

Database resources of the National Center for Biotechnology Information

Eric W. Sayers ¹*, Jeffrey Beck, Evan E. Bolton, Devon Bourexis, James R. Brister, Kathi Canese, Donald C. Comeau, Kathryn Funk, Sunghwan Kim ², William Klimke, Aron Marchler-Bauer, Melissa Landrum, Stacy Lathrop, Zhiyong Lu ³, Thomas L. Madden, Nuala O’Leary, Lon Phan, Sanjida H. Rangwala, Valerie A. Schneider, Yuri Skripchenko, Jiyao Wang, Jian Ye, Barton W. Trawick, Kim D. Pruitt and Stephen T. Sherry

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, 2020; Revised September 25, 2020; Editorial Decision September 28, 2020; Accepted October 08, 2020

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 34 distinct databases. The E-utilities serve as the programming interface for the Entrez system. Custom implementations of the BLAST program provide sequence-based searching of many specialized datasets. New resources released in the past year include a new PubMed interface and NCBI datasets. Additional resources that were updated in the past year include PMC, Bookshelf, Genome Data Viewer, SRA, ClinVar, dbSNP, dbVar, Pathogen Detection, BLAST, Primer-BLAST, IgBLAST, iCn3D and PubChem. All of these resources can be accessed through the NCBI home page at <https://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 (1). In this article we provide a brief overview of the NCBI Entrez system of databases, followed by a summary of resources that we either introduced or significantly updated in the past year. We provide more complete discussions of NCBI resources on the home pages of individual databases,

on the NCBI Learn page (<https://www.ncbi.nlm.nih.gov/learn/>) and in the NCBI Handbook (<https://www.ncbi.nlm.nih.gov/books/NBK143764/>).

The entrez system

Entrez (2) is an integrated database retrieval system that provides access to a diverse set of 34 databases that together contain 3.0 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 (<https://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 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 <https://eutils.ncbi.nlm.nih.gov/>.

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 (3) and participates with the EMBL-EBI European Nucleotide Archive (ENA) (4) and the DNA Data Bank of Japan (DDBJ) (5) as a partner in the International Nucleotide Sequence Database Collaboration (INSDC) (6). 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/). More information about the various collaborations, agreements, and curation efforts are also available through the home pages of the individual resources.

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

Table 1. The Entrez Databases (as of 9 September 2020)

Database	Records	Description
Literature		
PubMed	31 471 600	scientific and medical abstracts/citations
PubMed Central	6 447 271	full-text journal articles
NLM catalog	1 619 856	index of NLM collections
Books	825 385	books and reports
MeSH	300 500	ontology used for PubMed indexing
Genomes		
Nucleotide	429 731 711	DNA and RNA sequences
BioSample	14 628 076	descriptions of biological source materials
SRA	11 807 161	high-throughput DNA and RNA sequence read archive
Taxonomy	2 401 136	taxonomic classification and nomenclature catalog
Assembly	837 406	genome assembly information
BioProject	458 893	biological projects providing data to NCBI
Genome	55 580	genome sequencing projects by organism
BioCollections	8 138	museum, herbaria and other biorepository collections
Genes		
GEO Profiles	128 414 055	gene expression and molecular abundance profiles
Gene	28 377 759	collected information about gene loci
GEO datasets	4 002 373	functional genomics studies
PopSet	350 627	sequence sets from phylogenetic and population studies
HomoloGene	141 268	homologous gene sets for selected organisms
Genetics		
SNP	720 643 623	short genetic variations
dbVar	6 030 887	genome structural variation studies
ClinVar	845 008	human variations of clinical significance
MedGen	335 277	medical genetics literature and links
GTR	76 814	genetic testing registry
dbGaP	1 397	genotype/phenotype interaction studies
Proteins		
Protein	874 272 642	protein sequences
Identical protein groups	329 946 078	protein sequences grouped by identity
Protein clusters	1 137 329	sequence similarity-based protein clusters
Structure	167 650	experimentally-determined biomolecular structures
Sparcle	149 462	conserved domain architectures
Conserved domains	59 951	conserved protein domains
Chemicals		
PubChem substance	285 048 146	deposited substance and chemical information
PubChem compound	111 325 418	chemical information with structures, information and links
PubChem BioAssay	1 229 071	bioactivity screening studies
BioSystems	983 968	molecular pathways with links to genes, proteins and chemicals

RECENT DEVELOPMENTS

Literature updates

PubMed. After previewing an updated version of PubMed in 2019, we activated this updated version as the default system in May 2020. Among the numerous enhancements is a responsive layout that offers better support for accessing PubMed content on increasingly popular small-screen devices such as mobile phones and tablets. The interface is compatible with any screen size and provides a fresh, consistent look and feel throughout the application, no matter how one accesses it. Search results can now be sorted using ‘Best Match’ ordering that employs a machine learning algorithm to help users find relevant citations quickly (7). Search results also include ‘snippets’, highlighted text fragments from the article abstract that are selected based on their relatedness to the query. These snippets give users additional information to help them decide if an article is useful. Additional improvements to

the interface make it easier to discover related content such as similar articles, references and citations.

