D8–D20 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 27 November 2012*
doi:10.1093/nar/gks1189

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 October 2, 2012; Revised October 28, 2012; Accepted October 29, 2012

ABSTRACT

In addition to maintaining the GenBank® nucleic acid sequence database, the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov) provides analysis and retrieval resources for the data in GenBank and other biological data made available through the NCBI web site. NCBI resources include Entrez, the Entrez Programming Utilities, MyNCBI, PubMed, PubMed Central, Gene, the NCBI Taxonomy Browser, BLAST, BLAST Link (BLink), Primer-BLAST, COBALT, Splign, RefSeq, UniGene, HomoloGene, ProtEST, dbMHC, dbSNP, dbVar, Epigenomics, the Genetic Testing Registry, Genome and related tools, the Map Viewer, Model Maker, Evidence Viewer, Trace Archive, Sequence Read Archive, BioProject, BioSample, Retroviral Genotyping Tools, HIV-1/Human Protein Interaction Database, Gene Expression Omnibus, Probe, Online Mendelian Inheritance in Animals, the Molecular Modeling Database, the Conserved Domain Database, the Conserved Domain Architecture Retrieval Tool, Biosystems, Protein Clusters and the PubChem suite of small molecule 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.

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 the DNA Database of Japan (DDBJ) and the European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL-Bank) as well as from the scientific community, NCBI provides data retrieval systems and computational resources for the analysis of GenBank data and many other kinds of biological data. For the purposes of this article, after a summary of recent developments and an introduction to the Entrez system, the NCBI suite of resources is grouped into 10 broad categories based on those in the NCBI Guide. All resources discussed are available from the NCBI Guide 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

Genome database redesign

In 2012, NCBI completely redesigned the Genome database (www.ncbi.nlm.nih.gov/genome) to broaden its scope and better represent the complexity of modern genome sequencing data. Although an individual record in the previous Genome database represented a single chromosome, a record in the redesigned database represents an entire organism (usually a species) and contains information about the available genome sequence data for that organism and other related data such as transcriptome sequences, variation analyses and epigenetic studies. Also, organelle sequences for a species are now part of the corresponding Genome record rather than being separate records. The scope of the new database has broadened beyond the NCBI Reference Sequence (RefSeq) collection to include data from all submitters. Perhaps the most effective way to query the new database is with the name of a species. For example, the query 'homo sapiens' retrieves the record for the human genome with links to the individual chromosomes, a table of all available genome assemblies and links to a variety of other human projects in the BioProjects database (see later in the text). Links across the top of the page lead to

*To whom correspondence should be addressed. Eric W. Sayers. Tel: +30 1 49 62 475; Fax: +30 1 48 09 241; Email: sayers@ncbi.nlm.nih.gov
†The members of the NCBI Resource Coordinators group are listed in the Appendix.

detailed statistics for each assembly. For prokaryotes, the new Genome record will also represent the species and will include all available subspecies or strains (e.g. www.ncbi.nlm.nih.gov/167 for *Escherichia coli*).

Assembly database

The Assembly database (www.ncbi.nlm.nih.gov/assembly/) is a new resource that provides information about the structure of assembled genomes ranging from simple bacterial genome assemblies consisting of a single complete chromosome to complex assemblies for higher eukaryotes that include alternate locus group scaffolds and patches. The database unambiguously identifies the set of sequences in a particular version of an assembly using an assembly accession.version number and tracks changes to genome assemblies that are updated. The Assembly resource displays metadata about genome assemblies such as assembly names, simple statistical reports of the assembly (number of contigs and scaffolds, N50s and total sequence length) and its update history. The Assembly database also tracks the relationship between an assembly submitted to the International Nucleotide Sequence Database Consortium and the assembly represented in the NCBI RefSeq project. Assemblies of interest can be found either by searching in the resource or by browsing the assemblies available for a particular organism. More information about Assembly is at www.ncbi.nlm.nih.gov/assembly/help/model/.

Genetic testing registry

The Genetic Testing Registry (GTR) is a new resource that collects and displays information about genetic tests that have been voluntarily submitted by the test providers (www.ncbi.nlm.nih.gov/gtr/). This information includes a test's purpose, methodology, validity and evidence of its usefulness, as well as contacts, credentials and certifications for laboratories that perform the test (2). GTR also provides contextual access to data from NCBI's resources such as the Gene database, PubMed and Bookshelf in addition to clinical practice guidelines and clinical referral resources. Users can search GTR content about tests, conditions/phenotypes, genes, laboratories and GeneReviews. GTR content currently includes biochemical, cytogenetic and molecular tests for Mendelian disorders and drug responses.

The GTR web site supports access to GeneReviews, maintained by a team led by Roberta A. Pagon, MD at the University of Washington. GeneReviews (www.ncbi.nlm.nih.gov/books/NBK1116/) is a compendium of continually updated, expert-authored and peer-reviewed disease descriptions that relate genetic testing to the diagnosis, management and genetic counseling of patients and families with specific inherited conditions (3,4). The GTR web site also redisplays content from the GeneTests Laboratory Directory, and as a result, the latter site will be phased out in 2013. Thus, the GTR web site is a unified portal to information about disorders with a genetic component and available testing.

Submission portal

To streamline the process of submitting data to NCBI databases, NCBI is creating a unified submission portal (submit.ncbi.nlm.nih.gov) that will provide a single access point to the various submission interfaces. This portal should be particularly useful for submitters of complex high-throughput sequencing, genome-wide association studies (GWAS) or functional genomic data sets that involve the simultaneous submission of data to several NCBI resources. Submitters will be able to create accounts that will track and display all of their submissions and will facilitate communication with relevant NCBI staff. Currently, the portal fully supports submissions to BioSample, BioProjects and GTR, as well as to the TSA and whole genome shotgun (WGS) divisions of GenBank. It also provides links to the submission pages for 10 other databases.

