Published online 28 November 2015 *Nucleic Acids Research, 2016, Vol. 44, Database issue* ***D7–D19***
doi: 10.1093/nar/gkv1290

Database resources of the National Center for Biotechnology Information

NCBI Resource Coordinators[†]

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received September 16, 2015; Revised November 4, 2015; Accepted November 5, 2015

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 for published life science journals. Additional NCBI resources focus on literature (PubMed Central (PMC), Bookshelf and PubReader), health (ClinVar, dbGaP, dbMHC, the Genetic Testing Registry, HIV-1/Human Protein Interaction Database and MedGen), genomes (BioProject, Assembly, Genome, BioSample, dbSNP, dbVar, Epigenomics, the Map Viewer, Nucleotide, Probe, RefSeq, Sequence Read Archive, the Taxonomy Browser and the Trace Archive), genes (Gene, Gene Expression Omnibus (GEO), HomoloGene, PopSet and UniGene), proteins (Protein, the Conserved Domain Database (CDD), COBALT, Conserved Domain Architecture Retrieval Tool (CDART), the Molecular Modeling Database (MMDB) and Protein Clusters) and chemicals (Biosystems and the PubChem suite of small molecule databases). The Entrez system provides search and retrieval operations for most of these databases. Augmenting many of the web applications are custom implementations of the BLAST program optimized to search specialized datasets. All of these resources can be accessed through the NCBI home page at www.ncbi.nlm.nih.gov.

INTRODUCTION

The National Center for Biotechnology Information (NCBI) at the National Institutes of Health was created in 1988 to develop information systems for molecular biology. In addition to maintaining the GenBank® (1) nucleic acid sequence database, which receives data through an international collaboration with the DNA Data Bank of Japan (DDBJ) and the European Nucleotide Archive (ENA) as well as from the scientific community, NCBI provides many other kinds of biological data as well as retrieval systems and computational resources for the analysis of GenBank and other data. This article provides a summary of recent developments, including both new and updated resources, followed by an introduction to the Entrez system and a brief review of the suite of NCBI resources. All resources discussed are available through the NCBI home page at www.ncbi.nlm.nih.gov and can also be located using the *NCBI Web Site* database available in Entrez search menus. In most cases, the data underlying these resources and executables for the software described are available for download at ftp.ncbi.nlm.nih.gov.

RECENT DEVELOPMENTS

PubMed labs

PubMed Labs is an NCBI initiative for developing experimental projects by involving the user community throughout the process. Projects in PubMed Labs may be early versions of new services, proposed features for existing resources, or novel database content. PubMed Labs projects will be described on the NCBI Insights blog (ncbiinsights.ncbi.nlm.nih.gov) in a new category of posts labeled *PubMed Labs*. Each post will describe the project and provide instructions for how users can test its functions, and will also indicate what results to expect. We encourage users to share their experiences with us by commenting on these posts. Currently there are two initial projects in PubMed Labs: SmartBLAST and PubMed Also-Viewed.

SmartBLAST is an experimental tool that makes it easier to accomplish common protein sequence analysis tasks such as finding a candidate name for a protein, identifying highly conserved regions and locating segments that are present in closely related database sequences but that are missing from the query. SmartBLAST does this by performing two parallel BLASTp searches: one that retrieves the closest matching sequences available, and another that finds

To whom correspondence should be addressed. Eric W. Sayers. Tel: +1 301 496 2475; Fax: +1 301 480 9241; Email: sayers@ncbi.nlm.nih.gov
[†]The members of the NCBI Resource Coordinators group are listed in the Appendix.

the closest matches from well-annotated sequences from model organisms. The tool then constructs a multiple sequence alignment between the query and five of the closest matches, and displays this as a phylogenetic tree. SmartBLAST accomplishes all of this in much less time than it takes to run a typical BLASTp search. Links to SmartBLAST may be found on the main BLAST page as well as on BLASTp results pages.

PubMed Also-Viewed is a link on some PubMed abstract pages that shows PubMed records that other users have viewed with the current record. When present, the link is labeled *Articles frequently viewed together* and appears on the right side of the abstract page. At the time of this writing the link was available for about 5% of PubMed records.

Revised NCBI home page

In 2014 NCBI revised the main home page to include six prominent buttons that lead to pages focused on a particular set of services. The *Submit* button loads a page that helps submitters choose the correct mechanism(s) for submitting data to NCBI, while the *Download* page provides access to the FTP site and related tools. The *Learn* page introduces users to a variety of NCBI educational resources and programs, while the *Analyze* page leads to software tools developed by NCBI. Intended for software developers, the *Develop* page links to the several NCBI APIs and software toolkits, as well as to the NCBI GitHub repository. Finally, the *Research* page allows users to explore NCBI research projects and collaborations, along with NCBI research groups and associated staff. Most of these pages also feature a *Feedback* button so that users can easily provide comments and suggestions.

dbGaP data browser

A new addition to the Database of Genotypes and Phenotypes (dbGaP) is the dbGaP Data Browser (DDB; https://www.ncbi.nlm.nih.gov/gap/ddb/) that enables users to explore variant calls, genotype calls and supporting sequence read alignments for controlled access datasets in a genomic context. It is a companion to the dbGaP Beacon resource and provides a graphical interface for reviewing its results. The interface is modeled after the 1000 Genomes project browser; however, its content represents a combination of data from the dbGaP controlled access and public databases. Depending upon authentication credentials, users can access read-only views of selected datasets or download datasets to which they have been granted access. Within the browser, the *Subjects* widget supports faceted and free-text searching for browser tracks of Sequence Read Archive (SRA) alignments for samples of interest. Sample attributes such as the study accession, BioSample ID, sex, population, tumor/normal status and the availability of sample genotype data are just a few of the available search facets. The browser's graphical display allows users to view the read alignments from selected study participants alongside data from ClinVar, the Database of Single Nucleotide Polymorphisms (dbSNP) and Gene. The *Genotypes* table in the browser provides access to individual-level and aggregate General Research Use (GRU) dataset genotypes for studies having samples whose SRA alignments are displayed in the graphical view. Within the table, users can view frequencies or counts for either alleles or samples. DDB also includes functions found in other NCBI browsers, including feature search and support for user data uploads. Further documentation for DDB is available online (https://www.ncbi.nlm.nih.gov/gap/ddb/help/).

Pathogen detection

The NCBI Pathogen Detection project (www.ncbi.nlm.nih.gov/pathogens/) is a new system that facilitates real-time surveillance of bacterial pathogens and foodborne disease. This project is a collaboration with several public health agencies including the Centers for Disease Control (CDC), the Food and Drug Administration (FDA), the United States Department of Agriculture Food Safety and Inspection Service (USDA-FSIS), Public Health England (PHE) and several state and regional labs in the US. Samples collected by these agencies are sequenced and submitted to NCBI where an automated pipeline clusters and identifies the sequences and then quickly reports the results back. Collecting and analyzing pathogen sequence data obtained from food, the environment and human patients reveals potential sources of contamination and facilitates traceback investigations and responses to outbreaks. Currently the project focuses on the following bacterial groups: *Campylobacter, Listeria, Salmonella* and the combination of *Escherichia coli* and *Shigella.* Analysis results are available from the NCBI FTP site (ftp.ncbi.nlm.nih.gov/pathogen/).