Since 2012 PubMed has allowed users to search for a distinct author, and not merely for an author name, through an automatic author name disambiguation algorithm (8). We recently enhanced this algorithm by leveraging the significant growth of ORCID use in PubMed articles. Users can search PubMed directly with ORCIDs using the following syntax: 0000-0001-6166-3199[auid]. Additionally, to further strengthen how PubMed handles synonyms, we developed and implemented an updated algorithm to obtain word synonym pairs that improve PubMed indexing and retrieval (9).

The updated version of PubMed takes advantage of several new technologies to improve the user experience. The underlying document data indexed in the updated version is a merger of content from PubMed, Bookshelf and PubMed Central (PMC). This combined dataset allows us to display relevant information not previously available in a PubMed

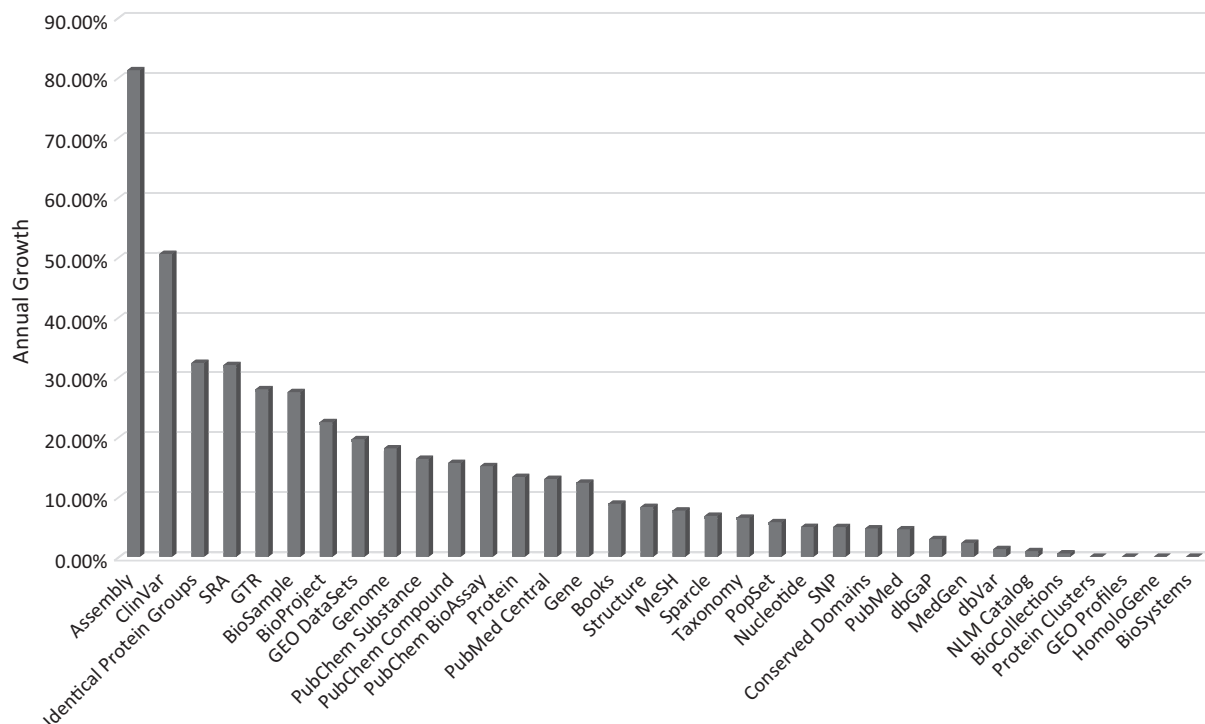


Figure 1. Annual growth rates of the number of records in each Entrez database as of 9 September 2020.

record, such as reference citations from PMC. While legacy PubMed limited the number of variants for a wildcard (“*”) search, PubMed is now capable of unlimited wildcard searches thanks to Solr (<https://lucene.apache.org/solr/>), the open-source enterprise search system that PubMed now uses for document indexing. Users will find that PubMed now has greater scalability and reliability, provided not only by Solr, but also by the MongoDB storage solution and the modern cloud architecture that together ensure both redundancy between data centers and also trustworthy backup environments. When visiting PubMed, users will enjoy a modern web experience using the latest web technologies and standards, all provided by the Django web framework.