BLAST updates

Domain Enhanced Lookup Time Accelerated BLAST (DELTA-BLAST) is a more sensitive BLAST algorithm for proteins that contain well-conserved domains (5). DELTA-BLAST begins by searching the query sequence using a Conserved Domain Search (CD-Search) and then constructs a position-specific scoring matrix from those results. This position-specific scoring matrix becomes the scoring matrix for DELTA-BLAST. DELTA-BLAST is particularly useful for proteins with significant matches to CD records, as the resulting alignments will tend to reflect the domain boundaries more accurately relative to those produced by BLASTp, and the results will tend to contain fewer hits that lack the target domain. DELTA-BLAST tends to outperform BLASTp when aligning sequences of low similarity, making it a potentially useful tool for exploring remote homologs. Once completed, DELTA-BLAST results can then be used to initiate a PSI-BLAST search. DELTA-BLAST is now an algorithm option on the standard protein BLAST page.

The microbial BLAST page (linked in the top section of the BLAST home page) has been redesigned and now conforms to the standard BLAST page formats. The pages allow the familiar algorithm and search options on the standard BLAST pages and also contain other familiar features such as 'Edit and Resubmit'. The new pages also include an 'Organism' select box with an auto-complete feature that allows users to include or exclude any taxonomic node.

My NCBI updates

My NCBI provides users with a wide range of services such as saving search queries, setting up automatic searches with e-mail alerts, storing and organizing NCBI database records, selecting preferred display formats, choosing filtering options and tracking recent usage history. In the past year, several improvements have been added to the My Bibliography component of My NCBI. My Bibliography can store a wide variety of citations and assist users with tracking compliance with the NIH Public Access Policy. NIH-funded scientists can now

link any grant in the system with citations in their bibliography, rather than being limited to grants associated with their profile. This is particularly helpful for large grants involving multiple sites or for new investigators wishing to link publications to grants from a previous laboratory. Citations linked to grants in this manner now automatically appear in the grant owner's My Bibliography collection, where the grant owner can review the associations made and edit them if necessary. New filter options have also been added that allow users to quickly find citations that were linked to their grants by other users or that have been processed as author manuscripts using the NIH Manuscript Submission System.

Updates to literature resources

In the past year, NCBI released several enhancements to PubMed including a new Filters sidebar that replaces the Limits page, a 'Citation manager' selection in the 'Send to' menu, and an updated Advanced Search page to provide users with a less cluttered, more intuitive way to build searches. Both the Filters sidebar and the new Advanced search page are further described in YouTube tutorials (see later in the text). The PubMed abstract display now includes a 'Save items' button that provides an easy way to add the citation to a MyNCBI collection. Moreover, when users click an author link in an abstract display, the resulting set of citations are sorted using an improved ranking algorithm. Finally, two new tools now appear in the Discovery column on the right side of PubMed search results: the PubMed 'Results by year' histogram and the PubMed Central (PMC) 'PMC images search'.

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 (6) is an integrated database retrieval system that provides access to a diverse set of 37 databases that together contain 690 million records (Table 1). 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 article in which it is reported or between a protein sequence and 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 'Submissions' column in Table 1 indicates those mechanisms by which each Entrez database receives data. The various collaborations, agreements and curation efforts are described throughout the remainder of this article.

Entrez programming utilities

The Entrez Programming Utilities (E-Utilities) constitute the Application Programming Interface (API) for the Entrez system. The API includes eight 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. Recently, NCBI released version 2.0 of EFetch, which adds support for the BioSample, BioSystems and Sequence Read Archive (SRA) databases and uses a standardized set of values for the *retmode* and *rettype* parameters. Also released are new version 2.0 XML formats available from ESummary. Detailed documentation for using these and the other E-Utilities are found at eutils.ncbi.nlm.nih.gov.

LITERATURE

PubMed

The PubMed database now contains >22 million citations from >24 000 life science journals. More than 12 million of these citations have abstracts, and 13 million have links to their full text articles, with 9.8 million having both an abstract and a link to full text. PubMed is heavily linked to other core NCBI databases, thereby providing a crucial

Table 1. The Entrez Databases (as of September 1, 2012)

Database	Section within this article	Records	Data source
Site search	Introduction	10 686	N
Assembly	Recent developments	9597	D, C, N
PubMed	Literature	22 076 132	C
PubMed central	Literature	2 523 284	D, C
NLM catalog	Literature	1 461 835	C, N
MeSH	Literature	236 253	N
Books	Literature	186 112	C, N
Taxonomy	Taxonomy	932 345	C, N
EST	DNA and RNA	73 666 909	D (GenBank)
Nucleotide	DNA and RNA	66 319 706	D (GenBank), C, N
GSS	DNA and RNA	34 533 114	D (GenBank)
BioSample	DNA and RNA	970 304	N
SRA	DNA and RNA	228 739	D
PopSet	DNA and RNA	159 345	D (GenBank)
Protein	Proteins	56 394 380	C, N
Protein clusters	Proteins	794 663	N
GEO profiles	Genes and expression	63 811 486	D
Probe	Genes and expression	14 248 527	D
Gene	Genes and expression	11 290 372	C, N
UniGene	Genes and expression	5 831 327	N
GEO data sets	Genes and expression	841 518	N
Biosystems	Genes and expression	396 029	C
Homologene	Genes and expression	133 012	N
Clone	Genomes	29 597 231	D, N
UniSTS	Genomes	545 353	D (dbSTS)
BioProject	Genomes	58 227	D
Genome	Genomes	8276	C, N
Epigenomics	Genomes	5484	D
SNP	Genetics and medicine	162 674 947	D (dbSNP), N
dbVar	Genetics and medicine	2 729 616	D
dbGaP	Genetics and medicine	143 624	D
Online mendelian inheritance in animals	Genetics and medicine	2810	C
PubChem substance	Chemicals and bioassays	100 157 112	D
PubChem compound	Chemicals and bioassays	35 545 766	N
PubChem bioassay	Chemicals and bioassays	621 642	D
Structure	Domains and structures	83 913	C, N
CDD	Domains and structures	46 389	C, N