PubMed updates

In response to user feedback, minor changes were made to the PubMed interface so that popular features are easier to find. The *Related citations* feature was renamed to *Similar articles* (the algorithm to generate the results was not modified). The *Save search* and *RSS* links that allow users to create My NCBI automatic email alerts were renamed to *Create alert* and *Create RSS*. Additionally, to provide a less-cluttered PubMed results display, the status tag lines (e.g. [PubMed—as supplied by publisher]) were removed from the summary results. The abstract display was also enhanced to show the transliterated title for citations originally published in a non-English language that do not include an English title.

PubMed Mobile was updated with a number of styling modifications and enhancements including a *Trending articles* feature on the home page. PubMed Mobile summary results now include the *Related searches* discovery tool as well as sorting options and additional filter selections. These Discovery tools appear below the results on mobile devices with smaller screen sizes. A *Show full citation* link was added to streamline the PubMed Mobile abstract page, and clicking this link displays the citation, author(s) and affiliation details. Clicking a linked author name in the resulting list displays a sorted set of citations for that author.

Updates to PubMed Central (PMC)

In response to requests from authors and users, PubMed Central (PMC) added several new features in the last year.

Among these features is the new citation exporter that makes it easy to retrieve pre-formatted citations in AMA, MLA or APA styles that users can then copy, paste or download into bibliographic reference manager software. Additionally, in order to facilitate text and data mining for articles in the Open Access Subset, PMC is now providing plain text files for these articles on the PMC FTP site. The files contain the full text of the article, extracted either from the XML or PDF source files.

PMC is also now serving as the public access repository for a number of federal agencies in addition to NIH that support scientific research. As of January 2015, the Centers for Disease Control and Prevention (CDC), the National Institute of Standards and Technology (NIST) and the Department of Veterans Affairs (VA) have implemented public access policies requiring researchers who are supported by these agencies to make the resulting manuscripts publicly available in PMC within 12 months of publication. The NIH manuscript submission (NIHMS) system has been extended to support researchers from these additional agencies.

SciENcv updates

My NCBI (www.ncbi.nlm.nih.gov/account/) provides a specialized biosketch tool called SciENcv (Science Experts Network Curriculum Vitae) for users who wish to create, download and share biosketches for NIH grant applications. SciENcv was updated to support the new NIH Biographical Sketch format that became mandatory for grant applications with due dates of May 2015 or later. SciENcv can upgrade biosketches stored in the previous NIH format to the new version and it also supports the NSF biosketch format on an alpha-release basis. SciENcv is integrated with the My Bibliography citation tool, which is required for NIH extramural grantees to demonstrate compliance with the NIH Public Access Policy. Among other improvements to the user interface, the process used to describe scientific accomplishments has been enhanced to allow easier import of citations.

Updates to medical genetics resources

GTR. The Genetic Testing Registry (GTR) collects and displays information that has been submitted by providers about their genetic tests (2). GTR accepts submissions for germline, somatic and research tests. Submission formats have been expanded from interactive wizards and Excel templates to include submission of tests as XML files. Submitted data appear on the GTR web site within 24–48 h (http://www.ncbi.nlm.nih.gov/gtr/docs/fulltest/). GTR submitters are required to review their laboratory and test data annually (www.ncbi.nlm.nih.gov/gtr/docs/annual_review/). To support automated test updates, submitters can download all of their submitted test data to an Excel template file to edit and upload. Records that have not been reviewed within a year of the previous review are marked *out of date* on the GTR site, and records that have not been reviewed in two years will be removed from display. This stringency in representing current information may result in differences in test counts between GTR and other public websites.

In addition to providing an advanced search to find tests by a variety of attributes, GTR now provides an *All GTR* display that not only allows users to interrogate content about tests, conditions, genes and laboratories simultaneously, but also organizes the data in each domain to show high-value information. For example, the *Tests* tab shows the name of the test, the name and location of the lab and the methods used, along with links to view details about conditions and test targets. The *Conditions* tab lists symptoms to support recognition and links to all tests, genes and *GeneReviews* for each condition. The *Genes* tab displays associated conditions for each gene and provides links to tests for each gene. Each of these tabs also allows users to select multiple items (e.g. genes, conditions, laboratories) and retrieve associated tests.

Genome updates

NCBI has continued to revise the genome area of the NCBI FTP site (ftp.ncbi.nlm.nih.gov/genomes/). The new directories within the genome area (*genbank*, *refseq* and *all*) now provide a standard set of data files for over 54 000 assemblies. The *genbank* directory contains data submitted directly to GenBank (or an INSDC database), while the *refseq* directory contains data that are part of the Reference Sequence (RefSeq) project. The common data unit within these three directories is a subdirectory corresponding to a record in the Assembly database (3) and given a name consisting of the Assembly accession followed by the name of the assembly. For example, the human GRCh38 release has assembly accession GCF_000001405.26, and so the subdirectory is named GCF_000001405.26_GRCh38. Users are encouraged to search the Assembly database directly to find these accessions, assembly names and other details about the datasets. The *genbank* and *refseq* directories collect the Assembly subdirectories within broad taxonomic directories (e.g. plant, bacteria and vertebrate_mammalian) and also directories for each species. Each Assembly subdirectory contains a standard set of files including FASTA and GenBank/GenPept data for genome, transcript and protein sequences, along with GFF3 and feature table files for annotated genome records. Older directories in the genome FTP area that are no longer being updated will be moved to an archive area by late 2015.

Gene updates

In response to the rapidly growing number of prokaryotic genomes being submitted to NCBI, the scope of the Gene database changed in 2014. Going forward, NCBI will create Gene records only for *reference* and selected *representative* genomes for a single prokaryotic species (www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/). Gene records for strains not included in these sets are being discontinued, and their record pages will contain messages providing more detail about each case (www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/faq/).

Protein updates

Protein database records now have a prominent link to the *Identical Protein Report* at the top of the record page. This

report displays the accessions of all other protein records whose sequences are identical to that of a given protein. The report also provides links to the CDS sequence in Nucleotide for each protein. While this report is available for all proteins, it is especially useful for the non-redundant RefSeq WP sequences introduced in 2013 (4). Because many WP sequences represent a set of identical proteins that may not have separate species- or strain-specific records, these WP sequences are connected to not one but a corresponding set of Nucleotide CDS sequences. The *Identical Protein Report* clarifies these relationships. For example, WP_002317106 collects over 40 thioredoxin sequences from several species and strains of *Enterococcus*. The report is also available through the E-utility EFetch with *&rettype = ipg*.

BLAST updates

A new search box on the BLAST home page makes it easy to find a genomic BLAST search page for a given organism. The box has an autocomplete feature that presents suggestions when a user starts to type. Choosing an organism loads a BLAST search page with the best genomic database for that organism preselected. The search box also produces metagenomic and microbial suggestions.

MOLE-BLAST is a new tool that classifies multiple nucleotide query sequences and displays their relationships. Ideal input for this tool is a set of sequences representing a specific locus from a group of organisms rather than the entire genome of an organism or a set of unannotated contigs. Example input would be 16S sequences from different bacteria or ITS sequences from fungi. MOLE-BLAST first assigns each query in the input set to a cluster using BLAST, thereby grouping the queries by locus. Second, it performs a database search to find top matches for each query. Third, it computes a multiple alignment (using MUSCLE) between the queries and their top matches, and presents this analysis as a phylogenetic tree.