PubMed Central (PMC). PMC continued to expand access to biomedical and life science literature over the past year, with the corpus now including more than 6 million journal articles and author manuscripts. Reflecting the NLM’s ongoing commitment to public access to research results supported by the NIH and other funding partners, we released a new NIH Manuscript Submission (NIHMS) system in January 2020. The new NIHMS system streamlines the author manuscript submission process to PMC and offers more transparent options to aid authors and investigators in avoiding processing delays and ensuring timely compliance with NIH policy. In the first 6 months after release, we received nearly 40 000 successful submissions (<https://www.nihms.nih.gov/about/statistics/>).

PMC launched the Public Health Emergency COVID-19 Initiative (<https://www.ncbi.nlm.nih.gov/pmc/about/covid-19/>) in March 2020, which helped enable the creation of the COVID-19 Open Research Dataset (CORD-19) hosted by the Allen Institute. This initiative resulted from a call

made by the national science and technology advisors of a dozen countries, including the US, to support ongoing public health emergency response efforts. Specifically, this call asked publishers and societies to agree voluntarily to make their publications and supporting data related to COVID-19 and the novel coronavirus immediately accessible in PMC and other appropriate public repositories. As of August, this initiative included more than 50 publishers and has added or updated the licenses on 80 000 articles in PMC to support secondary re-use and analysis.

On 9 June 2020, NLM launched a pilot project to test the viability of allowing users to obtain from PMC preprints resulting from NIH-funded research (<https://www.ncbi.nlm.nih.gov/pmc/about/nihpreprints/>). The primary goal of this NIH Preprint Pilot is to explore approaches to increase the discoverability of early NIH research results. Following standard NLM practice, citations for these preprint records are available in PubMed to increase the discoverability of this content. To ensure transparency, large banners clearly identify these preprint records in PMC and PubMed. The banners explain that the papers have not been peer reviewed, and link to information about the pilot for additional context. The pilot will run for a minimum of 12 months and will focus on preprints that relate to the current COVID-19 pandemic. Lessons we learn during that time will inform future NLM efforts involving preprints.

Bookshelf. The NCBI Bookshelf provides free online access to over 8500 books and documents in life science and healthcare from over 150 content providers. In the past year, we migrated the content management of two large full-text toxicology databases into its archive. The first, LactMed, is an NLM database of over 1500 peer-reviewed summaries

containing information on drugs and other chemicals to which breastfeeding mothers may be exposed. It also includes information on the levels of such substances in breast milk and infant blood, and the possible adverse effects in the nursing infant. The second, LiverTox, is an NIDDK database of over 1100 peer-reviewed documents containing information on the diagnosis, cause, frequency, clinical patterns and management of liver injury attributable to both prescription and nonprescription medications and also selected herbal and dietary supplements. Migrating these full-text databases into the Bookshelf increases their discoverability in two ways: the citations for each full-text LactMed and LiverTox summary record are now available in PubMed, where they link to the full-text in Bookshelf; and the toxicological data for these records are seamlessly integrated in PubChem with the scientific evidence summarized and cited in the full-text documents in Bookshelf. Users can still search within LactMed or LiverTox, and the full-text, machine-readable records for both are available through the NLM LitArch Open Access subset for data mining and reuse.

Genome updates

NCBI Datasets. We have continued to develop improvements to sequence search through the introduction of Datasets, a new resource that enables users to easily gather content from across NCBI databases (<https://www.ncbi.nlm.nih.gov/datasets/>). Developed with the FAIR (findability, accessibility, interoperability and reusability) principles of data management in mind, Datasets allows users to create custom datasets through web, RESTful API, and command-line interfaces, and download them in structured file packages. As of this writing, Datasets supports queries for genomes and genes across a wide range of taxonomies, including quick access to SARS-CoV-2 genomes and proteins. Users searching for genomic data can assemble a package consisting of genome, transcript and protein sequences as well as annotations. The package will also include a data report with rich metadata. The Datasets web interface (Figure 2) allows users to browse genomes on a taxonomic tree and choose any set of complete eukaryotic genomes in the NCBI Assembly database. The API and command-line interfaces provide access to prokaryotic and viral genomes in addition to eukaryotic genomes, and support searches with taxonomic identifiers or assembly accessions. For gene searches, Datasets allows users to construct data tables based on either an NCBI Gene ID or a combination of organism and gene symbol. On the web interface, users can continue editing the set of genes on the table and then select a set of desired data columns before downloading the data. The API and command-line interfaces provide similar tables programmatically. A Python library and corresponding Jupyter notebooks allow users to explore and learn the datasets API, and these are available on GitHub (<https://github.com/ncbi/datasets>).