D, direct submission; C, collaboration/agreement; N, internal NCBI/NLM curation.

bridge between the data of molecular biology and the scientific literature. PubMed records are also linked to one another as 'related citations' on the basis of computationally detected similarities using indexed Medical Subject Heading (MeSH) (7) terms and the text of titles and abstracts. Succinct descriptions of the top five related citations are shown on the default Abstract display.

PMC

PubMed Central (PMC) (8) is a digital archive of peer-reviewed journal articles in the life sciences and now contains >2.5 million full-text articles, having grown by 11% during the past year. More than 1600 journals, including *Nucleic Acids Research*, deposit the full text of their articles in PMC, and some 300 of these journals began depositing their data in the past year. Publisher participation in PMC requires a commitment to free access to full text, either immediately after publication or within a 12-month period. PMC also serves as the repository for all final peer-reviewed manuscripts arising from research using NIH funds and submitted through the NIH Manuscript Submission System. All PMC articles are identified in PubMed search results, and PMC itself can be searched using Entrez.

NLM catalog and MeSH

The NLM Catalog provides access to NLM bibliographic data for >1.4 million journals, books, audiovisuals, computer software, electronic resources and other materials. The NLM Catalog contains detailed indexing information for the 28 500 journals in PubMed and other NCBI databases.

The MeSH database includes information about the NLM-controlled vocabulary thesaurus used for indexing PubMed citations. Users may search the MeSH database, which contains >235 000 concepts, to find MeSH terms, including subheadings, publication types, supplementary concepts and pharmacological actions, and then build a PubMed search.

The NCBI bookshelf

The NCBI bookshelf (http://www.ncbi.nlm.nih.gov/books/) is an online service of the National Library of Medicine Literature Archive (NLM LitArch) that provides free access to the full text of >1300 books, reports, databases and documentation in the life sciences and health care fields. Approximately 28 new titles per month were added in 2012. Existing titles, especially those in the database and documentation category

continue to grow and receive regular updates. Information in Bookshelf is linked to and integrated with other NCBI resources, such as PubMed, Gene, GTR and PubChem. This integration enables the user reciprocal access to molecular genetic and structure information from the literature, offering further paths of discovery within this linked network of information. In 2012, a portion of Bookshelf content in NLM LitArch was made available in the NLM LitArch Open Access Subset, through which XML, images, PDF and supplementary files are available for download and reuse as permitted by the license agreements for individual titles.

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 (9). The taxonomy database reflects sequence data from almost 260 000 formally described species. This represents virtually all of the formally described species of prokaryotes and 10% of the eukaryotes. The Taxonomy Browser (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 NCBI RefSeq database (10) is a non-redundant set of curated and computationally derived sequences for transcripts, proteins and genomic regions. The number of nucleotide bases in the RefSeq collection has grown by 8% during the past year so that Release 54 (July, 2012) contains 176 billion bases representing 17 605 organisms. 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.

Sequences from GenBank and other sources

Sequences from GenBank can be searched in and retrieved from three Entrez databases: Nucleotide, EST (containing expressed sequence tags) and GSS (containing genome survey sequences), (Within E-utility calls, these databases should be specified as nuccore, nucest and nucgss.) The Nucleotide database contains all GenBank sequences except those within the EST or GSS GenBank divisions. The database also contains Whole Genome Shotgun sequences, Third Party Annotation sequences and sequences imported from the Structure database. In addition, those sequences that have been submitted as part of a population, phylogenetic or environmental study are placed in the PopSet database.

PopSet

The PopSet database (www.ncbi.nlm.nih.gov/popset/) 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. For PopSets with <100 sequences, links are provided to generate a BLAST alignment of the sequences or, if an alignment was submitted as part of the record, a distance tree view of the alignment.

The SRA and the trace archive

The SRA (11) is a repository for data generated by the latest generation of high-throughput nucleic acid sequencers. SRA contains >850 Terabasepairs of biological sequence data, adding more than a Terabase daily. SRA stores raw sequence reads and alignments. Data are deposited into SRA as supporting evidence for a wide range of study types including de novo genome assemblies, GWAS, single nucleotide polymorphism and structural variation analysis, pathogen identification, transcript assembly, metagenomic community profiling and epigenetics. SRA data are available for BLAST analysis and regular expression pattern matching. SRA provides back-end storage for sequence data deposited into the gene expression omnibus database (GEO) and the Database of Genotypes and Phenotypes (dbGaP). Further information about SRA data usage and submissions is available at www.ncbi.nlm.nih.gov/books/NBK47528/.

