D6–D17 *Nucleic Acids Research, 2015, Vol. 43, Database issue* *Published online 14 November 2014*
doi: 10.1093/nar/gku1130

Database resources of the National Center for Biotechnology Information

NCBI Resource Coordinators[*,†]

Received October 3, 2014; Accepted October 26, 2014

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 (Bookshelf, PubMed Central (PMC) and PubReader); medical genetics (ClinVar, dbMHC, the Genetic Testing Registry, HIV-1/Human Protein Interaction Database and MedGen); genes and genomics (BioProject, BioSample, dbSNP, dbVar, Epigenomics, Gene, Gene Expression Omnibus (GEO), Genome, HomoloGene, the Map Viewer, Nucleotide, PopSet, Probe, RefSeq, Sequence Read Archive, the Taxonomy Browser, Trace Archive and UniGene); and proteins and chemicals (Biosystems, COBALT, the Conserved Domain Database (CDD), the Conserved Domain Architecture Retrieval Tool (CDART), the Molecular Modeling Database (MMDB), Protein Clusters, Protein and the PubChem suite of small molecule databases). The Entrez system provides search and retrieval operations for many of these databases. Augmenting many of the Web applications are custom implementations of the BLAST program optimized to search specialized data sets. All of these resources can be accessed through the NCBI home page at http://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 the international collaboration with DDBJ and 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 http://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://ftp.ncbi.nlm.nih.gov.

RECENT DEVELOPMENTS

Variation Viewer

The Variation Viewer (http://www.ncbi.nlm.nih.gov/variation/view) is a new tool for displaying human variations from dbSNP, dbVar and ClinVar in the context of the Genome Reference Consortium (GRC) reference genomes, currently GRCh37 and GRCh38 (2). Based on the NCBI sequence viewer used in the Gene, Nucleotide and Protein databases, Variation Viewer displays the variations both graphically and in a table below a depiction of the ideogram of the appropriate chromosome. The variants can be filtered by a number of properties including clinical significance, source database, minor allele frequency and the type of variant. Once filtered as desired, the data shown in the table can be downloaded. Users can search for variants by gene symbol, variant ID or chromosomal coordinates and can also upload their own data in several popular formats.

PubMed Commons

PubMed Commons (http://www.ncbi.nlm.nih.gov/pubmedcommons/) was developed to enable 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 immediately in PubMed, below the publication's abstract, and are regularly monitored for adherence to guidelines (http://www.ncbi.nlm.nih.gov/pubmedcommons/help/guidelines/). Comments are citable and may be shared

[*] To whom correspondence should be addressed. Eric W. Sayers. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA. 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.

Published by Oxford University Press on behalf of Nucleic Acids Research 2014.

or adapted, with attribution, under a Creative Commons license (http://creativecommons.org/licenses/by/3.0/us/). Members are devising many uses for PubMed Commons including centralizing links to raw data and sources, seeking and providing clarifications of methods and results, updating prior findings, providing alternative interpretations, and contributing post-publication peer review. PubMed Commons will also be conducting a beta-test to allow journal clubs to participate. More information about joining and using PubMed Commons is available on the PubMed Commons Getting Started page (http://www.ncbi.nlm.nih.gov/pubmedcommons/get-started/).

BLAST in the cloud

NCBI is now providing an Amazon Machine Image (AMI) for the Basic Local Alignment Search Tool (BLAST) hosted at the Amazon Web Services Marketplace. This AMI is preconfigured with the latest BLAST+ applications and includes a Filesystem in Userspace (FUSE) client that can download BLAST databases from NCBI as needed. Users can also upload their own custom databases. The AMI also supports a subset of the BLAST URL Application Programming Interface (API) functions along with a simplified web page for launching searches. More information is available at the BLAST help page as well as in an archived webinar on this new AMI (see the NCBI webinar page at http://www.ncbi.nlm.nih.gov/education/webinars/).

Entrez Direct

Entrez Direct is a new set of tools that provides an interface to the Entrez Programming Utilities (E-Utilities) on the UNIX command line. Entrez Direct consists of several Perl scripts that can be executed directly and that take as input a variety of command-line arguments. These scripts 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. In addition to the standard E-Utility functions (see below), Entrez Direct also offers a utility named 'xtract' that parses the XML output of ESummary, EFetch and ELink 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 easy storage, sharing or use by other applications. Extensive documentation is available (http://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.