Graphical sequence viewers. NCBI's graphical sequence viewer tools visualize sequences, annotations and experimental data alignments archived in NCBI databases. These viewers include the NCBI Sequence Viewer (SV) and the

Multiple Sequence Alignment Viewer (MSAV). SV is available as a standalone application (<https://www.ncbi.nlm.nih.gov/projects/sviewer/>) and also appears as a fully-functional embed on many NCBI pages, including Gene and SNP records and the NCBI Variation Viewer. SV is also the graphical viewer for Nucleotide and Protein records and the BLAST (10) and Primer-BLAST (11) results pages. MSAV (<https://www.ncbi.nlm.nih.gov/projects/msaviewer/>) displays multiple sequence alignments created by NCBI BLAST, COBALT and NCBI Virus pages, and also displays custom alignments from researchers.

In the last year, we have further enhanced the sequence and data download capabilities within SV (<https://www.ncbi.nlm.nih.gov/tools/sviewer/>). Users can now download gene, feature and NCBI SNP annotation data from the graphical view tool in multiple common formats that include GFF3, VCF and BED. Users can also copy short strings of sequence data directly to the clipboard and can also download larger ranges in FASTA or GenBank flat file format. Improved tooltips now report positional information that changes dynamically depending on the position of the cursor. For gene features, including transcript and proteins, the tooltip provides the transcript, CDS, and/or protein position and exon or intron number. For BAM, SRA or BLAST alignments, the tooltip information also includes the ability to view any unaligned data, including insertions and 5' and 3' tails. Recent enhancements to NCBI's MSAV (<https://www.ncbi.nlm.nih.gov/tools/msaviewer/>) allow users to re-sort rows to find sequences of interest and then customize their display further by 'hiding' undesired sequences from view (such as partial or duplicate alignments) or in their SVG/PDF image output.

Genome data viewer. NCBI's flagship genome browser, Genome Data Viewer (GDV) (<https://www.ncbi.nlm.nih.gov/genome/gdv/>), integrates the SV graphical display with a robust search/retrieval/analysis console. Users can view their own data next to NCBI tracks as references. In our endeavor to better support our users' analyses and research needs, we released numerous enhancements in the past year. Of particular note is a dynamic sidebar that users can show or hide. Hiding the sidebar allows the graphical display to stretch for a wider view. Additional enhancements now allow navigation by assemblies, chromosomes and components directly from within the browser.

Users interested in non-human variation data can now add SNP data from the European Variation Archive as tracks directly to the view using the tracks configuration menu in SV. Researchers can also add their own or third-party data by streaming data hosted on external URLs. GDV also supports connectivity to UCSC-style track hubs directly from the EBI Track Hub Registry. GDV supports visualization of most common bioinformatics formats including GFF3, bigWig, multiWig, bigBED and indexed BAM and VCF files. Helpful tutorials to get started are available on the NCBI YouTube channel in the 'NCBI Genome Data Viewer' playlist. Additional documentation is available at <https://www.ncbi.nlm.nih.gov/genome/gdv/browser/help/>.

Welcome to NCBI Datasets BETA

NCBI Datasets is an experimental resource for finding and building datasets - and we're just getting started! Our web interface allows you to download genome sequence and annotation for eukaryotic organisms and our recently added SARS-CoV-2 genome and protein datasets. ... [more](#)

Programmatic access

Bacterial and viral data are not yet supported for online browsing. For access to data for all organisms, including bacteria and viruses, use our command line tool and RESTful APIs.

Command-line

Our Datasets command-line tool, is available for Windows, Mac, and Linux.

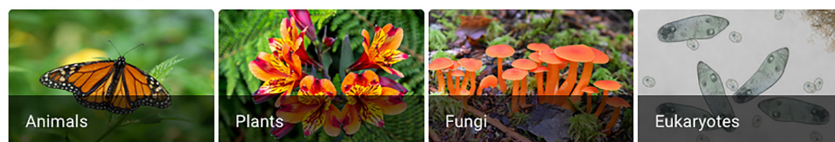
GitHub




Explore Datasets with our Python library and Jupyter notebooks.

Datasets API

Use our RESTful APIs to add functionality to your applications.