PubChem updates

Both the PubChem Compound and Substance record view pages were completely redesigned in the past year. The new pages use a responsive design approach that optimizes the display on a variety of screen sizes, including both touch- and mouse-based interfaces. These reports include an improved and expanded table of contents that makes navigation easier and users can now bookmark particular sections of the reports. Full details of the many improvements are discussed in posts on the PubChem blog (pubchemblog.ncbi.nlm.nih.gov).

THE ENTREZ SYSTEM

Entrez databases

Entrez (5) is an integrated database retrieval system that provides access to a diverse set of 39 databases that together contain 1.7 billion records (Table 1). Links to the web portal for each of these databases are provided on the Entrez GQuery page (www.ncbi.nlm.nih.gov/gquery/). Entrez supports text searching using simple Boolean queries, downloading of data in various formats and linking of records between databases based on asserted relationships. In their simplest form, these links may be cross-references between a sequence and the abstract of the paper in which it is reported, or between a protein sequence and either its coding DNA sequence or its 3D-structure. Computationally derived links between neighboring records, such as those based on computed similarities among sequences or among PubMed abstracts, allow rapid access to groups of related records. Several popular links are displayed as Discovery Components in the right column of Entrez search result or record view pages, making these connections easier to find and explore. The LinkOut service expands the range of links to include external resources, such as organism-specific genome databases. The records retrieved by Entrez can be displayed in many formats and downloaded singly or in batches.

Data sources and collaborations

NCBI receives data from three sources: direct submissions from external investigators, national and international collaborations or agreements with data providers and research consortia, and internal curation efforts. The *Data Source* column in Table 1 indicates those mechanisms by which each Entrez database receives data. More information about the various collaborations, agreements and curation efforts are available through the home pages of the individual resources.

Entrez programming utilities (E-Utilities)

The Entrez Programming Utilities (E-Utilities) constitute the Application Programming Interface (API) for the Entrez system. The API includes nine programs that support a uniform set of parameters used to search, link and download data from the Entrez databases. EInfo provides basic statistics on a given database, including the last update date, along with lists of all search fields and available links. ESearch returns the identifiers of records that match an Entrez text query and when combined with EFetch or ESummary, provides a mechanism for downloading the corresponding data records. ELink gives users access to the vast array of links within Entrez so that data related to an input set can be retrieved. By assembling URL calls to the E-utilities within simple scripts, users can create powerful applications to automate Entrez functions to accomplish batch tasks that are impractical using web browsers. Detailed documentation for using the E-Utilities is available at eutils.ncbi.nlm.nih.gov.

Entrez Direct is a set of executables that provides an interface to the E-utilities on the UNIX command line. These executables are designed so that the output of one can be passed directly as input to another using the UNIX pipe ('|'). In this way, it is straightforward to implement a diverse assortment of workflows. Entrez Direct also offers a utility named *xtract* that parses the XML output of ESummary and EFetch calls so that individual fields within records can be retrieved and formatted into custom tables, especially when combined with standard UNIX commands such as *grep*, *sort*, *cut*, *awk* or *sed*. Complex workflows can be conveniently saved as shell scripts for sharing or use

Table 1. The Entrez Databases (as of 1September 2015)

Database	Records	Section within this article	Data source[1]
Site Search	21 929	Introduction	N
PubMed	25 235 441	Literature	C
PubMed Central	3 633 245	Literature	D, C
NLM Catalog	1 530 854	Literature	C, N
MeSH	259 099	Literature	N
Books	446 888	Literature	C, N
MedGen*	272 979	Health	C, N
dbGaP	207 859	Health	D
ClinVar*	124 971	Health	D, N
PubMed Health	55 244	Health	C
GTR*	31 991	Health	D
SNP*	705 483 355	Genomes	D (dbSNP), N
Nucleotide*	199 827 994	Genomes	D (GenBank), C, N
GSS*	39 394 513	Genomes	D (GenBank)
Clone*	37 336 118	Genomes	D, N
Probe	32 379 570	Genomes	D
dbVar*	4 481 341	Genomes	D
BioSample	3 648 667	Genomes	D
SRA*	1 697 236	Genomes	D
Taxonomy*	1 426 896	Genomes	C, N
BioProject*	152 290	Genomes	D
Assembly*	59 566	Genomes	C, N
Genome*	13 532	Genomes	C, N
Epigenomics*	7789	Genomes	D
GEO Profiles*	108 708 851	Genes	D
EST*	75 992 479	Genes	D (GenBank)
Gene*	21 399 200	Genes	C, N
UniGene*	6 473 284	Genes	N
GEO Datasets*	1 645 202	Genes	D
PopSet*	231 877	Genes	D (GenBank)
Homologene*	141 268	Genes	N
Protein*	223 456 488	Proteins	C, N
Protein Clusters*	820 546	Proteins	N
Structure*	111 186	Proteins	C, N
CDD*	50 648	Proteins	C, N
PubChem Substance*	157 362 091	Chemicals	D
PubChem Compound*	60 774 418	Chemicals	N
PubChem Bioassay*	1 154 363	Chemicals	D
Biosystems*	805 473	Chemicals	C

[1]D = direct submission; C = collaboration/agreement; N = internal NCBI/NLM curation.
*Indicates that the data in this resource are available by FTP.

by other applications. Extensive documentation is available (www.ncbi.nlm.nih.gov/books/NBK179288/) that includes full descriptions of the many options and numerous examples spanning a wide variety of NCBI resources.

LITERATURE

PubMed

The PubMed database contains citations from life science journals, many of which include abstracts and links to their full text articles.

PubMed commons

PubMed Commons enables the community to share information and opinions on scientific publications. Any author of a publication indexed in PubMed is eligible to join PubMed Commons, and members may comment on any publication in PubMed. Comments appear below the publication's abstract, and are regularly monitored for adherence to guidelines (www.ncbi.nlm.nih.gov/pubmedcommons/help/guidelines/). Comments are citable and may be shared or adapted, with attribution, under a Creative Commons license (creativecommons.org/licenses/by/3.0/us/).

PubMed Central (PMC)

PMC (6) contains the full text of peer-reviewed journal articles in the life sciences, and is the repository for all manuscripts arising from NIH and other federal research funds (e.g. CDC, VA, NIST) that are submitted through the NIH manuscript submission system (NIHMS). Journals that have PMC-participation agreements provide free access to full-text articles in PMC either immediately after publication or after a set embargo period. Manuscripts that fall under the public access policies of participating funders must be made available in PMC within 12 months of publication. PMC articles are available as either HTML or PDF documents, or can be read using the PubReader viewer.

NLM catalog

The NLM Catalog contains bibliographic data for the various items in the NLM collections, including jour-

nals, books, audiovisuals, computer software, electronic resources and other materials.

Medical subject headings (MeSH)

The MeSH database (7) includes information about the NLM controlled vocabulary thesaurus used for indexing PubMed citations, and provides an interface for constructing PubMed queries using MeSH terms.

NCBI bookshelf

The NCBI Bookshelf is an online service of the National Library of Medicine Literature Archive (NLM LitArch) that provides free access to the full text of books, reports, databases and documentation in the life sciences and health care fields.