PubMed updates

In response to requests from authors and users, PubMed Abstract displays now contain several enhancements. Social media icons now appear below the abstract text and allow users to share PubMed citations on Facebook, Twitter and Google+. Also appearing below the abstract text are any comments from PubMed Commons, and these are visible to all PubMed users. Abstract displays for clinical trial citations may now include a 'Cited by systematic reviews' discovery panel if the trial is cited in a PubMed Health systematic review (e.g. PMID 15368038). Author affiliation data for PubMed citations have also been enhanced, in that an Abstract display may include multiple affiliations for authors, investigators or corporate authors if the publisher supplied these data. Previously, only the affiliation for the first author was available. Regarding PubMed searches, a new option in the 'Display Settings' menu allows results to be sorted by 'relevance'. This new algorithm calculates a weight for each PubMed citation based on both how many search terms are found in the PubMed record and in which fields they are found. Recently published articles are given somewhat higher weights. Finally, in order to provide access to the most recently published articles, the frequency at which new and updated citations are added to PubMed has been increased from 5 days per week to all 7 days.

Updates to medical genetics resources

ClinVar. ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/) supports users who want to determine what has been reported about the medical relevance of human sequence variation (3). To supplement the display of the record archive that has been provided since ClinVar was introduced in 2013, ClinVar now includes a view that organizes information around each variant. In both views, ClinVar aggregates data from multiple submitters to make it easier to evaluate the current status of interpretation. ClinVar maintains connections with dbSNP, dbVar, Gene, MedGen and PubMed using Entrez links, is accessible as annotations on chromosome and RefSeqGene sequences, and is included in the new 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 (4). As a partner of the ClinGen project, ClinVar encourages domain experts to apply for recognition as an expert panel (http://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 (http://www.ncbi.nlm.nih.gov/clinvar/docs/submit). ClinVar data are freely available for download from the web site, by FTP as VCF or XML files or using the E-Utilities.

GTR. The Genetic Testing Registry (GTR) (http://www.ncbi.nlm.nih.gov/gtr) collects and displays information that has been submitted by providers about their genetic tests (5). GTR now provides an advanced search feature to quickly find tests using a rich set of attributes (http://www.ncbi.nlm.nih.gov/gtr/tests/advanced/). In addition to condition name, gene symbol, lab location and name, searching for tests can be done based on specimen type, pharmacogenetic response conditions or services (e.g. whole exome or whole genome sequencing) either individually or in combinations. Complex tests can be found by their primary test methodology (e.g. next-generation sequencing or microarray), and panels can be identified by specifying the number of genes (e.g. greater than 4). GTR accepts submissions for germline, somatic and research tests. In response to requests from the community to make submission of panels and large test menus simpler, an Excel template is now available that allows submitters to upload a file with full clinical test information rather than entering tests one at a time. The uploaded data will then automatically appear on the GTR web site within 24–48 h (http://www.ncbi.nlm.nih.gov/gtr/docs/fulltest/). To make it easier for submitters to update information about their tests, they can now download all tests from their laboratory to an Excel file that they can then edit and resubmit (http://www.ncbi.nlm.nih.gov/gtr/docs/submit/download_tests/). GTR data can be downloaded in XML format using either FTP or the E-Utilities.

MedGen. MedGen (http://www.ncbi.nlm.nih.gov/medgen/) organizes information about human disorders that have a genetic component (http://www.ncbi.nlm.nih.gov/books/NBK159970/). Starting from freely available content in the semi-annual releases from the Unified Medical Language System (UMLS, http://www.nlm.nih.gov/research/umls/), MedGen adds recent content from several sources: Online Mendelian Inheritance in Man (OMIM), the Human Phenotype Ontology (http://www.human-phenotype-ontology.org/), the Orphanet Rare Disease Ontology (ORDO, http://www.orphadata.org/cgi-bin/inc/ordo_orphanet.inc.php), 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 NCBI's Gene database, variants submitted to ClinVar and whether testing is available in GTR. By reporting disorders, findings, clinical features and drugs, MedGen supports querying for disorders that share clinical features and drugs and their responses. MedGen data can be downloaded using FTP as pipe-delimited (RFF) or CSV text files and using the E-Utilities.