Browsing genome datasets



 Homo sapiens human	129 assemblies
 Mus musculus house mouse	22 assemblies
 Arabidopsis thaliana thale cress	30 assemblies

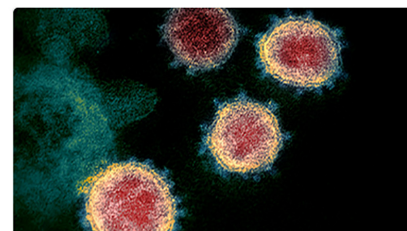
EDIT | DOWNLOAD | VIEW | SELECT COLUMNS

Gene ID	Symbol	Gene name	Chromosome	
<input type="checkbox"/>	6794	STK11	serine/threonine kinase 11	19
<input checked="" type="checkbox"/>	1499	CTNNB1	catenin beta 1	3
<input checked="" type="checkbox"/>	4089	SMAD4	SMAD family member 4	18
<input type="checkbox"/>	4436	MSH2	mutS homolog 2	2
<input checked="" type="checkbox"/>	207	AKT1	AKT serine/threonine kinase 1	14
<input type="checkbox"/>	11200	CHEK2	checkpoint kinase 2	22

Data tables

Build a table of genes or transcripts and choose from a variety of custom columns.

[GET STARTED](#)



Coronavirus datasets

Download SARS-CoV-2 genome and protein sequences, annotation and a data report for all complete genomes.

[GET DATA](#)

Figure 2. Landing page for the new NCBI Datasets product (<https://www.ncbi.nlm.nih.gov/datasets/>) that provides packaged downloads of genomic datasets using either a web interface, an API, or a UNIX/LINUX command-line tool.

Sequence Read Archive. NCBI maintains the NIH Sequence Read Archive (SRA), an archival database designed to support storage, retrieval and analysis of next-generation nucleotide sequence data. This archive now includes 8.8 Petabytes of publicly available data and another 4.6 Petabytes of controlled-access dbGaP data. While the archive holds tremendous promise for biomedical research, the sheer size of this dataset makes it difficult to store, retrieve and analyze. NCBI remains committed to continuing to expand SRA and improving access to the archive based on FAIR data principles.

As part of the NIH Science and Technology Research Infrastructure for Discovery, Experimentation and Sustainability Initiative (<https://datascience.nih.gov/strides/>), NCBI is now maintaining the entire SRA on two commercial cloud platforms, Amazon Web Services (AWS) and Google Cloud Platform (GCP). The cloud environment offers many advantages for the transfer and analysis of data (12), and the availability of SRA on cloud platforms should facilitate large-scale computational operations and collaborations. This idea has been supported by early pilot experiments in which thousands of metagenomic samples were analyzed for organismal content using cloud-adapted workflows and software (13). An additional example are COVID-focused datasets (including source and normalized SRA file formats) recently added to the AWS Public Dataset

Program. These datasets provide researchers easy, no-cost access to more than 13 000 SRA runs that include *Coronaviridae* content identified by a kmer-based approach using the SRA Taxonomy Analysis Tool.

We have made several additional updates that allow more effective use of cloud-hosted SRA data. By providing SRA run metadata and BioSample data on GCP BigQuery and AWS Athena, we give researchers more ways to identify sequence sets of interest. In addition to maintaining originally submitted source and SRA formatted data in the cloud, we are introducing normalized data formats that bin base quality scores to binary read assessments or align reads to references. These reduce file size and can expedite certain computations. We have also updated the SRA data location service and toolkit (<https://github.com/ncbi/sra-tools/wiki/02.-Installing-SRA-Toolkit>) to retrieve data from cloud locations in the desired data format.

Pathogens. The NCBI Pathogen Detection Project (<https://www.ncbi.nlm.nih.gov/pathogens/>) helps public health scientists investigate foodborne disease outbreaks by integrating pathogen genomic sequences obtained from cultured bacterial isolates and quickly clustering and identifying related sequences (14). As of August 2020, over 600 000 pathogen isolates covering 31 bacterial taxa and one emerging fungal pathogen, *Candida auris*, are actively be-

ing analyzed. We make these analysis results available in the Isolates Browser on a daily basis (<https://www.ncbi.nlm.nih.gov/pathogens/isolates>). As part of our effort to improve the compliance of the pipeline results (15) with FAIR principles, we annotate the generated assemblies using PGAP (16), submit them to GenBank and incorporate them into the NCBI Assembly resource. These assemblies link back to the Isolates Browser and also to the SNP Tree Viewer if that particular isolate is a member of a clonally related cluster. The Isolates Browser allows users to subset isolates and then download the assembled sequence and/or annotation of the submitted assemblies.