HEALTH

ClinVar

ClinVar supports users who want to determine what has been reported about the medical relevance of human sequence variation (8). ClinVar provides two major displays: a Record Report that aggregates submitted interpretations of a variation and a condition and a Variation Report that organizes information about each variant. In both views, ClinVar aggregates data from multiple submitters to make it easier to evaluate the current status of interpretation. ClinVar records maintain connections with dbSNP, dbVar, Gene, MedGen, and PubMed using Entrez links, and are accessible as annotations on chromosome and RefSeqGene sequences. They are also included in the Variation Viewer tool. ClinVar continues to add functions to facilitate retrieval, such as a query (http://www.ncbi.nlm.nih.gov/clinvar?term=%22gene%20acmg%20incidental%202013%22[Properties]) to retrieve all records of variants in genes for which investigators should report incidental findings as recommended by the American College of Medical Genetics and Genomics (9). As a partner of the ClinGen project, ClinVar encourages domain experts to apply for recognition as an expert panel (www.ncbi.nlm.nih.gov/clinvar/docs/expert_panel/) and submit their interpretations of human variants. ClinVar offers several options for submission, from simple spreadsheets to comprehensive XML files (www.ncbi.nlm.nih.gov/clinvar/docs/submit). ClinVar data are freely available for download from the website, by FTP as VCF or XML files or using the E-utilities.

MedGen

MedGen organizes information about human disorders that have a genetic component (www.ncbi.nlm.nih.gov/books/NBK159970/). Starting from freely available content in the semi-annual releases from UMLS (www.nlm.nih.gov/research/umls/), MedGen adds recent content from OMIM, terms and relationships from the Human Phenotype (www.human-phenotype-ontology.org/) and ORDO (www.orphadata.org/cgi-bin/inc/ordo_orphanet.inc.php) ontologies, and terms submitted from ClinVar and GTR. MedGen organizes terms from multiple sources by assigning them a concept ID, and then adds value by reporting practice guidelines, related genes from the Gene database, variants in ClinVar and available tests in GTR. MedGen supports querying for disorders that share clinical features as well as drugs and their responses. MedGen data can be downloaded using FTP as pipe-delimited (RFF) or CSV text files and using the E-utilities.

dbGaP

The Database of Genotypes and Phenotypes (dbGaP) (10) archives, distributes and supports submission of data that correlate genomic characteristics with observable traits. This database is a designated NIH repository for NIH-funded genome-wide association study (GWAS) results (grants.nih.gov/grants/gwas/). To protect the confidentiality of study subjects, dbGaP accepts only de-identified data and requires investigators to go through an authorization process in order to access individual-level data. Study documents, protocols and subject questionnaires are available without restriction.

PubMed health

PubMed Health provides information for consumers and clinicians about the prevention and treatment of diseases and conditions. The database specializes in reviews of clinical effectiveness research, containing both summaries for consumers and full technical reports.

dbMHC, dbLRC, dbRBC

NCBI maintains three databases for routine clinical applications: dbMHC, dbLRC and dbRBC. dbMHC focuses on the Major Histocompatibility Complex (MHC) and contains sequences and frequency distributions for MHC alleles, as well as genotype and clinical outcome information on hematopoietic cell transplants performed worldwide. dbLRC offers a comprehensive collection of alleles of the Leukocyte Receptor Complex (LRC) with an emphasis on KIR genes. dbRBC provides data on genes for Red Blood Cell (RBC) antigens along with access to the International Society of Blood Transfusion allele nomenclature of blood group alleles. dbRBC also hosts the Blood Group Antigen Gene Mutation Database (11) and integrates it with resources at NCBI. All three databases provide multiple sequence alignments, analysis tools to interpret homozygous or heterozygous sequencing results (12) and tools for DNA probe alignments.

GENOMES

BioProject

The BioProject database is a central access point for metadata about research projects whose data are deposited in an INSDC database. BioProject provides links to the primary data from these projects, which range from focused genome sequencing projects to large international collaborations. These larger projects may have multiple sub-projects incorporating experiments that produce nucleotide sequence sets,

genotype/phenotype data, sequence variants or epigenetic information.

Assembly

The Assembly database (3) collects metadata about genome assemblies that were either submitted to GenBank (or an INSDC database) or that are part of the RefSeq database. Assembly records also provide statistics about the genome as well as links to the sequence data in Entrez or in the genomes area of the FTP site.

Genome

The Genome database collects genomic sequencing projects for a given species and provides links to corresponding records in BioProject, Assembly, Nucleotide and Protein. Genome records collect genome assembly data at various levels of completion, ranging from genomes represented by scaffolds or contigs to fully assembled chromosomes with annotation. NCBI creates a Genome record for an organism if at least one assembly is available for that organism in the Assembly database. The Genome home page also provides links to an organism browser that lists the current status of all genomes annotated at NCBI.

RefSeq

The RefSeq database (13) is a non-redundant set of curated and computationally derived sequences for transcripts, proteins and genomic regions. RefSeq DNA and RNA sequences can be searched and retrieved from the Nucleotide database and the complete RefSeq collection is available in the RefSeq directory on the NCBI FTP site.

GenBank

GenBank (1) is the primary nucleotide sequence archive at NCBI and is a member of the International Nucleotide Sequence Database Collaboration (INSDC). Sequences from GenBank are available from three Entrez databases: Nucleotide, EST and GSS (specified as nuccore, nucest and nucgss within the E-utilities). The Nucleotide database contains all GenBank sequences except those within the EST or GSS GenBank divisions. The database also contains WGS sequences, Third Party Annotation (TPA) sequences and sequences imported from the Structure database.

PopSet

The PopSet database is a collection of related sequences and alignments derived from population, phylogenetic, mutation and ecosystem studies that have been submitted to GenBank. When available, PopSet alignments are shown in an embedded viewer on the PopSet record page.

Sequence Read Archive (SRA)

SRA (14) is a repository for raw sequence reads and alignments generated by high-throughput nucleic acid sequencers. Data are deposited into SRA as supporting evidence for a wide range of study types, including de novo genome assemblies, genome wide association studies, single nucleotide polymorphism and structural variation analysis, pathogen identification, transcript assembly, metagenomic community profiling and epigenetics studies.

Trace archive

The Trace Archive contains sequence traces from gel and capillary electrophoresis sequencers. These data arise from whole genomes of pathogens, organismal shotgun and BAC clone projects, and EST libraries. A companion resource, the Trace Assembly Archive, contains placements of individual trace reads on a GenBank sequence.

Clone database (CloneDB)

CloneDB is a resource for finding descriptions, sources, map positions and distributor information about available clones and libraries (15). For both genomic and cell-based clones and libraries, CloneDB contains information about the sequences themselves, such as their genomic mapping positions and associated markers, along with details about how the libraries were constructed.

Probe

The Probe database is a registry of nucleic acid reagents designed for use in a wide variety of biomedical research applications including genotyping, SNP discovery, gene expression, gene silencing and gene mapping. Probe also includes information on reagent distributors, probe effectiveness and computed sequence similarities.

BioSample

The BioSample database provides annotation for biological samples used in a variety of studies submitted to NCBI, including genomic sequencing, microarrays, genome wide association studies (GWAS) and epigenomics (16). The database promotes the use of structured and consistent attribute names and values that describe what the samples are as well as information about their provenance, where appropriate.