The Trace Archive contains >2 billion traces from gel and capillary electrophoresis sequencers. Data from >10 000 species are represented, including whole genomes of pathogens, organismal shotgun and bacterial artificial chromosome (BAC) clone projects and EST libraries. The Trace Archive was established after the conclusion of the Human Genome Sequencing Project, so only 12% of the traces are of human origin. The Trace Assembly Archive is a companion resource that contains placements of individual trace reads on a GenBank sequence. Using the Sequence Viewer, one can view multiple alignments of read placements at a given reference location. Many influenza virus genomes are presented in this way.

BioSample

The BioSample database (www.ncbi.nlm.nih.gov/biosample/) provides annotation for biological samples used in a variety of studies submitted to NCBI, including genomic sequencing, microarrays, GWAS and epigenomics (12). The primary aim of BioSamples is to address the problem of inconsistencies in annotation between similar samples from different studies so that investigators can more easily make connections between all of the available data for a particular sample. Currently BioSample contains >900 000 samples, with 90% of these coming from either SRA or dbGaP.

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. The number of amino acid residues in the RefSeq collection has grown by 24% during the past year so that Release 54 (July, 2012) contains 5.4 billion residues. 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.

Sequences from 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 38 million 'GenPept' sequences, the Protein database also contains sequences from Third Party Annotation, UniProtKB/Swiss-Prot (14), the Protein Research Foundation and the Protein Data Bank (PDB) (15).

Protein clusters

The Protein Clusters database contains >790 000 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 (16). These clusters are used as a basis for genome-wide comparison at NCBI and to provide simplified BLAST searches via Concise Microbial Protein BLAST (www.ncbi.nlm.nih.gov/genomes/prokhits.cgi). Protein Clusters provides annotations, publications, domains, structures, external links and analysis tools, including multiple sequence alignments and phylogenetic trees.

HIV-1/Human protein interaction database

The Division of Acquired Immunodeficiency Syndrome of the National Institute of Allergy and Infectious Diseases, in collaboration with the Southern Research Institute and NCBI, maintains a comprehensive HIV Protein-Interaction Database of documented interactions between HIV-1 proteins, host cell proteins, other HIV-1 proteins or proteins from disease organisms associated with HIV or AIDS (17). Summaries, including protein RefSeq accession numbers, Gene IDs, lists of interacting amino acids, brief descriptions of interactions, keywords and PubMed IDs for supporting journal articles are presented at www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/. All protein–protein interactions documented in the HIV Protein-Interaction Database are listed in Gene reports in the HIV-1 protein interactions section.

BLAST SEQUENCE ANALYSIS

BLAST

The BLAST programs (18–20) 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 to related transcript clusters (UniGene), annotated gene loci (Gene), 3D structures [Molecular Modeling Database (MMDB)] or microarray studies (GEO). The NCBI web interface for BLAST allows users to assign titles to searches, to review recent search results and to save parameter sets in MyNCBI for future use. The basic BLAST programs are also available as standalone command line programs, as network clients and as a local web-server package at 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, 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. Searches of this database generate a tabular display that partitions the BLAST hits by sequence type (genomic or transcript) and allows sorting by BLAST score, percent identity within the alignment and the percent of the query sequence contained in the alignment. 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 coding sequence (CDS) translations

Table 2. Selected NCBI software available for download

Software	Available binaries	Category within this article
BLAST (standalone)	Win, Mac, LINUX, Solaris	BLAST sequence analysis
BLAST (network client)	Win, Mac, LINUX, Solaris	BLAST sequence analysis
BLAST (web server)	Mac, LINUX, Solaris	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

from GenBank along with all RefSeq, UniProtKB/Swiss-Prot, PDB and Protein Research Foundation 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. The alignments returned can be limited by an Expectation Value threshold or range.

Genomic BLAST

NCBI maintains Genomic BLAST services for >120 organisms shown in the Map Viewer. These pages mirror the design of the standard BLAST forms and allow users access to apply the various BLAST algorithms to specialized databases for each particular genome. The default database contains the genomic sequence of an organism, but additional databases are also available such as the nucleotide and protein RefSeqs annotated on the genomic sequence as well as sets of sequences such as ESTs that are mapped to the genomic sequence. The default algorithm for the NCBI Genomic BLAST pages is MegaBLAST (21), 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 (22) 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 (23) 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, in that they do not generate valid PCR products on unintended targets. Users can also specify a forward or reverse primer in addition to a DNA template, in which case the other primer will be designed and analyzed. If both primers are specified along with a template, the tool performs only the final BLAST analysis. Users may also enter two primers without a template, in which case the BLAST analysis will display those templates in the chosen database that best match the primer pair. The available databases range from RefSeq mRNA or genomic sets for 1 of 12 model organisms to the entire BLAST nr database. An optional graphic result display allows users to view more details about the primers.

COBALT

COBALT (24) is a multiple alignment algorithm that finds a collection of pair-wise constraints derived from both the NCBI Conserved Domain database (CDD) and the sequence similarity programs RPS-BLAST, BLASTp and PHI-BLAST. These pair-wise constraints are then incorporated into a progressive multiple alignment. COBALT searches can be launched either from a BLASTp result page or from the main COBALT search page, where either FASTA sequences or accessions (or a combination thereof) may be entered into the query sequence box. 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 (25) provides an interface to curated sequences and descriptive information about genes with links to NCBI's Map Viewer, Evidence Viewer, BLink, protein domains from the CDD and other gene-related resources. Gene contains data for >10 million genes from almost 10 000 organisms. These data are accumulated and maintained through several international collaborations in addition to curation by in-house staff. Links within Gene to the newest citations in PubMed are maintained by curators and provided as Gene References into Function. 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.ncbi.nlm.nih.gov/toolbox/ncbi_tools/converters/by_program/gene2xml/.