SciENcv updates

SciENcv (Science Experts Network Curriculum Vitae) is a tool in the My NCBI suite (http://www.ncbi.nlm.nih.gov/account/) that assists NIH-funded researchers in creating and maintaining biosketches that are used in NIH grant applications. Based on user suggestions, SciENcv has received several recent enhancements. Researchers can now create and view multiple biosketches in a single My NCBI account, and these biosketches can be created in three ways: from scratch, from an external source or from an existing biosketch. Once created, biosketches can now be downloaded as XML, PDF or Microsoft Word documents. Researchers may also grant a delegate access to their My Bibliography, their SciENcv biosketches or both. These delegates then are able to create, edit or delete data in the applicable services.

SciENcv has also been upgraded to support biosketch profiles in the National Science Foundation (NSF) format. Users with NSF FastLane accounts can link them to My NCBI, which enables them to pre-populate SciENcv biosketches with their profile information stored in FastLane. In addition to FastLane profile information, SciENcv can also retrieve profile information, citations and award information from linked ORCID (Open Researcher and Contributor ID) accounts. Finally, SciENcv now also supports the new Enhanced NIH Biographical Sketch format, which allows researchers to describe their most significant contributions to scientific research along with the historical context that framed their research.

Genome updates

In late 2014 NCBI began releasing a major revision to the genome area of the NCBI FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/). This initial release provides a standard set of data files for over 45 000 genomic assemblies, and these are found in three new directories within the genome area: *genbank*, *refseq* and *all*. The *genbank* directory contains data submitted directly to GenBank (or an International Nucleotide Sequence Database Collaboration (INSDC) database), while the *refseq* directory contains data that are part of the RefSeq project. All RefSeq data sets will contain genome annotations. The common data unit within these three directories is a subdirectory corresponding to a record in the Assembly database 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 data sets. 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 files for annotated genome records. These new directories will co-exist with the older genome FTP data until 1 March 2015, when the older data files will be removed.

Gene updates

The full report pages in the Gene database have several new features designed to improve access to the annotation details for the given gene (6). Several new tracks are available on the sequence viewer interface, including RNA-seq alignments, differences between assemblies (for human genes only) and tracks that show the alignment of gene paralogs. A menu above the sequence viewer provides an easy way to load a previous assembly or an alternate assembly. For genomes annotated at NCBI, the 'Genomic context' section (above the sequence viewer) now includes a table that displays information about the corresponding Assembly record, including the location of the gene on the appropriate chromosome. For human genes, information about the previous assembly is also included, and this function will be expanded to other species in the future.

Protein updates

The Protein database now provides an 'Identical Protein Report' that displays the accessions of all other protein sequence records that are identical to a given protein along with 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 WP sequences introduced in 2013 in response to the anticpated rapid growth of highly redundant prokaryotic genome sequences from clinical samples (7). Each set of identical protein annotations on these many genomes will be represented by a single WP sequence, and therefore the individual protein annotations will not have separate records or GI numbers. WP sequences are thus connected to not one but a corresponding set of Nucleotide CDS sequences, potentially from multiple species. The Identical Protein Report clarifies these relationships. For example, WP_002317106 collects over 40 thioredoxin sequences from several species and strains of *Enterococcus.* To access the report, open the 'Display Settings' menu on a protein record page and choose 'Identical Protein Report'. The report is also available through the E-Utility EFetch with *&rettype = ipg*.

In 2014 NCBI also introduced VAST+, an enhanced version of the Vector Alignment Search Tool (VAST) that extends structural similarity searching to macromolecular complexes (8). While the original VAST algorithm identified structural neighbors for a single protein chain (or three-dimensional (3D) domains within that chain), VAST+ finds structural neighbors for the biological unit defined in the source Protein Data Bank (PDB) file. Biological units may be single chains, but often are several chains that fold together to form an active quaternary structure. For example, the biological unit for hemoglobin is a tetramer, and so VAST+ will retrieve other tetramers that have a similar quaternary structure. Precomputed VAST+ results are available for records in the NCBI structure database, and more information and examples are available at http://www.ncbi.nlm.nih.gov/Structure/vastplus/docs/vastplus_help.html.