Taxonomy

The NCBI taxonomy database is a central organizing principle for the Entrez biological databases and provides links to all data for each taxonomic node, from superkingdoms to subspecies (17). The taxonomy database reflects sequence data from virtually all of the formally described species of prokaryotes and about 10% of the eukaryotes. The Taxonomy Browser can be used to view the taxonomy tree or retrieve data from any of the Entrez databases for a particular organism or group. In 2013 the Taxonomy database began including type material for prokaryotic type strains and eukaryotic type specimens (18). As of January 2014 NCBI no longer assigns taxonomy IDs to bacterial strains that do not already have taxonomy IDs (19). Instead, such sequences will be assigned the taxonomy ID of the bacterial species, while the strain will be included in the source information

of the sequence record. In addition, the sequence record will be linked to a BioSample record that will contain strain information such as relevant culture collections and details about how the strain was isolated.

Genome reference consortium (GRC)

The Genome Reference Consortium (GRC) (www.genomereference.org) is an international collaboration between the Wellcome Trust Sanger Institute, the Genome Institute at Washington University, EMBL and NCBI. The GRC aims to produce assemblies of higher eukaryotic genomes that best reflect complex allelic diversity that is consistent with currently available data. The GRC currently produces assemblies for human (GRCh38), mouse (GRCm38) and zebrafish (GRCz10). Between major assembly releases the GRC provides minor patch releases that provide additional sequence scaffolds that either correct errors in the assembly (fix patches) or add an alternate loci (novel patches). GRC staff then incorporate these changes into the next major assembly release. GRC data are available for download from the NCBI FTP site (ftp.ncbi.nlm.nih.gov/pub/grc/) and the new genomes FTP area (see above).

dbSNP

The Database of Single Nucleotide Polymorphisms (dbSNP) (20) is a repository of all types of short genetic variations <50 bp in length, and so is a complement to dbVar (see below). dbSNP accepts submissions of both common and polymorphic variations, and contains both germline and somatic variations. In addition to archiving molecular details for each submission and calculating submitted variant locations on each genome assembly, dbSNP maintains information about population-specific allele frequencies and genotypes, reports the validation state of each variant and indicates if a variation call may be suspect because of paralogy (21).

dbVar

The Database of Genomic Structural Variation (dbVar) is an archive of large-scale genomic variants (generally >50 bp) such as insertions, deletions, translocations and inversions (22). These data are derived from several methods including computational sequence analysis and microarray experiments.

Variation viewer

The Variation Viewer (www.ncbi.nlm.nih.gov/variation/view) displays human variations from dbSNP, dbVar and ClinVar in the context of the current and previous human reference genome assemblies, now GRCh38 and GRCh37 (23). Variation Viewer provides users with genome-wide access to variations, both graphically and in tabular format. Users can search for variants by gene symbol, variant ID, or chromosomal coordinates, and can also upload their own data in several popular formats.

Virus variation resource

The Virus Variation Resource is an outgrowth of the Influenza virus and Dengue virus resources and has been updated to include West Nile virus, Middle Eastern Respiratory (MERS) coronavirus and Ebolavirus (23–26). The resource employs computational pipelines and manual curation to create consistent sequence annotations and metadata vocabularies across all sequences from constituent viruses. These standardized data are leveraged by a specialized search interface and a suite of tools designed to support the retrieval and display of large virus sequence datasets.

Epigenomics

The Epigenomics database collects data from studies examining epigenetic features such as post-translational modifications of histone proteins, genomic DNA methylation, chromatin organization and the expression of non-coding regulatory RNA (27). The Epigenomics database provides displays (genome tracks) of the raw data (stored in the Gene expression omnibus (GEO) and SRA databases) mapped to genomic coordinates. Data from the Roadmap Epigenomics project, currently stored in GEO (www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/), are being mirrored and are available for viewing and downloading.

GENES

Gene

Gene (28) provides an interface to curated sequences and descriptive information about genes with links to a wide variety of gene-related resources. These data are accumulated and maintained through several international collaborations in addition to curation by NCBI staff. The complete Gene dataset, as well as organism-specific subsets, is available in the compact NCBI Abstract Syntax Notation One (ASN.1) format on the NCBI FTP site. The gene2xml tool converts the native Gene ASN.1 format into XML and is available at ftp.ncbi.nlm.nih.gov/toolbox/ncbi_tools/converters/by_program/gene2xml/.

RefSeqGene

As part of the Locus Reference Genomic (LRG) collaboration (www.lrg-sequence.org), RefSeqGene provides stable, standard human genomic sequences annotated with mRNAs for well-characterized human genes (13). RefSeqGene records are part of the RefSeq collection and are used to establish numbering systems for exons and introns and for reporting and identifying genomic variants, especially those of clinical importance (29). RefSeqGene records can be retrieved from the Nucleotide database using the query 'refseqgene[keyword]', are available on corresponding Gene reports and can be downloaded from ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/RefSeqGene.

The conserved CDS database (CCDS)

The conserved CDS database (CCDS) project is a collaborative effort between NCBI, the European Bioinformatics Institute, the Wellcome Trust Sanger Institute (WTSI)

and the University of California, Santa Cruz (UCSC). The CCDS compiles a set of human and mouse protein coding regions that are consistently annotated and of high quality (30). The collaborators prepare the CCDS set by comparing the annotations they have independently determined and then identifying those coding regions that have identical coordinates on the genome. Those regions that pass quality evaluations are then added to the CCDS set. The CCDS sequence data are available at ftp.ncbi.nlm.nih.gov/pub/CCDS/.

Gene expression omnibus (GEO)

GEO (31) is a data repository and retrieval system for high-throughput functional genomic data generated by microarray and next-generation sequencing technologies. In addition to gene expression data, GEO accepts data from studies of genome copy number variation, genome-protein interaction surveys and methylation profiling studies. The repository can capture fully annotated raw and processed data, enabling compliance with reporting standards such as 'Minimum Information About a Microarray Experiment' (MIAME) (23,24). GEO data are housed in two Entrez databases: GEO Profiles, which contains quantitative gene expression measurements for one gene across an experiment, and GEO DataSets, which contains entire experiments.

UniGene

UniGene (32) is a system for partitioning transcript sequences (including ESTs) from GenBank into a non-redundant set of clusters, each of which contains sequences that seem to be produced by the same transcription locus. UniGene clusters are created for all organisms for which there are 70 000 or more ESTs in GenBank.

HomoloGene

HomoloGene is a system that automatically detects homologs, including paralogs and orthologs, among the genes of 21 completely sequenced eukaryotic genomes. HomoloGene reports include homology and phenotype information drawn from Online Mendelian Inheritance in Man (OMIM) (33), Mouse Genome Informatics (MGI) (34), the Zebrafish Information Network (ZFIN) (35), the Saccharomyces Genome Database (SGD) (36) and FlyBase (37). Information about the HomoloGene build procedure is provided at www.ncbi.nlm.nih.gov/HomoloGene/HTML/homologene_buildproc.html.

PROTEINS

RefSeq

In addition to genomic and transcript sequences, the RefSeq database (13) contains protein sequences that are curated and computationally derived from these DNA and RNA sequences. RefSeq protein sequences can be searched and retrieved from the Protein database, and the complete RefSeq collection is available in the RefSeq directory on the NCBI FTP site.