RefSeqGene

As part of the Locus Reference Genomic collaboration (www.lrg-sequence.org), RefSeqGene provides stable, standard human genomic sequences annotated with standard mRNAs for well-characterized human genes (13). RefSeqGene records are part of the RefSeq collection and are created in consultation with authoritative locus-specific databases or other experts on particular loci and provide a stable genomic sequence for establishing numbering systems for exons and introns and for reporting and identifying genomic variants, especially those of clinical importance (26). By default, a RefSeqGene record begins 5 kb upstream of the first exon of the gene and ends 2 kb downstream of the final exon, but those

positions will be adjusted on request. A RefSeqGene sequence may differ from the current genomic build so as to reflect standard alleles. RefSeqGene records can be retrieved from Nucleotide 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

The conserved CDS database (CCDS) project is a collaborative effort among NCBI, the European Bioinformatics Institute, the Wellcome Trust Sanger Institute and University of California, Santa Cruz (UCSC) to identify a set of human and mouse protein coding regions that are consistently annotated and of high quality (27). 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. To date, the CCDS database contains >26 400 human and 23 000 mouse CDS annotations. The web interface to the CCDS allows searches by gene or sequence identifiers and provides links to Gene, record revision histories, transcript and protein sequences, as well as gene views in Map Viewer, the Ensemble Genome Browser, the UCSC Genome Browser and the Sanger Institute Vega Browser. The CCDS sequence data are available at ftp.ncbi.nlm.nih.gov/pub/CCDS/.

GEO

GEO (28) 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 elements from other categories of experiments including studies of genome copy number variation [with optional joint submission to the Database of Genomic Structural Variation (dbVar)], genome-protein interaction surveys and methylation profiling studies (jointly submitted to Epigenomics). The repository can capture fully annotated raw and processed data, enabling compliance with major community-derived scientific reporting standards such as 'Minimum Information About a Microarray Experiment' (23,24). Several data deposit options and formats are supported, including web forms, spreadsheets, XML and plain text. 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. Additionally, this year, GEO released GEO2R, a web application that enables users to perform R-based analyses of GEO data (http://www.ncbi.nlm.nih.gov/geo/geo2r/). Currently, the GEO database hosts >32 000 studies submitted by 13 000 laboratories and comprising 800 000 samples and 70 billion individual abundance measurements for >1600 organisms. The distribution of study types contained within GEO can be viewed at www.ncbi.nlm.nih.gov/geo/summary/.

UniGene and ProtEST

UniGene (29) 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 ESTs in GenBank, and currently the database includes clusters for 142 eukaryotes. UniGene databases are updated weekly with new EST sequences, and bimonthly with newly characterized sequences. As an aid to identifying a UniGene cluster, ProtEST presents precomputed BLAST alignments between protein sequences from model organisms and the six-frame translations of nucleotide sequences in UniGene.

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 (30), Mouse Genome Informatics (31), Zebrafish Information Network (32), Saccharomyces Genome Database (33) and FlyBase (34). Information about the HomoloGene build procedure is provided at www.ncbi.nlm.nih.gov/HomoloGene/HTML/homologene_buildproc.html. The HomoloGene Downloader, appearing under the 'Download' link in HomoloGene displays, retrieves transcript, protein or genomic sequences for the genes in a HomoloGene group; in the case of genomic sequence, upstream and downstream regions may be specified.

Probe

The Probe database is a public registry of nucleic acid reagents designed for use in a wide variety of biomedical research applications, together with information on reagent distributors, probe effectiveness and computed sequence similarities. The Probe database archives >14 million probe sequences, among them probes for genotyping, SNP discovery, gene expression, gene silencing and gene mapping. The probe database also provides submission templates to simplify the process of depositing data (www.ncbi.nlm.nih.gov/genome/probe/doc/Submitting.shtml).

Biosystems

The Biosystems database collects together molecules 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 (35–37), BioCyc (38), Reactome (39), the Pathway Interaction Database (40), WikiPathways (41,42) and Gene Ontology (43). These source databases provide diagrams of pathways that display the various components with their substrates and products, as well as links to relevant literature. In addition to being linked to citations in PubMed, each component within a Biosystem

record is also linked to the corresponding records in Gene and Protein, whereas the substrates and products are linked to records in PubChem (see later in the text) so that the Biosystem record centralizes NCBI data related to the pathway, greatly facilitating computation on such systems.

GENOMES

BioProject

The BioProject database (www.ncbi.nlm.nih.gov/bioproject/) is a central access point for metadata about research projects whose data are deposited in databases maintained by members of the International Nucleotide Sequence Database Consortium. BioProject also provides links to the primary data from these projects, which range from focused genome sequencing projects to large international collaborations with multiple sub-projects incorporating experiments resulting in nucleotide sequence sets, genotype/phenotype data, sequence variants or epigenetic information. BioProject also allows users to search for and retrieve data sets that are often difficult to find due to inconsistent annotation, multiple independent submissions and the varied nature of diverse data types that are often stored in different databases. The Limits page provides a convenient overview of the types of projects available and allows users to retrieve a particular type of project easily. As part of the recent redesign of the Genome database (see earlier in the text), the 'Organism Overview' project type was removed from BioProject, and equivalent data records can now be found in Genome.

The genome reference consortium

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 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 (GRCh37), mouse (GRCm38) and zebrafish (Zv9). GRCm38 was a major update for mouse released in 2012. 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). In the next major assembly release, the changes represented by the fix patches will be incorporated into the new assembly, and the fix patches themselves will be removed from the release. Novel patches will become alternate loci integrated into the new assembly. The NCBI Map Viewer provides views of the most recent GRC releases for human and mouse and for Zv9 for zebrafish, and the GRC specific data are available for download from the NCBI FTP site (ftp.ncbi.nlm.nih.gov/pub/grc/) at links provided on the GRC web pages, and assembly data are available from the GenBank genomes ftp site (ftp.ncbi.nlm.nih.gov/genbank/genomes/).