Antimicrobial resistance (AMR) resources. The Pathogen Detection team has continued to improve and release updated resources for antimicrobial resistance (AMR) (<https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/>). The Reference Gene Catalog (<https://www.ncbi.nlm.nih.gov/pathogens/isolates#/refgene/>) now provides a searchable and browsable interface for two sets of genes: a ‘core’ AMR reference set of acquired genes and proteins as well as point mutations conferring AMR, and a ‘plus’ set of genes related to stress responses (acid, metal, heat and biocide) and virulence. The 16 July 2020 release included 6428 total proteins (5588 AMR proteins, 210 stress response proteins and 630 virulence proteins) as well as 682 point mutations. We also released an updated version of the AMRFinder software (17) called AMRFinderPlus that uses the reference set above (<https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/AMRFinder/>). All isolates in the Pathogen Detection Isolates Browser other than *C. auris* are analyzed with AMRFinderPlus, and the three categories of genes (AMR, stress and virulence) are available in the Isolates Browser. Currently over 590 000 isolates have at least one identified AMR gene, over 510 000 have at least one identified stress response gene and over 210 000 have at least one identified virulence gene.

For the subset of isolates from the Isolates Browser that are both deposited in GenBank and that have genes identified by AMRFinderPlus, a new tabular viewer called the Microbial Browser for Genetic and Genomic Elements is available (MicroBIGG-E, <https://www.ncbi.nlm.nih.gov/pathogens/isolates#/microbigge/>). Every row in the MicroBIGG-E viewer displays a gene or point mutation that has been identified. This new interface provides easy access to the gene and contig sequences to facilitate further analyses. The Isolate Browser and MicroBIGG-E allow cross-browser selection, so that selections of isolates in the Isolate Browser enable selections of genes they encode in MicroBIGG-E and *vice versa*.

Genetics updates

ClinVar. ClinVar is an archive of submitted reports of relationships among human variations and phenotypes with supporting evidence. In December 2019, ClinVar reached the milestone of 1 million submitted records, representing more than half a million variants. To support ClinVar’s continued growth, we have improved the submission processing pipeline so that submitted data can be published faster. From October 2019 to August 2020, we added validation

to the ClinVar Submission Portal so that all of the required fields in a file submission are validated before the file is submitted to ClinVar. As of August 2020, submitters also have the option to stop a submission and correct errors immediately, or to submit all records that pass validation and receive a report of any failures to correct later.

In January 2020, we added a new feature to ClinVar that allows a user to ‘follow’ a particular variant and be notified if the overall clinical interpretation in ClinVar changes, for example from a pathogenic category to a non-pathogenic one. This feature makes it easier for a laboratory to become aware of variants that may need to be re-evaluated, and for clinicians to know when they should contact their clinical testing laboratory and/or patient with new information.

dbSNP. The Database of Single Nucleotide Polymorphisms (dbSNP) is a repository of human genomic variations and frequency data that includes both common and rare single-nucleotide variations and other small-scale variations. In 2020 dbSNP released the new NCBI Allele Frequency Aggregator (ALFA) dataset (<https://www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/>). We calculated the ALFA frequency data from 98 500 dbGaP subjects for whom genotypes were available. Aggregated from 551 billion genotypes, the results include allele counts and frequencies for 443 million known variations and 4 million novel ones. Future ALFA releases will include additional dbGaP studies, and we expect the dataset to expand to over a billion variants from millions of subjects. The ALFA data are available as part of the dbSNP regular release (https://ftp.ncbi.nih.gov/snp/latest_release/) and also as a separate download (https://ftp.ncbi.nih.gov/snp/population_frequency/latest_release/). The ALFA data are also accessible through an API (<https://www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/#api-queries>).

dbVar. In the past year, dbVar added 20 new human structural variation studies, bringing the total in the database to 190 studies containing 6 million regions and 36 million variants. Some notable studies are the Decipher dataset (<https://www.ncbi.nlm.nih.gov/dbvar/studies/nstd183>), gnomAD (<https://www.ncbi.nlm.nih.gov/dbvar/studies/nstd166/>) and the NCBI Curated Common Structural Variants dataset (<https://www.ncbi.nlm.nih.gov/dbvar/studies/nstd186>) that includes frequency data. dbVar continues to make it easier to find and use structural variation data by making selected datasets available on TrackHub for viewing in the NCBI GDV and other genome browsers, such as the UCSC browser. Tracks are available for the structural variants imported from ClinVar (https://www.ncbi.nlm.nih.gov/dbvar/content/clinvar_summary/#homepage) and for the NCBI Curated Common Structural Variants (nstd186).