GenBank and other sources

As part of standard submission procedures, NCBI produces conceptual translations for any sequence in GenBank that contains a coding sequence and places these protein sequences in the Protein database. In addition to these GenPept sequences, the Protein database also contains sequences from TPA, UniProtKB/Swiss-Prot (38), the Protein Research Foundation (PRF) and the Protein Data Bank (PDB) (39).

Molecular modeling database (MMDB)

Molecular modeling database (MMDB) (40) contains experimentally determined coordinate sets from PDB (39) augmented with domain annotations and links to relevant literature, protein and nucleotide sequences, chemicals (PDB heterogens), and conserved domains in the Conserved Domain Database (CDD) (41). MMDB also provides interactive views of the data in Cn3D (42), the NCBI structure and alignment viewer. MMDB provides structural neighbors for each record based on similarities computed by the VAST algorithm between compact structural domains within protein structures (43,44).

Conserved domain database (CDD)

CDD (45) contains PSI-BLAST-derived Position Specific Score Matrices representing domains taken from the Simple Modular Architecture Research Tool (Smart) (46), Pfam (47), TIGRFAM (48) and from domain alignments derived from the Clusters of Orthologous Groups (COGs) database and the Protein Clusters database. In addition, CDD includes superfamily records that contain sets of CDs from one or more source databases that generate overlapping annotation on the same protein sequences.

Protein clusters

The Protein Clusters database contains sets of almost identical RefSeq proteins encoded by complete genomes from prokaryotes, eukaryotic organelles (mitochondria and chloroplasts), viruses and plasmids, as well as from some protozoans and plants. The clusters are organized in a taxonomic hierarchy and are created based on reciprocal best-hit protein BLAST scores (49).

HIV-1/human protein interaction database

The HIV-1/Human Protein Interaction Database is an online presentation of documented interactions between HIV-1 proteins, host cell proteins, other HIV-1 proteins or proteins from disease organisms associated with HIV or AIDS (50). These data are maintained by the Division of Acquired Immunodeficiency Syndrome of the National Institute of Allergy and Infectious Diseases (NIAID) in collaboration with the Southern Research Institute and NCBI.

BLAST SEQUENCE ANALYSIS

BLAST software

The BLAST programs (51–53) perform sequence-similarity searches against a variety of nucleotide and protein

databases, returning a set of gapped alignments with links to full sequence records and related NCBI resources. The basic BLAST programs are also available as standalone command line programs and network clients at ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/ (Table 2).

BLAST on the cloud

NCBI provides an experimental Amazon Machine Image (AMI) for BLAST hosted at the Amazon Web Services (AWS) Marketplace. This AMI is preconfigured with the latest BLAST+ applications and includes a FUSE client that can download BLAST databases from NCBI as needed. Users can also upload their own custom databases. For tools that access NCBI BLAST programmatically, the AMI also supports the BLAST URL API, making an AWS instance a drop-in replacement for the NCBI BLAST website. More information is available at the BLAST help page as well as in an archived webinar about this AMI (www.ncbi.nlm.nih.gov/education/webinars/).

BLAST databases

The default database for nucleotide BLAST searches (nr/nt) contains all RefSeq RNA records plus all GenBank sequences except for those from the EST, GSS, STS and HTG divisions. Another featured database is *Human genomic plus transcript* that contains human RefSeq transcript and genomic sequences arising from the NCBI annotation of the human genome. A similar database is available for mouse. Additional databases are also available and are described in links from the BLAST search form. Each of these databases can be limited to an arbitrary taxonomic node or those records satisfying any Entrez query.

For proteins the default database (nr) is a non-redundant set of all CDS translations from GenBank along with all sequences from RefSeq, UniProtKB/Swiss-Prot, PDB and the Protein Research Foundation (PRF). Subsets of this database are also available, such as the PDB or UniProtKB/Swiss-Prot sequences, along with separate databases for sequences from patents and environmental samples. Like the nucleotide databases, these collections can be limited by taxonomy or an arbitrary Entrez query.

BLAST output formats

Standard BLAST output formats include the default pairwise alignment, several query-anchored multiple sequence alignment formats, an easily-parsable Hit Table and a report that organizes the BLAST hits by taxonomy. A *pairwise with identities* mode better highlights differences between the query and a target sequence. A Tree View option for the Web BLAST service creates a dendrogram that clusters sequences according to their distances from the query sequence. Each alignment returned by BLAST is scored and assigned a measure of statistical significance called the Expectation Value (*E*-value). The alignments returned can be limited by an *E*-value threshold or range.

Genomic BLAST

NCBI maintains Genomic BLAST services that mirror the design of the standard BLAST forms and allow users access to specialized databases for each particular genome. The default database contains the genomic sequence of an organism, but additional databases are provided depending on the available data and annotations. The default algorithm for Genomic BLAST is MegaBLAST (54), a faster version of standard nucleotide BLAST designed to find alignments between nearly identical sequences, typically from the same species. For rapid cross-species nucleotide queries, NCBI offers Discontiguous MegaBLAST, which uses a non-contiguous word match (55) as the nucleus for its alignments. Discontiguous MegaBLAST is far more rapid than a translated search such as blastx, yet maintains a competitive degree of sensitivity when comparing coding regions.

Primer-BLAST

Primer-BLAST is a tool for designing and analyzing polymerase chain reaction (PCR) primers based on the existing program Primer3 (56) that designs PCR primers given a template DNA sequence. Primer-BLAST extends this functionality by running a BLAST search against a chosen database with the designed primers as queries, and then returns only those primer pairs specific to the desired target. If a user provides only one primer with the DNA template, the other primer will be designed and analyzed. If a user provides both primers and a template, the tool performs only the final BLAST analysis. If a user provides both primers but no template, primer-BLAST will display those templates that best match the primer pair. The available databases include the RefSeq mRNA collection, the BLAST nr database and genomic sets for one of twelve model organisms.

IgBLAST

IgBLAST is a specialized BLAST tool that facilitates the analysis of immunoglobulin variable domain sequences and T-cell receptor sequences (57). In addition to a standard BLAST analysis, IgBLAST reports the germline V, D and J gene matches to the query sequence, annotates immunoglobulin domains, reveals V(D)J junction details and indicates whether the rearrangement is in-frame or out-of-frame. IgBLAST is available both as a web tool and as a stand-alone package.

COBALT

The Constraint-based multiple protein Alignment Tool (COBALT) (58) is a multiple alignment algorithm for proteins that finds a collection of pair-wise constraints derived from both the NCBI CDD and the sequence similarity programs RPS-BLAST, BLASTp and PHI-BLAST. These pairwise constraints are then incorporated into a progressive multiple alignment. Links at the top of the COBALT report provide access to a phylogenetic tree view of the multiple alignment and allow users either to launch a modified search or download the alignment in several popular formats.