CloneDB

The Clone database (CloneDB) is a resource for finding descriptions, sources, map positions and distributor information about available clones and libraries (44). CloneDB contains information on both genomic and cell-based clones and libraries from >100 organisms. Clone records contain information about the sequences themselves as well as their genomic mapping positions and associated markers, whereas library records provide details about how the library was constructed. By default, these two record types are shown in separate tabs in CloneDB search results. The associated Library Browser allows users to filter either the genomic or cell-based library sets by organism, vector type, distributors and number of associated end or insert sequences.

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 (45). Raw data from these experiments, together with extensive metadata, are stored in the GEO and SRA databases. The Epigenomics database provides a higher-level view, allowing users to search and browse the data based on biological attributes such as cell type, tissue type, differentiation stage and health status, among many others. Data have been premapped to genomic coordinates (to make 'genome tracks'), so users are not required to be familiar with or manipulate the raw data. Tracks may be visualized in either the NCBI or UCSC genome viewers or may be downloaded to the user's computer for local analysis. Data from the Roadmap Epigenomics project, which are currently being hosted at GEO (www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/), are being mirrored and are available for viewing and downloading.

Influenza genome resources

The Influenza Genome Sequencing Project (46) provides researchers with a growing collection of >76 000 virus sequences essential to the identification of the genetic determinants of influenza pathogenicity. NCBI's Influenza Virus Resource links the Influenza Genome Sequencing Project data via PubMed to the most recent scientific literature on influenza and to a number of online analysis tools and databases. These databases include NCBI's Influenza Virus Sequence Database, comprising >215 000 influenza nucleotide sequences in the GenBank and RefSeq databases, as well as other Entrez databases containing 270 000 influenza protein sequences, 250 influenza protein structures and 870 influenza population studies. An online influenza genome annotation tool analyzes a novel sequence and produces output in a 'feature table' format that can be used by NCBI's GenBank submission tools such as tbl2asn (1).

NCBI now also provides the Virus Variation resource (www.ncbi.nlm.nih.gov/ genomes/VirusVariation/) that extends services available for Influenza to the dengue

and West Nile viruses. Virus Variation provides a portal for retrieving, downloading, analyzing and annotating virus sequences using pages customized to unique aspects of viral sequence data, including genotype, severity of the resulting disease and the year a sample was collected.

GENETICS AND MEDICINE

dbGaP

dbGaP (47) 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 (grants.nih.gov/grants/gwas/). The dbGaP collection contains >340 studies, each of which can be browsed by name or disease. To protect the confidentiality of study subjects, dbGaP accepts only de-identified data and requires investigators to go through an authorization process to access individual-level data. Study documents, protocols and subject questionnaires are available without restriction. Authorized access data distributed to primary investigators for use in approved research projects include de-identified phenotypes and genotypes for individual study subjects, pedigrees and some pre-computed associations between genotype and phenotype.

dbVar

dbVar is an archive of large-scale genomic variants (generally >50 bp) such as insertions, deletions, translocations and inversions (48). The number of dbVar studies increased by 50% during the past year to now >90 studies containing data from 11 eukaryotes. These data are derived from several methods including computational sequence analysis and microarray experiments. Each of the >2.7 million variants is linked to a graphical view showing its genomic context. Symbols for genes within variant regions are now displayed on search results, and users can also search for such genes directly in dbVar.

Database of short genetic variations

The Database of Short Genetic Variations (dbSNP) (49) is a repository of all types of short genetic variations <50 bp, and so it is a complement to dbVar. dbSNP accepts submissions of 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, indicates if a variation call may be suspect because of paralogy (50) and maintains links to related information in other NCBI databases. Moreover, dbSNP data are provided as annotation tracks on the NCBI 1000 Genomes Browser (www.ncbi.nlm.nih.gov/variation/tools/1000genomes/) that allows users to browse and download 1000 Genomes genotype and supporting sequence reads. dbSNP has two web-based portals for maintaining and analyzing human variations: the 'Human Variation: Search, Annotate, Submit' site (http://www.ncbi.nlm.nih.gov/projects/SNP/tranSNP/tranSNP.cgi) and the 'Human Variation: Annotation and Submit Batch Data with Clinical Impact' site (http://www.ncbi.nlm.nih.gov/projects/SNP/tranSNP/VarBatchSub.cgi). These portals accept single and batch submissions, respectively, and they both validate Human Genome Variation Society (HGVS) expressions and facilitate the submission of clinical significance data. Another tool, the Variation Reporter (http://www.ncbi.nlm.nih.gov/variation/tools/reporter), allows investigators to upload variant data in popular formats, such as browser extensible display (BED), genome variation format (GVF) and HGVS, and receive a report of matching records in dbSNP along with a report of functional consequences of each variant not already in dbSNP. An API for this tool is available at http://www.ncbi.nlm.nih.gov/variation/tools/reporter/docs/api/perl. Submissions of interpreted clinical significance to dbSNP are reported in collaboration with ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/), and include a file of common variants with no reported clinical significance (common_no_known_medical_impact.vcf.gz) developed specifically for those users wishing to narrow their list of variations to those that might warrant further evaluation for a novel disorder. These reports are provided in Variant Call Format (VCF) files (Table 3) and are documented on NCBI's variation site: http://www.ncbi.nlm.nih.gov/variation/docs/human_variation_vcf/.