BLAST updates

BLAST in the cloud. In the past year we have made some important changes to our BLAST databases. We released a new version of BLAST databases (v5) that are taxonomically aware, allowing users to limit a stand-alone BLAST+ search (18) by taxonomy using information stored in the database. To take advantage

of this feature, users will need a recent version of the BLAST+ package (2.9.0 release or later). The older BLAST database version (v4) has been deprecated. The new BLAST databases are available on the NCBI FTP site as well as on GCP and AWS cloud providers, and all three sites now offer the same 23 databases. These databases range from a *Betacoronavirus* collection (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&BLAST_SPEC=Betacoronavirus) to RefSeq representative genomes for eukaryotes, prokaryotes and viruses. Also included are the default NCBI nucleotide collection (nt), the non-redundant protein (nr) database, and genome assemblies for human and mouse. On AWS and GCP, users can access these databases with the BLAST+ Docker package described at https://github.com/ncbi/blast_plus_docs. Additionally, four databases based on the NCBI Targeted Loci Project (including collections of 18S, 28S and prokaryote 16S ribosomal RNA genes along with a database of fungal internal transcribed spacers) have been added to both the BLAST webpage and are also available as stand-alone databases on the NCBI FTP site and on GCP and AWS.

Primer-BLAST. Primer-BLAST (11) now allows users to design primers that are common for a group of highly similar sequences. This makes it easier for researchers to perform tasks such as amplifying multiple transcript variants for a single gene or detecting a group of highly related bacteria strains. Another new feature allows users to set the maximal nucleotide match at the 3' end of exon-exon junctions of transcripts to address non-specific polymerase chain reaction amplification.

IgBLAST. We have added several new features to IgBLAST (19). IgBLAST now supports annotations of the FWR4 region so that the entire V region can be annotated. Another new feature is that IgBLAST indicates if a sequence has a complete V(D)J region so that users can keep track of sequences with a full V region. Additional new functions include allowing users to extend J gene alignments at the 3' end and to analyze sequences from an organism of their choice.

Protein updates

NCBI released an updated version (2.19.0) of iCn3D (20), a three-dimensional (3D) molecular structure viewer that runs directly in web browsers. Interactive iCn3D views are embedded in structure summary pages of NCBI's Molecular Modeling Database (MMDB), and iCn3D visualizes the results of 3D structure comparisons computed by VAST+ as well as pairwise sequence-to-structure alignments computed by protein BLAST. iCn3D simultaneously displays 3D structures, 2D interaction schematics, alignments and protein/nucleotide sequences, as well as sequence annotations such as functional sites and conserved domain footprints. Sequence variations can be displayed for some human protein structures, and recently we have made sequence variation data accessible for selected structures of SARS-CoV-2 proteins. Other recently added features include extended visualization of 2D interaction networks

between proteins and ligands or other proteins, visualization of electrostatic potentials as computed by Delphi (21) and visualization of membrane bilayer location relative to membrane protein structures as provided by OPM (22). Finally, a Jupyter notebook widget version of iCn3D (called icn3dpy) enables users to view 3D structures in a Jupyter environment. iCn3D is available at <https://github.com/ncbi/icn3d>, and novel features are demonstrated on the gallery page at <https://www.ncbi.nlm.nih.gov/Structure/icn3d/icn3d.html#gallery>.

Chemical updates

PubChem (23–25) (<pubchem.ncbi.nlm.nih.gov>) is a public chemical data repository at NCBI. Over the past year, PubChem expanded the scope of its information content by integrating data from more than 50 new data sources. Notably, the World Intellectual Property Organization (WIPO) provided PubChem with more than 16 million chemical structures searchable in its patent database called PATENTSCOPE (<https://go.usa.gov/xdhfK>). In addition, SpringerMaterials provided links to hundreds of chemical and physical properties for more than 32 000 compounds, helping users to quickly locate articles for the property in question (<https://go.usa.gov/xvqfq>). Another new source of data came from ToxNet, a collection of NLM databases that provided a wide range of toxicological information. These databases were retired last year and their content was integrated into PubChem (<https://go.usa.gov/xfwyU>). They include the Genetic Toxicology Data Bank (GeneTox), the Chemical Carcinogenesis Research Information System (CCRIS), the Hazardous Substances Data Bank (HSDB), ChemIDplus, LactMed and LiverTox. Finally, in response to the COVID-19 pandemic, PubChem created a special data collection containing data related to COVID-19 and SARS-CoV-2 (<https://go.usa.gov/xfwmG>). We gathered the data in this collection from authoritative and curated sources, and a link to the collection appears on the PubChem home page.

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 <https://support.nlm.nih.gov>. Updates on NCBI resources and database enhancements are described on the NCBI Insights blog (<https://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.