Table 2. Selected NCBI software available for download

Software	Available binaries	Category within this article
BLAST (standalone)	Win, Mac, LINUX	BLAST sequence analysis
IgBLAST (standalone)	Win, Mac, LINUX	BLAST sequence analysis
CD-Tree	Win, Mac	Proteins
Cn3D	Win, Mac	Proteins
PC3D	Win, Mac, LINUX	Chemicals
gene2xml	Win, Mac, LINUX, Solaris	Genes
Genome Workbench	Win, Mac, LINUX	Genomes
splign	LINUX, Solaris	Genomes
tbl2asn	Win, Mac, LINUX, Solaris	Genomes

CHEMICALS

PubChem

PubChem (59,60) focuses on the chemical, structural and biological properties of small molecules, in particular their roles as diagnostic and therapeutic agents. A suite of three Entrez databases, PCSubstance, PCCompound and PCBioAssay, contain the structural and bioactivity data of the PubChem project. PubChem also provides a diverse set of three-dimensional (3D) conformers for 90% of the records in the PubChem Compound database.

Biosystems

The Biosystems database collects together molecules represented in Gene, Protein and PubChem that interact in a biological system, such as a biochemical pathway or disease. Currently Biosystems receives data from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (61–63), BioCyc (64), Reactome (65), the Pathway Interaction Database (66), WikiPathways (67,68) and Gene Ontology (69).

FOR FURTHER INFORMATION

The resources described here include documentation, other explanatory material and references to collaborators and data sources on their respective websites. 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 new second edition of 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/home/learn.shtml) 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. Updates on NCBI resources and database enhancements are described on the NCBI News site (www.ncbi.nlm.nih.gov/news/), NCBI social media sites (FaceBook, Twitter, Google+ and LinkedIn), the 'NCBI Insights' blog, 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 the NCBI News site.

FUNDING

Funding for open access charge: Intramural Research Program of the National Institutes of Health, National Library of Medicine.
Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Cavanaugh,M., Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2016) GenBank. *Nucleic Acids Res.*, doi:10.1093/nar/gkv1276.
2. Rubinstein,W.S., Maglott,D.R., Lee,J.M., Kattman,B.L., Malheiro,A.J., Ovetsky,M., Hem,V., Gorelenkov,V., Song,G., Wallin,C. *et al.* (2013) The NIH genetic testing registry: a new, centralized database of genetic tests to enable access to comprehensive information and improve transparency. *Nucleic Acids Res.*, **41**, D925–D935.
3. Kitts,P.A., Church,D.M., Choi,J., Hem,V., Smith,R., Tatusova,T., Thibaud-Nissen,F., DiCuccio,M., Murphy,T.D., Pruitt,K.D. *et al.* (2016) Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.*, doi:10.1093/nar/gkv1226.
4. NCBI Resource Coordinators. (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **43**, D6–D17.
5. 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.
6. Sequeira,E. (2003) PubMed Central—three years old and growing stronger. *ARL*, **228**, 5–9.
7. Sewell,W. (1964) Medical Subject Headings in Medlars. *Bull. Med. Libr. Assoc.*, **52**, 164–170.
8. Landrum,M.J., Lee,J.M., Riley,G.R., Jang,W., Rubinstein,W.S., Church,D.M. and Maglott,D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.
9. Green,R.C., Berg,J.S., Grody,W.W., Kalia,S.S., Korf,B.R., Martin,C.L., McGuire,A.L., Nussbaum,R.L., O'Daniel,J.M., Ormond,K.E. *et al.* (2013) ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.*, **15**, 565–574.
10. Manolio,T.A., Rodriguez,L.L., Brooks,L., Abecasis,G., Ballinger,D., Daly,M., Donnelly,P., Faraone,S.V., Frazer,K., Gabriel,S. *et al.* (2007) New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat. Genet.*, **39**, 1045–1051.
11. Blumenfeld,O.O. and Patnaik,S.K. (2004) Allelic genes of blood group antigens: a source of human mutations and cSNPs documented in the Blood Group Antigen Gene Mutation Database. *Hum. Mutat.*, **23**, 8–16.
12. Helmberg,W., Dunivin,R. and Feolo,M. (2004) The sequencing-based typing tool of dbMHC: typing highly polymorphic gene sequences. *Nucleic Acids Res.*, **32**, W173–W175.
13. Pruitt,K.D., Tatusova,T., Brown,G.R. and Maglott,D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.