Online mendelian inheritance in animals

Online Mendelian Inheritance in Animals is a database of genes, inherited disorders and traits in animal species other than human and mouse, and is authored by Professor Frank Nicholas and colleagues (51) of the University of Sydney, Australia. The database holds 2800 records containing textual information and references, as well as links to relevant records from Online Mendelian Inheritance in Man, PubMed and Gene.

Database cluster for routine clinical applications: dbMHC, dbLRC, dbRBC

dbMHC focuses on the Major Histocompatibility Complex (MHC) and contains sequences and frequency distributions for alleles of the MHC, an array of genes that play a central role in the success of organ transplants and an individual's susceptibility to infectious diseases. dbMHC also contains HLA 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 killer cell immunoglobulin-type receptor (KIR) genes. dbRBC represents data on genes and their sequences for red blood cell antigens or blood groups. It hosts the Blood Group Antigen Gene Mutation Database (52) and integrates it with resources at NCBI. dbRBC provides general information on individual genes and access to the International Society of Blood Transfusion allele nomenclature of blood group alleles. All three

Table 3. Summary of dbSNP FTP human VCF files

File name	Update frequency	dbSNP RefSNP count (based on build 137)
clinvar.vcf.gz	Weekly	36 K
A list of all human variations submitted through clinical channels that contain a mixture of variations asserted to be pathogenic and those known to be non-pathogenic		
00-All.vcf.gz	Once per dbSNP build	52 M
A comprehensive list of all short human variations based on the most recent dbSNP build		
common_all.vcf.gz	Once per dbSNP build	28 M
A subset of variations from 00-All.vcf.gz that are determined to be 'common' based on germline origin and a minor allele frequency of ≥0.01 in at least one major population, with at least two individuals from different families having the minor allele		
common_no_known_medical_impact.vcf.gz	Weekly	28 M
A list of all 'common' germline human variations that fall within the scope of VCF processing. To create this list, variation records of probable medical interest from clinvar.vcf.gz are removed from the list of common_all.vcf.gz		

databases dbMHC, dbLRC and dbRBC provide multiple sequence alignments, analysis tools to interpret homozygous or heterozygous sequencing results (53) and tools for DNA probe alignments.

CHEMICALS AND BIOASSAYS

PubChem (54,55) is the informatics backbone for the NIH Roadmap Initiative on molecular libraries and 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. The databases include records for 100 million substances containing 35 million unique chemical structures, and 2.3 million of these substances have bioactivity data in at least one of the 620 000 PubChem BioAssays. PubChem also provides a diverse set of three-dimensional (3D) conformers for 84% of the records in the PubChem Compound database. A viewing application, PC3D, is available to view both individual conformers and overlays of similar conformers. The PubChem databases link not only to other Entrez databases such as PubMed and PubMed Central but also to Structure and Protein to provide a bridge between the macromolecules of genomics and the small organic molecules of cellular metabolism. The PubChem databases are searchable using text queries as well as structural queries based on chemical SMILES, formulas or chemical structures provided in a variety of formats. The PubChem Sketcher, an online structure-drawing tool provides a simple way to construct a structure-based search (pubchem.ncbi.nlm.nih.gov/search/search.cgi).

DOMAINS AND STRUCTURES

MMDB

The NCBI MMDB (56) contains experimentally determined coordinate sets from the Protein Data Bank (15), augmented with domain annotations and links to relevant literature, protein and nucleotide sequences, chemicals (PDB heterogens) and conserved domains in CDD (57). The structure summary pages display these links along with thumbnail images of the biological and asymmetric units from the source PDB files. These images link to interactive views of the data in Cn3D (58), the NCBI structure and alignment viewer. Compact structural domains within protein structures are annotated on protein chains, and these graphic annotations link to structural neighbors computed by the VAST algorithm (59,60). Users can access MMDB structures either through direct text searches or through the 'Related Structures' link provided for all protein records.

CDD and Conserved Domain Architecture Retrieval Tool

The CDD (61) contains >46 000 PSI-BLAST-derived Position Specific Score Matrices representing domains taken from the Simple Modular Architecture Research Tool (62), Pfam (63), TIGRFAM (64) and from domain alignments derived from COGs and Protein Clusters. In addition, CDD includes 3300 superfamily records, each of which contains a set of CDs from one or more source databases that generate overlapping annotation on the same protein sequences. The NCBI Conserved Domain Search (CD-Search) service locates conserved domains within a protein sequence, and these results are available for all records in the Protein database through the 'Identify Conserved Domains' link in the upper right of a sequence record. Wherever possible, protein sequences with known 3D structures are included in CD alignments, which can be viewed along with these structures and also edited within Cn3D. CD alignments can be viewed online, edited or created *de novo* using CDTree (Table 2). CDTree uses PSI-BLAST to add new sequences to an existing CD alignment and provides an interface for exploring phylogenetic trends in domain architecture and for building hierarchies of alignment-based protein domains. The Conserved Domain Architecture Retrieval Tool searches protein databases with a query sequence and returns the domain architectures of database proteins containing the query domain.

FOR FURTHER INFORMATION

The resources described here include documentation, other explanatory material and references to collaborators and data sources on the respective web sites. An alphabetical list of NCBI resources is available from a link in the upper left of the NCBI home page. The NCBI Help Manual and the NCBI Handbook, both available as links in the common page footer, describe the principal NCBI resources in detail. The NCBI Education page (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 pages on Facebook, Twitter and YouTube. Several new training videos produced in the past year have been added to YouTube. A user-support staff is available to answer questions at info@ncbi.nlm.nih.gov. Updates on NCBI resources and database enhancements are described in the NCBI News newsletter (www.ncbi.nlm.nih.gov/books/NBK1969/). In addition, NCBI offers several mailing lists that provide updates on services and databases (www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html), as well as RSS feeds (www.ncbi.nlm.nih.gov/feed/).