FUNDING

Funding for open access charge: National Institutes of Health. Funding to pay the Open Access publication charges for this article was provided by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

- Sayers, E.W., Beck, J., Brister, J.R., Bolton, E.E., Canese, K., Comeau, D.C., Funk, K., Ketter, A., Kim, S., Kimchi, A. *et al.* (2020) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **48**, D9–D16.
- 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.
- Sayers, E.W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K.D. and Karsch-Mizrachi, I. (2020) GenBank. *Nucleic Acids Res.*, **48**, D84–D86.
- Amid, C., Alako, B.T.F., Balavenkataraman Kadhirvelu, V., Burdett, T., Burgin, J., Fan, J., Harrison, P.W., Holt, S., Hussein, A., Ivanov, E. *et al.* (2020) The European nucleotide archive in 2019. *Nucleic Acids Res.*, **48**, D70–D76.
- Ogasawara, O., Kodama, Y., Mashima, J., Kosuge, T. and Fujisawa, T. (2020) DDBJ database updates and computational infrastructure enhancement. *Nucleic Acids Res.*, **48**, D45–D50.
- 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.
- 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.
- Liu, W., Islamaj Dogan, R., Kim, S., Comeau, D.C., Kim, W., Yeganova, L., Lu, Z. and Wilbur, W.J. (2014) Author Name Disambiguation for PubMed. *J. Assoc. Inf. Sci. Technol.*, **65**, 765–781.
- Yeganova, L., Kim, S., Chen, Q., Balasanov, G., Wilbur, W.J. and Lu, Z. (2020) Better synonyms for enriching biomedical search. *J. Am. Med. Inform. Assoc.*, in press.
- Boratyn, G.M., Camacho, C., Cooper, P.S., Coulouris, G., Fong, A., Ma, N., Madden, T.L., Matten, W.T., McGinnis, S.D., Merezuk, Y. *et al.* (2013) BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.*, **41**, W29–W33.
- Ye, J., Coulouris, G., Zaretskaya, L., Cutcutache, I., Rozen, S. and Madden, T.L. (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, **13**, 134.
- Langmead, B. and Nellore, A. (2018) Cloud computing for genomic data analysis and collaboration. *Nat. Rev. Genet.*, **19**, 208–219.
- Connor, R., Brister, R., Buchmann, J.P., Deboutte, W., Edwards, R., Marti-Carreras, J., Tisza, M., Zalunin, V., Andrade-Martinez, J., Cantu, A. *et al.* (2019) NCBI's virus discovery Hackathon: Engaging research communities to identify cloud infrastructure requirements. *Genes (Basel)*, **10**, 714.
- NCBI Resource Coordinators. (2017) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **45**, D12–D17.
- Sayers, E.W., Agarwala, R., Bolton, E.E., Brister, J.R., Canese, K., Clark, K., Connor, R., Fiorini, N., Funk, K., Hefferon, T. *et al.* (2019) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **47**, D23–D28.
- Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E.P., Zaslavsky, L., Lomsadze, A., Pruitt, K.D., Borodovsky, M. and Ostell, J. (2016) NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.*, **44**, 6614–6624.
- Feldgarden, M., Brover, V., Haft, D.H., Prasad, A.B., Slotta, D.J., Tolstoy, I., Tyson, G.H., Zhao, S., Hsu, C.H., McDermott, P.F. *et al.* (2019) Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance Genotype-Phenotype correlations in a collection of Isolates. *Antimicrob. Agents Chemother.*, **63**, e00483-19.
- 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.
- 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.
- Wang, J., Youkharibache, P., Zhang, D., Lanczycki, C.J., Geer, R.C., Madej, T., Phan, L., Ward, M., Lu, S., Marchler, G.H. *et al.* (2020) iCn3D, a web-based 3D viewer for sharing 1D/2D/3D representations of biomolecular structures. *Bioinformatics*, **36**, 131–135.
- Li, L., Li, C., Sarkar, S., Zhang, J., Witham, S., Zhang, Z., Wang, L., Smith, N., Petukh, M. and Alexov, E. (2012) DelPhi: a comprehensive suite for DelPhi software and associated resources. *BMC Biophys.*, **5**, 9.
- Lomize, M.A., Pogozheva, I.D., Joo, H., Mosberg, H.I. and Lomize, A.L. (2012) OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res.*, **40**, D370–D376.
- Kim, S. (2016) Getting the most out of PubChem for virtual screening. *Expert. Opin. Drug Discov.*, **11**, 843–855.
- 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.
- Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B. *et al.* (2019) PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.*, **47**, D1102–D1109.