Taxonomy updates

Given the rapid increase in the number of bacterial genomes being submitted to INSDC, as of January 2014 NCBI no longer assigns taxonomy IDs to bacterial strains that do not already have taxonomy IDs (9). 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 information about the strain, such as culture collection and details about its isolation. For example, record CP006594 is the chromosome sequence of *Listeria monocytogenes* strain R2–502. The taxonomy link from this record leads to taxonomy ID 1639, the species-level record for *L. monocytogenes.* The source information in the feature table contains the value 'R2-502′ for the strain, and the DBLINK section of the record contains a link to BioSample SAMN02203126.

In 2013 the Taxonomy database began including type material for prokaryotic type strains and eukaryotic type specimens (10). For example, several type strains are indicated for *Escherichia coli* (http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=562&lvl=0). It is now possible to retrieve type sequences easily using the following Entrez query:

'sequence from type'[filter]

This query can be combined with organism terms or authors to restrict further the type sequences retrieved.

THE NCBI GUIDE AND THE ENTREZ SYSTEM

The NCBI Guide

The NCBI Guide serves not only as the NCBI home page but also as an interactive directory of the NCBI site. On the main page of the NCBI Guide, the categories in the Resource menu in the standard header are duplicated in a list on the left side of the page. Clicking on any category displays a list of relevant resources sorted into four groups: databases, downloads, submissions and tools. A list of how-to guides is also available via the 'How-To' tab on these pages. Popular resources are listed on the right under a 'Quick Links' heading, and on the main Guide page, a list of the most frequently used resources is provided in the 'Popular Resources' box and also as a list in the standard footer.

Entrez databases

Entrez (11) is an integrated database retrieval system that provides access to a diverse set of 40 databases that together contain 1.3 billion records (Table 1). Links to the web portal for each of these databases are provided on the Entrez GQuery page (http://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 in 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.

Table 1. The Entrez databases (as of 3 September 2014)

Database	Records	Section within this article	Data source[a]
Site search	21 929	Introduction	N
MedGen[b]	260 796	Recent developments	C, N
ClinVar[b]	124 671	Recent developments	D, N
GTR[b]	32 152	Recent developments	D
PubMed	24 157 837	Literature	C
PubMed Central	3 201 919	Literature	D, C
NLM Catalog	1 507 828	Literature	C, N
MeSH	253 057	Literature	N
Books	337 275	Literature	C, N
Taxonomy[b]	1 288 515	Taxonomy	C, N
Nucleotide[b]	146 035 069	DNA and RNA	D (GenBank), C, N
EST[b]	75 673 561	DNA and RNA	D (GenBank)
GSS[b]	37 613 795	DNA and RNA	D (GenBank)
BioSample	2 734 070	DNA and RNA	D
SRA[b]	963 108	DNA and RNA	D
PopSet[b]	207 794	DNA and RNA	D (GenBank)
Protein[b]	147 483 171	Proteins	C, N
Protein Clusters[b]	820 546	Proteins	N
Structure[b]	102 343	Proteins	C, N
CDD[b]	49 641	Proteins	C, N
GEO Profiles[b]	108 686 654	Genes and expression	D
Probe	31 887 935	Genes and expression	D
Gene[b]	17 530 632	Genes and expression	C, N
UniGene[b]	6 473 284	Genes and expression	N
GEO Data Sets[b]	1 295 573	Genes and expression	D
Biosystems[b]	619 468	Genes and expression	C
Homologene[b]	141 268	Genes and expression	N
Clone[b]	36 916 420	Genomes	D, N
BioProject[b]	134 582	Genomes	D
Assembly	32 501	Genomes	C, N
Genome[b]	10 244	Genomes	C, N
Epigenomics[b]	6634	Genomes	D
SNP[b]	394 164 715	Genetics and medicine	D (dbSNP), N
dbVar[b]	4 155 758	Genetics and medicine	D
dbGaP	163 310	Genetics and medicine	D
PubMed Health	49 278	Genetics and medicine	C
PubChem Substance[b]	157 203 085	Chemicals and bioassays	D
PubChem Compound[b]	53 371 491	Chemicals and bioassays	N
PubChem Bioassay[b]	1 091 044	Chemicals and bioassays	D

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

Entrez Programming Utilities (E-Utilities)

The E-Utilities constitute the 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 and 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 or SOAP 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 http://eutils.ncbi.nlm.nih.gov.