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., Karsch-Mizrachi,I., Clark,K., Lipman,D.J., Ostell,J. and Sayers,E.W. (2012) GenBank. *Nucleic Acids Res.*, **40**, D48–D53.
2. Rubinstein,W.S., Maglott,D.R., Lee,J., Kattman,B.L., Malheiro,A.J., Fomous,C. and Ostell,J.M. (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. Pagon,R.A. (2006) GeneTests: an online genetic information resource for health care providers. *J. Med. Libr. Assoc.*, **94**, 343–348.
4. Waggoner,D.J. and Pagon,R.A. (2009) Internet resources in medical genetics. *Curr. Protoc. Hum. Genet.*, **Chapter 9**, Unit 9 12.
5. Boratyn,G.M., Schaffer,A.A., Agarwala,R., Altschul,S.F., Lipman,D.J. and Madden,T.L. (2012) Domain enhanced lookup time accelerated BLAST. *Biol. Direct*, **7**, 12.
6. 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.
7. Sewell,W. (1964) Medical subject headings in Medlars. *Bull. Med. Libr. Assoc.*, **52**, 164–170.
8. Sequeira,E. (2003) PubMed Central - three years old and growing stronger. *ARL*, **228**, 5–9.
9. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
10. 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.
11. Kodama,Y., Shumway,M. and Leinonen,R. (2012) The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
12. Barrett,T., Clark,K., Gevorgyan,R., Gorelenkov,V., Gribov,E., Karsch-Mizrachi,I., Kimelman,M., Pruitt,K.D., Resenchuk,S., Tatusova,T. *et al.* (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.*, **40**, D57–D63.
13. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI reference sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, **37**, D32–D36.
14. Magrane,M. and Consortium,U. (2011) UniProt knowledgebase: a hub of integrated protein data. *Database (Oxford)*, **2011**, bar009.
15. 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.
16. 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.
17. 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.
18. 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.
19. 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.
20. Ye,J., McGinnis,S. and Madden,T.L. (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res.*, **34**, W6–W9.
21. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
22. Ma,B., Tromp,J. and Li,M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.
23. 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.
24. Papadopoulos,J.S. and Agarwala,R. (2007) COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics*, **23**, 1073–1079.
25. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.
26. 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.
27. Pruitt,K.D., Harrow,J., Harte,R.A., Wallin,C., Diekhans,M., Maglott,D.R., Searle,S., Farrell,C.M., Loveland,J.E., Ruef,B.J. *et al.* (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.
28. 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.
29. Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698.
30. 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.
31. 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.
32. 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.
33. 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.
34. 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.
35. 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.
36. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*., **28**, 27–30.
37. 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.
38. 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.
39. 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.
40. 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.
41. 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.
42. 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.
43. 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.
44. Schneider,V., Chen,H.C., Clausen,C., Meric,P., Zhou,G., Husain,N., Maglott,D. and Church,D.M. (2013) Clone DB: an integrated NCBI resource for clone-associated data. *Nucleic Acids Res*., **41**, D1070–D1078.
45. 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.
46. 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.
47. 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.
48. 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.
49. 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.
50. 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.
51. Lenffer,J., Nicholas,F.W., Castle,K., Rao,A., Gregory,S., Poidinger,M., Mailman,M.D. and Ranganathan,S. (2006) OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI. *Nucleic Acids Res*., **34**, D599–D601.
52. 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.
53. 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.
54. 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.
55. 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.
56. 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.
57. 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.
58. 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.
59. Gibrat,J.F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol*., **6**, 377–385.
60. Madej,T., Gibrat,J.F. and Bryant,S.H. (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.
61. 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.
62. 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.
63. 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.
64. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res*., **31**, 371–373.

APPENDIX

NCBI Resource Coordinators: Abigail Acland, Richa Agarwala, Tanya Barrett, Jeff Beck, Dennis A. Benson, Colleen Bollin, Evan Bolton, Stephen H. Bryant, Kathi Canese, Deanna M. Church, Karen Clark, Michael DiCuccio, Ilya Dondoshansky, Scott Federhen, Michael Feolo, Lewis Y. Geer, Viatcheslav Gorelenkov, Marilu Hoeppner, Mark Johnson, Christopher Kelly, Viatcheslav Khotomlianski, Avi Kimchi, Michael Kimelman, Paul Kitts, Sergey Krasnov, Anatoliy Kuznetsov, David Landsman, David J. Lipman, Zhiyong Lu, Thomas L. Madden, Tom Madej, Donna R. Maglott, Aron Marchler-Bauer, Ilene Karsch-Mizrachi, Terence Murphy, James Ostell, Christopher O'Sullivan, Anna Panchenko, Lon Phan, Don Preussm Kim D. Pruitt, Wendy Rubinstein, Eric W. Sayers, Valerie Schneider, Gregory D. Schuler, Edwin Sequeira, Stephen T. Sherry, Martin Shumway, Karl Sirotkin, Karanjit Siyan, Douglas Slotta, Alexandra Soboleva, Grigory Starchenko, Tatiana A. Tatusova, Bart Trawick, Denis Vakatov, Yanli Wang, Minghong Ward, W. John Wilbur, Eugene Yaschenko, Kerry Zbicz.