor repertoire (AIRR) community (docs.airr-community.org/en/latest/) and is specifically designed for repertoire studies using next generation sequencing technology.

Protein updates

iCn3D. In April 2018 NCBI released an updated iCn3D version (2.0) with more features and improved performance. iCn3D provides functionality similar to that of Cn3D, NCBI's standalone structure viewer, but runs directly in web browsers and does not require users to install an application. Interactive iCn3D views have been embedded in structure summary pages of NCBI's Molecular Modeling Database (MMDB), and iCn3D visualizes results of 3D structure comparisons computed by VAST+. iCn3D can simultaneously display 3D structures, 2D interaction schematics, and protein/nucleotide sequences, and can load annotations such as sequence variants, protein domains, and functional and binding sites. The displays interact with each other and facilitate a variety of selection, highlight, and analysis operations. iCn3D now supports the export of stereolithography (STL) or Virtual Reality Modeling Language (VRML) files for 3D printing, and can also generate shareable links for custom displays (e.g., d55qc.app.goo.gl/HDuWMFAVokxvHMKSA). The source code of iCn3D is available at https://github.com/ncbi/icn3d.

Chemical updates

PubChem (18–20) (pubchem.ncbi.nlm.nih.gov) now provides chemical information for more than 96 million compounds collected from over 620 data sources. In the past year, several important improvements have been made to PubChem. A data contribution from BioRad's SpectraBase provided over 630 000 spectra images for more than 225 000 compounds with pertinent metadata. In addition, the publisher Springer Nature graciously contributed more than 28 million links between more than 610 000 compounds and over four million scientific articles, with updates on a weekly basis. Of these, two million links point to over 350 000 open-access or free-to-read articles.

To improve access to bioactivity content, a set of dyad pages is now available (pubchemdocs.ncbi.nlm.nih.gov/bioactivity-dyad-pages). These pages provide quick access to bioactivity details for a given chemical tested in an assay, as well as useful information helpful in interpreting bioactivity of a compound or building a structure-activity relationship for a given gene or protein target.

New co-occurrence knowledge panels (pubchemdocs.ncbi.nlm.nih.gov/knowledge_panels) show a list of chemicals frequently co-occurring with a given compound in PubMed articles. Users can download the list of PubMed articles co-mentioning the two chemicals for further analysis. Additional details about these and other PubChem developments are available on the PubChem blog (pubchemblog.ncbi.nlm.nih.gov).

FOR FURTHER INFORMATION

The resources described here include documentation, other explanatory materials, and references to collaborators and data sources on their respective web sites. An alphabetical list of NCBI resources is available from a link above the category list on the left side of the NCBI home page. The NCBI Help Manual and the NCBI Handbook (www.ncbi.nlm.nih.gov/books/NBK143764/), both available as links in the common page footer, describe the principal NCBI resources in detail. The NCBI Learn page (www.ncbi.nlm.nih.gov/learn/) provides links to documentation, tutorials, webinars, courses, and upcoming conference exhibits. A variety of video tutorials are available on the NCBI YouTube channel that can be accessed through links in the standard NCBI page footer. A user-support staff is available to answer questions at info@ncbi.nlm.nih.gov, and users can view support articles at support.nlm.nih.gov. Updates on NCBI resources and database enhancements are described on the NCBI Insights blog (ncbiinsights.ncbi.nlm.nih.gov), NCBI social media sites (FaceBook, Twitter, and LinkedIn), and the several mailing lists and RSS feeds that provide updates on services and databases. Links to these resources are in the NCBI page footer and on NCBI Insights.

ACKNOWLEDGEMENTS

The authors would like to thank all of the NCBI staff who through their dedicated efforts continue to allow NCBI to provide our full collection of services to the community. In particular, the authors would like to thank these individuals for their contributions to the resources described in this work: Jeff Beck, Tiejun Cheng, John Garner, Lewis Geer, Asta Gindulyte, Siqian He, Anne Ketter, Michael Kimmelman, Anatoily Kuznetsov, Valerii Lashmanov, Martin Latterner, John Lopez, Peter Meric, Eric Moyer, Ben Shoemaker, Alexandre Souvorov, Tatiana Tatusova, Paul Thiessen, Jiyao Wang, Ming Ward, Chunlin Xiao, Eugene Yaschenko, Jian Ye, Bo Yu, Leonid Zaslavsky, Jian Zhang and Dachuan Zhang.

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. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. Fiorini,N., Lipman,D.J. and Lu,Z. (2017) Towards PubMed 2.0. *Elife*, **6**, e28801
7. Fiorini,N., Canese,K., Starchenko,G., Kireev,E., Kim,W., Miller,V., Osipov,M., Kholodov,M., Ismagilov,R., Mohan,S. *et al.* (2018) Best Match: new relevance search for PubMed. *PLoS Biol.*, **16**, e2005343.
8. Fiorini,N., Canese,K., Bryzgunov,R., Radetska,I., Gindulyte,A., Latterner,M., Miller,V., Osipov,M., Kholodov,M., Starchenko,G. *et al.* (2018) PubMed Labs: an experimental system for improving biomedical literature search. *Database (Oxford)*, doi:10.1093/database/bay094.
9. 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.
10. Brister,J.R., Ako-Adjei,D., Bao,Y. and Blinkova,O. (2015) NCBI viral genomes resource. *Nucleic Acids Res.*, **43**, D571–D577.
11. King,A.M.Q., Lefkowitz,E.J., Mushegian,A.R., Adams,M.J., Dutilh,B.E., Gorbalenya,A.E., Harrach,B., Harrison,R.L., Junglen,S., Knowles,N.J. *et al.* (2018) Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2018). *Arch. Virol.*, **163**, 2601–2631.
12. 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.
13. Federhen,S., Rossello-Mora,R., Klenk,H.P., Tindall,B.J., Konstantinidis,K.T., Whitman,W.B., Brown,D., Labeda,D., Ussery,D., Garrity,G.M. *et al.* (2016) Meeting report: GenBank microbial genomic taxonomy workshop (12-13 May, 2015). *Stand. Genomic Sci.*, **11**, 15.
14. Souvorov,A., Agarwala,R. and Lipman,D.J. (2018) SKESA: strategic kmer extension for scrupulous assemblies. *Genome Biol.*, **19**, 153.
15. Raney,B.J., Dreszer,T.R., Barber,G.P., Clawson,H., Fujita,P.A., Wang,T., Nguyen,N., Paten,B., Zweig,A.S., Karolchik,D. *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**, 1003–1005.
16. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
17. Ye,J., Ma,N., Madden,T.L. and Ostell,J.M. (2013) IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.*, **41**, W34–W40.
18. Kim,S., Thiessen,P.A., Bolton,E.E., Chen,J., Fu,G., Gindulyte,A., Han,L., He,J., He,S., Shoemaker,B.A. *et al.* (2016) PubChem Substance and Compound databases. *Nucleic Acids Res*, **44**, D1202–D1213.
19. Wang,Y., Bryant,S.H., Cheng,T., Wang,J., Gindulyte,A., Shoemaker,B.A., Thiessen,P.A., He,S. and Zhang,J. (2017) PubChem BioAssay: 2017 update. *Nucleic Acids Res.*, **45**, D955–D963.
20. Kim,S. (2016) Getting the most out of PubChem for virtual screening. *Expert Opin. Drug Discov.*, **11**, 843–855.