LITERATURE

PubMed

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

PubMed Central (PMC)

PMC (12) contains the full text of peer-reviewed journal articles in the life sciences, and is the repository for all manuscripts arising from research using NIH funds and submitted through the NIH manuscript submission system. 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 NIH Public Access Policy are required to be made available in PMC within 12 months of publication. PMC articles can be viewed either as traditional HTML, PDF or using the PubReader viewer.

NLM catalog

The NLM catalog contains bibliographic data for the various items in the NLM collections, including journals, books, audiovisuals, computer software, electronic resources and other materials.

Medical Subject Headings (MeSH)

The MeSH database (13) 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.

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 (14). 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 (http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi) can be used to view the taxonomy tree or retrieve data from any of the Entrez databases for a particular organism or group.

DNA AND RNA

RefSeq

The RefSeq database (15) 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 and other sources

GenBank (1) is the primary nucleotide sequence archive at NCBI and is a member of the INSDC. Sequences from GenBank are available from three Entrez databases: Nucleotide, Expressed Sequence Tags (EST) and Genome Survey Sequences (GSS). (These databases are 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 Whole Genome Shotgun (WGS) sequences, Third Party Annotation (TPA) sequences and sequences imported from the NCBI Structure database (see below).

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 (16) 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 (GWAS), single nucleotide polymorphism and structural variation analysis, pathogen identification, transcript assembly, metagenomic community profiling and epigenetics.

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.

BioSample

The BioSample database provides annotation for biological samples used in a variety of studies submitted to NCBI, including genomic sequencing, microarrays, GWAS and epigenomics (17). 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.

PROTEINS

RefSeq

In addition to genomic and transcript sequences, the RefSeq database (15) 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 (18), the Protein Research Foundation (PRF) and the PDB (19).

Molecular Modeling Database (MMDB)

MMDB (20) contains experimentally determined coordinate sets from PDB (19) 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) (21). MMDB also provides interactive views of the data in Cn3D (22), 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 (23,24).

Conserved Domain Database (CDD)

CDD (25) contains PSI-BLAST-derived Position Specific Score Matrices representing domains taken from the Simple Modular Architecture Research Tool (26), Pfam (27), TIGRFAM (28), and from domain alignments derived from Clusters of Orthologous Groups and Protein Clusters. 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 (29).

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 humanimmunodeficiency virus (HIV) or acquired immunedeficiency syndrome (30). These data are maintained by the Division of Acquired Immunodeficiency Syndrome of the National Institute of Allergy and Infectious Diseases in collaboration with the Southern Research Institute and NCBI.

BLAST SEQUENCE ANALYSIS

BLAST software

The BLAST programs (31–33) 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, as network clients, and as a local Web-server package at ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/ (Table 2).

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, Sequence Tagged Sites (STS) and High-Throughput Genomic (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 input 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 RefSeq, UniProtKB/Swiss-Prot, PDB and PRF proteins. 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 annotation. The default algorithm for Genomic BLAST is MegaBLAST (34), 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 Dis-contiguous MegaBLAST, which uses a non-contiguous word match (35) as the nucleus for its alignments. Dis-contiguous 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 PCR primers based on the existing program Primer3 (36) 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

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	Domains and structures
Cn3D	Win, Mac	Domains and structures
PC3D	Win, Mac, LINUX	Chemicals and bioassays
gene2xml	Win, Mac, LINUX, Solaris	Genes and expression
Genome Workbench	Win, Mac, LINUX	Genomes
splign	LINUX, Solaris	Genomes
tbl2asn	Win, Mac, LINUX, Solaris	Genomes

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 range from RefSeq mRNA or genomic sets for one of twelve model organisms to the entire BLAST nr database.