14. Kodama,Y., Shumway,M. and Leinonen,R. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
15. Schneider,V.A., Chen,H.C., Clausen,C., Meric,P.A., Zhou,Z., Bouk,N., Husain,N., Maglott,D.R. and Church,D.M. (2013) Clone DB: an integrated NCBI resource for clone-associated data. *Nucleic Acids Res.*, **41**, D1070–D1078.
16. 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.
17. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
18. Federhen,S. (2015) Type material in the NCBI Taxonomy Database. *Nucleic Acids Res.*, **43**, D1086–D1098.
19. Federhen,S., Clark,K., Barrett,T., Parkinson,H., Ostell,J., Kodama,Y., Mashima,J., Nakamura,Y., Cochrane,G. and Karsch-Mizrachi,I. (2014) Toward richer metadata for microbial sequences: replacing strain-level NCBI taxonomy taxids with BioProject, BioSample and Assembly records. *Stand. Genomic Sci.*, **9**, 1275–1277.
20. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
21. Sudmant,P.H., Kitzman,J.O., Antonacci,F., Alkan,C., Malig,M., Tsalenko,A., Sampas,N., Bruhn,L., Shendure,J. and Eichler,E.E. (2010) Diversity of human copy number variation and multicopy genes. *Science*, **330**, 641–646.
22. Church,D.M., Lappalainen,I., Sneddon,T.P., Hinton,J., Maguire,M., Lopez,J., Garner,J., Paschall,J., Dicuccio,M., Yaschenko,E. *et al.* (2010) Public data archives for genomic structural variation. *Nat. Genet.*, **42**, 813–814.
23. Brister,J.R., Bao,Y., Zhdanov,S.A., Ostapchuck,Y., Chetvernin,V., Kiryutin,B., Zaslavsky,L., Kimelman,M. and Tatusova,T.A. (2014) Virus Variation Resource–recent updates and future directions. *Nucleic Acids Res.*, **42**, D660–D665.
24. Brister,J.R., Ako-Adjei,D., Bao,Y. and Blinkova,O. (2015) NCBI viral genomes resource. *Nucleic Acids Res.*, **43**, D571–D577.
25. Resch,W., Zaslavsky,L., Kiryutin,B., Rozanov,M., Bao,Y. and Tatusova,T.A. (2009) Virus variation resources at the National Center for Biotechnology Information: dengue virus. *BMC Microbiol.*, **9**, 65–71.
26. Zaslavsky,L., Bao,Y. and Tatusova,T.A. (2008) Visualization of large influenza virus sequence datasets using adaptively aggregated trees with sampling-based subscale representation. *BMC Bioinformatics*, **9**, 237–243.
27. Fingerman,I.M., McDaniel,L., Zhang,X., Ratzat,W., Hassan,T., Jiang,Z., Cohen,R.F. and Schuler,G.D. (2011) NCBI Epigenomics: a new public resource for exploring epigenomic data sets. *Nucleic Acids Res.*, **39**, D908–D912.
28. Brown,G.R., Hem,V., Katz,K.S., Ovetsky,M., Wallin,C., Ermolaeva,O., Tolstoy,I., Tatusova,T., Pruitt,K.D., Maglott,D.R. *et al.* (2015) Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.*, **43**, D36–D42.
29. Gulley,M.L., Braziel,R.M., Halling,K.C., Hsi,E.D., Kant,J.A., Nikiforova,M.N., Nowak,J.A., Ogino,S., Oliveira,A., Polesky,H.F. *et al.* (2007) Clinical laboratory reports in molecular pathology. *Arch. Pathol. Lab. Med.*, **131**, 852–863.
30. Farrell,C.M., O'Leary,N.A., Harte,R.A., Loveland,J.E., Wilming,L.G., Wallin,C., Diekhans,M., Barrell,D., Searle,S.M., Aken,B. *et al.* (2014) Current status and new features of the Consensus Coding Sequence database. *Nucleic Acids Res.*, **42**, D865–D872.
31. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
32. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
33. Amberger,J., Bocchini,C.A., Scott,A.F. and Hamosh,A. (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.*, **37**, D793–D796.
34. Eppig,J.T., Blake,J.A., Bult,C.J., Kadin,J.A. and Richardson,J.E. (2007) The mouse genome database (MGD): new features facilitating a model system. *Nucleic Acids Res.*, **35**, D630–D637.
35. Sprague,J., Bayraktaroglu,L., Clements,D., Conlin,T., Fashena,D., Frazer,K., Haendel,M., Howe,D.G., Mani,P., Ramachandran,S. *et al.* (2006) The Zebrafish Information Network: the zebrafish model organism database. *Nucleic Acids Res.*, **34**, D581–D585.
36. Hong,E.L., Balakrishnan,R., Dong,Q., Christie,K.R., Park,J., Binkley,G., Costanzo,M.C., Dwight,S.S., Engel,S.R., Fisk,D.G. *et al.* (2008) Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res.*, **36**, D577–D581.
37. Crosby,M.A., Goodman,J.L., Strelets,V.B., Zhang,P. and Gelbart,W.M. (2007) FlyBase: genomes by the dozen. *Nucleic Acids Res.*, **35**, D486–D491.
38. Magrane,M. and Consortium,U. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)*, **2011**, bar009.
39. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
40. Madej,T., Addess,K.J., Fong,J.H., Geer,L.Y., Geer,R.C., Lanczycki,C.J., Liu,C., Lu,S., Marchler-Bauer,A., Panchenko,A.R. *et al.* (2012) MMDB: 3D structures and macromolecular interactions. *Nucleic Acids Res.*, **40**, D461–D464.
41. Marchler-Bauer,A., Derbyshire,M.K., Gonzales,N.R., Lu,S., Chitsaz,F., Geer,L.Y., Geer,R.C., He,J., Gwadz,M., Hurwitz,D.I. *et al.* (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res.*, **43**, D222–D226.
42. Wang,Y., Geer,L.Y., Chappey,C., Kans,J.A. and Bryant,S.H. (2000) Cn3D: sequence and structure views for Entrez. *Trends Biochem. Sci.*, **25**, 300–302.
43. Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.
44. Madej,T., Gibrat,J.F. and Bryant,S.H. (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.
45. Marchler-Bauer,A., Anderson,J.B., Chitsaz,F., Derbyshire,M.K., DeWeese-Scott,C., Fong,J.H., Geer,L.Y., Geer,R.C., Gonzales,N.R., Gwadz,M. *et al.* (2009) CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res.*, **37**, D205–D210.
46. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–D260.
47. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
48. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
49. Klimke,W., Agarwala,R., Badretdin,A., Chetvernin,S., Ciufo,S., Fedorov,B., Kiryutin,B., O'Neill,K., Resch,W., Resenchuk,S. *et al.* (2009) The National Center for Biotechnology Information's Protein Clusters Database. *Nucleic Acids Res.*, **37**, D216–D223.
50. Fu,W., Sanders-Beer,B.E., Katz,K.S., Maglott,D.R., Pruitt,K.D. and Ptak,R.G. (2009) Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Res.*, **37**, D417–D422.
51. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
52. 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.
53. 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.
54. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
55. Ma,B., Tromp,J. and Li,M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.
56. Rozen,S. and Skalestsky,H.J. (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz,S and Misener,S (eds). *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Humana Press, Totowa, NJ, pp. 365–386.

57. 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.
58. Papadopoulos,J.S. and Agarwala,R. (2007) COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics*, **23**, 1073–1079.
59. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J. and Bryant,S.H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.
60. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J., Zhou,Z., Han,L., Karapetyan,K., Dracheva,S., Shoemaker,B.A. *et al.* (2012) PubChem's BioAssay Database. *Nucleic Acids Res.*, **40**, D400–D412.
61. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
62. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
63. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
64. Keseler,I.M., Bonavides-Martinez,C., Collado-Vides,J., Gama-Castro,S., Gunsalus,R.P., Johnson,D.A., Krummenacker,M., Nolan,L.M., Paley,S., Paulsen,I.T. *et al.* (2009) EcoCyc: a comprehensive view of Escherichia coli biology. *Nucleic Acids Res.*, **37**, D464–D470.
65. Matthews,L., Gopinath,G., Gillespie,M., Caudy,M., Croft,D., de Bono,B., Garapati,P., Hemish,J., Hermjakob,H., Jassal,B. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, **37**, D619–D622.
66. Schaefer,C.F., Anthony,K., Krupa,S., Buchoff,J., Day,M., Hannay,T. and Buetow,K.H. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**, D674–D679.
67. Kelder,T., Pico,A.R., Hanspers,K., van Iersel,M.P., Evelo,C. and Conklin,B.R. (2009) Mining biological pathways using WikiPathways web services. *PLoS One*, **4**, e6447.
68. Pico,A.R., Kelder,T., van Iersel,M.P., Hanspers,K., Conklin,B.R. and Evelo,C. (2008) WikiPathways: pathway editing for the people. *PLoS Biol.*, **6**, e184.
69. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

APPENDIX

NCBI Resource Coordinators: Richa Agarwala, Tanya Barrett, Jeff Beck, Dennis A Benson, Colleen Bollin, Evan Bolton, Devon Bourexis, J Rodney Brister, Stephen H Bryant, Kathi Canese, Chad Charowhas, Karen Clark, Michael DiCuccio, Ilya Dondoshansky, Scott Federhen, Michael Feolo, Kathryn Funk, Lewis Y Geer, Viatcheslav Gorelenkov, Marilu Hoeppner, Brad Holmes, Mark Johnson, Viatcheslav Khotomlianski, Avi Kimchi, Michael Kimelman, Paul Kitts, William Klimke, Sergey Krasnov, Anatoliy Kuznetsov, Melissa J Landrum, David Landsman, Jennifer M Lee, David J Lipman, Zhiyong Lu, Thomas L Madden, Tom Madej, Aron Marchler-Bauer, Ilene Karsch-Mizrachi, Terence Murphy, Rebecca Orris, James Ostell, Christopher O'Sullivan, Anna Panchenko, Lon Phan, Don Preuss, Kim D Pruitt, Kurt Rodarmer, Wendy Rubinstein, Eric W Sayers, Valerie Schneider, Gregory D Schuler, Stephen T Sherry, Karl Sirotkin, Karanjit Siyan, Douglas Slotta, Alexandra Soboleva, Vladimir Soussov, Grigory Starchenko, Tatiana A Tatusova, Kamen Todorov, Bart W Trawick, Denis Vakatov, Yanli Wang, Minghong Ward, W John Wilbur, Eugene Yaschenko, Kerry Zbicz.