IgBLAST

IgBLAST is a specialized BLAST tool that facilitates the analysis of immunoglobulin variable domain sequences and T-cell receptor sequences (37). 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 standalone package.

COBALT

COBALT (38) is a multiple alignment algorithm for proteins that finds a collection of pairwise 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.

GENES AND EXPRESSION

Gene

Gene (39) 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 in-house staff. The complete Gene data set, 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://ftp.ncbi.nlm.nih.gov/toolbox/ncbi_tools/converters/by_program/gene2xml/.

RefSeqGene

As part of the Locus Reference Genomic collaboration (http://www.lrg-sequence.org), RefSeqGene provides stable, standard human genomic sequences annotated with standard mRNAs for well-characterized human genes (15). 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 (40). 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://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/RefSeqGene.

The Conserved CDS Database (CCDS)

The 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 to identify a set of human and mouse protein coding regions that are consistently annotated and of high quality (41). 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://ftp.ncbi.nlm.nih.gov/pub/CCDS/.

Gene Expression Omnibus (GEO)

GEO (42) 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’ (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 Data Sets, which contains entire experiments.

UniGene

UniGene (43) 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 OMIM (44), Mouse Genome Informatics (45), Zebrafish Information Network (46), Saccharomyces Genome Database (47) and FlyBase (48). Information about the HomoloGene build procedure is provided at http://www.ncbi.nlm.nih.gov/HomoloGene/HTML/homologene_buildproc.html.

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.

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 (49–51), BioCyc (52), Reactome (53), the Pathway Interaction Database (54), WikiPathways (55,56) and Gene Ontology (57).

GENOMES

BioProject

The BioProject database is a central access point for metadata about research projects whose data are deposited in databases maintained by members of the INSDC. BioProject provides links to the primary data from these projects, which range from focused genome sequencing projects to large international collaborations with multiple subprojects incorporating experiments resulting in nucleotide sequence sets, genotype/phenotype data, sequence variants or epigenetic information.

Genome

The Genome database collects genomic sequencing projects for a given species and provides links to corresponding records in BioProject, Assembly, Nucleotide and Protein. The Genome home page (http://www.ncbi.nlm.nih.gov/genome/) also provides links to eukaryotic and prokaryotic annotation reports that list the current status of all genomes annotated at NCBI.

Assembly

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

Genome Reference Consortium (GRC)

The GRC (http://www.genomereference.org) is an international collaboration between the WTSI, the Genome Institute at Washington University, EMBL and NCBI that aims to produce assemblies of higher eukaryotic genomes that best reflect complex allelic diversity 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://ftp.ncbi.nlm.nih.gov/pub/grc/) and the new genomes FTP area (see above).

Clone database (CloneDB)

CloneDB is a resource for finding descriptions, sources, map positions and distributor information about available clones and libraries (58). 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.

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 (59). The Epigenomics database provides displays ('genome tracks') of the raw data (stored in the GEO and SRA databases) mapped to genomic coordinates. Data from the Roadmap Epigenomics project, currently stored in GEO (http://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/), are being mirrored and are available for viewing and downloading.

Influenza Genome Resources

The Influenza Virus Resource links genome sequence data from the Influenza Genome Sequencing Project (60) to the most recent scientific literature in PubMed on influenza as well as to population studies and protein sequences and structures. The NCBI Virus Variation resource extends these services to the dengue and West Nile viruses.

GENETICS AND MEDICINE

dbGaP

The Database of Genotypes and Phenotypes (dbGaP) (61) archives, distributes and supports submission of data that correlate genomic characteristics with observable traits. This database is a designated NIH repository for NIH-funded GWAS results (http://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.

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 (62). These data are derived from several methods including computational sequence analysis and microarray experiments.

dbSNP

The Database of Short Genetic Variations (dbSNP) (63) is a repository of all types of short genetic variations less than 50 bp in length, and so is a complement to dbVar. dbSNP accepts submissions of common as well as 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 (64).

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. dbMHC also contains human leucocyte antigens genotype and clinical outcome information on hematopoietic cell transplants performed worldwide. dbLRC offers a comprehensive collection of alleles of the leukocyte receptor complex with an emphasis on KIR genes. dbRBC provides data on genes for red blood cell 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 (65) and integrates it with resources at NCBI. All three databases dbMHC, dbLRC and dbRBC provide multiple sequence alignments, analysis tools to interpret homozygous or heterozygous sequencing results (66) and tools for DNA probe alignments.

CHEMICALS AND BIOASSAYS

PubChem

PubChem (67,68) 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 3D conformers for 90% of the records in the PubChem Compound database.

FOR FURTHER INFORMATION

The resources described here include documentation, other explanatory material 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 new second edition of the NCBI Handbook (http://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 Education page (http://www.ncbi.nlm.nih.gov/Education/) lists links to documentation, tutorials and educational tools along with links to outreach initiatives, including Discovery Workshops, webinars and upcoming conference exhibits. The Education page, along with the standard NCBI page footer, contains links to the NCBI YouTube channel, which contains a variety of video tutorials. 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 (http://www.ncbi.nlm.nih.gov/news/), which also links to the 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.

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. (2015) GenBank. *Nucleic Acids Res.*, doi:10.1093/nar/gku1216.
2. 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.
3. 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.
4. 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.

5. 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.
6. Brown,G., Hem,V., Katz,K.S., Ovetsky,M., Wallin,C., Ermolaeva,O., Tolstoy,I., Tatusova,T., Pruitt,K.D., Maglott,D. *et al.* (2015) Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.*, doi:10.1093/nar/gku1055.
7. NCBI Resource Coordinators. (2014) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **42**, D7–D17.
8. Madej,T., Lanczycki,C.J., Zhang,D., Thiessen,P.A., Geer,R.C., Marchler-Bauer,A. and Bryant,S.H. (2014) MMDB and VAST+: tracking structural similarities between macromolecular complexes. *Nucleic Acids Res.*, **42**, D297–D303.
9. 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.
10. Federhen,S. (2015) Type material in the NCBI taxonomy database. *Nucleic Acids Res.*, doi:10.1093/nar/gku1127.
11. 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.
12. Sequeira,E. (2003) PubMed Central—three years old and growing stronger. *ARL*, **228**, 5–9.
13. Sewell,W. (1964) Medical Subject Headings in Medlars. *Bull. Med. Libr. Assoc.*, **52**, 164–170.
14. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
15. 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.
16. Kodama,Y., Shumway,M. and Leinonen,R. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
17. 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.
18. Magrane,M. and Consortium,U. (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database*, **2011**, bar009.
19. 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.
20. 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.
21. Marchler-Bauer,A., Lu,S., Anderson,J.B., Chitsaz,F., Derbyshire,M.K., DeWeese-Scott,C., Fong,J.H., Geer,L.Y., Geer,R.C., Gonzales,N.R. *et al.* (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.*, **39**, D225–D229.
22. 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.
23. Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.
24. Madej,T., Gibrat,J.F. and Bryant,S.H. (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.
25. 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.
26. 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.
27. 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.
28. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
29. 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.
30. 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.
31. 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.
32. 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.
33. 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.
34. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
35. Ma,B., Tromp,J. and Li,M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.
36. 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.
37. 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.
38. Papadopoulos,J.S. and Agarwala,R. (2007) COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics*, **23**, 1073–1079.
39. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.
40. 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.
41. 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.
42. 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.
43. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
44. 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.
45. 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.
46. 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.
47. 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.
48. 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.

49. 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.
50. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
51. 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.
52. 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.
53. 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.
54. 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.
55. 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.
56. 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.
57. 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.
58. 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.
59. 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.
60. Ghedin,E., Sengamalay,N.A., Shumway,M., Zaborsky,J., Feldblyum,T., Subbu,V., Spiro,D.J., Sitz,J., Koo,H., Bolotov,P. *et al.* (2005) Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature*, **437**, 1162–1166.
61. 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.
62. 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.
63. 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.
64. 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.
65. 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.
66. 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.
67. 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.
68. 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.

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, 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, 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, Bart W. Trawick, Denis Vakatov, Yanli Wang, Minghong Ward, W. John Wilbur, Eugene Yaschenko, Kerry Zbicz.